# *IBM SPSS Modeler 18.1.1 Nodos de modelado*

**Nota**

Antes de utilizar esta información y el producto al que da soporte, lea la información del apartado “Avisos” en la página 395.

**Información del producto**

Esta edición se aplica a la versión 18, release 1, modificación 1 de IBM SPSS Modeler y a todos los releases y las modificaciones posteriores, hasta que se indique lo contrario en nuevas ediciones.

# Contenido

# Prefacio

IBM® SPSS Modeler es conjunto de programas de minería de datos de IBM Corp. orientado a las empresas. SPSS Modeler ayuda a las organizaciones a mejorar la relación con sus clientes y los ciudadanos a través de la comprensión profunda de los datos. Las organizaciones utilizan la comprensión que les ofrece SPSS Modeler para retener a los clientes más rentables, identificar las oportunidades de venta cruzada, atraer a nuevos clientes, detectar el fraude, reducir el riesgo y mejorar la prestación de servicios del gobierno.

La interfaz visual de SPSS Modeler invita a los usuarios a que apliquen su experiencia empresarial específica, lo que deriva en modelos predictivos más potentes y reduce el tiempo necesario para encontrar soluciones. SPSS Modeler ofrece muchas técnicas de modelado tales como predicciones, clasificaciones, segmentación y algoritmos de detección de asociaciones. Una vez que se crean los modelos, IBM SPSS Modeler Solution Publisher permite su distribución en toda la empresa a los encargados de tomar las decisiones o a una base de datos.

## Acerca de IBM Business Analytics

IBM Business Analytics proporciona información completa, coherente y precisa en la que confían para mejorar el rendimiento de su negocio quienes toman las decisiones. Un conjunto integral de inteligencia empresarial, análisis predictivo, rendimiento financiero y gestión de estrategias y aplicaciones de análisis que ofrece una perspectiva clara, inmediata e interactiva del rendimiento actual y la capacidad de predecir resultados futuros. En combinación con extensas soluciones sectoriales, prácticas probadas y servicios profesionales, las organizaciones de cualquier tamaño pueden conseguir el máximo de productividad, automatizar las decisiones de forma fiable y alcanzar mejores resultados.

Como parte de estos documentos, IBM SPSS Predictive Analytics ayuda a las organizaciones a predecir situaciones futuras y a actuar de forma proactiva con esa información para mejorar sus resultados. Clientes comerciales, gubernamentales y académicos de todo el mundo confían en la tecnología IBM SPSS como mejora competitiva para atraer, conservar y aumentar la clientela reduciendo el fraude y los riesgos. Al incorporar el software de IBM SPSS en sus operaciones diarias, las organizaciones se convierten en empresas predictivas, capaces de dirigir y automatizar decisiones para alcanzar los objetivos comerciales y lograr una ventaja considerable sobre la competencia. Para obtener más información o contactar con un representante, visite http://www.ibm.com/spss.

## Asistencia técnica

El servicio de asistencia técnica está a disposición de todos los clientes de mantenimiento. Los clientes podrán ponerse en contacto con el servicio de asistencia técnica si desean recibir ayuda sobre la utilización de los productos de IBM Corp. o sobre la instalación en los entornos de hardware admitidos. Para ponerse en contacto con el servicio de asistencia técnica, consulte el sitio web de IBM Corp. en http://www.ibm.com/support. Tenga a mano su acuerdo de asistencia y esté preparado para identificarse a sí mismo y a su organización al solicitar ayuda.

# Capítulo 1. Acerca de IBM SPSS Modeler

IBM SPSS Modeler es un conjunto de herramientas de minería de datos que permite desarrollar rápidamente modelos predictivos mediante técnicas empresariales y desplegarlos en operaciones empresariales para mejorar la toma de decisiones. Con un diseño que sigue el modelo CRISP-DM, estándar del sector, IBM SPSS Modeler admite el proceso completo de minería de datos, desde los propios datos hasta obtener los mejores resultados empresariales.

IBM SPSS Modeler ofrece una gran variedad de métodos de modelado procedentes del aprendizaje automático, la inteligencia artificial y el estadístico. Los métodos disponibles en la paleta de modelado permiten derivar nueva información procedente de los datos y desarrollar modelos predictivos. Cada método tiene ciertos puntos fuertes y es más adecuado para determinados tipos de problemas.

SPSS Modeler puede adquirirse como producto independiente o utilizarse como cliente junto con SPSS Modeler Server. También hay disponible cierto número de opciones adicionales que se resumen en las siguientes secciones. Si desea obtener más información, consulte https://www.ibm.com/analytics/us/en/technology/spss/.

## Productos IBM SPSS Modeler

La familia de productos IBM SPSS Modeler y su software asociado se componen de lo siguiente:

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- Consola de administración de IBM SPSS Modeler (incluido con el Gestor de despliegue de IBM SPSS)
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- Adaptadores de IBM SPSS Modeler Server para IBM SPSS Collaboration and Deployment Services

## IBM SPSS Modeler

SPSS Modeler es una versión con todas las funcionalidades del producto que puede instalar y ejecutar en su ordenador personal. Puede ejecutar SPSS Modeler en modo local como un producto independiente o utilizarla en modo distribuido junto con IBM SPSS Modeler Server para mejorar el rendimiento a la hora de trabajar con grandes conjuntos de datos.

Con SPSS Modeler, puede crear modelos predictivos precisos de forma rápida e intuitiva sin necesidad de programación. Mediante su exclusiva interfaz visual, podrá visualizar fácilmente el proceso de minería de datos. Con ayuda del análisis avanzado incrustado en el producto podrá detectar patrones y tendencias en sus datos que anteriormente estaban ocultos. Podrá modelar los resultados y comprender los factores que influyen en ellos, lo que le permitirá aprovechar oportunidades comerciales y mitigar los riesgos.

SPSS Modeler está disponible en dos ediciones: SPSS Modeler Professional y SPSS Modeler Premium. Consulte el tema “Ediciones de IBM SPSS Modeler” en la página 2 para obtener más información.

## IBM SPSS Modeler Server

SPSS Modeler utiliza una arquitectura de cliente/servidor para distribuir peticiones de cliente para operaciones que requieren un uso intensivo de los recursos a un software de servidor de gran potencia, lo que proporciona un rendimiento más rápido con conjuntos de datos de mayor volumen.

SPSS Modeler Server es un producto con licencia independiente que se ejecuta de manera continua en modo de análisis distribuido en un host de servidor junto con una o más instalaciones de IBM SPSS

Modeler. De esta manera, SPSS Modeler Server ofrece un rendimiento superior cuando se trabaja con grandes conjuntos de datos, ya que las operaciones que requieren un uso intensivo de la memoria se pueden realizar en el servidor sin tener que descargar datos al equipo cliente. IBM SPSS Modeler Server también ofrece compatibilidad con la optimización SQL y la posibilidad de realizar el modelado interno de bases de datos, lo que ofrece ventajas adicionales de rendimiento y automatización.

## Consola de administración de IBM SPSS Modeler

El Consola de administración de Modeler región propietaria del archivos una interfaz gráfica de usuario para gestionar muchas de las opciones de configuración de SPSS Modeler Server, que también se pueden configurar a través de un archivo de opciones. La consola se incluye en el Gestor de despliegue de IBM SPSS, se puede utilizar para supervisar y configurar las instalaciones de SPSS Modeler Server y está disponible de forma gratuita para los clientes actuales de SPSS Modeler Server. La aplicación solamente se puede instalar en los ordenadores con Windows; sin embargo, puede administrar un servidor que esté instalado en cualquier plataforma compatible.

## IBM SPSS Modeler Batch

Aunque la minería de datos suele ser un proceso interactivo, también es posible ejecutar SPSS Modeler desde una línea de comandos, sin necesidad de la interfaz gráfica del usuario. Por ejemplo, puede que tenga tareas repetitivas o cuya ejecución sea de larga duración que quiera realizar sin intervención del usuario. SPSS Modeler Batch es una versión especial del producto que proporciona soporte para todas las prestaciones de análisis de SPSS Modeler sin acceso a la interfaz de usuario habitual. SPSS Modeler Server debe utilizar SPSS Modeler Batch.

## IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher es una herramienta que le permite crear una versión empaquetada de una ruta de SPSS Modeler que se puede ejecutar en un motor de tiempo de ejecución externo o incrustado en una aplicación externa. De este modo, podrá publicar y desplegar rutas completas de SPSS Modeler para utilizarlas en entornos que no tengan SPSS Modeler instalado. SPSS Modeler Solution Publisher se distribuye como parte del servicio IBM SPSS Collaboration and Deployment Services - Puntuación, para el que se necesita una licencia independiente. Con esta licencia, recibirá SPSS Modeler Solution Publisher Runtime, que le permite ejecutar las rutas publicadas.

Si desea más información sobre SPSS Modeler Solution Publisher, consulte la documentación de IBM SPSS Collaboration and Deployment Services. El Knowledge Center de IBM SPSS Collaboration and Deployment Services contiene secciones denominadas "IBM SPSS Modeler Solution Publisher" e "IBM SPSS Analytics Toolkit."

## Adaptadores de IBM SPSS Modeler Server para IBM SPSS Collaboration and Deployment Services

Tiene a su disposición un determinado número de adaptadores para IBM SPSS Collaboration and Deployment Services que permiten que SPSS Modeler y SPSS Modeler Server interactúen con un repositorio de IBM SPSS Collaboration and Deployment Services. De este modo, varios usuarios podrán compartir una ruta de SPSS Modeler desplegada en el repositorio, o bien se podrá acceder a ella desde la aplicación cliente de baja intensidad IBM SPSS Modeler Advantage. Debe instalar el adaptador en el sistema donde se aloje el repositorio.

---

# Ediciones de IBM SPSS Modeler

SPSS Modeler está disponible en las siguientes ediciones.

### SPSS Modeler Professional

SPSS Modeler Professional proporciona todas las herramientas que necesita para trabajar con la mayoría de los tipos de datos estructurados, como los comportamientos e interacciones registrados en los sistemas

de CRM, datos demográficos, comportamientos de compra y datos de ventas.

### SPSS Modeler Premium

SPSS Modeler Premium es un producto con licencia independiente que amplía SPSS Modeler Professional para trabajar con datos especializados y con datos de texto no estructurado. SPSS Modeler Premium incluye IBM SPSS Modeler Text Analytics:

**IBM SPSS Modeler Text Analytics** utiliza tecnologías de lingüística avanzada y Procesamiento del lenguaje natural (PLN) para procesar con rapidez una gran variedad de datos de texto sin estructurar, extraer y organizar los conceptos clave y agruparlos en categorías. Las categorías y conceptos extraídos se pueden combinar con los datos estructurados existentes, como pueden ser datos demográficos, y se pueden aplicar para modelar utilizando el conjunto completo de herramientas de minería de datos de IBM SPSS Modeler para tomar decisiones mejores y más certeras.

### IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription proporciona todas las mismas prestaciones de análisis predictivo que el cliente tradicional de IBM SPSS Modeler. Con la edición Subscription, puede descargar actualizaciones de producto con regularidad.

## Documentación

La documentación está disponible desde el menú Ayuda en SPSS Modeler. Así se abre el Knowledge Center, que está disponible públicamente fuera del producto.

También está disponible documentación completa para cada producto (incluyendo instrucciones de instalación) en formato PDF, en una carpeta comprimida separada, como parte de la descarga del producto. O bien, los documentos PDF se pueden descargar de la web en http://www.ibm.com/support/docview.wss?uid=swg27046871.

## SPSS Modeler ProfessionalDocumentación

El conjunto de documentación de SPSS Modeler Professional (excluidas las instrucciones de instalación) es el siguiente.

- **Guía del usuario de IBM SPSS Modeler.** Introducción general para utilizar SPSS Modeler, incluyendo cómo crear corrientes de datos, manejar valores que faltan, crear expresiones de CLEM, trabajar con proyectos e informes y empaquetar corrientes para el despliegue en IBM SPSS Collaboration and Deployment Services o IBM SPSS Modeler Advantage.
- **Nodos de origen, proceso y resultado de IBM SPSS Modeler.** Descripciones de todos los nodos utilizados para leer, procesar y dar salida a datos en diferentes formatos. En la práctica, esto implica todos los nodos que no sean nodos de modelado.
- **Nodos de modelado de IBM SPSS Modeler.** Descripciones de todos los nodos utilizados para crear modelos de minería de datos. IBM SPSS Modeler ofrece una gran variedad de métodos de modelado procedentes del aprendizaje automático, la inteligencia artificial y el estadístico.
- **Guía de aplicaciones de IBM SPSS Modeler.** Los ejemplos de esta guía ofrecen introducciones breves y concisas a métodos y técnicas de modelado específicos. También está disponible una versión en línea de esta guía desde el menú Ayuda. Consulte el tema “Ejemplos de aplicación” en la página 4 para obtener más información.
- **Scripts y automatización Python de IBM SPSS Modeler.** Información sobre la automatización del sistema mediante scripts de Python, incluidas las propiedades que se pueden utilizar para manipular nodos y rutas.
- **Guía de despliegue de IBM SPSS Modeler.** La información sobre cómo ejecutar rutas de IBM SPSS Modeler como pasos en el proceso de trabajos en el Gestor de despliegue de IBM SPSS.

- **Guía del desarrollador de IBM SPSS Modeler CLEF.** CLEF permite integrar programas de terceros, tales como rutinas de proceso de datos o algoritmos de modelado, como nodos en IBM SPSS Modeler.
- **Guía de minería interna de base de datos de IBM SPSS Modeler** Este manual incluye información sobre cómo utilizar la potencia de su base de datos, tanto para mejorar su rendimiento como para ampliar su oferta de capacidades analíticas a través de algoritmos de terceros.
- **Guía de administración y rendimiento de IBM SPSS Modeler Server.** Información sobre la configuración y administración de IBM SPSS Modeler Server.
- **Guía del usuario del Gestor de despliegue de IBM SPSS.** Información sobre cómo utilizar la interfaz de usuario de la consola de administración incluida en la aplicación Gestor de despliegue para supervisar y configurar IBM SPSS Modeler Server.
- **Guía de CRISP-DM de IBM SPSS Modeler.** Manual que explica paso a paso cómo utilizar la metodología de CRISP-DM en la minería de datos con SPSS Modeler.
- **Guía de usuario de IBM SPSS Modeler Batch.** Guía completa de cómo utilizar IBM SPSS Modeler en modo por lotes, incluida información detallada sobre la ejecución del modo por lotes y argumentos de línea de comandos. Esta guía está disponible únicamente en formato PDF.

## SPSS Modeler PremiumDocumentación

El conjunto de documentación de SPSS Modeler Premium (excluidas las instrucciones de instalación) es el siguiente.

- **Guía del usuario de SPSS Modeler Text Analytics .** Información sobre cómo utilizar el análisis de texto con SPSS Modeler, que cubre los nodos de minería de texto, programa interactivo, plantillas y otros recursos.

# Ejemplos de aplicación

Mientras que las herramientas de minería de datos de SPSS Modeler pueden ayudar a resolver una amplia variedad de problemas organizativos y empresariales, los ejemplos de la aplicación ofrecen introducciones breves y adaptadas de técnicas y métodos de modelado específicos. Los conjuntos de datos utilizados aquí son mucho más pequeños que los enormes almacenes de datos gestionados por otros analistas de datos, pero los conceptos y métodos implicados se pueden escalar a aplicaciones reales.

Para acceder a los ejemplos, pulse **Ejemplos de aplicación** en el menú Ayuda en SPSS Modeler.

Los archivos de datos y rutas de ejemplo se instalan en la carpeta `Demos` en el directorio de instalación del producto. Si desea obtener más información, consulte “Carpeta Demos”.

**Ejemplos de modelado de bases de datos.** Consulte los ejemplos que figuran en la *Guía de minería interna de bases de datos de IBM SPSS Modeler*.

**Ejemplos de scripts.** Consulte los ejemplos que figuran en la *Guía de scripts y automatización de IBM SPSS Modeler*.

# Carpeta Demos

Los archivos de datos y las corrientes de muestras que se utilizan con los ejemplos de aplicación se instalan en la carpeta `Demos` bajo el directorio de instalación del producto (por ejemplo: `C:\Archivos de programa\IBM\SPSS\Modeler\<versión>\Demos`). También se puede acceder a esta carpeta desde el grupo de programas de IBM SPSS Modeler en el menú Inicio de Windows, o pulsando `Demos` en la lista de directorios recientes en el recuadro de diálogo **Archivo** > **Abrir ruta**.

## Rastreo de licencias

Cuando se utiliza SPSS Modeler, el uso de las licencias se rastrea y registra a intervalos regulares. Las métricas de licencia que se registran son *AUTHORIZED_USER* y *CONCURRENT_USER*, y el tipo de métrica que se registra depende del tipo de licencia que tiene para SPSS Modeler.

IBM License Metric Tool puede procesar los archivos de registro que se generan, a partir de los cuales puede crear informes de uso de licencia.

Los archivos de registro de licencia se crean en el mismo directorio donde se registran los archivos de registro del cliente de SPSS Modeler (de forma predeterminada, `%ALLUSERSPROFILE%/IBM/SPSS/Modeler/<versión>/log`).

# Capítulo 2. Introducción al modelado

Un modelo es un conjunto de reglas, fórmulas o ecuaciones que puede utilizarse para predecir un resultado basándose en un conjunto de campos o variables de entrada. Por ejemplo, puede que una institución financiera utilice un modelo para predecir la probabilidad de que los solicitantes de un préstamo sean un riesgo bueno o malo, basándose en información que ya se conoce sobre solicitantes anteriores.

La capacidad de predecir un resultado es el objetivo central del análisis predictivo y la comprensión del proceso de modelado es la clave para utilizar IBM SPSS Modeler.

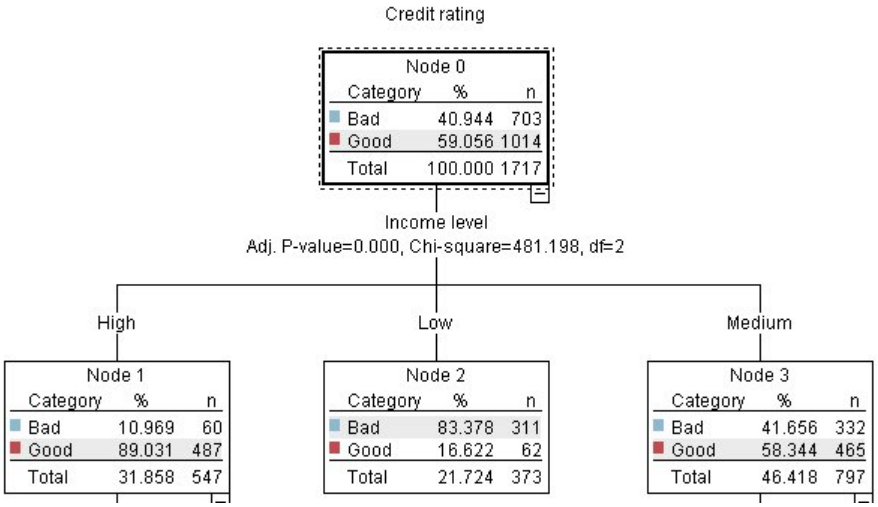



*Figura 1. Modelo de árbol de decisión sencillo*

Este ejemplo utiliza un modelo de **árbol de decisión** que clasifica los registros (y predice una respuesta) utilizando una serie de reglas de decisión, por ejemplo:

```
IF ingreso = Medio
AND tarjetas <5
THEN -> 'Bueno'
```

Aunque este ejemplo utiliza un modelo CHAID (Detección automática de interacciones mediante chi-cuadrado), se presenta como una introducción general y la mayoría de los conceptos se aplica de forma amplia en otros tipos de modelado de IBM SPSS Modeler.

Para comprender cualquier modelo, primero debe comprender los datos que incluye. Los datos de este ejemplo contienen información sobre los clientes de un banco. Se utilizan los siguientes campos:

| **Nombre de campo** | **Descripción** |
|---|---|
| Valoración_crédito | Valoración de crédito: 0=Malo, 1=Bueno, 9=valores perdidos |
| Edad | Edad en años |
| Ingresos | Nivel de ingresos: 1=Bajo, 2=Medio, 3=Alto |
| Tarjetas_crédito | Número de tarjetas de crédito en propiedad: 1=Menos de cinco, 2=Cinco o más |
| Educación | Nivel educativo: 1=Instituto, 2=Universidad |

| Nombre de campo | Descripción |
|---|---|
| Préstamo_coche | Número de préstamos de coche asumidos: 1= Ninguno o uno, 2=Más de dos |

El banco mantiene una base de datos con información histórica sobre los clientes a los que el banco ha concedido préstamos, incluido si los han reintegrado o no (Valoración de crédito = Bueno) o causado mora en el pago de dichos préstamos (Valoración de crédito = Malo). Con los datos existentes, el banco quiere generar un modelo que le permita predecir la probabilidad de mora del préstamo de los posibles solicitantes futuros de un préstamo.

Al utilizar un modelo de árbol de decisión, puede analizar las características de los dos grupos de clientes y predecir la probabilidad de mora del préstamo.

Este ejemplo utiliza la ruta denominada *modelingintro.str*, disponible en la carpeta *Demos* bajo la subcarpeta *streams*. El archivo de datos es *tree_credit.sav*. Consulte el tema "Carpeta Demos" en la página 4 para obtener más información.

Veamos la ruta más detenidamente.

1. Seleccione lo siguiente en el menú principal:
   **Archivo** > **Abrir ruta**
2. Pulse en el icono de nugget dorado de la barra de herramientas del cuadro de diálogo Abrir y seleccione la carpeta Demos.
3. Pulse dos veces en la carpeta *streams*.
4. Pulse dos veces en el archivo llamado *modelingintro.str*.

## Generación de la ruta

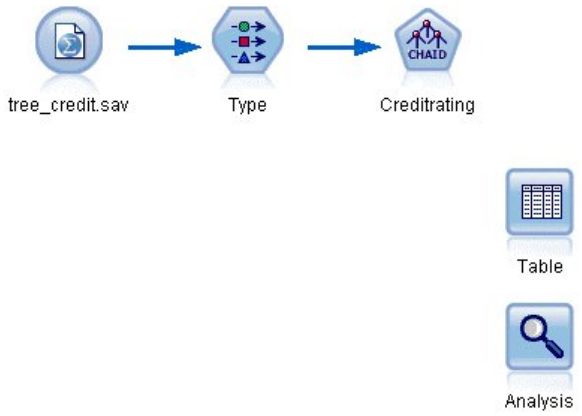



*Figura 2. Modelado de la ruta*

Para crear una ruta que cree un modelo, necesitamos al menos tres elementos:

- Un nodo de origen que lea los datos de un origen externo, en este caso, un archivo de datos IBM SPSS Statistics.
- Un nodo de origen o nodo Tipo que especifique propiedades de campo, como el nivel de medición (el tipo de datos que contiene el campo) y el rol de cada campo como objetivo o entrada en modelado.
- Un nodo de modelado que genera un nugget de modelo cuando se ejecuta la ruta.

En este ejemplo estamos usando un nodo de modelado CHAID. CHAID, o Detección automática de interacciones mediante chi-cuadrado, es un método de clasificación que genera árboles de decisión utilizando un tipo específico de estadísticos denominados estadísticos chi-cuadrado para determinar los mejores lugares para realizar las divisiones en el árbol de decisión.

Si se especifican niveles de medición en el nodo de origen, se puede eliminar el nodo Tipo independiente. Funcionalmente, el resultado es el mismo.

Esta ruta también tiene los nodos Tabla y Análisis que se utilizarán para ver los resultados de puntuación después de crear el nugget de modelo y añadirlo a la ruta.

El nodo de origen Archivo Statistics lee los datos en formato IBM SPSS Statistics del archivo de datos *tree_credit.sav*, que está instalado en la carpeta *Demos*. (Una variable especial denominada *$CLEO_DEMOS* se utiliza para hacer referencia a esta carpeta en la instalación actual de IBM SPSS Modeler. Esto garantiza que la ruta será válida independientemente de la carpeta o versión de la instalación actual.)


*Figura 3. Lectura de datos con un nodo de origen Archivo Statistics*

El nodo Tipo especifica el **nivel de medición** de cada campo. El nivel de medición es una categoría que indica el tipo de datos del campo. Nuestro archivo de datos de origen utiliza tres niveles de medición diferentes.

Un campo **Continuo** (como el campo *Edad*) contiene valores numéricos continuos, mientras que un campo **Nominal** (como el campo *Valoración de crédito*) tiene dos o más valores distintos, por ejemplo, *Malo*, *Bueno* o *Sin historial de crédito*. Un campo **Ordinal** (como el campo *Nivel de ingresos*) describe datos con varios valores distintos que tienen un orden inherente, en este caso *Bajo*, *Medio* y *Alto*.

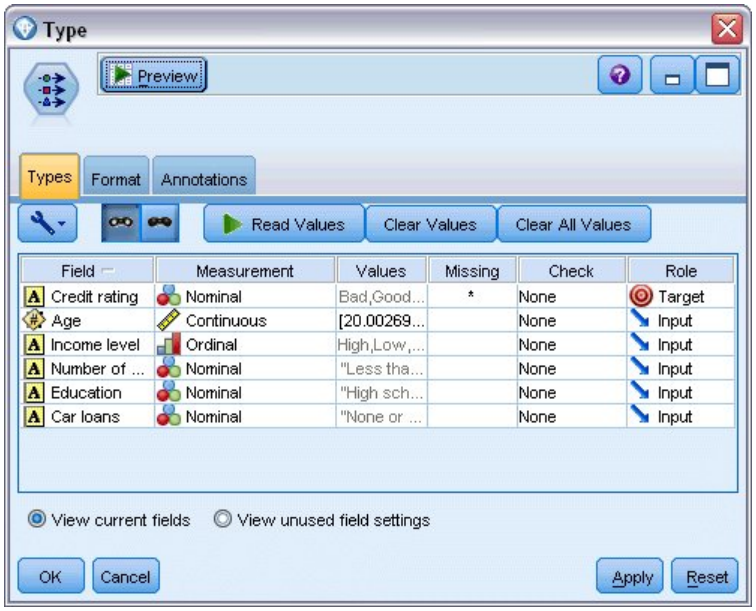



*Figura 4. Configuración de los campos de destino y entrada con el nodo Tipo*

Para cada campo, el nodo Tipo también especifica un **rol** para indicar el papel que desempeña cada campo en el modelado. El rol se define como *Objetivo* para el campo *Valoración de crédito*, que es el campo que indica si un cliente determinado ha causado mora en el pago del préstamo. Éste es el **objetivo** o campo cuyo valor queremos predecir.

El rol se define a *Entrada* para los otros campos. Los campos de entrada se conocen a menudo como **predictores**, o campos cuyos valores se utilizan en el algoritmo de modelado para predecir el valor del campo objetivo.

El nodo de modelado CHAID genera el modelo.

En la pestaña Campos del nodo de modelado está seleccionada la opción **Utilizar los roles predefinidos**, lo que significa que se utilizarán el objetivo y las entradas especificados en el nodo Tipo. En este punto podríamos cambiar los roles de campo, pero en este ejemplo las usaremos como están.

1. Pulse en la pestaña Opciones de generación.

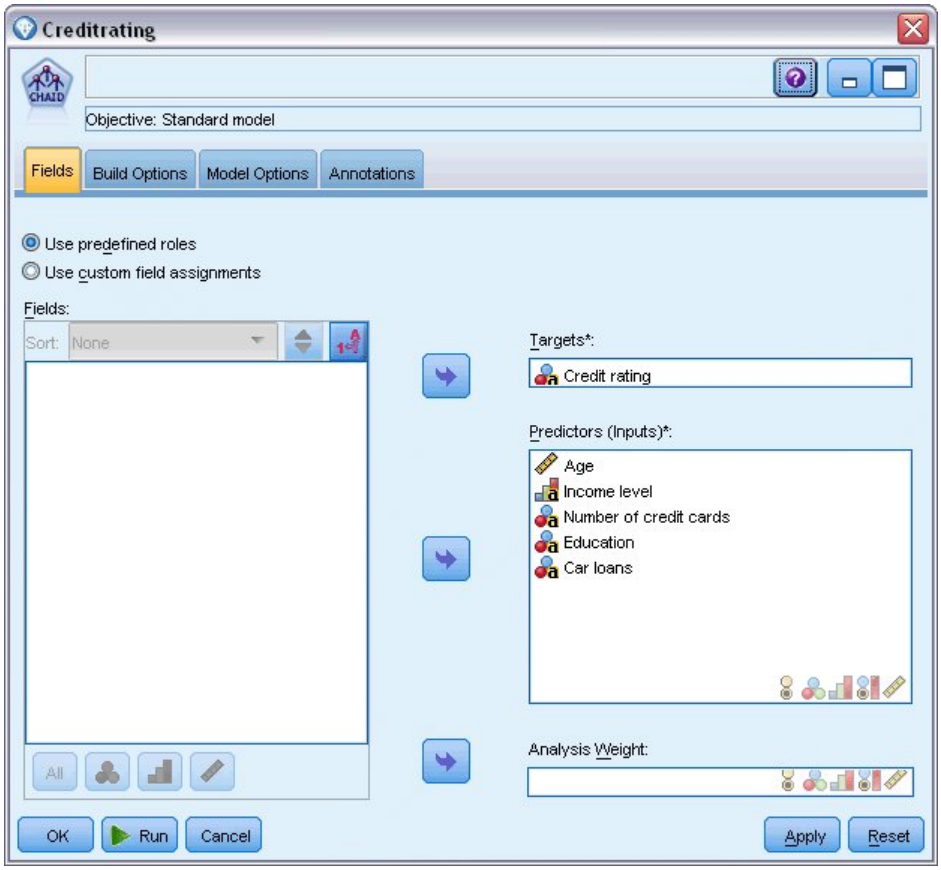



*Figura 5. Nodo de modelado CHAID, pestaña Campos*

Aquí hay varias opciones en las que podemos especificar el tipo de modelo que queremos generar.

Si queremos un modelo totalmente nuevo usaremos la opción predeterminada **Crear modelo nuevo**.

También deseamos un único modelo de árbol de decisión estándar sin mejoras, por lo que dejaremos la opción de objetivo predeterminada **Crear un árbol único**.

Aunque también podemos iniciar una sesión de modelado interactivo que nos permite ajustar con precisión el modelo, este ejemplo simplemente genera un modelo utilizando la configuración de modo predeterminada **Generar modelo**.

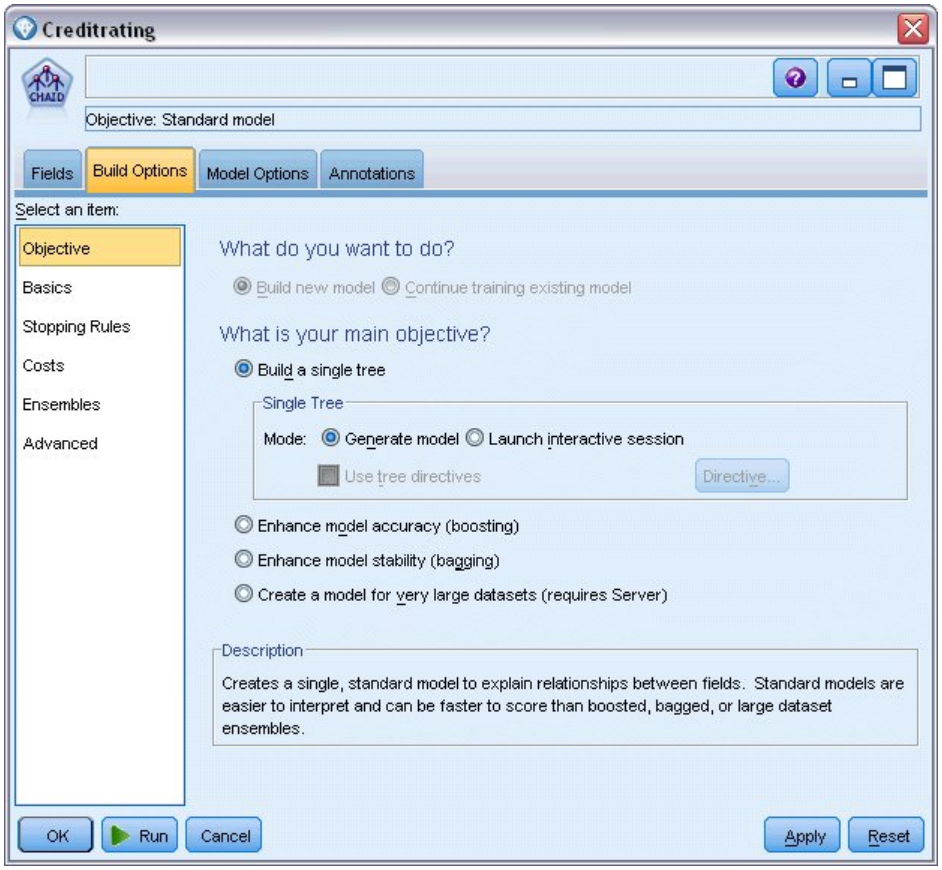



*Figura 6. Nodo de modelado CHAID, pestaña Opciones de generación*

Por ejemplo, queremos que el árbol sea bastante sencillo, así que limitaremos el crecimiento del árbol elevando el número mínimo de casos para los nodos padre e hijo.

2. En la pestaña Opciones de generación, seleccione **Reglas de parada** desde el panel de navegación de la izquierda.
3. Seleccione la opción **Utilizar valor absoluto**.
4. Establezca **Número mínimo de registros en rama padre** como 400.
5. Establezca **Número mínimo de registros por rama hija** como 200.

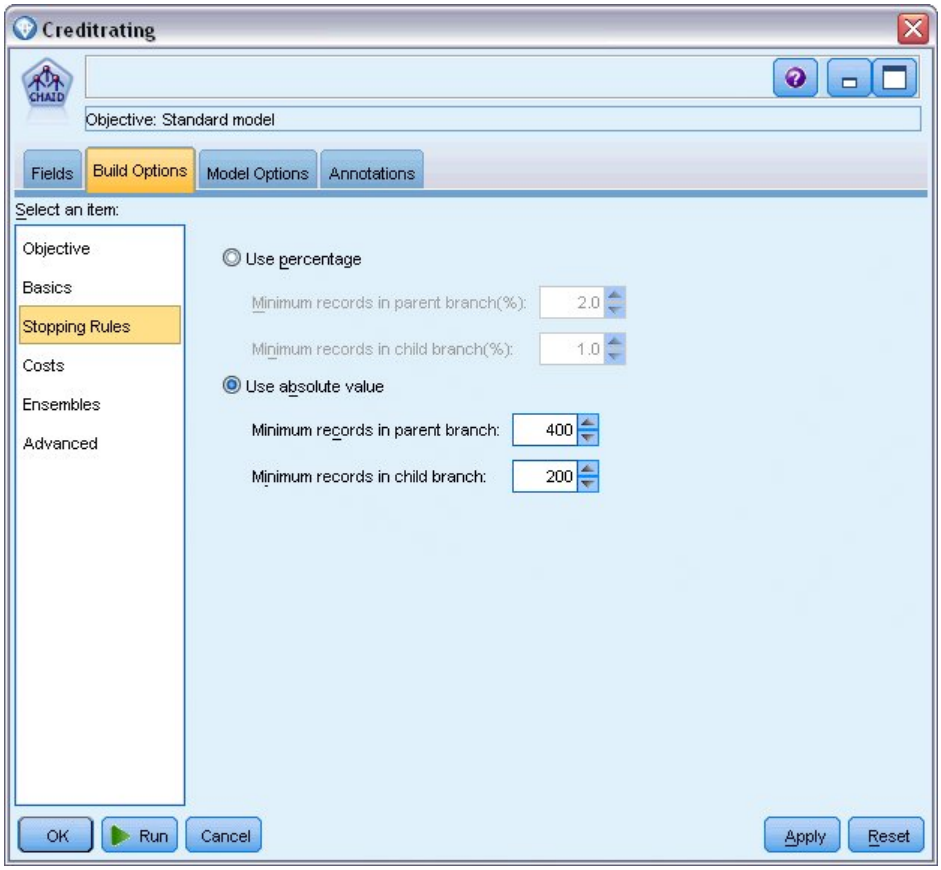



*Figura 7. Configuración de los criterios de parada para la generación de árboles de decisión*

Podemos usar todas las demás opciones predeterminadas para este ejemplo, por lo que pulse en **Ejecutar** para crear el modelo. (También puede pulsar con el botón derecho del ratón en el nodo y seleccionar **Ejecutar** del menú contextual o seleccionar el nodo y **Ejecutar** del menú Herramientas.)

## Exploración del modelo

Cuando finaliza la ejecución, se añade el nugget de modelo a la paleta Modelos en la esquina superior derecha de la ventana de aplicación, y también se coloca en el lienzo de rutas con un enlace al nodo de modelado desde el que se creó. Para ver los detalles del modelo, pulse con el botón derecho del ratón en el nugget y seleccione **Examinar** (en la paleta de modelos) o **Editar** (en el lienzo).

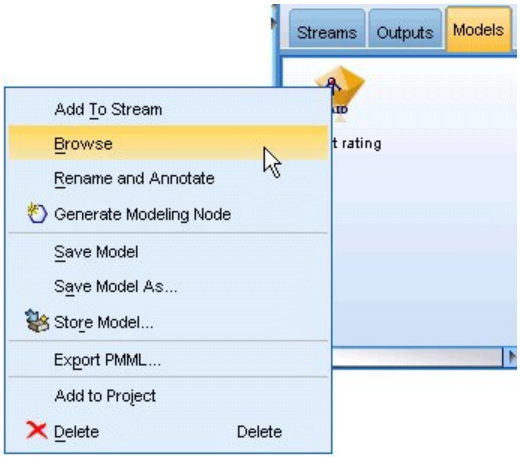



*Figura 8. Paleta de modelos*

En el caso del nugget CHAID, la pestaña Modelo muestra los detalles en forma de conjunto de reglas; éste se compone esencialmente de una serie de reglas que se pueden utilizar para asignar registros individuales a los nodos hijo basándose en los valores de distintos campos de entrada.

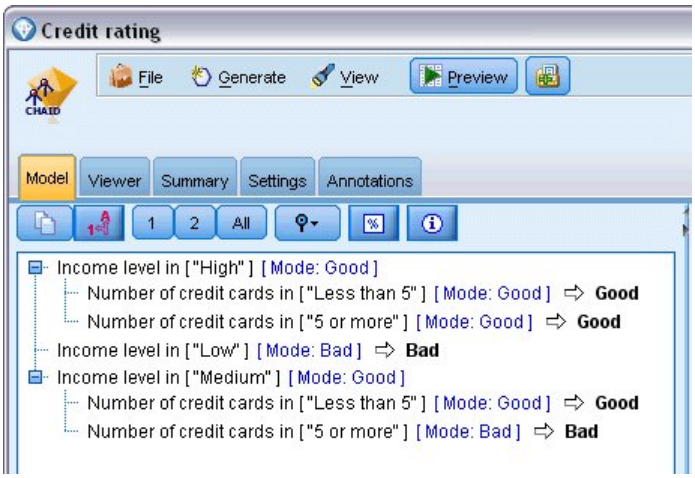



*Figura 9. Nugget de modelo CHAID, conjunto de reglas*

Por cada nodo terminal del árbol de decisión (aquellos nodos que no se dividen más) se devuelve la predicción *Bueno* o *Malo*. En cada caso, la predicción está determinada por el **modo** o, la respuesta más común, para registros que se incluyen en dicho nodo.

A la derecha del conjunto de reglas, la pestaña Predictor muestra el gráfico Importancia de variable, que muestra la importancia relativa de cada predictor en la estimación del modelo. A partir de aquí podemos determinar que *Nivel de ingresos* es fácilmente lo más significativo de este caso, y que el otro valor significativo es *Número de tarjetas de crédito en propiedad*.

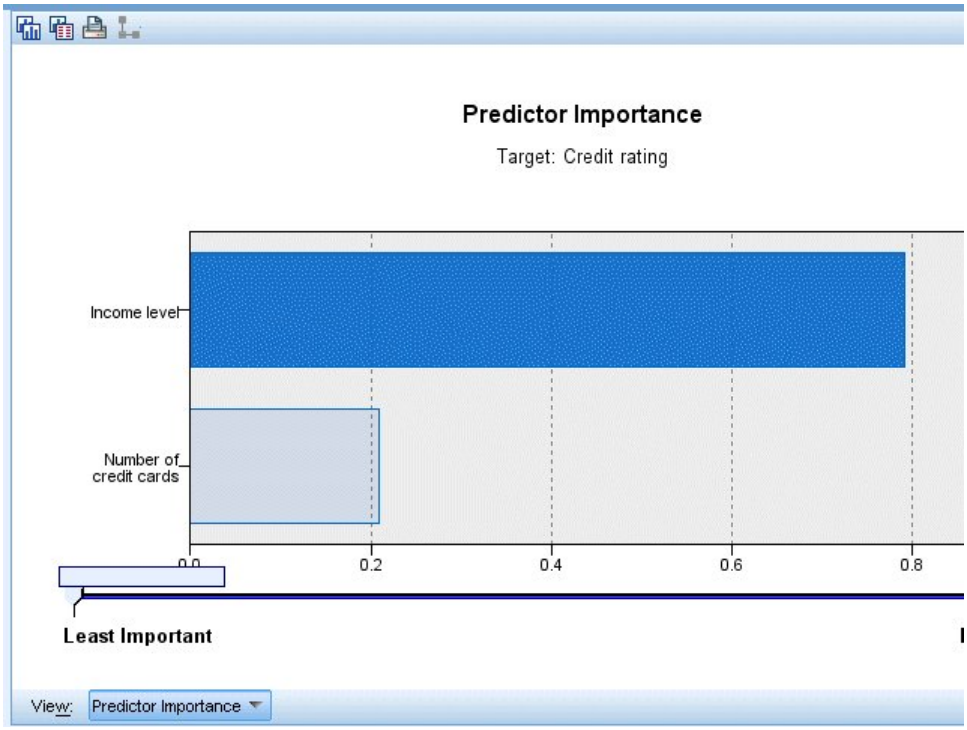



*Figura 10. Gráfico Importancia del predictor*

La pestaña Visor del nugget de modelo muestra el mismo modelo en forma de árbol, con un nodo en cada punto de decisión. Utilice los controles Zoom de la barra de herramientas para acercarse a un nodo específico o alejarse para ver una parte más amplia del árbol.

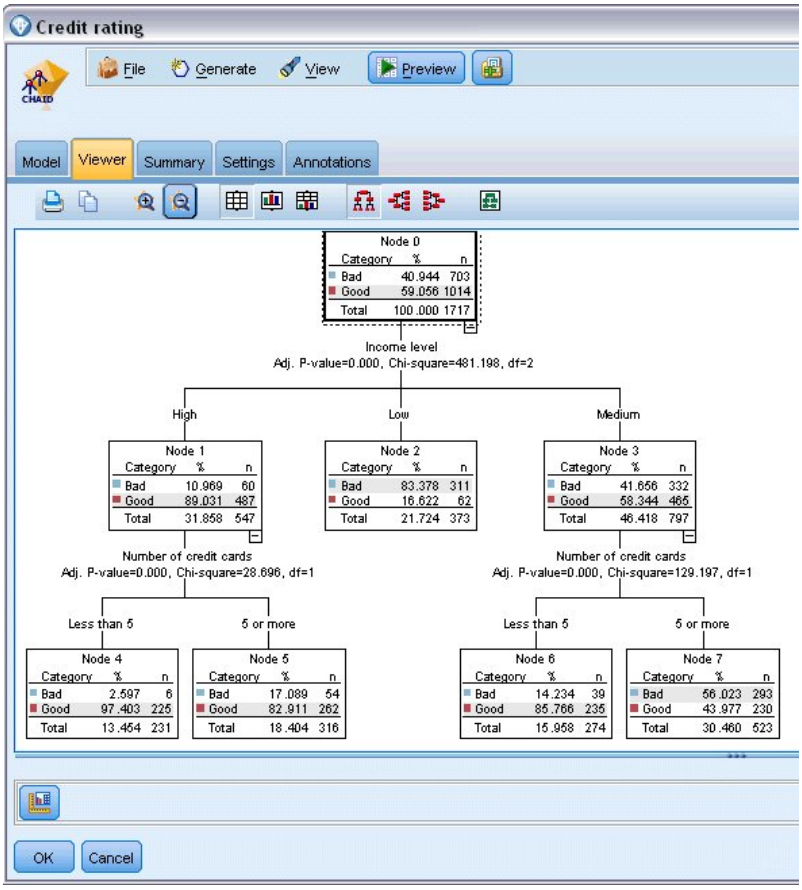



*Figura 11. Pestaña Visor del nugget de modelo, con la función alejar seleccionada*

Al observar la parte superior del árbol, el primer nodo (Nodo 0) nos ofrece un resumen de todos los registros del conjunto de datos. Algo más del 40% de los casos del conjunto de datos se clasifica como un riesgo malo. Es una proporción bastante alta, de modo que veamos si el árbol puede darnos más pistas sobre qué factores pueden ser los responsables.

Podemos ver que la primera división es por *Nivel de ingresos*. Los registros cuyo nivel de ingresos están en la categoría *Bajo* se asignan al Nodo 2, por lo que no es sorprendente que esta categoría contenga el mayor porcentaje de morosos de préstamos. Claramente, la concesión de un préstamo a clientes de esta categoría conlleva un alto riesgo.

Sin embargo, el 16% de los clientes de esta categoría *no* presentó mora en los pagos, por lo que la predicción no siempre será correcta. Ningún modelo puede predecir de manera fiable todas las respuestas, pero un buen modelo debe permitirnos predecir la respuesta *más probable* para cada registro basándonos en los datos disponibles.

Del mismo modo, si observamos a los clientes con ingresos elevados (Nodo 1), vemos que la amplia mayoría (89%) es un riesgo bueno. Sin embargo, también más de 1 de 10 de estos clientes ha cometido mora en los pagos. ¿Podemos refinar nuestros criterios de concesión de préstamos para minimizar estos riesgos?

Tenga en cuenta cómo ha dividido el modelo a estos clientes en dos subcategorías (Nodos 4 y 5) basándose en el número de tarjetas de crédito en propiedad. En el caso de clientes con ingresos elevados, si concedemos préstamos sólo a los que tengan menos de 5 tarjetas de crédito, podemos incrementar nuestra tasa de éxito del 89% al 97%, un resultado aun más satisfactorio.

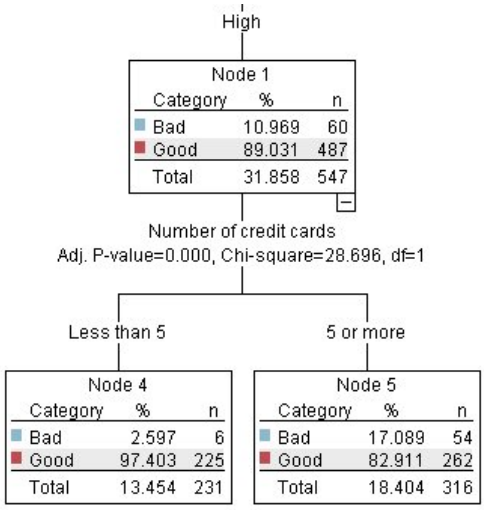



*Figura 12. Vista de árbol de clientes con ingresos elevados*

¿Qué ocurre con los clientes de la categoría de ingresos Medio (Nodo 3)? Están divididos mucho más homogéneamente entre las valoraciones Bueno y Malo.

De nuevo, las subcategorías (Nodos 6 y 7 en este caso) pueden ayudarnos. Esta vez, la concesión de préstamos sólo a los clientes con ingresos medios con menos de 5 tarjetas de crédito aumenta el porcentaje de valoraciones Bueno del 58% al 85%, lo cual es una mejora significativa.

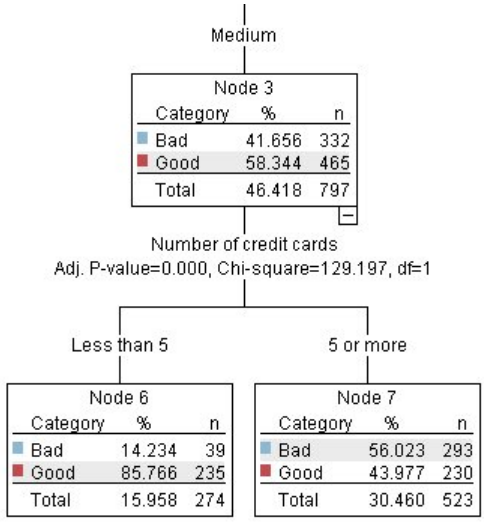



*Figura 13. Vista de árbol de clientes con ingresos medios*

Por lo tanto, hemos aprendido que cada registro que se entre en este modelo se le asignará a un nodo específico y se le asignará una predicción de *Bueno* o *Malo* según la respuesta más común de ese nodo.

Este proceso de asignar predicciones a registros individuales se conoce como **puntuación**. Al puntuar los mismos registros utilizados para calcular el modelo, podemos evaluar cuál es el rendimiento preciso en los datos de entrenamiento, es decir los datos para los que conocemos el resultado. Veamos cómo hacer esto.

## Evaluación del modelo

Hemos estado explorando el modelo para comprender cómo funciona la puntuación. Pero para evaluar *con qué precisión* trabaja, debemos puntuar varios registros y comparar las respuestas predichas por el modelo con los resultados reales. Vamos a puntuar los mismos registros que se utilizaron para estimar el modelo, lo que nos permite comparar las respuestas observadas y predichas.

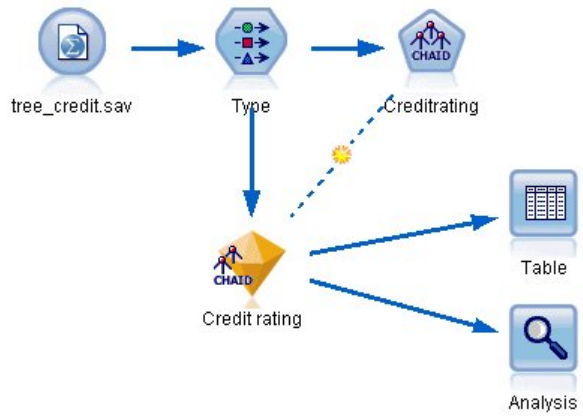
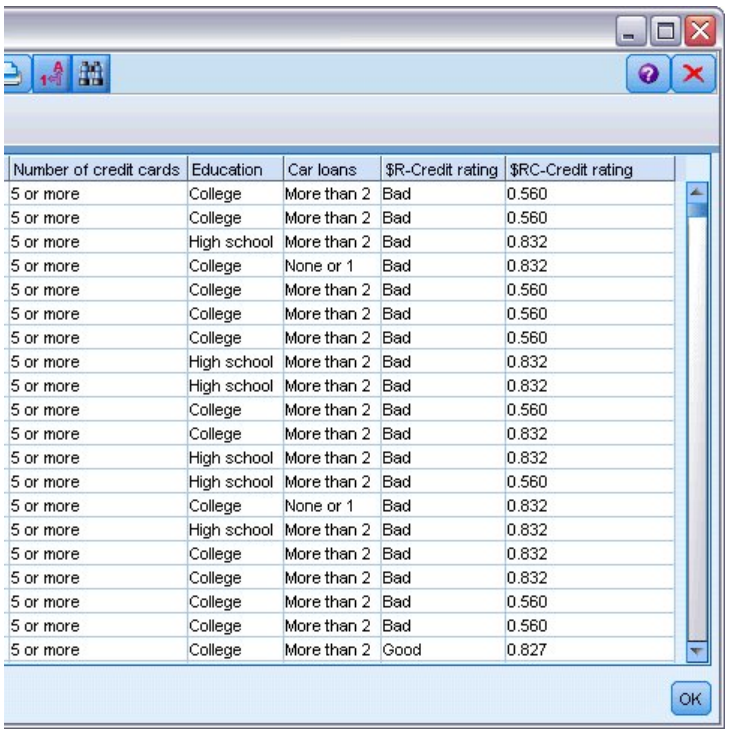



*Figura 14. Adición del nugget de modelo a los nodos de salida para la generación del modelo*

1. Para ver las puntuaciones o predicciones, adjunte el nodo Tabla al nugget de modelo, pulse dos veces en el nodo Tabla y pulse en **Ejecutar**.

   La tabla muestra las puntuaciones predichas en un campo denominado *$R-Valoración de crédito*, creado por el modelo. Podemos comparar estos valores con el campo *Valoración de crédito* original que contiene las respuestas reales.

   Por convención, los nombres de los campos generados durante la puntuación se basan en el campo objetivo, pero con un prefijo estándar. Los prefijos *$G* y *$GE* se generan mediante el modelo lineal generalizado, *$R* es el prefijo usado para la predicción generada por el modelo CHAID en este caso, *$RC* sirve para los valores de confianza, *$X* suele generarse utilizando un conjunto, y *$XR*, *$XS* y *$XF* se emplean como prefijos en casos en los que el campo de destino es un campo Continuo, Categórico, Conjunto o Marca, respectivamente. Los distintos tipos de modelo utilizan diferentes conjuntos de prefijos. Un **valor de confianza** es la estimación propia del modelo, en una escala de 0,0 a 1,0, sobre el grado de precisión de cada valor predicho.



| Number of credit cards | Education | Car loans | $R-Credit rating | $RC-Credit rating |
|---|---|---|---|---|
| 5 or more | College | More than 2 | Bad | 0.560 |
| 5 or more | College | More than 2 | Bad | 0.560 |
| 5 or more | High school | More than 2 | Bad | 0.832 |
| 5 or more | College | None or 1 | Bad | 0.832 |
| 5 or more | College | More than 2 | Bad | 0.560 |
| 5 or more | College | More than 2 | Bad | 0.560 |
| 5 or more | College | More than 2 | Bad | 0.560 |
| 5 or more | High school | More than 2 | Bad | 0.832 |
| 5 or more | High school | More than 2 | Bad | 0.832 |
| 5 or more | College | More than 2 | Bad | 0.560 |
| 5 or more | College | More than 2 | Bad | 0.832 |
| 5 or more | High school | More than 2 | Bad | 0.832 |
| 5 or more | High school | More than 2 | Bad | 0.560 |
| 5 or more | College | None or 1 | Bad | 0.832 |
| 5 or more | High school | More than 2 | Bad | 0.832 |
| 5 or more | College | More than 2 | Bad | 0.832 |
| 5 or more | College | More than 2 | Bad | 0.832 |
| 5 or more | College | More than 2 | Bad | 0.560 |
| 5 or more | College | More than 2 | Bad | 0.560 |
| 5 or more | College | More than 2 | Good | 0.827 |

OK

*Figura 15. Tabla que muestra las puntuaciones generadas y los valores de confianza*

Como se esperaba, el valor predicho coincide con las respuestas reales de muchos registros, pero no todos. El motivo es que cada nodo terminal CHAID tiene una mezcla de respuestas. La predicción coincide con la *más común*, pero es incorrecto para el resto de dicho nodo. (Recuerde la minoría del 16% de clientes con ingresos bajos que no cometió mora en los pagos.)

Para evitarlo, podemos seguir dividiendo el árbol en ramas cada vez más pequeñas, hasta que cada nodo sea 100 % todo puro *Bueno* o *Malo* sin respuestas mixtas. Pero dicho modelo sería extremadamente complicado y probablemente no se generalizaría bien en otros conjuntos de datos.

Para descubrir exactamente cuántas predicciones son correctas, podríamos observar la tabla y anotar el número de registros en los que el valor del campo predicho *$R-Valoración de crédito* coincida con el valor de *Valoración de crédito*. Afortunadamente, hay un modo más sencillo: podemos utilizar un nodo Análisis, que lo hace automáticamente.

2. Conecte el nugget de modelo al nodo Análisis.
3. Pulse dos veces en el nodo Análisis y pulse en **Ejecutar**.

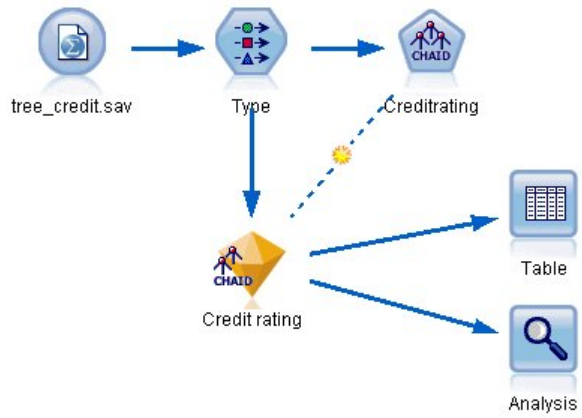



*Figura 16. Conexión del nodo Análisis*

El análisis muestra que para 1899 de 2464 registros (más del 77%), el valor predicho por el modelo coincidía con la respuesta real.

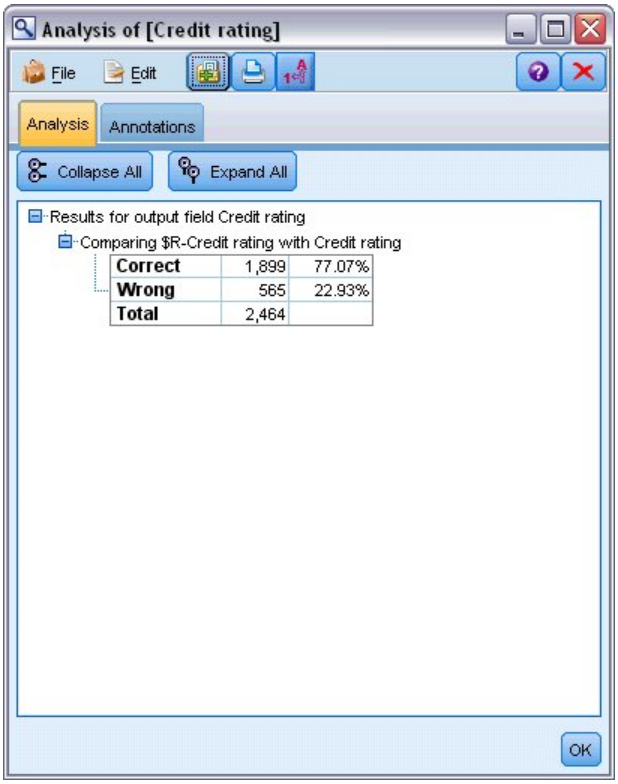



*Figura 17. Resultados de análisis que comparan respuestas observadas y predichas*

Este resultado está limitado por el hecho de que los registros que se están puntuando son los mismos utilizados para calcular el modelo. En una situación real, podría utilizar un nodo Partición para dividir los datos en muestras separadas para el entrenamiento y la evaluación.

Si utiliza una partición de muestra para generar el modelo y otra muestra para comprobarlo, podrá obtener una indicación mucho mejor de lo bien que se generalizará en otros conjuntos de datos.

El nodo Análisis nos permite comprobar el modelo frente a registros para los que ya conocemos el resultado real. La etapa siguiente muestra cómo podemos utilizar el modelo para puntuar registros cuyos

resultados no conocemos. Por ejemplo, esto podría incluir a personas que no son clientes actuales del banco, pero son posibles objetivos de correos promocionales.

## Puntuación de registros

Antes hemos puntuado los mismos registros utilizados para calcular el modelo con el fin de evaluar el grado de precisión del modelo. Ahora vamos a ver cómo puntuar un conjunto de registros diferentes de los utilizados para crear el modelo. Este es el objetivo para el modelado con un campo objetivo: estudie los registros de los que conoce los resultados para identificar patrones que le permitan predecir resultados que aún no conoce.

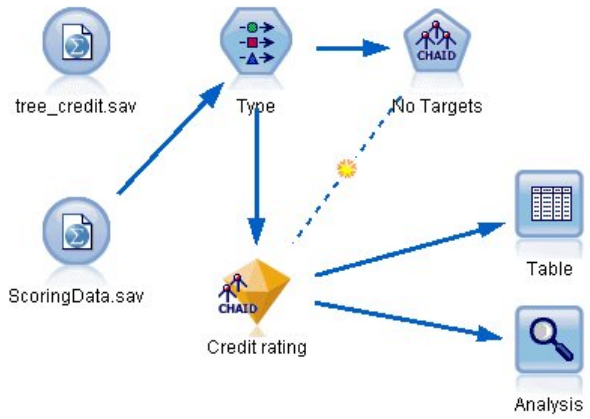



*Figura 18. Adición de nuevos datos para su puntuación*

Podría actualizar el nodo de origen Archivo Statistics para dirigirse a un archivo de datos diferente o podría añadir un nuevo nodo de origen que lea los datos que desea puntuar. En cualquier caso, el nuevo conjunto de datos debe contener los mismos campos de entrada utilizados por el modelo (*Edad*, *Nivel de ingresos*, *Educación*, etc.) pero no el campo objetivo *Valoración de crédito*.

También podría añadir el nugget de modelo a cualquier ruta que incluya los campos de entrada esperados. El tipo de origen no importa, tanto si se ha leído de un archivo o de una base de datos, siempre que los nombres y tipos de campo coincidan con los utilizados por el modelo.

También podría guardar el nugget de modelo como un archivo independiente, o exportar el modelo en formato PMML para su uso con otras aplicaciones que admitan este formato, o almacenar el modelo en un repositorio IBM SPSS Collaboration and Deployment Services, que ofrece despliegue, puntuación y gestión de modelos en toda la empresa.

Independientemente de la infraestructura utilizada, el propio modelo funciona del mismo modo.

## Resumen

Este ejemplo demuestra los pasos básicos para crear, evaluar y puntuar un modelo.

- El nodo de modelado calcula el modelo estudiando registros para los que se conoce el resultado y crea un nugget de modelo. Esto se denomina a veces entrenamiento del modelo.
- El nugget de modelo puede añadirse a cualquier ruta con los campos esperados para puntuar registros. Al puntuar los registros de los que ya conoce el resultado (como los clientes existentes), puede evaluar el grado de rendimiento.
- Una vez quede satisfecho con el rendimiento adecuado del modelo, podrá puntuar nuevos datos (como clientes potenciales) para predecir cómo responderán.

- Debe hacerse referencia a los datos utilizados para entrenar o calcular el modelo como los datos analíticos o históricos; también se puede hacer referencia a los datos de puntuación como los datos operativos.

# Capítulo 3. Conceptos básicos sobre modelado

## Visión general de nodos de modelado

IBM SPSS Modeler ofrece una gran variedad de métodos de modelado procedentes del aprendizaje automático, la inteligencia artificial y el estadístico. Los métodos disponibles en la paleta de modelado permiten derivar nueva información procedente de los datos y desarrollar modelos predictivos. Cada método tiene ciertos puntos fuertes y es más adecuado para determinados tipos de problemas.

La *Guía de aplicaciones de IBM SPSS Modeler* ofrece ejemplos para muchos de estos métodos, junto con una introducción general al proceso de modelado. Esta guía está disponible como una guía de aprendizaje en línea y, también, en formato PDF. Consulte el tema "Ejemplos de aplicación" en la página 4 para obtener más información.

Los métodos de modelado se dividen en estas categorías:

- Supervisado
- Asociación
- Segmentación.

Modelos supervisados

Los *modelos supervisados* utilizan los valores de uno o varios campos de **entrada** para predecir el valor de uno o varios campos de resultados o **destino**. Algunos ejemplos de estas técnicas son: árboles de decisiones (árbol C&R, QUEST, CHAID y algoritmos C5.0), regresión (lineal, logística, lineal generalizada y algoritmos de regresión de Cox), redes neuronales, máquinas de vectores de soporte y redes bayesianas.

Los modelos supervisados ayudan a las organizaciones a predecir un resultado conocido, por ejemplo si un cliente comprará o se irá o si una transacción se ajusta a un patrón conocido de fraude. Las técnicas de modelado incluyen aprendizaje automático de las máquinas, inducción de reglas, identificación de subgrupos, métodos estadísticos y generación de varios modelos.

*Nodos supervisados*

El nodo Clasificador automático crea y compara varios modelos diferentes para obtener resultados binarios (sí o no, abandono o no de clientes, etc.), lo que le permite seleccionar el mejor enfoque para un análisis determinado. Son compatibles varios algoritmos de modelado, por lo que es posible seleccionar los métodos que desee utilizar, las opciones específicas para cada uno y los criterios para comparar los resultados. El nodo genera un conjunto de modelos basado en las opciones especificadas y clasifica los mejores candidatos en función de los criterios que especifique.

El nodo Autonumérico calcula y compara modelos para resultados de rango numérico continuo utilizando cierto número de métodos diferentes. El nodo funciona de la misma manera que el nodo Clasificador automático, lo que le permite seleccionar los algoritmos que desee utilizar y experimentar con varias combinaciones de opciones en una única pasada de modelado. Los algoritmos admitidos incluyen redes neuronales, C&RT, CHAID, regresión lineal, regresión lineal generalizada y máquinas de vectores de soporte (SVM). Los modelos se pueden comparar basándose en la correlación, el error relativo o el número de variables utilizado.

El nodo de árbol de clasificación y regresión (C&R) genera un árbol de decisión que permite predecir o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera "puro" si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos (nominal, ordinal o marca). Todas las divisiones son binarias (sólo se crean dos subgrupos).

El nodo QUEST proporciona un método de clasificación binario para generar árboles de decisión; está diseñado para reducir el tiempo de procesamiento necesario para realizar los análisis de C&RT y reducir la tendencia de los métodos de clasificación de árboles para favorecer a las entradas que permitan realizar más divisiones. Los campos de entrada pueden ser continuos (rango numérico), sin embargo el campo objetivo debe ser categórico. Todas las divisiones son binarias.

El nodo CHAID genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. A diferencia de los nodos C&RT y QUEST, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles, aunque necesita más tiempo para realizar los cálculos.

El nodo C5.0 genera un árbol de decisión o un conjunto de reglas. El modelo divide la muestra basándose en el campo que ofrece la máxima ganancia de información en cada nivel. El campo objetivo debe ser categórico. Se permiten varias divisiones en más de dos subgrupos.

El nodo Lista de decisiones identifica subgrupos, o segmentos, que muestran una mayor o menor posibilidad de proporcionar un resultado binario relacionado con la población global. Por ejemplo, puede buscar clientes que tengan menos posibilidades de abandonar o más posibilidades de responder favorablemente a una campaña. Puede incorporar su conocimiento empresarial al modelo añadiendo sus propios segmentos personalizados y previsualizando modelos alternativos uno junto a otro para comparar los resultados. Los modelos de listas de decisiones constan de una lista de reglas en las que cada regla tiene una condición y un resultado. Las reglas se aplican en orden, y la primera regla que coincide determina el resultado.

Los modelos de regresión lineal predicen un objetivo continuo tomando como base las relaciones lineales entre el destino y uno o más predictores.

El nodo PCA/Factorial proporciona técnicas eficaces de reducción de datos para reducir la complejidad de los datos. Análisis de componentes principales (PCA) busca combinaciones lineales de los campos de entrada que realizan el mejor trabajo a la hora de capturar la varianza en todo el conjunto de campos, en el que los componentes son ortogonales (perpendiculares) entre ellos. Análisis factorial intenta identificar factores subyacentes que expliquen el patrón de correlaciones dentro de un conjunto de campos observados. Para los dos métodos, el objetivo es encontrar un número pequeño de campos derivados que resuma de forma eficaz la información del conjunto original de campos.

El nodo Selección de características filtra los campos de entrada para su eliminación en función de un conjunto de criterios (como el porcentaje de valores perdidos); a continuación, clasifica el grado de importancia del resto de entradas de acuerdo con un objetivo específico. Por ejemplo, a partir de un conjunto de datos dado con cientos de entradas potenciales, ¿cuáles tienen mayor probabilidad de ser útiles para el modelado de resultados de pacientes?

El análisis discriminante realiza más supuestos rigurosos que regresiones logísticas, pero puede ser una alternativa o un suplemento valioso al análisis de regresión logística si se cumplen dichos supuestos.

La regresión logística es una técnica de estadístico para clasificar los registros en función los valores de los campos de entrada. Es análoga a la regresión lineal pero toma un campo objetivo categórico en lugar de uno numérico.

El modelo lineal generalizado amplía el modelo lineal general, de manera que la variable dependiente está relacionada linealmente con los factores y las covariables mediante una determinada función de enlace. Además, el modelo permite que la variable dependiente tenga una distribución no normal. Cubre la funcionalidad de un amplio número de modelo estadísticos, incluyendo regresión lineal, regresión logística, modelos log lineales para recuento de datos y modelos de supervivencia censurados por intervalos.

Un modelo lineal mixto generalizado (GLMM) amplía el modelo lineal de modo que el objetivo pueda tener una distribución no normal, esté linealmente relacionado con los factores y covariables mediante una función de enlace especificada y las observaciones se puedan correlacionar. Los modelos lineales mixtos generalizados cubren una amplia variedad de modelos, desde modelos de regresión lineal simple hasta modelos multinivel complejos para datos longitudinales no normales.


El nodo Regresión de Cox le permite crear un modelo de supervivencia para datos de tiempo hasta el evento en presencia de registros censurados. El modelo produce una función de supervivencia que predice la probabilidad de que el evento de interés se haya producido en el momento dado (*t*) para valores determinados de las variables de entrada.

El nodo Máquina de vectores de soporte (SVM) le permite clasificar datos en uno o dos grupos sin que haya un ajuste por exceso. SVM funciona bien con conjuntos de datos grandes, como aquellos con un gran número de campos de entrada.

El nodo Red bayesiana le permite crear un modelo de probabilidad combinando pruebas observadas y registradas con conocimiento del mundo real para establecer la probabilidad de instancias. El nodo se centra en las redes Naïve Bayes aumentado a árbol (TAN) y de manto de Markov que se utilizan principalmente para la clasificación.

El nodo Modelo de respuesta de autoaprendizaje (SLRM) permite crear un modelo en el que un solo caso nuevo o un pequeño número de casos nuevos se pueden utilizar para volver a calcular el modelo sin tener que entrenar de nuevo el modelo utilizando todos los datos.

El nodo Serie temporal estima modelos de suavizado exponencial, modelos autorregresivos integrados de media móvil (ARIMA) univariados y modelos ARIMA (o de función de transferencia) multivariados para series temporales y genera previsiones. Este nodo Serie temporal es similar al nodo Serie temporal anterior que estaba en desuso en SPSS Modeler versión 18. Sin embargo, este nodo Serie temporal más reciente se ha diseñado para emplear la potencia de IBM SPSS Analytic Server para procesar grandes cantidades de datos y mostrar el modelo resultante en el visor de resultados que se ha añadido en SPSS Modeler versión 17.

El nodo *k* de modelado de vecino (KNN) asocia el nuevo caso con la categoría o valor de los objetos *k* junto a él en el espacio de predictores, donde *k* es un entero. Los casos parecidos están próximos y los que no lo son están alejados entre sí.

El nodo Predicción espacio-temporal (STP) utiliza datos que contienen datos de ubicación, campos de entrada para la predicción (predictores), un campo de hora y un campo de objetivo. Cada ubicación tiene muchas filas en los datos que representan los valores de cada predictor en cada tiempo de medición. Después de analizar los datos, se puede utilizar para predecir los valores de objetivo en cualquier ubicación dentro de los datos de forma que se utilizan en el análisis.

## Modelos de asociación

Los *modelos de asociación* encuentran patrones en los datos en los que una o más entidades (como eventos, compras o atributos) se asocian con una o más entidades. Los modelos construyen conjuntos de reglas que definen estas relaciones. Aquí los campos de los datos pueden funcionar como entradas y destinos. Podría encontrar estas asociaciones manualmente, pero los algoritmos de reglas de asociaciones lo hacen mucho más rápido, y pueden explorar patrones más complejos. Los modelos Apriori y Carma son ejemplos del uso de estos algoritmos. Otro tipo de modelo de asociación es el modelo de detección de secuencias, que encuentra patrones secuenciales en datos estructurados temporalmente.

Los modelos de asociación son los más útiles si se desean predecir varios resultados; por ejemplo, los clientes que adquirieron el producto X también adquirieron Y y Z. Los modelos de asociación relacionan una conclusión específica (como la decisión de adquirir un producto) con un conjunto de condiciones. La ventaja de los algoritmos de reglas de asociación sobre los algoritmos más estándar de árboles de decisión (C5.0 y Árbol C&R) es que las asociaciones pueden existir entre cualquiera de los atributos. Un algoritmo de árbol de decisión generará reglas con una única conclusión, mientras que los algoritmos de asociación tratan de buscar muchas reglas, cada una de las cuales puede tener una conclusión diferente.

### *Nodos de asociación*

El nodo Apriori extrae un conjunto de reglas de los datos y destaca aquellas reglas con un mayor contenido de información. Apriori ofrece cinco métodos diferentes para la selección de reglas y utiliza un sofisticado esquema de indización para procesar eficientemente grandes conjuntos de datos. En los problemas de mucho volumen, Apriori se entrena más rápidamente, no tiene un límite arbitrario para el número de reglas que puede retener y puede gestionar reglas que tengan hasta 32 precondiciones. Apriori requiere que todos los campos de entrada y salida sean categóricos, pero ofrece un mejor rendimiento ya que está optimizado para este tipo de datos.

El modelo CARMA extrae un conjunto de reglas de los datos sin necesidad de especificar campos de entrada ni de objetivo. A diferencia de Apriori el nodo CARMA ofrece configuraciones de generación basadas en el soporte de las reglas (soporte tanto para el antecedente como el consecuente) en lugar de hacerlo sólo respecto al soporte del antecedente. Esto significa que las reglas generadas se pueden utilizar en una gama de aplicaciones más amplia, por ejemplo, para buscar una lista de productos o servicios (antecedentes) cuyo consecuente es el elemento que se desea promocionar durante esta temporada de vacaciones.

El nodo Secuencia encuentra reglas de asociación en datos secuenciales o en datos ordenados en el tiempo. Una secuencia es una lista de conjuntos de elementos que tiende a producirse en un orden previsible. Por ejemplo, si un cliente compra una cuchilla y una loción para después del afeitado, probablemente comprará crema para afeitar la próxima vez que vaya a comprar. El nodo Secuencia se basa en el algoritmo de reglas de asociación de CARMA, que utiliza un método de dos pasos para encontrar las secuencias.

El nodo Reglas de asociación es parecido al nodo Apriori; sin embargo, a diferencia de Apriori, el nodo Reglas de asociación puede procesar datos de lista. Además, el nodo Reglas de asociación se puede utilizar con IBM SPSS Analytic Server para procesar big data y aprovechar el procesamiento paralelo.

## Modelos de segmentación

Los *modelos de segmentación* dividen los datos en segmentos o clústeres de registros que tienen patrones similares de campos de entrada. Como sólo se interesan por los campos de entrada, los modelos de segmentación no contemplan el concepto de campos de salida o destino. Ejemplos de modelos de segmentación son las redes Kohonen, la agrupación en clústeres de K-medias, la agrupación en clústeres en dos pasos y la detección de anomalías.

Los modelos de segmentación (también conocidos como "modelos de agrupación en clústeres") son útiles en aquellos casos en los que se desconoce el resultado específico (por ejemplo a la hora de detectar nuevos patrones de fraude o de identificar grupos de interés en la base de clientes). Los modelos de agrupación en clústeres se centran en la identificación de grupos de registros similares y en el etiquetado de registros según el grupo al que pertenecen. Esto se lleva a cabo sin la ventaja que ofrece el conocimiento previo sobre los grupos y sus características, y diferencia a los modelos de clústeres de otras técnicas de modelado en que no hay campos de salida u objetivo predefinidos para el modelo que se va a predecir. No hay respuestas correctas o incorrectas para estos modelos. Su valor viene determinado por su capacidad de capturar agrupaciones interesantes en los datos y proporcionar descripciones útiles de dichas agrupaciones. Los modelos de clúster se usan a menudo para crear clústeres o segmentos que se usan posteriormente como entradas en análisis posteriores, (por ejemplo mediante la segmentación de clientes potenciales en subgrupos homogéneos).

### *Nodos de segmentación*

El nodo Agrupación en clústeres automática calcula y compara los modelos de agrupación en clústeres que identifican grupos de registros con características similares. El nodo funciona de la misma manera que otros nodos de modelado automático, permitiéndole experimentar con múltiples combinaciones de opciones en una única pasada de modelado. Los modelos se pueden comparar utilizando medidas básicas con las que se intenta filtrar y definir la utilidad de los modelos de clúster y proporcionar una medida según la importancia de campos concretos.

El nodo K-medias agrupa conjuntos de datos en grupos distintos (o clústeres). El método define un número fijo de clústeres, de forma iterativa asigna registros a los clústeres y ajusta los centros de los clústeres hasta que no se pueda mejorar el modelo. En lugar de intentar predecir un resultado, los modelos de *k*-medias utilizan un proceso conocido como aprendizaje no supervisado para revelar los patrones del conjunto de campos de entrada.

El nodo Kohonen genera un tipo de red neuronal que se puede usar para agrupar un conjunto de datos en grupos distintos. Cuando la red se termina de entrenar, los registros que son similares se deberían cerrar juntos en el mapa de resultados, mientras que los registros que son diferentes aparecerían aparte. Puede observar el número de observaciones capturadas por cada unidad en el nugget de modelo para identificar unidades fuertes. Esto le proporcionará una idea del número apropiado de clústeres.

El nodo Bietápico es un método de agrupación en clústeres de dos pasos. El primer paso es hacer una única pasada por los datos para comprimir los datos de entrada de la fila en un conjunto de subclústeres administrable. El segundo paso utiliza un método de agrupación en clústeres jerárquica para fundir progresivamente los subclústeres en clústeres cada vez más grandes. El bietápico tiene la ventaja de estimar automáticamente el número óptimo de clústeres para los datos de entrenamiento. Puede gestionar tipos de campos mixtos y grandes conjuntos de datos eficazmente.

El nodo Detección de anomalías identifica casos extraños, o valores atípicos, que no se ajustan a patrones de datos "normales". Con este nodo, es posible identificar valores atípicos aunque no se ajusten a ningún patrón previamente conocido o no se realice una búsqueda exacta.

Modelos de minería interna de bases de datos

IBM SPSS Modeler admite la integración con herramientas de modelado y minería de datos que están disponibles en proveedores de bases de datos, incluido como Oracle Data Miner y Microsoft Analysis Services. Podrá crear, puntuar y almacenar modelos dentro de la base de datos, todo desde la aplicación IBM SPSS Modeler. Para obtener detalles completos, consulte el *Manual de minería interna de bases de datos de IBM SPSS Modeler*.

Modelos de IBM SPSS Statistics

Si dispone de una copia de IBM SPSS Statistics instalada y con la licencia necesaria en su ordenador, puede acceder y ejecutar determinadas rutinas de IBM SPSS Statistics en IBM SPSS Modeler para generar y puntuar modelos.

## Generación de modelos divididos

El modelado de divisiones le permite utilizar una sola ruta para crear modelos separados para cada posible valor de un campo de entrada de marca, nominal o continuo, con los modelos resultantes accesibles desde un nugget de modelo único. Los posibles valores de campos de entrada pueden tener efectos muy diferentes en el modelo. Con el modelado de divisiones se puede obtener el modelo que mejor se ajusta a cada valor de campo posible en una ejecución simple de la ruta.

Tenga en cuenta que las sesiones de modelado interactivo no utilizan división. Con el modelado interactivo es posible especificar cada modelo de forma individual, por lo que no supone ninguna ventaja en el uso de la división, que crea múltiples modelos de forma automática.

El modelado de divisiones funciona designando un campo de entrada concreto como un campo de división. Puede hacerlo definiendo el rol del campo como **Dividir** en la especificación de tipo.

Solo puede designar campos con un nivel de medición de **Marca**, **Nominal**, **Ordinal** o **Continuo** como campos de división.

Puede asignar más de un campo de entrada como campo de división. En este caso, sin embargo, el número de modelos que se crea se puede aumentar en gran medida. Se crea un modelo para cada combinación posible de los valores de los campos de división seleccionados. Por ejemplo, si tres campos de entrada, cada uno con tres valores posibles, se designan como campos de división, se crearán 27 modelos diferentes.

Incluso si asigna uno o más campos como campos de división, podrá seleccionar si desea crear modelos divididos o un modelo único, mediante un valor de casilla de verificación en el diálogo de nodo de modelado.

Si define los campos de división, pero no se ha seleccionado la casilla de verificación, solo se generará un modelo simple. De la misma forma, si se selecciona la casilla de verificación pero no se define el campo de división, la división se ignorará y se generará un modelo simple.

Si ejecuta la ruta, se genera un modelo diferente en segundo plano para cada valor posible del campo o campos de división, pero solo se coloca un nugget de modelo en la paleta de modelos y el lienzo de rutas. Un nugget de modelo dividido se indicado con el símbolo de división; son dos rectángulos grises recubiertos en el nugget de imagen.

Si navega por el nugget de modelo dividido, verá una lista de todos los modelos independientes que se han generado.

Puede investigar un modelo individual de una lista si pulsa dos veces en el icono del nugget en el visor. De esta forma se abre una ventana del navegador estándar del modelo individual. Cuando el nugget está en el lienzo, al pulsar dos veces en la miniatura de un gráfico se abre el gráfico a tamaño completo. Consulte el tema “Visor de modelos dividido” en la página 48 para obtener más información.

Una vez se ha creado un modelo como modelo dividido, no podrá eliminar el procesamiento de división, ni podrá deshacer la división posteriormente desde un nodo o nugget de modelado dividido.

**Ejemplo.** Una empresa nacional quiere realizar una estimación de ventas por categoría de producto en cada una de sus tiendas en el país. Mediante el modelado de divisiones, designan el campo Tienda de sus datos de entrada como campo de división, permitiendo crear modelos diferentes para cada categoría en cada tienda, en una sola operación. Posteriormente, podrán utilizar la información resultante para controlar los niveles de existencias de forma mucho más precisa que con un modelo simple.

## División y partición

La división tiene algunas características en común con la partición, pero se utilizan de formas muy diferentes.

**Partición** divide el conjunto de datos de forma aleatoria en dos o tres partes: formación, pruebas y (opcionalmente) validación, y se utiliza para comprobar el rendimiento de un modelo único.

**División** divide el conjunto de datos en tantas partes como valores posibles del campo de división y se utiliza para crear múltiples modelos.

La partición y división funcionan de manera completamente independiente entre sí. Puede seleccionar cualquiera de ellas, ambas o ninguna en un nodo de modelado.

## Nodos de modelado que admiten modelos divididos

Numerosos nodos de modelado pueden crear modelos divididos. Las excepciones son Clúster automático, Serie temporal, PCA/Factorial, Selección de características, SLRM, Árboles aleatorios, Árbol-AS, Lineal-AS, LSVM, los modelos de asociación (Apriori, Carma y Secuencia), los modelos de agrupación en clústeres (K-medias, Kohonen, Bietápico y Anomalía), los modelos Estadísticas y los nodos utilizados para el modelado interno de bases de datos.

Los nodos de modelado compatibles con el modelado de divisiones son:


Árbol C&R


Red bayesiana


Lineal


QUEST


GenLin


GLMM


CHAID


KNN


Serie temporal


C5.0


Cox


STP


Red neuronal


Clasificador automático


SVM de una clase


Lista de decisiones


Autonumérico


XGBoost Tree


Regresión


Logística


XGBoost Linear


Discriminante


SVM

## Características afectadas por la división

El uso de los modelos divididos afecta al número de características de IBM SPSS Modeler de varias formas. Esta sección proporciona orientación sobre cómo utilizar modelos de división con otros nodos en una ruta.

### Nodos Operaciones con registros

Cuando se utilizan modelos de división en una ruta que contiene un nodo Muestra, estratifique los registros a través del campo de división para conseguir un muestreo uniforme de registros. Esta opción está disponible cuando se selecciona *Complejo* como el método de muestra.

Si la ruta contiene un nodo Equilibrar, el equilibrado se aplica al conjunto global de registros de entrada, no al subconjunto de registros dentro de una división.

Al agregar registros mediante un nodo Agregar, establezca los campos de división para que sean campos clave, si desea calcular agregados para cada división.

### Nodos Operaciones con campos

El nodo Tipo es donde se especifica qué campo o campos utilizar como campos de división.

**Nota:** Mientras se utiliza el nodo Conjunto para combinar dos o más nuggets de modelo, no se puede utilizar para invertir la acción de la división, ya que los modelos de división están incluidos dentro de un nugget de modelo único.

### Nodos de modelado

Los modelos divididos no admiten el cálculo de importancia del predictor (la importancia relativa de los campos de entrada del predictor a la hora de calcular el modelo). Al crear modelos divididos, se ignora la configuración de importancia del predictor.

**Nota:** Los valores de puntuación de propensión ajustada se ignoran al utilizar un modelo de división.

El nodo KNN (vecino más próximo) soporta los modelos de división solo si está establecido para predecir un campo objetivo. El ajuste alternativo (identificar vecinos más cercanos únicamente) no crea un modelo. Si se ha elegido la opción **Seleccionar automáticamente k**, cada uno de los modelos de división podría tener un número diferente de vecinos más próximos. Además, el modelo global tiene un número de columnas generadas igual al mayor número de vecinos más próximos que se han encontrado en todos los modelos de división. Para estos modelos de división donde el número de vecinos más próximos es menor que este máximo, hay un número correspondiente de columnas llenadas con valores `$null`. Consulte el tema “Nodo KNN” en la página 361 para obtener más información.

### Nodos Modelado de bases de datos

Los nodos de modelado interno de bases de datos no admiten modelos divididos.

### Nuggets de modelo

No es posible **Exportar a PMML** desde un modelo dividido, ya que el nugget contiene múltiples modelos y PMML no admite ese empaquetado. Es posible exportar a texto o HTML.

## Opciones de los campos del nodo de modelado

Todos los nodos de modelado tienen una pestaña Campos en la que se pueden especificar los campos que se usarán para generar el modelo.

Para generar un modelo, primero se deben especificar los campos que se desea usar como objetivos y como entradas. Salvo algunas excepciones, todos los nodos de modelado usarán la información de los campos procedente de un nodo Tipo anterior en la ruta. Si utiliza un nodo Tipo para seleccionar campos de entrada y objetivo, no es necesario cambiar nada en esta pestaña. (Entre las excepciones se incluyen el nodo Secuencia y el nodo Extracción de texto, que requieren que la configuración del campo se especifique en el nodo de modelado.)

**Utilizar configuración del nodo Tipo.** Esta opción permite indicar al nodo que use la información de campo de un nodo Tipo situado en un punto anterior de la ruta. Este es el método predeterminado.

**Utilizar configuración personalizada.** Esta opción permite indicar al nodo que use la información de campo especificada aquí en lugar de la proporcionada en nodos Tipo situados en cualquier punto anterior de la ruta. Después de seleccionar esta opción, especifique los campos siguientes si es necesario.

*Nota: No todos los campos se visualizan para todos los nodos.*

- **Utilizar formato transaccional (Apriori, CARMA, Reglas de asociación MS y nodos Oracle Apriori únicamente).** Seleccione esta casilla de verificación si los datos de origen están en el **formato transaccional**. Los registros de este formato tienen dos campos, uno para una ID y otro para el contenido. Cada registro representa un único elemento o transacción y los elementos asociados se enlazan usando el mismo ID. Cancele esta selección si los datos están en **formato tabular**, en los que los elementos se representan por marcas separadas, donde cada campo de marca representa la presencia o ausencia de un elemento específico y cada registro representa un conjunto completo de elementos asociados. Consulte el tema "Datos tabulares frente a datos transaccionales" en la página 268 para obtener más información.
  - **ID.** Para los datos transaccionales, seleccione el campo de ID de la lista. Los campos numéricos o simbólicos se pueden utilizar como campo de ID. Cada valor exclusivo de este campo debe indicar una unidad de análisis específica. Por ejemplo, en una aplicación de la cesta de la compra, cada ID puede representar a un sólo cliente. Para una aplicación de análisis del registro Web, cada ID puede representar un equipo (con la dirección IP) o un usuario (con los datos de inicio de sesión).
  - **Los ID son contiguos.** (Nodos Apriori y CARMA únicamente) Si los datos se han clasificado previamente de forma que todos los registros con el mismo ID se agrupan en la ruta de datos, seleccione esta opción para que el procesamiento sea más rápido. Si los datos no se han clasificado previamente (o no lo sabe a ciencia cierta), no active esta opción y el nodo clasificará los datos automáticamente.

    *Nota*: si los datos no están clasificados y selecciona esta opción, es posible que obtenga resultados no válidos en el modelo.
  - **Contenido.** Especifique los campos de contenido del modelo. Estos campos contienen los elementos de interés del modelo de asociación. Se pueden especificar varios campos de marcas (si los datos están en formato tabular) o un sólo campo nominal (si los datos están en formato transaccional).
- **Objetivo.** En los modelos que requieran uno o varios campos objetivo, selecciónelos. Se trata de una acción similar a establecer el rol del campo en *Objetivo* en un nodo Tipo.
- **Evaluación.** (Para modelos de Autoclúster únicamente). No se ha especificado un objetivo para los modelos de clúster; sin embargo, puede seleccionar un campo de evaluación para identificar su nivel de importancia. Además, puede evaluar la calidad con la que los clústeres diferencian los valores de este campo, que a su vez indica si los clústeres se pueden utilizar para predecir este campo. *Nota* El campo de evaluación debe ser una cadena con más de un valor.
  - **Entradas.** Seleccione el campo(s) de entrada. Se trata de una acción similar a establecer el rol del campo en *Entrada* en un nodo Tipo.
  - **Partición.** Este campo permite especificar un campo usado para dividir los datos en muestras independientes para las fases de entrenamiento, prueba y validación en la generación del modelo. Si usa una muestra para generar el modelo y otra muestra distinta para comprobarlo, podrá obtener una buena indicación de la bondad del modelo a la hora de generar conjuntos de datos de mayor tamaño similares a los datos actuales. Si se han definido varios campos de partición mediante nodos Tipo o Partición, se deberá seleccionar un campo de partición simple en la pestaña Campos en todos

los nodos de modelado que usen la partición. (Si solamente hay una partición, se usará automáticamente siempre que se active la partición.) Debe tener en cuenta que al aplicar la partición seleccionada en su análisis, también debe activar la partición en la pestaña Opciones del modelo para el nodo. (Si se elimina la selección de esta opción, se posibilita la desactivación de la partición sin cambiar la configuración del campo.)

- **Divididos.** En modelos divididos, seleccione el campo o campos de división. Se trata de una acción similar a establecer el rol del campo en *Dividir* en un nodo Tipo. Sólo puede designar campos con un nivel de medición de **Marca**, **Nominal**, **Ordinal** o **Continuo** como campos de división. Los campos seleccionados como campos de división no se pueden utilizar como campos de destino, entrada, partición, frecuencia o ponderación. Consulte el tema "Generación de modelos divididos" en la página 28 para obtener más información.
- **Utilizar campo de frecuencia** Esta opción le permite seleccionar un campo como ponderación de frecuencia. Úsela si cada uno de los registros de sus datos de entrenamiento representan más de una unidad (por ejemplo, si está usando datos agregados). Los valores del campo deben ser el número de unidades representadas por cada registro. Consulte el tema "Uso de campos de frecuencia y ponderación" para obtener más información.

*Nota*: si ve el mensaje de error **Metadatos (en campos de entrada/salida) no válidos**, asegúrese de que ha especificado todos los campos necesarios, como el campo de frecuencia.

- **Utilizar campo de ponderación** Esta opción le permite seleccionar un campo como ponderación de casos. Las ponderaciones de casos se usan para contabilizar las diferencias existentes en la varianza entre los niveles del campo de salida. Consulte el tema "Uso de campos de frecuencia y ponderación" para obtener más información.
- **Consecuentes.** En el caso de nodos de reglas de inducción (Apriori), seleccione los campos que se deben usar como consecuentes en el conjunto de reglas resultante. (Se corresponde con los campos que tienen el rol *Objetivo* o *Ambas* de un nodo Tipo).
- **Antecedentes.** En el caso de nodos de reglas de inducción (Apriori), seleccione los campos que se deben usar como antecedentes en el conjunto de reglas resultante. (Se corresponde con los campos que tienen el rol tipo *Entrada* o *Ambas* de un nodo Tipo).

Algunos modelos presentan una pestaña denominada Campos que es diferente a lo descrito en esta sección.

- Consulte el tema "Opciones de campos para el nodo Secuencia" en la página 285 para obtener más información.
- Consulte el tema "Opciones de campos para el nodo CARMA" en la página 272 para obtener más información.

## Uso de campos de frecuencia y ponderación

Los campos de frecuencia y ponderación se utilizan para, por ejemplo, dar una importancia adicional a unos registros sobre otros, porque sabe que una sección de la población no está totalmente representada en los datos de entrenamiento (ponderación) o porque un registro representa un número de casos idénticos (frecuencia).

- los valores de un campo de frecuencia deben ser números enteros positivos. Los registros con una ponderación de frecuencias negativa o cero se excluyen del análisis. Las ponderaciones de frecuencias con valores no enteros se redondean al entero más cercano.
- Los valores de ponderación de casos deben ser positivos, pero no es necesario que sean enteros. Los registros con una ponderación de casos negativa o cero se excluyen del análisis.

Puntuación de campos de frecuencia y ponderación

Los campos de frecuencia y ponderación se utilizan en modelos de entrenamiento, pero no se utilizan en la puntuación porque la puntuación de cada registro se basa en sus características independientemente de cuántos casos represente. Por ejemplo, suponga que tiene los datos en la tabla siguiente.

*Tabla 1. Ejemplo de datos*

| **Casado** | **Respondido** |
|---|---|
| Sí | Sí |
| Sí | Sí |
| Sí | Sí |
| Sí | No |
| | |
| No | Sí |
| No | No |
| No | No |

Según esto, se llega a la conclusión de que tres de cada cuatro personas casadas responden a la promoción y dos de cada tres personas solteras no responden. Así se van a puntuar los nuevos registros en consecuencia, como se muestra en la siguiente tabla.

*Tabla 2. Ejemplo de registros puntuados*

| **Casado** | **$-Responded** | **$RP-Responded** |
|---|---|---|
| Sí | Sí | 0,75 (tres cuartos) |
| No | No | 0,67 (dos tercios) |

Como alternativa, puede almacenar los datos de entrenamiento de forma más compacta, utilizando un campo de frecuencia, tal y como se muestra en la tabla siguiente.

*Tabla 3. Ejemplo alternativo de registros puntuados*

| **Casado** | **Respondido** | **Frecuencia** |
|---|---|---|
| Sí | Sí | 3 |
| Sí | No | 1 |
| No | Sí | 1 |
| No | No | 2 |

Como esto representa exactamente el mismo conjunto de datos, creará el mismo modelo y predecirá respuestas basadas únicamente en el estado civil. Si tiene a diez personas casadas en sus datos de puntuación, predecirá *Sí* para cada una de ellas independientemente de si se presentan como diez registros separados o como uno con un valor de frecuencia de 10. La ponderación, aunque generalmente no es un número entero, se puede considerar que indica de igual modo la importancia de un registro. Éste es el motivo de por qué los campos de frecuencia y ponderación no se utilizan cuando se puntúan registros.

Evaluación y comparación de modelos

Algunos tipos de modelo admiten campos de frecuencia, algunos admiten campos de ponderación y otros admiten los dos. Sin embargo, en todos los casos en los que se aplican, solo se utilizan para la creación de modelos y no se tienen en cuenta cuando se evalúan modelos mediante los nodos Evaluación o Análisis o cuando se clasifican modelos mediante la mayoría de los métodos admitidos por los nodos Clasificador automático y Autonumérico.

- Al comparar modelos (por ejemplo, mediante diagramas de evaluación) se ignoran los valores de frecuencia y ponderación. Esto permite una comparación de nivel entre modelos que utilizan estos campos y modelos que no lo hacen, pero significa que, para una evaluación precisa, debe utilizarse un

conjunto de datos que represente la población de manera precisa sin depender de un campo de frecuencia o ponderación. En la práctica, puede hacerlo asegurándose de que los modelos se evalúan mediante una muestra de comprobación en la que el valor del campo de frecuencia o ponderación siempre sea nulo o 1. (Esta restricción solo se aplica al evaluar modelos; si los valores de frecuencia o ponderación siempre fueran 1 para las muestras de entrenamiento y comprobación, no habría necesidad de utilizar estos campos en primer lugar.)

- Si utiliza Clasificador automático, se puede tener en cuenta la frecuencia en caso de que se clasifiquen los modelos según Beneficio, de modo que este método se recomienda en ese caso.
- Si es necesario, puede dividir los datos en muestras de entrenamiento y comprobación utilizando el nodo Partición.

## Opciones de análisis del nodo de modelado

Muchos nodos de modelado incluyen la pestaña Analizar que le permite obtener información sobre la importancia de los predictores junto con puntuaciones de propensión ajustadas y en bruto.

Evaluación del modelo

**Calcular importancia del predictor.** En el caso de modelos que produzcan una medida adecuada de importancia, puede mostrar un gráfico que indique la importancia relativa de cada predictor al estimar el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Tenga en cuenta que puede tardarse más tiempo en calcular la importancia del predictor para algunos modelos, especialmente al trabajar con conjuntos de datos de gran tamaño; además, como resultado está desactivada para algunos modelos de forma predeterminada. La importancia del predictor no está disponible para modelos de listas de decisiones. Consulte el tema “Importancia del predictor” en la página 44 para obtener más información.

**Puntuaciones de propensión**

Las puntuaciones de propensión pueden activarse en el nodo de modelado y en la pestaña Configuración del nugget de modelo. Esta funcionalidad sólo está disponible cuando el objetivo seleccionado es un campo de marca. Consulte el tema “Puntuaciones de propensión” en la página 36 para obtener más información.

**Calcular puntuaciones de propensión en bruto.** Las puntuaciones de propensión en bruto están derivadas del modelo basado únicamente en los datos de entrenamiento. Si el modelo predice el valor *true* (responderá), la propensión es la misma que P, donde P es la probabilidad de la predicción. Si el modelo predice el valor false, la propensión se calcula como (1 - P).

- Si selecciona esta opción al crear el modelo, las puntuaciones de propensión se activarán en el nugget de modelo de forma predeterminada. Sin embargo, siempre puede activar las puntuaciones de propensión en bruto en el nugget de modelo independientemente de si las selecciona o no en el nodo de modelado.
- Al puntuar el modelo, se añadirán puntuaciones de propensión en bruto a un campo con las letras *RP* unidas al prefijo estándar. Por ejemplo, si las predicciones están en un campo denominado *$R-churn*, el nombre del campo de puntuación de propensión será *$RRP-churn*.

**Calcular puntuaciones de propensión ajustada.** Las propensiones brutas se basan totalmente en estimaciones proporcionadas por el modelo, las cuales pueden estar ajustadas excesivamente, lo que lleva a estimaciones de propensión demasiado optimistas. Las propensiones ajustadas intentan compensar este hecho observando el rendimiento del modelo en las particiones de comprobación o validación y ajustando las propensiones para proporcionar una mejor estimación en consecuencia.

- Esta configuración requiere que haya un campo de partición válido en la ruta.

- A diferencia de las puntuaciones brutas de confianza, las puntuaciones ajustadas de propensión deben calcularse al crear el modelo; de lo contrario, no estarán disponibles cuando se puntúe el nugget de modelo.
- Al puntuar el modelo, se añadirán puntuaciones ajustadas de propensión a un campo con las letras *AP* unidas al prefijo estándar. Por ejemplo, si las predicciones están en un campo denominado *$R-churn*, el nombre del campo de puntuación de propensión será *$RAP-churn*. Las puntuaciones ajustadas de propensión no están disponibles para modelos de regresión logística.
- Al calcular las puntuaciones ajustadas de propensión, la partición de comprobación o validación utilizada para el cálculo no debe haberse equilibrado. Para evitarlo, asegúrese de seleccionar la opción **Sólo datos de entrenamiento de equilibrado** en todos los nodos Equilibrar anteriores en la ruta. Además, si se ha llevado una muestra compleja a un punto anterior en la ruta, se invalidarán las puntuaciones ajustadas de propensión.
- Las puntuaciones ajustadas de propensión no están disponibles para modelos de árbol "aumentado" y de conjuntos de reglas. Consulte el tema “Modelos C5.0 aumentados” en la página 128 para obtener más información.

**Basado en.** Para que se calculen las puntuaciones ajustadas de propensión, debe haber un campo de partición en la ruta. Puede especificar si desea utilizar la partición de comprobación o validación para este cálculo. Para obtener los mejores resultados, la partición de comprobación o validación debe incluir al menos el mismo número de registros que la partición utilizada para entrenar el modelo original.

## Puntuaciones de propensión

En el caso de modelos que devuelven una predicción *sí* o *no*, puede solicitar puntuaciones de propensión además de los valores estándar de predicción y confianza. Las puntuaciones de propensión indican la verosimilitud de un resultado o respuesta específicos. La tabla siguiente contiene un ejemplo.

*Tabla 4. Puntuaciones de propensión*

| Cliente | Propensión de respuesta |
|---|---|
| Joe Smith | 35% |
| Jane Smith | 15% |

Las puntuaciones de propensión sólo están disponibles para modelos con objetivos de marca e indican la verosimilitud del valor *True* definido para el campo, como se especifica en un nodo de origen o nodo Tipo.

Puntuaciones de propensión frente a puntuaciones de confianza

Las puntuaciones de propensión se diferencian de las puntuaciones de confianza, que se aplican a la predicción actual, ya sea *sí* o *no*. En los casos en los que la predicción es *no*, por ejemplo una confianza elevada realmente significa una alta probabilidad de *no* en responder. Las puntuaciones de propensión eluden esta limitación para permitir una comparación más fácil entre todos los registros. Por ejemplo, una predicción *no* con una confianza de *0,85* se traduce en una propensión en bruto de *0,15* (o *1 menos 0,85*).

*Tabla 5. Puntuaciones de confianza*

| Cliente | Predicción | Confianza |
|---|---|---|
| Joe Smith | Responderá | 0,35 |
| Jane Smith | No responderá | 0,85 |

Obtención de puntuaciones de propensión

- Las puntuaciones de propensión pueden activarse en la pestaña Analizar del nodo de modelado o en la pestaña Configuración del nugget de modelo. Esta funcionalidad sólo está disponible cuando el

objetivo seleccionado es un campo de marca. Consulte el tema "Opciones de análisis del nodo de modelado" en la página 35 para obtener más información.

- El nodo Conjunto también puede calcular las puntuaciones de propensión, dependiendo del método de conjunto utilizado.

Cálculo de puntuaciones ajustadas de propensión

Las puntuaciones ajustadas de propensión se calculan como parte del proceso de creación del modelo y no estarán disponibles de otro modo. Una vez creado el modelo, se puntúa utilizando datos de la partición de comprobación o validación y se genera un nuevo modelo que proporcione puntuaciones ajustadas de propensión analizando el rendimiento del modelo original en dicha partición. Dependiendo del tipo de modelo, se puede utilizar uno de los dos métodos existentes para calcular las puntuaciones ajustadas de propensión.

- En el caso de modelos de conjuntos de reglas y de árbol, las puntuaciones ajustadas de propensión se generan volviendo a calcular la frecuencia de cada categoría en cada nodo de árbol (en modelos de árbol) o el soporte y la confianza de cada regla (en modelos de conjuntos de reglas). Esto da como resultado un nuevo modelo de conjuntos de reglas o de árbol que se almacena con el modelo original para su uso cuando sean necesarias las puntuaciones ajustadas de propensión. Cada vez que el modelo original se aplica a nuevos datos, el nuevo modelo puede aplicarse posteriormente a las puntuaciones de propensión en bruto para generar las puntuaciones ajustadas.
- En el caso de otros modelos, los registros producidos al puntuar el modelo original en la partición de comprobación o validación se establecen en intervalos por su puntuación de propensión en bruto. A continuación, se entrena un modelo de red neuronal que define una función no lineal que establece correlaciones entre la propensión en bruto media de cada intervalo y la propensión media observada del mismo intervalo. Como se ha indicado previamente en el caso de modelos de árbol, el modelo de red neuronal resultante se almacena con el modelo original y puede aplicarse a las puntuaciones de propensión en bruto cuando sean necesarias las puntuaciones ajustadas de propensión.

**Preste atención a los valores perdidos en la partición de comprobación.** El tratamiento de los valores de entrada perdidos en la partición de comprobación/validación varía según el modelo (consulte los algoritmos de puntuación de modelos individuales si desea información detallada al respecto). El modelo C5 no puede calcular propensiones ajustadas cuando faltan datos de entrada.

## Costes de clasificación errónea

En algunos contextos, ciertos tipos de errores son más costosos que otros. Por ejemplo, puede resultar más costoso clasificar a un solicitante de crédito de alto riesgo como de bajo riesgo (un tipo de error) que clasificar a un solicitante de crédito de bajo riesgo como de alto riesgo (otro tipo de error). Los costes de clasificación errónea permiten especificar la importancia relativa de los diversos tipos de errores de predicción.

Los costes de clasificación errónea son básicamente ponderaciones aplicadas a resultados específicos. Estas ponderaciones se extraen en el modelo y pueden realmente cambiar la predicción (como forma de protección frente a errores costosos).

Salvo los modelos C5.0, los costes de clasificación errónea no se aplican cuando se puntúa un modelo y no se tienen en cuenta cuando se clasifican o comparan modelos utilizando un nodo Clasificador automático, un diagrama de evaluación o un nodo Análisis. Es posible que un modelo que incluya costes no produzca menos errores que uno que no lo haga, y es posible que no ordene ningún valor mayor en términos de precisión global, pero es probable que funcione mejor en términos prácticos porque contiene un sesgo integrado en favor de errores *más baratos*.

La matriz de costes muestra el coste para cada combinación posible de categoría predicha y categoría real. De forma predeterminada, todos los costes de clasificación errónea se establecen en 1,0. Para

introducir valores de coste personalizados, seleccione **Utilizar costes de clasificación errónea** e introduzca valores personalizados en la matriz de costes.

Para cambiar un coste de clasificación errónea, seleccione la casilla correspondiente a la combinación deseada de valores predichos y reales, elimine el contenido existente de la casilla e introduzca en ella el coste deseado. Los costes no son simétricos automáticamente. Por ejemplo, si establece el coste de clasificación errónea *A* como *B* para que sea 2,0, el coste de clasificación errónea de *B* como *A* aún tendrá el valor predeterminado 1,0 hasta que también se modifique explícitamente.

**Nota:** Únicamente los modelos de árboles de decisión permiten especificar los costes durante la generación.

## Nuggets de modelo

*Figura 19. Nugget de modelo*

Un nugget de modelo es el recipiente de un modelo, es decir, es el conjunto de reglas, fórmulas o ecuaciones que representan los resultados de las operaciones de generación de modelos en SPSS Modeler. La finalidad principal de un nugget es puntuar datos para generar predicciones o permitir análisis adicionales de propiedades de modelos. Al abrir un nugget de modelo en la pantalla, podrá ver diversos datos del modelo, como la importancia relativa de los campos de entrada en la creación del modelo. Para ver las predicciones, necesitará adjuntar y ejecutar otro nodo proceso o de resultado. Consulte el tema “Uso de nugget de modelo en rutas” en la página 49 para obtener más información.


*Figura 20. Enlace de modelo del nodo de modelado al nugget de modelo*

Cuando se ejecuta satisfactoriamente un nodo de modelado, se coloca el nugget de modelo correspondiente en el lienzo de rutas, representado por un icono dorado en forma de diamante (de ahí el nombre de "nugget", que significa pepita de oro). En el lienzo de rutas, el nugget se muestra con una conexión (línea continua) al nodo adecuado más cercano previo al nodo de modelado, y con un enlace (línea discontinua) al nodo de modelado en sí.

El nugget también se coloca en la paleta de modelos de la esquina superior derecha de la ventana de IBM SPSS Modeler. Desde cualquiera de las ubicaciones, se pueden seleccionar y explorar los nuggets para ver los detalles del modelo.

Siempre se colocan nuggets en la paleta de modelos cuando se ejecuta correctamente un nodo de modelado. Puede establecer una opción de usuario para controlar si el nugget se coloca, además, en el lienzo de rutas.

Los siguientes temas proporcionan información acerca del uso de nuggets de modelo en IBM SPSS Modeler. Para obtener una mejor comprensión de los algoritmos utilizados, consulte la *Guía de algoritmos de IBM SPSS Modeler*, disponible como un archivo PDF como parte de la descarga del producto.

# Enlaces de modelo

De forma predeterminada, se muestra un nugget en el lienzo con un enlace al nodo de modelado que lo creó. Esto resulta especialmente útil en rutas complejas con varios nuggets, al permitir identificar el nugget que cada nodo de modelado actualizará. Cada enlace contiene un símbolo que indica si el modelo se sustituirá o no cuando se ejecute el nodo de modelado. Consulte el tema "Sustitución de un modelo" en la página 40 para obtener más información.

## Definición y eliminación de enlaces de modelo

Puede definir y eliminar enlaces manualmente en el lienzo. Cuando defina un enlace nuevo, el cursor cambiará al cursor de enlaces.

*Figura 21. Cursor de enlaces*

Definición de un nuevo enlace (menú contextual)

1. Pulse con el botón derecho en el nodo de modelado desde el que desea que empiece el enlace.
2. Elija **Definir enlace de modelo** en el menú contextual.
3. Pulse en el nugget en el que desea que acabe el enlace.

Definición de un nuevo enlace (menú principal)

1. Pulse en el nodo de modelado desde el que desea que empiece el enlace.
2. En el menú principal, elija:
   **Editar** > **Nodo** > **Definir enlace de modelo**
3. Pulse en el nugget en el que desea que acabe el enlace.

Eliminación de un enlace existente (menú contextual)

1. Pulse con el botón derecho en el nugget al final del enlace.
2. Elija **Eliminar enlace de modelo** en el menú contextual.

Asimismo:

1. Pulse con el botón derecho en el símbolo que aparece en mitad del enlace.
2. Elija **Eliminar enlace** en el menú contextual.

Eliminación de un enlace existente (menú principal)

1. Pulse en el nodo de modelado del que desea eliminar el enlace.
2. En el menú principal, elija:
   **Editar** > **Nodo** > **Eliminar enlace de modelo**

## Copiar y pegar enlaces de modelo

Si copia un nugget enlazado sin su nodo de modelado y lo pega en la misma ruta, éste se pegará con un enlace al nodo de modelado. El nuevo enlace tiene el mismo estado de sustitución de modelo (consulte "Sustitución de un modelo" en la página 40) que el enlace original.

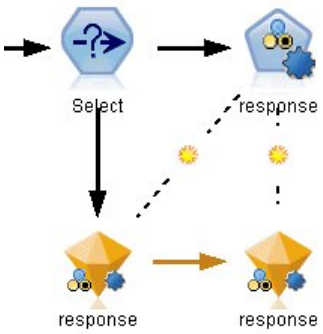



*Figura 22. Copiar y pegar un nugget enlazado*

Si copia y pega un nugget junto con su nodo de modelado enlazado, el enlace se mantiene tanto si los objetos se pegan en la misma ruta como si lo hacen en una ruta nueva.

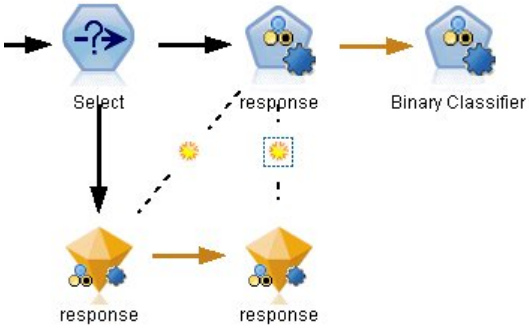



*Figura 23. Copiar y pegar un nugget enlazado*

*Nota*: si copia un nugget enlazado sin su nodo de modelado y lo pega en una nueva ruta (o en un Supernodo que no contenga el nodo de modelado), el enlace se rompe y solo se pega el nugget.

### Enlaces de modelo y Supernodos

Si define un Supernodo para que incluya el nodo de modelado o el nugget de modelo de un modelo enlazado (no ambos), se romperá el enlace. El enlace no se restaurará al expandir el Supernodo; solo podrá conseguirlo deshaciendo la creación del Supernodo.

## Sustitución de un modelo

Puede elegir si sustituir (es decir, actualizar) o no un nugget existente al volver a ejecutar el nodo de modelado que lo creó. Si desactiva la opción de sustitución, se creará un nuevo nugget cuando se ejecute de nuevo el nodo de modelado.

Cada enlace de un nodo de modelado a un nugget contiene un símbolo que indica si el modelo se sustituirá o no cuando se ejecute nuevamente el nodo de modelado.


*Figura 24. Enlace de modelo con sustitución de modelo activada*

El enlace se muestra inicialmente con la sustitución de modelo activada, representada por un pequeño símbolo en forma de resplandor solar. En este estado, al volver a ejecutar el nodo de modelado en un extremo del enlace se actualiza el nugget en el extremo contrario.


*Figura 25. Enlace de modelo con sustitución de modelo desactivada*

Si se desactiva la sustitución de modelo, se sustituye el símbolo de enlace por un punto gris. En este estado, al volver a ejecutar el nodo de modelado en un extremo del enlace, se añade una versión nueva y actualizada del nugget al lienzo.

En cualquiera de los casos, en la paleta de modelos se actualiza el nugget existente o se crea uno nuevo, según la configuración de la opción**Sustituir modelo anterior** del sistema.

### Orden de ejecución

Al ejecutar una ruta con múltiples ramas que contengan nuggets de modelo, la ruta se evalúa primero para asegurar que una rama con sustitución de modelos activada se ejecute antes que cualquier rama que use el nugget de modelo resultante.

Si sus requisitos son más complejos, puede definir el orden de ejecución manualmente mediante scripts.

### Modificación del valor de sustitución del modelo

1. Pulse con el botón derecho del ratón en el símbolo del enlace.
2. Elija **Activar (o Desactivar) sustitución de modelo,** según desee.

**Nota:** El valor de sustitución del modelo en un enlace de modelo altera temporalmente el valor en la pestaña Notificaciones del diálogo Opciones de usuario (Herramientas > Opciones > Opciones de usuario).

## La paleta de modelos

La paleta de modelos (en la pestaña Modelos de la ventana Gestores) le permite utilizar, examinar y modificar los nuggets de modelo de distintas maneras.

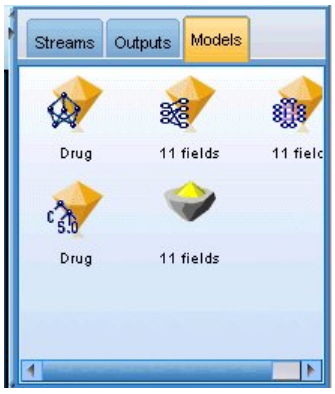



*Figura 26. Paleta de modelos*

Al pulsar con el botón derecho en un nugget de la paleta de modelos, se abre un menú contextual con las siguientes opciones:

- **Añadir a ruta.** Añade el nugget de modelo a la ruta activa actualmente. Si en la ruta hay un nodo seleccionado, el nugget de modelo se conectará al nodo seleccionado siempre que dicha conexión sea factible o, de lo contrario, al nodo más cercano posible. El nugget se muestra con un enlace al nodo de modelado que creó el modelo, si éste permanece en la ruta.
- **Examinar.** Abre el explorador de modelos del nugget.
- **Cambiar nombre y anotar.** Permite cambiar el nombre del nugget de modelo y/o modificar la anotación del mismo.
- **Generar nodo de modelado** Si tiene un nugget de modelo que desea modificar o actualizar y no está disponible la ruta que se utilizó para crear el modelo, puede usar esta opción para volver a generar el nodo de modelado con las mismas opciones que empleó para crear el modelo original.
- **Guardar modelo, Guardar modelo como.** Guarda el nugget de modelo en un archivo binario del modelo generado externo (.gm).
- **Almacenar modelo.** Guarda el nugget de modelo en Repositorio de IBM SPSS Collaboration and Deployment Services.
- **Exportar PMML.** Exporta el nugget de modelo como lenguaje de códigos para modelos predictivos (PMML), que se puede utilizar para puntuar nuevos datos fuera de IBM SPSS Modeler. **Exportar PMML** está disponible para todos los nodos de modelo generados.
- **Añadir al proyecto.** Guarda el nugget de modelo y lo añade al proyecto actual. En la pestaña Clases se añadirá el nugget a la carpeta Modelos generados. En la pestaña CRISP-DM se añadirá a la fase del proyecto predeterminada.
- **Eliminar.** Elimina el nugget de modelo de la paleta.

Al pulsar con el botón derecho en un área vacía de la paleta de modelos, se abre un menú contextual con las siguientes opciones:

- **Abrir modelo.** Carga un nugget de modelo creado anteriormente en IBM SPSS Modeler.
- **Recuperar modelo.** Recupera un modelo almacenado en un repositorio de IBM SPSS Collaboration and Deployment Services.
- **Cargar paleta.** Carga una paleta de modelos guardada en un archivo externo.
- **Recuperar paleta.** Recupera una paleta de modelos almacenada en un repositorio de IBM SPSS Collaboration and Deployment Services.
- **Guardar paleta.** Guarda todo el contenido de la paleta de modelos en un archivo externo de paleta de modelos (.gen).
- **Almacenar paleta.** Almacena todo el contenido de la paleta de modelos en un repositorio de IBM SPSS Collaboration and Deployment Services.
- **Borrar paleta.** Elimina todos los nugget de la paleta.
- **Añadir paleta al proyecto.** Guarda la paleta de modelos y la añade al proyecto actual. En la pestaña Clases se añadirá el nugget a la carpeta Modelos generados. En la pestaña CRISP-DM se añadirá a la fase del proyecto predeterminada.
- **Importar PMML.** Carga un modelo desde un archivo externo. Puede abrir, explorar y puntuar modelos PMML creados por IBM SPSS Statistics o por otras aplicaciones que admitan este formato. Consulte el tema “Cómo importar y exportar modelos como PMML” en la página 50 para obtener más información.

## Examen de nuggets de modelo

Los exploradores de nugget de modelo le permiten examinar y utilizar los resultados de los modelos. Desde el explorador se puede guardar, imprimir o exportar el modelo generado, examinar el resumen del modelo y ver o editar sus anotaciones. En algunos tipos de nugget de modelo también puede generar nuevos nodos, como nodos Filtrar o nodos de conjunto de reglas. En el caso de algunos modelos también

puede ver los parámetros del modelo, como las reglas o los centros de los clústeres. En algunos tipos de modelos (los modelos basados en árboles y los modelos de clúster) se puede mostrar una representación gráfica de la estructura del modelo. A continuación se describen los controles de uso de los exploradores de nugget de modelo.

## Menús

**Menú Archivo.** Todos los nuggets de modelo tienen un menú Archivo algunos subconjuntos con las siguientes opciones:

- **Guardar nodo.** Guarda el nugget de modelo en un archivo de nodo (`.nod`).
- **Almacenar nodo.** Almacena el nugget de modelo en un repositorio de IBM SPSS Collaboration and Deployment Services.
- **Cabecera y pie de página.** Permite editar la cabecera y el pie de página para la impresión desde el nugget.
- **Configurar página.** Permite cambiar la configuración de página para la impresión desde el nugget.
- **Presentación preliminar.** Muestra una presentación preliminar de la impresión del nugget. Desde el submenú, seleccione la información que desee mostrar en la presentación preliminar.
- **Imprimir.** Imprime el contenido del nugget. Desde el submenú, seleccione la información que desee imprimir.
- **Imprimir vista.** Imprime la vista actual o todas las vistas.
- **Exportar texto.** Exporta el contenido del nugget a un archivo de texto. Desde el submenú, seleccione la información que desee exportar.
- **Exportar HTML.** Exporta el contenido del nugget a un archivo HTML. Desde el submenú, seleccione la información que desee exportar.
- **Exportar PMML.** Exporta el modelo como lenguaje de códigos para modelos predictivos (PMML), que se puede utilizar con otro software compatible con PMML. Consulte el tema "Cómo importar y exportar modelos como PMML" en la página 50 para obtener más información.
- **Exportar SQL.** Exporta el modelo como lenguaje de consulta estructurado (SQL), que puede modificarse y utilizarse con otras bases de datos.

  **Nota:** La exportación SQL solo está disponible en los modelos siguientes: modelos C5, C&RT, CHAID, QUEST, Regresión lineal, Regresión logística, Red neuronal, PCA/Factor y Lista de decisiones.
- **Publicar en el adaptador de puntuación del servidor.** Publica el modelo en una base de datos con un adaptador de puntuación instalado, permitiendo que la puntuación de modelos tenga lugar dentro de la base de datos. Consulte el tema "Publicar modelos para un adaptador de puntuación" en la página 52 para obtener más información.

**Menú Generar.** La mayoría de los nugget de modelo también tienen un menú Generar, que permite generar nodos nuevos basados en el nugget de modelo. Las opciones disponibles de este menú variarán en función del tipo de modelo que se está examinando. Si desea obtener información más detallada sobre lo que se puede generar a partir de un determinado modelo, consulte el tipo específico de nugget de modelo.

**Menú Ver.** En la pestaña Modelo de un nugget, este menú permite mostrar u ocultar las diferentes barras de herramientas de visualización disponibles en el modo actual. Para que todas las barras de herramientas estén disponibles, seleccione Modo edición (en el icono de la brocha) de la barra de herramientas General.

**Botón Presentación preliminar.** Algunos nuggets de modelo tienen un botón Presentación preliminar, que permite ver una muestra de los datos del modelo, incluyendo los campos extra creados en el proceso de modelado. El número predeterminado de filas visualizadas es 10; sin embargo, puede cambiarlo en las propiedades de la ruta.

**Botón Añadir al proyecto actual.** Guarda el nugget de modelo y lo añade al proyecto actual. En la pestaña Clases se añadirá el nugget a la carpeta Modelos generados. En la pestaña CRISP-DM se añadirá a la fase del proyecto predeterminada.

## Información / Resumen de nugget de modelo

La pestaña Resumen o la vista Información de un nugget de modelo muestra información sobre los campos, la configuración de creación y el proceso de estimación del modelo. Los resultados se muestran en una vista de árbol que se puede expandir o contraer pulsando los elementos específicos.

**Análisis.** Muestra información sobre el modelo. Los detalles específicos varían en función del tipo de modelo y se tratan en la sección de cada nugget de modelo. Además, si ejecuta un nodo Análisis que está conectado a este nodo de modelado, la información de dicho análisis también se visualiza en esta sección.

**Campos.** Lista los campos que se utilizan como objetivo y las entradas en la generación del modelo. En modelos divididos, enumera los campos que determinan las divisiones.

**Nota:** En la vista Información para los modelos de redes neuronales, lineales y otros modelos con las modalidades de aumento o agregación indistintamente, el icono que se muestra es el mismo (icono nominal), independientemente de si el tipo es de marca, nominal u ordinal.

**Opciones / Configuración de creación.** Contiene información sobre los valores que se utilizan en la generación del modelo.

**Resumen de entrenamiento.** Muestra el tipo de modelo, la ruta que se utiliza para crearlo, el usuario que lo ha creado. cuándo se ha creado y el tiempo transcurrido para la generación del modelo. Tenga en cuenta que el tiempo transcurrido para la creación del modelo sólo está disponible en la pestaña Resumen, no en la vista Información.

## Importancia del predictor

Es normal centrar los esfuerzos de modelado en los campos predictores más importantes y valorar la omisión de aquellos con menor relevancia. El gráfico de importancia de los predictores le ayuda a hacerlo indicando la importancia relativa de cada predictor en la estimación del modelo. Como los valores son relativos, la suma de valores de todos los predictores de la visualización es 1.0. La importancia del predictor no está relacionada con la precisión del modelo. Sólo está relacionada con la importancia de cada predictor a la hora de realizar una predicción, no con si la predicción es o no precisa.

Importancia de predictor está disponible para modelos que producen una medida estadística adecuada de importancia, incluidas las redes neuronales, árboles de decisión (árbol C&R, C5.0, CHAID y QUEST), redes bayesianas, modelos discriminantes, SVM y SLRM, regresión lineal y logística, modelos lineales generalizados y de vecinos más próximos (KNN). Para la mayoría de estos modelos, la importancia de predictor puede activarse en la pestaña Analizar del nodo de modelado. Consulte el tema “Opciones de análisis del nodo de modelado” en la página 35 para obtener más información. Para los modelos KNN, consulte “Vecinos” en la página 363.

*Nota*: No se admite la importancia del predictor para modelos divididos. Al crear modelos divididos, se ignora la configuración de importancia del predictor. Consulte el tema “Generación de modelos divididos” en la página 28 para obtener más información.

El cálculo de importancia de predictor puede tardar significativamente más tiempo que la generación de modelos, especialmente al utilizar conjuntos de datos de gran tamaño. Se tarda más en calcular los modelos SVM y de regresión logística que otros modelos; de forma predeterminada, está desactivado

para estos modelos. Si utiliza un conjunto de datos con un gran número de predictores, un cribado inicial mediante un nodo de selección de característica puede proporcionar resultados más rápidamente (véase a continuación).

- La importancia del predictor se calcula a partir de la partición de comprobación si está disponible. De lo contrario se utilizan los datos de entrenamiento.
- En el caso de los modelos SLRM, la importancia del predictor está disponible, pero se calcula mediante el algoritmo SLRM. Consulte el tema "Nugget de modelo SLRM" en la página 348 para obtener más información.
- Puede utilizar las herramientas de gráfico de IBM SPSS Modeler para interactuar con el gráfico, así como editarlo y guardarlo.
- También puede generar el nodo Filtrar basado en la información del gráfico Importancia de predictor. Consulte el tema "Filtrado de variables basado en su importancia" para obtener más información.

Importancia de predictor y selección de características

Puede parecer que el gráfico Importancia de predictor que aparece en un nugget de modelo ofrece resultados parecidos a los del nodo Selección de características en algunos casos. Pero mientras Selección de características clasifica cada campo de entrada según la fuerza de su relación con el objetivo específico, independientemente de otras entradas, el gráfico Importancia de predictor indica la importancia relativa de cada entrada para *este* modelo en particular. Por lo tanto, Selección de características será más conservador al cribar entradas. Por ejemplo, si tanto *job title (puesto de trabajo)* como *job category (categoría laboral)* tienen una relación muy estrecha con el salario, Selección de características indicará que los dos son importantes. Sin embargo, en modelado, las interacciones y correlaciones también se tienen en cuenta. Por lo tanto, puede que descubra que solo se utilizan una o dos entradas si ambas duplican gran parte de la misma información. En la práctica, Selección de características es de gran utilidad para el cribado preliminar, especialmente cuando se trabaja con conjuntos de datos de gran tamaño con un gran número de variables; Importancia de predictor es de mayor utilidad en el ajuste con precisión del modelo.

Diferencias de importancia de predictor entre modelos únicos y nodos de modelado automatizado

En función de si está creando un modelo único desde un nodo individual, o si utiliza un nodo de modelado automatizado para generar resultados, es posible que vea ligeras diferencias en la importancia del predictor. Dichas diferencias en la implementación se deben a algunas restricciones de ingeniería.

Por ejemplo, con clasificadores únicos como, por ejemplo, CHAID el cálculo aplica una regla de parada y utiliza valores de probabilidad al calcular valores de importancia. A diferencia de esto, el clasificador automático no utiliza una regla de parada y utiliza etiquetas pronosticadas en el cálculo. Estas diferencias pueden significar que si genera un modelo único mediante el clasificador automático, el valor de importancia se puede considerar como una estimación aproximada, en comparación con lo calculado para un clasificador único. Para obtener los valores de importancia de predictor más precisos, se sugiere utilizar un único nodo, en lugar de los nodos de modelado automatizado.

## Filtrado de variables basado en su importancia

También puede generar el nodo Filtrar basado en la información del gráfico Importancia de predictor.

Marque los predictores que desee incluir en el gráfico si procede y seleccione en los menús:

**Generar** > **Nodo Filtrar (Importancia de predictor)**

O

> **Selección de campos (importancia del predictor)**

**Número especificado de variables.** Incluye o excluye los predictores más importantes hasta el número especificado.

**Importancia mayor que.** Incluye o excluye todos los predictores con una importancia relativa superior al valor especificado.

# Visor de conjuntos

## Modelos de conjuntos

El modelo de un conjunto ofrece información sobre los modelos de componente en el conjunto y el rendimiento del conjunto como un todo.

La barra de herramientas principal (que no depende de la vista) le permite seleccionar si desea usar el conjunto o un modelo de referencia para la puntuación. Si el conjunto se utiliza para la puntuación también puede seleccionar la regla de combinación. Estos cambios no requieren una segunda ejecución del modelo, sin embargo estas elecciones se guardan en el (nugget) de modelo para la puntuación y la evaluación del modelo posterior. También afectan al PMML exportado desde el visor de conjuntos.

**Reglas de combinación.** Al puntuar un conjunto, ésta es la regla utilizada para combinar los valores predichos a partir de los modelos básicos para calcular el valor de puntuación del conjunto.

- Los valores predichos de conjuntos de destinos **categóricos** pueden combinarse mediante votación, mayor probabilidad o mayor probabilidad media. **Votación** selecciona la categoría que tenga la mayor probabilidad más frecuentemente entre los modelos básicos. **La mayor probabilidad** selecciona la categoría que logra la mayor probabilidad individual entre todos los modelos básicos. **Mayor probabilidad media** selecciona la categoría con el valor más elevado cuando se calcula la media de las probabilidades de categoría entre los modelos básicos.
- Los valores pronosticados de conjunto para objetivos **continuos** pueden combinarse mediante la media o mediana de los valores pronosticados a partir de los modelos básicos.

El valor predeterminado se toma de las especificaciones realizadas durante la generación de modelos. Al cambiar la regla de combinación vuelve a calcularse la precisión del modelo y se actualizan todas las vistas de la precisión del modelo. El gráfico Importancia de predictor también se actualiza. Este control se desactiva si se selecciona el modelo de referencia para la puntuación.

**Mostrar todas las reglas de combinación.** Cuando se selecciona esta opción, los resultados de todas las reglas de combinación disponibles se muestran en el gráfico de calidad de modelos. El gráfico Precisión de modelo de componente también se actualiza para mostrar las líneas de referencia de cada método de votación.

**Resumen del modelo:** La vista Resumen de modelos es una instantánea, un resumen de un vistazo de la calidad y la diversidad de los conjuntos.

**Calidad.** El gráfico muestra la precisión del modelo final, en comparación con un modelo de referencia y un modelo naive. La precisión se presenta en un formato mientras más grande mejor: siendo el mejor modelo el que tendrá mayor precisión. Para un destino categórico, la precisión es simplemente el porcentaje de registros para los que el valor predicho concuerda con el observado. En el caso de un destino continuo, la precisión es 1 menos la relación entre el error absoluto promedio de la predicción (la media de los valores absolutos de los valores predichos menos los valores observados) y el rango de valores predichos (el valor predicho máximo menos el valor predicho mínimo).

Para empaquetar conjuntos, el modelo de referencia es un modelo estándar construido en la partición de entrenamiento al completo. Para los conjuntos potenciados, el modelo de referencia es el primer modelo de componente.

El modelo naive representa la precisión si no se construyó ningún modelo, y asigna todos los registros a la categoría modal. El modelo naive no se calcula para los destinos continuos.

**Diversidad.** El gráfico muestra la "diversidad de opiniones" entre los modelos de componente usados para construir el conjunto, presentados en un formato mayor y más diverso. Es una medida de cómo varían las predicciones entre los modelos básicos. La diversidad no está disponible para los modelos de conjuntos potenciados, ni aparece para los destinos continuos.

**Importancia del predictor:** Es normal centrar los esfuerzos de modelado en los campos predictores más importantes y valorar la omisión de aquellos con menor relevancia. El gráfico de importancia de los predictores le ayuda a hacerlo indicando la importancia relativa de cada predictor en la estimación del modelo. Como los valores son relativos, la suma de valores de todos los predictores de la visualización es 1.0. La importancia del predictor no está relacionada con la precisión del modelo. Sólo está relacionada con la importancia de cada predictor a la hora de realizar una predicción, no con si la predicción es o no precisa.

La importancia del predictor no se encuentra disponible para modelos de conjuntos. El conjunto de predictores puede variar entre modelos de componente, pero puede calcularse la importancia para los predictores usados en al menos un modelo de componente.

**Frecuencia de predictor:** El conjunto de predictores puede variar en distintos modelos de componente por la elección del método de modelado o la selección de predictores. El gráfico Frecuencia de predictor es un gráfico de puntos que muestra la distribución de predictores entre los modelos de componente del conjunto. Cada punto representa uno o más modelos de componente que contienen el predictor. Los predictores se representan en el eje y, y se ordenan en orden descendente de frecuencia, de forma que el predictor más alto es el que se usa en el mayor número de modelos de componente y el más bajo el que se utiliza en menos. Se muestran los 10 predictores principales.

Los predictores que aparecen más frecuentemente suelen ser los más importantes. Este gráfico no es útil para métodos en los que el conjunto de predictores varían entre los modelos de componente.

**Precisión de modelo de componente:** El gráfico es un gráfico de puntos de precisión predictiva para los modelos de componente. Cada punto representa uno o más modelos de componente con el nivel de precisión representado en el eje y. Pase el ratón sobre cualquier punto para obtener información sobre el modelo de componente individual correspondiente.

**Líneas de referencia.** El gráfico muestra líneas codificadas de color para el conjunto así como el modelo de referencia y los modelos naïve. Aparece una marca de selección junto a la línea correspondiente al modelo que se usará para la puntuación.

**Interactividad.** El gráfico se actualiza si cambia la regla de combinación.

**Conjuntos potenciados.** Se muestra un gráfico de líneas para los conjuntos potenciados.

**Detalles de modelo de componente:** La tabla muestra información sobre los modelos de componente, enumerados por fila. De forma predeterminada, los modelos de componente se orden en orden de número de modelo ascendente. Puede ordenar las filas en orden ascendente o descendente según los valores de cualquier columna.

**Modelo.** Número que representa el orden secuencial en el que se creó el modelo de componente.

**Precisión.** Precisión general con formato de porcentaje.

**Método.** Método de modelado.

**Predictores.** Número de predictores utilizados en el modelo de componente.

**Tamaño de modelo.** El tamaño de modelo depende del método de modelado: en los árboles, se trata del número de nodos del árbol; en los modelos lineales, corresponde al número de coeficientes; en las redes neuronales, es el número de sinapsis.

**Registros.** Número ponderado de registros de entrada de los datos de entrenamiento.

**Preparación de datos automática:**

Esta vista muestra información a cerca de qué campos se excluyen y cómo los campos transformados se derivaron en el paso de preparación automática de datos (ADP). Para cada campo que fue transformado o excluido, la tabla enumera el nombre del campo, su rol en el análisis y la acción tomada por el paso ADP. Los campos se clasifican por orden alfabético ascendente de nombres de campo.

La acción **Valores atípicos de recorte**, si aparece, indica que se han establecido valores de predictores continuos que se encuentran más allá de un valor de corte (3 desviaciones estándar de la media) para el valor de corte.

# Nuggets de modelo de modelos divididos

El nugget de modelo de un modelo dividido proporciona acceso a todos los modelos individuales que crean las divisiones.

Un nugget de modelo dividido contiene:

- una lista de todos los modelos divididos creados, junto con un conjunto de estadísticas de cada modelo
- información acerca del modelo total

En la lista de modelos divididos, puede abrir modelos individuales para examinarlos posteriormente.

## Visor de modelos dividido

La pestaña Modelo enumera todos los modelos del nugget y proporciona todas las estadísticas en diferentes formatos sobre los modelos divididos. Tiene dos formas generales, en función del nodo de modelado.

**Ordenar por**. Utilice esta lista para seleccionar el orden en que se mostrarán los modelos. Puede ordenar la lista según los valores de cualquiera de las columnas de visualización, en orden ascendente o descendente. También puede pulsar en una cabecera de columna para ordenar la lista por esa columna. El valor predeterminado es el orden descendente de precisión global.

**Mostrar/ocultar menú columnas**. Pulse este botón para ver un menú donde podrá seleccionar si ver u ocultar columnas individuales.

**Ver**. Si utiliza la partición, puede seleccionar visualizar los resultados por los datos de entrenamiento o los datos de comprobación.

En cada división se muestran los detalles de la siguiente forma:

**Gráfico**. Una miniatura indica la distribución de los datos del modelo. Cuando el nugget está en el lienzo, pulse dos veces en la miniatura para abrir el gráfico a tamaño completo.

**Modelo**. Un icono del tipo de modelo. Pulse dos veces en el icono para abrir el nugget de la división concreta.

**Campos de división**. Los campos designados en el nodo de modelado como campos de división, con sus posibles valores diferentes.

**Número de registros en la división**. El número de registros de la división concreta.

**Número de campos utilizados**. Clasifica los modelos divididos en función del número de campos de entrada utilizados.

**Precisión global (%)**. Porcentaje de registros predichos correctamente por el modelo dividido respecto al número total de registros en esa división.

**Dividir**. La cabecera de la columna muestra los campos usados para crear las divisiones, y las casillas son los valores de división. Pulse dos veces en cualquier segmentación para abrir un visor de modelos para el modelo construido para esa segmentación.

**Precisión.** Precisión general con formato de porcentaje.

**Tamaño de modelo.** El tamaño de modelo depende del método de modelado: en los árboles, se trata del número de nodos del árbol; en los modelos lineales, corresponde al número de coeficientes; en las redes neuronales, es el número de sinapsis.

**Registros.** Número ponderado de registros de entrada de los datos de entrenamiento.

## Uso de nugget de modelo en rutas

Los nuggets de modelo se colocan en rutas para permitir puntuar nuevos datos y generar nuevos nodos. La **puntuación** de datos le permite utilizar la información obtenida a partir de la generación de modelos para crear predicciones para nuevos registros. Para ver los resultados de la puntuación, necesitará adjuntar un nodo terminal (es decir, un nodo de procesamiento o de resultado) al nugget y ejecutar dicho nodo.

Para algunos modelos, los nugget de modelo también pueden proporcionar información adicional sobre la calidad de la predicción, como los valores de confianza o las distancias desde los centros de los clústeres. La generación de nuevos nodos le permite crear fácilmente nuevos nodos basados en la estructura del modelo generado. Por ejemplo, la mayoría de modelos que realizan la selección de campos de entrada le permiten generar nodos Filtrar que solo pasarán campos de entrada que el modelo haya identificado como importantes.

**Nota:** Puede haber pequeñas diferencias en las puntuaciones asignadas a un caso dado por un modelo determinado cuando se ejecutan en diferentes versiones de IBM SPSS Modeler. Suele ser el resultado de mejoras en el software entre versiones.

Uso de un nugget de modelo para puntuar datos

1. Conecte el nugget de modelo a una ruta u origen de datos que pasará datos al nugget.
2. Añada o conecte uno o más nodos de procesamiento o de resultados (como un nodo Tabla o Análisis) al nugget de modelo.
3. Ejecute uno de los nodos posteriores de la ruta desde el nugget de modelo.

*Nota*: no puede utilizar el nodo Reglas sin refinar. Para puntuar datos basados en un modelo de reglas de asociación, utilice el nodo de reglas sin refinar para generar un nugget de conjunto de reglas y utilice el nugget de conjunto de reglas para la puntuación. Consulte el tema "Generación de un conjunto de reglas desde un nugget de modelo de asociación" en la página 280 para obtener más información.

Uso de un nugget de modelo para generar nodos de procesamiento

1. En la paleta, examine el modelo o, en el lienzo de rutas, edite el modelo.

2. Seleccione el tipo de nodo deseado del menú Generar, en la ventana del explorador de nugget de modelo. Las opciones disponibles variarán en función del tipo del nugget de modelo. Si desea obtener información más detallada sobre lo que se puede generar a partir de un determinado modelo, consulte el tipo específico de nugget de modelo.

## Regeneración de un nodo de modelado

Si tiene un nugget de modelo que desea modificar o actualizar y no está disponible la ruta que se utilizó para crear el modelo, puede volver a generar el nodo de modelado con las mismas opciones que empleó para crear el modelo original.

Para volver a generar un modelo, pulse con el botón derecho en el modelo en la paleta de modelos y seleccione **Generar nodo de modelado**.

Si lo prefiere, al examinar un modelo, seleccione **Generar nodo de modelado** en el menú Generar.

El nodo de modelado que se ha vuelto a generar debería ser funcionalmente idéntico al utilizado para crear el modelo original en la mayoría de los casos.

- En los modelos de árboles de decisión, se puede almacenar con el nodo la configuración adicional especificada durante la sesión interactiva, y la opción **Utilizar directivas de árbol** aparecerá activada en el nodo de modelado regenerado.
- En los modelos de lista de decisiones, aparecerá activada la opción **Usar información de sesión interactiva guardada**. Consulte el tema "Opciones del modelo de la lista de decisiones" en la página 156 para obtener más información.
- En los modelos de serie temporal, la opción **Continuar con la estimación utilizando modelo(s) existente** está habilitada, lo que permite volver a generar el modelo anterior con los datos actuales. Consulte el tema Opciones del modelo de serie temporal para obtener más información.

## Cómo importar y exportar modelos como PMML

PMML, o lenguaje de códigos para modelos predictivos, es un formato XML para describir modelos estadísticos y de minería de datos, incluyendo entradas a modelos, transformaciones utilizadas para preparar los datos para minería de datos, y los parámetros que definen los propios modelos. IBM SPSS Modeler importa y exporta PMML, con lo que se permite compartir modelos con otras aplicaciones que admitan este formato, como IBM SPSS Statistics.

Si desea obtener más información sobre PMML, consulte el sitio Web del grupo de minería de datos (*http://www.dmg.org*).

Para exportar un modelo

La mayoría de tipos de modelos generados por IBM SPSS Modeler admite la exportación PMML. Consulte el tema "Tipos de modelo que admiten PMML" en la página 51 para obtener más información.

1. Pulse con el botón derecho del ratón en un nugget en la paleta de modelos. (también puede pulsar dos veces un nugget de modelo en el lienzo y seleccionar el menú Archivo.)
2. En el menú, pulse **Exportar PMML**.
3. En el cuadro de diálogo Exportar (o Guardar), especifique un directorio objetivo y un nombre exclusivo para el modelo.

*Nota*: Puede cambiar las opciones de exportación PMML en el cuadro de diálogo Opciones de usuario. En el menú principal, pulse en:

**Herramientas** > **Opciones** > **Opciones de usuario**

y pulse la pestaña PMML.

Para importar un modelo guardado como PMML

Los modelos exportados como PMML desde IBM SPSS Modeler o cualquier otra aplicación se pueden importar a la paleta de modelos. Consulte el tema "Tipos de modelo que admiten PMML" para obtener más información.

1. En la paleta de modelos, pulse con el botón derecho en la paleta y seleccione **Importar PMML** del menú.
2. Seleccione el archivo que desea importar y especifique las opciones de las etiquetas de valores y variables como desee.
3. Pulse en **Abrir**.

**Utilice las etiquetas de variables si están presentes en el modelo.** El lenguaje PMML puede especificar tanto nombres de variables como etiquetas de variables (como ID de referencia para *IDRef*) para las variables del diccionario de datos. Seleccione esta opción para utilizar etiquetas de variables si están presentes en el PMML exportado originalmente.

Si ha seleccionado las opciones anteriores de etiqueta pero en el PMML no hay ninguna etiqueta de variable o de valor, entonces los nombres de variables y valores literales se utilizarán como normales.

## Tipos de modelo que admiten PMML

### Exportación de PMML

**Modelos de IBM SPSS Modeler.** Los siguientes modelos creados en IBM SPSS Modeler pueden exportarse como PMML 4.0:

- Árbol C&R
- QUEST
- CHAID
- Regresión lineal
- Red neuronal
- C5.0
- Regresión Logística
- Genlin
- SVM
- A priori
- Carma
- K-medias
- Kohonen
- Dos fases
- GLMM (soporte únicamente para modelos GLMM de solo efecto fijo)
- Lista de decisiones
- Cox
- Secuencia (no se admite la puntuación para modelos PMML de secuencia)
- Estadísticas Modelo

**Modelos nativos de bases de datos.** Para modelos generados utilizando algoritmos nativos de bases de datos, la exportación PMML no está disponible. Los modelos creados mediante Analysis Services desde Microsoft o Oracle Data Miner no se pueden exportar.

### Importación de PMML

IBM SPSS Modeler puede importar y puntuar modelos PMML generados por versiones actuales de todos los productos de IBM SPSS Statistics, incluidos los modelos exportados desde IBM SPSS Modeler, así como cualquier modelo o transformación PMML generado mediante IBM SPSS Statistics 17.0 o posterior. Básicamente, esto significa cualquier PMML que pueda puntuar el motor de puntuación, con las siguientes excepciones:

- Los modelos Apriori, CARMA, de detección de anomalías, de secuencia y de reglas de asociación no pueden importarse.
- Es posible que no pueda navegar por los modelos de PMML después de importar a IBM SPSS Modeler aunque se puedan utilizar para la puntuación. (Tenga en cuenta que esto incluye los modelos que se exportaron de IBM SPSS Modeler para comenzar. Para evitar esta limitación, exporte el modelo como un archivo del modelo generado [`*.gm`] en lugar de como PMML.)
- La validación limitada se produce al importar, pero la validación completa se realiza al intentar puntuar el modelo. Por lo tanto es posible que la importación sea correcta pero que la puntuación falle o genere resultados incorrectos.

**Nota:** Para PMML de terceros importado en IBM SPSS Modeler, IBM SPSS Modeler se intentará puntar los PMML válidos que se puedan reconocer y puntuar. Sin embargo, no se garantiza que puntuarán todos los PMML o que lo harán de la misma manera que la aplicación que los ha generado.

## Publicar modelos para un adaptador de puntuación

Puede publicar modelos en un servidor de la base de datos que tenga instalado un adaptador de puntuación. Un adaptador de puntuación permite que la puntuación de modelos tenga lugar dentro de la base de datos, mediante el uso de las capacidades de las funciones definidas por el usuario (UDF) de la base de datos. Si realiza la puntuación en la base de datos ya no es necesario extraer los datos antes de la puntuación. Si publica en un adaptador de puntuación, también se genera algún SQL de ejemplo para ejecutar las UDF.

### Para publicar un adaptador de puntuación

1. Pulse dos veces en el nugget de modelo para abrirlo.
2. En el menú de nugget de modelo, seleccione:

   **Archivo** > **Publicar para adaptador de puntuación de servidor**
3. Cumplimente los campos pertinentes del cuadro de diálogo y pulse en **Aceptar**.

**Conexión a la base de datos.** Los detalles de la conexión a la base de datos que desea utilizar para el modelo.

**ID de publicación.** (Solo para bases de datos Db2 para z/OS) Un identificador para el modelo. Si vuelve a crear el mismo modelo y utiliza el mismo ID de publicación, el SQL generado no cambia, de modo que es posible volver a crear un modelo sin tener que cambiar la aplicación que utiliza el SQL generado previamente. (Para otras bases de datos, el SQL que se genera es exclusivo para el modelo.)

**Generar SQL de ejemplo.** Si se selecciona esta opción, se genera el SQL de ejemplo en el archivo especificado del campo **Archivo**.

## Modelos sin refinar

Un modelo sin refinar contiene información extraída de los datos, pero no está diseñado para generar predicciones directamente. Significa que no se puede añadir a las rutas. Los modelos sin refinar se visualizan como “diamantes en bruto” en la paleta de modelos generados.

*Figura 27. Icono de modelo sin refinar*

Si desea obtener información acerca del modelo de regla sin refinar, pulse con el botón derecho en el modelo y elija **Examinar** en el menú contextual. Al igual que otros modelos generados en IBM SPSS Modeler, las diferentes pestañas ofrecen información del resumen y las reglas del modelo creado.

**Generación de nodos.** El menú Generar le permite crear nuevos nodos basados en las reglas.

- **Nodo Seleccionar** Genera un nodo Seleccionar para elegir los registros a los que se aplica la regla actualmente seleccionada. Si no se selecciona ninguna regla, esta opción está desactivada.
- **Conjunto de reglas.** Genera un nodo de conjunto de reglas para predecir los valores de un campo objetivo único. Consulte el tema "Generación de un conjunto de reglas desde un nugget de modelo de asociación" en la página 280 para obtener más información.

# Capítulo 4. Modelos de cribado

## Cribado de campos y registros

Se pueden utilizar varios nodos de modelado durante las etapas preliminares de un análisis para buscar campos y registros que tienen más probabilidad de ser de interés para el modelado. Puede utilizar el nodo Selección de características para cribar y ordenar campos por rangos según la importancia, y el nodo Detección de anomalías, para buscar registros poco habituales que no cumplan los patrones conocidos de datos "normales".

El nodo Selección de características filtra los campos de entrada para su eliminación en función de un conjunto de criterios (como el porcentaje de valores perdidos); a continuación, clasifica el grado de importancia del resto de entradas de acuerdo con un objetivo específico. Por ejemplo, a partir de un conjunto de datos dado con cientos de entradas potenciales, ¿cuáles tienen mayor probabilidad de ser útiles para el modelado de resultados de pacientes?

El nodo Detección de anomalías identifica casos extraños, o valores atípicos, que no se ajustan a patrones de datos “normales”. Con este nodo, es posible identificar valores atípicos aunque no se ajusten a ningún patrón previamente conocido o no se realice una búsqueda exacta.

Tenga en cuenta de que la detección de anomalías identifica registros o casos extraños a través del análisis de clústeres según el conjunto de campos seleccionado en el modelo, sin considerar ningún campo objetivo específico (dependiente) ni si tales campos son relevantes para el patrón que intenta predecir. Por este motivo, puede que desee utilizar la detección de anomalías en combinación con la selección de características o con cualquier otra técnica de cribado y clasificación de campos. Así, puede utilizar la selección de características para identificar los campos más importantes relativos a un objetivo específico y, a continuación, utilizar la detección de anomalías para buscar los registros menos habituales con respecto a estos campos. (Un método alternativo sería crear un modelo de árbol de decisión y, a continuación, examinar los registros clasificados erróneamente como anomalías potenciales. Sin embargo, este método sería más difícil de replicar o automatizar a gran escala.)

## Nodo Selección de características

Puede que los problemas relacionados con la minería de datos impliquen cientos, o incluso miles, de campos que se pueden utilizar potencialmente como entradas. Por consiguiente, puede que se invierta mucho tiempo y esfuerzo en examinar qué campos o variables se incluirán en el modelo. Para limitar las opciones, se puede utilizar el algoritmo Selección de características para identificar los campos que son más importantes para un análisis específico. Por ejemplo, si está intentando predecir resultados de pacientes según un número de factores: ¿qué factores tienen la mayor probabilidad de ser importantes?

La selección de características se compone de tres pasos:

- **Cribado.** Elimina las entradas y registros, o casos, problemáticos y no importantes, como los campos de entrada con demasiados valores que faltan o con una variación demasiado grande o pequeña para ser útiles.
- **Clasificación.** Ordena las entradas restantes y les asigna un rango en función de la importancia.
- **Selección.** Identifica el subconjunto de características a utilizar en modelos posteriores, por ejemplo conservando sólo las entradas más importantes y filtrando o excluyendo el resto.

En una época en la que muchas organizaciones están sobrecargadas con demasiados datos, las ventajas de la selección de características al simplificar y agilizar el proceso de modelado pueden ser numerosas. Al

centrar la atención rápidamente en los campos más importantes, se puede reducir la cantidad de cálculos necesarios, localizar más fácilmente las relaciones pequeñas pero importantes que, de otra forma, se pasarían por alto y, por último, obtener modelos más sencillos, precisos y fáciles de explicar. Al reducir el número de campos utilizados en el modelo, verá que se puede reducir el tiempo de puntuación, así como la cantidad de datos recopilados en iteraciones futuras.

**Ejemplo.** Una compañía telefónica tiene un almacén de datos con información sobre las respuestas de 5.000 clientes en relación con una promoción especial. Los datos incluyen un gran número de campos que contienen los estadísticos del uso del teléfono, las edades de los clientes, el puesto de trabajo y los ingresos. Tres campos "objetivo" muestran si el cliente respondió a cada una de tres ofertas. La empresa desea utilizar estos datos para predecir qué clientes tienen más probabilidad de responder a ofertas similares en un futuro.

**Requisitos.** Un único campo de objetivo (uno con su rol definido a *Objetivo*), junto con múltiples campos de entrada que desee filtrar o clasificar de forma relativa a su objetivo. Ambos campos de objetivo y entrada pueden tener un nivel de medición de *Continuo* (rango numérico o *Categórico*.

## Configuración del modelo de selección de características

La configuración de la pestaña Modelo incluye las opciones de modelo estándar junto con la configuración que le permite ajustar los criterios para cribar campos de entrada.

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

Criba de campos de entrada

El cribado implica eliminar entradas o casos que no aportan ninguna información útil en cuanto a la relación entrada/objetivo. Las opciones de cribado se basan en atributos del campo en cuestión, sin contemplar la eficacia predictiva del campo objetivo seleccionado. Los campos cribados se excluyen de los cálculos utilizados para ordenar entradas por rangos y, opcionalmente, se pueden filtrar o eliminar de los datos utilizados en el modelado.

Los campos se pueden cribar en función de los siguientes criterios:

- **Porcentaje máximo de valores perdidos.** Criba campos con demasiados valores perdidos, expresados como un porcentaje del número total de registros. Los campos con un alto porcentaje de valores perdidos proporcionan poca información predictiva.
- **Porcentaje máximo de registros en una categoría única.** Criba campos con demasiados registros dentro de la misma categoría en relación con el número total de registros. Por ejemplo, si el 95% de los clientes de la base de datos conduce el mismo tipo de coche, no sería útil incluir esta información para distinguir a un cliente de otro. Cualquier campo que exceda el máximo especificado se criba. Esta opción sólo se aplica a campos categóricos.
- **Número máximo de categorías como un porcentaje de registros.** Criba campos con demasiadas categorías en relación con el número total de registros. Si un porcentaje elevado de las categorías contiene sólo un único caso, puede que el campo sea de uso limitado. Por ejemplo, si cada cliente lleva un sombrero diferente, será improbable que esta información sirva a la hora de modelar patrones de comportamiento. Esta opción sólo se aplica a campos categóricos.
- **Coeficiente mínimo de variación.** Criba campos con un coeficiente de varianza menor o igual que el mínimo especificado. Esta medida es el índice de la desviación estándar del campo de entrada a la media del campo de entrada. Si este valor es cercano a cero, no habrá mucha variabilidad en los valores de la variable. Esta opción sólo se aplica a campos continuos (rango numérico).
- **Desviación estándar mínima.** Criba campos con desviación estándar menor o igual que el mínimo especificado. Esta opción sólo se aplica a campos continuos (rango numérico).

**Registros con datos perdidos.** Los registros o casos que tienen valores perdidos en el campo objetivo, o bien valores perdidos en todas las entradas, se excluyen automáticamente de todos los cálculos utilizados en las clasificaciones.

## Opciones de la selección de características

La pestaña Opciones permite especificar la configuración predeterminada para seleccionar o excluir campos de entrada en el nugget de modelo. Tras ello, se puede añadir el modelo a una ruta para seleccionar un subconjunto de campos para usarlo en generaciones de modelos posteriores. Opcionalmente, se puede sobrescribir esta configuración seleccionando o anulando la selección de campos adicionales en el explorador de modelos cuando haya generado el modelo. Sin embargo, la configuración predeterminada permite aplicar el nugget de modelo sin más cambios, lo que puede ser especialmente útil para scripts.

Consulte el tema "Resultados del modelo de selección de características" en la página 58 para obtener más información.

Se encuentran disponibles las siguientes opciones:

**Todos los campos clasificados.** Selecciona los campos según la clasificación como *important*, *marginal* o *unimportant*. Se puede editar la etiqueta de clasificación, así como los valores de corte que se utilizan para asignar los registros a un rango u otro.

**Número especificado de campos.** Selecciona los *n* campos principales en función de su importancia.

**Importancia mayor que.** Selecciona todos los campos con una importancia superior al valor especificado.

El campo objetivo siempre se conserva, independientemente de la selección.

Opciones de clasificación de la importancia

**Todos categóricos.** Cuando todas las entradas y el objetivo son categóricos, la importancia se puede clasificar en función de cualquiera de las cuatro medidas siguientes:

- **Chi-cuadrado de Pearson.** Comprueba la independencia del objetivo y la entrada sin indicar la fuerza o la dirección de cualquier relación existente.
- **Chi-cuadrado de la razón de verosimilitud.** Parecida al chi-cuadrado de Pearson, pero también comprueba la independencia del objetivo y de la entrada entre sí.
- **V de Cramer.** Medida de asociación basada en el estadístico de chi-cuadrado de Pearson. Los valores oscilan entre 0 (que indica que no hay asociación) y 1 (que señala una asociación perfecta).
- **Lambda.** Una medida de asociación que refleja la reducción proporcional de error cuando se utiliza la variable para predecir el valor objetivo. Un valor de 1 indica que el campo de entrada predice perfectamente el objetivo, mientras que un valor de 0 denota que la entrada no proporciona información útil sobre el objetivo.

**Algunos categóricos.** Cuando algunas entradas, pero no todas, son categóricos y el objetivo también es categórico, la importancia se puede clasificar según los chi-cuadrado de Pearson o de la razón de verosimilitud. (La *V* de Cramer y lambda no estarán disponibles a menos que todas las entradas sean categóricas.)

**Categóricos frente a continuos.** Cuando se clasifica una entrada categórica al compararla con un objetivo continuo o a la inversa (uno de los dos es categórico, pero no ambos), se utiliza el estadístico *F*.

**Ambos continuos.** Cuando se clasifica una entrada continua al compararla con un objetivo continuo, se utiliza el estadístico *t* basado en el coeficiente de correlación.

## Nugget del modelo de selección de características

Los nugget de modelo de Selección de características muestran la importancia de cada entrada respecto al objetivo seleccionado, según la ordenación por rangos realizada a partir del nodo Selección de características. Se enumeran asimismo todos los campos que se hayan cribado antes de la clasificación. Consulte el tema "Nodo Selección de características" en la página 55 para obtener más información.

Cuando se ejecuta una ruta que contiene un nugget de modelo de Selección de características, el modelo actúa como un filtro que conserva sólo las entradas seleccionadas, tal y como se indica en la selección actual de la pestaña Modelo. Por ejemplo, podría seleccionar todos los campos con rango importante (una de las opciones predeterminadas) o bien seleccionar manualmente un subconjunto de campos en la pestaña Modelo. El campo objetivo se conserva también, independientemente de la selección. El resto de campos se excluye.

El filtrado se basa únicamente en el nombre del campo; por ejemplo, si selecciona *edad* e *ingresos*, se conservará cualquier campo que coincida con uno de estos nombres. El modelo no actualiza la clasificación de los campos a partir de datos nuevos, sino que simplemente filtra los campos en función de los nombres seleccionados. Por este motivo, deberá tener cuidado al aplicar el modelo a los datos nuevos o actualizados. Si no está seguro, se recomienda volver a generar el modelo.

## Resultados del modelo de selección de características

La pestaña Modelo de un nugget de modelo de selección de características muestra el rango y la importancia de todos las entradas en el panel superior y, asimismo, permite seleccionar los campos que se van a filtrar utilizando las casillas de verificación de la columna de la izquierda. Cuando se ejecuta la ruta, sólo los campos seleccionados se conservan; los demás campos se descartan. Las selecciones predeterminadas se basan en las opciones especificadas en el nodo de generación de modelos, pero se puede seleccionar o anular la selección de campos adicionales según sea necesario.

El panel inferior enumera las entradas que se han excluido de las clasificaciones en base al porcentaje de valores perdidos o a otros criterios especificados en el nodo de modelado. Al igual que con los campos con rangos, podrá optar por incluir o descartar estos campos utilizando las casillas de verificación de la columna de la izquierda. Consulte el tema "Configuración del modelo de selección de características" en la página 56 para obtener más información.

- Para ordenar la lista por rango, nombre del campo, importancia o cualquiera de las columnas que aparecen, pulse en la cabecera de la columna. También puede utilizar la barra de herramientas para seleccionar el elemento que desea de la lista Ordenar por y usar las flechas hacia arriba y hacia abajo para cambiar la dirección de la ordenación.
- Puede utilizar la barra de herramientas para seleccionar o deseleccionar todos los campos y para acceder al cuadro de diálogo Seleccionar campos, que le permite seleccionar campos por rango o importancia. También puede pulsar las teclas Mayús o Ctrl mientras pulsa en los campos para ampliar la selección y utilizar la barra espaciadora para activar o desactivar un grupo de campos seleccionados. Consulte el tema "Selección de campos por importancia" para obtener más información.
- Los valores de umbral para clasificar las entradas como importantes, marginales o sin importancia se muestran en la leyenda bajo la tabla. Estos valores se especifican en el nodo de modelado. Consulte el tema "Opciones de la selección de características" en la página 57 para obtener más información.

## Selección de campos por importancia

Al puntuar datos mediante un nugget de modelo de Selección de características, se conservarán todos los campos que se hayan seleccionado de la lista de campos cribados o con rango, que se indican mediante las casillas de verificación en la columna de la izquierda. El resto de campos se descartarán. Para cambiar la selección, puede utilizar la barra de herramientas para acceder al cuadro de diálogo Seleccionar campos, que permite seleccionar los campos por rango o importancia.

**Todos los campos marcados.** Selecciona todos los campos marcados como importantes, marginales o sin importancia.

**Número especificado de campos.** Le permite seleccionar los *n* campos principales en función de su importancia.

**Importancia mayor que.** Selecciona todos los campos con una importancia superior al umbral especificado.

## Generación de un filtro desde el modelo de selección de características

Basándose en los resultados de un modelo de selección de características, puede utilizar el cuadro de diálogo Generar filtro desde característica para generar uno o más nodos Filtro que incluyan o excluyan subconjuntos de campos basados en la importancia relativa al objetivo específico. Dado que el nugget de modelo también se puede utilizar como un filtro, proporciona flexibilidad para experimentar con diferentes subconjuntos de campos sin tener que copiar o modificar el modelo. Independientemente de si se ha optado por incluir o excluir, el filtro conserva siempre el campo objetivo.

**Incluir/Excluir.** Se puede elegir incluir campos o excluirlos, por ejemplo, incluir los 10 campos principales o excluir todos los campos marcados como sin importancia.

**Campos seleccionados.** Incluye o excluye todos los campos seleccionados actualmente en la tabla.

**Todos los campos marcados.** Selecciona todos los campos marcados como importantes, marginales o sin importancia.

**Número especificado de campos.** Le permite seleccionar los *n* campos principales en función de su importancia.

**Importancia mayor que.** Selecciona todos los campos con una importancia superior al umbral especificado.

# Nodo Detección de anomalías

Los modelos de detección de anomalías se utilizan para identificar valores atípicos, o casos extraños, en los datos. A diferencia de otros métodos de modelado que almacenan reglas acerca de casos extraños, los modelos de detección de anomalías almacenan información sobre el patrón de comportamiento normal. Esto permite identificar valores atípicos, incluso si no se ajustan a ningún patrón conocido, y puede ser especialmente útil en aplicaciones, como detección de fraudes, donde pueden surgir patrones nuevos constantemente. La detección de anomalías es un método no supervisado, lo que significa que no requiere un conjunto de datos de entrenamiento que contenga casos conocidos de fraudes para utilizarlos como punto de partida.

La detección de anomalías puede examinar un gran número de campos para identificar clústeres o grupos de homólogos en los que hay registros similares, mientras que los métodos tradicionales de identificación de valores atípicos observan una o dos variables a la vez. Así, se puede comparar cada registro con el resto del grupo de homólogos para identificar posibles anomalías. Cuanto más alejado esté un caso del centro normal, mayor será la probabilidad de que sea extraño. Por ejemplo, el algoritmo podría agrupar registros en tres clústeres distintos y marcar aquellos que se sitúen lejos del centro de cualquier clúster.

Se asigna un índice de anomalía a cada registro, que es el cociente del índice de desviación del grupo sobre su media sobre el clúster al que pertenece el caso. Cuanto mayor sea el valor de este índice, mayor será la desviación del caso sobre la media. En circunstancias normales, los casos con valores de índice de anomalía inferiores a 1 o incluso 1,5 no se considerarán anomalías, ya que su desviación es prácticamente

la misma o sólo un poco superior a la media. Sin embargo, los casos con un valor de índice superior a 2 se consideran anómalos por presentar una desviación que es al menos el doble de la media.

La detección de anomalías es un método exploratorio diseñado para detectar rápidamente casos o registros extraños que deberían someterse a un análisis más detallado. Éstos deben considerarse *sospechosos* de anomalía, los cuales tras un análisis más exhaustivo, puede que resulten anomalías reales. Aunque puede que un registro le parezca totalmente válido, debe analizarlo a partir de los datos para generar un modelo. Otra posibilidad es que, en el caso de que el algoritmo ofrezca repetidamente anomalías falsas, se trate de un error o artefacto en el proceso de recopilación de datos.

Tenga en cuenta de que la detección de anomalías identifica registros o casos extraños a través del análisis de clústeres según el conjunto de campos seleccionado en el modelo, sin considerar ningún campo objetivo específico (dependiente) ni si tales campos son relevantes para el patrón que intenta predecir. Por este motivo, puede que desee utilizar la detección de anomalías en combinación con la selección de características o con cualquier otra técnica de cribado y clasificación de campos. Así, puede utilizar la selección de características para identificar los campos más importantes relativos a un objetivo específico y, a continuación, utilizar la detección de anomalías para buscar los registros menos habituales con respecto a estos campos. (Un método alternativo sería crear un modelo de árbol de decisión y, a continuación, examinar los registros clasificados erróneamente como anomalías potenciales. Sin embargo, este método sería más difícil de replicar o automatizar a gran escala.)

**Ejemplo.** Al cribar subvenciones para el desarrollo agrícola para posibles casos de fraude, se puede utilizar la detección de anomalías para descubrir las desviaciones de la norma, resaltando aquellos registros que sean anómalos y dignos de una investigación más detallada. En particular, le interesan aquellas solicitudes de subvenciones que parezcan reclamar demasiado dinero teniendo en cuenta el tipo y tamaño de la granja.

**Requisitos.** Uno o varios campos de entrada. Tenga en cuenta que sólo se pueden usar como entrada aquellos campos con el rol definido como **Entrada** mediante un nodo de origen o un nodo Tipo. Se omitirán los campos objetivo (con el rol definido como **Objetivo** o **Ambos**).

**Puntos fuertes.** Si se marcan los casos que *no* cumplen con un conjunto de reglas conocido para diferenciarlos de los que sí lo hacen, los modelos de detección de anomalías podrán identificar casos poco habituales incluso cuando no sigan patrones conocidos anteriormente. Cuando la detección de anomalías se utiliza en combinación con la selección de características, permite cribar grandes cantidades de datos con el fin de identificar los registros de mayor interés de forma relativamente rápida.

## Opciones del modelo de detección de anomalías

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Determinar valor de corte para la anomalía basado en.** Especifica el método utilizado para determinar el valor de corte con el que se van a marcar anomalías. Se encuentran disponibles las siguientes opciones:

- **Nivel mínimo de índice de anomalía.** Especifica el valor de corte mínimo con el que se van a marcar anomalías. Se marcarán aquellos registros que cumplan o sobrepasen este umbral.
- **Porcentaje de registros más anómalos de los datos de entrenamiento.** Establece automáticamente el umbral en un nivel que marca el porcentaje de registros especificado en los datos de entrenamiento. El valor de corte resultante se incluye como un parámetro en el modelo. Tenga en cuenta que con esta opción se determina la manera en la que el valor de corte se establece, *no* el porcentaje real de registros que se va a marcar durante la puntuación. Los resultados de puntuación reales pueden variar en función de los datos.
- **Número de registros más anómalos de los datos de entrenamiento.** Establece automáticamente el umbral en un nivel que marca el número de registros especificado en los datos de entrenamiento. El umbral resultante se incluye como un parámetro en el modelo. Tenga en cuenta que con esta opción se

determina la manera en que se establece el valor de corte, *no* el número específico de registros que se va a marcar durante la puntuación. Los resultados de puntuación reales pueden variar en función de los datos.

*Nota*: independientemente de cómo se determine el valor de corte, esto no incidirá en el valor de índice de anomalía subyacente mostrado para cada registro. Simplemente define el umbral para marcar los registros como anómalos al calcular o puntuar el modelo. Si posteriormente desea examinar un número mayor o menor de registros, puede utilizar un nodo Seleccionar para identificar un subconjunto de registros a partir del valor de índice de anomalía (`$O-AnomalyIndex > X`).

**Número de campos de anomalía del informe.** Especifica el número de campos del informe como una indicación del motivo por el que un registro particular se ha marcado como una anomalía. Se informa de los campos más anómalos, definidos como aquellos que presentan la mayor desviación de la norma de campo relativa al clúster al que se ha asignado el registro.

## Opciones del experto de detección de anomalías

Para especificar las opciones para valores perdidos y otras configuraciones, establezca el modo en **Experto** en la pestaña Experto.

**Coeficiente de ajuste.** Valor utilizado para equilibrar la ponderación relativa asignada a los campos categóricos y continuos (rango numérico) al calcular la distancia. Los valores mayores aumentan la influencia de los campos continuos. Este valor debe ser distinto de cero.

**Calcular automáticamente número de grupos de homólogos.** La detección de anomalías se puede usar para analizar rápidamente un gran número de soluciones posibles mediante las que se puede elegir el número óptimo de grupos de homólogos para los datos de entrenamiento. Se puede ampliar o acotar el rango estableciendo el número de grupos de homólogos mínimo y máximo. Los valores mayores permitirán que el sistema explore un rango más amplio de posibles soluciones; sin embargo, el coste es un aumento del tiempo de proceso.

**Especificar número de grupos de homólogos.** Si sabe el número de clústeres que va a incluir en el modelo, seleccione esta opción e introduzca el número de grupos de homólogos. Normalmente, si selecciona esta opción obtendrá un mejor rendimiento.

**Nivel y Proporción de ruido.** Estas opciones determinan cómo se tratarán los valores atípicos durante un clúster de dos fases. En la primera fase, se utiliza un árbol de características de clústeres (CF) para reducir los datos de un gran número de registros individuales a un número de clústeres más manejable. El árbol se construye en base a medidas de similitud y, cuando un nodo del árbol obtiene numerosos registros, se divide en nodos hijo. En la segunda fase, comienza la agrupación en clústeres jerárquica en los nodos terminales del árbol CF. El tratamiento del ruido está activado en la primera lectura de datos y desactivado en la segunda. Los casos en el clúster de ruido de la primera lectura de datos se asignan a los clústeres habituales de la segunda lectura.

- **Nivel de ruido.** Especifique un valor entre 0 y 0,5. Esta configuración sólo es relevante cuando el árbol CF se llena durante la fase de crecimiento, lo que significa que no puede aceptar ningún caso más en un nodo hoja y, asimismo, que ningún nodo hoja se puede dividir.

  Si un árbol CF se llena y el nivel de ruido está establecido en 0, el umbral aumentará y el árbol CF volverá a crecer con todos los casos. Después del clúster final, aquellos valores que no se hayan podido asignar a un clúster se etiquetarán como atípicos. Al clúster de valores atípicos se le da un número de identificación de -1. El clúster de valores atípicos no está incluido en el recuento del número de clústeres; es decir, si especifica *n* clústeres y manejo de ruido, el algoritmo producirá *n* clústeres y un clúster de ruido. En la práctica, el aumento de este valor ofrece mayor latitud al algoritmo para ajustar registros poco habituales en el árbol en lugar de asignarlos a un clúster de valores atípicos diferente.

  Si el nivel de ruido es superior a 0 y el árbol CF se llena, éste volverá a crecer tras haber colocado los datos que se hallaban en hojas dispersas en su propia hoja de ruido correspondiente. Una hoja se considera dispersa cuando su cociente de número de casos en relación con el número de casos de la

hoja más grande es menor que el nivel de ruido. Cuando el árbol haya crecido, los valores atípicos se colocarán en el árbol CF si es posible. Si no, se descartarán para la segunda fase de agrupación en clústeres.

- **Proporción de ruido.** Especifica la parte de memoria asignada al componente que debería usarse para el almacenamiento en búfer de ruido. Este valor está comprendido ente 0,0 y 0,5. Si insertar un caso específico en una hoja del árbol produce una densidad inferior al umbral, la hoja no se dividirá. Si ésta excede el umbral, la hoja se dividirá, añadiendo un clúster pequeño más al árbol CF. En la práctica, el aumento de esta configuración puede provocar que el algoritmo se deje atraer más rápidamente por un árbol más sencillo.

**Imputar valores perdidos.** En relación con los campos continuos, sustituye la media del campo en lugar de algún valor perdido. En el caso de los campos categóricos, las categorías perdidas se combinan y se tratan como una categoría válida. Si está opción no está seleccionada, los registros con valores perdidos se excluirán del análisis.

## Nugget del modelo de detección de anomalías

Los nugget de modelo de detección de anomalías contienen toda la información que el modelo de detección de anomalías ha capturado, así como información acerca de los datos de entrenamiento y del proceso de estimación.

Cuando se ejecuta una ruta que contiene un nugget de modelo de detección de anomalías, se añade un número de campos nuevos a dicha ruta en función de las selecciones realizadas en la pestaña Configuración del nugget de modelo. Consulte el tema "Configuración del modelo de detección de anomalías" en la página 63 para obtener más información. Los nombres de campo de nuevos se basan en el nombre de modelo, precedido de *$O*, como se resume en la tabla siguiente.

*Tabla 6. Generación de nombre de campo nuevo.*

| Nombre de campo | Descripción |
|---|---|
| *$O-Anomaly* | Campo de marcas que indica si el registro es o no anómalo. |
| *$O-AnomalyIndex* | Valor de índice de anomalía del registro. |
| *$O-PeerGroup* | Especifica el grupo de homólogos al que el registro se asigna. |
| *$O-Field-n* | Nombre del *enésimo* campo más anómalo en términos de desviación de la norma del clúster. |
| *$O-FieldImpact-n* | Índice de desviación variable del campo. Este valor mide la desviación de la norma de campo relativa al clúster al que el registro se asigna. |

Si lo desea, puede suprimir las puntuaciones de los registros que no sean anómalos para hacer que los resultados sean más fáciles de leer. Consulte el tema "Configuración del modelo de detección de anomalías" en la página 63 para obtener más información.

## Detalles del modelo de detección de anomalías

La pestaña Modelo de un modelo de detección de anomalías generado muestra información sobre los grupos de homólogos del modelo.

Tenga en cuenta que los tamaños de los grupos de homólogos y los estadísticos de los que se ha informado son cálculos basados en los datos de entrenamiento y, por lo tanto, pueden diferir ligeramente de los resultados de puntuación, aun cuando se ejecuten en los mismos datos.

## Resumen del modelo de detección de anomalías

La pestaña Resumen de un nugget de modelo de detección de anomalías muestra información sobre los campos, la configuración de creación y el proceso de estimación de modelos. También se muestra el número de grupos de homólogos, además del valor de corte utilizado para marcar registros como anómalos.

## Configuración del modelo de detección de anomalías

Utilice la pestaña Configuración para especificar opciones para puntuar el nugget de modelo.

**Indicar registros anómalos con** Especifica el modo en que se tratan los registros anómalos en el resultado.

- **Marcar e indexar** Crea un campo de marcas que se establece en *True* para todos los registros que sobrepasen el valor de corte incluido en el modelo. También se informa del índice de anomalía de cada registro en un campo aparte. Consulte el tema "Opciones del modelo de detección de anomalías" en la página 60 para obtener más información.
- **Sólo marcar** Crea un campo de marcas, pero no se informa del índice de anomalía de cada registro.
- **Sólo índice** Informa del índice de anomalía sin crear un campo de marcas.

**Número de campos de anomalía a notificar** Especifica el número de campos del informe como una indicación del motivo por el que un registro particular se ha marcado como una anomalía. Se informa de los campos más anómalos, definidos como aquellos que presentan la mayor desviación de la norma de campo relativa al clúster al que se ha asignado el registro.

**Descartar registros** Seleccione esta opción para descartar todos los registros **No anómalos** de la ruta, ya que así es más fácil centrarse en las posibles anomalías en cualquier nodo posterior en la ruta. Si lo prefiere, puede elegir descartar todos los registros **Anómalos** para limitar los análisis posteriores a aquellos registros que no estén marcados como posibles anomalías en función del modelo.

**Nota:** Debido a ciertas diferencias en el redondeo, es posible que el número real de registros marcados durante la puntuación no sea exactamente igual al marcado durante el entrenamiento del modelo, aun cuando se ejecuten en los mismos datos.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

# Capítulo 5. Nodos de modelado automático

Los nodos de modelado automático calculan y comparan un número de diferentes enfoques de modelado, facilitando la prueba de una variedad de métodos de una única pasada. Puede seleccionar los algoritmos de modelado que se utilizarán y las opciones específicas de cada uno de ellos, incluyendo combinaciones que de otro modo serían excluyentes entre sí. Por ejemplo, en lugar de elegir entre los métodos rápido, dinámico o de poda de una red neuronal, puede probarlos todos. El nodo explora cada combinación posible de opciones, evalúa cada modelo candidato en función de la medida especificada y guarda los mejores para su uso en la puntuación o en futuros análisis.

Puede seleccionar entre tres nodos de modelado automáticos, dependiendo de las necesidades de su análisis:

El nodo Clasificador automático crea y compara varios modelos diferentes para obtener resultados binarios (sí o no, abandono o no de clientes, etc.), lo que le permite seleccionar el mejor enfoque para un análisis determinado. Son compatibles varios algoritmos de modelado, por lo que es posible seleccionar los métodos que desee utilizar, las opciones específicas para cada uno y los criterios para comparar los resultados. El nodo genera un conjunto de modelos basado en las opciones especificadas y clasifica los mejores candidatos en función de los criterios que especifique.

El nodo Autonumérico calcula y compara modelos para resultados de rango numérico continuo utilizando cierto número de métodos diferentes. El nodo funciona de la misma manera que el nodo Clasificador automático, lo que le permite seleccionar los algoritmos que desee utilizar y experimentar con varias combinaciones de opciones en una única pasada de modelado. Los algoritmos admitidos incluyen redes neuronales, C&RT, CHAID, regresión lineal, regresión lineal generalizada y máquinas de vectores de soporte (SVM). Los modelos se pueden comparar basándose en la correlación, el error relativo o el número de variables utilizado.

El nodo Agrupación en clústeres automática calcula y compara los modelos de agrupación en clústeres que identifican grupos de registros con características similares. El nodo funciona de la misma manera que otros nodos de modelado automático, permitiéndole experimentar con múltiples combinaciones de opciones en una única pasada de modelado. Los modelos se pueden comparar utilizando medidas básicas con las que se intenta filtrar y definir la utilidad de los modelos de clúster y proporcionar una medida según la importancia de campos concretos.

Los mejores modelos se guardan en un único nugget de modelo compuesto, permitiendo explorarlos y compararlos y seleccionar los modelos que se utilizarán en la puntuación.

- En objetivos numéricos, nominales y binarios únicamente, podrá seleccionar múltiples modelos de puntuación y combinar los resultados en un conjunto de modelos único. Al combinar predicciones de varios modelos, pueden evitarse las limitaciones en modelos individuales que suelen dar como resultado una precisión global superior que puede obtenerse de cualquiera de los modelos.
- También puede decidir profundizar en los resultados y generar nodos de modelado o nugget de modelo para cualquier modelo individual que desee utilizar o explorar más a fondo.

Modelos y tiempo de procesamiento

Dependiendo del conjunto de datos y del número de modelos, los nodos de modelado automático pueden tardar horas en ejecutarse. Preste atención al número de modelos que se están generando al

seleccionar opciones. Si le resulta práctico, puede programar modelados para que se ejecuten por las noches o durante los fines de semana, cuando la demanda de recursos del sistema suele ser menor.

- Si es necesario, puede usar los nodos Partición o Muestrear para reducir el número de registros incluidos en el paso de formación inicial. Cuando haya reducido las opciones a unos cuantos modelos candidatos, se puede restaurar el conjunto de datos completo.
- Para reducir el número de campos de entrada, utilice Selección de características. Consulte el tema "Nodo Selección de características" en la página 55 para obtener más información. También puede utilizar sus ejecuciones iniciales del modelado para identificar campos y opciones que merezca la pena explorar más a fondo. Por ejemplo, si todos sus modelos de mayor rendimiento parecen utilizar los tres mismos campos, es un claro indicador de que merece la pena mantener dichos campos.
- De manera opcional, puede limitar la cantidad de tiempo que se invierte al estimar cualquier modelo y especificar las medidas de evaluación utilizadas para cribar y clasificar modelos.

## Ajustes de algoritmo de nodos de modelado automático

Puede usar la configuración predeterminada o seleccionar las opciones para cada tipo de modelo. Las opciones específicas son parecidas a las disponibles en los nodos de modelado independientes, con la diferencia de que en lugar de elegir un ajuste u otro, puede seleccionar todos los que desee aplicar en la mayoría de los casos. Por ejemplo, si compara modelos Red neuronal, puede seleccionar varios métodos de entrenamiento diferentes y probar cada método con semilla aleatoria y sin ella. Se utilizarán todas las combinaciones posibles de las opciones seleccionadas, facilitando la generación de muchos modelos diferentes de una única pasada. No obstante, tenga cuidado, ya que la selección de varios ajustes puede hacer que el número de modelos se multiplique muy rápidamente.

Para seleccionar las opciones para cada tipo de modelo

1. En el nodo de modelado automatizado, seleccione la pestaña **Experto**.
2. Haga clic en la columna **Parámetros de modelo** para el tipo de modelo.
3. En el menú desplegable, seleccione **Especificar**.
4. En el cuadro de diálogo **Configuración de algoritmo**, seleccione las opciones de la columna **Opciones**.

*Nota*: Hay más opciones en la pestaña Experto del cuadro de diálogo **Configuración del algoritmo**.

## Reglas de parada de nodos de modelado automático

Las reglas de parada especificadas para los nodos de modelado automático están relacionadas con la ejecución global del nodo, no con la parada de modelos determinados generados por el nodo.

**Limitar tiempo de ejecución global a.** (Sólo modelos de Red neuronal, K-medias, Kohonen, Bietápico, SVM, KNN, Red bayesiana y C&R Tree) Detiene la ejecución tras un número específico de horas. Se incluirán en el nugget de modelo todos los modelos generados hasta ese momento, pero no se producirán más modelos.

**Deténgalo en cuanto se produzcan modelos válidos.** Detiene la ejecución cuando un modelo cumple todos los criterios especificados en la pestaña Descartar (para el nodo Clasificador automático o Autoclúster) o la pestaña Modelo (para el nodo Autonumérico). Consulte el tema "Opciones para descartar el nodo Clasificador automático" en la página 73 para obtener más información. Consulte el tema "Opciones para descartar del nodo Agrupación en clústeres automática" en la página 82 para obtener más información.

## Nodo Clasificador automático

El nodo Clasificador automático calcula y compara modelos de los objetivos nominales (conjuntos) o binarios (yes/no), utilizando varios métodos diferentes, permitiéndole probar diversos planteamientos en una sola ejecución de modelado. Puede seleccionar los algoritmos que se utilizarán y experimentar con múltiples combinaciones de opciones. Por ejemplo, en lugar de elegir entre los métodos de función de base radial, polinómico, sigmoide o lineal para una SVM, puede probarlos todos. El nodo explora cada combinación posible de opciones, evalúa cada modelo candidato basándose en la medida especificada y guarda los mejores modelos para utilizarlos en la puntuación o en futuros análisis. Si desea obtener más información, consulte Capítulo 5, "Nodos de modelado automático", en la página 65.

**Ejemplo**

Una empresa minorista contiene datos históricos en los que se registran las ofertas realizadas a determinados clientes en campañas anteriores. Ahora la empresa quiere obtener resultados más rentables realizando la mejor oferta para cada cliente.

**Requisitos**

Un campo de objetivo con un nivel de medición de *Nominal* o *Marca* (con el rol establecido a **Objetivo**), y al menos un campo de entrada (con el rol establecido a **Entrada**). En un campo de marca, el valor *Verdadero* definido para el campo objetivo se supone que representa un acierto al calcular beneficios, elevación y estadísticos relacionados. Los campos de entrada pueden tener un nivel de medición de *Continuo* o *Categórico*, con la limitación de que algunas entradas pueden no ser apropiadas para algunos tipos de modelos. Por ejemplo, los campos ordinales que se utilizan como entradas en los modelos C&RT, CHAID y QUEST deben tener almacenamiento numérico (no en cadenas); asimismo, estos modelos los omitirán si se especifica lo contrario. De igual modo, los campos de entrada continuos pueden establecerse en intervalos en algunos casos. Los requisitos son los mismos que cuando se utilizan los nodos de modelado individuales; por ejemplo, un modelo Red bayesiana funciona igual independientemente de si se ha generado desde el nodo Red bayesiana o el nodo Clasificador automático.

**Campos de frecuencia y ponderación**

La frecuencia y la ponderación se utilizan para proporcionar importancia adicional a ciertos registros sobre otros porque, por ejemplo, el usuario sabe que el conjunto de datos creado no representa totalmente una sección de la población principal (Ponderación) o porque un registro representa un número de casos idénticos (Frecuencia). Si se especifica, los modelos C&RT, CHAID, QUEST, Lista de decisiones y Red bayesiana pueden utilizar un campo de frecuencia. Los modelos C&RT, CHAID y C5.0 pueden utilizar un campo de ponderación. Otros tipos de modelo omitirán estos campos y crearán los modelos de todas formas. Los campos de frecuencia y ponderación sólo se utilizan para la creación de modelos y no se tienen en cuenta al evaluar o puntuar modelos. Si desea obtener más información, consulte "Uso de campos de frecuencia y ponderación" en la página 33.

**Prefijos**

Si conecta un nodo tabla al nugget para el nodo Clasificador automático, existen varias variables nuevas en la tabla con nombres que empiezan con un prefijo $.

Los nombres de los campos que se generan durante la puntuación se basan en el campo objetivo, pero con un prefijo estándar. Los distintos tipos de modelo utilizan diferentes conjuntos de prefijos.

Por ejemplo, los prefijos $G, $R, $C se utilizan como prefijo para las predicciones que han generado el modelo lineal generalizado, el modelo CHAID y el modelo C5.0, respectivamente. Normalmente, $X se genera utilizando un conjunto, y $XR, $XS y $XF se utilizan como prefijos en los casos donde el campo objetivo es un campo Continuo, Categórico o de Marca, respectivamente.

$.Los prefijos .C se utilizan para la confianza de predicción de un objetivo Categórico o de Marca; por ejemplo, $XFC se utiliza como prefijo para la confianza de predicción de Marca del conjunto. $RC y $CC son los prefijos para una sola predicción de confianza para un modelo CHAID y un modelo C5.0, respectivamente.

### Tipos de modelos admitidos

Los tipos de modelo soportados incluyen red neuronal, árbol C&R, QUEST, CHAID, C5.0, regresión logística, lista de decisiones, red bayesiana, discriminante, vecino más cercano, SVM, árbol XGBoost y XGBoost-AS. Consulte el tema "Opciones de experto para el nodo Clasificador automático" en la página 69 para obtener más información.

## Opciones de modelo para el nodo Clasificador automático

La pestaña Modelo del nodo Clasificador automático le permite especificar el número de modelos que se deben crear, junto con los criterios empleados para comparar modelos.

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Crear modelos divididos.** Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como campos de división. Consulte el tema "Generación de modelos divididos" en la página 28 para obtener más información.

**Clasificar modelos por.** Especifique los criterios utilizados para comparar y clasificar los modelos. Las opciones incluyen la precisión global, área debajo de la curva ROC, beneficio, elevación y número de campos. Tenga en cuenta que todas estas medidas estarán disponibles en el informe de resumen independientemente de lo que se seleccione aquí.

*Nota*: En el caso de un objetivo nominal (conjunto), la clasificación está restringida a **Precisión global** o **Número de campos**.

Al calcular beneficios, elevación y estadísticos relacionados, se supone que el valor *True* definido para el campo objetivo representa un acierto.

- **Precisión global** Porcentaje de registros predichos correctamente por el modelo respecto al número total de registros.
- **Área debajo de la curva COR** La curva COR proporciona el índice de rendimiento de un modelo. Cuanto más se encuentre la curva sobre la línea de referencia, más exacta será la prueba.
- **Beneficio (acumulado)** Suma de los beneficios de los percentiles acumulados (clasificados en términos de confianza para la predicción), calculados en base a los costes, ingresos y criterios de ponderación especificados. Normalmente, el beneficio comienza cerca del 0, aumenta rápidamente y, a continuación, desciende. Para obtener un modelo válido, los beneficios deben mostrar un pico bien definido junto con el percentil donde aparece. En el caso de un modelo que no proporciona ninguna información, la curva de beneficio será bastante recta y puede ascender, descender o permanecer en el mismo nivel en función de la estructura de costes/ingresos que se aplique.
- **Elevación (acumulado)** Tasa de aciertos en cantidades acumuladas con respecto a la muestra global (donde los cuantiles se clasifican en función de la confianza para la predicción). Por ejemplo, un valor de elevación de 3 para el cuantil superior indica una tasa de aciertos tres veces más alta que la de la muestra global. Para obtener un modelo válido, la elevación debe comenzar muy por encima de 1,0

para los cuantiles superiores y, a continuación, descender rápidamente hasta 1,0 para los cuantiles inferiores. En el caso de un modelo que no proporciona ninguna información, la elevación se mantendrá alrededor de 1,0.

- **Número de campos** Ordena los modelos en función del campos de entrada utilizados.

**Clasificar modelos usando.** Si se está usando una partición, puede especificar si los rangos se basan en el conjunto de datos de entrenamiento o en el conjunto de prueba. En conjuntos de datos de gran tamaño, si usa una partición para el cribado preliminar de modelos, puede mejorar rendimiento en gran medida.

**Número de modelos que se utilizarán.** Especifica el número máximo de modelos que aparecerán en el nugget de modelo generado por el nodo. Los primeros modelos de la clasificación se enumeran en función del criterio de clasificación especificado. Tenga en cuenta que si aumenta este límite puede ralentizarse el rendimiento. El valor máximo permitido es 100.

**Calcular importancia del predictor.** En el caso de modelos que produzcan una medida adecuada de importancia, puede mostrar un gráfico que indique la importancia relativa de cada predictor al estimar el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Tenga en cuenta que la importancia de predictor puede aumentar el tiempo necesario para calcular algunos modelos; además no se recomienda si sólo desea una amplia comparación entre varios modelos diferentes. Es de mayor utilidad una vez ha limitado su análisis a unos cuantos modelos que desee explorar más a fondo. Consulte el tema “Importancia del predictor” en la página 44 para obtener más información.

**Criterios de beneficio.** *Nota*: Sólo para objetivos de marca. El beneficio es igual a los ingresos de cada registro menos el coste del registro. Los beneficios de un cuantil son la suma de los beneficios de todos los registros del cuantil. Se asume que los beneficios se aplican sólo a los aciertos, pero los costes se aplican a todos los registros.

- **Costes.** Permite especificar el coste asociado con cada registro. Puede seleccionar costes **fijos** o **variables**. En el caso de los costes fijos, especifique el valor del coste. En el caso de los costes variables, pulse en el selector de campos para elegir un campo de costes. (**Costes** no está disponible para los gráficos ROC.)
- **Ingresos.** Permite especificar los ingresos asociados con cada registro que representa un acierto. Puede seleccionar costes **fijos** o **variables**. En el caso de los ingresos fijos, especifique el valor del ingreso. En el caso de los ingresos variables, pulse en el selector de campos para elegir un campo de ingresos. (**Ingresos** no está disponible para los gráficos ROC.)
- **Ponderación.** Si los registros de los datos representan más de una unidad, puede utilizar ponderaciones de frecuencias para ajustar los resultados. Especifique la ponderación asociada con cada registro mediante valores **fijos** o **variables**. En el caso de las ponderaciones fijas, especifique el valor de ponderación (el número de unidades por registro). En el caso de ponderaciones variables, pulse en el selector de campos para elegir un campo de ponderaciones. (**Ponderación** no está disponible para los gráficos ROC.)

**Criterios de elevación.** *Nota*: Sólo para objetivos de marca. Especifica el percentil que hay que utilizar para los cálculos de la elevación. Tenga en cuenta que también puede cambiar este valor al comparar los resultados. Consulte el tema “Nugget de modelo automático” en la página 82 para obtener más información.

## Opciones de experto para el nodo Clasificador automático

La pestaña Experto del nodo Clasificador automático le permite aplicar una partición (si está disponible), seleccionar los algoritmos que se deben utilizar y especificar las reglas de parada.

**Seleccionar modelos.** De forma predeterminada, todos los modelos están seleccionados para ser generados; sin embargo, si tiene Analytic Server, puede elegir restringir los modelos a aquellos que se

pueden ejecutar en Analytic Server y establecerlos previamente para que creen modelos de división o estén preparados para procesar conjuntos de datos muy grandes.

**Nota:** No se admite la creación local de modelos de Analytic Server en el nodo Clasificador automático.

**Modelos utilizados.** Use las casillas de verificación de la columna izquierda para seleccionar los tipos de modelo (algoritmos) que se van a incluir en la comparación. Cuantos más tipos seleccione, más modelos se crearán y más tardará el procesamiento.

**Tipo de modelo.** Enumera los algoritmos disponibles (consulte a continuación).

**Parámetros del modelo.** Puede usar la configuración predeterminada o seleccionar **Especificar** para elegir opciones para cada tipo de modelo. Las opciones específicas son parecidas a las disponibles en los nodos de modelado independientes, con la diferencia de que se pueden seleccionar varias opciones o combinaciones. Por ejemplo, si compara los modelos del nodo Red neuronal, puede seleccionar los seis modelos para entrenarlos de una vez en lugar de seleccionar uno de ellos.

**Número de modelos.** Enumera el número de modelos generados para cada algoritmo basados en la configuración actual. Al combinar opciones, puede aumentar rápidamente el número de modelos, por lo que se recomienda prestar especial atención a este número, especialmente si usa conjuntos de datos grandes.

**Limitar el tiempo máximo empleado en generar un único modelo.** (Sólo modelos de K-medias, Kohonen, bietápicos, SVM, KNN, de red bayesiana y de lista de decisiones) Establece un límite de tiempo máximo para cualquier modelo. Por ejemplo, si un modelo determinado necesita un período de tiempo más largo del esperado para entrenarse debido a una interacción compleja, es probable que no quiera detener la ejecución de todo el modelado.

**Nota:** Si el destino es un campo nominal (conjunto), la opción Lista de decisiones no está disponible.

## Algoritmos admitidos

El nodo Máquina de vectores de soporte (SVM) le permite clasificar datos en uno o dos grupos sin que haya un ajuste por exceso. SVM funciona bien con conjuntos de datos grandes, como aquellos con un gran número de campos de entrada.

El nodo *k* de modelado de vecino (KNN) asocia el nuevo caso con la categoría o valor de los objetos *k* junto a él en el espacio de predictores, donde *k* es un entero. Los casos parecidos están próximos y los que no lo son están alejados entre sí.

El análisis discriminante realiza más supuestos rigurosos que regresiones logísticas, pero puede ser una alternativa o un suplemento valioso al análisis de regresión logística si se cumplen dichos supuestos.

El nodo Red bayesiana le permite crear un modelo de probabilidad combinando pruebas observadas y registradas con conocimiento del mundo real para establecer la probabilidad de instancias. El nodo se centra en las redes Naïve Bayes aumentado a árbol (TAN) y de manto de Markov que se utilizan principalmente para la clasificación.

El nodo Lista de decisiones identifica subgrupos, o segmentos, que muestran una mayor o menor posibilidad de proporcionar un resultado binario relacionado con la población global. Por ejemplo, puede buscar clientes que tengan menos posibilidades de abandonar o más posibilidades de responder favorablemente a una campaña. Puede incorporar su conocimiento empresarial al modelo añadiendo sus propios segmentos personalizados y previsualizando modelos alternativos uno junto a otro para comparar los resultados. Los modelos de listas de decisiones constan de una lista de reglas en las que cada regla tiene una condición y un resultado. Las reglas se aplican en orden, y la primera regla que coincide determina el resultado.

La regresión logística es una técnica de estadístico para clasificar los registros en función los valores de los campos de entrada. Es análoga a la regresión lineal pero toma un campo objetivo categórico en lugar de uno numérico.


El nodo CHAID genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. A diferencia de los nodos C&RT y QUEST, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles, aunque necesita más tiempo para realizar los cálculos.


El nodo QUEST proporciona un método de clasificación binario para generar árboles de decisión; está diseñado para reducir el tiempo de procesamiento necesario para realizar los análisis de C&RT y reducir la tendencia de los métodos de clasificación de árboles para favorecer a las entradas que permitan realizar más divisiones. Los campos de entrada pueden ser continuos (rango numérico), sin embargo el campo objetivo debe ser categórico. Todas las divisiones son binarias.


El nodo de árbol de clasificación y regresión (C&R) genera un árbol de decisión que permite predecir o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera “puro” si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos (nominal, ordinal o marca). Todas las divisiones son binarias (sólo se crean dos subgrupos).


El nodo C5.0 genera un árbol de decisión o un conjunto de reglas. El modelo divide la muestra basándose en el campo que ofrece la máxima ganancia de información en cada nivel. El campo objetivo debe ser categórico. Se permiten varias divisiones en más de dos subgrupos.

El nodo Red neuronal utiliza un modelo simplificado que emula el modo en que el cerebro humano procesa la información: Funciona simultaneando un número elevado de unidades simples de procesamiento interconectadas que parecen versiones abstractas de neuronas. Las redes neuronales son dispositivos eficaces de cálculo de funciones generales y requieren un conocimiento matemático o estadístico mínimo para entrenarlas o aplicarlas.

Los modelos de regresión lineal predicen un objetivo continuo tomando como base las relaciones lineales entre el destino y uno o más predictores.

El nodo Máquina de vectores de soporte lineal (LSVM) le permite clasificar datos en uno o dos grupos sin que haya un ajuste por exceso. LSVM es lineal y funciona bien con conjuntos de datos grandes, como aquellos con un gran número de registros.

El nodo Árboles aleatorios es similar al nodo C&RT existente; el nodo Árboles aleatorios se diseñado para procesar grandes cantidades de datos (Big Data) para crear un único árbol y mostrar el modelo resultante en el visor de la salida que se ha añadido en SPSS Modeler versión 17. El nodo del árbol Árboles aleatorios genera un árbol de decisiones que se utiliza para predecir o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera *puro* si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos (nominal, ordinal o marca). Todas las divisiones son binarias (sólo se crean dos subgrupos).

El nodo Tree-AS es similar al nodo CHAID existente, el nodo Tree-AS se ha diseñado para procesar grandes cantidades de datos (Big Data) para crear un solo árbol y muestra el modelo resultante en el visor de la salida que se ha añadido en SPSS Modeler versión 17. El nodo genera un árbol de decisiones usando un estadístico de chi-cuadrado (CHAID) para identificar las divisiones óptimas. Este uso de CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles, aunque necesita más tiempo para realizar los cálculos.

XGBoost Tree© es una implementación avanzada de un algoritmo de aumento de gradiente con un modelo de árbol como modelo base. Los algoritmos de aumento conocen de forma iterativa los clasificadores débiles y, a continuación, los añaden a un clasificador fuerte final. XGBoost Tree es muy flexible y proporciona muchos parámetros que pueden ser abrumadores para la mayoría de usuarios, de modo que el nodo XGBoost Tree en SPSS Modeler expone las características principales y los parámetros utilizados comúnmente. El nodo se implementa en Python.


XGBoost© es una implementación avanzada de un algoritmo de aumento de gradiente. Los algoritmos de aumento conocen de forma iterativa los clasificadores débiles y, a continuación, los añaden a un clasificador fuerte final. XGBoost es muy flexible y proporciona muchos parámetros que pueden resultar abrumadores para la mayoría de los usuarios, así que el nodo XGBoost-AS en SPSS Modeler expone las características principales y los parámetros utilizados normalmente. El nodo XGBoost-AS se implementa en Spark.

**Nota:** Si selecciona Tree-AS para ejecutarse en un servidor de análisis, fallará y no podrá crear un modelo si hay una nodo de partición en sentido ascendente. En este caso, para conseguir que el clasificador automático funcione con otros nodos de modelado en el servidor de análisis, deseleccione el tipo de modelo Tree-AS.

## Costes de clasificación errónea

En algunos contextos, ciertos tipos de errores son más costosos que otros. Por ejemplo, puede resultar más costoso clasificar a un solicitante de crédito de alto riesgo como de bajo riesgo (un tipo de error) que clasificar a un solicitante de crédito de bajo riesgo como de alto riesgo (otro tipo de error). Los costes de clasificación errónea permiten especificar la importancia relativa de los diversos tipos de errores de predicción.

Los costes de clasificación errónea son básicamente ponderaciones aplicadas a resultados específicos. Estas ponderaciones se extraen en el modelo y pueden realmente cambiar la predicción (como forma de protección frente a errores costosos).

Salvo los modelos C5.0, los costes de clasificación errónea no se aplican cuando se puntúa un modelo y no se tienen en cuenta cuando se clasifican o comparan modelos utilizando un nodo Clasificador automático, un diagrama de evaluación o un nodo Análisis. Es posible que un modelo que incluya costes no produzca menos errores que uno que no lo haga, y es posible que no ordene ningún valor mayor en términos de precisión global, pero es probable que funcione mejor en términos prácticos porque contiene un sesgo integrado en favor de errores *más baratos*.

La matriz de costes muestra el coste para cada combinación posible de categoría predicha y categoría real. De forma predeterminada, todos los costes de clasificación errónea se establecen en 1,0. Para introducir valores de coste personalizados, seleccione **Utilizar costes de clasificación errónea** e introduzca valores personalizados en la matriz de costes.

Para cambiar un coste de clasificación errónea, seleccione la casilla correspondiente a la combinación deseada de valores predichos y reales, elimine el contenido existente de la casilla e introduzca en ella el coste deseado. Los costes no son simétricos automáticamente. Por ejemplo, si establece el coste de clasificación errónea *A* como *B* para que sea 2,0, el coste de clasificación errónea de *B* como *A* aún tendrá el valor predeterminado 1,0 hasta que también se modifique explícitamente.

## Opciones para descartar el nodo Clasificador automático

La pestaña Descartar del nodo Clasificador automático le permite descartar automáticamente los modelos que no cumplen determinados criterios. Estos modelos no aparecerán enumerados en el informe de resumen.

Puede especificar un umbral mínimo para la precisión global y un umbral máximo para el número de variables usadas en el modelo. Además, en el caso de objetivos de marca, puede especificar un umbral mínimo para la elevación, los beneficios y un área debajo de la curva; la elevación y los beneficios se determinan según lo especificado en la pestaña Modelo. Consulte el tema "Opciones de modelo para el nodo Clasificador automático" en la página 68 para obtener más información.

Si lo desea, puede configurar el nodo para que se detenga la ejecución la primera vez que se genere un modelo que cumpla todos los criterios especificados. Consulte el tema "Reglas de parada de nodos de modelado automático" en la página 66 para obtener más información.

## Opciones de configuración del nodo Clasificador automático

La pestaña Configuración del nodo Clasificador automático permite preconfigurar las opciones de puntuación de tiempo disponibles en el nugget.

**Filtrar campos generados por modelos de conjunto.** Elimina los resultados de todos los campos adicionales que generan los modelos individuales que contiene el nodo Conjunto. Seleccione esta casilla de verificación si solamente está interesado en la puntuación combinada de todos los modelos de entrada. Asegúrese de que esta opción no está seleccionada si, por ejemplo, desea utilizar un nodo Análisis o Evaluación para comparar la precisión de la puntuación combinada con la de cada uno de los modelos de entrada individuales.

## Nodo Autonumérico

El nodo Autonumérico calcula y compara los modelos de resultados de rango numérico continuo utilizando varios métodos diferentes, permitiéndole probar una gran variedad de planteamientos en una única ejecución de modelado. Puede seleccionar los algoritmos que se utilizarán y experimentar con múltiples combinaciones de opciones. Por ejemplo, puede predecir valores de viviendas utilizando los modelos Red neuronal, Regresión lineal, C&RT y CHAID para ver cuál tiene el mejor rendimiento; asimismo, puede probar diferentes combinaciones de métodos de regresión Por pasos, Adelante y Hacia atrás. El nodo explora cada combinación posible de opciones, evalúa cada modelo candidato en función de la medida especificada y guarda los mejores para su uso en la puntuación o en futuros análisis. Consulte el tema Capítulo 5, “Nodos de modelado automático”, en la página 65 para obtener más información.

**Ejemplo**

Un municipio desea calcular de forma más precisa el impuesto sobre la propiedad y ajustar los valores de propiedades específicas del modo necesario sin tener que inspeccionar cada propiedad. Mediante el nodo Autonumérico, el analista puede generar y comparar un número de modelos que predicen valores de propiedad basándose en el tipo de edificación, vecindario, tamaño y otros factores conocidos.

**Requisitos**

Un único campo objetivo (con el rol establecido a **Objetivo**), y al menos un campo de entrada (con el rol establecido a **Entrada**). El objetivo debe ser un campo continuo (rango numérico), como *edad* o *ingresos*. Los campos de entrada pueden ser continuos o categóricos, con la limitación de que puede que algunas entradas no sean adecuadas para algunos tipos de modelo. Por ejemplo, los modelos C&RT pueden utilizar campos de cadena categóricos como entradas, mientras que los modelos Regresión lineal no pueden utilizar estos campos y los omitirán si se especifica. Los requisitos son los mismos que cuando se utilizan los nodos de modelado individuales. Por ejemplo, un modelo CHAID funciona igual independientemente de si se ha generado desde el nodo CHAID o el nodo Autonumérico.

**Campos de frecuencia y ponderación**

La frecuencia y la ponderación se utilizan para proporcionar importancia adicional a ciertos registros sobre otros porque, por ejemplo, el usuario sabe que el conjunto de datos creado no representa totalmente una sección de la población principal (Ponderación) o porque un registro representa un número de casos idénticos (Frecuencia). Si se especifica, los algoritmos C&RT y CHAID pueden utilizar un campo de frecuencia. Los algoritmos C&RT, CHAID, Regresión y GenLin pueden utilizar un campo de ponderación. Otros tipos de modelo omitirán estos campos y crearán los modelos de todas formas. Los campos de frecuencia y ponderación sólo se utilizan para la creación de modelos y no se tienen en cuenta al evaluar o puntuar modelos. Consulte el tema “Uso de campos de frecuencia y ponderación” en la página 33 para obtener más información.

**Prefijos**

Si conecta un nodo de tabla al nugget para el nodo Autonumérico, existe varias variables nuevas en la tabla con nombres que empiezan con un prefijo $.

Los nombres de los campos que se generan durante la puntuación se basan en el campo objetivo, pero con un prefijo estándar. Los distintos tipos de modelo utilizan diferentes conjuntos de prefijos.

Por ejemplo, los prefijos $G, $R, $C se utilizan como prefijo para las predicciones que han generado el modelo lineal generalizado, el modelo CHAID y el modelo C5.0, respectivamente. Normalmente, $X se genera utilizando un conjunto, y $XR, $XS y $XF se utilizan como prefijos en los casos donde el campo objetivo es un campo Continuo, Categórico o de Marca, respectivamente.

$.Los prefijos .E se utilizan para la confianza de predicción de un objetivo Continuo; por ejemplo, $XRE se utiliza como prefijo para la confianza de predicción Continua de conjunto. $GE es el prefijo para una sola predicción de confianza para un modelo lineal generalizado.

### Tipos de modelo soportados

Los tipos de modelo soportados incluyen red neuronal, árbol C&R, CHAID, regresión, GenLin, vecino más cercano, SVM, XGBoost Linear, GLE y XGBoost-AS. Si desea obtener más información, consulte "Opciones de experto para el nodo Autonumérico" en la página 76.

## Opciones de modelo para el nodo Autonumérico

La pestaña Modelo del nodo Autonumérico le permite especificar el número de modelos que se van a guardar, junto con los criterios empleados para compararlos.

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Crear modelos divididos.** Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como campos de división. Consulte el tema "Generación de modelos divididos" en la página 28 para obtener más información.

**Clasificar modelos por.** Especifique los criterios utilizados para comparar modelos.

- **Correlación.** Correlación de Pearson entre el valor observado para cada registro y el valor predicho por el modelo. La correlación es una medida de asociación lineal entre dos variables, con valores cercanos a 1 que indican una relación más fuerte. (Los valores de correlación se encuentran en el rango de -1, para una relación negativa perfecta, y +1, para una relación positiva perfecta. El valor 0 indica la ausencia de relaciones lineales, mientras que un modelo con una correlación negativa estaría en el último puesto de la lista.)
- **Número de campos.** Número de campos utilizados como predictores en el modelo. La selección de modelos que utilizan menos campos puede simplificar la preparación de datos y mejorar el rendimiento en algunos casos.
- **Error relativo.** El error relativo es el cociente de la varianza de los valores observados de aquellos predichos por el modelo a la varianza de los valores observados de la media. En la práctica, compara el buen rendimiento del modelo con respecto a un modelo **nulo** o **de interceptación** que simplemente devuelve el valor medio del campo objetivo como la predicción. En un buen modelo, este valor debe ser inferior a 1, lo que indica que el modelo es más preciso que el modelo nulo. Un modelo con un error relativo superior a 1 es menos preciso que el modelo nulo y por lo tanto no es útil. En el caso de modelos Regresión lineal, el error relativo es igual al cuadrado de la correlación y no añade información nueva. En el caso de modelos no lineales, el error relativo no está relacionado con la correlación y proporciona una medida adicional para valorar el rendimiento del modelo.

**Clasificar modelos usando.** Si se está usando una partición, puede especificar si los rangos se basan en la partición de entrenamiento o en la partición de comprobación. En conjuntos de datos de gran tamaño, si usa una partición para el cribado preliminar de modelos, puede mejorar rendimiento en gran medida.

**Número de modelos que se utilizarán.** Especifica el número máximo de modelos que aparecerán en el nugget de modelo generado por el nodo. Los primeros modelos de la clasificación se enumeran en función del criterio de clasificación especificado. El aumento de este límite le permitirá comparar resultados de más modelos pero puede ralentizar el rendimiento. El valor máximo permitido es 100.

**Calcular importancia del predictor.** En el caso de modelos que produzcan una medida adecuada de importancia, puede mostrar un gráfico que indique la importancia relativa de cada predictor al estimar el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Tenga en cuenta que la importancia de predictor puede aumentar el tiempo necesario para calcular algunos modelos; además no se recomienda si sólo desea una amplia comparación entre varios modelos diferentes. Es de mayor utilidad una vez ha limitado su análisis a unos cuantos modelos que desee explorar más a fondo. Consulte el tema "Importancia del predictor" en la página 44 para obtener más información.

**No conservar modelos si.** Especifica valores de umbral para la correlación, el error relativo y el número de campos utilizados. Los modelos que no cumplen alguno de estos criterios se descartarán y no se incluirán en el informe de resumen.

- **Correlación menor que.** Correlación mínima (en cuanto a valor absoluto) para que un modelo se incluya en el informe de resumen.
- **Número de campos utilizados mayor que.** Número máximo de campos que puede utilizar cualquier modelo que vaya a incluirse.
- **Error relativo mayor que.** Error relativo máximo para cualquier modelo que vaya a incluirse.

Si lo desea, puede configurar el nodo para que se detenga la ejecución la primera vez que se genere un modelo que cumpla todos los criterios especificados. Consulte el tema "Reglas de parada de nodos de modelado automático" en la página 66 para obtener más información.

## Opciones de experto para el nodo Autonumérico

La pestaña Experto del nodo Autonumérico le permite seleccionar los algoritmos y opciones que se van a usar y especificar las reglas de parada.

**Seleccionar modelos.** De forma predeterminada, todos los modelos se seleccionan para ser generados; sin embargo, si tiene Analytic Server, puede elegir restringir los modelos a los que se pueden ejecutar en Analytic Server y establecerlos previamente para que creen modelos de división o para que estén listos para procesar conjuntos de datos muy grandes.

**Nota:** No se soporta la creación local de modelos de Analytic Server con el nodo Autonumérico.

**Modelos utilizados.** Use las casillas de verificación de la columna izquierda para seleccionar los tipos de modelo (algoritmos) que se van a incluir en la comparación. Cuantos más tipos seleccione, más modelos se crearán y más tardará el procesamiento.

**Tipo de modelo.** Enumera los algoritmos disponibles (consulte a continuación).

**Parámetros del modelo.** Puede usar la configuración predeterminada o seleccionar **Especificar** para elegir opciones para cada tipo de modelo. Las opciones específicas son parecidas a las disponibles en los nodos de modelado independientes, con la diferencia de que se pueden seleccionar varias opciones o combinaciones. Por ejemplo, si compara los modelos del nodo Red neuronal, puede seleccionar los seis modelos para entrenarlos de una vez en lugar de seleccionar uno de ellos.

**Número de modelos.** Enumera el número de modelos generados para cada algoritmo basados en la configuración actual. Al combinar opciones, puede aumentar rápidamente el número de modelos, por lo que se recomienda prestar especial atención a este número, especialmente si usa conjuntos de datos grandes.

**Limitar el tiempo máximo empleado en generar un único modelo.** (Sólo modelos de K-medias, Kohonen, bietápicos, SVM, KNN, de red bayesiana y de lista de decisiones) Establece un límite de tiempo máximo para cualquier modelo. Por ejemplo, si un modelo determinado necesita un período de tiempo más largo del esperado para entrenarse debido a una interacción compleja, es probable que no quiera detener la ejecución de todo el modelado.

## Algoritmos soportados

El nodo Red neuronal utiliza un modelo simplificado que emula el modo en que el cerebro humano procesa la información: Funciona simultaneando un número elevado de unidades simples de procesamiento interconectadas que parecen versiones abstractas de neuronas. Las redes neuronales son dispositivos eficaces de cálculo de funciones generales y requieren un conocimiento matemático o estadístico mínimo para entrenarlas o aplicarlas.


El nodo de árbol de clasificación y regresión (C&R) genera un árbol de decisión que permite predecir o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera "puro" si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos (nominal, ordinal o marca). Todas las divisiones son binarias (sólo se crean dos subgrupos).


El nodo CHAID genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. A diferencia de los nodos C&RT y QUEST, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles, aunque necesita más tiempo para realizar los cálculos.

La regresión lineal es una técnica de estadístico común utilizada para resumir datos y realizar predicciones ajustando una superficie o línea recta que minimice las discrepancias existentes entre los valores de salida reales y los predichos.

El modelo lineal generalizado amplía el modelo lineal general, de manera que la variable dependiente está relacionada linealmente con los factores y las covariables mediante una determinada función de enlace. Además, el modelo permite que la variable dependiente tenga una distribución no normal. Cubre la funcionalidad de un amplio número de modelo estadísticos, incluyendo regresión lineal, regresión logística, modelos log lineales para recuento de datos y modelos de supervivencia censurados por intervalos.

El nodo *k* de modelado de vecino (KNN) asocia el nuevo caso con la categoría o valor de los objetos *k* junto a él en el espacio de predictores, donde *k* es un entero. Los casos parecidos están próximos y los que no lo son están alejados entre sí.

El nodo Máquina de vectores de soporte (SVM) le permite clasificar datos en uno o dos grupos sin que haya un ajuste por exceso. SVM funciona bien con conjuntos de datos grandes, como aquellos con un gran número de campos de entrada.

Los modelos de regresión lineal predicen un objetivo continuo tomando como base las relaciones lineales entre el destino y uno o más predictores.

El nodo Máquina de vectores de soporte lineal (LSVM) le permite clasificar datos en uno o dos grupos sin que haya un ajuste por exceso. LSVM es lineal y funciona bien con conjuntos de datos grandes, como aquellos con un gran número de registros.

El nodo Árboles aleatorios es similar al nodo C&RT existente; el nodo Árboles aleatorios se diseñado para procesar grandes cantidades de datos (Big Data) para crear un único árbol y mostrar el modelo resultante en el visor de la salida que se ha añadido en SPSS Modeler versión 17. El nodo del árbol Árboles aleatorios genera un árbol de decisiones que se utiliza para predecir o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera *puro* si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos (nominal, ordinal o marca). Todas las divisiones son binarias (sólo se crean dos subgrupos).

El nodo Tree-AS es similar al nodo CHAID existente, el nodo Tree-AS se ha diseñado para procesar grandes cantidades de datos (Big Data) para crear un solo árbol y muestra el modelo resultante en el visor de la salida que se ha añadido en SPSS Modeler versión 17. El nodo genera un árbol de decisiones usando un estadístico de chi-cuadrado (CHAID) para identificar las divisiones óptimas. Este uso de CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles, aunque necesita más tiempo para realizar los cálculos.

XGBoost Linear© es una implementación avanzada de un algoritmo de aumento de gradiente con un modelo lineal como modelo base. Los algoritmos de aumento conocen de forma iterativa los clasificadores débiles y, a continuación, los añaden a un clasificador fuerte final. El nodo XGBoost Linear en SPSS Modeler se implementa en Python.

Un GLE amplía el modelo lineal de forma que el objetivo puede tener una distribución no normal, está relacionado de forma lineal con los factores y las covariables a través de una función de enlace especificada y las observaciones se pueden correlacionar. Los modelos lineales mixtos generalizados cubren una amplia variedad de modelos, desde modelos de regresión lineal simple hasta modelos multinivel complejos para datos longitudinales no normales.


XGBoost© es una implementación avanzada de un algoritmo de aumento de gradiente. Los algoritmos de aumento conocen de forma iterativa los clasificadores débiles y, a continuación, los añaden a un clasificador fuerte final. XGBoost es muy flexible y proporciona muchos parámetros que pueden resultar abrumadores para la mayoría de los usuarios, así que el nodo XGBoost-AS en SPSS Modeler expone las características principales y los parámetros utilizados normalmente. El nodo XGBoost-AS se implementa en Spark.

## Opciones de configuración para el nodo Autonumérico

La pestaña Configuración del nodo Autonumérico permite preconfigurar las opciones de puntuación de tiempo disponibles en el nugget.

**Filtrar campos generados por modelos de conjunto.** Elimina los resultados de todos los campos adicionales que generan los modelos individuales que contiene el nodo Conjunto. Seleccione esta casilla de verificación si solamente está interesado en la puntuación combinada de todos los modelos de entrada.

Asegúrese de que esta opción no está seleccionada si, por ejemplo, desea utilizar un nodo Análisis o Evaluación para comparar la precisión de la puntuación combinada con la de cada uno de los modelos de entrada individuales.

**Calcular error estándar.** Para un objetivo continuo (rango numérico), se ejecuta un error estándar de forma predeterminada para calcular la diferencia entre los valores medidos o estimados y los valores true; y para mostrar si las estimaciones coinciden.

## Nodo Autoclúster

El nodo Agrupación en clústeres automática calcula y compara los modelos de agrupación en clústeres que identifican grupos de registros con características similares. El nodo funciona de la misma manera que otros nodos de modelado automático, permitiéndole experimentar con varias combinaciones de opciones en una única pasada de modelado. Los modelos se pueden comparar utilizando medidas básicas con las que se intenta filtrar y definir la utilidad de los modelos de clúster y proporcionar una medida según la importancia de campos concretos.

Los modelos de agrupación en clústeres se suelen identificar con grupos que se pueden utilizar como entradas en futuros análisis. Por ejemplo, es posible que desee dirigirse a grupos de clientes según sus características demográficas como ingresos, o según los servicios que hayan contratado en el pasado. Esto puede hacerse si se tiene un conocimiento previo de los grupos y sus características; es posible que no sepa en cuántos grupos buscar o las características que debe utilizar para definirlos. Los modelos de agrupación en clústeres se suelen definir como modelos de aprendizaje no supervisado, ya que no utilizan un campo de destino y no devuelven una predicción específica que se pueda evaluar como true o false. El valor de un modelo de agrupación en clústeres viene determinado por su capacidad de capturar agrupaciones interesantes en los datos y proporcionar descripciones útiles de dichas agrupaciones. Consulte Capítulo 11, “Modelos de agrupación en clústeres”, en la página 243 si desea obtener más información.

**Requisitos.** Uno o más campos que definen las características de interés. Los modelos de clúster no utilizan campos objetivo de la misma manera que otros modelos, porque no realizan predicciones específicas que se pueden evaluar como true o false. En su lugar, se utilizan para identificar grupos de casos que pueden estar relacionados. Por ejemplo, no puede utilizar un modelo de clúster para predecir si un cliente concreto abandonará o responderá a una oferta. Pero puede utilizar un modelo de clúster para asignar clientes a grupos en función de su tendencia a hacer determinadas cosas. Los campos de ponderación y frecuencia no se usan.

**Campos de evaluación.** Mientras no utilice un objetivo, puede especificar uno o más campos de evaluación que se utilizarán en la comparación de modelos. La utilidad de un modelo de clúster se puede evaluar lo bien (o mal) que los clústeres diferencian los campos.

### Tipos de modelo soportados

Los tipos de modelo soportados incluyen Dos fases, K-Medias, Kohonen, SVM de una clase y K-Medias-AS.

## Opciones de modelo para el nodo Agrupación en clústeres automática

La pestaña Modelo del nodo Agrupación en clústeres automática le permite especificar el número de modelos que se deben guardar, junto con los criterios empleados para comparar modelos.

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Clasificar modelos por.** Especifique los criterios utilizados para comparar y clasificar los modelos.

- **Silueta.** Un índice que mide la cohesión y separación del clúster. Consulte *Medida de clasificación de siluetas* a continuación para obtener más información.
- **Número de clústeres.** El número de clústeres que se utilizan en el modelo.
- **Tamaño del clúster más pequeño.** El menor tamaño de clúster.
- **Tamaño del clúster mayor.** El mayor tamaño de clúster.
- **Clúster mayor / menor.** La razón del tamaño del clúster menor y el mayor.
- **Importancia.** La importancia del campo **Evaluación** en la pestaña **Campos**. Tenga en cuenta que sólo se puede calcular si se ha especificado un campo **Evaluación**.

**Clasificar modelos usando.** Si se está usando una partición, puede especificar si los rangos se basan en el conjunto de datos de entrenamiento o en el conjunto de prueba. En conjuntos de datos de gran tamaño, si usa una partición para el cribado preliminar de modelos, puede mejorar rendimiento en gran medida.

**Número de modelos que se mantendrán.** Especifica el número máximo de modelos que aparecerán en el nugget generado por el nodo. Los primeros modelos de la clasificación se enumeran en función del criterio de clasificación especificado. Tenga en cuenta que si aumenta este límite puede ralentizarse el rendimiento. El valor máximo permitido es 100.

Medida de clasificación de siluetas

La medida de clasificación predeterminada, Silueta, tiene un valor predeterminado de 0 porque un valor inferior a 0 (es decir, negativo) indica que la distancia media entre un caso y los puntos de su clúster asignado es mayor que la distancia media mínima hasta los puntos de otro clúster. Por lo tanto, los modelos con una Silueta negativa pueden descartarse de manera segura.

La medida de clasificación es de hecho un coeficiente de silueta modificado, que combina los conceptos de cohesión de clústeres (favoreciendo a los modelos que contengan clústeres cohesivos) y separación de clústeres (favoreciendo a los modelos que contengan clústeres altamente separados). El coeficiente de Silueta medio es simplemente la media de todos los casos del siguiente cálculo por cada caso individual:

$(B - A) / \max(A, B)$

donde $A$ es la distancia desde el caso hasta el centroide del clúster al que pertenece el caso; y $B$ es la distancia mínima desde el caso hasta el centroide de cada uno de los otros clústeres.

El coeficiente de Silueta (y su media) van desde -1 (lo que indica un modelo muy pobre) hasta 1 (lo que indica un modelo excelente). La media puede realizarse a nivel de casos totales (lo cual produce una silueta total) o a nivel de clústeres (lo cual produce una silueta de clústeres). Las distancias pueden calcularse utilizando distancias euclídeas.

## Opciones de experto para el nodo Agrupación en clústeres automática

La pestaña Experto del nodo Agrupación en clústeres automática le permite aplicar una partición (si está disponible), seleccionar los algoritmos que se deben utilizar y especificar las reglas de parada.

**Seleccionar modelos.** De forma predeterminada, todos los modelos se seleccionan para ser generados; sin embargo, si tiene Analytic Server, puede elegir restringir los modelos a los que se pueden ejecutar en Analytic Server y establecerlos previamente para que creen modelos de división o para que estén listos para procesar conjuntos de datos muy grandes.

**Nota:** No está soportada la creación local de modelos de Analytic Server dentro del nodo Clúster automático. .

**Modelos utilizados.** Use las casillas de verificación de la columna izquierda para seleccionar los tipos de modelo (algoritmos) que se van a incluir en la comparación. Cuantos más tipos seleccione, más modelos se crearán y más tardará el procesamiento.

**Tipo de modelo.** Enumera los algoritmos disponibles (consulte a continuación).

**Parámetros del modelo.** Puede usar la configuración predeterminada o seleccionar **Especificar** para elegir opciones para cada tipo de modelo. Las opciones específicas son parecidas a las disponibles en los nodos de modelado independientes, con la diferencia de que se pueden seleccionar varias opciones o combinaciones. Por ejemplo, si compara los modelos del nodo Red neuronal, puede seleccionar los seis modelos para entrenarlos de una vez en lugar de seleccionar uno de ellos.

**Número de modelos.** Enumera el número de modelos generados para cada algoritmo basados en la configuración actual. Al combinar opciones, puede aumentar rápidamente el número de modelos, por lo que se recomienda prestar especial atención a este número, especialmente si usa conjuntos de datos grandes.

**Limitar el tiempo máximo empleado en generar un único modelo.** (Sólo modelos de K-medias, Kohonen, bietápicos, SVM, KNN, de red bayesiana y de lista de decisiones) Establece un límite de tiempo máximo para cualquier modelo. Por ejemplo, si un modelo determinado necesita un período de tiempo más largo del esperado para entrenarse debido a una interacción compleja, es probable que no quiera detener la ejecución de todo el modelado.

## Algoritmos soportados

El nodo K-medias agrupa conjuntos de datos en grupos distintos (o clústeres). El método define un número fijo de clústeres, de forma iterativa asigna registros a los clústeres y ajusta los centros de los clústeres hasta que no se pueda mejorar el modelo. En lugar de intentar predecir un resultado, los modelos de *k*-medias utilizan un proceso conocido como aprendizaje no supervisado para revelar los patrones del conjunto de campos de entrada.

El nodo Kohonen genera un tipo de red neuronal que se puede usar para agrupar un conjunto de datos en grupos distintos. Cuando la red se termina de entrenar, los registros que son similares se deberían cerrar juntos en el mapa de resultados, mientras que los registros que son diferentes aparecerían aparte. Puede observar el número de observaciones capturadas por cada unidad en el nugget de modelo para identificar unidades fuertes. Esto le proporcionará una idea del número apropiado de clústeres.

El nodo Bietápico es un método de agrupación en clústeres de dos pasos. El primer paso es hacer una única pasada por los datos para comprimir los datos de entrada de la fila en un conjunto de subclústeres administrable. El segundo paso utiliza un método de agrupación en clústeres jerárquica para fundir progresivamente los subclústeres en clústeres cada vez más grandes. El bietápico tiene la ventaja de estimar automáticamente el número óptimo de clústeres para los datos de entrenamiento. Puede gestionar tipos de campos mixtos y grandes conjuntos de datos eficazmente.

El nodo SVM de una clase utiliza un algoritmo de aprendizaje no supervisado. El nodo se puede utilizar para la detección de novedad. Detectará el límite flexible de un conjunto de muestras proporcionado, para clasificar a continuación los puntos nuevos como pertenecientes o no a dicho conjunto. Este nodo de modelado SVM de una clase en SPSS Modeler se implementa en Python y necesita la biblioteca `scikit-learn`© de Python.

k-medias es uno de los algoritmos de agrupación en clúster utilizado con más frecuencia. Agrupa en clúster puntos de datos en un número predefinido de clústeres. El nodo K-Medias-AS en SPSS Modeler se implementa en Spark. Si desea más detalles sobre algoritmos de k-medias, consulte https://spark.apache.org/docs/2.2.0/ml-clustering.html.Tenga en cuenta que el nodo K-Medias-AS realiza una codificación "one-hot" automáticamente para variables categóricas.

## Opciones para descartar del nodo Agrupación en clústeres automática

La pestaña Descartar del nodo Agrupación en clústeres automática le permite descartar automáticamente los modelos que no cumplen determinados criterios. Estos modelos no aparecerán enumerados en el nugget de modelo.

Puede especificar el valor mínimo de silueta, número de clústeres, tamaños de clústeres y la importancia del campo de evaluación en el modelo. La silueta y el número y tamaño de clústeres están determinados en la especificación del nodo de modelado. Consulte el tema “Opciones de modelo para el nodo Agrupación en clústeres automática” en la página 79 para obtener más información.

Si lo desea, puede configurar el nodo para que se detenga la ejecución la primera vez que se genere un modelo que cumpla todos los criterios especificados. Consulte el tema “Reglas de parada de nodos de modelado automático” en la página 66 para obtener más información.

## Nugget de modelo automático

Cuando se ejecuta el nodo de modelado automático, el nodo estima modelos candidatos de cada combinación de opciones posible, clasifica cada modelo de candidato en función de la medida que especifique y guarda los mejores modelos en un nugget de modelo automático compuesto. Este nugget de modelo contiene un conjunto de uno o más modelos que genera el nodo, que se pueden examinar o seleccionar individualmente para la puntuación. El tipo de modelo y el tiempo de generación se incluyen para cada modelo, junto con un número de otras mediciones que resulten adecuadas para el modelo. Puede ordenar la tabla en cualquiera de estas columnas para identificar rápidamente los modelos más interesantes.

- Para examinar cualquiera de los nugget de modelo individuales, pulse dos veces en el icono del nugget. A partir de aquí, puede generar un nodo de modelado para ese modelo en el lienzo de rutas, o una copia del nugget de modelo en la paleta de modelos.
- Los gráficos de miniatura ofrecen una rápida evaluación visual de cada modelo, tal y como se resume a continuación. Puede pulsar dos veces en una miniatura para generar un gráfico a tamaño completo. El gráfico a tamaño completo muestra hasta 1.000 puntos y se basará en una muestra si el conjunto de datos contiene más. (Sólo en el caso de los diagramas de dispersión, el gráfico se regenera cada vez que se muestra, de modo que los cambios en los datos anteriores en la ruta, por ejemplo la actualización de una muestra aleatoria o partición si **Establecer semilla aleatoria** no está seleccionado, puedan reflejarse cada vez que se vuelva a dibujar el diagrama de dispersión.]
- Use la barra de herramientas para mostrar u ocultar columnas específicas de la pestaña Modelo o cambiar la columna usada para ordenar la tabla. (También puede cambiar el orden pulsando en las cabeceras de columna.)
- Utilice el botón Eliminar para eliminar permanentemente los modelos no utilizados.
- Para reorganizar columnas, pulse en la cabecera de una columna y arrastre la columna a la ubicación que desee.
- Si se está usando una partición, puede optar por ver los resultados de la partición de comprobación o entrenamiento según proceda.

Las columnas específicas dependen del tipo de modelos que se estén comparando, como se indica a continuación.

Objetivos binarios

- En el caso de modelos binarios, la miniatura muestra la distribución de valores reales, superpuestos con los valores previstos, para ofrecer una rápida indicación visual de cuántos registros se predeciron correctamente en cada categoría.
- Los criterios de clasificación coinciden con las opciones del nodo de modelado Clasificador automático. Consulte el tema "Opciones de modelo para el nodo Clasificador automático" en la página 68 para obtener más información.
- Para obtener el máximo beneficio, también aparece en el informe el percentil en el que se produce el valor máximo.
- En el caso de la elevación acumulada, puede cambiar el percentil seleccionado mediante la barra de herramientas.

Objetivos nominales

- En el caso de modelos nominales (conjunto), la miniatura muestra la distribución de valores reales, superpuestos con los valores previstos, para ofrecer una rápida indicación visual de cuántos registros se predeciron correctamente en cada categoría.
- Los criterios de clasificación coinciden con las opciones del nodo de modelado Clasificador automático. Consulte el tema "Opciones de modelo para el nodo Clasificador automático" en la página 68 para obtener más información.

Objetivos continuos

- En el caso de modelos continuos (rango numérico), el gráfico representa los valores predichos frente a los valores observados de cada modelo, lo que ofrece una rápida indicación visual de la correlación entre ellos. En el caso de un buen modelo, los puntos tienden a agruparse en la diagonal en lugar de estar dispersos aleatoriamente por el gráfico.
- Los criterios de clasificación coinciden con las opciones del nodo de modelado Autonumérico. Consulte el tema "Opciones de modelo para el nodo Autonumérico" en la página 75 para obtener más información.

Objetivos de clúster

- En el caso de modelos de clúster, el gráfico representa los recuentos frente a los clústeres para cada modelo, lo que ofrece una rápida indicación visual de la distribución de clústeres.
- Los criterios de clasificación coinciden con las opciones del nodo de autoclúster. Consulte el tema "Opciones de modelo para el nodo Agrupación en clústeres automática" en la página 79 para obtener más información.

Selección de modelos para puntuación

La columna **Uso?** le permite seleccionar los modelos que se utilizarán en la puntuación.

- En objetivos numéricos, nominales y binarios, podrá seleccionar múltiples modelos de puntuación y combinar los resultados en el nugget de modelo de conjunto único. Al combinar predicciones de varios modelos, pueden evitarse las limitaciones en modelos individuales que suelen dar como resultado una precisión global superior que puede obtenerse de cualquiera de los modelos.
- En modelos de clúster sólo puede seleccionar un modelo de puntuación cada vez. De forma predeterminada, el primer clasificado se selecciona primero.

## Generación de nodos y modelos

Puede generar una copia del nugget de modelo automático compuesto o el nodo de modelado automático a partir del que se generó. Por ejemplo, esto puede ser de utilidad si no tiene la ruta original a partir de la que se generó el nugget de modelo automático. También puede generar un nugget o nodo de modelado para cualquiera de los modelos individuales enumerados en el nugget de modelo automático.

Nugget de modelado automático

En el menú Generar, seleccione **Modelo(s) a paleta** para añadir el nugget de modelo automático a la paleta de modelos. Se puede guardar o usar cada modelo generado tal cual sin volver a ejecutar la ruta.

Si lo prefiere, puede seleccionar **Generar nodo de modelado** en el menú Generar para añadir el nodo de modelado al lienzo de rutas. Se puede para volver a estimar los modelos seleccionados sin repetir la ejecución de todo el modelado del clasificador binario.

Nugget de modelado individual

1. En el menú **Modelo**, pulse dos veces en el nugget individual que necesite. Una copia de dicho nugget se abrirá en un nuevo cuadro de diálogo.
2. En el menú Generar del nuevo cuadro de diálogo, seleccione **Modelo(s) a paleta** para añadir el nugget de modelado individual a la paleta de modelos.
3. Si lo prefiere, puede seleccionar **Generar nodo de modelado** en el menú Generar del nuevo cuadro de diálogo para añadir el nodo de modelado individual al lienzo de rutas.

## Generación de diagramas de evaluación

Sólo en el caso de modelos binarios, puede generar diagramas de evaluación que ofrecen un modo visual de valorar y comparar el rendimiento de cada modelo. Los diagramas de evaluación no están disponibles para modelos generados por los nodos Autonumérico o Autoclúster.

1. En la columna *Uso?* del nugget de modelo automático Clasificador automático, seleccione los modelos que desee evaluar.
2. En el menú Generar, seleccione **Diagrama(s) de evaluación**. Se visualiza el cuadro de diálogo Diagrama de evaluación.
3. Seleccione el tipo de diagrama y otras opciones que desee.

## Diagramas de evaluación

En la pestaña Modelo del nugget de modelo automático, puede profundizar para visualizar gráficos individuales de cada uno de los modelos que se muestran. En el caso de los nugget Clasificador automático y Autonumérico, la pestaña Gráfico muestra tanto un gráfico como la importancia de predictor que reflejan los resultados de todos los modelos combinados. Consulte el tema “Importancia del predictor” en la página 44 para obtener más información.

En el caso de Clasificador automático, se muestra un gráfico de distribución, mientras que para Autonumérico, se muestra un gráfico múltiple (también denominado diagrama de dispersión).

# Capítulo 6. Árboles de decisión

## Modelos de árboles de decisión

Utilice modelos de árbol de decisión para desarrollar sistemas de clasificación que predicen o clasifican observaciones futuras basándose en un conjunto de reglas de decisión. Si dispone de datos divididos en clases que le interesan (por ejemplo, préstamos de alto riesgo frente a préstamos de bajo riesgo, suscriptores frente a no suscriptores, votantes frente a no votantes o tipos de bacterias), puede usar los datos para generar reglas que pueda usar para clasificar casos antiguos o recientes con la máxima precisión. Por ejemplo, podría generar un árbol que clasificara el riesgo de crédito o la intención de compra basándose en la edad y otros factores.

Este método, a veces conocido como *inducción de regla*, presenta varias ventajas. Primero, el proceso de razonamiento detrás del modelo resulta claramente evidente cuando se examina el árbol. Esto contrasta con otras técnicas de modelado de *caja negra*, en las que la lógica interna puede resultar difícil de averiguar.

En segundo lugar, el proceso incluye automáticamente en su regla únicamente los atributos que realmente importan en la toma decisiones. Los atributos que no contribuyan a la precisión del árbol se omiten. Esto puede proporcionar información de gran utilidad acerca de los datos y se puede usar para reducir los datos a campos relevantes antes de entrenar otra técnica de aprendizaje, como una red neuronal.

Los nugget de modelo de árbol de decisión se pueden convertir en una colección de reglas if-then (un *conjunto de reglas*), que en muchos casos muestra la información de forma más comprensible. La presentación del árbol de decisión resulta útil cuando se desea ver el modo en que los atributos de los datos pueden *dividir* o *particionar* la población en subconjuntos relevantes para el problema. La salida del nodo Tree-AS es diferente de los otros nodos Árbol de decisión porque incluye una lista de reglas directamente en el nugget sin tener que crear un conjunto de reglas. La presentación del conjunto de reglas resulta de utilidad si se desea ver el modo en que determinados grupos de elementos se vinculan a una conclusión particular. Por ejemplo, la siguiente regla proporciona un *perfil* de un grupo de vehículos que merece la pena comprar:

```
IF tested = 'yes'
AND kilometraje = 'bajo'
THEN -> 'BUY'.
```

### Algoritmos de generación de árboles

Hay varios algoritmos disponibles para realizar un análisis de segmentación y clasificación. Todos estos algoritmos son básicamente similares: examinan todos los campos del conjunto de datos para detectar el que proporciona la mejor clasificación o predicción dividiendo los datos en subgrupos. El proceso se aplica de forma recursiva, dividiendo los subgrupos en unidades cada vez más pequeñas hasta completar el árbol (según defina determinados criterios de parada). Los campos objetivo y de entrada utilizados en la generación del árbol pueden ser continuos (rango numérico) o categóricos, dependiendo del algoritmo que se utilice. Si se usa un objetivo continuo, se genera un árbol de regresión; si se usa un objetivo categórico, se genera un árbol de clasificación.


El nodo de árbol de clasificación y regresión (C&R) genera un árbol de decisión que permite predecir o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera "puro" si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos (nominal, ordinal o marca). Todas las divisiones son binarias (sólo se crean dos subgrupos).


El nodo CHAID genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. A diferencia de los nodos C&RT y QUEST, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles, aunque necesita más tiempo para realizar los cálculos.


El nodo QUEST proporciona un método de clasificación binario para generar árboles de decisión; está diseñado para reducir el tiempo de procesamiento necesario para realizar los análisis de C&RT y reducir la tendencia de los métodos de clasificación de árboles para favorecer a las entradas que permitan realizar más divisiones. Los campos de entrada pueden ser continuos (rango numérico), sin embargo el campo objetivo debe ser categórico. Todas las divisiones son binarias.


El nodo C5.0 genera un árbol de decisión o un conjunto de reglas. El modelo divide la muestra basándose en el campo que ofrece la máxima ganancia de información en cada nivel. El campo objetivo debe ser categórico. Se permiten varias divisiones en más de dos subgrupos.

El nodo Tree-AS es similar al nodo CHAID existente, el nodo Tree-AS se ha diseñado para procesar grandes cantidades de datos (Big Data) para crear un solo árbol y muestra el modelo resultante en el visor de la salida que se ha añadido en SPSS Modeler versión 17. El nodo genera un árbol de decisiones usando un estadístico de chi-cuadrado (CHAID) para identificar las divisiones óptimas. Este uso de CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles, aunque necesita más tiempo para realizar los cálculos.

El nodo Árboles aleatorios es similar al nodo C&RT existente; el nodo Árboles aleatorios se diseñado para procesar grandes cantidades de datos (Big Data) para crear un único árbol y mostrar el modelo resultante en el visor de la salida que se ha añadido en SPSS Modeler versión 17. El nodo del árbol Árboles aleatorios genera un árbol de decisiones que se utiliza para predecir o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera *puro* si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos (nominal, ordinal o marca). Todas las divisiones son binarias (sólo se crean dos subgrupos).

## Usos generales del análisis basado en árboles

A continuación se detallan algunos usos generales del análisis basado en árboles:

*Segmentación* Identifique personas que probablemente sean miembros de una clase concreta.

*Estratificación* Asigna los casos a una categoría de entre varias, por ejemplo, grupos de alto riesgo, bajo riesgo y riesgo intermedio.

*Predicción* Crea reglas y las utiliza para predecir eventos futuros. Las predicciones también pueden significar intentos de relacionar atributos predictivos con valores de una variable continua.

*Reducción de datos y clasificación de variables* Selecciona un subconjunto útil de predictores a partir de un gran conjunto de variables para utilizarlo en la creación de un modelo paramétrico formal.

*Identificación de interacción* Identifique relaciones que solo pertenecen a subgrupos específicos y especifique estas relaciones en un modelo paramétrico formal.

*Fusión de categorías y creación de tramos de variables continuas* Vuelve a codificar las variables continuas y las categorías de los predictores del grupo, con una pérdida mínima de información.

## El Generador de árboles interactivos

Puede generar un modelo de árbol automáticamente, en el que el algoritmo decida la división más adecuada para cada nivel, o bien, puede utilizar el generador de árboles interactivos para tomar el control, aplicando sus conocimientos empresariales para refinar o simplificar el árbol antes de guardar el nugget de modelo.

1. Cree una ruta y añada uno de los nodos de los árboles de decisión C&R, CHAID o QUEST.

   **Nota:** los árboles C5.0 y Tree-AS no admiten la generación de árboles interactivos.
2. Abra el nodo y, en la pestaña Campos, seleccione los campos de destino y predictores y especifique las opciones de modelo adicionales que sean necesarias. Para obtener instrucciones específicas, consulte la documentación de los distintos nodos de generación de árboles.
3. En el panel Objetivos de la pestaña Opciones de generación, seleccione **Iniciar sesión interactiva**.
4. Pulse en **Ejecutar** para iniciar el generador de árboles.

Se muestra el árbol actual desde el nodo raíz. Antes de generar uno o varios modelos, puede editar y podar el árbol nivel a nivel y acceder a ganancias, riesgos e información relacionada.

### Comentarios

- Con los nodos Árbol C&R, CHAID y QUEST, todos los campos ordinales que se utilizan en el modelo deberán contar con un almacenamiento numérico (y no en cadenas). Si lo considera necesario, utilice a continuación el nodo Reclasificar para realizar las conversiones.
- Si lo desea, puede utilizar un campo de partición para separar los datos de las muestras de comprobación y entrenamiento.
- Como alternativa al generador de árboles, también puede generar el modelo directamente desde el nodo de modelado, al igual que con otros modelos de IBM SPSS Modeler. Consulte el tema "Creación directa de un modelo de árbol" en la página 98 para obtener más información.

## Desarrollo y poda del árbol

La pestaña Visor del generador de árboles le permite ver el árbol actual, empezando con el nodo raíz.

1. Para hacer crecer el árbol, seleccione en los menús:

   **Árbol** > **Hacer crecer árbol**

   El sistema genera el árbol mediante la división recursiva de las distintas ramas hasta que se cumplen uno o varios criterios de parada. En cada división se selecciona automáticamente el mejor predictor, dependiendo del método de modelado utilizado.
2. Si lo prefiere, también puede seleccionar **Hacer crecer árbol un nivel** para añadir un solo nivel.
3. Para añadir una rama bajo un nodo específico, seleccione el nodo y, a continuación, **Hacer crecer rama**.
4. Para seleccionar el predictor que debe utilizarse en una división, seleccione el nodo que desee y, a continuación, **Hacer crecer rama con división personalizada**. Consulte el tema "Definición de divisiones personalizadas" en la página 88 para obtener más información.
5. Para podar una rama, seleccione un nodo y seleccione **Eliminar rama** para borrar el nodo seleccionado.
6. Para eliminar el nivel inferior del árbol, seleccione **Eliminar un nivel**.

7. Exclusivamente para los nodos Árbol C&R y QUEST, seleccione **Hacer crecer árbol y podar** para podar de acuerdo con un algoritmo de coste-complejidad que ajusta la estimación del riesgo en función del número de nodos terminales, y que suele generar un árbol más simple. Consulte el tema "Nodo Árbol C&R" en la página 100 para obtener más información.

Lectura de reglas divididas en la pestaña Visor

Al visualizar reglas divididas en la pestaña Visor, los corchetes significan que el valor adyacente se incluye en el rango mientras que los paréntesis indican que el valor adyacente se excluye del rango. Por lo tanto, la expresión (23,37] significa que el rango va desde el 23 exclusive hasta el 37 inclusive; es decir, desde el 24 hasta el 37. En la pestaña Modelo, la misma situación se mostraría como:

```
Edad > 23 y Edad <= 37
```

**Interrupción del crecimiento de los árboles.** Para interrumpir una operación de crecimiento de árboles (cuando tarda más de lo esperado, por ejemplo), pulse en el botón Detener ejecución de la barra de herramientas.

*Figura 28. Botón Detener ejecución*

El botón solamente se activa durante la generación del árbol. Detiene la operación de desarrollo en curso en el punto en que se encuentra: se conservan los nodos añadidos, no se guardan cambios, se cierra la ventana, etc. El generador de árboles permanece abierto y permite generar un modelo, actualizar directivas o exportar resultados en el formato adecuado, según sea necesario.

## Definición de divisiones personalizadas

En el cuadro de diálogo Definir división le permite seleccionar el predictor y especificar las condiciones de cada división.

1. En el generador de árboles, seleccione un nodo en la pestaña Visor y, en los menús, seleccione:
   **Árbol** > **Hacer crecer rama con división personalizada**
2. Seleccione el predictor que desee en la lista desplegable, o bien, pulse en el botón **Predictores** para ver detalles acerca de los distintos predictores. Consulte el tema "Visualización de detalles de predictores" en la página 89 para obtener más información.
3. Puede aceptar las condiciones predeterminadas de las divisiones, o bien, seleccionar **Personalizado** para especificar las condiciones que considere adecuadas para las divisiones.
   - En predictores continuos (rangos numérico), puede utilizar los campos **Editar valores de rango** para especificar el intervalo de valores que caen en cada nuevo nodo.
   - En predictores categóricos, puede utilizar los campos **Editar valores de conjunto** o **Editar valores ordinales** para especificar los valores (o intervalo de valores en caso de un predictor ordinal) que se correlacionen a cada nuevo nodo.
4. Seleccione **Hacer crecer** para que la rama vuelva a crecer con el predictor seleccionado.

Por lo general, el árbol puede dividirse con cualquiera de los predictores, independientemente de las reglas de parada. Las únicas excepciones se producen si un nodo es puro (donde el 100% de los casos corresponde a la misma clase objetivo y no quedan elementos para dividir) o cuando el predictor seleccionado es constante (no quedan elementos frente a los cuales dividir).

**Valores perdidos en.** Únicamente con los árboles CHAID, cuando existen valores perdidos disponibles para un predictor determinado, tiene la opción de definir una división personalizada para asignar estos valores a un nodo hijo específico. (Con C&RT y QUEST, los valores perdidos se gestionan mediante sustitutos, tal como se define en el algoritmo.) Consulte el tema "Sustitutos y detalles de la división" en la página 89 para obtener más información. )

### Visualización de detalles de predictores

El cuadro de diálogo Seleccionar predictor muestra los estadísticos de los predictores disponibles (o también a veces denominados "competidores") que se pueden utilizar en la división actual.

- Para CHAID y CHAID exhaustivo, se muestran los estadísticos de chi-cuadrado para los distintos predictores categóricos, si un predictor es un rango numérico, se muestra el estadístico *F*. El estadístico de chi-cuadrado es una medida de la independencia del campo objetivo respecto al campo de división. Unas estadísticas de chi-cuadrado elevado normalmente están relacionadas con una probabilidad más baja, lo que significa que es poco probable que ambos campos sean independientes; lo que indicaría que es una buena división. También se incluyen los grados de libertad porque estos tienen en cuenta el hecho de que es más fácil que una división de tres factores tenga un estadístico elevado y una probabilidad pequeña que una división de dos factores.
- Para C&RT y QUEST, se muestra la mejora de los distintos predictores. Cuanto mayor es la mejora, más se reduce la impureza entre los nodos padre e hijo cuando se utiliza dicho predictor. (Un nodo puro es aquel en que todos los casos corresponden a una sola categoría objetivo; cuanto menor es la impureza a través del árbol, mejor se ajusta el modelo a los datos.) En otras palabras, una cifra de gran mejora normalmente indica una división de utilidad para este tipo de árbol. La medida de impureza utilizada se especifica en el nodo de generación de árboles.

## Sustitutos y detalles de la división

Puede seleccionar cualquier nodo de la pestaña Visor y seleccionar el botón Mostrar información de división en la parte derecha de la barra de herramientas para ver los detalles acerca de la división del nodo. Se muestra la regla de división utilizada junto con los estadísticos relevantes. Para los árboles categóricos C&RT, se muestran la mejora y la asociación. La asociación es una medida de correspondencia entre un sustituto y el campo de división principal, donde normalmente el “mejor” sustituto es aquel que mejor imita al campo de división. Para C&RT y QUEST, también se mostrarán todos los sustitutos que se hayan utilizado en lugar del predictor principal.

Para editar la división del nodo seleccionado, pulse en el icono situado en la parte izquierda del panel de sustitutos para abrir el cuadro de diálogo Definir división. (Cómo atajo, puede seleccionar un sustituto de la lista antes de pulsar en el icono para seleccionarlo como campo de división principal.)

**Sustitutos.** Si procede, se muestra cualquier sustituto del campo de división principal para el nodo seleccionado. Los sustitutos son campos alternativos que se usan en caso de que el valor predictor principal no esté presente en un determinado registro. El número máximo de sustitutos permitido para una división en particular se especifica en el nodo de generación de árbol, pero el número real depende de los datos de entrenamiento. En general, cuanto mayor sea la cantidad de datos perdidos, mayor será la probabilidad de usar sustitutos. En otros modelos de árboles de decisión esta pestaña está vacía.

**Nota:** Para que se incluya en el modelo, los sustitutos se deben identificar durante la fase de entrenamiento. Si la muestra de entrenamiento no tiene valores perdidos, no se identificarán sustitutos. Los registros con valores perdidos que se encuentren durante la comprobación o puntuación pasarán automáticamente al nodo hijo que tenga un mayor número de registros. Si se esperan valores perdidos durante la comprobación o puntuación, asegúrese de que los valores no están presentes en la muestra de entrenamiento. No hay sustitutos disponibles para los árboles CHAID.

Aunque con los árboles CHAID no se utilizan sustitutos, podrá asignarlos a un nodo hijo específico al definir una división personalizada. Consulte el tema “Definición de divisiones personalizadas” en la página 88 para obtener más información.

## Personalización de la vista del árbol

La pestaña Visor del generador de árboles muestra el árbol actual. Todas las ramas del árbol se encuentran expandidas de forma predeterminada, sin embargo, puede expandir y contraer las ramas y personalizar la configuración restante según considere necesario.

- Pulse en el signo menos (–) situado en la esquina inferior derecha de un nodo padre para ocultar todos sus nodos hijo. Pulse en el signo más (+) situado en la esquina inferior izquierda de un nodo padre para mostrar sus nodos hijo.
- Utilice el menú Ver o la barra de herramientas para cambiar la orientación del árbol (de arriba a abajo, de izquierda a derecha o de derecha a izquierda).
- Pulse en el botón "Mostrar etiquetas de valor y de campo" en la barra de herramientas principal para mostrar u ocultar las etiquetas de campo y valor.
- Utilice los botones de lupa para acercar o alejar la vista, o bien, pulse en el botón de mapa del árbol situado en la parte derecha de la barra de herramientas para ver un diagrama del árbol completo.
- Cuando utilice un campo de partición, podrá intercambiar la vista del árbol entre las particiones de comprobación y entrenamiento (**Ver > Partición**). Mientras se visualiza la muestra de comprobación, el árbol puede verse pero no editarse. (La partición actual se muestra en la barra de estado situada en la esquina inferior derecha de la ventana.)
- Pulse en el botón de información (el botón "i" más a la derecha de la barra de herramientas) para ver los detalles de la división actual. Consulte el tema “Sustitutos y detalles de la división” en la página 89 para obtener más información.
- En cada nodo puede mostrar los estadísticos y los gráficos, juntos o por separado (consulte los siguientes apartados).

Visualización de estadísticos y gráficos

**Estadísticos de nodo.** Para un campo objetivo categórico, la tabla de cada nodo muestra el número y el porcentaje de registros de cada categoría, junto con el porcentaje de la muestra completa que el nodo representa. Para un campo objetivo continuo (rango numérico), la tabla muestra la media, la desviación estándar, el número de registros y el valor predicho del campo objetivo.

**Gráficos de nodo.** En el caso de un campo objetivo categórico, el gráfico consistirá en un diagrama de barras de los porcentajes de las distintas categorías del campo objetivo. Delante de cada fila de la tabla, aparece una muestra de color que corresponde al color representado por cada una de las categorías de campo objetivo en los gráficos de nodo. Para un campo objetivo continuo (rango numérico), el gráfico muestra un histograma del campo objetivo de los registros del nodo.

## Ganancias

La pestaña Ganancias muestra los estadísticos de todos los nodos terminales de un árbol. Las ganancias proporcionan una medida para considerar la distancia de la media o proporción de un nodo determinado respecto a la media global. Por lo general, cuanto mayor es esta diferencia, más útil resulta el árbol como herramienta para la toma de decisiones. Por ejemplo, un valor de índice de 148% en un nodo indica que los registros de dicho nodo tienen una probabilidad 1,5 veces superior de corresponder a la categoría objetivo que el conjunto de datos como un todo.

Para los nodos C&RT y QUEST en los que se haya especificado un conjunto de prevención sobreajustado, se muestran dos conjuntos de estadísticos:

- conjunto de desarrollo de árboles: la muestra de entrenamiento con el conjunto de prevención sobreajustado eliminado
- conjunto de prevención sobreajustado

Para otros árboles interactivos C&RT y QUEST, así como para todos los árboles interactivos CHAID, solamente se muestran los tres estadísticos del conjunto de desarrollo de árboles.

La pestaña Ganancias le permite:

- Mostrar los estadísticos de cuantiles, acumulados o nodo por nodo.
- Mostrar ganancias o beneficios.
- Intercambiar la vista entre tablas y gráficos.

- Seleccionar la categoría objetivo (sólo objetivos categóricos).
- Ordenar la tabla en orden ascendente o descendente según el porcentaje de índice. Cuando se muestran estadísticos de varias particiones, las ordenaciones se aplican siempre en la muestra de entrenamiento, no en la muestra de comprobación.

Por lo general, las selecciones realizadas en la tabla de ganancias se actualizan en la vista del árbol, y viceversa. Por ejemplo, si se selecciona una fila de la tabla, se seleccionará en el árbol el nodo correspondiente.

## Ganancias de clasificación

En el caso de los árboles de clasificación (que cuentan con una variable objetivo categórica), el índice porcentual de ganancias indicará la diferencia entre la proporción de la categoría objetivo determinada de cada nodo respecto a la proporción global.

Estadísticos nodo por nodo

En esta vista, la tabla muestra una fila por cada nodo terminal. Por ejemplo, si la respuesta global a una campaña publicitaria por correo ha sido del 10%, pero el 20% de los registros correspondientes al nodo X ha emitido una respuesta positiva, el índice porcentual del nodo sería del 200%, lo que indica que los encuestados de este grupo tienen el doble de probabilidades de comprar respecto a la población global.

Para los nodos C&RT y QUEST en los que se haya especificado un conjunto de prevención sobreajustado, se muestran dos conjuntos de estadísticos:

- conjunto de desarrollo de árboles: la muestra de entrenamiento con el conjunto de prevención sobreajustado eliminado
- conjunto de prevención sobreajustado

Para otros árboles interactivos C&RT y QUEST, así como para todos los árboles interactivos CHAID, solamente se muestran los tres estadísticos del conjunto de desarrollo de árboles.

**Nodos.** El ID del nodo actual (tal como se muestra en la pestaña Visor).

**Nodo: n.** El número total de registros en ese nodo.

**Nodo (%).** El porcentaje de todos los registros del conjunto de datos correspondiente a este nodo.

**Ganancia: n.** El número de registros con la categoría objetivo seleccionada que corresponden a este nodo. Dicho de otro modo: de todos los registros del conjunto de datos correspondientes a la categoría objetivo, ¿cuántos se encuentran en este nodo?

**Ganancia (%).** El porcentaje de todos los registros de la categoría objetivo (del conjunto de datos completo) correspondiente a este nodo.

**Respuesta (%).** El porcentaje de registros del nodo actual correspondiente a la categoría objetivo. En ocasiones, a las respuestas se les denomina "aciertos" en este contexto.

**Índice (%).** El porcentaje de respuestas para el nodo actual expresado como porcentaje de respuesta del conjunto de datos completo. Por ejemplo, un valor de índice de 300% indica que los registros del nodo tienen una probabilidad tres veces superior de corresponder a la categoría objetivo que el conjunto de datos como un todo.

Estadísticos acumulados

En la vista acumulada, la tabla muestra un nodo por fila. Sin embargo, los estadísticos están acumulados y dispuestos en orden ascendente o descendente por porcentaje de índice. Por ejemplo, si se aplica un

orden descendente, se mostrará en primer lugar el nodo con el índice porcentual más elevado, y se acumularán los estadísticos que se muestran en las filas subsiguientes para esta fila y las superiores.

El índice porcentual acumulado disminuye fila por fila cuando se añaden nodos con porcentajes de respuesta cada vez más reducidos. El índice acumulado de la fila final es siempre del 100%, porque en este punto se incluye el conjunto de datos completo.

Cuantiles

En esta vista, cada una de las filas de la tabla representa un cuantil en lugar de un nodo. Los cuantiles pueden ser cuartiles, quintiles (quintos), deciles (décimos), veintiles (vigésimos) o percentiles (centésimos). Se pueden indicar varios nodos en un único cuantil cuando es necesario más de un nodo para constituir tal porcentaje (por ejemplo, si se muestran los cuartiles, pero los dos nodos superiores contienen menos del 50% de todos los casos). El resto de la tabla se acumula y se puede interpretar como la vista acumulada.

## Rentabilidad de la inversión (ROI) y beneficios de la clasificación

En el caso de los árboles de clasificación, también se pueden mostrar los estadísticos de ganancias en términos de beneficios y ROI (rentabilidad de la inversión). En el cuadro de diálogo Definir beneficios puede especificar los ingresos y los gastos de cada categoría.

1. En la pestaña Ganancias, pulse en el botón Beneficio (con la etiqueta $/$) de la barra de herramientas para acceder al cuadro de diálogo.
2. Introduzca los valores de beneficios y gastos a las distintas categorías del campo objetivo.

Por ejemplo, si cuesta $0.48 enviar por correo una oferta a cada cliente y los ingresos de una respuesta positiva son $9,95 para una suscripción de tres meses, cada respuesta *no* cuesta $0,48 y cada *sí* tiene un beneficio de $9,47 (calculado como 9,95–0,48).

En la tabla de ganancias, los **beneficios** se calculan como la suma de los ingresos menos los gastos de cada uno de los registros ubicados en un nodo terminal. **ROI** es el total de beneficios dividido entre el total de gastos de un nodo.

Comentarios

- Los valores de beneficio sólo afectan al beneficio promedio y a los valores de ROI que se muestran en la tabla de ganancias, como un modo de visualización de estadísticos más aplicable a sus prioridades. No afectan, sin embargo, a la estructura básica del modelo del árbol. Los beneficios no deben confundirse con los costes de clasificación errónea especificados en el nodo de generación de árboles. Sus factores se extraen en el modelo como protección frente a errores muy costosos.
- Las especificaciones de los beneficios no se conservan entre distintas sesiones de generación de árboles interactivos.

## Ganancias de regresión

En el caso de los árboles de regresión, puede elegir entre las vistas de cuantil, nodo por nodo o acumulada. En la tabla se muestran los valores promedio. Únicamente hay gráficos disponibles para los cuantiles.

## Ganancias

Los gráficos pueden mostrarse en la pestaña Ganancias como alternativa a las tablas.

1. En la pestaña Ganancias, seleccione el icono Cuantiles (el tercero comenzando por la izquierda en la barra de herramientas). (No se muestran gráficos para estadísticos acumulados o nodo por nodo.)
2. Seleccione el icono de gráficos.
3. Seleccione las unidades que desea (percentiles, deciles, etc.) en la lista desplegable.
4. Seleccione **Ganancias**, **Respuesta** o **Elevación** para cambiar la medida que se muestra.

Ganancias

Los gráficos de ganancias representan los valores de la columna *Ganancia (%)* en la tabla. Las ganancias se definen como la proporción de aciertos en cada uno de los incrementos en relación con el número total de aciertos en el árbol, y se obtienen mediante la ecuación:

(aciertos del incremento / número total de aciertos) x 100%

El gráfico muestra de forma eficaz la difusión necesaria para una red cuando se desea capturar un porcentaje determinado de todos los aciertos del árbol. La línea diagonal representa la respuesta esperada para la muestra completa, si no se ha utilizado el modelo. En este caso la tasa de respuesta debería ser constante, ya que una persona tiene la misma probabilidad de responder que otra. Para duplicar los resultados deberá preguntar dos veces al mismo número de personas. La línea curvada indica hasta qué punto se puede mejorar la respuesta incluyendo únicamente elementos situados en los percentiles superiores en función de las ganancias. Por ejemplo, si incluye el 50% superior, obtendrá más del 70% de respuestas positivas. Cuanto más pronunciada es la curva, mayor es la ganancia.

Gráfico de elevación (índice)

El gráfico de elevación representa los valores de la columna *Índice (%)* en la tabla. Este gráfico compara el porcentaje de registros en cada uno de los incrementos considerado de aciertos con el porcentaje global de aciertos del conjunto de datos de entrenamiento, y se obtiene mediante la ecuación:

(aciertos del incremento / registros del incremento) / (número total de aciertos / número total de registros)

Gráfico de respuestas

El gráfico de respuestas representa los valores de la columna *Respuesta (%)* en la tabla. La respuesta es un porcentaje de registros en el incremento considerado de aciertos, y se obtiene mediante la ecuación:

(respuestas del incremento / registros del incremento) x 100%

## Selección basada en ganancias

El cuadro de diálogo Selección basada en ganancias permite seleccionar automáticamente los nodos terminales con las mayores (o menores) ganancias en función de una regla o un umbral especificado. Seguidamente, puede generar un nodo Seleccionar de acuerdo con su selección.

1. En la pestaña Ganancias, seleccione la vista acumulada o nodo por nodo y, a continuación, la categoría objetivo en la que desea basar la selección. (Las selecciones se basan en la representación de tabla actual se encuentran disponibles para cuantiles.)
2. En la pestaña Ganancias, seleccione en los menús:

   **Editar** > **Seleccionar nodos terminales** > **Selección basada en ganancias**

   **Seleccionar solamente.** Puede seleccionar nodos coincidentes *o* nodos no coincidentes, por ejemplo para seleccionar *todo excepto* los 100 registros superiores.

   **Coincidir por información de ganancias.** Establece la coincidencia entre los nodos en función de los estadísticos de ganancias para la categoría objetivo actual, e incluye:

   - Nodos en los que la ganancia, la respuesta o la elevación (índice) coincide con un umbral especificado, por ejemplo, respuesta mayor o igual que el 50 %.
   - Los *n* mejores nodos basados en la ganancia establecida para la categoría objetivo.
   - Los mejores nodos hasta un número especificado de registros.
   - Los mejores nodos hasta un porcentaje especificado de datos de entrenamiento.
3. Pulse en **Aceptar** para actualizar la selección de la pestaña Visor.

4. Para crear un nuevo nodo Seleccionar en función de la selección actual de la pestaña Visor, seleccione **Nodo Seleccionar** en el menú Generar. Consulte el tema "Generación nodos Seleccionar y Filtro" en la página 97 para obtener más información.

*Nota*: dado que está seleccionando nodos en lugar de registros o porcentajes, es posible que en algunos casos no logre una coincidencia perfecta con el criterio de selección. El sistema selecciona nodos completos *hasta* el nivel especificado. Por ejemplo, si selecciona los 12 casos superiores y dispone de 10 en el primer nodo y de dos en el segundo, únicamente se seleccionará el primer nodo.

## Riesgos

El riesgo le indica la posibilidad de aparición de errores de clasificación en cualquier nivel. La pestaña Riesgos muestra la estimación del riesgo de un punto y, para los resultados categóricos, también una tabla de errores de clasificación.

- En el caso de las predicciones numéricas, el riesgo es una estimación combinada de la varianza de cada uno de los nodos terminales.
- En el caso de las predicciones categóricas, el riesgo es la proporción de casos clasificados incorrectamente ajustada a los costes de los errores de clasificación o previas.

## Almacenamiento de resultados y modelos de árbol

Puede guardar o exportar los resultados de las sesiones de generación de árboles interactivos mediante distintos procedimientos, como:

- Generar un modelo basado en el árbol actual (**Generar > Generar modelo**).
- Guardar las directivas utilizadas para hacer crecer el árbol actual. La próxima vez que se ejecute el nodo de generación de árboles, el árbol actual volverá a crecer automáticamente e incluirá todas las divisiones personalizadas que se hayan definido.
- Exportar la información de riesgo, ganancias y modelo. Consulte el tema "Exportación de la información de riesgo, ganancias y modelo" en la página 96 para obtener más información.

Desde el generador de árboles o un modelo de árbol generado, puede:

- Generar un nodo Filtrar o Seleccionar en base al árbol actual. Consulte el tema "Generación nodos Seleccionar y Filtro" en la página 97 para obtener más información.
- Generar un nugget de conjunto de reglas que representa la estructura del árbol como un conjunto de reglas y define las ramas terminales del árbol. Consulte el tema "Generación de un conjunto de reglas desde un árbol de decisión" en la página 97 para obtener más información.
- Además, en el caso de los nugget de árbol generado, puede exportar el modelo en formato PMML. Consulte el tema "La paleta de modelos" en la página 41 para obtener más información. Si el modelo incluye divisiones personalizadas, esta información no se conserva en el PMML exportado. (La división se conserva, pero no el hecho de que sea personalizada y no seleccionada por el algoritmo.)
- Generar un gráfico basado en una parte seleccionada del árbol actual. *Nota*: sólo funciona para un nugget si se vincula a otros nodos en una ruta. Consulte el tema "Generación de gráficos" en la página 129 para obtener más información.

*Nota*: el árbol interactivo no puede guardarse propiamente. Para que no se pierda su trabajo, genere un modelo o actualice las directivas de árbol antes de cerrar la ventana del generador de árboles.

### Generación de un modelo desde el Generador de árboles

Para generar un modelo basado en el árbol actual, seleccione en el generador de árboles:

**Generar** > **Modelo**

En el cuadro de diálogo Generar nuevo modelo, puede elegir entre las opciones siguientes:

**Nombre del modelo.** Puede especificar un nombre personalizado, o bien, generar el nombre automáticamente basado en el nombre del nodo de modelado.

**Crear nodo en.** Puede añadir el nodo a las paletas **Lienzo**, **Paleta de modelos generados** o **Ambas**.

**Incluir directivas de árbol.** Para incluir las directivas desde el árbol actual en el modelo generado, seleccione esta casilla. De este modo podrá volver a generar el árbol si es necesario. Consulte el tema "Directivas de desarrollo de árboles" para obtener más información.

## Directivas de desarrollo de árboles

En el caso de los modelos Árbol C&R, CHAID y QUEST, las directivas de árbol especifican las condiciones de desarrollo del árbol en un nivel cada vez. Las directivas se aplican siempre que se inicia el Generador de árboles interactivos desde el nodo.

- Constituyen un método seguro para volver a generar un árbol creado durante una sesión interactiva anterior. Consulte el tema "Actualización de directivas de árbol" en la página 96 para obtener más información. También puede editar las directivas de forma manual, aunque deberá proceder con precaución.
- Las directivas son muy específicas con respecto a la estructura del árbol que describen. Por lo tanto, cualquier cambio de los datos subyacentes o de las opciones de modelado puede provocar que falle un conjunto de directivas válido hasta el momento. Por ejemplo, si el algoritmo CHAID cambia una división de dos factores a otra de tres basándose en los datos actualizados, podrían generarse errores en cualquiera de las directivas basadas en la división de dos factores anterior.

*Nota*: si decide generar un modelo directamente (sin utilizar el generador de árboles) se ignorarán las directivas de árbol.

Edición de directivas

1. Para ver o editar las directivas guardadas, abra el nodo de generación de árboles y seleccione el panel Objetivo de la pestaña Opciones de generación.
2. Seleccione **Iniciar sesión interactiva** para activar los controles, **Utilizar directivas de árbol** y, a continuación, **Directivas**.

Sintaxis de las directivas

Las directivas especifican condiciones para el desarrollo de un árbol, comenzando por el nodo raíz. Por ejemplo, para hacer crecer el árbol un nivel:

```
Grow Node Index 0 Children 1 2
```

Como no se ha especificado ningún predictor, el algoritmo seleccionará la división más adecuada.

Observe que la primera división siempre debe encontrarse en el nodo raíz (`Índice 0`) y que es necesario especificar los valores de índice para ambos hijos (en este caso, `1` y `2`). A menos que en primer lugar haga crecer la raíz que ha creado el nodo 2, no podrá especificar `Grow Node Index 2 Children 3 4`.

Para hacer crecer el árbol:

```
Hacer crecer árbol
```

Para hacer crecer el árbol y podarlo (solamente Árbol C&R):

```
Grow_And_Prune Tree
```

Si desea especificar una división personalizada para un predictor continuo:

```
Grow Node Index 0 Children 1 2 Spliton
  ( "EDUCATE", Interval ( NegativeInfinity, 12.5)
     Interval ( 12.5, Infinity ))
```

Para realizar divisiones en un predictor nominal con dos valores:

```
Grow Node Index 2 Children 3 4 Spliton
  ( "GENDER", Group( "0.0" )Group( "1.0" ))
```

Para realizar divisiones en un predictor nominal con varios valores:

```
Grow Node Index 6 Children 7 8 Spliton
   ( "ORGS", Group( "2.0","4.0" )
             Group( "0.0","1.0","3.0","6.0" ))
```

Para realizar divisiones en un predictor ordinal:

```
Grow Node Index 4 Children 5 6 Spliton
  ( "CHILDS", Interval ( NegativeInfinity, 1.0)
              Interval ( 1.0, Infinity ))
```

*Nota*: Al especificar divisiones personalizadas, los valores y los nombres de campos (`EDUCATE`, `GENDER`, `CHILDS`, etc.) son sensibles a las mayúsculas y las minúsculas.

Directivas para árboles CHAID

Las directivas para los árboles CHAID son especialmente sensibles a los cambios en los datos o el modelo, ya que, a diferencia de C&RT y QUEST, no se restringen para permitir el uso de divisiones binarias. Por ejemplo, aunque la siguiente sintaxis parece absolutamente válida, generaría errores si el algoritmo dividiera el nodo raíz en más de dos hijos:

```
Grow Node Index 0 Children 1 2
Grow Node Index 1 Children 3 4
```

En el caso de CHAID, es posible que el nodo 0 cuente con 3 o 4 nodos hijo que podrían generar errores en la segunda línea de la sintaxis.

Uso de directivas en scripts

Las directivas también pueden incrustarse en scripts mediante comillas triples.

## Actualización de directivas de árbol

Si desea conservar el trabajo de una sesión de generación de árboles interactivos, puede guardar las directivas que ha utilizado para generar el árbol actual. A diferencia de lo que ocurre cuando se guarda un nugget de modelo, que no puede volver a editarse, este procedimiento permite volver a generar el árbol en su estado actual para realizar ediciones adicionales.

Para actualizar directivas, seleccione en los menús del generador de árboles:

**Archivo** > **Actualizar directivas**

Las directivas se guardarán en el nodo de modelado utilizado para crear el árbol (independientemente de si se trata de Árbol C&R, QUEST o CHAID), y podrán utilizarse para volver a generar el árbol actual. Consulte el tema “Directivas de desarrollo de árboles” en la página 95 para obtener más información.

## Exportación de la información de riesgo, ganancias y modelo

En el generador de árboles, puede exportar los estadísticos de riesgo, ganancias y modelo a texto, HTML o formatos de imagen, según considere necesario.

1. En la ventana del generador de árboles, seleccione la pestaña o la vista que desea exportar.

2. Seleccione en los menús:

   **Archivo** > **Exportar**
3. Seleccione **Texto**, **HTML** o **Gráfico** según considere necesario y, a continuación, seleccione los elementos específicos que desea exportar en el submenú.

Siempre que resulta aplicable, la exportación se basa en las selecciones actuales.

**Exportación de formatos HTML o texto.** Puede exportar estadísticos de riesgo o ganancias para la partición de comprobación o entrenamiento (si se ha definido). La exportación se basa en las selecciones actuales de la pestaña Ganancias. Por ejemplo, puede seleccionar estadísticos de cuantiles, acumuladas o nodo por nodo.

**Exportación de gráficos.** Puede exportar el árbol actual tal como se muestra en la pestaña Visor, o bien, exportar los gráficos de ganancias para la partición de comprobación o entrenamiento. Entre los formatos disponibles se incluyen *.JPEG*, *.PNG* y *.BMP*. En el caso de las ganancias, la exportación se basa en las selecciones actuales de la pestaña Ganancias (que sólo está disponible cuando se muestra un gráfico).

## Generación nodos Seleccionar y Filtro

En la ventana del generador de árboles, o bien, al buscar un nugget de modelo de árbol de decisión, seleccione en los menús:

**Generar** > **Nodo Filtrar**

*o*

> **Nodo Seleccionar**

**Nodo Filtrar** Genera un nodo que filtra los campos no utilizados en el árbol actual. Constituye un método rápido para reducir el conjunto de datos para incluir únicamente los campos que el algoritmo considera importantes. Si existe un nodo Tipo anterior de la ruta al nodo Árbol de decisión, todos los campos con el rol *Objetivo* pasan por el nugget de modelo Filtro.

**Nodo Seleccionar** Genera un nodo que selecciona todos los registros correspondientes al nodo actual. Para ejecutar esta opción, deberá seleccionar una o varias ramas en la pestaña Visor.

El nugget de modelo se coloca en el lienzo de rutas.

## Generación de un conjunto de reglas desde un árbol de decisión

Puede generar un nugget de modelo Conjunto de reglas que represente la estructura del árbol como un conjunto de reglas que define las ramas terminales del árbol. Por lo general, los conjuntos de reglas pueden retener la mayor parte de la información significativa de un árbol de decisión completo, aunque utilizan un modelo menos complejo. La diferencia más importante consiste en que con un conjunto de reglas, puede aplicarse más de una regla a cualquier registro específico o no aplicar ninguna regla. Por ejemplo, podría ver todas las reglas que predicen un resultado *negativo*, seguidas de todas aquellas que predicen un resultado *positivo*. Al aplicar varias reglas, cada una de ellas obtiene un "voto" ponderado basado en la confianza que se asocia a dicha regla. La predicción final se alcanza mediante la combinación de los votos ponderados de todas las reglas que se aplican al registro en cuestión. Si no se aplica ninguna regla, se asignará al registro una predicción predeterminada.

**Nota:** Al puntuar un conjunto de reglas, puede que observe diferencias en comparación con la puntuación de un árbol; esto se debe a que cada rama terminal de un árbol se puntúa de forma independiente. Un área en la que puede ser observable esta diferencia es cuando faltan valores en los datos.

Los conjuntos de reglas solamente se pueden generar a partir de árboles con campos objetivo categóricos (y no en árboles de regresión).

En la ventana del generador de árboles, o bien, al buscar un nugget de modelo de árbol de decisión, seleccione en los menús:

**Generar** > **Conjunto de reglas**

**Nombre de conjuntos de reglas** Especifique el nombre del nuevo nugget de modelo Conjunto de reglas.

**Crear nodo en** Controla la ubicación del nuevo nugget de modelo Conjunto de reglas. Seleccione **Lienzo**, **Paleta de modelos generados** o **Ambas**.

**Instancias mínimas** Especifique el número mínimo de instancias (número de registros a los que se aplica la regla) que desea guardar en el nugget de modelo Conjunto de reglas. El nuevo conjunto de reglas no incluirá reglas con un soporte inferior al valor especificado.

**Confianza mínima** Especifique la confianza mínima para las reglas que desea conservar en el nugget de modelo Conjunto de reglas. El nuevo conjunto de reglas no incluirá reglas con una confianza inferior al valor especificado.

## Creación directa de un modelo de árbol

Como alternativa al uso del generador de árboles interactivo, también puede generar un modelo de árbol directamente desde el nodo cuando se ejecuta la ruta. Es coherente con la mayoría del resto de nodos de generación de modelos. Para los modelos de árbol C5.0 y Tree-AS, que no son compatibles con el generador de árboles interactivo, este es el único método que se puede utilizar.

1. Cree una ruta y añada uno de los nodos de árboles de decisión: Árbol C&R, CHAID, QUEST, C5.0 o Tree-AS.
2. En Árbol C&R, QUEST o CHAID, en el panel Objetivo de la pestaña Opciones de generación, seleccione uno de los objetivos principales. Si elige **Generar un único árbol**, asegúrese de que el **Modo** está establecido en **Generar modelo**.

   En C5.0, en la pestaña Modelo, defina **Tipo de resultado** a **Árbol de decisión**.

   Para Tree-AS, en el panel Información básica de la pestaña Opciones de generación, seleccione el tipo **Algoritmo de desarrollo de árboles**.
3. Seleccione los campos objetivo y predictor y especifique las opciones del modelo adicionales que considere necesario. Para obtener instrucciones específicas, consulte la documentación de los distintos nodos de generación de árboles.
4. Ejecute la ruta para generar el modelo.

Comentarios acerca de la generación de árboles.

- Cuando se generan árboles con este método, las directivas de desarrollo de árboles se omiten.
- Independientemente de si se utiliza un método de creación de árboles de decisión directo o interactivo, finalmente se obtendrán modelos similares. Se trata simplemente de considerar el control que se desea mantener.

## Nodos de árbol de decisión

Los nodos de árboles de decisión de IBM SPSS Modeler proporcionan acceso a los siguientes algoritmos de generación de árboles:

- Árbol C&R
- QUEST
- CHAID

- C5.0
- Tree-AS
- Árboles aleatorios

Consulte el tema “Modelos de árboles de decisión” en la página 85 para obtener más información.

Los algoritmos son similares porque pueden generar un árbol de decisiones mediante la división recursiva de los datos en subgrupos cada vez más pequeños. Sin embargo, existen algunas diferencias importantes.

**Campos de entrada.** Los campos de entrada (predictores) pueden ser cualquiera de los siguientes tipos (niveles de medición): continuo, categórico, de marca, nominal u ordinal.

**Campos objetivo.** Solamente es posible especificar un campo objetivo. Para el Árbol C&R, CHAID, Tree-AS y Árboles aleatorios, el objetivo puede ser continuo, categórico, de marca, nominal u ordinal. Para QUEST puede ser categórico, de marca o nominal. Para C5.0 el objetivo puede ser de marca, nominal u ordinal.

**Tipo de división.** El Árbol C&R, QUEST y Árboles aleatorios solo soportan divisiones binarias (es decir, cada nodo del árbol no se puede dividir en más de dos ramas). Por contra, CHAID, C5.0 y Tree-AS admiten la división en más de dos ramas al mismo tiempo.

**Método utilizado para la división.** Los algoritmos son diferentes según los criterios utilizados para decidir las divisiones. Cuando Árbol C&R predice un resultado categórico, se utiliza una medida de dispersión (de forma predeterminada, el coeficiente Gini, aunque se puede modificar). En el caso de objetivos continuos, se utiliza el método de desviación cuadrática mínima. CHAID y Tree-AS utilizan una prueba de chi-cuadrado; QUEST utiliza una prueba de chi-cuadrado para predictores categóricos y análisis de varianza de entradas continuas. En C5.0 se utiliza una medida de teoría de información, el cociente de ganancia de información.

**Gestión de valores perdidos.** Todos los algoritmos permiten valores perdidos para los campos del predictor, porque utilizan métodos diferentes para gestionarlos. Los nodos Árbol C&R y QUEST utilizan campos de predicción de sustitución, donde sea necesario, para avanzar un registro con los valores perdidos en el árbol durante la formación. CHAID convierte los valores perdidos en una categoría independiente y permite utilizarlos en la generación de árbol. C5.0 utiliza un método de fracción, que transmite una parte fraccional de un registro a cada rama del árbol desde un nodo en el que la división se basa en un campo con un valor perdido.

**Poda del árbol.** Árbol C&R, QUEST y C5.0 ofrecen la opción de hacer crecer el árbol y volver a podarlo eliminando divisiones de nivel inferior que no contribuyen de forma significativa a la precisión del árbol. Sin embargo, todos los algoritmos de árbol de decisión permiten controlar el tamaño mínimo del subgrupo, que ayuda a evitar las ramas con pocos registros de datos.

**Generación de árboles interactivos.** Árbol C&R, QUEST y CHAID permiten iniciar una sesión interactiva. Esto permite crear un nivel de árbol cada vez, editar las divisiones y podar el árbol antes de crear el modelo. C5.0, Tree-AS y Árboles aleatorios no tienen una opción interactiva.

**Probabilidades previas.** Árbol C&R y QUEST admiten la especificación de probabilidades previas de categorías al predecir un campo de destino categórico. Probabilidades previas son estimaciones de la frecuencia relativa general para cada categoría objetivo en la población de la que se trazan los datos de entrenamiento. En otras palabras, son las estimaciones de probabilidad que se obtendrían para cada campo objetivo antes de conocer nada acerca de los valores predictores. CHAID, C5.0, Tree-AS y Árboles aleatorios no soportan especificar probabilidades previas.

**Conjuntos de reglas.** No disponible para Tree-AS o Árboles aleatorios. En modelos con campos de destinos categóricos, los nodos de árboles de decisión para crear el modelo en forma de un conjunto de reglas, que en ocasiones puede ser más fácil de interpretar que un árbol de decisión complejo. En Árbol C&R, QUEST y CHAID puede generar un conjunto de reglas de una sesión interactiva; en C5.0 puede especificar esta opción en el nodo de modelado. Además, todos los modelos de árboles permiten generar un conjunto de reglas desde el nugget de modelo. Consulte el tema "Generación de un conjunto de reglas desde un árbol de decisión" en la página 97 para obtener más información.

## Nodo Árbol C&R

El nodo Árbol de clasificación y regresión (C&R) es un método de predicción y clasificación basado en árboles. Similar a C5.0, este método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos con valores de campo de salida similares. El nodo C&RT comienza por realizar un examen de los campos de entrada para buscar la mejor división, que se ha medido mediante la reducción del índice de impureza resultado de la división. La división define dos subgrupos, que se siguen dividiendo en otros dos subgrupos sucesivamente hasta que se activa un criterio de parada. Todas las divisiones son binarias (solamente se crean dos subgrupos).

Poda del árbol

Los árboles C&RT ofrecen la opción de hacer crecer el árbol en primer lugar y, a continuación, podar según un algoritmo de complejidad de costes que ajusta la estimación de riesgo en función del número de nodos terminales. Este método, que permite al árbol crecer enormemente antes de la poda a partir de criterios más complejos, puede generar árboles más pequeños con mejores propiedades de validación cruzada. Al aumentar el número de nodos terminales, por lo general se reduce el riesgo sobre los datos (de entrenamiento) actuales, pero se puede aumentar el riesgo real si el modelo se generaliza a datos no mostrados. Supongamos un caso extremo en que exista un nodo terminal independiente para cada registro del conjunto de entrenamiento. La estimación del riesgo sería del 0%, ya que cada registro correspondería a su propio nodo. Sin embargo, el riesgo de clasificación errónea para los datos (de comprobación) no mostrados sería con una certeza casi absoluta mayor que 0. La medida de coste-complejidad intenta compensar este punto.

**Ejemplo.** Una empresa de televisión ha solicitado un estudio de marketing para determinar qué clientes contratarían una suscripción a un servicio de noticias interactivo por cable. A partir de los datos del estudio, puede crear una ruta en la que el campo objetivo sea la intención de suscribirse y los campos predictores incluyan edad, sexo, educación, nivel de ingresos, horas invertidas en ver la televisión cada día y número de hijos. Aplicando un nodo Árbol CR a la ruta podrá predecir y clasificar las respuestas para obtener la tasa de respuesta más alta para su campaña.

**Requisitos.** Para entrenar un modelo de Árbol C&R, se precisan uno o varios campos de *Entrada* y exactamente uno de *Objetivo*. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. Se ignorarán los campos establecidos en *Ambos* o *Ninguno*. Los tipos de los campos utilizados en el modelo deben estar completamente instanciados y cualquier campo ordinal (conjunto ordenado) que se utilice en el modelo debe disponer de almacenamiento numérico (no en cadena). Si lo considera necesario, utilice a continuación el nodo Reclasificar para realizar las conversiones.

**Puntos fuertes.** Los modelos de Árbol C&R son bastante más robustos cuando aparecen problemas como datos perdidos y un número elevado de campos. Por lo general no precisan de largos tiempos de entrenamiento para calcular las estimaciones. Además, los modelos de Árbol C&R suelen ser más fáciles de comprender que algunos tipos de modelos: la interpretación de las reglas derivadas del modelo es muy directa. A diferencia de C5.0, Árbol C&R puede adaptar continuos como campos de salida categóricos.

## Nodo CHAID

CHAID, o detección automática de interacciones mediante chi-cuadrado (del inglés Chi-squared Automatic Interaction Detection), es un método de clasificación para generar árboles de decisión mediante estadísticos de chi-cuadrado para identificar divisiones óptimas.

CHAID examina en primer lugar las tablas de tabulación cruzada entre los campos de entrada y los resultados para, a continuación, comprobar la significación mediante una comprobación de independencia de chi-cuadrado. Si varias de estas relaciones son estadísticamente importantes, CHAID seleccionará el campo de entrada de mayor relevancia (el valor *P* más pequeño). Si una entrada cuenta con más de dos categorías, se compararán estas categorías y se contraerán las que no presenten diferencias en los resultados. Para ello, se unirá el par de categorías que presenten menor diferencia, y así sucesivamente. Este proceso de fusión de categorías se detiene cuando todas las categorías restantes difieren entre sí en el nivel de comprobación especificado. En el caso de campos de entrada nominales, pueden fundirse todas las categorías. Sin embargo, en los conjuntos ordinales, únicamente podrán fundirse las categorías contiguas.

CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles para cada predictor, aunque necesita más tiempo pare realizar los cálculos.

**Requisitos.** Los campos objetivo y de entrada pueden ser continuos o categóricos. Los nodos pueden dividirse en dos o más subgrupos en cada nivel. Todos los campos ordinales utilizados en el modelo deben disponer de almacenamiento numérico (no en cadenas). Si lo considera necesario, utilice a continuación el nodo Reclasificar para realizar las conversiones.

**Puntos fuertes.** A diferencia de los nodos C&RT y QUEST, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Es por ello que tiende a crear un árbol más extenso que los métodos de desarrollo binarios. CHAID admite todos los tipos de entradas y acepta tanto variables de frecuencia como ponderaciones de casos.

## Nodo QUEST

QUEST, o árbol estadístico eficiente insesgado y rápido, es un método de clasificación binario para generar árboles de decisión. Una de las principales motivaciones para su desarrollo ha sido la reducción del tiempo de procesamiento necesario para los análisis de C&RT de gran tamaño con varias variables o varios casos. Un segundo objetivo de QUEST consiste en reducir la tendencia de los métodos de clasificación de árboles para favorecer a las entradas que permiten realizar más divisiones, es decir, los campos de entrada continuos (rango numérico) o los correspondientes a varias categorías.

- QUEST utiliza una secuencia de reglas basada en comprobaciones de significación para evaluar los campos de entrada de un nodo. A efectos de selección, únicamente deberá realizar una sola comprobación en las distintas entradas de un nodo. A diferencia de lo que ocurre con C&RT, no se examinan todas las divisiones y, a diferencia de los casos de C&RT y CHAID, las combinaciones de categorías no se comprueban al evaluar un campo de entrada para su selección. Así se aumenta la velocidad del análisis.
- Para determinar las divisiones, se ejecuta un análisis de discriminante cuadrático mediante la entrada seleccionada en los grupos formados por las categorías objetivo. Este método vuelve a mejorar la velocidad de las búsquedas exhaustivas (C&RT) para determinar la división óptima.

**Requisitos.** Los campos de entrada pueden ser continuos (rango numérico), sin embargo el campo objetivo debe ser categórico. Todas las divisiones son binarias. No podrá utilizar los campos de ponderación. Todos los campos ordinales (conjunto ordenado) utilizados en el modelo deben disponer de almacenamiento numérico (no en cadenas). Si lo considera necesario, utilice a continuación el nodo Reclasificar para realizar las conversiones.

**Puntos fuertes.** Al igual que CHAID (pero a diferencia de C&RT), QUEST utiliza comprobaciones estadísticas para decidir si se ha de utilizar un campo de entrada o no. También separa las cuestiones

relacionadas con la división y la selección de entradas, y aplica criterios distintos a ambos casos. Esto contrasta con los casos de CHAID, donde el resultado de la comprobación de estadísticos que determina la selección de variables también genera la división. De un modo similar, C&RT emplea la medida de impureza-cambio tanto para seleccionar un campo de entrada como para determinar la división.

## Opciones de campos de nodo de árbol de decisión

En la pestaña Campos, puede seleccionar si desea utilizar la configuración de rol de campo ya definida en nodos anteriores o realizar las asignaciones de campos manualmente.

**Utilizar roles predefinidos** Esta opción utiliza las definiciones de roles (objetivos, predictores, etcétera) desde un nodo Tipo anterior (o la pestaña Tipo de un nodo de origen anterior).

**Utilizar asignaciones de campos personalizadas.** Para asignar manualmente objetivos, predictores y otros roles, seleccione esta opción.

**Campos.** Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los diferentes campos de roles en la parte derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol.

Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

**Objetivo.** Seleccione un campo como el objetivo de la predicción.

**Predictores (Entradas).** Seleccione uno o más campos como entradas de la predicción.

**Ponderación de análisis.** (sólo CHAID, C&RT y Trees-AS) Para utilizar un campo como ponderación de casos, especifique el campo aquí. Las ponderaciones de casos se usan para contabilizar las diferencias existentes en la varianza entre los niveles del campo de salida. Consulte el tema “Uso de campos de frecuencia y ponderación” en la página 33 para obtener más información.

## Opciones de generación de nodo de árbol de decisión

La pestaña Opciones de generación es la ubicación en la que se definen todas las opciones para crear el modelo. Puede pulsar en el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero normalmente querrá personalizar la generación de sus tareas.

La pestaña contiene varios paneles en los que puede establecer las personalizaciones específicas del modelo.

### Nodos de árbol de decisión: Objetivos

Para los nodos de Árbol C&R, QUEST y CHAID, en el panel Objetivos de la pestaña Opciones de generación, puede seleccionar si desea generar un nuevo modelo o actualizar uno existente. Establezca también el objetivo principal del nodo: generar un modelo estándar, para crear uno con una precisión o estabilidad mejorada o para generar uno para utilizarlo para conjuntos de datos de grandes dimensiones.

#### ¿Qué desea hacer?

**Crear modelo nuevo.** (Valor predeterminado) Crea un nuevo modelo completamente nuevo cada vez que ejecute una ruta con este nodo de modelado.

**Continuar entrenando modelo existente.** De forma predeterminada, cada vez que se ejecuta un nodo de modelado, se crea un modelo completamente nuevo. Si esta opción está seleccionada, el entrenamiento continúa con el último modelo generado correctamente por el nodo. Esto permite actualizar un modelo

existente sin tener que acceder a los datos originales. Además, puede dar como resultado un rendimiento significativamente más rápido ya que *sólo* se introducen en la ruta los registros nuevos o actualizados. Los detalles del modelo anterior se almacenan con el nodo de modelado, lo que permite utilizar esta opción incluso si el nugget de modelo anterior ya no está disponible en la ruta o la paleta de modelos.

**Nota:** Esta opción solo se activa si selecciona **Crear un árbol único** (para C&R Tree, CHAID y QUEST), **Crear un modelo estándar** (para Neural Net y Linear) o **Crear un modelo para conjuntos de datos muy grandes** como objetivo.

### ¿Cuál es su objetivo principal?

- **Crear un árbol único.** Crea un modelo de árbol de decisión estándar único. Los modelos estándar suelen ser más fáciles de interpretar y más rápidos de puntuar que los modelos creados utilizando el resto de opciones de objetivos.

  **Nota:** Para modelos segmentados, para utilizar esta opción con **Continuar entrenando un modelo existente** debe estar conectado a Analytic Server.

  **Modo.** Especifica el método utilizado para generar el modelo. **Generar modelo** crea un modelo automáticamente al ejecutar la ruta. **Iniciar sesión interactiva** abre el generador de árboles, que le permite generar un nivel de árbol cada vez, editar divisiones y podar según se considere necesario antes de crear el nugget de modelo.

  **Utilizar directivas de árbol.** Seleccione esta opción para especificar las directivas que desea aplicar al generar un árbol interactivo desde el nodo. Por ejemplo, puede especificar divisiones en los niveles primero y segundo y aplicarlas automáticamente al iniciar el generador de árboles. También puede guardar las directivas de una sesión de generación de árboles interactivos para volver a crear el árbol posteriormente. Consulte el tema “Actualización de directivas de árbol” en la página 96 para obtener más información.
- **Mejorar la precisión del modelo (aumento).** Seleccione esta opción si desea utilizar un método especial, conocido como **aumento**, para mejorar la tasa de precisión del modelo. El aumento funciona generando varios modelos en una secuencia. El primer modelo se crea con el procedimiento habitual. A continuación, se crea un segundo modelo que se centra en los registros que el primer modelo clasificó erróneamente. Seguidamente se crea un tercer modelo que se basará en los errores del segundo modelo, y así sucesivamente. Por último, para clasificar los casos, se les aplica todo el conjunto de modelos de acuerdo con un procedimiento de votación ponderada para combinar las distintas predicciones en una predicción global. El aumento puede mejorar significativamente la precisión del modelo de árbol de decisión, aunque también precisa de un entrenamiento más largo.
- **Mejorar la estabilidad del modelo (agregación autodocimante).** Seleccione esta opción si desea utilizar un método especial, conocido como **agregación autodocimante**, para mejorar la estabilidad del modelo para evitar sobreajustes. Esta opción crea múltiples modelos y los combina, con objeto de obtener predicciones más fiables. Los modelos obtenidos utilizando esta opción pueden tardar más en crearse y en puntuarse que los modelos estándar.
- **Crear un modelo para conjuntos de datos muy grandes.** Seleccione esta opción cuando trabaje con conjuntos de datos que son demasiado grandes para crear un modelo utilizando cualquiera del resto de opciones de objetivos. Esta opción divide los datos en bloques de datos más pequeños y genera un modelo en cada bloque. Los modelos más precisos se seleccionan automáticamente y se combinan en un único nugget de modelo. Puede ejecutar actualizaciones de modelos incrementales si selecciona la opción **Continuar el entrenamiento del modelo existente** en esta pantalla.

  **Nota:** esta opción para conjuntos de datos de grandes dimensiones requiere una conexión a IBM SPSS Modeler Server.

## Nodos de árbol de decisión: conceptos básicos

Especifique las opciones básicas sobre cómo se crea el árbol de decisiones.

**Algoritmo de desarrollo de árboles** (sólo CHAID y Tree-AS) Seleccione el tipo de algoritmo de **CHAID** que desee utilizar. **CHAID exhaustivo** es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles para cada predictor, aunque necesita más tiempo pare realizar los cálculos.

**Profundidad máxima del árbol** Especifique el número máximo de niveles bajo el nodo raíz (el número de veces que la muestra se dividirá repetidamente). El valor predeterminado es 5; seleccione **Personalizado** y entre un valor para especificar un número diferente de niveles.

### Poda del árbol (C&RT y QUEST únicamente)

**Poda del árbol para evitar sobreajustes** La poda consiste en eliminar las divisiones de nivel inferior que no aportan demasiado a la precisión del árbol. La poda puede ayudar a simplificar un árbol, que resultará más fácil de interpretar y, en determinados casos, mejora la generalización. Deje esta opción sin seleccionar para conservar un árbol completo sin podar.

- **Establecer diferencia máxima en riesgo (en errores estándar)** Permite especificar una regla de poda más liberal. La regla de error estándar permite al algoritmo seleccionar el árbol más simple cuya estimación de riesgo es próxima (pero posiblemente superior) a la del subárbol con el riesgo menor. El valor indica el tamaño de la diferencia admisible en la estimación del riesgo entre el árbol podado y el árbol con el riesgo menor en términos de estimación del riesgo. Por ejemplo, si se especifica 2, podría seleccionarse un árbol cuya estimación de riesgo sea (2 x error estándar) mayor que la del árbol completo.

**Máximo de sustitutos**. Los sustitutos constituyen un método de gestión de valores perdidos. Para cada una de las divisiones del árbol, el algoritmo identifica los campos de entrada más parecidos al campo de división seleccionado. Estos campos serán los *sustitutos* de la división. Cuando debe clasificarse un registro que presenta un valor perdido para un campo de división, puede utilizarse su valor en un campo de sustituto para realizar la división. Si se aumenta este valor, se permitirá una mayor flexibilidad para la gestión de los valores perdidos. Sin embargo, pueden aumentar el uso de memoria y los tiempos de entrenamiento.

## Nodos Árbol de decisión: Reglas de parada

Estas opciones controlan la construcción del árbol. Las reglas de parada determinan cuándo debe detenerse la división de ramas específicas del árbol. Establezca los tamaños de rama mínimos para evitar las divisiones a partir de las cuales se crearían subgrupos muy pequeños. **Número mínimo de registros en rama padre** impide una división si el número de registro del nodo que se va a dividir (el *padre*) es menor que el valor especificado. **Número mínimo de registros en rama hija** impide una división cuando el número de registros de cualquiera de las ramas creadas por la división (*hijo*) resulte inferior al valor especificado.

- **Utilizar porcentaje** Especifique los tamaños en términos de un porcentaje de datos de entrenamiento globales.
- **Utilizar valor absoluto** Especifique tamaños como los números absolutos de registros.

## Nodos de árbol de decisión: Conjuntos

Estos ajustes determinan el comportamiento de la agrupación que se produce cuando los conjuntos de datos de gran tamaño o de aumento o agregación autodocimante son obligatorios en Objetivos. Las opciones no aplicables al objetivo seleccionado se ignorarán.

**Agregación autodocimante y conjuntos de datos muy grandes.** Al puntuar un conjunto, ésta es la regla utilizada para combinar los valores predichos a partir de los modelos básicos para calcular el valor de puntuación del conjunto.

- **Regla de combinación predeterminada para objetivos categóricos.** Los valores predichos de conjunto para objetivos categóricos pueden combinarse mediante votación, la mayor probabilidad o la mayor probabilidad media. **Votación** selecciona la categoría que tenga la mayor probabilidad más

frecuentemente entre los modelos básicos. **La mayor probabilidad** selecciona la categoría que logra la mayor probabilidad individual entre todos los modelos básicos. **Mayor probabilidad media** selecciona la categoría con el valor más elevado cuando se calcula la media de las probabilidades de categoría entre los modelos básicos.

- **Regla de combinación predeterminada para objetivos continuos.** Los valores predichos de conjunto para objetivos continuos pueden combinarse mediante la media o mediana de los valores predichos a partir de los modelos básicos.

Tenga en cuenta que cuando el objetivo es mejorar la precisión del modelo, se ignoran las selecciones de reglas de combinación. El aumento siempre utiliza un voto de mayoría ponderada para puntuar objetivos categóricos y una mediana ponderada para puntuar objetivos continuos.

**Aumento y agregación autodocimante.** Especifique el número de modelos básicos que debe generarse cuando el objetivo es mejorar la precisión o estabilidad del modelo; en el caso de la agregación autodocimante, se trata del número de muestras de bootstrap. Debe ser un número entero positivo.

## Nodos Árbol C&R y QUEST - Costes y Previas

### Costes de clasificación errónea

En algunos contextos, ciertos tipos de errores son más costosos que otros. Por ejemplo, puede resultar más costoso clasificar a un solicitante de crédito de alto riesgo como de bajo riesgo (un tipo de error) que clasificar a un solicitante de crédito de bajo riesgo como de alto riesgo (otro tipo de error). Los costes de clasificación errónea permiten especificar la importancia relativa de los diversos tipos de errores de predicción.

Los costes de clasificación errónea son básicamente ponderaciones aplicadas a resultados específicos. Estas ponderaciones se extraen en el modelo y pueden realmente cambiar la predicción (como forma de protección frente a errores costosos).

Salvo los modelos C5.0, los costes de clasificación errónea no se aplican cuando se puntúa un modelo y no se tienen en cuenta cuando se clasifican o comparan modelos utilizando un nodo Clasificador automático, un diagrama de evaluación o un nodo Análisis. Es posible que un modelo que incluya costes no produzca menos errores que uno que no lo haga, y es posible que no ordene ningún valor mayor en términos de precisión global, pero es probable que funcione mejor en términos prácticos porque contiene un sesgo integrado en favor de errores *más baratos*.

La matriz de costes muestra el coste para cada combinación posible de categoría predicha y categoría real. De forma predeterminada, todos los costes de clasificación errónea se establecen en 1,0. Para introducir valores de coste personalizados, seleccione **Utilizar costes de clasificación errónea** e introduzca valores personalizados en la matriz de costes.

Para cambiar un coste de clasificación errónea, seleccione la casilla correspondiente a la combinación deseada de valores predichos y reales, elimine el contenido existente de la casilla e introduzca en ella el coste deseado. Los costes no son simétricos automáticamente. Por ejemplo, si establece el coste de clasificación errónea *A* como *B* para que sea 2,0, el coste de clasificación errónea de *B* como *A* aún tendrá el valor predeterminado 1,0 hasta que también se modifique explícitamente.

### Previas

Estas opciones permiten especificar probabilidades previas para categorías durante la predicción de un campo objetivo categórico. **Probabilidades previas** son estimaciones de la frecuencia relativa general para cada categoría objetivo en la población de la que se trazan los datos de entrenamiento. En otras palabras, son las estimaciones de probabilidad que se obtendrían para cada campo objetivo *antes* de conocer nada acerca de los valores predictores. Hay tres métodos de configuración de probabilidades previas:

- **Basadas en datos de entrenamiento.** Este es el método predeterminado. Las probabilidades previas se basan en las frecuencias relativas de las categorías en los datos de entrenamiento.
- **Igual para todas las clases.** Las probabilidades previas de todas las categorías se definen como $1/k$, donde $k$ es el número de categorías objetivo.
- **Personalizado.** Puede especificar sus propias probabilidades previas. Los valores iniciales de las probabilidades previas se configuran como iguales para todas las clases. Puede ajustar las probabilidades de cada categoría individualmente con valores personalizados. Para ajustar la probabilidad de una categoría específica, seleccione la casilla de la tabla correspondiente a la probabilidad que desee, elimine el contenido de la casilla e introduzca el valor que desee.

Las probabilidades previas de todas las categorías deben sumar 1,0 (la **restricción de probabilidad**). En caso contrario, aparecerá una advertencia con una opción para normalizar los valores automáticamente. Este ajuste automático conserva las proporciones en todas las categorías a la vez que fuerza la restricción de probabilidad. Puede llevar a cabo este ajuste en cualquier momento pulsando en el botón **Normalizar**. Para restablecer la tabla de modo que todas las categorías tengan el mismo valor, pulse en el botón **Igualar**.

**Ajustar previas utilizando costes de clasificación errónea**. Esta opción le permite ajustar las previas, basándose en costes de clasificación errónea (especificados en la pestaña Costes). De este modo puede incorporar información de costes directamente al proceso de desarrollo de los árboles que utilizan la medida de impureza binaria. (Si no se selecciona esta opción, la información de costes se utilizará únicamente para la clasificación de registros y el cálculo de estimaciones de riesgo para los árboles, en función de la medida binaria.)

## Nodo CHAID: Costes

En algunos contextos, ciertos tipos de errores son más costosos que otros. Por ejemplo, puede resultar más costoso clasificar a un solicitante de crédito de alto riesgo como de bajo riesgo (un tipo de error) que clasificar a un solicitante de crédito de bajo riesgo como de alto riesgo (otro tipo de error). Los costes de clasificación errónea permiten especificar la importancia relativa de los diversos tipos de errores de predicción.

Los costes de clasificación errónea son básicamente ponderaciones aplicadas a resultados específicos. Estas ponderaciones se extraen en el modelo y pueden realmente cambiar la predicción (como forma de protección frente a errores costosos).

Salvo los modelos C5.0, los costes de clasificación errónea no se aplican cuando se puntúa un modelo y no se tienen en cuenta cuando se clasifican o comparan modelos utilizando un nodo Clasificador automático, un diagrama de evaluación o un nodo Análisis. Es posible que un modelo que incluya costes no produzca menos errores que uno que no lo haga, y es posible que no ordene ningún valor mayor en términos de precisión global, pero es probable que funcione mejor en términos prácticos porque contiene un sesgo integrado en favor de errores *más baratos*.

La matriz de costes muestra el coste para cada combinación posible de categoría predicha y categoría real. De forma predeterminada, todos los costes de clasificación errónea se establecen en 1,0. Para introducir valores de coste personalizados, seleccione **Utilizar costes de clasificación errónea** e introduzca valores personalizados en la matriz de costes.

Para cambiar un coste de clasificación errónea, seleccione la casilla correspondiente a la combinación deseada de valores predichos y reales, elimine el contenido existente de la casilla e introduzca en ella el coste deseado. Los costes no son simétricos automáticamente. Por ejemplo, si establece el coste de clasificación errónea *A* como *B* para que sea 2,0, el coste de clasificación errónea de *B* como *A* aún tendrá el valor predeterminado 1,0 hasta que también se modifique explícitamente.

## Nodo Árbol C&R - Opciones avanzadas

Las opciones avanzadas permiten ajustar el proceso de generación de árboles.

**Cambio mínimo en la impureza.** Especifique el cambio mínimo en la impureza para crear una nueva división en el árbol. Con **impureza** nos referimos al punto hasta el cual los subgrupos definidos por el árbol presentan una amplia variedad de valores de campo de salida dentro de cada grupo. En lo que respecta a los objetivos categóricos, un nodo se considera "puro" si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. El objetivo de la generación de árboles es crear subgrupos con valores de salida similares, es decir, minimizar la impureza dentro de cada nodo. Si la mejor división de una rama no reduce la impureza hasta el punto especificado, no se realizará dicha división.

**Medida de impureza para objetivos categóricos.** Especifique el método que desea utilizar para los campos objetivo categóricos para medir la impureza del árbol. (En el caso de objetivos continuos, esta opción se ignorará y se utilizará siempre la **desviación cuadrática mínima** como medida de impureza.)

- **Gini** es una medida de impureza general que aplica a la rama probabilidades de pertenencia a categorías.
- **Binario** es una medida de impureza que enfatiza la división binaria y con la que es más probable obtener ramas con un tamaño similar a partir de una división.
- La opción **Ordinal** impone la restricción adicional de que únicamente pueden agruparse las clases objetivo contiguas, ya que únicamente resulta aplicable con objetivos ordinales. Cuando esta opción se selecciona para un objetivo nominal, se utiliza de forma predeterminada la medida binaria estándar.

**Conjunto de prevención sobreajustado.** El algoritmo divide los registros de manera interna en un conjunto de creación de modelos y un conjunto de prevención sobreajustado, el cual es un conjunto independiente de registros de datos utilizado para realizar un seguimiento de errores durante la formación para evitar que el método modele una variación atribuible al azar en los datos. Especifique un porcentaje de registros. El valor por omisión es 30.

**Replicar resultados.** Al establecer una semilla aleatoria podrá replicar los análisis. Especifique un entero o pulse en **Generar**, lo que creará un entero pseudo-aleatorio entre 1 y 2147483647, ambos inclusive.

## Nodo QUEST: Avanzado

Las opciones avanzadas permiten ajustar el proceso de generación de árboles.

**Nivel de significancia para división.** Especifica el nivel de significación (alfa) para la división de nodos. El valor debe estar comprendido entre 0 y 1. Los valores inferiores tienden a producir árboles con menos nodos.

**Conjunto de prevención sobreajustado.** El algoritmo divide los registros de manera interna en un conjunto de creación de modelos y un conjunto de prevención sobreajustado, el cual es un conjunto independiente de registros de datos utilizado para realizar un seguimiento de errores durante la formación para evitar que el método modele una variación atribuible al azar en los datos. Especifique un porcentaje de registros. El valor predeterminado es 30.

**Replicar resultados.** Al establecer una semilla aleatoria podrá replicar los análisis. Especifique un entero o pulse en **Generar**, lo que creará un entero pseudo-aleatorio entre 1 y 2147483647, ambos inclusive.

## Nodo CHAID: Avanzado

Las opciones avanzadas permiten ajustar el proceso de generación de árboles.

**Nivel de significancia para división.** Especifica el nivel de significación (alfa) para la división de nodos. El valor debe estar comprendido entre 0 y 1. Los valores inferiores tienden a producir árboles con menos nodos.

**Nivel de significancia para fusión.** Especifica el nivel de significación (alfa) para la fusión de categorías. El valor debe ser superior a 0 e inferior o igual que 1. Para impedir todas las fusiones de categorías,

especifique un valor de 1. En el caso de los objetivos continuos, hará referencia al número de categorías de la variable en el árbol final que coincide con el número especificado de intervalos. Esta opción no se encuentra disponible para CHAID exhaustivo.

**Los valores de significancia de ajuste utilizando el método de Bonferroni.** Ajusta los valores de significación al comprobar las distintas combinaciones de categorías de un predictor. Los valores se ajustan en función del número de comprobaciones, directamente relacionado con el número de categorías y el nivel de medición de un predictor. Suele resultar conveniente porque el control ejercido es mejor y ofrece el cociente de error de falsos positivos. Desactive esta opción para aumentar la potencia del análisis y buscar diferencias reales, con un cociente aumentado de falsos positivos en contrapartida. Concretamente se recomienda desactivar esta opción para muestras pequeñas.

**Permitir nuevas divisiones de categorías fusionadas en un nodo.** El algoritmo CHAID intenta fusionar categorías para producir el árbol más simple que describe el modelo. Si se selecciona, esta opción permite que las categorías fusionadas vuelvan a dividirse si esto puede considerarse una solución mejor.

**Chi-cuadrado para objetivos categóricos.** Especifique el método que desea utilizar para calcular los estadísticos de chi-cuadrado con objetivos categóricos.

- **Pearson.** Este método proporciona cálculos más rápidos pero se debe utilizar con precaución en muestras pequeñas.
- **Razón de verosimilitud.** Este método es más robusto que el método Pearson, pero tarda más tiempo en realizar los cálculos. Es el método preferido para muestras pequeñas. Para objetivos continuos, siempre se utiliza este método.

**Cambio mínimo en frecuencias de casillas esperadas.** Al calcular las frecuencias de casilla (tanto para el modelo nominal como para el modelo ordinal de efectos de fila), se utiliza un procedimiento iterativo (épsilon) para convergir en la estimación óptima que se haya utilizado en la prueba de chi-cuadrado para una división específica. Épsilon determina el nivel de cambio que debe producirse para que continúen las iteraciones. Si el cambio de la última iteración es menor que el valor especificado, las iteraciones se detendrán. Si tiene algún problema con la convergencia del algoritmo, aumente este valor o reduzca el número máximo de iteraciones hasta que tenga lugar la convergencia.

**Número máximo de iteraciones para la convergencia.** Especifica el número máximo de iteraciones que deben producirse antes de la parada, haya o no tenido lugar la convergencia.

**Conjunto de prevención sobreajustado.** (Esta opción sólo está disponible cuando se utiliza el generador de árboles interactivos.) El algoritmo divide los registros de manera interna en un conjunto de creación de modelos y un conjunto de prevención sobreajustado, el cual es un conjunto independiente de registros de datos utilizado para realizar un seguimiento de errores durante la formación para evitar que el método modele una variación atribuible al azar en los datos. Especifique un porcentaje de registros. El valor predeterminado es 30.

**Replicar resultados.** Al establecer una semilla aleatoria podrá replicar los análisis. Especifique un entero o pulse en **Generar**, lo que creará un entero pseudo-aleatorio entre 1 y 2147483647, ambos inclusive.

## Opciones de modelo de nodo de árboles de decisión

En la pestaña Opciones del modelo, puede seleccionar si desea especificar un nombre para el modelo o generar un nombre automáticamente. También puede seleccionar obtener información sobre la importancia del predictor, así como puntuaciones de propensión ajustadas y brutas de objetivos de marca.

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Calcular importancia del predictor.** En el caso de modelos que produzcan una medida adecuada de importancia, puede mostrar un gráfico que indique la importancia relativa de cada predictor al estimar el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Tenga en cuenta que puede tardarse más tiempo en calcular la importancia del predictor para algunos modelos, especialmente al trabajar con conjuntos de datos de gran tamaño; además, como resultado está desactivada para algunos modelos de forma predeterminada. La importancia del predictor no está disponible para modelos de listas de decisiones. Consulte el tema “Importancia del predictor” en la página 44 para obtener más información.

**Puntuaciones de propensión**

Las puntuaciones de propensión pueden activarse en el nodo de modelado y en la pestaña Configuración del nugget de modelo. Esta funcionalidad sólo está disponible cuando el objetivo seleccionado es un campo de marca. Consulte el tema “Puntuaciones de propensión” en la página 36 para obtener más información.

**Calcular puntuaciones de propensión en bruto.** Las puntuaciones de propensión en bruto están derivadas del modelo basado únicamente en los datos de entrenamiento. Si el modelo predice el valor *true* (responderá), la propensión es la misma que P, donde P es la probabilidad de la predicción. Si el modelo predice el valor false, la propensión se calcula como (1 - P).

- Si selecciona esta opción al crear el modelo, las puntuaciones de propensión se activarán en el nugget de modelo de forma predeterminada. Sin embargo, siempre puede activar las puntuaciones de propensión en bruto en el nugget de modelo independientemente de si las selecciona o no en el nodo de modelado.
- Al puntuar el modelo, se añadirán puntuaciones de propensión en bruto a un campo con las letras *RP* unidas al prefijo estándar. Por ejemplo, si las predicciones están en un campo denominado *$R-churn*, el nombre del campo de puntuación de propensión será *$RRP-churn*.

**Calcular puntuaciones de propensión ajustada.** Las propensiones brutas se basan totalmente en estimaciones proporcionadas por el modelo, las cuales pueden estar ajustadas excesivamente, lo que lleva a estimaciones de propensión demasiado optimistas. Las propensiones ajustadas intentan compensar este hecho observando el rendimiento del modelo en las particiones de comprobación o validación y ajustando las propensiones para proporcionar una mejor estimación en consecuencia.

- Esta configuración requiere que haya un campo de partición válido en la ruta.
- A diferencia de las puntuaciones brutas de confianza, las puntuaciones ajustadas de propensión deben calcularse al crear el modelo; de lo contrario, no estarán disponibles cuando se puntúe el nugget de modelo.
- Al puntuar el modelo, se añadirán puntuaciones ajustadas de propensión a un campo con las letras *AP* unidas al prefijo estándar. Por ejemplo, si las predicciones están en un campo denominado *$R-churn*, el nombre del campo de puntuación de propensión será *$RAP-churn*. Las puntuaciones ajustadas de propensión no están disponibles para modelos de regresión logística.
- Al calcular las puntuaciones ajustadas de propensión, la partición de comprobación o validación utilizada para el cálculo no debe haberse equilibrado. Para evitarlo, asegúrese de seleccionar la opción **Sólo datos de entrenamiento de equilibrado** en todos los nodos Equilibrar anteriores en la ruta. Además, si se ha llevado una muestra compleja a un punto anterior en la ruta, se invalidarán las puntuaciones ajustadas de propensión.
- Las puntuaciones ajustadas de propensión no están disponibles para modelos de árbol "aumentado" y de conjuntos de reglas. Consulte el tema “Modelos C5.0 aumentados” en la página 128 para obtener más información.

**Basado en.** Para que se calculen las puntuaciones ajustadas de propensión, debe haber un campo de partición en la ruta. Puede especificar si desea utilizar la partición de comprobación o validación para

este cálculo. Para obtener los mejores resultados, la partición de comprobación o validación debe incluir al menos el mismo número de registros que la partición utilizada para entrenar el modelo original.

## Nodo C5.0

Esta característica está disponible en SPSS Modeler Professional y SPSS Modeler Premium.

Este nodo utiliza el algoritmo C5.0 para generar un **árbol de decisión** o un **conjunto de reglas**. Los modelos C5.0 dividen la muestra en función del campo que ofrece la máxima **ganancia de información**. Las distintas submuestras definidas por la primera división se vuelven a dividir, por lo general basándose en otro campo, y el proceso se repite hasta que resulta imposible dividir las submuestras de nuevo. Por último se vuelven a examinar las divisiones del nivel inferior, y se eliminan o **podan** las que no contribuyen significativamente con el valor del modelo.

*Nota*: el nodo C5.0 solamente puede predecir un objetivo categórico. Al analizar datos con campos categóricos (nominales u ordinales), el nodo tiene mayor probabilidad de agrupar categorías que las versiones de C5.0 anteriores a la versión 11.0.

C5.0 puede generar dos tipos de modelos. Un **árbol de decisión** es una descripción sencilla de las divisiones que se han encontrado en el algoritmo. Los distintos nodos terminales (o "de hoja") describen un subconjunto de datos de entrenamiento, y cada uno de los casos incluidos en los datos de entrenamiento pertenece exactamente a un nodo terminal del árbol. En otras palabras, es posible realizar exactamente una predicción para cada registro de datos específico presente en un árbol de decisión.

En cambio, un **conjunto de reglas** es, como su propio nombre indica, un conjunto de reglas que intenta realizar predicciones de registros individuales. Los conjuntos de reglas derivan de los árboles de decisión y, en cierto modo, representan una versión simplificada de la información que se incluye en estos árboles. Por lo general, los conjuntos de reglas pueden retener la mayor parte de la información significativa de un árbol de decisión completo, aunque utilizan un modelo menos complejo. Debido a las diferencias de funcionamiento de los conjuntos de reglas, sus propiedades son distintas de las de los árboles de decisión. La diferencia más importante consiste en que con un conjunto de reglas, puede aplicarse más de una regla a cualquier registro específico o no aplicar ninguna regla. Al aplicar varias reglas, cada una de ellas obtiene un "voto" ponderado basado en la confianza que se asocia a dicha regla. La predicción final se alcanza mediante la combinación de los votos ponderados de todas las reglas que se aplican al registro en cuestión. Si no se aplica ninguna regla, se asignará al registro una predicción predeterminada.

**Ejemplo.** Un investigador médico ha recopilado información sobre un conjunto de pacientes, de los cuales todos sufrieron la misma enfermedad. Durante el curso del tratamiento, cada paciente respondió a un medicamento de un total de cinco. Puede utilizar un modelo C5.0 combinado con otros nodos para averiguar qué medicamento es el adecuado para un futuro paciente con la misma enfermedad.

**Requisitos.** Para entrenar un modelo C5.0, debe existir un campo categórico (por ejemplo, nominal u ordinal) *Objetivo* y uno o más campos *Entrada* de cualquier tipo. Se ignorarán los campos establecidos en *Ambos* o *Ninguno*. Los tipos de los campos utilizados en el modelo deben estar completamente instanciados. También se puede especificar un campo de ponderación.

**Puntos fuertes.** Los modelos C5.0 son bastante más robustos cuando aparecen problemas como datos perdidos y un número elevado de campos de entrada. Por lo general no precisan de largos tiempos de entrenamiento para calcular las estimaciones. Además, los modelos C5.0 suelen ser más fáciles de comprender que algunos tipos de modelos, ya que la interpretación de las reglas derivadas del modelo es muy directa. C5.0 también ofrece el eficaz método del **aumento** para obtener una mayor precisión en tareas de clasificación.

*Nota*: la velocidad de generación de modelos de C5.0 puede mejorarse al activar el procesamiento paralelo.

## Opciones de modelo para el nodo C5.0

**Nombre del modelo.** Especifique el nombre del modelo que desea generar.

- **Automático.** Seleccione esta opción para generar automáticamente el nombre del modelo de acuerdo con los nombres de los campos objetivos. Este es el método predeterminado.
- **Personalizado.** Seleccione esta opción para especificar el nombre que desea para el nugget de modelo que se creará en este nodo.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Crear modelos divididos.** Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como campos de división. Consulte el tema "Generación de modelos divididos" en la página 28 para obtener más información.

**Tipo de resultados.** Especifique si desea que el nugget de modelo sea un **árbol de decisión** o un **conjunto de reglas**.

**Agrupar simbólicos.** Seleccione esta opción para que C5.0 intente combinar los valores simbólicos que cuentan con patrones similares respecto al campo de salida. Seleccione esta opción para que C5.0 cree un nodo hijo para cada uno de los valores del campo simbólico utilizado para dividir el nodo padre. Por ejemplo, si C5.0 realiza divisiones en un campo *COLOR* (con valores *ROJO*, *VERDE* y *AZUL*), se creará de forma predeterminada una división de tres factores. No obstante, si selecciona esta opción y los registros donde *COLOR = ROJO* son muy similares a los registros donde *COLOR = AZUL*, se creará una división de dos factores, con los registros correspondientes a *VERDE* en un grupo y los registros para *AZUL* y *ROJO* en otro.

**Utilizar aumento.** El algoritmo C5.0 cuenta con un método especial para mejorar su precisión denominado **aumento**. Este método genera varios modelos en una secuencia. El primer modelo se crea con el procedimiento habitual. A continuación, se crea un segundo modelo que se centra en los registros que el primer modelo clasificó erróneamente. Seguidamente se crea un tercer modelo que se basará en los errores del segundo modelo, y así sucesivamente. Por último, para clasificar los casos, se les aplica todo el conjunto de modelos de acuerdo con un procedimiento de votación ponderada para combinar las distintas predicciones en una predicción global. El aumento puede mejorar significativamente la precisión del modelo C5.0, aunque también precisa de un entrenamiento más largo. La opción **Número de ensayos** permite controlar el número de modelos que deben utilizarse para el modelo aumentado. Esta característica se basa en la investigación de Freund y Schapire, con ciertas mejoras propietarias para gestionar mejor los datos con ruido.

**Efectuar validación cruzada.** Seleccione esta opción para que C5.0 utilice un conjunto de modelos creado a partir de subconjuntos de datos de entrenamiento para calcular una estimación de la precisión de un modelo creado a partir de un conjunto de datos completo. Esta función resulta de utilidad cuando el conjunto de datos es demasiado pequeño para dividirlo en conjuntos tradicionales de comprobación o entrenamiento. Los modelos de validación cruzada se descartan una vez calculada la estimación de precisión. Puede especificar el **número de veces** o el número de modelos que desea aplicar a la validación cruzada. Observe que, en versiones anteriores de IBM SPSS Modeler, la creación del modelo y la validación cruzada eran dos operaciones independientes. En la versión actual, no se precisa ningún otro paso para generar el modelo. La validación cruzada y la generación del modelo se realizan al mismo tiempo.

**Modo.** En un entrenamiento **Simple**, la mayoría de los parámetros de C5.0 se establecen automáticamente. El entrenamiento **Experto** permite ejercer un control más directo sobre los parámetros de entrenamiento.

Opciones de modo Simple

**Favorecer.** De forma predeterminada, C5.0 intentará producir el árbol más preciso posible. En algunos casos, puede producirse un sobreajuste que puede ocasionar un rendimiento pobre al aplicar el modelo a nuevos datos. Seleccione **Generalización** para utilizar la configuración de algoritmo menos propensa a este problema.

*Nota*: los modelos generados con la opción **Generalización** no tienen por qué generalizar mejor que el resto de modelos necesariamente. Cuando la generalización resulta fundamental, valide siempre el modelo con una muestra de comprobación reservada.

**Ruido esperado (%).** Especifique la proporción esperada de datos con ruido o erróneos en el conjunto de entrenamiento.

Opciones de modo Experto

**Gravedad de la poda.** Determina hasta qué punto se debe podar el árbol de decisión o conjunto de reglas. Aumente este valor para obtener un árbol más pequeño y resumido. Disminúyalo para obtener un árbol más preciso. Este parámetro afecta únicamente a la poda local (consulte "Utilizar poda global" a continuación).

**Número mínimo de registros por rama hija.** Puede utilizar el tamaño de los subgrupos para limitar el número de divisiones de cualquier rama del árbol. Una rama se dividirá únicamente si dos o más de las subramas resultantes pueden contener al menos este número de registros del conjunto de entrenamiento. El valor predeterminado es 2. Auméntelo para impedir el **sobreentrenamiento** con los datos con ruido.

**Utilizar poda global.** Los árboles se podan en dos etapa: La primera, una fase de poda local, que examina los subárboles y contrae las ramas para aumentar la precisión del modelo. La segunda es una fase de poda global en que se considera el árbol como un todo y pueden contraerse los subárboles. De forma predeterminada, se realiza la poda global. Anule la selección de esta opción para omitir esta fase.

**Valoración inicial de atributos.** Seleccione esta opción para que C5.0 examine la utilidad de los predictores antes de iniciar la generación del modelo. A continuación, se excluyen de este proceso de generación los predictores que no se consideran importantes. Esta opción puede resultar útil para los modelos con varios campos predictores y puede ayudar a impedir el sobreajuste.

*Nota*: la velocidad de generación de modelos de C5.0 puede mejorarse al activar el procesamiento paralelo.

## Nodo Tree-AS

El nodo Tree-AS se puede utilizar con datos en un entorno distribuido. En este nodo puede elegir generar árboles de decisión utilizando un modelo CHAID o CHAID exhaustivo.

CHAID, o detección automática de interacciones mediante chi-cuadrado (del inglés Chi-squared Automatic Interaction Detection), es un método de clasificación para generar árboles de decisión mediante estadísticos de chi-cuadrado para identificar divisiones óptimas.

CHAID examina en primer lugar las tablas de tabulación cruzada entre los campos de entrada y los resultados para, a continuación, comprobar la significación mediante una comprobación de independencia de chi-cuadrado. Si varias de estas relaciones son estadísticamente importantes, CHAID seleccionará el campo de entrada de mayor relevancia (el valor *P* más pequeño). Si una entrada cuenta con más de dos categorías, se compararán estas categorías y se contraerán las que no presenten diferencias en los resultados. Para ello, se unirá el par de categorías que presenten menor diferencia, y así sucesivamente. Este proceso de fusión de categorías se detiene cuando todas las categorías restantes difieren entre sí en el nivel de comprobación especificado. En el caso de campos de entrada nominales, pueden fundirse todas las categorías. Sin embargo, en los conjuntos ordinales, únicamente podrán fundirse las categorías contiguas.

CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles para cada predictor, aunque necesita más tiempo pare realizar los cálculos.

**Requisitos.** Los campos objetivo y de entrada pueden ser continuos o categóricos. Los nodos pueden dividirse en dos o más subgrupos en cada nivel. Todos los campos ordinales utilizados en el modelo deben disponer de almacenamiento numérico (no en cadenas). Si es necesario, utilice el nodo Reclasificar para convertirlos.

**Puntos fuertes.** CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Es por ello que tiende a crear un árbol más extenso que los métodos de desarrollo binarios. CHAID admite todos los tipos de entradas y acepta tanto variables de frecuencia como ponderaciones de casos.

## Opciones de campos del nodo Tree-AS

En la pestaña Campos, puede seleccionar si desea utilizar la configuración de rol de campo ya definida en nodos anteriores o realizar las asignaciones de campos manualmente.

**Utilizar roles predefinidos** Esta opción utiliza las definiciones de roles (objetivos, predictores, etcétera) desde un nodo Tipo anterior (o la pestaña Tipo de un nodo de origen anterior).

**Utilizar asignaciones de campos personalizadas.** Para asignar manualmente objetivos, predictores y otros roles, seleccione esta opción.

**Campos.** Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los diferentes campos de roles en la parte derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol.

Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

**Objetivo.** Seleccione un campo como el objetivo de la predicción.

**Predictores** Seleccione uno o más campos como entradas de la predicción.

**Ponderación de análisis** Para utilizar un campo como ponderación de casos, especifique el campo aquí. Las ponderaciones de casos se usan para contabilizar las diferencias existentes en la varianza entre los niveles del campo de salida. Si desea obtener más información, consulte “Uso de campos de frecuencia y ponderación” en la página 33.

## Opciones de generación del nodo Tree-AS

La pestaña Opciones de generación es la ubicación en la que se definen todas las opciones para crear el modelo. Puede pulsar en el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero normalmente querrá personalizar la generación de sus tareas.

La pestaña contiene varios paneles en los que puede establecer las personalizaciones específicas del modelo.

### Nodo Tree-AS - conceptos básicos

Especifique las opciones básicas sobre cómo se crea el árbol de decisiones.

**Algoritmo de desarrollo de árboles** Seleccione el tipo de algoritmo de **CHAID** que desee utilizar. **CHAID exhaustivo** es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles para cada predictor, aunque necesita más tiempo pare realizar los cálculos.

**Profundidad máxima del árbol** Especifique el número máximo de niveles bajo el nodo raíz (el número de veces que la muestra se dividirá repetidamente); el valor predeterminado es 5. El número máximo de niveles (también denominados *nodos*) es de 50.000.

**En intervalos** Si utiliza datos continuos, debe agrupar las entradas. Puede hacerlo en un nodo precedente; sin embargo, el nodo Tree-AS agrupa automáticamente las entradas continuas. Si utiliza el nodo Tree-AS para agrupar automáticamente los datos, seleccione el **Número de intervalos** en el que deben dividirse las entradas. Los datos se dividen en intervalos de igual frecuencia; las opciones disponibles son 2, 4, 5, 10, 20, 25, 50 o 100.

## Nodo Tree-AS - crecimiento

Utilice las opciones de crecimiento para ajustar el proceso de generación de árboles.

**Registrar umbral para conmutar de valores p a tamaños de efecto** Especifique el número de registros a partir del cual el modelo pasará de utilizar los **Valores P** a los **Valores de tamaño de efecto** al generar el árbol. El valor predeterminado es 1.000.000.

**Nivel de significación para división** Especifique el nivel de significación (alfa) para la división de nodos. El valor debe estar entre 0,01 y 0,99. El valor debe estar comprendido entre 0 y 1. Los valores inferiores tienden a producir árboles con menos nodos.

**Nivel de significación para fusión** Especifique el nivel de significación (alfa) para la fusión de categorías. El valor debe estar entre 0,01 y 0,99. Esta opción no se encuentra disponible para CHAID exhaustivo.

**Ajustar valores de significación utilizando el método Bonferroni** Ajuste los valores de significación al probar las diversas combinaciones de categorías de un predictor. Los valores se ajustan en función del número de comprobaciones, directamente relacionado con el número de categorías y el nivel de medición de un predictor. Suele resultar conveniente porque el control ejercido es mejor y ofrece el cociente de error de falsos positivos. Desactive esta opción para aumentar la potencia del análisis y buscar diferencias reales, con un cociente aumentado de falsos positivos en contrapartida. Concretamente se recomienda desactivar esta opción para muestras pequeñas.

**Umbral de tamaño del efecto (sólo objetivos continuos)** Establece el umbral de tamaño del efecto que debe utilizarse al dividir nodos y fusionar categorías al utilizar un objetivo continuo. El valor debe estar entre 0,01 y 0,99.

**Umbral de tamaño del efecto (sólo objetivos categóricos)** Establece el umbral de tamaño del efecto que debe utilizarse al dividir nodos y fusionar categorías al utilizar un objetivo categórico. El valor debe estar entre 0,01 y 0,99.

**Permitir nuevas divisiones de categorías fusionadas en un nodo** El algoritmo CHAID intenta fusionar categorías para producir el árbol más simple que describe el modelo. Si se selecciona, esta opción permite que las categorías fusionadas vuelvan a dividirse si esto puede considerarse una solución mejor.

**Nivel de significación para agrupar nodos hoja** Especifique el nivel de significación que determina cómo se forman grupos de nodos hoja o cómo se identifican nodos hoja inusuales.

**Chi-cuadrado para objetivos categóricos** Para objetivos categóricos, puede especificar el método utilizado para calcular el estadístico de chi-cuadrado.

- **Pearson** Este método proporciona cálculos más rápidos pero se debe utilizar con precaución en muestras pequeñas.
- **Razón de verosimilitud** Este método es más robusto que el método Pearson, pero tarda más tiempo en realizar los cálculos. Es el método preferido para muestras pequeñas. Para objetivos continuos, siempre se utiliza este método.

## Nodo Tree-AS - reglas de parada

Estas opciones controlan la construcción del árbol. Las reglas de parada determinan cuándo debe detenerse la división de ramas específicas del árbol. Establezca los tamaños de rama mínimos para evitar las divisiones a partir de las cuales se crearían subgrupos muy pequeños. **Número mínimo de registros en rama padre** impide una división si el número de registro del nodo que se va a dividir (el *padre*) es menor que el valor especificado. **Número mínimo de registros en rama hija** impide una división cuando el número de registros de cualquiera de las ramas creadas por la división (*hijo*) resulte inferior al valor especificado.

- **Utilizar porcentaje** Especifique los tamaños en términos de un porcentaje de datos de entrenamiento globales.
- **Utilizar valor absoluto** Especifique tamaños como los números absolutos de registros.

**Cambio mínimo en frecuencias de casillas esperadas** Al calcular las frecuencias de casilla (tanto para el modelo nominal como para el modelo ordinal de efectos de fila), se utiliza un procedimiento iterativo (épsilon) para convergir en la estimación óptima que se haya utilizado en la prueba de chi-cuadrado para una división específica. Épsilon determina el nivel de cambio que debe producirse para que continúen las iteraciones. Si el cambio de la última iteración es menor que el valor especificado, las iteraciones se detendrán. Si tiene algún problema con la convergencia del algoritmo, aumente este valor o reduzca el número máximo de iteraciones hasta que tenga lugar la convergencia.

**Número máximo de iteraciones para la convergencia** Especifica el número máximo de iteraciones que deben producirse antes de la parada, haya o no tenido lugar la convergencia.

## Nodo Tree-AS - Costes

En algunos contextos, ciertos tipos de errores son más costosos que otros. Por ejemplo, puede resultar más costoso clasificar a un solicitante de crédito de alto riesgo como de bajo riesgo (un tipo de error) que clasificar a un solicitante de crédito de bajo riesgo como de alto riesgo (otro tipo de error). Los costes de clasificación errónea permiten especificar la importancia relativa de los diversos tipos de errores de predicción.

Los costes de clasificación errónea son básicamente ponderaciones aplicadas a resultados específicos. Estas ponderaciones se extraen en el modelo y pueden realmente cambiar la predicción (como forma de protección frente a errores costosos).

Es posible que un modelo que incluya costes no produzca menos errores que uno que no lo haga, y es posible que no ordene ningún valor mayor en términos de precisión global, pero es probable que funcione mejor en términos prácticos porque contiene un sesgo integrado en favor de errores menos costosos.

La matriz de costes muestra el coste para cada combinación posible de categoría predicha y categoría real. De forma predeterminada, todos los costes de clasificación errónea se establecen en 1,0. Para introducir valores de coste personalizados, seleccione **Utilizar costes de clasificación errónea** e introduzca valores personalizados en la matriz de costes.

Para cambiar un coste de clasificación errónea, seleccione la casilla correspondiente a la combinación deseada de valores predichos y reales, elimine el contenido existente de la casilla e introduzca en ella el coste deseado. Los costes no son simétricos automáticamente. Por ejemplo, si establece el coste de clasificación errónea *A* como *B* para que sea 2,0, el coste de clasificación errónea de *B* como *A* aún tendrá el valor predeterminado 1,0 hasta que también se modifique explícitamente.

Sólo para objetivos ordinales, puede seleccionar el **Aumento de coste predeterminado para objetivos ordinales** y establecer valores predeterminados en la matriz de costes. Las opciones disponibles se describen en la lista siguiente.

- **Sin aumento** - Un valor predeterminado de 1,0 para cada predicción correcta.

- **Lineal** - Cada predicción incorrecta sucesiva aumenta el coste en 1.
- **Cuadrado** - Cada predicción incorrecta sucesiva es el cuadrado del valor lineal. En este caso, los valores podrían ser: 1, 4, 9 y, así, sucesivamente.
- **Personalizado** - Si edita manualmente los valores de la tabla, la opción desplegable cambia automáticamente a **Personalizado**. Si cambia la selección desplegable por alguna de las demás opciones, los valores editados se sustituirán por los valores de la opción seleccionada.

# Opciones de modelo del nodo Tree-AS

En la pestaña Opciones del modelo, puede seleccionar si desea especificar un nombre para el modelo o generar un nombre automáticamente. También puede elegir calcular valores de confianza y añadir un ID de identificación durante la puntuación del modelo.

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Calcular confianzas** Para añadir un campo de confianza al puntuar el modelo, marque esta casilla de verificación.

**Identificador de regla** Para añadir un campo al puntuar el modelo que contenga el ID del nodo hoja al que se ha asignado un registro, marque esta casilla de verificación.

# Nugget de modelo Tree-AS

## Salida del nugget de modelo Tree-AS

Después de crear un modelo Tree-AS, la información siguiente está disponible en el visor de la salida.

### Tabla de información de modelo

La tabla de información de modelo proporciona información clave acerca del modelo. La tabla identifica algunos ajustes de modelo de alto nivel, por ejemplo:

- El tipo de algoritmo utilizado: CHAID o CHAID exhaustivo.
- El nombre del campo objetivo seleccionado en la pestaña Campos de nodo Tipo o nodo Tree-AS.
- Los nombres de los campos seleccionados como predictores en la pestaña Campos de nodo Tipo o nodo Tree-AS.
- El número de registros de los datos. Si crea un modelo con un peso de frecuencia, este valor es el recuento ponderado válido que representa los registros en los que se basa el árbol.
- El número de *nodos hoja* del árbol generado.
- El número de niveles del árbol; es decir, la profundidad del árbol.

### Importancia del predictor

El gráfico Importancia de predictor muestra la importancia de las 10 entradas (predictores) principales del modelo en forma de diagrama de barras.

Si hay más de 10 campos en el gráfico, puede cambiar la selección de los predictores que se incluyen en el gráfico mediante el control deslizante situado debajo del mismo. Las marcas indicadores del control deslizante son de anchura fija, y cada marca del control deslizante presenta 10 campos. Puede mover las marcas indicadoras a lo largo del control deslizante para visualizar los 10 ca,pos anteriores o siguientes, ordenados por importancia de predictor.

Puede efectuar una doble pulsación en el gráfico para abrir un cuadro de diálogo independiente en el que se pueden editar los valores del gráfico. Por ejemplo, puede corregir elementos tales como el tamaño del gráfico y el tamaño y color de los fonts utilizados. Al cerrar este cuadro de diálogo de edición independiente, los cambios se aplicarán al gráfico que se visualiza en la pestaña Salida.

### Tabla Reglas de decisión principales

De forma predeterminada, esta tabla interactiva muestra las estadísticas de las reglas de los cinco nodos hoja principales de la salida, en función del porcentaje del total de registros contenidos en el nodo hoja.

Puede efectuar una doble pulsación en la tabla para abrir un cuadro de diálogo independiente en el que puede editar la información de reglas que se muestra en la tabla. La información que se muestra y las opciones que están disponibles en el cuadro de diálogo dependen del tipo de datos del destino; por ejemplo, categórico o continuo.

En la tabla se muestra la siguiente información de reglas:

- ID de regla
- Los detalles de cómo se aplica y forma la regla
- Recuento de registros para cada regla. Si crea un modelo con un peso de frecuencia, este valor es el recuento ponderado válido que representa los registros en los que se basa el árbol.
- Porcentaje de registros para cada regla

Además, para un destino continuo, una columna adicional de la tabla muestra el valor **Medio** para cada regla.

Puede modificar el diseño de la tabla de reglas mediante las siguientes opciones de **Contenido de la tabla**:

- **Reglas de decisión principales** Las cinco reglas de decisión principales se clasifican por el porcentaje del total de registros contenidos en los nodos hoja.
- **Todas las reglas** La tabla contiene todos los nodos hoja generados por el modelo, pero sólo muestra 20 reglas por página. Si selecciona este diseño, puede buscar una regla utilizando las opciones adicionales **Buscar regla por ID** y **Página**.

Además, para un destino categórico, puede modificar el diseño de la tabla utilizando la opción **Reglas principales por categoría**. Las cinco reglas de decisión principales se clasifican por el porcentaje del total de registros para una **Categoría de destino** seleccionada.

Si cambia el diseño de la tabla de reglas, puede copiar de nuevo la tabla de reglas modificada en el visor de la salida pulsando el botón Copiar en el visor situado en la parte superior izquierda del cuadro de diálogo.

## Configuración del nugget de modelo Tree-AS

La pestaña Configuración de un nugget de modelo Tree-AS, se especifican las opciones para confianzas y para la generación de SQL durante la puntuación de modelos. Esta pestaña sólo está disponible después de añadir el nugget de modelo a una ruta.

**Calcular confianzas** Para incluir confianzas en operaciones de puntuación, seleccione esta casilla de verificación. Al puntuar modelos en la base de datos, si excluye confianzas puede generar SQL más eficaz. Para los árboles de regresión, no se asignan las confianzas.

**Identificador de regla** Para añadir un campo en el resultado de puntuación que indique el ID para el nodo terminal al que se asigna cada registro, seleccione esta casilla de verificación.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se genera SQL:

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

## Nodo Árboles aleatorios

El nodo Árboles aleatorios se puede utilizar con datos en un entorno distribuido. En este nodo, se genera un modelo de conjunto que está formado por varios árboles de decisiones.

El nodo Árboles aleatorios es un método de predicción y clasificación basado en árbol que se basa en la metodología de Árbol de clasificación y regresión. Al igual que con el Árbol C&R, este método de predicción utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos con valores de campo de salida similares. El nodo empieza examinando los campos de entrada disponibles para el mismo para buscar la mejor división, que se mide a través de la reducción en un índice de impureza resultado de la división. La división define dos subgrupos, cada uno de los cuales se divide después en dos subgrupos más y, así, sucesivamente, hasta que se desencadena uno de los criterios de parada. Todas las divisiones son binarias (solamente se crean dos subgrupos).

Los Árboles aleatorios añaden dos características en comparación con el Árbol C&R:

- La primera característica es la *agregación* donde las réplicas del conjunto de datos de entrenamiento se crean mediante muestreos con sustitución del conjunto de datos original. Esta acción crea muestras de programa de arranque que son de tamaño igual al conjunto de datos original, después de lo cual se basa un *modelo de componente* en cada réplica. Juntos, estos modelos de componentes forman un modelo de conjunto.
- La segunda característica es que, en cada división del árbol, solo se tiene en cuenta un muestreo de los campos de entrada para la medida de la impureza.

**Requisitos.** Para entrenar un modelo Árboles aleatorios, necesitará uno o más campos de *entrada* y un campo *objetivo*. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. Los campos que se establecen en *Ambos* o *Ninguno* se ignoran. Los campos que se utilizan en el modelo debe tener sus tipos completamente instanciados, y cualquier campo ordinal (conjunto ordenado) que se utiliza en el modelo debe tener almacenamiento numérico (no cadena). Si lo considera necesario, utilice a continuación el nodo Reclasificar para realizar las conversiones.

**Puntos fuertes.** Los modelos Árboles aleatorios son sólidos cuando se trata con grandes cantidades de datos y números elevados de campos. Debido al uso de la agregación y al muestreo de campos, son mucho menos propensos al sobreajuste y, lo más probable, es que los resultados que se ven en las pruebas se repitan cuando se utilizan datos nuevos.

## Opciones de campos del nodo Árboles aleatorios

En la pestaña Campos, puede seleccionar si desea utilizar la configuración de rol de campo ya definida en nodos anteriores o realizar las asignaciones de campos manualmente.

**Utilizar roles predefinidos** Esta opción utiliza las definiciones de roles (objetivos, predictores, etcétera) desde un nodo Tipo anterior (o la pestaña Tipo de un nodo de origen anterior).

**Utilizar asignaciones de campos personalizadas.** Para asignar manualmente objetivos, predictores y otros roles, seleccione esta opción.

**Campos.** Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los diferentes campos de roles en la parte derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol.

Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

**Objetivo.** Seleccione un campo como el objetivo de la predicción.

**Predictores** Seleccione uno o más campos como entradas de la predicción.

**Ponderación de análisis** Para utilizar un campo como ponderación de casos, especifique el campo aquí. Las ponderaciones de casos se usan para contabilizar las diferencias existentes en la varianza entre los niveles del campo de salida. Si desea obtener más información, consulte "Uso de campos de frecuencia y ponderación" en la página 33.

## Opciones de generación del nodo Árboles aleatorios

La pestaña Opciones de generación es la ubicación en la que se definen todas las opciones para crear el modelo. Puede pulsar en el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero normalmente querrá personalizar la generación de sus tareas.

La pestaña contiene varios paneles en los que puede establecer las personalizaciones específicas del modelo.

### Nodo Árboles aleatorios - Aspectos básicos

Especifique las opciones básicas sobre cómo se crea el árbol de decisiones.

**Número de modelos a construir** Especifique el número máximo de modelos que el nodo puede construir.

**Tamaño de muestra** De forma predeterminada, el tamaño de la muestra de simulación de muestreo es igual a los datos de entrenamiento originales. Al tratar con conjuntos de datos grandes, la reducción del tamaño de muestra puede aumentar el rendimiento.

**Manejar datos desequilibrados** Si el destino del modelo es un resultado de distintivo (por ejemplo, comprar o no comprar) y la proporción del resultado deseado para no deseado es muy pequeña, los datos están desequilibrados y el muestreo de simulación de muestreo realizado por el modelo puede afectar la precisión del modelo. Para mejorar la precisión marque esta casilla de verificación; entonces el modelo captura una proporción mayor de los resultados deseados y genera un modelo mejor.

**Utilizar muestreo ponderado para la selección de variables** De forma predeterminada, las variables para cada nodo hoja se seleccionan aleatoriamente con la misma probabilidad. Para aplicar ponderación a las variables y mejorar el proceso de selección, seleccione esta casilla de verificación.

**Número máximo de nodos** Especifique el número máximo de nodos hoja que están permitidas en árboles individuales. Si el número supera en la división siguiente, el crecimiento del árbol se detiene antes de que se produzca la división.

**Profundidad máxima del árbol** Especifique el número máximo de niveles *nodos hoja* bajo el nodo raíz; es decir, el número de veces que la muestra se divide repetidamente).

**Tamaño mínimo de nodo hijo** Especifique el número mínimo de registros que deben estar contenidos en un nodo hijo después de que se haya dividido el nodo padre. Si un nodo hijo contiene menos registros que los que se especifican, el nodo padre no se dividirá.

**Especifique el número de predictores para utilizar para la división** Si está creando modelos de división, establezca el número mínimo de predictores que se van a utilizar para crear cada división. Así se evita que la división cree un número excesivo de subgrupos pequeños.

**Nota:** El número de predictores para la división no puede ser mayor que el número total de predictores en los datos.

**Dejar de crear cuando ya no se puede mejorar la precisión** Para mejorar los tiempos de generación de modelo, seleccione esta opción para detener el proceso de generación del modelo cuando no se puede mejorar la precisión del resultado.

## Nodo Árboles aleatorios - Costes

En algunos contextos, ciertos tipos de errores son más costosos que otros. Por ejemplo, puede resultar más costoso clasificar a un solicitante de crédito de alto riesgo como de bajo riesgo (un tipo de error) que clasificar a un solicitante de crédito de bajo riesgo como de alto riesgo (otro tipo de error). Los costes de clasificación errónea permiten especificar la importancia relativa de los diversos tipos de errores de predicción.

Los costes de clasificación errónea son básicamente ponderaciones aplicadas a resultados específicos. Estas ponderaciones se extraen en el modelo y pueden realmente cambiar la predicción (como forma de protección frente a errores costosos).

Es posible que un modelo que incluya costes no produzca menos errores que uno que no lo haga, y es posible que no ordene ningún valor mayor en términos de precisión global, pero es probable que funcione mejor en términos prácticos porque contiene un sesgo integrado en favor de errores menos costosos.

La matriz de costes muestra el coste para cada combinación posible de categoría predicha y categoría real. De forma predeterminada, todos los costes de clasificación errónea se establecen en 1,0. Para introducir valores de coste personalizados, seleccione **Utilizar costes de clasificación errónea** e introduzca valores personalizados en la matriz de costes.

Para cambiar un coste de clasificación errónea, seleccione la casilla correspondiente a la combinación deseada de valores predichos y reales, elimine el contenido existente de la casilla e introduzca en ella el coste deseado. Los costes no son simétricos automáticamente. Por ejemplo, si establece el coste de clasificación errónea *A* como *B* para que sea 2,0, el coste de clasificación errónea de *B* como *A* aún tendrá el valor predeterminado 1,0 hasta que también se modifique explícitamente.

Sólo para objetivos ordinales, puede seleccionar el **Aumento de coste predeterminado para objetivos ordinales** y establecer valores predeterminados en la matriz de costes. Las opciones disponibles se describen en la lista siguiente.

- **Sin aumento** - Un valor predeterminado de 1.0 para cada predicción incorrecta.
- **Lineal** - Cada predicción incorrecta sucesiva aumenta el coste en 1.
- **Cuadrado** - Cada predicción incorrecta sucesiva es el cuadrado del valor lineal. En este caso, los valores podrían ser: 1, 4, 9 y, así, sucesivamente.
- **Personalizado** - Si edita manualmente los valores de la tabla, la opción desplegable cambia automáticamente a **Personalizado**. Si cambia la selección desplegable por alguna de las demás opciones, los valores editados se sustituirán por los valores de la opción seleccionada.

## Nodo Árboles aleatorios - Avanzado

Especifique las opciones avanzadas sobre cómo se debe crear el árbol de decisiones.

**Porcentaje máximo de valores perdidos.** Especifique el porcentaje máximo de valores perdidos permitidos en cualquier entrada. Si el porcentaje supera este número, la entrada se excluye de la generación de modelos.

**Excluir campos con una sola categoría por mayoría.** Especifique el porcentaje máximo de registros que puede tener una sola categoría dentro de un campo. Si cualquier un valor de categoría representa un porcentaje más alto de registros que el especificado, todo el campo se excluye de la generación de modelos.

**Número máximo de categorías de campo.** Especifique el número máximo de categorías que se pueden incluir dentro de un campo. Si el número de categorías supera este número, el campo se excluye de la generación de modelos.

**Variación mínima de campo.** Si el coeficiente de variación de un campo continuo es menor que el valor que especifique aquí, el campo se excluye de la creación del modelo.

**Número de intervalos.** Especifique el número de enlaces de frecuencia iguales que se van a utilizar para entradas continuas. Las opciones disponibles son: 2, 4, 5, 10, 20, 25 ,50 o 100.

**Número de reglas interesantes de las que informar** Especifique el número de reglas sobre las que informar (mínimo de 1, máximo de 1000, con un valor predeterminado 50).

# Opciones del modelo del modo Árboles aleatorios

En la pestaña Opciones del modelo, puede seleccionar si desea especificar un nombre para el modelo o generar un nombre automáticamente. También puede optar por calcular la importancia de los predictores durante la puntuación del modelo.

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

# Nugget del modelo Árboles aleatorios

## Salida del nugget del modelo Árboles aleatorios

Tras crear un modelo Árboles aleatorios, la información siguiente está disponible en el visor de la salida:

### Tabla de información del modelo

La tabla de información del modelo proporciona información clave sobre el modelo. La tabla siempre incluye los valores del modelo de alto nivel siguientes:

- El nombre del campo objetivo que se ha seleccionado en el nodo Tipo o bien en la pestaña Campos del nodo Árboles aleatorios.
- El método de generación de modelos - Árboles aleatorios.
- El número de entrada de predictores en el modelo.

Los detalles adicionales que se muestran en la tabla dependen de si genera un modelo de clasificación o regresión, y de si el modelo se genera para manejar datos desequilibrados:

- Modelo de clasificación (valores predeterminados)
  - Precisión de modelo
  - Regla de clasificación errónea
- Modelo de clasificación (**Manejar datos desequilibrados** seleccionado)
  - Media geométrica
  - Tasa de verdaderos positivos, que se subdivide en clases.
- Modelo de regresión
  - Error cuadrático promedio raíz
  - Error relativo

– Varianza explicada

### Resumen de registros

El resumen muestra cuántos registros se han utilizado para ajustarse al modelo y cuántos se han excluido. Se muestran ambos, el número de registros y el porcentaje del número entero. Si el modelo se ha creado para incluir la ponderación de frecuencia, también se muestra el número no ponderado de registros que se han incluido y excluido.

### Importancia del predictor

El gráfico Importancia de predictor muestra la importancia de las 10 entradas (predictores) principales del modelo en forma de diagrama de barras.

Si hay más de 10 campos en el gráfico, puede cambiar la selección de los predictores que se incluyen en el gráfico mediante el control deslizante situado debajo del mismo. Las marcas indicadores del control deslizante son de anchura fija, y cada marca del control deslizante presenta 10 campos. Puede mover las marcas indicadoras a lo largo del control deslizante para visualizar los 10 ca,pos anteriores o siguientes, ordenados por importancia de predictor.

Puede efectuar una doble pulsación en el gráfico para abrir un recuadro de diálogo separado en el cual puede editar el tamaño del gráfico. Al cerrar este cuadro de diálogo de edición independiente, los cambios se aplicarán al gráfico que se visualiza en la pestaña Salida.

### Tabla Reglas de decisión principales

De forma predeterminada, esta tabla interactiva muestra las estadísticas de las reglas principales, que se clasifican según el grado de interés.

Puede efectuar una doble pulsación en la tabla para abrir un cuadro de diálogo independiente en el que puede editar la información de reglas que se muestra en la tabla. La información que se muestra y las opciones que están disponibles en el cuadro de diálogo dependen del tipo de datos del destino; por ejemplo, categórico o continuo.

En la tabla se muestra la siguiente información de reglas:

- Los detalles de cómo se aplica y forma la regla
- Si los resultados están en la categoría más frecuente
- Precisión de la regla
- Precisión de los árboles
- Índice del grado de interés

El índice del grado de interés se calcula utilizando la fórmula siguiente:

$$I_{index}(t) = P(A(t)) * P(B(t)) * (P(B(t)|A(t)) + P(\bar{B}(t)|\bar{A}(t)))$$

En esta fórmula:

- P(A(t)) es la precisión de los árboles
- P(B(t)) es la precisión de la regla
- P(B(t) | A(t)) representa las predicciones correctas según los árboles y, también, el nodo
- La parte restante de la fórmula representa predicciones incorrectas de acuerdo con los árboles y el nodo.

Puede alterar el diseño de la tabla de reglas utilizando las opciones siguientes de **Contenido de tabla**:

- **Reglas de decisión principales** Las cinco reglas de decisión principales, que se orden de acuerdo con el índice del grado de interés.
- **Todas las reglas** La tabla contiene todas las reglas generadas por el modelo, pero solo muestra 20 reglas por página. Si selecciona este diseño, puede buscar una regla utilizando las opciones adicionales **Buscar regla por ID** y **Página**.

Además, para un objetivo categórico, puede alterar el diseño de la tabla de reglas utilizando la opción **Reglas principales por categoría**. Las cinco reglas de decisión principales se clasifican por el porcentaje del total de registros para una **Categoría de destino** seleccionada.

**Nota:** Para objetivos categóricos, la tabla solo está disponible cuando **Manejar datos desequilibrados** no está seleccionado en la pestaña Información básica de las Opciones de generación.

Si cambia el diseño de la tabla de reglas, puede volver a copiar la tabla de reglas modificada en el visor de la salida pulsando el botón Copiar en visor en el extremo superior izquierdo del recuadro de diálogo.

### Matriz de confusión

Para modelos de clasificación, la matriz de confusión muestra el número de resultados pronosticados con respecto a los resultados reales observados, incluyendo la proporción de predicciones correctas.

**Nota:** La matriz de confusión no está disponible para los modelos de regresión, ni cuando **Manejar datos desequilibrados** está seleccionado en la pestaña Información básica de las Opciones de generación.

## Configuración del nugget del modelo Árboles aleatorios

En la pestaña Configuración para un nugget del modelo Árboles aleatorios, especifique opciones para confianzas y para la generación de SQL durante la puntuación del modelo. Esta pestaña sólo está disponible después de añadir el nugget de modelo a una ruta.

**Calcular confianzas** Para incluir confianzas en operaciones de puntuación, seleccione esta casilla de verificación. Al puntuar modelos en la base de datos, si excluye confianzas puede generar SQL más eficaz. Para los árboles de regresión, no se asignan las confianzas.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se genera SQL:

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

# Nuggets de modelo de árbol de decisión de Árbol C&R, CHAID, QUEST y C5.0

Los nuggets del modelo de árbol de decisión representan las estructuras de árbol para predecir u n campo de salida concreto descubierto por uno de los nodos de modelado de árbol de decisión (árbol C&R, CHAID, QUEST o C5.0). Se pueden generar tres modelos directamente desde el nodo de generación de árboles o indirectamente desde el Generador de árboles interactivos. Consulte el tema “El Generador de árboles interactivos” en la página 87 para obtener más información.

### Puntuación de modelos de árbol

Cuando se ejecuta una ruta que contiene un nugget de modelo de árbol, el resultado determinado depende del tipo de árbol.

- Para los árboles de clasificación (objetivo categórico), se añaden dos nuevos campos a los datos, estos contienen el valor predicho y la confianza para cada registro. La predicción se basa en la categoría más frecuente para el nodo terminal al que se asigna el registro; si una mayoría de encuestados de un nodo determinado es *sí*, la predicción para todos los registros asignados a dicho nodo es sí.
- En los árboles de regresión, solamente los valores predichos se generan; las confianzas no se asignan.
- Si lo prefiere, en los modelos CHAID, QUEST y C&RT, se puede añadir un campo adicional que indica el ID del nodo al que se asigna cada registro.

Los nuevos nombres de campos se derivan del nombre del modelo añadiendo prefijos. En el caso de C&RT, CHAID y QUEST, los prefijos son *$R-* para el campo de predicción, *$RC-* para el campo de confianza y *$RI-* para el campo identificador del nodo. En el caso de los árboles C5.0, los prefijos son *$C-* para el campo de predicción y *$CC-* para el campo de confianza. i hay presentes varios nodos del modelo de árbol, los nuevos nombres de campos incluirán números en el *prefijo* para distinguirlos si fuera necesario; por ejemplo *$R1-*, *$RC1-* y *$R2-*.

### Trabajo con nugget de modelo de árboles

Existen varias maneras de guardar o exportar información relacionada con el modelo.

**Nota:** Muchas de estas opciones también están disponibles en la ventana del generador de árboles.

Desde el generador de árboles o un modelo de árbol generado, puede:

- Generar un nodo Filtrar o Seleccionar en base al árbol actual. Consulte el tema "Generación nodos Seleccionar y Filtro" en la página 97 para obtener más información.
- Generar un nugget de conjunto de reglas que representa la estructura del árbol como un conjunto de reglas y define las ramas terminales del árbol. Consulte el tema "Generación de un conjunto de reglas desde un árbol de decisión" en la página 97 para obtener más información.
- Además, en el caso de los nugget de árbol generado, puede exportar el modelo en formato PMML. Consulte el tema "La paleta de modelos" en la página 41 para obtener más información. Si el modelo incluye divisiones personalizadas, esta información no se conserva en el PMML exportado. (La división se conserva, pero no el hecho de que sea personalizada y no seleccionada por el algoritmo.)
- Generar un gráfico basado en una parte seleccionada del árbol actual. *Nota*: sólo funciona para un nugget si se vincula a otros nodos en una ruta. Consulte el tema "Generación de gráficos" en la página 129 para obtener más información.
- Solamente en el caso de los modelos C5.0 aumentados, puede seleccionar **Árbol de decisión único (Lienzo)** o **Árbol de decisión único (Paleta de modelos generados)** para crear un nuevo conjunto de reglas único derivado de la regla seleccionada actualmente. Consulte el tema "Modelos C5.0 aumentados" en la página 128 para obtener más información.

**Nota:** si bien se sustituyó el nodo Crear regla por el nodo C&RT, los nodos Árbol de decisión de las rutas existentes creadas originalmente a partir de un nodo Crear regla seguirán funcionando correctamente.

## Nugget de modelo de árboles únicos

Si selecciona **Crear un árbol único** como objetivo principal del nodo de modelado, el nugget de modelo resultante contendrá las siguientes pestañas.

*Tabla 7. Pestañas del nugget de árbol único*

| Tabulación | Descripción | Más información |
|---|---|---|
| Modelo | Muestra las reglas que definen el modelo. | Consulte el tema "Reglas de modelo de árbol de decisión" para obtener más información. |
| Visor | Muestra la vista de árbol del modelo. | Consulte el tema "Visor de modelo de árbol de decisión" en la página 127 para obtener más información. |
| Resumen | Muestra información sobre los campos, los ajustes de versión y el proceso de estimación del modelo. | Consulte el tema "Información / Resumen de nugget de modelo" en la página 44 para obtener más información. |
| Configuración | Le permite especificar las opciones de confianzas y generación SQL durante la puntuación de modelos. | Consulte el tema "Configuración del nugget de modelo árbol de decisiones/conjunto de reglas" en la página 127 para obtener más información. |
| Anotación | Le permite añadir anotaciones descriptivas, especificar un nombre personalizado, añadir texto de información sobre herramientas y especificar las palabras clave de búsqueda para el modelo. | |

## Reglas de modelo de árbol de decisión

La pestaña Modelo de un nugget de árbol de decisión muestra las reglas que definen el modelo. Opcionalmente, también se pueden mostrar un gráfico de importancia de los predictores y un tercer panel con información acerca del historial, frecuencias y sustitutos.

**Nota:** Cuando selecciona la opción **Crear un modelo para conjuntos de datos de grandes dimensiones** en la pestaña Opciones de generación del nodo CHAID (panel Objetivo), la pestaña Modelo sólo muestra los detalles de regla de árbol.

### Reglas de árbol

El panel izquierdo muestra una lista de condiciones que definen la partición de los datos descubiertos por el algoritmo; esencialmente es una serie de reglas que se pueden utilizar para asignar registros individuales a los nodos hijo basándose en los valores de distintos predictores.

Los árboles de decisión funcionan particionando de forma recursiva los datos basados en valores de campos de entrada. Las particiones de los datos se denominan *ramas*. La rama inicial (a veces denominada *raíz*) engloba a todos los registros de datos. La raíz se divide en subconjuntos, o *ramas hijas*, basados en el valor de un determinado campo de entrada. Cada rama hija se puede dividir en subramas, que pueden, a su vez, volver a dividirse, y así sucesivamente. En el nivel inferior del árbol están las ramas que ya no tienen más divisiones. Dichas ramas se conocen como *ramas terminales* (u *hojas*).

### Detalles de regla de árbol

El explorador de reglas muestra los valores de entrada que definen cada partición o rama y un resumen de los valores de los campos de salida para los registros de dicha división. Para obtener información general sobre cómo utilizar el explorador de modelos, consulte "Examen de nuggets de modelo" en la página 42.

En el caso de las divisiones basadas en campos numéricos, la rama se representa mediante una línea con la forma:

```
nombrecampo relación valor [resumen]
```

donde *relación* es una relación numérica. Por ejemplo, una rama definida por valores mayores que 100 para el campo *ingresos* tendría la forma:

```
ingresos > 100 [resumen]
```

En el caso de divisiones basadas en campos simbólicos, la rama se representa mediante una línea con la forma:

```
nombrecampo = valor [resumen] o nombrecampo en [valores] [resumen]
```

donde *valores* representa los valores del campo que definen la rama. Por ejemplo, una rama que incluya registros donde el valor de *región* puede ser *Norte*, *Oeste* o *Sur* quedaría representada de la siguiente forma:

```
región en ["Norte" "Oeste" "Sur"] [resumen]
```

En el caso de las ramas terminales también se proporciona una predicción agregando una flecha y el valor predicho al final de la condición de la regla. Por ejemplo, una hoja definida por *ingresos* > 100 que predice un valor *alto* para el campo de salida quedaría representada de la siguiente forma:

```
ingresos > 100 [Modo: alto] → alto
```

El *resumen* de la rama se define de forma diferente a los campos de salida numéricos y simbólicos. En el caso de los árboles con campos de salida numéricos, el resumen será el valor *promedio* de la rama y el *efecto* de la rama consistirá en la diferencia entre el promedio de la rama y el promedio de su rama padre. En el caso de árboles con campos de salida simbólicos, el resumen será la *moda*, o el valor más frecuente, si se trata de los registros de la rama.

Para describir completamente una rama, necesita incluir la condición que define la rama más las condiciones que definen las divisiones en la parte superior del árbol. Por ejemplo, en el árbol:

```
ingresos > 100
  región = "Norte"
  región en ["Sur" "Este" "Oeste"]
    ingresos < = 200
```

la rama representada por la segunda línea viene definida por las condiciones *ingresos > 100* y *región = "Norte"*.

Si pulsa en **Mostrar u ocultar las cifras de ocurrencias y confianzas** en la barra de herramientas, cada regla también mostrará información acerca del número de registros a los que se aplica la regla (*Ocurrencias*), así como la proporción de registros para los que la regla es verdadera (*Confianza*).

## Importancia del predictor

Opcionalmente, en la pestaña Modelo también se puede mostrar un gráfico que indique la importancia relativa de cada predictor cuando se calcule el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos.

**Nota:** Este gráfico solo está disponible si **Calcular importancia de predictor** está seleccionado en la pestaña Analizar antes de generar el modelo. Consulte el tema "Importancia del predictor" en la página 44 para obtener más información.

## Información adicional del modelo

Si pulsa en la barra de herramientas la opción de **mostrar el panel de información adicional**, verá un panel con información detallada de la regla seleccionada en la parte inferior de la ventana. El panel de información contiene tres pestañas.

**Historial.** Esta pestaña rastrea las condiciones de división desde el nodo raíz hasta el nodo seleccionado. Así se obtiene una lista de condiciones que determina cuándo se asigna un registro al nodo seleccionado. Los registros para los que todas las condiciones sean verdaderas se asignarán a este nodo.

**Frecuencias.** En el caso de los modelos con campos objetivo simbólicos, esta pestaña muestra (para cada valor objetivo posible) el número de registros asignados a este nodo (en los datos de entrenamiento) que tienen dicho valor objetivo. La cifra de frecuencia, expresada como porcentaje (expresada con un máximo de tres decimales) también se muestra. En otros modelos con objetivos numéricos, esta pestaña está vacía.

**Sustitutos.** Si procede, se muestra cualquier sustituto del campo de división principal para el nodo seleccionado. Los sustitutos son campos alternativos que se usan en caso de que el valor predictor principal no esté presente en un determinado registro. El número máximo de sustitutos permitido para una división en particular se especifica en el nodo de generación de árbol, pero el número real depende de los datos de entrenamiento. En general, cuanto mayor sea la cantidad de datos perdidos, mayor será la probabilidad de usar sustitutos. En otros modelos de árboles de decisión esta pestaña está vacía.

**Nota:** Para que se incluya en el modelo, los sustitutos se deben identificar durante la fase de entrenamiento. Si la muestra de entrenamiento no tiene valores perdidos, no se identificarán sustitutos. Los registros con valores perdidos que se encuentren durante la comprobación o puntuación pasarán automáticamente al nodo hijo que tenga un mayor número de registros. Si se esperan valores perdidos durante la comprobación o puntuación, asegúrese de que los valores no están presentes en la muestra de entrenamiento. No hay sustitutos disponibles para los árboles CHAID.

## Visor de modelo de árbol de decisión

La pestaña Visor para un nugget de modelo de árbol de decisión se parece a la pantalla que aparece en el generador de árboles. La diferencia principal es que al examinar el nugget de modelo, no se puede hacer crecer ni modificar el árbol. El resto de opciones para visualizar y personalizar la presentación son similares en los dos componentes. Consulte el tema “Personalización de la vista del árbol” en la página 89 para obtener más información.

**Nota:** La pestaña Visor no se muestra para los nuggets de modelo CHAID incorporados, si selecciona la opción **Crear un modelo para conjuntos de datos muy grandes** en la pestaña Opciones de generación del panel Objetivo.

Al visualizar reglas divididas en la pestaña Visor, los corchetes significan que el valor adyacente se incluye en el rango mientras que los paréntesis indican que el valor adyacente se excluye del rango. Por lo tanto, la expresión (23,37] significa que el rango va desde el 23 exclusive hasta el 37 inclusive; es decir, desde el 24 hasta el 37. En la pestaña Modelo, la misma situación se mostraría como:

```
Edad > 23 y Edad <= 37
```

## Configuración del nugget de modelo árbol de decisiones/conjunto de reglas

La pestaña Configuración de un nugget de modelo de árbol de decisión o conjunto de reglas permite especificar opciones para confianzas y para la generación de SQL durante la puntuación de modelos. Esta pestaña está sólo disponible después de que el nugget de modelo se haya añadido a una ruta.

**Calcular confianzas** Seleccione incluir confianzas al puntuar operaciones. Al puntuar modelos en la base de datos, si excluye confianzas puede generar SQL más eficaz. Para los árboles de regresión, no se asignan las confianzas.

**Nota:** si selecciona la opción **Crear un modelo para conjuntos de datos de grandes dimensiones** en la pestaña Opciones de generación del panel Método para modelos CHAID, esta casilla de verificación solamente está disponible en los nuggets de modelos para objetivos categóricos para nominales o marcadores.

**Calcular puntuaciones de propensión bruta** En el caso de modelos con un objetivo de marca (que devuelve una predicción de sí o no), puede solicitar puntuaciones de propensión que indican la

probabilidad del resultado true especificado para el campo objetivo. Éstas se añaden a otros valores de predicción y confianza que pueden generarse durante la puntuación.

**Nota:** Si selecciona la opción **Crear un modelo para conjuntos de datos muy grandes** en la pestaña Opciones de generación - panel Método para modelos CHAID, esta casilla de verificación solo está disponible en nuggets de modelo con un objetivo categórico de marca.

**Calcular puntuaciones de propensión ajustada** Las puntuaciones de propensión en bruto se basan sólo en los datos de entrenamiento y pueden ser demasiado optimistas debido a la tendencia de muchos modelos a sobreajustar estos datos. Las propensiones ajustadas intentan compensar evaluando el rendimiento del modelo frente a una partición de comprobación o validación. Esta opción requiere que se haya definido un campo de partición en la ruta y que se hayan activado puntuaciones ajustadas de propensión en el modo de modelado antes de generar el modelo.

**Nota:** las puntuaciones ajustadas de propensión no están disponibles para modelos de árbol aumentado y de conjuntos de reglas. Consulte el tema "Modelos C5.0 aumentados" para obtener más información.

**Identificador de regla** Para modelos CHAID, QUEST y Árbol C&R, esta opción añade un campo en la salida de puntuación que indica el ID para el nodo de terminal al que se asigna cada registro.

**Nota:** cuando se selecciona esta opción, no está disponible la generación de SQL.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar convirtiendo a SQL nativo sin Soporte para valores perdidos** Si selecciona esta opción, se genera SQL nativo para puntuar el modelo dentro de la base de datos, sin la sobrecarga de manejar valores perdidos. Esta opción simplemente establece la predicción como nula (`$null$`) cuando se encuentra un valor perdido durante la puntuación de un caso.

  **Nota:** Esta opción no está disponible para modelos CHAID. Para otros tipos de modelo, solamente está disponible para árboles de decisión (no conjuntos de reglas).
- **Puntuar convirtiendo a SQL nativo con Soporte para valores perdidos** En el caso de los modelos CHAID, QUEST y de árbol C&R, puede generar SQL nativo para puntuar el modelo dentro de la base de datos con una compatibilidad completa de valores perdidos. Esto significa que se genera SQL de manera que los valores perdidos se gestionan tal y como se especificó en el modelo. Por ejemplo, los árboles C&RT utilizan reglas basadas en sustitutos garantizadas por el descendiente mayor.

  **Nota:** Para modelos C5.0, esta opción solo está disponible para conjuntos de reglas (no árboles de decisión).
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

## Modelos C5.0 aumentados

Esta característica está disponible en SPSS Modeler Professional y SPSS Modeler Premium.

Cuando se crea un modelo C5.0 aumentado (ya sea un conjunto de reglas o un árbol de decisión), realmente se crea un conjunto de modelos relacionados. El explorador de reglas del modelo para un

modelo C5.0 aumentado muestra la lista de modelos en el nivel superior de la jerarquía, junto con la precisión estimada de cada modelo y la precisión global de los modelos aumentados. Para examinar las reglas o divisiones de un determinado modelo, seleccione el modelo y expándalo como haría con una regla o rama en un modelo individual.

También puede extraer un determinado modelo del conjunto de modelos aumentados y crear un nuevo nugget de modelo Conjunto de reglas que solamente contenga dicho modelo. Para crear un nuevo conjunto de reglas a partir de un modelo C5.0 aumentado, seleccione el conjunto de reglas o árbol de interés y seleccione **Árbol de decisión único (Paleta de modelos generados)** o **Árbol de decisión único (Lienzo)** del menú Generar.

### Generación de gráficos

Los nodos Árbol proporcionan gran cantidad de información, sin embargo, es posible que no estén en un formato fácilmente accesible para usuarios empresariales. Puede producir gráficos de datos seleccionados para ofrecerlos de una forma que puedan incorporarse fácilmente en informes comerciales, presentaciones, etc. Por ejemplo, en la pestaña Modelo o Visor de un nugget de modelo o desde la pestaña Visor de un árbol interactivo, puede generar un gráfico para una parte seleccionada del árbol, creando únicamente un gráfico para los casos del árbol o nodo de rama seleccionado.

*Nota*: solamente puede generar un gráfico desde un nugget si se adjunta a otros nodos en una ruta.

Generación de un gráfico

El primer paso es seleccionar la información que se mostrará en el gráfico:

- En la pestaña Modelo de un nugget, expanda la lista de condiciones y reglas en el panel izquierdo y seleccione la lista que le interese.
- En la pestaña Visor de un nugget, expanda la lista de las ramas y seleccione el nodo que le interese.
- En la pestaña Visor de un árbol interactivo, expanda la lista de las ramas y seleccione el nodo que le interese.

  *Nota*: no puede seleccionar el nodo superior en una pestaña Visor.

La forma en la que cree un gráfico es la misma, con independencia de la forma en que seleccione que se muestren los datos:

1. En el menú Generar, seleccione **Gráfico (desde selección)**; de forma alternativa, puede pulsar en el botón **Gráfico (desde selección)** en la esquina inferior izquierda de la pestaña Visor. Aparecerá la pestaña Tablero básico.

   *Nota*: cuando abre la pestaña Tablero de esta forma, las únicas pestañas disponibles son Básico y Detallado.
2. Si utiliza la configuración de la pestaña Básico o Detallado, especifique los detalles que se mostrarán en el gráfico.
3. Pulse en Aceptar para generar el gráfico.

La cabecera del gráfico identifica los nodos o reglas seleccionadas que se incluirán.

## Nuggets de modelo para aumentar, realizar una agregación autodocimante y conjuntos de datos muy grandes

Si selecciona **Mejorar la precisión del modelo (aumento)**, **Mejorar la estabilidad del modelo (agregación autodocimante)** o **Crear un modelo para conjuntos de datos muy grandes** como objetivo principal del nodo de modelado, IBM SPSS Modeler crea un conjunto de múltiples modelos. Consulte el tema "Modelos de conjuntos" en la página 46 para obtener más información.

El nugget de modelo resultante contiene las siguientes pestañas. La pestaña Modelo ofrece una serie de vistas del modelo.

*Tabla 8. Pestañas disponibles en el nugget de modelo*

| Tabulación | Ver | Descripción | Más información |
|---|---|---|---|
| Modelo | Resumen del modelo | Muestra un resumen de la calidad y diversidad (excepto para los modelos aumentados y objetivos continuos) del conjunto, un medida de cuánto varían las predicciones en los diferentes modelos. | Consulte el tema "Resumen del modelo" en la página 46 para obtener más información. |
| | Importancia del predictor | Muestra un gráfico que indica la importancia relativa de cada predictor (campo de entrada) cuando se calcula el modelo. | Consulte el tema "Importancia del predictor" en la página 47 para obtener más información. |
| | Frecuencia de predictor | Muestra un gráfico en el que se indica la frecuencia relativa con la que se usa cada predictor en el conjunto de modelos. | Consulte el tema "Frecuencia de predictor" en la página 47 para obtener más información. |
| | Precisión de los modelos componentes | Traza un gráfico de la precisión predictiva de cada uno de los distintos modelos del conjunto. | |
| | Detalles de los modelos componentes | Muestra información sobre cada uno de los distintos modelos del conjunto. | Consulte el tema "Detalles de modelo de componente" en la página 47 para obtener más información. |
| | Información | Muestra información sobre los campos, los ajustes de versión y el proceso de estimación del modelo. | Consulte el tema "Información / Resumen de nugget de modelo" en la página 44 para obtener más información. |
| Configuración | | Le permite incluir confianzas en operaciones de puntuación. | Consulte el tema "Configuración del nugget de modelo árbol de decisiones/conjunto de reglas" en la página 127 para obtener más información. |
| Anotación | | Le permite añadir anotaciones descriptivas, especificar un nombre personalizado, añadir texto de información sobre herramientas y especificar las palabras clave de búsqueda para el modelo. | |

## Nuggets de modelo de conjunto de reglas de Árbol C&R, CHAID, QUEST, C5.0 y Apriori

Un nugget de modelo Conjunto de reglas representa las reglas para predecir un determinado campo de salida descubierto por el nodo de modelado de reglas de asociación (Apriori) o por uno de los nodos de generación de árboles (Árbol C&RT, CHAID, QUEST o C5.0). Para las reglas de asociación, el conjunto de reglas se debe generar desde un nugget de reglas sin refinar. En el caso de los árboles, se puede generar un conjunto de reglas desde el generador de árboles interactivo, desde un nodo de generación de modelos C5.0 o desde cualquier nugget de modelo de árbol. Al contrario que los nugget de reglas sin refinar, los nugget Conjunto de reglas pueden ubicarse en rutas a fin de generar predicciones.

Cuando se ejecuta una ruta que contiene el nugget Conjunto de reglas, se añaden a la ruta dos nuevos campos que contienen el valor predicho y la confianza para cada registro de los datos. Los nuevos nombres de campos se derivan del nombre del modelo añadiendo prefijos. Para los conjuntos de reglas de asociación, los prefijos son *$A-* para el campo de predicción y *$AC-* para el campo de confianza. Para los conjuntos de reglas C5.0, los prefijos son *$C-* para el campo de predicción y *$CC-* para el campo de confianza. En el caso de los conjuntos de reglas de C&RT, los prefijos son *$R-* para el campo de predicción y *$RC-* para el campo de confianza. En una ruta con varios nugget Conjunto de reglas en una serie que predicen los mismos campos de salida, los nuevos nombres de campos contendrán números en el *prefijo* para que se puedan distinguir entre sí. El primer nugget Conjunto de reglas de asociación de la ruta utilizará los nombres comunes, el segundo usará los nombres que comiencen por *$A1-* y *$AC1-*, mientras que el tercer nodo utilizará nombres que comiencen por *$A2-* y *$AC2-*, y así sucesivamente.

**Cómo se aplican las reglas.** Los conjuntos de reglas generados a partir de reglas de asociación son diferentes a otros nugget de modelo porque, en el caso de cualquier registro particular, se pueden generar varias predicciones y puede que no coincidan. Existen dos métodos para generar predicciones a partir de conjuntos de reglas.

**Nota:** los conjuntos de reglas de los árboles de decisión devuelven los mismos resultados independientemente del método que se utilice, ya que las reglas derivadas de un árbol de decisión se excluyen entre sí.

- **Elección.** Este método intenta combinar las predicciones de todas las reglas que se aplican al registro. Para cada registro, se examinan todas las reglas, y se utiliza cada regla que se aplica al registro a fin de generar una predicción y una confianza asociada. Se calcula la suma de cifras de confianza para cada valor de resultado, y se elige el valor con la mayor suma de confianza como predicción final. La confianza para la predicción final es la suma de confianzas para ese valor dividida por el número de reglas se activaron para ese registro.
- **Primer acierto.** Este método simplemente comprueba las reglas en orden. La primera regla que se aplica al registro es la que se utiliza para generar la predicción.

El método utilizado puede controlarse en las opciones de ruta.

**Generación de nodos.** El menú Generar permite crear nuevos nodos basados en el conjunto de reglas.

- **Nodo Filtrar** Crea un nuevo nodo Filtrar para filtrar los campos que no han utilizado las reglas en el conjunto de reglas.
- **Nodo Seleccionar** Crea un nuevo nodo Seleccionar para seleccionar los registros a los que se aplica la regla seleccionada. El nodo generado seleccionará los registros para los que todos los antecedentes de la regla son verdaderos. Esta opción requiere que se seleccione una regla.
- **Nodo Seguimiento de regla** Crea un nuevo Supernodo que calcula un campo que indica qué regla se utilizó para realizar la predicción para cada registro. Cuando se evalúa un conjunto de reglas mediante el método del primer acierto, éste es simplemente un símbolo que indica la primera regla que se activaría. Cuando se evalúa el conjunto de reglas mediante el método de votación, éste es una cadena más compleja que muestra la entrada al mecanismo de votación.
- **Árbol de decisión único** (Lienzo) / **Árbol de decisión único** (Paleta de modelos generados). Crea un nuevo nugget Conjunto de reglas único derivado de la regla seleccionada actualmente. Sólo está disponible para los modelos C5.0 **aumentados**. Consulte el tema “Modelos C5.0 aumentados” en la página 128 para obtener más información.
- **Modelo a paleta** Devuelve el modelo a la paleta de modelos. Esto resulta útil en las situaciones en que un colega le haya enviado una ruta que contenga un modelo y no el modelo en sí.

**Nota:** Las pestañas Configuración y Resumen del nugget Conjunto de reglas son idénticas a las de los modelos de árbol de decisión.

## Pestaña Modelo del conjunto de reglas

La pestaña Modelo de un nugget Conjunto de reglas muestra una lista de reglas que el algoritmo extrae de los datos.

El consecuente descompone las reglas (categoría predicha), que se presentan en el siguiente formato:

```
if antecedente_1
and antecedente_2
...
and antecedente_n
then valor predicho
```

donde `consecuente` y `antecedente_1` hasta `antecedente_n` son condiciones. La regla se interpreta como "para los registros donde `antecedente_1` hasta `antecedente_n` son todos verdaderos, `consecuente` es también probablemente verdadero." Si pulsa en el botón **Mostrar u ocultar las cifras de ocurrencias y confianzas** de la barra de herramientas, cada regla también mostrará información acerca del número de registros a los que se aplica la regla, es decir, para los que los antecedentes son verdaderos (**Ocurrencias**), así como la proporción de registros para los que toda la regla es verdadera (**Confianza**).

Recuerde que la confianza se calcula de manera levemente diferente en el caso de los conjuntos de reglas C5.0. C5.0 utiliza la siguiente fórmula para calcular la confianza de una regla:

```
(1 + número de registros donde la regla es correcta)
/
(2 + número de registros para los que los antecedentes de la regla son verdaderos)
```

Este cálculo de la confianza estima los ajustes para el proceso de generalización de las reglas a partir de un árbol de decisión (es decir, lo que hace C5.0 cuando crea un conjunto de reglas).

---

## Importación de proyectos desde AnswerTree 3.0

IBM SPSS Modeler puede importar proyectos guardados en AnswerTree 3.0 ó 3.1 desde el cuadro de diálogo estándar Archivo > Abrir, tal como se indica:

1. Seleccione en los menús de IBM SPSS Modeler:

   **Archivo** > **Abrir ruta**
2. En la lista desplegable Archivos de tipo, seleccione **Archivos de proyecto AT (*.atp, *.ats)**.

   Los proyectos importados se convierten en rutas de IBM SPSS Modeler con los siguientes nodos:
   - Un nodo de origen que define el origen de datos utilizado (por ejemplo, un origen de base de datos o un archivo de datos de IBM SPSS Statistics).
   - Para cada árbol del proyecto (puede haber varios) se crea un nodo Tipo que define las propiedades de los distintos campos (variables), incluidos el tipo, el rol (campo predictor o de entrada frente a campo predicho o de salida), los valores perdidos y otras opciones.
   - Para cada árbol del proyecto, se creará un nodo Partición, donde se realizan particiones de los datos para una muestra de comprobación o entrenamiento, y también se creará un nodo de generación de árboles, donde se definen los parámetros de generación del árbol (un nodo C&RT, QUEST o CHAID).
3. Para ver los árboles generados, ejecute la ruta.

Comentarios

- Por lo general, no es posible exportar a AnswerTree los árboles de decisión generados en IBM SPSS Modeler; la importación de AnswerTree a IBM SPSS Modeler es una acción monodireccional.
- Los beneficios definidos en AnswerTree no se conservan una vez importado el proyecto a IBM SPSS Modeler.

# Capítulo 7. Modelos de redes bayesianas

## Nodo Red bayesiana

El nodo **Red bayesiana** le permite crear un modelo de probabilidad combinando pruebas observadas y registradas con conocimiento del mundo real de "sentido común" para establecer la probabilidad de instancias utilizando atributos aparentemente no vinculados. El nodo se centra en las redes Naïve Bayes aumentado a árbol (TAN) y de manto de Markov que se utilizan principalmente para la clasificación.

Las redes bayesianas se utilizan para realizar predicciones en diferentes situaciones; algunos ejemplos son los siguientes:

- Selección de oportunidades de crédito con poco riesgo de fracaso.
- Estimación cuando se necesite reparar el equipo o piezas de recambio, en función de los datos de los sensores y los registros existentes.
- Solución de problemas de los clientes mediante herramientas de solución de problemas en línea.
- Diagnóstico y solución de problemas de redes de telefonía móvil en tiempo real.
- Evaluación de los riesgos potenciales y recompensas de proyectos de investigación y desarrollo para centrar los recursos en las mejores oportunidades.

Una red bayesiana es un modelo gráfico que muestra variables (que se suelen denominar **nodos**) en un conjunto de datos y las independencias probabilísticas o condicionales entre ellas. Las relaciones causales entre los nodos se pueden representar por una red bayesiana; sin embargo, los enlaces en la red (también denominados **arcos**) no representan necesariamente una relación directa de causa y efecto. Por ejemplo, una red bayesiana se puede utilizar para calcular la probabilidad de un paciente con una enfermedad concreta, con la presencia o no de algunos síntomas y otros datos relevantes, si las independencias probabilísticas entre síntomas y enfermedad son verdaderas, tal y como se muestra en el gráfico. Las redes son muy robustas en los puntos en los que falta información y realizan las mejores predicciones posibles utilizando la información disponible.

Lauritzen y Spiegelhalter crearon un ejemplo común y básico de una red bayesiana en 1988. También se conoce como modelo "Asia" y es una versión simplificada de una red que se puede utilizar para diagnosticar a los nuevos pacientes de un médico; la dirección de los enlaces corresponde por lo general a la causalidad. Cada nodo representa una faceta que se puede relacionar con el estado de un paciente; por ejemplo, "fumador" indica que se trata de un fumador habitual y "VisitaAsia" muestra que recientemente ha viajado a Asia. Los enlaces entre los nodos muestran las relaciones probabilísticas; por ejemplo, fumar aumenta las posibilidades de que el paciente padezca bronquitis y cáncer de pulmón, mientras que la edad parece estar relacionada únicamente con la posibilidad de desarrollar cáncer de pulmón. De la misma forma, las anomalías detectadas en una radiografía de los pulmones pueden estar causadas por tuberculosis o cáncer de pulmón, mientras que las posibilidades de que un paciente tenga dificultades respiratorias (disnea) aumentan si también padece bronquitis o cáncer de pulmón.

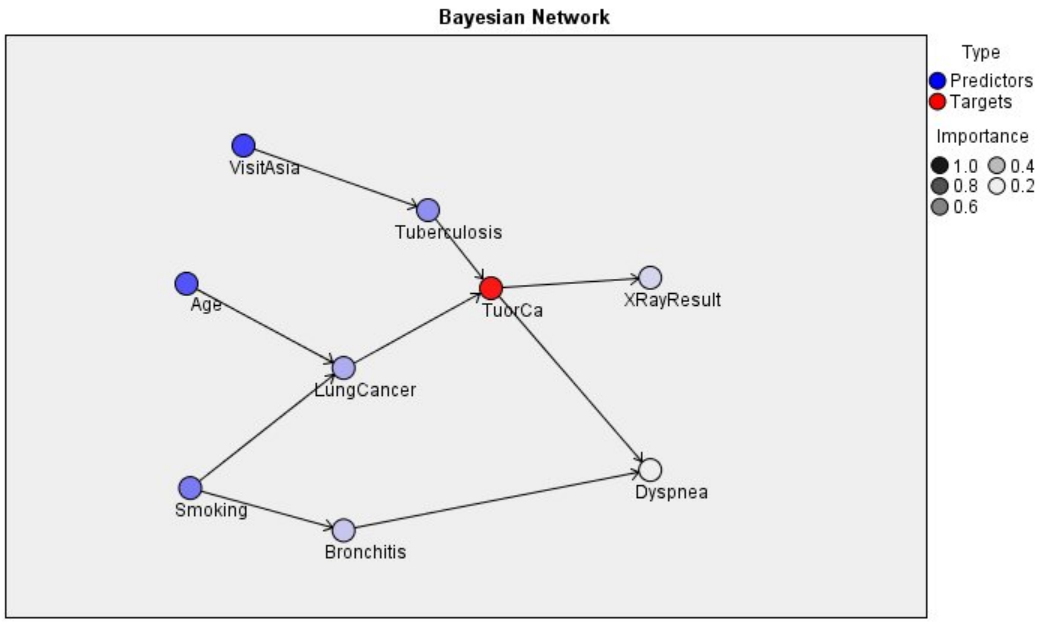



*Figura 29. Ejemplo de red Asia de Lauritzen y Spegelhalter*

Existen diferentes razones para elegir utilizar una red bayesiana:

- Es de gran ayuda para obtener información acerca de las relaciones causales. Permite conocer un área problemática y predecir las consecuencias de cualquier intervención.
- La red proporciona un método eficaz sin ajustar los datos en exceso.
- Puede obtener una vista clara de las relaciones que intervienen.

**Requisitos**. Los campos objetivo deben ser categóricos y pueden tener un nivel de medición *Nominal*, *Ordinal* o *Marca*. Las entradas pueden ser campos de cualquier tipo. Los campos de entrada continuos (rangos numéricos) se clasifican en intervalos de forma automática; sin embargo, si la distribución es asimétrica, puede obtener mejores resultados clasificando los campos en intervalos de forma manual, utilizando un nodo Intervalos antes del nodo Red bayesiana. Por ejemplo, utilice Intervalos óptimos si **Campo Supervisor** es el mismo que el campo **Objetivo** del nodo Red bayesiana.

**Ejemplo.** Un analista de un banco quiere poder predecir clientes o clientes potenciales, con posibilidades de impago de sus créditos. Puede utilizar un modelo de red bayesiana para identificar las características de los clientes con más posibilidades de impago y generar varios tipos diferentes de modelo para establecer el mejor método para predecir los clientes con más posibilidades de impago.

**Ejemplo.** Un operador de telecomunicaciones quiere reducir el número de clientes que dejan el negocio (denominado "abandono") y actualizar el modelo mensualmente utilizando los datos del mes anterior. Puede utilizar un modelo de red bayesiana para identificar las características de los clientes con más posibilidades de abandono y continuar formando el modelo cada mes con nuevos datos.

## Opciones de modelo de nodo de red bayesiana

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Construir modelo para cada división.** Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como campos de división. Consulte el tema "Generación de modelos divididos" en la página 28 para obtener más información.

**Partición.** Este campo permite especificar un campo usado para dividir los datos en muestras independientes para las fases de entrenamiento, prueba y validación de la generación de modelos. Si usa una muestra para generar el modelo y otra muestra distinta para comprobarlo, podrá obtener una buena indicación de la bondad del modelo a la hora de generar conjuntos de datos de mayor tamaño similares a los datos actuales. Si se han definido varios campos de partición mediante nodos Tipo o Partición, se deberá seleccionar un campo de partición simple en la pestaña Campos en todos los nodos de modelado que usen la partición. (Si solamente hay una partición, se usará automáticamente siempre que se active la partición.) Debe tener en cuenta que al aplicar la partición seleccionada en su análisis, también debe activar la partición en la pestaña Opciones del modelo para el nodo. (Si se elimina la selección de esta opción, se posibilita la desactivación de la partición sin cambiar la configuración del campo.)

**Divididos.** En modelos divididos, seleccione el campo o campos de división. Se trata de una acción similar a establecer el rol del campo en *Dividir* en un nodo Tipo. Sólo puede designar campos con un nivel de medición de **Marca**, **Nominal**, **Ordinal** o **Continuo** como campos de división. Los campos seleccionados como campos de división no se pueden utilizar como campos de destino, entrada, partición, frecuencia o ponderación. Consulte el tema "Generación de modelos divididos" en la página 28 para obtener más información.

**Continuar entrenando modelo existente.** Si selecciona esta opción, los resultados mostrados en la pestaña Modelo del nugget de modelo se generan y actualizan cada vez que se ejecuta el modelo. Por ejemplo, puede hacerlo cuando se haya añadido un origen de datos nuevo o actualizado a un modelo existente.

*Nota*: Sólo puede actualizar la red existente; no puede añadir o eliminar nodos o conexiones. Cada vez que entrena el modelo, la red tendrá la misma forma, sólo cambiarán las probabilidades condicionales y la importancia del predictor. No importa si los nuevos datos son muy similares a los datos antiguos, ya que espera que los mismos elementos sean significativos; sin embargo, si desea comprobar o actualizar *los elementos* significativos (en oposición a su significancia), deberá crear un nuevo modelo, es decir una nueva red.

**Tipo de estructura.** Seleccione la estructura que desea utilizar cuando cree la red bayesiana:

- **TAN.** El modelo de redes Naïve Bayes aumentado a árbol (TAN) crea un modelo de red bayesiana simple que es una mejora respecto al modelo Naïve Bayes estándar. Se debe a que cada predictor depende de otro predictor además de la variable objetivo, aumentando la precisión de la clasificación.
- **Manto de Markov.** Esto selecciona el conjunto de nodos del conjunto de datos que contiene los padres de la variable objetivo, sus hijos y los padres de sus hijos. Esencialmente, un manto de Markov identifica todas las variables de la red que son necesarias para predecir la variable objetivo. Este método de generar redes se considera más preciso; sin embargo, con conjuntos de datos más grandes se necesita más tiempo, debido al elevado número de variables implicadas. Para reducir el procesamiento, puede utilizar las opciones de **selección de características** de la pestaña Experto para seleccionar las variables que están muy relacionadas con la variable objetivo.

**Incluir paso de procesamiento previo de selección de características.** Si selecciona esta casilla podrá utilizar las opciones de **selección de características** de la pestaña Experto.

**Método de aprendizaje de parámetro.** Los parámetros de la red bayesiana hacen referencia a las probabilidades condicionales de cada nodo teniendo en cuenta los valores de sus padres. Son dos selecciones posibles que puede utilizar para controlar la tarea de estimar las tablas de probabilidades condicionales entre los nodos si se conocen los valores de los padres:

- **Verosimilitud máxima.** Seleccione esta casilla si utiliza un conjunto de datos grande. Ésta es la selección predeterminada.

- **Ajuste bayesiano de recuentos de casillas de tamaño reducido.** En conjuntos de datos más pequeños, existe el peligro de ajustar el modelo en exceso, así como la posibilidad de un elevado número de recuentos cero. Seleccione esta opción para evitar estos problemas, aplicando suavizado para reducir el efecto de cualquier recuento cero y los efectos de cálculos no fiables.

## Opciones de experto del nodo de red bayesiana

Las opciones de experto de nodo permiten ajustar el proceso de generación de modelos. Para acceder a estas opciones, active el modo **Experto** en la pestaña Experto.

**Valores perdidos.** De forma predeterminada, IBM SPSS Modeler utiliza sólo los registros que dispongan de valores válidos en todos los campos utilizados en el modelo (Esto se denomina a veces **eliminación según lista** de los valores perdidos). Si tiene muchos datos perdidos, descubrirá que este método elimina muchos registros, dejándole sin los datos suficientes para generar un buen modelo. En estos casos, puede anular la selección de la opción **Sólo usar registros completos**. IBM SPSS Modeler intenta utilizar tanta información como sea posible para estimar el modelo, incluidos los registros en los que algunos campos tienen valores perdidos. (Esto se denomina a veces *eliminación por pareja* de los valores perdidos.) No obstante, en algunas situaciones, el uso de registros incompletos de esta forma puede dar lugar a problemas computacionales a la hora de calcular el modelo.

**Añadir todas las probabilidades.** Especifica si se añaden las probabilidades de cada categoría del campo de resultados a cada registro procesado por el nodo. Si no se selecciona esta opción, sólo se añadirá la probabilidad de la categoría predicha.

**Comprobación de independencia.** Una comprobación de independencia determina si las observaciones relacionadas de dos variables son independientes entre sí. Seleccione el tipo de comprobación que se utilizará. Las opciones disponibles son:

- **Razón de verosimilitud.** Comprueba la independencia del objetivo y del predictor calculando una proporción entre la probabilidad máxima de un resultado en dos hipótesis diferentes.
- **Chi-cuadrado de Pearson.** Comprueba la independencia del objetivo y del predictor utilizando una hipótesis nula en la que las frecuencias relativas de las instancias de eventos observados siguen a una distribución de frecuencia especificada.

Los modelos de redes bayesianas realizan comprobaciones condicionales de independencia que utilizan variables adicionales más allá de los pares comprobados. Además, los modelos no sólo exploran las relaciones entre los objetivos y predictores, sino también las relaciones entre los predictores mismos.

**Nota:** Las opciones de prueba de Independencia solo están disponibles si selecciona **Incluir paso de procesamiento previo de selección de características** o un **Tipo de estructura** de Markov Blanket en la pestaña Modelo.

**Nivel de significación.** Utilizada junto con la configuración de comprobaciones de independencia, permite definir un valor de corte que se utilizará cuando realice las comprobaciones. Cuanto menor sea el valor, menos enlaces quedarán en la red; el valor predeterminado es 0,01.

**Nota:** Esta opción solo está disponible si selecciona **Incluir paso de procesamiento previo de selección de características** o un **Tipo de estructura** de Markov Blanket en la pestaña Modelo.

**Tamaño del conjunto de condición máxima.** El algoritmo para crear una estructura de manto de Markov utiliza conjuntos de condiciones de aumento de tamaño para realizar comprobaciones de independencia y eliminar enlaces innecesarios de la red. Como las comprobaciones con un alto número de variables de condición requieren más tiempo y memoria de procesamiento, puede limitar el número de variables que se van a incluir. Puede ser de gran utilidad si procesa datos con grandes dependencias entre muchas variables. Tenga en cuenta, sin embargo, que la red resultante puede contener algunos enlaces innecesarios.

Establezca el número máximo de variables de condición que se utilizarán para la comprobación de la independencia. El valor predeterminado es 5.

**Nota:** Esta opción solo está disponible si selecciona **Incluir paso de procesamiento previo de selección de características** o un **Tipo de estructura** de Markov Blanket en la pestaña Modelo.

**Selección de características.** Estas opciones permiten limitar el números de entradas utilizadas cuando procese el modelo para acelerar el proceso de generación de modelo. Es un método especialmente útil cuando cree una estructura de un manto de Markov, debido a un posible número elevado de entradas potenciales; permite seleccionar las entradas significativas en función de su variable objetivo.

**Nota:** Las opciones de selección de características solo están disponibles si selecciona **Incluir paso de procesamiento previo de selección de características** en la pestaña Modelo.

- **Entradas siempre seleccionadas.** Utilizando el selector de campos (botón a la derecha del campo de texto), seleccione los campos en el conjunto de datos que siempre se van a utilizar al generar el modelo de red bayesiana. El campo objetivo siempre está seleccionado. Tenga en cuenta que la red bayesiana todavía puede descartar elementos de esta lista durante el proceso de generación del modelo, si otras pruebas no los consideran significativos. De esta forma, esta opción simplemente garantiza que los elementos de la lista se utilizan en el propio proceso de generación del modelo, no que aparecerán por completo en el modelo bayesiano resultante.
- **Número máximo de entradas.** Especifique el número total de entradas del conjunto de datos que se utilizarán cuando genere el modelo de red bayesiana. El mayor número que introduzca es el número total de entradas en el conjunto de datos.

  **Nota:** Si el número de campos seleccionados en **Entradas siempre seleccionadas** excede el valor de **Número máximo de entradas**, se muestra un mensaje de error.

## Nugget de modelo de red bayesiana

**Nota:** si ha seleccionado **Continuar entrenando modelo existente** en la pestaña Modelo del nodo de modelado, la información que aparece en la pestaña Modelo del nugget de modelo se actualiza cada vez que se vuelve a generar el modelo.

La pestaña Modelo del nugget de modelo se dividirá en dos paneles.

### Panel izquierdo

**Básica** Esta vista contiene una red de gráficos de nodos que muestra la relación entre el objetivo y sus predictores más importantes, y la relación entre los predictores. La importancia de cada predictor se muestra según la densidad del color; un color más fuerte muestra un predictor importante y viceversa.

Los valores de agrupación en intervalos de los nodos que representan un rango se muestran en una ayuda contextual cuando pasa el puntero sobre el nodo.

Puede utilizar las herramientas gráficas de IBM SPSS Modeler para interactuar con el gráfico, así como editarlo y guardarlo. Por ejemplo, para su uso en otras aplicaciones como MS Word.

**Consejo:** si la red contiene muchos nodos, puede pulsar en un nodo para seleccionarlo y arrastrarlo para que el gráfico sea más legible.

**Distribución** Esta vista muestra las probabilidades condicionales de cada nodo de la red como un minigráfico. Pase el puntero por encima de un gráfico para visualizar sus valores en una ayuda contextual.

### Panel derecho

**Importancia de predictor** Muestra un gráfico que indica la importancia relativa de cada predictor cuando se calcula el modelo. Si desea obtener más información, consulte “Importancia del predictor” en la página 44.

**Probabilidades condicionales** Si selecciona un nodo o un minigráfico de distribución en el panel izquierdo, se muestra la tabla de probabilidades condicionales asociadas en el panel derecho. Esta tabla contiene el valor de probabilidad condicional de cada valor de nodo y combinación de valores en sus nodos padre. Además, incluye el número de registros observados para cada valor de registro y cada combinación de valores en los nodos padre.

## Parámetros de modelo de red bayesiana

La pestaña Configuración de un nugget de modelo de red bayesiana especifica las opciones para modificar el modelo generado. Por ejemplo, puede utilizar el nodo de red bayesiana para generar varios modelos diferentes con los mismos datos y la misma configuración y, a continuación, usar esta pestaña en cada modelo para modificar ligeramente la configuración y comprobar cómo afecta eso a los resultados.

**Nota:** Esta pestaña está sólo disponible después de que el nugget de modelo se haya añadido a una ruta.

**Calcular puntuaciones de propensión en bruto.** En el caso de modelos con un objetivo de marca (que devuelve una predicción de sí o no), puede solicitar puntuaciones de propensión que indican la probabilidad del resultado true especificado para el campo objetivo. Éstas se añaden a otros valores de predicción y confianza que pueden generarse durante la puntuación.

**Calcular puntuaciones de propensión ajustada.** Las puntuaciones de propensión en bruto se basan sólo en los datos de entrenamiento y pueden ser demasiado optimistas debido a la tendencia de muchos modelos a sobreajustar estos datos. Las propensiones ajustadas intentan compensar evaluando el rendimiento del modelo frente a una partición de comprobación o validación. Esta opción requiere que se haya definido un campo de partición en la ruta y que se hayan activado puntuaciones ajustadas de propensión en el modo de modelado antes de generar el modelo.

**Añadir todas las probabilidades** Especifica si se añaden las probabilidades de cada categoría del campo de resultados a cada registro procesado por el nodo. Si no se selecciona esta opción, sólo se añadirá la probabilidad de la categoría predicha.

El valor predeterminado de esta casilla de verificación está determinado por la casilla de verificación correspondiente en la pestaña Experto del nodo de modelado. Consulte el tema “Opciones de experto del nodo de red bayesiana” en la página 136 para obtener más información.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

## Resumen de modelo de red bayesiana

La pestaña Resumen de un nugget de modelo muestra información sobre el propio modelo (*Análisis*), los campos utilizados en el modelo (*Campos*), la configuración utilizada al generar el modelo (*Configuración de creación*) y el entrenamiento del modelo (*Resumen de entrenamiento*).

Cuando se examina el nodo por primera vez, los resultados de la pestaña Resumen aparecen contraídos. Para ver los resultados de interés, utilice el control de expansión situado a la izquierda de un elemento con objeto de desplegarlo, o bien pulse en el botón **Expandir todo** para mostrar todos los resultados. Para ocultar los resultados cuando haya terminado de consultarlos, utilice el control de expansión con objeto de contraer los resultados específicos que desee ocultar o pulse en el botón **Contraer todo** para contraer todos los resultados.

**Análisis.** Muestra información sobre el modelo específico.

**Campos.** Enumera los campos utilizados como objetivo y entradas en la generación del modelo.

**Configuración de creación.** Contiene información sobre la configuración que se utiliza en la generación del modelo.

**Resumen de entrenamiento.** Muestra el tipo del modelo, la ruta utilizada para crearlo, el usuario que lo creó, cuándo se generó y el tiempo que se tardó en generar el modelo.

# Capítulo 8. Redes neuronales

Una **red neuronal** puede aproximar una amplia gama de modelos predictivos con demandas mínimas sobre la estructura y asunción de modelos. La forma de las relaciones está determinada durante el proceso de aprendizaje. Si una relación lineal entre el objetivo y los predictores es apropiada, los resultados de la red neuronal deben aproximarse mucho a los del modelo lineal tradicional. Si una relación no lineal es más apropiada, la red neuronal aproximará automáticamente la estructura de modelo "correcta".

El equilibrio de esta flexibilidad es que la red neuronal no es fácilmente interpretable. Si intenta explicar un proceso subyacente que genera las relaciones entre el objetivo y los predictores, se debería utilizar mejor un modelo estadístico más tradicional. Sin embargo, si la interpretabilidad del modelo no es importante, puede obtener buenas predicciones utilizando una red neuronal.

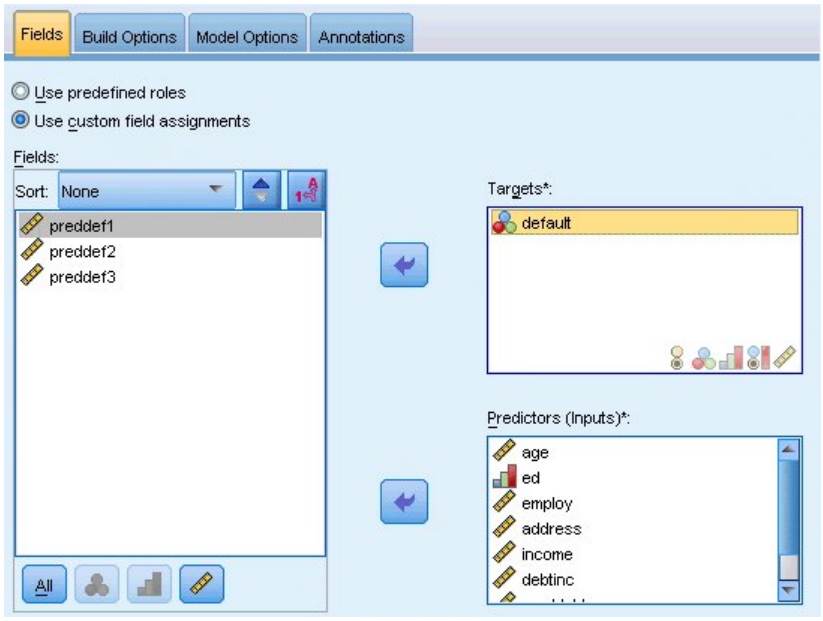



*Figura 30. pestaña Campos*

**Requisitos del campo.** Debe haber al menos un objetivo y una entrada. Se ignorarán los campos establecidos en Ambos o Ninguno. No hay restricciones de nivel de medición en los objetivos o en los predictores (entradas). Consulte el tema “Opciones de los campos del nodo de modelado” en la página 31 para obtener más información.

## El modelo de redes neuronales

Las redes neuronales son modelos simples del funcionamiento del sistema nervioso. Las unidades básicas son las **neuronas**, que generalmente se organizan en **capas**, como se muestra en la siguiente ilustración.

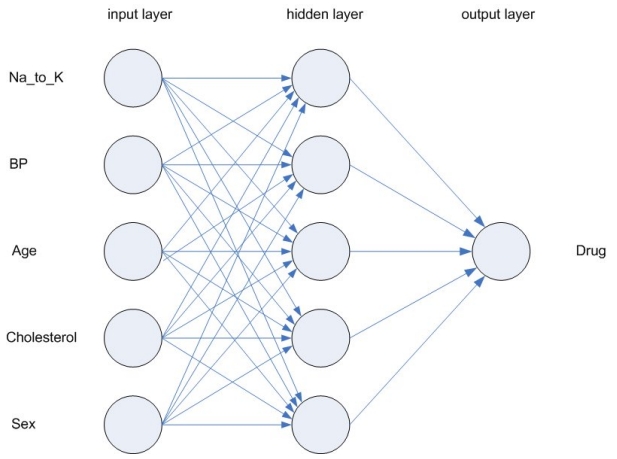



*Figura 31. Estructura de una red neuronal*

Una **red neuronal** es un modelo simplificado que emula el modo en que el cerebro humano procesa la información: Funciona simultaneando un número elevado de unidades de procesamiento interconectadas que parecen versiones abstractas de neuronas.

Las unidades de procesamiento se organizan en capas. Hay tres partes normalmente en una red neuronal : una **capa de entrada**, con unidades que representan los campos de entrada; una o varias **capas ocultas**; y una **capa de salida**, con una unidad o unidades que representa el campo o los campos de destino. Las unidades se conectan con fuerzas de conexión variables (o **ponderaciones**). Los datos de entrada se presentan en la primera capa, y los valores se propagan desde cada neurona hasta cada neurona de la capa siguiente. al final, se envía un resultado desde la capa de salida.

La red aprende examinando los registros individuales, generando una predicción para cada registro y realizando ajustes a las ponderaciones cuando realiza una predicción incorrecta. Este proceso se repite muchas veces y la red sigue mejorando sus predicciones hasta haber alcanzado uno o varios criterios de parada.

Al principio, todas las ponderaciones son aleatorias y las respuestas que resultan de la red son, posiblemente, disparatadas. La red aprende a través del **entrenamiento**. Continuamente se presentan a la red ejemplos para los que se conoce el resultado, y las respuestas que proporciona se comparan con los resultados conocidos. La información procedente de esta comparación se pasa hacia atrás a través de la red, cambiando las ponderaciones gradualmente. A medida que progresa el entrenamiento, la red se va haciendo cada vez más precisa en la replicación de resultados conocidos. Una vez entrenada, la red se puede aplicar a casos futuros en los que se desconoce el resultado.

## Uso de redes neuronales con rutas heredadas

La versión 14 de IBM SPSS Modeler ha introducido un nuevo nodo de red neuronal, que admite técnicas de aumento y agregación autodocimante y optimización para conjuntos de datos de grandes dimensiones. Las rutas existentes contienen en nodo anterior seguirán generando y puntuando modelos en esta versión. Sin embargo, esta compatibilidad se eliminará en versiones futuras, por lo que se recomienda utilizar la nueva versión desde ahora.

Desde la versión 13 en adelante, los campos con valores desconocidos (es decir, valores que no están presentes en los datos de entrenamiento) ya no son tratados automáticamente como valores perdidos y se puntúan con el valor `$null$`. Por lo tanto, si desea puntuar campos con valores desconocidos como no nulos mediante un modelo de red neuronal anterior (anterior a 13) a la versión 13 o posterior, debería marcar los valores desconocidos como valores perdidos (por ejemplo, por medio del nodo Tipo).

Tenga en cuenta que, por motivos de compatibilidad, las rutas de legado que todavía contienen el nodo antiguo todavía pueden estar utilizando la opción *Limitar tamaño de conjunto* en **Herramientas > Propiedades de ruta > Opciones**; esta opción sólo se aplica a las redes de Kohonen y los nodos *K*-Medias de la versión 14 en adelante5.

## Objetivos

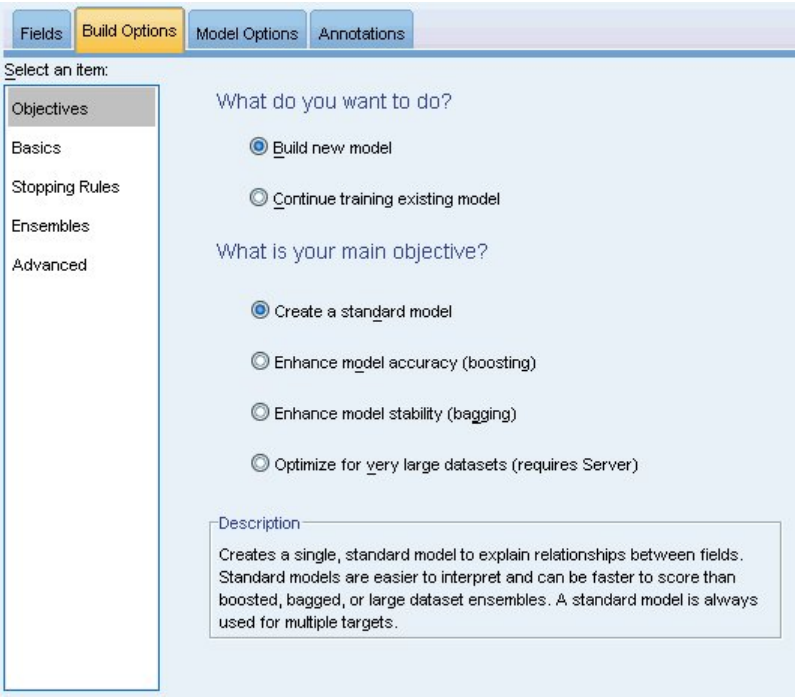



*Figura 32. Configuración de objetivos*

### ¿Qué desea hacer?

- **Crear un modelo nuevo.** Crear un modelo totalmente nuevo. Éste es el funcionamiento habitual del nodo.
- **Continuar entrenando un modelo existente.** El entrenamiento continúa con el último modelo creado correctamente por el nodo. Esto permite actualizar un modelo existente sin tener que acceder a los datos originales. Además, puede dar como resultado un rendimiento significativamente más rápido ya que sólo se introducen en la ruta los registros nuevos o actualizados. Los detalles del modelo anterior se almacenan con el nodo de modelado, lo que permite utilizar esta opción incluso si el nugget de modelo anterior ya no está disponible en la ruta o la paleta de modelos.

  **Nota:** Cuando esta opción está habilitada, todos los demás controles de las pestañas Campos y Opciones de generación están inhabilitados.

**¿Cuál es su objetivo principal?** Seleccione el objetivo apropiado.

- **Crear un modelo estándar.** El método genera un modelo simple para predecir el destino mediante los predictores. Por lo general, los modelos estándar son más fáciles de interpretar y pueden puntuarse más rápido que los conjuntos por aumento, agregación auntodocimante o los conjuntos de datos muy grandes.

  **Nota:** Para modelos segmentados, para utilizar esta opción con **Continuar entrenando un modelo existente** debe estar conectado a Analytic Server.

- **Mejorar la precisión del modelo (aumento).** El método genera un modelo de conjunto mediante el aumento, que genera una secuencia de modelos para obtener predicciones más precisas. Se puede tardar más tiempo en generar y puntuar conjuntos que un modelo estándar.

  El aumento produce una sucesión de "modelos de componente", cada uno de ellos basados en el conjunto de datos completo. Antes de crear cada modelo de componente sucesivo, los registros se ponderan en función de los residuos del modelo del componente anterior. Los casos con residuos de grandes dimensiones tienen ponderaciones de análisis relativamente superiores para que el siguiente modelo de componente se centre en predecir correctamente estos registros. Juntos, estos modelos de componentes forman un modelo de conjunto. El modelo de conjunto puntúa algunos registros con una regla de combinación; las reglas disponibles dependen del nivel de medición del destino.
- **Mejorar la estabilidad del modelo (agregación autodocimante).** El método genera un modelo de conjunto mediante la agregación autodocimante, que genera varios modelos para obtener predicciones más fiables. Se puede tardar más tiempo en generar y puntuar conjuntos que un modelo estándar.

  La agregación autodocimante produce replicaciones del conjunto de datos de entrenamiento mediante muestreo con repetición del conjunto de datos original. Crea muestras de bootstrap de igual tamaño al conjunto de datos original. Es decir, se crea un "modelo de componente" de cada replicación. Juntos, estos modelos de componentes forman un modelo de conjunto. El modelo de conjunto puntúa algunos registros con una regla de combinación; las reglas disponibles dependen del nivel de medición del destino.
- **Crear un modelo para conjuntos de datos muy grandes.** El método genera un modelo de conjunto dividiendo el conjunto de datos en bloques de datos independientes. Seleccione esta opción si su conjunto de datos es demasiado grande para construir cualquiera de los modelos anteriores o para la generación incremental de modelos. Puede que se tarde menos tiempo en generar esta opción, pero se puede tardar más tiempo en puntuarla que un modelo estándar.

Cuando existen objetivos múltiples, este método sólo creará un modelo estándar, sin importar el objetivo seleccionado.

## Conceptos básicos

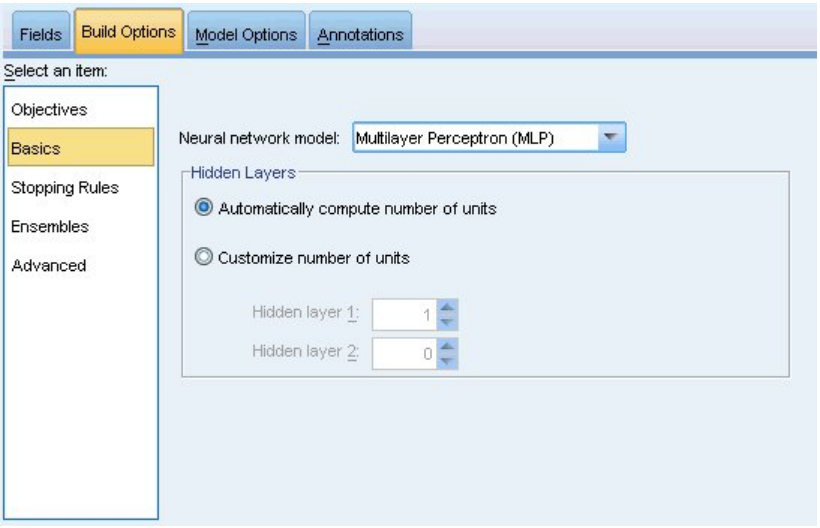



*Figura 33. Configuración básica*

**Modelo de red neuronal.** El tipo de modelo determina cómo la red conecta los predictores con los objetivos a través de las capas ocultas. Los **perceptrones multicapa (PMC)** permiten relaciones más complejas con el coste posible de aumentar el tiempo de entrenamiento y puntuación. La **función de base**

**radial (RBF)** puede tener tiempos de entrenamiento y puntuación inferiores, con el coste posible de una potencia de predicción reducida en comparación con PMC.

**Capas ocultas.** Las capas ocultas de una red neuronal contienen unidades no observables. El valor de cada unidad oculta es alguna función de los predictores; la forma exacta de la función depende en parte del tipo de red. Los perceptrones multicapa pueden tener una o dos capas ocultas; la red de función de base radial puede tener una capa oculta.

- **Calcular automáticamente el número de unidades.** Esta opción construye una red con una capa oculta y calcula el "mejor" número de unidades en la capa oculta.
- **Personalizar el número de unidades.** Esta opción le permite especificar el número de unidades en cada capa oculta. La primera capa oculta debe tener al menos una unidad. La especificación de 0 unidades para la segunda capa oculta construye perceptrones multicapa con una única capa oculta.

  *Note:* debe seleccionar valores de modo que el número de nodos no supere el número de predictores continuos junto con el número total de categorías en todos los predictores categóricos (marca, nominal u ordinal).

## Reglas de parada


*Figura 34. Configuración de reglas de parada*

Son las reglas que determinan cuándo detener el entrenamiento de las redes de perceptrones multicapa; esta configuración se ignora cuando se utiliza el algoritmo de función de base radial. El entrenamiento continúa al menos un ciclo (lectura de datos) y puede detenerse luego según los siguientes criterios.

**Emplear el tiempo de entrenamiento máximo (por modelo de componente).** Seleccione si se especifica un número máximo de minutos para ejecutar el algoritmo. Especificar un número superior a 0. Cuando se construye un modelo de conjunto, es el tiempo de entrenamiento permitido para cada modelo de componente del conjunto. Tenga en cuenta que el entrenamiento puede superar ligeramente el límite de tiempo especificado para completar el ciclo actual.

**Personalizar el número máximo de ciclos de entrenamiento.** El número máximo de ciclos de entrenamiento permitidos. Si se supera el número máximo de ciclos, el entrenamiento se detiene. Especifique un entero mayor que 0.

**Utilizar precisión mínima.** Seleccione esta opción para que el entrenamiento continúe hasta alcanzar la precisión especificada. Aunque no debería ocurrir, puede interrumpir el entrenamiento en cualquier momento y guardar la red con la mejor precisión obtenida hasta el momento.

El algoritmo de entrenamiento también se detendrá si el error en el conjunto de prevención sobreajustado no disminuye tras cada ciclo, si el campo relativo en el error de entrenamiento es pequeño, o si el índice del error de entrenamiento actual es pequeño comparado con el error inicial.

## Conjuntos

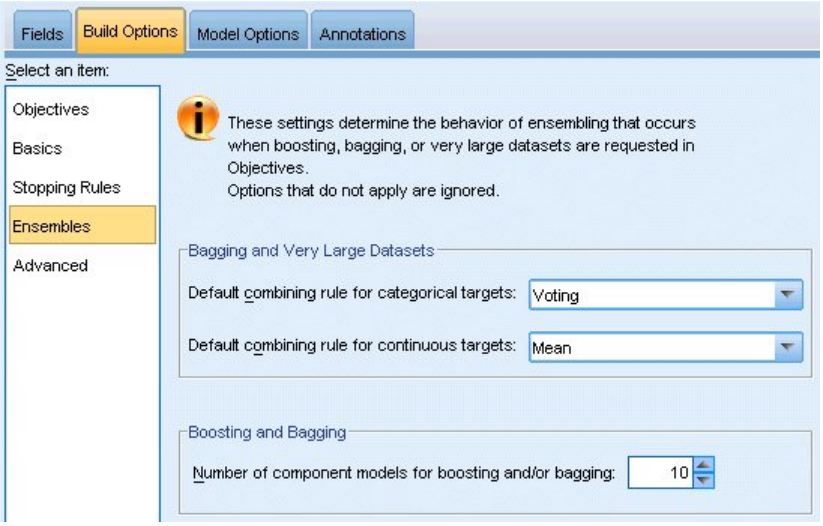



*Figura 35. Configuración de conjuntos*

Estos ajustes determinan el comportamiento de la agrupación que se produce cuando los conjuntos de datos de gran tamaño o de aumento o agregación autodocimante son obligatorios en Objetivos. Las opciones no aplicables al objetivo seleccionado se ignorarán.

**Agregación autodocimante y conjuntos de datos muy grandes.** Al puntuar un conjunto, ésta es la regla utilizada para combinar los valores predichos a partir de los modelos básicos para calcular el valor de puntuación del conjunto.

- **Regla de combinación predeterminada para objetivos categóricos.** Los valores predichos de conjunto para objetivos categóricos pueden combinarse mediante votación, la mayor probabilidad o la mayor probabilidad media. **Votación** selecciona la categoría que tenga la mayor probabilidad más frecuentemente entre los modelos básicos. **La mayor probabilidad** selecciona la categoría que logra la mayor probabilidad individual entre todos los modelos básicos. **Mayor probabilidad media** selecciona la categoría con el valor más elevado cuando se calcula la media de las probabilidades de categoría entre los modelos básicos.
- **Regla de combinación predeterminada para objetivos continuos.** Los valores predichos de conjunto para objetivos continuos pueden combinarse mediante la media o mediana de los valores predichos a partir de los modelos básicos.

Tenga en cuenta que cuando el objetivo es mejorar la precisión del modelo, se ignoran las selecciones de reglas de combinación. El aumento siempre utiliza un voto de mayoría ponderada para puntuar objetivos categóricos y una mediana ponderada para puntuar objetivos continuos.

**Aumento y agregación autodocimante.** Especifique el número de modelos básicos que debe generarse cuando el objetivo es mejorar la precisión o estabilidad del modelo; en el caso de la agregación autodocimante, se trata del número de muestras de bootstrap. Debe ser un número entero positivo.

## Avanzados

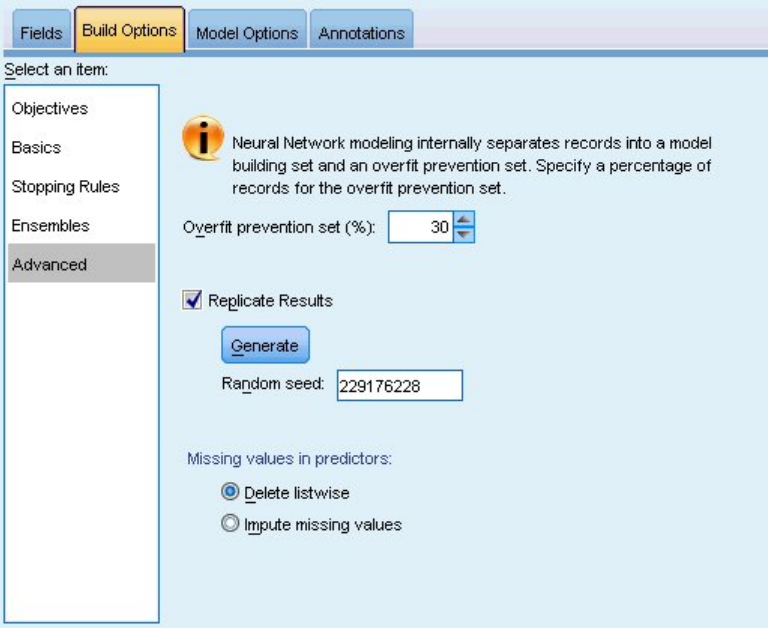



*Figura 36. Configuración avanzada*

La configuración avanzada controla las opciones que no se ajustan bien en otros grupos de configuraciones.

**Conjunto de prevención sobreajustado.** El método de red neuronal divide los registros de manera interna en un conjunto de creación de modelos y un conjunto de prevención sobreajustado, el cual es un conjunto independiente de registros de datos utilizado para realizar un seguimiento de errores durante la formación para evitar que el método modele una variación atribuible al azar en los datos. Especifique un porcentaje de registros. El valor por omisión es 30.

**Replicar resultados.** Al establecer una semilla aleatoria podrá replicar análisis. Especifique un entero o pulse en **Generar**, lo que creará un entero pseudo-aleatorio entre 1 y 2147483647, ambos inclusive. De forma predeterminada, los análisis se replican con la semilla 229176228.

**Valores perdidos en predictores.** Especifica cómo tratar los valores perdidos. **Eliminar de lista** retira los registros con valores perdidos en predictores de la creación de modelos. **Imputar valores perdidos** sustituirá los valores perdidos de los predictores y utilizará esos registros en el análisis. Los campos continuos imputan la media de los valores observados mínimos y máximos; los campos categóricos imputan la categoría que se produce con mayor frecuencia. Tenga en cuenta que los registros con valores perdidos en cualquier otro campo especificado en la pestaña Campos se eliminan siempre de la creación de modelos.

# Opciones de modelo

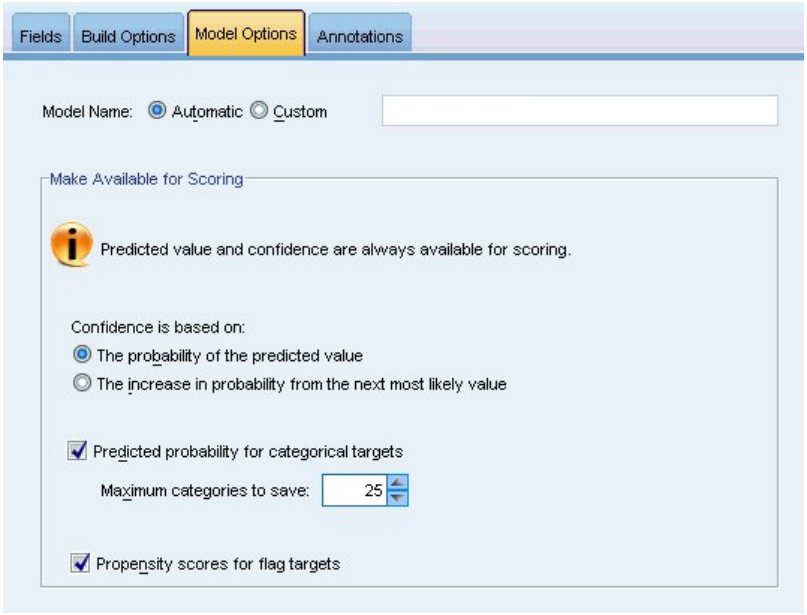



*Figura 37. Pestaña Opciones de modelo*

**Nombre del modelo.** Puede generar el nombre del modelo automáticamente tomando como base los campos objetivo o especificar un nombre personalizado. El nombre generado automáticamente es el nombre del campo objetivo. Si existen objetivos múltiples, el nombre del modelo se forma con los nombres de campos en orden, conectados por símbolos &. Por ejemplo, si *campo1 campo2 campo3* son objetivos, el nombre de modelo es: *campo1 & campo2 & campo3*.

**Dejar disponible para puntuación.** Cuando se puntúa el modelo, se crearán los elementos seleccionados en este grupo. El valor predicho (para todos los objetivos) y la confianza (para objetivos categóricos) se calculan siempre cuando se puntúa el modelo. La confianza calculada puede basarse en la probabilidad del valor predicho (la probabilidad predicha más alta) o la diferencia entre la probabilidad predicha más alta y la segunda probabilidad predicha más alta.

- **Probabilidad predicha para objetivos categóricos.** Genera las probabilidades predichas para objetivos categóricos. Se crea un campo para cada categoría.
- **Puntuaciones de propensión para objetivos de marca.** En el caso de modelos con un objetivo de marca (que devuelve una predicción de sí o no), puede solicitar puntuaciones de propensión que indican la probabilidad del resultado true especificado para el campo objetivo. El modelo produce puntuaciones de propensión en bruto; si hay particiones activas, el modelo también producirá puntuaciones de propensión ajustadas en función de la partición de prueba.

## Resumen del modelo

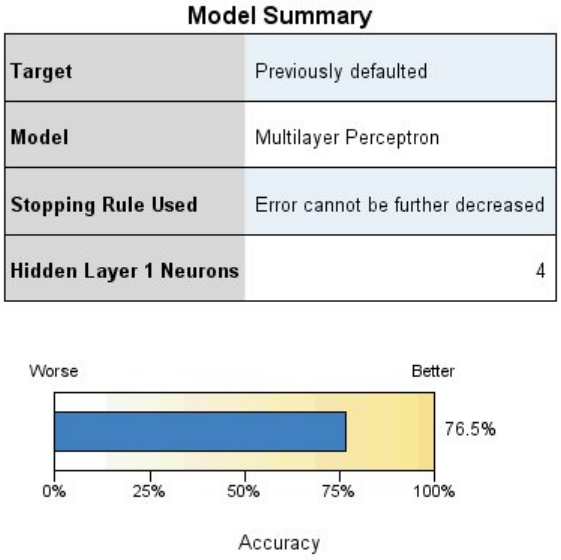



*Figura 38. Vista Resumen del modelo de red neuronal*

La vista Resumen del modelo es una instantánea, un resumen visual de la precisión de la clasificación o de la predicción de la red neuronal.

**Resumen del modelo.** La tabla identifica el objetivo, el tipo de red neuronal entrenada, la regla de parada que ha detenido el entrenamiento (mostrado si se ha entrenado una red de perceptrones multicapa) y el número de neuronas en cada capa oculta de la red.

**Calidad de la red neuronal.** El gráfico muestra la precisión del modelo final, que se presenta en el formato mayor es mejor. Para un objetivo categórico, es simplemente el porcentaje de registros para el que el valor predicho hace coincidir el valor observado. Para un objetivo continuo, la precisión se especifica como el valor $R^2$.

**Objetivos múltiples.** Si hay objetivos múltiples, cada objetivo se muestra en la fila **Objetivo** de la tabla. La precisión mostrada en el gráfico es la media de las precisiones de objetivos individuales.

# Importancia del predictor

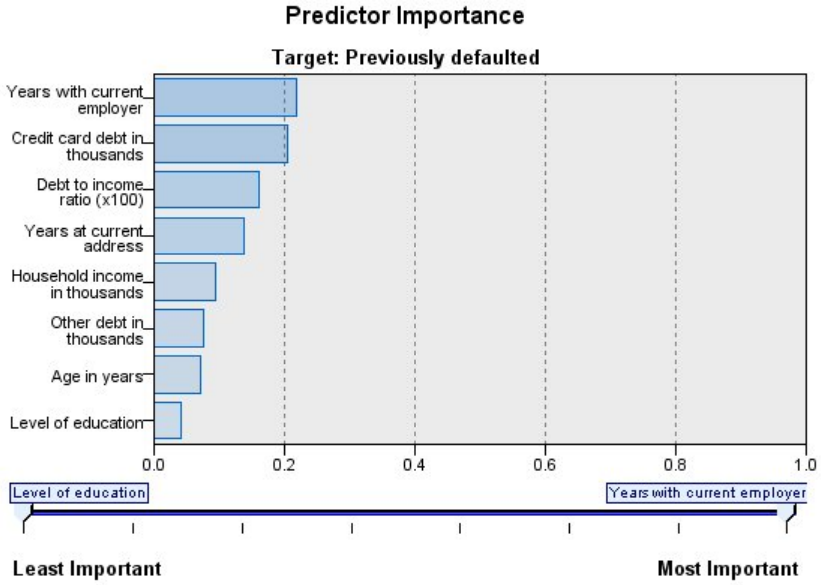



*Figura 39. Vista de importancia del predictor*

Es normal centrar los esfuerzos de modelado en los campos predictores más importantes y valorar la omisión de aquellos con menor relevancia. El gráfico de importancia de los predictores le ayuda a hacerlo indicando la importancia relativa de cada predictor en la estimación del modelo. Como los valores son relativos, la suma de valores de todos los predictores de la visualización es 1.0. La importancia del predictor no está relacionada con la precisión del modelo. Sólo está relacionada con la importancia de cada predictor a la hora de realizar una predicción, no con si la predicción es o no precisa.

**Objetivos múltiples.** Si existen varios objetivos, cada objetivo se muestra en un gráfico separado y hay una lista desplegable de **Objetivos** que controla qué objetivos mostrar.

## Predicho por observado

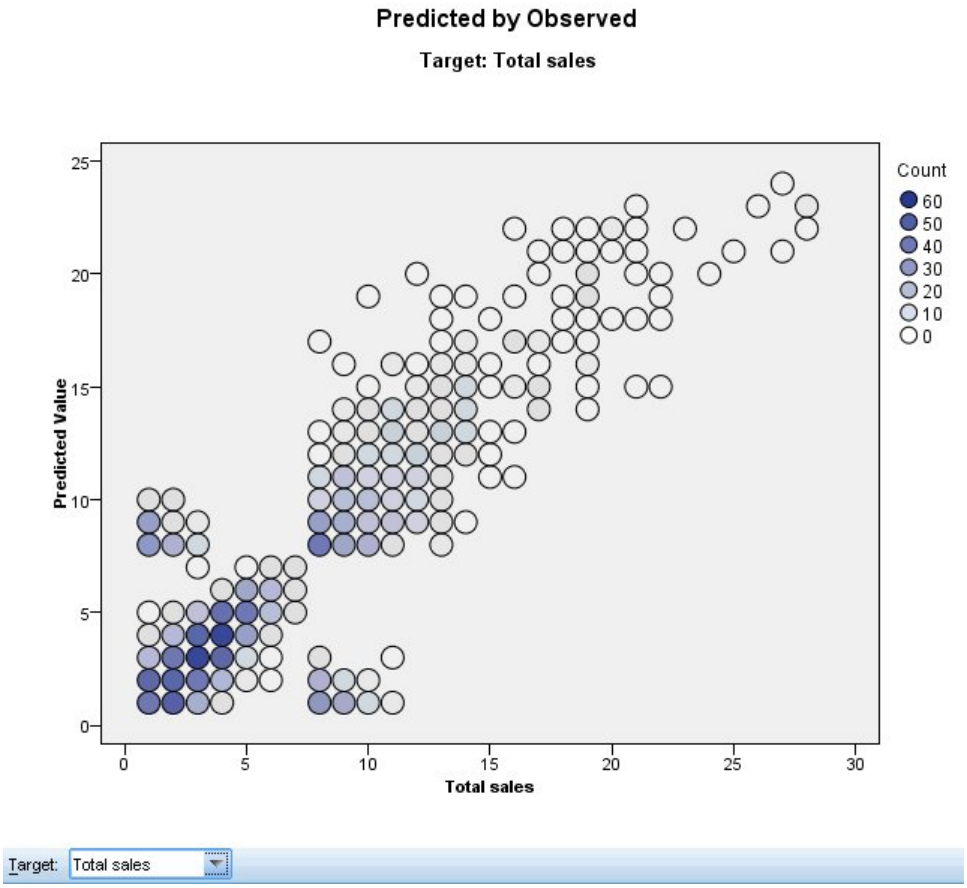



*Figura 40. Vista Predicho por observado*

Para objetivos continuos, muestra un diagrama de dispersión en intervalos de los valores predichos en el eje vertical por los valores observados en el eje horizontal.

**Objetivos múltiples.** Si existen varios objetivos continuos, cada objetivo se muestra en un gráfico separado y hay una lista desplegable de **Objetivos** que controla qué objetivos mostrar.

## Clasificación

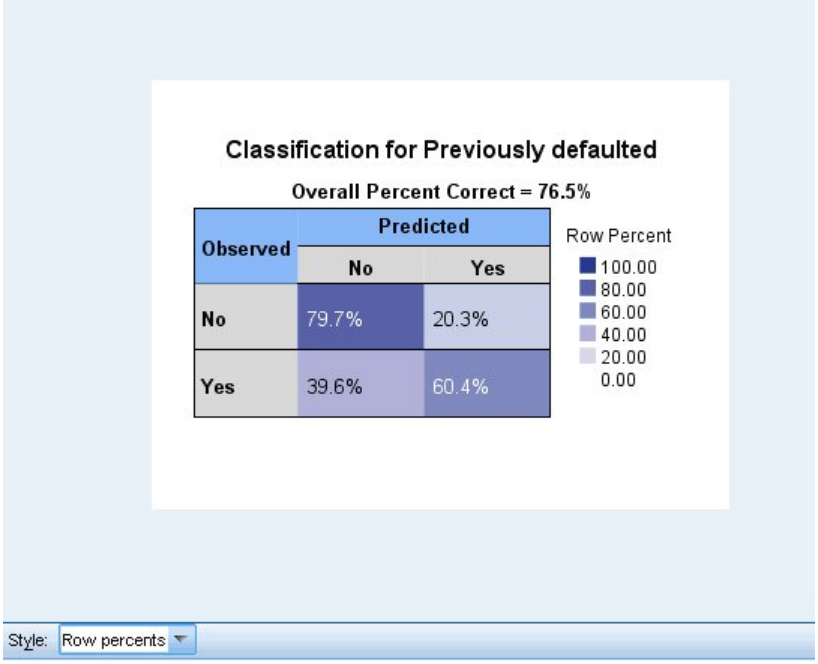



*Figura 41. Vista de clasificación, estilo de porcentajes en filas*

Para los objetivos categóricos, muestra la clasificación cruzada de los valores observados en contraposición a los predichos en el mapa de calor, junto con el porcentaje global correcto.

**Estilos de tabla.** Existen varios estilos de visualización diferentes, que son accesibles desde la lista desplegable **Estilo**.

- **Porcentajes de fila.** Muestra los porcentajes de filas (la casilla cuenta lo expresado como un porcentaje de los totales de filas) en las casillas. Este es el método predeterminado.
- **Recuentos de casillas.** Muestra los recuentos de casillas en las casillas. El sombreado del mapa de calor se basa aún en los porcentajes de filas.
- **Mapa de calor.** No muestra valores en las casillas, solamente el sombreado.
- **Comprimido.** No muestra cabeceras de filas o columnas, ni valores en las casillas. Puede ser útil cuando el objetivo tiene muchas categorías.

**Perdidos.** Si cualquier registro tiene valores perdidos en el objetivo, se muestran en una fila **(Perdidos)** bajo todas las filas válidas. Los registros con valores perdidos no contribuyen al porcentaje global correcto.

**Objetivos múltiples.** Si existen varios objetivos categóricos, cada objetivo se muestra en una tabla separada y hay una lista desplegable de **Objetivos** que controla qué objetivos mostrar.

**Tablas grandes.** Si el objetivo mostrado tiene más de 100 categorías, no se mostrará ninguna tabla.

## Red

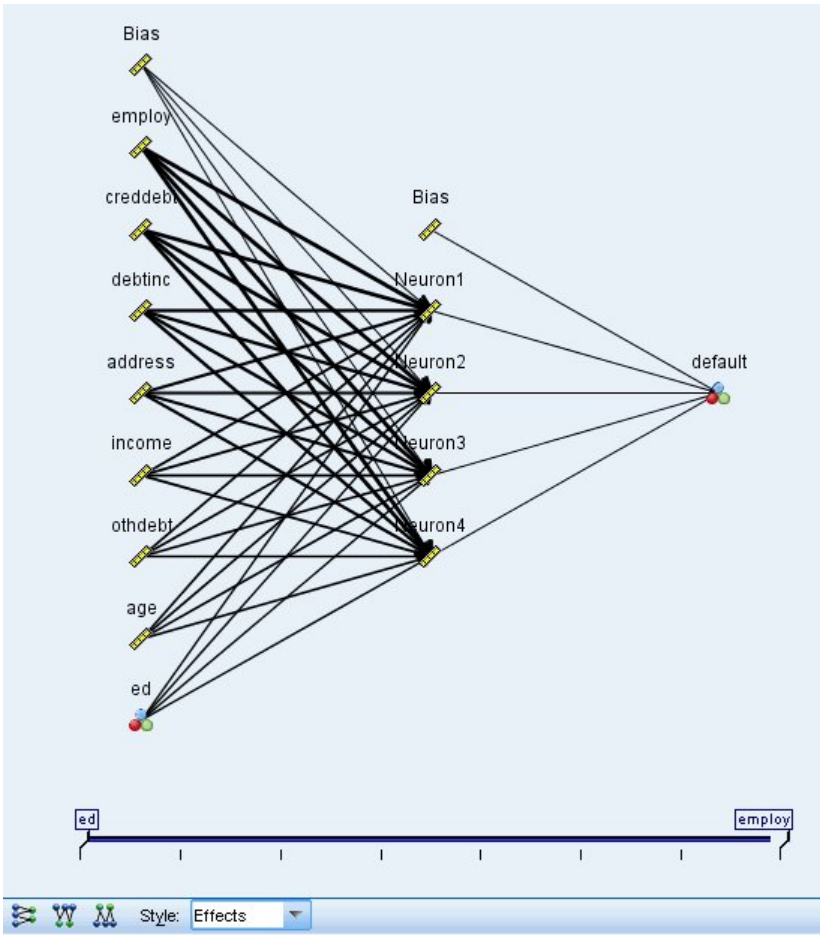



*Figura 42. Vista de red, entradas a la izquierda, estilo de efectos*

Muestra una representación gráfica de la red neuronal.

**Estilos de gráfico.** Existen dos estilos de visualización diferentes, que son accesibles desde la lista desplegable **Estilo**.

- **Efectos.** Muestra cada predictor y objetivo como un nodo en el diagrama sin importar si la escala de medición es continua o categórica. Este es el método predeterminado.
- **Coeficientes.** Muestra nodos indicadores múltiples para predictores y objetivos categóricos. Las líneas de conexión en el diagrama de estilo de coeficientes están coloreadas tomando como base el valor estimado de la ponderación sináptica.

**Orientación del diagrama.** De forma predeterminada, el diagrama de la red está dispuesto con las entradas a la izquierda y los objetivos a la derecha. Utilizando los controles de la barra de herramientas puede cambiar la orientación, de modo que las entradas estén en la parte superior y los objetivos en la parte inferior, o las entradas en la parte inferior y los objetivos en la parte superior.

**Importancia del predictor.** Las líneas de conexión del diagrama se ponderan tomando como base la importancia de predictores, con un grosor de línea mayor correspondiente a una importancia mayor. Existe un control deslizante Importancia del predictor en la barra de herramientas que controla qué predictores se muestran en el diagrama de red. Esto no cambia el modelo, simplemente le permite centrarse en los predictores más importantes.

**Objetivos múltiples.** Si hay objetivos múltiples, se muestran todos en el gráfico.

## Configuración

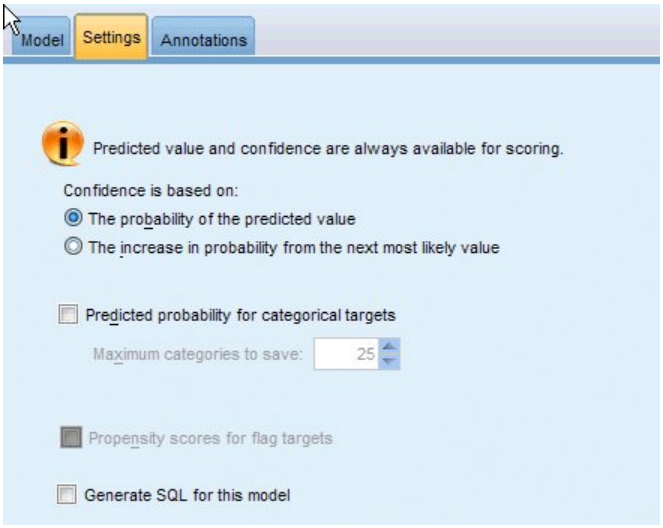



*Figura 43. Pestaña Configuración*

Cuando se puntúa el modelo, se crearán los elementos seleccionados en esta pestaña. El valor predicho (para todos los objetivos) y la confianza (para objetivos categóricos) se calculan siempre cuando se puntúa el modelo. La confianza calculada puede basarse en la probabilidad del valor predicho (la probabilidad predicha más alta) o la diferencia entre la probabilidad predicha más alta y la segunda probabilidad predicha más alta.

- **Probabilidad predicha para objetivos categóricos.** Genera las probabilidades predichas para objetivos categóricos. Se crea un campo para cada categoría.
- **Puntuaciones de propensión para objetivos de marca.** En el caso de modelos con un objetivo de marca (que devuelve una predicción de sí o no), puede solicitar puntuaciones de propensión que indican la probabilidad del resultado true especificado para el campo objetivo. El modelo produce puntuaciones de propensión en bruto; si hay particiones activas, el modelo también producirá puntuaciones de propensión ajustadas en función de la partición de prueba.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

**Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

**Puntuar convirtiendo a SQL nativo** Si selecciona esta opción, se genera SQL para puntuar el modelo dentro de la base de datos.

**Nota:** Aunque esta opción puede proporcionar resultados más rápidos, el tamaño y la complejidad del SQL nativo aumenta a medida que lo hace la complejidad del modelo.

**Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

# Capítulo 9. Lista de decisiones

Los modelos de Lista de decisiones identifican subgrupos o **segmentos** que muestran una mayor o menor posibilidad de proporcionar un resultado binario (sí o no) relacionado con la muestra global. Por ejemplo, puede buscar clientes con menos posibilidades de abandono o con más posibilidades de decir sí a una campaña u oferta determinada. El Visor de lista de decisiones proporciona control total sobre el modelo, ya que le permite editar segmentos, añadir sus propias reglas de negocio, especificar la forma de puntuación de cada segmento y personalizar el modelo de distintas maneras para optimizar la proporción de aciertos en todos los segmentos. Gracias a ello, se adapta especialmente bien a la generación de listas de mailing y a cualquier otro tipo de identificación de los registros a los que hay que dirigir una determinada campaña. También puede utilizar varias **tareas de minería** para combinar enfoques de modelado, por ejemplo identificando segmentos de alto y bajo rendimiento en el mismo modelo e incluyendo o excluyendo cada uno en la etapa de puntuación como corresponda.

Segmentos, reglas y condiciones

Un modelo consta de una lista de segmentos, cada uno de los cuales está definido por una regla que selecciona los registros coincidentes. Una regla determinada puede tener varias condiciones; por ejemplo:

```
RFM_SCORE > 10 y
MONTHS_CURRENT <= 9
```

Las reglas se aplican en el orden indicado; la primera regla coincidente determina el resultado de un registro dado. Si se toman de forma independiente, las reglas o condiciones se pueden solapar, pero el orden de las reglas resuelve la ambigüedad. Si ninguna regla coincide, el registro se asigna a la regla restante.

Control total sobre la puntuación

El Visor de lista de decisiones permite ver, modificar y reorganizar segmentos, así como seleccionar lo que se va a incluir o excluir para la puntuación. Por ejemplo, puede optar por excluir un grupo de clientes de futuras ofertas e incluir otros y ver inmediatamente cómo afecta a su tasa de aciertos global. Los modelos de Lista de decisiones devuelven una puntuación de *Sí* para los segmentos incluidos y *$null$* para todo lo demás, incluido el resto. Este control directo sobre la puntuación hace que los modelos Lista de decisiones sean ideales para generar listas de mailing, por lo que se utilizan con frecuencia en la gestión de relaciones con los clientes, incluidos los centros de llamadas y las aplicaciones de marketing.

Tareas de minería, medidas y selecciones

El proceso de modelado se lleva a cabo mediante las **tareas de minería**. Cada tarea de minería inicia eficazmente una nueva ejecución de modelado y devuelve un nuevo conjunto de modelos alternativos para escoger. La tarea predeterminada se basa en las especificaciones iniciales del nodo Lista de decisiones, pero puede definir cualquier número de tareas personalizadas. También puede aplicar tareas de forma iterativa; por ejemplo, puede ejecutar una búsqueda de alta probabilidad en todo el conjunto de entrenamiento y, a continuación, ejecutar una búsqueda de baja probabilidad en el resto para eliminar los segmentos de bajo rendimiento.

Selecciones de datos

Puede definir selecciones de datos y medidas de modelo personalizadas para generar y evaluar modelos. Por ejemplo, puede especificar una selección de datos en una tarea de minería para adaptar el modelo a una región determinada y crear una medida personalizada para evaluar cómo funciona ese modelo en

todo el país. Al contrario que las tareas de minería, las medidas no cambian el modelo subyacente, sino que proporcionan otra perspectiva para evaluar su funcionamiento.

Incorporación de conocimiento empresarial

Al ajustar y ampliar segmentos identificados por el algoritmo, el Visor de lista de decisiones permite incorporar su conocimiento empresarial al modelo. Puede editar los segmentos generados por el modelo o añadir otros segmentos según las reglas especificadas. A continuación, puede aplicar los cambios y previsualizar los resultados.

Para obtener una mejor comprensión, un enlace dinámico con Excel permite exportar datos a Excel, donde se pueden utilizar para crear gráficos para presentaciones y calcular medidas personalizadas, como medidas de beneficio completo y rendimiento de la inversión, que se pueden ver en el Visor de lista de decisiones mientras se genera el modelo.

**Ejemplo.** El departamento de marketing de una entidad financiera desea obtener resultados más rentables en las futuras campañas adaptando la oferta adecuada a cada cliente. Puede utilizar un modelo de lista de decisiones para identificar las características de los clientes que es más probable que respondan favorablemente teniendo en cuenta las promociones anteriores y generar una lista de mailing a partir de estos resultados.

**Requisitos.** Un único campo objetivo categórico con un nivel de medición del tipo *Marca* o *Nominal* que indica el resultado binario que desea predecir (sí/no) y al menos un campo de entrada. Cuando el tipo de campo objetivo es *Nominal*, deberá elegir manualmente un único valor para tratarlo como **acierto** o **respuesta**. Todos los demás valores se agruparán como **no acierto**. También se puede especificar un campo de frecuencia opcional. Los campos de fecha/hora continuos se ignorarán. El algoritmo agrupa automáticamente las entradas de rango numérico continuo según se haya especificado en la pestaña Experto del nodo de modelado. Para disponer de un mayor control sobre los intervalos, puede añadir un nodo Intervalos en un punto anterior de la ruta y utilizar el campo agrupado como entrada con un nivel de medición de *Ordinal*.

## Opciones del modelo de la lista de decisiones

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Crear modelos divididos.** Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como campos de división. Consulte el tema "Generación de modelos divididos" en la página 28 para obtener más información.

**Modo.** Especifica el método utilizado para generar el modelo.

- **Generar modelo.** Genera automáticamente un modelo en la paleta de modelos al ejecutar el nodo. El modelo resultante se puede añadir a las rutas para obtener puntuaciones, pero no se puede seguir editando.
- **Iniciar sesión interactiva.** Abre la ventana de modelado (salida) interactiva de Visor de lista de decisiones, que le permite elegir entre varias alternativas y aplicar repetidamente el algoritmo con diferentes configuraciones para hacer crecer o modificar progresivamente el modelo. Consulte el tema "Visor de lista de decisiones" en la página 159 para obtener más información.
- **Usar información de sesión interactiva guardada.** Inicia una sesión interactiva utilizando una configuración previamente guardada. La configuración interactiva se puede guardar desde el visor de

listas de Visor de lista de decisiones utilizando el menú Generar (para crear un modelo o un nodo de modelado) o el menú Archivo (para actualizar el nodo desde el que se inició la sesión).

**Valor objetivo.** Especifica el valor del campo objetivo que indica el resultado que desea modelar. Por ejemplo, si el abandono del campo objetivo se codifica `0 = no` y `1 = sí`, especifique `1` para identificar reglas que indiquen qué registros tienen más probabilidad de perderse.

**Buscar segmentos con.** Indica si la búsqueda de la variable objetivo debe buscar cada vez que aparezca una **alta probabilidad** o **baja probabilidad**. Encontrarlos y excluirlos puede ser una manera útil de mejorar el modelo, especialmente, cuando el resto tiene una baja probabilidad.

**Número máximo de segmentos.** Especifica el número máximo de segmentos que se van a devolver. Se crean los *N* segmentos superiores, donde el mejor segmento es el que tiene mayor probabilidad o, si más de un modelo tiene la misma probabilidad, la mayor cobertura. El valor mínimo permitido es 1, no hay ningún valor máximo.

**Tamaño mínimo del segmento.** Los dos parámetros inferiores dictan el tamaño mínimo del segmento. El mayor de los dos valores tiene preferencia. Por ejemplo, si el valor de porcentaje iguala un número mayor que el valor absoluto, el parámetro de porcentaje tiene preferencia.

- **Como porcentaje del segmento previo (%).** Especifica el tamaño mínimo de grupo como porcentaje de registros. El valor mínimo permitido es 0, el valor máximo permitido es 99,9.
- **Como valor absoluto (N).** Especifica el tamaño mínimo de grupo como número absoluto de registros. El valor mínimo permitido es 1, no hay ningún valor máximo.

**Reglas de segmentación.**

**Número máximo de atributos.** Especifica el número máximo de condiciones por regla de segmentación. El valor mínimo permitido es 1, no hay ningún valor máximo.

- **Permitir reutilización de atributos.** Cuando están activados, cada ciclo puede considerar todos los atributos, incluso aquellos utilizados en ciclos anteriores. Las condiciones para un segmento se crean en ciclos, donde cada ciclo añade una nueva condición. El número de ciclos se define utilizando el parámetro **Número máximo de atributos**.

**Intervalo de confianza para nuevas condiciones (%).** Especifica el nivel de confianza para comprobar la significación del segmento. Este parámetro juega un rol importante en el número de segmentos (si los hay) que se devuelven así como el número de condiciones por regla de segmentación. Cuanto mayor sea el valor, menor será el conjunto de resultados devueltos. El parámetro mínimo permitido es 50, el máximo es 99,9.

## Opciones de experto del nodo Lista de decisiones

Las opciones de experto permiten ajustar de manera precisa el proceso de generación de modelos.

**Método de intervalos.** Método utilizado para crear campos continuos de intervalos (igual frecuencia o igual amplitud).

**Número de intervalos.** Número de intervalos a crear para los campos continuos. El valor mínimo permitido es 2; no hay ningún valor máximo.

**Amplitud de búsqueda de modelo.** Número máximo de resultados de modelo por ciclo que se puede utilizar para el siguiente ciclo. El valor mínimo permitido es 1, no hay ningún valor máximo.

**Amplitud de búsqueda de reglas.** Número máximo de resultados de regla por ciclo que se pueden utilizar para el siguiente ciclo. El valor mínimo permitido es 1, no hay ningún valor máximo.

**Factor de fusión de intervalos.** Cantidad mínima que un segmento debe crecer cuando se funde con un segmento cercano. El parámetro mínimo permitido es 1,01, no hay parámetro máximo.

- **Permitir valores perdidos en condiciones.** `True` para permitir la prueba `IS MISSING` en las reglas.
- **Descartar resultados intermedios.** Cuando es `True`, sólo se devuelven los resultados finales del proceso de búsqueda. Un resultado final es un resultado que no se refina más en el proceso de búsqueda, Cuando es `False`, los resultados intermedios también se devuelven.

  **Número máximo de alternativas.** Especifica el número máximo de alternativas que se devolverán tras ejecutar la tarea de minería. El valor mínimo permitido es 1, no hay ningún valor máximo.

  Tenga en cuenta que la tarea de minería sólo devolverá el número real de alternativas, hasta el número máximo especificado. Por ejemplo, si el máximo está definido a 100 y sólo se encuentran 3 alternativas, únicamente se muestran 3.

## Nugget del modelo de la lista de decisiones

Un modelo consta de una lista de **segmentos**, cada uno de los cuales está definido por una **regla** que selecciona los registros coincidentes. Puede ver o modificar fácilmente los segmentos antes de generar el modelo y elegir los segmentos que quiere incluir o excluir. Cuando se utilizan para obtener puntuaciones, los modelos de listas de decisiones devuelven *Yes* para los segmentos incluidos y *$null$* para todo los demás, incluido el resto. Este control directo sobre la puntuación hace que los modelos de listas de decisiones sean ideales para generar listas de mailing, por lo que se utilizan con frecuencia en la gestión de relaciones con los clientes, incluidos los centros de llamadas y las aplicaciones de marketing.

Al ejecutar una ruta que contiene un modelo de lista de decisiones, el nodo añade tres nuevos campos que contienen la puntuación, que puede ser *1* (para indicar *Sí*) para los campos incluidos o *$null$* para los campos excluidos, la probabilidad (tasa de aciertos) del segmento al que corresponde el registro y el número de ID del segmento. Los nombres de los nuevos campos se derivan del nombre del campo de salida que se está predicho, con el prefijo *$D-* para la puntuación, *$DP-* para la probabilidad y *$DI-* para el ID del segmento.

El modelo se puntúa teniendo en cuenta el valor objetivo especificado cuando se generó el modelo. Se pueden excluir segmentos manualmente, de manera que obtengan la puntuación *$null$*. Por ejemplo, si ejecuta una búsqueda de baja probabilidad para buscar segmentos con valores menores que el promedio de las tasas de acierto, estos segmentos "bajos" recibirán la puntuación *Sí* a menos que los excluya manualmente. Si es necesario, puede recodificar los valores nulos como *No* utilizando un nodo Derivar o Rellenar.

PMML

Un modelo de lista de decisiones se puede almacenar como un modelo de conjunto de reglas PMML con un criterio de selección de "primer acierto". No obstante, se esperará que todas las reglas tengan la misma puntuación. Para permitir que se realicen cambios en el campo objetivo o en el valor objetivo, es posible almacenar varios modelos de conjuntos de reglas en un archivo para aplicarlos en orden, de manera que los casos que no coincidan con el primer modelo se pasen al segundo, etc. El nombre del algoritmo *DecisionList* se utiliza para indicar este comportamiento especial y únicamente los modelos de conjuntos de reglas con este nombre serán reconocidos como modelos de listas de decisiones y se puntuarán como tales.

## Configuración de nugget del modelo de la lista de decisiones

La pestaña Configuración de un nugget de modelo de listas de decisiones le permite obtener puntuaciones de propensión y activar o desactivar la optimización de SQL. Esta pestaña sólo está disponible después de haber añadido el nugget de modelo a una ruta.

**Calcular puntuaciones de propensión en bruto.** En el caso de modelos con un objetivo de marca (que devuelve una predicción de sí o no), puede solicitar puntuaciones de propensión que indican la

probabilidad del resultado true especificado para el campo objetivo. Éstas se añaden a otros valores de predicción y confianza que pueden generarse durante la puntuación.

**Calcular puntuaciones de propensión ajustada.** Las puntuaciones de propensión en bruto se basan sólo en los datos de entrenamiento y pueden ser demasiado optimistas debido a la tendencia de muchos modelos a sobreajustar estos datos. Las propensiones ajustadas intentan compensar evaluando el rendimiento del modelo frente a una partición de comprobación o validación. Esta opción requiere que se haya definido un campo de partición en la ruta y que se hayan activado puntuaciones ajustadas de propensión en el modo de modelado antes de generar el modelo.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar convirtiendo a SQL nativo** Si selecciona esta opción, se genera SQL para puntuar el modelo dentro de la base de datos.

  **Nota:** aunque esta opción puede proporcionar resultados más rápidos, el tamaño y la complejidad del SQL nativo aumenta a medida que lo hace la complejidad del modelo.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

---

## Visor de lista de decisiones

La interfaz gráfica de Visor de lista de decisiones, basada en tareas y de fácil uso, elimina la complejidad del proceso de generación de modelos, evitando que tenga que ocuparse de los detalles de bajo nivel de las técnicas de minería de datos y le permite dedicar toda su atención a las partes del análisis que requieren la intervención del usuario, como el establecimiento de objetivos, la selección de los grupos objetivo, el análisis de los resultados y la elección del modelo óptimo.

## Panel de modelo de trabajo

El panel de modelo de trabajo muestra el modelo actual, incluidas las tareas de minería y cualquier otra acción que se aplique al modelo de trabajo.

**ID.** Identifica el orden secuencial del segmento. Los segmentos del modelo se calculan, de manera secuencial, utilizando el número de ID.

**Reglas de segmentación.** Indica el nombre del segmento y las condiciones de segmento definidas. De forma predeterminada, el nombre del segmento es el nombre de campo o los nombres de campo concatenados utilizados en las condiciones, separados por comas.

**Puntuación.** Representa el campo que se desea predecir, cuyo valor se supone que está relacionado con los valores de otros campos (los predictores).

*Nota*: las siguientes opciones se pueden activar o desactivar en el cuadro de diálogo “Organización de las medidas del modelo” en la página 169.

**Cobertura.** El gráfico circular muestra visualmente la cobertura de cada segmento respecto a la cobertura total.

**Cobertura (n).** Muestra la cobertura de cada segmento respecto a la cobertura total.

**Frecuencia.** Muestra el número de aciertos recibidos respecto a la cobertura. Por ejemplo, si la cobertura es 79 y la frecuencia es 50, 50 de los 79 habrán respondido según el segmento seleccionado.

**Probabilidad.** Indica la probabilidad del segmento. Por ejemplo, si la cobertura es 79 y la frecuencia 50, la probabilidad del segmento será de 63,29% (50 dividido entre 79).

**Error.** Indica el error del segmento.

La información que aparece en la parte inferior del panel indica la cobertura, la frecuencia y la probabilidad del modelo entero.

Barra de herramientas del modelo de trabajo

El panel del modelo de trabajo ofrece las siguientes funciones mediante una barra de herramientas.

*Nota*: también es posible acceder a estas funciones pulsando con el botón derecho del ratón en un segmento del modelo.

*Tabla 9. Botones de la barra de herramientas del modelo de trabajo.*

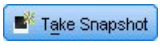
| **Botón de la barra de herramientas** | **Descripción** |
|---|---|
|  | Inicia el diálogo Generar nuevo modelo, que proporciona opciones para crear un nugget de modelo nuevo. |
|  | Guarda el estado actual de la sesión interactiva. El nodo de modelado de lista de decisiones se actualizará con la configuración que esté utilizando, incluidas tareas de minería, instantáneas de modelos, selecciones de datos y medidas personalizadas. Para restaurar una sesión a este estado, marque la casilla **Usar información de sesión guardada** en la pestaña Modelo del nodo de modelado y pulse en **Ejecutar**. |
|  | Muestra el cuadro de diálogo Organizar medidas del modelo. Consulte el tema "Organización de las medidas del modelo" en la página 169 para obtener más información. |
|  | Muestra el cuadro de diálogo Organizar selecciones de datos. Consulte el tema "Organización de selecciones de datos" en la página 165 para obtener más información. |
|  | Muestra la pestaña Instantáneas. Consulte el tema "Pestaña Instantáneas" en la página 161 para obtener más información. |
|  | Muestra la pestaña Alternativas. Consulte el tema "Pestaña Alternativas" en la página 161 para obtener más información. |
|   | Toma una instantánea de la estructura del modelo actual. Las instantáneas aparecen en la pestaña Instantáneas y suelen utilizarse para comparar modelos. |
|  | Inicia el diálogo Inserción de segmentos, que proporciona opciones para crear nuevos segmentos del modelo. |
|  | Inicia el diálogo Edición de reglas de segmentación que proporciona opciones para añadir condiciones o cmbiar condiciones de segmento de modelo definidas anteriormente. |

*Tabla 9. Botones de la barra de herramientas del modelo de trabajo (continuación).*

| | |
|---|---|
|  | Sube el segmento seleccionado en la jerarquía del modelo. |
|  | Baja el segmento seleccionado en la jerarquía del modelo. |
|  | Elimina el segmento seleccionado. |
|  | Incluye o excluye el segmento seleccionado del modelo. Si se excluye, los resultados del segmento se añadirán al resto. Esta opción se diferencia de la eliminación de un segmento en que es posible reactivar el segmento en otro momento. |

## Pestaña Alternativas

Se genera si pulsa en **Buscar segmentos**, la pestaña Alternativas muestra todos los resultados de minería alternativos del modelo seleccionado o del segmento en el panel de modelo de trabajo.

Para promocionar una alternativa que sea el modelo de trabajo, resalte la alternativa necesaria y pulse en **Cargar**; el modelo alternativo se muestra en el panel del modelo de trabajo.

*Nota*: la pestaña Alternativas sólo se muestra si ha definido **Número máximo de alternativas** en el nodo de modelado Lista de decisiones de la pestaña Experto para crear más de una alternativa.

Cada alternativa de modelo generada muestra información sobre el modelo específico:

**Nombre.** Cada alternativa está numerada secuencialmente. La primera alternativa suele contener los mejores resultados.

**Objetivo.** Indica el valor objetivo. Por ejemplo: 1, que es igual a "true".

**Número de segmentos.** El número de reglas de segmentos que se utilizan en el modelo alternativo.

**Cobertura.** La cobertura del modelo alternativo.

**Frecuencia.** Muestra el número de aciertos respecto a la cobertura.

**P.** Indica es el porcentaje de probabilidad del modelo alternativo.

*Nota*: los resultados alternativos no se guardan con el modelo, sino que sólo son válidos durante la sesión activa.

## Pestaña Instantáneas

Una instantánea es una vista de un modelo en un momento determinado. Por ejemplo, puede tomar una instantánea de modelo cuando desee cargar otro modelo alternativo en el panel Modelo de trabajo, pero no quiera perder el trabajo realizado con el modelo actual. La pestaña Instantáneas muestra todas las instantáneas de modelos tomadas manualmente para todos los estados de modelo de trabajo que se deseen.

*Nota*: las instantáneas se guardan con el modelo. Es recomendable que tome una instantánea cuando cargue el primer modelo. Esta instantánea conservará la estructura original del modelo, lo que le permite

volver en cualquier momento al estado original del modelo. El nombre de la instantánea generada se muestra como la marca de tiempo, lo que indica el momento en que se generó.

Creación de una instantánea del modelo

1. Seleccione el modelo o la alternativa que desea que aparezca en el panel Modelo de trabajo.
2. Realice todos los cambios necesarios al modelo de trabajo.
3. Pulse en **Tomar instantánea**. Aparecerá una nueva instantánea en la pestaña Instantáneas.

   **Nombre.** Nombre de la instantánea. Puede cambiar el nombre de una instantánea pulsando dos veces en el nombre de la instantánea.

   **Objetivo.** Indica el valor objetivo. Por ejemplo: 1, que es igual a "true".

   **Número de segmentos.** El número de reglas de segmentos que se utilizan en el modelo.

   **Cobertura.** La cobertura del modelo.

   **Frecuencia.** Muestra el número de aciertos respecto a la cobertura.

   **P.** Indica es el porcentaje de probabilidad del modelo.
4. Para promocionar una instantánea que sea el modelo de trabajo, resalte la instantánea necesaria y pulse en **Cargar**; la instantánea alternativa se muestra en el panel del modelo de trabajo.
5. Puede eliminar una instantánea pulsando en **Eliminar** o pulsando con el botón derecho del ratón en la instantánea y seleccionando **Eliminar** en el menú.

# Utilización de Visor de lista de decisiones

La generación de un modelo que prediga de manera óptima la respuesta y el comportamiento de los clientes se realiza en varias fases. Al iniciar Visor de lista de decisiones, el modelo de trabajo se rellena con los segmentos y medidas del modelo definido y el usuario podrá empezar una tarea de minería, modificar los segmentos o las medidas según sea necesario y generar un nuevo modelo o nodo de modelado.

Puede añadir una o más reglas de segmentación hasta que haya desarrollado un modelo satisfactorio. Puede añadir reglas de segmentación al modelo ejecutando tareas de minería o utilizando la función **Editar regla de segmentación**.

Durante el proceso de generación del modelo, puede evaluar el rendimiento del modelo validando el modelo respecto a los datos de medidas, visualizando el modelo en un diagrama o generando medidas de Excel personalizadas.

Cuando esté seguro de la calidad del modelo, puede generar un nuevo modelo y colocarlo en el lienzo de IBM SPSS Modeler o en la paleta de modelos.

## Tareas de minería

Una **tarea de minería** es una colección de parámetros que determina la manera en que se generan nuevas reglas. Algunos de estos parámetros se pueden seleccionar, lo que le ofrece la flexibilidad necesaria para adaptar modelos a nuevas situaciones. Una tarea consta de una plantilla de tarea (tipo), un objetivo y una selección de generación (conjunto de datos de minería).

Las siguientes secciones detallan las diferentes operaciones de tareas de minería:

- "Ejecución de tareas de minería"
- "Creación y edición de una tarea de minería" en la página 163
- "Organización de selecciones de datos" en la página 165

**Ejecución de tareas de minería:** Visor de lista de decisiones le permite añadir manualmente reglas de segmento a un modelo ejecutando tareas de minería o copiando y pegando reglas de segmento entre modelos. Una tarea de minería contiene información sobre cómo generar nuevas reglas (la configuración de los parámetros de minería de datos, como la estrategia de búsqueda, los atributos de origen, la

amplitud de búsqueda, el nivel de confianza, etc.), el comportamiento de los clientes que se desea predecir y los datos que se desea investigar. El objetivo de una tarea de minería es buscar las mejores reglas posibles de segmento.

Para generar un segmento de regla del modelo ejecutando una tarea de minería:

1. Pulse en la fila **Resto**. Si ya hay segmentos que aparecen en el panel de modelo de trabajo, también puede seleccionar uno de ellos para buscar reglas adicionales basadas en el segmento seleccionado. Tras seleccionar el resto o un segmento, utilice uno de los siguientes métodos para generar el modelo, o los modelos alternativos.

- En el menú Herramientas, seleccione **Buscar segmentos**.
- Pulse con el botón derecho del ratón en la fila/segmento **Resto** y elija **Buscar segmentos**.
- Pulse en **Buscar segmentos** en el panel de modelo de trabajo.

Mientras se está procesando la tarea, el progreso aparece en la parte inferior del espacio de trabajo donde también se le informará cuando termine la tarea. El tiempo exacto que tarda una tarea en terminarse depende de la complejidad de la tarea de minería y del tamaño del conjunto de datos. Si sólo hay un modelo en los resultados que se muestran en el panel de modelo de trabajo en cuanto se completa la tarea; sin embargo, si los resultados contienen más de un modelo que se muestran en la pestaña Alternativas.

*Nota*: el resultado de una tarea puede ser terminada con modelos, terminada sin modelos o error.

El proceso de búsqueda de nuevas reglas del segmento se puede repetir hasta que no se añada ninguna regla nueva al modelo, lo que indicará que se han encontrado todos los grupos de clientes significativos.

Es posible ejecutar una tarea de minería en cualquier segmento del modelo existente. Si el resultado de una tarea no es el que busca, puede optar por iniciar otra tarea de minería en el mismo segmento, lo que le proporcionará reglas encontradas adicionales basadas en el segmento seleccionado. Los segmentos que se encuentran "por debajo" del segmento seleccionado (es decir, que se han añadido al modelo posteriormente al segmento seleccionado) serán sustituidos por los nuevos segmentos, ya que cada segmento depende de sus predecesores.

**Creación y edición de una tarea de minería:** Una tarea de minería es el mecanismo que busca la colección de reglas que constituyen un modelo de datos. Junto a los criterios de búsqueda definidos en la plantilla seleccionada, una tarea también define el objetivo (la pregunta real que ha motivado el análisis, como cuántos clientes es posible que respondan a un mailing) e identifica los conjuntos de datos que se utilizarán. El objetivo de una tarea de minería es buscar los mejores modelos posibles.

Crear una tarea de minería

Para crear una tarea de minería:

1. Seleccione el segmento a partir del que desea buscar condiciones de segmento adicionales.
2. Pulse en **Configuración**. Aparecerá el cuadro de diálogo Crear/editar tarea de minería. El cuadro de diálogo ofrece las opciones para definir la tarea de minería.
3. Realice todos los cambios necesarios y pulse en **Aceptar** para volver al cuadro de diálogo Organizar tareas de minería. Visor de lista de decisiones utiliza la configuración como ajustes predeterminados para ejecutar todas las tareas hasta que se selecciona una tarea o configuración alternativa.
4. Pulse en **Buscar segmentos** para iniciar la tarea de minería en el segmento seleccionado.

Editar una tarea de minería

El cuadro de diálogo Crear/editar tarea de minería incluye opciones que permiten definir una nueva tarea de minería o editar una existente.

La mayoría de los parámetros disponibles para las tareas de minería son similares a los que aparecen en el nodo Lista de decisiones. Las excepciones se muestran a continuación. Consulte el tema "Opciones del modelo de la lista de decisiones" en la página 156 para obtener más información.

**Configuración de carga:** Cuando haya creado más de una tarea de minería, seleccione la tarea requerida.

**Nuevo...** Pulse para crear una nueva tarea de minería en función de los ajustes de la tarea que se muestra.

Objetivo

**Campo Objetivo:** Representa el campo que se desea predecir, cuyo valor se supone que está relacionado con los valores de otros campos (los predictores)

**Valor objetivo.** Especifica el valor del campo objetivo que indica el resultado que desea modelar. Por ejemplo, si el abandono del campo objetivo se codifica `0 = no` y `1 = sí`, especifique `1` para identificar reglas que indiquen qué registros tienen más probabilidad de perderse.

Configuración simple

**Número máximo de alternativas.** Especifica el número de alternativas que aparecerán tras ejecutar la tarea de minería. El valor mínimo permitido es 1, no hay ningún valor máximo.

Configuración de experto

**Editar...** Abre el diálogo **Editar parámetros avanzados** que le permite definir los valores avanzados. Consulte el tema "Editar parámetros avanzados" para obtener más información.

Datos

**Selección de generación.** Incluye opciones que permiten especificar la medida de evaluación que Visor de lista de decisiones analizará para buscar nuevas reglas. Las medidas de evaluación de la lista se crean y editan en el cuadro de diálogo Organizar selecciones de datos.

**Campos disponibles.** Incluye opciones que permiten mostrar todos los campos o seleccionar manualmente los campos que se mostrarán.

**Editar...** Si se selecciona la opción **Personalizado**, se abre el diálogo **Personalizar campos disponibles** que le permite seleccionar qué campos están disponibles como atributos de segmento encontrados por la tarea de minería. Consulte el tema "Personalizar campos disponibles" en la página 165 para obtener más información.

*Editar parámetros avanzados:* El cuadro de diálogo Editar parámetros avanzados ofrece las siguientes opciones de configuración.

**Método de intervalos.** Método utilizado para crear campos continuos de intervalos (igual frecuencia o igual amplitud).

**Número de intervalos.** Número de intervalos a crear para los campos continuos. El valor mínimo permitido es 2; no hay ningún valor máximo.

**Amplitud de búsqueda de modelo.** Número máximo de resultados de modelo por ciclo que se puede utilizar para el siguiente ciclo. El valor mínimo permitido es 1, no hay ningún valor máximo.

**Amplitud de búsqueda de reglas.** Número máximo de resultados de regla por ciclo que se pueden utilizar para el siguiente ciclo. El valor mínimo permitido es 1, no hay ningún valor máximo.

**Factor de fusión de intervalos.** Cantidad mínima que un segmento debe crecer cuando se funde con un segmento cercano. El parámetro mínimo permitido es 1,01, no hay parámetro máximo.

- **Permitir valores perdidos en condiciones.** `True` para permitir la prueba `IS MISSING` en las reglas.
- **Descartar resultados intermedios.** Cuando es `True`, sólo se devuelven los resultados finales del proceso de búsqueda. Un resultado final es un resultado que no se refina más en el proceso de búsqueda, Cuando es `False`, los resultados intermedios también se devuelven.

*Personalizar campos disponibles:* El diálogo Personalizar campos disponibles le permite seleccionar los campos que estarán disponibles como atributos de segmentos encontrados por la tarea de minería.

**Disponible.** Muestra los campos que están disponibles en este momento como atributos de segmentos. Para eliminar campos de la lista, seleccione los campos pertinentes y pulse en **Quitar >>**. Los campos seleccionados pasarán de la lista Disponibles a la lista No disponibles.

**No disponible.** Muestra los campos que no están disponibles en este momento como atributos de segmentos. Para incluir los campos en la lista disponibles, seleccione los campos pertinentes y pulse en **<< Añadir**. Los campos seleccionados pasarán de la lista No disponibles a la lista Disponibles.

**Organización de selecciones de datos:** Mediante la organización de las selecciones de datos (el conjunto de datos de minería), puede especificar las medidas de evaluación que Visor de lista de decisiones debe analizar para buscar nuevas reglas y elegir las selecciones de datos que se utilizarán como base de las medidas.

Para organizar las selecciones de datos:

1. En el menú Herramientas, elija **Organizar selecciones de datos** o pulse con el botón derecho del ratón en un segmento y seleccione la opción. Aparecerá el cuadro de diálogo Organizar selecciones de datos.

   *Note:* el cuadro de diálogo Organizar selecciones de datos también le permite editar o eliminar selecciones de datos existentes.
2. Pulse en el botón **Añadir nueva selección de datos**. Se añadirá una nueva entrada de selección de datos a la tabla existente.
3. Pulse en **Nombre** y escriba el nombre de la selección.
4. Pulse en **Partición** y seleccione el tipo de partición.
5. Pulse en **Condición** y seleccione la opción de condición. Cuando se selecciona **Especificar**, aparece el cuadro de diálogo Especificar condición de selección, que incluye opciones que permiten definir condiciones de campos específicas.
6. Defina la condición adecuada y pulse en **Aceptar**.

Las selecciones de datos están disponibles en la lista desplegable Selección de generación en el cuadro de diálogo Crear/editar tarea de minería. La lista le permite seleccionar la medida de evaluación que se utilizará para una determinada tarea de minería.

## Reglas de segmentación

Puede buscar las reglas de segmentación del modelo ejecutando tareas de minería basadas en plantillas de tareas. Puede añadir manualmente reglas de segmentación a un modelo mediante las funciones Insertar segmento o Editar regla de segmentación.

Si opta por buscar nuevas reglas de segmentación, los resultados, si los hay, aparecerán en la pestaña Visor del cuadro de diálogo Lista interactiva. Puede refinar rápidamente su modelo seleccionando uno de los resultados alternativos desde el cuadro de diálogo Álbumes de modelos y pulsando en **Cargar**. De esta manera, puede experimentar con diferentes resultados hasta que esté listo para generar un modelo que describa con precisión el grupo objetivo óptimo.

**Inserción de segmentos:** Puede añadir manualmente reglas de segmentación a un modelo mediante la función Segmentar.

Para añadir una condición de regla de segmentación a un modelo:

1. En el cuadro de diálogo Lista interactiva, seleccione una ubicación donde desee añadir un nuevo segmento. El nuevo segmento se insertará directamente sobre el segmento seleccionado.
2. En el menú Edición, elija **Insertar segmento** o acceda a esta selección pulsando con el botón derecho del ratón en un segmento.

   Se abrirá el cuadro de diálogo Insertar segmento, permitiéndole insertar nuevas condiciones de regla de segmentación.
3. Pulse en **Insertar**. El cuadro de diálogo Insertar condición se abrirá, permitiéndole definir los atributos para la nueva condición de regla.
4. Seleccione un campo y un operador en las listas desplegables.

   *Nota*: Si selecciona el operador **No en**, la condición seleccionada actuará como condición de exclusión y aparecerá en rojo en el cuadro de diálogo Insertar regla. Por ejemplo, cuando la condición `región = 'CIUDAD'` aparece en rojo, indica que `CIUDAD` se excluye del conjunto de resultados.
5. Introduzca uno o más valores o pulse en el icono **Insertar valor** para acceder al cuadro de diálogo Insertar valor. El diálogo permite elegir un valor definido para el campo seleccionado. Por ejemplo, el campo **casado** ofrecerá los valores **sí** y **no**.
6. Pulse en **Aceptar** para volver al cuadro de diálogo Insertar segmento. Pulse en **Aceptar** de nuevo para añadir el segmento creado al modelo.

El nuevo segmento aparecerá en la ubicación de modelo especificada.

**Edición de reglas de segmentación:** La funcionalidad Editar regla de segmentación permite añadir, cambiar o eliminar condiciones de regla de segmentación.

Para cambiar una condición de regla de segmentación:

1. Seleccione el segmento del modelo que desea editar.
2. En el menú Edición, elija **Editar regla de segmentación** o pulse con el botón derecho del ratón en la regla para acceder a esta selección.

   Aparecerá el cuadro de diálogo Editar regla de segmentación.
3. Seleccione la condición adecuada y pulse en **Editar**.

   El cuadro de diálogo Editar condición se abrirá, permitiéndole definir los atributos para la condición de regla seleccionada.
4. Seleccione un campo y un operador en las listas desplegables.

   *Nota*: si selecciona el operador **No en**, la condición seleccionada actuará como condición de exclusión y aparecerá en rojo en el cuadro de diálogo Editar segmento. Por ejemplo, cuando la condición `región = 'CIUDAD'` aparece en rojo, indica que `CIUDAD` se excluye del conjunto de resultados.
5. Introduzca uno o más valores o pulse en el botón **Insertar valor** para acceder al cuadro de diálogo Insertar valor. El diálogo permite elegir un valor definido para el campo seleccionado. Por ejemplo, el campo **casado** ofrecerá los valores **sí** y **no**.
6. Pulse en **Aceptar** para volver al cuadro de diálogo Editar regla de segmentación. Pulse nuevamente en **Aceptar** para volver al modelo de trabajo.

El segmento seleccionado aparecerá con las condiciones de reglas actualizadas.

*Eliminación de condiciones de reglas de segmentación:* **Para eliminar una condición de regla de segmentación:**

1. Seleccione el segmento del modelo que contiene las condiciones de reglas que desea eliminar.

2. En el menú Edición, elija **Editar regla de segmentación** o pulse con el botón derecho del ratón en el segmento para acceder a esta selección.

   Aparecerá el diálogo Editar regla de segmentación, que le permite eliminar una o más condiciones de reglas de segmentación.
3. Seleccione la condición de regla adecuada y pulse en **Eliminar**.
4. Pulse en **Aceptar**.

Al eliminar una o más condiciones de reglas de segmentación, se actualizan las métricas de medidas en el panel del modelo de trabajo.

**Copia de segmentos:** Visor de lista de decisiones ofrece una cómoda manera de copiar los segmentos del modelo. Cuando quiera aplicar un segmento de un modelo a otro modelo, sólo tendrá que copiar (o cortar) el segmento en un modelo y pegarlo en otro modelo. También puede copiar un segmento de un modelo que aparezca en el panel Presentación preliminar de alternativa y pegarlo en el modelo que aparece en el panel Modelo de trabajo. Las funciones que permiten cortar, copiar y pegar utilizan un portapapeles del sistema para almacenar o recuperar los datos temporales. Es decir, las condiciones y el objetivo se copian en el portapapeles. El contenido del portapapeles no está reservado exclusivamente para su uso en Visor de lista de decisiones, sino que también se puede pegar en otras aplicaciones. Cuando, por ejemplo, el contenido del portapapeles se pega en un editor de texto, las condiciones y el objetivo se pegan en formato XML.

Para copiar o cortar los segmentos del modelo:
1. Seleccione el segmento del modelo que desea utilizar en otro modelo.
2. En el menú Edición, seleccione **Copiar** (o **Cortar**) o pulse con el botón derecho del ratón en el segmento del modelo y seleccione **Copiar** o **Cortar**.
3. Abra el modelo en el que desea pegar el segmento del modelo.
4. Seleccione uno de los segmentos del modelo y pulse en **Pegar**.

*Nota*: En lugar de los mandatos **Cortar**, **Copiar** y **Pegar** también puede utilizar las combinaciones de teclas : **Ctrl+X**, **Ctrl+C**y **Ctrl+V**.

El segmento copiado (o cortado) se insertará encima del segmento del modelo anteriormente seleccionado. Se volverán a calcular las medidas del segmento pegado y de los segmentos inferiores.

*Nota*: en este procedimiento, ambos modelos deben basarse en la misma plantilla de modelo de escenario de y deben contener el mismo objetivo. En otro caso, aparecerá un mensaje de error.

**Modelos alternativos:** Si hay más de un resultado, la pestaña Alternativas muestra los resultados de cada tarea de minería. Cada resultado consta de las condiciones de los datos seleccionados que más se adaptan al objetivo, así como todas las alternativas "suficientemente buenas". El número total de alternativas mostradas depende de los criterios de búsqueda que se hayan utilizado durante el proceso de análisis.

Para ver los modelos alternativos:
1. Pulse en un modelo alternativo en la pestaña Alternativas. Aparecerán los segmentos del modelo alternativos y sustituirán a los segmentos del modelo actual en el panel Presentación preliminar de alternativa.
2. Para trabajar con un modelo alternativo en el panel del modelo de trabajo, pulse en **Cargar** en el panel Presentación preliminar de alternativa o pulse con el botón derecho en el nombre de la alternativa en la pestaña Alternativas y seleccione **Cargar**.

*Nota*: los modelos alternativos no se guardan al generar un nuevo modelo.

## Personalización de un modelo

Los datos no son estáticos. Los clientes cambian de residencia, se casan o cambian de trabajo. Los productos dejan de estar enfocados al mercado y se quedan obsoletos.

Visor de lista de decisiones ofrece a los usuarios empresariales la flexibilidad necesaria para adaptar los modelos a nuevas situaciones de manera rápida y sencilla. Puede cambiar un modelo editando, eliminando, desactivando o cambiando la prioridad de determinados segmentos del modelo.

**Asignación de prioridades a los segmentos:** Puede clasificar las reglas del modelo en el orden que desee. De forma predeterminada, los segmentos del modelo aparecen por orden de prioridad, siendo el primer segmento el que tiene la mayor prioridad. Cuando se asigna unaprioridad diferente a uno o más de los segmentos, se cambiará el modelo de la manera correspondiente. Puede modificar el modelo de la manera necesaria moviendo los segmentos a una posición con unaprioridad mayor o menor.

Para asignar unaprioridad a los segmentos del modelo:

1. Seleccione el segmento del modelo al que desea asignar unaprioridad diferente.
2. Pulse en uno de los dos botones de flecha de la barra de herramientas del panel del modelo de trabajo para subir o bajar en la lista el segmento del modelo seleccionado.

Tras asignar la prioridad, se volverán a calcular todos los resultados de evaluación anteriores y se mostrarán los nuevos valores.

**Eliminación de segmentos:** Para eliminar uno o más segmentos:

1. Seleccione un segmento del modelo.
2. En el menú Edición, seleccione **Eliminar segmento** o pulse en el botón Eliminar de la barra de herramientas del panel del modelo de trabajo.

Se volverán a calcular las medidas del modelo modificado y se cambiará el modelo de la manera correspondiente.

**Exclusión de segmentos:** Mientras se buscan grupos concretos, es posible que se basen acciones empresariales en una selección de segmentos del modelo. Al desplegar un modelo, es posible que desee excluir determinados segmentos de un modelo. Los segmentos excluidos se puntúan con valores nulos. La exclusión de un segmento no implica que no se utilice el segmento, sino que todos los registros que cumplan dicha regla se excluirán de la lista de mailing. La regla sigue aplicándose, pero de manera diferente.

Para excluir determinados segmentos del modelo:

1. Seleccione un segmento en el panel del modelo de trabajo.
2. Pulse en el botón **Conmutar exclusión de segmentos** de la barra de herramientas del panel del modelo de trabajo. Aparecerá **Excluido** en la columna Objetivo seleccionada del segmento elegido.

*Nota*: a diferencia de los segmentos eliminados, sigue siendo posible reutilizar los segmentos excluidos en el modelo final. Los segmentos excluidos afectan a los resultados de los diagramas.

**Cambiar valor objetivo:** El diálogo Cambiar valor objetivo le permite cambiar el valor objetivo del campo objetivo actual.

Las instantáneas y los resultados de la sesión con un valor objetivo diferente del modelo de trabajo se pueden identificar cambiando el color de fondo de dicha fila de la tabla a amarillo, lo que indica que dicha instantánea/resultado de la sesión es obsoleto.

El cuadro de diálogo **Crear/Editar tarea de minería** muestra el valor objetivo del modelo de trabajo actual. El valor objetivo no se guarda con la tarea de minería, sino que en su lugar se toma del valor del modelo de trabajo.

Cuando se asciende a modelo de trabajo un modelo guardado que tiene un valor objetivo diferente del modelo de trabajo actual (por ejemplo, si se edita un resultado de la alternativa o se edita una copia de una instantánea), el valor objetivo del modelo guardado se cambiará para que coincida con el del modelo de trabajo (por tanto, no se cambiará el valor objetivo que aparece en el panel Modelo de trabajo). ⌂Las métricas del modelo se volverán a evaluar con el nuevo objetivo.

## Generar nuevo modelo

El cuadro de diálogo Generar nuevo modelo incluye opciones que permiten asignar un nombre al modelo y seleccionar dónde se creará el nuevo nodo.

**Nombre del modelo.** Seleccione **Personalizado** para ajustar el nombre generado automáticamente o crear un nombre exclusivo para el nodo como se muestra en el lienzo de rutas.

**Crear nodo en.** Si se selecciona **Lienzo** se colocará el nuevo modelo en el lienzo de trabajo; si se selecciona **Paleta de modelos generados** se colocará el nuevo modelo en la paleta de modelos; seleccionando **Ambos** se colocará el nuevo modelo tanto en el lienzo de trabajo como en la paleta de modelos.

**Incluir estado de sesión interactiva.** Cuando se activa, el estado de la sesión interactiva se conserva en el modelo generado. Cuando posteriormente se genera un nodo de modelado a partir del modelo, se transfiere el estado y se utiliza para inicializar la sesión interactiva. Independientemente de si selecciona esta opción, el modelo puntuará los nuevos datos de manera idéntica. Cuando no se selecciona esta opción, el modelo seguirá pudiendo crear un nodo de generación, pero será un nodo de generación más genérico que iniciará una nueva sesión interactiva en vez que continuar a partir del punto en el que se abandonó la sesión anterior. Si cambia la configuración del nodo pero se ejecuta con un estado guardado, se ignorará la configuración que se ha cambiado y se utilizará la configuración del estado guardado.

*Nota*: las métricas estándar son las únicas métricas que permanecen con el modelo. Las métricas adicionales se conservan con el estado interactivo. El modelo generado no representa el estado guardado de la tarea de minería interactiva. Tras iniciar Visor de lista de decisiones, aparecerá la configuración realizada en un principio utilizando el Visor.

Consulte el tema “Regeneración de un nodo de modelado” en la página 50 para obtener más información.

## Evaluación de modelos

Para crear con éxito un modelo, es necesario evaluar cuidadosamente el modelo antes de implementarlo en el entorno de producción. Visor de lista de decisiones ofrece varias medidas estadísticas y empresariales que se pueden utilizar para evaluar el impacto de un modelo en el mundo real. Entre éstas se incluyen los gráficos de ganancias y la total interoperabilidad con Excel, lo que permite simular escenarios de coste/beneficio para evaluar el impacto del despliegue.

Puede evaluar el modelo de las siguientes formas:

- Utilizando las medidas estadísticas y empresariales predefinidas que ofrece Visor de lista de decisiones (probabilidad, frecuencia).
- Evaluando las medidas importadas de Microsoft Excel.
- Visualizando el modelo mediante un gráfico de ganancias.

**Organización de las medidas del modelo:** Visor de lista de decisiones incluye opciones que permiten definir las medidas que se calculan y muestran como columnas. Cada segmento puede incluir la cobertura predeterminada, la frecuencia, la probabilidad y las medidas de error representadas como columnas. También puede crear nuevas medidas que aparecerán como columnas.

Definición de las medidas del modelo

Para añadir una medida al modelo o definir una medida existente:

1. En el menú Herramientas, elija **Organizar medidas del modelo** o pulse con el botón derecho del ratón en el modelo para realizar esta selección. Aparecerá el cuadro de diálogo Organizar medidas del modelo.
2. Pulse en el botón **Añadir nueva medida de modelo** (a la derecha de la columna Mostrar). Se mostrará una nueva medida en la tabla.
3. Especifique el nombre de la medida y seleccione el tipo, opción de visualización y selección. La columna Mostrar indica si se mostrará la medida para el modelo del trabajo. Al definir una medida existente, seleccione una métrica y una selección adecuadas y especifique si se mostrará la medida para el modelo de trabajo.
4. Pulse en **Aceptar** para volver al espacio de trabajo de Visor de lista de decisiones. Si se activó la columna Mostrar para la nueva medida, aparecerá la nueva medida para el modelo de trabajo.

Métricas personalizadas en Excel

Consulte el tema "Evaluación en Excel" para obtener más información.

*Actualización de medidas:* En algunos casos, es posible que sea necesario volver a calcular las medidas del modelo, como cuando se aplica un modelo existente a un nuevo conjunto de clientes.

Para volver a calcular (actualizar) las medidas del modelo:

En el menú Edición, elija **Actualizar todas las medidas**.

*o*

Pulse F5.

Se volverán a calcular todas las medidas y se mostrarán los nuevos valores para el modelo de trabajo.

**Evaluación en Excel:** Visor de lista de decisiones puede integrarse con Microsoft Excel, lo que le permite utilizar sus propios cálculos de valores y fórmulas de beneficios directamente en el proceso de generación del modelo para simular escenarios de coste/beneficio. El enlace con Excel permite exportar datos a Excel, donde se pueden utilizar para crear gráficos de presentaciones, calcular medidas personalizadas, por ejemplo medidas de beneficio complejo y rendimiento de la inversión), y verlos en Visor de lista de decisiones mientras se genera el modelo.

*Nota*: para poder trabajar con una hoja de cálculo de Excel, el experto de análisis de CRM debe definir la información de configuración para la sincronización de Visor de lista de decisiones con Microsoft Excel. La configuración figura en un archivo de hoja de cálculo de Excel y especifica la información que se transfiere de Visor de lista de decisiones a Excel y viceversa.

Los siguientes pasos son válidos sólo si se ha instalado MS Excel. Si no se ha instalado Excel, no aparecerán las opciones que permiten sincronizar modelos con Excel.

Para sincronizar modelos con MS Excel:

1. Abra el modelo, ejecute una sesión interactiva y seleccione **Organizar medidas del modelo** desde el menú Herramientas.
2. Seleccione **Sí** para la opción **Calcular medidas personalizadas en Excel**. El campo **Libro** se activa, lo que permite seleccionar una plantilla de libro de trabajo de Excel preconfigurada.
3. Pulse en el botón **Conectar a Excel**. Se abre el diálogo Abrir, que permite navegar hasta la ubicación de plantilla preconfigurada en el sistema de archivos de red o local.

4. Seleccione la plantilla de Excel adecuada y pulse en **Abrir**. Se abrirá la plantilla de Excel seleccionada; utilice la barra de tareas de Windows (o pulse Alt-Tab) para volver al cuadro de diálogo Seleccionar entradas para medidas personalizadas.
5. Seleccione las correlaciones adecuadas entre los nombres de las métricas definidas en la plantilla de Excel y los nombres de las métricas del modelo y pulse en **Aceptar**.

Una vez establecido este enlace, se inicia Excel con la plantilla de Excel preconfigurada que muestra las reglas del modelo en la hoja de cálculo. Los resultados calculados en Excel se mostrarán como nuevas columnas en Visor de lista de decisiones.

*Nota*: las métricas de Excel no se guardan con el modelo, sino que sólo son válidas durante la sesión activa. No obstante, puede crear instantáneas que incluyan métricas de Excel. Las métricas de Excel guardadas en las vistas de instantáneas sólo son válidas para realizar comparaciones históricas y no se actualizan cuando se vuelven a abrir. Consulte el tema "Pestaña Instantáneas" en la página 161 para obtener más información. Las métricas de Excel no aparecerán en las instantáneas hasta que se vuelva a establecer una conexión con la plantilla de Excel.

*Configuración de integración de MS Excel:* La integración entre Visor de lista de decisiones y Microsoft Excel se realiza utilizando una plantilla de hoja de cálculo de Excel preconfigurada. Esta plantilla consta de tres hojas de trabajo:

**Medidas del modelo.** Muestra las medidas de Visor de lista de decisiones importadas, las medidas de Excel personalizadas y los totales de los cálculos (definidos en la hoja de trabajo de configuración).

**Configuración.** Proporciona las variables que generan los cálculos basados en las medidas de Visor de lista de decisiones importadas y las medidas de Excel personalizadas.

**Configuración.** Incluye opciones que permiten especificar las medidas que se importarán de Visor de lista de decisiones y definir las medidas de Excel personalizadas.

*AVISO*: La estructura de la hoja de trabajo Configuración está estrictamente definida. **NO** edite ninguna casilla en la zona verde sombreada.

- **Métricas del modelo.** Indica las métricas de Visor de lista de decisiones que se utilizarán en los cálculos.
- **Métricas al modelo.** Indica las métricas generadas en Excel que se devolverán a Visor de lista de decisiones. Las métricas generadas por Excel aparecen en Visor de lista de decisiones como nuevas columnas de medidas.

*Nota*: las métricas de Excel no se conservan con el modelo cuando se genera un nuevo modelo, sino que sólo son válidas durante la sesión activa.

*Cambio de medidas del modelo:* Los siguientes ejemplos explican cómo cambiar Medidas del modelo de varias formas:

- Cambiar una medida existente.
- Importar una medida estándar adicional desde el modelo.
- Exportar una medida personalizada adicional al modelo.

Cambiar una medida existente

1. Abra la plantilla y seleccione la hoja de trabajo Configuración.
2. Edite cualquier **Nombre** o **Descripción** resaltándolo e introduciendo encima el nuevo valor.

   Tenga en cuenta que si desea cambiar una medida, por ejemplo para solicitar al usuario la probabilidad en lugar de la frecuencia, sólo tendrá que cambiar el nombre y la descripción en **Métricas del modelo**; entonces esto se mostrará en el modelo y el usuario podrá seleccionar la medida apropiada para el mapa.

Importar una medida estándar adicional desde el modelo

1. Abra la plantilla y seleccione la hoja de trabajo Configuración.
2. Seleccione en los menús:

   **Herramientas** > **Protección** > **Hoja no protegida**
3. Seleccione la casilla A5, que se encuentra sombreada y contiene la palabra **End**.
4. Seleccione en los menús:

   **Insertar** > **Filas**
5. Introduzca el **Nombre** y **Descripción** de la nueva medida. Por ejemplo, **Error** se convierte **Error asociado con segmento**.
6. En la casilla C5, introduzca la fórmula **=COLUMN('Medidas del modelo'!N3)**.
7. En la casilla D5, introduzca la fórmula **=ROW('Medidas del modelo'!N3)+1**.

   Estas fórmulas harán que la nueva medida aparezca en la columna N de la hoja de trabajo Medidas del modelo, que actualmente está vacía.
8. Seleccione en los menús:

   **Herramientas** > **Protección** > **Hoja protegida**
9. Pulse en **Aceptar**.
10. En la hoja de trabajo Medidas del modelo, asegúrese de que la casilla N3 tiene **Error** como título de la nueva columna.
11. Seleccione toda la columna N.
12. Seleccione en los menús:

    **Formato** > **Casillas**
13. De forma predeterminada, todas las casillas tienen una categoría de número **General**. Pulse en **Porcentaje** para cambiar las cifras que se muestran. Le ayuda a comprobar sus cifras en Excel; además, le permite utilizar los datos para otros fines, por ejemplo, como resultado para un gráfico.
14. Pulse en **Aceptar**.
15. Guarde la hoja de cálculo como una plantilla de Excel 2003, con un nombre exclusivo y la extensión de archivo *.xlt*. Para localizar fácilmente la nueva plantilla, recomendamos que la guarde en la ubicación de la plantilla preconfigurada en el sistema de archivos de red o local.

Exportación de una medida personalizada adicional al modelo

1. Abra la plantilla a la que ha añadido la columna Error en el ejemplo anterior; seleccione la hoja de trabajo Configuración.
2. Seleccione en los menús:

   **Herramientas** > **Protección** > **Hoja no protegida**
3. Seleccione la casilla A14, que se encuentra sombreada y contiene la palabra **End**.
4. Seleccione en los menús:

   **Insertar** > **Filas**
5. Introduzca el **Nombre** y **Descripción** de la nueva medida. Por ejemplo, **Error escalado** y **Escala aplicada a un error de Excel**.
6. En la casilla C14, introduzca la fórmula **=COLUMN('Medidas del modelo'!O3)**.
7. En la casilla D14, introduzca la fórmula **=ROW('Medidas del modelo'!O3)+1**.

   Estas fórmulas especifican que la columna O proporcionará la nueva medida al modelo.
8. Seleccione la hoja de trabajo Parámetros.
9. En la casilla A17, introduzca la descripción **'- Error escalado**.
10. En la casilla B17, introduzca el factor de escala **10**.
11. En la hoja de trabajo Medidas del modelo, introduzca la descripción **Error escalado** en la casilla O3 como título de la nueva columna.

12. En la casilla O4, introduzca la fórmula **=N4*Settings!$B$17**.
13. Seleccione la esquina de la casilla O4 y arrástrela a la casilla O22 para copiar la fórmula en cada casilla.
14. Seleccione en los menús:

    **Herramientas** > **Protección** > **Hoja protegida**
15. Pulse en **Aceptar**.
16. Guarde la hoja de cálculo como una plantilla de Excel 2003, con un nombre exclusivo y la extensión de archivo *.xlt*. Para localizar fácilmente la nueva plantilla, recomendamos que la guarde en la ubicación de la plantilla preconfigurada en el sistema de archivos de red o local.

Cuando se conecte a Excel mediante esta plantilla, el valor Error estará disponible como una nueva medida personalizada.

## Visualización de modelos

La mejor manera de comprender el impacto de un modelo es visualizarlo. Mediante un gráfico de ganancias, puede lograr valiosos datos acerca de la evolución diaria de la empresa y aprovechar técnicamente el modelo estudiando el efecto de varias alternativas en tiempo real. La sección "Ganancias" muestra las ventajas de un modelo respecto a la toma aleatoria de decisiones y permite comparar directamente varios gráficos cuando hay modelos alternativos.

**Ganancias:** Los gráficos de ganancias representan los valores de la columna *% ganancia* en la tabla. Las ganancias se definen como la proporción de aciertos en cada uno de los incrementos en relación con el número total de aciertos en el árbol, y se obtienen mediante la ecuación:

(aciertos del incremento / número total de aciertos) x 100%

El gráfico de ganancias ilustra de manera eficaz la difusión necesaria para una red cuando se desea capturar un porcentaje determinado de todos los aciertos del árbol. La línea diagonal representa la respuesta esperada para la muestra completa, si no se utilizase el modelo. En este caso la tasa de respuesta debería ser constante, ya que una persona tiene la misma probabilidad de responder que otra. Para duplicar los resultados deberá preguntar dos veces al mismo número de personas. La línea curvada indica hasta qué punto se puede mejorar la respuesta incluyendo únicamente elementos situados en los percentiles superiores en función de las ganancias. Por ejemplo, si incluye el 50% superior, obtendrá más del 70% de respuestas positivas. Cuanto más pronunciada es la curva, mayor es la ganancia.

Para ver un gráfico de ganancias:

1. Abra una ruta que contenga un nodo Lista de decisiones e inicie una sesión interactiva desde dicho nodo.
2. Pulse en la pestaña **Ganancias**. Según las particiones que se hayan especificado, es posible que se muestren dos gráficos (por ejemplo, aparecerán dos gráficos si se ha definido una partición de entrenamiento y otra de prueba para las medidas del modelo).

De forma predeterminada, los gráficos aparecen como segmentos. Si desea que los gráficos aparezcan como cuantiles, seleccione **Cuantiles** y, a continuación, seleccione el método de cuantiles en el menú desplegable.

*Opciones del diagrama:* La característica Opciones del diagrama incluye opciones que permiten seleccionar los modelos y las instantáneas que se incluirán en el diagrama, las particiones que se representarán y si se mostrarán las etiquetas de los segmentos.

Modelos para representar

**Modelos actuales.** Permite seleccionar los modelos que desea representar. Puede seleccionar el modelo de trabajo o cualquier modelo de instantánea creado.

Particiones para representar

**Particiones para diagrama a la izquierda.** La lista desplegable incluye opciones que permiten mostrar todas las particiones definidas o todos los datos.

**Particiones para diagrama a la derecha.** La lista desplegable incluye opciones que permiten mostrar todas las particiones definidas, todos los datos o sólo el diagrama a la izquierda. Cuando se selecciona **Gráfico sólo izquierda**, sólo se muestra el diagrama izquierdo.

**Mostrar etiquetas de segmento.** Cuando esta opción está activada, se muestran todas las etiquetas de segmento en los diagramas.

# Capítulo 10. Modelos estadísticos

Los modelos estadísticos utilizan ecuaciones matemáticas para codificar información extraída de los datos. En algunos casos, las técnicas de modelado estadístico pueden proporcionar modelos adecuados de forma rápida. Incluso en el caso de problemas en los que las técnicas más flexibles de aprendizaje de las máquinas (como redes neuronales) pueden ofrecer a la postre mejores resultados, es posible usar algunos modelos estadísticos como modelos predictivos de línea base para juzgar el rendimiento de técnicas más avanzadas.

Están disponibles los siguientes nodos de modelado estadístico.

Los modelos de regresión lineal predicen un objetivo continuo tomando como base las relaciones lineales entre el destino y uno o más predictores.

La regresión logística es una técnica de estadístico para clasificar los registros en función los valores de los campos de entrada. Es análoga a la regresión lineal pero toma un campo objetivo categórico en lugar de uno numérico.

El nodo PCA/Factorial proporciona técnicas eficaces de reducción de datos para reducir la complejidad de los datos. Análisis de componentes principales (PCA) busca combinaciones lineales de los campos de entrada que realizan el mejor trabajo a la hora de capturar la varianza en todo el conjunto de campos, en el que los componentes son ortogonales (perpendiculares) entre ellos. Análisis factorial intenta identificar factores subyacentes que expliquen el patrón de correlaciones dentro de un conjunto de campos observados. Para los dos métodos, el objetivo es encontrar un número pequeño de campos derivados que resuma de forma eficaz la información del conjunto original de campos.

El análisis discriminante realiza más supuestos rigurosos que regresiones logísticas, pero puede ser una alternativa o un suplemento valioso al análisis de regresión logística si se cumplen dichos supuestos.

El modelo lineal generalizado amplía el modelo lineal general, de manera que la variable dependiente está relacionada linealmente con los factores y las covariables mediante una determinada función de enlace. Además, el modelo permite que la variable dependiente tenga una distribución no normal. Cubre la funcionalidad de un amplio número de modelo estadísticos, incluyendo regresión lineal, regresión logística, modelos log lineales para recuento de datos y modelos de supervivencia censurados por intervalos.

Un modelo lineal mixto generalizado (GLMM) amplía el modelo lineal de modo que el objetivo pueda tener una distribución no normal, esté linealmente relacionado con los factores y covariables mediante una función de enlace especificada y las observaciones se puedan correlacionar. Los modelos lineales mixtos generalizados cubren una amplia variedad de modelos, desde modelos de regresión lineal simple hasta modelos multinivel complejos para datos longitudinales no normales.

El nodo Regresión de Cox le permite crear un modelo de supervivencia para datos de tiempo hasta el evento en presencia de registros censurados. El modelo produce una función de supervivencia que predice la probabilidad de que el evento de interés se haya producido en el momento dado (*t*) para valores determinados de las variables de entrada.

# Nodo Lineal

La regresión lineal es una técnica de estadístico común para clasificar los registros en función los valores de los campos de entrada numérica. La regresión lineal se ajusta a una línea recta o una superficie que minimiza las discrepancias entre los valores de resultados predichos y reales.

**Requisitos.** Sólo se pueden utilizar campos numéricos en un modelo de regresión lineal. Debe tener exactamente un campo objetivo (con el rol definido a **Objetivo**) y uno o más predictores (con el rol definido a **Entrada**). Los campos con un rol **Ambos** o **Ninguno** se ignoran, ya que no son campos numéricos. (Si es necesario, los campos no numéricos se pueden recodificar mediante un nodo Derivar.)

**Puntos fuertes.** Los modelos de regresión lineal son relativamente simples y proporcionan una fórmula matemática fácil de interpretar para la creación de predicciones. Debido a que la regresión lineal es un procedimiento estadístico consolidado desde hace tiempo, las propiedades de estos modelos se conocen con mucho detalle. Normalmente, los modelos lineales se entrenan muy rápidamente. El nodo Lineal proporciona métodos para la selección automática de campos con el fin de eliminar de la ecuación los campos de entrada no significativos.

**Nota:** en los casos en que el campo objetivo es categórico en lugar de ser un rango continuo, como **sí/no** o **abandonar/no abandonar**, puede utilizarse la regresión logística como una alternativa. La regresión logística también admite las entradas no numéricas, por lo que no es necesario recodificar estos campos. Consulte el tema “Nodo Logística” en la página 187 para obtener más información.

# Modelos lineales

Los modelos lineales predicen un objetivo continuo basándose en relaciones lineales entre el objetivo y uno o más predictores.

Los modelos lineales son relativamente simples y proporcionan una fórmula matemática fácil de interpretar para la puntuación. Las propiedades de estos modelos se entienden bien y normalmente pueden crearse con bastante rapidez en comparación con otros tipos de modelos (como redes neuronales o árboles de decisión) del mismo conjunto de datos.

**Ejemplo.** Una correduría de seguros con recursos limitados para investigar las reclamaciones de seguros de los asegurados desea crear un modelo para estimar el coste de las reclamaciones. Al desplegar este modelo en centros de servicios, los representantes pueden introducir información sobre reclamaciones mientras atienden por teléfono al cliente y obtienen inmediatamente el coste "esperado" de la reclamación en función de los datos pasados.

**Requisitos de campo.** Debe haber un Objetivo y, al menos, una Entrada. De forma predeterminada, los campos con los roles predefinidos Ambos o Ninguno no se utilizan. El destino debe ser continuo (escala). No hay restricciones del nivel de medición de los predictores (entradas); los campos categóricos (marca, nominal y ordinal) se utilizan como factores en el modelo y los campos continuos se usan como covariables.

## Objetivos

### ¿Qué desea hacer?

- **Crear un modelo nuevo.** Crear un modelo totalmente nuevo. Éste es el funcionamiento habitual del nodo.

- **Continuar entrenando un modelo existente.** El entrenamiento continúa con el último modelo creado correctamente por el nodo. Esto permite actualizar un modelo existente sin tener que acceder a los datos originales. Además, puede dar como resultado un rendimiento significativamente más rápido ya que sólo se introducen en la ruta los registros nuevos o actualizados. Los detalles del modelo anterior se almacenan con el nodo de modelado, lo que permite utilizar esta opción incluso si el nugget de modelo anterior ya no está disponible en la ruta o la paleta de modelos.

  **Nota:** Cuando esta opción está habilitada, todos los demás controles de las pestañas Campos y Opciones de generación están inhabilitados.

**¿Cuál es su objetivo principal?** Seleccione el objetivo apropiado.

- **Crear un modelo estándar.** El método genera un modelo simple para predecir el destino mediante los predictores. Por lo general, los modelos estándar son más fáciles de interpretar y pueden puntuarse más rápido que los conjuntos por aumento, agregación auntodocimante o los conjuntos de datos muy grandes.

  **Nota:** Para modelos segmentados, para utilizar esta opción con **Continuar entrenando un modelo existente** debe estar conectado a Analytic Server.
- **Mejorar la precisión del modelo (aumento).** El método genera un modelo de conjunto mediante el aumento, que genera una secuencia de modelos para obtener predicciones más precisas. Se puede tardar más tiempo en generar y puntuar conjuntos que un modelo estándar.

  El aumento produce una sucesión de "modelos de componente", cada uno de ellos basados en el conjunto de datos completo. Antes de crear cada modelo de componente sucesivo, los registros se ponderan en función de los residuos del modelo del componente anterior. Los casos con residuos de grandes dimensiones tienen ponderaciones de análisis relativamente superiores para que el siguiente modelo de componente se centre en predecir correctamente estos registros. Juntos, estos modelos de componentes forman un modelo de conjunto. El modelo de conjunto puntúa algunos registros con una regla de combinación; las reglas disponibles dependen del nivel de medición del destino.
- **Mejorar la estabilidad del modelo (agregación autodocimante).** El método genera un modelo de conjunto mediante la agregación autodocimante, que genera varios modelos para obtener predicciones más fiables. Se puede tardar más tiempo en generar y puntuar conjuntos que un modelo estándar.

  La agregación autodocimante produce replicaciones del conjunto de datos de entrenamiento mediante muestreo con repetición del conjunto de datos original. Crea muestras de bootstrap de igual tamaño al conjunto de datos original. Es decir, se crea un "modelo de componente" de cada replicación. Juntos, estos modelos de componentes forman un modelo de conjunto. El modelo de conjunto puntúa algunos registros con una regla de combinación; las reglas disponibles dependen del nivel de medición del destino.
- **Crear un modelo para conjuntos de datos muy grandes.** El método genera un modelo de conjunto dividiendo el conjunto de datos en bloques de datos independientes. Seleccione esta opción si su conjunto de datos es demasiado grande para construir cualquiera de los modelos anteriores o para la generación incremental de modelos. Puede que se tarde menos tiempo en generar esta opción, pero se puede tardar más tiempo en puntuarla que un modelo estándar.

Consulte “Conjuntos” en la página 179 para los valores relacionados con el aumento, la agregación autodocimante y conjuntos de datos muy grandes.

## Conceptos básicos

**Preparar automáticamente datos.** Esta opción permite que el procedimiento transforme internamente el destino y los predictores para aprovechar al máximo el poder predictivo del modelo; cualquier transformación se guarda con el modelo y se aplica a los nuevos datos para su puntuación. Las versiones originales de los campos transformados se excluyen del modelo. De forma predeterminada, se realiza la siguiente preparación automática de datos.

- **Fecha y hora.** Cada predictor de fecha se transforma en un nuevo predictor continuo que contiene el tiempo transcurrido desde una fecha de referencia (01-01-1970). Cada predictor de hora se transforma en un nuevo predictor continuo que contiene el tiempo transcurrido desde una hora de referencia (00:00:00).
- **Ajustar nivel de medición.** Los predictores continuos con menos de 5 valores distintos se reestructuran como predictores ordinales. Los predictores ordinales con más de 10 valores distintos se reestructuran como predictores continuos.
- **Tratamiento de valores atípicos.** Los valores de los predictores continuos que recaen más allá de un valor de corte (3 desviaciones estándar de la media) se establecen con el valor de corte.
- **Gestión de valores perdidos.** Los valores perdidos de los predictores nominales se sustituyen por el modo de la partición de entrenamiento. Los valores perdidos de los predictores ordinales se sustituyen por la mediana de la partición de entrenamiento. Los valores perdidos de los predictores continuos se sustituyen por la media de la partición de entrenamiento.
- **Fusión supervisada.** Hace un modelo más parsimonioso reduciendo el número de campos que deben procesarse junto con el destino. Las categorías similares se identifican en función de la relación entre la entrada y destino. Las categorías que no son significativamente diferentes; es decir, que tienen un valor p superior al valor 0,1, se fusionan. Tenga en cuenta que si todas las categorías se combinan en una, las versiones original y derivada del campo se excluyen del modelo porque no tienen ningún valor como predictor.

**Nivel de confianza.** Éste es el nivel de confianza que se utiliza para calcular las estimaciones de intervalos de los coeficientes de modelos en la vista Coeficientes. Especifique un valor mayor que 0 y menor que 100. El valor predeterminado es 95.

## Selección de modelos

**Método de selección de modelos.** Seleccione uno de los métodos de selección de modelos (a continuación se encuentran los detalles) o **Incluir todos los predictores**, que simplemente introduce todos los predictores disponibles como términos del modelo de efectos principales. De forma predeterminada, se utiliza **Pasos sucesivos hacia adelante**.

**Selección de Pasos sucesivos hacia adelante.** Comienza sin efectos en el modelo y añade y elimina efectos paso por paso hasta que ya no se puedan añadir o eliminar según los criterios de los pasos sucesivos.

- **Criterios para entrada/eliminación.** Ésta es la estadística utilizada para determinar si debe añadirse o eliminarse un efecto del modelo. **Criterio de información (AICC)** se basa en la similitud del conjunto de entrenamiento que se le da al modelo, y se ajusta para penalizar modelos excesivamente complejos. **Estadísticos de F** se utiliza en una prueba estadística de la mejora en el error de modelo. **R cuadrado corregida** se basa en el ajuste del conjunto de entrenamiento, y se ajusta para penalizar modelos excesivamente complejos. **Criterio de prevención sobreajustado (ASE)** se basa en el ajuste del conjunto (error cuadrático medio, ASE) de prevención sobreajustado. El conjunto de prevención sobreajustado es una submuestra aleatoria de aproximadamente 30% del conjunto de datos original que no se utiliza para entrenar el modelo.

  Si se selecciona otro criterio que no sea **Estadísticos de F**, se añadirá al modelo cada paso del efecto que se corresponda con el aumento positivo mayor en el criterio. Se eliminará cualquier efecto en el modelo que se corresponda con una disminución en el criterio.

  Si se selecciona **Estadísticos de F** como criterio, cada paso en el efecto que tenga el valor *p* más pequeño inferior al umbral especificado, se añadirá **Incluir efectos con valores p inferiores a** al modelo. El valor predeterminado es 0,05. Cualquier efecto en el modelo con un valor *p* superior al umbral especificado, **Eliminar efectos con valores p mayores que**, será eliminado. El valor predeterminado es 0.10.
- **Personalizar número máximo de efectos en el modelo final.** De forma predeterminada, pueden introducirse todos los efectos disponibles en el modelo. Del mismo modo, si el algoritmo por pasos sucesivos termina con un paso con el número máximo de efectos especificado, el algoritmo se detiene con el conjunto actual de efectos.

- **Personalizar número máximo de pasos.** El algoritmo por pasos sucesivos termina tras un cierto número de pasos. De forma predeterminada, es 3 veces el número de efectos disponibles. Del mismo modo, especifique un entero positivo para el número máximo de pasos.

**Selección de mejores subconjuntos.** Comprueba "todos los modelos posibles", o al menos el subconjunto más grande de los modelos posibles que los pasos sucesivos hacia adelante, para seleccionar el mejor según el criterio de mejores subconjuntos. **Criterio de información (AICC)** se basa en la similitud del conjunto de entrenamiento que se le da al modelo, y se ajusta para penalizar modelos excesivamente complejos. **R cuadrado corregida** se basa en el ajuste del conjunto de entrenamiento, y se ajusta para penalizar modelos excesivamente complejos. **Criterio de prevención sobreajustado (ASE)** se basa en el ajuste del conjunto (error cuadrático medio, ASE) de prevención sobreajustado. El conjunto de prevención sobreajustado es una submuestra aleatoria de aproximadamente 30% del conjunto de datos original que no se utiliza para entrenar el modelo.

Se selecciona el modelo con el valor mayor del criterio como el mejor modelo.

**Nota:** la selección de mejores subconjuntos requiere más trabajo computacional que la selección por pasos sucesivos hacia adelante. Cuando los mejores subconjuntos se procesan junto con aumento, agregación autodocimante y conjuntos de datos de gran tamaño, la generación de un modelo estándar generado mediante una selección por pasos sucesivos hacia delante puede tardar considerablemente más tiempo.

## Conjuntos

Estos ajustes determinan el comportamiento de la agrupación que se produce cuando los conjuntos de datos de gran tamaño o de aumento o agregación autodocimante son obligatorios en Objetivos. Las opciones no aplicables al objetivo seleccionado se ignorarán.

**Agregación autodocimante y conjuntos de datos muy grandes.** Al puntuar un conjunto, ésta es la regla utilizada para combinar los valores predichos a partir de los modelos básicos para calcular el valor de puntuación del conjunto.

- **Regla de combinación predeterminada para objetivos continuos.** Los valores predichos de conjunto para objetivos continuos pueden combinarse mediante la media o mediana de los valores predichos a partir de los modelos básicos.

Tenga en cuenta que cuando el objetivo es mejorar la precisión del modelo, se ignoran las selecciones de reglas de combinación. El aumento siempre utiliza un voto de mayoría ponderada para puntuar objetivos categóricos y una mediana ponderada para puntuar objetivos continuos.

**Aumento y agregación autodocimante.** Especifique el número de modelos básicos que debe generarse cuando el objetivo es mejorar la precisión o estabilidad del modelo; en el caso de la agregación autodocimante, se trata del número de muestras de bootstrap. Debe ser un número entero positivo.

## Avanzado

**Replicar resultados.** Al establecer una semilla aleatoria podrá replicar análisis. El generador de números aleatorios se utiliza para seleccionar qué registros se encuentran en el conjunto de prevención sobreajustado. Especifique un entero o pulse en **Generar**, lo que creará un entero pseudo-aleatorio entre 1 y 2147483647, ambos inclusive. El valor predeterminado es 54752075.

## Opciones de modelos

**Nombre del modelo.** Puede generar el nombre del modelo automáticamente tomando como base los campos objetivo o especificar un nombre personalizado. El nombre generado automáticamente es el nombre del campo objetivo.

Tenga en cuenta que el valor predicho se calcula siempre cuando se puntúa el modelo. El nombre del nuevo campo es el nombre del campo objetivo con el prefijo *$L-*. Por ejemplo, para un campo objetivo llamado *ventas*, el nuevo campo se llamaría*$L-ventas*.

## Resumen del modelo

La vista Resumen del modelo es una instantánea, un resumen visual del modelo y su ajuste.

**Tabla.** La tabla identifica algunos parámetros de modelo superior, incluyendo:

- El nombre del destino especificado en la pestaña Campos,
- Si la preparación de datos automática se ha realizado tal como se especifica en la configuración de Básicos,
- El criterio de selección y el método de selección de modelo especificados en la configuración de Selección de modelo. También se muestra el valor del criterio de selección del modelo final y se presenta en un formato más reducido y mejor.

**Gráfico.** El gráfico muestra la precisión del modelo final, que se presenta en el formato mayor es mejor. El valor es $100 \times R^2$ ajustado para el modelo final.

## Preparación automática de datos

Esta vista muestra información a cerca de qué campos se excluyen y cómo los campos transformados se derivaron en el paso de preparación automática de datos (ADP). Para cada campo que fue transformado o excluido, la tabla enumera el nombre del campo, su rol en el análisis y la acción tomada por el paso ADP. Los campos se clasifican por orden alfabético ascendente de nombres de campo. Las posibles acciones que se toman en cada campo incluyen:

- **Derivar duración: meses** calcula el tiempo transcurrido en meses a partir de los valores de un campo que contiene las fechas hasta la fecha actual del sistema.
- **Derivar duración: horas** calcula el tiempo transcurrido en horas a partir de los valores de un campo que contiene las horas hasta la hora actual del sistema.
- **Cambiar nivel de medición de continuo a ordinal** reestructura los campos continuos con menos de 5 valores exclusivos como campos ordinales.
- **Cambiar nivel de medición de ordinal a continuo** reestructura los campos continuos con más de 10 valores exclusivos como campos continuos.
- **Recortar valores atípicos** define los valores de los predictores continuos que recaen más allá de un valor de corte (3 desviaciones estándar de la media) se establecen con el valor de corte.
- **Sustituir valores perdidos** sustituye los valores perdidos de los campos nominales por el modo, los campos ordinales por la mediana y los campos continuos por la media.
- **Combinar categorías para aumentar al máximo la asociación con el destino** identifica categorías de predictor "similares" basadas en la relación entre la entrada y el destino. Las categorías que no son significativamente diferentes; es decir, que tienen un valor *p* superior al valor 0,05, se fusionan.
- **Excluir predictor constante / después del tratamiento de los valores atípicos / después de fusionar categorías** elimina los predictores que tiene un único valor, posiblemente después de tomar otras acciones ADP.

## Importancia de predictor

Es normal centrar los esfuerzos de modelado en los campos predictores más importantes y valorar la omisión de aquellos con menor relevancia. El gráfico de importancia de los predictores le ayuda a hacerlo indicando la importancia relativa de cada predictor en la estimación del modelo. Como los valores son relativos, la suma de valores de todos los predictores de la visualización es 1.0. La importancia del predictor no está relacionada con la precisión del modelo. Sólo está relacionada con la importancia de cada predictor a la hora de realizar una predicción, no con si la predicción es o no precisa.

## Predicho por observado

Muestra un diagrama de dispersión en intervalos de los valores predichos en el eje vertical por los valores observados en el eje horizontal. Idealmente, los puntos deben basarse en una línea de 45 grados; esta vista indica si hay algún registro predicho de manera incorrecta en el modelo.

## Residuos

Muestra un gráfico de diagnosis de los residuos del modelo.

**Estilos de gráfico.** Existen varios estilos de visualización diferentes, que son accesibles desde la lista desplegable **Estilo**.

- **Histograma.** Se trata de un histograma en intervalos de los residuos estudentizados de una superposición de la distribución normal. Los modelos lineales asumen que los residuos tienen una distribución normal, de forma que el histograma debería estar muy cercano a la línea continua.
- **Gráfico p-p.** Se trata de un gráfico probabilidad-probabilidad en intervalos que compara los residuos estudentizados con una distribución normal. Si la curva de los puntos representados es menos pronunciada que la línea normal, los residuos muestran una variabilidad mayor que una distribución normal; si la curva es más pronunciada, los residuos muestran una variabilidad inferior que una distribución normal. Si los puntos representados tienen una curva con forma en S, la distribución de los residuos es asimétrica.

## Valores atípicos

Esta tabla enumera los registros que ejercen una influencia excesiva sobre el modelo, y muestra el ID de registro (si se especifica en la pestaña Campos), el valor objetivo y la distancia de Cook. La distancia de Cook es una medida de cuánto cambiarían los residuos de todos los registros si un registro en particular se excluyera del cálculo de los coeficientes del modelo. Una distancia de Cook grande indica que la exclusión de un registro cambia sustancialmente los coeficientes, y por lo tanto debe considerarse relevante.

Los registros relevantes deben examinarse cuidadosamente para determinar si puede darles menos importancia en la estimación del modelo, truncar los valores atípicos a algún umbral aceptable o eliminar los registros relevantes completamente.

## Efectos

Esta vista muestra el tamaño de cada efecto en el modelo.

**Estilos.** Existen varios estilos de visualización diferentes, que son accesibles desde la lista desplegable **Estilo**.

- **Diagrama.** Es un gráfico en el que los efectos se clasifican desde arriba hacia abajo con una importancia de predictores descendente. Las líneas de conexión del diagrama se ponderan tomando como base la significación del efecto, con un grosor de línea mayor correspondiente a efectos con mayor significación (valores *p* inferiores). Al pasar el ratón por encima de una línea de conexión muestra la información sobre herramientas que muestra el valor de *p* y la importancia del efecto. Este es el método predeterminado.
- **Tabla.** Se trata de una tabla ANOVA para el modelo completo y los efectos de modelo individuales. Los efectos individuales se clasifican desde arriba hacia abajo con una importancia de predictores descendente. Tenga en cuenta que, de forma predeterminada, la tabla se contrae para mostrar únicamente los resultados del modelo global. Para ver los resultados de los efectos del modelo individual, pulse la casilla **Modelo corregido** en la tabla.

**Importancia del predictor.** Existe un control deslizante Importancia del predictor que controla qué predictores se muestran en la vista. Esto no cambia el modelo, simplemente le permite centrarse en los predictores más importantes. De forma predeterminada, se muestran los 10 efectos más importantes.

**Significación.** Existe un control deslizante Significación que controla aún más qué efectos se muestran en la vista, a parte de los que se muestran tomando como base la importancia de predictor. Se ocultan los efectos con valores de significación superiores al valor del control deslizante. Esto no cambia el modelo, simplemente le permite centrarse en los efectos más importantes. El valor predeterminado es 1.00, de modo que no se filtran efectos tomando como base la significación.

## Coeficientes

Esta vista muestra el valor de cada coeficiente en el modelo. Tenga en cuenta que los factores (predictores categóricos) tienen codificación de indicador dentro del modelo, de modo que los **efectos** que contienen los factores generalmente tendrán múltiples **coeficientes** asociados: uno por cada categoría exceptuando la categoría que corresponde al parámetro (de referencia) redundante.

**Estilos.** Existen varios estilos de visualización diferentes, que son accesibles desde la lista desplegable **Estilo**.

- **Diagrama.** Es un gráfico que muestra la interceptación primero, y luego clasifica los efectos desde arriba hacia abajo con una importancia de predictores descendente. Dentro de los efectos que contienen factores, los coeficientes se clasifican en orden ascendente de valores de datos. Las líneas de conexión del diagrama se colorean en función del coeficiente (consulte la clave del diagrama) y se ponderan tomando como base la significación del coeficiente, con un grosor de línea mayor correspondiente a coeficientes con mayor significación (valores *p* inferiores). Al pasar el ratón por encima de una línea de conexión se muestra la información sobre herramientas que muestra el valor del coeficiente, su valor *p* y la importancia del efecto con el que se asocia el parámetro. Este es el estilo predeterminado.
- **Tabla.** Muestra los valores, las pruebas de significación y los intervalos de confianza para los coeficientes de modelos individuales. Tras la interceptación, los efectos se clasifican desde arriba hacia abajo con una importancia de predictores descendente. Dentro de los efectos que contienen factores, los coeficientes se clasifican en orden ascendente de valores de datos. Tenga en cuenta que, de forma predeterminada, la tabla se contrae para mostrar únicamente el coeficiente, la significación y la importancia de cada parámetro del modelo. Para ver el error estándar, estadístico *t* y el intervalo de confianza, pulse la casilla **Coeficiente** en la tabla. Al pasar el ratón por encima del nombre de un parámetro de modelo en la tabla se muestra la información sobre herramientas con el nombre del parámetro, el efecto con el que se asocia el parámetro y (en los predictores categóricos) las etiquetas de valor asociadas con el parámetro del modelo. Puede ser especialmente útil para ver las nuevas categorías creadas cuando la preparación de datos automática fusiona categorías similares de un predictor categórico.

**Importancia del predictor.** Existe un control deslizante Importancia del predictor que controla qué predictores se muestran en la vista. Esto no cambia el modelo, simplemente le permite centrarse en los predictores más importantes. De forma predeterminada, se muestran los 10 efectos más importantes.

**Significación.** Existe un control deslizante Significación que controla aún más qué coeficientes se muestran en la vista, a parte de los que se muestran tomando como base la importancia de predictor. Se ocultan los coeficientes con valores de significación superiores al valor del control deslizante. Esto no cambia el modelo, simplemente le permite centrarse en los coeficientes más importantes. El valor predeterminado es 1.00, de modo que no se filtran coeficientes tomando como base la significación.

## Medias estimadas

Son gráficos representados para predictores significativos. El gráfico muestra el valor estimado de modelo del objetivo en el eje vertical de cada valor del predictor en el eje horizontal, que alberga el resto de los predictores constantes. Proporciona una visualización útil de los efectos de los coeficientes de cada predictor en el objetivo.

*Note:* si no hay predictores significativos, no se generan medias estimadas.

## Resumen de creación de modelos

Cuando se selecciona un algoritmo de selección de modelos que no sea **Ninguno**, proporciona algunos detalles del proceso de creación del modelo.

**Pasos sucesivos hacia adelante.** Cuando la selección por pasos hacia adelante es el algoritmo de selección, la tabla muestra los últimos 10 pasos en el algoritmo de selección por pasos hacia adelante. Para cada paso, se muestran el valor del criterio de selección y los efectos en el modelo en ese paso. Esto

ofrece el sentido del grado de contribución de cada paso al modelo. Cada columna le permite clasificar las filas, de modo que es posible ver con mayor facilidad qué efectos hay en un paso en particular.

**Mejores subconjuntos.** Cuando Mejores subconjuntos es el algoritmo de selección, la tabla muestra los 10 modelos principales. Para cada modelo, se muestran el valor del criterio de selección y los efectos en el modelo. Esto ofrece un sentido de la estabilidad de los modelos principales; si tienden a tener muchos efectos similares con pocas diferencias, puede tenerse una confianza casi completa en el modelo "principal"; si tienden a tener muchos efectos diferentes, algunos efectos pueden ser demasiado parecidos y deberían combinarse (o eliminar uno). Cada columna le permite clasificar las filas, de modo que es posible ver con mayor facilidad qué efectos hay en un paso en particular.

### Configuración

Tenga en cuenta que el valor predicho se calcula siempre cuando se puntúa el modelo. El nombre del nuevo campo es el nombre del campo objetivo con el prefijo *$L-*. Por ejemplo, para un campo objetivo llamado *ventas*, el nuevo campo se llamaría*$L-ventas*.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar convirtiendo a SQL nativo** Si selecciona esta opción, se genera SQL para puntuar el modelo dentro de la base de datos.

  **Nota:** aunque esta opción puede proporcionar resultados más rápidos, el tamaño y la complejidad del SQL nativo aumenta a medida que lo hace la complejidad del modelo.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

## Nodo Linear-AS

IBM SPSS Modeler tiene dos versiones distintas del nodo Lineal:

- **Lineal** es el nodo tradicional que se ejecuta en IBM SPSS Modeler Server.
- **AS lineal** se puede ejecutar cuando está conectado a IBM SPSS Analytic Server.

La regresión lineal es una técnica de estadístico común para clasificar los registros en función los valores de los campos de entrada numérica. La regresión lineal se ajusta a una línea recta o una superficie que minimiza las discrepancias entre los valores de resultados predichos y reales.

**Requisitos.** Sólo se pueden utilizar campos numéricos y predictores categóricos en un modelo de regresión lineal. Debe tener exactamente un campo objetivo (con el rol definido a **Objetivo**) y uno o más predictores (con el rol definido a **Entrada**). Los campos con un rol **Ambos** o **Ninguno** se ignoran, ya que no son campos numéricos. (Si es necesario, los campos no numéricos se pueden recodificar mediante un nodo Derivar.)

**Puntos fuertes.** Los modelos de regresión lineal son relativamente simples y proporcionan una fórmula matemática fácil de interpretar para la creación de predicciones. Debido a que la regresión lineal es un procedimiento estadístico consolidado desde hace tiempo, las propiedades de estos modelos se conocen

con mucho detalle. Normalmente, los modelos lineales se entrenan muy rápidamente. El nodo Lineal proporciona métodos para la selección automática de campos con el fin de eliminar de la ecuación los campos de entrada no significativos.

**Nota:** en los casos en que el campo objetivo es categórico en lugar de ser un rango continuo, como **sí/no** o **abandonar/no abandonar**, puede utilizarse la regresión logística como una alternativa. La regresión logística también admite las entradas no numéricas, por lo que no es necesario recodificar estos campos. Consulte el tema "Nodo Logística" en la página 187 para obtener más información.

# Modelos Linear-AS

Los modelos lineales predicen un objetivo continuo basándose en relaciones lineales entre el objetivo y uno o más predictores.

Los modelos lineales son relativamente simples y proporcionan una fórmula matemática fácil de interpretar para la puntuación. Las propiedades de estos modelos se entienden bien y normalmente pueden crearse con bastante rapidez en comparación con otros tipos de modelos (como redes neuronales o árboles de decisión) del mismo conjunto de datos.

**Ejemplo.** Una correduría de seguros con recursos limitados para investigar las reclamaciones de seguros de los asegurados desea crear un modelo para estimar el coste de las reclamaciones. Al desplegar este modelo en centros de servicios, los representantes pueden introducir información sobre reclamaciones mientras atienden por teléfono al cliente y obtienen inmediatamente el coste "esperado" de la reclamación en función de los datos pasados.

**Requisitos de campo.** Debe haber un Objetivo y, al menos, una Entrada. De forma predeterminada, los campos con los roles predefinidos Ambos o Ninguno no se utilizan. El destino debe ser continuo (escala). No hay restricciones del nivel de medición de los predictores (entradas); los campos categóricos (marca, nominal y ordinal) se utilizan como factores en el modelo y los campos continuos se usan como covariables.

## Información básica

**Incluir interceptación.** Esta opción incluye un desplazamiento en el eje y cuando el eje x es 0. La interceptación se incluye generalmente en el modelo. Si se puede dar por supuesto la lectura de datos en el origen, se puede excluir la interceptación.

**Considerar interacción de dos factores** Esta opción indica al modelo que debe comparar cada par de entradas posibles para ver si la tendencia de una afecta a la otra. Si es así, entonces es más probable que esas entradas se incluyan en la matriz del diseño.

**Intervalo de confianza para estimaciones de coeficientes (%).** Éste es el intervalo de confianza que se utiliza para calcular las estimaciones de los coeficientes de modelos en la vista Coeficientes. Especifique un valor mayor que 0 y menor que 100. El valor predeterminado es 95.

**Orden de clasificación para predictores categóricos** Estos controles determinan el orden de las categorías para los factores (entradas categóricas) para determinar la "última" categoría. El valor del orden de clasificación se ignora si la entrada no es categórica o si se especifica una categoría de referencia personalizada.

## Selección de modelos

**Método de selección de modelos.** Seleccione uno de los métodos de selección de modelos (a continuación se encuentran los detalles) o **Incluir todos los predictores**, que simplemente introduce todos los predictores disponibles como términos del modelo de efectos principales. De forma predeterminada, se utiliza **Pasos sucesivos hacia adelante**.

**Selección de Pasos sucesivos hacia adelante.** Comienza sin efectos en el modelo y añade y elimina efectos paso por paso hasta que ya no se puedan añadir o eliminar según los criterios de los pasos sucesivos.

- **Criterios para entrada/eliminación.** Ésta es la estadística utilizada para determinar si debe añadirse o eliminarse un efecto del modelo. **Criterio de información (AICC)** se basa en la similitud del conjunto de entrenamiento que se le da al modelo, y se ajusta para penalizar modelos excesivamente complejos. **Estadísticos de F** se utiliza en una prueba estadística de la mejora en el error de modelo. **R cuadrado corregida** se basa en el ajuste del conjunto de entrenamiento, y se ajusta para penalizar modelos excesivamente complejos. **Criterio de prevención sobreajustado (ASE)** se basa en el ajuste del conjunto (error cuadrático medio, ASE) de prevención sobreajustado. El conjunto de prevención sobreajustado es una submuestra aleatoria de aproximadamente 30% del conjunto de datos original que no se utiliza para entrenar el modelo.

  Si se selecciona otro criterio que no sea **Estadísticos de F**, se añadirá al modelo cada paso del efecto que se corresponda con el aumento positivo mayor en el criterio. Se eliminará cualquier efecto en el modelo que se corresponda con una disminución en el criterio.

  Si se selecciona **Estadísticos de F** como criterio, cada paso en el efecto que tenga el valor *p* más pequeño inferior al umbral especificado, se añadirá **Incluir efectos con valores p inferiores a** al modelo. El valor predeterminado es 0,05. Cualquier efecto en el modelo con un valor *p* superior al umbral especificado, **Eliminar efectos con valores p mayores que**, será eliminado. El valor predeterminado es 0.10.
- **Personalizar número máximo de efectos en el modelo final.** De forma predeterminada, pueden introducirse todos los efectos disponibles en el modelo. Del mismo modo, si el algoritmo por pasos sucesivos termina con un paso con el número máximo de efectos especificado, el algoritmo se detiene con el conjunto actual de efectos.
- **Personalizar número máximo de pasos.** El algoritmo por pasos sucesivos termina tras un cierto número de pasos. De forma predeterminada, es 3 veces el número de efectos disponibles. Del mismo modo, especifique un entero positivo para el número máximo de pasos.

**Selección de mejores subconjuntos.** Comprueba "todos los modelos posibles", o al menos el subconjunto más grande de los modelos posibles que los pasos sucesivos hacia adelante, para seleccionar el mejor según el criterio de mejores subconjuntos. **Criterio de información (AICC)** se basa en la similitud del conjunto de entrenamiento que se le da al modelo, y se ajusta para penalizar modelos excesivamente complejos. **R cuadrado corregida** se basa en el ajuste del conjunto de entrenamiento, y se ajusta para penalizar modelos excesivamente complejos. **Criterio de prevención sobreajustado (ASE)** se basa en el ajuste del conjunto (error cuadrático medio, ASE) de prevención sobreajustado. El conjunto de prevención sobreajustado es una submuestra aleatoria de aproximadamente 30% del conjunto de datos original que no se utiliza para entrenar el modelo.

Se selecciona el modelo con el valor mayor del criterio como el mejor modelo.

**Nota:** la selección de mejores subconjuntos requiere más trabajo computacional que la selección por pasos sucesivos hacia adelante. Cuando los mejores subconjuntos se procesan junto con aumento, agregación autodocimante y conjuntos de datos de gran tamaño, la generación de un modelo estándar generado mediante una selección por pasos sucesivos hacia delante puede tardar considerablemente más tiempo.

## Opciones de modelo

**Nombre del modelo.** Puede generar el nombre del modelo automáticamente tomando como base los campos objetivo o especificar un nombre personalizado. El nombre generado automáticamente es el nombre del campo objetivo.

Tenga en cuenta que el valor predicho se calcula siempre cuando se puntúa el modelo. El nombre del nuevo campo es el nombre del campo objetivo con el prefijo *$L-*. Por ejemplo, para un campo objetivo llamado *ventas*, el nuevo campo se llamaría*$L-ventas*.

## Resultado interactivo

Después de ejecutar un modelo Lineal-AS, están disponibles los siguientes resultados.

### Información del modelo

La vista de información de modelo proporciona información clave acerca del modelo. La tabla identifica algunos ajustes de modelo de alto nivel, por ejemplo:

- El nombre del objetivo especificado en la pestaña Campos
- El campo de ponderación de regresión
- El método de generación de modelos especificado en la configuración de Selección de modelos
- El número de entrada de predictores
- El número de predictores del modelo final
- Criterio de información de Akaike corregido (AICC). AICC es una medida para seleccionar y comparar modelos mixtos basada en el logaritmo de la verosimilitud -2 (restringida). Los valores menores indican modelos mejores. El AICC "corrige" el AIC respecto a tamaños muestrales pequeños. A medida que aumenta el tamaño de la muestra, el AICC converge con el AIC.
- R Cuadrado. Medida de la bondad de ajuste de un modelo lineal; en ocasiones recibe el nombre de coeficiente de determinación. Es la proporción de la variación de la variable dependiente explicada por el modelo de regresión. Puede tomar un valor entre 0 y 1. Un valor pequeño indica que el modelo no se ajusta bien a los datos.
- R cuadrado ajustado

### Resumen de registros

La vista Resumen de registros proporciona información acerca del número y porcentaje de registros (casos) incluidos y excluidos del modelo.

### Importancia del predictor

Es normal centrar los esfuerzos de modelado en los campos predictores más importantes y valorar la omisión de aquellos con menor relevancia. El gráfico de importancia de los predictores le ayuda a hacerlo indicando la importancia relativa de cada predictor en la estimación del modelo. Como los valores son relativos, la suma de valores de todos los predictores de la visualización es 1.0. La importancia del predictor no está relacionada con la precisión del modelo. Sólo está relacionada con la importancia de cada predictor a la hora de realizar una predicción, no con si la predicción es o no precisa.

### Predicho por observado

Muestra un diagrama de dispersión en intervalos de los valores predichos en el eje vertical por los valores observados en el eje horizontal. Idealmente, los puntos deben basarse en una línea de 45 grados; esta vista indica si hay algún registro predicho de manera incorrecta en el modelo.

## Configuración

Tenga en cuenta que el valor predicho se calcula siempre cuando se puntúa el modelo. El nombre del nuevo campo es el nombre del campo objetivo con el prefijo *$L-*. Por ejemplo, para un campo objetivo llamado *ventas*, el nuevo campo se llamaría*$L-ventas*.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso.** Si se conecta a una base de datos con un adaptador de puntuación instalado,

se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

- **Puntuar fuera de la base de datos.** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

## Nodo Logística

La **regresión logística**, también denominada **regresión nominal**, es una técnica estadística para clasificar los registros a partir de los valores de los campos de entrada. Es análoga a la regresión lineal pero utiliza un campo objetivo categórico en lugar de uno numérico. Se admiten tanto los modelos binomiales (para objetivos con dos categorías discretas) como los multinomiales (para objetivos con más de dos categorías).

La regresión logística trabaja creando un conjunto de ecuaciones que relacionan los valores de los campos de entrada con las probabilidades asociadas a cada una de las categorías de los campos de salida. Una vez se ha generado el modelo, se puede utilizar para calcular las probabilidades de datos nuevos. Para cada registro, se calcula una probabilidad de pertenencia a cada categoría posible de salida. La categoría objetivo con la probabilidad más alta se asigna como el valor de salida predicho para cada registro.

**Ejemplo binomial.** Un proveedor de telecomunicaciones está preocupado por el número de clientes que se están pasando a la competencia. Mediante los datos de uso de servicio puede crear un modelo binomial para predecir qué clientes tienen más probabilidad de contratar otro proveedor y personalizar las ofertas para retener el mayor número de clientes posible. Se utiliza un modelo binomial porque el objetivo tiene dos categorías distintas (probabilidad de pasar a la competencia o no).

**Nota:** Solo para modelos binomiales, los campos de serie están limitados a ocho caracteres. Si es necesario, se pueden recodificar series más largas utilizando un nodo Reclasificar o utilizando el nodo Anonimizar.

**Ejemplo multinomial.** Un proveedor de telecomunicaciones ha segmentado su base de clientes por patrones de uso de servicio, categorizando los clientes en cuatro grupos. Al utilizar datos demográficos para predecir la pertenencia a un grupo, puede crear un modelo multinomial para clasificar a los clientes potenciales en grupos y personalizar las ofertas a los clientes individuales.

**Requisitos.** Uno o más campos de entrada y exactamente un campo objetivo categórico con dos o más categorías. Para un modelo binomial el objetivo debe tener un nivel de medición de *Marca*. Para un modelo multinomial, el objetivo puede tener un nivel de medición de *Marca* o *Nominal* con dos o más categorías. Se ignorarán los campos establecidos en *Ambos* o *Ninguno*. Los tipos de los campos utilizados en el modelo deben estar completamente instanciados.

**Puntos fuertes.** Los modelos de regresión logística suelen ser bastante exactos. Pueden gestionar campos de entrada simbólicos y numéricos. Pueden proporcionar probabilidades predichas para todas las categorías objetivo, de forma que la "segunda mejor predicción" sea fácil de identificar. Los modelos logísticos son más eficaces cuando la pertenencia a grupos es un campo categórico verdadero; si la pertenencia a un grupo está basada en los valores de un campo de rango continuo (por ejemplo, "CI alto" frente a "CI bajo"), debería considerar la posibilidad de utilizar una regresión lineal para aprovechar la mayor riqueza de información que ofrece el rango completo de valores. Los modelos logísticos también pueden realizar la selección de campos automática, aunque otros métodos, como los modelos de árboles o la selección de características, pueden hacerlo de forma más rápida en conjuntos de datos grandes. Por último, ya que los modelos logísticos son bien conocidos por muchos analistas y analistas de datos, se pueden utilizar como línea de base con la que comparar otras técnicas de modelado.

Al procesar conjuntos grandes de datos, puede mejorar sensiblemente el rendimiento desactivando el contraste de razón de verosimilitud, una opción avanzada de los resultados. Consulte el tema “Salida avanzada de regresión logística” en la página 193 para obtener más información.

**Importante:** Si el espacio del disco temporal es bajo, la regresión logística binomial puede fallar y no crearse y se mostrará un error. Cuando se construye a partir de un conjunto de datos muy grande (10 GB o más), es necesaria la misma cantidad de espacio libre en disco. Puede utilizar la variable de entorno `SPSSTMPDIR` para establecer la ubicación del directorio temporal.

## Opciones de modelo para el nodo Logística

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Crear modelos divididos.** Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como campos de división. Consulte el tema "Generación de modelos divididos" en la página 28 para obtener más información.

**Procedimiento.** Especifica si se ha creado un modelo binomial o multinomial. Las opciones disponibles en el cuadro de diálogo varían en función del tipo de procedimiento de modelado seleccionado.

- **Binomial.** Se utiliza cuando el campo objetivo es un campo marca o nominal con dos valores discretos (dicotómicos), como *sí/no*, *encendido/apagado*, *hombre/mujer*.
- **Multinomial.** Se utiliza cuando el campo objetivo es un campo nominal con más de dos valores. Puede especificar **Efectos principales**, **Factorial completo** o **Personalizada**.

**Incluir la constante en la ecuación.** Esta opción determina si las ecuaciones resultantes incluirán un término constante. En la mayoría de las situaciones, debe dejar esta opción seleccionada.

### Modelos binomiales

Para los modelos binomiales, están disponibles los siguientes métodos y opciones:

**Método.** Especifique el método que se va a utilizar para la creación del modelo de regresión logística.

- **Intro.** Éste es el método predeterminado que introduce directamente todos los términos en la ecuación. No se realiza ninguna selección de campos en la creación del modelo.
- **Por pasos hacia adelante.** El método Por pasos hacia adelante de selección de campo crea la ecuación por pasos, como su nombre indica. El modelo inicial es el más simple que puede haber, sin ningún término del modelo (excepto el constante) en la ecuación. En cada paso, se evalúan los términos que no se han añadido aún al modelo y si el mejor de dichos términos se suma de forma significativa a la eficacia predictiva del modelo, se añadirá a éste. Además, los términos que se encuentran actualmente en el modelo se vuelven a evaluar para determinar si se puede eliminar alguno de ellos sin que afecte al modelo de forma significativa. Si es así, se eliminan. El proceso se repite y se añaden y/o eliminan otros términos. Cuando no se puedan añadir más términos para mejorar el modelo, y no se puedan eliminar más sin que le afecte, se creará el modelo final.
- **Por pasos hacia atrás.** El método Por pasos hacia atrás es fundamentalmente lo contrario al método Por pasos hacia adelante. Con este método, el modelo inicial contiene todos los términos como predictores. En cada paso, se evalúan los términos del modelo y se eliminan los que no afecten al modelo de forma significativa. Además, los términos eliminados anteriormente se vuelven a evaluar para determinar si el mejor de dichos términos se añade de forma significativa a la eficacia predictiva del modelo. Si es así, se volverá a añadir al modelo. Cuando no se puedan añadir más términos para mejorar el modelo y no se puedan eliminar más sin que le afecte, se creará el modelo final.

**Entradas categóricas.** Enumera los campos que están identificados como categóricos, o sea, los que tienen un nivel de medición de marca, nominal u ordinal. Puede especificar el contraste y la categoría base para cada campo categórico.

- **Nombre de campo.** Esta columna contiene los nombres de campo de las entradas categóricas. Para añadir entradas continuas o numéricas a esta columna, pulse en el icono de agregar campos situado a la derecha de la lista y, a continuación, seleccione las entradas requeridas.
- **Contraste.** La interpretación de los coeficientes de regresión para un campo categórico depende de los contrastes utilizados. El contraste determina como se configuran los contrastes de hipótesis para comparar las medias estimadas. Por ejemplo, si sabe que un campo categórico tiene un orden implícito (como un patrón o agrupación), puede utilizar el contraste para modelar dicho orden. Los contrastes disponibles son:

  **Indicador.** Los contrastes indican la presencia o ausencia de la pertenencia a una categoría. Éste es el método predeterminado.

  **simples.** Se compara cada categoría del campo predictor, excepto la categoría de referencia, con la categoría de referencia.

  **Diferencia.** Se compara cada categoría del campo predictor, excepto la primera categoría, con el efecto promedio de las categorías anteriores. También se conoce como contrastes de Helmert inversos.

  **Helmert.** Se compara cada categoría del campo predictor, excepto la última categoría, con el efecto promedio de las categorías posteriores.

  **Repetidas.** Se compara cada categoría del campo predictor, excepto la primera categoría, con la categoría que la precede.

  **Polinómico.** Contrastes polinómicos ortogonales. Se supone que las categorías están espaciadas de forma equidistante. Los contrastes polinómicos están disponibles sólo para los campos numéricos.

  **Desviación.** Se compara cada categoría del campo predictor, excepto la categoría de referencia, con el efecto global.
- **Categoría base.** Especifica de qué forma se determina la categoría de referencia para el tipo de contraste seleccionado. Seleccione **Primera** para utilizar la primera categoría para el campo de entrada, -ordenado alfabéticamente, o seleccione **Última** para utilizar la última categoría. La categoría base predeterminada se aplica a las variables que están listadas en el área **Entradas categóricas**.

  **Nota:** Este campo no está disponible si el valor de contraste se Diferencia, Helmert, Repetido o Polinómico.

La estimación del efecto de cada campo sobre la respuesta global se calcula como un aumento o disminución en la verosimilitud de cada una de las otras categorías relativas a la categoría de referencia. Esto puede ayudarle a identificar los campos y valores con más posibilidades de producir una respuesta específica.

La categoría base se muestra en el resultado como 0,0. Esto se debe a que, al compararse consigo misma, produce un resultado vacío. El resto de categorías se muestran como ecuaciones relevantes para la categoría base. Consulte el tema "Detalles del nugget de modelo logístico" en la página 195 para obtener más información.

## Modelos multinomiales

Para los modelos multinomiales, están disponibles los siguientes métodos y opciones:

**Método.** Especifique el método que se va a utilizar para la creación del modelo de regresión logística.
- **Intro.** Éste es el método predeterminado que introduce directamente todos los términos en la ecuación. No se realiza ninguna selección de campos en la creación del modelo.
- **Por pasos.** El método de selección de campos Por pasos crea la ecuación por pasos, como su nombre indica. El modelo inicial es el más simple que puede haber, sin ningún término del modelo (excepto el constante) en la ecuación. En cada paso, se evalúan los términos que no se han añadido aún al modelo y si el mejor de dichos términos se suma de forma significativa a la eficacia predictiva del modelo, se añadirá a éste. Además, los términos que se encuentran actualmente en el modelo se vuelven a evaluar para determinar si se puede eliminar alguno de ellos sin que afecte al modelo de forma significativa. Si

es así, se eliminan. El proceso se repite y se añaden y/o eliminan otros términos. Cuando no se puedan añadir más términos para mejorar el modelo, y no se puedan eliminar más sin que le afecte, se creará el modelo final.

- **Adelante.** El método Adelante de la selección de campos se parece al método Por pasos en el que el modelo se crea por pasos. No obstante, con este método, el modelo inicial es el más simple y sólo se pueden añadir los términos y la constante al modelo. En cada paso, los términos que no están aún en el modelo se prueban en función de lo que puedan mejorarlo y el que resulte ser el mejor de esos términos es el que se añade al modelo. Cuando no se puedan añadir más términos, o el mejor candidato no produzca una mejora lo suficientemente grande en el modelo, se creará el modelo final.
- **Hacia atrás.** El método Hacia atrás es fundamentalmente lo contrario al método Adelante. Con este método, el modelo inicial contiene todos los términos como predictores y sólo se pueden eliminar los términos del modelo. Los términos del modelo que contribuyen poco al modelo se eliminan uno a uno hasta que no se puedan eliminar más sin que lo perjudiquen de forma significativa, dando lugar al modelo final.
- **Por pasos hacia atrás.** El método Por pasos hacia atrás es fundamentalmente lo contrario al método Por pasos. Con este método, el modelo inicial contiene todos los términos como predictores. En cada paso, se evalúan los términos del modelo y se eliminan los que no afecten al modelo de forma significativa. Además, los términos eliminados anteriormente se vuelven a evaluar para determinar si el mejor de dichos términos se añade de forma significativa a la eficacia predictiva del modelo. Si es así, se volverá a añadir al modelo. Cuando no se puedan añadir más términos para mejorar el modelo y no se puedan eliminar más sin que le afecte, se creará el modelo final.

**Nota:** Los métodos automáticos, incluidos Por pasos, Hacia adelante y Hacia atrás, son métodos de aprendizaje altamente adaptables y tienen una fuerte tendencia a sobreajustarse a los datos de entrenamiento. Cuando se utilicen estos métodos, es muy importante comprobar la validez del modelo resultante, bien con datos nuevos o con una muestra de comprobación reservada mediante el nodo Partición.

**Categoría base para el objetivo.** Especifica cómo se determina la categoría de referencia. Se utiliza como línea de base con la que se estiman las ecuaciones de regresión para todas las otras categorías del objetivo. Seleccione **Primera** para utilizar la primera categoría para el campo objetivo actual, ordenado alfabéticamente, o seleccione **Última** para utilizar la última categoría. Si lo prefiere, puede seleccionar **Especificar** para seleccionar una categoría específica y elegir el valor deseado de la lista. Se pueden definir los valores disponibles para cada campo en un nodo Tipo.

Normalmente se especifica la categoría en la que se está menos interesado como categoría base, por ejemplo, un producto líder con pérdidas. A continuación, se relaciona con la categoría base el resto de categorías de forma relativa para identificar la probabilidad de que estén en su propia categoría. Esto puede ayudarle a identificar los campos y valores con más posibilidades de producir una respuesta específica.

La categoría base se muestra en el resultado como 0,0. Esto se debe a que, al compararse consigo misma, produce un resultado vacío. El resto de categorías se muestran como ecuaciones relevantes para la categoría base. Consulte el tema “Detalles del nugget de modelo logístico” en la página 195 para obtener más información.

**Tipo de modelo.** Hay tres opciones para definir los términos del modelo. Los modelos **Efectos principales** sólo incluyen los campos de entrada de forma individual y no comprueban las interacciones (efectos multiplicativos) entre los campos de entrada. Los modelos del tipo **Factorial completo** incluyen todas las interacciones, así como los efectos principales de los campos de entrada. Los modelos factoriales completos están más capacitados para capturar relaciones complejas pero son mucho más difíciles de interpretar y tienen más posibilidades de sufrir sobreajuste. Debido a la posibilidad de que haya un gran número de combinaciones posibles, los métodos de selección automática de campos (métodos distintos de Introducir) se desactivarán para los modelos factoriales completos. Los modelos **Personalizados** sólo

incluyen los términos que se especifiquen (efectos principales e interacciones). Cuando seleccione esta opción, utilice la lista Términos del modelo para añadir términos al modelo o eliminarlos.

**Términos del modelo.** Al crear un modelo personalizado, deberá especificar explícitamente los términos del modelo. La lista muestra el conjunto actual de términos para el modelo. Los botones situados en el lado derecho de la lista Términos del modelo le permiten añadir y eliminar términos de modelo.

- Para añadir términos al modelo, pulse en el botón *Añadir nuevos términos del modelo*.
- Seleccione los términos deseados para eliminarlos y pulse en el botón *Eliminar los términos del modelo seleccionado*.

## Adición de términos a un modelo de regresión logística

Al solicitar un modelo de regresión logística personalizado, puede añadirle términos pulsando en el botón *Añadir nuevos términos del modelo* de la pestaña Modelo de regresión logística. Se abrirá un cuadro de diálogo Nuevos términos en el que podrá especificar los términos.

**Tipo de término que se va a añadir.** Hay varias formas de añadir términos al modelo, según la selección de los campos de entrada de la lista Campos disponibles.

- **Interacción sencilla.** Inserta el término que representa la interacción de todos los campos seleccionados.
- **Efectos principales.** Inserta un término de efectos principales (el propio campo) para cada campo de entrada seleccionado.
- **Todas las interacciones de dos factores.** Inserta un término de interacción de 2 factores (el producto de los campos de entrada) para cada posible par de campos de entrada seleccionados. Por ejemplo, si ha seleccionado los campos de entrada *A*, *B* y *C* en la lista Campos disponibles, este método insertará los términos *A * B*, *A * C* y *B * C*.
- **Todas las interacciones de tres factores.** Inserta un término de interacción de 3 factores (el producto de los campos de entrada) para cada posible combinación de campos de entrada seleccionados, tomando tres al mismo tiempo. Por ejemplo, si ha seleccionado los campos de entrada *A*, *B*, *C* y *D* en la lista Campos disponibles, este método insertará los términos *A * B * C*, *A * B * D*, *A * C * D* y *B * C * D*.
- **Todas las interacciones de cuatro factores.** Inserta un término de interacción de 4 factores (el producto de los campos de entrada) para cada posible combinación de campos de entrada seleccionados, tomando cuatro al mismo tiempo. Por ejemplo, si ha seleccionado los campos de entrada *A*, *B*, *C*, *D* y *E* en la lista Campos disponibles, este método insertará los términos *A * B * C * D*, *A * B * C * E*, *A * B * D * E*, *A * C * D * E* y *B * C * D * E*.

**Campos disponibles.** Muestra los campos de entrada disponibles que se van a utilizar en la construcción de términos del modelo.

**Presentación preliminar.** Muestra los términos que se añadirán al modelo si pulsa en **Insertar**, según los campos seleccionados y el tipo de término.

**Insertar.** Inserta los términos del modelo (según la selección actual de los campos y el tipo de término) y cierra el cuadro de diálogo.

## Opciones de experto para el nodo Logística

Las opciones de experto le permiten ajustar el proceso de entrenamiento, si tiene conocimientos detallados de regresión logística. Para acceder a las opciones de experto, en la pestaña Experto, establezca Modo como **Experto**.

**Escalas (sólo modelos multinomiales).** Puede especificar un valor para el escalamiento de la dispersión que será utilizado para corregir la estimación de la matriz de covarianzas de los parámetros. **Pearson** calcula el valor de escalamiento utilizando el estadístico chi-cuadrado de Pearson. **Desvianza** calcula el

valor de escalamiento utilizando el estadístico de la función de desvianza (chi-cuadrado de la razón de verosimilitud). También puede especificar su propio valor de escalamiento definido por el usuario. Debe ser un valor numérico positivo.

**Añadir todas las probabilidades.** Si esta opción se selecciona, se añadirán a cada registro procesado por el nodo las probabilidades para cada una de las categorías del campo de salida. Si no se selecciona esta opción, sólo se añadirá la probabilidad de la categoría predicha.

Por ejemplo, una tabla que contenga los resultados de un modelo multinomial con tres categorías incluirá cinco nuevas columnas. Una columna mostrará la probabilidad de que el resultado se prediga correctamente, la siguiente mostrará la probabilidad de que la predicción sea un acierto o un valor perdido, y las siguientes tres columnas mostrarán la probabilidad de que cada predicción de la categoría sea un acierto o un valor perdido. Consulte el tema "Nugget de modelo logístico" en la página 194 para obtener más información.

*Nota*: esta opción siempre está seleccionada para los modelos binomiales.

**Tolerancia para la singularidad.** Especifica la tolerancia utilizada en la comprobación de singularidades.

**Convergencia.** Estas opciones le permiten controlar los parámetros de la convergencia del modelo. Cuando se ejecuta el modelo, la configuración de la convergencia controla cuántas veces se ejecutan los distintos parámetros a través de éste para comprobar si se ajustan. Cuanta más veces se prueben los parámetros, más próximos estarán los resultados (es decir, los resultados convergirán). Consulte el tema "Opciones de convergencia de regresión logística" para obtener más información.

**Resultados.** Estas opciones le permiten solicitar estadísticos adicionales que aparecerán en el resultado avanzado del nugget de modelo construido por el nodo. Consulte el tema "Salida avanzada de regresión logística" en la página 193 para obtener más información.

**Método por pasos.** Estas opciones le permiten controlar los criterios para añadir y eliminar los campos con los métodos de estimación Por pasos, Adelante, Hacia atrás o Por pasos hacia atrás. (Si el método Introducir está seleccionado, el botón estará desactivado.) Consulte el tema "Opciones del método por pasos de regresión logística" en la página 193 para obtener más información.

## Opciones de convergencia de regresión logística

Puede establecer los parámetros de convergencia para la estimación del modelo de regresión logística.

**Iteraciones máximas.** Especifica el número máximo de iteraciones para la estimación del modelo.

**Máxima subdivisión por pasos.** La regresión logística utiliza la técnica de subdivisión por pasos para gestionar las complejidades en el proceso de estimación. En circunstancias normales, debe utilizar el valor predeterminado.

**Convergencia del logaritmo de la verosimilitud.** Las iteraciones se detendrán si el cambio relativo del logaritmo de la verosimilitud es menor que este valor. Este criterio no se aplica si el valor es igual a 0.

**Convergencia de los parámetros.** Las iteraciones se detendrán si el cambio absoluto o relativo de las estimaciones de los parámetros es menor que este valor. Este criterio no se aplica si el valor es igual a 0.

**Delta (sólo modelos multinomiales).** Puede especificar un valor entre 0 y 1 para añadirlo a cada casilla vacía (combinación de valores de campos de entrada y de salida). Esto puede ayudar al algoritmo de estimación a gestionar los datos en los que hay muchas combinaciones posibles de valores de los campos respecto al número de registros existente en los datos. El valor por omisión es 0.

## Salida avanzada de regresión logística

Seleccione el resultado opcional que desee mostrar en el resultado avanzado del nugget de modelo de regresión. Para ver el resultado avanzado, examine el nugget de modelo y pulse en la pestaña **Avanzado**. Consulte el tema "Resultado avanzado del nugget de modelo logístico" en la página 197 para obtener más información.

Opciones binomiales

Seleccione los tipos de resultados que se generarán para el modelo. Consulte el tema "Resultado avanzado del nugget de modelo logístico" en la página 197 para obtener más información.

**Mostrar.** Seleccione si desea mostrar los resultados a cada paso o esperar hasta que terminen todos los pasos.

**CI para exp(B).** Seleccione los intervalos de confianza para cada coeficiente (mostrado como Beta) de la expresión. Especifica el nivel del intervalo de confianza (el valor predeterminado es el 95%).

**Diagnóstico de residuos.** Solicita una tabla de diagnósticos por caso de los residuos.

- **Valores atípicos (desv. est.).** Muestra sólo los casos residuales para los que el valor absoluto estandarizado de la variable de la lista es como mínimo tan grande como el valor especificado. El valor por omisión es 2.
- **Todos los casos.** Incluye todos los casos en la tabla de diagnósticos por caso de los residuos.

  *Nota*: como esta opción enumera todos los registros de entrada, puede producir una tabla extraordinariamente grande en el informe, con una línea por cada registro.

**Punto de corte para la clasificación.** Esto le permite determinar el punto de corte para clasificar casos. Los casos con valores predichos que han sobrepasado el punto de corte para la clasificación se clasifican como positivos, mientras que aquéllos con valores predichos menores que el punto de corte se clasifican como negativos. Para cambiar el valor predeterminado, introduzca un valor entre 0,01 y 0,99.

Opciones multinomiales

Seleccione los tipos de resultados que se generarán para el modelo. Consulte el tema "Resultado avanzado del nugget de modelo logístico" en la página 197 para obtener más información.

*Nota*: si se selecciona la opción **Contrastes de razón de verosimilitud**, aumenta en gran medida el tiempo de procesamiento necesario para generar un modelo de regresión logística. Si la generación del modelo está tardando demasiado, puede desactivar esta opción o utilizar, en su lugar, los estadísticos Wald o de puntuación. Consulte el tema "Opciones del método por pasos de regresión logística" para obtener más información.

**Historial de iteraciones para cada.** Seleccione el intervalo de pasos para la impresión del estado de iteración en el resultado avanzado.

**Intervalo de confianza.** Intervalos de confianza para los coeficientes de las ecuaciones. Especifica el nivel del intervalo de confianza (el valor predeterminado es el 95%).

## Opciones del método por pasos de regresión logística

Estas opciones le permiten controlar los criterios para añadir y eliminar los campos con los métodos de estimación Por pasos, Adelante, Hacia atrás o Por pasos hacia atrás.

**Número de términos en el modelo (sólo modelos multinomiales).** Puede especificar el número mínimo de términos para los modelos Hacia atrás y Por pasos hacia atrás y el número máximo para los modelos Adelante y Por pasos. Si especifica un valor mínimo mayor que 0, el modelo incluirá dicho número de

términos, incluso cuando se habrían eliminado algunos de los términos basándose en los criterios estadísticos. La especificación del mínimo será ignorada en los modelos Adelante, Por pasos e Introducir. Si especifica un valor máximo, puede que se omitan algunos términos del modelo, incluso cuando habrían sido seleccionados basándose en los criterios estadísticos. La configuración **Especificar máximo** será ignorada en los modelos Hacia atrás, Por pasos hacia atrás e Introducir.

**Criterio de entrada (sólo modelos multinomiales).** Seleccione **Puntuación** para maximizar la velocidad de procesamiento. La opción **Razón de verosimilitud** puede proporcionar estimaciones algo más robustas, pero tarda más tiempo en realizar los cálculos. La configuración predeterminada es el uso del estadístico Puntuación.

**Criterio de exclusión.** Seleccione **Razón de verosimilitud** para un modelo más robusto. Si desea reducir el tiempo necesario para generar el modelo, puede intentar seleccionar **Wald**. Sin embargo, si tiene una separación completa o casi completa en los datos (que puede determinar con la pestaña Avanzado del nugget de modelo) el estadístico de Wald pasará a ser particularmente inestable y no se debería utilizar. La configuración predeterminada es el uso del estadístico razón de verosimilitud. Para los modelos binomiales existe la opción adicional **Condicional**. Esta opción permite una comprobación de eliminación en función de la probabilidad del estadístico de razón de verosimilitud basado en estimaciones de parámetros condicionales.

**Umbrales de significación para los criterios de RL.** Esta opción le permite especificar criterios de selección según la probabilidad estadística (el valor *p*) asociada a cada campo. Los campos se añadirán al modelo sólo si el valor *p* asociado es más pequeño que el valor **Entrada** y se eliminarán sólo si el valor *p* es mayor que el valor **Eliminación**. El valor **Entrada** debe ser menor que el valor **Eliminación**.

**Requisitos para la introducción o eliminación (sólo modelos multinomiales).** Para algunas aplicaciones, no es recomendable desde el punto de vista matemático añadir términos de interacción al modelo a no ser que éste también contenga los términos de orden inferior para los campos implicados en el término de interacción. Por ejemplo, no tendrá sentido incluir *A * B* en el modelo a no ser que *A* y *B* también se incluyan en el mismo. Estas opciones le permiten determinar cómo se gestionan estas dependencias durante la selección del término por pasos.

- **Jerarquía para efectos discretos.** Los efectos de orden superior (interacciones que implican más campos) se introducirán en el modelo sólo si ya están en el modelo todos los efectos de orden inferior (efectos principales o interacciones que implican menos campos) de los campos pertinentes y los efectos de orden inferior no se eliminarán si los efectos de orden superior que implican los mismos campos están en el modelo. Esta opción sólo se aplica a campos categóricos.
- **Jerarquía para todos los efectos.** Esta opción funciona de la misma manera que la opción previa, excepto que se aplica a todos los campos de entrada.
- **Contención para todos los efectos.** Los efectos pueden incluirse en el modelo sólo si todos los efectos que se encuentran en el efecto se incluyen también en el modelo. Esta opción es similar a la de **Jerarquía para todos los efectos**, excepto que los campos continuos se tratan de forma diferente. Para que un efecto contenga otro efecto, el que está contenido (orden inferior) debe incluir *todos* los campos continuos implicados en el contenedor (orden superior) y los campos categóricos del que está contenido deben ser un subconjunto de los que están en el efecto contenedor. Por ejemplo, si *A* y *B* son campos categóricos y *X* es un campo continuo, el término *A * B * X* contendrá los términos *A * X* y *B * X*.
- **Ninguno.** No hay ninguna relación forzosa. Los términos se añaden al modelo y se eliminan de forma independiente.

## Nugget de modelo logístico

Un nugget de modelo logístico representa la ecuación calculada por un nodo Logística. Contiene toda la información capturada por el modelo de regresión logística, así como información acerca del rendimiento y la estructura del modelo. Otros modelos como Oracle SVM también pueden generar este tipo de ecuación.

Cuando se ejecuta una ruta que contiene un nugget de modelo logístico, el nodo añade dos nuevos campos que contienen la predicción del modelo y la probabilidad asociada. Los nombres de los nuevos campos se derivan del nombre del campo de salida que se está prediciendo, con el prefijo *$L-* para la categoría predicha y *$LP-* para la probabilidad asociada. Por ejemplo, para un campo de salida llamado *colorpref*, los nuevos campos se llamarían *$L-colorpref* y *$LP-colorpref*. Además, si ha seleccionado la opción **Añadir todas las probabilidades** en el nodo Logística, se añadirá un campo adicional para cada categoría del campo de salida, que contiene la probabilidad perteneciente a la categoría correspondiente de cada registro. Los nombres de estos campos adicionales se asignan en función de los valores del campo de salida, con el prefijo *$LP-*. Por ejemplo, si los valores legales de *colorpref* son *Rojo*, *Verde*y *Azul*, se añadirán tres nuevos campos: *$LP-Rojo*, *$LP-Verde*y *$LP-Azul*.

**Generación de un nodo Filtrar.** El menú Generar permite crear un nuevo nodo Filtrar para pasar los campos de entrada en función de los resultados del modelo. El nodo generado filtrará los campos que se eliminan del modelo debido a la multicolinealidad y los campos que no se utilizan en el modelo.

## Detalles del nugget de modelo logístico

Para los modelos multinomiales, la pestaña Modelo de un nugget de modelo logístico tiene una visualización dividida con ecuaciones de modelo en el panel izquierdo y la importancia del predictor en el derecho. Para los modelos binomiales, la pestaña sólo muestra la importancia del predictor. Consulte el tema "Importancia del predictor" en la página 44 para obtener más información.

Ecuaciones de modelo

Para modelos multinomiales, el panel izquierdo muestra las ecuaciones reales calculadas para el modelo de regresión logística. Hay una ecuación por cada categoría en el campo objetivo, excepto la categoría de línea base. Las ecuaciones se muestran en un formato de árbol. Otros tipos de modelos como Oracle SVM también pueden generar este tipo de ecuación.

**Ecuación para.** Muestra las ecuaciones de regresión utilizadas para derivar las probabilidades de la categoría objetivo, dado un conjunto de valores de predicción. La última categoría del campo objetivo se considera la **categoría de línea de base**; las ecuaciones mostradas ofrecen los logaritmos de probabilidades para el resto de categorías de destino relativas a la categoría de línea de base para un conjunto de valores de predicción concretos. La probabilidad predicha para cada categoría del patrón de predicción dado se deriva de estos valores de logaritmo de probabilidades.

¿Cómo se calculan las probabilidades?

Cada ecuación calcula el logaritmo de probabilidades de una categoría objetivo particular, relativa a la categoría de línea de base. El **logaritmo de probabilidades**, también llamado **logit**, es el cociente de la probabilidad de la categoría objetivo especificada a la de la categoría de la línea base, con la función de logaritmo natural aplicada al resultado. En el caso de la categoría de línea de base, la probabilidad de la categoría relativa a sí misma es de 1,0 y, por lo tanto, el logaritmo de probabilidades es 0. Esto se puede interpretar como una ecuación implícita de la categoría de línea de base donde todos los coeficientes son 0.

Para derivar la probabilidad a partir de los logaritmos de probabilidades de una categoría objetivo particular, tome el valor de logit calculado por la ecuación para esa categoría y aplique la siguiente fórmula:

$$P(grupo_i) = \exp(g_i) \;/\; \Sigma_k \exp(g_k)$$

donde $g$ es el logaritmo de probabilidades calculado, $i$ es el índice de categoría y $k$ varía entre 1 y el número de categorías objetivo.

Importancia del predictor

Opcionalmente, en la pestaña Modelo también se puede mostrar un gráfico que indique la importancia relativa de cada predictor cuando se calcule el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Tenga en cuenta que este gráfico sólo está disponible si se ha seleccionado **Calcular importancia de predictor** en la pestaña Analizar antes de generar el modelo. Consulte el tema "Importancia del predictor" en la página 44 para obtener más información.

*Nota:* la importancia del predictor puede requerir más tiempo de cálculo para la regresión logística que para otros tipos de modelos y no está seleccionada en la pestaña Analizar de forma predeterminada. Si selecciona esta opción se reducirá el rendimiento, especialmente con conjuntos de datos más grandes.

## Resumen de nugget de modelo logístico

El resumen de un modelo de regresión logística muestra los campos y ajustes utilizados para generar el modelo. Además, si ha ejecutado un nodo Análisis conectado a este nodo de modelado, la información de dicho análisis también se mostrará en esta sección. Para obtener información general sobre cómo utilizar el explorador de modelos, consulte "Examen de nuggets de modelo" en la página 42.

## Configuración del nugget de modelo logístico

La pestaña Configuración de un nugget de modelo logístico especifica opciones para confianzas, probabilidades, puntuaciones de propensión y generación de SQL durante la puntuación de modelos. Esta pestaña sólo está disponible después de añadir el nugget de modelo a una ruta y muestra diferentes opciones dependiendo del tipo de modelo y objetivo.

### Modelos multinomiales

Para modelos multinomiales, están disponibles las opciones siguientes:

**Calcular confianzas** Especifica si las confianzas se van a calcular durante la puntuación.

**Calcular puntuaciones de propensión en bruto (sólo objetivos de marca)** En el caso de modelos sólo con objetivos de marca, puede solicitar puntuaciones de propensión en bruto que indican la probabilidad del resultado `true` especificado para el campo objetivo. Éstas se añaden a los valores estándar de predicción y confianza. Las puntuaciones ajustadas de propensión no están disponibles. Consulte el tema "Opciones de análisis del nodo de modelado" en la página 35 para obtener más información.

**Añadir todas las probabilidades** Especifica si se añaden las probabilidades de cada categoría del campo de resultados a cada registro procesado por el nodo. Si no se selecciona esta opción, sólo se añadirá la probabilidad de la categoría predicha. Por ejemplo, para un objetivo nominal con tres categorías, el resultado de puntuación incluirá una columna para cada una de las tres categorías, además de una cuarta columna indicando la probabilidad para cualquier categoría que se prediga. Por ejemplo, si las probabilidades de las categorías *Rojo*, *Verde* y *Azul* son 0,6, 0,3 y 0,1 respectivamente, la categoría predicha sería *Rojo*, con una probabilidad de 0,6.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

- **Puntuar convirtiendo a SQL nativo** Si selecciona esta opción, se genera SQL para puntuar el modelo dentro de la base de datos.

  **Nota:** aunque esta opción puede proporcionar resultados más rápidos, el tamaño y la complejidad del SQL nativo aumenta a medida que lo hace la complejidad del modelo.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

**Nota:** Para modelos multinomiales, la generación de SQL no está disponible si se ha seleccionado **Añadir todas las probabilidades**, o para modelos con objetivos nominales, si se ha seleccionado **Calcular confianzas**. La generación de SQL con cálculos de confianza sólo se admite para modelos multinomiales con objetivos de marca. La generación de SQL no se encuentra disponible para modelos binomiales.

### Modelos binomiales

Para modelos binomiales, las confianzas y probabilidades siempre están habilitadas y los ajustes que le permitirían desactivar dichas opciones no están disponibles. La generación de SQL no se encuentra disponible para modelos binomiales. El único ajuste que se puede cambiar para modelos binomiales es la capacidad de calcular puntuaciones de propensión en bruto. Como se ha indicado anteriormente para los modelos multinomiales, esto sólo es aplicable en modelos con objetivos de marca. Consulte el tema "Opciones de análisis del nodo de modelado" en la página 35 para obtener más información.

## Resultado avanzado del nugget de modelo logístico

La salida avanzada para la regresión logística (también denominada **regresión nominal**) ofrece información detallada sobre el modelo estimado y su rendimiento. La mayoría de la información contenida en la salida avanzada es bastante técnica, por lo que se precisan amplios conocimientos sobre análisis de regresión logística para interpretar correctamente estos resultados.

**Advertencias.** Muestra advertencias o problemas potenciales relacionados con los resultados.

**Resumen del procesamiento de los casos.** Muestra el número de registros procesados, descompuestos por cada campo simbólico del modelo.

**Resumen de pasos (opcional).** Muestra los efectos añadidos o eliminados en cada paso de la creación del modelo cuando se utiliza la selección de campos automática.

*Nota*: solo se muestra para los métodos Por pasos, Hacia atrás o Por pasos hacia atrás.

**Historial de iteraciones (opcional).** Muestra el historial de iteraciones de las estimaciones de los parámetros para cada *n* iteraciones que empiecen por las estimaciones iniciales, donde *n* es el valor del intervalo de impresión. El valor predeterminado es que se impriman todas las iteraciones (n=1).

**Información de ajuste de los modelos (modelos multinomiales).** Muestra los contrastes sobre razones de verosimilitud de su modelo (final) en comparación con uno en el que todos los coeficientes de parámetros son 0 (sólo interceptación).

**Clasificación (opcional).** Muestra la matriz de los valores de campo de salida reales y predichos con porcentajes.

**Estadísticos de bondad de ajuste de chi-cuadrado (opcional).** Muestra los estadísticos de Pearson y de chi-cuadrado de razón de verosimilitud. Estos estadísticos comprueban el ajuste global del modelo en los datos de entrenamiento.

**Bondad de ajuste de Hosmer-Lemeshow (opcional).** Muestra los resultados de agrupar casos en deciles de riesgo y compara la probabilidad observada con la probabilidad esperada dentro de cada decil. Este

estadístico de bondad de ajuste es más robusto que los estadísticos de bondad de ajuste tradicionales que se utilizan en los modelos multinomiales, especialmente en modelos con covariables continuas y en estudios con tamaños de muestra pequeños.

**Pseudo R cuadrado (opcional).** Muestra las medidas de *R* cuadrado de Cox y Snell, Nagelkerke y McFadden para el ajuste del modelo. Estos estadísticos son de alguna forma análogos al estadístico de *R*-cuadrado en la regresión lineal.

**Medidas de monotonía (opcional).** Muestra el número de pares concordantes, pares discordantes y empates en los datos, así como el porcentaje del número total de pares que representa cada uno. La D de Somers, la gamma de Goodman y Kruskal, la tau-a de Kendall y el índice de concordancia C también se muestran en esta tabla.

**Criterios de información (opcional).** Muestra el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC) de Schwarz.

**Contrastes sobre razón de verosimilitud (opcional).** Muestra los estadísticos comprobando si los coeficientes de los efectos del modelo son estadísticamente diferentes de 0. Los campos de entrada importantes son los que tienen niveles de significación muy pequeños en el resultado (con la etiqueta *Sig.*).

**Estimaciones de los parámetros (opcional).** Muestra estimaciones de los coeficientes de ecuación, comprobaciones de dichos coeficientes, razones de las ventajas derivadas de los coeficientes, con etiqueta *Exp(B)*, e intervalos de confianza para las razones de las ventajas.

**Matriz de covarianzas/correlaciones asintóticas (opcional).** Muestra las covarianzas y/o correlaciones asintóticas de las estimaciones de los coeficientes.

**Frecuencias observadas y predichas (opcional).** En cada patrón de covariables, muestra las frecuencias observadas y predichas para cada valor de campo de salida. Esta tabla puede ser bastante grande, especialmente para modelos con campos de entrada numéricos. Si la tabla resultante es demasiado grande para ser práctica, se omite y se muestra una advertencia.

## Nodo PCA/Factorial

El nodo PCA/Factorial proporciona técnicas eficaces de reducción de datos para reducir la complejidad de los datos. Se indican dos métodos similares pero distintos.

- **Análisis de componentes principales (PCA)** encuentra las combinaciones lineales de los campos de entrada que mejor realizan la tarea de capturar la varianza disponible en la totalidad del conjunto de campos, de manera que los componentes son ortogonales (perpendiculares) unos de otros. PCA se centra en todas las varianzas, incluyendo tanto las varianzas comunes como las exclusivas. PCA se centra en todas las varianzas, incluidas las compartidas y las exclusivas.
- **Análisis factorial** intenta identificar conceptos subyacentes o **factores** que expliquen el patrón de correlaciones dentro de un conjunto de campos observados. El análisis factorial sólo se centra en las varianzas compartidas. La varianza que es exclusiva a campos específicos no se tiene en cuenta a la hora de estimar el modelo. El nodo PCA/Factorial proporciona varios métodos de análisis factorial.

Para los dos métodos, el objetivo es encontrar un número pequeño de campos derivados que resuman de forma eficaz la información del conjunto original de campos.

**Requisitos.** Sólo se pueden utilizar campos numéricos en un modelo PCA-factorial. Para estimar un análisis factorial o PCA, son necesarios uno o más campos con el rol definido a campos de *Entrada*. Los campos el rol establecido a *Objetivo*, *Ambos* o *Ninguno* son ignorados, al igual que los campos no numéricos.

**Puntos fuertes.** Los análisis factorial y PCA pueden reducir de forma eficaz la complejidad de los datos sin llegar a sacrificar una parte sustancial del contenido de información. Estas técnicas pueden ayudarle a crear modelos más robustos que realicen ejecuciones de forma más rápida que con los campos de entrada iniciales.

## Opciones de modelo para el nodo PCA/Factorial

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Método de extracción.** Especifica el método que se va a utilizar para la reducción de datos.

- **Componentes principales.** Método predeterminado que utiliza PCA para encontrar los componentes que resumen los campos de entrada.
- **Mínimos cuadrados no ponderados.** Este método de análisis factorial busca el conjunto de factores que mejor reproducen el patrón de relaciones (correlaciones) entre los campos de entrada.
- **Mínimos cuadrados generalizados.** Este método de análisis factorial es similar al de mínimos cuadrados no ponderados, con la diferencia de que utiliza una ponderación para restar importancia a los campos con gran cantidad de varianza exclusiva (no compartidas).
- **Número máximo de verosimilitudes.** Este método de análisis factorial genera las ecuaciones factoriales que pueden haber dado lugar, con mayor probabilidad, al patrón observado de relaciones (correlaciones) entre los campos de entrada, basándose en ciertos supuestos sobre la forma de dichas relaciones. Específicamente, el método supone que los datos de entrenamiento siguen una distribución normal multivariante.
- **Factorización de ejes principales.** Este método de análisis factorial es muy similar al de componentes principales, con la diferencia de que se centra sólo en la varianza compartida.
- **Factorización alfa.** Este método de análisis factorial considera que los campos del análisis son una muestra del universo de campos de entrada potenciales. Maximiza la fiabilidad estadística de los factores.
- **Factorización imagen.** Este método de análisis factorial utiliza la estimación de los datos para aislar la varianza común y encontrar los factores que la describan.

## Opciones de experto para el nodo PCA/Factorial

Las opciones de experto le permiten ajustar el proceso de entrenamiento, si tiene conocimientos sobre análisis factorial y PCA. Para acceder a las opciones de experto, en la pestaña Experto, establezca Modo como **Experto**.

**Valores perdidos.** De forma predeterminada, IBM SPSS Modeler utiliza sólo los registros que dispongan de valores válidos en todos los campos utilizados en el modelo (Esto se denomina a veces **eliminación según lista** de los valores perdidos). Si tiene muchos datos perdidos, descubrirá que este método elimina muchos registros, dejándole sin los datos suficientes para generar un buen modelo. En estos casos, puede anular la selección de la opción **Sólo utilizar registros completos**. IBM SPSS Modeler intenta utilizar tanta información como sea posible para estimar el modelo, incluidos los registros en los que algunos campos tienen valores perdidos. (Esto se denomina a veces **eliminación por pareja** de los valores perdidos.) No obstante, en algunas situaciones, el uso de registros incompletos de esta forma puede dar lugar a problemas computacionales a la hora de calcular el modelo.

**Campos.** Especifica si se debe utilizar la matriz de correlaciones (valor predeterminado) o la matriz de covarianzas de los campos de entrada para estimar el modelo.

**Número máximo de iteraciones para la convergencia.** Especifica el número máximo de iteraciones para la estimación del modelo.

**Extraer factores.** Hay dos formas de seleccionar el número de factores que se deben extraer de los campos de entrada.

- **Autovalores mayores que.** Esta opción retendrá todos los factores o componentes con autovalores mayores que el criterio especificado. Los **autovalores** miden la capacidad de cada factor o componente para resumir la varianza disponible en el conjunto de los campos de entrada. El modelo retendrá todos los factores o componentes con autovalores mayores que el valor especificado cuando se utilice la matriz de correlaciones. Al utilizar la matriz de covarianzas, el criterio corresponde al número de veces que debe ser mayor que el autovalor promedio. Este escalamiento consigue que la opción tenga un significado similar en los dos tipos de matriz.
- **Número máximo.** Esta opción retendrá el número especificado de factores o componentes en orden descendente de autovalores. Es decir, los factores o componentes que corresponden a los *n* autovalores más altos están retenidos, donde *n* es el criterio especificado. El criterio de extracción predeterminado es de cinco factores/componentes.

**Formato de la matriz de componentes/factores.** Estas opciones controlan el formato de la matriz de factores (o matriz de componentes para los modelos PCA).

- **Ordenar valores.** Si esta opción está seleccionada, se ordenarán numéricamente las saturaciones factoriales en los resultados del modelo.
- **Ocultar valores por debajo de.** Si esta opción está seleccionada, las puntuaciones por debajo del umbral especificado se ocultarán en la matriz para permitir apreciar con mayor facilidad el patrón existente en la matriz.

**Rotación.** Estas opciones le permiten controlar el método de rotación para el modelo. Consulte el tema "Opciones de rotación para el nodo PCA/Factorial" para obtener más información.

## Opciones de rotación para el nodo PCA/Factorial

En muchos casos, la rotación matemática del conjunto de factores retenidos puede aumentar la utilidad y sobre todo, la interpretabilidad. Seleccione un método de rotación:

- **Sin rotación.** Opción predeterminada. No se utiliza ninguna rotación.
- **Varimax.** Método de rotación ortogonal que minimiza el número de campos con altas cargas en cada factor. Simplifica la interpretación de los factores.
- **Oblimin directa.** Método para rotación oblicua (no ortogonal). Cuando **Delta** sea igual a 0 (valor predeterminado), las soluciones serán oblicuas. A medida que delta se va haciendo más negativo, los factores son menos oblicuos. Para anular el valor predeterminado 0 para delta, introduzca un número menor o igual que 0,8.
- **Quartimax.** Método ortogonal que minimiza el número de factores necesarios para explicar los campos. Simplifica la interpretación de los campos observados.
- **Equamax.** Método de rotación que es una combinación del método Varimax, que simplifica los factores, y el método Quartimax, que simplifica los campos. Se minimiza el número de campos que saturan alto en un factor y el número de factores necesarios para explicar un campo.
- **Promax.** Rotación oblicua, que permite que los factores estén correlacionados. Se puede calcular de forma más rápida que una rotación oblimin directa, por lo que resulta útil para conjuntos grandes de datos. **Kappa** controla la oblicuidad de la solución (el punto hasta el cual los factores pueden correlacionar).

## Nugget de modelo PCA/Factorial

Un nugget de modelo PCA/Factorial representa los modelos de análisis factorial y de análisis de componentes principales (PCA, del inglés "principal component analysis") creados por un nodo PCA/Factorial. Contienen toda la información capturada por el modelo entrenado, así como información acerca del rendimiento y las características del modelo.

Cuando se ejecuta una ruta que contiene un modelo de ecuación factorial, el nodo añade un nuevo campo para cada factor o componente del modelo. Los nuevos nombres de campos se derivan del nombre del modelo, con el prefijo *$F-* y el sufijo *-n*, donde *n* es el número del factor o componente. Por ejemplo, si el modelo se denomina *Factor* y contiene tres factores, los nuevos campos se llamarían *$F-Factor-1*, *$F-Factor-2* y *$F-Factor-3*.

Para obtener una impresión más acertada de qué ha codificado el modelo, puede realizar un análisis hacia abajo más profundo. Una forma útil de consultar el resultado del modelo factorial es ver las correlaciones entre los campos de factores y entradas mediante un nodo Estadísticos. De este modo se detectan los campos de entrada que se cargan con exceso y los factores en que lo hacen y se puede descubrir si los factores tienen un significado o una interpretación subyacente.

También se puede evaluar el modelo factorial mediante la información disponible en la salida avanzada. Para ver la salida avanzada, pulse en la pestaña **Avanzado** del explorador de nugget de modelo. La salida avanzada contiene mucha información detallada y está destinada a usuarios con amplios conocimiento sobre análisis factorial o PCA. Consulte el tema “Resultado avanzado del nugget de modelo PCA/Factorial” para obtener más información.

## Ecuaciones de nugget de modelo PCA/Factorial

La pestaña Modelo para un nugget de modelo Factorial muestra la ecuación factorial para cada factor. Las puntuaciones factoriales o del componente se calculan multiplicando cada valor de campo de entrada por su coeficiente y, a continuación, sumando los resultados.

## Resumen de nugget de modelo PCA/Factorial

La pestaña Resumen de un modelo factorial muestra el número de factores que se retienen en el modelo PCA/factorial, junto con información adicional de los campos y ajustes utilizados para generar el modelo. Consulte el tema “Examen de nuggets de modelo” en la página 42 para obtener más información.

## Resultado avanzado del nugget de modelo PCA/Factorial

La salida avanzada para el análisis factorial ofrece información detallada sobre el modelo estimado y su rendimiento. La mayoría de la información contenida en la salida avanzada es bastante técnica y es necesario tener amplios conocimientos sobre análisis factorial para interpretar correctamente estos resultados.

**Advertencias.** Muestra advertencias o problemas potenciales relacionados con los resultados.

**Comunalidades.** Muestra la proporción de cada varianza de campo que los factores o componentes tienen en cuenta. *Inicial* otorga las comunalidades iniciales con el conjunto completo de factores (el modelo comienza con tantos factores como campos de entrada) y *Extracción* proporciona las comunalidades basadas en el conjunto de factores retenido.

**Varianza total explicada.** Muestra la varianza total explicada por los factores en el modelo. *Autovalores iniciales* muestra la varianza explicada por todo el conjunto de factores iniciales. *Sumas de extracción de las saturaciones al cuadrado* muestra la varianza explicada por los factores retenidos en el modelo. *Sumas de rotación de las saturaciones al cuadrado* muestra la varianza explicada por los factores rotados. Recuerde que, en el caso de las rotaciones oblicuas, las *sumas de rotación de las saturaciones al cuadrado* muestran sólo las sumas de saturaciones al cuadrado, pero no muestran los porcentajes de la varianza.

**Matriz de factor (o componente).** Muestra las correlaciones entre los campos de entrada y los factores sin rotar.

**Matriz de factor rotado (o componente).** Muestra las correlaciones entre los campos de entrada y los factores rotados para las rotaciones ortogonales.

**Matriz de patrón.** Muestra las correlaciones parciales entre los campos de entrada y los factores rotados para las rotaciones oblicuas.

**Matriz de estructura.** Muestra las correlaciones simples entre los campos de entrada y los factores rotados para las rotaciones oblicuas.

**Matriz de correlación factorial.** Muestra las correlaciones entre los factores para las rotaciones oblicuas.

## Nodo Discriminante

El análisis discriminante genera un modelo predictivo para la pertenencia a un grupo. El modelo se compone de una función discriminante (o, para más de dos grupos, un conjunto de funciones discriminantes) basada en combinaciones lineales de las variables de predictor que ofrecen la mejor discriminación entre los grupos. Las funciones se generan a partir de una muestra de casos cuya pertenencia al grupo se conoce; las funciones se pueden aplicar entonces a nuevos casos con mediciones para las variables de predictor pero con una pertenencia a grupo desconocida.

**Ejemplo.** Una empresa de telecomunicaciones puede usar el análisis discriminante para clasificar a los clientes en grupos basados en datos de uso. Esto permite puntuar a los clientes potenciales y dirigirse a los que tienen más posibilidades de estar incluidos en los grupos más valiosos.

**Requisitos.** Son necesarios uno o más campos de entrada y exactamente un campo de objetivo. El objetivo debe ser un campo categórico (con un nivel de medición de *Marca* o *Nominal*) con un almacenamiento de cadena o entero. (Si es necesario, es posible convertir el almacenamiento mediante un nodo Rellenar o Derivar. ) Se ignorarán los campos establecidos en *Ambos* o *Ninguno*. Los tipos de los campos utilizados en el modelo deben estar completamente instanciados.

**Puntos fuertes.** Análisis discriminante y Regresión logística son modelos de clasificación adecuados. Sin embargo, el análisis discriminante realiza más supuestos sobre los campos de entrada, por ejemplo que suelen distribuirse y deben ser continuos, y ofrecen mejores resultados si se cumplen esos requisitos, especialmente si el tamaño de muestra es pequeño.

## Opciones de modelo del nodo Discriminante

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Crear modelos divididos.** Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como campos de división. Consulte el tema “Generación de modelos divididos” en la página 28 para obtener más información.

**Método.** Éstas son las opciones disponibles para introducir predictores en el modelo:

- **Intro.** Éste es el método predeterminado que introduce directamente todos los términos en la ecuación. No se añaden los términos que no aumentan de forma significativa el poder predictivo del modelo.

- **Por pasos.** El modelo inicial es el más simple que puede haber, sin ningún término del modelo (excepto el constante) en la ecuación. En cada paso, se evalúan los términos que no se han añadido aún al modelo y si el mejor de dichos términos se suma de forma significativa a la eficacia predictiva del modelo, se añadirá a éste.

*Nota*: El método Por pasos tiene una fuerte tendencia a sobreajustarse a los datos de entrenamiento. Cuando se utilicen estos métodos, es muy importante comprobar la validez del modelo resultante con una muestra de comprobación reservada o datos nuevos.

## Opciones de experto del nodo Discriminante

Las opciones de experto le permiten ajustar el proceso de entrenamiento si tiene conocimientos sobre análisis discriminante. Para acceder a las opciones de expero, establezca **Modo** en **Experto** en la pestaña Experto.

**Probabilidades previas** Esta opción determina si se corrigen los coeficientes de clasificación teniendo en cuenta la información previa sobre la pertenencia a los grupos.

- **Todos los grupos iguales.** Se asumen probabilidades previas iguales para todos los grupos; esto no tiene efecto en los coeficientes.
- **Calcular según tamaños de grupos.** Los tamaños de grupo observados en la muestra determinan las probabilidades previas de la pertenencia a grupo. Por ejemplo, el 50% de las observaciones incluidas en el análisis entran en el primer grupo, el 25% en el segundo y el 25% en el tercero, los coeficientes de clasificación se ajustan para aumentar la verosimilitud de pertenencia en el primer grupo relativa a los otros dos.

**Usar matriz de covarianzas.** Puede elegir clasificar casos utilizando una matriz de covarianza intra-grupos o una matriz de covarianzas de los grupos separados.

- *Intra-grupos*. Se utiliza la matriz de covarianza intra-grupos combinada para clasificar los casos.
- *Grupos separados*. Para la clasificación se utilizan las matrices de covarianza de los grupos separados. Dado que la clasificación se basa en las funciones discriminantes y no en las variables originales, esta opción no siempre es equivalente a la discriminación cuadrática.

**Resultados.** Estas opciones le permiten solicitar estadísticos adicionales que aparecerán en el resultado avanzado del nugget de modelo construido por el nodo. Consulte el tema “Opciones de resultados del nodo Discriminante” para obtener más información.

**Método por pasos.** Estas opciones le permiten controlar los criterios para añadir y eliminar los campos con el método de estimación Por pasos. (Si el método Introducir está seleccionado, el botón estará desactivado.) Consulte el tema “Opciones del método por pasos del nodo Discriminante” en la página 204 para obtener más información.

## Opciones de resultados del nodo Discriminante

Seleccione el resultado opcional que desee mostrar en el resultado avanzado del nugget de modelo de regresión logística. Para ver el resultado avanzado, examine el nugget de modelo y pulse en la pestaña **Avanzado**. Consulte el tema “Resultados avanzados del nugget de modelo Discriminante” en la página 205 para obtener más información.

**Descriptivos.** Las opciones disponibles son medias (incluyendo las desviaciones estándar), ANOVAs univariadas y prueba *M* de Box.

- *Medias*. Muestra la media y desviación estándar totales y las medias y desviaciones estándar de grupo, para las variables independientes.
- *ANOVAs univariados*. Realiza un análisis de varianza de un factor sobre la igualdad de las medias de grupo para cada variable independiente.

- *M de Box*. Contraste sobre la igualdad de las matrices de covarianza de los grupos. Para tamaños de muestras suficientemente grandes, un valor de p no significativo quiere decir que no hay suficiente evidencia de que las varianzas sean diferentes. Esta prueba es sensible a las desviaciones de la normalidad multivariada.

**Coeficientes de función.** Las opciones disponibles son coeficientes de clasificación de Fisher y coeficientes no estandarizados.

- *De Fisher*. Muestra los coeficientes de la función de clasificación de Fisher que pueden utilizarse directamente para la clasificación. Se obtiene un conjunto de coeficientes para cada grupo, y se asigna un caso al grupo para el que tiene una mayor puntuación discriminante (valor de función de clasificación).
- *No tipificados*. Muestra los coeficientes de la función discriminante sin estandarizar.

**Matrices.** Las matrices disponibles de coeficientes para variables independientes son la matriz de correlaciones intra-grupos, la matriz de covarianza intra-grupos, la matriz de covarianzas de los grupos separados y matriz de covarianzas total.

- *Correlación intra-grupos*. Muestra la matriz de correlaciones intra-grupos combinada, que se obtiene de promediar las matrices de covarianza individuales para todos los grupos antes de calcular las correlaciones.
- *Covarianza intra-grupos*. Muestra la matriz de covarianza intra-grupos combinada, la cual puede diferir de la matriz de covarianza total. La matriz se obtiene de promediar, para todos los grupos, las matrices de covarianza individuales.
- *Covarianza de grupos separados*. Muestra las matrices de covarianza de cada grupo por separado.
- *Covarianza total*. Muestra la matriz de covarianza para todos los casos, como si fueran una única muestra.

**Clasificación.** El resultado siguiente pertenece a los resultados de clasificación.

- *Resultados para cada caso*. Se muestran, para cada caso, los códigos del grupo real de pertenencia, el grupo pronosticado, las probabilidades posteriores y las puntuaciones discriminantes.
- *Tabla de resumen*. Número de casos correcta e incorrectamente asignados a cada uno de los grupos, basándose en el análisis discriminante. En ocasiones se denomina "Matriz de Confusión".
- *Clasificación dejando uno fuera*. Se clasifica cada caso del análisis mediante la función derivada de todos los casos, excepto el propio caso. También se conoce como método U.
- *Mapa territorial*. Gráfico de las fronteras utilizadas para clasificar los casos en grupos a partir de los valores en las funciones. Los números corresponden a los grupos en los que se clasifican los casos. La media de cada grupo se indica mediante un asterisco situado dentro de sus fronteras. No se mostrará el mapa si sólo hay una función discriminante.
- *Grupos combinados*. Crea un diagrama de dispersión, con todos los grupos, de los valores en las dos primeras funciones discriminantes. Si sólo hay una función, en su lugar se muestra un histograma.
- *Grupos separados*. Crea diagramas de dispersión, de los grupos por separado, para los valores en las dos primeras funciones discriminantes. Si sólo hay una función, en su lugar se muestra un histograma.

**Por pasos. Resumen de pasos** visualiza estadísticas para todas las variables después de cada paso; **F para distancias por parejas** visualiza una matriz de razones *F* por parejas para cada pareja de grupos. Las razones *F* se pueden usar para comprobaciones de significación de las distancias de Mahalanobis entre grupos.

## Opciones del método por pasos del nodo Discriminante

**Método.** Seleccione el estadístico que se va usar para introducir o eliminar variables nuevas. Las alternativas disponibles son lambda de Wilks, varianza no explicada, distancia de Mahalanobis, razón *F* más pequeña y *V* de Rao. Con la *V* de Rao, se puede especificar el aumento mínimo en *V* para una variable que se vaya a introducir.

- *Lambda de Wilks*. Método para la selección de variables por pasos del análisis discriminante que selecciona las variables para su introducción en la ecuación basándose en cuánto contribuyen a disminuir la lambda de Wilks. En cada paso se introduce la variable que minimiza la lambda de Wilks global.
- *Varianza no explicada*. En cada paso se introduce la variable que minimiza la suma de la variación no explicada entre los grupos.
- *Distancia de Mahalanobis*. Medida de cuánto difieren del promedio para todos los casos los valores en las variables independientes de un caso dado. Una distancia de Mahalanobis grande identifica un caso que tenga valores extremos en una o más de las variables independientes.
- *Razón F más pequeña*. Método para la selección de variables en los análisis por pasos que se basa en maximizar la razón F, calculada a partir de la distancia de Mahalanobis entre los grupos.
- *V de Rao*. Medida de las diferencias entre las medias de los grupos. También se denomina la traza de Lawley-Hotelling. En cada paso, se incluye la variable que maximiza el incremento de la V de Rao. Después de seleccionar esta opción, introduzca el valor mínimo que debe tener una variable para poder incluirse en el análisis.

**Criterios** Las alternativas disponibles son **Utilizar el valor F** y **Utilizar probabilidad de F**. Entre valores para entrar y eliminar variables.

- *Utilizar el valor F*. Una variable se introduce en el modelo si su valor de F es mayor que el valor de entrada, y se elimina si su valor de F es menor que el valor de Eliminación. La entrada debe ser mayor que la eliminación y ambos valores deben ser positivos. Para introducir más variables en el modelo, disminuya el valor de entrada. Para eliminar más variables del modelo, eleve el valor de eliminación.
- *Utilizar probabilidad de F*. Una variable se introduce en el modelo si el nivel de significación de su valor de F es menor que el valor de entrada, y se elimina si el nivel de significación de su valor de F es mayor que el valor de Eliminación. La entrada debe ser menor que la eliminación y ambos valores deben ser positivos. Para introducir más variables en el modelo, aumente el valor de entrada. Para eliminar más variables del modelo, disminuya el valor de eliminación.

## Nugget de modelo Discriminante

Los nugget de modelo Discriminante representan las ecuaciones estimadas por los nodos Discriminante. Contienen toda la información capturada por el modelo discriminante, así como información acerca del rendimiento y la estructura del modelo.

Cuando se ejecuta una ruta que contiene un nugget de modelo Discriminante, el nodo añade dos nuevos campos que contienen la predicción del modelo y la probabilidad asociada. Los nombres de los nuevos campos se derivan del nombre del campo de salida que se está prediciendo, con el prefijo *$D-* para la categoría predicha y *$DP-* para la probabilidad asociada. Por ejemplo, para un campo de salida llamado *colorpref*, los nuevos campos se llamarían *$D-colorpref* y *$DP-colorpref*.

**Generación de un nodo Filtrar.** El menú Generar permite crear un nuevo nodo Filtrar para pasar los campos de entrada en función de los resultados del modelo.

Importancia del predictor

Opcionalmente, en la pestaña Modelo también se puede mostrar un gráfico que indique la importancia relativa de cada predictor cuando se calcule el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Tenga en cuenta que este gráfico sólo está disponible si se ha seleccionado **Calcular importancia de predictor** en la pestaña Analizar antes de generar el modelo. Consulte el tema “Importancia del predictor” en la página 44 para obtener más información.

### Resultados avanzados del nugget de modelo Discriminante

Los resultados avanzados del análisis discriminante ofrecen información detallada sobre el modelo estimado y su rendimiento. La mayoría de la información contenida en los resultados avanzados es

bastante técnica y es necesario tener amplios conocimientos sobre análisis discriminante para interpretar correctamente estos resultados. Consulte el tema “Opciones de resultados del nodo Discriminante” en la página 203 para obtener más información.

### Configuración de nugget de modelo Discriminante

La pestaña Configuración de un nugget de modelo Discriminante le permite obtener puntuaciones de propensión al puntuar el modelo. Esta pestaña está disponible sólo para modelos con objetivos de marca y sólo después de que el nugget de modelo se haya añadido a una ruta.

**Calcular puntuaciones de propensión en bruto.** En el caso de modelos con un objetivo de marca (que devuelve una predicción de sí o no), puede solicitar puntuaciones de propensión que indican la probabilidad del resultado true especificado para el campo objetivo. Éstas se añaden a otros valores de predicción y confianza que pueden generarse durante la puntuación.

**Calcular puntuaciones de propensión ajustada.** Las puntuaciones de propensión en bruto se basan sólo en los datos de entrenamiento y pueden ser demasiado optimistas debido a la tendencia de muchos modelos a sobreajustar estos datos. Las propensiones ajustadas intentan compensar evaluando el rendimiento del modelo frente a una partición de comprobación o validación. Esta opción requiere que se haya definido un campo de partición en la ruta y que se hayan activado puntuaciones ajustadas de propensión en el modo de modelado antes de generar el modelo.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

### Resumen de nugget de modelo Discriminante

La pestaña Resumen de un nugget de modelo Discriminante muestra los campos y ajustes utilizados para generar el modelo. Además, si ha ejecutado un nodo Análisis conectado a este nodo de modelado, la información de dicho análisis también se mostrará en esta sección. Para obtener información general sobre cómo utilizar el explorador de modelos, consulte “Examen de nuggets de modelo” en la página 42.

## Nodo GenLin

El modelo lineal generalizado amplía el modelo lineal general, de manera que la variable dependiente está relacionada linealmente con los factores y las covariables mediante una determinada función de enlace. Además, el modelo permite que la variable dependiente tenga una distribución no normal. Cubre los modelos estadísticos más comunes, como la regresión lineal para respuestas distribuidas normalmente, los modelos logísticos para datos binarios, el modelo lineal de logaritmo para datos de frecuencias, modelos log-log complementarios para datos de supervivencia censurados por intervalos y numerosos modelos estadísticos a través de su formulación general de modelos.

**Ejemplos.** Una compañía de transporte puede utilizar modelos lineales generalizados para ajustar una regresión de Poisson a las frecuencias de daños de varios tipos de barcos construidos en varios períodos de tiempo. El modelo resultante puede ayudar a determinar cuales son los tipos de barcos más propensos a sufrir daños.

Una compañía de seguros de automóviles puede utilizar modelos lineales generalizados para ajustar una regresión gamma a las reclamaciones por daños de los automóviles. El modelo resultante puede ayudar a determinar los factores que más contribuyen al tamaño de la reclamación.

Los investigadores médicos pueden utilizar modelos lineales generalizados para ajustar una regresión log-log complementario a los datos de supervivencia censurados por intervalos para pronosticar el tiempo que tardará en reaparecer una enfermedad.

Los modelos lineales generalizados funcionan generando una ecuación que relaciona los valores de los campos de entrada con los valores de los campos de salida. Una vez se ha generado el modelo, se puede utilizar para calcular los valores de datos nuevos. Para cada registro, se calcula una probabilidad de pertenencia a cada categoría posible de salida. La categoría objetivo con la probabilidad más alta se asigna como el valor de salida predicho para cada registro.

**Requisitos.** Necesita uno o más campos de entrada y exactamente un campo objetivo (que puede tener un nivel de medición *Continuo* o *Marca*) con dos o más categorías. Los tipos de los campos utilizados en el modelo deben estar completamente instanciados.

**Puntos fuertes.** El modelo lineal generalizado es extremadamente flexible, pero el proceso de selección de la estructura del modelo no está automatizado y, por tanto, requiere cierta familiaridad con los datos que no es necesaria en los algoritmos de "caja negra".

## Opciones de los campos del nodo GenLin

Además de las opciones personalizadas de objetivo, entrada y partición que se suelen ofrecer en las pestañas Campos del nodo de modelado (consulte “Opciones de los campos del nodo de modelado” en la página 31) el nodo GenLin ofrece la siguiente funcionalidad adicional.

**Utilizar campo de ponderación** El parámetro de escala es un parámetro de modelo estimado relacionado con la varianza de la respuesta. Las ponderaciones de escala son valores "conocidos" que pueden variar de una observación a otra. Si se especifica una variable de ponderación de escala, el parámetro de escala, que está relacionado con la varianza de la respuesta, se divide por ella para cada observación. Los registros con valores de ponderación de escala que sean inferiores o iguales a 0 o que sean valores perdidos no se utilizarán en el análisis.

**El campo Objetivo representa el número de eventos que se producen en un conjunto de ensayos.** Cuando la respuesta es un número de eventos que se producen en un conjunto de ensayos, el campo objetivo contiene el número de eventos y puede seleccionar una variable adicional que contenga el número de ensayos. Otra posibilidad, si el número de ensayos es el mismo en todos los sujetos, consiste en especificar los ensayos mediante un valor fijo. El número de ensayos debe ser superior o igual al número de eventos de cada registro. Los eventos deben ser enteros no negativos y los ensayos deben ser enteros positivos.

## Opciones de modelo del nodo GenLin

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Crear modelos divididos.** Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como campos de división. Consulte el tema “Generación de modelos divididos” en la página 28 para obtener más información.

**Tipo de modelo.** Hay dos opciones para generar el modelo. **Sólo efectos principales** hace que el modelo sólo incluya los campos de entrada de forma individual y no comprueban las interacciones (efectos multiplicativos) entre los campos de entrada. **Efectos principales y todas las interacciones de dos factores** incluye todas las interacciones de dos factores y los efectos principales de los campos de entrada.

**Desplazamiento.** El término desplazamiento es un predictor "estructural". Su coeficiente no se estima por el modelo pero se supone que tiene el valor 1. Por tanto, los valores del desplazamiento se suman sencillamente al predictor lineal del destino. Esto resulta especialmente útil en los modelos de regresión de Poisson, en los que cada caso puede tener diferentes niveles de exposición al evento de interés.

Por ejemplo, al modelar las tasas de accidente de diferentes conductores, hay una importante diferencia entre un conductor que ha sido el culpable de 1 accidente en 3 años y un conductor que ha sido el culpable de 1 accidente en 25 años. El número de accidentes se puede modelar como una respuesta Poisson o una binomial negativa con un enlace de registro si el registro natural de la experiencia del conductor se incluye como un término de desplazamiento.

Otras combinaciones de distribución y tipos de enlaces necesitarían otras transformaciones de la variable de desplazamiento.

*Note:* Si se utiliza un campo de desplazamiento de variable, el campo especificado no debe utilizarse también como una entrada. Defina el rol del campo de desplazamiento como **Ninguno** en un origen anterior de la ruta o nodo Tipo si es necesario.

**Categoría base para el objetivo de marca.**

Para la respuesta binaria, puede seleccionar la categoría de referencia para la variable dependiente. Esto puede afectar a determinados resultados, como las estimaciones de los parámetros y los valores guardados, pero no debería cambiar el ajuste del modelo. Por ejemplo, si la respuesta binaria toma los valores 0 y 1.

- De forma predeterminada, el procedimiento convierte la última categoría (la de mayor valor), o 1, en categoría de referencia. En esta situación, las probabilidades guardadas por el modelo calculan la posibilidad de que un caso determinado tome el valor 0 y los cálculos del parámetro deberían interpretarse como si estuvieran relacionados con la verosimilitud de categoría 0.
- Si especifica la primera categoría (la de menor valor), o 0, como categoría de referencia, entonces las probabilidades guardadas por el modelo calculan la posibilidad de que un caso determinado tome el valor 1.
- Si especifica la categoría personalizada y su variable tiene etiquetas definidas, puede establecer la categoría de referencia seleccionando un valor de la lista. Esto puede resultar cómodo cuando, al especificar un modelo, no se recuerda exactamente cómo se codificó una determinada variable.

**Incluir la interceptación en el modelo.** La interceptación se incluye normalmente en el modelo. Si se puede dar por supuesto la lectura de datos en el origen, se puede excluir la interceptación.

## Opciones de experto del nodo GenLin

Las opciones de experto le permiten ajustar el proceso de entrenamiento, si tiene conocimientos sobre modelos lineales generalizados. Para acceder a las opciones de expero, establezca **Modo** en **Experto** en la pestaña Experto.

Distribución de campo objetivo y Función de enlace

**Distribución.**

Esta selección especifica la distribución de la variable dependiente. La posibilidad de especificar una distribución que no sea la normal y una función de enlace que no sea la identidad es la principal mejora que aporta el modelo lineal generalizado respecto al modelo lineal general. Hay muchas combinaciones

posibles de distribución y función de enlace, varias de las cuales pueden ser adecuadas para un determinado conjunto de datos, por lo que su elección puede estar guiada por consideraciones teóricas a priori y por las combinaciones que parezcan funcionar mejor.

- **Binomial.** Esta distribución es apropiada únicamente para variables que representan una respuesta binaria o un número de eventos.
- **Gamma.** Esta distribución es adecuada para las variables con valores de escala positivos que se desvían hacia valores positivos más grandes. Si un valor de datos es menor o igual que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis.
- **De Gauss inversa.** Esta distribución es adecuada para las variables con valores de escala positivos que se desvían hacia valores positivos más grandes. Si un valor de datos es menor o igual que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis.
- **Binomial negativa.** Esta distribución considera el número de intentos necesarios para lograr *k* éxitos y es adecuada para variables que tengan valores enteros que no sean negativos. Si un valor de datos no es entero, es menor que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis. El valor del parámetro auxiliar de la distribución binomial negativa puede ser cualquier número mayor o igual que 0. Cuando el parámetro auxiliar se establece como 0, utilizar esta distribución equivale a utilizar la distribución de Poisson.
- **Normal.** Es adecuada para variables de escala cuyos valores adoptan una distribución simétrica con forma de campana en torno a un valor central (la media). La variable dependiente debe ser numérica.
- **Poisson.** Esta distribución considera el número de instancias de un evento de interés en un período fijo de tiempo y es apropiada para variables que tengan valores enteros que no sean negativos. Si un valor de datos no es entero, es menor que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis.
- **Tweedie.** Esta distribución es adecuada para las variables que se pueden representar mediante mezclas de Poisson de distribuciones gamma; la distribución está "mezclada" en el sentido de que combina propiedades de distribuciones continuas (toma valores reales no negativos) y discretas (masa de probabilidad positiva en un único valor, 0). La variable dependiente debe ser numérica, con valores de datos mayores o iguales que cero. Si un valor de datos es menor que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis. El valor fijo del parámetro de distribución Tweedie puede ser cualquier número mayor que uno y menor que dos.
- **Multinomial.** Esta distribución es la adecuada para las variables que representan una respuesta ordinal. La variable dependiente puede ser numérica o una cadena y debe tener al menos dos valores de datos válidos distintos.

**Funciones de enlace.**

La función de enlace es una transformación de la variable dependiente que permite una estimación del modelo. Se encuentran disponibles las siguientes funciones:

- **Identidad.** $f(x)=x$. No se transforma la variable dependiente. Este enlace se puede utilizar con cualquier distribución.
- **log-log complementario.** $f(x)=\log(-\log(1-x))$. Adecuado solamente con la distribución binomial.
- **Cauchit acumulado.** $f(x) = \tan(\pi\ (x - 0{,}5))$, se aplica a la probabilidad acumulada de cada categoría de la respuesta. Adecuado sólo con la distribución multinomial.
- **Log-log complementario acumulado.** $f(x)=\ln(-\ln(1-x))$, se aplica a la probabilidad acumulada de cada categoría de la respuesta. Adecuado sólo con la distribución multinomial.
- **Logit acumulado** $f(x)=\ln(x\ /\ (1-x))$, se aplica a la probabilidad acumulada de cada categoría de la respuesta. Adecuado sólo con la distribución multinomial.
- **Log-log negativo acumulado.** $f(x)=-\ln(-\ln(x))$, se aplica a la probabilidad acumulada de cada categoría de la respuesta. Adecuado sólo con la distribución multinomial.
- **Probit acumulado.** $f(x)=\Phi^{-1}(x)$, se aplica a la probabilidad acumulada de cada categoría de la respuesta, donde $\Phi^{-1}$ es la función de distribución acumulada normal estándar inversa. Adecuado sólo con la distribución multinomial.

- **Logaritmo.** $f(x)=\log(x)$. Este enlace se puede utilizar con cualquier distribución.
- **Complemento Log.l** $f(x)=\log(1-x)$. Adecuado solamente con la distribución binomial.
- **Logit.** $f(x)=\log(x\ /\ (1-x))$. Adecuado solamente con la distribución binomial.
- **Binomial negativa.** $f(x)=\log(x\ /\ (x+k^{-1}))$, donde $k$ el parámetro auxiliar de la distribución binomial negativa. Adecuado sólo con la distribución binomial negativa.
- **log-log negativa.** $f(x)=-\log(-\log(x))$. Adecuado solamente con la distribución binomial.
- **Potencia de las ventajas.** $f(x)=[(x/(1-x))^{\alpha}-1]/\alpha$, si $\alpha \neq 0$. $f(x)=\log(x)$, si $\alpha=0$. $\alpha$ es la especificación de número necesaria y debe ser un número real. Adecuado solamente con la distribución binomial.
- **Probit.** $f(x)=\Phi^{-1}(x)$, donde $\Phi^{-1}$ es la función de distribución acumulada normal estándar inversa. Adecuado solamente con la distribución binomial.
- **Potencia.** $f(x)=x^{\alpha}$, si $\alpha \neq 0$. $f(x)=\log(x)$, si $\alpha=0$. $\alpha$ es la especificación de número necesaria y debe ser un número real. Este enlace se puede utilizar con cualquier distribución.

**Parámetros.** Los controles de este grupo le permiten especificar los valores de parámetros si se seleccionan algunas opciones de distribución.

- **Parámetro de binomial negativa.** Para distribución binomial negativa, seleccione si desea especificar un valor o permitir que el sistema proporcione un valor estimado.
- **Parámetro de Tweedie.** Para la distribución de Tweedie, especifique un número entre 1,0 y 2,0 para el valor fijo.

  **Estimación de parámetros.** Los controles de este grupo le permiten especificar los métodos de estimación y proporcionar los valores iniciales para las estimaciones de los parámetros.

  – **Método.** Puede seleccionar el método de estimación del parámetro de escala. Los métodos disponibles son Newton-Raphson, Scoring de Fisher o un método híbrido en el que las iteraciones de Scoring de Fisher se realizan antes de cambiar al método de Newton-Raphson. Si se logra la convergencia durante la fase de Scoring de Fisher del método híbrido antes de que se lleven a cabo el número máximo de iteraciones de Fisher, el algoritmo continúa con el método de Newton-Raphson.

  – **Método de parámetro de escala.** Puede seleccionar el método de estimación del parámetro de escala. La máxima verosimilitud estima conjuntamente el parámetro de escala y los efectos del modelo. Tenga en cuenta que esta opción no es válida si la respuesta tiene una distribución binomial negativa, de Poisson o binomial . Las opciones de desvianza y de chi-cuadrado de Pearson estiman el parámetro de escala a partir del valor de dichos estadísticos. Otra posibilidad consiste en especificar un valor corregido para el parámetro de escala.
- **Matriz de covarianzas.** El estimador basado en el modelo es el negativo de la inversa generalizada de la matriz hessiana. El estimador robusto (también llamado de Huber/White/sandwich) es un estimador "corregido" basado en el modelo que proporciona una estimación coherente de la covarianza, incluso cuando se ha especificado incorrectamente la varianza y las funciones de enlace.

**Iteraciones.** Estas opciones le permiten controlar los parámetros de la convergencia del modelo. Consulte el tema “Iteraciones de modelos lineales generalizados” en la página 211 para obtener más información.

**Resultados.** Estas opciones le permiten solicitar estadísticos adicionales que aparecerán en el resultado avanzado del nugget de modelo construido por el nodo. Consulte el tema “Resultados avanzados de modelos lineales generalizados” en la página 211 para obtener más información.

**Tolerancia para la singularidad.** Las matrices singulares (que no se pueden invertir) tienen columnas linealmente dependientes, lo que puedo causar graves problemas al algoritmo de estimación. Incluso las matrices casi singulares pueden generar resultados deficientes, por lo que el procedimiento tratará una matriz cuyo determinante es menor que la tolerancia como singular. Especifique un valor positivo.

## Iteraciones de modelos lineales generalizados

Puede establecer los parámetros de convergencia para la estimación del modelo lineal generalizado.

**Iteraciones.** Se encuentran disponibles las siguientes opciones:

- **Iteraciones máximas.** Número máximo de iteraciones que se ejecutará el algoritmo. Especifique un número entero no negativo.
- **Máxima subdivisión por pasos.** En cada iteración, se reduce el tamaño del paso mediante un factor de 0,5 hasta que aumenta el logaritmo de la verosimilitud o se alcanza la máxima subdivisión por pasos. Especifique un número entero positivo.
- **Comprobar si hay separación completa de los puntos de los datos.** Si se activa, el algoritmo realiza una prueba para garantizar que las estimaciones de los parámetros tienen valores exclusivos. Se produce una separación cuando el procedimiento pueda generar un modelo que clasifique cada caso de forma correcta. Esta opción no está disponible para respuestas binomiales con formato binario .

**Criterios de convergencia.** Las opciones disponibles son las siguientes:

- **Convergencia de los parámetros.** Si se activa, el algoritmo se detiene tras una iteración en la que las modificaciones absolutas o relativas en las estimaciones de los parámetros sean inferiores que el valor especificado, que debe ser positivo.
- **Convergencia del logaritmo de la verosimilitud.** Si se activa, el algoritmo se detiene tras una iteración en la que las modificaciones absolutas o relativas en la función de log-verosimilitud sean inferiores que el valor especificado, que debe ser positivo.
- **Convergencia hessiana.** En el caso de la especificación Absoluta, se supone la convergencia si un estadístico basado en la convergencia hessiana es menor que el valor positivo especificado. En el caso de la especificación Relativa, se supone la convergencia si el estadístico es menor que el producto del valor positivo especificado y el valor absoluto del logaritmo de la verosimilitud.

## Resultados avanzados de modelos lineales generalizados

Seleccione el resultado opcional que desee mostrar en el resultado avanzado del nugget de modelo lineal generalizado. Para ver el resultado avanzado, examine el nugget de modelo y pulse en la pestaña **Avanzado**. Consulte el tema “Resultado avanzado del nugget de modelo GenLin” en la página 213 para obtener más información.

Los siguientes resultados están disponibles:

- **Resumen del procesamiento de los casos.** Muestra el número y el porcentaje de los casos incluidos y excluidos del análisis y la tabla Resumen de datos correlacionados.
- **Estadísticos descriptivos.** Muestra estadísticos descriptivos e información resumida acerca de los factores, las covariables y la variable dependiente.
- **Información del modelo.** Muestra el nombre del conjunto de datos, la variable dependiente o las variables de eventos y ensayos, la variable de desplazamiento, la distribución de probabilidad y la función de enlace.
- **Estadísticos de bondad de ajuste.** Muestra la desvianza y la desvianza escalada, chi-cuadrado de Pearson y chi-cuadrado de Pearson escalado, log-verosimilitud, criterio de información de Akaike (AIC), AIC corregido para muestras finitas (AICC), criterio de información bayesiano (BIC) y AIC consistente (CAIC).
- **Estadísticos de resumen del modelo.** Muestra contraste de ajuste del modelo, incluidos los estadísticos de la razón de la verosimilitud para el contraste Omnibus del ajuste del modelo y los estadísticos para los contrastes de Tipo I o III para cada efecto.
- **Estimaciones de los parámetros.** Muestra las estimaciones de los parámetros y los correspondientes estadísticos de contraste e intervalos de confianza. Si lo desea, puede mostrar las estimaciones exponenciadas de los parámetros además de las estimaciones brutas de los parámetros.
- **Matriz de covarianzas de las estimaciones de los parámetros.** Muestra la matriz de covarianzas de los parámetros estimados.

- **Matriz de correlaciones de las estimaciones de los parámetros.** Muestra la matriz de correlaciones de los parámetros estimados.
- **Matrices (L) de los coeficientes de contraste.** Muestra los coeficientes de los contrastes para los efectos predeterminados y para las medias marginales estimadas, si se solicitaron en la pestaña Medias marginales estimadas.
- **Funciones estimables generales.** Muestra las matrices para generar las matrices (L) de los coeficientes de contraste.
- **Historial de iteraciones.** Muestra el historial de iteraciones de las estimaciones de los parámetros y el log-verosimilitud, e imprime la última evaluación del vector de gradiente y la matriz hessiana. En la tabla de historial de iteraciones se muestran las estimaciones de parámetro para cada $n^{\text{-ésima}}$ iteración que comienza con la iteración $0^{\text{-ésima}}$ (la estimación inicial), donde *n* es el valor del intervalo de impresión. Si se solicita el historial de iteraciones, la última iteración siempre se muestra independientemente de *n*.
- **Contraste de multiplicador de Lagrange.** Muestra los estadísticos de contraste de multiplicadores de Lagrange que permite evaluar la validez de un parámetro de escala calculado utilizando la desvianza o chi-cuadrado de Pearson, así como establecer un número fijo para las distribuciones normal, gamma y de Gauss inversa. Para la distribución binomial negativa, se contrasta el parámetro auxiliar fijo.

**Efectos del modelo.** Se encuentran disponibles las siguientes opciones:

- **Tipo de análisis.** Especifique el tipo de análisis que se va a producir. El análisis de tipo I suele ser apropiado cuando tiene motivos a priori para ordenar los predictores del modelo, mientras que el tipo III es de aplicación más general. Los estadísticos de Wald o de la razón de la verosimilitud se calculan según la selección del grupo de estadísticos de chi-cuadrado.
- **Intervalos de confianza.** Especifique un nivel de confianza mayor que 50 y menor que 100. Los intervalos de Wald se basan en el supuesto de que los parámetros tienen una distribución normal asintótica; los intervalos de verosimilitud de perfil son más precisos pero pueden suponer un mayor esfuerzo computacional. El nivel de tolerancia de intervalos de verosimilitud de perfil son los criterios utilizados para detener el algoritmo iterativo utilizado para calcular los intervalos.
- **Función de log-verosimilitud.** Esto controla el formato de presentación de la función de log-verosimilitud. La función completa incluye un término adicional constante con respecto a las estimaciones de los parámetros; no tiene efecto en la estimación de parámetros y no se muestra en algunos productos de software.

## Nugget de modelo GenLin

Un nugget de modelo GenLin representa la ecuación calculada por un nodo GenLin. Contienen toda la información capturada por el modelo, así como información acerca del rendimiento y la estructura del modelo.

Cuando se ejecuta una ruta que contiene un nugget de modelo GenLin, el nodo añade nuevos campos cuyo contenido depende de la naturaleza del campo objetivo:

- **Marcar objetivo.** Añade campos que contienen la categoría predicha y la probabilidad asociada, así como las probabilidades de cada categoría. Los nombres de los dos primeros nuevos campos se derivan del nombre del campo de salida que se está prediciendo, con el prefijo *$G-* para la categoría predicha y *$GP-* para la probabilidad asociada. Por ejemplo, para un campo de salida llamado *valor-predeterminado*, los nuevos campos se llamarían *$G-predeterminado* y *$GP-predeterminado*. Los nombres de los posteriores dos campos adicionales se asignan en función de los valores del campo de salida, con el prefijo *$GP-*. Por ejemplo, si los valores correctos de *valor-predeterminado* son *Sí* y *No*, los nuevos campos se denominarán *$GP-Sí* y *$GP-No*.
- **Objetivo continuo.** Añade campos que contienen la medida predicha y el error estándar.
- **Objetivo continuo, representando el número de eventos en una serie de ensayos.** Añade campos que contienen la medida predicha y el error estándar.

- **Objetivo ordinal.** Añade campos que contienen la categoría predicha y la probabilidad asociada para cada valor del conjunto ordenado. Los nombres de los campos se derivan del valor del conjunto ordenado que se está prediciendo, con el prefijo *$G-* para la categoría predicha y *$GP-* para la probabilidad asociada.

**Generación de un nodo Filtrar.** El menú Generar permite crear un nuevo nodo Filtrar para pasar los campos de entrada en función de los resultados del modelo.

Importancia del predictor

Opcionalmente, en la pestaña Modelo también se puede mostrar un gráfico que indique la importancia relativa de cada predictor cuando se calcule el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Tenga en cuenta que este gráfico sólo está disponible si se ha seleccionado **Calcular importancia de predictor** en la pestaña Analizar antes de generar el modelo. Consulte el tema "Importancia del predictor" en la página 44 para obtener más información.

## Resultado avanzado del nugget de modelo GenLin

El resultado avanzado del modelo lineal generalizado ofrece información detallada sobre el modelo estimado y su rendimiento. La mayoría de la información contenida en los resultados avanzados es bastante técnica y es necesario tener amplios conocimientos sobre este tipo de análisis para interpretar correctamente estos resultados. Consulte el tema "Resultados avanzados de modelos lineales generalizados" en la página 211 para obtener más información.

## Configuración de nugget de modelo GenLin

La pestaña Configuración de un nugget de modelo GenLin le permite obtener puntuaciones de propensión al puntuar el modelo, y también para la generación de SQL durante la puntuación de modelos. Esta pestaña está disponible sólo para modelos con objetivos de marca y sólo después de que el nugget de modelo se haya añadido a una ruta.

**Calcular puntuaciones de propensión en bruto.** En el caso de modelos con un objetivo de marca (que devuelve una predicción de sí o no), puede solicitar puntuaciones de propensión que indican la probabilidad del resultado true especificado para el campo objetivo. Éstas se añaden a otros valores de predicción y confianza que pueden generarse durante la puntuación.

**Calcular puntuaciones de propensión ajustada.** Las puntuaciones de propensión en bruto se basan sólo en los datos de entrenamiento y pueden ser demasiado optimistas debido a la tendencia de muchos modelos a sobreajustar estos datos. Las propensiones ajustadas intentan compensar evaluando el rendimiento del modelo frente a una partición de comprobación o validación. Esta opción requiere que se haya definido un campo de partición en la ruta y que se hayan activado puntuaciones ajustadas de propensión en el modo de modelado antes de generar el modelo.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

### Resumen de nugget de modelo GenLin

La pestaña Resumen de un nugget de modelo GenLin muestra los campos y ajustes utilizados para generar el modelo. Además, si ha ejecutado un nodo Análisis conectado a este nodo de modelado, la información de dicho análisis también se mostrará en esta sección. Para obtener información general sobre cómo utilizar el explorador de modelos, consulte "Examen de nuggets de modelo" en la página 42.

# Modelos lineales mixtos generalizados

# Nodo GLMM

Utilice este nodo para crear un modelo lineal mixto generalizado (GLMM).

## Modelos lineales mixtos generalizados

Los modelos lineales mixtos generalizados amplían el modelo lineal de modo que:

- El objetivo está linealmente relacionado con los factores y covariables mediante una función de enlace especificada.
- El objetivo puede tener una distribución no normal.
- Las observaciones se pueden correlacionar.

Los modelos lineales mixtos generalizados cubren una amplia variedad de modelos, desde modelos de regresión lineal simple hasta modelos multinivel complejos para datos longitudinales no normales.

**Ejemplos.** El consejo escolar del distrito puede utilizar un modelo lineal mixto generalizado para determinar si un método educativo experimental es eficaz para mejorar las notas en matemáticas. Los estudiantes de la misma clase deberían correlacionarse dado que les enseña el mismo maestro. Asimismo, las clases del mismo colegio también deberían correlacionarse. De este modo, podemos incluir efectos aleatorios a nivel de colegio y de clase para explicar las diferentes fuentes de variabilidad.

Los investigadores médicos pueden utilizar un modelo lineal mixto generalizado para determinar si un nuevo medicamento antiepiléptico puede reducir la tasa de crisis epilépticas de un paciente. Las mediciones repetidas del mismo paciente normalmente se correlacionan de forma positiva, de modo que sería adecuado utilizar un modelo mixto con algunos efectos aleatorios. El campo objetivo, el número de convulsiones, utiliza valores enteros positivos, por lo que podría ser adecuado utilizar un modelo lineal mixto generalizado con una distribución de Poisson y enlace log.

Los ejecutivos de un proveedor de servicios de televisión, teléfono e Internet por cable pueden utilizar un modelo lineal mixto generalizado para saber más sobre los posibles clientes. Como las posibles respuestas tienen niveles de medición nominal, el analista de la empresa utiliza un modelo logit mixto generalizado con una interceptación aleatoria para capturar la correlación entre las respuestas a las preguntas del uso de los distintos tipos de servicios (televisión, teléfono e Internet) dentro de las respuestas de una persona a una encuesta específica.

La pestaña Estructura de datos le permite especificar las relaciones estructurales entre los registros de su conjunto de datos cuando se correlacionan observaciones. Si los registros del conjunto de datos representan observaciones independientes, no necesita especificar nada en esta pestaña.

**Sujetos.** La combinación de valores de los campos categóricos especificados debe definir de forma exclusiva los sujetos del conjunto de datos. Por ejemplo, un único campo *ID de paciente* debería ser suficiente para definir los sujetos de un único hospital, pero puede que sea necesario combinar *ID de hospital* e *ID de paciente* si los números de identificación de paciente no son exclusivos entre varios hospitales. En una configuración de medidas repetidas, se registran varias observaciones para cada sujeto, de manera que cada sujeto puede ocupar varios registros del conjunto de datos.

Un **sujeto** es una unidad de observación que puede considerarse independiente de otros sujetos. Por ejemplo, las lecturas de la tensión arterial de un paciente en un estudio médico pueden considerarse independientes de las lecturas de otros pacientes. La definición de sujetos resulta especialmente importante cuando se producen mediciones repetidas por sujeto y quiere modelar la correlación entre estas observaciones. Por ejemplo, cabría esperar que las lecturas de la tensión arterial de un único paciente durante visitas consecutivas al médico estén correlacionadas.

Todos los campos especificados como Sujetos en la pestaña Estructura de datos se utilizan para definir sujetos para la estructura de la covarianza residual y proporcionan la lista de posibles campos para definir sujetos para estructuras de covarianza de los efectos aleatorios en el Bloque de efectos aleatorios.

**Medidas repetidas.** Los campos especificados aquí se utilizan para identificar observaciones repetidas. Por ejemplo, una única variable *Semana* podría identificar las 10 semanas de observaciones en un estudio médico, o *Mes* y *Día* podrían utilizarse en conjunto para identificar observaciones diarias a lo largo de un año.

**Definir grupos de covarianzas por.** Los campos categóricos especificados aquí definen conjuntos independientes de parámetros de covarianza de efectos repetidos; uno para cada categoría definida por la clasificación cruzada de los campos de agrupación. Todos los sujetos tienen el mismo tipo de covarianza; los sujetos con la misma agrupación de covarianza tendrán los mismos valores para los parámetros.

**Coordenadas de covarianza espaciales.** Las variables de esta lista especifican las coordenadas de las observaciones repetidas, cuando se selecciona uno de los tipos de covarianza espaciales para el tipo de covarianza repetido.

**Tipo de covarianza repetido.** Esto especifica la estructura de la covarianza de los residuos. Las estructuras disponibles son:

- Autorregresiva de primer orden (AR1)
- Media móvil autorregresiva (1,1) (ARMA11)
- Simetría compuesta
- Diagonal
- Identidad escalada
- Espacial: potencia
- Espacial: exponencial
- Espacial: gausiano
- Espacial: Lineal
- Espacial: log-lineal
- Espacial: esférico
- Toeplitz
- Sin estructura
- Componentes de la varianza

**Objetivo:** Estos ajustes definen el objetivo, su distribución y su relación con los predictores mediante la función de enlace.

**Objetivo.** El objetivo es obligatorio. Puede tener cualquier nivel de medición. Dicho nivel de medición del objetivo restringe las distribuciones y funciones de enlace que son adecuadas.

- **Utilice el número de ensayos como denominador.** Cuando la respuesta objetivo es un número de eventos que se producen en un conjunto de ensayos, el campo objetivo contiene el número de eventos y puede seleccionar un campo adicional que contenga el número de ensayos. Por ejemplo, al probar un nuevo pesticida puede que exponga muestras de hormigas a diferentes concentraciones del pesticida y, a continuación, registre el número de hormigas muertas y el número de hormigas de cada muestra. En

este caso, el campo que registra el número de hormigas muertas debe especificarse como el campo objetivo (eventos) y el campo que registre el número de hormigas de cada muestra debe especificarse como el campo de ensayos. Si el número de hormigas es el mismo para cada muestra, entonces el número de ensayos puede especificarse mediante un valor fijo.

El número de ensayos debe ser superior o igual al número de eventos de cada registro. Los eventos deben ser enteros no negativos y los ensayos deben ser enteros positivos.

- **Personalice la categoría de referencia.** Para un objetivo categórico, puede seleccionar la categoría de referencia. Esto puede afectar a determinados resultados, como las estimaciones de los parámetros, pero no debería cambiar el ajuste del modelo. Por ejemplo, si su objetivo toma los valores 0, 1 y 2, de forma predeterminada, el procedimiento convierte la última categoría (el valor más alto), o 2, en la categoría de referencia. En esta situación, las estimaciones de los parámetros deben interpretarse como relacionadas con la verosimilitud de la categoría 0 o 1 *relativa* a la verosimilitud de la categoría 2. Si especifica una categoría personalizada y su objetivo tiene etiquetas definidas, puede establecer la categoría de referencia seleccionando un valor de la lista. Esto puede resultar cómodo cuando, al especificar un modelo, no se recuerda exactamente cómo se codificó un determinado campo.

**Distribución de objetivos y relación (enlace) con el modelo lineal.** Dados los valores de los predictores, el modelo espera que la distribución de valores del objetivo siga la forma especificada y que los valores de objetivo estén linealmente relacionados con los predictores mediante la función de enlace especificada. Se proporcionan los accesos directos de varios modelos comunes o seleccione un ajuste **Personalizado** si hay una combinación específica de distribución y función de enlace que desee ajustar y que no esté en la lista corta.

- **Modelo lineal.** Especifica una distribución normal con un enlace de identidad, que resulta de utilidad cuando se puede predecir el objetivo mediante un modelo de regresión lineal o ANOVA.
- **Regresión Gamma.** Especifica una distribución Gamma con un enlace log, que debe utilizarse cuando todos los valores que contiene el objetivo son positivos y el objetivo se desvía hacia valores más grandes.
- **Loglineal.** Especifica una distribución de Poisson con un enlace log, que debe utilizarse cuando el objetivo representa un recuento de instancias en un período de tiempo fijo.
- **Regresión binomial negativa.** Especifica una distribución binomial negativa con un enlace log, que debe utilizarse cuando el objetivo y el denominador representan el número de ensayos necesarios para lograr *k* éxitos.
- **Regresión logística multinomial.** Especifica una distribución multinomial, que debe utilizarse cuando el objetivo es una respuesta de categorías múltiples. Utiliza un enlace logit acumulado (resultados ordinales) o un enlace logit generalizado (respuestas nominales de categorías múltiples).
- **Regresión logística binaria.** Especifica una distribución binomial con un enlace logit, que debe utilizarse cuando el objetivo es una respuesta binaria predicha por un modelo de regresión logística.
- **Probit binario.** Especifica una distribución binomial con un enlace probit, que debe utilizarse cuando el objetivo es una respuesta binaria con una distribución normal subyacente.
- **Supervivencia censurada por intervalos.** Especifica una distribución binomial con un enlace log-log complementario, que resulta de utilidad en el análisis de supervivencia cuando algunas observaciones no tienen evento de terminación.

Distribución

Esta selección especifica la distribución del objetivo. La posibilidad de especificar una distribución que no sea la normal y una función de enlace que no sea la identidad es la principal mejora que aporta el modelo lineal mixto generalizado respecto al modelo lineal mixto. Hay muchas combinaciones posibles de distribución y función de enlace, varias de las cuales pueden ser adecuadas para un determinado conjunto de datos, por lo que su elección puede estar guiada por consideraciones teóricas a priori y por las combinaciones que parezcan funcionar mejor.

- **Binomial.** Esta distribución es apropiada únicamente para un objetivo que represente una respuesta binaria o un número de eventos.

- **Gamma.** Esta distribución es adecuada para un objetivo con valores de escala positivos que se desvían hacia valores positivos más grandes. Si un valor de datos es menor o igual que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis.
- **De Gauss inversa.** Esta distribución es adecuada para un objetivo con valores de escala positivos que se desvían hacia valores positivos más grandes. Si un valor de datos es menor o igual que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis.
- **Multinomial.** Esta distribución es adecuada para un objetivo que representa una respuesta de categorías múltiples. La forma del modelo dependerá del nivel de medición del objetivo.

  Un objetivo **nominal** dará como resultado un modelo multinomial nominal en el que se calcula un conjunto independiente de parámetros del modelo para cada categoría del objetivo (excepto la categoría de referencia). Las estimaciones de parámetros de un predictor determinado muestran la relación entre ese predictor y la verosimilitud de cada categoría del objetivo, relativa a la categoría de referencia.

  Un objetivo **ordinal** dará como resultado un modelo multinomial ordinal en el que el término de interceptación tradicional se sustituye por un conjunto de parámetros de **umbral** que se relacionan con la probabilidad acumulada de las categorías objetivo.
- **Binomial negativa.** La regresión binomial negativa utiliza una distribución binomial negativa con un enlace log, que debe utilizarse cuando el objetivo representa un recuento de instancias con varianza elevada.
- **Normal.** Es adecuada para un objetivo continuo cuyos valores adoptan una distribución simétrica con forma de campana en torno a un valor central (la media).
- **Poisson.** Esta distribución considera el número de instancias de un evento de interés en un período fijo de tiempo y es apropiada para variables que tengan valores enteros que no sean negativos. Si un valor de datos no es entero, es menor que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis.

Funciones de enlace

La función de enlace es una transformación del objetivo que permite una estimación del modelo. Se encuentran disponibles las siguientes funciones:

- **Identidad.** $f(x)=x$. El destino no se transforma. Este enlace se puede utilizar con cualquier distribución, excepto la multinomial.
- **log-log complementario.** $f(x)=\log(-\log(1-x))$. Adecuado solamente con la distribución binomial o multinomial.
- **Cauchit.** $f(x) = \tan(\pi\ (x - 0.5))$. Adecuado solamente con la distribución binomial o multinomial.
- **Logaritmo.** $f(x)=\log(x)$. Este enlace se puede utilizar con cualquier distribución, excepto la multinomial.
- **Complemento Log.l** $f(x)=\log(1-x)$. Adecuado solamente con la distribución binomial.
- **Logit.** $f(x)=\log(x\ /\ (1-x))$. Adecuado solamente con la distribución binomial o multinomial.
- **log-log negativa.** $f(x)=-\log(-\log(x))$. Adecuado solamente con la distribución binomial o multinomial.
- **Probit.** $f(x)=\Phi^{-1}(x)$, donde $\Phi^{-1}$ es la función de distribución acumulada normal estándar inversa. Adecuado solamente con la distribución binomial o multinomial.
- **Potencia.** $f(x)=x^{\alpha}$, si $\alpha \neq 0$. $f(x)=\log(x)$, si $\alpha=0$. $\alpha$ es la especificación de número necesaria y debe ser un número real. Este enlace se puede utilizar con cualquier distribución, excepto la multinomial.

**Efectos fijos:** Los factores de efectos fijos suelen considerarse campos cuyos valores de interés están todos representados en el conjunto de datos y pueden utilizarse para la puntuación. De forma predeterminada, los campos con el rol de entrada predefinido que no se especifican en ningún otro sitio del cuadro de diálogo se introducen en la parte de efectos fijos del modelo. Los campos categóricos (marca, nominal y ordinal) se utilizan como factores en el modelo y los campos continuos se utilizan como covariables.

Introduzca efectos en el modelo seleccionando uno o más campos en la lista de orígenes y arrastrándolos a la lista de efectos. El tipo de efecto creado depende de la zona activa en la que suelte la selección.

- **Principal.** Los campos que suelte aparecen como efectos principales independientes en la parte inferior de la lista de efectos.
- **2 factores.** Todos los pares posibles de los campos que suelte aparecen como interacciones de 2 factores en la parte inferior de la lista de efectos.
- **3 factores.** Todos los triples posibles de los campos que suelte aparecen como interacciones de 3 factores en la parte inferior de la lista de efectos.
- ***.** La combinación de todos los campos eliminados aparece como una interacción única en la parte inferior de la lista de efectos.

Los botones a la derecha del generador de efectos le permiten realizar diversas acciones.

*Tabla 10. Descripciones de botones de generador de efectos.*

| Icono | Descripción |
|---|---|
|  | Eliminar términos del modelo de efectos fijos seleccionando los términos que desea eliminar y pulsando en el botón eliminar. |
|  | Reordenar los términos dentro del modelo de efectos fijos seleccionando los términos que quiera reordenar y pulsando en la flecha arriba o abajo. |
|  | Añadir términos anidados al modelo utilizando el diálogo "Añadir un término personalizado", pulsando el botón Añadir un término personalizado. |

**Incluir interceptación.** La interceptación se incluye normalmente en el modelo. Si se puede dar por supuesto la lectura de datos en el origen, se puede excluir la interceptación.

*Añadir un término personalizado:* Puede generar términos anidados para su modelo en este procedimiento. Los términos anidados son útiles para modelar el efecto de un factor o covariable cuyos valores no interactúen con los niveles de otro factor. Por ejemplo, una cadena de supermercados puede seguir los hábitos de consumo de sus clientes en varias ubicaciones de sus tiendas. Dado que cada cliente frecuenta tan solamente una de estas ubicaciones, se puede decir que el efecto de *Cliente* está **anidado dentro** del efecto de *Ubicación de la tienda*.

Además, puede incluir efectos de interacción, como términos polinómicos que implican a la misma covariable, o añadir varios niveles de anidación al término anidado.

**Limitaciones.** Los términos anidados tienen las siguientes restricciones:

- Todos los factores incluidos en una interacción deben ser exclusivos entre sí. Por consiguiente, si *A* es un factor, no es válido especificar *A*A*.
- Todos los factores incluidos en un efecto anidado deben ser exclusivos entre sí. Por consiguiente, si *A* es un factor, no es válido especificar *A(A)*.
- No se puede anidar ningún efecto dentro de una covariable. Por consiguiente, si *A* es un factor y *X* es una covariable, no es válido especificar *A(X)*.

Creación de un término anidado

1. Seleccione un factor o covariable que esté anidado en otro factor y, a continuación, pulse en el botón de flecha.

2. Pulse en **(Dentro)**.
3. Seleccione el factor dentro del cual el factor o covariable anterior se anida y, a continuación, pulse en el botón de flecha.
4. Pulse en **Añadir término**.

Si lo desea, puede incluir efectos de interacción o añadir varios niveles de anidación al término anidado.

**Efectos aleatorios:** Los factores de efectos aleatorios son campos cuyos valores en el archivo de datos pueden considerarse una muestra aleatoria de una población de valores más grande. Son de utilidad para explicar la variabilidad excesiva en el objetivo. De forma predeterminada, si ha seleccionado más de un sujeto en la pestaña Estructura de datos, se creará un bloque de efectos aleatorios para cada sujeto más allá del sujeto más al interior. Por ejemplo, si ha seleccionado Colegio, Clase y Estudiante como sujetos en la pestaña Estructura de datos, se crearán los siguientes bloques de efectos aleatorios:

- Efecto aleatorio 1: el sujeto es colegio (sin efectos, sólo interceptación)
- Efecto aleatorio 2: el sujeto es colegio * clase (sin efectos, sólo interceptación)

Puede trabajar con bloques de efectos aleatorios de las maneras siguientes:

1. Para añadir un nuevo bloque, pulse en **Añadir bloque...** Esta acción abre el diálogo "Bloque de efectos aleatorios".
2. Para editar un bloque existente, seleccione el bloque que desee editar y pulse en **Editar bloque...** Esta acción abre el diálogo "Bloque de efectos aleatorios".
3. Para eliminar uno o más bloques, seleccione los bloques que quiera eliminar y pulse en el botón Eliminar.

*Bloque de efectos aleatorios:* Introduzca efectos en el modelo seleccionando uno o más campos en la lista de orígenes y arrastrándolos a la lista de efectos. El tipo de efecto creado depende de la zona activa en la que suelte la selección. Los campos categóricos (marca, nominal y ordinal) se utilizan como factores en el modelo y los campos continuos se utilizan como covariables.

- **Principal.** Los campos que suelte aparecen como efectos principales independientes en la parte inferior de la lista de efectos.
- **2 factores.** Todos los pares posibles de los campos que suelte aparecen como interacciones de 2 factores en la parte inferior de la lista de efectos.
- **3 factores.** Todos los triples posibles de los campos que suelte aparecen como interacciones de 3 factores en la parte inferior de la lista de efectos.
- ***.** La combinación de todos los campos eliminados aparece como una interacción única en la parte inferior de la lista de efectos.

Los botones a la derecha del generador de efectos le permiten realizar diversas acciones.

*Tabla 11. Descripciones de botones de generador de efectos.*

| Icono | Descripción |
|---|---|
|  | Eliminar términos del modelo seleccionando los términos que quiera eliminar y pulsando en el botón Eliminar. |
|  | Reordenar los términos dentro del modelo seleccionando los términos que quiera reordenar y pulsando la flecha arriba o abajo. |

*Tabla 11. Descripciones de botones de generador de efectos (continuación).*

| Icono | Descripción |
|---|---|
|  | Añadir términos anidados al modelo utilizando el diálogo "Añadir un término personalizado" en la página 218, pulsando el botón Añadir un término personalizado. |

**Incluir interceptación.** La interceptación no está incluida en el modelo de efectos aleatorios de forma predeterminada. Si se puede dar por supuesto la lectura de datos en el origen, se puede excluir la interceptación.

**Mostrar predicciones de parámetro para este bloque.** Especifica que se van a mostrar estimaciones de parámetro de efectos aleatorios.

**Definir grupos de covarianzas por.** Los campos categóricos especificados aquí definen conjuntos independientes de parámetros de covarianza de efectos aleatorios; uno para cada categoría definida por la clasificación cruzada de los campos de agrupación. Se puede especificar un conjunto diferente de campos de agrupación para cada bloque de efectos aleatorios. Todos los sujetos tienen el mismo tipo de covarianza; los sujetos con la misma agrupación de covarianza tendrán los mismos valores para los parámetros.

**Combinación de sujetos.** Esto le permite especificar sujetos de efectos aleatorios a partir de combinaciones de sujetos predefinidas desde la pestaña Estructura de datos. Por ejemplo, si *Colegio*, *Clase* y *Estudiante* se definen como sujetos en la pestaña Estructura de datos, en ese orden, entonces la lista desplegable Combinación de sujetos tendrá las opciones **Ninguno**, **Colegio**, **Colegio * Clase** y **Colegio * Clase * Estudiante**.

**Tipo de covarianza de efectos aleatorios.** Esto especifica la estructura de la covarianza de los residuos. Las estructuras disponibles son:

- Autorregresiva de primer orden (AR1)
- Media móvil autorregresiva (1,1) (ARMA11)
- Simetría compuesta
- Diagonal
- Identidad escalada
- Toeplitz
- Sin estructura
- Componentes de la varianza

**Ponderación y desplazamiento: Ponderación de análisis.** El parámetro de escala es un parámetro del modelo estimado relacionado con la varianza de la respuesta. Las ponderaciones de análisis son valores "conocidos" que pueden variar de una observación a otra. Si se especifica el campo Ponderación de análisis, el parámetro de escala, que está relacionado con la varianza de la respuesta, se divide entre los valores de ponderación de análisis para cada observación. Los registros con valores de ponderación de análisis que sean inferiores o iguales a 0 o que sean valores perdidos no se utilizarán en el análisis.

**Desplazamiento.** El término desplazamiento es un predictor "estructural". Su coeficiente no se estima por el modelo pero se supone que tiene el valor 1. Por tanto, los valores del desplazamiento se suman sencillamente al predictor lineal del destino. Esto resulta especialmente útil en los modelos de regresión de Poisson, en los que cada caso puede tener diferentes niveles de exposición al evento de interés.

Por ejemplo, al modelar las tasas de accidente de diferentes conductores, hay una importante diferencia entre un conductor que ha sido el culpable de 1 accidente en 3 años y un conductor que ha sido el culpable de 1 accidente en 25 años. El número de accidentes se puede modelar como una respuesta

Poisson o una binomial negativa con un enlace de registro si el registro natural de la experiencia del conductor se incluye como un término de desplazamiento.

Otras combinaciones de distribución y tipos de enlaces necesitarían otras transformaciones de la variable de desplazamiento.

**Opciones de generación general:** Estas selecciones especifican algunos criterios más avanzados utilizados para generar el modelo.

**Ordenación.** Estos controles determinan el orden de las categorías del objetivo y los factores (entradas categóricas) para determinar la "última" categoría. La configuración de orden de clasificación del objetivo se ignora si el objetivo no es categórico o si se especifica una categoría de referencia personalizada en la configuración de “Objetivo” en la página 215.

**Reglas de parada.** Puede especificar el número máximo de iteraciones que se ejecutará el algoritmo. El algoritmo utiliza un proceso iterativo doble que consta de un bucle interno y un bucle externo. El valor especificado para el número máximo de iteraciones se aplica a ambos bucles. Especifique un número entero no negativo. El valor predeterminado es 100.

**Configuración de estimación posterior.** Estos ajustes determinan el modo en que algunos de los resultados de modelo se calculan para su visualización.

- **Nivel de confianza.** Éste es el nivel de confianza que se utiliza para calcular las estimaciones de intervalos de los coeficientes de modelos. Especifique un valor mayor que 0 y menor que 100. El valor predeterminado es 95.
- **Grados de libertad.** Esto especifica cómo se calculan los grados de libertad para las pruebas de significación. Seleccione **Fijo para todas las pruebas (método residual)** si el tamaño de su muestra es suficientemente grande, si los datos están equilibrados o si el modelo utiliza un tipo de covarianza más simple; por ejemplo, identidad escalada o diagonal. Este es el método predeterminado. Seleccione **Variado en las pruebas (aproximación de Satterthwaite)** si el tamaño de su muestra es pequeño, si los datos no están equilibrados o si el modelo utiliza un tipo de covarianza complicado; por ejemplo, sin estructura.
- **Pruebas de efectos fijos y coeficientes.** Éste es el método para calcular la matriz de covarianzas de estimaciones de parámetros. Seleccione la estimación robusta si le preocupa que se incumplan los supuestos de modelo.

**Estimación:** El algoritmo de generación de modelo utiliza un proceso iterativo que consta de un bucle interno y un bucle externo. Los valores siguientes se aplican al bucle interno.

**Convergencia de parámetros.**
Se asume la convergencia si el cambio máximo absoluto o el cambio máximo relativo en las estimaciones de parámetro es menor que el valor especificado, que no debe ser negativo. El criterio no se utiliza si el valor especificado es igual a 0.

**Convergencia de log-verosimilitud.**
Se asume la convergencia si el cambio absoluto o el cambio relativo en la función de log-verosimilitud es menor que el valor especificado, que no debe ser negativo. El criterio no se utiliza si el valor especificado es igual a 0.

**Convergencia hessiana.**
Para la especificación **Absoluta**, se asume la convergencia si un estadístico basado en la convergencia hessiana es menor que el valor especificado. En el caso de la especificación **Relativa**, se supone la convergencia si el estadístico es menor que el producto del valor especificado y el valor absoluto del logaritmo de la verosimilitud. El criterio no se utiliza si el valor especificado es igual a 0.

**Pasos máximos de puntuación de Fisher.**
Especifique un número entero no negativo. Un valor de 0 especifica el método Newton-Raphson.

Los valores mayores que 0 especifican el uso del algoritmo de puntuación de Fisher hasta el número de iteración *n*, donde *n* es el entero especificado y después de Newton-Raphson.

**Tolerancia para la singularidad.**
Este valor se utiliza como la tolerancia en la comprobación de la singularidad. Especifique un valor positivo.

**Nota:** De forma predeterminada, se utiliza la convergencia de parámetros, donde se marca el cambio **Absoluto** máximo a una tolerancia de 1E-6. Este valor podría generar resultados que difieren de los resultados obtenidos en versiones anteriores a la versión 22. Para generar los resultados desde versiones anteriores a la 22, utilice **Relativo** para el criterio de convergencia de parámetros y mantener el valor de tolerancia predeterminado de 1E-6.

**General: Nombre del modelo.** Puede generar el nombre del modelo automáticamente tomando como base los campos objetivo o especificar un nombre personalizado. El nombre generado automáticamente es el nombre del campo objetivo. Si existen objetivos múltiples, el nombre del modelo se forma con los nombres de campos en orden, conectados por símbolos &. Por ejemplo, si *campo1 campo2 campo3* son objetivos, el nombre de modelo es: *campo1 & campo2 & campo3*.

**Dejar disponible para puntuación.** Cuando se puntúa el modelo, se crearán los elementos seleccionados en este grupo. El valor predicho (para todos los objetivos) y la confianza (para objetivos categóricos) se calculan siempre cuando se puntúa el modelo. La confianza calculada puede basarse en la probabilidad del valor predicho (la probabilidad predicha más alta) o la diferencia entre la probabilidad predicha más alta y la segunda probabilidad predicha más alta.

- **Probabilidad predicha para objetivos categóricos.** Genera las probabilidades predichas para objetivos categóricos. Se crea un campo para cada categoría.
- **Puntuaciones de propensión para objetivos de marca.** En el caso de modelos con un objetivo de marca (que devuelve una predicción de sí o no), puede solicitar puntuaciones de propensión que indican la probabilidad del resultado true especificado para el campo objetivo. El modelo produce puntuaciones de propensión en bruto; si hay particiones activas, el modelo también producirá puntuaciones de propensión ajustadas en función de la partición de prueba.

**Medias estimadas:** Esta pestaña le permite mostrar las medias marginales estimadas para niveles de factores e interacciones de factores. Las medias marginales estimadas no están disponibles para modelos multinomiales.

**Términos.** Los términos de modelo de Efectos fijos que se componen exclusivamente de campos categóricos se enumeran aquí. Compruebe cada término para el que quiera que el modelo produzca medias marginales estimadas.

- **Tipo de contraste.** Esto especifica el tipo de contraste que debe utilizarse para los niveles del campo Contraste. Si se selecciona **Ninguno**, no se produce ningún contraste. **Por parejas** produce comparaciones por parejas para todas las combinaciones de niveles de los factores especificados. Este contraste es el único disponible para las interacciones de los factores. **Contrastes de desviación** comparan cada nivel del factor con la media global. **Contrastes simples** comparan cada nivel del factor, excepto el último, con el último nivel. El "último" nivel está determinado por la ordenación de los factores especificada en Opciones de generación. Tenga en cuenta que todos estos tipos de contrastes no son ortogonales.
- **Campo Contraste.** Esto especifica un factor, cuyos niveles se comparan mediante el tipo de contraste seleccionado. Si se selecciona **Ninguno** como tipo de contraste, no se puede (o no es necesario) seleccionar ningún campo Contraste.

**Campos continuos.** Los campos continuos enumerados se extraen de los términos de Efectos fijos que utilizan campos continuos. Al calcular medias marginales estimadas, las covariables están fijas en los valores especificados. Seleccione la media o especifique un valor personalizado.

**Mostrar medias estimadas en cuanto a.** Esto especifica si las medias marginales estimadas se calculan basándose en la escala original del objetivo o basándose en la transformación de la función de enlace. **Escala de objetivo original** calcula las medias marginales estimadas para el objetivo. Tenga en cuenta que cuando el objetivo se especifica mediante la opción Eventos/Ensayos, proporciona la media marginal estimada de la proporción de eventos/ensayos en lugar de la del número de eventos. **Transformación de la función de enlace** calcula la media marginal estimada del predictor lineal.

**Ajuste de comparaciones múltiples mediante.** Al realizar contrastes de hipótesis con varios contrastes, el nivel de significación global se puede ajustar utilizando los niveles de significación de los contrastes incluidos. Esto le permite seleccionar el método de ajuste.

- **Diferencia menos significativa.** Este método no controla la probabilidad general de rechazar las hipótesis de que algunos contrastes lineales son diferentes a los valores de hipótesis nula.
- *Bonferroni secuencial*. Éste es un procedimiento de Bonferroni de rechazo secuencial decreciente que es mucho menos conservador en cuanto al rechazo de hipótesis individuales pero que mantiene el mismo nivel de significación global.
- *Sidak secuencial*. Este es un procedimiento de Sidak de rechazo secuencial decreciente que es mucho menos conservador en términos de rechazar las hipótesis individuales pero que mantiene el mismo nivel de significación global.

El método de diferencia menos significativa es menos conservador que el método Sidak secuencial, que a su vez es menos conservador que Bonferroni secuencial; es decir, la diferencia menos significativa rechazará al menos tantas hipótesis individuales como Sidak secuencial, que a su vez rechazará al menos tantas hipótesis individuales como Bonferroni secuencial.

**Vista de modelo:** De forma predeterminada, se muestra la vista Resumen del modelo. Para ver otra vista de modelo, selecciónela entre las vistas en miniatura.

*Resumen del modelo:* Esta vista es una instantánea, un resumen visual del modelo y su ajuste.

**Tabla.** La tabla identifica el objetivo, la distribución de probabilidad y la función de enlace especificados en la Configuración de objetivo. Si el objetivo se define mediante eventos y ensayos, la casilla se divide para mostrar el campo Eventos y el campo Ensayos o el número fijo de ensayos. Además, se muestran el criterio de información de Akaike corregido para muestras finitas (AICC) y el criterio de información bayesiano (BIC).

- *Akaike corregido*. Una medida para seleccionar y comparar modelos mixtos basada en la log-verosimilitud -2 (restringida). Los valores menores indican modelos mejores. El AICC "corrige" el AIC respecto a tamaños muestrales pequeños. A medida que aumenta el tamaño de la muestra, el AICC converge con el AIC.
- *Bayesiano*. Una medida para seleccionar y comparar modelos basados en el logaritmo de la verosimilitud -2. Los valores menores indican modelos mejores. El BIC también "penaliza" modelos sobreparametrizados (modelos complejos con un gran número de entradas, por ejemplo), pero de forma más estricta que el AIC.

**Gráfico.** Si el objetivo es categórico, un gráfico muestra la precisión del modelo final, que es el porcentaje de clasificaciones correctas.

*Estructura de datos:* Esta vista proporciona un resumen de la estructura de datos que especifique y le ayuda a comprobar que los sujetos y las medidas repetidas se han especificado correctamente. La información observada para el primer sujeto se muestra para cada campo de sujeto y campo de medidas repetidas, así como el objetivo. Además, se muestra el número de niveles de cada campo de sujeto y campo de medidas repetidas.

*Predicho por observado:* Para objetivos continuos, incluidos objetivos especificados como eventos/ensayos, muestra un diagrama de dispersión en intervalos de los valores predichos en el eje vertical por los

valores observados en el eje horizontal. Idealmente, los puntos deben basarse en una línea de 45 grados; esta vista indica si hay algún registro predicho de manera incorrecta en el modelo.

*Clasificación:* Para los objetivos categóricos, muestra la clasificación cruzada de los valores observados en contraposición a los predichos en el mapa de calor, junto con el porcentaje global correcto.

**Estilos de tabla.** Existen varios estilos de visualización diferentes, que son accesibles desde la lista desplegable **Estilo**.

- **Porcentajes de fila.** Muestra los porcentajes de filas (la casilla cuenta lo expresado como un porcentaje de los totales de filas) en las casillas. Este es el método predeterminado.
- **Recuentos de casillas.** Muestra los recuentos de casillas en las casillas. El sombreado del mapa de calor se basa aún en los porcentajes de filas.
- **Mapa de calor.** No muestra valores en las casillas, solamente el sombreado.
- **Comprimido.** No muestra cabeceras de filas o columnas, ni valores en las casillas. Puede ser útil cuando el objetivo tiene muchas categorías.

**Perdidos.** Si cualquier registro tiene valores perdidos en el objetivo, se muestran en una fila **(Perdidos)** bajo todas las filas válidas. Los registros con valores perdidos no contribuyen al porcentaje global correcto.

**Objetivos múltiples.** Si existen varios objetivos categóricos, cada objetivo se muestra en una tabla separada y hay una lista desplegable de **Objetivos** que controla qué objetivos mostrar.

**Tablas grandes.** Si el objetivo mostrado tiene más de 100 categorías, no se mostrará ninguna tabla.

*Efectos fijos:* Esta vista muestra el tamaño de cada efecto fijo en el modelo.

**Estilos.** Existen varios estilos de visualización diferentes, que son accesibles desde la lista desplegable **Estilo**.

- **Diagrama.** Éste es un gráfico en el que los efectos están ordenados de arriba a abajo en el orden en que se especificaron en la configuración de Efectos fijos. Las líneas de conexión del diagrama se ponderan tomando como base la significación del efecto, con un grosor de línea mayor correspondiente a efectos con mayor significación (valores *p* inferiores). Este es el método predeterminado.
- **Tabla.** Se trata de una tabla ANOVA para el modelo completo y los efectos de modelo individuales. Los efectos individuales están ordenados de arriba a abajo en el orden en que se especificaron en la configuración de Efectos fijos.

**Significación.** Existe un control deslizante Significación que controla qué efectos se muestran en la vista. Se ocultan los efectos con valores de significación superiores al valor del control deslizante. Esto no cambia el modelo, simplemente le permite centrarse en los efectos más importantes. El valor predeterminado es 1.00, de modo que no se filtran efectos tomando como base la significación.

*Coeficientes fijos:* Esta vista muestra el valor de cada coeficiente fijo en el modelo. Tenga en cuenta que los factores (predictores categóricos) tienen codificación de indicador dentro del modelo, de modo que los **efectos** que contienen los factores generalmente tendrán múltiples **coeficientes** asociados: uno por cada categoría exceptuando la categoría que corresponde al coeficiente redundante.

**Estilos.** Existen varios estilos de visualización diferentes, que son accesibles desde la lista desplegable **Estilo**.

- **Diagrama.** Éste es un gráfico que muestra la interceptación en primer lugar y luego ordena los efectos de arriba a abajo en el orden en que se especificaron en la configuración de Efectos fijos. Dentro de los efectos que contienen factores, los coeficientes se clasifican en orden ascendente de valores de datos. Las líneas de conexión del diagrama se colorean y se ponderan tomando como base la significación del

coeficiente, con un grosor de línea mayor correspondiente a coeficientes con mayor significación (valores *p* inferiores). Este es el estilo predeterminado.

- **Tabla.** Muestra los valores, las pruebas de significación y los intervalos de confianza para los coeficientes de modelos individuales. Después de la interceptación, los efectos están ordenados de arriba a abajo en el orden en que se especificaron en la configuración de Efectos fijos. Dentro de los efectos que contienen factores, los coeficientes se clasifican en orden ascendente de valores de datos.

**Multinomial.** Si la distribución multinomial es efectiva, la lista desplegable Multinomial controla qué categoría objetivo se muestra. La ordenación de los valores de la lista está determinada por la especificación en la configuración de Opciones de generación.

**Exponencial.** Esto muestra estimaciones de coeficientes exponenciales e intervalos de confianza para determinados tipos de modelos, incluidos la regresión logística binaria (distribución binomial y enlace logit), la regresión logística nominal (distribución multinomial y enlace logit), la regresión binomial negativa (distribución binomial negativa y enlace log) y el modelo lineal del logaritmo (distribución de Poisson y enlace log).

**Significación.** Existe un control deslizante Significación que controla qué coeficientes se muestran en la vista. Se ocultan los coeficientes con valores de significación superiores al valor del control deslizante. Esto no cambia el modelo, simplemente le permite centrarse en los coeficientes más importantes. El valor predeterminado es 1.00, de modo que no se filtran coeficientes tomando como base la significación.

*Covarianzas de efectos aleatorios:* Esta vista muestra la matriz de covarianzas de efectos aleatorios (**G**).

**Estilos.** Existen varios estilos de visualización diferentes, que son accesibles desde la lista desplegable **Estilo**.

- **Valores de covarianzas.** Éste es un mapa de calor de la matriz de covarianzas en el que los efectos están ordenados de arriba a abajo en el orden en que se especificaron en la configuración de Efectos fijos. Los colores del Corrgram se corresponden con los valores de las casillas que se muestran en la leyenda. Este es el método predeterminado.
- **Corrgram.** Éste es un mapa de calor de la matriz de covarianzas.
- **Comprimido.** Éste es un mapa de calor de la matriz de covarianzas sin las cabeceras de fila y columna.

**Bloques.** Si hay varios bloques de efectos aleatorios, existe una lista desplegable Bloque para seleccionar el bloque que se muestra.

**Grupos.** Si un bloque de efectos aleatorios tiene una especificación de grupo, existe una lista desplegable Grupo para seleccionar el nivel de grupo que se muestra.

**Multinomial.** Si la distribución multinomial es efectiva, la lista desplegable Multinomial controla qué categoría objetivo se muestra. La ordenación de los valores de la lista está determinada por la especificación en la configuración de Opciones de generación.

*Parámetros de covarianza:* Esta vista muestra las estimaciones de parámetros de covarianza y los estadísticos relacionados para los efectos residuales y aleatorios. Estos son resultados avanzados, pero fundamentales, que proporcionan información sobre si la estructura de la covarianza es adecuada.

**Tabla de resumen.** Ésta es una referencia rápida al número de parámetros en las matrices de covarianza de efectos residuales (**R**) y aleatorios (**G**), el rango (número de columnas) en las matrices de diseño de efectos fijos (**X**) y efectos aleatorios (**Z**) y el número de sujetos definidos por los campos de sujeto que definen la estructura de datos.

**Tabla Parámetro de covarianza.** Para el efecto seleccionado, la estimación, el error estándar y el intervalo de confianza se muestran para cada parámetro de covarianza. El número de parámetros que se muestra

depende de la estructura de la covarianza del efecto y, en el caso de bloques de efectos aleatorios, el número de efectos del bloque. Si observa que los parámetros de fuera de la diagonal no son significativos, tal vez pueda utilizar una estructura de covarianza más simple.

**Efectos.** Si hay bloques de efectos aleatorios, existe una lista desplegable Efecto para seleccionar el bloque de efectos residuales o aleatorios que se muestra. El efecto residual siempre está disponible.

**Grupos.** Si un bloque de efectos residuales o aleatorios tiene una especificación de grupo, existe una lista desplegable Grupo para seleccionar el nivel de grupo que se muestra.

**Multinomial.** Si la distribución multinomial es efectiva, la lista desplegable Multinomial controla qué categoría objetivo se muestra. La ordenación de los valores de la lista está determinada por la especificación en la configuración de Opciones de generación.

*Medias estimadas: Efectos significativos:* Estos son gráficos que se muestran para los 10 efectos de todos los factores fijos "más significativos", comenzando por las interacciones de 3 factores, seguidas de las interacciones de 2 factores y, por último, los efectos principales. El gráfico muestra el valor estimado por el modelo del objetivo en el eje vertical para cada valor del efecto principal (o el primer efecto enumerado en una interacción) en el eje horizontal; se genera una línea independiente para cada valor del segundo efecto enumerado en una interacción; se genera un gráfico independiente para cada valor del tercer efecto enumerado en una interacción de 3 factores; el resto de predictores se mantiene constante. Proporciona una visualización útil de los efectos de los coeficientes de cada predictor en el objetivo. Tenga en cuenta que si no hay predictores significativos, no se generan medias estimadas.

**Confianza.** Esto muestra los límites de confianza superior e inferior para las medias marginales, mediante el nivel de confianza especificado como parte de Opciones de generación.

*Medias estimadas: Efectos personalizados:* Estas son tablas y gráficos para efectos de todos los factores fijos solicitados por los usuarios.

**Estilos.** Existen varios estilos de visualización diferentes, que son accesibles desde la lista desplegable **Estilo**.

- **Diagrama.** Este estilo muestra un gráfico de líneas del valor estimado por el modelo del objetivo en el eje vertical para cada valor del efecto principal (o el primer efecto enumerado en una interacción) en el eje horizontal; se genera una línea independiente para cada valor del segundo efecto enumerado en una interacción; se genera un gráfico independiente para cada valor del tercer efecto enumerado en una interacción de 3 factores; el resto de predictores se mantiene constante.

  Si se solicitan contrastes, se muestra otro gráfico para comparar los niveles del campo Contraste; para las interacciones, se muestra un gráfico para cada combinación de niveles de los efectos distintos del campo Contraste. Para contrastes **por parejas**, es un gráfico de red de distancia; es decir, una representación gráfica de la tabla de comparaciones en la que las distancias entre los nodos de la red se corresponden con las diferencias entre las muestras. Las líneas amarillas se corresponden con diferencias significativas estadísticamente; las líneas negras se corresponden con diferencias no significativas. Al pasar el ratón por encima de una línea de la red, aparece información sobre herramientas con el significado ajustado de la diferencia entre los nodos conectados por la línea.

  Para **Contrastes de desviación**, se muestra un gráfico de barras con el valor estimado por el modelo del objetivo en el eje vertical y los valores del campo Contraste en el eje horizontal; para las interacciones, se muestra un gráfico para cada combinación de niveles de los efectos distintos del campo Contraste. Las barras muestran la diferencia entre cada nivel del campo Contraste y la media global, que está representada por una línea horizontal negra.

  Para **Contrastes simples**, se muestra un gráfico de barras con el valor estimado por el modelo del objetivo en el eje vertical y los valores del campo Contraste en el eje horizontal; para las interacciones, se muestra un gráfico para cada combinación de niveles de los efectos distintos del campo Contraste. Las barras muestran la diferencia entre cada nivel del campo Contraste (excepto el último) y el último nivel, que está representado por una línea horizontal negra.

- **Tabla.** Este estilo muestra una tabla del valor estimado por el modelo del objetivo, su error estándar y el intervalo de confianza para cada combinación de niveles de los campos del efecto; el resto de predictores se mantiene constante.

  Si se solicitan contrastes, se muestra otra tabla con la estimación, el error estándar, la prueba de significación y el intervalo de confianza para cada contraste; para las interacciones, hay un conjunto de filas independiente para cada combinación de niveles de los efectos distintos del campo Contraste. Además, se muestra una tabla con los resultados de las pruebas globales; para las interacciones, hay una prueba global independiente para cada combinación de niveles de los efectos distintos del campo Contraste.

**Confianza.** Esto cambia la visualización de los límites de confianza superior e inferior para las medias marginales, mediante el nivel de confianza especificado como parte de Opciones de generación.

**Diseño.** Esto cambia el diseño del diagrama de contrastes por parejas. El diseño circular muestra menos de los contrastes que el diseño de red, pero evita que se superpongan las líneas.

*Configuración:* Cuando se puntúa el modelo, se crearán los elementos seleccionados en esta pestaña. El valor predicho (para todos los objetivos) y la confianza (para objetivos categóricos) se calculan siempre cuando se puntúa el modelo. La confianza calculada puede basarse en la probabilidad del valor predicho (la probabilidad predicha más alta) o la diferencia entre la probabilidad predicha más alta y la segunda probabilidad predicha más alta.

- **Probabilidad predicha para objetivos categóricos.** Genera las probabilidades predichas para objetivos categóricos. Se crea un campo para cada categoría.
- **Puntuaciones de propensión para objetivos de marca.** En el caso de modelos con un objetivo de marca (que devuelve una predicción de sí o no), puede solicitar puntuaciones de propensión que indican la probabilidad del resultado true especificado para el campo objetivo. El modelo produce puntuaciones de propensión en bruto; si hay particiones activas, el modelo también producirá puntuaciones de propensión ajustadas en función de la partición de prueba.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

## Nodo GLE

El modelo GLE identifica la variable dependiente que está relacionada linealmente a los factores y covariables a través de una función de enlace especificada. Además, el modelo permite que la variable dependiente tenga una distribución no normal. Cubre los modelos estadísticos más comunes, como la regresión lineal para respuestas distribuidas normalmente, los modelos logísticos para datos binarios, el modelo lineal de logaritmo para datos de frecuencias, modelos log-log complementarios para datos de supervivencia censurados por intervalos y numerosos modelos estadísticos a través de su formulación general de modelos

**Ejemplos.** Una compañía de transporte puede utilizar modelos lineales generalizados para ajustar una regresión de Poisson a las frecuencias de daños de varios tipos de barcos construidos en varios períodos de tiempo. El modelo resultante puede ayudar a determinar cuales son los tipos de barcos más propensos a sufrir daños.

Una compañía de seguros de automóviles puede utilizar modelos lineales generalizados para ajustar una regresión gamma a las reclamaciones por daños de los automóviles. El modelo resultante puede ayudar a determinar los factores que más contribuyen al tamaño de la reclamación.

Los investigadores médicos pueden utilizar modelos lineales generalizados para ajustar una regresión log-log complementario a los datos de supervivencia censurados por intervalos para predecir el tiempo que tardará en reaparecer una enfermedad.

Los modelos GLE funcionan generando una ecuación que relaciona los valores de campo de entrada con los valores de campo de salida. Una vez se ha generado el modelo, se puede utilizar para calcular los valores de datos nuevos.

Para un objetivo categórico, para cada registro, se calcula una probabilidad de pertenencia para cada categoría de salida posible. La categoría objetivo con la probabilidad más alta se asigna como el valor de salida predicho para cada registro.

**Requisitos.** Necesita uno o más campos de entrada y exactamente un campo objetivo (que puede tener un nivel de medición de *Continuo*, *Categórico* o *Marca*) con dos o más categorías. Los tipos de los campos utilizados en el modelo deben estar completamente instanciados.

## Objetivo

Estos ajustes definen el objetivo, su distribución y su relación con los predictores mediante la función de enlace.

**Objetivo** El objetivo es obligatorio. Puede tener cualquier nivel de medición y el nivel de medición del destino afecta a las distribuciones y funciones de enlace que son adecuadas.

- **Utilizar destino predefinido** Para utilizar los valores de destino desde un nodo Tipo situado en un punto anterior de la ruta (o la pestaña Tipos de un nodo de origen situado en un punto anterior de la ruta), seleccione esta opción.
- **Utilizar destino personalizado** Para asignar manualmente un destino, seleccione esta opción.
- **Utilice el número de ensayos como denominador** Cuando la respuesta prevista es un número de eventos que se producen en un conjunto de ensayos, el campo objetivo contiene el número de eventos y puede seleccionar un campo adicional que contiene el número de ensayos. Por ejemplo, al probar un nuevo pesticida puede que exponga muestras de hormigas a diferentes concentraciones del pesticida y, a continuación, registre el número de hormigas muertas y el número de hormigas de cada muestra. En este caso, el campo que registra el número de hormigas muertas debe especificarse como el campo objetivo (eventos) y el campo que registre el número de hormigas de cada muestra debe especificarse como el campo de ensayos. Si el número de hormigas es el mismo para cada muestra, entonces el número de ensayos puede especificarse mediante un valor fijo.

  El número de ensayos debe ser superior o igual al número de eventos de cada registro. Los eventos deben ser enteros no negativos y los ensayos deben ser enteros positivos.
- **Personalizar categoría de referencia**. Para un objetivo categórico, puede seleccionar la categoría de referencia. Esto puede afectar a determinados resultados, como las estimaciones de los parámetros, pero no debería cambiar el ajuste del modelo. Por ejemplo, si su objetivo toma los valores 0, 1 y 2, de forma predeterminada, el procedimiento convierte la última categoría (el valor más alto), o 2, en la categoría de referencia. En esta situación, las estimaciones de los parámetros deben interpretarse como relacionadas con la verosimilitud de la categoría 0 o 1 *relativa* a la verosimilitud de la categoría 2. Si especifica una categoría personalizada y su objetivo tiene etiquetas definidas, puede establecer la

categoría de referencia seleccionando un valor de la lista. Esto puede resultar cómodo cuando, al especificar un modelo, no se recuerda exactamente cómo se codificó un determinado campo.

**Distribución de destino y relación (enlace) con el modelo lineal** Teniendo en cuenta los valores de los predictores, el modelo espera que la distribución de los valores del objetivo siga la forma especificada, y que los valores de objetivo estén relacionados de forma lineal con los predictores a través de la función de enlace especificada. Se proporcionan los accesos directos de varios modelos comunes o seleccione un ajuste **Personalizado** si hay una combinación específica de distribución y función de enlace que desee ajustar y que no esté en la lista corta.

- **Modelo lineal** Especifica una distribución normal con un enlace de identidad, que es útil cuando el objetivo se puede pronosticar utilizando una regresión lineal o un modelo ANOVA.
- **Regresión gamma** Especifica una distribución gamma con un enlace log, que se deberá utilizar cuando el objetivo contiene todos los valores positivos y es asimétrico a valores mayores.
- **Loglinear** Especifica una distribución Poisson con un enlace log,que se deberá utilizar cuando el objetivo representa un recuento de apariciones en un periodo de tiempo fijo.
- **Regresión binomial negativa** Especifica una distribución binomial negativa con un enlace log, que se deberá utilizar cuando el objetivo y el denominador representan el número de ensayos necesarios para observar *k* éxitos.
- **Regresión Tweedie** Especifica una distribución Tweedie con funciones de enlace de potencia, logaritmo o identidad y son útiles para modelar respuestas que son una combinación de ceros y valores reales positivos. Estas distribuciones también se denominan *Poisson compuesto*, *gamma compuesto* y *Poisson-gamma*.
- **Regresión logística multinomial** Especifica una distribución multinomial, que se deberá utilizar cuando el objetivo es una respuesta de varias categorías. Utiliza un enlace logit acumulado (resultados ordinales) o un enlace logit generalizado (respuestas nominales de categorías múltiples).
- **Regresión logística binaria** Especifica una distribución binomial con un enlace logit, que se deberá utilizar cuando el objetivo es una repuesta binaria pronosticada por un modelo de regresión logística.
- **Probit binario** Especifica una distribución binomial con un enlace probit, que se deberá utilizar cuando el objetivo es una respuesta binaria con una distribución normal subyacente.
- **Supervivencia censurada por intervalos** Especifica una distribución binomial con un enlace log-log complementario, que es útil en el análisis de supervivencia cuando algunas observaciones no tienen ningún evento de terminación.
- **Personalizado** Especifique su propia combinación de la función de distribución y enlace.

## Distribución

Esta selección especifica la **Distribución** del objetivo. La capacidad de especificar una distribución no normal y una función de enlace que no sea la identidad es la mejora principal del modelo lineal generalizado con respecto al modelo lineal. Hay muchas combinaciones posibles de distribución y función de enlace, varias de las cuales pueden ser adecuadas para un determinado conjunto de datos, por lo que su elección puede estar guiada por consideraciones teóricas a priori y por las combinaciones que parezcan funcionar mejor.

- **Automático** Si no está seguro de qué distribución a utilizar, seleccione esta opción; el nodo analiza los datos para calcular y aplicar el mejor método de distribución.
- **Binomial** Esta distribución solo es apropiada para un objetivo que representa una respuesta binaria o un número de eventos.
- **Gamma** Esta distribución es apropiada para un objetivo con valores de escala positivos que se desvían hacia valores positivos mayores. Si un valor de datos es menor o igual que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis.
- **De Gauss inversa** Esta distribución es apropiada para un objetivo con valores de escala positivos que se desvían hacia valores positivos mayores. Si un valor de datos es menor o igual que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis.

- **Multinomial** Esta distribución es apropiada para un objetivo que representa una respuesta de varias categorías. La forma del modelo dependerá del nivel de medición del objetivo.

  Un objetivo **nominal** dará como resultado un modelo multinomial nominal en el que se calcula un conjunto independiente de parámetros del modelo para cada categoría del objetivo (excepto la categoría de referencia). Las estimaciones de parámetros de un predictor determinado muestran la relación entre ese predictor y la verosimilitud de cada categoría del objetivo, relativa a la categoría de referencia.

  Un objetivo **ordinal** dará como resultado un modelo multinomial ordinal en el que el término de interceptación tradicional se sustituye por un conjunto de parámetros de **umbral** que se relacionan con la probabilidad acumulada de las categorías objetivo.
- **Binomial negativa** La regresión binomial negativa utiliza una distribución binomial negativa con un enlace de logaritmo, que se debe utilizar cuando el objetivo representa un recuento de apariciones con una varianza elevada.
- **Normal** Esto es apropiado para un objetivo continuo cuyos valores adoptan una distribución simétrica en forma de campana en torno a un valor central (media).
- **Poisson** Esta distribución se puede considerar el número de apariciones de un evento de interés en un periodo de tiempo fijo y es apropiado para variables con valores de entero que no sean negativos. Si un valor de datos no es entero, es menor que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis.
- **Tweedie** Esta distribución es adecuada para las variables que se pueden representar mediante mezclas de Poisson de distribuciones gamma; la distribución está "mezclada" en el sentido de que combina propiedades de distribuciones continuas (toma valores reales no negativos) y discretas (masa de probabilidad positiva en un único valor, 0). La variable dependiente debe ser numérica, con valores de datos mayores o iguales que cero. Si un valor de datos es menor que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis. El valor fijo del parámetro de distribución Tweedie puede ser cualquier número mayor que uno y menor que dos.

## Funciones de enlace

La **función de enlace** es una transformación del objetivo que permite la estimación del modelo. Se encuentran disponibles las siguientes funciones:

- **Automático** Si no está seguro de qué enlace utilizar, seleccione esta opción; el nodo analiza los datos para calcular y aplicar la mejor función de enlace.
- **Identidad** $f(x)=x$. El destino no se transforma. Este enlace se puede utilizar con cualquier distribución, excepto la multinomial.
- **log-log complementario** $f(x)=\log(-\log(1-x))$. Adecuado solamente con la distribución binomial o multinomial.
- **Cauchit** $f(x) = \tan(\pi\ (x - 0.5))$. Adecuado solamente con la distribución binomial o multinomial.
- **Log** $f(x)=\log(x)$. Este enlace se puede utilizar con cualquier distribución, excepto la multinomial.
- **Complemento log** $f(x)=\log(1-x)$. Adecuado solamente con la distribución binomial.
- **Logit** $f(x)=\log(x\ /\ (1-x))$. Adecuado solamente con la distribución binomial o multinomial.
- **log-log negativo** $f(x)=-\log(-\log(x))$. Adecuado solamente con la distribución binomial o multinomial.
- **Probit** $f(x)=\Phi^{-1}(x)$, donde $^{-1}$ es la función de distribución acumulada normal estándar inversa. Adecuado solamente con la distribución binomial o multinomial.
- **Potencia** $f(x)=x^{\alpha}$, if $\alpha \neq 0$. $f(x)=\log(x)$, si $\alpha=0$. $\alpha$ es la especificación de número necesaria y debe ser un número real. Este enlace se puede utilizar con cualquier distribución, excepto la multinomial.

**Parámetro para Tweedie** Sólo disponible si ha seleccionado el botón de selección **Regresión Tweedie** o Tweedie como el método de **Distribución**. Seleccione un valor entre 1 y 2.

## Efectos de modelo

Los factores de efectos fijos suelen considerarse campos cuyos valores de interés están todos representados en el conjunto de datos y pueden utilizarse para la puntuación. De forma predeterminada, los campos con el rol de entrada predefinido que no se especifican en ningún otro sitio del cuadro de diálogo se introducen en la parte de efectos fijos del modelo. Los campos categóricos (distintivo, nominal y ordinal) se utilizan como factores en el modelo y los campos continuos se utilizan como covariables.

Introduzca efectos en el modelo seleccionando uno o más campos en la lista de orígenes y arrastrándolos a la lista de efectos. El tipo de efecto creado depende de la zona activa en la que suelte la selección.

- **Principal** Los campos descartados aparecen como efectos principales independientes en la parte inferior de la lista de efectos.
- **2 factores** Todos los pares posibles de los campos descartados aparecen como interacciones de 2 factores en la parte inferior de la lista de efectos.
- **3 factores** Todos los triples posibles de los campos descartados aparecen como interacciones de 3 factores en la parte inferior de la lista de efectos.
- * La combinación de todos los campos descartados aparecen como una sola interacción en la parte inferior de la lista de efectos.

Los botones a la derecha del generador de efectos le permiten realizar diversas acciones.

*Tabla 12. Descripciones de botones de generador de efectos*

| Icono | Descripción |
|---|---|
|  | Eliminar términos del modelo de efectos fijos seleccionando los términos que desea eliminar y pulsando en el botón eliminar. |
|  | Reordenar los términos dentro del modelo de efectos fijos seleccionando los términos que quiera reordenar y pulsando en la flecha arriba o abajo. |
|  | Añadir términos anidados al modelo utilizando el diálogo Añadir una término personalizado, pulsando el botón Añadir un término personalizado. |

**Incluir interceptación** La interceptación normalmente se incluye en el modelo. Si se puede dar por supuesto la lectura de datos en el origen, se puede excluir la interceptación.

### Añadir un término personalizado

Puede generar términos anidados para su modelo en este procedimiento. Los términos anidados son útiles para modelar el efecto de un factor o covariable cuyos valores no interactúen con los niveles de otro factor. Por ejemplo, una cadena de supermercados puede seguir los hábitos de consumo de sus clientes en varias ubicaciones de sus tiendas. Puesto que cada cliente solo frecuenta una de estas ubicaciones, se puede decir que el efecto del Cliente *está anidado* dentro del efecto de la ubicación de la tienda.

Además, puede incluir efectos de interacción, como términos polinómicos que implican a la misma covariable, o añadir varios niveles de anidación al término anidado.

**Limitaciones.** Los términos anidados tienen las siguientes restricciones:

- Todos los factores incluidos en una interacción deben ser exclusivos entre sí. Por consiguiente, si *A* es un factor, no es válido especificar *A*A*.

- Todos los factores incluidos en un efecto anidado deben ser exclusivos entre sí. Por consiguiente, si *A* es un factor, no es válido especificar *A(A)*.
- No se puede anidar ningún efecto dentro de una covariable. Por consiguiente, si *A* es un factor y *X* es una covariable, no es válido especificar *A(X)*.

### Creación de un término anidado

1. Seleccione un factor o covariable que esté anidado en otro factor y, a continuación, pulse en el botón de flecha.
2. Pulse en **(Dentro)**.
3. Seleccione el factor dentro del cual el factor o covariable anterior se anida y, a continuación, pulse en el botón de flecha.
4. Pulse en **Añadir término**.

Si lo desea, puede incluir efectos de interacción o añadir varios niveles de anidación al término anidado.

## Ponderación y desplazamiento

**Ponderación de análisis** El parámetro de escala es un parámetro de modelo estimado relacionado con la varianza de la respuesta. Las ponderaciones de análisis son valores "conocidos" que pueden variar de una observación a otra. Si se especifica el campo **Ponderación de análisis**, el parámetro de escala, que está relacionado con la varianza de la respuesta, se divide por los valores de ponderación de análisis para cada observación. Los registros con valores de ponderación de análisis que sean inferiores o iguales a 0, o que falten, no se utilizan en el análisis.

**Desplazamiento** El término desplazamiento es un predictor *estructural*. Su coeficiente no se estima por el modelo pero se supone que tiene el valor 1. Por tanto, los valores del desplazamiento se suman sencillamente al predictor lineal del destino. Esto resulta especialmente útil en los modelos de regresión de Poisson, en los que cada caso puede tener diferentes niveles de exposición al evento de interés.

Por ejemplo, al modelar las tasas de accidente de diferentes conductores, hay una importante diferencia entre un conductor que ha sido culpable de 1 accidente en 3 años y un conductor que ha sido culpable de 1 accidente en 25 años. El número de accidentes se puede modelar como una respuesta Poisson o una binomial negativa con un enlace de registro si el registro natural de la experiencia del conductor se incluye como un término de desplazamiento.

Otras combinaciones de distribución y tipos de enlaces necesitarían otras transformaciones de la variable de desplazamiento.

## Opciones de generación

Estas selecciones especifican algunos criterios más avanzados utilizados para generar el modelo.

**Orden de clasificación** Estos controles determinan el orden de las categorías para el objetivo y los factores (entradas categóricas) para determinar la "última categoría". La configuración de orden de clasificación del objetivo se ignora si el objetivo no es categórico o si se especifica una categoría de referencia personalizada en la configuración de “Objetivo” en la página 228.

**Configuración de estimación posterior** Estos valores determinan cómo se calcula algún resultado del modelo para la visualización.

- **% de nivel de confianza** Este es el nivel de confianza utilizado para calcular estimaciones de intervalo de los coeficientes de modelo. Especifique un valor mayor que 0 y menor que 100. El valor predeterminado es 95.
- **Grados de libertad** Esto especifica cómo se calculan los grados de libertad para pruebas de significancia. Seleccione **Fijo para todas las pruebas (método residual)** si el tamaño de su muestra es suficientemente grande, si los datos están equilibrados o si el modelo utiliza un tipo de covarianza más

simple; por ejemplo, identidad escalada o diagonal. Este es el método predeterminado. Seleccione **Variado en las pruebas (aproximación de Satterthwaite)** si el tamaño de su muestra es pequeño, si los datos no están equilibrados o si el modelo utiliza un tipo de covarianza complicado; por ejemplo, sin estructura.

- **Pruebas de efectos fijos y coeficientes.**. Éste es el método para calcular la matriz de covarianzas de estimaciones de parámetros. Seleccione la estimación robusta si le preocupa que se incumplan los supuestos de modelo.

**Detectar valores atípicos influyentes** Para todas las distribuciones excepto la distribución multinomial, seleccione esta opción para identificar valores atípicos influyentes.

**Realizar análisis de tendencias** Para un diagrama de dispersión, seleccione esta opción para realizar análisis de tendencias.

## Estimación

**Método** Seleccione el método de estimación de probabilidad máxima que se debe utilizar; las opciones disponibles son:

- Puntuación de Fisher
- Newton-Raphson
- Híbrido

**Número máximo de iteraciones Fisher** Especifique un entero no negativo. Un valor de 0 especifica el método Newton-Raphson. Los valores mayores que 0 especifican el uso del algoritmo de puntuación de Fisher hasta el número de iteración *n*, donde *n* es el entero especificado y después de Newton-Raphson.

**Método de parámetro de escala** Seleccione el método para la estimación del parámetro de escala; las opciones disponibles son:

- Estimación de máxima verosimilitud
- Valor fijo. Establezca también el **Valor** a utilizar.
- Desvianza
- Chi-cuadrado de Pearson

**Método binomial negativa** Seleccione el método para la estimación del parámetro auxiliar de binomial negativa; las opciones disponibles son:

- Estimación de máxima verosimilitud
- Valor fijo. Establezca también el **Valor** a utilizar.

**Convergencia de los parámetros** Se asume la convergencia si el cambio máximo absoluto o el cambio máximo relativo en las estimaciones de parámetro es menor que el valor especificado, que no debe ser negativo. El criterio no se utiliza si el valor especificado es igual a 0.

**Convergencia de log-verosimilitud** Se asume la convergencia si el cambio absoluto o el cambio relativo en la función de log-verosimilitud es menor que el valor especificado, que no debe ser negativo. El criterio no se utiliza si el valor especificado es igual a 0.

**Convergencia hessiana** Para la especificación **Absoluta**, se asume la convergencia si un estadístico basado en la convergencia hessiana es menor que el valor especificado. En el caso de la especificación **Relativa**, se supone la convergencia si el estadístico es menor que el producto del valor especificado y el valor absoluto del logaritmo de la verosimilitud. El criterio no se utiliza si el valor especificado es igual a 0.

**Número máximo de iteraciones** Puede especificar el número máximo de iteraciones que ejecutará el algoritmo. El algoritmo utiliza un proceso iterativo doble que consta de un bucle interno y un bucle

externo. El valor especificado para el número máximo de iteraciones se aplica a ambos bucles. Especifique un número entero no negativo. El valor predeterminado es 100.

**Tolerancia para la singularidad** Este valor se utiliza como la tolerancia en la comprobación de la singularidad. Especifique un valor positivo.

**Nota:** De forma predeterminada, se utiliza **Convergencia de parámetros**, donde se comprueba el cambio máximo **Absoluto** en una tolerancia de 1E-6. Este valor podía generar resultados que difieren de los resultados obtenidos en versiones anteriores a la versión 17. Para reproducir los resultados de versiones anteriores a la 17, utilice **Relativo** para el criterio de convergencia de parámetros y conserve el valor de tolerancia predeterminado de 1E-6.

## Selección de modelos

**Utilizar selección de modelo o regularización** Para activar los controles de este panel, seleccione esta casilla de verificación.

**Método** Seleccione el método de selección de modelo o (si utiliza **Ridge**) la regularización que se va a utilizar. Puede elegir ente las opciones siguientes:

- **Lazo** También se conoce como la regularización L1, este método es más rápido que Pasos sucesivos hacia adelante si hay un gran número de predictores. Este método evita el sobreajuste mediante reducción (es decir, imponiendo una penalización) en los parámetros. Puede reducir algunos parámetros a cero, realizando un lazo de selección de variables.
- **Ridge** También se conoce como regularización L2, este método evita el sobreajuste mediante reducción (es decir, imponiendo una penalización) en los parámetros. Reduce todos los parámetros en las mismas proporciones pero no elimina ninguno y no es un método de selección de variables.
- **Red elástica** También conocido como regularización L1 + L2, este método evita el sobreajuste mediante reducción (es decir, imponiendo una penalización) en los parámetros. Puede reducir algunos parámetros a cero, realizando selección de variables.
- **Pasos sucesivos hacia adelante** Este método se inicia sin efectos en el modelo y añade o elimina efectos de paso en paso hasta que no se pueden añadir o eliminar ninguno más, según los criterios de pasos sucesivos.

**Detectar automáticamente interacciones bidireccionales** Para detectar automáticamente interacciones bidireccionales, seleccione esta opción.

**Parámetros de penalización**

Estas opciones sólo están disponibles si selecciona el **Método** de Lazo o Red elástica.

**Seleccionar automáticamente parámetros de penalización** Si no está seguro de qué penalizaciones de parámetro se deben establecer, seleccione esta casilla de verificación y el nodo identificará y aplicará las penalizaciones.

**Parámetro de penalización de lazo** Entre el parámetro de penalización que el **Método** de selección de modelo de lazo debe utilizar.

**Parámetro de penalización de red elástica 1** Entre el parámetro de penalización L1 que el **Método** de selección de modelo de red elástica debe utilizar.

**Parámetro de penalización de red elástica 2** Entre el parámetro de penalización L2 que **Método** de selección de modelo de red elástica debe utilizar.

**Paso adelante**

Estas opciones sólo están disponibles si selecciona el **Método** Pasos sucesivos hacia adelante.

**Incluir efectos con valores p no inferiores a** Especifique el valor de probabilidad mínimo que pueden tener los efectos que se deben incluir en el cálculo.

**Eliminar efectos con valores p superiores a** Especifique el valor de probabilidad máximo que pueden tener los efectos que se deben incluir en el cálculo.

**Personalizar el número máximo de efectos en el modelo final** Para activar la opción **Número máximo de efectos**, seleccione esta casilla de verificación.

**Número máximo de efectos** Especifique el número máximo de efectos cuando se utiliza el método de construcción de pasos sucesivos hacia adelante.

**Personalizar número máximo de pasos** Para activar la opción **Número máximo de pasos** , seleccione esta casilla de verificación.

**Número máximo de pasos** Especifique el número máximo de pasos cuando se utiliza el método de construcción de pasos sucesivos hacia adelante.

## Opciones de modelo

**Nombre de modelo** Puede generar el nombre del modelo automáticamente basándose en el campo objetivo, o especificar un nombre **personalizado**. El nombre generado automáticamente es el nombre del campo objetivo. Si existen objetivos múltiples, el nombre del modelo se forma con los nombres de campos en orden, conectados por símbolos &. Por ejemplo, si field1, field2 y field3 son objetivos, el nombre de modelo es: *field1 & field2 & field3*.

**Calcular importancia de predictor** Para los modelos que generan una medida adecuada de importancia, puede mostrar un gráfico que indique la importancia relativa de cada predictor al estimar el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Tenga en cuenta que puede tardarse más tiempo en calcular la importancia del predictor para algunos modelos, especialmente al trabajar con conjuntos de datos de gran tamaño; además, como resultado está desactivada para algunos modelos de forma predeterminada.

Si desea obtener más información, consulte “Importancia del predictor” en la página 44.

## Nugget de modelo GLE

### Salida de nugget de modelo GLE

Después de crear un modelo GLE, está disponible la información siguiente en el visor de salida.

#### Tabla de información de modelo

La tabla de información de modelo proporciona información clave acerca del modelo. La tabla identifica algunos ajustes de modelo de alto nivel, por ejemplo:

- El nombre del campo objetivo seleccionado en la pestaña Campos de nodo Tipo o nodo GLE.
- Los porcentajes de categoría de destino de referencia y modelados.
- La distribución de probabilidad y la función de enlace asociada.
- El método de generación de modelos utilizado.
- El número de entrada de predictores y el número en el modelo final.
- El porcentaje de precisión de la clasificación.
- El tipo de modelo.
- El porcentaje de precisión del modelo, si el destino es continuo.

#### Resumen de registros

La tabla de resumen muestra cuántos registros se han utilizado para ajustarse al modelo y cuántos se han excluido. Los detalles mostrados incluyen el número y porcentaje de los registros incluidos y excluidos, así como el número no ponderado si ha utilizado ponderación de frecuencia.

### Importancia del predictor

El gráfico Importancia de predictor muestra la importancia de las 10 entradas (predictores) principales del modelo en forma de diagrama de barras.

Si hay más de 10 campos en el gráfico, puede cambiar la selección de los predictores que se incluyen en el gráfico mediante el control deslizante situado debajo del mismo. Las marcas indicadores del control deslizante son de anchura fija, y cada marca del control deslizante presenta 10 campos. Puede mover las marcas indicadoras a lo largo del control deslizante para visualizar los 10 ca,pos anteriores o siguientes, ordenados por importancia de predictor.

Puede efectuar una doble pulsación en el gráfico para abrir un cuadro de diálogo independiente en el que se pueden editar los valores del gráfico. Por ejemplo, puede corregir elementos tales como el tamaño del gráfico y el tamaño y color de los fonts utilizados. Al cerrar este cuadro de diálogo de edición independiente, los cambios se aplicarán al gráfico que se visualiza en la pestaña Salida.

### Gráfico de residuos por pronosticados

Puede utilizar este gráfico para identificar valores atípicos o para diagnosticar la varianza de error no constante o de no linealidad. Un gráfico ideal mostrará los puntos dispersos aleatoriamente acerca de la línea cero.

El patrón esperado es que la distribución de los residuos de desvianza estandarizados en valores pronosticados del predictor lineal tenga un valor medio de cero y un rango constante. El patrón esperado es una línea horizontal a través de cero.

## Valores de nugget de modelo GLE

En la pestaña Configuración de un nugget de modelo GLE, especifique las opciones para la propensión bruta y para la generación de SQL durante la puntuación de modelos. Esta pestaña sólo está disponible después de añadir el nugget de modelo a una ruta.

**Calcular puntuaciones de propensión bruta** Para modelos sólo con objetivos de marca, puede solicitar puntuaciones de propensión bruta que indiquen la probabilidad del resultado verdadero especificado para el campo objetivo. Éstas se añaden a los valores estándar de predicción y confianza. Las puntuaciones ajustadas de propensión no están disponibles.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se genera SQL:

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

# Nodo Cox

La regresión de Cox crea un modelo predictivo para datos de tiempo hasta el evento. El modelo genera una función de supervivencia que predice la probabilidad de que se haya producido el evento de interés en un momento dado *t* para determinados valores de las variables predictoras. La forma de la función de supervivencia y los coeficientes de regresión para los predictores se calculan a partir de los sujetos

observados; a continuación, el modelo puede aplicarse a nuevos casos que tengan mediciones para las variables del predictor. Tenga en cuenta que la información de sujetos censurados, es decir, los que no experimentan el evento de interés durante el tiempo de observación, contribuye de manera útil al cálculo del modelo.

**Ejemplo.** Como parte de su esfuerzo por reducir el abandono de clientes, una empresa de telecomunicaciones se ha interesado en el modelado del "tiempo de abandono" para determinar los factores que se asocian a los clientes que están a punto de cambiarse de servicio. Para este propósito, se ha seleccionado una muestra aleatoria de clientes y se ha extraído de la base de datos su duración como cliente (si aún son o no clientes activos) y distintos campos demográficos

**Requisitos.** Necesita uno o más campos de entrada, exactamente un campo objetivo y debe especificar un campo de tiempo de supervivencia dentro del nodo Cox. El campo objetivo debe estar codificado de manera que el valor "false" indique supervivencia y el valor "true" indique que se ha producido el evento de interés; debe tener un nivel de medición de *Marca* con almacenamiento de cadena o entero. (Si es necesario, es posible convertir el almacenamiento mediante un nodo Rellenar o Derivar. ) Se ignorarán los campos establecidos en *Ambos* o *Ninguno*. Los tipos de los campos utilizados en el modelo deben estar completamente instanciados. El tiempo de supervivencia puede ser cualquier campo numérico.

**Nota:** Al puntuar un modelo de regresión de Cox, se ha notificado un error si las series vacías en variables categóricas se utilizan como entrada para la generación del modelo. Evite utilizar series vacías como entrada.

**Fechas y horas.** Los campos Fecha y Hora no se pueden utilizar para definir directamente el tiempo de supervivencia; si tiene campos Fecha y Hora debe utilizarlos para crear un campo que contenga tiempos de supervivencia, basados en la diferencia entre la fecha de entrada en el estudio y la fecha de observación.

**Análisis Kaplan-Meier.** La regresión de Cox se puede realizar sin campos de entrada. Equivale a un análisis de Kaplan-Meier.

## Opciones de campos del nodo Cox

**Tiempo de supervivencia.** Seleccione un campo numérico (uno con un nivel de medición de *Continuo*) para que el nodo se pueda ejecutar. El tiempo de supervivencia indica la vida útil del registro que se está prediciendo. Por ejemplo, si modela el tiempo de abandono de cliente, éste será el campo que registra el tiempo que el cliente ha estado en la organización. La fecha en la que el cliente se una o abandone no afectará al modelo; sólo será importante la duración del periodo del cliente.

El tiempo de supervivencia debe ser una duración sin unidades. Debe asegurarse que los campos de entrada coinciden con el tiempo de convivencia. Por ejemplo, en un estudio para medir los abandonos por meses, utilizaría las ventas por meses como entrada en lugar de las ventas por año. Si sus datos tienen fechas de inicio y de fin en lugar de una duración, debe recodificar esas fechas a una duración anterior del nodo Cox.

Los campos restantes de este cuadro de diálogo son los que se utilizan normalmente en IBM SPSS Modeler. Consulte el tema “Opciones de los campos del nodo de modelado” en la página 31 para obtener más información.

## Opciones de modelo para el nodo Cox

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Crear modelos divididos.** Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como campos de división. Consulte el tema "Generación de modelos divididos" en la página 28 para obtener más información.

**Método.** Éstas son las opciones disponibles para introducir predictores en el modelo:

- **Intro.** Éste es el método predeterminado que introduce directamente todos los términos en el modelo. No se realiza ninguna selección de campos en la creación del modelo.
- **Por pasos.** El método de selección de campos Por pasos crea el modelo por pasos, como su nombre indica. El modelo inicial es el más simple, sin ningún término del modelo (excepto el constante) en el modelo. En cada paso, se evalúan los términos que no se han añadido aún al modelo y si el mejor de dichos términos se suma de forma significativa a la eficacia predictiva del modelo, se añadirá a éste. Además, los términos que se encuentran actualmente en el modelo se vuelven a evaluar para determinar si se puede eliminar alguno de ellos sin que afecte al modelo de forma significativa. Si es así, se eliminan. El proceso se repite y se añaden y/o eliminan otros términos. Cuando no se puedan añadir más términos para mejorar el modelo, y no se puedan eliminar más sin que le afecte, se creará el modelo final.
- **Por pasos hacia atrás.** El método Por pasos hacia atrás es fundamentalmente lo contrario al método Por pasos. Con este método, el modelo inicial contiene todos los términos como predictores. En cada paso, se evalúan los términos del modelo y se eliminan los que no afecten al modelo de forma significativa. Además, los términos eliminados anteriormente se vuelven a evaluar para determinar si el mejor de dichos términos se añade de forma significativa a la eficacia predictiva del modelo. Si es así, se volverá a añadir al modelo. Cuando no se puedan añadir más términos para mejorar el modelo y no se puedan eliminar más sin que le afecte, se creará el modelo final.

*Nota*: Los métodos automáticos (incluidos Por pasos y Por pasos hacia atrás) son métodos de aprendizaje altamente adaptables y tienen una fuerte tendencia a ajustar los datos de entrenamiento. Cuando se utilicen estos métodos, es muy importante comprobar la validez del modelo resultante, bien con datos nuevos o con una muestra de comprobación reservada mediante el nodo Partición.

**Grupos.** La especificación de un campo de grupos hace que el nodo calcule modelos separados para cada categoría del campo. El objetivo debe ser cualquier campo categórico (marca o nominal) con almacenamiento de cadena o entero.

**Tipo de modelo.** Hay dos opciones para definir los términos del modelo. Los modelos **Efectos principales** sólo incluyen los campos de entrada de forma individual y no comprueban las interacciones (efectos multiplicativos) entre los campos de entrada. Los modelos **Personalizados** sólo incluyen los términos que se especifiquen (efectos principales e interacciones). Cuando seleccione esta opción, utilice la lista Términos del modelo para añadir términos al modelo o eliminarlos.

**Términos del modelo.** Al crear un modelo personalizado, deberá especificar explícitamente los términos del modelo. La lista muestra el conjunto actual de términos para el modelo. Los botones situados en la parte derecha de la lista Términos del modelo le permitirán añadir y eliminar los términos del modelo.

- Para añadir términos al modelo, pulse en el botón *Añadir nuevos términos del modelo*.
- Seleccione los términos deseados para eliminarlos y pulse en el botón *Eliminar los términos del modelo seleccionado*.

## Adición de términos a un modelo de regresión de Cox

Al solicitar un modelo personalizado, puede añadirle términos pulsando en el botón *Añadir nuevos términos del modelo* de la pestaña Modelo. Se abrirá un nuevo cuadro de diálogo en el que podrá especificar los términos.

**Tipo de término que se va a añadir.** Hay varias formas de añadir términos al modelo, según la selección de los campos de entrada de la lista Campos disponibles.

- **Interacción sencilla.** Inserta el término que representa la interacción de todos los campos seleccionados.

- **Efectos principales.** Inserta un término de efectos principales (el propio campo) para cada campo de entrada seleccionado.
- **Todas las interacciones de dos factores.** Inserta un término de interacción de 2 factores (el producto de los campos de entrada) para cada posible par de campos de entrada seleccionados. Por ejemplo, si ha seleccionado los campos de entrada *A*, *B* y *C* en la lista Campos disponibles, este método insertará los términos *A * B*, *A * C* y *B * C*.
- **Todas las interacciones de tres factores.** Inserta un término de interacción de 3 factores (el producto de los campos de entrada) para cada posible combinación de campos de entrada seleccionados, tomando tres al mismo tiempo. Por ejemplo, si ha seleccionado los campos de entrada *A*, *B*, *C* y *D* en la lista Campos disponibles, este método insertará los términos *A * B * C*, *A * B * D*, *A * C * D* y *B * C * D*.
- **Todas las interacciones de cuatro factores.** Inserta un término de interacción de 4 factores (el producto de los campos de entrada) para cada posible combinación de campos de entrada seleccionados, tomando cuatro al mismo tiempo. Por ejemplo, si ha seleccionado los campos de entrada *A*, *B*, *C*, *D* y *E* en la lista Campos disponibles, este método insertará los términos *A * B * C * D*, *A * B * C * E*, *A * B * D * E*, *A * C * D * E* y *B * C * D * E*.

**Campos disponibles.** Muestra los campos de entrada disponibles que se van a utilizar en la construcción de términos del modelo. Tenga en cuenta que la lista puede incluir campos que no son campos de entrada correctos, por lo que asegúrese de que todos los términos del modelo incluyen sólo campos de entrada.

**Presentación preliminar.** Muestra los términos que se añadirán al modelo si pulsa en **Insertar**, según los campos seleccionados y el tipo de término seleccionado anteriormente.

**Insertar.** Inserta los términos del modelo (según la selección actual de los campos y el tipo de término) y cierra el cuadro de diálogo.

## Opciones de experto para el nodo Cox

**Convergencia.** Estas opciones le permiten controlar los parámetros de la convergencia del modelo. Cuando se ejecuta el modelo, la configuración de la convergencia controla cuántas veces se ejecutan los distintos parámetros a través de éste para comprobar si se ajustan. Cuanta más veces se prueben los parámetros, más próximos estarán los resultados (es decir, los resultados convergirán). Consulte el tema "Criterios de convergencia del nodo Cox" para obtener más información.

**Resultados.** Estas opciones le permiten solicitar estadísticos adicionales y gráficos, incluida la curva de supervivencia, que aparecerán en el resultado avanzado del modelo generado construido por el nodo. Consulte el tema "Opciones de resultados avanzados del nodo Cox" en la página 240 para obtener más información.

**Método por pasos.** Estas opciones le permiten controlar los criterios para añadir y eliminar los campos con el método de estimación Por pasos. (Si el método Introducir está seleccionado, el botón estará desactivado.) Consulte el tema "Criterios del método por pasos del nodo Cox" en la página 240 para obtener más información.

### Criterios de convergencia del nodo Cox

**Iteraciones máximas.** Permite especificar las iteraciones máximas del modelo, que controla durante cuánto tiempo el procedimiento buscará una solución.

**Convergencia del logaritmo de la verosimilitud.** Las iteraciones se detendrán si el cambio relativo del logaritmo de la verosimilitud es menor que este valor. Este criterio no se aplica si el valor es igual a 0.

**Convergencia de los parámetros.** Las iteraciones se detendrán si el cambio absoluto o relativo de las estimaciones de los parámetros es menor que este valor. Este criterio no se aplica si el valor es igual a 0.

## Opciones de resultados avanzados del nodo Cox

**Estadísticos.** Puede obtener estadísticos para sus parámetros del modelo, incluidos los intervalos de confianza para exp(*B*) y correlaciones de estimaciones. Puede solicitar estos estadísticos en cada paso o sólo en el último paso.

**Mostrar la función de línea base.** Permite visualizar la función de riesgo de línea base y la supervivencia acumulada en la media de las covariables.

Gráficos

Los gráficos pueden ayudarle a evaluar el modelo estimado e interpretar los resultados. Puede trazar las funciones de supervivencia, de riesgo, de registro menos registro y de uno menos supervivencia.

- *Supervivencia*. Muestra la función de supervivencia acumulada, en una escala lineal.
- *Riesgo*. Muestra la función de riesgo acumulado en una escala lineal.
- **Log menos log.** Muestra la estimación de supervivencia acumulada después de aplicar la transformación In(-In) a la estimación.
- *Uno menos la supervivencia*. Representa la función uno menos la supervivencia en una escala lineal.

**Trace una línea diferente para cada valor.** Esta opción sólo se encuentra disponible para los campos categóricos.

**Valor para utilizar con los gráficos.** Dado que estas funciones dependen de los valores de los predictores, debe utilizar valores constantes en los predictores para trazar las funciones frente al tiempo. El valor predeterminado es utilizar la media de cada predictor como un valor constante, pero puede introducir sus propios valores para el gráfico utilizando la cuadrícula. Para las entradas categóricas se utiliza la codificación de indicador, de manera que hay un coeficiente de regresión para cada categoría (excepto para la última). Así, una entrada categórica tiene un valor medio para cada contraste de indicador, igual a la proporción de casos en la categoría correspondiente al contraste del indicador.

## Criterios del método por pasos del nodo Cox

**Criterio de exclusión.** Seleccione **Razón de verosimilitud** para un modelo más robusto. Si desea reducir el tiempo necesario para generar el modelo, puede intentar seleccionar **Wald**. Existe la opción adicional **Condicional** que permite una comprobación de eliminación en función de la probabilidad del estadístico de razón de verosimilitud basado en estimaciones de parámetros condicionales.

**Umbrales de significación para los criterios de RL.** Esta opción le permite especificar criterios de selección según la probabilidad estadística (el valor *p*) asociada a cada campo. Los campos se añadirán al modelo sólo si el valor *p* asociado es más pequeño que el valor **Entrada** y se eliminarán sólo si el valor *p* es mayor que el valor **Eliminación**. El valor **Entrada** debe ser menor que el valor **Eliminación**.

# Opciones de configuración para el nodo Cox

**Predecir supervivencia en el futuro.** Seleccione uno o varios tiempos futuros. La supervivencia, es decir, si cada caso puede sobrevivir al menos durante ese período de tiempo (desde ahora) sin que se haya producido el evento terminal, se predice para cada registro en cada valor de tiempo, una predicción por valor de tiempo. Tenga en cuenta que esa supervivencia es el valor "false" del campo objetivo.

- **Intervalos regulares.** Los valores de supervivencia se generan a partir del **Intervalo de tiempo** y **Número de períodos de tiempo que se van a puntuar**. Por ejemplo, si se solicitan períodos de 3 tiempos con un intervalo de 2 cada vez, la supervivencia se predecirá en los tiempos futuros 2, 4, 6. Cada registro se evalúa en los mismos valores de tiempo.
- **Campos de tiempo.** Los tiempos de supervivencia se proporcionan con cada cambio de tiempo seleccionado (se genera un campo de predicción), así cada registro puede evaluarse en momentos diferentes.

**Tiempo de supervivencia pasado.** Especifique el tiempo de supervivencia del registro hasta ahora; por ejemplo, el cargo de un cliente existente como un campo. La puntuación de la probabilidad de supervivencia en un tiempo futuro será condicional en el tiempo de supervivencia pasado.

*Nota:* Los valores de los tiempos de supervivencia futuros y pasados deben estar en el rango de tiempos de supervivencia en los datos utilizados para entrenar el modelo. Los registros cuyos tiempos no estén comprendidos dentro de este rango se puntúan como nulos.

**Añadir todas las probabilidades.** Especifica si se añaden las probabilidades de cada categoría del campo de resultados a cada registro procesado por el nodo. Si no se selecciona esta opción, sólo se añadirá la probabilidad de la categoría predicha. Las probabilidades se calculan para cada tiempo futuro.

**Calcular función de riesgo acumulado.** Especifique si el valor del riesgo acumulado se añade a cada registro. El riesgo acumulado se calcula para cada tiempo futuro.

## Nugget de modelo de Cox

Los modelos de regresión de Cox representan las ecuaciones calculadas por los nodos Cox. Contienen toda la información capturada por el modelo, así como información acerca del rendimiento y la estructura del modelo.

Cuando se ejecuta una ruta que contiene un modelo de regresión Cox generado, el nodo añade dos nuevos campos que contienen la predicción del modelo y la probabilidad asociada. Los nombres de los nuevos campos se derivan del nombre del campo de salida que se está prediciendo, con el prefijo *$C-* para la categoría predicha y *$CP-* para la probabilidad asociada y con el sufijo del número del intervalo de tiempo futuro o el nombre del campo de tiempo que define el intervalo de tiempo. Por ejemplo, para un campo de salida denominado *abandono* y dos intervalos de tiempo futuro definidos regularmente, los nuevos campos se denominarán *$C-abandono-1*, *$CP-abandono-1*, *$C-abandono-2* y *$CP-abandono-2*. Si los tiempos futuros se definen con un *cargo* de campo de tiempo, los nuevos campos serán *$C-abandono_cargo* y *$CP-abandono_cargo*.

Si ha seleccionado la opción de configuración **Añadir todas las probabilidades** en el nodo Cox, se añadirán dos campos adicionales para cada tiempo futuro, que contenga las probabilidades de supervivencia y fallo para cada registro. Estos campos adicionales se denominan en base al nombre del campo de salida, con el prefijo *$CP-<valor falso>-* para la probabilidad de supervivencia y *$CP-<valor verdadero>-* para la probabilidad del caso que se ha producido y con el sufijo del número del intervalo de tiempos futuros. Por ejemplo, para un campo de salida donde el valor "falso" es 0 y el valor "verdadero" es 1 y dos intervalos de tiempo futuro definidos regularmente, los nuevos campos se denominarán *$CP-0-1*, *$CP-1-1*, *$CP-0-2* y *$CP-1-2*. Si los tiempos futuros se definen con un *cargo* de campo de tiempo, los nuevos campos serán *$CP-0-1* y *$CP-1-1*, dado que existe un único intervalo futuro.

Si ha seleccionado la opción de configuración **Calcular función de riesgo acumulado** en el nodo Cox, se añadirá un campo adicional para cada tiempo futuro, que contenga la función de riesgo acumulado para cada registro. Estos campos adicionales se denominan en base al nombre del campo de salida, con el prefijo *$CH-* y con el sufijo del número del intervalo de tiempos futuros o el nombre del campo de tiempo que define el intervalo de tiempo. Por ejemplo, para un campo de salida denominado *abandono* y dos intervalos de tiempo futuro definidos regularmente, los nuevos campos se denominarán *$CH-abandono-1* y *$CH-abandono-2*. Si los tiempos futuros se definen con un *cargo* de campo de tiempo, el nuevo campo será *$CH-abandono-1*.

### Configuración de resultados de regresión de Cox

Excepto para la generación de SQL, la pestaña Configuración del nugget contiene el mismo control que la pestaña Configuración del nodo del modelo. Los valores predeterminados de los controles del nugget se determinan por el conjunto de valores definidos en el nodo de modelo. Consulte el tema “Opciones de configuración para el nodo Cox” en la página 240 para obtener más información.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

## Resultado avanzado de regresión de Cox

Los resultados avanzados de la regresión de Cox ofrecen información detallada sobre el modelo estimado y su rendimiento, incluida la curva de supervivencia. La mayoría de la información contenida en la salida avanzada es bastante técnica y es necesario tener amplios conocimientos sobre la regresión de Cox para interpretar correctamente estos resultados.

# Capítulo 11. Modelos de agrupación en clústeres

Los modelos de agrupación en clústeres se centran en la identificación de grupos de registros similares y en el etiquetado de registros según el grupo al que pertenecen. Esto se lleva a cabo sin la ventaja de disponer de conocimientos previos sobre los grupos y sus características. De hecho, puede que ni siquiera sepa exactamente cuántos grupos va a buscar. Esto es lo que diferencia a los modelos de agrupación en clústeres de otras técnicas de aprendizaje de máquinas: no hay campo objetivo o de salida predefinidos para el modelo que se va a predecir. A menudo se hace referencia a estos modelos como modelos de **aprendizaje no supervisado**, ya que no hay ningún estándar externo con el que juzgar el rendimiento de la clasificación del modelo. No hay respuestas *correctas* o *incorrectas* para estos modelos. Su valor viene determinado por su capacidad de capturar agrupaciones interesantes en los datos y proporcionar descripciones útiles de dichas agrupaciones.

Los métodos de agrupación en clústeres se basan en la medición de distancias entre registros y entre clústeres. Los registros se asignan a los clústeres de un modo que tiende a minimizar la distancia entre los registros pertenecientes al mismo clúster.

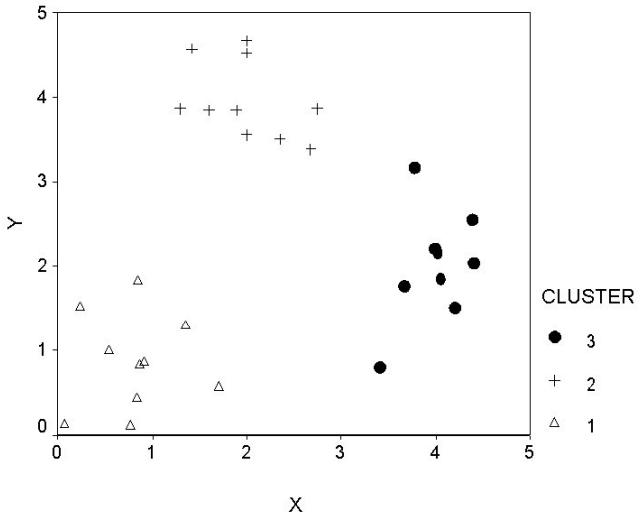



*Figura 44. Modelo de agrupación en clústeres simple*

Se ofrecen tres métodos de agrupación en clústeres:

El nodo K-medias agrupa conjuntos de datos en grupos distintos (o clústeres). El método define un número fijo de clústeres, de forma iterativa asigna registros a los clústeres y ajusta los centros de los clústeres hasta que no se pueda mejorar el modelo. En lugar de intentar predecir un resultado, los modelos de *k*-medias utilizan un proceso conocido como aprendizaje no supervisado para revelar los patrones del conjunto de campos de entrada.

El nodo Bietápico es un método de agrupación en clústeres de dos pasos. El primer paso es hacer una única pasada por los datos para comprimir los datos de entrada de la fila en un conjunto de subclústeres administrable. El segundo paso utiliza un método de agrupación en clústeres jerárquica para fundir progresivamente los subclústeres en clústeres cada vez más grandes. El bietápico tiene la ventaja de estimar automáticamente el número óptimo de clústeres para los datos de entrenamiento. Puede gestionar tipos de campos mixtos y grandes conjuntos de datos eficazmente.

El nodo Kohonen genera un tipo de red neuronal que se puede usar para agrupar un conjunto de datos en grupos distintos. Cuando la red se termina de entrenar, los registros que son similares se deberían cerrar juntos en el mapa de resultados, mientras que los registros que son diferentes aparecerían aparte. Puede observar el número de observaciones capturadas por cada unidad en el nugget de modelo para identificar unidades fuertes. Esto le proporcionará una idea del número apropiado de clústeres.

Los modelos de agrupación en clústeres se usan a menudo para crear clústeres o segmentos que se usan posteriormente como entradas en análisis posteriores. Un ejemplo común lo ilustran los segmentos del mercado que usan los comerciantes para dividir su mercado en subgrupos homogéneos. Cada segmento tiene unas características especiales que afectan al éxito de los esfuerzos de mercado orientados a ello. Si utiliza la minería de datos para optimizar su estrategia de mercado, normalmente podrá mejorar el modelo de forma significativa identificando los segmentos apropiados y utilizando esa información sobre los segmentos en sus modelos predictivos.

## Nodo Kohonen

Las redes de Kohonen son un tipo de red neuronal que realiza agrupación en clústeres, también conocidas como **knet** o como un **mapa autoorganizativo**. Este tipo de redes se puede utilizar para agrupar el conjunto de datos en grupos distintos cuando no se sabe lo que son al principio. Los registros se agrupan de manera que los de un mismo grupo o clúster tiendan a ser similares entre ellos y que los de otros grupos sean distintos.

Las unidades básicas son **neuronas** y se organizan en dos capas: la **capa de entrada** y la **capa de salida** (también denominada **mapa de resultados**). Todas las neuronas de entrada están conectadas a todas las neuronas de salida, y estas conexiones tienen **fuerzas** o **ponderaciones** asociadas a ellas. Durante el entrenamiento, cada unidad compite con las demás para "ganar" cada registro.

El mapa de resultados es una red de neuronas bidimensional sin conexiones entre las unidades.

Los datos de entrada se presentan en la capa de entrada y los valores se propagan a la capa de salida. La neurona de salida con la respuesta más fuerte se considera la **ganadora** y constituye la respuesta para dicha entrada.

Al comienzo, todas las ponderaciones son aleatorias. Cuando una unidad gana un registro, sus fuerzas (junto con las de las unidades más próximas, colectivamente conocidas como **vecindad**) se ajustan para coincidir mejor con el patrón de los valores predictores de dicho registro. Se presentan todos los registros de entrada y se actualizan todas las ponderaciones consecuentemente. Este proceso se repite varias veces hasta que las modificaciones sean muy pequeñas. A medida que avanza el entrenamiento, las ponderaciones en las unidades de la tabla se ajustan para formar un "mapa" bidimensional de los clústeres (de ahí el término **mapa autoorganizativo**).

Cuando la red se termina de entrenar, los registros que son similares se deberían cerrar juntos en el mapa de resultados, mientras que los registros que son muy diferentes aparecerían aparte.

A diferencia de la mayoría de los métodos de aprendizaje de IBM SPSS Modeler, las redes de Kohonen *no* utilizan un campo objetivo. Este tipo de aprendizaje, sin campo objetivo, se denomina **aprendizaje no supervisado**. En lugar de intentar predecir un resultado, las redes de Kohonen intentan revelar los patrones en el conjunto de campos de entrada. Por lo general, una red de Kohonen termina con unas pocas unidades que resumen muchas observaciones (unidades **fuertes**) y varias unidades que no corresponden realmente con ninguna de las observaciones (unidades **débiles**). Las unidades fuertes (y, en ocasiones, algunas unidades consecutivas a ellas en la cuadrícula) representan posibles centros de clústeres.

Otro uso de las redes de Kohonen es el de la **reducción de dimensión**. La característica espacial de las cuadrículas bidimensionales permite una correlación desde los predictores originales *k* a dos características derivadas que conservan las relaciones de similitud de los predictores originales. En algunos casos, esto puede ofrecer el mismo tipo de ventaja que el análisis factorial o PCA.

Observe que el método para calcular el tamaño predeterminado de la cuadrícula de salida ha cambiado con respecto a versiones anteriores de IBM SPSS Modeler. El nuevo método suele generar capas de salida más pequeñas, que se entrenan más rápidamente y se generalizan mejor. Si obtiene unos resultados no significativos con el tamaño predeterminado, intente aumentar el tamaño de la cuadrícula de salida en la pestaña Experto. Consulte el tema "Opciones de experto para el nodo Kohonen" en la página 246 para obtener más información.

**Requisitos.** Para entrenar una red de Kohonen, necesita uno o más campos con su rol establecido como *Entrada*. Se ignorarán los campos con el rol establecido como *Objetivo*, *Ambos* o *Ninguno*.

**Puntos fuertes.** No es necesario tener los datos en pertenencia a grupos para crear un modelo de red de Kohonen. Ni siquiera necesita saber el número de grupos que buscar. Las redes de Kohonen comienzan con un número elevado de unidades y, según avanza el entrenamiento, las unidades se dejan atraer por los clústeres naturales de los datos. Puede mirar el número de observaciones capturadas por cada unidad en el nugget de modelo para identificar las unidades fuertes, las cuales pueden darle una idea del número adecuado de clústeres.

## Opciones de modelo para el nodo Kohonen

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Continuar entrenando modelo existente.** De forma predeterminada, cada vez que se ejecuta un nodo Kohonen, se crea una red completamente nueva. Si selecciona esta opción, el entrenamiento continúa con la última red generada correctamente por el nodo.

**Mostrar gráfico de retroalimentación.** Seleccione esta opción para mostrar una representación visual de la matriz bidimensional durante el entrenamiento. La fuerza de cada nodo se representa con un color diferente. El rojo denota una unidad que está ganando muchos registros (**fuerte**) y el blanco, una unidad que está ganando pocos registros, o ninguno (**débil**). Los comentarios pueden no mostrarse si el tiempo para construir el modelo es relativamente corto. Tenga en cuenta que esta característica puede ralentizar el tiempo de entrenamiento. Para acelerarlo, anule la selección de esta opción.

**Detener cuando.** El criterio de parada predeterminado detiene el entrenamiento en base a unos parámetros internos. También puede especificar una hora como criterio de parada. Introduzca la hora (en minutos) para la red a entrenar.

**Establecer semilla aleatoria.** Si no se establece ninguna semilla aleatoria, cada vez que se ejecute el nodo se obtendrán ponderaciones distintas de la secuencia de valores aleatorios utilizada para inicializar la red. Esto puede hacer que el nodo cree diferentes modelos en distintas ejecuciones, incluso si la configuración del nodo y los valores de los datos son exactamente los mismos. Si selecciona esta opción, puede establecer la semilla aleatoria en un valor específico para que el modelo resultante se pueda reproducir con exactitud. Una semilla aleatoria específica siempre genera la misma secuencia de valores aleatorios, en cuyo caso la ejecución del nodo siempre da como resultado el mismo modelo generado.

*Nota*: cuando se utiliza la opción **Establecer semilla aleatoria** con registros leídos de una base de datos, puede ser necesario un nodo Ordenar, antes del muestreo con el fin de garantizar el mismo resultado

cada vez que se ejecute el nodo. Esto se debe a que la semilla aleatoria depende del orden de registros, sin estar garantizado que sea el mismo en una base de datos relacional.

*Nota*: Si desea incluir campos nominales (conjuntos) en su modelo pero tiene problemas de memoria en su creación o le está llevando demasiado tiempo, considere la opción de volver a codificar campos de conjuntos grandes para reducir el número de valores o de utilizar un campo distinto con menos valores como proxy para los conjuntos grandes. Por ejemplo, si tiene un problema con un campo *id_producto* que contiene valores para productos individuales, podría considerar la opción de eliminarlo del modelo y añadir un campo *categoría_producto* menos detallado en su lugar.

**Optimizar.** Seleccione opciones diseñadas para aumentar el rendimiento durante la generación de modelos según sus necesidades específicas.

- Seleccione **Velocidad** para indicar al algoritmo que nunca debe recurrir al volcado en disco para mejorar el rendimiento.
- Seleccione **Memoria** para indicar al algoritmo que utilice el volcado en disco cuando lo considere oportuno en detrimento de la velocidad. Esta opción está seleccionada de forma predeterminada.

  *Nota*: cuando se ejecuta en modo distribuido, esta configuración puede quedar anulada por las opciones del administrador especificadas en *options.cfg*.

**Añadir etiqueta de clúster.** Seleccionada de forma predeterminada para nuevos modelos, pero no seleccionada para modelos cargados desde versiones anteriores de IBM SPSS Modeler, crea un único campo de puntuación categórico del mismo tipo que crean los nodos K-medias y Bietápico. Este campo de cadena es el utilizado en el nodo Agrupación en clústeres automática cuando se calculan medidas de clasificación de los diferentes tipos de modelos. Consulte el tema “Nodo Autoclúster” en la página 79 para obtener más información.

## Opciones de experto para el nodo Kohonen

Para aquéllos que tengan conocimientos avanzados sobre las redes de Kohonen, las opciones de experto permiten ajustar con precisión el proceso de entrenamiento. Para acceder a estas opciones, active el modo **Experto** en la pestaña Experto.

**Ancho y Longitud.** Especifique el tamaño (ancho y longitud) del mapa de resultados bidimensional como número de unidades de resultados en cada dimensión.

**Decrecimiento de tasas de aprendizaje.** Seleccione el decrecimiento de tasas de aprendizaje lineal o exponencial. La **tasa de aprendizaje** es un factor de ponderación que decrece con el tiempo, de tal modo que la red comienza codificando características generales de los datos y se va centrando gradualmente en detalles más precisos.

**Fase 1 y fase 2.** El entrenamiento de la red de Kohonen se divide en dos fases. La fase 1 es de estimación a grandes rasgos y se usa para capturar los patrones generales de los datos. La fase 2 es de precisión y se usa para ajustar el mapa con el fin de modelar las características más precisas de los datos. Cada fase presenta tres parámetros que definir:

- **Vecindad.** Establece el tamaño (radio) inicial de la vecindad, determinando el número de unidades "cercanas" que se actualizan junto con la unidad ganadora durante el entrenamiento. Durante la fase 1, el tamaño de vecindad comienza con *Vecindad de Fase 1* y decrece a (*Vecindad de Fase 2* + 1). Durante la fase 2, el tamaño de vecindad comienza con *Vecindad de Fase 2* y decrece a 1,0. *Vecindad de Fase 1* debería ser mayor que *Vecindad de Fase 2*.
- **Eta inicial.** Establece el valor inicial para la **eta** de tasas de aprendizaje. Durante la fase 1, eta comienza con *Eta inicial de Fase 1* y decrece a *Eta inicial de Fase 2*. Durante la fase 2, eta comienza con *Eta inicial de Fase 2* y decrece a 0. *Eta inicial de Fase 1* debería ser mayor que *Eta inicial de Fase 2*.
- **Ciclos.** Establece el número de ciclos para cada fase de entrenamiento. Cada fase continúa durante la cantidad especificada de pasadas por los datos.

## Nugget de modelo Kohonen

Los nugget de modelo Kohonen contienen toda la información capturada por la red Kohonen entrenada, así como la información acerca de la arquitectura de red de Kohonen.

Al ejecutar una ruta que contiene un nugget de modelo Kohonen, el nodo añade los dos nuevos campos que contienen las coordenadas *X* e *Y* de la unidad en la cuadrícula de resultados Kohonen que mejor respondieron a ese registro. Los nuevos nombres de campos se derivan del nombre del modelo, con el prefijo *$KX-* y *$KY-*. Por ejemplo, si el modelo se llama *Kohonen*, los nuevos campos se llamarían *$KX-Kohonen* y *$KY-Kohonen*.

Para tener una impresión más acertada de lo que la red de Kohonen ha codificado, pulse en la pestaña Modelo del explorador de nugget de modelo. Se mostrará el visor de clústeres, que ofrece una representación gráfica de los clústeres, campos y niveles de importancia. Consulte el tema "Visor de clústeres - Pestaña Modelo" en la página 259 para obtener más información.

Si prefiere visualizar los clústeres como una cuadrícula, puede consultar el resultado de la red de Kohonen trazando los campos *$KX-* y *$KY-* mediante un nodo Gráfico. (Debe seleccionar **agitación X** y **agitación Y** en el nodo Gráfico para evitar que los registros de cada unidad se representen superpuestos.) En el gráfico, también puede superponer un campo simbólico para investigar la forma en que la red de Kohonen aglomera los datos.

Otra buena forma de comprender la red de Kohonen es utilizar la inducción de reglas para descubrir las características que distinguen los clústeres que la red ha encontrado. Consulte el tema "Nodo C5.0" en la página 110 para obtener más información.

Para obtener información general sobre la utilización del explorador del modelo, consulte "Examen de nuggets de modelo" en la página 42

## Resumen de modelo Kohonen

La pestaña Resumen para un nugget de modelo Kohonen muestra información acerca de la arquitectura o topología de la red. La longitud y el ancho del mapa bidimensional de características de Kohonen (la capa de salida) se muestran como **$KX-** *nombre_del_modelo* y **$KY-** *nombre_del_modelo*. En el caso de las capas de entrada y salida, se enumera el número de unidades en esa capa.

## Nodo K-medias

El nodo K-medias ofrece un método de **análisis de clústeres**. Se puede utilizar para agrupar el conjunto de datos en grupos distintos cuando no se sabe lo que son al principio. A diferencia de la mayoría de los métodos de aprendizaje de IBM SPSS Modeler, los modelos de K-medias *no* utilizan un campo objetivo. Este tipo de aprendizaje, sin campo objetivo, se denomina **aprendizaje no supervisado**. En lugar de intentar predecir un resultado, los modelos de K-medias intentan revelar los patrones en el conjunto de campos de entrada. Los registros se agrupan de manera que los de un mismo grupo o clúster tiendan a ser similares entre ellos, y que los de otros grupos sean distintos.

K-medias empieza definiendo un conjunto de centros de clústeres iniciales derivados de datos. Después asigna cada registro al clúster de registros más similares, basándose en los valores de los campos de entrada de registros. Una vez asignados todos los casos, los centros de clústeres se actualizan para reflejar el nuevo conjunto de registros asignados a cada clúster. Los registros se vuelven a comprobar para ver si se deben reasignar a otro clúster, y el proceso de iteración de clúster/asignación continúa hasta que se alcanza el número máximo de iteraciones o el cambio entre una iteración y otra no sobrepasa el umbral especificado.

*Nota*: el modelo resultante depende, hasta cierto punto, del orden de los datos de entrenamiento. Reordenar los datos y regenerar el modelo puede dar como resultado un modelo de clústeres final distinto.

**Requisitos.** Para entrenar un modelo K-Means, necesita uno o más campos con su rol establecido como *Entrada*. Se ignorarán los campos con el rol establecido como *Resultado*, *Ambos* o *Ninguno*.

**Puntos fuertes.** No es necesario tener los datos en pertenencia a grupos para crear un modelo de K-medias. Este modelo suele ser el método más rápido de agrupación en clústeres para conjuntos de datos grandes.

## Opciones de modelo para el nodo K-medias

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Número de clústeres especificado.** Especifique el número de clústeres que generar. El valor predeterminado es 5.

**Generar campo de distancia.** Seleccione esta opción para que el nugget de modelo incluya un campo con la distancia de cada registro desde el centro del clúster que le ha sido asignado.

**Etiqueta de clúster.** Especifique el formato de los valores del campo de pertenencia al clúster generado. La pertenencia a un clúster se puede indicar con una **Cadena** con el **Prefijo de etiqueta** especificado (por ejemplo `"Clúster 1"`, `"Clúster 2"`, etc.) o con un **Número**.

*Nota*: Si desea incluir campos nominales (conjuntos) en su modelo pero tiene problemas de memoria en su creación o le está llevando demasiado tiempo, considere la opción de volver a codificar campos de conjuntos grandes para reducir el número de valores o de utilizar un campo distinto con menos valores como proxy para los conjuntos grandes. Por ejemplo, si tiene un problema con un campo *id_producto* que contiene valores para productos individuales, podría considerar la opción de eliminarlo del modelo y añadir un campo *categoría_producto* menos detallado en su lugar.

**Optimizar.** Seleccione opciones diseñadas para aumentar el rendimiento durante la generación de modelos según sus necesidades específicas.

- Seleccione **Velocidad** para indicar al algoritmo que nunca debe recurrir al volcado en disco para mejorar el rendimiento.
- Seleccione **Memoria** para indicar al algoritmo que utilice el volcado en disco cuando lo considere oportuno en detrimento de la velocidad. Esta opción está seleccionada de forma predeterminada.

  *Nota*: cuando se ejecuta en modo distribuido, esta configuración puede quedar anulada por las opciones del administrador especificadas en *options.cfg*.

## Opciones de experto para el nodo K-medias

Para aquellos con conocimientos avanzados de los clústeres de *K*-Medias, las opciones de experto permiten ajustar con precisión el proceso de entrenamiento. Para acceder a estas opciones, active el modo **Experto** en la pestaña Experto.

**Detener cuando.** Especifique los criterios de parada que utilizar en el entrenamiento del modelo. El criterio de parada **predeterminado** es 20 iteraciones o cambiar en < 0.000001 (lo que ocurra primero). Seleccione **Personalizado** para especificar sus propios criterios de parada.

- **Número máximo de iteraciones.** Esta opción permite detener el entrenamiento del modelo después del número de iteraciones especificado.
- **Cambio en la tolerancia.** Esta opción permite detener el entrenamiento del modelo cuando el cambio más grande de los centros de clústeres para una iteración sea inferior al nivel especificado.

**Valor codificado para conjuntos.** Especifique un valor entre 0 y 1,0, y utilícelo para volver a codificar los campos de conjuntos como grupos de campos numéricos. El valor predeterminado es la raíz cuadrada de 0,5 (0,707107 aproximadamente), que proporciona la ponderación adecuada de los campos de marcas que se han recodificado. Los valores cercanos a 1,0 ponderarán los campos de conjuntos más profundamente que los numéricos.

## Nugget de modelo de K-medias

Los nugget de modelo de K-medias contienen toda la información capturada por el modelo de agrupación en clúster, así como información acerca de los datos de entrenamiento y el proceso de estimación.

Cuando ejecuta una ruta que contiene un nugget de modelo de K-medias, el nodo añade dos nuevos campos que contienen la pertenencia del clúster y la distancia a partir del centro del clúster asignado para ese registro. Del nombre del modelo se derivan los nuevos nombres de campos con el prefijo *$KM-* para la pertenencia del clúster y *$KMD-* para la distancia desde el centro del clúster. Por ejemplo, si el modelo se llama *Kmeans*, los nuevos campos se llamarían *$KM-Kmeans* y *$KMD-Kmeans*.

Una buena forma de comprender el modelo K-medias es utilizar la inducción de reglas para descubrir las características que distinguen los clústeres que el modelo ha encontrado. Consulte el tema "Nodo C5.0" en la página 110 para obtener más información. Asimismo, puede pulsar en la pestaña Modelo del explorador de nugget de modelo para ver el visor de clústeres, que proporciona una representación gráfica de clústeres, campos y niveles de importancia. Consulte el tema "Visor de clústeres - Pestaña Modelo" en la página 259 para obtener más información.

Para obtener información general sobre la utilización del explorador del modelo, consulte "Examen de nuggets de modelo" en la página 42

## Resumen de modelo de K-medias

La pestaña Resumen de un nugget de modelo de K-medias contiene información acerca de los datos de entrenamiento, el proceso de estimación y los clústeres definidos por el modelo. Se muestra el número de clústeres y el historial de iteración. Si ha ejecutado un nodo Análisis conectado a este nodo de modelado, la información de dicho análisis también se mostrará en esta sección.

## Nodo de clúster bietápico

El nodo de clúster bietápico ofrece un método de **análisis de clústeres**. Se puede utilizar para agrupar el conjunto de datos en grupos distintos cuando no se sabe lo que son al principio. Al igual que los nodos Kohonen y K-medias, los modelos de clústeres bietápicos *no* utilizan un campo objetivo. En lugar de intentar predecir un resultado, el clúster Bietápico intenta revelar los patrones en el conjunto de campos de entrada. Los registros se agrupan de manera que los de un mismo grupo o clúster tiendan a ser similares entre ellos, y que los de otros grupos sean distintos.

El clúster Bietápico es un método de agrupación en clústeres de dos pasos. El primer paso es hacer una única pasada por los datos, durante la cual se comprimen los datos de entrada iniciales en un conjunto de subclústeres que se puede administrar. El segundo paso utiliza un método de agrupación en clústeres jerárquico para fundir progresivamente los subclústeres en clústeres cada vez más grandes, sin necesidad de realizar otra pasada por los datos. La agrupación en clústeres jerárquica tiene la ventaja de que no es necesario seleccionar el número de clústeres por adelantado. Muchos métodos de clúster jerárquico comienzan con registros individuales como clústeres iniciales y los van fundiendo sucesivamente para

generar clústeres más grandes. Aunque estos métodos suelen no funcionar bien con grandes cantidades de datos, la agrupación en clústeres previa inicial Bietápica permite que la agrupación en clústeres jerárquica sea rápida incluso con grandes conjuntos de datos.

**Nota:** el modelo resultante depende, hasta cierto punto, del orden de los datos de entrenamiento. Reordenar los datos y regenerar el modelo puede dar como resultado un modelo de clústeres final distinto.

**Requisitos.** Para entrenar un modelo de clúster Bietápico, necesita uno o más campos con su rol establecido como *Entrada*. Se ignorarán los campos con el rol establecido como *Objetivo*, *Ambos* o *Ninguno*. El algoritmo de clústeres bietápicos no gestiona los valores perdidos. Los registros con elementos vacíos para cualquiera de los campos de entrada se ignorarán al crear el modelo.

**Puntos fuertes.** El clúster Bietápico puede gestionar distintos tipos de campos mezclados y conjuntos de datos grandes con eficacia. También tiene capacidad para comprobar varias soluciones de clústeres y seleccionar la mejor, por lo que no tendrá que saber el número de clústeres que hay que pedir al comienzo. El clúster Bietápico se puede configurar para que excluya automáticamente **valores atípicos** o casos muy extraños que puedan contaminar sus resultados.

**Importante:**

IBM SPSS Modeler tiene dos versiones distintas del nodo de clúster TwoStep:

- **Clúster TwoStep** es el nodo tradicional que se ejecuta en IBM SPSS Modeler Server.
- **Clúster TwoStep-AS** se puede ejecutar cuando está conectado a IBM SPSS Analytic Server.

## Opciones de modelo para el nodo de clúster bietápico

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Estandarizar campos numéricos.** De forma predeterminada, el nodo Bietápico estandariza todos los campos de entrada numéricos a la misma escala, con una media de 0 y una varianza de 1. Para conservar la escala original de los campos numéricos, anule la selección de esta opción. Los campos simbólicos no se ven afectados.

**Excluir valores atípicos.** Seleccione esta opción para que los registros que no parezcan encajar en un clúster significativo se excluyan automáticamente del análisis. De este modo evitará que estos casos distorsionen los resultados.

La detección de valores atípicos se produce durante el paso de preclúster. Cuando se selecciona esta opción, los subclústeres con pocos registros relativos a otros subclústeres se consideran valores atípicos potenciales y se vuelve a crear el árbol de subclústeres excluyendo esos registros. El tamaño por debajo del cual se considera que los subclústeres contienen posibles valores atípicos está controlado por la opción **Porcentaje**. Algunos de esos registros de valores atípicos potenciales pueden añadirse a los subclústeres creados de nuevo, si son lo suficientemente similares a alguno de los nuevos perfiles de subclústeres. Los demás valores atípicos potenciales que no se puedan fundir se considerarán valores atípicos, se añadirán a un clúster "ruido" y se excluirán del paso de agrupación en clústeres jerárquica.

Al *puntuar* datos con un modelo Bietápico que utiliza el tratamiento de los valores atípicos, los nuevos casos que estén a una distancia de umbral excesiva (basándose en el logaritmo de la verosimilitud) del clúster significativo más cercano, se consideran valores atípicos y se asignan al clúster "ruido" con el nombre -1.

**Etiqueta de clúster.** Especifique el formato del campo de pertenencia al clúster generado. La pertenencia a un clúster se puede indicar con una **Cadena** con el **Prefijo de etiqueta** especificado (por ejemplo `"Clúster 1"`, `"Clúster 2"`, etc.) o con un **Número**.

**Calcular automáticamente número de clústeres.** El clúster Bietápico puede analizar rápidamente un gran número de soluciones de clúster para seleccionar el número óptimo de clústeres para los datos de entrenamiento. Especifique un rango de soluciones que comprobar estableciendo el número **Máximo** y **Mínimo** de clústeres. Bietápico utiliza un proceso de dos etapas para determinar el número óptimo de clústeres. En la primera etapa se selecciona un límite superior para el número de clústeres del modelo en función del cambio del Criterio de información bayesiano (BIC) a medida que se van añadiendo clústeres. En la segunda etapa se encuentra el cambio de la distancia mínima entre clústeres para todos los modelos con menor número de clústeres que el mínimo de la solución de BIC. El mayor cambio de distancia se utiliza para identificar el modelo de clúster final.

**Especificar número de clústeres.** Si conoce el número de clústeres que incluir en el modelo, seleccione esta opción e introduzca dicho número.

**Medida de distancia.** Esta opción determina cómo se calcula la similaridad entre dos clústeres.

- **Log-verosimilitud.** La medida de la verosimilitud realiza una distribución de probabilidad entre las variables. Las variables continuas se supone que tienen una distribución normal, mientras que las variables categóricas se supone que son multinomiales. Se supone que todas las variables son independientes.
- **Euclídea.** La medida euclídea es la distancia según una "línea recta" entre dos clústeres. Sólo se puede utilizar cuando todas las variables son continuas.

**Criterio de agrupación en clústeres.** Esta opción determina cómo el algoritmo de agrupación en clústeres determina el número de clústeres. Se puede especificar tanto el criterio de información bayesiano (BIC) como el criterio de información de Akaike (AIC).

## Nugget de modelo de clúster Bietápico

Los nugget de modelo de clúster Bietápico contienen toda la información capturada por el modelo de agrupación en clústeres, así como información acerca de los datos de entrenamiento y el proceso de estimación.

Cuando se ejecuta una ruta que contiene un nugget de modelo Bietápico, el nodo añade un nuevo campo que contiene la pertenencia al clúster para ese registro. El nuevo nombre de campo se deriva del nombre del modelo, con el prefijo *$T-*. Por ejemplo, si el modelo se llama *Bietápico*, los nuevos campos se llamarían *$T-Bietápico*.

Una buena forma de comprender el modelo Bietápico es utilizar la inducción de reglas para descubrir las características que distinguen los clústeres que el modelo ha encontrado. Consulte el tema “Nodo C5.0” en la página 110 para obtener más información. Asimismo, puede pulsar en la pestaña Modelo del explorador de nugget de modelo para ver el visor de clústeres, que proporciona una representación gráfica de clústeres, campos y niveles de importancia. Consulte el tema “Visor de clústeres - Pestaña Modelo” en la página 259 para obtener más información.

Para obtener información general sobre la utilización del explorador del modelo, consulte “Examen de nuggets de modelo” en la página 42

## Resumen de modelo bietápico

La pestaña Resumen de un nugget de modelo de clúster Bietápico muestra el número de clústeres encontrados, junto con la información acerca de los datos de entrenamiento, el proceso de estimación y la configuración de creación utilizada.

Consulte el tema “Examen de nuggets de modelo” en la página 42 para obtener más información.

# Nodo de clúster TwoStep-AS

IBM SPSS Modeler tiene dos versiones distintas del nodo de clúster TwoStep:

- **Clúster TwoStep** es el nodo tradicional que se ejecuta en IBM SPSS Modeler Server.
- **Clúster TwoStep-AS** se puede ejecutar cuando está conectado a IBM SPSS Analytic Server.

# Análisis de clúster Bietápico-AS

El clúster Bietápico es una herramienta de exploración diseñada para revelar agrupaciones naturales (o clústeres) dentro de un conjunto de datos que de otro modo no serían evidentes. El algoritmo que emplea este procedimiento incluye varias atractivas características que lo hacen diferente de las técnicas de agrupación en clústeres tradicionales:

- **Tratamiento de variables categóricas y continuas.** Al suponer que las variables son independientes, es posible aplicar una distribución normal multinomial conjunta en las variables continuas y categóricas.
- **Selección automática del número de clústeres.** Mediante la comparación de los valores de un criterio de selección del modelo para diferentes soluciones de agrupación en clústeres, el procedimiento puede determinar automáticamente el número óptimo de clústeres.
- **Escalabilidad.** Mediante la construcción de un árbol de características de clústeres (CF) que resume los registros, el algoritmo bietápico puede analizar archivos de datos de gran tamaño.

Por ejemplo, las empresas minoristas y de productos de consumo aplican habitualmente técnicas para obtener información que describe los hábitos de consumo, el sexo, la edad, el nivel de ingresos y otros atributos de sus clientes. Estas empresas adaptan sus estrategias de marketing y desarrollo de productos a cada grupo de consumidores para aumentar las ventas y el nivel de fidelidad a la marca.

## Pestaña Campos

La pestaña Campos especifica los campos que se utilizan en el análisis.

**Utilizar roles predefinidos**. Se seleccionan todos los campos con un rol definido de entrada.

**Utilizar asignaciones de campos personalizadas**. Añadir y eliminar campos independientemente de sus asignaciones de rol definidas. Puede seleccionar campos con cualquier rol y ponerlos o sacarlos de la lista **Predictores (entradas)**.

## Información básica

### Número de clústeres

**Determinar automáticamente**
: El procedimiento determinas el mejor número de clústeres dentro del intervalo especificado. El valor **mínimo** debe ser superior a 1. Esta es la opción predeterminada.

**Especificar número fijo**
: El procedimiento genera el número especificado de clústeres. El **número** debe ser mayor que 1.

### Criterio de agrupación en clústeres

Esta opción controla cómo el algoritmo de agrupación en clústeres determina el número de clústeres.

**Criterio de información bayesiano (BIC)**
: Una medida para seleccionar y comparar modelos basados en el logaritmo de la verosimilitud -2. Los valores menores indican modelos mejores. El BIC también "penaliza" modelos sobreparametrizados (modelos complejos con un gran número de entradas, por ejemplo), pero de forma más estricta que el AIC.

**Criterio de información de Akaike (AIC)**
Una medida para seleccionar y comparar modelos basados en el logaritmo de la verosimilitud -2. Los valores menores indican modelos mejores. El AIC "penaliza" modelos sobreparametrizados (modelos complejos con un gran número de entradas, por ejemplo).

### Método de agrupación en clústeres automático

Si selecciona **Determinar automáticamente**, seleccione uno de los siguientes métodos de agrupación en clúster utilizados para determinar automáticamente el número de clústeres:

**Utilización de la configuración del criterio de agrupación en clústeres**
La convergencia de los criterios de información es la tasa de criterios de información correspondiente a dos soluciones de clúster actuales y a la primera solución de clúster. El criterio usado es el seleccionado en el grupo Criterio de agrupación en clústeres.

**Salto de distancia**
El salto de distancia es la tasa de distancias correspondientes a dos soluciones de clúster consecutivas.

**Máximo**
Combine los resultados del método de convergencia de los criterios de información y el método de salto de distancia para producir el número de clústeres que corresponde al segundo salto.

**Mínimo**
Combine los resultados del método de convergencia de los criterios de información y el método de salto de distancia para producir el número de clústeres que corresponde al primer salto.

### Método de importancia de características

El **método de importancia de características** determina la importancia de las características (los campos) en la solución en clúster. La salida incluye información sobre la importancia global de las características y la importancia de cada campo de característica en cada clúster. Las características que no cumplen con un umbral mínimo se excluyen.

**Utilización de la configuración del criterio de agrupación en clústeres**
Este es el método predeterminado basado en el criterio seleccionado en el grupo Criterio de agrupación en clústeres.

**Tamaño del efecto**
La importancia de la característica se basa en el tamaño del efecto en lugar de en los valores de significación.

## Función de tres criterios

Esta configuración determina cómo se genera el árbol de características del clúster. Generando un árbol de características del clúster y resumiendo los registros, el algoritmo Bietápico puede analizar grandes archivos de datos. Dicho de otro modo, el clúster Bietápico utiliza un árbol de características de clúster para generar clústeres, lo cual le permite procesar numerosos casos.

### Medida de distancia

Esta opción determina cómo se calcula la similaridad entre dos clústeres.

**Log-verosimilitud**

La medida de la verosimilitud realiza una distribución de probabilidad entre los campos. Los campos continuos se supone que tienen una distribución normal, mientras que los campos categóricos se supone que son multinominales. Se supone que todos los campos son independientes.

**Euclidean**

La medida euclídea es la distancia según una "línea recta" entre dos clústeres. La medida euclídea cuadrada y el método Ward se usan para calcular la similitud entre clústeres. Sólo se puede utilizar cuando todos los campos son continuos.

### Clústeres atípicos

**Incluir clústeres atípicos**
Incluya clústeres para casos que son atípicos respecto a los clústeres habituales. Si no se selecciona esta opción, en los clústeres habituales se incluyen todos los casos.

**El número de casos en la hoja del árbol de características es menor que.**
Si el número de casos de la hoja de árbol de características es menor que el valor indicado, la hoja se considera un valor atípico. El valor debe ser un entero mayor que 1. Si cambia este valor, es posible que los valores superiores den como resultado más clústeres atípicos.

**Porcentaje superior de valores atípicos.**
Cuando se genera el modelo de clústeres, los valores atípicos se clasifican según la fuerza de los valores atípicos. La fuerza de los valores atípicos que se necesita para estar en el porcentaje superior de valores atípicos se utiliza como umbral para determinar si un caso se clasifica como atípico. Los valores superiores significan que hay más casos clasificados como atípicos. El valor debe estar entre 1 y 100.

### Valores adicionales

**Umbral de cambio de distancia inicial**

Umbral inicial que se utiliza para hacer crecer el árbol de características del clúster. Si la inserción de una hoja en una hoja del árbol produce una densidad inferior a este umbral, la hoja no se divide. Si la densidad supera este umbral, se dividirá la hoja.

**Máximo de ramas del nodo de hoja**

Número máximo de nodos hijo que puede tener un nodo hoja.

**Máximo de ramas de un nodo que no sea hoja**

Número máximo de nodos hijo que puede tener un nodo que no sea hoja.

**Máxima profundidad de árbol**

Número máximo de niveles que puede tener un árbol de clúster.

**Peso de ajuste en el nivel de medición**
Reduce la influencia de los campos categóricos aumentando el peso de los campos continuos. Este valor representa un denominador para reducir el peso de los campos categóricos. Por lo tanto, un valor predeterminado de 6, por ejemplo, da a los campos categóricos un peso de 1/6.

**Asignación de memoria**

La cantidad máxima de memoria en megabytes (MB) que el algoritmo de agrupación en clústeres utiliza. Si el procedimiento supera este máximo, utiliza el disco para almacenar la información que no se pueda colocar en la memoria.

**División atrasada**
Nueva creación atrasada del árbol de características del clúster. El algoritmo de agrupación en clústeres vuelve a generar el árbol de características de clústeres varias veces mientras evalúa nuevos casos. Esta opción puede mejorar el rendimiento atrasando esa operación y reduciendo el número de veces que se vuelve a genera el árbol.

## Estandarizar

El algoritmo de agrupación en clústeres trabaja con campos continuos estandarizados. De forma predeterminada, todos los campos continuos están estandarizados. Para ahorrar tiempo y esfuerzo de cálculo, puede mover campos continuos que ya están estandarizados a la lista **No estandarizar**.

## Sel. características

En la pantalla Selección de características, puede establecer reglas que determinen cuándo se excluyen los campos. Por ejemplo, puede excluir campos en los que falten numerosos valores.

### Reglas para excluir campos

**El porcentaje de valores que faltan es mayor que.**
Los campos con un porcentaje de valores que faltan superior al valor indicado se excluyen del análisis. El valor debe ser un número positivo mayor que cero y menor que 100.

**El número de categorías para campos categóricos es mayor que.**
Los campos categóricos con un número de categorías superior al especificado se excluyen del análisis. El valor debe ser un entero positivo mayor que 1.

**Campos con una tendencia hacia un valor único**

**El coeficiente de variación para los campos continuos es menor que.**
Los campos continuos con un coeficiente de variación menor que el valor especificado se excluyen del análisis. El coeficiente de variación es la tasa del desvío estándar respecto de la media. Los valores inferiores tienden a indicar una variación de valores inferior. El valor debe estar entre 0 y 1.

**El porcentaje de casos en una categoría única para los campos categóricos es mayor que.**
Los campos categóricos con un porcentaje de casos en una categoría única superior al valor indicado se excluyen del análisis. El valor debe ser mayor que 0 y menor que 100.

### Selección de características adaptativas

Esta opción realiza una lectura adicional de datos para buscar y eliminar los campos menos importantes.

## Resultado de modelo

### Resumen de generación de modelos

**Especificaciones de modelos**
Resumen de las especificaciones de modelo, número de clústeres en el modelo final, y entradas (campos) incluidos en el modelo final.

**Resumen de registros**
Número y porcentaje de registros (casos) incluidos y excluidos del modelo.

**Entradas excluidas**
Para los campos no incluidos en el modelo final, la razón por la que el campo se ha excluido.

### Evaluación

**Calidad del modelo**
Tabla de bondad e importancia para cada clúster y bondad de ajuste del modelo global.

**Diagrama de barras de importancia de la característica**
Diagrama de barras de la importancia de la característica (campo) entre todos los clúster. Las características (campos) con barras más largas en el diagrama son más importantes que los campos con barras más cortas. También se clasifican en orden descendente de importancia (la barra en la parte superior es la más importante).

**Nube de palabras de importancia de la característica**
Nube de palabras de la importancia de la característica (campo) entre todos los clúster. Las características (campos) con el texto más grande son más importantes que los que tienen el texto más pequeño.

**Clústeres de valores atípicos**
Estas opciones están inhabilitadas si elige no incluir valores atípicos.

**Tabla y gráfico interactivos**
Tabla y gráfico de fuerza de valor atípico y la similitud relativa de clústeres de valor atípico y clústeres regulares. Al seleccionar filas diferentes en la tabla se visualiza información para los distintos clústeres de valor atípico en el gráfico.

**Tabla dinámica**
Tabla de fuerza de valor atípico y la similitud relativa de clústeres de valor atípico y clústeres regulares. Esta tabla contiene la misma información que la visualización interactiva. Esta tabla soporta todas las funciones estándar para tablas dinámicas y de edición.

**Número máximo**
El número máximo de valores atípicos a visualizar en la salida. Si hay más de veinte clústeres de valor atípico, en su lugar se muestra una tabla dinámica.

### Interpretación

**Perfiles de importancia de la característica entre clústeres**

**Tabla y gráfico interactivos.**
Tabla y gráficos de importancia de característica y centros de clúster para cada entrada (campo) utilizada en la solución de clúster. Al seleccionar filas diferentes en la tabla se visualiza un gráfico distinto. En los campos categóricos se muestra un gráfico de barras. Para los campos continuos, se visualiza un gráfico de medias y desviaciones estándar.

**Tabla dinámica.**
Tabla de importancia de característica y centros de clúster para cada entrada (campo). Esta tabla contiene la misma información que la visualización interactiva. Esta tabla soporta todas las funciones estándar para tablas dinámicas y de edición.

**Importancia de la característica intra-clúster**
Para cada clúster, el centro del clúster y la importancia de la característica para cada entrada (campo). Existe una tabla aparte para cada clúster.

**Distancias entre clústeres**
Un gráfico con panel que muestra las distancias entre los clústeres. Existe un panel aparte para cada clúster.

### Etiqueta de clúster

**Texto** La etiqueta para cada clúster es el valor que se ha especificado para **Prefijo**, seguido de un número secuencial.

**Número**
La etiqueta para cada clúster es un número secuencial.

## Opciones de Modelo

**Nombre del modelo.** Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

---

# Nugget de modelo de clúster TwoStep-AS

El nugget de modelo TwoStep-AS muestra los detalles del modelo en la pestaña Modelo del Visor de salida. Para obtener más información sobre cómo utilizar el visor, consulte la sección titulada "Cómo trabajar con la salida" en la Guía del usuario de Modeler (ModelerUsersGuide.pdf).

Los nugget de modelo de clúster TwoStep-AS contienen toda la información capturada por el modelo de agrupación en clústeres, así como información acerca de los datos de entrenamiento y el proceso de estimación.

Cuando se ejecuta una ruta que contiene un nugget de modelo TwoStep-AS, el nodo añade un nuevo campo que contiene la pertenencia al clúster para ese registro. El nuevo nombre de campo se deriva del nombre del modelo, con el prefijo ***$AS-***. Por ejemplo, si el modelo se denomina TwoStep, el campo nuevo se llamará ***$AS-TwoStep***.

Una buena forma de comprender el modelo TwoStep-AS es utilizar la inducción de reglas para descubrir las características que distinguen los clústeres que el modelo ha encontrado. Consulte el tema "Nodo C5.0" en la página 110 para obtener más información.

Si desea obtener información general sobre el explorador de modelos, consulte "Examen de nuggets de modelo" en la página 42

## Configuración de nugget de modelo Clúster TwoStep-AS

La pestaña Configuración proporciona opciones adicionales para el nugget de modelo TwoStep-AS.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar convirtiendo a SQL nativo** Si selecciona esta opción, se genera SQL para puntuar el modelo dentro de la base de datos.

  **Nota:** aunque esta opción puede proporcionar resultados más rápidos, el tamaño y la complejidad del SQL nativo aumenta a medida que lo hace la complejidad del modelo.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

---

# Nodo de K-Medias-AS

k-medias es uno de los algoritmos de agrupación en clúster utilizado con más frecuencia. Agrupa en clúster puntos de datos en un número predefinido de clústeres.[1] El nodo K-Medias-AS en SPSS Modeler se implementa en Spark.

Si desea más detalles sobre algoritmos de k-medias, consulte https://spark.apache.org/docs/2.2.0/ml-clustering.html.

Tenga en cuenta que el nodo K-Medias-AS realiza una codificación "one-hot" automáticamente para variables categóricas.

[1] "Clustering." *Apache Spark*. MLlib: Main Guide. Web. 3 Oct 2017.

## Campos de nodo K-Medias-AS

La pestaña Campos especifica los campos que se utilizan en el análisis.

**Utilizar roles predefinidos.** Esta opción permite indicar al nodo que use la información de campo de un nodo Tipo situado en un punto anterior de la ruta. Está seleccionado de forma predeterminada.

**Utilizar asignaciones de campos personalizadas.** Si desea asignar manualmente campos de entrada, seleccione esta opción y, después, seleccione el campo o los campos de entrada. El uso de esta opción es similar a establecer el rol del campo en **Input** en un nodo Tipo.

# Opciones de generación del nodo K-Medias-AS

Utilice la pestaña Opciones de generación para especificar opciones de generación para el nodo K-Medias-AS, que incluye opciones regulares para la creación de modelos, opciones de inicialización para inicializar centros de clúster y opciones avanzadas para la semilla aleatoria y la iteración de cálculo. Si desea más información, consulte el JavaDoc para K-Medias en SparkML.[1]

## Regular

**Nombre de modelo.** El nombre del campo generado después de puntuar un clúster específico. Seleccione **Auto** (valor predeterminado) o seleccione **Personalizado** y escriba un nombre.

**Número de clústeres.** Especifique el número de clústeres que generar. El valor predeterminado es **5** y el mínimo es **2**.

## Inicialización

**Modalidad de inicialización.** Especifique el método para inicializar los centros de clúster. **K-Means\|\|** es el valor predeterminado. Si desea más detalles sobre estos dos métodos, consulte Scalable K-Means++. [2]

**Pasos de inicialización.** Si está seleccionada la modalidad de inicialización **K-Means\|\|**, especifique el número de pasos de inicialización. **2** es el valor predeterminado.

## Avanzado

**Configuración avanzada.** Seleccione esta opción si desea establecer opciones avanzada del modo siguiente.

**Iteración máx.** Especifique el número máximo de iteraciones para realizar al buscar en centros de clúster. **20** es el valor predeterminado.

**Tolerancia.** Especifique la tolerancia de convergencia para algoritmo iterativos. **1.0E-4** es el valor predeterminado.

**Establecer semilla aleatoria.** Seleccione esta opción y pulse **Generar** para generar la semilla utilizada por el generador de números aleatorios.

## Visualización

**Mostrar gráfico.** Seleccione esta opción si desea que se incluya un gráfico en la salida.

La tabla siguiente muestra la relación entre los valores de los parámetros Spark del nodo K-Medias-AS y K-Medias de SPSS Modeler.

*Tabla 13. Propiedades de nodo correlacionadas con parámetros Spark*

| Valor de SPSS Modeler | Nombre de script (nombre de propiedad) | Parámetro SparkML de k-medias |
|---|---|---|
| Campos de entrada | `features` | |
| Número de clústeres | `clustersNum` | `k` |
| Modalidad de inicialización | `initMode` | `initMode` |
| Pasos de inicialización | `initSteps` | `initSteps` |

*Tabla 13. Propiedades de nodo correlacionadas con parámetros Spark (continuación)*

| Valor de SPSS Modeler | Nombre de script (nombre de propiedad) | Parámetro SparkML de k-medias |
|---|---|---|
| Iteración máx | maxIter | maxIter |
| Tolerancia | toleration | tol |
| Semilla aleatoria | randomSeed | seed |

[1] "Class KMeans." *Apache Spark*. JavaDoc. Web. 3 Oct 2017.

[2] Bahmani, Moseley, et al. "Scalable K-Means++." Feb 28, 2012. http://theory.stanford.edu/%7Esergei/papers/vldb12-kmpar.pdf.

## El visor de clústeres

Los modelos de clústeres se suelen utilizar para buscar grupos (o clústeres) de registros similares basados en las variables examinadas, donde la similitud entre los miembros del mismo grupo es alta y es baja entre miembros de grupos diferentes. Los resultados pueden utilizarse para identificar asociaciones que de otra forma no serían evidentes. Por ejemplo, es posible utilizar un análisis de clústeres de preferencias de clientes, nivel de ingresos y hábitos de compra para identificar los tipos de clientes más propensos a responder a una campaña de marketing concreta.

Existen dos métodos para interpretar los resultados de una visualización de clústeres:

- Examinar los clústeres para determinar las características exclusivas de cada clúster. *¿Contiene uno de los clústeres todos los socios con un alto nivel de ingresos? ¿Contiene este clúster más registros que los demás?*
- Examinar los campos de todos los clústeres para determinar la forma en que los valores se distribuyen en ellos. *¿Afecta el nivel de educación a la inclusión en un clúster? ¿Sirve una puntuación alta de crédito para distinguir entre la pertenencia a un clúster o a otro?*

Puede utilizar las vistas principales y las diferentes vistas vinculadas en el visor de clústeres para obtener una mayor perspectiva que le ayuda a responder a estas preguntas.

En IBM SPSS Modeler es posible generar los siguientes nuggets de modelos de clúster:

- Nugget de modelo de red de Kohonen
- Nugget de modelo de K-medias
- Nugget de modelo de clúster Bietápico

Para ver información sobre los nuggets de modelos de clúster, pulse con el botón derecho en el nodo del modelo y seleccione **Examinar** en el menú contextual (o **Modificar** en los nodos de una transmisión). Además, si está utilizando el nodo de modelado de clúster automático, pulse dos veces en el nugget de clústeres que proceda, dentro del nugget de modelo Clúster automático. Consulte el tema "Nodo Autoclúster" en la página 79 para obtener más información.

## Visor de clústeres - Pestaña Modelo

La pestaña Modelos de los modelos de clúster muestra una visualización gráfica de estadísticas y distribuciones de resúmenes para campos entre clústeres, que se conoce como el **Visor de clústeres**.

*Nota*: La pestaña Modelo no está disponible para modelos creados en versiones de IBM SPSS Modeler anteriores a la 13.

El Visor de clústeres se compone de dos paneles, la vista principal en la parte izquierda y la vista relacionada o auxiliar de la derecha. Hay dos vistas principales:

- Resumen del modelo (valor predeterminado). Consulte el tema “Vista Resumen del modelo” para obtener más información.
- Clústeres. Consulte el tema “Vista de clústeres” para obtener más información.

Hay cuatro vistas relacionadas/auxiliares:

- Importancia del predictor. Consulte el tema “Vista Importancia del predictor de clústeres” en la página 262 para obtener más información.
- Tamaños de clústeres (valor predeterminado) Consulte el tema “Vista de tamaños de clústeres” en la página 262 para obtener más información.
- Distribución de casillas. Consulte el tema “Vista Distribución de casillas” en la página 262 para obtener más información.
- Comparación de clústeres. Consulte el tema “Vista Comparación de clústeres” en la página 262 para obtener más información.

## Vista Resumen del modelo

La vista Resumen del modelo muestra una instantánea o resumen del modelo de clúster, incluyendo una medida de silueta de la cohesión y separación de clústeres sombreada para indicar resultados pobres, correctos o buenos. Esta instantánea le permite comprobar rápidamente si la calidad es insuficiente, en cuyo caso puede optar por volver al nodo de modelado para cambiar los ajustes del modelo de clúster para producir mejores resultados.

Los resultados serán pobres, correctos o buenos de acuerdo con el trabajo de Kaufman y Rousseeuw (1990) sobre la interpretación de estructuras de clústeres. En la vista Resumen del modelo, un resultado "bueno" indica que los datos reflejan una evidencia razonable o sólida de que existe una estructura de clústeres, de acuerdo con la valoración Kaufman y Rousseeuw; una resultado "correcto" indica que ese evidencia es débil, y un resultado "pobre" significa que, según esa valoración, no hay evidencias obvias.

Los promedios de medida de silueta, en todos los registros, (B−A) / max(A,B), donde A es la distancia del registro al centro de clúster y B es la distancia del registro al centro del clúster más cercano al que no pertenece. Un coeficiente de silueta de 1 podría implicar que todos los casos están ubicados directamente en sus centros de clúster. Un valor de -1 significaría que todos los casos se encuentran en los centros de clúster de algún otro clúster. Un valor de 0 implica, de media, que los casos están equidistantes entre el centro de su propio clúster y el siguiente clúster más cercano.

El resumen incluye una tabla que contiene la siguiente información:

- **Algoritmo.** El algoritmo de agrupación en clústeres utilizado, por ejemplo, "Dos fases".
- **Características de entrada.** El número de campos, también conocidos como **entradas** o **predictores**.
- **Clústeres.** Número de clústeres de la solución.

## Vista de clústeres

La vista Clústeres contiene una cuadrícula de clústeres por características que incluye nombres de clústeres, tamaños y perfiles para cada clúster.

Las columnas de la cuadrícula contienen la siguiente información:

- **Clúster.** Números de clústeres creados por el algoritmo.
- **Etiqueta.** Etiquetas aplicadas a cada clúster (está en blanco de forma predeterminada). Pulse dos veces la casilla para introducir una etiqueta que describa el contenido del clúster, por ejemplo "Compradores de automóviles de lujo".
- **Descripción.** Cualquier descripción de los contenidos de los clústeres (está en blanco de forma predeterminada). Pulse dos veces la casilla para introducir una descripción del clúster, por ejemplo "Más de 55 años de edad, profesionales, con ingresos superiores a 100.000 €".

- **Tamaño.** El tamaño de cada clúster como porcentaje de la muestra general del clúster. Cada casilla de tamaño de la cuadrícula muestra una barra vertical que muestra el porcentaje de tamaño del clúster, un porcentaje de tamaño en formato numérico y los recuentos de casos de clúster.
- **Características.** Los predictores o entradas individuales, ordenados por importancia general de forma predeterminada. Si hay columnas con tamaños iguales, se muestran en orden ascendente en función de los miembros del clúster.

  La importancia general de la característica se indica por el color del sombreado del fondo de la casilla, siendo más oscuro cuanto más importante sea la característica. Una guía sobre la tabla indica la importancia vinculada a cada color de casilla de característica.

Cuando pasa el ratón por una casilla, se muestra el nombre completo/etiqueta de la característica y el valor de importancia de la casilla. Es posible que aparezca más información, en función de la vista y tipo de característica. En la vista Centros de clústeres, esto incluye la estadística de casilla y el valor de la casilla, por ejemplo: "Media: 4.32". En las características categóricas, la casilla muestra el nombre de la categoría (modal) más frecuente y su porcentaje.

En la vista Clústeres, puede seleccionar varias formas de mostrar la información de clústeres:
- Transponer clústeres y características Consulte el tema "Transponer clústeres y características" para obtener más información.
- Clasificar características Consulte el tema "Clasificar características" para obtener más información.
- Clasificar clústeres Consulte el tema "Clasificar clústeres" para obtener más información.
- Seleccionar contenido de casilla Consulte el tema "Contenido de casilla" para obtener más información.

**Transponer clústeres y características:** De forma predeterminada, los clústeres se muestran como columnas y las características aparecen como filas. Para invertir esta visualización, pulse el botón **Transponer clústeres y características** a la izquierda de los botones **Clasificar características**. Por ejemplo, puede que desea hacer esto cuando se muestren muchos clústeres para reducir la cantidad de desplazamiento horizontal necesario para visualizar los datos.

**Clasificar características:** Los botones **Clasificar características por** le permiten seleccionar la cantidad de casillas de características:
- **Importancia global** Este es el orden de clasificación predeterminado. Las características se clasifican en orden descendente de importancia general y el orden de clasificación es el mismo entre los distintos clústeres. Si hay características que empatan en valores de importancia, éstas se muestran en orden de clasificación ascendente según el nombre.
- **Importancia dentro del clúster** Las características se clasifican con respecto de su importancia para cada clúster. Si hay características que empatan en valores de importancia, éstas se muestran en orden de clasificación ascendente según el nombre. Si esta opción está seleccionada, el orden de clasificación suele variar en los diferentes clústeres.
- **Nombre.** Las características se clasifican por nombre en orden alfabético.
- **Orden de los datos** Las características se clasifican por orden en el conjunto de datos.

**Clasificar clústeres:** De forma predeterminada, los clústeres se clasifican en orden de tamaño descendente. Los botones **Clasificar clústeres por** le permiten ordenarlos por nombre en orden alfabético o, si ha creado etiquetas exclusivas, por orden de etiqueta alfanumérico.

Las características con la misma etiqueta se clasifican por nombre de clúster. Si los clústeres se clasifican por etiqueta y modifica la etiqueta de un clúster, el orden de clasificación se actualiza automáticamente.

**Contenido de casilla:** Los botones **Casillas** le permiten cambiar la visualización del contenido de casillas de características y campos de evaluación.

- **Centros de los clústeres.** De forma predeterminada, las casillas muestran nombres/etiquetas de características y la tendencia central para cada combinación de clúster/característica. La media se muestra para los campos continuos y el modo (categoría más frecuente) con porcentaje de categoría para los campos categóricos.
- **Distribuciones absolutas.** Muestra nombres/etiquetas de características y distribuciones absolutas de las características de cada clúster. En el caso de las características categóricas, la visualización muestra gráficos de barras superpuestas con las categorías ordenadas en orden ascendente de valores de datos. En las características continuas, la visualización muestra un gráfico de densidad suave que utiliza los mismos puntos finales e intervalos para cada clúster.

  La visualización en color rojo oscuro muestra la distribución de clústeres, mientras que la más clara representa los datos generales.
- **Distribuciones relativas** Muestra los nombres/etiquetas de características y las distribuciones relativas en las casillas. En general, las visualizaciones son similares a las mostradas para las distribuciones absolutas, sólo que en su lugar se muestran distribuciones relativas.

  La visualización en color rojo oscuro muestra la distribución de clústeres, mientras que la más clara representa los datos generales.
- **Vista básica.** Si hay muchos clústeres, puede resultar difícil ver todos los detalles sin desplazarse. Para reducir la cantidad de desplazamiento, seleccione esta vista para cambiar la visualización a una versión más compacta de la tabla.

## Vista Importancia del predictor de clústeres

La vista Importancia del predictor muestra la importancia relativa de cada campo en la estimación del modelo.

## Vista de tamaños de clústeres

La vista Tamaños de clústeres muestra el gráfico circular que contiene cada clúster. El tamaño de porcentaje de cada clúster se muestra en cada porción, pase el ratón sobre cada porción para mostrar el recuento de esa porción.

Bajo el gráfico, una tabla enumera la siguiente información de tamaño:

- El tamaño del clúster más pequeño (un recuento y porcentaje del conjunto).
- El tamaño del clúster mayor (un recuento y porcentaje del conjunto).
- La proporción entre el tamaño del mayor clúster y el del menor.

## Vista Distribución de casillas

La vista Distribución de casillas muestra un gráfico expandido y más detallado de la distribución de los datos para cualquier casilla de característica que seleccione en el panel principal Clústeres.

## Vista Comparación de clústeres

La vista Comparación de clústeres se compone de un diseño en estilo de cuadrícula, con características en las filas y clústeres seleccionados en las columnas. Esta vista le ayuda a entender mejor los factores de los que se componen los clústeres, y le permite ver las diferencias entre los clústeres no sólo con respecto a los datos generales, sino entre sí.

Para seleccionar clústeres para su visualización, pulse en la parte superior de la columna del clúster en el panel principal Clústeres. Pulse las teclas Ctrl o Mayús y pulse para seleccionar o cancelar la selección de más de un clúster para su comparación.

*Nota*: Puede seleccionar hasta cinco clústeres para su visualización.

Los clústeres se muestran en el orden en que se seleccionaron, mientras que el orden de los campos viene determinado por la opción **Clasificar características por**. Si selecciona **Importancia dentro del clúster**, los campos siempre se clasifican por importancia general.

Los gráficos de fondo muestran las distribuciones generales de cada característica:

- Las características categóricas aparecen como gráficos de puntos, donde el tamaño del punto indica la categoría más frecuente/modal para cada clúster (por característica).
- Las características continuas se muestran como diagramas de caja, que muestran las medianas globales y los rangos intercuartiles.

En estas vistas de fondo aparecen superpuestos diagramas de caja para los clústeres seleccionados:

- En las características continuas hay marcadores de puntos cuadrados y líneas horizontales que indican la mediana y el rango intercuartil de cada clúster.
- Cada clúster viene representado por un color distinto, que se muestra en la parte superior de la vista.

## Navegación en el Visor de clústeres

El visor de clústeres es una pantalla interactiva. Puede:

- Seleccionar un campo o clúster para ver más detalles.
- Comparar clústeres para seleccionar elementos de interés.
- Alterar la visualización.
- Transponer ejes.
- Generar nodos Derivar, Filtrar y Seleccionar mediante el menú Generar.

Uso de las barras de herramientas

Puede controlar la información que aparece en los paneles izquierdo y derecho mediante las opciones de la barra de herramientas. Puede cambiar la orientación de la pantalla (de arriba a abajo, de izquierda a derecha, o de derecha a izquierda) mediante los controles de la barra de herramientas. Además, también puede restablecer el visor a los ajustes predeterminados, y abrir un cuadro de diálogo para especificar el contenido de la vista Clústeres en el panel principal.

Las opciones **Clasificar características por**, **Clasificar clústeres por**, **Casillas** y **Mostrar** sólo están disponibles cuando selecciona la vista **Clústeres** en el panel principal. Consulte el tema "Vista de clústeres" en la página 260 para obtener más información.

*Tabla 14. Iconos de barra de herramientas.*

| Icono | Tema |
|---|---|
|  | Consulte Transponer clústeres y características |
|  | Consulte Clasificar características por |
|  | Consulte Clasificar clústeres por |
|  | Consulte Casillas |

Generación de nodos a partir de modelos de clústeres

El menú Generar le permite crear nuevos nodos basados en el modelo de clúster. Esta opción está disponible desde la pestaña Modelo del modelo generado, y le permite generar nodos en función de la visualización o selección actual (es decir, todos los clústeres visibles o todos los seleccionados). Por

ejemplo, puede seleccionar una única característica y generar un nodo Filtrar para descartar todas las demás (no visibles). Los nodos generados aparecen sin conexión en el lienzo. Además, puede generar una copia del nugget de modelo en la paleta de modelos. No olvide conectar los nodos y realizar las modificaciones que desee antes de la ejecución.

- **Generar nodo de modelado** Crea un nodo de modelado en el lienzo de rutas. Esto puede resultar útil, por ejemplo, si tiene una transmisión en la que desea utilizar estos ajustes de modelo, pero ya no tiene el nodo de modelado utilizado para generarlos.
- **Modelo a paleta** Crea un nugget en la paleta Modelos. Esto resulta útil en las situaciones en que un colega le haya enviado una ruta que contenga un modelo y no el modelo en sí.
- **Nodo Filtrar** Crea un nuevo nodo Filtrar para filtrar los campos que no se utilicen en el modelo de clúster o que no sea visible en la visualización Visor de clústeres actual. Si hay un nodo Tipo anterior a este nodo Clúster, cualquier campo con el rol *destino* queda descartado por el nodo Filtrar generado.
- **Nodo Filtrar (desde la selección)** Crea un nuevo nodo Filtrar para filtrar los campos en función de las selecciones del Visor de clústeres. Seleccione varios campos manteniendo pulsada la tecla Ctrl mientras selecciona los campos. Los campos seleccionados en el Visor de clústeres se descartan posteriormente, aunque puede cambiar este comportamiento editando el nodo Filtrar antes de su ejecución.
- **Nodo Seleccionar** Crea un nuevo nodo Seleccionar para seleccionar registros en función de su pertenencia a cualquiera de los clústeres visibles en la visualización actual del Visor de clústeres. Se genera de manera automática una condición de selección.
- **Nodo Seleccionar (desde la selección)** Crea un nuevo nodo Seleccionar para seleccionar registros en función de la pertenencia a clústeres seleccionados en el Visor de clústeres. Seleccione varios clústeres manteniendo pulsada la tecla Ctrl mientras selecciona los clústeres.
- **Nodo Derivar** Crea un nuevo nodo Derivar, que deriva un campo de marca que asigna a los registros un valor de *Verdadero* o *Falso* en función de su pertenencia a todos los clústeres visibles en el Visor de clústeres. Se genera de manera automática una condición de derivación.
- **Nodo Derivar (desde la selección)** Crea un nuevo nodo Derivar que deriva un campo de marca en función de la pertenencia a clústeres seleccionados en el Visor de clústeres. Seleccione varios clústeres manteniendo pulsada la tecla Ctrl mientras selecciona los clústeres.

Además de generar nodos, también puede crear gráficos desde el menú Generar. Consulte el tema "Generación de gráficos desde los modelos de clúster" en la página 265 para obtener más información.

Control de la visualización de clústeres

Para controlar qué se muestra en la vista Clústeres del panel principal, pulse el botón **Mostrar** y se abrirá el cuadro de diálogo Mostrar.

**Características.** Está seleccionado de forma predeterminada. Para ocultar todas las características de entrada, cancele la selección de la casilla de verificación.

**Campos de evaluación** Seleccione los campos de evaluación (campos que no se usan para crear el modelo de clúster, sino que se envían al visor de modelos para evaluar los clústeres) que desea mostrar, ya que ninguno se muestra de forma predeterminada. *Nota* El campo de evaluación debe ser una cadena con más de un valor. Esta casilla de verificación no está disponible si no hay ningún campo de evaluación disponible.

**Descripciones de clústeres** Está seleccionado de forma predeterminada. Para ocultar todas las casillas de descripción de clúster, cancele la selección de la casilla de verificación.

**Tamaños de clúster** Está seleccionado de forma predeterminada. Para ocultar todas las casillas de tamaño de clúster, cancele la selección de la casilla de verificación.

**Número máximo de categorías** Especifique el número máximo de categorías que se mostrarán en gráficos de características categóricas. El valor predeterminado es 20.

## Generación de gráficos desde los modelos de clúster

Los modelos de clúster ofrecen mucha información, aunque no siempre en un formato accesible para los usuarios empresariales. Puede producir gráficos de datos seleccionados para ofrecerlos de una forma que puedan incorporarse fácilmente en informes comerciales, presentaciones, etc. Por ejemplo, en el Visor de clústeres puede generar un gráfico para un clúster seleccionado, creando así un gráfico únicamente para los casos de ese clúster.

*Nota*: Sólo puede generar un gráfico desde el Visor de clústeres si el nugget de modelo está vinculado a otros nodos de una transmisión.

Generación de un gráfico

1. Abra el nugget de modelo que contiene el Visor de clústeres.
2. En la pestaña Modelo, seleccione *Clústeres* en la lista desplegable **Vista**.
3. En la vista principal, seleccione el clúster o clústeres de los que desea crear un gráfico.
4. En el menú Generar, seleccione **Gráfico (desde selección)** y se mostrará la pestaña Tablero básico.

   *Nota*: cuando abre la pestaña Tablero de esta forma, las únicas pestañas disponibles son Básico y Detallado.
5. Si utiliza la configuración de la pestaña Básico o Detallado, especifique los detalles que se mostrarán en el gráfico.
6. Pulse en Aceptar para generar el gráfico.

La cabecera del gráfico identifica el tipo de modelo y el clúster o clústeres que se seleccionaron para su inclusión.

# Capítulo 12. Reglas de asociación

Las **reglas de asociación** relacionan una determinada conclusión (por ejemplo, la compra de un producto dado) con un conjunto de condiciones (por ejemplo, la compra de otros productos). Por ejemplo, la regla

```
cerveza <= lata_veg & congelados (173, 17,0%, 0,84)
```

indica que, a menudo, se da el caso de *cerveza* cuando *lata_veg* y *congelados* ocurren al mismo tiempo. La regla es fiable en un 84 % y se aplica al 17 % de los datos (o 173 registros). Los algoritmos de reglas de asociación buscan automáticamente las asociaciones que se podrían encontrar manualmente usando técnicas de visualización, como en el nodo Malla .

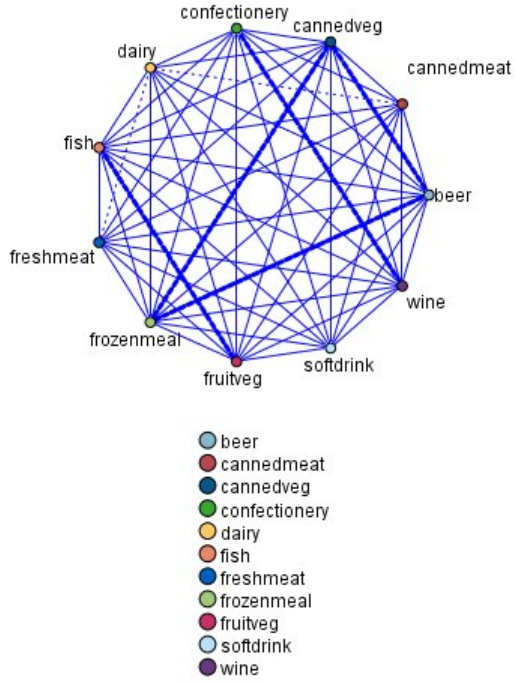



*Figura 45. Nodo Malla mostrando asociaciones entre elementos de la cesta de la compra*

La ventaja de los algoritmos de reglas de asociación sobre los algoritmos más estándar de árboles de decisión (C5.0 y Árbol C&R) es que las asociaciones pueden existir entre *cualquiera* de los atributos. Un algoritmo de árbol de decisión generará reglas con una única conclusión, mientras que los algoritmos de asociación tratan de buscar muchas reglas, cada una de las cuales puede tener una conclusión diferente.

La desventaja de los algoritmos de asociación es que tratan de encontrar patrones en un espacio de búsqueda potencialmente muy amplio y, por tanto, pueden necesitar mucho más tiempo de ejecución que un algoritmo de árbol de decisión. Los algoritmos usan un método de **generación y comprobación** para buscar reglas: se generan inicialmente reglas sencillas que se validan basándose en el conjunto de datos. Las buenas reglas se almacenan y todas las reglas, sujetas a varias restricciones, se especializan posteriormente. **La especialización** es el proceso de añadir condiciones a una regla. Estas nuevas reglas se validan basándose en los datos y el proceso almacena de forma iterativa las mejores reglas o las más interesantes que se encuentren. El usuario proporciona generalmente alguna limitación al número posible de antecedentes que permitir en una regla, y se usan diversas técnicas basadas en la teoría de la información o esquemas de indización eficientes para reducir el potencialmente amplio espacio de la búsqueda.

al final del procesamiento se presenta una tabla con las mejores reglas. A diferencia de un árbol de decisión, este conjunto de reglas de asociación no se puede usar directamente para realizar predicciones de mismo modo que puede hacerlo un modelo estándar (como un árbol de decisión o una red neuronal). Esto se debe a las diversas conclusiones diferentes posibles de las reglas. Otro nivel de transformación es preciso para transformar las reglas de asociación en un conjunto de reglas de clasificación. Por tanto, las reglas de asociación producidas por algoritmos de asociación se conocen como **modelos sin refinar**. Aunque el usuario puede examinar estos modelos sin definir, éstos no se pueden usar explícitamente como modelos de clasificación a menos que el usuario indique al sistema que genere un modelo de clasificación a partir del modelo sin definir. Este se lleva a cabo desde el explorador a través de una opción del menú Generar.

Se admiten dos algoritmos de reglas de asociación:

El nodo Apriori extrae un conjunto de reglas de los datos y destaca aquellas reglas con un mayor contenido de información. Apriori ofrece cinco métodos diferentes para la selección de reglas y utiliza un sofisticado esquema de indización para procesar eficientemente grandes conjuntos de datos. En los problemas de mucho volumen, Apriori se entrena más rápidamente, no tiene un límite arbitrario para el número de reglas que puede retener y puede gestionar reglas que tengan hasta 32 precondiciones. Apriori requiere que todos los campos de entrada y salida sean categóricos, pero ofrece un mejor rendimiento ya que está optimizado para este tipo de datos.

El nodo Secuencia encuentra reglas de asociación en datos secuenciales o en datos ordenados en el tiempo. Una secuencia es una lista de conjuntos de elementos que tiende a producirse en un orden previsible. Por ejemplo, si un cliente compra una cuchilla y una loción para después del afeitado, probablemente comprará crema para afeitar la próxima vez que vaya a comprar. El nodo Secuencia se basa en el algoritmo de reglas de asociación de CARMA, que utiliza un método de dos pasos para encontrar las secuencias.

## Datos tabulares frente a datos transaccionales

Los datos utilizados por modelos de reglas de asociación pueden estar en formato tabular o transaccional, como se describe a continuación. Lo siguiente son sólo descripciones generales, los requisitos específicos pueden variar como se discute en la documentación para cada tipo de modelo. Tenga en cuenta que al puntuar modelos, los datos que se van a puntuar deben reflejar el formato de los datos utilizados para generar el modelo. Los modelos generados utilizando datos tabulares se pueden utilizar para puntuar sólo datos tabulares; los modelos generados utilizando datos transaccionales sólo pueden puntuar datos transaccionales.

### Formato transaccional

Los datos transaccionales tienen un registro diferente para cada transacción o elemento. Si un cliente realiza varias compras, por ejemplo, cada una sería un registro diferente, con elementos asociados vinculados por un ID de cliente. Esto a veces se conoce como formato **anidado**.

| Cliente | Compra |
|---|---|
| 1 | mermelada |
| 2 | leche |
| 3 | mermelada |
| 3 | pan |
| 4 | mermelada |
| 4 | pan |
| 4 | leche |

Los nodos Apriori, CARMA y Secuencia pueden todos utilizar datos transaccionales.

### Datos tabulares

Los datos tabulares (también conocidos como datos **de la cesta** o **de la tabla de verdad**) tienen elementos representados por marcadores diferentes, donde cada campo de marcas representa la presencia o ausencia de un elemento específico. Cada registro representa un conjunto completo de elementos asociados. Los campos de marcas pueden ser categóricos o numéricos, aunque ciertos modelos pueden tener requisitos más específicos.

| Cliente | Mermelada | Pan | Leche |
|---|---|---|---|
| 1 | T | F | F |
| 2 | F | F | T |
| 3 | T | T | F |
| 4 | T | T | T |

Apriori, CARMA, Tanto GSAR como los nodos Secuencia pueden utilizar datos tabulares.

## Nodo Apriori

El nodo Apriori también encuentra reglas de asociación en los datos. Apriori ofrece cinco métodos diferentes para la selección de reglas y utiliza un sofisticado esquema de indización para procesar grandes conjuntos de datos de forma eficaz.

**Requisitos.** Para crear un conjunto de reglas de Apriori, se precisan uno o varios campos de *Entrada* y uno o varios campos de *Objetivo*. Los campos de entrada y de salida (con el rol *Entrada*, *Objetivo* o *Ambos*) deben ser simbólicos. Los campos con el rol *Ninguno* se omiten. Los tipos de campo deben estar completamente instanciados antes de ejecutar el nodo. Los datos pueden estar en formato tabular o transaccional. Consulte el tema “Datos tabulares frente a datos transaccionales” en la página 268 para obtener más información.

**Puntos fuertes.** En los problemas de grandes dimensiones, Apriori se entrena más rápidamente. Tampoco tiene un límite arbitrario para el número de reglas que puede retenerse y puede gestionar reglas que tengan hasta 32 precondiciones. Apriori ofrece cinco métodos de entrenamiento distintos, lo que permite una mayor flexibilidad para asociar el método de minería de datos con el problema en cuestión.

## Opciones de modelo para el nodo Apriori

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Soporte mínimo de las reglas.** Se puede especificar un criterio de soporte para mantener las reglas en el conjunto de reglas. **Soporte** hace referencia al porcentaje de registros de los datos de entrenamiento en los que los antecedentes (la parte de la regla "si") son verdaderos. (Observe que esta definición de soporte es diferente a la que se utiliza en los nodos CARMA y Secuencia. Consulte el tema “Opciones de modelo para el nodo Secuencia” en la página 285 para obtener más información. ) Si las reglas obtenidas se aplican a subconjuntos de datos muy pequeños, pruebe a aumentar el valor de este parámetro.

**Nota:** la definición de soporte para Apriori se basa en el número de registros con los antecedentes. Sucede de forma contraria que en los algoritmos CARMA y Secuencia, en los que la definición de soporte se basa en el número de registros con todos los elementos de una regla (es decir, los antecedentes y consecuentes). Los resultados de los modelos de asociación muestran tanto el soporte (antecedente) como las medidas de soporte de reglas.

**Confianza mínima de las reglas.** También se puede especificar un criterio de confianza. **La confianza** se basa en los registros por los que los antecedentes de la regla son verdaderos y es el porcentaje de esos mismos registros en los que los consecuentes también son verdaderos. Es decir, es el porcentaje de predicciones basadas en la regla que son correctas. Las reglas con una confianza inferior a la especificada en el criterio de precisión se descartan. Si se obtienen demasiadas reglas, pruebe a aumentar el valor de este parámetro. Si se obtienen muy pocas reglas (o casi ninguna), pruebe a disminuir el valor de este parámetro.

**Nota:** si es necesario, puede resaltar el valor y el tipo en su propio valor. Tenga en cuenta que, si reduce el valor de confianza por debajo de 1,0, además de que el proceso requiere gran cantidad de memoria libre, puede que las reglas tarden un tiempo extremadamente largo en generarse.

**Número máximo de antecedentes.** Se puede especificar el número máximo de precondiciones de cualquier regla. Se trata de una forma de limitar la complejidad de las reglas. Si las reglas son demasiado complejas o específicas, pruebe a disminuir el valor de este parámetro. Esta configuración también tiene mucha influencia en el tiempo de entrenamiento. Si el entrenamiento del conjunto de reglas que ha creado se toma demasiado tiempo, pruebe a disminuir el valor de este parámetro.

**Sólo valores verdaderos para las marcas.** Si selecciona esta opción para los datos en formato tabular (tabla de verdad), sólo se incluirán los valores verdaderos en las reglas resultantes. Esto puede ayudar a que las reglas se entiendan con más facilidad. La opción no se aplica a los datos en formato transaccional. Consulte el tema “Datos tabulares frente a datos transaccionales” en la página 268 para obtener más información.

**Nota:** El nodo de generación de modelos de CARMA ignora los registros vacíos al generar un modelo si el tipo de campo es una marca, mientras que el nodo de generación de modelos Apriori incluye los registros vacíos. Los registros vacíos son registros en los que todos los campos utilizados en la generación de modelo tienen un valor de false.

**Optimizar.** Seleccione opciones diseñadas para aumentar el rendimiento durante la generación de modelos según sus necesidades específicas.

- Seleccione **Velocidad** para indicar al algoritmo que nunca debe recurrir al volcado en disco para mejorar el rendimiento.
- Seleccione **Memoria** para indicar al algoritmo que utilice el volcado en disco cuando lo considere oportuno en detrimento de la velocidad. Esta opción está seleccionada de forma predeterminada.

  **Nota:** cuando se ejecuta en modo distribuido, esta configuración puede quedar anulada por las opciones del administrador especificadas en el archivo *options.cfg* Consulte la *Guía del administrador de IBM SPSS Modeler Server* para obtener más información.

## Opciones de experto para el nodo Apriori

Las opciones de experto siguientes permiten ajustar el proceso de inducción a los usuarios con conocimientos sobre redes neuronales. Para acceder a estas opciones, active el modo **Experto** en la pestaña Experto.

**Medida de evaluación.** Apriori admite cinco métodos de evaluación de reglas potenciales.

- **Confianza de la regla.** El método predeterminado utiliza la confianza (o precisión) de la regla para evaluar reglas. Para esta medida, se desactiva el **Límite inferior de la medida de evaluación**, ya que es redundante con la opción **Confianza mínima de las reglas** de la pestaña Modelo. Consulte el tema “Opciones de modelo para el nodo Apriori” en la página 269 para obtener más información.
- **Diferencia de confianza.** (También denominada **diferencia de confianza mínima absoluta con la previa**.) Esta medida de evaluación es la diferencia absoluta entre la confianza de la regla y su confianza apriori. Esta opción evita sesgos cuando los resultados no se distribuyen uniformemente. Ayuda a evitar que se conserven reglas "obvias". Por ejemplo, podría darse el caso de que el 80% de

los clientes comprasen su producto más popular. Una regla que predice la venta de ese producto tan popular con un 85% de precisión no aporta demasiados datos, a pesar de que una precisión del 85% es un porcentaje bastante alto en una escala absoluta. Establezca el límite inferior de la medida de evaluación en relación a la diferencia mínima de confianza o probabilidad con la que desea que se conserven las reglas.

- **Cociente de confianza.** (También denominada **diferencia de cociente de confianza establecida en 1**.) Esta medida de evaluación es igual a 1 menos el cociente de la confianza de la regla con respecto a la anterior (o si el cociente es superior a uno, su inverso). Al igual que la Diferencia de confianza, este método tiene en cuenta las distribuciones que no son homogéneas. Es especialmente apropiado para encontrar reglas que predicen eventos raros. Por ejemplo, supongamos que hay una enfermedad que sólo se da en el 1% de los pacientes. Una regla que puede predecir esta enfermedad un 10% de las veces constituye un gran avance respecto al pronóstico al azar, a pesar de que en una escala absoluta un 10% de precisión no destaque demasiado. Establezca el límite inferior de la medida de evaluación en función de la diferencia con la que desea que se conserven las reglas.
- **Diferencia de información.** (También denominada **diferencia de información respecto a la previa**.) Esta medida se basa en la medida de la **ganancia de información**. Si se considera la probabilidad de un consecuente determinado como un valor lógico (un **bit**), la ganancia de información es la proporción que puede determinarse de ese bit en función de los antecedentes. La diferencia de información es la diferencia existente entre la ganancia de información, dados los antecedentes, y la ganancia de información, dada sólo la confianza previa del consecuente. Una característica importante de este método es que tiene en cuenta el soporte de forma que son preferibles aquellas reglas que cubren más registros para un nivel de confianza determinado. Establezca el límite inferior de la medida de evaluación en función de la diferencia de información con la que desea que se conserven las reglas.

  *Nota*: Puesto que la escala para esta medida es algo menos intuitiva que las demás escalas, es posible que necesite probar con límites inferiores diferentes para obtener un conjunto de reglas satisfactorio.
- **Chi-cuadrado normalizada.** (También denominada **medida de chi-cuadrado normalizada**.) Esta medida es un índice estadístico de asociación entre antecedentes y consecuentes. La medida se normaliza para adquirir valores entre 0 y 1 y depende aún más del soporte que la medida de la diferencia de información. Establezca el límite inferior de la medida de evaluación en función de la diferencia de información con la que desea que se conserven las reglas.

  *Nota*: Al igual que sucede con la medida de la diferencia de información, la escala de esta medida es algo menos intuitiva que otras escalas, por lo que puede ser necesario experimentar con distintos límites inferiores para obtener un conjunto de reglas satisfactorio.

**Permitir reglas sin antecedentes.** Seleccione esta opción para permitir las reglas que sólo incluyen el consecuente (elemento o conjunto de elementos). Esto resulta de utilidad para determinar elementos o conjuntos de elementos comunes. Por ejemplo, `cannedveg` es una regla compuesta por un único elemento que carece de antecedentes e indica que la adquisición de *cannedveg* es una instancia común en los datos. En algunos casos, se pueden incluir estas reglas si sólo le interesan las predicciones de mayor confianza. Esta opción esta desactivada de forma predeterminada. Por convención, el soporte de antecedentes para las reglas que carecen de antecedentes se expresa con el 100% y el soporte de reglas es el mismo que la confianza.

## Nodo CARMA

El nodo CARMA utiliza un algoritmo de descubrimiento de reglas de asociación para encontrar reglas de asociación existentes en los datos. Las reglas de asociación son instrucciones del tipo

**if** *antecedente(s)* **then** *consecuente(s)*

Por ejemplo, si un cliente del sitio Web adquiere una tarjeta y un enrutador de gama alta inalámbricos, es muy probable que también adquiera un servidor de música inalámbrico si se le ofrece. El modelo CARMA extrae un conjunto de reglas de los datos sin necesidad de especificar campos de entrada ni de objetivo. Esto significa que las reglas generadas se pueden utilizar en una variedad de aplicaciones mucho más amplia. Por ejemplo, las reglas que ha generado este nodo se pueden utilizar para buscar una

lista de productos o servicios (antecedentes) cuyo consecuente es el elemento que desea promocionar durante esta temporada de vacaciones. Con IBM SPSS Modeler, puede determinar los clientes que han adquirido los productos antecedentes y diseñar una campaña de marketing destinada a la promoción del producto consecuente.

**Requisitos.** A diferencia de Apriori, el nodo CARMA no requiere que los campos sean de *Entrada* o de *Objetivo*. Esto es esencial para el modo en que funciona el algoritmo y equivale a la generación de un modelo de Apriori con todos los campos establecidos en *Ambas*. Se pueden restringir los elementos que aparecen sólo como antecedentes o como consecuentes activando el filtrado del modelo una vez generado éste. Por ejemplo, se puede utilizar el explorador de modelos para buscar una lista de productos o servicios (antecedentes) cuyo consecuente es el elemento que se desea promocionar durante esta temporada de vacaciones.

Para crear un conjunto de reglas de CARMA, es necesario especificar un campo de ID y uno o varios campos de contenido. El campo de ID puede tener cualquier rol o nivel de medición. Los campos con el rol *Ninguno* se omiten. Los tipos de campo deben estar completamente instanciados antes de ejecutar el nodo. Al igual que en Apriori, los datos pueden estar en formato tabular o transaccional. Consulte el tema "Datos tabulares frente a datos transaccionales" en la página 268 para obtener más información.

**Puntos fuertes.** El nodo CARMA se basa en el algoritmo de reglas de asociación de CARMA. A diferencia de Apriori, el nodo CARMA ofrece opciones de construcción basadas en el soporte de la regla (soporte tanto para el antecedente como el consecuente) en lugar de hacerlo sólo respecto al soporte del antecedente. CARMA también permite reglas con varios consecuentes. Como sucede con Apriori, los modelos que genera un nodo CARMA se pueden insertar en una ruta de datos para crear predicciones. Consulte el tema "Nuggets de modelo" en la página 38 para obtener más información.

## Opciones de campos para el nodo CARMA

Antes de ejecutar un nodo CARMA se deben especificar los campos de entrada en la pestaña Campos del nodo CARMA. Mientras que la mayoría de los nodos de Modelado comparten las mismas opciones de la pestaña Campos, el nodo CARMA contiene muchas opciones particulares. A continuación se describen todas las opciones.

**Utilizar configuración del nodo Tipo.** Esta opción permite indicar al nodo que use la información de campo de un nodo Tipo situado en un punto anterior de la ruta. Este es el método predeterminado.

**Utilizar configuración personalizada.** Esta opción permite indicar al nodo que use la información de campo especificada aquí en lugar de la proporcionada en nodos Tipo situados en cualquier punto anterior de la ruta. Una vez seleccionada esta opción, especifique los campos en función del formato (transaccional o tabular) en el que desee leer los datos.

**Utilizar formato transaccional.** Esta opción modifica los controles de campo del resto de este cuadro de diálogo en función de que el formato de los datos sea transaccional o tabular. Si se utilizan varios campos con datos transaccionales, se asume que los elementos especificados en estos campos para un registro determinado representan los elementos encontrados en una sola transacción con una sola marca de tiempo. Consulte el tema "Datos tabulares frente a datos transaccionales" en la página 268 para obtener más información.

Datos tabulares

Si no se selecciona **Utilizar formato transaccional**, se muestran los siguientes campos.

- **Entradas.** Seleccione el campo(s) de entrada. Se trata de una acción similar a establecer el rol del campo en *Entrada* en un nodo Tipo.
- **Partición.** Este campo permite especificar un campo usado para dividir los datos en muestras independientes para las fases de entrenamiento, prueba y validación en la generación del modelo. Si usa una muestra para generar el modelo y otra muestra distinta para comprobarlo, podrá obtener una

buena indicación de la bondad del modelo a la hora de generar conjuntos de datos de mayor tamaño similares a los datos actuales. Si se han definido varios campos de partición mediante nodos Tipo o Partición, se deberá seleccionar un campo de partición simple en la pestaña Campos en todos los nodos de modelado que usen la partición. (Si solamente hay una partición, se usará automáticamente siempre que se active la partición.) Debe tener en cuenta que al aplicar la partición seleccionada en su análisis, también debe activar la partición en la pestaña Opciones del modelo para el nodo. (Si se elimina la selección de esta opción, se posibilita la desactivación de la partición sin cambiar la configuración del campo.)

Datos transaccionales

Si selecciona **Utilizar formato transaccional**, se muestran los siguientes campos.

- **ID.** Para los datos transaccionales, seleccione el campo de ID de la lista. Los campos numéricos o simbólicos se pueden utilizar como campo de ID. Cada valor exclusivo de este campo debe indicar una unidad de análisis específica. Por ejemplo, en una aplicación de la cesta de la compra, cada ID puede representar a un sólo cliente. Para una aplicación de análisis del registro Web, cada ID puede representar un equipo (con la dirección IP) o un usuario (con los datos de inicio de sesión).
- **Los ID son contiguos.** (Nodos Apriori y CARMA únicamente) Si los datos se han clasificado previamente de forma que todos los registros con el mismo ID se agrupan en la ruta de datos, seleccione esta opción para que el procesamiento sea más rápido. Si los datos no se han clasificado previamente (o no lo sabe a ciencia cierta), no active esta opción y el nodo clasificará los datos automáticamente.

  *Nota*: si los datos no están clasificados y selecciona esta opción, es posible que obtenga resultados no válidos en el modelo.
- **Contenido.** Especifique los campos de contenido del modelo. Estos campos contienen los elementos de interés del modelo de asociación. Se pueden especificar varios campos de marcas (si los datos están en formato tabular) o un sólo campo nominal (si los datos están en formato transaccional).

## Opciones de modelo para el nodo CARMA

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Soporte mínimo de las reglas (%).** Puede especificar un criterio de soporte. **Soporte de la regla** hace referencia a la proporción de campos de ID existente en los datos de entrenamiento que contienen la regla completa. (Tenga en cuenta que esta definición de soporte es diferente al soporte del antecedente utilizado en los nodos Apriori.) Si desea centrarse en las reglas más comunes, aumente el valor de este parámetro.

**Confianza mínima de las reglas (%).** Se puede especificar un criterio de confianza para mantener las reglas en el conjunto de reglas. **La confianza** hace referencia al porcentaje de campos de ID en los que se realiza una predicción correcta (de todos los campos de ID para los que la regla realiza una predicción). Se calcula como la cantidad de ID en los que se encuentra la regla completa dividido por la cantidad de ID en los que se encuentran los antecedentes, basado en los datos de entrenamiento. Las reglas con una confianza inferior a la especificada en el criterio de precisión se descartan. Si se obtienen demasiadas reglas o reglas de poco interés, pruebe a aumentar el valor de este parámetro. Si se obtienen muy pocas reglas, pruebe a disminuir el valor de este parámetro.

**Nota:** si es necesario, puede resaltar el valor y el tipo en su propio valor. Tenga en cuenta que, si reduce el valor de confianza por debajo de 1,0, además de que el proceso requiere gran cantidad de memoria libre, puede que las reglas tarden un tiempo extremadamente largo en generarse.

**Tamaño máximo de regla.** Se puede configurar el número máximo de `conjuntos de elementos` (a diferencia de los `elementos`) distintos en una regla. Si las reglas de interés resultantes son pocas, se puede disminuir el valor del parámetro para que el conjunto de reglas se genere más rápido.

**Nota:** El nodo de generación de modelos de CARMA ignora los registros vacíos al generar un modelo si el tipo de campo es una marca, mientras que el nodo de generación de modelos Apriori incluye los registros vacíos. Los registros vacíos son registros en los que todos los campos utilizados en la generación de modelo tienen un valor de false.

## Opciones de experto para el nodo CARMA

Las opciones de experto siguientes permiten ajustar el proceso de generación de modelos a los usuarios con conocimientos sobre el funcionamiento del nodo CARMA. Para acceder a estas opciones, active el modo **Experto** en la pestaña Experto.

**Excluir reglas con varios consecuentes.** Seleccione excluir los consecuentes de "dos direcciones", es decir los consecuentes que contienen dos elementos. Por ejemplo, la regla `pan y queso y pescado -> vino y fruta` contiene un consecuente de dos direcciones `vino y fruta`. Estas reglas se incluyen de forma predeterminada.

**Definir valor de poda.** Para conservar la memoria, el algoritmo CARMA utilizado periódicamente elimina (**poda**) los conjuntos de elementos poco frecuentes de una lista de conjuntos de elementos potenciales durante el procesamiento. Seleccione esta opción para ajustar la frecuencia de poda; el número especificado determina la frecuencia de la misma. Introduzca un valor más pequeño para disminuir los requisitos de memoria del algoritmo (pero aumentar potencialmente el tiempo de entrenamiento necesario) o introduzca un valor mayor para que el entrenamiento sea más rápido (pero aumentar potencialmente los requisitos de memoria). El valor por omisión es 500.

**Variar soporte.** Seleccione esta opción para aumentar la eficacia mediante la exclusión de conjuntos de elementos poco frecuentes que aparentan ser frecuentes cuando se incluyen de forma irregular. Esto se consigue comenzando con un nivel de soporte superior que después se disminuye hasta el nivel especificado en la pestaña Modelo. Especifique un valor en **Número de transacciones estimado** para especificar la velocidad con la que debe disminuirse el nivel de soporte.

**Permitir reglas sin antecedentes.** Seleccione esta opción para permitir las reglas que sólo incluyen el consecuente (elemento o conjunto de elementos). Esto resulta de utilidad para determinar elementos o conjuntos de elementos comunes. Por ejemplo, `cannedveg` es una regla compuesta por un único elemento que carece de antecedentes e indica que la adquisición de *cannedveg* es una instancia común en los datos. En algunos casos, se pueden incluir estas reglas si sólo le interesan las predicciones de mayor confianza. Esta opción esta desactivada de forma predeterminada.

## Nugget del modelo de reglas de asociación

Los nugget de modelo de reglas de asociación representan las reglas descubiertas por uno de los siguientes nodos de modelado de reglas de asociación:

- A priori
- CARMA

Los nugget de modelo contienen información acerca de las reglas extraídas a partir de los datos durante la generación de modelos.

**Nota:** La puntuación de nugget de regla de asociación podría ser incorrecta si no clasifica los datos transaccionales por ID.

Visualización de los resultados

Para examinar las reglas generadas por los modelos de asociación (Apriori y CARMA) y modelos de secuencias, se puede utilizar la pestaña Modelo del cuadro de diálogo. Al examinar un nugget de modelo, se muestra la información acerca de las reglas y se ofrecen opciones para filtrar y ordenar los resultados antes de generar nuevos nodos o puntuar el modelo.

Puntuación de modelos

Los nugget de modelo refinado (Apriori, CARMA y Secuencia) se pueden añadir a una ruta y utilizarse para puntuar. Consulte el tema “Uso de nugget de modelo en rutas” en la página 49 para obtener más información. Los modelos de nuggets utilizados para puntuar incluyen una pestaña Configuración adicional en los cuadros de diálogo respectivos. Consulte el tema “Configuración del nugget de modelo de reglas de asociación” en la página 278 para obtener más información.

Un nugget de modelo sin refinar no se puede utilizar para puntuar en su formato sin procesar. En su lugar, se puede generar un conjunto de reglas y utilizar dicho conjunto para puntuar. Consulte el tema “Generación de un conjunto de reglas desde un nugget de modelo de asociación” en la página 280 para obtener más información.

## Detalles del nugget de modelo de reglas de asociación

En la pestaña Modelo de un nugget de modelo Regla de asociación se incluye una tabla con las reglas que ha extraído el algoritmo. Cada fila de la tabla representa una regla. La primera columna representa los consecuentes (la parte "entonces" de una regla), mientras que la siguiente columna representa los antecedentes (la parte "si" de la regla). Las siguientes columnas contienen información de las reglas, como la confianza, el soporte y la elevación.

Las reglas de asociación a menudo se muestran en el formato de la tabla siguiente.

*Tabla 15. Ejemplo de norma de asociación*

| Consecuente | Antecedente |
|---|---|
| `Drug = medicamentoA` | `Sexo = F`<br>`PS = ALTA` |

La regla de ejemplo se interpreta como *si Sexo = "F" y PS = "ALTA", entonces Medicamento probablemente sea drugY*; o, dicho de otro modo, *para los registros en los que Sexo = "F" y PS = "ALTA", es muy probable que Medicamento sea drugY*. Mediante la barra de herramientas del cuadro de diálogo, puede seleccionar presentar información adicional, como la confianza, el soporte y las instancias.

**Menú Ordenar.** El botón del menú Ordenar en la barra de herramientas controla la ordenación de las reglas. Es posible cambiar la dirección de ordenación (ascendente o descendente) mediante el botón de dirección de la ordenación (flecha arriba y abajo).

Los valores se pueden ordenar por:

- Asistencia
- Confianza
- Soporte de regla
- Consecuente
- Lift
- Capacidad de despliegue

**Menú Mostrar/ocultar.** El menú Mostrar/ocultar (botón de criterios de la barra de herramientas) controla las opciones de visualización de las reglas.

*Figura 46. Botón Mostrar/ocultar*

Están disponibles las siguientes opciones de presentación:

- **ID de regla** muestra el identificador de regla asignado durante la generación del modelo. Un ID de regla permite identificar qué reglas se están aplicando para una determinada predicción. Los ID de regla también permiten, más adelante, fundir información adicional de las reglas, como capacidad de despliegue, información del producto o antecedentes.
- **Instancias** muestra información acerca del número de ID exclusivos a los que se aplica la regla; es decir, para los que se cumple la condición de los antecedentes. Por ejemplo, dada la regla `pan > queso`, se hace referencia al número de registros en los datos de entrenamiento que incluyan el antecedente *pan* como **instancias**.
- **Soporte** visualiza soporte de antecedentes; es decir, la proporción de ID para los que los antecedentes son verdaderos, basándose en los datos de entrenamiento. Por ejemplo, si el 50% de los datos de entrenamiento incluyen la compra de pan, entonces la regla `pan > queso` tendrá un soporte de antecedentes del 50%. *Nota*: el soporte, tal y como se define aquí, es igual que las instancias, pero se representa como un porcentaje.
- **Confianza** muestra el cociente entre el soporte de regla y el soporte de antecedentes. Esto indica la proporción de ID con el antecedente (o los antecedentes) especificado para el que el consecuente (o los consecuentes) es también verdadero. Por ejemplo, si el 50% de los datos de entrenamiento contiene pan (indicando así el soporte de antecedentes), pero sólo el 20% contiene pan y queso (lo que indica el soporte de regla), la confianza de la regla `pan > queso` sería `Soporte de regla / Soporte de antecedentes` o, en este caso, 40%.
- **Soporte de regla** muestra la proporción de ID para los que se cumple que toda la regla (los antecedentes y consecuentes) son verdaderos. Por ejemplo, si el 20% de los datos de entrenamiento incluyen tanto pan como queso, el soporte de la regla `pan > queso` será el 20%.
- **Elevación** muestra el cociente de confianza de la regla en la probabilidad previa de disponer del consecuente. Por ejemplo, si el 10% de toda la población compra pan, una regla que predice si la gente va a comprar pan con un 20% de confianza tendrá una elevación de 20/10 = 2. Si otra regla indica que la gente va a comprar pan con el 11% de confianza, la regla tiene una elevación cercana a 1, lo que significa que disponer de antecedente (o antecedentes) no supone una gran diferencia en el caso de disponer de consecuente. Por lo general, las reglas con una elevación distinta a 1 serán más interesantes que las reglas con elevación cercana a 1.
- **La capacidad de despliegue** mide qué porcentaje de los datos de entrenamiento satisface las condiciones del antecedente pero no satisface el consecuente. En términos de compra de producto, básicamente significa qué porcentaje de la base total de clientes posee (o ha comprado) el antecedente (o los antecedentes) pero no ha comprado aún el consecuente. El estadístico de capacidad de despliegue se define como `((Soporte de antecedentes en número de registros - Soporte de regla en número de registros) / Número de registros) * 100`, donde *Soporte de antecedentes* hace referencia al número de registros para los que los antecedentes son verdaderos y *Soporte de regla* hace referencia al número de registros para los que tanto los antecedentes como los consecuentes son verdaderos.

**Botón de filtrado.** En el menú, el botón de filtrado (icono del embudo) expande el botón del cuadro de diálogo para mostrar un panel en el que aparecen los filtros de regla activos. Los filtros se utilizan para contraer el número de reglas que se muestran en la pestaña Modelos.

*Figura 47. Botón de filtrado*

Para crear un filtro, pulse en el icono de filtrado, en la parte derecha del panel expandido. Esta operación abre un cuadro de diálogo independiente en el que se pueden especificar las restricciones a la hora de

mostrar reglas. Tenga en cuenta que el botón de filtrado se suele utilizar junto con el menú Generar para, en primer lugar, filtrar las reglas y, a continuación, generar un modelo que contenga ese subconjunto de reglas. Si desea obtener más información, consulte "Especificación de filtros para reglas" a continuación.

**Botón Buscar regla.** El botón Buscar regla (el icono de los prismáticos) permite examinar las reglas mostradas para un ID de regla especificado. El cuadro de diálogo adyacente indica el número de reglas que actualmente se muestran con respecto al número disponible. El modelo asigna los ID de reglas inmediatamente en el orden de descubrimiento y los añade a los datos durante la puntuación.

*Figura 48. Botón Buscar regla*

Para ordenar de nuevo los ID de la regla:

1. Es posible reorganizar los ID de regla en IBM SPSS Modeler ordenando, en primer lugar, la tabla de representación de reglas según la medición que desee, como la confianza o la elevación.
2. A continuación, puede crear un modelo filtrado mediante las opciones del menú Generar.
3. En el cuadro de diálogo Modelo filtrado, seleccione **Volver a numerar reglas consecutivamente, comenzando por** y especifique un número de inicio.

Consulte el tema "Generación de un modelo filtrado" en la página 280 para obtener más información.

## Especificación de filtros para reglas

De forma predeterminada, los algoritmos de regla, como Apriori, CARMA y Secuencia, pueden generar un número de reglas grande y engorroso. Para mejorar la claridad a la hora de examinar o para simplificar la puntuación de la regla, debería contemplar las reglas de filtrado para que las consecuencias y los antecedentes de interés destaquen más. El uso de las opciones de filtrado en la pestaña Modelo, situada en el explorador de reglas, permite abrir un cuadro de diálogo para especificar las calificaciones del filtro.

**Consecuentes.** Seleccione **Activar filtro** para activar las opciones de filtrado de reglas basadas en la inclusión o exclusión de consecuentes especificados. Seleccione **Incluir cualquiera de** para crear un filtro donde las reglas contienen al menos uno de los consecuentes especificados. También puede seleccionar **Excluye** para crear un filtro que excluya los consecuentes especificados. Puede seleccionar los consecuentes mediante el icono del selector en la parte derecha del cuadro de lista. Esta acción abre un cuadro de diálogo que enumera todos los consecuentes presentes en las reglas generadas.

*Nota*: los consecuentes pueden contener más de un elemento. Los filtros sólo comprobarán que un consecuente contenga uno de los elementos especificados.

**Antecedentes.** Seleccione **Activar filtro** para activar las opciones de filtrado de reglas basadas en la inclusión o exclusión de antecedentes especificados. Puede seleccionar elementos mediante el icono del selector en la parte derecha del cuadro de lista. Esta acción abre un cuadro de diálogo que enumera todos los antecedentes presentes en las reglas generadas.

- Seleccione **Incluir todos** para definir el filtro como de inclusión, de manera que todos los antecedentes especificados deban incluirse en una regla.
- Seleccione **Incluir cualquiera de** para crear un filtro donde las reglas contengan al menos uno de los antecedentes especificados.
- Seleccione **Excluye** para crear un filtro que excluya las reglas que contengan un antecedente especificado.

**Confianza.** Seleccione **Activar filtro** para activar las opciones de filtrado de reglas basadas en el nivel de confianza de una regla. Puede utilizar los controles **Mín.** y **Máx.** para especificar un intervalo de

confianza. Al examinar los modelos generados, la confianza se enumera como porcentaje. Al puntuar los resultados, la confianza se expresa como un número entre 0 y 1.

**Soporte de antecedentes.** Seleccione **Activar filtro** para activar las opciones de filtrado de reglas basadas en el nivel de soporte de antecedentes de una regla. El soporte de antecedentes indica la proporción de los datos de entrenamiento que contienen los mismos antecedentes que la regla actual, lo que lo convierte en análogo al índice de popularidad. Puede utilizar los controles **Mín.** y **Máx.** para especificar un intervalo utilizado para filtrar las reglas basadas en el nivel del soporte.

**Elevación.** Seleccione **Activar filtro** para activar las opciones de filtrado de reglas basadas en la medición de elevación de una regla. *Nota*: el filtrado de la elevación sólo está disponible para los modelos de asociación creados posteriormente a la versión 8.5 o para modelos anteriores que contengan una medición de elevación. Los modelos de secuencias no contienen esta opción.

Pulse en **Aceptar** para aplicar todos los filtros que se han activado en este cuadro de diálogo.

### Generación de gráficos para reglas

Los nodos Asociación proporcionan gran cantidad de información, sin embargo, es posible que no estén en un formato fácilmente accesible para usuarios empresariales. Puede producir gráficos de datos seleccionados para ofrecerlos de una forma que puedan incorporarse fácilmente en informes comerciales, presentaciones, etc. En la pestaña Modelo, puede crear un gráfico de la regla seleccionada, creando únicamente un gráfico para los casos de esa regla.

1. En la pestaña Modelo, selecciona la regla que le interese.
2. En el menú Generar, seleccione **Gráfico (desde selección)**. Aparecerá la pestaña Tablero básico.

   *Nota*: cuando abre la pestaña Tablero de esta forma, las únicas pestañas disponibles son Básico y Detallado.
3. Si utiliza la configuración de la pestaña Básico o Detallado, especifique los detalles que se mostrarán en el gráfico.
4. Pulse en Aceptar para generar el gráfico.

La cabecera del gráfico identifica la regla y los detalles de antecedentes seleccionados para incluir.

## Configuración del nugget de modelo de reglas de asociación

Esta pestaña Configuración se utiliza para especificar las opciones de puntuación de los modelos de asociación (Apriori y CARMA). Esta pestaña sólo estará disponible después de que el nugget de modelo se haya añadido a una ruta para la puntuación.

**Nota:** El cuadro de diálogo para examinar un modelo no refinado no incluye la pestaña Configuración, porque no se puede puntuar. Para puntuar el modelo "sin refinar", primero debe generar un conjunto de reglas. Consulte el tema “Generación de un conjunto de reglas desde un nugget de modelo de asociación” en la página 280 para obtener más información.

**Número máximo de predicciones** Especifique el número máximo de predicciones que se incluyen para cada conjunto de elementos de la cesta. Esta opción se utiliza conjuntamente con el criterio de regla siguiente para producir las previsiones “superiores”, donde *superiores* indica el nivel más alto de confianza, soporte, elevación, etc., como se especifica más abajo.

**Criterio de regla** Seleccione la medida utilizada para determinar la fuerza de las reglas. Las reglas se clasifican según la fuerza de los criterios aquí seleccionados a fin de devolver las mejores predicciones para un conjunto de elementos. Los criterios disponibles se muestran en la lista siguiente.

- Confianza
- Asistencia
- Soporte de regla (Soporte * Confianza)

- Lift
- Capacidad de despliegue

**Permitir predicciones repetidas** Selecciónelo para incluir varias reglas con el mismo consecuente a la hora de puntuar. Por ejemplo, si se selecciona esta opción se pueden puntuar las siguientes reglas:

```
pan y vino y queso
queso y fruta y vino
```

Desactive esta opción para excluir las predicciones repetidas a la hora de puntuar.

**Nota:** las reglas con varios consecuentes (`pan & queso & fruta -> vino & paté`) se consideran predicciones repetidas sólo si todos los consecuentes (`vino & paté`) se han predicho anteriormente.

**Ignorar elementos de cesta no coincidentes** Seleccione esta opción para omitir la presencia de elementos adicionales en el conjunto de elementos. Por ejemplo, al seleccionar esta opción para una cesta que contiene `[tienda de campaña & saco de dormir & tetera]`, se aplicará la regla `tienda de campaña & saco de dormir > cocina_gas`, a pesar del elemento adicional (`tetera`) incluido en la cesta.

Existen algunas circunstancias en las que los elementos deberían excluirse. Por ejemplo, es probable que alguien que compra una tienda de campaña, un saco de dormir y una tetera pueda disponer ya de una cocina de gas, hecho indicado por la presencia de la tetera. Dicho de otro modo, una cocina de gas puede no ser la mejor predicción. En estos casos, debería eliminar la selección **Ignorar elementos de cesta no coincidentes** para asegurarse de que los antecedentes de la regla coinciden exactamente con el contenido de una cesta. De forma predeterminada, se omiten los elementos no coincidentes.

**Comprobar que las predicciones no están en la cesta.** Seleccione esta opción para asegurarse de que los antecedentes no se encuentran en la cesta. Por ejemplo, si el propósito de la puntuación es recomendar mobiliario para el hogar, no es probable que una cesta que ya contiene una mesa de comedor recomiende comprar otra. En este caso, debería seleccionar esta opción. Por otra parte, si los productos son perecederos o desechables (como el queso, un biberón o un pañuelo), las reglas en las que el consecuente ya se encuentra en la cesta pueden ser de valor. En este último caso, la opción más útil puede ser **No comprobar las predicciones de la cesta**, situada más abajo.

**Comprobar que las predicciones están en la cesta** Seleccione esta opción para asegurarse de que los consecuentes también se encuentran en la cesta. Este método es útil cuando se intenta comprender mejor los clientes o las transacciones existentes. Por ejemplo, es posible que desee identificar las reglas con la mayor elevación y, a continuación, explorar qué clientes se ajustan a estas reglas.

**No comprobar las predicciones de la cesta** Seleccione esta opción para incluir todas las reglas a la hora de puntuar, independientemente de la presencia o ausencia de consecuentes en la cesta.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

## Resumen del nugget de modelo de reglas de asociación

La pestaña Resumen para un nugget de modelo de reglas de una asociación muestra el número de reglas descubiertas y el mínimo y máximo de soporte y capacidad de despliegue de las reglas en el conjunto de reglas.

## Generación de un conjunto de reglas desde un nugget de modelo de asociación

Los nugget de modelo de asociación, como Apriori y CARMA, pueden utilizarse para puntuar los datos directamente; sin embargo, también es posible generar en primer lugar un subconjunto de reglas conocido como **conjunto de reglas**. Los conjuntos de reglas son de gran utilidad si trabaja con un modelo , que no se puede utilizar directamente para la puntuación. Consulte el tema "Modelos sin refinar" en la página 52 para obtener más información.

Para generar un conjunto de reglas, seleccione **Conjunto de reglas** en el menú Generar, situado en el explorador de nugget de modelo. Es posible especificar las siguientes opciones para trasladar las reglas a un conjunto de reglas:

**Nombre de conjunto de reglas.** Permite especificar el nombre del nuevo nodo de conjunto de reglas generado.

**Crear nodo en.** Controla la ubicación del nuevo nodo de conjunto de reglas generado. Seleccione **Lienzo**, **Paleta de modelos generados** o **Ambas**.

**Campo objetivo.** Determina qué campo de salida se utilizará para el nodo de conjunto de reglas generado. Seleccione un único campo de salida de la lista.

**Soporte mínimo.** Especifique el soporte mínimo para las reglas que desea conservar en el conjunto de reglas generado. El nuevo conjunto de reglas no incluirá reglas con un soporte inferior al valor especificado.

**Confianza mínima.** Especifique la confianza mínima para las reglas que desea conservar en el conjunto de reglas generado. El nuevo conjunto de reglas no incluirá las reglas con un valor de confianza inferior al especificado.

**Valor predeterminado.** Permite especificar un valor predeterminado para el campo objetivo que se asigna a los registros puntuados para los que no se activa una regla.

## Generación de un modelo filtrado

Para generar un modelo filtrado desde un nugget de modelo de asociación, como un nodo Apriori, CARMA o de conjunto de reglas de secuencias, seleccione **Modelo filtrado** en el menú Generar del explorador de nugget de modelo. Esto crea un modelo de subconjuntos que incluye sólo las reglas actualmente mostradas en el explorador. *Nota*: No puede generar modelos filtrados para modelos no refinados.

Puede especificar las siguientes opciones para filtrar las reglas:

**Nombre para nuevo modelo.** Permite especificar el nombre del nuevo nodo del modelo filtrado.

**Crear nodo en.** Controla la ubicación del nuevo nodo del modelo filtrado. Seleccione **Lienzo**, **Paleta de modelos generados** o **Ambas**.

**Numeración de reglas.** Especifique cómo se numerarán los ID de reglas en el subconjunto de reglas contenido en el modelo filtrado.

- **Conservar números originales de ID de regla.** Seleccione esta opción para mantener la numeración original de las reglas. De forma predeterminada, se otorga un ID a las reglas que corresponde con el orden en que el algoritmo las haya descubierto. Ese orden puede variar dependiendo del algoritmo empleado.
- **Volver a numerar reglas consecutivamente, comenzando por.** Seleccione esta opción para asignar nuevos ID de reglas para las reglas filtradas. Se asignan los nuevos ID basados en el orden de clasificación mostrados en la tabla de exploración de reglas, situada en la pestaña Modelo y comenzando por el número que se especifica aquí. Utilice las flechas de la derecha para especificar el número de inicio para los ID.

## Reglas de asociación de la puntuación

Las puntuaciones obtenidas al ejecutar nuevos datos mediante un nugget de modelo de reglas de asociación se devuelven en campos independientes. Se añaden tres nuevos campos para cada predicción: *P* representa la predicción, *C* la confianza e *I* representa el ID. La organización de estos campos de salida depende de si los datos de entrada están en formato transaccional o tabular. Consulte "Datos tabulares frente a datos transaccionales" en la página 268 para obtener una visión general de estos formatos.

Por ejemplo, suponga que está puntuando datos de la cesta de la compra con un modelo que genera predicciones basadas en estas tres reglas:

```
Rule_15 pan, vino y carne (confianza 54%)
Rule_22 queso y fruta (confianza 43%)
Rule_15 pan, queso -> verdura (confianza 24%)
```

**Datos tabulares.** En el caso de los datos tabulares, se devuelven las tres predicciones (3 es el valor predeterminado) en un único registro.

*Tabla 16. Puntuaciones en formato tabular.*

| ID | Pan | Vino | Queso | P1 | C1 | I1 | P2 | C2 | I2 | P3 | C3 | I3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Fred | 1 | 1 | 1 | carne | 0,54 | 15 | fruta | 0,43 | 22 | verdura | 0,24 | 5 |

**Datos transaccionales.** En los datos transaccionales, se genera un registro independiente para cada predicción. Las predicciones se siguen añadiendo en columnas independientes, pero las puntuaciones se devuelven cuando se calculan. Así se producen registros con predicciones incompletas, como se muestra en el siguiente resultado de muestra. La segunda y la tercera predicción (P2 y P3) están en blanco en el primer registro, junto con las confianzas asociadas y los ID de regla. Sin embargo, a medida que se devuelven las puntuaciones, el registro final contiene las tres predicciones.

*Tabla 17. Puntuaciones en formato transaccional.*

| ID | Elem. | P1 | C1 | I1 | P2 | C2 | I2 | P3 | C3 | I3 |
|---|---|---|---|---|---|---|---|---|---|---|
| Fred | pan | carne | 0,54 | 14 | $null$ | $null$ | $null$ | $null$ | $null$ | $null$ |
| Fred | queso | carne | 0,54 | 14 | fruta | 0,43 | 22 | $null$ | $null$ | $null$ |
| Fred | vino | carne | 0,54 | 14 | fruta | 0,43 | 22 | verdura | 0,24 | 5 |

Para incluir sólo las predicciones completas con propósito de notificación o de despliegue, utilice un nodo Seleccionar para seleccionar los registros completos.

*Nota*: los nombres de campos utilizados en estos ejemplos están abreviados para mayor claridad. Durante el uso real, los campos de resultados para los modelos de asociación se denominan como se muestra en la tabla siguiente.

*Tabla 18. Nombres de campos de resultados para modelos de asociación.*

| Campo nuevo | Nombre del campo de ejemplo |
|---|---|
| Predicción | *$A-TRANSACCIÓN_NÚMERO-1* |
| Confianza (u otro criterio) | *$AC-TRANSACCIÓN_NÚMERO-1* |
| ID de regla | *$A-ID_de_regla-1* |

Reglas con varios consecuentes

El algoritmo CARMA permite reglas con varios consecuentes, por ejemplo:

```
pan > vino y queso
```

Al puntuar reglas de “dos direcciones”, se devuelven predicciones en el formato visualizado en la tabla siguiente.

*Tabla 19. Puntuación de resultados incluido una predicción con varios consecuentes.*

| ID | Pan | Vino | Queso | P1 | C1 | I1 | P2 | C2 | I2 | P3 | C3 | I3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Fred | 1 | 1 | 1 | carne y verd | 0,54 | 16 | fruta | 0,43 | 22 | verdura | 0,24 | 5 |

En ciertos casos, es posible que sea necesario dividir dichas puntuaciones antes de realizar el despliegue. Para dividir una predicción con varios consecuentes, deberá analizar el campo mediante las funciones de cadena de CLEM.

## Despliegue de modelos de asociación

Al puntuar modelos de asociación, las predicciones y las confianzas se muestran en columnas independientes (donde *P* representa la predicción, *C* representa la confianza y *I* representa el ID de regla). En este caso los datos de entrada pueden ser tabulares o transaccionales. Consulte el tema “Reglas de asociación de la puntuación” en la página 281 para obtener más información.

Al preparar puntuaciones para el despliegue, puede que la aplicación requiera transponer los datos de salida a un formato con predicciones en filas en lugar de columnas (una predicción por fila, que en ocasiones se conoce como formato "anidado").

Transposición de puntuaciones tabulares

Puede transponer puntuaciones tabulares de columnas a filas utilizando una combinación de pasos en IBM SPSS Modeler, como se describe en los siguientes pasos.

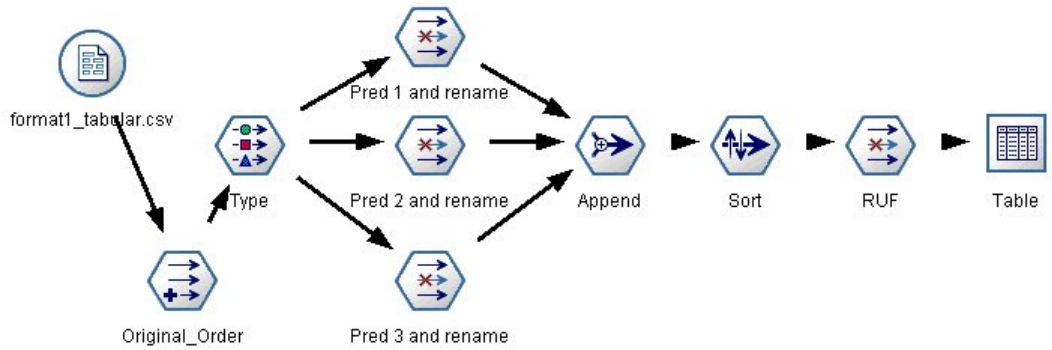



*Figura 49. La ruta de ejemplo utilizada para transponer los datos tabulares en formato anidado*

1. Utilice la función `@INDEX` de un nodo Derivar para comprobar el orden actual de las predicciones y guardar este indicador en un nuevo campo, como *Orden_original*.
2. Añada un nodo Tipo para asegurarse de que todos los campos están instanciados.
3. Utilice un nodo Filtrar para cambiar el nombre predeterminado de la predicción, la confianza y los campos de ID (*P1, C1, I1*) en campos comunes, como *Pred*, *Crit* e *ID_de_regla*, que se utilizará para añadir registros más tarde. Necesitará un nodo Filtrar para cada predicción generada.

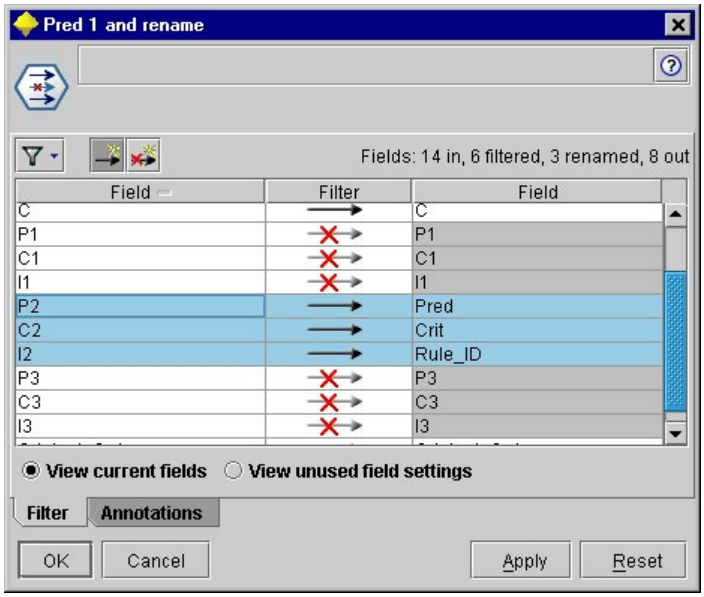



*Figura 50. Filtrado de campos para las predicciones 1 y 3 a la vez que se cambia el nombre de los campos para la predicción 2.*

4. Utilice un nodo Agregar para añadir valores para los datos compartidos *Pred*, *Crit* e *ID_de_regla*.
5. Conecte un nodo Ordenar para clasificar registros en orden ascendente para el campo *Orden_original* y en orden descendente para *Crit*, que es el campo utilizado para clasificar las predicciones por criterios como confianza, elevación y soporte.
6. Utilice otro nodo Filtrar para filtrar el campo *Orden_original* desde el resultado.

En este punto, los datos ya están listos para el despliegue.

Transposición de puntuaciones transaccionales

El proceso es similar para la transposición de puntuaciones transaccionales. Por ejemplo, la ruta que se muestra a continuación transpone puntuaciones a un formato con una única predicción en cada fila según sea necesario para el despliegue.

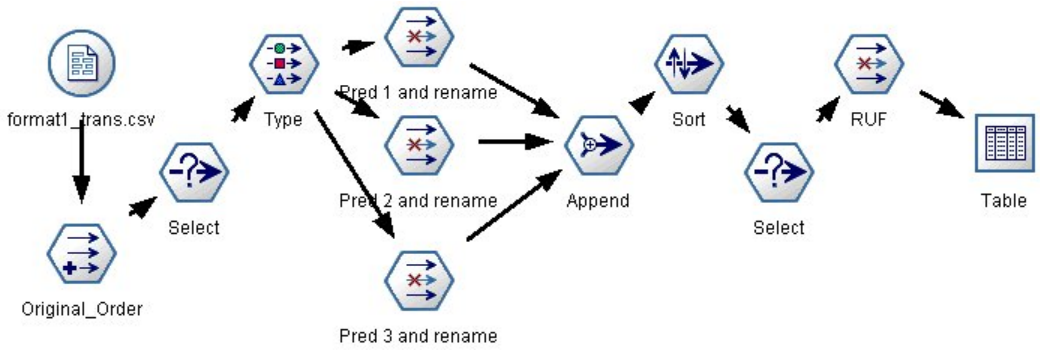



*Figura 51. La ruta de ejemplo utilizada para transponer los datos transaccionales en formato anidado*

Con la adición de dos nodos Seleccionar, el proceso es idéntico al detallado anteriormente para los datos tabulares.

- El primer nodo Seleccionar se utiliza para comparar los ID de regla de todos los registros adyacentes e incluye sólo registros exclusivos o no definidos. Este nodo Seleccionar utiliza la expresión CLEM para seleccionar registros: `ID /= @OFFSET(ID,-1) or @OFFSET(ID,-1) = undef`.
- El segundo nodo Seleccionar se utiliza para descartar reglas no pertinentes, o reglas donde `ID_de_regla` tiene un valor nulo. Este nodo Seleccionar utiliza la siguiente expresión CLEM para descartar registros: `not(@NULL(Rule_ID))`.

Si desea obtener más información acerca de la transposición de puntuaciones para el despliegue, contacte el servicio de asistencia técnica.

## Nodo Secuencia

El nodo Secuencia descubre patrones, en datos secuenciales u ordenados en el tiempo, con el formato `pan > queso`. Los elementos de una secuencia son **conjuntos de elementos** que constituyen una única transacción. Por ejemplo, si una persona va a la tienda y compra pan y leche y, varios días después, vuelve a la tienda para comprar un poco de queso, la actividad de compras de esa persona se puede representar como dos conjuntos de elementos. El primer conjunto de elementos contiene pan y leche y el segundo contiene queso. Una **secuencia** es una lista de conjuntos de elementos que tiende a producirse en un orden previsible. El nodo Secuencia detecta secuencias frecuentes y crea un nodo de modelo generado que se puede utilizar para realizar predicciones.

**Requisitos.** Para crear un conjunto de reglas del nodo Secuencia, es necesario especificar un campo de ID, un campo de tiempo opcional y uno o varios campos de contenido. Observe que estos ajustes se deben realizar en la pestaña Campos del nodo de modelado; no se pueden leer en el nodo Tipo anterior de la ruta. El campo de ID puede tener cualquier rol o nivel de medición. Si se especifica un campo de tiempo, éste puede tener cualquier rol, aunque su almacenamiento debe ser numérico, una fecha, una hora o una marca de tiempo. Si no se especifica un campo de tiempo, el nodo Secuencia utilizará una marca de tiempo implícita que se activará utilizando los números de fila como valores de tiempo. Los campos de contenido pueden tener cualquier nivel de medición y rol, pero todos los campos de contenido deben ser del mismo tipo. Si son numéricos, deben ser rangos enteros (no rangos reales).

**Puntos fuertes.** El nodo Secuencia se basa en el algoritmo de reglas de asociación de CARMA, que utiliza un método de dos pasos para encontrar las secuencias. Además, el nodo de modelo generado que ha creado el nodo Secuencia se puede insertar en una ruta de datos con el fin de crear predicciones. El nodo de modelo generado también puede generar Supernodos para detectar y contar las secuencias específicas y realizar predicciones basadas en secuencias específicas.

## Opciones de campos para el nodo Secuencia

Antes de ejecutar un nodo de Secuencia, se deben especificar los campos de ID y de contenido en la pestaña Campos del nodo Secuencia. Si desea utilizar un campo de tiempo, también será necesario especificarlo aquí.

**Campo de ID.** Seleccione un campo de ID de la lista. Los campos numéricos o simbólicos se pueden utilizar como campo de ID. Cada valor exclusivo de este campo debe indicar una unidad de análisis específica. Por ejemplo, en una aplicación de la cesta de la compra, cada ID puede representar a un sólo cliente. Para una aplicación de análisis del registro Web, cada ID puede representar un equipo (con la dirección IP) o un usuario (con los datos de inicio de sesión).

- **Los ID son contiguos.** Si los datos se han clasificado previamente de forma que todos los registros con el mismo ID se agrupan en la ruta de datos, seleccione esta opción para que el procesamiento sea más rápido. Si los datos no se han clasificado previamente (o no lo sabe a ciencia cierta), no active esta opción y el nodo Secuencia clasificará los datos automáticamente.

*Nota*: si los datos no están clasificados y selecciona esta opción, es posible que obtenga resultados no válidos en el modelo de secuencias.

**Campo de tiempo.** Si desea utilizar un campo existente en los datos para indicar el momento de los eventos, seleccione **Utilizar campo de tiempo** y especifique el campo que desea utilizar. El campo de tiempo debe ser un número, una fecha, una hora o una marca de tiempo. Si no se especifica un campo de tiempo, se asume que los registros llegan del origen de datos en orden secuencial y los números del registro se utilizan como valores de tiempo (el primer registro se produce en el tiempo `"1"`; el segundo, en el tiempo `"2"`; y así sucesivamente).

**Campos de contenido.** Especifique los campos de contenido del modelo. Estos campos contienen los eventos de interés del modelado de secuencia.

El nodo Secuencia puede tratar datos en formato tabular o transaccional. Si se utilizan varios campos con datos transaccionales, se asume que los elementos especificados en estos campos para un registro determinado representan los elementos encontrados en una sola transacción con una sola marca de tiempo. Consulte el tema “Datos tabulares frente a datos transaccionales” en la página 268 para obtener más información.

**Partición.** Este campo permite especificar un campo usado para dividir los datos en muestras independientes para las fases de entrenamiento, prueba y validación en la generación del modelo. Si usa una muestra para generar el modelo y otra muestra distinta para comprobarlo, podrá obtener una buena indicación de la bondad del modelo a la hora de generar conjuntos de datos de mayor tamaño similares a los datos actuales. Si se han definido varios campos de partición mediante nodos Tipo o Partición, se deberá seleccionar un campo de partición simple en la pestaña Campos en todos los nodos de modelado que usen la partición. (Si solamente hay una partición, se usará automáticamente siempre que se active la partición.) Debe tener en cuenta que al aplicar la partición seleccionada en su análisis, también debe activar la partición en la pestaña Opciones del modelo para el nodo. (Si se elimina la selección de esta opción, se posibilita la desactivación de la partición sin cambiar la configuración del campo.)

## Opciones de modelo para el nodo Secuencia

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Soporte mínimo de reglas (%)** Puede especificar un criterio de soporte. *Soporte de la regla* hace referencia a la proporción de campos de ID existentes en los datos de entrenamiento que contienen la secuencia completa. Si desea centrarse en las secuencias más comunes, aumente el valor de este parámetro.

**Confianza de reglas mínima (%)** Se puede especificar un criterio de confianza para mantener las secuencias en el conjunto de secuencias. *La confianza* hace referencia al porcentaje de campos de ID en el que se realiza una predicción correcta a partir de todos los campos de ID para los que la regla realiza una predicción. Se calcula como la cantidad de ID en los que se encuentra toda la secuencia dividido por la cantidad de ID en los que se encuentran los antecedentes, basado en los datos de entrenamiento. Las secuencias con una confianza inferior a la del criterio especificado se descartan. Si se obtienen demasiadas secuencias o secuencias sin interés, pruebe a aumentar el valor de este parámetro. Si se obtienen muy pocas secuencias, pruebe a disminuir el valor de este parámetro.

**Nota:** si es necesario, puede resaltar el valor y el tipo en su propio valor. Tenga en cuenta que, si reduce el valor de confianza por debajo de 1,0, además de que el proceso requiere gran cantidad de memoria libre, puede que las reglas tarden un tiempo extremadamente largo en generarse.

**Tamaño máximo de secuencias** Puede establecer el número máximo de elementos distintos de una secuencia. Si las secuencias de interés resultantes son pocas, se puede disminuir el valor del parámetro para que el conjunto de secuencias se genere más rápido.

**Predicciones para añadir a la ruta** Especifique el número de predicciones que desea que el nodo Modelo resultante añada a la ruta. Si desea obtener más información, consulte "Nugget del modelo de secuencia" en la página 287.

## Opciones de experto para el nodo Secuencia

Las siguientes opciones de experto permiten a los usuarios con conocimientos sobre la operación del nodo Secuencia ajustar el proceso de generación de modelos. Para acceder a estas opciones, active el modo **Experto** en la pestaña Experto.

**Definir duración máxima.** En caso de seleccionar esta opción, las secuencias estarán limitadas a aquellas que tengan una duración (tiempo entre el primer conjunto de elementos y el último) inferior o igual al valor especificado. Si no se ha especificado un campo de tiempo, la duración se expresa en términos de filas (registros) existentes en los datos sin procesar. Si el campo de tiempo utilizado es una hora, una fecha o una marca de tiempo, la duración se expresa en segundos. En el caso de los campos numéricos, la duración se expresa con las mismas unidades que el campo en sí.

**Definir valor de poda.** El algoritmo CARMA utilizado en el nodo Secuencia elimina periódicamente (**poda**) los conjuntos de elementos poco frecuentes de la lista de conjuntos de elementos potenciales durante el procesamiento para conservar la memoria. Seleccione esta opción para ajustar la frecuencia de poda. El número especificado determina la frecuencia de poda. Introduzca un valor más pequeño para disminuir los requisitos de memoria del algoritmo (pero aumentar potencialmente el tiempo de entrenamiento necesario) o introduzca un valor mayor para que el entrenamiento sea más rápido (pero aumentar potencialmente los requisitos de memoria).

**Definir secuencias máximas en memoria.** Si selecciona esta opción, el algoritmo CARMA limitará el almacenamiento en memoria de secuencias de candidatos durante la generación del modelo al número de secuencias especificado. Seleccione esta opción si IBM SPSS Modeler utiliza demasiada memoria durante la generación de modelos de Secuencia. Observe que el valor máximo de secuencias que se especifica aquí es el número de secuencias de candidatos registrados internamente cuando se genera el modelo. Este número debe ser mucho mayor que el número de secuencias previsto para el modelo final.

**Restringir discontinuidades entre conjuntos de elementos.** Esta opción permite especificar las restricciones en las discontinuidades de tiempo que separan los conjuntos de elementos. Si se selecciona esta opción, los conjuntos de elementos con discontinuidades de tiempo inferiores a la **Discontinuidad**

**mínima** o superiores a la **Discontinuidad máxima** que se especifiquen no se considerarán como parte integrante de una secuencia. Utilice esta opción para evitar el recuento de secuencias que incluyen intervalos de tiempo largos o intervalos que se producen en un marco temporal muy corto.

*Nota*: si el campo de tiempo utilizado es una hora, una fecha o una marca de tiempo, la discontinuidad de tiempo se expresa en segundos. Para los campos numéricos, la discontinuidad de tiempo se expresa con las mismas unidades que el campo de tiempo.

Por ejemplo, observe la siguiente lista de transacciones.

*Tabla 20. Ejemplo de lista de transacciones.*

| ID | Hora | Contenido |
|---|---|---|
| 1001 | 1 | manzanas |
| 1001 | 2 | pan |
| 1001 | 5 | queso |
| 1001 | 6 | ropa |

Si se genera un modelo sobre estos datos con la discontinuidad mínima establecida en 2, se obtendrían las siguientes secuencias:

```
manzanas > queso

manzanas > ropa

pan > queso

pan > ropa
```

No aparecerían secuencias tales como `manzanas > pan`, porque la discontinuidad entre `manzanas` y `pan` es inferior a la discontinuidad mínima. Del mismo modo, tenga en cuenta los siguientes datos alternativos.

*Tabla 21. Ejemplo de lista de transacciones.*

| ID | Hora | Contenido |
|---|---|---|
| 1001 | 1 | manzanas |
| 1001 | 2 | pan |
| 1001 | 5 | queso |
| 1001 | 20 | ropa |

Si la discontinuidad máxima se hubiese establecido en 10, no aparecería ninguna secuencia con `ropa`, porque la discontinuidad entre `queso` y `ropa` es demasiado amplia para que se consideren parte de la misma secuencia.

## Nugget del modelo de secuencia

Los nuggets de modelo de ruta representan las rutas que se encuentran para un campo de salida determinado descubierto por el nodo Secuencia y pueden añadirse a rutas para generar predicciones.

Al ejecutar una ruta que contenga un nodo de ruta, dicho nodo añade en los datos un par de campos con las predicciones y los valores de confianza asociados para cada predicción desde el modelo de ruta. De forma predeterminada, se añaden tres pares de campos con las tres mejores predicciones (y los valores de confianza asociados). Puede cambiar el número de predicciones generadas al crear el modelo mientras configura las opciones de modelo para el nodo Secuencia durante la creación, así como en la pestaña

Configuración después de añadir el nugget de modelo a una ruta. Consulte el tema "Configuración del nugget de modelo de secuencia" en la página 291 para obtener más información.

Los nuevos nombres de campos se derivan del nombre del modelo. Los nombres de campos son *$S-secuencia-n* para el campo de predicción (donde *n* indica la *enésima* predicción) y *$SC-secuencia-n* para el campo de la confianza. En una ruta con varios nodos de reglas de ruta en una serie, los nuevos nombres de campos incluirán números en el prefijo para distinguirse entre sí. El primer nodo de conjuntos de ruta de la ruta utilizará los nombres normales, el segundo usará los nombres que comiencen por *$S1-* y *$SC1-*, mientras que el tercer nodo utilizará nombres que comiencen por *$S2-* y *$SC2-*, y así sucesivamente. Las predicciones se muestran en orden de confianza, por lo que *$S-secuencia-1* contiene la predicción con la mayor confianza, mientras que*$S-secuencia-2* contiene la siguiente mayor confianza, y así sucesivamente. Para los registros en los que el número de predicciones disponibles es menor que el número de predicciones solicitadas, las predicciones restantes contienen el valor `$null$`. Por ejemplo, si solo se pueden realizar dos predicciones para un registro particular, los valores de *$S-secuencia-3* y *$SC-secuencia-3* serán `$null$`.

Para cada registro, se comparan las reglas del modelo con el conjunto de transacciones para el ID actual hasta ese momento, incluido el registro actual y cualquier registro previo con el mismo ID y cadena de tiempo previa. Se utilizan las reglas *k* con los mayores valores de confianza que se aplican a este conjunto de transacciones para generar las predicciones *k* para el registro, donde *k* es el número de predicciones especificado en la pestaña Configuración después de añadir el modelo a la ruta. (Si varias reglas predicen el mismo resultado para el conjunto de transacciones, sólo se utilizará la regla con la mayor confianza.) Consulte el tema "Configuración del nugget de modelo de secuencia" en la página 291 para obtener más información.

Como ocurre con otros tipos de modelos de reglas de asociación, el formato de datos debe coincidir con el formato utilizado al crear el modelo de secuencias. Por ejemplo, los modelos creados con datos tabulares pueden utilizarse para puntuar sólo datos tabulares. Consulte el tema "Reglas de asociación de la puntuación" en la página 281 para obtener más información.

*Nota*: al puntuar los datos mediante un nodo de conjuntos de ruta generado en una ruta, cualquier configuración de tolerancia o discontinuidad seleccionada en la creación del modelo se omite para efectuar la puntuación.

Predicciones de las reglas de secuencia

El nodo trata los registros dependiendo del tiempo (o del orden, si no se ha utilizado ninguna marca de tiempo para crear el modelo). Los registros deben ordenarse por el campo ID y el campo de marca de tiempo (si hay alguno). Sin embargo, las predicciones no concuerdan con la marca de tiempo del registro al que se añaden. Sencillamente se refieren a los elementos que más probabilidades tienen de producirse *en algún momento futuro*, dado el historial de transacciones para el ID actual hasta el registro actual.

Tenga en cuenta que las predicciones para cada registro no dependen necesariamente de las transacciones de ese registro. Si las transacciones de registro actuales no activan una regla específica, las reglas se seleccionarán en función de las transacciones previas para el ID actual. En otras palabras, si el registro actual no añade ninguna información de predicción útil a la secuencia, la predicción se transfiere al registro actual desde la última transacción útil de este ID.

Por ejemplo, imagine que tiene un modelo de secuencia con una única regla

```
Mermelada > Pan (0.66)
```

y le pasa los siguientes registros.

*Tabla 22. Registros de ejemplo.*

| ID | Compra | Predicción |
|---|---|---|
| 001 | mermelada | pan |
| 001 | leche | pan |

Observe que el primer registro genera una predicción de *pan*, como se esperaba. El segundo registro también contiene una predicción de *pan*, porque no existe ninguna regla para la *mermelada* seguida de *leche*. Por lo tanto, la transacción de la *leche* no añade información útil y la regla `Mermelada > Pan` se sigue aplicando.

Generación de nuevos nodos

El menú Generar permite crear nuevos Supernodos basados en el modelo de secuencia.

- **Supernodo Regla.** Crea un Supernodo que puede detectar y contar las instancias de las secuencias en los datos puntuados. Si no se selecciona ninguna regla, esta opción está desactivada. Consulte el tema "Generación de un Supernodo Regla a partir de un nugget de modelo de secuencia" en la página 291 para obtener más información.
- **Modelo a paleta** Devuelve el modelo a la paleta de modelos. Esto resulta útil en las situaciones en que un colega le haya enviado una ruta que contenga un modelo y no el modelo en sí.

## Detalles del nugget de modelo de secuencia

La pestaña Modelo para un nugget de secuencia muestra las reglas exactamente por el algoritmo. Cada fila de la tabla representa una regla, con el antecedente (la parte "si" de la regla) en la primera columna seguida del consecuente (la parte "entonces" de la regla) en la segunda columna.

Cada regla se muestra en el siguiente formato.

*Tabla 23. Formato de regla*

| Antecedente | Consecuente |
|---|---|
| cerveza y lata_veg | cerveza |
| pescado<br>pescado | pescado |

La primera regla de ejemplo se interpreta como *para los ID que tenían "cerveza" y "lata_veg" en la misma transacción, donde hay probabilidad de una futura aparición de "cerveza."* La segunda regla de ejemplo se puede interpretar como *para los ID que tenían "pescado" en una transacción y, a continuación, "pescado" en otra, donde hay probabilidad de una futura aparición de "pescado"*. Tenga en cuenta que en la primera regla, *cerveza* y *lata_veg* se adquieren al mismo tiempo y, en la segunda, *pescado* se adquiere en dos transacciones independientes.

**Menú Ordenar.** El botón del menú Ordenar en la barra de herramientas controla la ordenación de las reglas. Es posible cambiar la dirección de ordenación (ascendente o descendente) mediante el botón de dirección de la ordenación (flecha arriba y abajo).

Los valores se pueden ordenar por:

- % de soporte
- Confianza
- Soporte de regla %
- Consecuente
- Primer antecedente
- Último antecedente

- Número de elementos (antecedentes)

Por ejemplo, la siguiente tabla se clasifica en orden descendente por el número de elementos. Las reglas con varios artículos en el conjunto de antecedentes preceden a las que incluyen pocos artículos.

*Tabla 24. Reglas clasificadas por número de elementos*

| Antecedente | Consecuente |
|---|---|
| cerveza y lata_veg y congelados | congelados |
| cerveza y lata_veg | cerveza |
| pescado<br>pescado | pescado |
| refrescos | refrescos |

**Mostrar/ocultar criterios.** El botón Mostrar/ocultar criterios (icono de cuadrícula) controla las opciones de visualización de las reglas. Están disponibles las siguientes opciones de presentación:

- **Instancias** muestra información acerca del número de ID exclusivos para el que se produce la *secuencia completa*, antecedentes y consecuentez. (Tenga en cuenta que esto difiere de los modelos de asociación, para los que el número de instancias hace referencia al número de ID para los que *sólo* se aplican los antecedentes.) Por ejemplo, dada la regla `pan > queso`, se hace referencia al número de ID en los datos de entrenamiento que incluyan *pan* y *queso* como **instancias**.
- **Soporte** muestra la proporción de ID en los datos de entretenimiento para los que los antecedentes son verdaderos. Por ejemplo, si el 50% de los datos de entrenamiento incluye el antecedente *pan*, el soporte para la regla `pan > queso` sería del 50%. (A diferencia de los modelos de asociación, el soporte *no* se basa en el número de instancias, como se indicó anteriormente.)
- **La confianza** muestra el porcentaje de campos de ID en el que se realiza una predicción correcta a partir de todos los campos de ID para los que la regla realiza una predicción. Se calcula como la cantidad de ID en los que se encuentra toda la secuencia dividido por la cantidad de ID en los que se encuentran los antecedentes, basado en los datos de entrenamiento. Por ejemplo, si el 50% de los datos de entrenamiento contienen `lata_veg` (indicando así soporte del antecedente), pero sólo el 20% contiene tanto `lata_veg` como `congelados`, la confianza para la regla `lata_veg > congelados` sería `Soporte de regla / Soporte de antecedentes` o, en este caso, 40%.
- **Soporte de regla** para los modelos de secuencias está basado en instancias y muestra la proporción de registros de entrenamiento para los que se cumple que toda la regla, los antecedentes y consecuentes, son verdaderos. Por ejemplo, si el 20% de los datos de entrenamiento incluyen tanto *pan* como *queso*, el soporte de la regla `pan > queso` será el 20%.

Recuerde que las proporciones se basan en transacciones válidas (transacciones con al menos un elemento observado o valor verdadero) en lugar de transacciones totales. Las transacciones no válidas, las que no tienen elementos o valores verdaderos, se descartan para estos cálculos.

**Botón de filtrado.** En el menú, el botón de filtrado (icono del embudo) expande el botón del cuadro de diálogo para mostrar un panel en el que aparecen los filtros de regla activos. Los filtros se utilizan para contraer el número de reglas que se muestran en la pestaña Modelos.

*Figura 52. Botón de filtrado*

Para crear un filtro, pulse en el icono de filtrado, en la parte derecha del panel expandido. Esta operación abre un cuadro de diálogo independiente en el que se pueden especificar las restricciones a la hora de mostrar reglas. Tenga en cuenta que el botón de filtrado se suele utilizar junto con el menú Generar para,

en primer lugar, filtrar las reglas y, a continuación, generar un modelo que contenga ese subconjunto de reglas. Si desea obtener más información, consulte “Especificación de filtros para reglas” en la página 277 a continuación.

## Configuración del nugget de modelo de secuencia

La pestaña Configuración de un modelo de secuencia muestra opciones de puntuación para el modelo. Esta pestaña sólo estará disponible después de que el modelo se haya añadido al lienzo de rutas para la puntuación.

**Número máximo de predicciones.** Especifique el número máximo de predicciones incluidas para cada conjunto de elementos de la cesta. Las reglas con los valores de confianza más altos que se aplican a este conjunto de transacciones se utilizan para generar predicciones para el registro hasta el límite especificado.

## Resumen de nugget de modelo de secuencia

La pestaña Resumen para un nugget de modelo de reglas de secuencia muestra el número de reglas descubiertas y el mínimo y máximo de soporte y confianza en las reglas. Si ha ejecutado un nodo Análisis conectado a este nodo de modelado, la información de dicho análisis también se mostrará en esta sección.

Consulte el tema “Examen de nuggets de modelo” en la página 42 para obtener más información.

## Generación de un Supernodo Regla a partir de un nugget de modelo de secuencia

Para generar un Supernodo Regla basado en una regla de secuencia:

1. En la pestaña Modelo para el nugget de modelo de regla de secuencia, pulse en una fila de la tabla para seleccionar la regla deseada.
2. Seleccione en los menús del explorador de reglas:

   **Generar** > **Supernodo Regla**

*Importante*: Para utilizar el Supernodo generado, debe clasificar los datos por el campo de ID (y por el campo de tiempo, si fuera necesario) antes de pasarlos al Supernodo. En los datos sin clasificar, el Supernodo no detectará secuencias de manera adecuada.

Puede especificar las siguientes opciones para generar un Supernodo regla:

**Detectar.** Especifica cómo se definen las coincidencias para los datos que han pasado al Supernodo.

- **Sólo antecedentes.** El Supernodo identifica una coincidencia cada vez que encuentra los antecedentes para la regla seleccionada en el orden correcto dentro de un conjunto de registros con el mismo ID, sin importar si se ha encontrado también el antecedente. Recuerde que esto no tiene en cuenta la tolerancia de marca de tiempo o la configuración de la restricción de la discontinuidad del elemento del nodo de modelado Secuencia original. Cuando se detecta el último conjunto de elementos de antecedentes en la ruta (y se ha encontrado el resto de antecedentes en el orden correcto), los registros posteriores con el ID actual contienen el resumen seleccionado más abajo.
- **Secuencia completa.** El Supernodo identifica una coincidencia cada vez que encuentra los antecedentes y los consecuentes para la regla seleccionada en el orden correcto dentro de un conjunto de registros con el mismo ID, sin importar si se ha encontrado también el antecedente. Esto no tiene en cuenta la tolerancia de marca de tiempo o la configuración de la restricción de la discontinuidad del elemento del nodo de modelado Secuencia original. Cuando se detecta el consecuente en la ruta (y, asimismo, se han encontrado todos los antecedentes en orden correcto), el registro actual y los registros posteriores con el ID actual contiene el resumen seleccionado más abajo.

**Mostrar.** Controla la forma en que los resúmenes de coincidencias se añaden a los datos en el resultado del Supernodo Regla.

- **Valor consecuente para la primera instancia.** El valor añadido a los datos es el valor consecuente predicho según la primera instancia de la coincidencia. Los valores se añaden como un nuevo campo llamado *rule_n_consequent*, donde *n* es el número de regla (basado en el orden de creación de los Supernodos Regla de la ruta).
- **Valor verdadero para la primera instancia.** El valor que se añade a los datos es verdadero si hay al menos una coincidencia para el ID, mientras que será falso si no existe coincidencia alguna. Los valores se añaden como un nuevo campo llamado *rule_n_flag*.
- **Recuento de ocurrencias.** El valor añadido a los datos es el número de coincidencias para el ID. Los valores se añaden como un nuevo campo llamado *rule_n_count*.
- **Número de regla.** El valor añadido es el número de regla para la regla seleccionada. Los **números de regla** se asignan en función del orden en que se añadió el Supernodo a la ruta. Por ejemplo, el primer Supernodo Regla se considera *regla 1*, el segundo Supernodo Regla se considera *regla 2*, y así sucesivamente. Esta opción resulta especialmente útil al incluir varios Supernodos Regla en la ruta. Los valores se añaden como un nuevo campo llamado *rule_n_number*.
- **Incluir cifras de confianza.** Si se selecciona esta opción, añadirá la confianza de reglas a la ruta de los datos, así como el otro resumen seleccionado. Los valores se añaden como un nuevo campo llamado *rule_n_confidence*.

## Nodo de reglas de asociación

Las reglas de asociación son instrucciones del formato siguiente

Por ejemplo, "Si un cliente compra una cuchilla y una loción para después del afeitado, entonces hay un 80% de posibilidades de que el cliente compre también crema de afeitado". El nodo Reglas de asociación extrae un conjunto de reglas de los datos y destaca aquellas reglas con un mayor contenido de información. El nodo Reglas de asociación es muy parecido al nodo Apriori. Sin embargo, hay algunas diferencias notables:

- El nodo Reglas de asociación no puede procesar los datos transaccionales.
- El nodo Reglas de asociación puede procesar datos que tienen el tipo de almacenamiento Lista y el nivel de medición Colección.
- El nodo Reglas de asociación se puede utilizar con IBM SPSS Analytic Server. Esto proporciona escalabilidad y significa que se pueden procesar big data y aprovechar un procesamiento paralelo más rápido.
- El nodo Reglas de asociación proporciona valores adicionales, como la capacidad de limitar el número de reglas que se generan, aumentando así la velocidad de procesamiento.
- La salida del nugget de modelo se muestra en el Visor de salidas.

**Nota:** El nodo Reglas de asociación no soporte los pasos Evaluación del modelo o Ganador-Comparativo en IBM SPSS Collaboration and Deployment Services.

**Nota:** El nodo Reglas de asociación ignora registros vacíos al generar un modelo si el tipo de campo es de marca. Los registros vacíos son registros en los que todos los campos utilizados en la generación de modelo tienen un valor de false.

Una ruta que muestra un ejemplo de trabajo de cómo utilizar Reglas de asociación, llamado `geospatial_association.str`, y que hace referencia a los archivos de datos `InsuranceData.sav`, `CountyData.sav` y `ChicagoAreaCounties.shp` está disponible en el directorio `Demos` de su instalación de IBM SPSS Modeler. Puede acceder al directorio `Demos` desde el grupo de programas de IBM SPSS Modeler en el menú Inicio de Windows. El archivo `geospatial_association.str` se encuentra en el directorio `streams`.

# Reglas de asociación - Opciones de campos

En la pestaña **Campos**, puede seleccionar si desea utilizar la configuración de rol de campo ya definida en nodos anteriores o realizar las asignaciones de campos manualmente.

**Utilizar roles predefinidos**
Esta opción utiliza las definiciones de roles (como objetivos o predictores) desde un nodo Tipo anterior (o la pestaña Tipo de un nodo de origen anterior). Los campos con un rol de entrada se consideran Condiciones, los campos con un rol de destino se consideran Predicciones, y los campos usados como entradas y objetivos se considera que tienen ambos roles.

**Utilizar asignaciones de campos personalizadas**
Seleccione esta opción si desea asignar objetivos, predictores y otros roles manualmente en esta pantalla.

**Campos**
Si ha seleccionado **Utilizar asignaciones de campos personalizadas**, utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los recuadros de la parte derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo.

**Ambos (condición o predicción)**
Los campos añadidos a esta lista pueden tomar el rol de condición o de predicción en las reglas generadas por el modelo. Esto se hace regla por regla, de manera que un campo puede ser una condición en una regla y una predicción en otra.

**Solo predicción**
Los campos añadidos a esta lista pueden aparecer solo como una predicción (también conocida como un "consecuente") de una regla. La presencia de un campo en esta lista no significa que el campo se utilice en reglas, sino solamente que si se utiliza puede ser solo una predicción.

**Solo condición**
Los campos añadidos a esta lista pueden aparecer solo como una condición (también conocida como un "antecedente") de una regla. La presencia de un campo en esta lista no significa que el campo se utilice en reglas, sino solamente que si se utiliza puede ser solo una condición.

# Reglas de asociación - Generación de reglas

## Elementos por regla

Use estas opciones para especificar cuántos elementos o valores se pueden usar en cada regla.

**Nota:** El total combinado de estos dos campos no puede superar 10.

**Número máximo de condiciones**
Seleccione el número máximo de condiciones que se pueden incluir en una sola regla.

**Número máximo de predicciones**
Seleccione el número máximo de predicciones que se pueden incluir en una sola regla.

## Generación de reglas

Use estas opciones para especificar el número y tipo de reglas que se va a generar.

**Número máximo de reglas**
Especifique el número máximo de reglas que se pueden considerar para usar para generar reglas para su modelo.

**Criterio de reglas N superiores**
Seleccione el criterio que se utiliza para establecer cuáles son las reglas N superiores, donde N es el valor introducido en el campo **Maximum number of rules**. Puede elegir entre los criterios siguientes.

- **Confianza**

- **Soporte de regla**
- **Soporte de condición**
- **Lift**
- **Capacidad de despliegue**

**Sólo valores verdaderos para las marcas**
Cuando sus datos estén en formato tabular, seleccione esta opción para incluir solo los valores verdaderos para los campos de distintivo en las reglas resultantes. Seleccionar valores verdaderos puede ayudar a que las reglas se entiendan con más facilidad. La opción no se aplica a los datos en formato transaccional. Si desea obtener más información, consulte "Datos tabulares frente a datos transaccionales" en la página 268.

### Criterio de regla

Si selecciona **Habilitar criterio de regla**, puede usar estas opciones para seleccionar la fuerza mínima que las reglas deben cumplir para considerarse para el uso en su modelo.

- **Confianza** Especifique el valor porcentual mínimo para el nivel de Confianza para una regla producida por el modelo. Si el modelo produce una regla con un nivel inferior a este valor, la regla se descarta.
- **Soporte de regla** Especifique el valor porcentual mínimo para el nivel de soporte de regla para una regla producida por el modelo. Si el modelo produce una regla con un nivel inferior a este valor, la regla se descarta.
- **Soporte de condición** Especifique el valor porcentual mínimo para el nivel de soporte de regla para una regla producida por el modelo. Si el modelo produce una regla con un nivel inferior al valor indicado, la regla se descarta.
- **Elevación** Especifique el valor de elevación mínimo permitido para una regla producida por el modelo. Si el modelo produce una regla con un valor inferior al indicado, la regla se descarta.

### Excluir reglas

En algunos casos, la asociación entre dos o más campos es conocida o evidente. En estos casos, puede excluir reglas cuando los campos se predicen mutuamente. Al excluir reglas que contienen ambos valores, se reduce la entrada irrelevante de datos y se aumentan las posibilidades de encontrar resultados útiles.

**Campos**
Seleccione los campos asociados que no desee usar juntos en la generación de reglas. Por ejemplo, algunos campos asociados pueden ser Fabricante de coche y Modelo de coche, o bien Año escolar y Edad del alumno. Cuando el modelo crea reglas, si la regla contiene al menos uno de los campos seleccionados en un lado de la regla (condición o predicción), la regla se descarta.

## Reglas de asociación - Transformaciones

### En intervalo

Use estas opciones para especificar cómo se agrupan los campos continuos (intervalo numérico).

**Número de intervalos**
Los campos continuos configurados para que se agrupen automáticamente están divididos en el número de intervalos igualmente espaciados que se indique. Puede seleccionar cualquier número del intervalo 2 - 10.

### Campos de la lista

**Longitud máxima de la lista**
Para restringir el número de elementos que se van a incluir en el modelo si no se conoce la longitud de un campo de la lista, indique la longitud máxima de la lista. Puede seleccionar

cualquier número del intervalo 1 - 100. Si una lista es más larga que el número que indique, el modelo seguirá usando el campo pero solo incluirá valores hasta este número; los valores extra del campo se omitirán.

# Reglas de asociación - Salida

Use las opciones de este panel para controlar la salida que se genera cuando se genera el modelo.

## Tablas de reglas

Use estas opciones para crear uno o más tipos de tablas que muestren el mejor número de reglas (basándose en un número que indique) para cada criterio seleccionado.

**Confianza**
Confianza es el cociente entre el soporte de regla y el soporte de condiciones. De los elementos con los valores de la lista, el porcentaje que tiene los valores consecuentes pronosticados. Crea una tabla que contiene las mejores reglas de asociación N basadas en la confianza que se van a incluir en la salida (donde N es el valor **Reglas para mostrar**).

**Soporte de regla**
La proporción de elementos para los que toda la regla, las condiciones y las predicciones son verdaderas. Para todos los elementos del conjunto de datos, el porcentaje que la regla contabiliza correctamente y predice. Esta medida da una importancia general para la regla. Crea una tabla que contiene las mejores reglas de asociación N basadas en el soporte de reglas que se van a incluir en la salida (donde N es el valor **Reglas para mostrar**).

**Elevación**
La tasa de confianza de reglas y la probabilidad anterior de tener la predicción. La tasa del valor de Confianza para una regla frente al porcentaje de que se produzcan los valores consecuentes en la población general. Esta tasa da una medida de cuánto mejora la regla respecto a la probabilidad. Crea una tabla que contiene las mejores reglas de asociación N basadas en la elevación que se van a incluir en la salida (donde N es el valor **Reglas para mostrar**).

**Soporte de condición**
La proporción de elementos para los que las condiciones son verdaderas. Crea una tabla que contiene las mejores reglas de asociación N basadas en el soporte de antecedentes que se van a incluir en la salida (donde N es el valor **Reglas para mostrar**).

**Capacidad de despliegue**
Medida del porcentaje en la que los datos de formación satisfacen la condición pero no la predicción. Esta medida muestra la frecuencia con la que la regla falla. Es de hecho el opuesto de la Confianza. Crea una tabla que contiene las mejores reglas de asociación N basadas en la capacidad de despliegue que se van a incluir en la salida (N es el valor **Reglas para mostrar**).

**Reglas para mostrar**
Establezca el número máximo de reglas para mostrar en las tablas.

## Tablas de información de modelos

Use una o más de estas opciones para seleccionar qué tablas de modelos incluir en la salida.

- **Transformaciones de campos**
- **Resumen de registros**
- **Estadísticas de regla**
- **Valores más frecuentes**
- **Campos más frecuentes**

### Nube de palabras de reglas clasificable.

Use estas opciones para crear una nube de palabras que muestre las salidas de reglas. Las palabras se muestran en tamaños de texto en aumento para indicar su importancia.

**Cree una nube de palabras clasificable.**
Marque esta casilla de verificación para crear una nube de palabras clasificable en su salida.

**Clasificación predeterminada**
Seleccione el tipo de clasificación que usar al crear inicialmente la nube de palabras. La nube de palabras es interactiva y puede cambiar el criterio en el Visor de modelos para ver las diversas reglas y clasificaciones. Puede elegir entre las siguientes opciones de clasificación:

- Confianza.
- Soporte de regla
- Elevación
- Soporte de condición.
- Capacidad de despliegue

**Máximo de reglas para mostrar**
Establezca el número de reglas que se van a mostrar en el mapa; el máximo que puede seleccionarse es 20.

## Reglas de asociación - Opciones de modelos

Use los valores de esta pestaña para indicar las opciones de puntuación para los modelos de reglas de asociación.

**Nombre del modelo** Puede generar el nombre de modelo que se base automáticamente en el campo de destino (o el tipo de modelo en casos en los que no se haya indicado ese campo), o especificar un nombre personalizado.

**Número máximo de predicciones** Indique el número máximo de predicciones incluidas en el resultado de puntuación. Esta opción se utiliza con la opción **Criterio de regla** para producir las “mejores” predicciones, en las que "mejores" indica el nivel más alto de confianza, soporte, elevación, etc.

**Criterio de regla** Seleccione la medida utilizada para determinar la fuerza de las reglas. Las reglas se clasifican según la fuerza de los criterios aquí seleccionados a fin de devolver las mejores predicciones para un conjunto de elementos. Puede elegir entre cinco criterios diferentes.

- **Confianza** Confianza es el cociente entre el soporte de regla y el soporte de condición. De los elementos con los valores de la lista, el porcentaje que tiene los valores consecuentes pronosticados.
- **Soporte de condición** La proporción de elementos para los que las condiciones son verdaderas.
- **Soporte de regla** La proporción de elementos para los que toda la regla, las condiciones y las predicciones son verdaderas. Calculado multiplicando el valor **Soporte de condición** por el valor de **Confianza**.
- **Elevación** La tasa de confianza de reglas y la probabilidad anterior de tener la predicción.
- **Capacidad de despliegue** Medida del porcentaje en la que los datos de formación satisfacen la condición pero no la predicción.

**Permitir predicciones repetidas** Seleccione esta opción para incluir varias reglas con la misma predicción a la hora de puntuar. Por ejemplo, si se selecciona esta opción se pueden puntuar las siguientes reglas.

```
pan y vino y queso
queso y fruta y vino
```

**Nota:** Las reglas con varias predicciones (`pan & queso & fruta -> vino & paté`) se consideran predicciones repetidas solo si todas las predicciones (`vino & paté`) se habían pronosticado antes.

**Solo puntuar reglas cuando las predicciones no están presentes en la entrada** Seleccione esta opción para asegurarse de que las predicciones no estén también presentes en la entrada. Por ejemplo, si el propósito de la puntuación es recomendar mobiliario para el hogar, no es probable que una entrada que ya contiene una mesa de comedor recomiende comprar otra. En este caso, seleccione esta opción. Sin embargo, si los productos son perecederos o desechables (como el queso, un biberón o un pañuelo), las reglas en las que el consecuente ya se encuentra en la entrada pueden ser de valor. En el último caso, la opción más útil puede ser **Puntuar todas las reglas**.

**Solo puntuar reglas cuando las predicciones están presentes en la entrada** Seleccione esta opción para asegurarse de que las predicciones estén también presentes en la entrada. Este método es útil cuando se intenta comprender mejor los clientes o las transacciones existentes. Por ejemplo, es posible que desee identificar las reglas con la mayor elevación y, a continuación, explorar qué clientes se ajustan a estas reglas.

**Puntuar todas las reglas** Seleccione esta opción para incluir todas las reglas a la hora de puntuar, independientemente de la presencia o ausencia de predicciones.

# Nugget del modelo de reglas de asociación

El nugget de modelos contiene información acerca de las reglas extraídas a partir de los datos durante la generación de modelos.

## Visualización de los resultados

Puede examinar las reglas generadas por los modelos de Association Rules Reglas de asociación utilizando la pestaña Modelo del cuadro de diálogo. Al examinar un nugget de modelo, se muestra la información acerca de las reglas antes de generar nuevos nodos o puntuar el modelo.

## Puntuación de modelos

Los nuggets de modelos refinados se pueden añadir a una ruta y utilizarse para puntuar. Consulte el tema “Uso de nugget de modelo en rutas” en la página 49 para obtener más información. Los modelos de nuggets utilizados para puntuar incluyen una pestaña Configuración adicional en los cuadros de diálogo respectivos. Consulte el tema “Configuración del nugget de modelo de reglas de asociación” en la página 298 para obtener más información.

## Detalles del nugget de modelos de reglas de asociación

El nugget de modelo de reglas de asociación muestra detalles del modelo en la pestaña Modelo del Visor de salidas. Para obtener más información sobre cómo utilizar el visor, consulte el apartado titulado "Cómo trabajar con la salida" en la Guía del usuario del modelador (ModelerUsersGuide.pdf).

La operación de modelado de GSAR crea varios campos nuevos con el prefijo $A como se indica en la tabla siguiente.

*Tabla 25. Campos nuevos creados por la operación de modelado de reglas de asociación*

| Nombre de campo | Descripción |
|---|---|
| $A-<predicción># | Este campo contiene la predicción del modelo para los registros puntuados.<br><br><predicción> es el nombre del campo incluido en el rol Predicciones del modelo, y # es una secuencia de números para las reglas de salida (por ejemplo, si se configura que la puntuación incluya 3 reglas, la secuencia de números será del 1 al 3). |

*Tabla 25. Campos nuevos creados por la operación de modelado de reglas de asociación (continuación)*

| | |
|---|---|
| $AC-<predicción># | Este campo contiene la confianza en la predicción.<br><br><predicción> es el nombre del campo incluido en el rol Predicciones del modelo, y # es una secuencia de números para las reglas de salida (por ejemplo, si se configura que la puntuación incluya 3 reglas, la secuencia de números será del 1 al 3). |
| $A-Rule_ID# | Esta columna contiene el ID de la regla pronosticada para cada registro del conjunto de datos puntuado.<br><br># es una secuencia de números para las reglas de salida (por ejemplo, si se configura que la puntuación incluya 3 reglas, la secuencia de números será del 1 al 3). |

## Configuración del nugget de modelo de reglas de asociación

La pestaña Configuración de un modelo de reglas de asociación muestra opciones de puntuación para el modelo. Esta pestaña sólo estará disponible después de que el modelo se haya añadido al lienzo de rutas para la puntuación.

**Número máximo de predicciones** Especifique el número máximo de predicciones que se incluyen para cada conjunto de elementos. Las reglas con los valores de confianza más altos que se aplican a este conjunto de transacciones se utilizan para generar predicciones para el registro hasta el límite especificado. Utilice esta opción con la opción **Criterio de regla** para producir las "mejores" predicciones, en las que *mejores* indica el nivel más alto de confianza, soporte, elevación, etc.

**Criterio de regla** Seleccione la medida utilizada para determinar la fuerza de las reglas. Las reglas se clasifican según la fuerza de los criterios aquí seleccionados a fin de devolver las mejores predicciones para un conjunto de elementos. Puede elegir entre los criterios siguientes.

- **Confianza**
- **Soporte de regla**
- **Lift**
- **Soporte de condición**
- **Capacidad de despliegue**

**Permitir predicciones repetidas** Seleccione esta opción para incluir varias reglas con el mismo consecuente a la hora de puntuar. Por ejemplo, seleccionar esta opción significa que se puede puntuar la siguiente regla:

```
pan y vino y queso
queso y fruta y vino
```

Para excluir predicciones repetidas al puntuar, deje la casilla de verificación en blanco.

**Nota:** las reglas con varios consecuentes (`pan & queso & fruta -> vino & paté`) se consideran predicciones repetidas sólo si todos los consecuentes (`vino & paté`) se han predicho anteriormente.

**Solo puntuar reglas cuando las predicciones no están presentes en la entrada** Seleccionar para garantizar que los consecuentes no estén presentes también en la entrada. Por ejemplo, si el propósito de la puntuación es recomendar mobiliario para el hogar, no es probable que una entrada que ya contiene una mesa de comedor recomiende comprar otra. En este caso, seleccione esta opción. Por otra parte, si los productos son perecederos o desechables (como el queso, un biberón o un pañuelo), las reglas en las que el consecuente ya se encuentra en la entrada pueden ser de valor. En el último caso, la opción más útil puede ser **Puntuar todas las reglas**.

**Solo puntuar reglas cuando las predicciones están presentes en la entrada** Seleccionar para garantizar que los consecuentes estén presentes también en la entrada. Este método es útil cuando se intenta comprender mejor los clientes o las transacciones existentes. Por ejemplo, es posible que desee identificar las reglas con la mayor elevación y, a continuación, explorar qué clientes se ajustan a estas reglas.

**Puntuar todas las reglas** Seleccione esta opción para incluir todas las reglas a la hora de puntuar, independientemente de la presencia o ausencia de consecuentes en la entrada.

# Capítulo 13. Modelos de series temporales

## ¿Por qué es importante hacer previsiones?

Hacer previsiones consiste en predecir los valores de una o varias series a lo largo del tiempo. Por ejemplo, puede que desee predecir la demanda esperada de una línea de productos o servicios con la finalidad de poder asignar recursos para su fabricación o distribución. Como para implementar las decisiones de planificación es necesario cierto tiempo, las previsiones son una herramienta esencial en muchos procesos de planificación.

Los métodos de modelado de series temporales suponen que la historia se repite, si no exactamente, de una manera lo suficientemente parecida como para que estudiando el pasado sea posible tomar decisiones mejores en el futuro. Para predecir las ventas del año que viene, por ejemplo, es probable que empiece examinando las ventas de este año y después las de años anteriores para averiguar las tendencias o los patrones, si los hay, que se han desarrollado en los últimos años. No obstante, los patrones pueden ser difíciles de calcular. Si las ventas aumentan durante varias semanas seguidas, por ejemplo, ¿forma esto parte de un ciclo estacional o se trata del principio de una tendencia a largo plazo?

Con las técnicas de modelado estadístico, puede analizar los patrones de los datos del pasado y proyectar dichos patrones para determinar el rango en el que probablemente se incluirán los valores futuros de la serie. Como resultado, se obtienen previsiones más precisas en las que podrá basar sus decisiones.

## Datos de series temporales

Una **serie temporal** es una colección ordenada de mediciones tomadas en intervalos regulares; por ejemplo, los precios diarios de las acciones o los datos de ventas semanales. Las mediciones pueden estar relacionadas con cualquier cosa que le interese, y cada serie se suele clasificar en una de las siguientes categorías:

- **Dependiente.** Una serie para la que desea hacer previsiones.
- **Predictora.** Serie que puede ayudar a explicar el objetivo, por ejemplo, el presupuesto de publicidad para predecir las ventas. Las series predictoras sólo se pueden usar con modelos ARIMA.
- **Evento.** Serie predictora especial que se utiliza para tener en cuenta incidentes recurrentes predecibles, por ejemplo promociones de ventas.
- **Intervención.** Serie predictora especial que se utiliza para tener en cuenta incidentes puntuales del pasado como, por ejemplo, apagones o huelgas.

Los intervalos pueden representar cualquier unidad de tiempo, pero debe utilizarse un mismo intervalo para todas las mediciones. Además, si algún intervalo no tiene ninguna medición, debe definirse en el valor perdido. De esta forma, el número de intervalos con mediciones (incluidos los que tienen valores perdidos) define la duración del período histórico de los datos.

## Características de las series temporales

Estudiar el comportamiento pasado de una serie le ayudará a identificar los patrones y realizar mejores previsiones. Cuando se representan, muchas series temporales muestran una o varias de estas características:

- Tendencias
- Ciclos estacionales y no estacionales
- Pulsos y pasos
- Valores atípicos

## Tendencias

Una **tendencia** es un cambio gradual ascendente o descendente en el nivel de la serie o la trayectoria que siguen los valores de la serie de aumentar o disminuir a lo largo del tiempo.

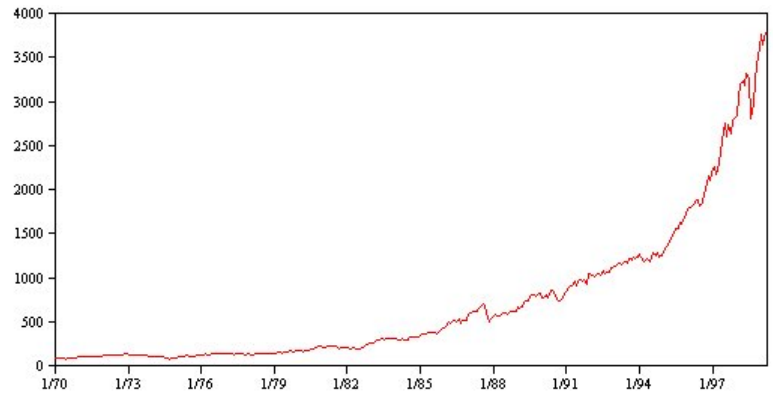



*Figura 53. Tendencia*

Las tendencias pueden ser **locales** o **globales**, pero una misma serie puede mostrar ambas. Históricamente, los gráficos de series del índice del mercado de valores muestran una tendencia global ascendente. Han aparecido tendencias descendentes locales en épocas de recesión y tendencias ascendentes locales en épocas de prosperidad.

Las tendencias también pueden ser **lineales** o **no lineales**. Las tendencias lineales son incrementos aditivos positivos o negativos en el nivel de la serie, comparables al efecto del interés simple sobre el principal. Las tendencias no lineales suelen ser multiplicativas, con incrementos proporcionales a los valores de series anteriores.

Las tendencias lineales globales son adecuadas y hacen previsiones correctas tanto con los modelos ARIMA como con los de suavizado exponencial. Al generar modelos ARIMA, suelen diferenciarse las series que muestran tendencias para eliminar el efecto de éstas.

## Ciclos estacionales

Un **ciclo estacional** es un patrón repetitivo y predecible de los valores de las series.

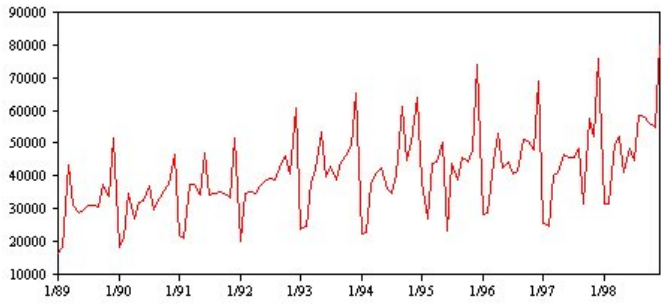



*Figura 54. Ciclo estacional*

Los ciclos estacionales están ligados al intervalo de la serie. Por ejemplo, los datos mensuales suelen mostrar un comportamiento cíclico a lo largo de trimestres y años. Una serie mensual puede mostrar un ciclo trimestral significativo con un mínimo en el primer trimestre o un ciclo anual con un pico en cada mes de diciembre. Se dice que las series con un ciclo estacional muestran **estacionalidad**.

Los patrones estacionales resultan útiles para obtener buenos ajustes y previsiones. Hay modelos ARIMA y de suavizado exponencial que capturan la estacionalidad.

## Ciclos no estacionales

Un **ciclo no estacional** es un patrón repetitivo y posiblemente impredecible de los valores de las series.

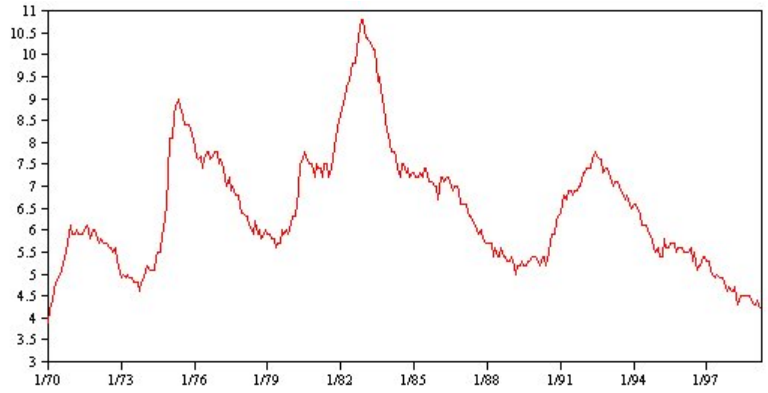



*Figura 55. Ciclo no estacional*

Algunas series, como la tasa de desempleo, muestran un claro comportamiento cíclico; no obstante, la periodicidad del ciclo varía a lo largo del tiempo, por lo que resulta difícil predecir cuándo se van a producir máximos o mínimos. Otras series pueden tener ciclos predecibles, pero no se ajustan exactamente al calendario gregoriano o tienen ciclos que se prolongan más de un año. Por ejemplo, las mareas siguen el calendario lunar, los viajes y el comercio internacionales relacionados con los Juegos Olímpicos aumentan cada cuatro años, y hay muchas festividades religiosas cuyas fechas gregorianas cambian de un año a otro.

Los patrones cíclicos no estacionales son difíciles de modelar y suelen aumentar la incertidumbre de las previsiones. El mercado de valores, por ejemplo, proporciona numerosos ejemplos de series que han desafiado el trabajo de los que hacen las previsiones. No obstante, los patrones no estacionales se deben tener en cuenta cuando existen. En muchos casos, aún así es posible identificar un modelo que se ajuste a los datos históricos razonablemente bien, lo que le ofrece una oportunidad excelente para minimizar la incertidumbre en las previsiones.

## Pulsos y pasos

Muchas series experimentan cambios bruscos de nivel. Normalmente son de dos tipos:

- Un cambio repentino y *temporal*, o **pulso**, en el nivel de la serie.
- Un cambio repentino y *permanente*, o **paso**, en el nivel de la serie.

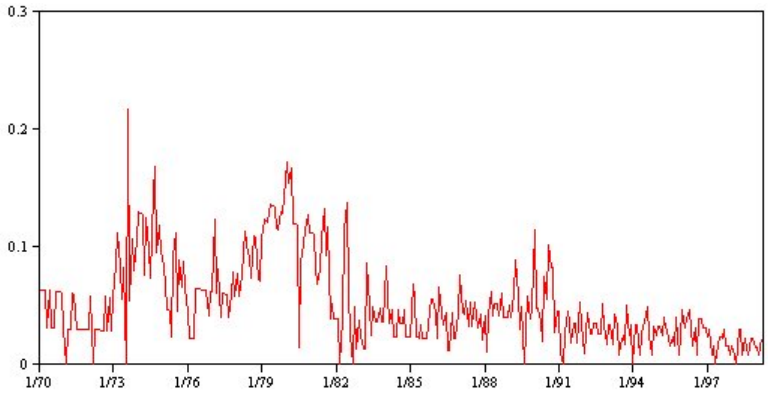



*Figura 56. Series con pulsos*

Cuando se observan pasos o pulsos, es importante encontrar una explicación convincente. Los modelos de series temporales están diseñados para explicar cambios graduales y no repentinos. Por tanto, suelen subestimar los pulsos y pueden quedar inutilizados por los pasos, lo que da como resultado modelos poco ajustados y previsiones imprecisas. (Es posible que algunos casos de estacionalidad parezcan presentar cambios repentinos de nivel, pero que el nivel sea constante de un período estacional a otro.)

Si puede explicarse una alteración, se puede modelar mediante una **intervención** o un **evento**. Por ejemplo, en agosto de 1973, la Organización de Países Exportadores de Petróleo (OPEP) impuso un embargo sobre el petróleo que cambió drásticamente la tasa de inflación, aunque recuperó sus niveles normales en los meses siguientes. Si especifica una **intervención por puntos** para el mes del embargo, puede mejorar el ajuste del modelo, lo que mejorará las previsiones indirectamente. Por ejemplo, puede que un comercio minorista descubra que sus ventas se incrementaron mucho más de lo normal un día que todos los artículos se rebajaron un 50%. Si se especifica una promoción de rebajas del 50% como **evento** recurrente, puede mejorar el ajuste del modelo y estimar la repercusión que tendría esa misma promoción en el futuro.

## Valores atípicos

Los desplazamientos en el nivel de una serie temporal que no se pueden explicar se denominan **valores atípicos**. Estas observaciones no coinciden con el resto de las series y pueden influir considerablemente en el análisis y, por lo tanto, afectar a la capacidad de previsión del modelo de serie temporal.

En la siguiente figura se muestran los distintos tipos de valores atípicos que se producen normalmente en las series temporales. Las líneas azules representan una serie sin valores atípicos. Las líneas rojas sugieren un patrón que podría estar presente si la serie contuviera valores atípicos. Estos valores atípicos se clasifican todos como **deterministas** porque afectan únicamente al nivel de la media de la serie.

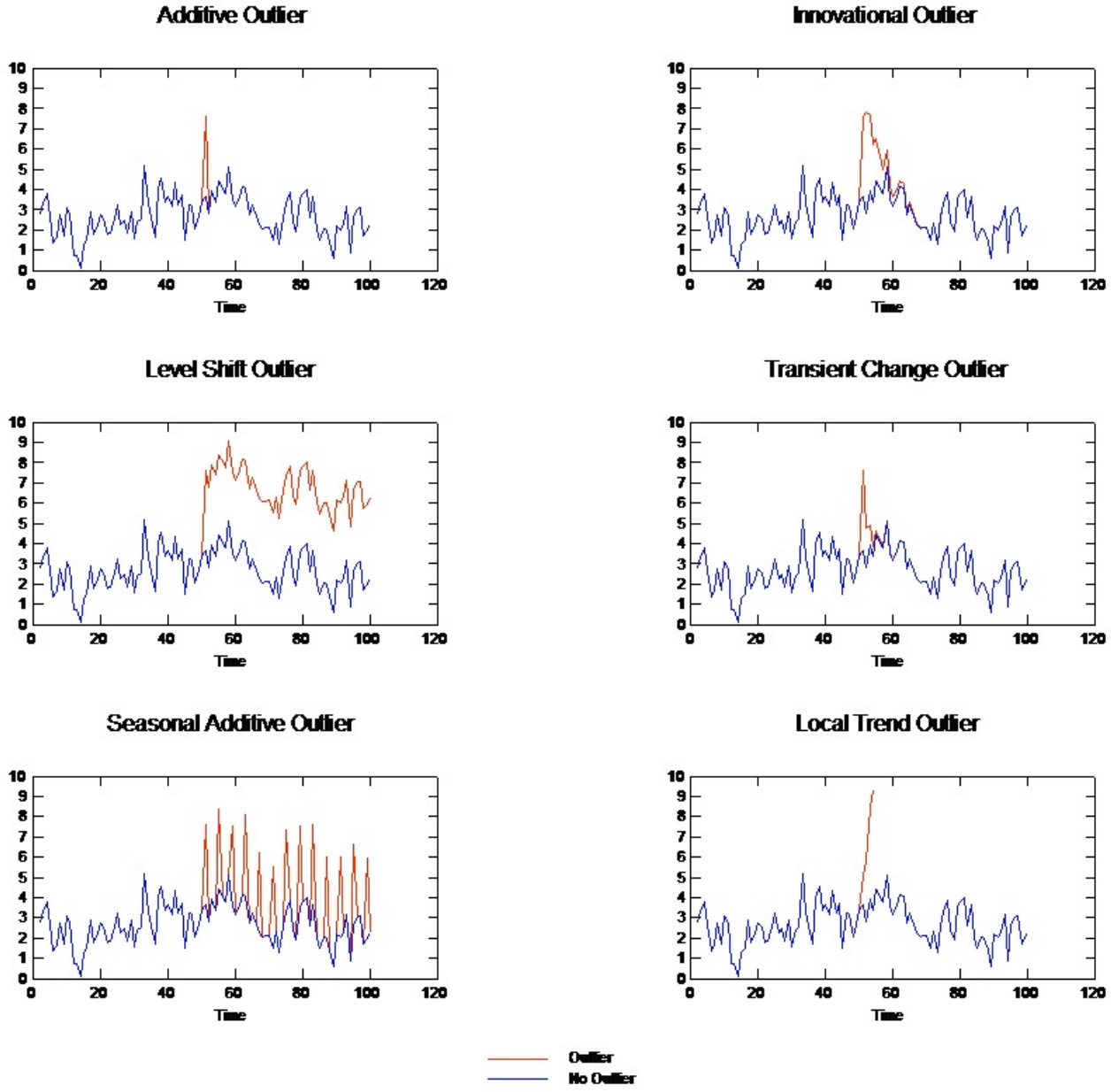



*Figura 57. Tipos de valor atípico*

- **Valor atípico aditivo.** Un valor atípico aditivo aparece como un valor inesperadamente alto o bajo que se produce para una única observación. Las siguientes observaciones no se ven afectadas por un valor atípico aditivo. Los valores atípicos adictivos consecutivos se denominan normalmente **parches de valores atípicos aditivos**.
- **Valor atípico innovador.** Un valor atípico innovador se caracteriza por un impacto inicial con efectos que se extienden sobre las siguientes observaciones. La influencia de los valores atípicos puede aumentar mientras avanza el tiempo.
- **Valor atípico de cambio de nivel.** En el cambio de nivel, todas las observaciones que aparecen después del valor atípico se desplazan a un nuevo nivel. A diferencia de los valores atípicos aditivos, un valor atípico de cambio de nivel afecta a diversas observaciones y tiene un efecto permanente.

- **Valor atípico de cambio transitorio.** Los valores atípicos de cambio transitorio son similares a los valores atípicos de cambio de nivel, pero su efecto se reduce exponencialmente en las siguientes observaciones. Finalmente, las series vuelven a su nivel normal.
- **Valor atípico aditivo estacional.** Un valor atípico aditivo estacional aparece como un valor inesperadamente alto o bajo que se produce repetidamente en intervalos regulares.
- **Valor atípico de tendencia local.** Un valor atípico de tendencia local produce un cambio general en la serie causado por un patrón en los valores atípicos después de la aparición del valor atípico inicial.

La detección de valores atípicos en una serie temporal implica determinar la ubicación, tipo y magnitud de todos los valores atípicos presentes. Tsay (1998) propuso un procedimiento iterativo para detectar el cambio del nivel de la media con el fin de identificar los valores atípicos deterministas. Este proceso implica la comparación de un modelo de serie temporal que supone que no hay presentes valores atípicos con otro modelo que incorpore valores atípicos. Las diferencias entre modelos permiten calcular el efecto de tratar cualquier punto como un valor atípico.

## Funciones de autocorrelación y autocorrelación parcial

La autocorrelación y la autocorrelación parcial son medidas de asociación entre valores de series actuales y pasadas e indican cuáles son los valores de series pasadas más útiles para predecir valores futuros. Con estos datos podrá determinar el orden de los procesos en un modelo ARIMA. Más concretamente,

- **Función de autocorrelación (FAS).** En el retardo *k*, es la autocorrelación entre los valores de las series que se encuentran a *k* intervalos de distancia.
- **Función de autocorrelación parcial (FAP).** En el retardo *k*, es la autocorrelación entre los valores de las series que se encuentran a *k* intervalos de distancia, teniendo en cuenta los valores de los intervalos intermedios.

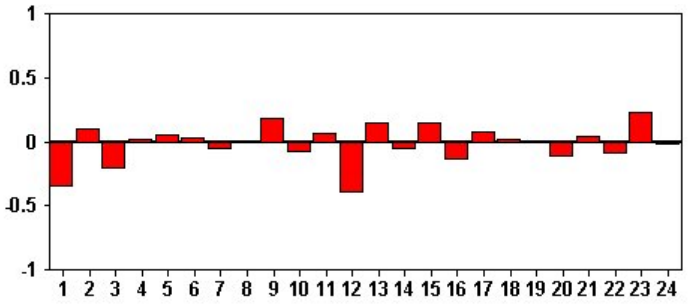



*Figura 58. Gráfico de FAS de una serie*

El eje *x* del gráfico de FAS indica el retardo en el que se calcula la autocorrelación; el eje *y* indica el valor de la correlación (entre -1 y 1). Por ejemplo, un trazo de unión en el retardo 1 de un gráfico de FAS indica que existe una fuerte correlación entre el valor de cada serie y el valor anterior, un trazo de unión en el retardo 2 indica que existe una fuerte correlación entre el valor de cada serie y el valor que aparece dos puntos anteriores, etc.

- Una correlación positiva indica que los valores grandes actuales se corresponden con valores grandes en el retardo especificado; una correlación negativa indica que los valores grandes actuales se corresponden con valores pequeños en el retardo especificado.
- El valor absoluto de una correlación es una medida de la fuerza de la asociación, con valores absolutos mayores que indican relaciones más fuertes.

## Transformaciones de series

Las transformaciones suelen ser útiles para estabilizar una serie antes de estimar modelos. Esto es especialmente importante para modelos ARIMA, que necesitan que las series sean **estacionarias** antes de estimar los modelos. Una serie es estacionaria si el nivel global (media) y la desvianza media del nivel (varianza) son constantes a lo largo de la serie.

Aunque la mayoría de las series interesantes no son estacionarias, ARIMA es eficaz siempre y cuando la serie se pueda convertir en estacionaria mediante la aplicación de transformaciones tales como el logaritmo natural, la diferenciación o la diferenciación estacional.

**Transformaciones de estabilización de la varianza.** Las series en las que la varianza cambia a lo largo del tiempo con frecuencia se pueden estabilizar con una transformación logarítmica natural o de raíz cuadrada. También reciben el nombre de transformaciones funcionales.

- **Log natural.** El logaritmo natural se aplica a los valores de las series.
- **Raíz cuadrada.** La función de raíz cuadrada se aplica a los valores de las series.

No se pueden usar las transformaciones logarítmica natural o de raíz cuadrada para series con valores negativos.

**Transformaciones de estabilización del nivel.** Un suave descenso de los valores de la FAS indica que todos los valores de la serie están estrechamente correlacionados con el valor anterior. Si analiza el cambio de los valores de la serie, obtendrá un nivel estable.

- **Diferenciación simple.** Se calculan las diferencias existentes entre cada valor y el anterior de la serie, a excepción del valor más antiguo de la serie. Por tanto, la serie diferenciada tendrá un valor menos que la serie original.
- **Diferenciación estacional.** Es idéntica a la diferenciación simple, excepto en que se calculan las diferencias existentes entre cada valor y el valor estacional anterior.

Si se usa la diferenciación simple o estacional de forma simultánea con la transformación logarítmica o de raíz cuadrada, siempre se aplicará primero la transformación de estabilización de la varianza. Si se usan la diferenciación simple y estacional, los valores de la serie resultante son iguales independientemente de si se aplica primero una diferenciación u otra.

## Serie predictora

La serie predictora contiene datos relacionados que pueden ayudar a explicar el comportamiento de la serie para la que se van a realizar previsiones. Por ejemplo, un minorista de venta por catálogo o por Internet podría prever el número de ventas en función del número de catálogos enviados, el número de líneas telefónicas abiertas o el número de entradas a la página Web de su empresa.

Cualquier serie puede utilizarse como un predictor siempre que se extienda en el tiempo que desea prever y tenga los datos completos, sin valores perdidos.

Tenga cuidado al añadir predictores a un modelo, ya que añadir un gran número de predictores aumentará el tiempo necesario para calcular los modelos. Aunque añadir predictores puede mejorar la capacidad del modelo para ajustarse a los datos históricos, no significa necesariamente que el modelo vaya a realizar una mejor previsión, por lo que una mayor complejidad puede no valer la pena. Lo ideal sería identificar el modelo más simple que mejores previsiones realice.

Como norma general, se recomienda que el número de predictores sea inferior al tamaño de la muestra divido entre 15 (como mucho, un predictor por 15 casos).

**Predictores con datos perdidos.** Los predictores con datos incompletos o perdidos no pueden utilizarse para la previsión. Esto es aplicable tanto a los datos históricos como a los valores futuros. En algunos

casos, puede evitar esta limitación mediante la configuración de la amplitud de estimación del modelo para excluir los datos más antiguos a la hora de calcular los modelos.

## Nodo de modelado Predicción espacio-temporal

La predicción espacio-temporal (STP) tiene muchas posibles aplicaciones, tales como la gestión de energía para edificios o instalaciones, el análisis y previsión del rendimiento para ingenieros de servicios mecánicos o la planificación del transporte público. En estas aplicaciones, las mediciones, tales como el uso de la energía, a menudo se obtienen a lo largo del espacio y el tiempo. Las cuestiones que pueden ser importantes para el registro de estas mediciones incluyen determinar qué factores afectarán a las observaciones futuras, y qué se puede hacer para producir un cambio deseado o gestionar mejor el sistema. Para abordar estas cuestiones, puede utilizar técnicas estadísticas para predecir valores futuros en ubicaciones diferentes, y puede ajustar explícitamente factores para realizar análisis de hipótesis.

El análisis de STP utiliza datos de ubicación, campos de entrada para predicciones (predictores), un campo de tiempo y un campo de destino. Cada ubicación tiene muchas filas de datos que representan los valores de cada predictor para cada tiempo de medición. Después de analizar los datos, se pueden utilizar para predecir valores en cualquier ubicación dentro de los datos de forma utilizados en el análisis. El análisis de STP también puede realizar previsiones cuando se conocen los datos de entrada para puntos específicos en el tiempo.

**Nota:** El nodo STP no soporta los pasos Evaluación de modelo o Ganador-comparativo en IBM SPSS Collaboration and Deployment Services.

El directorio `Demos` de la instalación de IBM SPSS Modeler contiene un ejemplo de utilización de STP denominado `server_demo.str`. El ejemplo utiliza los archivos de datos `room_data.csv` y `score_data.csv`. Puede acceder al directorio `Demos` desde el grupo de programas de IBM SPSS Modeler en el menú Inicio de Windows. El archivo `server_demo.str` está situado en el directorio `streams`.

## Predicción espacio-temporal - Opciones de campos

En la pestaña Campos, puede seleccionar si desea utilizar la configuración de rol de campo ya definida en nodos anteriores o realizar las asignaciones de campos manualmente.

**Utilizar roles predefinidos**
Esta opción utiliza las definiciones de roles (solo objetivos y predictores) desde un nodo Tipo anterior (o la pestaña Tipo de un nodo de origen anterior).

**Utilizar asignaciones de campos personalizadas**
Para asignar objetivos, predictores y otros roles manualmente en esta pantalla, seleccione esta opción.

**Campos**
Muestra todos los campos en los datos que se pueden seleccionar. Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los diferentes recuadros en la parte derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo.

**Nota:** STP requiere 1 registro por ubicación, por intervalo de tiempo para funcionar correctamente; por lo tanto, éstos son campos obligatorios.

En la parte inferior del panel **Campos**, pulse en el botón **Todos** para seleccionar todos los campos, independientemente del nivel de medición, o pulse en un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

**Objetivo**
Seleccione un campo como el objetivo de la predicción.

**Nota:** Sólo puede seleccionar campos con un nivel de medición de continuo.

**Ubicación**
Seleccione el tipo de ubicación que se utilizará en el modelo.

**Nota:** Sólo puede seleccionar campos con un nivel de medición de geoespacial.

**Etiqueta de ubicación**
Los datos de forma a menudo incluyen un campo que muestra los nombres de las características en la capa, por ejemplo, pueden ser los nombres de estados o países. Utilice este campo para asociar un nombre, o una etiqueta, con una ubicación seleccionando un campo categórico para etiquetar el campo **Ubicación** elegido en la salida.

**Campo de tiempo**
Seleccione los campos de tiempo a utilizar en sus predicciones.

**Nota:** Puede seleccionar campos sólo con un nivel de medición de continuo y un tipo de almacenamiento de hora, fecha, indicación de fecha y hora o entero.

**Predictores (entradas)**
Seleccione uno o más campos como entradas de la predicción.

**Nota:** sólo puede seleccionar campos con un nivel de medición de continuo.

## Predicción espacio-temporal - Intervalos temporales

En el panel Intervalos de tiempo puede seleccionar las opciones para establecer el intervalo de tiempo y la agregación necesaria a lo largo del tiempo.

Es necesario preparar los datos para convertir los campos de tiempo en un índice antes de poder generar un modelo STP; para que la conversión sea posible, el campo de tiempo debe tener un intervalo constante entre los registros. Si sus datos no contienen ya esta información, use las opciones de este panel para establecer este intervalo antes de poder usar el nodo de modelado.

**Intervalo de tiempo** Seleccione el intervalo al que desee que se convierta el conjunto de datos. Las opciones disponibles dependen del tipo de almacenamiento del campo elegido como **Campo de hora** para el modelo en la pestaña Campos.

- **Periodos** Solo disponible para los campos de tiempo en valores enteros; es una serie de intervalos con un intervalo uniforme entre cada medición, que no coincide con ninguno de los otros intervalos.
- **Años** Solo disponible para los campos temporales Fecha o Marca de tiempo.
- **Trimestres** Solo disponible para los campos temporales Fecha o Marca de tiempo. Si selecciona esta opción, se le solicitará que seleccione el **Mes inicial** del primer trimestre.
- **Meses** Solo disponible para los campos temporales Fecha o Marca de tiempo.
- **Semanas** Solo disponible para los campos temporales Fecha o Marca de tiempo.
- **Días** Solo disponible para los campos temporales Fecha o Marca de tiempo.
- **Horas** Solo disponible para los campos temporales Fecha o Marca de tiempo.
- **Minutos** Solo disponible para los campos temporales Fecha o Marca de tiempo.
- **Segundos** Solo disponible para los campos temporales Fecha o Marca de tiempo.

Cuando se selecciona **Intervalo de tiempo**, se le solicitará que complete más campos. Los campos disponibles dependen del intervalo de tiempo y el tipo de almacenamiento. Los campos que pueden mostrarse aparecen en la lista siguiente.

- **Número de días por semana**
- **Número de horas de un día**
- **La semana comienza el** El primer día de la semana
- **El día comienza a las** Hora en la que se considera que empieza un nuevo día.

- **Valor del intervalo** Puede elegir una de las opciones siguientes: 1, 2, 3, 4, 5, 6, 10, 12, 15, 20 o 30.
- **Mes inicial** El mes en el que empieza el nuevo año fiscal.
- **Periodo inicial** Si utiliza **Periodos**, seleccione el periodo inicial.

**Los datos coinciden con el intervalo temporal especificado** Si sus datos ya contienen información correcta sobre intervalos temporales y no es necesario convertirlos, seleccione esta casilla de verificación. Cuando se selecciona esta casilla, los campos del área **Agregación** no están disponibles.

### Agregación

Solo disponible si se desmarca la casilla de verificación **Los datos coinciden con el intervalo temporal especificado**; especifique las opciones para agregar campos para que coincidan con el intervalo indicado. Por ejemplo, si tiene una mezcla de datos semanales y mensuales, puede agregar o "acumular" los valores semanales con el fin de obtener un intervalo mensual uniforme. Seleccione la configuración predeterminada para usarla como agregación de diversos tipos de campos y cree la configuración personalizada que desee para los campos específicos.

- **Continuo** Configure el método de agregación predeterminado para aplicarlo a todos los campos continuos que no se indiquen individualmente. Puede elegir entre varios métodos:
  - Suma
  - Media
  - Mínimo
  - Máximo
  - Mediana
  - Primer cuartil
  - Tercer cuartil

**Valores personalizados para campos especificados** Para aplicar una función de agregación especial a campos individuales, selecciónelos en esta tabla y elija el método de agregación.

- **Campo** Use el botón **Añadir campo** para mostrar el cuadro de diálogo Seleccionar campos y elija los campos necesarios. Los campos elegidos se muestran en esta columna.
- **Función de agregación** En la lista desplegable, seleccione la función de agregación para convertir el campo al intervalo de tiempo especificado.

## Predicción espacio-temporal - Opciones básicas de generación

Utilice los valores de este cuadro de diálogo para establecer las opciones básicas de generación de modelos.

### Configuración del modelo

**Incluir ordenada en el origen**

Incluir la interceptación (el término constante en el modelo) puede aumentar la precisión global de la solución. Si se puede dar por supuesto la lectura de datos en el origen, se puede excluir la interceptación.

**Órdenes autorregresivos máximos**

Los órdenes autorregresivos especifican los valores previos utilizados para predecir los valores actuales. Utilice esta opción para especificar el número de registros anteriores que se utilizan para calcular un valor nuevo. Puede elegir cualquier entero entre 1 y 5.

### Covarianza espacial

**Método de estimación**
Seleccione el método de estimación que se va a utilizar; puede elegir **Parámetrico** o **No paramétrico**. Para el método **Parámetrico** puede elegir entre uno de los tres tipos de **Modelo**:

- **Gauss**
- **Exponential**
- **Powered Exponential** Si selecciona esta opción, también debe especificar el nivel de **Potencia** a utilizar. Este nivel puede ser cualquier valor entre 1 y 2, en incrementos de 0,1.

## Predicción espacio-temporal - Opciones avanzadas de generación

Los usuarios con conocimiento detallado de STP pueden utilizar las opciones siguientes para ajustar el proceso de generación de modelos.

**Porcentaje máximo de valores perdidos**
Especifique el porcentaje máximo de registros que contienen valores que faltan que pueden incluirse en el modelo.

**Nivel de significación para prueba de hipótesis en generación de modelos**
Especifique el valor de nivel de significación que se va a utilizar para todas las pruebas de estimación de modelo STP, incluyendo dos pruebas de bondad de ajuste, pruebas F de efecto y pruebas T de coeficiente. Este nivel puede ser cualquier valor de 0 a 1, en incrementos de 0,01.

## Predicción espacio temporal: datos de salida

Antes de crear el modelo, utilice las opciones de este panel para seleccionar los datos de salida que desee incluir en el visor de la salida.

### Información del modelo

**Especificaciones de modelos**
Seleccione esta opción para incluir información de especificación del modelo en los datos de salida del modelo.

**Resumen de información temporal**
Seleccione esta opción para incluir un resumen de información temporal en los datos de salida del modelo.

### Evaluación

**Calidad del modelo**
Seleccione esta opción para incluir la calidad del modelo en los datos de salida del modelo.

**Prueba de efectos en el modelo de estructura de media**
Seleccione esta opción para incluir información de prueba de efectos en los datos de salida del modelo.

### Interpretación

**Coeficientes de modelo de estructura media**
Seleccione esta opción para incluir información de coeficientes de modelo de estructura media en los datos de salida del modelo.

**Coeficientes autorregresivos**
Seleccione esta opción para incluir información de coeficientes autorregresivos en los datos de salida del modelo.

**Pruebas de extinción en el espacio**
Seleccione esta opción para incluir información de la prueba de covarianza espacial, o prueba de extinción en el espacio, en los datos de salida del modelo.

**Representación de parámetros del modelo de covarianza espacial paramétrico**
Seleccione esta opción para incluir información sobre la representación de parámetros del modelo de covarianza espacial paramétrico en los datos de salida del modelo.

**Nota:** esta opción sólo está disponible si ha seleccionado el método de estimación **Paramétrica** en el panel Aspectos básicos.

**Mapa de calor de correlaciones**
Seleccione esta opción para incluir un mapa de los valores de destino en los datos de salida del modelo.

**Nota:** si existen más de 500 ubicaciones en el modelo, el mapa no se creará en los datos de salida.

**Mapa de correlaciones**
Seleccione esta opción para incluir un mapa de correlaciones en los datos de salida del modelo.

**Nota:** si existen más de 500 ubicaciones en el modelo, el mapa no se creará en los datos de salida.

**Clústeres de ubicaciones**
Seleccione esta opción para incluir información de agrupación de ubicaciones en los datos de salida del modelo. La información de salida sobre agrupaciones en clúster sólo incluirá datos que no necesitan acceso a datos de mapa.

**Nota:** estos datos de salida sólo se pueden crear para un modelo de covarianza espacial no parámetrico.

Si selecciona esta opción, puede definir lo siguiente:

- **Umbral de similitud** Seleccione el valor umbral para el cual los clústeres de salida se considera que son lo suficientemente semejantes para ser fusionados y crear un clúster individual.
- **Número máximo de clústeres para visualizar** Defina el límite superior para el número de clústeres que se pueden incluir en los datos de salida del modelo.

## Opciones del modelo de predicción espacio temporal

**Nombre de modelo** Puede generar el nombre del modelo automáticamente, basándose en los campos de destino, o especificar un nombre personalizado. El nombre generado automáticamente es el nombre del campo de destino.

**Factor de incertidumbre (%)** El factor de incertidumbre es un valor porcentual que representa el crecimiento de la incertidumbre cuando realiza una predicción. El límite superior e inferior de la incertidumbre de la predicción aumenta de acuerdo con este porcentaje en cada paso que se avanza hacia el futuro. Defina el factor de incertidumbre que se debe aplicar a los resultados del modelo. Esto definirá los límites superior e inferior para los valores predichos.

## Nugget del modelo de predicción espacio temporal

El nugget del modelo de predicción espacio temporal (modelo PST) muestra detalles del modelo en el panel Modelo, del Visor de la salida. Para obtener más información sobre cómo utilizar el visor, consulte la sección titulada "Utilización de los datos de salida" en la Guía del usuario de Modeler (ModelerUsersGuide.pdf).

El funcionamiento del modelo de predicción espacio temporal (modelo STP) crea varios campos nuevos con el prefijo $STP- tal como se muestra en la tabla siguiente.

*Tabla 26. Campos nuevos creados por el funcionamiento del modelo STP*

| Nombre de campo | Descripción |
|---|---|

*Tabla 26. Campos nuevos creados por el funcionamiento del modelo STP (continuación)*

| | |
|---|---|
| $STP-<Hora> | Campo de hora que se crea como parte de la creación del modelo. Los valores contenidos en el panel Intervalos de tiempo, de la pestaña Opciones de generación, determinan cómo se crea este campo.<br><br><Hora> es el nombre original del campo seleccionado como **Campo de hora** en el panel Campos.<br>**Nota:** este campo sólo se crea si ha convertido el **Campo de hora** original como parte de la creación del modelo. |
| $STP-<Destino> | Este campo contiene las predicciones para el valor de destino.<br><br><Destino> es el nombre del campo de **Destino** original correspondiente al modelo |
| $STPVAR-<Destino> | Este campo contiene los valores de VarianceOfPointPrediction.<br><br><Destino> es el nombre del campo de **Destino** original correspondiente al modelo |
| $STPLCI-<Destino> | Este campo contiene los valores de LowerOfPredictionInterval, es decir, el límite inferior de confianza.<br><br><Destino> es el nombre del campo de **Destino** original correspondiente al modelo |
| $STPUCI-<Destino> | Este campo contiene los valores de UpperOfPredictionInterval, es decir, el límite superior de confianza.<br><br><Destino> es el nombre del campo de **Destino** original correspondiente al modelo |

### Valores del modelo de predicción espacio temporal

Utilice el panel Valores para controlar el nivel de incertidumbre que considera aceptable en la operación de modelado.

**Factor de incertidumbre (%)** El factor de incertidumbre es un valor porcentual que representa el crecimiento de la incertidumbre cuando realiza una predicción. El límite superior e inferior de la incertidumbre de la predicción aumenta de acuerdo con este porcentaje en cada paso que se avanza hacia el futuro. Defina el factor de incertidumbre que se debe aplicar a los resultados del modelo. Esto definirá los límites superior e inferior para los valores predichos.

## Nodo TCM

Utilice este nodo para crear un modelo causal temporal (TCM).

## Modelos causales temporales

El modelado causal temporal intenta descubrir relaciones causales clave en datos de series temporales. En el modelado causal temporal, especifique un conjunto de series de objetivos y un conjunto de entradas candidato para estos objetivos. El procedimiento crea un modelo de serie temporal autorregresivo para cada objetivo y solo incluye las entradas que tienen una relación causal con el objetivo. Este enfoque difiere del modelado de serie temporal tradicional donde debe especificar explícitamente los predictores para una serie de objetivos. Puesto que el modelado causal temporal normalmente implica crear modelos para varias series temporal relacionadas, se hace referencia al resultado como un *sistema de modelo*.

En el contexto del modelado causal temporal, el término *causal* hace referencia a la causalidad Granger. En una serie temporal X se indica que la "causa Granger" provoca otra serie temporal Y si realiza una regresión para Y en términos de valores pasados de ambos resultados, X e y, en un modelo mejor para Y que realiza una regresión solo en los valores pasados en Y.

**Nota:** El nodo de modelado Causal temporal no da soporte a los pasos Evaluación del modelo o Ganador-Comparativo en IBM SPSS Collaboration and Deployment Services.

## Ejemplos

Los tomadores de decisiones empresariales pueden utilizar el modelado causal temporal para descubrir relaciones causales en un conjunto grande de métricas basadas en tiempo que describen el negocio. El análisis puede revelar unas pocas entradas controlables, que tienen el mayor impacto en los indicadores de rendimiento clave.

Los gestores de sistemas de TI grandes pueden utilizar el modelado causal temporal para detectar anomalías en un conjunto grande de métricas operativas interrelacionadas. El modelo causal permite ir más allá de la detección de anomalías y descubrir las causas principales más probables de las anomalías.

## Requisitos de campo

Debe haber al menos un objetivo. De forma predeterminada, no se utilizan los campos con un papel predefinido de Ninguno.

## Estructura de datos

El modelado causal temporal soporta dos tipos de estructuras de datos.

**Datos basados en columna**
: Para datos basados en columna, cada campo de serie temporal contiene los datos para una sola serie temporal. Esta estructura es la estructura tradicional de los datos de serie temporal, tal como se utiliza en el modelador de series temporales.

**Datos multidimensionales**
: Para los datos multidimensionales, cada campo de serie temporal contiene los datos para varias series temporales. Las series temporales separadas, en un campo particular, se identifican mediante un conjunto de valores de campos categóricos a los que se hace referencia como campos de *dimensión*. Por ejemplo, los datos de ventas para dos canales de ventas diferentes (minorista y web) se podrían almacenar en un único campo *ventas*. Un campo de dimensión que se llama *canal*, con los valores 'minorista' y 'web', identifica los registros que están asociados a cada uno de los dos canales de ventas.

**Nota:** Para generar un modelo causal temporal, necesita suficientes puntos de datos. El producto utiliza la restricción:

```
m>(L + KL + 1)
```

donde `m` es el número de puntos de datos, `L` es el número de retardos y `K` es el número de predictores. Asegúrese de que el conjunto de datos es suficientemente grande, de forma que el número de puntos de datos (`m`) satisface la condición.

## Series temporales para modelar

En la pestaña Campos, utilice el valor de **Serie temporal** para especificar la serie que desee incluir en el sistema de modelo.

Seleccione la opción para la estructura de datos que se aplica a sus datos. Para datos multidimensionales, pulse **Seleccionar dimensiones** para especificar los campos de dimensión. El orden especificado de los campos de dimensión define el orden en el cual aparecen en todos los diálogos y resultados posteriores. Utilice los botones de flecha hacia arriba y abajo en el subdiálogo Seleccionar dimensiones para reordenar los campos de dimensión.

Para datos basados en columnas, el término *serie* tiene el mismo significado que el término *campo*. Para datos multidimensionales, los campos que contienen series temporales se denominan campos *de métrica*. Una serie temporal, para datos multidimensionales, se define mediante un campo de métrica y un valor para cada uno de los campos de dimensión. Las consideraciones siguientes se aplican a ambos datos, basados en columna y multidimensionales.

- Las series especificadas como entradas candidato, o como objetivo y entrada, se consideran para la inclusión en el modelo de cada objetivo. El modelo para cada objetivo siempre incluye valores retardados del propio objetivo.
- Las series especificadas como entradas forzadas siempre se incluyen en el modelo de cada objetivo.
- Al menos, se debe especificar una serie como un objetivo o como un objetivo y también entrada.
- Cuando está seleccionado **Utilizar roles predefinidos**, los campos que tienen un rol de Entrada se establecen como entradas candidato. Ningún rol predefinido se correlaciona con una entrada forzada.

## Datos multidimensionales

Para datos multidimensionales, especifique campos de métrica y papeles asociados en una cuadrícula, donde cada fila de la cuadrícula especifica una única métrica y un único papel. De forma predeterminada, el sistema del modelo incluye series para todas las combinaciones de los campos de dimensión para cada fila de la cuadrícula. Por ejemplo, si hay dimensiones para *región* y *marca*, de forma predeterminada, especificar la métrica *ventas* como un objetivo significa que hay una serie de objetivos de ventas individual para cada combinación de *región* y *marca*.

Para cada fila de la cuadrícula, puede personalizar el conjunto de valores para cualquiera de los campos de dimensión pulsando el botón de puntos suspensivos para una dimensión. Esta acción abre el subdiálogo Seleccionar valores de dimensión. También puede añadir, suprimir o copiar filas de cuadrícula.

La columna **Recuento de series** muestra el número de conjuntos de valores de dimensión que se han especificado actualmente para la métrica asociada. El valor visualizado puede ser mayor que el número real de series (una serie por conjunto). Esta condición se produce cuando algunas de las combinaciones especificadas de valores de dimensión no corresponden a las series incluidas por la métrica asociada.

**Seleccionar valores de dimensión:** Para datos multidimensionales, puede personalizar los análisis especificando qué valores de dimensión aplicar a un campo de métrica particular con un papel determinado. Por ejemplo si *ventas* es un campo de métrica y *canal* es una dimensión con los valores 'minorista' y 'web,' puede especificar que ventas 'web' es una entrada y ventas 'minorista' es un destino. También puede especificar subconjuntos de dimensiones que se aplican a todos los campos de métrica utilizados en el análisis. Por ejemplo, si *región* es un campo de dimensión que indica región geográfica, puede limitar el análisis a regiones particulares.

**Todos los valores**
: Especifica que todos los valores del campo de dimensión actual se incluyen. Esta opción es la predeterminada.

**Seleccionar valores que se van a incluir o excluir**
: Utilice esta opción para especificar el conjunto de valores para el campo de dimensión actual. Cuando está seleccionado **Incluir** para el **Modo**, solo se incluyen los valores que se han especificado en la lista **Valores seleccionados**. Cuando está seleccionado **Excluir** para el **Modo**, se incluyen todos los valores distintos a los valores que se han especificado en la lista **Valores seleccionados**.

  Puede filtrar el conjunto de valores entre los que elegir. Los valores que cumplen la condición de filtro aparecen en la pestaña **Coincidentes** y los valores que no cumplen la condición de filtro aparecen en la pestaña **No coincidentes** de la lista **Valores no seleccionados**. La pestaña **Todos** lista todos los valores no seleccionados, independientemente de cualquier condición de filtro.

  - Puede utilizar asteriscos (*) para indicar caracteres comodín cuando especifique un filtro.
  - Para borrar el filtro actual, especifique un valor vacío para el término de búsqueda en el diálogo Filtrar valores visualizados.

## Observaciones

En la pestaña Campos, utilice el valor **Observaciones** para especificar los campos que definen las observaciones.

**Observaciones que se definen por fecha/horas**

Puede especificar que las observaciones se definen mediante un campo de fecha, hora o fecha/hora. Además del campo que define las observaciones, seleccione el intervalo de tiempo apropiado que describe las observaciones. En función del intervalo de tiempo especificado, también puede especificar otros valores como, por ejemplo, el intervalo entre observaciones (incremento) o el número de días por semana. Las consideraciones siguientes se aplican al intervalo de tiempo:

- Utilice el valor **Irregular** cuando las observaciones se asignan de forma irregular en el tiempo como, por ejemplo, la hora a la que se procesa un pedido de compra. Si está seleccionado **Irregular**, debe especificar el intervalo de tiempo que se utiliza para el análisis, desde la configuración de **Intervalo de tiempo** en la pestaña Especificaciones de datos.
- Cuando las observaciones representan una fecha y una hora y el intervalo de tiempo es horas, minutos o segundos, utilice **Horas del día**, **Minutos del día** o **Segundos del día**. Cuando las observaciones representan un periodo de tiempo (duración) sin referencia a una fecha y el intervalo de tiempo es horas, minutos o segundos, utilice **Horas (no periódico)**, **Minutos (no periódico)** o **Segundos (no periódico)**.
- Basándose en el intervalo de tiempo seleccionado, el procedimiento puede detectar observaciones que faltan. La detección de observaciones que faltan es necesaria porque el procedimiento presupone que todas las observaciones se han espaciado de forma uniforme en el tiempo y que no falta ninguna observación. Por ejemplo, si el intervalo de tiempo es días y la fecha 2014-10-27 viene seguida por 2014-10-29, falta una observación para 2014-10-28. Los valores se imputan para cualquier observación que falta. La configuración para manejar los valores que faltan se puede especificar en la pestaña Especificaciones de datos.
- El intervalo de tiempo especificado permite al procedimiento detectar varias observaciones en el mismo intervalo de tiempo que se deben agregar juntas y alinear observaciones en un límite de intervalo como, por ejemplo, el primer día del mes, para garantizar que las observaciones se espacian de forma uniforme. Por ejemplo, si el intervalo de tiempo es Meses, varias fechas del mismo mes se pueden agregar juntas. Se hace referencia a este tipo de agregación como *agrupación*. De forma predeterminada, las observaciones se suman cuando se agrupan. Puede especificar un método diferente para la agrupación, como la media de las observaciones, desde la configuración **Agregación y distribución** en la pestaña Especificaciones de datos.
- Para algunos intervalos de tiempo, los valores adicionales pueden definir saltos en los intervalos normales espaciados de forma uniforme. Por ejemplo, si el intervalo de tiempo es Días, pero solo son válidos los fines de semana, puede especificar que hay cinco días en una semana y que la semana empieza el lunes.

**Observaciones que se definen mediante periodos o periodos cíclicos**

Las observaciones se pueden definir mediante uno o más campos de enteros que representan periodos o ciclos repetitivos de periodos, hasta un número arbitrario de niveles de ciclo. Con esta estructura, puede describir una serie de observaciones que no caben en uno de los intervalos de tiempo estándar. Por ejemplo, un año fiscal con solo 10 meses se puede describir con un campo de ciclo que representa años y un capo de periodo que representa meses, donde la longitud de un ciclo es 10.

Los campos que especifican periodos cíclicos definen una jerarquía de niveles periódicos, donde el nivel más bajo se define mediante el campo **Periodo**. El siguiente nivel más alto se especifica mediante un campo de ciclo cuyo nivel es 1, seguido de un campo de ciclo cuyo nivel es 2 y, así, sucesivamente. Los valores de campo para cada nivel, excepto el más alto, deben ser periódicos con respecto al siguiente nivel superior. Los valores para el nivel superior no pueden ser periódicos. Por ejemplo, en el caso de un año fiscal de 10 meses, los meses son periódicos dentro de los años y los años no son periódicos.

- La longitud de un ciclo en un nivel particular es la periodicidad del siguiente nivel inferior. Para el ejemplo del año fiscal, solo hay un nivel de ciclo y la longitud del ciclo es 10 porque el siguiente nivel inferior representa meses y hay 10 meses en el año fiscal especificado.
- Especifique el valor inicial para cualquier campo periódico que no empiece desde el 1.Este valor es necesario para detectar valores que faltan. Por ejemplo, si un campo periódico empieza en el 2, pero el valor inicial se especifica como 1, el procedimiento supone que falta un valor para el primer periodo de cada ciclo de dicho campo.

## Intervalo temporal para el análisis

El intervalo de tiempo que se utiliza para el análisis puede diferir del intervalo de tiempo de las observaciones. Por ejemplo, si el intervalo de tiempo de las observaciones es Días, podría elegir Meses para el intervalo de tiempo del análisis. Los datos se añaden después a partir de datos de diarios a mensuales antes de que se genere el modelo. También puede elegir distribuir los datos de un intervalo de tiempo más largo a uno más corto. Por ejemplo, si las observaciones son trimestrales, puede distribuir los datos de datos trimestrales a datos mensuales.

Las opciones disponibles para el intervalo de tiempo en el que se realiza el análisis dependen de cómo se han definido las observaciones y del intervalo de tiempo de estas observaciones. En particular, cuando las observaciones se han definido mediante periodos cíclicos, solo se da soporte a la agregación. En este caso, el intervalo de tiempo del análisis debe ser mayor o igual que el intervalo de tiempo de las observaciones.

El intervalo de tiempo para el análisis se especifica desde la configuración de **Intervalo de tiempo** en la pestaña Especificaciones de datos. El método mediante el cual los datos se agregan o distribuyen se especifica en la configuración de **Agregación y distribución** en la pestaña Especificaciones de datos.

## Agregación y distribución

**Funciones de agregación**

Cuando el intervalo de tiempo que se utiliza para el análisis es más largo que el intervalo de tiempo de las observaciones, se agregan los datos de entrada. Por ejemplo, la agregación se realiza cuando el intervalo de tiempo de las observaciones es Días y el intervalo de tiempo para el análisis es Meses. Están disponibles las funciones de agregación siguientes: media, suma, moda, mín o máx.

**Funciones de distribución**

Cuando el intervalo de tiempo que se utiliza para el análisis es más corto que el intervalo de tiempo de las observaciones, se distribuyen los datos de entrada. Por ejemplo, la distribución se realiza cuando el intervalo de tiempo de las observaciones es Trimestres y el intervalo de tiempo para el análisis es Meses. Están disponibles las funciones de distribución siguientes: media o suma.

**Funciones de agrupación**

La agrupación se aplica cuando las observaciones se definen mediante fecha/horas y se producen varias observaciones en el mismo intervalo de tiempo. Por ejemplo, si el intervalo de tiempo de las observaciones es Meses, se agrupan varias fechas del mismo mes y se asocian al mes en el cual se producen. Están disponibles las funciones de agrupación siguientes: media, suma, moda, mín o máx. La agrupación siempre se realiza cuando las observaciones se definen por fecha/horas y el intervalo de tiempo de las observaciones se especifica como Irregular.

**Nota:** Aunque la agrupación es una forma de agregación, se realiza antes de manejar los valores que faltan, mientras que la agregación formal se realiza después de que se manejen los valores que faltan. Cuando el intervalo de tiempo de las observaciones se especifica como irregular, la agregación solo se realiza con la función de agrupación.

**Agregar observaciones del día al día anterior**

Especifica si las observaciones con horas que pasan el límite de un día se agregan a los valores del día anterior. Por ejemplo para las observaciones por hora con un día de ocho horas que

empieza a las 20:00, este valor especifica si las observaciones entre 00:00 y 04:00 se incluyen en los resultados agregados para el día anterior. Este valor se aplica solo si el intervalo de tiempo de las observaciones es Horas por día, Minutos por día o Segundos por día y el intervalo de tiempo para el análisis es Días.

**Valores personalizados para campos especificados**
Puede especificar funciones de agregación, distribución y agrupación en un campo por campo. Estos valores alteran temporalmente los valores predeterminados para las funciones de agregación, distribución y agrupación.

## Valores que faltan

Los valores que faltan en los datos de entrada se sustituyen con un valor imputado. Están disponibles los métodos de sustitución siguientes:

**Interpolación lineal**
Sustituye los valores que faltan utilizando una interpolación lineal. Se utilizan para la interpolación el último valor válido antes del valor perdido y el primer valor válido después del valor perdido. Si la primera o la última observación de la serie tiene un valor que falta, se utilizan los dos valores que no faltan más cercanos al principio o al final de la serie.

**Media de la serie**
Sustituye los valores perdidos con la media de la serie completa.

**Media de los puntos adyacentes**
Sustituye los valores perdidos por la media de los valores válidos circundantes. La amplitud de los puntos adyacentes es el número de valores válidos anteriores y posteriores del valor que falta que se utilizan para calcular la media.

**Mediana de puntos adyacentes**
Sustituye los valores perdidos por la mediana de los valores válidos circundantes. La amplitud de los puntos adyacentes es el número de valores válidos anteriores y posteriores del valor que falta que se utilizan para calcular la mediana.

**Tendencia lineal**
Esta opción utiliza todas las observaciones que no faltan en la serie para ajustar un modelo de regresión lineal simple, que se utiliza para imputar los valores que faltan.

Otros valores:

**Porcentaje máximo de valores perdidos (%)**
Especifica el porcentaje máximo de valores perdidos que están permitidos para cualquier serie. Las series con más valores perdidos que el máximo especificado se excluyen del análisis.

## Opciones de datos generales

**Número máximo de valores distintos por campo de dimensión**
Este valor se aplica a datos multidimensionales y especifica el número máximo de valores distintos que están permitios para cualquier campo de dimensión. De forma predeterminada, este límite se establece en 10000, pero se puede aumentar a un número arbitrariamente grande.

## Opciones generales de generación

**Ancho de intervalo de confianza (%)**
Este valor controla los intervalos de confianza para ambos parámetros, de previsiones y de modelo. Puede especificarse cualquier valor positivo inferior a 100. De forma predeterminada, se utiliza un intervalo de confianza de 95 %.

**Número máximo de entradas para cada objetivo**
Este valor especifica el número máximo de entradas que están permitidas en el modelo para cada objetivo. Puede especificar un entero dentro del rango de 1 a 20. El modelo para cada objetivo siempre incluye valores retardados de sí mismo, de modo que establecer este valor en 1 especifica que la entrada solo es el objetivo propio.

**Tolerancia del modelo**
Este valor controla el proceso iterativo que se utiliza para determinar el mejor conjunto de entradas para cada objetivo. Se puede especificar cualquier valor mayor que cero. El valor predeterminado es 0,001. La tolerancia del modelo es un criterio de parada para la selección del predictor. Puede afectar al número de predictores que se incluyen en el modelo final. Pero si un objetivo se puede predecir a sí mismo muy bien, otros predictores pueden no incluirse en el modelo final. Puede ser necesario algún ejercicio de prueba y error (por ejemplo, si tiene este valor establecido en un valor alto, puede intentar establecerlo en un valor inferior para ver si otros predictores se pueden incluir o no).

**Umbral de valor atípico (%)**
Una observación se marca como un valor atípico si la probabilidad, tal como se calcula a partir del modelo, que es un valor atípico excede este umbral. Puede especificar un valor dentro del rango de 50 a 100.

**Número de retardos de cada entrada**
Este valor especifica el número de términos de retardo de cada entrada en el modelo para cada objetivo. De forma predeterminada, el número de términos de retardo se determina automáticamente a partir del intervalo temporal usado en el análisis. Por ejemplo, si el intervalo temporal es de meses (con incremento de un mes), el número de retardos será 12. De forma opcional, se puede especificar explícitamente el número de retardos. El valor especificado debe ser un entero comprendido en el rango 1 - 20.

**Continuar estimación utilizando modelos existentes**
Si ya ha generado un modelo causal temporal, seleccione esta opción para reutilizar los valores de criterios que se han especificado para dicho modelo, en lugar de generar un modelo nuevo. De esta forma, puede ahorrar tiempo volviendo a estimar y generando una nueva predicción que se basa en los mismos valores de modelo como antes, pero utilizando datos más recientes.

## Serie para mostrar

Estas opciones especifican las series (de objetivos o entradas) para los cuales se visualiza el resultado. El contenido del resultado para las series especificadas se determina mediante el valor de **Opciones de resultado**.

**Mostrar objetivos asociados con modelos que mejor se ajustan**
De forma predeterminada, se visualiza el resultado para los objetivos que están asociados a los 10 modelos de mejor ajuste, según lo que determina el valor cuadrado R. Puede especificar un número fijo diferente de modelos de mejor ajuste o puede especificar un porcentaje de modelos de mejor ajuste. También puede elegir entre las siguientes medidas de bondad de ajuste:

**R cuadrado**
Medida de la bondad de ajuste de un modelo lineal; en ocasiones recibe el nombre de coeficiente de determinación. Es la proporción de la variación de la variable de destino explicada por el modelo. Puede tomar un valor entre 0 y 1. Un valor pequeño indica que el modelo no se ajusta bien a los datos.

**Raíz de la media cuadrática de los errores porcentuales**
Una medida de cuánto difieren los valores pronosticados por el modelo de los valores observados de la serie. Es independiente de las unidades utilizadas y, por tanto, se puede utilizar para comparar series con distintas unidades.

**Raíz de la media cuadrática de los errores**
La raíz cuadrada del error cuadrático promedio. Una medida de cuánto se desvía la serie dependiente del nivel pronosticado por el modelo, expresado en las mismas unidades que la serie dependiente.

**BIC** Criterio de información Bayesiano. Una medida para seleccionar y comparar modelos basados en el logaritmo de la verosimilitud -2 reducida. Los valores menores indican

modelos mejores. El BIC también "penaliza" modelos sobreparametrizados (modelos complejos con un gran número de entradas, por ejemplo), pero de forma más estricta que el AIC.

**AIC** Criterio de información de Akaike. Una medida para seleccionar y comparar modelos basados en el logaritmo de la verosimilitud -2 reducida. Los valores menores indican modelos mejores. El AIC "penaliza" modelos sobreparametrizados (modelos complejos con un gran número de entradas, por ejemplo).

**Especifique series individuales**

Puede especificar series individuales para los cuales desea un resultado.

- Para los datos basados en columna, especifique los campos que contienen las series que desea. EL orden de los campos especificados define el orden en el cual aparecen en el resultado.
- Para los datos multidimensionales, especifique una serie particular añadiendo una entrada a la cuadrícula para el campo de métrica que contiene la serie. Especifique los valores de los campos de dimensión que definen las series.
  - Puede especificar el valor para cada campo de dimensión directamente en la cuadrícula o puede seleccionar en la lista de valores de dimensión disponibles. Para seleccionar en la lista de valores de dimensión disponibles, pulse el botón de puntos suspensivos en la casilla para la dimensión que desee. Esta acción abre el subdiálogo Seleccionar valor de dimensión.
  - Puede buscar en la lista de valores de dimensión, en el subdiálogo Seleccionar valor de dimensión, pulsando el icono de los prismáticos y especificando un término de búsqueda. Los espacios se tratan como parte del término de búsqueda. Los asteriscos (*) en el término de búsqueda no indican caracteres comodín.
  - El orden de las series de la cuadrícula define el orden en el cual aparecen en el resultado.

Para ambos datos, los que se basan en columna y los multidimensionales, el resultado está limitado a 30 series. Este límite incluye series individuales (entradas u objetivos) que especifique y los objetivos que están asociados a los modelos de mejor ajuste. Las series especificadas de forma individual tienen prioridad sobre los objetivos que están asociados a los modelos de mejor ajuste.

## Opciones de salida

Estas opciones especifican el contenido del resultado. Las opciones del grupo **Objetivos de resultado** generan el resultado para los objetivos que están asociados a los modelos de mejor ajuste en los valores **Series para visualizar**. Las opciones del grupo **Resultado para series** generan el resultado para las series individuales que se han especificado en los valores **Series para visualizar**.

**Sistema de modelo global**

Muestra una representación gráfica de las relaciones causales entre series en el sistema del modelo. Las tablas tanto en las estadísticas de ajuste del modelo y los valores atípicos para los objetivos visualizado se incluyen como parte del elemento de resultado. Cuando esta opción está seleccionada en el grupo **Resultado para series**, se crea un elemento de resultado separado para cada serie individual que se ha especificado en los valores de **Series para visualizar**.

Las relaciones causales entre las series tienen un nivel de significación asociado, donde un nivel de significación más pequeño indica un conexión más significativa. Puede elegir ocultar las relaciones con un nivel de significación que es mayor que un valor especificado.

**Estadísticas y valores atípicos del ajuste de modelo**

Las tablas de estadísticas y valores atípicos del ajuste del modelo para las series de objetivos que se han seleccionado para su visualización. Estas tablas contienen la misma información que las tablas en la visualización de Sistema global de modelo. Estas tablas soportan todas las funciones estándar para tablas dinámicas y de edición.

**Efectos del modelo y parámetros del modelo**
Tablas de pruebas de efectos del modelo y parámetros de modelo para las series de objetivos que se han seleccionado para su visualización. Las pruebas de efectos del modelo incluyen la estadística F y un valor de significación asociado para cada entrada incluida en el modelo.

**Diagrama de impacto**

Muestra una representación gráfica de las relaciones causales entre una serie de interés y otras series a las que afecta o que le afectan. Las series que afectan a las series de interés se denominan *causes*. Al seleccionar **Efectos** se genera un diagrama de impactos que se inicializa para visualizar los efectos. Si se selecciona **Causas**, se genera un diagrama de impacto que se inicializa para visualizar causas. Al seleccionar **Causas y efectos** se generan dos diagramas de impactos separados, uno que se inicializa para las causas y uno que se inicializa para los efectos. Se puede alternar de forma interactiva entre causas y efectos en el elemento de resultado que muestra el diagrama de impacto.

Puede especificar el número de niveles de causas o efectos para mostrar, donde el primer nivel es simplemente la serie de interés. Cada nivel adicional muestra más causas o efectos indirectas de las series de interés. Por ejemplo, el tercer nivel de la visualización de efectos está formado por la series que contienen series del segundo nivel como una entrada directa. Las series del tercer nivel son afectadas directamente por las series de interés puesto que las series de interés es una entrada directa en las series del segundo nivel.

**Gráfico de series**
Gráficos de valores observados y pronosticados para las series de objetivos que se han seleccionado para la visualización. Cuando se solicitan previsiones, el gráfico también muestra los valores pronosticados y los intervalos de confianza para las previsiones.

**Gráfico de residuos**
Gráficos de los residuos del modelo para las series de objetivos que se han seleccionado para la visualización.

**Entradas superiores**
Gráficos para cada objetivo visualizado, a lo largo del tiempo, junto con las 3 entradas superiores del objetivo. Las entradas superiores son las entradas con el valor de significación menor. Para acomodar distintas escalas para las entradas y el objetivo, el eje Y representa la puntuación Z para cada serie.

**Tabla de previsiones**
Tablas de valores pronosticados e intervalos de confianza de estas previsiones para las series de objetivos que se han seleccionado para la visualización.

**Análisis de causa principal de valor atípico**
Determina qué series tienen la mayor probabilidad de ser la causa de cada valor atípico en una series de interés. El análisis de causa principal de valor atípico se realiza para cada serie de objetivos que se incluye en la lista de series individuales en el valor **Series para visualizar**.

**Resultado**

**Tabla y gráfico de valores atípicos interactivos**
Tabla y gráfico de valores atípicos y causas principales de estos valores atípicos para cada serie de interés. La tabla contiene una sola fila para cada valor atípico. El gráfico es un diagrama de impactos. Seleccionar una fila en la tabla resalta la vía de acceso, en el diagrama de impactos, de la serie de interés en la serie que es más probable que provoque el valor atípico asociado.

**Tabla dinámica de valores atípicos**
Tabla de valores atípico y causas principales de estos valores atípico para cada series de interés. Esta tabla contiene la misma información que la tabla en la visualización interactiva. Esta tabla soporta todas las funciones estándar para tablas dinámicas y de edición.

**Niveles causales**
Puede especificar el número de niveles para incluir en la búsqueda para las causas principales. El concepto de niveles que se utiliza aquí es el mismo que el que se describe para diagramas de impactos.

**Ajuste de modelo en todos los modelos**
Histograma del ajuste del modelo para todos los modelos y paras las estadísticas de ajuste seleccionadas. Están disponibles las estadísticos de ajuste siguientes:

**R cuadrado**
Medida de la bondad de ajuste de un modelo lineal; en ocasiones recibe el nombre de coeficiente de determinación. Es la proporción de la variación de la variable de destino explicada por el modelo. Puede tomar un valor entre 0 y 1. Un valor pequeño indica que el modelo no se ajusta bien a los datos.

**Raíz de la media cuadrática de los errores porcentuales**
Una medida de cuánto difieren los valores pronosticados por el modelo de los valores observados de la serie. Es independiente de las unidades utilizadas y, por tanto, se puede utilizar para comparar series con distintas unidades.

**Raíz de la media cuadrática de los errores**
La raíz cuadrada del error cuadrático promedio. Una medida de cuánto se desvía la serie dependiente del nivel pronosticado por el modelo, expresado en las mismas unidades que la serie dependiente.

**BIC** Criterio de información Bayesiano. Una medida para seleccionar y comparar modelos basados en el logaritmo de la verosimilitud -2 reducida. Los valores menores indican modelos mejores. El BIC también "penaliza" modelos sobreparametrizados (modelos complejos con un gran número de entradas, por ejemplo), pero de forma más estricta que el AIC.

**AIC** Criterio de información de Akaike. Una medida para seleccionar y comparar modelos basados en el logaritmo de la verosimilitud -2 reducida. Los valores menores indican modelos mejores. El AIC "penaliza" modelos sobreparametrizados (modelos complejos con un gran número de entradas, por ejemplo).

**Valores atípicos en el tiempo**
Diagrama de barras del número de valores atípicos, en todos los objetivos, para cada intervalo de tiempo en el periodo de estimación.

**Transformaciones de series**
Tabla de cualquier transformación que se ha aplicado a la serie en el sistema del modelo. Las posibles transformaciones son una imputación, agregación y distribución de un valor que falta.

## Periodo de estimación

De forma predeterminada, el período de estimación inicia la hora de la primera observación y finaliza la hora de la última observación en todas las series.

**Por hora de inicio y de finalización**
Puede especificar el inicio y la finalización del período de estimación o puede especificar simplemente el inicio o solo el final. Si omite el inicio o la finalización del período de estimación, se utiliza el valor predeterminado.

- Si las observaciones se han definido mediante un campo de fecha/hora, especifique los valores para el inicio y el final con el mismo formato que se utiliza para el campo de fecha/hora.
- Para las observaciones que se han definido mediante periodos cíclicos, especifique un valor para cada uno de los campos de periodos cíclicos. Cada campo se visualiza en una columna separada.

**Por los primeros o los últimos intervalos de tiempo**

Define el período de estimación como un número especificado de intervalos de tiempo que se inician en el primer intervalo de tiempo o que finalizan en el último intervalo de tiempo en los datos, con un desplazamiento opcional. En este contexto, el intervalo de tiempo hace referencia al intervalo de tiempo del análisis. Por ejemplo, asuma que las observaciones son mensuales, pero el intervalo de tiempo del análisis es trimestres. Especificar **Último** y un valor de 24 para el **Número de intervalos de tiempo** significa los últimos 24 trimestres.

De forma opcional, puede excluir un número especificado de intervalos de tiempo. Por ejemplo, especificar los últimos 24 intervalos de tiempo y 1 para el número para excluir significa que el periodo de estimación consta de los 24 intervalos que preceden al último.

## Opciones de modelo

**Nombre del modelo**
Puede especificar un nombre personalizado para el modelo o aceptar el nombre generado automáticamente, que es *TCM*.

**Previsión**
La opción de **Extender registros en el futuro** define el número de intervalos de tiempo para prever más allá del final del periodo de estimación. El intervalo de tiempo en este caso es el intervalo de tiempo del análisis, que se especifica en la pestaña Especificaciones de datos. Cuando se solicitan previsiones, se generan modelos autorregresivos automáticamente para cualquier serie de entrada que tampoco son objetivos. Estos modelos se utilizan para generar valores para estas series de entrada en el periodo de previsión. No hay límite máximo para este valor.

## Salida interactiva

La salida de un modelado causal temporal incluye una serie de objetos de resultados interactivos. Las funcionalidades interactivas están disponibles activando el objeto (efectuando una doble pulsación sobre él) en el Visor de salidas.

**Sistema de modelo global**
Visualiza las relaciones causales entre las series en el sistema de modelo. Todas las líneas que conectan a un determinado destino con sus entradas tienen el mismo color. El grosor de la línea indica la importancia de la conexión causal, donde una línea más gruesa representa una conexión más importante. Las entradas que no son asimismo objetivos se indican con un cuadrado negro.

- Pueden visualizarse las relaciones de los modelos de nivel superior, de una serie concreta, de todas las series o de modelos sin entradas. Los modelos superiores son los modelos que cumplen los criterios que se han especificado para modelos de mejor ajuste en la configuración de **Series para visualizar**.
- Puede generar diagramas de impacto para una o más series seleccionando los nombres de serie en el gráfico, pulsando con el botón derecho del ratón y seleccionando **Crear diagrama de impacto** en el menú contextual.
- Puede optarse por ocultar las relaciones causales que tengan un nivel de significación mayor que un valor especificado. Un nivel de significación pequeño indica una relación causal más significativa.
- Puede visualizar las relaciones de una serie determinada seleccionando el nombre de la serie en el gráfico, pulsando con el botón derecho y seleccionando **Resaltar las relaciones de la serie** en el menú contextual.

**Diagrama de impacto**
Muestra una representación gráfica de las relaciones causales entre una serie de interés y otras series a las que afecta o que le afectan. Las series que afectan a las series de interés se denominan *causes*.

- Las series de interés pueden cambiarse especificando el nombre de la serie que se desee. Al efectuar una doble pulsación en cualquier nodo del diagrama de impacto se cambia la serie de interés a la serie asociada a dicho nodo.

- Pueden alternarse las visualizaciones de causas y efectos, y puede cambiarse el número de niveles de causas o efectos mostrados.
- Si se efectúa una única pulsación en cualquier nodo, se abre un diagrama de secuencia detallado de la serie asociada a dicho nodo.

**Análisis de causa principal de valor atípico**
Determina qué series tienen la mayor probabilidad de ser la causa de cada valor atípico en una series de interés.

- Puede visualizarse la causa principal de cualquier valor atípico seleccionando la fila del valor atípico en la tabla Valores atípicos. También puede mostrarse la causa principal pulsando en el icono del valor atípico en el gráfico de secuencia.
- Si se efectúa una única pulsación en cualquier nodo, se abre un diagrama de secuencia detallado de la serie asociada a dicho nodo.

**Calidad del modelo global**
Histograma de ajuste de modelo de todos los modelos para un determinado estadístico de ajuste. Si se pulsa en una barra del gráfico de barras, se filtra el gráfico de puntos de forma que solo se muestran los modelos asociados a la barra seleccionada. Puede encontrarse el modelo de una determinada serie objetivo en gráfico de puntos especificando el nombre de dicha serie.

**Distribución de valores atípicos**
Diagrama de barras del número de valores atípicos, en todos los objetivos, para cada intervalo de tiempo en el periodo de estimación. Si se pulsa en una barra del gráfico de barras, se filtra el gráfico de puntos de forma que solo se visualizan los valores atípicos asociados a la barra seleccionada.

# Nugget de modelo TCM

La operación de modelado TCM crea varios campos nuevos con el prefijo $TCM- como se indica en la tabla siguiente.

*Tabla 27. Campos nuevos creados por la operación de modelado TCM*

| Nombre de campo | Descripción |
|---|---|
| $TCM-*nombrecol* | Valor previsto por el modelo para cada serie objetivo. |
| $TCMLCI-*nombrecol* | Los intervalos de confianza más bajos para cada serie para la que se han hecho previsiones. |
| $TSUCI-*nombrecol* | Los intervalos de confianza más altos para cada serie para la que se han hecho previsiones. |
| $TCMResidual-*nombrecol* | Valor de residuo de ruido para cada columna de datos del modelo generado. |

## Configuración del nugget de modelo TCM

La pestaña Configuración proporciona opciones adicionales para el nugget de modelo TCM.

### Previsión

La opción de **Extender registros en el futuro** define el número de intervalos de tiempo para prever más allá del final del periodo de estimación. El intervalo de tiempo en este caso es el intervalo de tiempo del análisis, que se especifica en la pestaña Especificaciones de datos del nodo TCM. Cuando se solicitan previsiones, se generan modelos autorregresivos automáticamente para cualquier serie de entrada que tampoco son objetivos. Estos modelos se utilizan para generar valores para estas series de entrada en el periodo de previsión.

### Dejar disponible para puntuación

**Crear nuevos campos para cada modelo que se puntuará.** Permite especificar los nuevos campos que se crearán para cada modelo que se puntuará.

- **Residuos de ruido.** Si se selecciona, creará un nuevo campo (con el prefijo predeterminado $TCM-) para los residuos del modelo de cada campo objetivo, junto con un total de estos valores.
- **Límites de confianza superior e inferior.** Si está seleccionada, se crearán nuevos campos (con el prefijo predeterminado $TCM-) para los intervalos de confianza superior e inferior, respectivamente, de cada campo objetivo, además de los totales de estos valores.

**Objetivos incluidos para puntuación.** Seleccione los objetivos disponibles que se incluirán en la puntuación del modelo.

## Escenarios de modelos causales temporales

El procedimiento de escenarios de modelos causales temporales ejecuta escenarios definidos por el usuario para un sistema de modelo causal temporal, con los datos del conjunto de datos activo. Un *escenario* se define mediante una serie temporal, a la que se hace referencia como *serie raíz* y un conjunto de valores definidos por el usuario para dicha serie a lo largo de un rango de tiempo especificado. Los valores especificados se utilizan para generar predicciones para las series temporales afectadas por la serie raíz. El procedimiento requiere un archivo de sistema de modelo que ha creado el procedimiento de modelado causal temporal. Se presupone que el conjunto de datos activo son los mismos datos que se han utilizado para crear el archivo del sistema de modelo.

### Ejemplo

Mediante el uso del procedimiento de modelado causal temporal, un tomador de decisiones empresariales ha descubierto una métrica clave afecta a un número de indicadores de rendimiento importantes. La métrica se puede controlar, de forma que el tomador de decisiones desea investigar el efecto de distintos conjuntos de valores para la métrica durante el próximo trimestre. La investigación se puede realizar fácilmente cargando el archivo del sistema de modelo en el procedimiento de escenarios de modelos causales temporales y especificando los conjuntos de valores para la métrica clave.

### Definición del periodo de escenario

El periodo del escenario es el periodo durante el cual especifica los valores que se utilizan para ejecutar los escenarios. Puede empezar antes o después del final del período de estimación. De forma opcional, puede especificar predecir más allá del final del periodo del escenario. De forma predeterminada, las predicciones se generan hasta el final del periodo del escenario. Todos los escenarios utilizan el mismo periodo de escenario y especificaciones por lo que concierne a durante cuánto tiempo predecir.

**Nota:** Las predicciones empiezan en el primer período de tiempo posterior al comienzo del período de escenario. Por ejemplo, si el periodo de escenario empieza en 2014-11-01 y el intervalo de tiempo es de meses, la primera predicción será para 2014-12-01.

**Especificar por hora de inicio, hora de finalización o tiempo de pronóstico**

- Si las observaciones se definen mediante un campo de fecha/hora, especifique valores de inicio, final y tiempo de pronóstico con el mismo formato que se utiliza para el campo de fecha/hora. Los valores para los campos de fecha/hora se alinean al principio del intervalo de tiempo asociado. Por ejemplo, si el intervalo de tiempo del análisis es meses, el valor 10/10/2014 se ajusta a 10/01/2014, que es el inicio del mes.
- Para las observaciones que se han definido mediante periodos cíclicos, especifique un valor para cada uno de los campos de periodos cíclicos. Cada campo se visualiza en una columna separada.

**Especificar por intervalos de tiempo relativos al final del periodo de estimación**

Define el inicio y el final en términos del número de intervalos de tiempo relativos al final del periodo de estimación, donde el intervalo de tiempo es el intervalo de tiempo del análisis. El final del período de estimación se define como el intervalo de tiempo 0. Los intervalos de tiempo antes del final del período de estimación tienen valores negativos y los intervalos después del período de estimación tienen valores positivos. También puede especificar cuántos intervalos predecir más allá del final del periodo del escenario. El valor por omisión es 0.

Por ejemplo, suponga que el intervalo de tiempo del análisis es meses y que especifica 1 para iniciar el intervalo y 3 para finalizarlo, y 1 para cuánto más ampliar la previsión. El periodo del escenario sería de 3 meses que siguen al final del periodo de estimación. Las predicciones se generan para los meses segundo y tercero del periodo de escenario y para 1 mes más después del final periodo de escenario.

## Adición de escenarios y grupos de escenarios

La pestaña Escenarios especifica los escenarios que se van a ejecutar. Para definir escenarios, en primer lugar, debe definir el periodo del escenario pulsando **Definir periodo del escenario**. Los escenarios y los grupos de escenarios (solo se aplica a datos multidimensionales) se crean pulsando el botón asociado **Añadir escenario** o **Añadir grupo de escenarios**. Al seleccionar un escenario particular o un grupo de escenarios en la cuadrícula asociada, puede editarlo, hacer una copia o suprimirlo.

### Datos basados en columna

La columna **Campo raíz** de la cuadrícula especifica el campo de serie temporal cuyos valores se sustituyen con los valores de escenario. La columna **Valores de escenario** muestra los valores de escenario especificados en el orden del primero al último. Si los valores del escenario se definen mediante una expresión, la columna muestra la expresión.

### Datos multidimensionales

**Escenarios individuales**

Cada fila de la cuadrícula Escenarios individuales especifica una serie temporal cuyos valores se sustituyen por los valores de escenario especificados. La serie se define mediante la combinación del campo que se especifica en la columna **Métrica raíz** y el valor especificado para cada uno de los campos de dimensión. El contenido de la columna **Valores de escenario** es el mismo que para los datos basados en columnas.

**Grupos de escenarios**

Un *grupo de escenarios* define un conjunto de escenarios que se basan en un solo campo de métrica raíz y varios conjuntos de valores de dimensión. Cada conjunto de valores de dimensión (un valor por campo de dimensión), para el campo de métrica especificado, define una serie temporal. Un escenario individual se genera para cada una de dichas series temporales, cuyos valores se sustituyen después por los valores de escenario. Los valores de escenario para un grupo de escenarios se especifican mediante una expresión, que se aplica después a cada serie temporal del grupo.

La columna **Recuento de series** muestra el número de conjuntos de valores de dimensión que están asociados a un grupo de escenarios. El valor visualizado puede ser mayor que el número real de series temporales que están asociadas al grupo de escenarios (una serie por conjunto). Esta condición se produce cuando algunas de las combinaciones especificadas de valores de dimensión no corresponden a las series incluidas por la métrica raíz para el grupo.

Como ejemplo de un grupo de escenarios, considere un campo de métrica *publicidad* y dos campos de dimensión *región* y *marca*. Puede definir un grupo de escenarios que se basa en *publicidad* como la métrica raíz y que incluye todas las combinaciones de *región* y *marca*. Es posible que después especifique `publicidad*1.2` como la expresión para investigar el efecto de aumentar *publicidad* en un 20 % para cada una de las series temporales que están asociadas al campo *publicidad*. Si hay 4 valores de *región* y 2 valores de *marca*, existen 8 series temporales de ese tipo y, por lo tanto, 8 escenarios definidos mediante el grupo.

**Definición de escenario:** Los valores para definir un escenario dependen de si los datos se basan en columnas o son multidimensionales.

**Serie raíz**

Especifica la serie raíz para el escenario. Cada escenario se basa en un única serie raíz. Para los datos basados en columna, seleccione el campo que define la serie raíz. Para datos multidimensionales, especifique la serie raíz añadiendo una entrada a la cuadrícula para el campo de métrica que contiene la serie. A continuación, especifique los valores de los campos de dimensión que definen la serie raíz. Se aplica lo siguiente para especificar los valores de dimensión:

- Puede especificar el valor para cada campo de dimensión directamente en la cuadrícula o puede seleccionar en la lista de valores de dimensión disponibles. Para seleccionar en la lista de valores de dimensión disponibles, pulse el botón de puntos suspensivos en la casilla para la dimensión que desee. Esta acción abre el subdiálogo Seleccionar valor de dimensión.
- Puede buscar en la lista de valores de dimensión, en el subdiálogo Seleccionar valor de dimensión, pulsando el icono de los prismáticos y especificando un término de búsqueda. Los espacios se tratan como parte del término de búsqueda. Los asteriscos (*) en el término de búsqueda no indican caracteres comodín.

**Especificar objetivos afectados**

Utilice esta opción cuando sepa los objetivos específicos afectados por la serie raíz y cuando desee investigar los efectos solo en estos objetivos. De forma predeterminada, los objetivos afectados por la serie raíz se determinan automáticamente. Puede especificar la amplitud de la serie afectada por el escenario con los valores de la pestaña Opciones.

Para datos basados en columnas, seleccione los objetivos que desea. Para datos multidimensionales, especifique series de objetivo añadiendo una entrada a la cuadrícula para el campo de métrica de objetivo que contiene la serie. De forma predeterminada, se incluyen todas las series que se incluyen en el campo de métrica especificado. Puede personalizar el conjunto de series incluidas personalizando los valores incluidos para uno o más de los campos de dimensión. Para personalizar los valores de dimensión que se incluyen, pulse el botón de los puntos suspensivos para la dimensión que desee. Esta acción abre el diálogo Seleccionar valores de dimensión.

La columna **Recuento de series** (para datos multidimensionales) muestra el número de conjuntos de valores de dimensión especificados actualmente para la métrica de objetivo asociado. El valor visualizado puede ser mayor que el número real de series de objetivos afectados (una serie por conjunto). Esta condición se produce cuando algunas de las combinaciones especificadas de valores de dimensión no corresponden a las series incluidas en la métrica de objetivo asociado.

**ID de escenario**

Cada escenario debe tener un identificador exclusivo. El identificador se visualiza en el resultado asociado al escenario. No hay ninguna restricción, que no sea la exclusividad, en el valor del identificador.

**Especificar valores del escenario para la serie raíz**

Utilice esta opción para especificar valores explícitos de la serie raíz en el periodo del escenario. Debe especificar un valor numérico para cada intervalo de tiempo listado en la cuadrícula. Puede obtener los valores de la serie raíz (real o previsto) para cada intervalo del periodo del escenario pulsando **Leer**, **Previsión** o **Leer\Previsión**.

**Especificar expresión de los valores del escenario de la serie raíz**

Puede definir una expresión para calcular los valores de la serie raíz en el periodo de escenario. Puede especificar la expresión directamente o pulsar el botón de la calculadora y crear la expresión desde el Generador de expresiones de valores de escenario.

- La expresión puede contener cualquier objetivo o entrada en el sistema modelo.
- Cuando el periodo del escenario se extiende más allá de los datos existentes, la expresión se aplica a valores previstos de los campos en la expresión.

- Para datos multidimensionales, cada campo de la expresión especifica una serie temporal definida mediante los valores del campo y de dimensión que se han especificado para la métrica raíz. Es una de esas series temporales que se utilizan para evaluar la expresión.

Como ejemplo, suponga que el campo raíz es *publicidad* y la expresión es `publicidad*1.2`. Los valores para *publicidad* que se utilizan en el escenario representan un aumento del 20 % con respecto a los valores existentes.

**Nota:** Se crean escenarios pulsando **Añadir escenario** en la pestaña Escenarios.

*Seleccionar valores de dimensión:* Para datos multidimensionales, puede personalizar los valores de dimensión que definen los objetivos afectados por un escenario o un grupo de escenarios. También puede personalizar los valores de dimensión que definen el conjunto de series raíz para un grupo de escenarios.

**Todos los valores**
Especifica que todos los valores del campo de dimensión actual se incluyen. Esta opción es la predeterminada.

**Seleccionar valores**
Utilice esta opción para especificar el conjunto de valores para el campo de dimensión actual. Puede filtrar el conjunto de valores entre los que elegir. Los valores que cumplen la condición de filtro aparecen en la pestaña **Coincidentes** y los valores que no cumplen la condición de filtro aparecen en la pestaña **No coincidentes** de la lista **Valores no seleccionados**. La pestaña **Todos** lista todos los valores no seleccionados, independientemente de cualquier condición de filtro.

- Puede utilizar asteriscos (*) para indicar caracteres comodín cuando especifique un filtro.
- Para borrar el filtro actual, especifique un valor vacío para el término de búsqueda en el diálogo Filtrar valores visualizados.

Para personalizar valores de dimensión para objetivos afectados:

1. En el diálogo Definición de escenario o Definición de grupo de escenarios, seleccione la métrica de objetivo para la cual desea personalizar valores de dimensión.
2. Pulse el botón de puntos suspensivos en la columna para la dimensión que desee personalizar.

Para personalizar valores de dimensión para las series raíz de un grupo de escenarios:

1. En el diálogo Definición de grupo de escenarios, pulse el botón de puntos suspensivos (en la cuadrícula de serie raíz) para la dimensión que desea personalizar.

**Definición de grupo de escenario:**

**Serie raíz**
Especifica el conjunto de series raíz para el grupo de escenarios. Un escenario individual se genera para cada serie temporal del conjunto. Especifique la serie raíz añadiendo una entrada a la cuadrícula para el campo de métrica que contiene la serie que desea. A continuación, especifique los valores de los campos de dimensión que definen el conjunto. De forma predeterminada, todas las series incluidas en el campo de métrica raíz especificado se incluyen. Puede personalizar el conjunto de series incluidas personalizando los valores incluidos para uno o más de los campos de dimensión. Para personalizar los valores de dimensión que se incluyen, pulse el botón de los puntos suspensivos para una dimensión. Esta acción abre el diálogo Seleccionar valores de dimensión.

La columna **Recuento de series** muestra el número de conjuntos de valores de dimensión que están incluidos actualmente para la métrica raíz asociada. El valor visualizado puede ser mayor que el número real de series raíz para el grupo de escenarios (una serie por conjunto). Esta condición se produce cuando algunas de las combinaciones especificadas de valores de dimensión no corresponden a las series incluidas por la métrica raíz.

**Especifique series objetivo afectadas**
Utilice esta opción cuando sepa los objetivos específicos afectados por el conjunto de series raíz y

desea investigar los efectos solo en estos objetivos. De forma predeterminada, los objetivos afectados por cada serie raíz se determinan automáticamente. Puede especificar el ancho de las series afectadas por cada escenario individual con la configuración en la pestaña Opciones.

Especifique las series objetivo añadiendo una entrada a la cuadrícula para el campo de métrica que contiene las series. De forma predeterminada, se incluyen todas las series que se incluyen en el campo de métrica especificado. Puede personalizar el conjunto de series incluidas personalizando los valores incluidos para uno o más de los campos de dimensión. Para personalizar los valores de dimensión que se incluyen, pulse el botón de los puntos suspensivos para la dimensión que desee. Esta acción abre el diálogo Seleccionar valores de dimensión.

La columna **Recuento de series** muestra el número de conjuntos de valores de dimensión que están especificados actualmente para la métrica de objetivo asociada. El valor visualizado puede ser mayor que el número real de series de objetivos afectados (una serie por conjunto). Esta condición se produce cuando algunas de las combinaciones especificadas de valores de dimensión no corresponden a las series incluidas en la métrica de objetivo asociado.

**Prefijo del ID de escenario**
Cada grupo de escenarios debe tener un prefijo exclusivo. El prefijo se utiliza para construir un identificador que se visualiza en el resultado que está asociado a cada escenario individual e n el grupo de escenarios. El identificador para un escenario individual es el prefijo, seguido por un subrayado, seguido por el valor de cada campo de dimensión que identifica la serie raíz. Los valores de dimensión se separan mediante subrayados. No hay ninguna restricción, que no sea la exclusividad, en el valor del prefijo.

**Expresión para valores de escenario para series raíz**
Los valores de escenario para un grupo de escenarios se especifican mediante una expresión, que se utiliza después para calcular los valores para cada una de las series raíz del grupo. Puede especificar una expresión directamente o pulsar el botón de la calculadora y crear la expresión a partir del Generador de expresiones de valores de escenario.

- La expresión puede contener cualquier objetivo o entrada en el sistema modelo.
- Cuando el periodo del escenario se extiende más allá de los datos existentes, la expresión se aplica a valores previstos de los campos en la expresión.
- Para cada serie raíz del grupo, los campos de la expresión especifican series temporales definidas mediante estos campos y los valores de dimensión que definen las series raíz. Es una de esas series temporales que se utilizan para evaluar la expresión. Por ejemplo, si una serie raíz se define mediante `región='north'` y `marca='X'`, las series temporales que se utilizan en la expresión se definen mediante estos mismos valores de dimensión.

Como ejemplo, suponga que el campo de métrica raíz es *publicidad* y que hay dos campos de dimensión *región* y *marca*. Además, suponga que el grupo de escenarios incluye todas las combinaciones de los valores de campo de dimensión. Es posible que después especifique `publicidad*1.2` como la expresión para investigar el efecto de aumentar *publicidad* en un 20 % para cada una de las series temporales que están asociadas al campo *publicidad*.

**Nota:** Los grupos de escenarios se aplican a datos multidimensionales y se crean pulsando **Añadir grupo de escenarios** en la pestaña Escenarios.

## Opciones

**Nivel máximo de objetivos afectados**
Especifica el número máximo de niveles de objetivos afectados. Cada nivel sucesivo, hasta el máximo de 5, incluye objetivos a los que las series raíz afectan indirectamente. Específicamente, el primer nivel incluye objetivos que tienen las series raíz como entrada directa. Los objetivos del segundo nivel tienen objetivos del primer nivel como una entrada directa, y así sucesivamente. Aumentar este valor aumenta la complejidad del cálculo y podría afectar al rendimiento.

**Número máximo de objetivos detectados automáticamente**
Especifica el número máximo de objetivos afectados que se han detectado automáticamente para cada serie raíz. Aumentar este valor aumenta la complejidad del cálculo y podría afectar al rendimiento.

**Diagrama de impacto**
Muestra una representación gráfica de las relaciones causales entre las series raíz para cada escenario y las series de objetivos a que afecta. Las tablas de los valores de escenario y de los valores pronosticados de los objetivos afectados se incluyen como parte del elemento de resultado. El gráfico incluye gráficos de los valores pronosticados de los objetos afectados. Si pulsa una vez en cualquier nodo del diagrama de impactos se abre un diagrama de secuencia detallada para la serie que está asociada al nodo. Se genera un diagrama de impactos separado para cada escenario.

**Gráficos de series**
Genera gráficos de series de los valores pronosticados para cada uno de los objetivos afectados en cada escenario.

**Tablas de previsión y escenario**
Tablas de valores pronosticados y valores de escenario para cada escenario. Estas tablas contienen la misma información que las tablas del diagrama de impactos. Estas tablas soportan todas las funciones estándar para tablas dinámicas y de edición.

**Incluir intervalos de confianza en gráficos y tablas**
Especifica si los intervalos de confianza para las predicciones de escenario se incluyen en los resultados del gráfico y, también, de tabla.

**Ancho de intervalo de confianza (%)**
Este valor controla los intervalos de confianza para las predicciones de escenario. Puede especificarse cualquier valor positivo inferior a 100. De forma predeterminada, se utiliza un intervalo de confianza del 95%.

## Nodo Serie temporal

El nodo Serie temporal se puede utilizar con datos en un entorno local o distribuido; en un entorno distribuido puede emplear la potencia de IBM SPSS Analytic Server. Con este nodo, puede elegir estimar y generar modelos de suavizado exponencial, modelos autorregresivos integrados de media móvil (ARIMA) univariados o ARIMA multivariados (o función de transferencia) para la serie temporal, y generar previsiones basándose en los datos de serie temporal.

El **suavizado exponencial** es un método de previsión que utiliza los valores ponderados de las observaciones anteriores de la serie para predecir los valores futuros. Como tal, el suavizado exponencial no se basa en una comprensión teórica de los datos. Prevé un punto cada vez, corrigiendo las previsiones a medida que entran nuevos datos. La técnica es útil para hacer previsiones de las series que muestran una tendencia o estacionalidad. Puede elegir entre varios modelos de suavizado exponencial que difieren en su tratamiento de la tendencia y la estacionalidad.

Los modelos **ARIMA** proporcionan métodos más sofisticados para crear modelos de los componentes de tendencia y estacionales que los modelos de suavizado exponencial y, en concreto, ofrecen la ventaja adicional de incluir variables independientes (predictoras) en el modelo. Esto implica la especificación explícita de órdenes autorregresivos y de media móvil además del grado de diferenciación. Puede incluir variables del predictor y definir funciones de transferencia para algunas o todas ellas, así como especificar la detección automática de valores atípicos o especificar un conjunto explícito de valores atípicos.

**Nota:** En términos prácticos, los modelos ARIMA son especialmente útiles si desea incluir predictores que podrían ayudar a explicar el comportamiento de la serie que se está previendo como, por ejemplo, el número de catálogos que se han enviado por correo electrónico o el número de visitas en la página web

de una empresa. Los modelos de suavizado exponencial describen el comportamiento de la serie temporal sin tratar de comprender el motivo de su comportamiento. Por ejemplo, una serie que históricamente llega a su máximo cada 12 meses probablemente seguirá haciéndolo, incluso aunque se desconozca el motivo.

También está disponible una opción **Modelizador experto**, que intenta identificar y estimar automáticamente el modelo de suavizado exponencial o ARIMA de mejor ajuste para una o más variables de objetivo, así se elimina la necesidad de identificar un modelo apropiado mediante el sistema de prueba y error. En caso de duda, utilice la opción Modelizador experto.

Si se especifican variables de predictor, el Modelizador experto selecciona estas variables que tienen una relación estadísticamente significativa con la serie dependiente para la inclusión en modelos ARIMA. Las variables del modelo se transforman cuando es necesario mediante una diferenciación y/o una raíz cuadrada o una transformación logarítmica natural. De forma predeterminada, el modelizador experto tiene en cuenta todos los modelos de suavizado exponencial y todos los modelos ARIMA y elige el mejor modelo para cada campo objetivo. Sin embargo, puede limitar el modelizador experto para que sólo elija el mejor modelo de suavizado exponencial o para que sólo elija el mejor modelo ARIMA. Además, puede especificar la detección automática de valores atípicos.

## Nodo Serie temporal - Opciones de campo

En la pestaña Campos, puede seleccionar si desea utilizar la configuración de rol de campo ya definida en nodos anteriores o realizar las asignaciones de campos manualmente.

**Utilizar roles predefinidos** Esta opción utiliza las definiciones de roles (objetivos, predictores, etcétera) desde un nodo Tipo anterior (o la pestaña Tipo de un nodo de origen anterior).

**Utilizar asignaciones de campos personalizadas.** Para asignar manualmente objetivos, predictores y otros roles, seleccione esta opción.

**Campos.** Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los diferentes campos de roles en la parte derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol.

Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

**Objetivos.** Seleccione uno o varios de los campos como el destino de la predicción.

**Entradas candidatas.** Seleccione uno o más campos como entradas de la predicción.

**Eventos e intervenciones.** Utilice esta área para designar determinados campos de entrada como campos de eventos o intervención. Esta designación identifica un campo como que contiene datos de series temporales que se ven afectados por eventos (situaciones recurrentes predecibles; por ejemplo, promociones de ventas) o intervenciones (incidentes puntuales; por ejemplo, apagones o huelgas).

## Nodo Serie temporal - Opciones de especificación de datos

La pestaña Especificación de datos es donde se establecen todas las opciones para los datos que se van a incluir en el modelo. Siempre que especifique ambos campos, **Campo de fecha/hora** e **Intervalo de tiempo**, puede pulsar el botón **Ejecutar** para crear un modelo con todas las opciones predeterminadas pero, normalmente, querrá personalizar la creación para sus propios objetivos.

La pestaña contiene varios paneles en los que puede establecer las personalizaciones específicas del modelo.

## Nodo Serie temporal - Observaciones

Utilice los valores de este panel para especificar los campos que definen las observaciones.

### Observaciones que se especifican mediante un campo de fecha/hora.

Puede especificar que las observaciones se definen mediante un campo de fecha, hora o de indicación de fecha y hora. Además del campo que define las observaciones, seleccione el intervalo de tiempo apropiado que describe las observaciones. En función del intervalo de tiempo especificado, también puede especificar otros valores como, por ejemplo, el intervalo entre observaciones (incremento) o el número de días por semana. Las consideraciones siguientes se aplican al intervalo de tiempo:

- Utilice el valor **Irregular** cuando las observaciones se asignan de forma irregular en el tiempo como, por ejemplo, la hora a la que se procesa un pedido de compra. Si está seleccionado **Irregular**, debe especificar el intervalo de tiempo que se utiliza para el análisis, desde la configuración de **Intervalo de tiempo** en la pestaña Especificaciones de datos.
- Cuando las observaciones representan una fecha y una hora y el intervalo de tiempo es horas, minutos o segundos, utilice **Horas del día**, **Minutos del día** o **Segundos del día**. Cuando las observaciones representan un periodo de tiempo (duración) sin referencia a una fecha y el intervalo de tiempo es horas, minutos o segundos, utilice **Horas (no periódico)**, **Minutos (no periódico)** o **Segundos (no periódico)**.
- Basándose en el intervalo de tiempo seleccionado, el procedimiento puede detectar observaciones que faltan. La detección de observaciones que faltan es necesaria, porque el procedimiento presupone que todas las observaciones se han espaciado de forma uniforme en el tiempo y que no falta ninguna observación. Por ejemplo, si el intervalo de tiempo es Días y la fecha 2015-10-27 está seguida por 2015-10-29, falta una observación para 2015-10-28. Se imputan valores para cualquier observación que falta; utilice el área **Manejo de valores de perdidos** de la pestaña Especificación de datos para especificar valores para el manejo de valores perdidos.
- El intervalo de tiempo especificado permite al procedimiento detectar varias observaciones en el mismo intervalo de tiempo que se deben agregar juntas y alinear observaciones en un límite de intervalo como, por ejemplo, el primer día del mes, para garantizar que las observaciones se espacian de forma uniforme. Por ejemplo, si el intervalo de tiempo es Meses, varias fechas del mismo mes se pueden agregar juntas. Se hace referencia a este tipo de agregación como *agrupación*. De forma predeterminada, las observaciones se suman cuando se agrupan. Puede especificar un método diferente para la agrupación, como la media de las observaciones, desde la configuración **Agregación y distribución** en la pestaña Especificaciones de datos.
- Para algunos intervalos de tiempo, los valores adicionales pueden definir saltos en los intervalos normales espaciados de forma uniforme. Por ejemplo, si el intervalo de tiempo es Días, pero solo son válidos los fines de semana, puede especificar que hay cinco días en una semana y que la semana empieza el lunes.

### Observaciones que están definidas como periodos o periodos cíclicos

Las observaciones se pueden definir mediante uno o más campos de enteros que representan periodos o ciclos repetitivos de periodos, hasta un número arbitrario de niveles de ciclo. Con esta estructura, puede describir series de observaciones que no caben en uno delos intervalos de tiempo estándar. Por ejemplo, un año fiscal con solo 10 meses se puede describir con un campo de ciclo que representa años y un capo de periodo que representa meses, donde la longitud de un ciclo es 10.

Los campos que especifican periodos cíclicos definen una jerarquía de niveles periódicos, donde el nivel más bajo se define mediante el campo **Periodo**. El siguiente nivel más alto se especifica mediante un campo de ciclo cuyo nivel es 1, seguido de un campo de ciclo cuyo nivel es 2 y, así, sucesivamente. Los valores de campo para cada nivel, excepto para el más alto, deben ser periódicos con respecto al siguiente nivel superior. Los valores para el nivel superior no pueden ser periódicos. Por ejemplo, en el caso del año fiscal de 10 meses, los meses son periódicos dentro de los años y los años no son periódicos.

- La longitud de un ciclo en un nivel particular es la periodicidad del siguiente nivel inferior. Para el ejemplo del año fiscal, solo hay un nivel de ciclo y la longitud del ciclo es 10 porque el siguiente nivel inferior representa meses y hay 10 meses en el año fiscal especificado.
- Especifique el valor de inicio para cualquier campo periódico que no empiece desde el 1. Este valor es necesario para detectar valores perdidos. Por ejemplo, si un campo periódico empieza en el 2, pero el valor inicial se especifica como 1, el procedimiento supone que falta un valor para el primer periodo de cada ciclo de dicho campo.

## Nodo Serie temporal - Intervalo de tiempo para análisis

El intervalo de tiempo que se utiliza para el análisis puede diferir del intervalo de tiempo de las observaciones. Por ejemplo, si el intervalo de tiempo de las observaciones es Días, podría elegir Meses para el intervalo de tiempo del análisis. A continuación, se agregan los datos de datos diarios a mensuales antes de que se genere el modelo. También puede elegir distribuir los datos de un intervalo de tiempo más largo a uno más corto. Por ejemplo, si las observaciones son trimestrales, puede distribuir los datos de datos trimestrales a datos mensuales.

Utilice los valores de este panel para especificar el intervalo de tiempo para el análisis. El método mediante el cual los datos se agregan o distribuyen se especifica en los valores de **Agregación y distribución** en la pestaña Especificaciones de datos.

Las opciones disponibles para el intervalo de tiempo en el que se realiza el análisis dependen de cómo se han definido las observaciones y del intervalo de tiempo de estas observaciones. En particular, cuando las observaciones se definen mediante periodos cíclicos, solo está soportada la agregación. En este caso, el intervalo de tiempo del análisis debe ser mayor o igual que el intervalo de tiempo de las observaciones.

## Nodo Serie temporal - Opciones de agregación y distribución

Utilice los valores de este panel para especificar los valores para agregar o distribuir los datos de entrada con respecto a los intervalos de tiempo de las observaciones.

**Funciones de agregación**

Cuando el intervalo de tiempo que se utiliza para el análisis es más largo que el intervalo de tiempo de las observaciones, se agregan los datos de entrada. Por ejemplo, la agregación se realiza cuando el intervalo de tiempo de las observaciones es Días y el intervalo de tiempo para el análisis es Meses. Están disponibles las funciones de agregación siguientes: media, suma, moda, mín o máx.

**Funciones de distribución**

Cuando el intervalo de tiempo que se utiliza para el análisis es más corto que el intervalo de tiempo de las observaciones, se distribuyen los datos de entrada. Por ejemplo, la distribución se realiza cuando el intervalo de tiempo de las observaciones es Trimestres y el intervalo de tiempo para el análisis es Meses. Están disponibles las funciones de distribución siguientes: media o suma.

**Funciones de agrupación**

La agrupación se aplica cuando las observaciones se definen mediante fecha/horas y se producen varias observaciones en el mismo intervalo de tiempo. Por ejemplo, si el intervalo de tiempo de las observaciones es Meses, se agrupan varias fechas del mismo mes y se asocian al mes en el cual se producen. Están disponibles las funciones de agrupación siguientes: media, suma, moda, mín o máx. La agrupación siempre se realiza cuando las observaciones se definen por fecha/horas y el intervalo de tiempo de las observaciones se especifica como Irregular.

**Nota:** Aunque la agrupación es una forma de agregación, se realiza antes de manejar los valores que faltan, mientras que la agregación formal se realiza después de que se manejen los valores que faltan. Cuando el intervalo de tiempo de las observaciones se especifica como irregular, la agregación solo se realiza con la función de agrupación.

**Agregar observaciones del día al día anterior**

Especifica si las observaciones con horas que cruzan el límite de un día se agregan a los valores

del día anterior. Por ejemplo, para observaciones por hora con un día de ocho horas que empieza a las 20:00, este valor especifica si las observaciones entre las 00:00 y las 04:00 se incluyen en los resultados agregados para el día anterior. Este valor se aplica solo si el intervalo de tiempo de las observaciones es Horas por día, Minutos por día o Segundos por día y el intervalo de tiempo para el análisis es Días.

**Valores personalizados para campos especificados**
Puede especificar funciones de agregación, distribución y agrupación en un campo por campo. Estos valores alteran temporalmente los valores predeterminados para las funciones de agregación, distribución y agrupación.

## Nodo Serie temporal - Opciones de valor perdido

Utilice los valores de este panel para especifica cómo se van a sustituir los posibles valores perdidos en los datos de entrada con un valor imputado. Están disponibles los métodos de sustitución siguientes:

**Interpolación lineal**
Sustituye los valores perdidos utilizando una interpolación lineal. Se utilizan para la interpolación el último valor válido antes del valor perdido y el primer valor válido después del valor perdido. Si la primera o la última observación de la serie tiene un valor que falta, se utilizan los dos valores que no faltan más cercanos al principio o al final de la serie.

**Media de la serie**
Sustituye los valores perdidos con la media de toda la serie.

**Media de los puntos adyacentes**
Sustituye los valores perdidos por la media de los valores válidos circundantes. La amplitud de los puntos adyacentes es el número de valores válidos anteriores y posteriores del valor que falta que se utilizan para calcular la media.

**Mediana de puntos adyacentes**
Sustituye los valores perdidos por la mediana de los valores válidos circundantes. La amplitud de los puntos adyacentes es el número de valores válidos anteriores y posteriores del valor que falta que se utilizan para calcular la mediana.

**Tendencia lineal**
Esta opción utiliza todas las observaciones que no faltan en la serie para ajustarse a un modelo de regresión lineal simple, que se utiliza después para imputar los valores perdidos.

Otros valores:

**Puntuación más baja de calidad de datos (%)**
Calcula las medidas de la calidad de datos para la variable de tiempo y para los datos de entrada correspondientes a cada serie temporal. Si la puntuación de la calidad de datos es inferior a este umbral, se descartará la serie temporal correspondiente.

## Nodo Serie temporal - Período de estimación

En el panel Período de estimación, puede especificar el rango de registros que se van a utilizar en la estimación del modelo. De forma predeterminada, el período de estimación inicia la hora de la primera observación y finaliza la hora de la última observación en todas las series.

**Por hora de inicio y de finalización**
Puede especificar el inicio y la finalización del período de estimación o puede especificar simplemente el inicio o solo el final. Si omite el inicio o la finalización del período de estimación, se utiliza el valor predeterminado.

- Si las observaciones se han definido mediante un campo de fecha/hora, especifique valores para el inicio y el final con el mismo formato que se ha utilizado para el campo de fecha/hora.
- Para las observaciones que se han definido mediante periodos cíclicos, especifique un valor para cada uno de los campos de periodos cíclicos. Cada campo se visualiza en una columna separada.

**Por los primeros o los últimos intervalos de tiempo**

Define el período de estimación como un número especificado de intervalos de tiempo que empiezan en el primer intervalo de tiempo o finalizan en el último intervalo de tiempo de los datos, con un desplazamiento opcional. En este contexto, el intervalo de tiempo hace referencia al intervalo de tiempo del análisis. Por ejemplo, asuma que las observaciones son mensuales, pero el intervalo de tiempo del análisis es trimestres. Especificar **Último** y un valor de 24 para el **Número de intervalos de tiempo** significa los últimos 24 trimestres.

De forma opcional, puede excluir un número especificado de intervalos de tiempo. Por ejemplo, especificar los últimos 24 intervalos de tiempo y 1 para el número para excluir significa que el periodo de estimación consta de los 24 intervalos que preceden al último.

# Nodo Serie temporal - Opciones de generación

La pestaña Opciones de generación es donde se establecen todas las opciones para generar el modelo. Puede pulsar en el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero normalmente querrá personalizar la generación de sus tareas.

La pestaña contiene dos paneles diferentes en los cuales se establecen las personalizaciones que son específicas al modelo.

## Nodo Serie temporal - Opciones generales de generación

Las opciones disponibles en este panel dependen de cuál de los tres valores siguientes seleccione en la lista **Método**:

- **Modelizador experto.** Seleccione esta opción para utilizar el modelizador experto, que busca automáticamente el modelo que mejor se ajusta a cada serie dependiente.
- **Suavizado exponencial.** Utilice esta opción para especificar un modelo de suavizado exponencial personalizado.
- **ARIMA.** Utilice esta opción para especificar un modelo ARIMA personalizado.

### Modelizador experto

Bajo **Tipo de modelo**, seleccione el tipo de modelos que desea generar:

- **Todos los modelos.** El modelizador experto tiene en cuenta tanto los modelos ARIMA como los modelos de suavizado exponencial.
- **Sólo modelos de suavizado exponencial.** El Modelizador experto solo tiene en cuenta los modelos de suavizado exponencial.
- **Sólo modelos ARIMA.** El Modelizador experto solo tiene en cuenta los modelos ARIMA.

**El modelizador experto considera modelos estacionales.** Esta opción sólo se habilita si se ha definido una periodicidad para el conjunto de datos activo. Si esta opción está seleccionada, el modelizador experto tiene en cuenta los modelos tanto estacionales como no estacionales. Si esta opción no está seleccionada, el Modelizador experto solo tiene en cuenta modelos no estacionales.

**El Modelizador experto tiene en cuenta los modelos de suavizado exponencial sofisticados.** Cuando se selecciona esta opción, el Modelizador experto busca un total de 13 modelos de suavizado exponencial (7 de ellos existían en el nodo Serie temporal original y 6 de ellos se han añadido en la versión 18.1). Si no se selecciona esta opción, el Modelizador experto sólo busca los 7 modelos de suavizado exponencial originales.

Bajo **Valores atípicos**, seleccione las opciones siguientes

**Detectar automáticamente los valores atípicos.** De forma predeterminada, no se realiza la detección automática de valores atípicos. Seleccione esta opción para realizar una detección automática de valores atípicos y, a continuación, seleccione los tipos de valores atípicos que desee.

Los campos de entrada deben tener un nivel de medición de *Marca*, *Nominal* u *Ordinal* y deben ser numéricos (por ejemplo, 1/0, no Verdadero/Falso, para un campo de marca), antes de que se puedan incluir en esta lista.

El Modelizador experto solo tiene en cuenta las funciones de regresión simple y nos las de transferencia arbitraria para entradas identificadas como campos de evento o intervención en la pestaña **Campos**.

## suavizado exponencial

**Tipo de modelo.** Los modelos de suavizado exponencial se clasifican como estacionales o no estacionales.[1] Los modelos estacionales sólo están disponibles si la periodicidad definida utilizando el panel Intervalos de tiempo en la pestaña Especificaciones de datos es estacional. Las periodicidades estacionales son las siguientes: periodos cíclicos, años, trimestres, meses, días por semana, horas por día, minutos por día y segundos por día. Están disponibles los tipos de modelo siguientes:

- **simples.** Este modelo es adecuado para las series sin tendencia ni estacionalidad. Su único parámetro de suavizado relevante es el nivel. El suavizado exponencial simple es muy similar a un ARIMA con cero órdenes de autorregresión, un orden de diferenciación, un orden de media móvil y ninguna constante.
- **Tendencia lineal de Holt.** Este modelo es adecuado para las series con una tendencia lineal y sin estacionalidad. Sus parámetros de suavizado relevantes son el nivel y la tendencia y, en este modelo, no están restringidos por sus valores respectivos. El modelo de Holt es más general que el de Brown, pero puede tardar más en calcular estimaciones para series grandes. El suavizado exponencial de Holt es muy similar a un ARIMA con cero órdenes de autorregresión, dos órdenes de diferenciación y dos órdenes de media móvil.
- **Tendencia amortiguada.** Este modelo es adecuado para las series con una tendencia lineal que va desapareciendo y sin estacionalidad. Sus parámetros de suavizado relevantes son el nivel, la tendencia y la tendencia de amortiguación. El suavizado exponencial amortiguado es muy similar a un ARIMA con cero órdenes de autorregresión, un orden de diferenciación y dos órdenes de media móvil.
- **Tendencia multiplicativa.** Este modelo es adecuado para una serie en la que hay una tendencia que cambia con la magnitud de la serie y sin estacionalidad. Sus parámetros de suavizado relevantes son el nivel y la tendencia. El suavizado exponencial de tendencia multiplicativa no es similar a ningún modelo ARIMA.
- **Tendencia lineal de Brown.** Este modelo es adecuado para las series con una tendencia lineal y sin estacionalidad. Sus parámetros de suavizado relevantes son el nivel y la tendencia, pero, en este modelo, se supone que son iguales. Por ello, el modelo de Brown es un caso especial del modelo de Holt. El suavizado exponencial de Brown es muy similar a un ARIMA con cero órdenes de autorregresión, dos órdenes de diferenciación y dos órdenes de media móvil, siendo el coeficiente del segundo orden de la media móvil igual a la mitad del coeficiente del primer orden al cuadrado.
- **Estacional simple.** Este modelo es adecuado para las series sin una tendencia y un efecto estacional constante a lo largo del tiempo. Sus parámetros de suavizado relevantes son el nivel y la estacionalidad. El suavizado exponencial estacional es muy similar a un ARIMA con cero órdenes de autorregresión, un orden de diferenciación, un orden de diferenciación estacional y los órdenes 1, $p$ y $p$+1 de media móvil, donde $p$ es el número de períodos contenidos en un intervalo estacional. En el caso de los datos mensuales, $p = 12$.
- **Aditivo de Winters.** Este modelo es adecuado para las series con una tendencia lineal y un efecto estacional constante a lo largo del tiempo. Sus parámetros de suavizado relevantes son el nivel, la tendencia y la estacionalidad. El suavizado exponencial aditivo de Winters es muy similar a un ARIMA con cero órdenes de autorregresión, un orden de diferenciación, un orden de diferenciación estacional y los órdenes $p$+1 de media móvil, donde $p$ es el número de períodos contenidos en un intervalo estacional. En el caso de los datos mensuales, $p = 12$.

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting*, 4, 1-28.

- **Tendencia amortiguada con estacional aditiva.** Este modelo es adecuado para una serie en la que hay una tendencia lineal que va desapareciendo y un efecto estacional que es constante a lo largo del tiempo. Sus parámetros de suavizado relevantes son nivel, tendencia, tendencia de amortiguación y estacionalidad. Tendencia amortiguada y suavizado exponencial estacional aditivo no es similar a ningún modelo ARIMA.
- **Tendencia multiplicativa con estacional aditiva.** Este modelo es adecuado para una serie en la que hay una tendencia que cambia con la magnitud de la serie y un efecto estacional que es constante a lo largo del tiempo. Sus parámetros de suavizado relevantes son el nivel, la tendencia y la estacionalidad. Tendencia multiplicativa y suavizado exponencial estacional aditivo no es similar a ningún modelo ARIMA.
- **Estacional multiplicativa.** Este modelo es adecuado para una serie en la que no hay tendencia y un efecto estacional que cambia con la magnitud de la serie. Sus parámetros de suavizado relevantes son el nivel y la estacionalidad. El suavizado exponencial estacional multiplicativo no es similar a ningún modelo ARIMA.
- **Multiplicativo de Winters.** Este modelo es adecuado para series en las que haya una tendencia lineal y con un efecto estacional que cambie en función de la magnitud de las series. Sus parámetros de suavizado relevantes son el nivel, la tendencia y la estacionalidad. El modelo de suavizado exponencial multiplicativo de Winters no es similar a ningún modelo ARIMA.
- **Tendencia amortiguada con estacional multiplicativa.** Este modelo es adecuado para una serie en la que hay una tendencia lineal que va desapareciendo y un efecto estacional que cambia con la magnitud de la serie. Sus parámetros de suavizado relevantes son nivel, tendencia, tendencia de amortiguación y estacionalidad. Tendencia amortiguada y suavizado exponencial estacional multiplicativo no es similar a ningún modelo ARIMA.
- **Tendencia multiplicativa con estacional multiplicativo.** Este modelo es adecuado para una serie en la que hay una tendencia y un efecto estacional que cambian con la magnitud de la serie. Sus parámetros de suavizado relevantes son el nivel, la tendencia y la estacionalidad. Tendencia multiplicativa y suavizado exponencial estacional multiplicativo no es similar a ningún modelo ARIMA.

**Transformación de objetivo.** Puede especificar una transformación para que se lleve a cabo en cada variable dependiente antes de su modelado.

- **Ninguno.** No se lleva a cabo ninguna transformación.
- **Raíz cuadrada.** Se realiza una transformación de raíz cuadrada.
- **Log natural.** Se realiza una transformación logarítmica natural.

## ARIMA

Especifique la estructura de un modelo ARIMA personalizado.

**Órdenes ARIMA.** Escriba valores para los distintos componentes ARIMA del modelo en las casillas correspondientes de la cuadrícula. Todos los valores deben ser enteros no negativos. Para los componentes autorregresivos y de media móvil, el valor representa el orden máximo. Todos los órdenes inferiores positivos se incluyen en el modelo. Por ejemplo, si especifica 2, el modelo incluye los órdenes 2 y 1. Las casillas de la columna Estacional solo están habilitadas si se ha definido una periodicidad para el conjunto de datos activo.

- **Autorregresivo (p).** Es el número de órdenes autorregresivos del modelo. Los órdenes autorregresivos especifican los valores previos de la serie utilizados para predecir los valores actuales. Por ejemplo, un orden autorregresivo de 2 especifica que se utiliza el valor de dos periodos de tiempo de la serie en el pasado para predecir el valor actual.
- **Diferencia (d).** Especifica el orden de diferenciación aplicado a la serie antes de estimar los modelos. La diferenciación es necesaria si hay tendencias (las series con tendencias suelen ser no estacionarias y el modelado de ARIMA asume la estacionariedad) y se utiliza para eliminar su efecto. El orden de

diferenciación corresponde al grado de la tendencia de la serie; la diferenciación de primer orden se tiene en cuenta para tendencias lineales; la diferenciación de segundo orden se tiene en cuenta para tendencias cuadráticas y, así, sucesivamente.

- **Media móvil (q).** Es el número de órdenes de media móvil presentes en el modelo. Los órdenes de media móvil especifican el modo en que se utilizan las desviaciones respecto a la media de la serie para los valores previos con el fin de predecir los valores actuales. Por ejemplo, los órdenes de media móvil de 1 y 2 especifican que las desviaciones del valor medio de la serie de cada uno de los dos últimos períodos de tiempo se tienen en cuenta al predecir los valores actuales de la serie.

**Estacional.** Los componentes estacionales autorregresivos, de media móvil y de diferenciación tienen los mismos roles que los componentes no estacionales correspondientes. Sin embargo, para los órdenes estacionales, los valores de la serie actual se ven afectados por valores de la serie anterior que están separados por uno o más periodos estacionales. Por ejemplo, para datos mensuales (periodo estacional de 12), un orden estacional de 1 significa que el valor de la serie actual se ve afectado por el valor de los 12 periodos de la serie antes del actual. Un orden estacional de 1 para los datos mensuales equivale a la especificación de un orden no estacional de 12.

**Detectar automáticamente los valores atípicos.** Seleccione esta opción para realizar una detección automática de valores atípicos y seleccione uno o más de los tipos de valor atípico disponibles.

**Tipos de valores atípicos que se detectarán.** Seleccione los tipos de valores atípicos que desea detectar. Los tipos admitidos son:

- Aditivo (valor predeterminado)
- Cambio de nivel (valor predeterminado)
- Innovador
- Transitorio
- Aditivo estacional
- Tendencia local
- Parche aditivo

**Órdenes de función de transferencia y transformaciones.** Para especificar transformaciones y definir funciones de transferencia para alguno o todos los campos de entrada del modelo ARIMA, pulse **Establecer**; se muestra un recuadro de diálogo independiente en el que puede especificar los detalles de transferencia y transformación.

**Incluir constante en el modelo.** La inclusión de una constante es estándar a menos que esté seguro de que el valor de la media global de la serie es 0. Se recomienda la exclusión de la constante si se aplica la diferenciación.

## Detalles adicionales

- Si desea más información sobre tipos de valores atípicos, consulte “Valores atípicos” en la página 304.
- Si desea más información sobre funciones de transferencia y transformación, consulte “Funciones de transferencia y transformación”.

**Funciones de transferencia y transformación:** Utilice el recuadro de diálogo Órdenes de función de transferencia y transformaciones para especificar transformaciones y para definir funciones de transferencia para alguno o todos los campos de entrada del modelo ARIMA.

**Transformaciones de destino.** En este panel puede especificar una transformación que se deberá realizar en cada variable de destino antes de que se modele.

- **Ninguno.** No se lleva a cabo ninguna transformación.
- **Raíz cuadrada.** Se realiza una transformación de raíz cuadrada.
- **Log natural.** Se realiza una transformación logarítmica natural.

**Funciones de transferencia de entradas candidatas y transformación.** Utilice las funciones de transferencia para especificar una forma en que los valores pasados de los campos de entrada se utilizan para predecir los valores futuros de la serie de destino. La lista situada a la izquierda del panel muestra todos los campos de entrada. La información restante de este panel es específica al campo de entrada que seleccione.

**Órdenes de la función de transferencia.** Especifique los valores para los diversos componentes de la función de transferencia en las casillas correspondientes de la cuadrícula de **Estructura** . Todos los valores deben ser enteros no negativos. Para los componentes de numerador y denominador, el valor representa el orden máximo. Todos los órdenes inferiores positivos se incluyen en el modelo. Además, el orden 0 siempre se incluye para los componentes de numerador. Por ejemplo, si especifica 2 para el numerador, el modelo incluye los órdenes 2, 1 y 0. Si especifica 3 para el denominador, el modelo incluye los órdenes 3, 2 y 1. Las casillas de la columna Estacional solo están habilitadas si se ha definido una periodicidad para el conjunto de datos activo.

**Numerador.** El orden de numerador de la función de transferencia especifica los valores previos de la serie independiente (predictora) seleccionada que se utilizan para predecir los valores actuales de la serie dependiente. Por ejemplo, un orden de numerador de 1 especifica que se utiliza el valor de un periodo de tiempo de una serie independiente en el pasado, además del valor actual de la serie independiente, para predecir el valor actual de cada serie dependiente.

**Denominador.** El orden de denominador de la función de transferencia especifica cómo se utilizan las desviaciones respecto a la media de la serie para los valores previos de la serie independiente (predictora) seleccionada para predecir los valores actuales de la serie dependiente. Por ejemplo, se tiene en cuenta un orden de denominador de 1 especifica las desviaciones del valor medio de un periodo de tiempo de una serie independiente en el pasado al predecir el valor actual de cada serie dependiente.

**Diferencia.** Especifica el orden de diferenciación aplicado a la serie independiente (predictora) seleccionada antes de estimar los modelos. La diferenciación es necesaria si hay tendencias y se utiliza para eliminar su efecto.

**Estacional.** Los componentes estacionales de numerador, denominador y diferenciación tienen los mismoss roles que los componentes no estacionales correspondientes. Sin embargo, para los órdenes estacionales, los valores de la serie actual se ven afectados por valores de la serie anterior que están separados por uno o más periodos estacionales. Por ejemplo, para los datos mensuales (período estacional de 12), un orden estacional de 1 significa que el valor de la serie actual se ve afectado por el valor de la serie 12 períodos antes del actual. Un orden estacional de 1 para los datos mensuales equivale a la especificación de un orden no estacional de 12.

**Retardo.** Establecer un retardo provoca que la influencia del campo de entrada se retrase según el número de intervalos especificados. Por ejemplo, si el retardo se establece en 5, el valor de la variable de entrada en el tiempo *t* no afecta a las previsiones hasta que han transcurrido cinco períodos (*t* + 5).

**Transformación.** La especificación de una función de transferencia para un conjunto de variables independientes también incluye una transformación opcional que se puede aplicar a dichas variables.

- **Ninguno.** No se lleva a cabo ninguna transformación.
- **Raíz cuadrada.** Se realiza una transformación de raíz cuadrada.
- **Log natural.** Se realiza una transformación logarítmica natural.

## Nodo Serie temporal - Opciones de salida de generación

**Número máximo de retardos en resultados de las FAS y FAP.** La autocorrelación (ACF) y la autocorrelación parcial (PACF) son medidas de asociación entre valores de serie actuales y pasadas e indican cuáles son los valores de series pasadas más útiles para predecir valores futuros. Puede establecer el número máximo de retardos que se muestran en las tablas y gráficos de autocorrelaciones y autocorrelaciones parciales.

**Calcular importancia del predictor.** En el caso de modelos que produzcan una medida adecuada de importancia, puede mostrar un gráfico que indique la importancia relativa de cada predictor al estimar el modelo. Por regla general, el usuario desea centrar sus esfuerzos de modelado en los predictores que le importan más y considera descartar o ignorar los predictores que le preocupan menos. La importancia de predictor puede tardar más en calcularse para algunos modelos, especialmente, cuando se trabaja con conjuntos de datos de gran tamaño y está desactivado, de forma predeterminada, para algunos modelos como resultado.

## Nodo Serie temporal - Opciones de modelo

**Nombre del modelo.** Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

**Amplitud de límite de confianza (%).** Los intervalos de confianza se calculan para las predicciones del modelo y las autocorrelaciones residuales. Puede especificarse cualquier valor positivo inferior a 100. De forma predeterminada, se utiliza un intervalo de confianza del 95%.

**Continuar con la estimación utilizando modelo(s) existente.** Si ya ha generado un modelo de serie temporal, seleccione esta opción para reutilizar los valores de criterios que se han especificado para dicho modelo y genere un nuevo nodo de modelo en la paleta de modelos, en lugar de crear un modelo nuevo desde el principio. De esta forma, puede ahorrar tiempo volviendo a estimar y generando una nueva predicción que se base en los mismos valores de modelo que antes pero utilizando datos más recientes. Así, si el modelo original de una serie temporal determinada era, por ejemplo, Tendencia lineal de Holt, se utilizará el mismo tipo de modelo para volver a estimar esos datos y realizar una previsión con ellos. El sistema no volverá a intentar buscar el mejor tipo de modelo para los nuevos datos.

**Construir únicamente modelo de puntuación.** Para reducir la cantidad de datos que se almacenan en el modelo, seleccione esta casilla. El uso de esta opción puede mejorar el rendimiento al generar modelos con un gran número de series temporales (decenas de miles). Puede seguir puntuando los datos de la forma habitual.

**Extender registros en el futuro.** Habilita la siguiente sección **Valores futuros que se utilizarán en la previsión**, donde puede establecer el número de intervalos de tiempo a prever más allá del final del periodo de estimación. El intervalo de tiempo en este caso es el intervalo de tiempo del análisis, que se especifica en la pestaña Especificaciones de datos. No hay límite máximo para este valor. Mediante las opciones siguientes, puede calcular automáticamente los valores futuros de las entradas o puede especificar manualmente los valores de previsión para uno o más predictores.

### Valores futuros que se utilizarán en la previsión

- **Calcular valores futuros de entradas** Si selecciona esta opción, los valores de previsión para predictores, predicciones de ruido, estimación de varianza y futuros valores de tiempo se calculan automáticamente. Cuando se solicitan previsiones, se generan modelos autorregresivos automáticamente para cualquier serie de entrada que tampoco son objetivos. Estos modelos se utilizan para generar valores para estas series de entrada en el periodo de previsión.
- **Seleccionar los campos cuyos valores se desean añadir a los datos.** Para cada registro que desea prever (excluyendo los casos reservados), si está utilizando campos predictores (con el rol establecido en `Entrada`), puede especificar los valores estimados para el periodo de previsión para cada predictor. Puede introducir los valores manualmente o seleccionarlos de una lista.
  - **Campo.** Pulse en el botón selector de campos y seleccione los campos que se utilizarán como predictores. Tenga en cuenta que los campos aquí seleccionados podrán utilizarse o no en el modelado; para utilizar un campo como predictor debe seleccionarse en un nodo de modelado posterior. Este cuadro de diálogo permite especificar los valores futuros de forma cómoda para que puedan compartirlos varios nodos de modelado posteriores sin tener que especificarlos por separado en cada nodo. También tenga en cuenta que la lista de campos disponibles puede restringirse mediante las selecciones realizadas en la pestaña Opciones de generación.

Tenga en cuenta que si se especifican valores futuros para un campo que ya no está disponible en la ruta (porque se ha eliminado o debido a las selecciones actualizadas realizadas en la pestaña Opciones de generación), el campo se muestra en rojo.

– **Valores.** Para cada campo, puede seleccionar en una lista de funciones o pulsar en **Especificar** para introducir los valores manualmente o desde una lista de valores predefinidos. Si los campos predictores se refieren a los elementos que están bajo su control, o que se conocen de otra forma, debe introducir los valores manualmente. Por ejemplo, si está haciendo previsiones de los ingresos del próximo mes para un hotel a partir del número de reservas, puede especificar el número de reservas que tiene para ese período. Por el contrario, si un campo predictor está relacionado con algo que escapa a su control, como el precio de las acciones, puede utilizar una función como el valor más reciente o la media de los puntos recientes.

  Las funciones disponibles dependen del nivel de medición del campo.

*Tabla 28. Funciones disponibles para los niveles de medición*

| Nivel de medición | Funciones |
|---|---|
| Campo continuo o nominal | Vacío<br>Media de los puntos recientes<br>Valor más reciente<br>Especifica |
| Campo de marcas | Vacío<br>Valor más reciente<br>True<br>False<br>Especifica |

**Media de los puntos recientes** calcula el valor futuro a partir de la media de los tres últimos puntos de datos.

**Valor más reciente** establece el valor futuro en el del punto de datos más reciente.

**Verdadero/falso** establece el valor futuro de un campo de marcas en Verdadero o Falso según lo especificado.

**Especificar** abre un cuadro de diálogo para especificar los valores futuros manualmente o eligiéndolos en una lista predefinida.

## Dejar disponible para puntuación

Puede establecer aquí los valores predeterminados para las opciones de puntuación que aparecen en el cuadro de diálogo del nugget de modelo.

- **Calcular límites de confianza superior e inferior.** Si se selecciona, esta opción crea campos nuevos (con los prefijos predeterminados $TSLCI- y $TSUCI-) para los intervalos de confianza inferior y superior, para cada campo objetivo.
- **Calcular residuos de ruido.** Si se selecciona, esta opción crea un campo nuevo (con el prefijo predeterminado $TSResidual-) para los residuos de modelo para cada campo objetivo, junto con un total de estos valores.

## Configuración del modelo

**Número máximo de modelos que se mostrarán en la salida.** Especifique el número máximo de modelos que desea incluir en el resultado. Tenga en cuenta que si el número de modelos creados supera este umbral, los modelos no se muestran en la salida pero siguen estando disponibles para la puntuación. El valor predeterminado es 10. La visualización de un gran número de modelos puede producir un rendimiento pobre o inestabilidad.

# Nugget del modelo Serie temporal

## Salida del nugget del modelo Serie temporal

Tras crear un modelo de serie temporal, está disponible la información siguiente en el visor de resultados. Tenga en cuenta que existe un límite de 10 modelos que se pueden mostrar en el visor de resultados para los modelos de serie temporal.

### Resumen de información temporal

El resumen muestra la información siguiente:

- El campo Hora
- El incremento
- El punto de inicio y final
- El número de puntos exclusivos

El resumen se aplica a todos los objetivos.

### Tabla de información del modelo

Se repite para cada objetivo, la tabla Modelo de información proporciona información clave sobre el modelo. La tabla siempre incluye los valores del modelo de alto nivel siguientes:

- El nombre del campo objetivo que está seleccionado en el nodo Tipo o la pestaña Campos del nodo Serie temporal.
- El método de generación de modelos, por ejemplo, suavizado exponencial o ARIMA.
- El número de entrada de predictores en el modelo.
- El número de registros que se han utilizado para ajustar el tipo de modelo. Los ejemplos de los distintos tipos de modelos podrían incluir: RMSE, MAE, AIC, BIC y R cuadrado.

Además, las estadísticas Ljunq-Box Q también se podrían mostrar si los datos cumplen las condiciones necesarias. Esta estadística **no** está disponible bajo las condiciones siguientes:

- Si el número de puntos de datos que no faltan es menor o igual que el número de términos de suma deseado (fijado en 18).
- Si el número de parámetros es mayor o igual que el número de términos de suma deseado.
- Si el número de términos de suma que se han calculado es menor que el valor más pequeño de k aceptable (fijado en 7).
- Si la tabla se repite para cada objetivo.

### Importancia del predictor

Se repite para cada objetivo, el gráfico de importancia de predictor muestra la importancia de las 10 primeras entradas (predictores) en el modelo como un gráfico de barras.

Si hay más de 10 campos en el gráfico, puede cambiar la selección de los predictores que se incluyen en el gráfico mediante el control deslizante situado debajo del mismo. Las marcas indicadores del control deslizante son de anchura fija, y cada marca del control deslizante presenta 10 campos. Puede mover las marcas indicadoras a lo largo del control deslizante para visualizar los 10 ca,pos anteriores o siguientes, ordenados por importancia de predictor.

Puede efectuar una doble pulsación en el gráfico para abrir un cuadro de diálogo independiente en el que se pueden editar los valores del gráfico. Por ejemplo, puede corregir elementos tales como el tamaño del gráfico y el tamaño y color de los fonts utilizados. Al cerrar este cuadro de diálogo de edición independiente, los cambios se aplicarán al gráfico que se visualiza en la pestaña Salida.

### Correlograma

Se muestra un correlograma, o un gráfico de autocorrelación, para cada objetivo y muestra la función de autocorrelación (ACF), o la función de autocorrelación parcial (PACF), de los residuos (la diferencia entre valores esperados y reales) con respecto a los retardos temporales. El intervalo de confianza se muestra como un resaltado en el diagrama.

### Estimaciones de parámetro

Se repite para cada objetivo, la tabla Estimaciones de parámetro muestra (donde procede) los detalles siguientes:

- Nombre de objetivo
- La transformación aplicada
- Los retardos utilizados para este parámetro en el modelo (ARIMA)
- El valor de coeficiente
- El error estándar de la estimación de parámetro
- El valor de la estimación de parámetro divido por el error estándar
- Nivel de significación para la estimación del parámetro.

## Configuración del nugget del modelo Serie temporal

La pestaña Configuración proporciona opciones adicionales para el nugget del modelo Serie temporal.

### Previsión

La opción para **Extender registros en el futuro** establece el número de intervalos de tiempo para prever más allá del final del periodo de estimación. El intervalo de tiempo en este caso es el intervalo de tiempo del análisis, que se especifica en la pestaña Especificaciones de datos del nodo Serie temporal. Cuando se solicitan previsiones, se generan modelos autorregresivos automáticamente para cualquier serie de entrada que tampoco son objetivos. Estos modelos se utilizan para generar valores para estas series de entrada en el periodo de previsión.

**Calcular valores futuros de entradas.** Si selecciona esta opción, se calculan los valores de previsión para predictores, predicciones de ruido, estimación de varianza y futuros valores de tiempo.

### Valores futuros que se utilizarán en la previsión

- **Calcular valores futuros de entradas** Si selecciona esta opción, los valores de previsión para predictores, predicciones de ruido, estimación de varianza y futuros valores de tiempo se calculan automáticamente. Cuando se solicitan previsiones, se generan modelos autorregresivos automáticamente para cualquier serie de entrada que tampoco son objetivos. Estos modelos se utilizan para generar valores para estas series de entrada en el periodo de previsión.
- **Seleccionar los campos cuyos valores se desean añadir a los datos.** Para cada registro que desea prever (excluyendo los casos reservados), si está utilizando campos predictores (con el rol establecido en `Entrada`), puede especificar los valores estimados para el periodo de previsión para cada predictor. Puede introducir los valores manualmente o seleccionarlos de una lista.
  - **Campo.** Pulse en el botón selector de campos y seleccione los campos que se utilizarán como predictores. Tenga en cuenta que los campos aquí seleccionados podrán utilizarse o no en el modelado; para utilizar un campo como predictor debe seleccionarse en un nodo de modelado posterior. Este cuadro de diálogo permite especificar los valores futuros de forma cómoda para que puedan compartirlos varios nodos de modelado posteriores sin tener que especificarlos por separado en cada nodo. También tenga en cuenta que la lista de campos disponibles puede restringirse mediante las selecciones realizadas en la pestaña Opciones de generación.

Tenga en cuenta que si se especifican valores futuros para un campo que ya no está disponible en la ruta (porque se ha eliminado o debido a las selecciones actualizadas realizadas en la pestaña Opciones de generación), el campo se muestra en rojo.

– **Valores.** Para cada campo, puede seleccionar en una lista de funciones o pulsar en **Especificar** para introducir los valores manualmente o desde una lista de valores predefinidos. Si los campos predictores se refieren a los elementos que están bajo su control, o que se conocen de otra forma, debe introducir los valores manualmente. Por ejemplo, si está haciendo previsiones de los ingresos del próximo mes para un hotel a partir del número de reservas, puede especificar el número de reservas que tiene para ese período. Por el contrario, si un campo predictor está relacionado con algo que escapa a su control, como el precio de las acciones, puede utilizar una función como el valor más reciente o la media de los puntos recientes.

  Las funciones disponibles dependen del nivel de medición del campo.

*Tabla 29. Funciones disponibles para los niveles de medición*

| Nivel de medición | Funciones |
|---|---|
| Campo continuo o nominal | Vacío<br>Media de los puntos recientes<br>Valor más reciente<br>Especifica |
| Campo de marcas | Vacío<br>Valor más reciente<br>True<br>False<br>Especifica |

**Media de los puntos recientes** calcula el valor futuro a partir de la media de los tres últimos puntos de datos.

**Valor más reciente** establece el valor futuro en el del punto de datos más reciente.

**Verdadero/falso** establece el valor futuro de un campo de marcas en Verdadero o Falso según lo especificado.

**Especificar** abre un cuadro de diálogo para especificar los valores futuros manualmente o eligiéndolos en una lista predefinida.

## Dejar disponible para puntuación

**Crear nuevos campos para cada modelo que se puntuará.** Permite especificar los nuevos campos que se crearán para cada modelo que se puntuará.

- **Residuos de ruido.** Si se selecciona, creará un nuevo campo (con el prefijo predeterminado $TSResidual-) para los residuos de modelo para cada campo objetivo, junto con un total de estos valores.
- **Límites de confianza superior e inferior.** Si se selecciona, esta opción crea campos nuevos (con los prefijos predeterminados $TSLCI- y $TSUCI-) para los intervalos de confianza inferior y superior, respectivamente, para cada campo objetivo, junto con los totales de estos valores.

**Objetivos incluidos para puntuación.** Seleccione los objetivos disponibles que se incluirán en la puntuación del modelo.

# Capítulo 14. Nodos de modelo de respuesta de autoaprendizaje

## Nodo SLRM

El nodo de **modelo de respuesta de autoaprendizaje** (SLRM) permite generar un modelo que se puede actualizar o volver a estimar continuamente a medida que crece el conjunto de datos, sin necesidad de volver a generarlo cada vez con el conjunto de datos completo. Por ejemplo, esta posibilidad es útil cuando se tienen varios productos y se desea identificar el producto que es más probable que compre un cliente si se le ofrece. Este modelo permite predecir qué ofertas son las más apropiadas para los clientes y la probabilidad de que sean aceptadas.

El modelo se puede generar inicialmente con un conjunto de datos pequeño con ofertas realizadas aleatoriamente y las respuestas a éstas. A medida que crece el conjunto de datos, el modelo se puede actualizar y, de ese modo, aumenta su capacidad para predecir las ofertas más adecuadas para los clientes y la probabilidad de aceptación basada en otros campos de entrada como edad, sexo, trabajo e ingresos. Las ofertas disponibles se pueden cambiar añadiéndolas al cuadro de diálogo del nodo o eliminándolas de éste, en lugar de tener que cambiar el campo objetivo del conjunto de datos.

Junto con IBM SPSS Collaboration and Deployment Services, puede configurar actualizaciones periódicas automáticas del modelo. Este proceso, que no necesita supervisión humana, proporciona una solución rentable y flexible para las organizaciones y aplicaciones en las que no es necesaria o posible una intervención personalizada del analizador de datos.

**Ejemplo.** Una institución financiera desea obtener resultados más rentables adaptando a cada cliente la oferta que es más posible que acepte. Puede utilizar el modelo de autoaprendizaje para identificar las características de los clientes que es más probable que respondan favorablemente teniendo en cuenta las promociones anteriores y actualizar el modelo en tiempo real en función de las últimas respuestas de los clientes.

## Opciones de los campos del nodo SLRM

Antes de ejecutar un nodo SLRM, debe especificar los campos objetivo y de respuesta objetivo en la pestaña Campos del nodo.

**Campo objetivo.** Selecciona el campo objetivo de la lista; por ejemplo, un campo nominal (conjunto) que contiene diversos productos que desea ofrecer a los clientes.

*Nota*: El campo objetivo debe tener almacenamiento de cadena, no numérico.

**Campo de respuesta objetivo.** Seleccione el campo de respuesta objetivo de la lista. Por ejemplo, Aceptado o Rechazado.

*Nota*: este campo debe ser de marcas. El valor para verdadero de la marca indica la aceptación de la oferta y el valor para falso el rechazo de la oferta.

Los campos restantes de este cuadro de diálogo son los que se utilizan normalmente en IBM SPSS Modeler. Consulte el tema “Opciones de los campos del nodo de modelado” en la página 31 para obtener más información.

*Nota*: si los datos de origen incluyen rangos que se van a utilizar como campos de entrada continuos (rango numérico), debe asegurarse de que los metadatos incluyen los datos mínimos y máximos para cada rango.

## Opciones de modelo del nodo SLRM

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Continuar entrenando modelo existente.** De forma predeterminada, cada vez que se ejecuta un nodo de modelado, se crea un modelo completamente nuevo. Si esta opción está seleccionada, el entrenamiento continúa con el último modelo generado correctamente por el nodo. Esto permite actualizar un modelo existente sin tener que acceder a los datos originales. Además, puede dar como resultado un rendimiento significativamente más rápido ya que *sólo* se introducen en la ruta los registros nuevos o actualizados. Los detalles del modelo anterior se almacenan con el nodo de modelado, lo que permite utilizar esta opción incluso si el nugget de modelo anterior ya no está disponible en la ruta o la paleta de modelos.

**Valores del campo objetivo** De forma predeterminada, está establecido en **Utilizar todos**, lo que indica que se generará un modelo que contenga todas las ofertas asociadas al valor del campo objetivo seleccionado. Si desea generar un modelo que únicamente contenga algunas de las ofertas del campo objetivo, pulse en **Especificar** y utilice los botones **Añadir**, **Editar** y **Eliminar** para introducir o modificar los nombres de las ofertas para las que desea generar un modelo. Por ejemplo, si elige un objetivo que incluya todos los productos que suministra, puede utilizar este campo para limitar los productos ofrecidos a unos pocos que introducirá aquí.

**Evaluación del modelo.** Los campos de este panel son independientes del modelo, ya que no afectan a la puntuación. En su lugar, permiten crear una representación visual de la forma en la que el modelo predecirá los resultados.

*Nota*: Para mostrar los resultados de evaluación del modelo en el nugget de modelo, debe seleccionar también la casilla **Mostrar evaluación del modelo**.

- **Incluir evaluación de modelos.** Seleccione esta casilla para crear gráficos que muestren la precisión predicha del modelo para cada oferta seleccionada.
- **Establecer semilla aleatoria.** Cuando se estima la precisión de un modelo a partir de un porcentaje aleatorio, esta opción permite duplicar los mismos resultados en otra sesión. Al especificar el valor inicial utilizado por el generador de números aleatorios, puede garantizar que se asignan los mismos registros cada vez que se ejecuta el nodo. Introduzca el valor deseado. Si no se selecciona esta opción, se generará una muestra diferente cada vez que se ejecute el nodo.
- **Tamaño de muestra simulada.** Especifique el número de registros que se utilizarán en la muestra al evaluar el modelo. El valor predeterminado es 100.
- **Número de iteraciones.** Permite detener la generación de la evaluación de modelos tras un número de iteraciones especificado. Especifique el número máximo de iteraciones, el valor predeterminado es 20.

  *Nota*: tenga en cuenta que los tamaños grandes de muestra y los números elevados de iteraciones incrementarán el tiempo necesario para la generación del modelo.

**Mostrar evaluación del modelo.** Seleccione esta opción para mostrar una representación gráfica de los resultados en el nugget de modelo.

## Opciones de configuración del nodo SLRM

Las opciones de configuración del nodo permiten ajustar el proceso de generación de modelos.

**Número máximo de predicciones por registro** Esta opción le permite limitar el número de predicciones realizadas para cada registro del conjunto de datos. El valor predeterminado es 3.

Por ejemplo, si tiene seis ofertas (por ejemplo, ahorros, hipoteca, préstamo para coche, pensión, tarjeta de crédito y seguro) pero sólo quiere saber las dos que es preferible recomendar, deberá establecer este campo como 2. Cuando genere el modelo y lo conecte a una tabla, deberá ver dos columnas de predicciones (y la confianza asociada a la probabilidad de que se acepte la oferta) por registro. Las predicciones puedan estar compuestas por cualquiera de las seis posibles ofertas.

**Nivel de aleatorización** Para evitar cualquier sesgo, por ejemplo en un conjunto de datos pequeño o incompleto, y tratar todas las posibles ofertas por igual, puede añadir un nivel de aleatorización a la selección de las ofertas y la probabilidad de que éstas se incluyan como ofertas recomendadas. La aleatorización se expresa como un porcentaje, mostrado como valores decimales entre 0,0 (sin aleatorización) y 1,0 (completamente aleatorio). El valor predeterminado es 0,0.

**Establecer semilla aleatoria** Al añadir un nivel de aleatorización a la selección de una oferta, es posible duplicar los mismos resultados obtenidos en otra sesión. Al especificar el valor inicial utilizado por el generador de números aleatorios, puede garantizar que se asignan los mismos registros cada vez que se ejecuta el nodo. Introduzca el valor deseado. Si no se selecciona esta opción, se generará una muestra diferente cada vez que se ejecute el nodo.

**Nota:** Cuando se utiliza la opción **Establecer semilla aleatoria** con registros leídos de una base de datos, puede ser necesario un nodo Ordenar, antes del muestreo con el fin de garantizar el mismo resultado cada vez que se ejecute el nodo. Esto se debe a que la semilla aleatoria depende del orden de registros, sin estar garantizado que sea el mismo en una base de datos relacional.

**Orden de clasificación** Seleccione el orden en el que deben mostrarse las ofertas en el modelo generado:

- **Descendente** El modelo muestra primero las ofertas con las puntuaciones más altas. Éstas son las ofertas que tienen la mayor probabilidad de ser aceptadas.
- **Ascendente** El modelo muestra primero las ofertas con las puntuaciones más bajas. Éstas son las ofertas que tienen la mayor probabilidad de ser rechazadas. Por ejemplo, puede ser útil al decidir que clientes se deben eliminar de una campaña de marketing para una determinada oferta.

**Preferencias de campos objetivo** Al generar un modelo, es posible que haya determinadas características de los datos que desee eliminar o dar más importancia activamente. Por ejemplo, si está generando un modelo para seleccionar la mejor oferta financiera para enviar publicidad a un cliente, tal vez desee asegurarse de que se incluye siempre dicha oferta concreta, independientemente de la puntuación que obtenga para cada cliente.

Para incluir una oferta en este panel y editar sus preferencias, pulse en **Añadir**, escriba el nombre de la oferta (por ejemplo, Ahorros o Hipoteca) y pulse en **Aceptar**.

- **Valor** Muestra el nombre de la oferta que ha añadido.
- **Preferencia** Especifique el nivel de preferencia que se aplicará a la oferta. La preferencia expresada como porcentaje, mostrado como valores decimales entre 0,0 (no preferido) y 1,0 (el más preferido). El valor predeterminado es 0,0.
- **Incluir siempre** Para asegurarse de que se incluye siempre una oferta específica en las predicciones, active esta casilla.

  **Nota:** si la **Preferencia** se establece en 0,0, se ignorará la configuración de **Incluir siempre**.

**Tener en cuenta fiabilidad del modelo** Un modelo bien estructurado y con suficientes datos que se haya ajustado con precisión generándolo varias veces proporcionará siempre resultados más precisos que un modelo totalmente nuevo con pocos datos. Para aprovechar la mayor fiabilidad de un modelo más maduro, active esta casilla.

## Nugget de modelo SLRM

*Nota*: los resultados sólo aparecen en esta pestaña si selecciona **Incluir evaluación del modelo** y **Mostrar evaluación del modelo** en la pestaña de opciones de Modelo.

Cuando se ejecuta una ruta que contiene un modelo SLRM, el nodo estima la precisión de las predicciones de cada valor del campo objetivo (oferta) y la importancia de cada predictor utilizado.

*Nota*: Si ha seleccionado **Continuar entrenando modelo existente** en la pestaña Modelo del nodo de modelo, la información que aparece en el nugget de modelo se actualiza cada vez que se vuelve a generar el modelo.

Para modelos generados mediante IBM SPSS Modeler 12.0 o posterior, la pestaña Modelo del nugget de modelo está dividida en dos columnas:

**Columna izquierda.**

- **Ver.** Cuando se tiene más de una oferta, seleccione la oferta para la que desea mostrar los resultados.
- **Rendimiento del modelo.** Muestra la precisión estimada del modelo para cada oferta. Este conjunto de prueba se genera mediante simulación.

**Columna derecha.**

- **Ver.** Seleccione si desea visualizar los detalles de **Asociación con respuesta** o **Importancia de variable**.
- **Asociación con respuesta.** Muestra la asociación (correlación) de cada predictor con la variable objetivo.
- **Importancia del predictor.** Indica la importancia relativa de cada predictor cuando se calcula el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Este gráfico puede interpretarse de la misma manera que otros modelos que muestran la importancia de predictor, aunque en el caso de SLRM el gráfico se genera mediante simulación por el algoritmo SLRM. Se realiza eliminando sucesivamente del modelo cada predictor y comprobando cómo afecta esto a la precisión del modelo. Consulte el tema "Importancia del predictor" en la página 44 para obtener más información.

## Configuración del modelo SLRM

La pestaña Configuración de un nugget de modelo SLRM especifica las opciones para modificar el modelo generado. Por ejemplo, puede utilizar el nodo SLRM para generar varios modelos diferentes con los mismos datos y la misma configuración y, a continuación, usar esta pestaña en cada modelo para modificar ligeramente la configuración y comprobar cómo afecta eso a los resultados.

**Nota:** Esta pestaña está sólo disponible después de que el nugget de modelo se haya añadido a una ruta.

**Número máximo de predicciones por registro** Esta opción le permite limitar el número de predicciones realizadas para cada registro del conjunto de datos. El valor predeterminado es 3.

Por ejemplo, si tiene seis ofertas (por ejemplo, ahorros, hipoteca, préstamo para coche, pensión, tarjeta de crédito y seguro) pero sólo quiere saber las dos que es preferible recomendar, deberá establecer este campo como 2. Cuando genere el modelo y lo conecte a una tabla, deberá ver dos columnas de predicciones (y la confianza asociada a la probabilidad de que se acepte la oferta) por registro. Las predicciones puedan estar compuestas por cualquiera de las seis posibles ofertas.

**Nivel de aleatorización** Para evitar cualquier sesgo, por ejemplo en un conjunto de datos pequeño o incompleto, y tratar todas las posibles ofertas por igual, puede añadir un nivel de aleatorización a la selección de las ofertas y la probabilidad de que éstas se incluyan como ofertas recomendadas. La aleatorización se expresa como un porcentaje, mostrado como valores decimales entre 0,0 (sin aleatorización) y 1,0 (completamente aleatorio). El valor predeterminado es 0,0.

**Establecer semilla aleatoria** Al añadir un nivel de aleatorización a la selección de una oferta, es posible duplicar los mismos resultados obtenidos en otra sesión. Al especificar el valor inicial utilizado por el generador de números aleatorios, puede garantizar que se asignan los mismos registros cada vez que se ejecuta el nodo. Introduzca el valor deseado. Si no se selecciona esta opción, se generará una muestra diferente cada vez que se ejecute el nodo.

**Nota:** Cuando se utiliza la opción **Establecer semilla aleatoria** con registros leídos de una base de datos, puede ser necesario un nodo Ordenar, antes del muestreo con el fin de garantizar el mismo resultado cada vez que se ejecute el nodo. Esto se debe a que la semilla aleatoria depende del orden de registros, sin estar garantizado que sea el mismo en una base de datos relacional.

**Orden de clasificación** Seleccione el orden en el que deben mostrarse las ofertas en el modelo generado:

- **Descendente** El modelo muestra primero las ofertas con las puntuaciones más altas. Éstas son las ofertas que tienen la mayor probabilidad de ser aceptadas.
- **Ascendente** El modelo muestra primero las ofertas con las puntuaciones más bajas. Éstas son las ofertas que tienen la mayor probabilidad de ser rechazadas. Por ejemplo, puede ser útil al decidir que clientes se deben eliminar de una campaña de marketing para una determinada oferta.

**Preferencias de campos objetivo** Al generar un modelo, es posible que haya determinadas características de los datos que desee eliminar o dar más importancia activamente. Por ejemplo, si está generando un modelo para seleccionar la mejor oferta financiera para enviar publicidad a un cliente, tal vez desee asegurarse de que se incluye siempre dicha oferta concreta, independientemente de la puntuación que obtenga para cada cliente.

Para incluir una oferta en este panel y editar sus preferencias, pulse en **Añadir**, escriba el nombre de la oferta (por ejemplo, Ahorros o Hipoteca) y pulse en **Aceptar**.

- **Valor** Muestra el nombre de la oferta que ha añadido.
- **Preferencia** Especifique el nivel de preferencia que se aplicará a la oferta. La preferencia expresada como porcentaje, mostrado como valores decimales entre 0,0 (no preferido) y 1,0 (el más preferido). El valor predeterminado es 0,0.
- **Incluir siempre** Para asegurarse de que se incluye siempre una oferta específica en las predicciones, active esta casilla.

  **Nota:** si la **Preferencia** se establece en 0,0, se ignorará la configuración de **Incluir siempre**.

**Tener en cuenta fiabilidad del modelo** Un modelo bien estructurado y con suficientes datos que se haya ajustado con precisión generándolo varias veces proporcionará siempre resultados más precisos que un modelo totalmente nuevo con pocos datos. Para aprovechar la mayor fiabilidad de un modelo más maduro, active esta casilla.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

# Capítulo 15. Modelos de máquina de vectores de soporte

## Acerca de SVM

SVM (máquina de vectores de soporte) es una técnica de clasificación y regresión que aprovecha al máximo la precisión de las predicciones de un modelo sin ajustar excesivamente los datos de entrenamiento. SVM es ideal para analizar datos con un gran número de campos de predictores (por ejemplo, miles).

SVM tiene aplicaciones en multitud de disciplinas, incluyendo la gestión de relaciones con los clientes (CRM), el reconocimiento facial y de otras imágenes, bioinformática, extracción de conceptos de minería de texto, detección de intrusiones, predicción de estructura de proteínas y reconocimiento de la voz.

## Funcionamiento de SVM

SVM funciona correlacionando datos a un espacio de características de grandes dimensiones de forma que los puntos de datos se puedan categorizar, incluso si los datos no se puedan separar linealmente de otro modo. Se detecta un separador entre las categorías y los datos se transforman de forma que el separador se puede extraer como un hiperplano. Tras ello, las características de los nuevos datos se pueden utilizar para predecir el grupo al que pertenece el nuevo registro.

Por ejemplo, imagine la siguiente figura, en la que los puntos de datos corresponden a dos categorías diferentes.

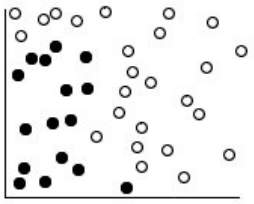


*Figura 59. Conjunto de datos original*

Las dos categorías se pueden separar con una curva, como se muestra en la siguiente figura.

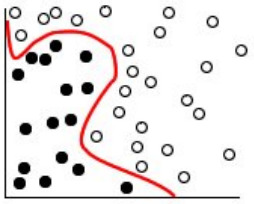


*Figura 60. Datos con un separador añadido*

Tras la transformación, el límite entre las dos categorías se puede definir por un hiperplano, como se muestra en la siguiente figura.

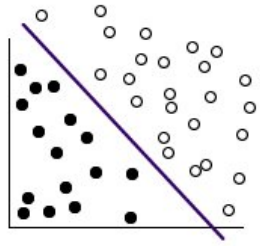


*Figura 61. Datos transformados*

La función matemática utilizada para la transformación se conoce como función **kernel**. SVM en IBM SPSS Modeler admite los siguientes tipos de kernel:

- Lineal
- Polinómico
- Función de base radial (RBF)
- Sigmoide

Una función kernel lineal es recomendable si la separación lineal de los datos es sencilla. En otros casos, se debe utilizar una del resto de las funciones. Deberá experimentar con las diferentes funciones para obtener el mejor modelo en cada caso, ya que utilizan algoritmos y parámetros diferentes.

## Ajuste de un modelo SVM

Además de la línea de separación entre las categorías, un modelo SVM de clasificación también encuentra líneas marginales que definen el espacio entre las dos categorías.

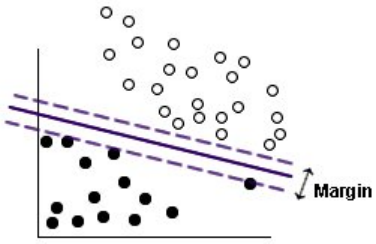



*Figura 62. Datos con un modelo preliminar*

Los puntos de datos que están en los márgenes se conocen como **vectores de soporte**.

Cuanto más amplio sea el margen entre las dos categorías, mejor será el modelo para predecir la categoría de nuevos registros. En el ejemplo anterior, el margen no es muy amplio y el modelo se conoce como **ajuste en exceso**. Se puede aceptar una pequeña cantidad de clasificación errónea para ampliar el margen; de muestra un ejemplo de esto en la figura siguiente.

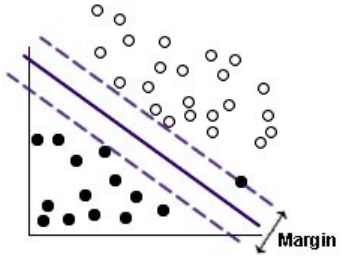



*Figura 63. Datos con un modelo mejorado*

En algunos casos, la separación lineal es más difícil; un ejemplo de ello se muestra en la siguiente figura.

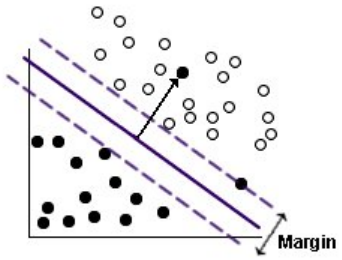



*Figura 64. Un problema de separación lineal*

En un caso como este, el objetivo es encontrar el equilibrio óptimo entre un margen amplio y un pequeño número de puntos de datos clasificados erróneamente. La función kernel tiene un **parámetro de regularización** (denominado C) que controla el equilibrio entre estos dos valores. Probablemente necesitará experimentar con diferentes valores de este y de otros parámetros kernel para encontrar el mejor modelo.

## Nodo SVM

El nodo SVM permite utilizar una máquina de vectores de soporte para clasificar los datos. SVM es ideal para conjuntos de datos grandes, es decir, con un gran número de campos predictores. Puede utilizar la configuración predeterminada en el nodo para producir un modelo básico relativamente rápido o puede utilizar la configuración de Experto para experimentar con tipos diferentes del modelo SVM.

Cuando haya creado el modelo, podrá:

- Explorar el nugget de modelo para representar la importancia relativa de los campos de entrada en la generación del modelo.
- Añadir un nodo Tabla al nugget del modelo para ver el resultado del modelo.

**Ejemplo.** Un investigador médico ha obtenido un conjunto de datos con las características de un número de muestras de células humanas extraídas de pacientes con riesgo de desarrollar un cáncer. El análisis de los datos originales demostró que muchas de las características de las muestras benignas y malignas eran muy diferentes. El investigador quiere desarrollar un modelo SVM que pueda utilizar los valores de características de casillas similares en las muestras de otros pacientes para indicar si las muestras pueden ser benignas o malignas.

## Opciones de modelo del nodo SVM

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Crear modelos divididos.** Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como campos de división. Consulte el tema "Generación de modelos divididos" en la página 28 para obtener más información.

## Opciones de experto del nodo SVM

Las opciones de experto le permiten ajustar el proceso de entrenamiento, si tiene amplios conocimientos sobre máquinas de vectores de soporte. Para acceder a estas opciones, active el modo **Experto** en la pestaña Experto.

**Añadir todas las probabilidades (válido para objetivos categóricos únicamente).** Si está seleccionada, especifica que las posibilidades de cada valor posible de un campo objetivo nominal o marca se representan para cada registro que procesa el nodo. Si no se ha seleccionado esta opción, sólo se representa la probabilidad del valor predicho de campos objetivo nominal o marca. El ajuste de esta casilla de verificación determina el estado predeterminado de la casilla de verificación correspondiente de la representación de nugget de modelo.

**Criterios de detención**. Determina cuándo detener el algoritmo de optimización. Los valores van del 1.0E-1 a 1.0E-6; el valor predeterminado es 1.0E-3. Reducir el valor da como resultado un modelo más preciso, pero el modelo tardará más tiempo en entrenar.

**Parámetro de regularización (C)**. Controla el equilibrio entre la maximización del margen y la minimización del término de error de entrenamiento. El valor normalmente debe estar entre 1 y 10 inclusive; el valor predeterminado es 10. Si aumenta el valor, mejorará la precisión (o reducirá el error de regresión) de los datos de formación, pero también puede suponer un ajuste por exceso.

**Precisión de regresión (epsilon)**. Sólo se utiliza si el nivel de medición del campo objetivo es *Continuo*. Causa errores que serán aceptables si son inferiores al valor especificado. Si aumenta el valor, el modelado será más rápido, pero en detrimento de la precisión.

**Tipo Kernel**. Determina el tipo de función kernel utilizada para la transformación. Los diferentes tipos de kernel causan que el separador se calcule de diferentes formas, por lo que es aconsejable que experimente con las diferentes opciones. El valor predeterminado es **RBF** (Función de base radial).

**RBF gamma**. Sólo se activa si el tipo de kernel está definido como **RBF**. El valor suele oscilar entre $3/k$ y $6/k$, donde $k$ es el número de campos de entrada. Por ejemplo, si hay 12 campos de entrada, los valores entre 0,25 y 0,5 serán significativos. Si aumenta el valor, mejorará la precisión (o reducirá el error de regresión) de los datos de formación, pero también puede suponer un ajuste por exceso.

**Gamma**. Sólo se activa si el tipo de kernel está definido como **Polinómico** o **Sigmoide**. Si aumenta el valor, mejorará la precisión (o reducirá el error de regresión) de los datos de formación, pero también puede suponer un ajuste por exceso.

**Sesgo**. Sólo se activa si el tipo de kernel está definido como **Polinómico** o **Sigmoide**. Define el valor `coef0` en la función kernel. El valor predeterminado 0 es el adecuado en la mayoría de los casos.

**Grado**. Sólo se activa si el tipo de kernel está definido como **Polinómico**. Controla la complejidad (dimensión) del espacio de correlación. Normalmente no utilizará un valor superior a 10.

## Nugget de modelo SVM

El modelo SVM crea nuevos campos. El campo más importante es **$S-nombrecampo**, que muestra el valor de campo objetivo que predice el modelo.

El número y los nombres de los nuevos campos que crea el modelo dependen del nivel de medición del campo objetivo (este campo se indica en las siguientes tablas como *nombredecampo*).

Para ver estos campos y sus valores, añada un nodo Tabla al nugget de modelo SVM y ejecute el nodo Tabla.

*Tabla 30. El nivel de medición del campo objetivo es "Nominal" o "Marca".*

| **Nombre del campo nuevo** | **Descripción** |
|---|---|
| $S-*nombredecampo* | Valor predicho del campo objetivo. |
| $SP-*nombredecampo* | Probabilidad del valor predicho. |
| $SP-*valor* | La probabilidad de cada valor posible nominal o marca (sólo se representa si **Añadir todas las probabilidades** está seleccionada en la pestaña Configuración del nugget de modelo). |
| $SRP-*valor* | (Sólo objetivos de marca) Las puntuaciones de propensión en bruto (SRP) y ajustadas (SAP), que indican la posibilidad de un resultado "true" del campo objetivo. Estas puntuaciones sólo se representan si se han seleccionado las casillas de verificación correspondientes en la pestaña Analizar del nodo de modelado SVM antes de que se genere el modelo. Consulte el tema “Opciones de análisis del nodo de modelado” en la página 35 para obtener más información. |
| $SAP-*valor* | |

*Tabla 31. El nivel de medición del campo objetivo es "Continuo"*

| **Nombre del campo nuevo** | **Descripción** |
|---|---|
| $S-*nombredecampo* | Valor predicho del campo objetivo. |

Importancia del predictor

Opcionalmente, en la pestaña Modelo también se puede mostrar un gráfico que indique la importancia relativa de cada predictor cuando se calcule el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Tenga en cuenta que este gráfico sólo está disponible si se ha seleccionado **Calcular importancia de predictor** en la pestaña Analizar antes de generar el modelo. Consulte el tema “Importancia del predictor” en la página 44 para obtener más información.

*Nota:* La importancia del predictor puede requerir más tiempo de cálculo para SVM que otros tipos de modelos y de forma predeterminada no se selecciona en la pestaña Analizar. Si selecciona esta opción se reducirá el rendimiento, especialmente con conjuntos de datos más grandes.

## Configuración de modelo SVM

La pestaña Configuración permite especificar campos extra que se mostrarán cuando se visualizan los resultados (por ejemplo ejecutando un nodo Tabla adjunto al nugget). Puede ver el efecto de cada una de estas opciones seleccionándolas y pulsando en el botón Presentación preliminar (desplácese a la derecha de los resultados de la presentación preliminar para ver los campos extra).

**Añadir todas las probabilidades (válido para objetivos categóricos únicamente).** Si está seleccionada, especifica que las posibilidades de cada valor posible de un campo objetivo nominal o marca se representan para cada registro que procesa el nodo. Si no está seleccionada, únicamente se representará el valor predicho y su probabilidad para campos objetivo nominal o marca.

El valor predeterminado de esta casilla de verificación está determinado por la casilla de verificación correspondiente en el nodo de modelado.

**Calcular puntuaciones de propensión en bruto.** En el caso de modelos con un objetivo de marca (que devuelve una predicción de sí o no), puede solicitar puntuaciones de propensión que indican la probabilidad del resultado true especificado para el campo objetivo. Éstas se añaden a otros valores de predicción y confianza que pueden generarse durante la puntuación.

**Calcular puntuaciones de propensión ajustada.** Las puntuaciones de propensión en bruto se basan sólo en los datos de entrenamiento y pueden ser demasiado optimistas debido a la tendencia de muchos modelos a sobreajustar estos datos. Las propensiones ajustadas intentan compensar evaluando el rendimiento del modelo frente a una partición de comprobación o validación. Esta opción requiere que se haya definido un campo de partición en la ruta y que se hayan activado puntuaciones ajustadas de propensión en el modo de modelado antes de generar el modelo.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se lleva a cabo la generación de SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.
- **Puntuar fuera de la base de datos** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

# Nodo LSVM

El nodo LSVM permite utilizar una máquina de vectores de soporte lineal para clasificar los datos. LSVM es especialmente adecuado para utilizarse con conjuntos de datos amplios, es decir, aquellos con un gran número de campos de predictor. Puede utilizar los valores predeterminados en el nodo para producir un modelo básico de forma relativamente rápida o puede utilizar las opciones de generación para experimentar con valores diferentes.

El nodo LSVM es similar al nodo SVM, pero es lineal y es mejor en el manejo de un gran número de registros.

Cuando haya creado el modelo, podrá:

- Explorar el nugget de modelo para representar la importancia relativa de los campos de entrada en la generación del modelo.
- Añadir un nodo Tabla al nugget del modelo para ver el resultado del modelo.

**Ejemplo.** Un investigador médico ha obtenido un conjunto de datos con las características de un número de muestras de células humanas extraídas de pacientes con riesgo de desarrollar un cáncer. El análisis de los datos originales demostró que muchas de las características de las muestras benignas y malignas eran muy diferentes. El investigador quiere desarrollar un modelo LSVM que pueda utilizar los valores de características de células similares en las muestras de otros pacientes para dar una indicación temprana de si las muestras pueden ser benignas o malignas.

## Opciones de modelo de nodo LSVM

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Calcular importancia del predictor.** En el caso de modelos que produzcan una medida adecuada de importancia, puede mostrar un gráfico que indique la importancia relativa de cada predictor al estimar el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Tenga en cuenta que puede tardarse más tiempo en calcular la importancia del predictor para algunos modelos, especialmente al trabajar con conjuntos de datos de gran tamaño; además, como resultado está desactivada para algunos modelos de forma predeterminada. La importancia del predictor no está disponible para modelos de listas de decisiones. Consulte el tema “Importancia del predictor” en la página 44 para obtener más información.

# Opciones de generación de LSVM

## Configuración del modelo

**Incluir interceptación**. Incluir la interceptación (el término constante en el modelo) puede aumentar la precisión global de la solución. Si se puede dar por supuesto la lectura de datos en el origen, se puede excluir la interceptación.

**Orden de clasificación para destino categórico**. Especifica el orden de clasificación del objetivo categórico. Este valor se ignora para los objetivos continuos.

**Precisión de regresión (epsilon)**. Sólo se utiliza si el nivel de medición del campo objetivo es *Continuo*. Causa errores que serán aceptables si son inferiores al valor especificado. Si aumenta el valor, el modelado será más rápido, pero en detrimento de la precisión.

**Excluir registros con cualquier valor perdido**. Cuando se establece en **True**, un registro se excluye si falta algún valor único.

## Valores de penalización

**Función de penalización**. Especifica el tipo de función de penalización que se utiliza para reducir la probabilidad de sobreajuste. Las opciones son **L1** o **L2**.

**L1** y **L2** reducen la posibilidad de sobreajuste añadiendo una penalización en los coeficientes. La diferencia entre estos es que cuando existe un gran número de características, **L1** utiliza la selección de características estableciendo algunos coeficientes en 0 durante la creación del modelo. **L2** no tiene esta prestación, así que no se debería utilizar, si se tiene un gran número de características.

**Pena de parámetro (lambda)**. Especifica el parámetro de penalización (regularización). Este valor está habilitado si se ha establecido **Función de penalización**.

---

# Nugget de modelo LSVM (resultado interactivo)

Después de ejecutar un modelo LSVM, está disponible la salida siguiente.

## Información del modelo

La vista de información de modelo proporciona información clave acerca del modelo. La tabla identifica algunos ajustes de modelo de alto nivel, por ejemplo:

- El nombre del destino especificado en la pestaña Campos
- El método de generación de modelos especificado en la configuración de Selección de modelos
- El número de entrada de predictores
- El número de predictores del modelo final
- El tipo de regularización (L1 o L2)
- El parámetro de penalización (lambda). Este es el parámetro de regularización.

- La precisión de regresión (épsilon). Los errores se aceptan si son menores que este valor. Un valor más alto puede producir un modelado más rápido, pero en detrimento de la precisión. Sólo se utiliza sólo si el nivel de medición del campo objetivo es *Continuo*.
- El porcentaje de precisión de la clasificación. Esto sólo es aplicable para la clasificación.
- El error cuadrático de promedio. Esto sólo es aplicable para Regresión.

### Resumen de registros

La vista Resumen de registros proporciona información acerca del número y porcentaje de registros (casos) incluidos y excluidos del modelo.

### Importancia del predictor

Es normal centrar los esfuerzos de modelado en los campos predictores más importantes y valorar la omisión de aquellos con menor relevancia. El gráfico de importancia de los predictores le ayuda a hacerlo indicando la importancia relativa de cada predictor en la estimación del modelo. Como los valores son relativos, la suma de valores de todos los predictores de la visualización es 1.0. La importancia del predictor no está relacionada con la precisión del modelo. Sólo está relacionada con la importancia de cada predictor a la hora de realizar una predicción, no con si la predicción es o no precisa.

### Predicho por observado

Muestra un diagrama de dispersión en intervalos de los valores predichos en el eje vertical por los valores observados en el eje horizontal. Idealmente, los puntos deben basarse en una línea de 45 grados; esta vista indica si hay algún registro predicho de manera incorrecta en el modelo.

**Nota:** La importancia de predictor puede tardar más en el cálculo para LSVM y SVM que para otros tipos de modelos. Si selecciona esta opción se reducirá el rendimiento, especialmente con conjuntos de datos más grandes.

### Matriz de confusión

La matriz de confusión, que a veces se conoce como tabla de resumen, muestra el número de casos correcta e incorrectamente asignados a cada uno de los grupos basados en el análisis LSVM.

## Configuración de modelo LSVM

En la pestaña Configuración de un nugget de modelo SVLM, especifique las opciones para la propensión bruta y para la generación de SQL durante la puntuación de modelos. Esta pestaña sólo está disponible después de añadir el nugget de modelo a una ruta.

**Calcular puntuaciones de propensión bruta** Para modelos sólo con objetivos de marca, puede solicitar puntuaciones de propensión bruta que indiquen la probabilidad del resultado verdadero especificado para el campo objetivo. Éstas se añaden a los valores estándar de predicción y confianza. Las puntuaciones ajustadas de propensión no están disponibles.

**Generar SQL para este modelo** Cuando se utilizan datos de una base de datos, se puede devolver código SQL a la base de datos para su ejecución, lo que proporciona un mayor rendimiento para muchas operaciones.

Seleccione una de las siguientes opciones para especificar cómo se genera SQL.

- **Valor predeterminado: Puntuar utilizando el adaptador de puntuación del servidor (si está instalado) de lo contrario en curso.** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación y las funciones definidas por el usuario (UDF) asociadas y se puntúa el modelo dentro de la base de datos. Si no hay ningún adaptador de puntuación disponible, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

- **Puntuar fuera de la base de datos.** Si se selecciona, esta opción capta los datos de la base de datos y los puntúa en SPSS Modeler.

# Capítulo 16. Modelos de vecinos más próximos

## Nodo KNN

Análisis de vecino más próximo es un método de clasificación de casos basado en su similaridad con otros casos. En aprendizaje de máquinas, se ha desarrollado como una forma de reconocer patrones de datos sin requerir una coincidencia exacta con patrones o casos almacenados. Los casos parecidos están próximos y los que no lo son están alejados entre sí. Además, la distancia entre dos casos es una medida de sus diferencias.

Los casos próximos entre sí se denominan "vecinos". Cuando se presenta un nuevo caso (reserva), se calcula su distancia desde cada caso del modelo. Las clasificaciones de los casos más parecidos, los vecinos más próximos, se cuadran y el nuevo caso se incluye en la categoría que contiene el mayor número de vecinos más próximos.

Puede especificar el número de vecinos más próximos a examinar; este valor se denomina *k*. Las imágenes muestran cómo se clasificará un nuevo caso utilizando dos valores diferentes de *k*. Cuando $k = 5$, el nuevo caso se coloca en la categoría *1* porque una mayoría de los vecinos más próximos pertenecen a esa categoría *1*. Sin embargo, cuando $k = 9$, el nuevo caso se coloca en la categoría *0* porque una mayoría de los vecinos más próximos pertenecen a la categoría *0*.

El análisis de vecino más próximo también se puede utilizar para calcular los valores de un objetivo continuo. En esta situación, la media o el valor objetivo medio de los vecinos más próximos se utiliza para obtener el valor predicho del nuevo caso.

## Opciones de objetivos del nodo KNN

En la pestaña Objetivos podrá seleccionar si desea generar un modelo que prediga el valor de un campo objetivo en sus datos de entrada en función de los valores de sus vecinos más cercanos, o simplemente para encontrar los vecinos más cercanos de un caso concreto.

¿Qué tipo de análisis desea realizar?

**Predecir un campo de destino.** Seleccione esta opción si desea predecir el valor de un campo de destino en función de los valores de sus vecinos más próximos.

**Identificar sólo los vecinos más próximos.** Seleccione esta opción si sólo desea ver los vecinos más próximos de un campo de entrada concreto.

Si selecciona identificar únicamente los vecinos más cercanos, se desactivan las opciones restantes de esta pestaña relativas a la velocidad y precisión, ya que sólo son relevantes para destinos de predicciones.

¿Cuál es su objetivo?

Al predecir un campo objetivo, este grupo de opciones permite decidir si la velocidad, precisión o una mezcla de ambas son los factores más importantes a la hora de predecir un campo objetivo. También puede seleccionar personalizar la configuración por sí mismo.

Si selecciona la opción Equilibrar, Velocidad o Precisión, el algoritmo preselecciona la combinación de configuraciones más adecuada para esa opción. Es posible que los usuarios avanzados deseen sobrescribir estas selecciones; esta acción se puede realizar en los diferentes paneles de la pestaña Configuración.

**Equilibrar velocidad y precisión.** Selecciona el número de vecinos más adecuado en un rango pequeño.

**Velocidad.** Busca un número fijo de vecinos.

**Precisión.** Selecciona el mejor número de vecinos en un mayor rango y utiliza la importancia de predictor al calcular las distancias.

**Análisis personalizado.** Seleccione esta opción para ajustar con precisión el algoritmo en la pestaña Configuración.

**Nota:** El tamaño del modelo KNN resultante, a diferencia de la mayoría de los demás modelos, aumenta con la cantidad de datos de entrenamiento. Si, al intentar crear un modelo KNN, aparece un error que le informa de que se ha quedado sin memoria, pruebe a aumentar la memoria máxima del sistema utilizada por IBM SPSS Modeler. Para ello, seleccione

**Herramientas** > **Opciones** > **Opciones de sistema**

e introduzca el tamaño nuevo en el campo **Memoria máxima**. Los cambios realizados en el cuadro de diálogo Opciones de sistema no surtirán efecto hasta que no reinicie IBM SPSS Modeler.

## Ajustes del nodo KNN

En la pestaña Configuración puede especificar las opciones específicas del análisis de vecino más próximo. La barra de la izquierda contiene la lista de paneles que utiliza para especificar las opciones.

### Modelo

El panel Modelo proporcionan las opciones que controlan cómo se va a generar el modelo, por ejemplo, si se utilizarán modelos de partición o divididos, si se transformarán los campos de entrada numéricos para que todos estén dentro del mismo rango y cómo se gestionarán los casos de interés. También puede seleccionar un nombre personalizado para el modelo.

**Nota:** Las opciones **Utilizar los datos en particiones** y **Usar etiquetas de caso** no pueden usar el mismo campo.

**Nombre de modelo** Puede generar el nombre del modelo de forma automática basándose en el campo de destino o de ID (o en el nombre del tipo de modelo si se especifica ningún campo de destino), o bien especificar un nombre personalizado.

**Utilizar los datos en particiones.** Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen los datos de la partición de entrenamiento para la generación del modelo.

**Crear modelos divididos.** Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como campos de división. Consulte el tema "Generación de modelos divididos" en la página 28 para obtener más información.

**Para seleccionar campos manualmente...** De forma predeterminada, el nodo utiliza la partición y divide la configuración de campo (si existe) del nodo Tipo pero puede alterar temporalmente estos valores aquí. Para activar los campos **Partición** y **Divisiones**, seleccione la pestaña **Campos** y pulse **Utilizar configuración personalizada** para volver aquí.

- **Partición.** Este campo permite especificar un campo usado para dividir los datos en muestras independientes para las fases de entrenamiento, prueba y validación en la generación del modelo. Si usa una muestra para generar el modelo y otra muestra distinta para comprobarlo, podrá obtener una buena indicación de la bondad del modelo a la hora de generar conjuntos de datos de mayor tamaño similares a los datos actuales. Si se han definido varios campos de partición mediante nodos Tipo o Partición, se deberá seleccionar un campo de partición simple en la pestaña Campos en todos los nodos de modelado que usen la partición. (Si solamente hay una partición, se usará automáticamente siempre que se active la partición.) Debe tener en cuenta que al aplicar la partición seleccionada en su

análisis, también debe activar la partición en la pestaña Opciones del modelo para el nodo. (Si se elimina la selección de esta opción, se posibilita la desactivación de la partición sin cambiar la configuración del campo.)

- **Divididos.** En modelos divididos, seleccione el campo o campos de división. Se trata de una acción similar a establecer el rol del campo en *Dividir* en un nodo Tipo. Sólo puede diseñar campos de tipo **Marca**, **Nominal** u **Ordinal** como campos de división. Los campos seleccionados como campos de división no se pueden utilizar como campos de destino, entrada, partición, frecuencia o ponderación. Consulte el tema “Generación de modelos divididos” en la página 28 para obtener más información.

**Normalizar entradas de rango.** Seleccione esta casilla para normalizar los valores de campos de entrada continuos. Las características normalizadas tienen el mismo rango de valores, lo que puede mejorar el rendimiento del algoritmo de estimación. Se utilizará la normalización ajustada [2*(x−min)/(max−min)]-1. Los valores normalizados ajustados están entre -1 y 1.

**Utilizar etiquetas de casos.** Seleccione esta casilla para activar la lista desplegable desde la que podrá elegir un campo cuyos valores se usarán como etiquetas para identificar los casos de interés en el gráfico de espacio de predictores, gráfico de homólogos y mapa de cuadrantes en el visor de modelos. Puede seleccionar cualquier campo con un nivel de medición de *Nominal*, *Ordinal* o *Marca* para su uso como campo de etiqueta. Si no elige ningún campo aquí, los registros se mostrarán en los gráficos del visor de modelos con los vecinos más próximos identificados por número de fila en los datos de origen. Si va a manipular los datos después de crear el modelo, utilice etiquetas de caso para evitar tener que volver a los datos de origen cada vez que necesite identificar los casos en la visualización.

**Identificar registro focal.** Marque esta casilla para activar la lista desplegable, que le permite marcar un campo de entrada de un interés en particular (sólo para campos marca). Si especifica un campo aquí, los puntos que representan ese campo se seleccionan inicialmente en el visor de modelos cuando se construye el modelo. La selección de un registro focal aquí es opcional; cualquier punto puede convertirse temporalmente en un registro focal cuando se selecciona manualmente en el visor de modelos.

## Vecinos

El panel Vecinos tiene un conjunto de opciones que controlan cómo se calculará el número de vecinos más próximos.

**Número de vecinos más próximos (k)** Especifique el número de vecinos más próximos de un caso concreto. Tenga en cuenta que el uso de un número mayor de vecinos no implica que el modelo resultante sea más preciso.

Si el objetivo es predecir un objetivo, dispone de dos opciones:

- **Especificar k fijo.** Utilice esta opción si desea especificar un número fijo de vecinos más próximos que se buscarán.
- **Seleccionar k automáticamente.** También puede utilizar los campos **Mínimo** y **Máximo** para especificar un rango de valores y permitir que el procedimiento seleccione el "mejor" número de vecinos en ese rango. El método para determinar el número de vecinos más cercanos depende de si la selección de características se solicita en el panel Selección de características:

  Si la selección de características está activada, ésta se realizará para cada valor de *k* en el rango solicitado, y se seleccionará la *k* y el conjunto de características compañero con la menor tasa de error (o el menor error cuadrático si el destino es continuo).

  Si la selección de características no está activada, se utilizará la validación cruzada de pliegue en *V* para seleccionar el “mejor” número de vecinos. Consulte el panel Validación cruzada para tener más control sobre las asignaciones de veces.

**Cálculo de distancias.** Es la métrica utilizada para especificar la métrica de distancia empleada para medir la similitud de los casos.

- **Métrica euclídea.** La distancia entre dos casos, x e y, es la raíz cuadrada de la suma, sobre todas las dimensiones, de las diferencias cuadradas entre los valores de esos casos.

- **Métrica de bloques de ciudad.** La distancia entre dos casos es la suma, en todas las dimensiones, de las diferencias absolutas entre los valores de esos casos. También se conoce como la distancia de Manhattan.

Opcionalmente, si el objetivo es predecir un objetivo, podrá seleccionar ponderar características según la importancia normalizada al calcular las distancias. La importancia de características de un predictor se calcula por la relación de la tasa de errores, o el error de la suma de los cuadrados del modelo sin el predictor del modelo para la tasa de errores, o el error de la suma de los cuadrados para el modelo completo. La importancia normalizada se calcula volviendo a ponderar los valores de importancia de la característica para que sumen 1.

**Ponderar características por importancia al calcular distancias.** (Se muestra sólo si el objetivo es predecir un objetivo.) Active esta casilla de verificación para que la importancia del predictor se utilice al calcular las distancias entre los vecinos. La importancia del predictor se mostrará en el nugget del modelo y se utilizará en las predicciones (y afectará a la puntuación). Consulte el tema "Importancia del predictor" en la página 44 para obtener más información.

**Predicciones del destino de rango.** (Se muestra sólo si el objetivo es predecir un objetivo.) Si se especifica un objetivo continuo (rango numérico), defines si el valor predicho se calcula en función de la mediana o el valor de la mediana de los vecinos más próximos.

## Sel. características

Este panel se activa sólo si el objetivo es predecir un objetivo. Permite solicitar y especificar opciones de selección de características. De forma predeterminada, todas las características se tienen en cuenta para la selección de características, pero es posible seleccionar un subconjunto de características para forzarlas en el modelo.

**Realizar selección de características.** Seleccione esta casilla para activar las opciones de selección de características.

- **Entrada forzada.** Pulse el botón de selección de campos junto a esta casilla y seleccione una o más características que se forzarán en el modelo.

**Criterio de parada.** En cada paso, la característica cuya suma al modelo dé lugar al menor error (calculado como la tasa de error de un destino categórico y el error cuadrático de un objetivo continuo) se tiene en cuenta para su inclusión en el conjunto de modelos. La selección continúa hasta que se cumple la condición especificada.

- **Parar cuando se ha seleccionado el número especificado de características.** El algoritmo añade un número fijo de características además de las forzadas en el modelo. Especifique un número entero positivo. Si se disminuyen los valores de número que se puede seleccionar se obtiene un modelo más reducido, lo que supone el riesgo de perder importantes características. Si se aumentan los valores de número que se puede seleccionar se incluirán todas las características importantes, pero se corre el riesgo de añadir características que aumenten el error del modelo.
- **Parar cuando el cambio en el índice de errores absolutos sea inferior o igual al mínimo.** El algoritmo se detiene cuando el cambio de la tasa de errores absolutos indica que el modelo no puede mejorarse más añadiendo nuevas características. Especifique un número positivo. Los valores decrecientes del cambio mínimo tienden a incluir más características, con el riesgo de incluir características que no añaden demasiado valor al modelo. Si se aumentan los valores del cambio mínimo se excluirán más características, pero puede que se pierdan características importantes para el modelo. El valor "óptimo" del cambio mínimo dependerá de los datos y de la aplicación. Consulte el Registro de errores de selección de características en los resultados para poder evaluar qué características son más importantes. Consulte el tema "Registro de errores de selección de predictores" en la página 369 para obtener más información.

## Validación cruzada

Este panel se activa sólo si el objetivo es predecir un objetivo. Las opciones de este panel controlan si se utilizarán validación cruzada cuando se calculen los vecinos más próximos.

La validación cruzada divide la muestra en un número de submuestras o **pliegues**. A continuación, se generan los modelos de vecino más próximo, que no incluyen los datos de cada submuestra. El primer modelo se basa en todos los casos excepto los correspondientes al primer pliegue de la muestra; el segundo modelo se basa en todos los casos excepto los del segundo pliegue de la muestra y así sucesivamente. Para cada modelo se calcula el error aplicando el modelo a la submuestra que se excluyó al generarse este. El "mejor" número de vecinos más próximos será el que produzca el menor error entre los pliegues.

**Pliegues de validación cruzada.** La validación cruzada de pliegue en *V* se utiliza para determinar el "mejor" número de vecinos. Por razones de rendimiento, no está disponible con la selección de características.

- **Asignar casos a pliegues aleatoriamente.** Especifique el número de pliegues que se utilizarán para la validación cruzada. El procedimiento asigna aleatoriamente casos a los pliegues, numerados de 1 a *V*, que es el número de pliegues.
- **Establecer semilla aleatoria.** Cuando se estima la precisión de un modelo a partir de un porcentaje aleatorio, esta opción permite duplicar los mismos resultados en otra sesión. Al especificar el valor inicial utilizado por el generador de números aleatorios, puede garantizar que se asignan los mismos registros cada vez que se ejecuta el nodo. Introduzca el valor deseado. Si no se selecciona esta opción, se generará una muestra diferente cada vez que se ejecute el nodo.
- **Utilizar campo para asignar los casos.** Especifique un campo numérico que asigna cada caso en el conjunto de datos activo a un pliegue. El campo debe ser numérico y adoptar valores de 1 a *V*. Si falta algún valor de este rango, y en cualquier campo de segmento si los modelos de segmentación están en vigor, se producirá un error.

## Analizar

El panel Analizar se activa sólo si el objetivo es predecir un objetivo. Puede utilizarlo para especificar si el modelo incluirá variables adicionales que contienen:

- probabilidades para cada valor de campo objetivo posible
- distancias entre un caso y sus vecinos más próximos
- puntuaciones ajustadas y de propensión en bruto (sólo para objetivos de marca)

**Añadir todas las probabilidades.** Si está seleccionada, especifica que las posibilidades de cada valor posible de un campo objetivo nominal o marca se representan para cada registro que procesa el nodo. Si no está seleccionada, únicamente se representará el valor predicho y su probabilidad para campos objetivo nominal o marca.

**Guardar distancias entre casos focales y vecinos k más próximos.** En cada registro focal, se crea una variable diferente para cada uno de los vecinos más próximos *k* del registro focal (de la muestra de entrenamiento) y las distancias más cercanas *k* correspondientes.

**Puntuaciones de propensión**

Las puntuaciones de propensión pueden activarse en el nodo de modelado y en la pestaña Configuración del nugget de modelo. Esta funcionalidad sólo está disponible cuando el objetivo seleccionado es un campo de marca. Consulte el tema “Puntuaciones de propensión” en la página 36 para obtener más información.

**Calcular puntuaciones de propensión en bruto.** Las puntuaciones de propensión en bruto están derivadas del modelo basado únicamente en los datos de entrenamiento. Si el modelo predice el valor

*true* (responderá), la propensión es la misma que P, donde P es la probabilidad de la predicción. Si el modelo predice el valor false, la propensión se calcula como (1 - P).

- Si selecciona esta opción al crear el modelo, las puntuaciones de propensión se activarán en el nugget de modelo de forma predeterminada. Sin embargo, siempre puede activar las puntuaciones de propensión en bruto en el nugget de modelo independientemente de si las selecciona o no en el nodo de modelado.
- Al puntuar el modelo, se añadirán puntuaciones de propensión en bruto a un campo con las letras *RP* unidas al prefijo estándar. Por ejemplo, si las predicciones están en un campo denominado *$R-churn*, el nombre del campo de puntuación de propensión será *$RRP-churn*.

**Calcular puntuaciones de propensión ajustada.** Las propensiones brutas se basan totalmente en estimaciones proporcionadas por el modelo, las cuales pueden estar ajustadas excesivamente, lo que lleva a estimaciones de propensión demasiado optimistas. Las propensiones ajustadas intentan compensar este hecho observando el rendimiento del modelo en las particiones de comprobación o validación y ajustando las propensiones para proporcionar una mejor estimación en consecuencia.

- Esta configuración requiere que haya un campo de partición válido en la ruta.
- A diferencia de las puntuaciones brutas de confianza, las puntuaciones ajustadas de propensión deben calcularse al crear el modelo; de lo contrario, no estarán disponibles cuando se puntúe el nugget de modelo.
- Al puntuar el modelo, se añadirán puntuaciones ajustadas de propensión a un campo con las letras *AP* unidas al prefijo estándar. Por ejemplo, si las predicciones están en un campo denominado *$R-churn*, el nombre del campo de puntuación de propensión será *$RAP-churn*. Las puntuaciones ajustadas de propensión no están disponibles para modelos de regresión logística.
- Al calcular las puntuaciones ajustadas de propensión, la partición de comprobación o validación utilizada para el cálculo no debe haberse equilibrado. Para evitarlo, asegúrese de seleccionar la opción **Sólo datos de entrenamiento de equilibrado** en todos los nodos Equilibrar anteriores en la ruta. Además, si se ha llevado una muestra compleja a un punto anterior en la ruta, se invalidarán las puntuaciones ajustadas de propensión.
- Las puntuaciones ajustadas de propensión no están disponibles para modelos de árbol "aumentado" y de conjuntos de reglas. Consulte el tema “Modelos C5.0 aumentados” en la página 128 para obtener más información.

## Nugget de modelo KNN

El modelo KNN crea un número de campos nuevos, tal y como se muestra en la tabla siguiente. Para ver estos campos y sus valores, añada un nodo Tabla al nugget de modelo KNN y ejecute el nodo Tabla, o pulse en el botón Presentación preliminar en el nugget.

*Tabla 32. Campos de modelo KNN*

| Nombre del campo nuevo | Descripción |
|---|---|
| $KNN-*nombredecampo* | Valor predicho del campo objetivo. |
| $KNNP-*nombredecampo* | Probabilidad del valor predicho. |
| $KNNP-*valor* | La probabilidad de cada valor posible de un campo nominal o marca. Sólo se incluye si **Añadir todas las probabilidades** está seleccionada en la pestaña Configuración del nugget de modelo. |
| $KNN-vecino-*n* | El nombre del vecino más cercano *n* del registro focal. Se incluye sólo si **Mostrar más cercano** de la pestaña Configuración del nugget de modelo está definida a un valor distinto de cero. |
| $KNN-distancia-*n* | La distancia relativa del registro focal del vecino *n* más cercano al registro focal. Se incluye sólo si **Mostrar más cercano** de la pestaña Configuración del nugget de modelo está definida a un valor distinto de cero. |

# Vista de modelo de vecino más próximo

## Vista de modelo

La vista de modelos tiene una ventana con dos paneles:

- El primer panel muestra una descripción general del modelo denominado vista principal.
- El segundo panel muestra uno de los dos tipos de vistas:

  Una vista de modelos auxiliar muestra más información sobre el modelo, pero no se centra en el propio modelo.

  Una vista enlazada es una vista que muestra detalles sobre una característica del modelo cuando el usuario desglosa parte de la vista principal.

De forma predeterminada, el primer panel muestra el espacio predictor y el segundo muestra el gráfico de importancia de los predictores. Si el gráfico de importancia de predictor no está disponible, es decir cuando **Ponderar características por importancia** no se ha seleccionado en el panel Vecinos de la pestaña Configuración, se muestra la primera vista disponible en el desplegable Ver.

Cuando una vista no tiene información disponible, se omite del cuadro desplegable Ver.

**Espacio predictor:** El gráfico Espacio predictor es un gráfico interactivo del espacio predictor (o un subespacio, si hay más de 3 predictores). Cada eje representa un predictor del modelo, y la ubicación de los puntos del gráfico muestra los valores de dichos predictores para casos de las particiones de formación y reserva.

**Claves.** Además de los valores predictores, los puntos del gráfico indican otra información.

- La forma indica la partición a la que pertenece un punto, ya sea Entrenamiento o Reserva.
- El color y el sombreado de un punto indican el valor del destino de ese caso: cada valor de color distinto representa las categorías de un destino categórico y las sombras indican el rango de valores de un objetivo continuo. El valor indicado para la partición de entrenamiento es el valor observado, mientras que en el caso de la partición de reserva, representa el valor predicho. Si no se especifica ningún destino, esta clave no aparece.
- Los titulares más gruesos indican que un caso es focal. Los registros focales se muestran en relación con sus *k* vecinos más próximos.

**Controles e interactividad.** Una serie de controles del gráfico le permite explorar el espacio predictor.

- Puede seleccionar qué subconjunto de predictores mostrar en el gráfico y modificar qué predictores se representan en las dimensiones.
- Los "registros focales" son simplemente puntos seleccionados en el gráfico Espacio predictor. Si ha especificado una variable de registro focal, los puntos que representan los registros focales se seleccionarán inicialmente. Sin embargo, cualquier punto puede convertirse en un registro focal si lo selecciona. Se aplican los controles "normales" para la selección de puntos; al pulsar en un punto éste se selecciona y se cancela la selección de todos los demás; si pulsa Control y el ratón sobre un punto éste se añade al conjunto de puntos seleccionados. Las vistas enlazadas, como el gráfico Homólogos, se actualizarán automáticamente en función de los casos seleccionados en el espacio predictor.
- Puede modificar el número de vecinos más próximos (*k*) para mostrar registros focales.
- Al pasar el ratón sobre un punto del gráfico se mostrará una sugerencia con el valor de la etiqueta de caso o un número de caso si las etiquetas de caso no se definen, así como los valores de destino observados y predichos.
- Un botón "Restablecer" le permite devolver el espacio predictor a su estado original.

*Cambio de los ejes en el gráfico Espacio predictor:* Puede controlar qué características se muestran en los ejes del gráfico Espacio predictor.

Para cambiar la configuración de los ejes:

1. Pulse en el botón Modo edición (icono de pincel) en el panel de la izquierda para seleccionar el Modo edición de Espacio predictor.
2. Cambie la vista (a cualquier cosa) en el panel derecho. Aparecerá el panel **Mostrar zonas** entre los dos paneles principales.
3. Pulse en la casilla de verificación **Mostrar zonas**.
4. Pulse en cualquier punto de datos en Espacio predictor.
5. Para sustituir un eje por un predictor del mismo tipo de datos:
   - Arrastre el nuevo predictor sobre la etiqueta de zona (la que tiene el botón X pequeño) del que desee sustituir.
6. Para sustituir un eje por un predictor de un tipo de datos diferente:
   - En la etiqueta de zona del predictor que desea sustituir, pulse en el botón X pequeño. El espacio predictor cambiará a una visión bidimensional.
   - Arrastre el nuevo predictor sobre la etiqueta de zona **Añadir dimensión**.
7. Pulse en el botón Explorar modo (icono de punta de flecha) en el panel de la izquierda para salir del Modo edición.

**Importancia del predictor:** Es normal centrar los esfuerzos de modelado en los campos predictores más importantes y valorar la omisión de aquellos con menor relevancia. El gráfico de importancia de los predictores le ayuda a hacerlo indicando la importancia relativa de cada predictor en la estimación del modelo. Como los valores son relativos, la suma de valores de todos los predictores de la visualización es 1.0. La importancia del predictor no está relacionada con la precisión del modelo. Sólo está relacionada con la importancia de cada predictor a la hora de realizar una predicción, no con si la predicción es o no precisa.

**Distancias de vecinos más próximos:** Esta tabla muestra los *k* vecinos más próximos y las distancias de registros focales únicamente. Está disponible si se especifica un identificador de registro focal en la nodo de modelado y sólo muestra los registros focales identificados por esta variable.

Cada fila de:

- La columna **Registro focal** contiene el valor de la variable de etiqueta de caso del registro focal; si las etiquetas de caso no se definen, esta columna contendrá el número de caso del caso focal.
- La columna *i* del grupo **Vecinos más próximos** contiene el valor de la variable de etiquetas de casos del vecino más cercano *i* del registro focal; si las etiquetas del caso no están definidas, esta columna contiene el número de caso del vecino más cercano *i* del registro focal.
- La columna *i* del grupo **Distancias más cercanas** contiene la distancia de los vecinos más próximos *i* al registro focal

**Homólogos:** Este gráfico muestra los casos focales y sus *k* vecinos más próximos en cada predictor y en el destino. Está disponible si se selecciona un caso focal en Espacio predictor.

El gráfico Homólogos se vincula a Espacio predictor de dos maneras.

- Los casos seleccionados (focal) en Espacio predictor se muestran en el gráfico Homólogos, juntos con sus *k* vecinos más próximos.
- El valor de *k* seleccionado en Espacio predictor se utiliza en el gráfico Homólogos.

**Seleccionar predictores.** Permite seleccionar predictores en la visualización del gráfico Homólogos.

**Mapa de cuadrantes:** Este gráfico muestra los casos focales y sus *k* vecinos más próximos en un diagrama de dispersión (o gráfico de puntos, dependiendo del nivel de medición del destino) con el destino en el eje *y* y un predictor de escala en el eje *x*, panelado por predictores. Está disponible si hay un destino y se selecciona un caso focal en Espacio predictor.

- Se dibujan líneas de referencia para las variables continuas en las medias variables en la partición de entrenamiento.

**Seleccionar predictores.** Permite seleccionar predictores en la visualización del mapa de cuadrantes.

**Registro de errores de selección de predictores:** Señala en la vista de gráfico el error (la tasa de error o de error cuadrático, dependiendo del nivel de medición del destino) en el eje *y* para el modelo con el predictor enumerado en el eje *x* (además de todas las características a la izquierda del eje *x*). Este gráfico está disponible si hay un destino y la selección de características está activada.

**Tabla de clasificación:** Esta tabla muestra la clasificación cruzada de los valores observados en comparación con los valores predichos del destino, en función de la partición. Está disponible si hay un destino y es categórico (marca, nominal u ordinal).

- La fila **(Perdidos)** de la partición de reserva contiene casos de reserva con los valores perdidos en el destino. Estos casos contribuyen a los valores de Muestra de reserva: Porcentaje global, pero no a los valores de Porcentaje correcto.

**Resumen de error:** Esta tabla está disponible si hay una variable de destino. Muestra el error asociado con el modelo; la suma de cuadrados de un destino continuo y la tasa de error (100% - porcentaje global correcto) de un destino categórico.

## Ajustes de modelo KNN

La pestaña Configuración permite especificar campos extra que se mostrarán cuando se visualizan los resultados (por ejemplo ejecutando un nodo Tabla adjunto al nugget). Puede ver el efecto de cada una de estas opciones seleccionándolas y pulsando en el botón Presentación preliminar (desplácese a la derecha de los resultados de la presentación preliminar para ver los campos extra).

**Añadir todas las probabilidades (válido para objetivos categóricos únicamente).** Si está seleccionada, especifica que las posibilidades de cada valor posible de un campo objetivo nominal o marca se representan para cada registro que procesa el nodo. Si no está seleccionada, únicamente se representará el valor predicho y su probabilidad para campos objetivo nominal o marca.

El valor predeterminado de esta casilla de verificación está determinado por la casilla de verificación correspondiente en el nodo de modelado.

**Calcular puntuaciones de propensión en bruto.** En el caso de modelos con un objetivo de marca (que devuelve una predicción de sí o no), puede solicitar puntuaciones de propensión que indican la probabilidad del resultado true especificado para el campo objetivo. Éstas se añaden a otros valores de predicción y confianza que pueden generarse durante la puntuación.

**Calcular puntuaciones de propensión ajustada.** Las puntuaciones de propensión en bruto se basan sólo en los datos de entrenamiento y pueden ser demasiado optimistas debido a la tendencia de muchos modelos a sobreajustar estos datos. Las propensiones ajustadas intentan compensar evaluando el rendimiento del modelo frente a una partición de comprobación o validación. Esta opción requiere que se haya definido un campo de partición en la ruta y que se hayan activado puntuaciones ajustadas de propensión en el modo de modelado antes de generar el modelo.

**Mostrar más cercano.** Si define este valor a *n*, siendo *n* un entero positivo distinto de cero, los *n* vecinos más cercanos al registro focal se incluyen en el modelo, junto con sus distancias relativas al registro focal.

# Capítulo 17. Nodos Python

SPSS Modeler ofrece nodos para utilizar algoritmos nativos Python. La pestaña **Python** de la Paleta de nodos contiene los nodos siguientes que puede utilizar para ejecutar algoritmos Python. Estos nodos se soportan en Windows 64, Linux64 y Mac.

El nodo SMOTE (Synthetic Minority Over-sampling Technique) proporciona un algoritmo de sobremuestreo para tratar con conjuntos de datos desequilibrados. Proporciona un método avanzado para equilibrar los datos. El nodo de proceso SMOTE en SPSS Modeler se implementa en Python y necesita la biblioteca de Python `imbalanced-learn`©.

XGBoost Linear© es una implementación avanzada de un algoritmo de aumento de gradiente con un modelo lineal como modelo base. Los algoritmos de aumento conocen de forma iterativa los clasificadores débiles y, a continuación, los añaden a un clasificador fuerte final. El nodo XGBoost Linear en SPSS Modeler se implementa en Python.

XGBoost Tree© es una implementación avanzada de un algoritmo de aumento de gradiente con un modelo de árbol como modelo base. Los algoritmos de aumento conocen de forma iterativa los clasificadores débiles y, a continuación, los añaden a un clasificador fuerte final. XGBoost Tree es muy flexible y proporciona muchos parámetros que pueden ser abrumadores para la mayoría de usuarios, de modo que el nodo XGBoost Tree en SPSS Modeler expone las características principales y los parámetros utilizados comúnmente. El nodo se implementa en Python.

t-SNE (vecino estocástico con t distribuida incorporado) es una herramienta para visualizar datos de alta dimensión. Convierte afinidades de puntos de datos a probabilidades. Este nodo t-SNE en SPSS Modeler se implementa en Python y requiere la biblioteca `scikit-learn`© Python.

El nodo Bosque aleatorio utiliza una implementación avanzada de un algoritmo de agregación autodocimante con un modelo de árbol como modelo base. Este nodo de modelado de bosque aleatorio en SPSS Modeler se implementa en Python y requiere la biblioteca `scikit-learn`© Python.

El nodo SVM de una clase utiliza un algoritmo de aprendizaje no supervisado. El nodo se puede utilizar para la detección de novedad. Detectará el límite flexible de un conjunto de muestras proporcionado, para clasificar a continuación los puntos nuevos como pertenecientes o no a dicho conjunto. Este nodo de modelado SVM de una clase en SPSS Modeler se implementa en Python y necesita la biblioteca `scikit-learn`© de Python.

# Nodo SMOTE

El nodo SMOTE (Synthetic Minority Over-sampling Technique) proporciona un algoritmo de sobremuestreo para tratar con conjuntos de datos desequilibrados. Proporciona un método avanzado para equilibrar los datos. El nodo de proceso SMOTE se implementa en Python y necesita la biblioteca de Python `imbalanced-learn`©. Para obtener detalles sobre la biblioteca imbalanced-learn, consulte http://contrib.scikit-learn.org/imbalanced-learn/about.html[1].

La pestaña Python de la Paleta de nodos contiene el nodo SMOTE y otros nodos Python.

[1]Lemaître, Nogueira, Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research*, vol. 18, no. 17, 2017, pp. 1-5. (http://jmlr.org/papers/v18/16-365.html)

# Configuración de nodo SMOTE

Defina los valores siguientes en la pestaña **Configuración** del nodo SMOTE.

## Valor de destino

**Campo de destino.** Seleccione el campo de destino. Se soportan todos los tipos de medición de marca, nominal, ordinal y discreta. Si la opción **Utilizar los datos en particiones** está seleccionada en la sección Partición, sólo se sobremuestrearán los datos de entrenamiento.

## Razón de sobremuestreo

Seleccione **Automático** para seleccionar automáticamente una razón de sobremuestreo o seleccione **Establecer cociente (minoritario con respecto a mayoritario)** para establecer un valor de cociente personalizado. El cociente es el número de muestras en la clase de minoría sobre el número de muestras en la clase de mayoría. El valor debe ser mayor que **0** y menor que o igual a **1**.

## Semilla aleatoria

**Establecer semilla aleatoria.** Seleccione esta opción y pulse **Generar** para generar la semilla utilizada por el generador de números aleatorios.

## Métodos

**Tipo de algoritmo.** Seleccione el tipo de algoritmo SMOTE que desea utilizar.

## Reglas de muestras

**Vecinos K.** Especifique el número de vecinos más próximos a utilizar para construir muestras sintéticas.

**Vecinos M.** Especifique el número de vecinos más próximos a utilizar para determinar si una muestra de minoría está en peligro. Esto sólo se utilizará si se selecciona el tipo de algoritmo SMOTE **Borderline1** o **Borderline1**.

## Partición

**Utilice datos en particiones.** Seleccione esta opción si sólo desea que se sobremuestreen datos de entrenamiento.

El nodo SMOTE necesita la biblioteca de Python `imbalanced-learn`©. La tabla siguiente muestra la relación entre los valores del diálogo de nodo SMOTE de SPSS Modeler y el algoritmo de Python.

*Tabla 33. Propiedades de nodo correlacionadas con parámetros de biblioteca de Python*

| Valor de SPSS Modeler | Nombre de script (nombre de propiedad) | Nombre de parámetro de API Python |
|---|---|---|
| Razón de sobremuestreo (control de entrada de número) | sample_ratio_value | ratio |
| Semilla aleatoria | random_seed | random_state |
| Vecinos K | k_neighbours | k |
| Vecinos M | m_neighbours | m |
| Tipo de algoritmo | algorithm_kind | kind |

## Nodo XGBoost Linear

XGBoost Linear© es una implementación avanzada de un algoritmo de aumento de gradiente con un modelo lineal como modelo base. Los algoritmos de aumento conocen de forma iterativa los clasificadores débiles y, a continuación, los añaden a un clasificador fuerte final. El nodo XGBoost Linear en SPSS Modeler se implementa en Python.

Para obtener más información sobre los algoritmo de aumento, consulte los tutoriales de XGBoost disponibles en http://xgboost.readthedocs.io/en/latest/tutorials/index.html. [1]

Tenga en cuenta que la función de validación cruzada de XGBoost no se soporta en SPSS Modeler. Puede utilizar el nodo Partición de SPSS Modeler para esta funcionalidad. También tenga en cuenta que XGBoost en SPSS Modeler ejecuta la codificación One-Hot automáticamente para variables categóricas.

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

## Campos de nodo XGBoost Linear

La pestaña Campos especifica los campos que se utilizan en el análisis.

**Utilizar roles predefinidos.** Esta opción utiliza la configuración de rol (objetivos, predictores, etc) de un nodo de Tipo anterior (o la pestaña Tipos de un nodo de origen anterior).

**Utilizar asignaciones de campos personalizadas.** Para asignar manualmente un objetivo y predictores, seleccione esta opción.

**Campos.** Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los campos de roles de Objetivo y Predictores a la derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol. Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

**Objetivo.** Seleccione un campo para utilizarlo como objetivo de la predicción.

**Predictores.** Seleccione uno o más campos como entradas de la predicción.

## Opciones de generación de nodo XGBoost Linear

Utilice la pestaña Opciones de generación para especificar opciones de generación para el nodo XGBoost Linear, incluyendo **opciones básicas** como parámetros de aumento lineal y generación de modelos, y **opciones de tareas de aprendizaje** para objetivos. Para obtener información adicional sobre estas opciones, consulte los recursos en línea siguientes:

- Referencia de parámetro XGBoost[1]
- API Python XGBoost[2]

- Página de inicio de XGBoost[3]

## Básico

**Optimización de hiper-parámetro (basada en Rbfopt).** Seleccione esta opción para habilitar la optimización de hiper-parámetro basada en Rbfopt, que descubre automáticamente la combinación óptima de parámetros, de forma que el modelo conseguirá el índice de error previsto o inferior en las muestras. Si desea detalles sobre Rbfopt, consulte http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html.

**Alfa.** Término de regularización L1 en ponderaciones. Al aumentar este valor, el modelo será más conservador.

**Lambda.** Término de regularización L2 en ponderaciones. Al aumentar este valor, el modelo será más conservador.

**Lambda bias.** Término de regularización L2 en sesgo. (No hay ningún término de regularización L1 en sesgo porque no es importante.)

**Número de ronda de aumento.** Número de iteraciones de aumento.

## Tarea de aprendizaje

**Objetivo.** Seleccione en los siguientes tipos de objetivo de tarea de aprendizaje: **reg:linear**, **reg:logistic**, **reg:gamma**, **reg:tweedie**, **count:poisson**, **rank:pairwise**, **binary:logistic** o **multi**.

**Semilla aleatoria.** Puede pulsar **Generar** para generar la semilla utilizada por el generador de números aleatorios.

La tabla siguiente muestra la relación entre los valores del diálogo de nodo XGBoost Linear de SPSS Modeler y los parámetros de biblioteca XGBoost de Python.

*Tabla 34. Propiedades de nodo correlacionadas con parámetros de biblioteca de Python*

| Valor de SPSS Modeler | Nombre de script (nombre de propiedad) | Parámetro de XGBoost |
|---|---|---|
| Objetivo | `TargetField` | |
| Predictores | `InputFields` | |
| Lambda | `lambda` | `lambda` |
| Alfa | `alpha` | `alpha` |
| Lambda bias | `lambdaBias` | `lambda_bias` |
| Redondeo de aumento de número | `numBoostRound` | `num_boost_round` |
| Objetivo | `objectiveType` | `objective` |
| Semilla aleatoria | `random_seed` | `seed` |

[1] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[2] "Plotting API" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[3] "Scalable and Flexible Gradient Boosting." Web. © 2015-2016 DMLC.

## Opciones de modelo de nodo XGBoost Linear

**Nombre del modelo.** Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

---

# Nodo XGBoost Tree

XGBoost Tree© es una implementación avanzada de un algoritmo de aumento de gradiente con un modelo de árbol como modelo base. Los algoritmos de aumento conocen de forma iterativa los clasificadores débiles y, a continuación, los añaden a un clasificador fuerte final. XGBoost Tree es muy flexible y proporciona muchos parámetros que pueden ser abrumadores para la mayoría de usuarios, de modo que el nodo XGBoost Tree en SPSS Modeler expone las características principales y los parámetros utilizados comúnmente. El nodo se implementa en Python.

Para obtener más información sobre los algoritmo de aumento, consulte los tutoriales de XGBoost disponibles en http://xgboost.readthedocs.io/en/latest/tutorials/index.html. [1]

Tenga en cuenta que la función de validación cruzada de XGBoost no se soporta en SPSS Modeler. Puede utilizar el nodo Partición de SPSS Modeler para esta funcionalidad. También tenga en cuenta que XGBoost en SPSS Modeler ejecuta la codificación One-Hot automáticamente para variables categóricas.

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

## Campos de nodo XGBoost Tree

La pestaña Campos especifica los campos que se utilizan en el análisis.

**Utilizar roles predefinidos.** Esta opción utiliza la configuración de rol (objetivos, predictores, etc) de un nodo de Tipo anterior (o la pestaña Tipos de un nodo de origen anterior).

**Utilizar asignaciones de campos personalizadas.** Para asignar manualmente un objetivo y predictores, seleccione esta opción.

**Campos.** Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los campos de roles de Objetivo y Predictores a la derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol. Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

**Objetivo.** Seleccione un campo para utilizarlo como objetivo de la predicción.

**Predictores.** Seleccione uno o más campos como entradas de la predicción.

## Opciones de generación de nodo XGBoost Tree

Utilice la pestaña Opciones de generación para especificar opciones de generación para el nodo XGBoost Tree, incluyendo **opciones básicas** para la generación de modelos y el crecimiento del árbol, **opciones de tarea de aprendizaje** para objetivos y **opciones avanzadas** para el sobreajuste de control y el manejo de conjuntos de datos desequilibrados. Para obtener información adicional sobre estas opciones, consulte los recursos en línea siguientes:

- Referencia de parámetro XGBoost[1]
- API Python XGBoost[2]
- Página de inicio de XGBoost[3]

## Básico

**Optimización de hiper-parámetro (basada en Rbfopt).** Seleccione esta opción para habilitar la optimización de hiper-parámetro basada en Rbfopt, que descubre automáticamente la combinación óptima de parámetros, de forma que el modelo conseguirá el índice de error previsto o inferior en las muestras. Si desea detalles sobre Rbfopt, consulte http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html.

**Método de árbol.** Seleccione el algoritmo de construcción de árbol de XGBoost a utilizar.

**Redondeo de aumento de número.** Especifique el número de iteraciones de aumento.

**Profundidad máxima** Especifique la profundidad máxima para árboles. Al aumentar este valor, el modelo se hará más complejo y será más probable que se sobreajuste.

**Ponderación hijo mínima.** Especifique la suman mínima de ponderación de instancia (hessiana) necesaria en un hijo. Si el paso de partición de árbol produce un nodo hoja con la suma de la ponderación de instancia menor que esta **Ponderación hijo mínima**, el proceso de generación detendrá el particionamiento adicional. En modo de regresión lineal, este simplemente corresponde al número mínimo de instancias necesarias en cada nodo. Cuando mayor es la ponderación, más conservador será el algoritmo.

**Paso delta máximo.** Especifique el paso delta máximo que se debe permitir para la estimación ponderada de cada árbol. Si se establece en **0**, no hay ninguna restricción. Si se establece en un valor positivo, puede ayudar a que el paso de actualización sea más conservador. Generalmente este parámetro no es necesario, pero puede ayudar en la regresión logística cuando una clase es extremadamente desequilibrada.

## Tarea de aprendizaje

**Objetivo.** Seleccione en los siguientes tipos de objetivo de tarea de aprendizaje: **reg:linear**, **reg:logistic**, **reg:gamma**, **reg:tweedie**, **count:poisson**, **rank:pairwise**, **binary:logistic** o **multi**.

**Semilla aleatoria.** Puede pulsar **Generar** para generar la semilla utilizada por el generador de números aleatorios.

## Avanzado

**Submuestra.** Submuestra en la razón de la instancia de entrenamiento. Por ejemplo, si establece esto en **0,5**, XGBoost recolectará aleatoriamente la mitad de las instancias de datos para hacer crecer árboles y esto evitará el sobreajuste.

**Eta.** La reducción de tamaño de paso utilizada durante el paso de actualización para evitar el sobreajuste. Después de cada paso de aumento, se pueden obtener directamente las ponderaciones de nuevas características. Eta también reduce las ponderaciones de característica para que el proceso de aumento sea más conservador.

**Gamma.** La reducción de pérdida mínima necesaria para realizar una partición adicional en un nodo hoja del árbol. Cuanto mayor es el valor de gamma, más conservador será el algoritmo.

**Muestra de columna por árbol.** Razón de submuestra de columnas cuando se construye cada árbol.

**Muestra de columna por nivel.** Razón de submuestra de columnas para cada división, en cada nivel.

**Lambda.** Término de regularización L2 en ponderaciones. Al aumentar este valor, el modelo será más conservador.

**Alfa.** Término de regularización L1 en ponderaciones. Al aumentar este valor, el modelo será más conservador.

**Ponderación de posición de escala.** Controlar el equilibrio de ponderaciones positivas y negativas. Esto es útil para clases desequilibradas.

La tabla siguiente muestra la relación entre los valores del diálogo de nodo XGBoost Tree de SPSS Modeler y los parámetros de biblioteca XGBoost de Python.

*Tabla 35. Propiedades de nodo correlacionadas con parámetros de biblioteca de Python*

| Valor de SPSS Modeler | Nombre de script (nombre de propiedad) | Parámetro de XGBoost |
|---|---|---|
| Objetivo | `TargetField` | |
| Predictores | `InputFields` | |
| Método de árbol | `treeMethod` | `tree_method` |
| Redondeo de aumento de número | `numBoostRound` | `num_boost_round` |
| Profundidad máxima | `maxDepth` | `max_depth` |
| Ponderación hijo mínima | `minChildWeight` | `min_child_weight` |
| Paso delta máximo | `maxDeltaStep` | `max_delta_step` |
| Objetivo | `objectiveType` | `objective` |
| Semilla aleatoria | `random_seed` | `seed` |
| Submuestra | `sampleSize` | `subsample` |
| Eta | `eta` | `eta` |
| Gamma | `gamma` | `gamma` |
| Muestra de columna por árbol | `colsSampleRatio` | `colsample_bytree` |
| Muestra de columna por nivel | `colsSampleLevel` | `colsample_bylevel` |
| Lambda | `lambda` | `lambda` |
| Alfa | `alpha` | `alpha` |
| Ponderación de posición de escala | `scalePosWeight` | `scale_pos_weight` |

[1] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[2] "Plotting API" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[3] "Scalable and Flexible Gradient Boosting." Web. © 2015-2016 DMLC.

## Opciones de modelo de nodo XGBoost Tree

**Nombre del modelo.** Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

# Nodo t-SNE

La incorporación de un vecino estocástico con t distribuida (t-SNE)© es una herramienta para visualizar datos de alta dimensión. Convierte afinidades de puntos de datos a probabilidades. Las afinidades del espacio original se representan mediante probabilidades conjuntas gausianas y las afinidades del espacio incorporado se representan mediante distribuciones de t de Student. Esto permite que t-SNE sea especialmente sensible a la estructura local y tenga otras ventajas sobre técnicas existentes: [1]

- Revelar la estructura en muchas escalas en una sola correlación

- Revelar datos que residen en muchos y distintos colectores o clústeres
- Reducir la tendencia de amontonar puntos juntos en el centro

El nodo t-SNE en SPSS Modeler se implementa en Python y requiere la biblioteca `scikit-learn`© Python. Si desea detalles sobre t-SNE y la biblioteca scikit-learn, consulte:

- https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html#sklearn.manifold.TSNE
- https://scikit-learn.org/stable/modules/manifold.html#t-sne

La pestaña Python en la Paleta de nodos contiene este nodo y otros nodos Python. El nodo t-SNE también está disponible en la pestaña Gráficos.

[1] Referencias:

van der Maaten, L.J.P.; Hinton, G. "Visualizing High-Dimensional Data using t-SNE." Journal of Machine Learning Research. 9:2579-2605, 2008.

van der Maaten, L.J.P. "t-Distributed Stochastic Neighbor Embedding."

van der Maaten, L.J.P. "Accelerating t-SNE using Tree-Based Algorithms." Journal of Machine Learning Research. 15(Oct):3221-3245, 2014.

## Opciones de Experto del nodo t-SNE

Elija la modalidad **Simple** o la modalidad **Experto** en función de las opciones que desea establecer para el nodo t-SNE.

**Tipo de visualización.** Seleccione **2D** o **3D** para especificar si desea trazar el gráfico como dos dimensiones o tres dimensiones.

**Método.** Seleccione **Barnes Hut** o **Exacto**. De forma predeterminada, el algoritmo de cálculo de gradiente utiliza la aproximación Barnes-Hut que se ejecuta mucho más rápido que el método Exacto. La aproximación Barnes-Hut permite que la técnica t-SNE se aplique a conjuntos de datos grandes y del mundo real. El algoritmo Exacto realizará un mejor trabajo al impedir errores de vecino más cercano.

**Inic.** Seleccione **Aleatorio** o **PCA** para la inicialización de la incorporación.

**Campo de destino.** Seleccione el campo objetivo para mostrar como un mapa de colores en el gráfico de salida. El gráfico utilizará un color si aquí no se ha especificado ningún campo objetivo.

### Optimización

**Perplejidad.** La perplejidad está relacionada con el número de vecinos más cercanos que se utilizan en otros algoritmos de aprendizaje de colector. Normalmente, los conjuntos de datos más grandes necesitaban una perplejidad mayor. Considere seleccionar un valor entre **5** y **50**. El valor predeterminado es **30** y el rango es **2 - 9999999**.

**Exageración temprana.** Este valor controla cómo de ajustados están los clústeres naturales del espacio original en el espacio incorporado, y cuánto espacio habrá entre ellos. El valor predeterminado es **12**, y el rango es **2 - 9999999**.

**Índice de aprendizaje.** Si el índice de aprendizaje es demasiado alto, los datos podrían tener un aspecto de "pelota" con todos los puntos aproximadamente equidistantes con respecto a sus vecinos más cercanos. Si el índice de aprendizaje es demasiado bajo, la mayoría de los puntos pueden parecer comprimidos en

una nube densa con unos pocos valores atípicos. Si la función de coste se atasca en un mínimo local incorrecto, aumentar el índice de aprendizaje puede ayudar. El valor predeterminado es **200**, y el rango es **0 - 9999999**.

**Máx de iteraciones.** El número máximo de iteraciones para la optimización. El valor predeterminado es **1000**, y el rango es **250 - 9999999**.

**Tamaño angular.** El tamaño angular de un nodo distante se mide a partir de un punto. Especifique un valor entre **0** y **1**. El valor predeterminado es **0.5**.

## Semilla aleatoria

**Establecer semilla aleatoria.** Seleccione esta opción y pulse **Generar** para generar la semilla utilizada por el generador de números aleatorios.

## Condición de detención de optimización

**Máx de iteraciones sin progreso.** El número máximo de iteraciones sin progreso para realizar antes de detener la optimización, se utiliza después de 250 iteraciones iniciales con una exageración temprana. Tenga en cuenta que el progreso solo se comprueba cada 50 iteraciones, así que este valor se redondea hasta el siguiente múltiplo de 50. El valor predeterminado es **300**, y el rango es **0 - 9999999**.

**Mín de norma gradiente.** Si la norma gradiente está por debajo de este umbral mínimo, la optimización se detendrá. El valor predeterminado es **1.0E-7**.

**Métrica.** La métrica para utilizar al calcular la distancia entre instancias en una matriz de características. Si la métrica es una cadena, debe ser una de las opciones permitidas por `scipy.spatial.distance.pdist` para su parámetro de métrica, o una métrica listada en `pairwise.PAIRWISE_DISTANCE_FUNCTIONS`. Seleccione uno de los tipos de métrica disponibles. El valor predeterminado es **euclidean**.

**Cuando el número de registros es mayor que.** Especifique un método para trazar conjuntos de datos grandes. Puede especificar un tamaño de conjunto de datos máximo o utilizar los 2.000 puntos predeterminados. El rendimiento se mejora para conjuntos de datos grandes cuando se selecciona las opciones **Intervalo** o **Muestra**. De forma alternativa, puede elegir trazar todos los puntos de datos seleccionando **Utilizar todos los datos**, pero debe tener en cuenta que esto puede degradar significativamente el rendimiento del software.

- **Intervalo.** Seleccione esta opción para habilitar los intervalos cuando el conjunto de datos contiene más registros que el número especificado. Los intervalos dividen el gráfico en cuadrículas finas antes de trazar y contar realmente el número de conexiones que aparecerían en cada una de las celdas de la cuadrícula. En el gráfico final, se utiliza una conexión por celda en el centroide del intervalo (promedio de todos los puntos de conexión del intervalo).
- **Muestra.** Seleccione esta opción para muestrear aleatoriamente los datos en el número especificado de registros.

La tabla siguiente muestra las relaciones entre los valores de la pestaña Experto del diálogo del nodo t-SNE de SPSS Modeler y los parámetros de biblioteca t-SNE de Python.

*Tabla 36. Propiedades de nodo correlacionadas con parámetros de biblioteca de Python*

| Valor de SPSS Modeler | Nombre de script (nombre de propiedad) | Parámetro t-SNE Python |
|---|---|---|
| Modo | `mode_type` | |
| Tipo de visualización | `n_components` | `n_components` |
| Método | `method` | `method` |
| Inicialización de incorporación | `init` | `init` |

*Tabla 36. Propiedades de nodo correlacionadas con parámetros de biblioteca de Python (continuación)*

| Valor de SPSS Modeler | Nombre de script (nombre de propiedad) | Parámetro t-SNE Python |
|---|---|---|
| Objetivo | target_field | target_field |
| Perplejidad | perplexity | perplexity |
| Exageración temprana | early_exaggeration | early_exaggeration |
| Índice de aprendizaje | learning_rate | learning_rate |
| Máx de iteraciones | n_iter | n_iter |
| Tamaño angular | angle | angle |
| Establecer semilla aleatoria | enable_random_seed | |
| Semilla aleatoria | random_seed | random_state |
| Máx de iteraciones sin progreso | n_iter_without_progress | n_iter_without_progress |
| Mín de norma gradiente | min_grad_norm | min_grad_norm |
| Realizar t-SNE con varias perplejidades | isGridSearch | |

## Opciones de salida de nodo t-SNE

Especifique opciones para la salida del nodo t-SNE en la pestaña **Salida**.

**Nombre de salida.** Especifique el nombre de la salida que se genera cuando se ejecuta el nodo. Si selecciona **Auto**, el nombre de salida se establece automáticamente.

**Resultados en pantalla.** Seleccione esta opción para generar y mostrar los resultados en una nueva ventana. El resultado también se añade al gestor de salidas.

**Resultados a archivo.** Seleccione esta opción para guardar el resultado en un archivo. Esto habilita los campos **Nombre de archivo** y **Tipo de archivo**. El nodo t-SNE requiere acceso a este archivo de salida si desea crear gráficos utilizando otros campos para fines de comparación – o para utilizar su salida como predictores en modelos de clasificación o de regresión. El modelo t-SNE crea un archivo de resultado de los campos de coordenadas x, y (y z) al que se acceso más fácilmente mediante un nodo de origen Archivo fijo. Si desea más información, consulte @@@@@@.

La tabla siguiente muestra la relación entre los valores en la pestaña Salida del diálogo del nodo t-SNE de SPSS Modeler y los parámetros de biblioteca t-SNE de Python.

*Tabla 37. Propiedades de nodo correlacionadas con parámetros de biblioteca de Python*

| Valor de SPSS Modeler | Nombre de script (nombre de propiedad) | Parámetro t-SNE Python |
|---|---|---|
| Nombre de resultado | output_Rename | output_Rename |
| Modalidad de salida | output_to | output_to |
| Nombre de archivo | full_filename | full_filename |
| Tipo de archivo | output_file_type | output_file_type |
| Objetivo | target_field | target_field |

## Nuggets de modelo t-SNE

Los nuggets de modelo t-SNE contienen toda la información capturada por el modelo t-SNE. Las pestañas siguientes están disponibles.

### Gráfico

La pestaña **Gráfico** muestra la salida del gráfico para el nodo t-SNE. Un gráfico de dispersión pyplot muestra el resultado de bajas dimensiones. Si no ha seleccionado la opción **Realizar t-SNE con varias perplejidades** en la pestaña Experto del nodo t-SNE, solo se incluye un gráfico, en lugar de seis gráficos con distintas perplejidades.

### Salida de texto

La pestaña **Salida de texto** muestra los resultados del algoritmo t-SNE. Si selecciona el tipo de visualización **2D** en la pestaña Experto del nodo t-SNE, aquí el resultado es el valor de punto en dos dimensiones. Si elige **3D**, el resultado es el valor de punto en tres dimensiones.

---

## Nodo Bosque aleatorio

Random Forest© es una implementación avanzada de un algoritmo de agregación autodocimante con un modelo de árbol como modelo base. En los bosques aleatorios, cada árbol del conjunto se genera a partir de una muestra trazada con sustitución (por ejemplo, una muestra de programa de arranque) a partir del conjunto de entrenamiento. Al dividir un nodo durante la construcción del árbol, la división elegida deja de ser la mejor división entre todas las características. En su lugar, la división que es elige es la mejor división entre un conjunto aleatorio de las características. Debido a esta aleatoriedad, por regla general el sesgo del bosque aumenta ligeramente (con respecto al sesgo de un único árbol no aleatorio) pero, debido al promedio, su varianza también disminuye, normalmente más que la compensación para el aumento en el sesgo, por este motivo genera un modelo mejor en general.[1]

El nodo Bosque aleatorio en SPSS Modeler se implementa en Python. La pestaña Python en la Paleta de nodos contiene este nodo y otros nodos Python.

Si desea más información sobre algoritmos de bosque aleatorio, consulte https://scikit-learn.org/stable/modules/ensemble.html#forest.

[1]L. Breiman, "Random Forests," Machine Learning, 45(1), 5-32, 2001.

## Campos de nodo Bosque aleatorio

La pestaña Campos especifica los campos que se utilizan en el análisis.

**Utilizar roles predefinidos.** Esta opción utiliza la configuración de rol (objetivos, predictores, etc) de un nodo de Tipo anterior (o la pestaña Tipos de un nodo de origen anterior).

**Utilizar asignaciones de campos personalizadas.** Para asignar manualmente un objetivo y predictores, seleccione esta opción.

**Campos.** Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los campos de roles de Objetivo y Predictores a la derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol. Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

**Objetivo.** Seleccione un campo para utilizarlo como objetivo de la predicción.

**Predictores.** Seleccione uno o más campos como entradas de la predicción.

# Opciones de generación del nodo Bosque aleatorio

Utilice la pestaña Opciones de generación para especificar opciones de generación para el nodo Bosque aleatorio, incluyendo **opciones básicas** y **opciones avanzadas**. Si desea más información sobre estas opciones, consulte https://scikit-learn.org/stable/modules/ensemble.html#forest

## Básico

**Número de árboles que se crearán.** Seleccione el número de árboles del bosque.

**Especificar máxima profundidad.** Si no está seleccionado, los nodos se expanden hasta llegar a las hojas que no contienen nada o hasta que todas las hojas contienen menos de `min_samples_split` muestras.

**Profundidad máxima** La profundidad máxima del árbol.

**Tamaño mínimo de nodo de hoja.** El número mínimo de muestras que deben estar en un nodo de hoja.

**Número de características para utilizar para la división.** El número de características para tener en cuenta al buscar la mejor división:

- Si `auto`, `max_features=sqrt(n_features)` para el clasificador y `max_features=sqrt(n_features)` para la regresión.
- Si `sqrt`, `max_features=sqrt(n_features)`.
- Si `log2`, `max_features=log2(n_features)`.

## Avanzado

**Utilizar muestras de programa de arranque al generar árboles.** Si está seleccionado, las muestras de programa de arranque se utilizan al generar árboles.

**Utilizar muestras aleatorias para estimar la precisión de la generalización.** Si está seleccionado, las muestras aleatorias se utilizan para estimar la precisión de la generalización.

**Utilizar árboles extremadamente aleatorizados.** Si está seleccionado, se utilizan árboles extremadamente aleatorizados en lugar de bosques aleatorios generales. En árboles extremadamente aleatorizados, la aleatoriedad va un paso más allá en la forma cómo se calculan las divisiones. Como en los bosques aleatorios, se utiliza un subconjunto aleatorio de características candidatas, pero en lugar de buscar los umbrales más discriminatorios, los umbrales se trazan de forma aleatoria para cada característica candidata y el mejor de estos umbrales generados al azar se selecciona como regla de división. Normalmente, esto permite que la varianza del modelo se reduzca un poco más, a expensas de un ligero mayor incremento en el sesgo. [1]

**Replicar resultados.** Si está seleccionado, el proceso de generación de modelo se replica para conseguir los mismos resultados de puntuación.

**Semilla aleatoria.** Puede pulsar **Generar** para generar la semilla utilizada por el generador de números aleatorios.

**Optimización de hiper-parámetro (basada en Rbfopt).** Seleccione esta opción para habilitar la optimización de hiper-parámetro basada en Rbfopt, que descubre automáticamente la combinación óptima de parámetros, de forma que el modelo conseguirá el índice de error previsto o inferior en las muestras. Si desea detalles sobre Rbfopt, consulte http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html.

**Objetivo.** El valor de función de objetivo (índice de errores del modelo en las muestras) que desea alcanzar (es decir, el valor del óptimo desconocido). Establezcalo en un valor aceptable como, por ejemplo, `0.01`.

**Máx de iteraciones.** El número máximo de iteraciones para intentar el modelo. El valor predeterminado es `1000`.

**Máx de evaluaciones.** El número máximo de evaluaciones de función de forma precisa, para volver a intentar el modelo. El valor predeterminado es `300`.

La tabla siguiente muestra la relación entre los valores del diálogo del nodo Bosque aleatorio de SPSS Modeler y los parámetros de la biblioteca de Bosque aleatorio Python.

*Tabla 38. Propiedades de nodo correlacionadas con parámetros de biblioteca de Python*

| Valor de SPSS Modeler | Nombre de script (nombre de propiedad) | Parámetro de bosque aleatorio |
|---|---|---|
| Objetivo | `target` | |
| Predictores | `inputs` | |
| Número de árboles que se generarán | `n_estimators` | `n_estimators` |
| Especificar máxima profundidad | `specify_max_depth` | `specify_max_depth` |
| Profundidad máxima | `max_depth` | `max_depth` |
| Tamaño mínimo de nodo de hoja | `min_samples_leaf` | `min_samples_leaf` |
| Número de características para utilizar para la división | `max_features` | `max_features` |
| Utilizar muestras de programa de arranque al generar árboles | `bootstrap` | `bootstrap` |
| Utilizar muestras aleatorias para estimar la precisión de generalización | `oob_score` | `oob_score` |
| Utilizar árboles extremadamente aleatorizados | `extreme` | |
| Replicar resultados | `use_random_seed` | |
| Semilla aleatoria | `random_seed` | `random_seed` |
| Optimización de hiper parámetro (basada en Rbfopt) | `enable_hpo` | |
| Objetivo (para HPO) | `target_objval` | |
| Máx de iteraciones (para HPO) | `max_iterations` | |
| Máx de evaluaciones (para HPO) | `max_evaluations` | |

[1]L. Breiman, "Random Forests," Machine Learning, 45(1), 5-32, 2001.

## Opciones de modelo del nodo Bosque aleatorio

**Nombre del modelo.** Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

## Nuggets de modelo de bosque aleatorio

Los nuggets de modelo de bosque aleatorio contienen toda la información capturada por el modelo de bosque aleatorio. Están disponibles las secciones siguientes.

### Información del modelo

Esta vista proporciona información clave sobre el modelo, incluyendo campos de entrada, valores de codificación One-Hot (un solo bit significativo) y parámetros de modelo.

### Importancia del predictor

Esta vista muestra un gráfico que indica la importancia relativa de cada predictor cuando se estima el modelo. Si desea obtener más información, consulte “Importancia del predictor” en la página 44.

---

## Nodo SVM de una clase

El nodo SVM© de una clase utiliza un algoritmo de aprendizaje no supervisado. El nodo se puede utilizar para la detección de novedad. Detectará el límite flexible de un conjunto de muestras proporcionado, para clasificar a continuación los puntos nuevos como pertenecientes o no a dicho conjunto. Este nodo de modelado SVM de una clase se implementa en Python y necesita la biblioteca `scikit-learn`© de Python. Para obtener detalles sobre la biblioteca scikit-learn, consulte http://contrib.scikit-learn.org/imbalanced-learn/about.html[1].

La pestaña Python de la Paleta de nodos contiene el nodo SVM de una clase y otros nodos Python.

**Nota:** SVM de una clase se utiliza para la detección de novedades y valores atípicos sin supervisar. En la mayoría de los casos, recomendamos utilizar un conjunto de datos "normal" conocido para crear el modelo de forma que el algoritmo pueda establecer un límite correcto para la muestras proporcionadas. Los parámetros para el modelo – como nu, gamma y kernel – tienen un impacto significativo en el resultado. Por consiguiente, es posible que necesite experimentar con estas opciones hasta que encuentre la configuración óptima para la situación.

[1]Smola, Schölkopf. "A Tutorial on Support Vector Regression." *Statistics and Computing Archive*, vol. 14, no. 3, Agosto 2004, páginas 199-222. (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.114.4288)

## Campos de nodo SVM de una clase

La pestaña Campos especifica los campos que se utilizan en el análisis.

**Utilizar roles predefinidos.** Seleccione esta opción para seleccionar todos los campos con un rol definido de Entrada.

**Utilizar asignaciones de campos personalizadas.** Para seleccionar campos manualmente, seleccione esta opción y elija campos de entrada y campos de división:

**Entradas.** Seleccione los campos de entrada a utilizar en el análisis. Se soportan todos los tipos de almacenamiento y tipos de medición, excepto sin tipo y desconocido. Si un campo tiene un tipo de almacenamiento de Cadena, los valores de este campo se convertirán en binarios de una manera de uno frente a todos a través de un algoritmo de codificación One-Hot (un solo bit significativo).

**Dividir**. Seleccione qué campo o campos se deben utilizar como campos de división. Se soportan todos los tipos de medición de marca, nominal, ordinal y discreta.

**Utilizar los datos en particiones** Si se define un campo de partición, esta opción asegura que sólo se utilicen los datos de la partición de entrenamiento para generar el modelo.

## Experto de nodo SVM de una clase

En la pestaña Experto del nodo SVM de una clase, puede elegir entre modo **Simple** o modo **Experto**. Si elige **Simple**, todos los parámetros se establecen con los valores predeterminados como se muestra a continuación. Si selecciona **Experto**, puede especificar valores personalizados para estos parámetros. Para obtener detalles adicionales sobre estas opciones, consulte http://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html#sklearn.svm.OneClassSVM.

**Criterios de parada.** Especifique la tolerancia para los criterios de parada. El valor predeterminado es **1.0E-3** (0,001).

**Precisión de regresión (nu).** Límite en la fracción de errores de entrenamiento y vectores de soporte. El valor predeterminado es **0,1**.

**Tipo de kernel.** Tipo de kernel a utilizar en el algoritmo. Las opciones incluyen **RBF**, **Polinómico**, **Sigmoide**, **Lineal** o **Precalculado**. El valor predeterminado es **RBF**.

**Especificar Gamma.** Seleccione esta opción para especificar Gamma. De lo contrario, se aplicará gamma automática.

**Gamma.** El valor Gamma solo está disponible para los tipos de kernel RBF, polinómico y Sigmoide.

**Coef0.** Coef0 solo está disponible para los tipos de kernel Polinómico y Sigmoide.

**Grado.** Grado solo está disponible para el tipo de kernel Polinómico.

**Utilizar la heurística de reducción.** Seleccione esta opción para utilizar la heurística de reducción. Esta opción está deseleccionada de forma predeterminada.

**Establecer semilla aleatoria.** Seleccione esta opción para establecer la semilla de aleatorización a utilizar cuando se reorganizan los datos para la estimación de probabilidad. Esta opción está deseleccionada de forma predeterminada.

**Especificar el tamaño de caché de kernel (e MB).** Seleccione esta opción para especificar el tamaño de la memoria caché de kernel. Esta opción está deseleccionada de forma predeterminada. Cuando está seleccionada, el valor predeterminado es **200** MB.

**Optimización de hiper-parámetro (basada en Rbfopt).** Seleccione esta opción para habilitar la optimización de hiper-parámetro basada en Rbfopt, que descubre automáticamente la combinación óptima de parámetros, de forma que el modelo conseguirá el índice de error previsto o inferior en las muestras. Si desea detalles sobre Rbfopt, consulte http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html.

**Objetivo.** El valor de función de objetivo (índice de error del modelo en las muestras) que se desea alcanzar (por ejemplo, el valor del óptimo desconocido). Establezcalo en un valor aceptable como, por ejemplo, **0.01**.

**Máx de iteraciones.** Número máximo de iteraciones para intentar el modelo. El valor predeterminado es **1000**.

**Máx de evaluaciones.** Número máximo de evaluaciones de función para intentar el modelo, donde el foco es la precisión en la velocidad. El valor predeterminado es **300**.

El nodo SVM de una clase necesita la biblioteca de Python `scikit-learn`©. La tabla siguiente muestra la relación entre los valores del diálogo de nodo SMOTE de SPSS Modeler y el algoritmo de Python.

*Tabla 39. Propiedades de nodo correlacionadas con parámetros de biblioteca de Python*

| Nombre de parámetro | Nombre de script (nombre de propiedad) | Nombre de parámetro de API Python |
|---|---|---|
| Criterios de parada | `stopping_criteria` | `tol` |
| Precisión de regresión | `precision` | `nu` |
| Tipo de kernel | `kernel` | `kernel` |
| Gamma | `gamma` | `gamma` |
| Coef0 | `coef0` | `coef0` |
| Grado | `degree` | `degree` |

*Tabla 39. Propiedades de nodo correlacionadas con parámetros de biblioteca de Python (continuación)*

| Nombre de parámetro | Nombre de script (nombre de propiedad) | Nombre de parámetro de API Python |
|---|---|---|
| Utilizar la heurística de reducción | shrinking | shrinking |
| Especificar el tamaño de caché de kernel (cuadro de entrada de número) | cache_size | cache_size |
| Semilla aleatoria | random_seed | random_state |

## Opciones de nodo SVM de una clase

En la pestaña Opciones del nodo SVM de una clase, puede establecer las opciones siguientes.

**Tipo de gráfico de coordenadas paralelas** SPSS Modeler traza gráficos de coordenadas paralelas para presentar el modelo generado. A veces, los valores para algunas columnas de datos/características se visualizarán mucho más tiempo que otros, lo que puede hacer que algunas otras partes del gráfico sean difíciles de ver. Para casos como éste, puede elegir la opción **Ejes verticales independientes** para proporcionar a todos los ejes verticales una escala de eje autónoma o seleccionar **Ejes verticales generales** para forzar que todos los ejes verticales compartan la misma escala de ejes.

**Líneas máximas en el gráfico.** Especifique el número máximo de filas de datos (líneas) a visualizar en el resultado gráfico. El valor predeterminado es 100. Por razones de rendimiento, se visualizará un máximo de 20 campos.

**Trazar todos los campos de entrada en el gráfico.** Seleccione esta opción para mostrar todos los campos de entrada del resultado gráfico. De forma predeterminada, cada campo de datos se trazará como un eje vertical. Por razones de rendimiento, se visualizará un máximo de 30 campos.

**Campos personalizados a trazar en el gráfico.** En lugar de mostrar todos los campos de entrada en el resultado gráfico, puede seleccionar esta opción y elegir un subconjunto de campos a mostrar. Esto puede mejorar el rendimiento. Por razones de rendimiento, se visualizará un máximo de 20 campos.

# Capítulo 18. Nodos Spark

SPSS Modeler ofrece nodos para utilizar algoritmos nativos Spark. La pestaña **Spark** en la Paleta de nodos contiene los nodos siguientes que puede utilizar para ejecutar algoritmos Spark. Estos nodos están soportados en Windows 64 y Mac.

XGBoost© es una implementación avanzada de un algoritmo de aumento de gradiente. Los algoritmos de aumento conocen de forma iterativa los clasificadores débiles y, a continuación, los añaden a un clasificador fuerte final. XGBoost es muy flexible y proporciona muchos parámetros que pueden resultar abrumadores para la mayoría de los usuarios, así que el nodo XGBoost-AS en SPSS Modeler expone las características principales y los parámetros utilizados normalmente. El nodo XGBoost-AS se implementa en Spark.

La regresión isotónica pertenece a la familia de algoritmos de regresión. El nodo Isotónica-AS en SPSS Modeler se implementa en Spark. Si desea detalles sobre algoritmos de regresión isotónica, consulte https://spark.apache.org/docs/2.2.0/mllib-isotonic-regression.html.

k-medias es uno de los algoritmos de agrupación en clúster utilizado con más frecuencia. Agrupa en clúster puntos de datos en un número predefinido de clústeres. El nodo K-Medias-AS en SPSS Modeler se implementa en Spark. Si desea más detalles sobre algoritmos de k-medias, consulte https://spark.apache.org/docs/2.2.0/ml-clustering.html.Tenga en cuenta que el nodo K-Medias-AS realiza una codificación "one-hot" automáticamente para variables categóricas.

## Nodo Isotónica-AS

La regresión isotónica pertenece a la familia de algoritmos de regresión. El nodo Isotónica-AS en SPSS Modeler se implementa en Spark.

Si desea detalles sobre algoritmos de regresión isotónica, consulte https://spark.apache.org/docs/2.2.0/mllib-isotonic-regression.html. [1]

[1] "Regression - RDD-based API." *Apache Spark*. MLlib: Main Guide. Web. 3 Oct 2017.

## Campos del nodo Isotónica-AS

La pestaña Campos especifica los campos que se utilizan en el análisis.

**Campos.** Lista todos los campos en la fuente de origen. Utilice los botones de flecha para asignar elementos manualmente de esta lista a los campos Objetivo, Entrada y Ponderación a la derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol. Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

**Objetivo.** Seleccione un campo para utilizar como objetivo.

**Entrada.** Seleccione el campo(s) de entrada.

**Ponderación.** Seleccione un campo de ponderación para la ponderación exponencial. Si no está establecido, se utilizará el valor de ponderación predeterminado de **1**.

## Opciones de generación de nodo Isotónica-AS

Utilice la pestaña Opciones de generación para especificar opciones de generación para el nodo Isotónica-AS, que incluye el índice de características y el tipo isotónico. Si desea más información, consulte http://spark.apache.org/docs/latest/api/java/org/apache/spark/ml/regression/IsotonicRegression.html.[1]

**Índice de campos de entrada.** Especifique el índice de los campos de entrada. El valor predeterminado es **0**.

**Tipo isotónico.** Este valor determina si la secuencia de salida debe ser isotónica/creciente o antitónica/decreciente. El valor predeterminado es **Isotonic**.

[1] "Class IsotonicRegression." *Apache Spark*. JavaDoc. Web. 3 Oct 2017.

## Nuggets del modelo Isotónica-AS

Los nuggets del modelo Isotónica-AS contienen toda la información capturada por el modelo de regresión isotónica. Están disponibles las secciones siguientes.

### Resumen del modelo

Esta vista proporciona información clave sobre el modelo, incluyendo campos de entrada, campo objetivo y opciones de generación de modelo.

### Gráfico de modelo

Esta vista muestra un diagrama de dispersión.

---

## Nodo XGBoost-AS

XGBoost© es una implementación avanzada de un algoritmo de aumento de gradiente. Los algoritmos de aumento conocen de forma iterativa los clasificadores débiles y, a continuación, los añaden a un clasificador fuerte final. XGBoost es muy flexible y proporciona muchos parámetros que pueden resultar abrumadores para la mayoría de los usuarios, así que el nodo XGBoost-AS en SPSS Modeler expone las características principales y los parámetros utilizados normalmente. El nodo XGBoost-AS se implementa en Spark.

Para obtener más información sobre los algoritmo de aumento, consulte los tutoriales de XGBoost disponibles en http://xgboost.readthedocs.io/en/latest/tutorials/index.html. [1]

Tenga en cuenta que la función de validación cruzada de XGBoost no se soporta en SPSS Modeler. Puede utilizar el nodo Partición de SPSS Modeler para esta funcionalidad. También tenga en cuenta que XGBoost en SPSS Modeler ejecuta la codificación One-Hot automáticamente para variables categóricas.

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

## Campos de nodo XGBoost-AS

La pestaña Campos especifica los campos que se utilizan en el análisis.

**Utilizar roles predefinidos.** Esta opción utiliza la configuración de rol (objetivos, predictores, etc) de un nodo de Tipo anterior (o la pestaña Tipos de un nodo de origen anterior).

**Utilizar asignaciones de campos personalizadas.** Para asignar manualmente un objetivo y predictores, seleccione esta opción.

**Campos.** Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los campos de roles de Objetivo y Predictores a la derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol. Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

**Objetivo.** Seleccione un campo para utilizarlo como objetivo de la predicción.

**Predictores.** Seleccione uno o más campos como entradas de la predicción.

# Opciones de generación de nodo XGBoost-AS

Utilice la pestaña Opciones de generación para especificar opciones de generación para el nodo XGBoost-AS, que incluyen **opciones generales** para la generación de modelo y el manejo de conjuntos de datos desequilibrados, **opciones de tarea de aprendizaje** para objetivos y métricas de evaluación y **parámetros de amplificador** para amplificadores específicos. Si desea más información sobre estas opciones, consulte los recursos en línea siguientes:

- Página de inicio de XGBoost[1]
- Referencia de parámetro XGBoost[2]
- API Spark XGBoost[3]

## General

**Número de trabajadores.** Número de trabajadores utilizados para entrenar el modelo XGBoost.

**Número de hebras.** Número de hebras utilizadas por trabajador.

**Utilizar memoria externa.** Indica si se va a utilizar memoria externa como memoria caché.

**Tipo de amplificador.** El amplificador para utilizar (**gbtree**, **gblinear** o **dart**).

**Número de rondas de amplificador.** El número de rondas para aumentar.

**Ponderación de posición de escala.** Este valor controla el equilibrio de ponderaciones positivas y negativas y es útil para clases desequilibradas.

**Semilla aleatoria.** Pulse **Generar** para generar la semilla utilizada por el generador de números aleatorios.

## Tarea de aprendizaje

**Objetivo.** Seleccione en los siguientes tipos de objetivo de tarea de aprendizaje: **reg:linear**, **reg:logistic**, **reg:gamma**, **reg:tweedie**, **count:poisson**, **rank:pairwise**, **binary:logistic** o **multi**.

**Métricas de evaluación.** Métricas de evaluación para datos de validación. Se asignará una métrica predeterminada de acuerdo con el objetivo (**rmse** para regresión **error** para clasificación o **mean average precision** para clasificar). Las opciones disponibles son **rmse**, **mae**, **logloss**, **error**, **merror**, **mlogloss**, **uac**, **ndcg**, **map** o **gamma-deviance** (el valor predeterminado es **rmse**).

## Parámetros de amplificador

**Lambda.** Término de regularización L2 en ponderaciones. Al aumentar este valor, el modelo será más conservador.

**Alfa.** Término de regularización L1 en ponderaciones. Al aumentar este valor, el modelo será más conservador.

**Lambda bias.** Término de regularización L2 en sesgo. (No hay ningún término de regularización L1 en sesgo porque no es importante.)

**Método de árbol.** Seleccione el algoritmo de construcción de árbol de XGBoost a utilizar.

**Profundidad máxima** Especifique la profundidad máxima para árboles. Al aumentar este valor, el modelo se hará más complejo y será más probable que se sobreajuste.

**Ponderación hijo mínima.** Especifique la suman mínima de ponderación de instancia (hessiana) necesaria en un hijo. Si el paso de partición de árbol produce un nodo hoja con la suma de la ponderación de instancia menor que esta **Ponderación hijo mínima**, el proceso de generación detendrá el particionamiento adicional. En modo de regresión lineal, este simplemente corresponde al número mínimo de instancias necesarias en cada nodo. Cuando mayor es la ponderación, más conservador será el algoritmo.

**Paso delta máximo.** Especifique el paso delta máximo que se debe permitir para la estimación ponderada de cada árbol. Si se establece en **0**, no hay ninguna restricción. Si se establece en un valor positivo, puede ayudar a que el paso de actualización sea más conservador. Generalmente este parámetro no es necesario, pero puede ayudar en la regresión logística cuando una clase es extremadamente desequilibrada.

**Submuestra.** Submuestra en la razón de la instancia de entrenamiento. Por ejemplo, si establece esto en **0,5**, XGBoost recolectará aleatoriamente la mitad de las instancias de datos para hacer crecer árboles y esto evitará el sobreajuste.

**Eta.** La reducción de tamaño de paso utilizada durante el paso de actualización para evitar el sobreajuste. Después de cada paso de aumento, se pueden obtener directamente las ponderaciones de nuevas características. Eta también reduce las ponderaciones de característica para que el proceso de aumento sea más conservador.

**Gamma.** La reducción de pérdida mínima necesaria para realizar una partición adicional en un nodo hoja del árbol. Cuanto mayor es el valor de gamma, más conservador será el algoritmo.

**Muestra de columna por árbol.** Razón de submuestra de columnas cuando se construye cada árbol.

**Muestra de columna por nivel.** Razón de submuestra de columnas para cada división, en cada nivel.

**Algoritmo de normalización** El algoritmo de normalización para utilizar cuando está seleccionado el tipo de amplificador Dart bajo Opciones generales. Las opciones disponibles son **tree** o **forest** (el valor predeterminado es **tree**).

**Algoritmo de muestreo.** El algoritmo de muestreo para utilizar cuando está seleccionado el tipo de amplificador Dart bajo Opciones generales. El algoritmo **uniform** selecciona de forma uniforme árboles descartados. El algoritmo **weighted** selecciona árboles descartados en proporción con la ponderación. El valor predeterminado es **uniform**.

**Índice de descarte.** El índice de descarte para utilizar cuando está seleccionado el tipo de amplificador Dart bajo Opciones generales.

**Probabilidad de omitir descarte.** La probabilidad de omitir descarte para utilizar cuando está seleccionado el tipo de amplificador Dart bajo Opciones generales. Si se omite un descarte, se añaden árboles nuevos de la misma forma que **gbtree**.

La tabla siguiente muestra la relación entre los valores del diálogo del nodo XGBoost-AS de SPSS Modeler y los parámetros Spark deXGBoost.

*Tabla 40. Propiedades de nodo correlacionadas con parámetros Spark*

| Valor de SPSS Modeler | Nombre de script (nombre de propiedad) | Parámetro Spark de XGBoost |
|---|---|---|
| Objetivo | target_fields | |
| Predictores | input_fields | |
| Lambda | lambda | lambda |
| Número de trabajadores | nWorkers | nWorkers |
| Número de hebras | numThreadPerTask | numThreadPerTask |
| Utilizar memoria externa | useExternalMemory | useExternalMemory |
| Tipo de amplificador | boosterType | boosterType |
| Número de ronda de aumento | numBoostRound | round |
| Ponderación de posición de escala | scalePosWeight | scalePosWeight |
| Objetivo | objectiveType | objective |
| Métricas de evaluación | evalMetric | evalMetric |
| Lambda | lambda | lambda |
| Alfa | alpha | alpha |
| Lambda bias | lambdaBias | lambdaBias |
| Método de árbol | treeMethod | treeMethod |
| Profundidad máx | maxDepth | maxDepth |
| Ponderación hijo mínima | minChildWeight | minChildWeight |
| Paso delta máximo | maxDeltaStep | maxDeltaStep |
| Submuestra | sampleSize | sampleSize |
| Eta | eta | eta |
| Gamma | gamma | gamma |
| Muestra de columna por árbol | colsSampleRation | colSampleByTree |
| Muestra de columna por nivel | colsSampleLevel | colsSampleLevel |
| Algoritmo de normalización | normalizeType | normalizeType |
| Algoritmo de muestreo | sampleType | sampleType |
| Índice de descarte | rateDrop | rateDrop |
| Probabilidad de omitir descarte | skipDrop | skipDrop |

[1] "Scalable and Flexible Gradient Boosting." Web. © 2015-2016 DMLC.

[2] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[3] "ml.dmlc.xgboost4j.scala.spark Params." *DMLC for Scalable and Reliable Machine Learning*. Web. 3 Oct 2017.

## Opciones de modelo de nodo XGBoost-AS

**Nombre del modelo.** Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

# Nodo de K-Medias-AS

k-medias es uno de los algoritmos de agrupación en clúster utilizado con más frecuencia. Agrupa en clúster puntos de datos en un número predefinido de clústeres.[1] El nodo K-Medias-AS en SPSS Modeler se implementa en Spark.

Si desea más detalles sobre algoritmos de k-medias, consulte https://spark.apache.org/docs/2.2.0/ml-clustering.html.

Tenga en cuenta que el nodo K-Medias-AS realiza una codificación "one-hot" automáticamente para variables categóricas.

[1] "Clustering." *Apache Spark*. MLlib: Main Guide. Web. 3 Oct 2017.

# Campos de nodo K-Medias-AS

La pestaña Campos especifica los campos que se utilizan en el análisis.

**Utilizar roles predefinidos.** Esta opción permite indicar al nodo que use la información de campo de un nodo Tipo situado en un punto anterior de la ruta. Está seleccionado de forma predeterminada.

**Utilizar asignaciones de campos personalizadas.** Si desea asignar manualmente campos de entrada, seleccione esta opción y, después, seleccione el campo o los campos de entrada. El uso de esta opción es similar a establecer el rol del campo en **Input** en un nodo Tipo.

# Opciones de generación del nodo K-Medias-AS

Utilice la pestaña Opciones de generación para especificar opciones de generación para el nodo K-Medias-AS, que incluye opciones regulares para la creación de modelos, opciones de inicialización para inicializar centros de clúster y opciones avanzadas para la semilla aleatoria y la iteración de cálculo. Si desea más información, consulte el JavaDoc para K-Medias en SparkML.[1]

## Regular

**Nombre de modelo.** El nombre del campo generado después de puntuar un clúster específico. Seleccione **Auto** (valor predeterminado) o seleccione **Personalizado** y escriba un nombre.

**Número de clústeres.** Especifique el número de clústeres que generar. El valor predeterminado es **5** y el mínimo es **2**.

## Inicialización

**Modalidad de inicialización.** Especifique el método para inicializar los centros de clúster. **K-Means\|\|** es el valor predeterminado. Si desea más detalles sobre estos dos métodos, consulte Scalable K-Means++. [2]

**Pasos de inicialización.** Si está seleccionada la modalidad de inicialización **K-Means\|\|**, especifique el número de pasos de inicialización. **2** es el valor predeterminado.

## Avanzado

**Configuración avanzada.** Seleccione esta opción si desea establecer opciones avanzada del modo siguiente.

**Iteración máx.** Especifique el número máximo de iteraciones para realizar al buscar en centros de clúster. **20** es el valor predeterminado.

**Tolerancia.** Especifique la tolerancia de convergencia para algoritmo iterativos. **1.0E-4** es el valor predeterminado.

**Establecer semilla aleatoria.** Seleccione esta opción y pulse **Generar** para generar la semilla utilizada por el generador de números aleatorios.

## Visualización

**Mostrar gráfico.** Seleccione esta opción si desea que se incluya un gráfico en la salida.

La tabla siguiente muestra la relación entre los valores de los parámetros Spark del nodo K-Medias-AS y K-Medias de SPSS Modeler.

*Tabla 41. Propiedades de nodo correlacionadas con parámetros Spark*

| Valor de SPSS Modeler | Nombre de script (nombre de propiedad) | Parámetro SparkML de k-medias |
|---|---|---|
| Campos de entrada | `features` | |
| Número de clústeres | `clustersNum` | `k` |
| Modalidad de inicialización | `initMode` | `initMode` |
| Pasos de inicialización | `initSteps` | `initSteps` |
| Iteración máx | `maxIter` | `maxIter` |
| Tolerancia | `toleration` | `tol` |
| Semilla aleatoria | `randomSeed` | `seed` |

[1] "Class KMeans." *Apache Spark*. JavaDoc. Web. 3 Oct 2017.

[2] Bahmani, Moseley, et al. "Scalable K-Means++." Feb 28, 2012. http://theory.stanford.edu/%7Esergei/papers/vldb12-kmpar.pdf.

# Avisos

Esta información se ha desarrollado para productos y servicios ofrecidos en los EE.UU. Este material puede estar disponible en IBM en otros idiomas. Sin embargo, es posible que deba ser propietario de una copia del producto o de la versión del producto en dicho idioma para acceder a él.

Es posible que IBM no ofrezca los productos, servicios o características que se tratan en este documento en otros países. El representante local de IBM le puede informar sobre los productos y servicios que están actualmente disponibles en su localidad. Cualquier referencia a un producto, programa o servicio de IBM no pretende afirmar ni implicar que solamente se pueda utilizar ese producto, programa o servicio de IBM. En su lugar, se puede utilizar cualquier producto, programa o servicio funcionalmente equivalente que no infrinja los derechos de propiedad intelectual de IBM. Sin embargo, es responsabilidad del usuario evaluar y comprobar el funcionamiento de todo producto, programa o servicio que no sea de IBM.

IBM puede tener patentes o solicitudes de patente en tramitación que cubran la materia descrita en este documento. Este documento no le otorga ninguna licencia para estas patentes. Puede enviar preguntas acerca de las licencias, por escrito, a:

*IBM Director of Licensing*
*IBM Corporation*
*North Castle Drive, MD-NC119*
*Armonk, NY 10504-1785*
*EE.UU.*

Para consultas sobre licencias relacionadas con información de doble byte (DBCS), póngase en contacto con el departamento de propiedad intelectual de IBM de su país o envíe sus consultas, por escrito, a:

*Intellectual Property Licensing*
*Legal and Intellectual Property Law*
*IBM Japan Ltd.*
*19-21, Nihonbashi-Hakozakicho, Chuo-ku*
*Tokio 103-8510, Japón*

INTERNATIONAL BUSINESS MACHINES CORPORATION PROPORCIONA ESTA PUBLICACIÓN "TAL CUAL", SIN GARANTÍAS DE NINGUNA CLASE, NI EXPLÍCITAS NI IMPLÍCITAS, INCLUYENDO, PERO SIN LIMITARSE A, LAS GARANTÍAS IMPLÍCITAS DE NO VULNERACIÓN, COMERCIALIZACIÓN O ADECUACIÓN A UN PROPÓSITO DETERMINADO. Algunas jurisdicciones no permiten la renuncia a las garantías explícitas o implícitas en determinadas transacciones; por lo tanto, es posible que esta declaración no sea aplicable a su caso.

Esta información puede incluir imprecisiones técnicas o errores tipográficos. Periódicamente, se efectúan cambios en la información aquí y estos cambios se incorporarán en nuevas ediciones de la publicación. IBM puede realizar en cualquier momento mejoras o cambios en los productos o programas descritos en esta publicación sin previo aviso.

Las referencias hechas en esta publicación a sitios web que no son de IBM se proporcionan sólo para la comodidad del usuario y no constituyen de modo alguno un aval de esos sitios web. La información de esos sitios web no forma parte de la información de este producto de IBM y la utilización de esos sitios web se realiza bajo la responsabilidad del usuario.

IBM puede utilizar o distribuir la información que se le proporcione del modo que considere adecuado sin incurrir por ello en ninguna obligación con el remitente.

Los titulares de licencias de este programa que deseen tener información sobre el mismo con el fin de permitir: (i) el intercambio de información entre programas creados independientemente y otros programas (incluido este) y (ii) el uso mutuo de la información que se ha intercambiado, deberán ponerse en contacto con:

*IBM Director of Licensing*
*IBM Corporation*
*North Castle Drive, MD-NC119*
*Armonk, NY 10504-1785*
*EE.UU.*

Esta información estará disponible, bajo las condiciones adecuadas, incluyendo en algunos casos el pago de una cuota.

El programa bajo licencia que se describe en este documento y todo el material bajo licencia disponible los proporciona IBM bajo los términos de las Condiciones Generales de IBM, Acuerdo Internacional de Programas Bajo Licencia de IBM o cualquier acuerdo equivalente entre las partes.

Los ejemplos de datos de rendimiento y de clientes citados se presentan solamente a efectos ilustrativos. Los resultados reales de rendimiento pueden variar en función de las configuraciones específicas y condiciones de operación.

La información relacionada con productos no IBM se ha obtenido de los proveedores de esos productos, de sus anuncios publicados o de otras fuentes disponibles públicamente. IBM no ha probado esos productos y no puede confirmar la exactitud del rendimiento, la compatibilidad ni ninguna otra afirmación relacionada con productos no IBM. Las preguntas sobre las posibilidades de productos que no son de IBM deben dirigirse a los proveedores de esos productos.

Las declaraciones sobre el futuro rumbo o intención de IBM están sujetas a cambio o retirada sin previo aviso y representan únicamente metas y objetivos.

Esta información contiene ejemplos de datos e informes utilizados en operaciones comerciales diarias. Para ilustrarlos lo máximo posible, los ejemplos incluyen los nombres de las personas, empresas, marcas y productos. Todos estos nombres son ficticios y cualquier parecido con personas o empresas comerciales reales es pura coincidencia.

## Marcas comerciales

IBM, el logotipo de IBM e ibm.com son marcas registradas o marcas comerciales de International Business Machines Corp., registradas en muchas jurisdicciones en todo el mundo. Otros nombres de productos y servicios podrían ser marcas registradas de IBM u otras compañías. En Internet hay disponible una lista actualizada de las marcas registradas de IBM, en "Copyright and trademark information", en www.ibm.com/legal/copytrade.shtml.

Adobe, el logotipo Adobe, PostScript y el logotipo PostScript son marcas registradas o marcas comerciales de Adobe Systems Incorporated en Estados Unidos y/o otros países.

Intel, el logotipo de Intel, Intel Inside, el logotipo de Intel Inside, Intel Centrino, el logotipo de Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium y Pentium son marcas comerciales o marcas registradas de Intel Corporation o sus filiales en Estados Unidos y otros países.

Linux es una marca registrada de Linus Torvalds en Estados Unidos, otros países o ambos.

Microsoft, Windows, Windows NT, y el logotipo de Windows son marcas comerciales de Microsoft Corporation en Estados Unidos, otros países o ambos.

UNIX es una marca registrada de The Open Group en Estados Unidos y otros países.

Java y todas las marcas comerciales y los logotipos basados en Java son marcas comerciales o registradas de Oracle y/o sus afiliados.

## Términos y condiciones para la documentación del producto

Los permisos para utilizar estas publicaciones se otorgan de acuerdo con los términos y condiciones siguientes.

### Aplicabilidad

Estos términos y condiciones son adicionales a los términos de uso del sitio web de IBM.

### Uso personal

Estas publicaciones se pueden reproducir para uso personal no comercial siempre que se conserven todos los avisos de propiedad. No puede distribuir, visualizar ni realizar trabajos derivados de estas publicaciones, ni de partes de las mismas, sin el consentimiento expreso de IBM.

### Uso comercial

Puede reproducir, distribuir y visualizar estas publicaciones únicamente dentro de la empresa a condición de que se conserven todos los avisos de propiedad. No puede realizar trabajos derivados de estas publicaciones, ni de partes de las mismas, ni reproducirlas, distribuirlas o visualizarlas fuera de su empresa sin el consentimiento expreso de IBM.

### Derechos

Excepto de la forma explícitamente otorgada en este permiso, no se otorga ningún permiso, licencia ni derecho, ni explícito ni implícito, sobre las publicaciones ni a ninguna otra información, datos, software u otra propiedad intelectual contenida en ellas.

IBM se reserva el derecho de retirar los permisos aquí otorgados siempre que, a su discreción, el uso de las publicaciones sea perjudicial para su interés o cuando, según determine IBM, las instrucciones anteriores no se sigan correctamente.

No puede descargar, exportar ni volver a exportar esta información si no es cumpliendo totalmente todas las leyes y regulaciones aplicables, incluyendo las leyes y regulaciones de exportación de los Estados Unidos.

IBM NO GARANTIZA EL CONTENIDO DE ESTAS PUBLICACIONES. LAS PUBLICACIONES SE PROPORCIONAN "TAL CUAL" Y SIN GARANTÍA DE NINGUNA CLASE, NI EXPLÍCITA NI IMPLÍCITA, INCLUYENDO PERO SIN LIMITARSE A LAS GARANTÍAS IMPLÍCITAS DE COMERCIALIZACIÓN, NO VULNERACIÓN E IDONEIDAD PARA UN FIN DETERMINADO.

# Glosario

## A

*AICC* . Una medida para seleccionar y comparar modelos mixtos basada en la log-verosimilitud -2 (restringida). Los valores menores indican modelos mejores. El AICC "corrige" el AIC respecto a tamaños muestrales pequeños. A medida que aumenta el tamaño de la muestra, el AICC converge con el AIC.

## B

*Criterio de información bayesiano (BIC)* . Una medida para seleccionar y comparar modelos basados en el logaritmo de la verosimilitud -2. Los valores menores indican modelos mejores. El BIC también "penaliza" modelos sobreparametrizados (modelos complejos con un gran número de entradas, por ejemplo), pero de forma más estricta que el AIC.

*Prueba M de Box* . Contraste sobre la igualdad de las matrices de covarianza de los grupos. Para tamaños de muestras suficientemente grandes, un valor de p no significativo quiere decir que no hay suficiente evidencia de que las varianzas sean diferentes. Esta prueba es sensible a las desviaciones de la normalidad multivariada.

## C

*Casos* . Se muestran, para cada caso, los códigos del grupo real de pertenencia, el grupo pronosticado, las probabilidades posteriores y las puntuaciones discriminantes.

*Resultados de clasificación* . Número de casos correcta e incorrectamente asignados a cada uno de los grupos, basándose en el análisis discriminante. En ocasiones se denomina "Matriz de Confusión".

*Gráfico de los grupos combinados* . Crea un diagrama de dispersión, con todos los grupos, de los valores en las dos primeras funciones discriminantes. Si sólo hay una función, en su lugar se muestra un histograma.

*Covarianza* . Medida no tipificada del grado de asociación entre dos variables igual a la desviación del producto vectorial dividido por N-1.

## F

*De Fisher* . Muestra los coeficientes de la función de clasificación de Fisher que pueden utilizarse directamente para la clasificación. Se obtiene un conjunto de coeficientes para cada grupo, y se asigna un caso al grupo para el que tiene una mayor puntuación discriminante (valor de función de clasificación).

## H

*Gráfico de riesgo* . Muestra la función de riesgo acumulado en una escala lineal.

## K

*Curtosis* . Una medida con respecto al número de valores atípicos que hay. Para una distribución normal, el valor del estadístico de curtosis es 0. Una curtosis positiva indica que los datos muestran más valores atípicos extremos que una distribución normal. Una curtosis negativa indica que los datos muestran menos valores atípicos extremos que una distribución normal.

## L

*Clasificación dejando uno fuera* . Se clasifica cada caso del análisis mediante la función derivada de todos los casos, excepto el propio caso. También se conoce como método U.

# M

*MAE* . Error absoluto promedio (Mean Absolute Error). Mide la desviación de la serie del nivel pronosticado por el modelo. El MAE se informa en las unidades originales de la serie.

*Distancia de Mahalanobis* . Medida de cuánto difieren del promedio para todos los casos los valores en las variables independientes de un caso dado. Una distancia de Mahalanobis grande identifica un caso que tenga valores extremos en una o más de las variables independientes.

*MAPE* . Error absoluto porcentual promedio (Mean Average Percentage Error). Medida de la desviación de la serie dependiente del nivel pronosticado por el modelo. Es independiente de las unidades utilizadas y se puede utilizar para comparar series con distintas unidades.

*MaxAE* . Error absoluto máximo. El mayor error previsto, expresado en las mismas unidades que la variable dependiente. Al igual que el MaxAPE, es útil para imaginar el peor escenario de los casos en las previsiones. El error absoluto máximo y el error absoluto porcentual máximo pueden darse en distintos puntos de la serie. Por ejemplo, si el error absoluto de un valor de la serie grande es ligeramente mayor que el error absoluto de un valor de la serie pequeño. En ese caso el error absoluto máximo se obtendrá en el valor de la serie mayor y el error absoluto porcentual máximo corresponderá al valor de la serie menor.

*MaxAPE* . Error absoluto porcentual máximo (Maximum Absolute Percentage Error). El mayor error previsto, expresado como porcentaje. Esta medida es útil para imaginar el peor escenario de un caso en las previsiones.

*Método de entrada por maximización de la mínima razón F* . Método para la selección de variables en los análisis por pasos que se basa en maximizar la razón F, calculada a partir de la distancia de Mahalanobis entre los grupos.

*Máximo* . Se trata del valor mayor de una variable numérica.

*Media* . Una medida de tendencia central. El promedio aritmético, la suma dividida por el número de casos.

*Medias* . Muestra la media y desviación estándar totales y las medias y desviaciones estándar de grupo, para las variables independientes.

*Mediana* . Es el valor por encima y por debajo del cual se encuentran la mitad de los casos, el percentil 50. Si hay un número par de casos, la mediana es la media de los dos valores centrales, cuando los casos se ordenan en orden ascendente o descendente. La mediana es una medida de tendencia central que no es sensible a los valores atípicos (a diferencia de la media, que puede resultar afectada por unos pocos valores extremadamente altos o bajos).

*Minimizar la lambda de Wilks* . Método para la selección de variables por pasos del análisis discriminante que selecciona las variables para su introducción en la ecuación basándose en cuánto contribuyen a disminuir la lambda de Wilks. En cada paso se introduce la variable que minimiza la lambda de Wilks global.

*Mínimo* . Se trata del valor menor de una variable numérica.

*Moda* . El valor que ocurre con mayor frecuencia. Si varios valores comparten la mayor frecuencia de aparición, cada uno de ellos es un modo.

# N

*BIC normalizado* . Criterio de información Bayesiano normalizado (Normalized Bayesian Information Criterion). Una medida general del ajuste global del modelo que intenta tener en cuenta la complejidad del modelo. Es una medida basada en el error cuadrático promedio que incluye una penalización para el número de parámetros presentes en el modelo y la longitud de la serie. La penalización elimina la ventaja de los modelos con mayor número de parámetros, haciendo que el estadístico sea fácil de comparar entre distintos modelos para la misma serie.

# O

*Uno menos la supervivencia* . Representa la función uno menos la supervivencia en una escala lineal.

# R

*Rango* . Diferencia entre los valores mayor y menor de una variable numérica; el máximo menos el mínimo.

*V de Rao (análisis discriminante)* . Medida de las diferencias entre las medias de los grupos. También se denomina la traza de Lawley-Hotelling. En cada paso, se incluye la variable que maximiza el incremento de la V de Rao. Después de seleccionar esta opción, introduzca el valor mínimo que debe tener una variable para poder incluirse en el análisis.

*RMSE* . Raíz del error cuadrático promedio (Root Mean Square Error). La raíz cuadrada del error cuadrático promedio. Una medida de cuánto se desvía la serie dependiente del nivel pronosticado por el modelo, expresado en las mismas unidades que la serie dependiente.

*R cuadrado* . Medida de la bondad de ajuste de un modelo lineal; en ocasiones recibe el nombre de coeficiente de determinación. Es la proporción de la variación de la variable dependiente explicada por el modelo de regresión. Puede tomar un valor entre 0 y 1. Un valor pequeño indica que el modelo no se ajusta bien a los datos.

# S

*Grupos separados* . Para la clasificación se utilizan las matrices de covarianza de los grupos separados. Dado que la clasificación se basa en las funciones discriminantes y no en las variables originales, esta opción no siempre es equivalente a la discriminación cuadrática.

*Covarianza de los grupos separados* . Muestra las matrices de covarianza de cada grupo por separado.

*Gráficos de los grupos separados* . Crea diagramas de dispersión, de los grupos por separado, para los valores en las dos primeras funciones discriminantes. Si sólo hay una función, en su lugar se muestra un histograma.

*Bonferroni secuencial* . Éste es un procedimiento de Bonferroni de rechazo secuencial decreciente que es mucho menos conservador en cuanto al rechazo de hipótesis individuales pero que mantiene el mismo nivel de significación global.

*Sidak secuencial* . Este es un procedimiento de Sidak de rechazo secuencial decreciente que es mucho menos conservador en términos de rechazar las hipótesis individuales pero que mantiene el mismo nivel de significación global.

*Asimetría* . Medida de la asimetría de una distribución. La distribución normal es simétrica y tiene un valor de sesgo de 0. Una distribución con un sesgo positivo significativo tiene una cola derecha larga. Una distribución que tenga una asimetría negativa significativa tiene una cola izquierda larga. Como regla aproximada, un valor de la asimetría mayor que el doble de su error estándar se asume que indica una desviación de la simetría.

*desviación estándar* . Una medida de la dispersión en torno a la media, igual a la raíz cuadrada de la varianza. La desviación estándar se mide en las mismas unidades que la variable original.

*Desviación estándar* . Una medida de dispersión sobre la media. En una distribución normal, el 68% de los casos se encuentra dentro de una desviación estándar de la media y el 95% queda entre dos desviaciones estándar. Por ejemplo, si la edad media es de 45 años, con una desviación estándar de 10, el 95% de los casos estaría entre los 25 y 65 en una distribución normal.

*Error estándar* . Medida de cuánto puede variar el valor de un estadístico de contraste de muestra en muestra. Es la desviación estándar de la distribución muestral de un estadístico. Por ejemplo, el error estándar de la media es la desviación estándar de las medias muestrales.

*Error estándar de curtosis* . La razón de la curtosis sobre su error estándar puede utilizarse como prueba de normalidad (es decir, se puede rechazar la normalidad si la razón es menor que -2 o mayor que +2). Un valor grande y positivo para la curtosis indica que las colas son más largas que las de una distribución normal; por el contrario, un valor extremo y negativo indica que las colas son más cortas (llegando a tener forma de caja como en la distribución uniforme).

*Error estándar de la media* . Medida de cuánto puede variar el valor de la media de una muestra a otra, extraídas éstas de la misma distribución. Puede utilizarse para comparar de forma aproximada la media observada respecto a un valor hipotetizado (es decir, se puede concluir que los dos valores son distintos si la diferencia entre ellos, dividida por el error estándar, es menor que -2 o mayor que +2).

*Error estándar de asimetría* . La razón de la asimetría sobre su error estándar puede utilizarse como una prueba de normalidad (es decir, se puede rechazar la normalidad si la razón es menor que -2 o mayor que +2). Un valor grande y positivo para la asimetría indica una cola larga a la derecha; un valor extremo y negativo indica una cola larga por la izquierda

*R cuadrado estacionaria* . Una medida que compara la parte estacionaria del modelo con un modelo de promedio simple. Esta medida es preferible al R-cuadrado ordinario cuando existe tendencia o patrón estacional. El valor de R cuadrado estacionario puede ser negativo con un rango de infinidad negativa hasta 1. Los valores negativos significan que el modelo estudiado es peor que el modelo basal. Los valores positivos significan que el modelo estudiado es mejor que el modelo basal.

*Suma* . Suma o total de todos los valores, a lo largo de todos los casos con valores no perdidos.

*Gráfico de supervivencia* . Muestra la función de supervivencia acumulada, en una escala lineal.

## T

*Mapa territorial* . Gráfico de las fronteras utilizadas para clasificar los casos en grupos a partir de los valores en las funciones. Los números corresponden a los grupos en los que se clasifican los casos. La media de cada grupo se indica mediante un asterisco situado dentro de sus fronteras. No se mostrará el mapa si sólo hay una función discriminante.

*Covarianza total* . Muestra la matriz de covarianza para todos los casos, como si fueran una única muestra.

## U

*Varianza no explicada* . En cada paso se introduce la variable que minimiza la suma de la variación no explicada entre los grupos.

*Exclusivo* . Evalúa todos los efectos simultáneamente, corrigiendo cada efecto por todos los demás efectos de cualquier tipo.

*ANOVAs univariados* . Realiza un análisis de varianza de un factor sobre la igualdad de las medias de grupo para cada variable independiente.

*No tipificados* . Muestra los coeficientes de la función discriminante sin estandarizar.

*Usar valor de F* . Una variable se introduce en el modelo si su valor de F es mayor que el valor de entrada, y se elimina si su valor de F es menor que el valor de Eliminación. La entrada debe ser mayor que la eliminación y ambos valores deben ser positivos. Para introducir más variables en el modelo, disminuya el valor de entrada. Para eliminar más variables del modelo, eleve el valor de eliminación.

*Usar probabilidad de F* . Una variable se introduce en el modelo si el nivel de significación de su valor de F es menor que el valor de entrada, y se elimina si el nivel de significación de su valor de F es mayor que el valor de Eliminación. La entrada debe ser menor que la eliminación y ambos valores deben ser positivos. Para introducir más variables en el modelo, aumente el valor de entrada. Para eliminar más variables del modelo, disminuya el valor de eliminación.

## V

*Válidos* . Los casos válidos que no tienen un valor perdido del sistema ni un valor definido como perdido del usuario.

*Varianza* . Medida de dispersión sobre la media, igual a la suma de las desviaciones al cuadrado de la media divida por el número de casos menos uno. La varianza se mide en unidades que son el cuadrado de las de la variable en cuestión.

## W

*Intra grupos* . Se utiliza la matriz de covarianza intra-grupos combinada para clasificar los casos.

*Correlación intra-grupos* . Muestra la matriz de correlaciones intra-grupos combinada, que se obtiene de promediar las matrices de covarianza individuales para todos los grupos antes de calcular las correlaciones.

*Covarianza intra-grupos* . Muestra la matriz de covarianza intra-grupos combinada, la cual puede diferir de la matriz de covarianza total. La matriz se obtiene de promediar, para todos los grupos, las matrices de covarianza individuales.

# Índice

# E



# F



# G



# H



# I

# K



# L



# M

# N



# O



# P

# R



# S



# T

# U



# V



# X

Impreso en España