Nodos de modelado de IBM SPSS Modeler 15

Nota: Antes de utilizar esta información y el producto que admite, lea la información general en Avisos el p. .

Esta edición se aplica a IBM SPSS Modeler 15 y a todas las versiones y modificaciones posteriores hasta que se indique lo contrario en nuevas ediciones.

Capturas de pantalla de productos de Adobe reimpresas con permiso de Adobe Systems Incorporated.

Capturas de pantalla de productos de Microsoft reimpresas con permiso de Microsoft Corporation.

Materiales bajo licencia: Propiedad de IBM

Prefacio

IBM® SPSS® Modeler es el 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 la pericia empresarial específica de los usuarios, lo que deriva en modelos predictivos más eficaces y la reducción del tiempo necesario para encontrar soluciones. SPSS Modeler ofrece muchas técnicas de modelado tales como pronósticos, 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

El software IBM Business Analytics ofrece información completa, coherente y precisa en la que los órganos de toma de decisiones confían para mejorar el rendimiento comercial. Un conjunto integral de inteligencia empresarial, análisis predictivo,, rendimiento comercial y gestión de estrategias, así como de aplicaciones de análisis le ofrece una información clara, inmediata e interactiva del rendimiento actual y la capacidad para 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 y alcanzar mejores resultados.

Como parte de esta familia, el software de análisis predictivo de IBM SPSS ayuda a las organizaciones a predecir eventos futuros y actuar proactivamente según esa información para lograr mejores resultados comerciales. Los clientes comerciales, gubernamentales y académicos de todo el mundo confían en la tecnología de IBM SPSS como ventaja ante la competencia para atraer, retener y hacer crecer los clientes, reduciendo al mismo tiempo el fraude y reduciendo el riesgo. 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

La asistencia técnica está disponible para el mantenimiento de los clientes. 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, visite elIBM Corp. sitio Web en *http://www.ibm.com/support*. Prepárese para identificarse, identificar a su organización y su acuerdo de asistencia al solicitar asistencia.

Contenido

Apéndice

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 utilizarlos en operaciones empresariales para mejorar la toma de decisiones. Con un diseño que sigue el modelo CRISP-DM, estándar del sector, SPSS Modeler admite el proceso completo de minería de datos, desde los propios datos hasta obtener los mejores resultados empresariales.

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 *http://www.ibm.com/software/analytics/spss/products/modeler/*.

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
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server adaptadores 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. Si desea obtener más información, consulte el tema Ediciones de IBM SPSS Modeler en *Manual de usuario de IBM SPSS Modeler 15*.

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 este modo, SPSS Modeler Server ofrece un mejor rendimiento cuando se trabaja con grandes conjuntos de datos, ya que las operaciones que requieren un uso intensivo de memoria se pueden realizar en el servidor sin tener que descargar datos al equipo cliente. IBM® SPSS® Modeler Server también ofrece asistencia para las capacidades de optimización de SQL y modelado interno de la base de datos, lo que proporciona mayores ventajas en cuanto al rendimiento y la automatización.

IBM SPSS Modeler Administration Console

Modeler Administration Console es una aplicación gráfica para administrar muchas de las opciones de configuración de SPSS Modeler Server, las cuales también pueden configurarse a través de un archivo de opciones. La aplicación proporciona una interfaz de usuario de la consola para supervisar y configurar las instalaciones de SPSS Modeler Server y está disponible de forma completamente 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 por parte del usuario. SPSS Modeler Batch es una versión especial del producto que ofrece asistencia para todas las capacidades analíticas de SPSS Modeler sin acceder a la interfaz de usuario habitual. Es necesario disponer de una licencia de SPSS Modeler Server para 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 distribuir 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 - Scoring, para el que se necesita una licencia independiente. Con esta licencia, recibirá SPSS Modeler Solution Publisher Runtime, que le permite ejecutar las rutas publicadas.

IBM SPSS Modeler Server Adaptadores 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 distribuida 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 poder trabajar con datos especializados, como los utilizados para el análisis de entidades o las redes sociales, así como con datos de texto no estructurados. SPSS Modeler Premium está formado por los siguientes componentes:

IBM® SPSS® Modeler Entity Analytics incorpora una dimensión completamente nueva al análisis predictivo de IBM® SPSS® Modeler. Mientras que el análisis predictivo trata de predecir comportamientos futuros a partir de datos del pasado, el análisis de entidades se centra en mejorar la coherencia de los datos actuales mediante la resolución de conflictos de identidades dentro de los propios registros. La identidad de un individuo, una organización, un objeto o cualquier otra entidad puede estar expuesta a ambigüedades. La resolución de identidades puede ser vital en diversos campos, entre los que se incluyen la gestión de la relación con el cliente, la detección de fraudes, la lucha contra el blanqueo de dinero y la seguridad nacional e internacional.

IBM SPSS Modeler Social Network Analysis transforma la información sobre relaciones en campos que caracterizan el comportamiento social de individuos y grupos. Mediante el uso de datos que describen las relaciones subyacentes de las redes sociales, IBM® SPSS® Modeler Social Network Analysis identifica a los líderes sociales que influyen en el comportamiento de otros en la red. Además, puede determinar qué personas se ven más afectadas por otros participantes de la red. Al combinar estos resultados con otras medidas, puede crear perfiles completos de individuos en los que basar sus modelos predictivos. Los modelos que incluyan esta información social tendrán un mejor rendimiento que los modelos que no la incluyan.

Text Analytics for IBM® SPSS® Modeler 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 SPSS Modeler para tomar decisiones mejores y más certeras.

Documentación de IBM SPSS Modeler

Tiene a su disposición documentación en formato de ayuda en línea desde el menú Ayuda de SPSS Modeler. Se incluye documentación para SPSS Modeler, SPSS Modeler Server y SPSS Modeler Solution Publisher, así como el Manual de aplicaciones y otros materiales de apoyo.

La documentación completa de cada producto (incluidas las instrucciones de instalación) en formato PDF está disponible en la carpeta *\Documentation* en cada DVD del producto. También es posible descargar los documentos de instalación en Internet en *http://www-01.ibm.com/support/docview.wss?uid=swg27023172*.

La documentación en ambos formatos también está disponible desde el centro de información de SPSS Modeler en *http://publib.boulder.ibm.com/infocenter/spssmodl/v15r0m0/*.

Documentación de SPSS Modeler Professional

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

- **Manual del usuario de IBM SPSS Modeler.** Introducción general sobre cómo usar SPSS Modeler, incluyendo cómo crear rutas de datos, tratar valores perdidos, crear expresiones CLEM, trabajar con proyectos e informes y empaquetas rutas para su distribución en IBM SPSS Collaboration and Deployment Services, Predictive Applications o IBM SPSS Modeler Advantage.
- **Nodos Fuente, 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 variedad de métodos de modelado tomados del aprendizaje de las máquinas, la inteligencia artificial y

la estadística. Si desea obtener más información, consulte el tema Conceptos básicos sobre nodos de modelado en el capítulo 3 el p. 26.

- **Manual de algoritmos de IBM SPSS Modeler.** Descripciones de los fundamentos matemáticos de los métodos de modelado que se utilizan en SPSS Modeler. Esta guía está disponible únicamente en formato PDF.
- **Manual de aplicaciones de IBM SPSS Modeler.** Los ejemplos de esta guía ofrecen introducciones breves y concisas a métodos y técnicas de modelado específicos. También tiene a su disposición una versión en línea de este manual en el menú Ayuda. Si desea obtener más información, consulte el tema Ejemplos de aplicaciones en *Manual de usuario de IBM SPSS Modeler 15*.
- **Procesos y automatización de IBM SPSS Modeler.** Información sobre la automatización del sistema a través de procesos, incluidas las propiedades que se pueden utilizar para manipular nodos y rutas.
- **IBM SPSS Modeler Manual de distribución.** Información sobre la ejecución de rutas y escenarios de SPSS Modeler como pasos en trabajos de procesamiento en IBM® SPSS® Collaboration and Deployment Services Deployment Manager.
- **Guía del desarrollador de IBM SPSS Modeler CLEF.**CLEF permite integrar programas de otros fabricantes, como rutinas de procesamiento de datos o algoritmos de modelado como nodos en SPSS Modeler.
- **Manual de minería interna de bases 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.
- **Manual del usuario de IBM SPSS Modeler Administration Console.** Información sobre cómo instalar y utilizar la interfaz de usuario de la consola para supervisar y configurar SPSS Modeler Server. La consola se implementa como complemento de la aplicación Deployment Manager.
- **Manual de IBM SPSS Modeler Solution Publisher.** SPSS Modeler Solution Publisher es un componente complementario que permite a las organizaciones publicar rutas para su uso fuera del entorno estándar de SPSS Modeler.
- **Manual 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.
- **Manual del 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.

Documentación de SPSS Modeler Premium

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

- **IBM SPSS Modeler Entity Analytics Manual del usuario.** Información sobre cómo utilizar el análisis de entidades con SPSS Modeler, que cubre la instalación y configuración de repositorios, nodos de análisis de entidades y tareas administrativas.
- **IBM SPSS Modeler Social Network Analysis Manual del usuario.** Una guía para realizar análisis de redes sociales con SPSS Modeler, incluido el análisis de grupos y el análisis de difusión.
- **Text Analytics for SPSS Modeler Manual del usuario.** 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.
- Manual del usuario de **Text Analytics for IBM SPSS Modeler Administration Console.** Información sobre cómo instalar y utilizar la interfaz de usuario de la consola para supervisar y configurar IBM® SPSS® Modeler Server para su uso con Text Analytics for SPSS Modeler. La consola se implementa como complemento de la aplicación Deployment Manager.

Ejemplos de aplicaciones

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 algunos analizadores de datos, pero los conceptos y métodos implicados deberían ser escalables a las aplicaciones reales.

Para acceder a los ejemplos pulsando Ejemplos de aplicación en el menú Ayuda de SPSS Modeler. Los archivos de datos y rutas de muestra se instalan en la carpeta *Demos* en el directorio de instalación del producto. Si desea obtener más información, consulte el tema Carpeta Demos en *Manual de usuario de IBM SPSS Modeler 15*.

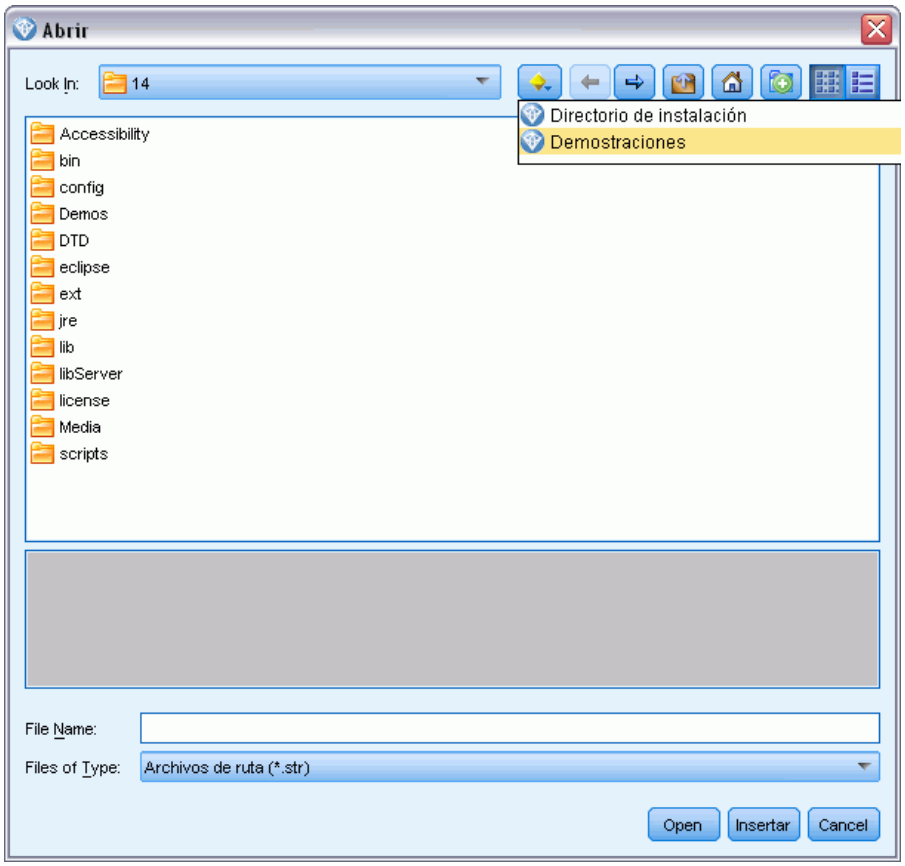
Ejemplos de modelado de base de datos. Consulte los ejemplos que figuran en el Manual de minería interna de bases de datos de *IBM SPSS Modeler*.

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

Carpeta Demos

Los archivos de datos y rutas de muestra utilizados con los ejemplos de la aplicación se instalan en la carpeta *Demos* en el directorio de instalación del producto. También puede acceder a esta carpeta desde el grupo de programas IBM SPSS Modeler 15 del menú Inicio de Windows o pulsando *Demos* de la lista de directorios recientes en el cuadro de diálogo Abrir archivo.

Figura 1-1
Selección de la carpeta Demos desde la lista de directorios utilizados recientemente



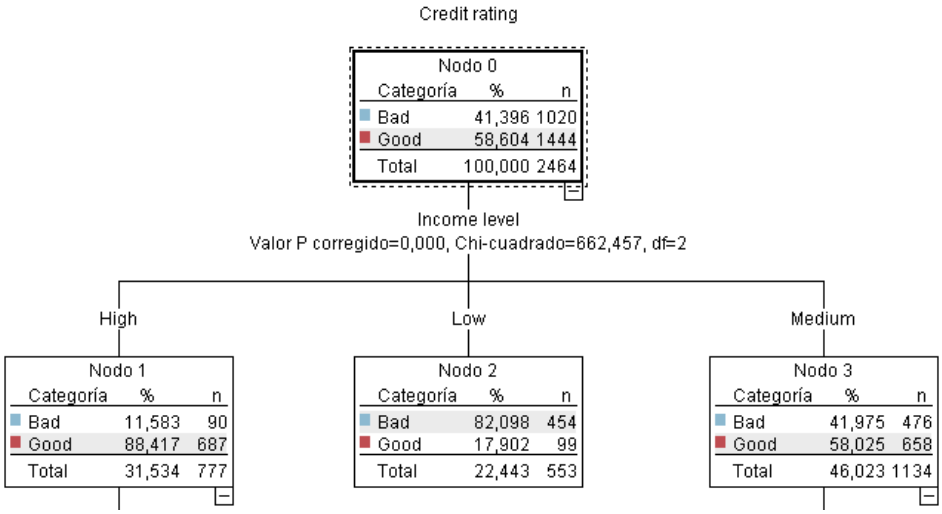
Capítulo
2

Introducción al modelado

Un modelo es un conjunto de reglas, fórmulas o ecuaciones que puede utilizarse para pronosticar 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 pronosticar 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 2-1
Modelo de árbol de decisión sencillo



Este ejemplo utiliza un modelo de **árbol de decisión** que clasifica los registros (y pronostica 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 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
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*. Si desea obtener más información, consulte el tema Carpeta Demos en el capítulo 1 en *Manual de usuario de IBM SPSS Modeler 15*.

Veamos la ruta más detenidamente.

- Seleccione lo siguiente en el menú principal:
  File > Abrir ruta

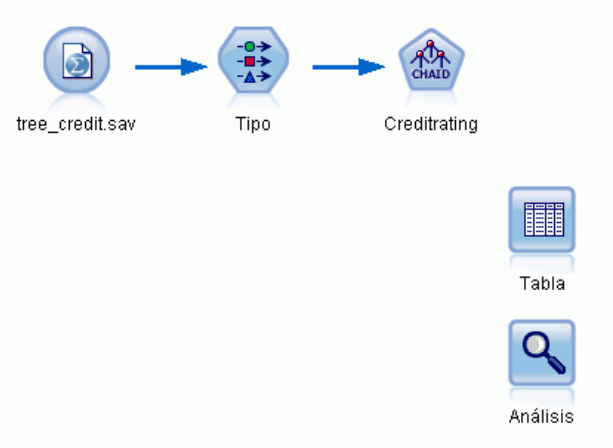
- Pulse en el icono de nugget dorado de la barra de herramientas del cuadro de diálogo Abrir y seleccione la carpeta Demos.

- Pulse dos veces en la carpeta *streams*.

- Pulse dos veces en el archivo llamado *modelingintro.str*.

Generación de la ruta

Figura 2-2
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 papel 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.

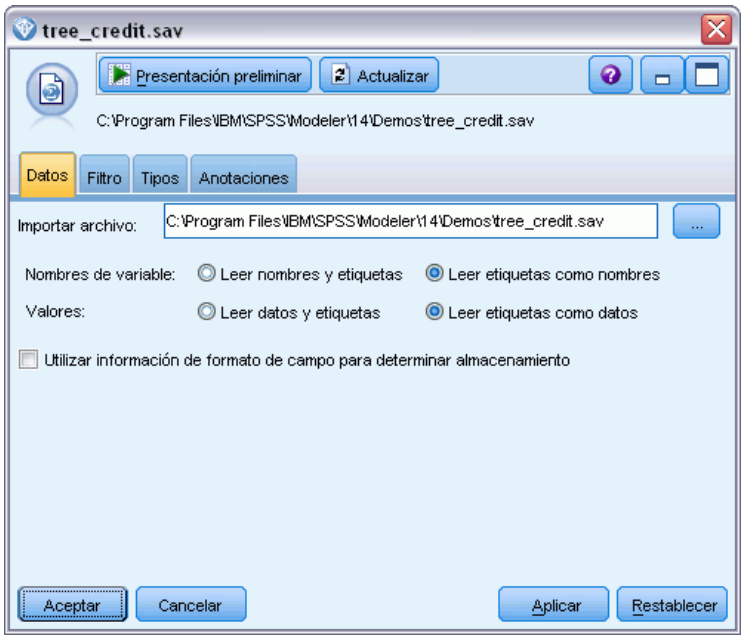
Si se especifican niveles de medición en el nodo de origen, se puede eliminar el nodo Tipo independiente. Funcionalmente, el resultado es el mismo.

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 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 2-3
Lectura de datos con un nodo de origen Archivo Statistics

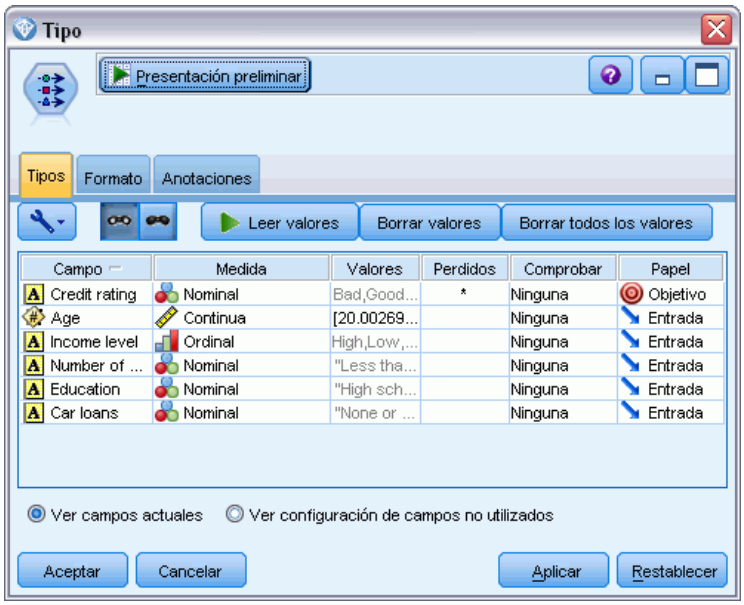


El nodo Tipo especifica el **nivel de medición** de cada campo. El nivel de medición es una categoría que indica el tipo de datos del campo. Nuestro archivo de datos de origen utiliza tres niveles de medición diferentes.

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 2-4
Configuración de los campos de destino y entrada con el nodo Tipo



Para cada campo, el nodo Tipo también especifica un **papel** para indicar la función que desempeña cada campo en el modelado. El papel 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 pronosticar.

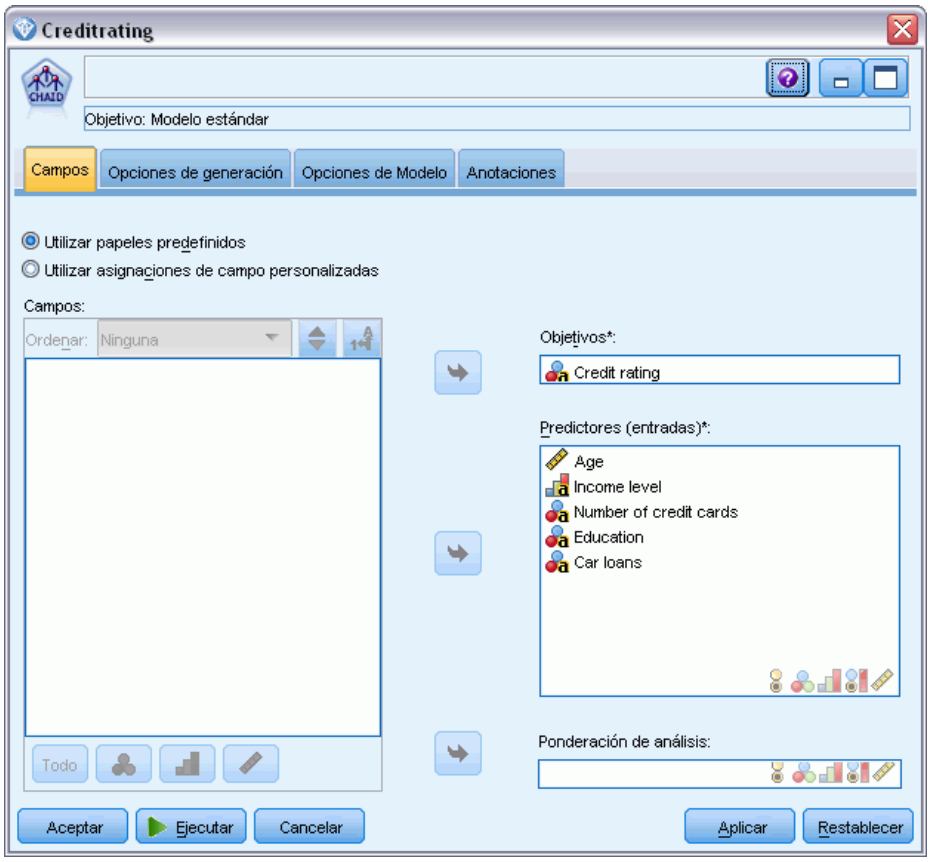
El papel se define a *Entrada* para los otros campos. Los campos de entrada se conocen a menudo como **predictores**, o campos cuyos valores se utilizan en el algoritmo de modelado para predecir el valor del campo objetivo.

El nodo de modelado CHAID genera el modelo.

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

▶ Pulse en la pestaña Crear opciones.

Figura 2-5
Nodo de modelado CHAID, pestaña Campos



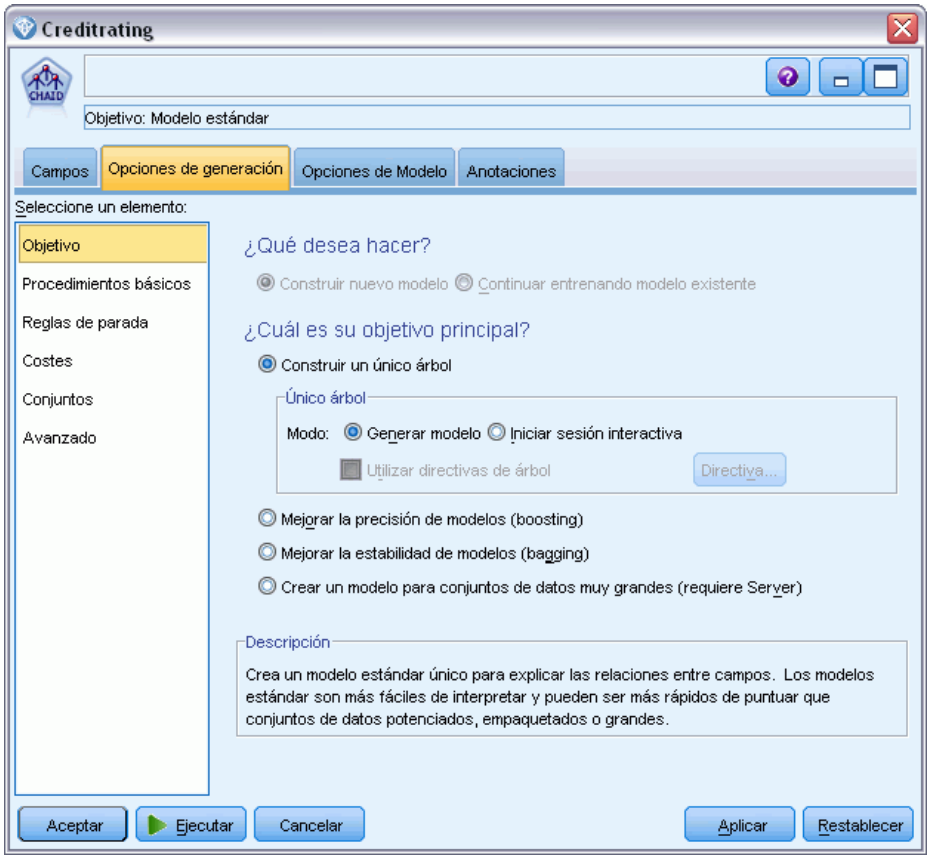
Aquí hay varias opciones en las que podemos especificar el tipo de modelo que queremos generar.

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 predeterminado 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 por defecto Generar modelo.

Figura 2-6
Nodo de modelado CHAID, pestaña Opciones de generación

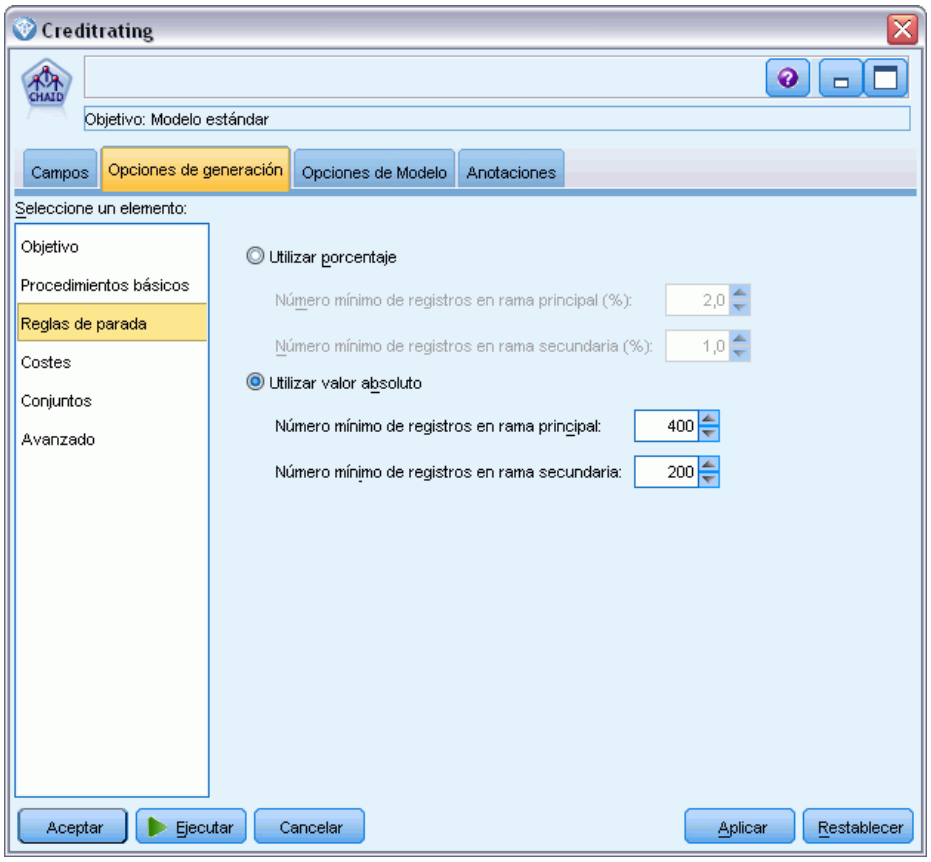


Por ejemplo, queremos que el árbol sea bastante sencillo, así que limitaremos el crecimiento del árbol elevando el número mínimo de casos para los nodos principales y filiales.

- En la pestaña Opciones de generación, seleccione Reglas de parada desde el panel de navegación de la izquierda.
- Seleccione la opción Utilizar valor absoluto.
- Establezca Número mínimo de registros en rama parental como 400.

▶ Establezca Número mínimo de registros por rama filial como 200.

Figura 2-7
Configuración de los criterios de parada para la generación de árboles de decisión



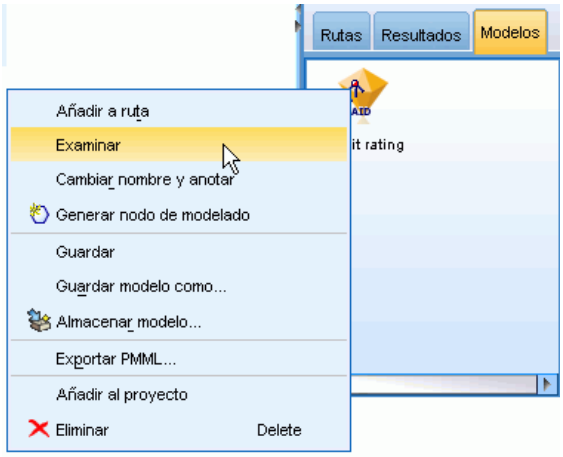
Podemos usar todas las demás opciones predeterminadas para este ejemplo, por lo que pulse en Ejecutar para crear el modelo. (También puede pulsar con el botón derecho del ratón en el nodo y seleccionar Ejecutar del menú contextual o seleccionar el nodo y Ejecutar del menú Herramientas.)

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

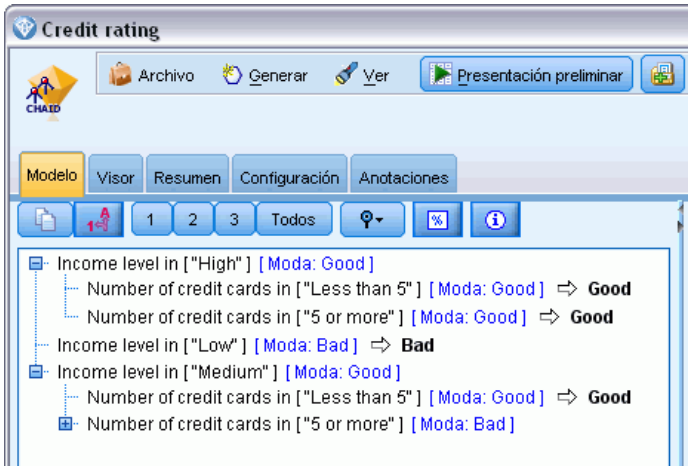
botón derecho del ratón en el nugget y seleccione Examinar (en la paleta de modelos) o Editar (en el lienzo).

Figura 2-8
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 filiales basándose en los valores de distintos campos de entrada.

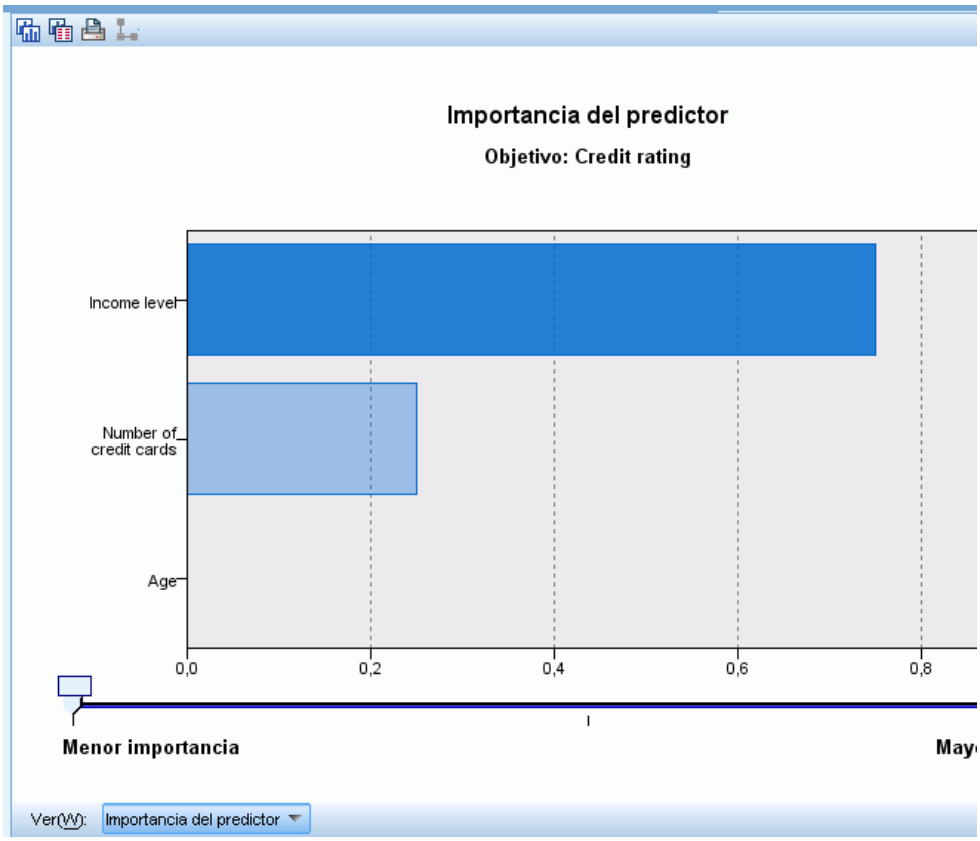
Figura 2-9
Nugget de modelo CHAID, conjunto de reglas



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, el pronóstico está determinado por el **modo** o, la respuesta más común, para registros que se incluyen en dicho nodo.

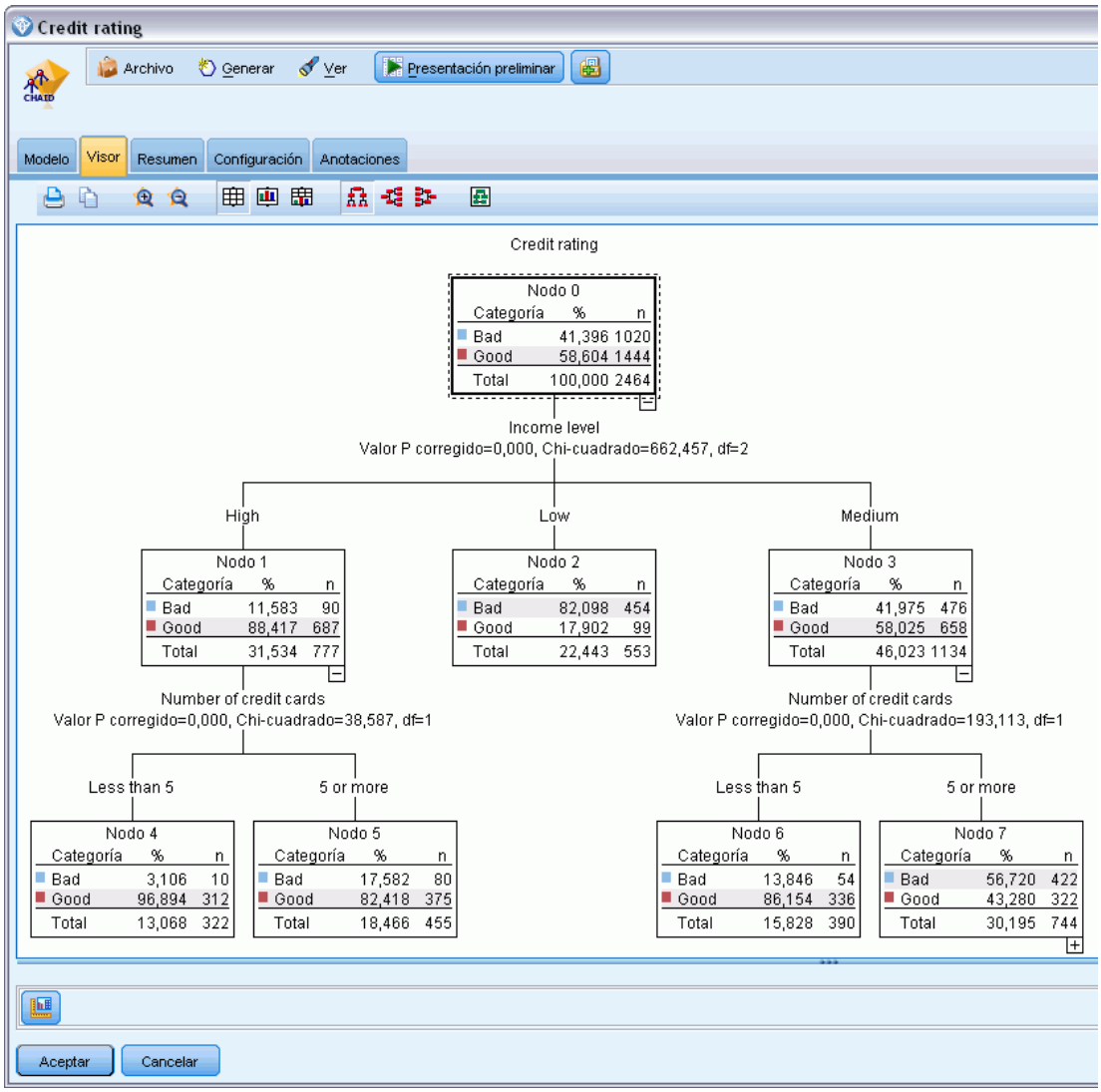
A la derecha del conjunto de reglas, la pestaña Predictor muestra el gráfico Importancia de variable, que muestra la importancia relativa de cada predictor en la estimación del modelo. A partir de aquí podemos determinar que *Nivel de ingresos* es fácilmente lo más significativo de este caso, y que el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

Figura 2-10
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 2-11
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.

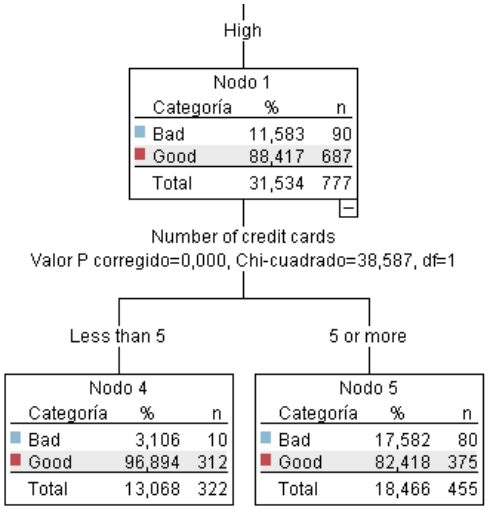
Podemos ver que la primera división es por *Nivel de ingresos*. Los registros cuyo nivel de ingresos están en la categoría *Bajo* se asignan al Nodo 2, por lo que no es sorprendente que esta categoría contenga el mayor porcentaje de morosos de préstamos. Claramente, la concesión de un préstamo a clientes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

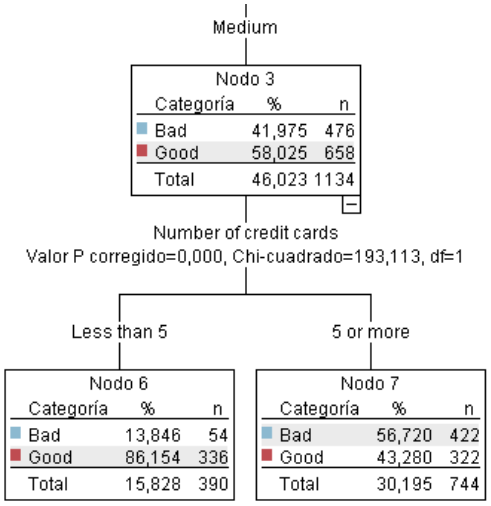
Figura 2-12
Vista de árbol de clientes con ingresos elevados



¿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 2-13
Vista de árbol de clientes con ingresos medios



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

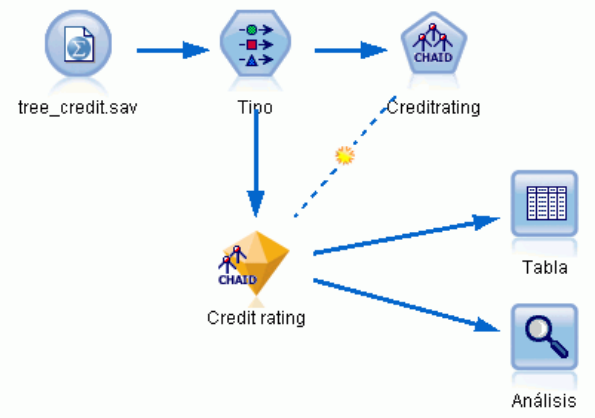
Este proceso de asignar pronósticos a registros individuales se conoce como **puntuación**. Al puntuar los mismos registros utilizados para calcular el modelo, podemos evaluar cuál es el rendimiento preciso en los datos de entrenamiento, es decir, los datos para los que conocemos el resultado. Veamos 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 pronosticadas 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 2-14
Adición del nugget de modelo a los nodos de salida para la generación del modelo



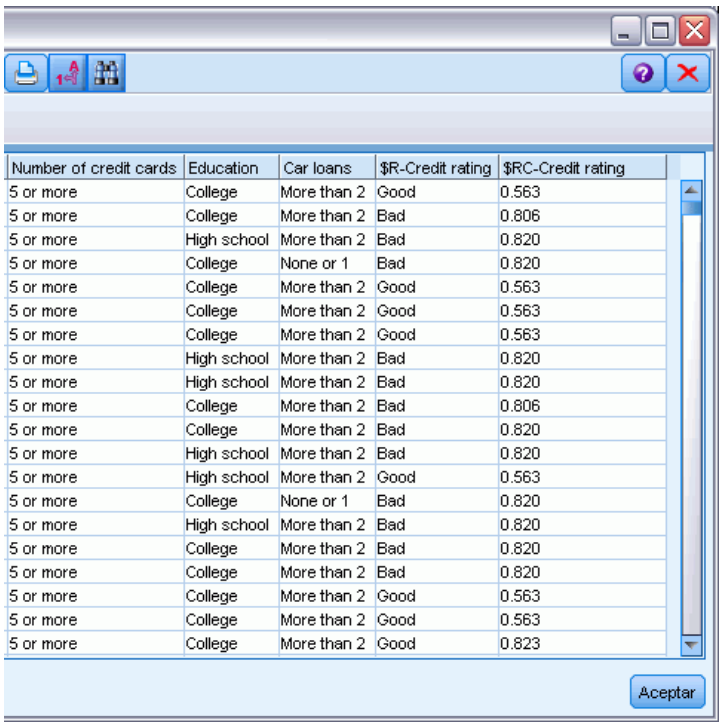
- Para ver las puntuaciones o pronósticos, adjunte el nodo Tabla al nugget de modelo, pulse dos veces en el nodo Tabla y pulse en Ejecutar.

  La tabla muestra las puntuaciones pronosticadas 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 como *$R-* para pronósticos o *$RC-* para valores de confianza. Los distintos tipos de modelo utilizan diferentes conjuntos de prefijos. Un **valor de**

confianza es la estimación del propio modelo, en una escala de 0,0 a 1,0, sobre el grado de precisión de cada valor pronosticado.

Figura 2-15
Tabla que muestra las puntuaciones generadas y los valores de confianza



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Bad	0.806
5 or more	High school	More than 2	Bad	0.820
5 or more	College	None or 1	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.806
5 or more	College	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Good	0.563
5 or more	College	None or 1	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.823

Aceptar

Como se esperaba, el valor pronosticado coincide con las respuestas reales de muchos registros, pero no todos. El motivo es que cada nodo terminal CHAID tiene una mezcla de respuestas. El pronóstico 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.)

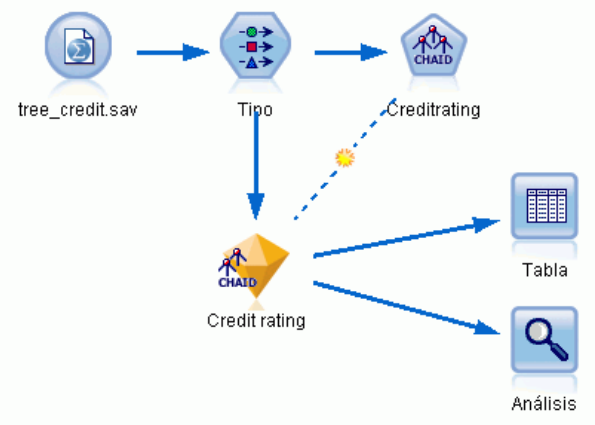
Para evitarlo, podemos seguir dividiendo el árbol en ramas cada vez más pequeñas, hasta que cada nodo sea 100 % puro: todas las respuestas son *Bueno* o *Malo* sin respuestas mezcladas. Pero dicho modelo sería extremadamente complicado y probablemente no se generalizaría bien en otros conjuntos de datos.

Para descubrir exactamente cuántas predicciones son correctas, podríamos observar la tabla y anotar el número de registros en los que el valor del campo pronosticado *$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.

- Conecte el nugget de modelo al nodo Análisis.

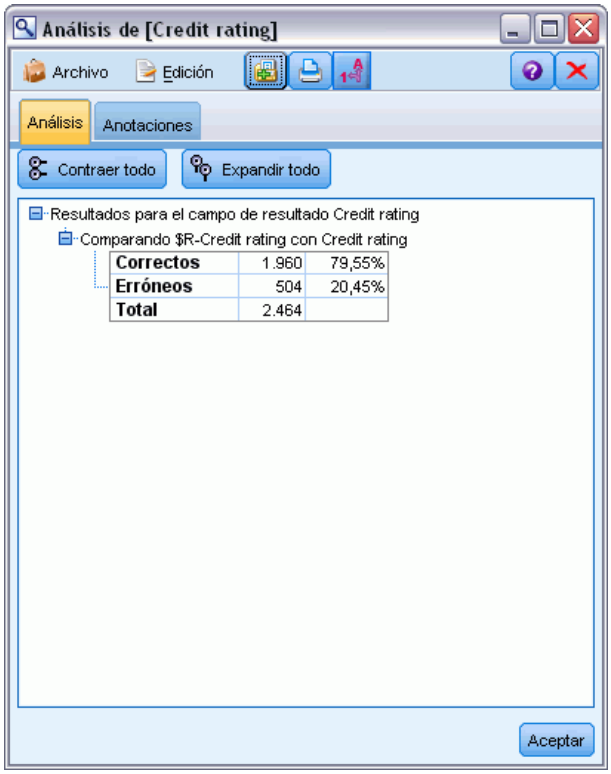
▶ Pulse dos veces en el nodo Análisis y pulse en Ejecutar.

Figura 2-16
Conexión del nodo Análisis



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

Figura 2-17
Resultados de análisis que comparan respuestas observadas y pronosticadas



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.

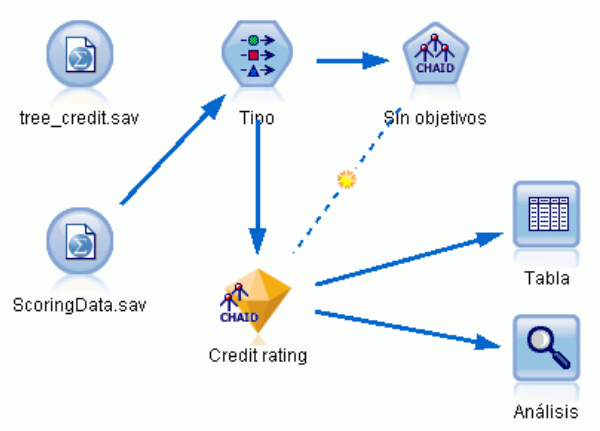
Si utiliza una partición de muestra para generar el modelo y otra muestra para comprobarlo, podrá obtener una indicación mucho mejor de lo bien que se generalizará en otros conjuntos de datos.

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. Ésta es la meta del modelado con un campo objetivo: Estudie los registros de los que conoce los resultados para identificar patrones que le permitirán pronosticar resultados que todavía no conoce.

Figura 2-18
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, 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 distribución, 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 pronosticar 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

Conceptos básicos sobre 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.

El *Manual de aplicaciones de SPSS Modeler* ofrece ejemplos para muchos de estos métodos, junto con una introducción general al proceso de modelado. Este manual está disponible como tutorial en línea y también en formato PDF. Si desea obtener más información, consulte el tema Ejemplos de aplicaciones en el capítulo 1 en *Manual de usuario de IBM SPSS Modeler 15*.

Los métodos de modelado se dividen en tres categorías:

- Clasificación
- Asociación
- Segmentación.

Modelos de clasificación

Los *modelos de clasificación* usan el valor de uno o más campos de **entrada** para predecir el valor de uno o más resultados o campos de **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 de clasificación ayudan a las organizaciones a pronosticar un resultado conocido, como saber 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 de clasificación

El nodo Clasificador automático crea y compara varios modelos diferentes para obtener resultados binarios (sí o no, pérdida 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. Si desea obtener más información, consulte el tema Nodo Clasificador automático en el capítulo 5 el p. 97.

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. Si desea obtener más información, consulte el tema Nodo Autonumérico en el capítulo 5 el p. 107.

El nodo de árbol de clasificación y regresión (C&R) genera un árbol de decisión que permite pronosticar 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). Si desea obtener más información, consulte el tema Nodo Árbol C&R en el capítulo 6 el p. 152.

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. Si desea obtener más información, consulte el tema Nodo QUEST en el capítulo 6 el p. 154.

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. Si desea obtener más información, consulte el tema Nodo CHAID en el capítulo 6 el p. 153.

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. Si desea obtener más información, consulte el tema Nodo C5.0 en el capítulo 6 el p. 174.

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 perder clientes 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. Si desea obtener más información, consulte el tema Lista de decisiones en el capítulo 9 el p. 219.

Los modelos de regresión lineal predicen un destino continuo tomando como base las relaciones lineales entre el destino y uno o más predictores. Si desea obtener más información, consulte el tema Modelos lineales en el capítulo 10 el p. 258.

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. Si desea obtener más información, consulte el tema Nodo PCA/Factorial en el capítulo 10 el p. 299.

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? Si desea obtener más información, consulte el tema Nodo Selección de características en el capítulo 4 el p. 76.

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. Si desea obtener más información, consulte el tema Nodo Discriminante en el capítulo 10 el p. 307.

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. Si desea obtener más información, consulte el tema Nodo Logística en el capítulo 10 el p. 278.

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 que no sea 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. Si desea obtener más información, consulte el tema Nodo GenLin en el capítulo 10 el p. 316.

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. Si desea obtener más información, consulte el tema Nodo GLMM en el capítulo 10 el p. 331.

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 pronostica la probabilidad de que el evento de interés se haya producido en el momento dado (*t*) para valores determinados de las variables de entrada. Si desea obtener más información, consulte el tema Nodo Cox en el capítulo 10 el p. 359.

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. Si desea obtener más información, consulte el tema Nodo SVM en el capítulo 15 el p. 489.

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 cadena de Markov que se utilizan principalmente para la clasificación. Si desea obtener más información, consulte el tema Nodo Red bayesiana en el capítulo 7 el p. 194.

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. Si desea obtener más información, consulte el tema Nodo SLRM en el capítulo 14 el p. 476.

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 datos de predicciones. Un nodo Serie temporal debe ir siempre precedido por un nodo Intervalos de tiempo. Si desea obtener más información, consulte el tema Nodo Modelos de series temporales en el capítulo 13 el p. 453.

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. Si desea obtener más información, consulte el tema Nodo KNN en el capítulo 16 el p. 495.

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 pronosticar 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 A priori extrae un conjunto de reglas de los datos y destaca aquellas reglas con un mayor contenido de información. A priori 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, A priori 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. A priori 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. Si desea obtener más información, consulte el tema Nodo A priori en el capítulo 12 el p. 405.

El modelo CARMA extrae un conjunto de reglas de los datos sin necesidad de especificar campos de entrada ni de objetivo. A diferencia de A priori, 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. Si desea obtener más información, consulte el tema Nodo CARMA en el capítulo 12 el p. 409.

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, es probable que un cliente que compra una cuchilla y una loción para después del afeitado compre 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. Si desea obtener más información, consulte el tema Nodo Secuencia en el capítulo 12 el p. 431.

Modelos de segmentación

Los *modelos de segmentación* dividen los datos en segmentos o conglomerados 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, los conglomerados de K-medias, los conglomerados en dos pasos y la detección de anomalías.

Los modelos de segmentación (también conocidos como "modelos de conglomerados") 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 conglomerados 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 conglomerados de otras técnicas de modelado en que no hay campos de salida u objetivo predefinidos para el modelo que se va a pronosticar. 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 conglomerado se usan a menudo para crear conglomerados 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 Autoconglomeración calcula y compara los modelos de conglomerado que identifican grupos de registros con características similares. El nodo funciona de la misma manera que otros nodos de modelado de conglomerado, 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 conglomerado y proporcionar una medida según la importancia de campos concretos. Si desea obtener más información, consulte el tema Nodo Autoconglomeración en el capítulo 5 el p. 113.

El nodo K-medias agrupa conjuntos de datos en grupos distintos (o conglomerados). El método define un número fijo de conglomerados, de forma iterativa asigna registros a los conglomerados y ajusta los centros de los conglomerados hasta que no se pueda mejorar el modelo. En lugar de intentar pronosticar 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. Si desea obtener más información, consulte el tema Nodo K-medias en el capítulo 11 el p. 378.

El nodo Kohonen genera un tipo de red neuronal que se puede usar para conglomerar 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

conglomerados. Si desea obtener más información, consulte el tema Nodo Kohonen en el capítulo 11 el p. 371.

El nodo Bietápico es un método de conglomerado 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 subconglomerados administrable. El segundo paso utiliza un método de conglomerado jerárquico para fundir progresivamente los subconglomerados en conglomerados cada vez más grandes. El bietápico tiene la ventaja de estimar automáticamente el número óptimo de conglomerados para los datos de entrenamiento. Puede gestionar tipos de campos mixtos y grandes conjuntos de datos eficazmente. Si desea obtener más información, consulte el tema Nodo de conglomerado Bietápico en el capítulo 11 el p. 383.

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. Si desea obtener más información, consulte el tema Nodo Detección de anomalías en el capítulo 4 el p. 84.

Modelos de minería interna de la base de datos

SPSS Modeler admite la integración con herramientas de modelado y minería de datos que están disponibles en proveedores de bases de datos como Oracle Data Miner, IBM DB2 InfoSphere Warehouse y Microsoft Analysis Services. Podrá crear, puntuar y almacenar modelos dentro de la base de datos, todo desde la aplicación SPSS Modeler. Para obtener los detalles completos, consulte el *SPSS Modeler Manual de minería interna de bases de datos*, disponible en DVD.

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 SPSS Statistics en SPSS Modeler para generar y puntuar modelos. Si desea obtener más información, consulte el tema Conceptos básicos de nodos de IBM SPSS Statistics en el capítulo 8 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Más información

También hay disponible información detallada sobre el modelado de algoritmos. Si desea obtener más información, consulte el *Manual de algoritmos de SPSS Modeler*, disponible en el DVD del producto.

Generación de modelos divididos

Los modelos divididos permiten utilizar una sola ruta para generar modelos diferentes para cada posible valor de una marca o campos de entrada nominales o continuos, con todos los modelos resultantes accesibles desde un nugget de modelo simple. Los posibles valores de campos de entrada pueden tener efectos muy diferentes en el modelo. Con los modelos divididos 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.

Los modelos divididos funcionan designando un campo de entrada concreto como un campo de división. Puede hacerlo definiendo el papel del campo como Dividir en la especificación de tipo:

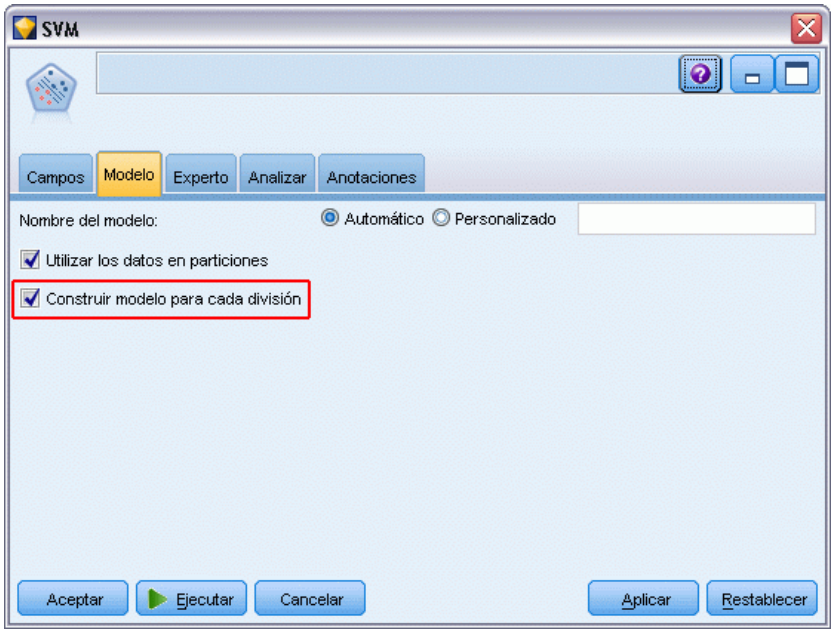
Figura 3-1
Designación de campo de entrada como campo de división

Sólo puede designar campos con un nivel de medida de Marca, Nominal, Ordinal o Continuo como campos divididos.

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 divididos 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 de división o un modelo único, mediante una casilla de verificación en el cuadro de diálogo del nodo de modelado:

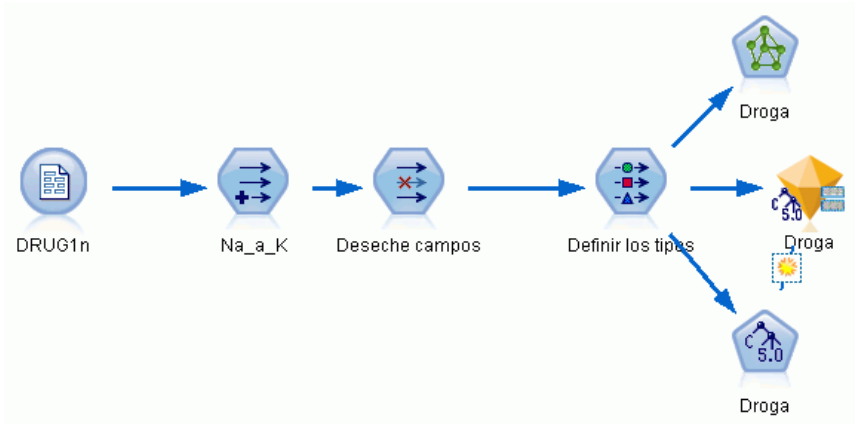
Figura 3-2
Selección para generar modelos de división



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.

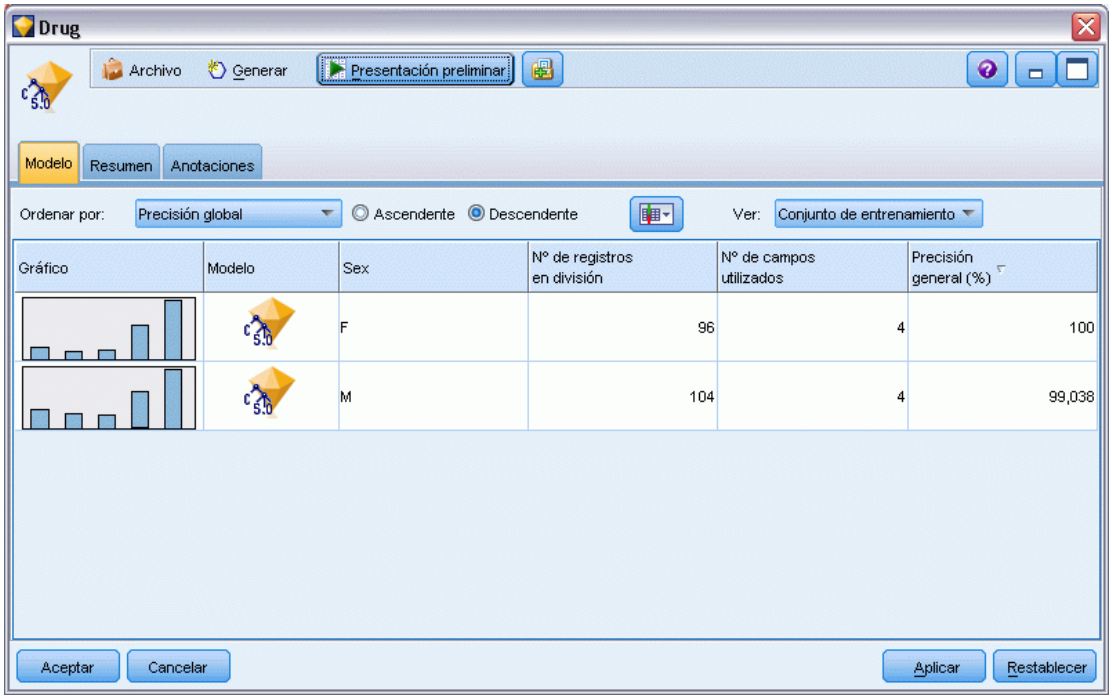
Si ejecuta la ruta, se genera un modelo diferente en segundo plano para cada valor posible del campo o campos de división, pero solo se coloca un nugget de modelo en la paleta de modelos y el lienzo de rutas. Un nugget de modelo dividido se distingue por el símbolo de división:

Figura 3-3
Nugget de modelo dividido en una ruta



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

Figura 3-4
Visor del modelo de división



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. Si desea obtener más información, consulte el tema Visor del modelo de división el p. 67.

Una vez se ha creado un modelo como modelo de división, 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 dividido, 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 funciones 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, comprobación 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 de división

Numerosos nodos de modelado pueden crear modelos divididos. Las excepciones son Autoconglomerado, Serie temporal, PCA/Factor, Selección de características, SLRM, los modelos de asociación (A priori, Carma y Sequence), los modelos de conglomerado ((K-medias, Kohonen, Bietápica y Anomalías), Modelo Statistics y los nodos utilizados en el modelado de base de datos.

Los nodos de modelado compatibles con modelos divididos son:

Árbol C&R

Red bayesiana

QUEST

GenLin

CHAID

KNN

C5.0

Cox

Red neuronal

Clasificador automático

Lista de decisiones

Autonumérico

Regresión

Logística

Discriminante

SVM

Funciones afectadas por la división

El uso de los modelos divididos afecta al número de funciones de IBM® SPSS® Modeler de varias formas. Esta sección proporciona ayuda acerca del uso de modelos divididos, junto con otros nodos en una ruta.

Nodos Operaciones con registros

Si utiliza modelos de división en una ruta que contiene un nodo **Muestra**, estratifique los registros por el campo de división para lograr un muestreo de registro uniforme. Está opción está disponible si selecciona Complejo como el método de muestra. Si desea obtener más información, consulte el tema Configuración de Conglomerado y estratificación en el capítulo 3 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Si la ruta contiene un nodo **Equilibrar**, tenga en cuenta que el equilibrado se aplica al conjunto de registros de entrada, no al subconjunto de registros de una división.

SI agrega registros mediante un nodo **Agregar**, defina los campos de división como campos clave si desea calcular los agregados de cada división. Si desea obtener más información, consulte el tema Nodo Agregar en el capítulo 3 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Nodos Operaciones con campos

El nodo **Tipo** es donde especifica los campos que se usarán como campos de división. Si desea obtener más información, consulte el tema Nodo Tipo en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Tenga en cuenta que, aunque el nodo **Conjunto** se utiliza para combinar dos o más nuggets de modelo, no se puede utilizar para invertir la acción de división, ya que los modelos de división están contenidos en 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.

El nodo **KNN** (vecino más cercano) admite modelos divididos solo si se define para pronosticar un campo objetivo. El ajuste alternativo (identificar vecinos más cercanos únicamente) no crea un modelo. Si la opción “Seleccionar k automáticamente” está seleccionada, cada uno de los modelos de división puede tener un número diferente de vecinos más cercanos. Además, el modelo final tendrá un número de columnas generadas igual al mayor número de vecinos más cercanos en todos los modelos de división. Para los modelos divididos en los que el número de vecinos más cercanos es inferior al máximo, existirá un número de columnas correspondiente con valores $null. Si desea obtener más información, consulte el tema Nodo KNN en el capítulo 16 el p. 495.

Nodos Modelado de base de datos

Los nodos de modelado interno de la base de datos no admiten modelos de división.

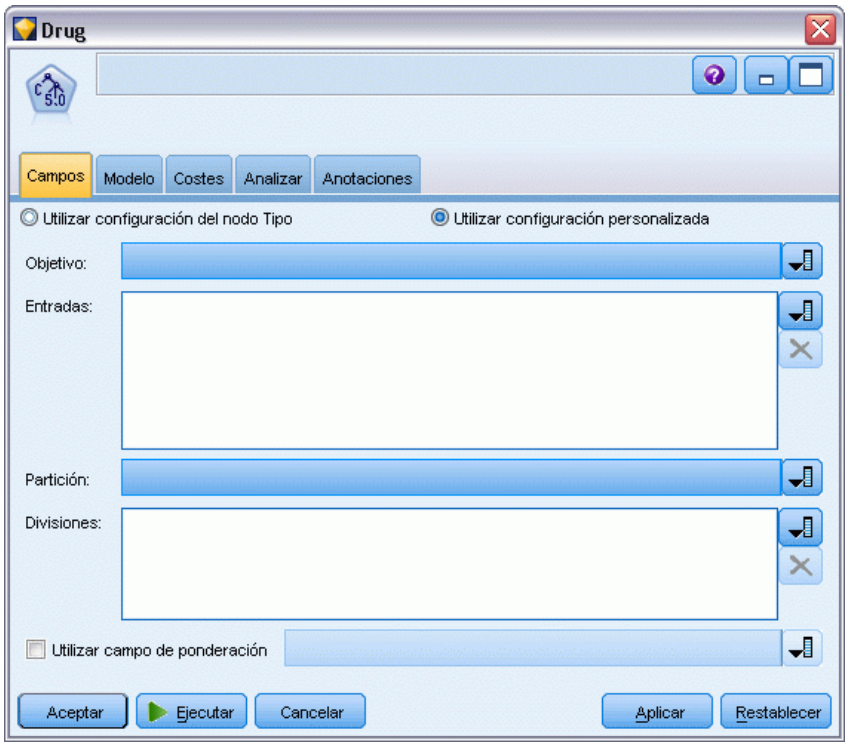
Nuggets de modelo

No es posible **Exportar a PMML** desde un modelo de división, ya que el nugget contiene múltiples modelos y PMML no admite ese empaquetado. Sin embargo, 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.

Figura 3-5
Ejemplo de una pestaña Campos de nodo de modelado



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 por defecto.

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 muestran 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. Si desea obtener más información, consulte el tema Datos tabulares frente a datos transaccionales en el capítulo 12 el p. 404.
- **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 único de este campo debe indicar una unidad de análisis específica. Por ejemplo, en una aplicación de la cesta del supermercado, 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 papel del campo en *Objetivo* en un nodo Tipo.
- **Evaluación.** (Para modelos de Autoconglomerado únicamente). No se ha especificado un objetivo para los modelos de conglomerado; sin embargo, puede seleccionar un campo de evaluación para identificar su nivel de importancia. Además, puede evaluar la calidad con la que los conglomerados diferencian los valores de este campo, que a su vez indica si los conglomerados se pueden utilizar para pronosticar este campo.
- **Entradas.** Seleccione el campo(s) de entrada. Se trata de una acción similar a establecer el papel 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 sobre la adecuación 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 ficha Campos en todos los nodos de modelado que usen la partición. (Si sólo hay una partición, se usará automáticamente siempre que se active la partición.) Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*. Debe tener en cuenta que al aplicar la partición seleccionada en su análisis, también debe activar la partición en la ficha 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 dividido. Se trata de una acción similar a establecer el papel del campo en *Segmentar* en un nodo Tipo. Sólo puede designar campos con un nivel de medida de Marca, Nominal, Ordinal o Continuo como campos divididos. Los campos seleccionados como campos divididos no se pueden utilizar como campos de destino, entrada, partición, frecuencia o ponderación. Si desea obtener más información, consulte el tema Generación de modelos divididos el p. 32.
- **Utilizar campo de frecuencia** Esta opción permite seleccionar un campo como ponderación de frecuencias. 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 de campo deben corresponder al número de unidades que representa cada registro. Si desea obtener más información, consulte el tema Uso de campos de frecuencia y ponderación el p. 41.

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 permite seleccionar un campo como ponderación de casos. Las ponderaciones de caso se utilizan para contabilizar las diferencias existentes en la varianza entre los niveles del campo de salida. Si desea obtener más información, consulte el tema Uso de campos de frecuencia y ponderación el p. 41.
- **Consecuentes.** Para los nodos de reglas de inducción (A priori), seleccione los campos que se deben utilizar como consecuentes en el conjunto de reglas resultante. (Se corresponde con los campos de tipo *Objetivo* o *Ambas* de un nodo Tipo).
- **Antecedentes.** Para los nodos de reglas de inducción (A priori), seleccione los campos que se deben utilizar como antecedentes en el conjunto de reglas resultante. (Se corresponde con los campos de tipo *Entrada* o *Ambas* de un nodo Tipo).

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

- Si desea obtener más información, consulte el tema Opciones de campos para el nodo Secuencia en el capítulo 12 el p. 432.
- Si desea obtener más información, consulte el tema Opciones de campos para el nodo CARMA en el capítulo 12 el p. 410.

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 se 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, imagine que tiene los datos siguientes:

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 respondieron a la promoción y dos de cada tres personas solteras no respondieron. De este modo, puntuará los nuevos registros en consecuencia:

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

También puede almacenar sus datos de entrenamiento de forma más compacta utilizando un campo de frecuencia:

Casado	Respondido	Frequency
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 pronosticará respuestas basadas únicamente en el estado civil. Si tiene a diez personas casadas en sus datos de puntuación, pronosticará *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

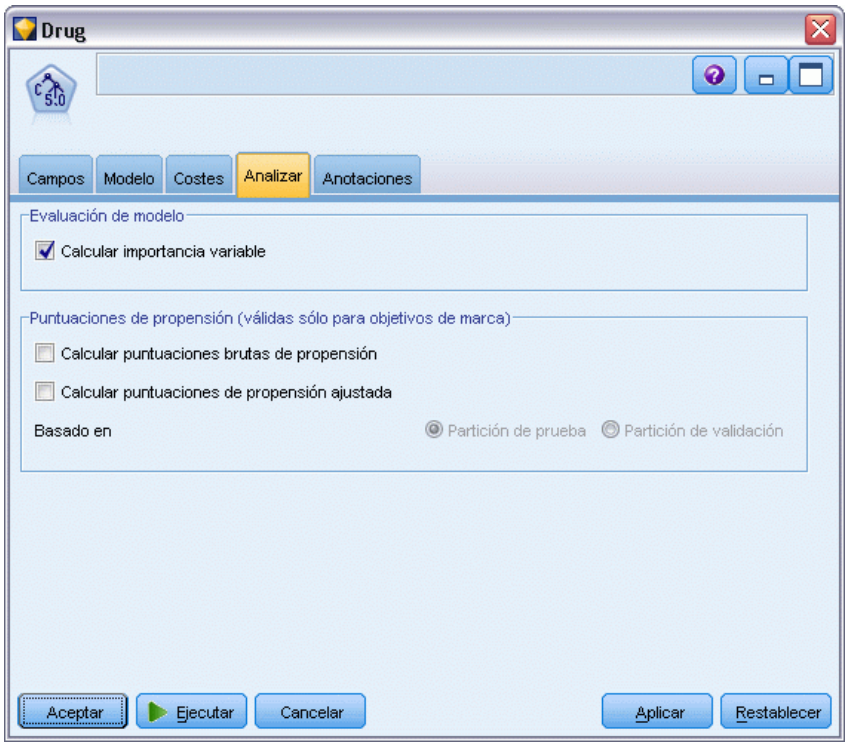
Algunos tipos de modelo admiten campos de frecuencia, algunos admiten campos de ponderación y otros admiten los dos. Sin embargo, en todos los casos en los que se aplican, solo se utilizan para la creación de modelos y no se tienen en cuenta cuando se evalúan modelos mediante los nodos Evaluación o Análisis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

- Al comparar modelos (por ejemplo, mediante gráficos 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 ordenen 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. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

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 ajustadas y brutas de propensión.

Figura 3-6
Pestaña Analizar del nodo de modelado



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 manera predeterminada. La importancia del predictor no está disponible para modelos de listas de decisiones. Si desea obtener más información, consulte el tema Importancia del predictor el p. 55.

Puntuaciones de propensión

Las puntuaciones de propensión pueden activarse en el nodo de modelado y en la ficha Configuración del nugget de modelo. Esta funcionalidad sólo está disponible cuando el objetivo seleccionado es un campo de marca. Si desea obtener más información, consulte el tema Puntuaciones de propensión el p. 45.

Calcular puntuaciones brutas de propensión. Las puntuaciones brutas de propensión 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 del pronóstico. Si el modelo predice el valor false, la propensión se calculará como (1 – P).

- Si selecciona esta opción al crear el modelo, las puntuaciones de propensión se activarán en el nugget de modelo por defecto. Sin embargo, siempre puede activar las puntuaciones brutas de propensión en el nugget de modelo independientemente de si las selecciona o no en el nodo de modelado.
- Al puntuar el modelo, se añadirán puntuaciones brutas de propensión a un campo con las letras *RP* unidas al prefijo estándar. Por ejemplo, si los pronósticos 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. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.
- 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 los pronósticos 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. Si desea obtener más información, consulte el tema Opciones de configuración del nodo Equilibrar en el capítulo 3 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*. 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. Si desea obtener más información, consulte el tema Modelos C5.0 aumentados en el capítulo 6 el p. 186.

  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. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Puntuaciones de propensión

En el caso de modelos que devuelven un pronóstico *sí* o *no*, puede solicitar puntuaciones de propensión además de los valores estándar de pronóstico y confianza. Las puntuaciones de propensión indican la verosimilitud de un resultado o respuesta específicos. Por ejemplo:

Tabla 3-1
Puntuaciones de propensión

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

Las puntuaciones de propensión solo 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. Si desea obtener más información, consulte el tema Especificación de valores para una marca en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Puntuaciones de propensión frente a puntuaciones de confianza

Las puntuaciones de propensión se diferencian de las puntuaciones de confianza, que se aplican al pronóstico actual, ya sea *sí* o *no*. Por ejemplo, en los casos en los que el pronóstico sea *no*, una confianza elevada realmente significa una mayor posibilidad de *no* responder. Las puntuaciones de propensión eluden esta limitación para permitir una comparación más fácil entre todos los registros. Por ejemplo, un pronóstico *no* con una confianza de *0,85* se traduce en una propensión bruta de *0,15* (o *1 menos 0,85*).

Tabla 3-2
Puntuaciones de confianza

Cliente	Pronóstico	Confianza
Joe Smith	Responderá	.35
Jane Smith	No responderá	.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 solo está disponible cuando el objetivo seleccionado es un campo de marca. Si desea obtener más información, consulte el tema Opciones de análisis del nodo de modelado el p. 42.
- El nodo Conjunto también puede calcular las puntuaciones de propensión, dependiendo del método de conjunto utilizado. Si desea obtener más información, consulte el tema Nodo Conjunto en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

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 brutas de propensión 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 bruta de propensión. A continuación, se entrena un modelo de red neuronal que define una función no lineal que establece correspondencias entre la propensión bruta 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 brutas de propensión 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.

Nuggets de modelo

Figura 3-7
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 IBM® SPSS® Modeler. La finalidad principal de un nugget es puntuar datos para generar pronósticos o permitir nuevos análisis 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 los pronósticos, necesitará adjuntar y ejecutar otro nodo proceso o de resultado. Si desea obtener más información, consulte el tema Uso de nugget de modelo en rutas el p. 68.

Figura 3-8
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 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. Si desea obtener más información, consulte el tema Opciones de configuración de notificación en el capítulo 12 en *Manual de usuario de IBM SPSS Modeler 15*.

Los siguientes temas proporcionan información acerca del uso de nuggets de modelo en SPSS Modeler. Si desea comprender mejor los algoritmos utilizados, consulte el *Manual de algoritmos de SPSS Modeler*, disponible en la carpeta *Documentation*, en elDVD de IBM® SPSS® Modeler.

Enlaces de modelo

Por defecto, 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. Si desea obtener más información, consulte el tema Sustitución de un modelo el p. 49.

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 3-9
Cursor de enlaces

Definición de un nuevo enlace (menú contextual)

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

Definición de un nuevo enlace (menú principal)

- Pulse en el nodo de modelado desde el que desea que empiece el enlace.

- En el menú principal, elija:
  Editar > Nodo > Definir enlace de modelo
- Pulse en el nugget en el que desea que acabe el enlace.

Eliminación de un enlace existente (menú contextual)

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

Asimismo:

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

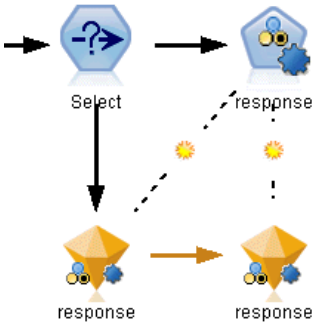
Eliminación de un enlace existente (menú principal)

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

Copiar y pegar enlaces de modelo

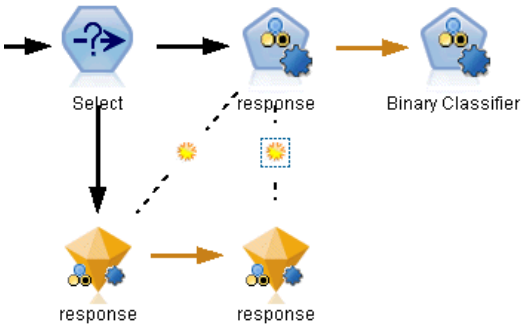
Si copia un nugget enlazado sin su nodo de modelado y lo pega en la misma ruta, éste se pegará con un enlace al nodo de modelado. El nuevo enlace tiene el mismo estado de sustitución de modelo (consulteSustitución de un modelo el p. 49) que el original:

Figura 3-10
Copiar y pegar un nugget enlazado



Si copia y pega un nugget junto con su nodo de modelado enlazado, el enlace se mantiene, al margen de que los objetos se peguen en la misma ruta o en una nueva:

Figura 3-11
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.

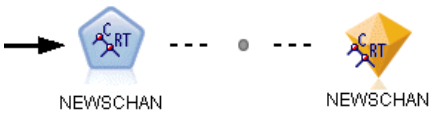
Nota: sustituir un modelo no es lo mismo que actualizarlo, que hace referencia a la actualización de un modelo en un escenario. Si desea obtener más información, consulte el tema Actualización de modelos en el capítulo 9 en *Manual de usuario de IBM SPSS Modeler 15*.

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 3-12
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 3-13
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 opciónSustituir modelo anterior del sistema. Si desea obtener más información, consulte el tema Opciones de configuración de notificación en el capítulo 12 en *Manual de usuario de IBM SPSS Modeler 15*.).

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 un procesamiento.

Cambio de la configuración de sustitución de modelo

Para cambiar la configuración de sustitución de modelo.

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

Nota: el ajuste de sustitución de modelo de un enlace de modelo reemplaza al ajuste de la pestaña Notificaciones del cuadro de diálogo Opciones de usuario (Herramientas > Opciones > Opciones de usuario).

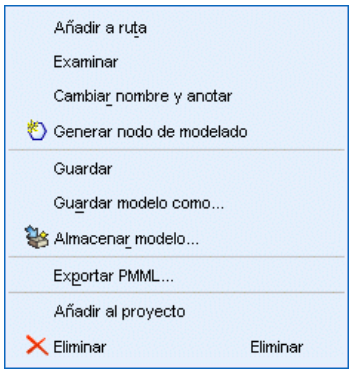
La paleta de modelos

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

Figura 3-14
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:

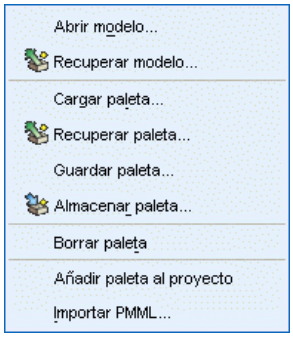
Figura 3-15
Menú contextual de nugget de modelo



- **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 de modelo (.gm) externo.

- **Almacenar modelo.** Guarda el nugget de modelo en IBM® SPSS® Collaboration and Deployment Services Repository. Si desea obtener más información, consulte el tema Acerca de IBM SPSS Collaboration and Deployment Services Repository en el capítulo 9 en *Manual de usuario de IBM SPSS Modeler 15*.
- **Exportar PMML.** Exporta el nugget de modelo como lenguaje de marcas 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. *Nota*: para utilizar esta función, se necesita una licencia de IBM® SPSS® Modeler Server. Si desea obtener más información, consulte el tema Opciones de configuración de exportación de PMML en el capítulo 12 en *Manual de usuario de IBM SPSS Modeler 15*.
- **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 por defecto. (Si desea obtener más información sobre cómo cambiar la fase del proyecto por defecto, consulte Establecimiento de la fase del proyecto por defecto.)
- **Eliminar.** Elimina el nugget de modelo de la paleta.

Figura 3-16
Menú contextual de la paleta de modelos



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 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 por defecto.
- **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. Si desea obtener más información, consulte el tema Cómo importar y exportar modelos como PMML en *Manual de usuario de IBM SPSS Modeler 15*.

Exploración de nugget 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 Filtro 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 conglomerados. En algunos tipos de modelos (los modelos basados en árboles y los modelos de conglomerados) 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.
- **Encabezado y pie de página.** Permite editar el encabezado 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 marcas para modelos predictivos (PMML), que se puede utilizar con otro software compatible con PMML. Si desea obtener más información, consulte el tema Cómo importar y exportar modelos como PMML en *Manual de usuario de IBM SPSS Modeler 15*.*Nota*: para utilizar esta función, se necesita una licencia de IBM® SPSS® Modeler Server. Si desea obtener más información, consulte el

tema Opciones de configuración de exportación de PMML en el capítulo 12 en *Manual de usuario de IBM SPSS Modeler 15*.

- **Exportar SQL.** Exporta el modelo como lenguaje de consulta estructurado (SQL), que puede modificarse y utilizarse con otras bases de datos.

  Nota: la opción Exportar SQL solo está disponible en los siguientes modelos: C5, C&RT, CHAID, QUEST, Regresión lineal, Regresión logística, Red neuronal, PCA/Factorial y modelos de listas de decisiones. Si desea obtener más información, consulte el tema Nodos que admiten la generación de SQL en el capítulo 6 en *Guía de rendimiento y administración de IBM SPSS Modeler Server 15*.

- **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 del modelo tenga lugar dentro de la base de datos. Si desea obtener más información, consulte el tema Publicación de modelos para un adaptador de puntuación el p. 74.

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 por defecto de filas visualizadas es 10; sin embargo, puede cambiarlo en las propiedades de la ruta. Si desea obtener más información, consulte el tema Configuración de opciones generales de las rutas en el capítulo 5 en *Manual de usuario de IBM SPSS Modeler 15*.

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 por defecto. (Si desea obtener más información sobre cómo cambiar la fase del proyecto por defecto, consulte Establecimiento de la fase del proyecto por defecto.)

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 ha ejecutado un nodo Análisis conectado a este nodo de modelado, la información de dicho análisis también se mostrará en esta sección. Si desea obtener más información, consulte el tema Nodo Análisis en el capítulo 6 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Campos. Enumera los campos utilizados como objetivo y entradas en la generación del modelo. En modelos de división, enumera los campos que determinan las divisiones.

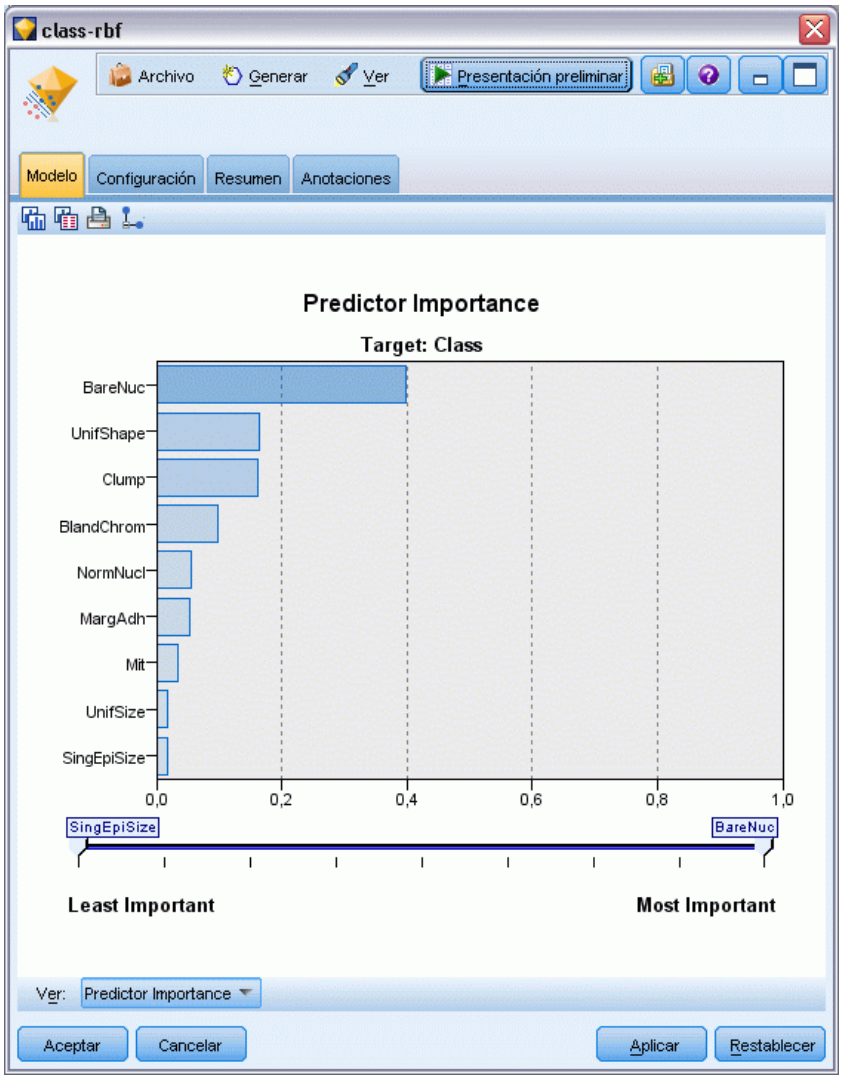
Opciones / Configuración de creación. Contiene información sobre la configuración que se utiliza en la generación del modelo.

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.

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.

Figura 3-17
Gráfico Importancia del predictor



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. Si desea obtener más información, consulte el tema Opciones de análisis del nodo de modelado el p. 42.Si desea obtener más información sobre los modelos KNN, consulte Vecinos el p. 500.

Nota: No se admite la importancia del predictor para modelos divididos. Al crear modelos divididos, se ignora la configuración de importancia del predictor. Si desea obtener más información, consulte el tema Generación de modelos divididos el p. 32.

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; por defecto, está desactivado para estos modelos. Si está utilizando un conjunto de datos con un gran número de predictores, un cribado inicial mediante el nodo Selección de características puede dar resultados más rápidos (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. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.
- En el caso de los modelos SLRM, la importancia del predictor está disponible, pero se calcula mediante el algoritmo SLRM. Si desea obtener más información, consulte el tema Nugget de modelo SLRM en el capítulo 14 el p. 481.
- 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 Filtro basado en la información del gráfico Importancia de predictor. Si desea obtener más información, consulte el tema Filtrado de variables basado en su importancia el p. 57.

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 analizar 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 análisis 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.

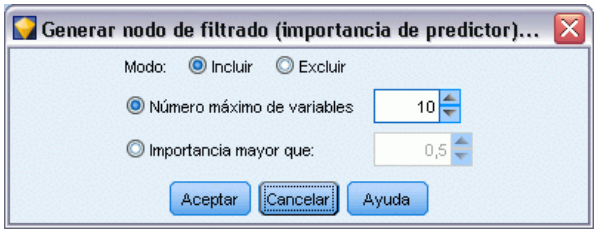
Filtrado de variables basado en su importancia

También puede generar el nodo Filtro 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 Filtro (Importancia de predictor)

O
> Selección de campo (Importancia de predictor)

Figura 3-18
Filtrado de predictores basado en su importancia



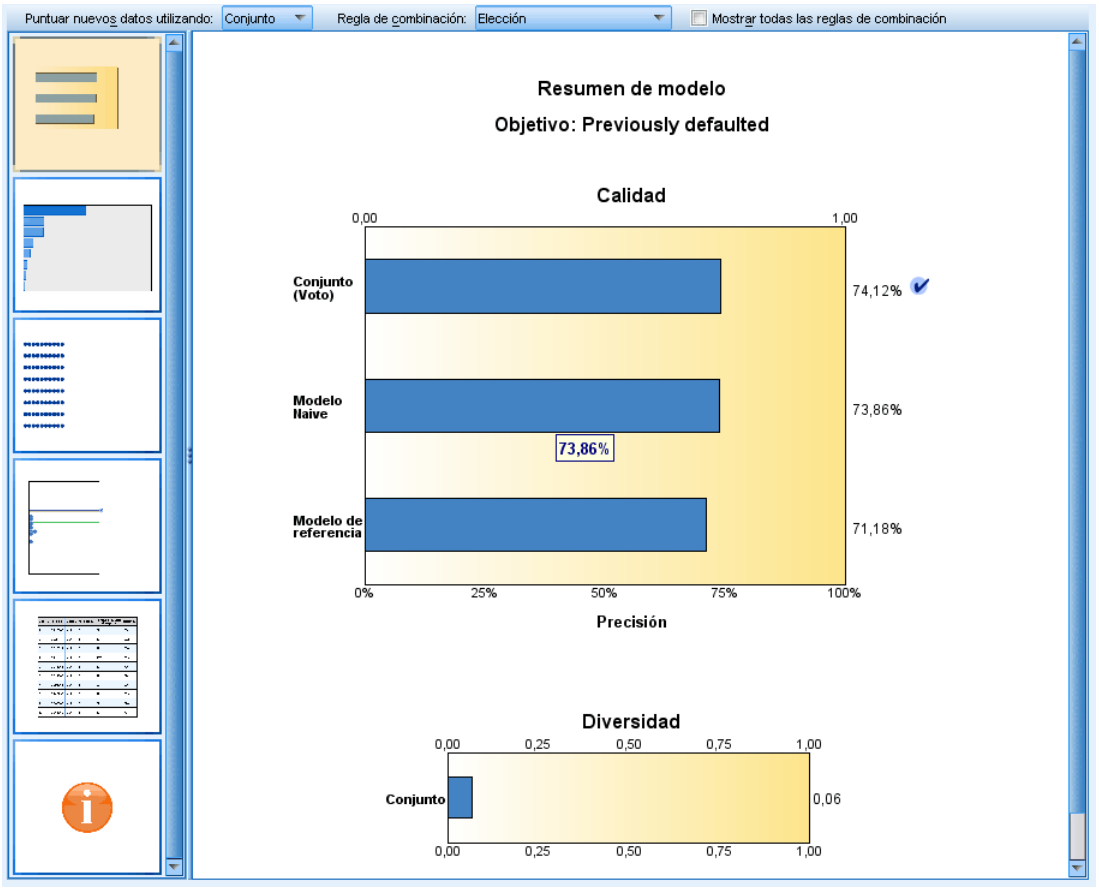
Número especificado de variables. Incluye o excluye los predictores más importantes hasta el número especificado.

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

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.

Figura 3-19
Vista Resumen del modelo



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 pronosticados a partir de los modelos básicos para calcular el valor de puntuación del conjunto.

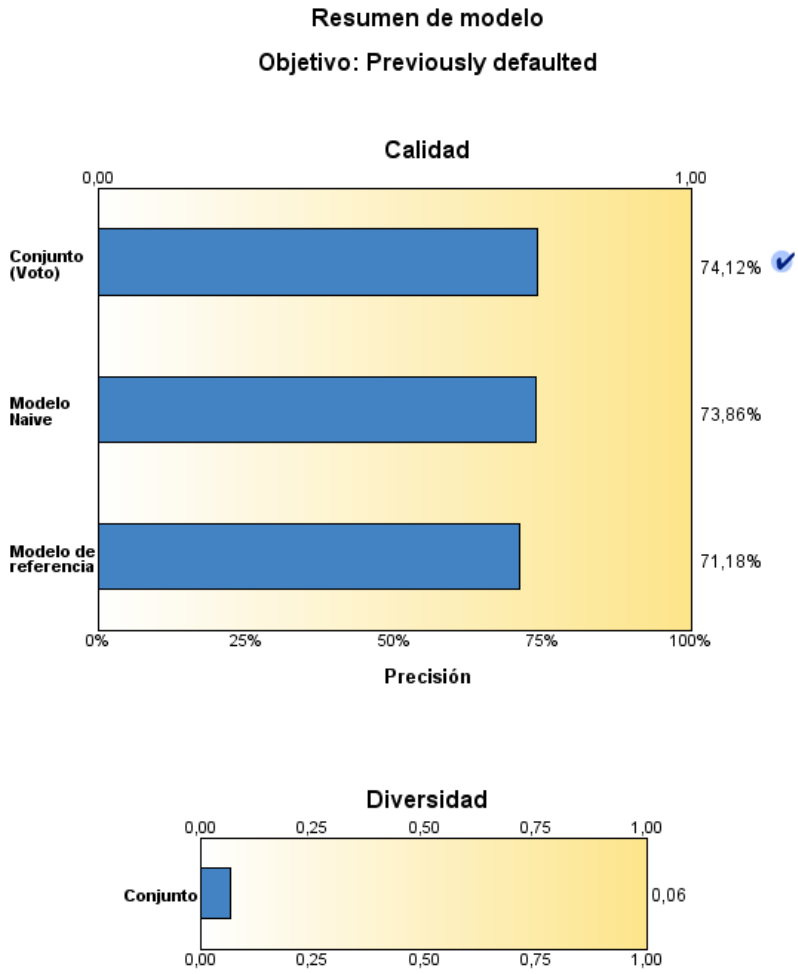
- Los valores predichos de conjuntos de destinos **categóricos** pueden combinarse mediante votación, mayor probabilidad o mayor probabilidad media. La **votación** selecciona la categoría con mayor probabilidad más común en los distintos modelos básicos. La **mayor probabilidad** selecciona la categoría que consigue la mayor probabilidad más alta en los distintos modelos básicos. La **mayor probabilidad media** selecciona la categoría con el valor más alto cuando se realiza una media de las probabilidades de categoría en los distintos 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 construcción del modelo. 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

Figura 3-20
Vista 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).

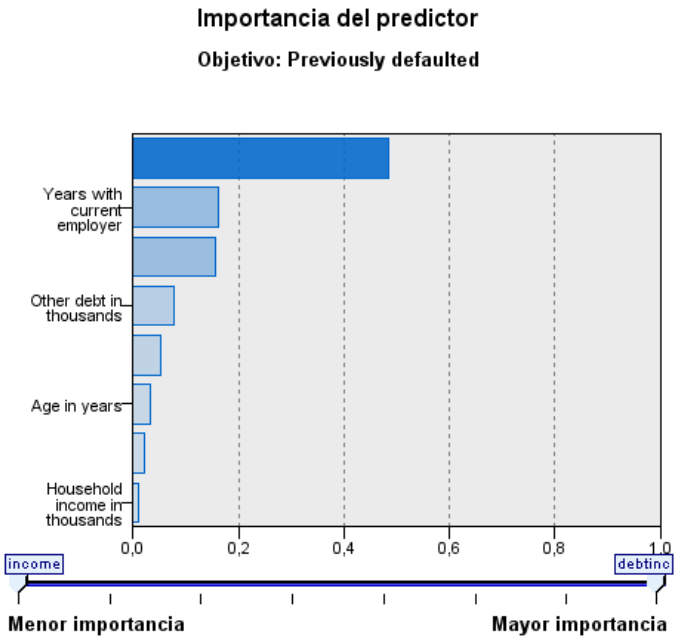
Para empaquetar conjuntos, el modelo de referencia es un modelo estándar construido en la partición de entrenamiento al completo. Para los conjuntos potenciados, el modelo de referencia es el primer modelo de componente.

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

Figura 3-21
Vista de importancia del predictor

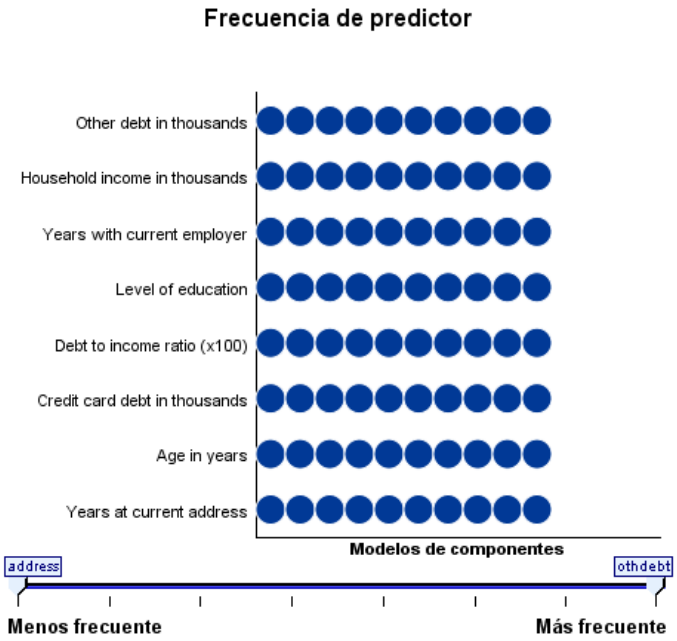


Es normal centrar los esfuerzos de modelado en los campos predictores más importantes y valorar la omisión de aquellos con menor relevancia. El gráfico de importancia de los predictores le ayuda a hacerlo indicando la importancia relativa de cada predictor en la estimación del modelo. Como los valores 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.

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

Figura 3-22
Vista Frecuencia de predictor

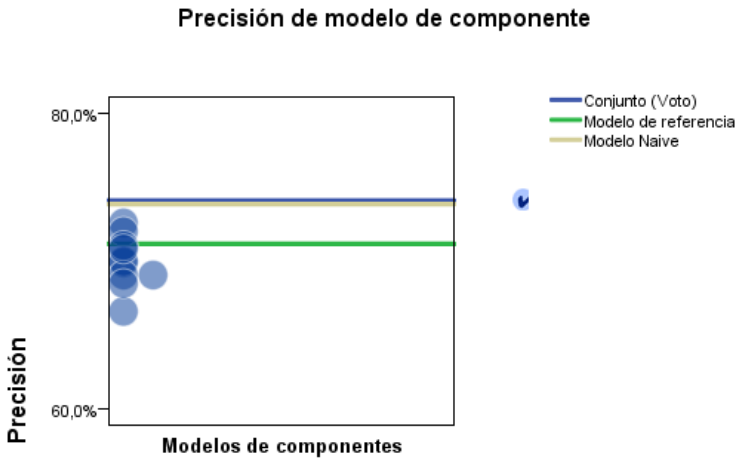


El conjunto de predictores puede variar en distintos modelos de componente por la elección del método de modelado o la selección de predictores. El gráfico Frecuencia de predictor es un gráfico de puntos que muestra la distribución de predictores entre los modelos de componente del conjunto. Cada punto 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

Figura 3-23
Vista Precisión de modelo de componente



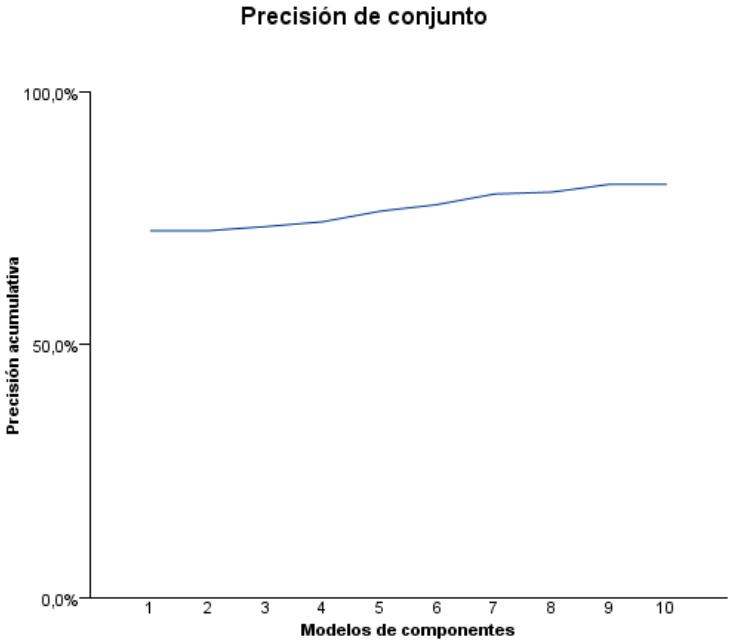
El gráfico es un gráfico de puntos de precisión predictiva para los modelos de componente. Cada punto representa uno o más modelos de componente con el nivel de precisión representado en el eje y. Pase el ratón sobre cualquier punto para obtener información sobre el modelo de componente individual correspondiente.

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.

Figura 3-24
Vista de precisión de conjuntos, conjunto potenciado



Detalles de modelo de componente

Figura 3-25
Vista Detalles de modelo de componente

Detalles de modelo de componente

Modelo	Precisión	Método	Predictores	Tamaño de modelo (Sinapsis)	Registros
1	72,6%		8	77	700
2	72,0%		8	107	692
3	69,9%		8	92	708
4	70,0%		8	107	685
5	71,1%		8	107	706
6	69,1%		8	92	690
7	69,1%		8	92	696
8	70,8%		8	122	703
9	66,6%		8	62	726
10	68,5%		8	107	701

La tabla muestra información sobre los modelos de componente, enumerados por fila. Por defecto, 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.

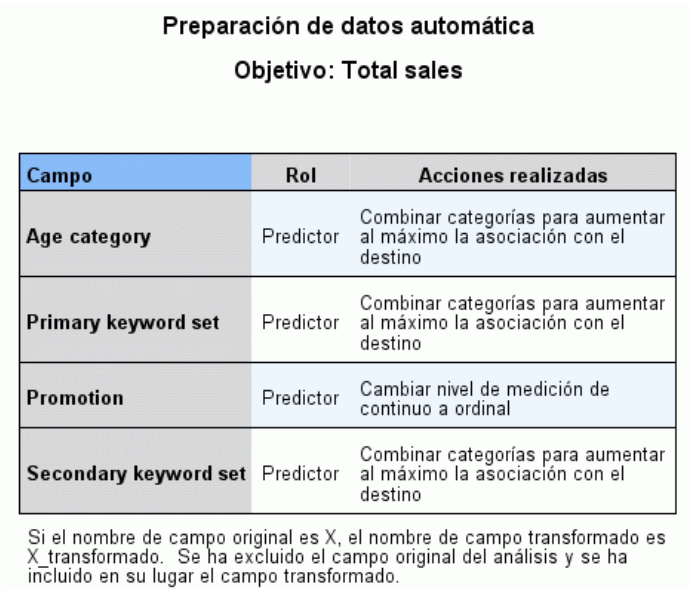
Predictores. Número de predictores utilizados en el modelo de componente.

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

Registros. El número ponderado de registros de entrada en la muestra de entrenamiento.

Preparación automática de datos

Figura 3-26
Vista Preparación de datos automática



Preparación de datos automática

Objetivo: Total sales

Campo	Rol	Acciones realizadas
Age category	Predictor	Combinar categorías para aumentar al máximo la asociación con el destino
Primary keyword set	Predictor	Combinar categorías para aumentar al máximo la asociación con el destino
Promotion	Predictor	Cambiar nivel de medición de continuo a ordinal
Secondary keyword set	Predictor	Combinar categorías para aumentar al máximo la asociación con el destino

Si el nombre de campo original es X, el nombre de campo transformado es X_transformado. Se ha excluido el campo original del análisis y se ha incluido en su lugar el campo transformado.

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 papel 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 Recortar valores atípicos, si se muestra, indica que los valores de predictores continuos que caen más allá de un valor de corte (3 desviaciones típicas de la media) se han ajustado al valor de corte.

Nuggets de modelo de modelos de división

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

Un nugget de modelo de división contiene:

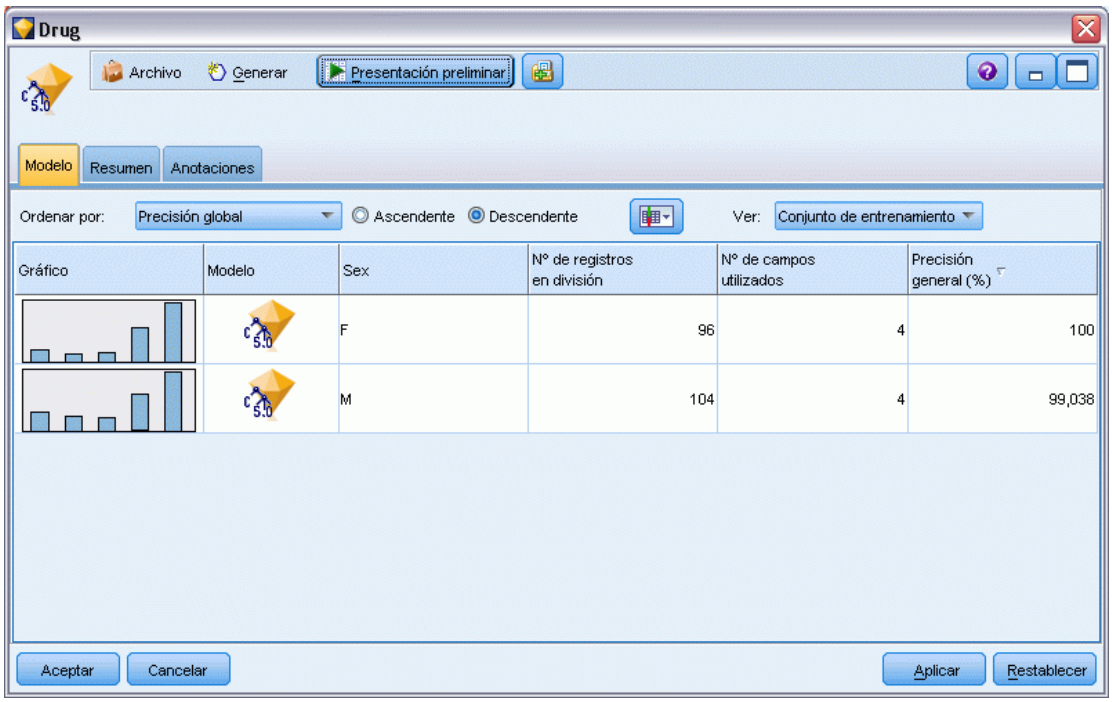
- una lista de todos los modelos de división creados, junto con un conjunto de estadísticas de cada modelo
- información acerca del modelo total

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

Visor del modelo de división

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

Figura 3-27
Visor del modelo de división



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 un encabezado de columna para ordenar la lista por esa columna. El valor por defecto 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. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

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 de división en función del número de campos de entrada utilizados.

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

Figura 3-28
Segmentar visor de modelo

Dividir grupos

ed	Precisión	Tamaño de modelo (Sinapsis)	Registros
Did not complete high school	68,3%	62	372
High school degree	66,8%	42	198
Some college	72,3%	42	87
College degree	59,2%	12	38
Post-undergraduate degree	.	.	.

No se han podido crear los modelos para uno o más grupos de división.

Segmentar. El encabezado de columna muestra los campos usados para crear divisiones y las celdas son los valores de segmentación. Pulse dos veces en cualquier segmentación para abrir un visor de modelo 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 en el árbol; en los modelos lineales, es el número de coeficientes; en las redes neuronales, es el número de sinapsis.

Registros. El número ponderado de registros de entrada en la muestra de entrenamiento.

Uso de nugget de modelo en rutas

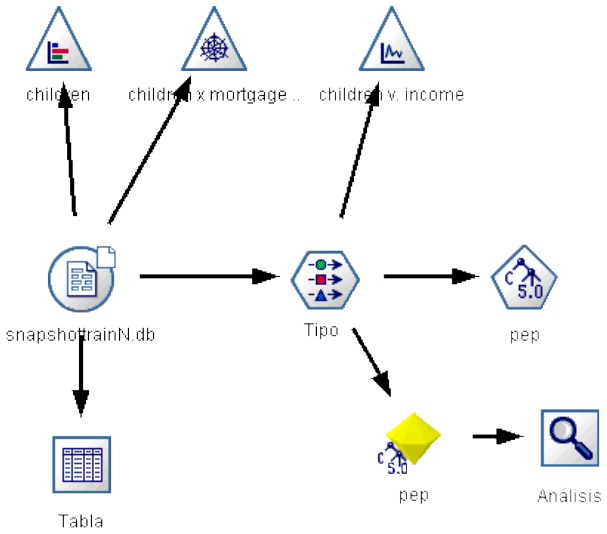
Los nuggets de modelo se colocan en rutas para permitir puntuar nuevos datos y generar nuevos nodos. La **puntuación** de datos le permite utilizar la información obtenida a partir de la generación de modelos para crear pronósticos para nuevos registros. Para ver los resultados de la puntuación, necesitará 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 del pronóstico, como los valores de confianza o las distancias desde los centros de los conglomerados. 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 una selección de campos de entrada le permiten generar nodos Filtro que solo permiten el paso de aquellos campos de entrada que el modelo haya identificado como importantes.

Uso de un nugget de modelo para puntuar datos

- Conecte el nugget de modelo a una ruta u origen de datos que pasará datos al nugget.

Figura 3-29
Uso de un nugget de modelo para puntuar



- Añada o conecte uno o más nodos de procesamiento o de resultados (como un nodo Tabla o Análisis) al nugget de modelo.
- 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 los datos. Para almacenar datos basados en un , utilice el nodo de reglas sin refinar para generar un nugget de conjunto de reglas y utilice este nugget para la puntuación. Si desea obtener más información, consulte el tema Generación de un conjunto de reglas desde un nugget de modelo de asociación en el capítulo 12 el p. 425.

Uso de un nugget de modelo para generar nodos de procesamiento

- En la paleta, examine el modelo o, en el lienzo de ruta, edite el modelo.
- 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. Si desea obtener más información, consulte el tema Opciones del modelo de la lista de decisiones en el capítulo 9 el p. 224.
- En los modelos de serie temporal, está activada la opción Continuar con la estimación utilizando modelo(s) existente, que permite volver a generar el modelo anterior con los datos actuales. Si desea obtener más información, consulte el tema Opciones del modelo de serie temporal en el capítulo 13 el p. 457.

Cómo importar y exportar modelos como PMML

PMML, o lenguaje de marcas 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 puede importar y exportar PMML, permitiendo compartir modelos con otras aplicaciones que admitan este formato, como IBM® SPSS® Statistics.

Nota: Para exportar PMML, se necesita una licencia de IBM® SPSS® Modeler Server.

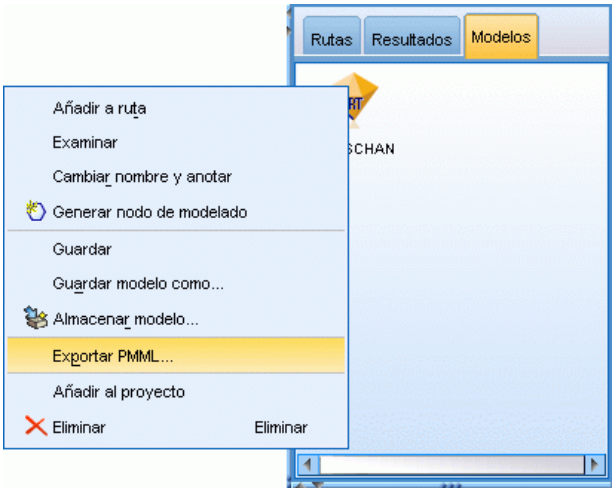
Si desea obtener más información sobre PMML, consulte el sitio Web del grupo de minería de datos (*http://www.dmg.org*).

Para exportar un modelo

La mayoría de tipos de modelos generados por SPSS Modeler admite la exportación PMML. Si desea obtener más información, consulte el tema Tipos de modelos que admiten PMML en *Manual de usuario de IBM SPSS Modeler 15*.

- 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.)
- En el menú, pulse Exportar PMML.

Figura 3-30
Exportación de un modelo en formato PMML



- En el cuadro de diálogo Exportar (o Guardar), especifique un directorio objetivo y un nombre único 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.

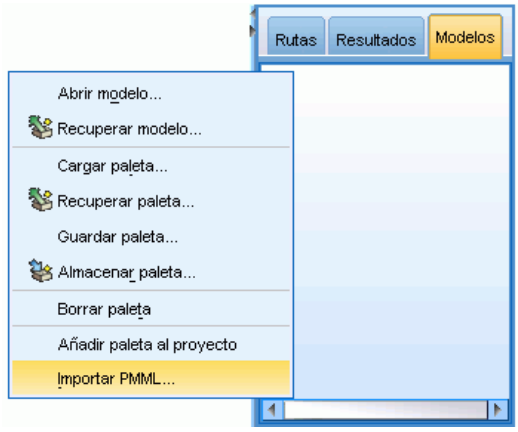
  Si desea obtener más información, consulte el tema Opciones de configuración de exportación de PMML en el capítulo 12 en *Manual de usuario de IBM SPSS Modeler 15*.

Para importar un modelo guardado como PMML

Los modelos exportados como PMML desde SPSS Modeler o cualquier otra aplicación se pueden importar a la paleta de modelos. Si desea obtener más información, consulte el tema Tipos de modelos que admiten PMML en *Manual de usuario de IBM SPSS Modeler 15*.

- En la paleta de modelos, pulse con el botón derecho en la paleta y seleccione Importar PMML del menú.

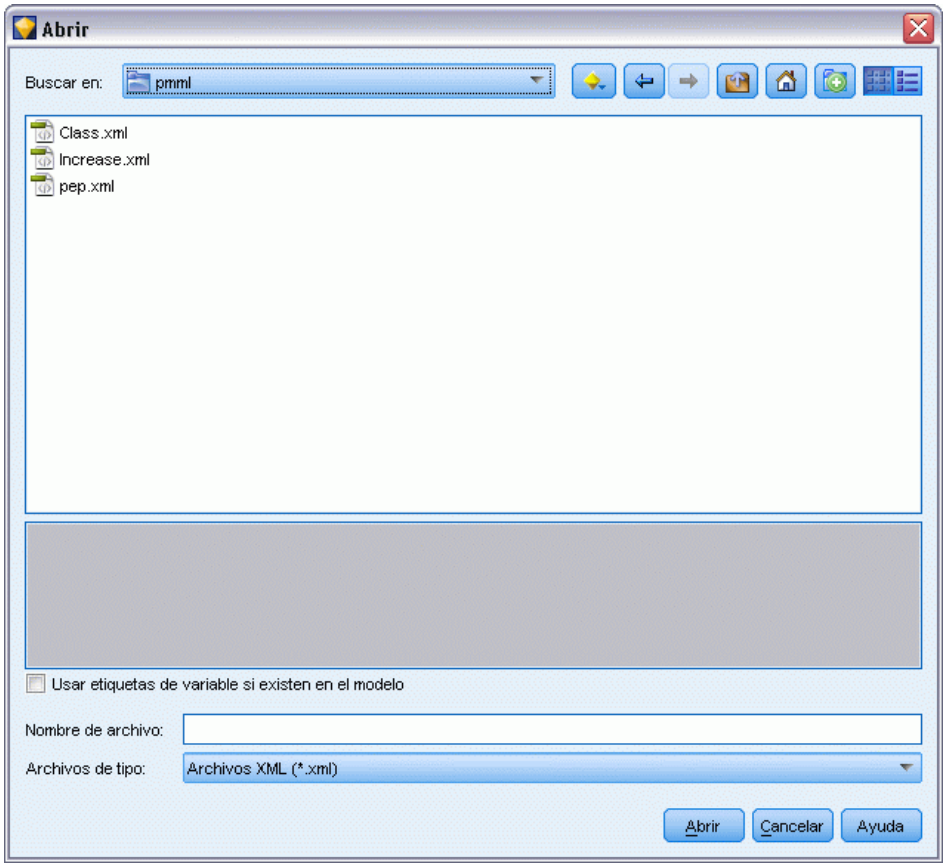
Figura 3-31
Importación de un modelo en formato PMML



- Seleccione el archivo que desea importar y especifique las opciones de las etiquetas de valores y variables como desee.

- Pulse en Abrir.

Figura 3-32
Selección del archivo XML para un modelo guardado utilizando PMML



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 modelos que admiten PMML

Exportación de PMML

SPSS Modeler modelos ALSCAL asimétricos. 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
- Red bayesiana
- A priori
- Carma
- K-medias
- Kohonen
- Dos fases
- KNN
- Statistics Modelo

El siguiente modelo creado en SPSS Modeler puede exportarse como PMML 3.2:

- Lista de decisiones

Modelos nativos de bases de datos. Para modelos generados mediante algoritmos nativos de bases de datos, la exportación PMML está disponible solamente para modelos de IBM InfoSphere Warehouse. Los modelos creados mediante Analysis Services desde Microsoft u Oracle Data Miner no se pueden exportar. Tenga en cuenta también que los modelos IBM exportados como PMML no se pueden volver a importar a SPSS Modeler. Si desea obtener más información,

consulte el tema Conceptos básicos del modelado de la base de datos en el capítulo 2 en *Manual de minería interna de la base de datos de IBM SPSS Modeler 15*.

Importación de PMML

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 SPSS Modeler, así como cualquier modelo o transformación PMML generado mediante 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 y de secuencia no pueden importarse.
- Es posible que no pueda navegar por los modelos de PMML después de importar a SPSS Modeler aunque se puedan utilizar para la puntuación. (Tenga en cuenta que esto incluye los modelos que se exportaron de SPSS Modeler para comenzar. Para evitar esta limitación, exporte el modelo como un archivo del modelo generado [**.gm*] en lugar de como PMML.)
- Los modelos de IBM InfoSphere Warehouse exportados como PMML no se pueden importar.
- 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.

Publicación de 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 del modelo 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 en un adaptador de puntuación

▶ Pulse dos veces en el nugget de modelo para abrirlo.

▶ En el menú de nugget de modelo, seleccione:
File > Publicar en el adaptador de puntuación del servidor.

▶ 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 las bases de datos de 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 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 aparecen como "diamantes en bruto" en la paleta de modelos generados.

Figura 3-33
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 pronosticar los valores de un campo objetivo único. Si desea obtener más información, consulte el tema Generación de un conjunto de reglas desde un nugget de modelo de asociación en el capítulo 12 el p. 425.

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? Si desea obtener más información, consulte el tema Nodo Selección de características el p. 76.

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. Si desea obtener más información, consulte el tema Nodo Detección de anomalías el p. 84.

Tenga en cuenta de que la detección de anomalías identifica registros o casos extraños a través del análisis de conglomerados 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 pronosticar. 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 orden de campos por rangos. 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 pronosticar 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:

- **Filtrado.** 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.
- **Asignación de rangos.** Ordena las entradas restantes y les asigna un rango en función de la importancia.
- **Selección.** Identifica el subconjunto de características que se utilizará en modelos posteriores, por ejemplo, conservando solamente 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.

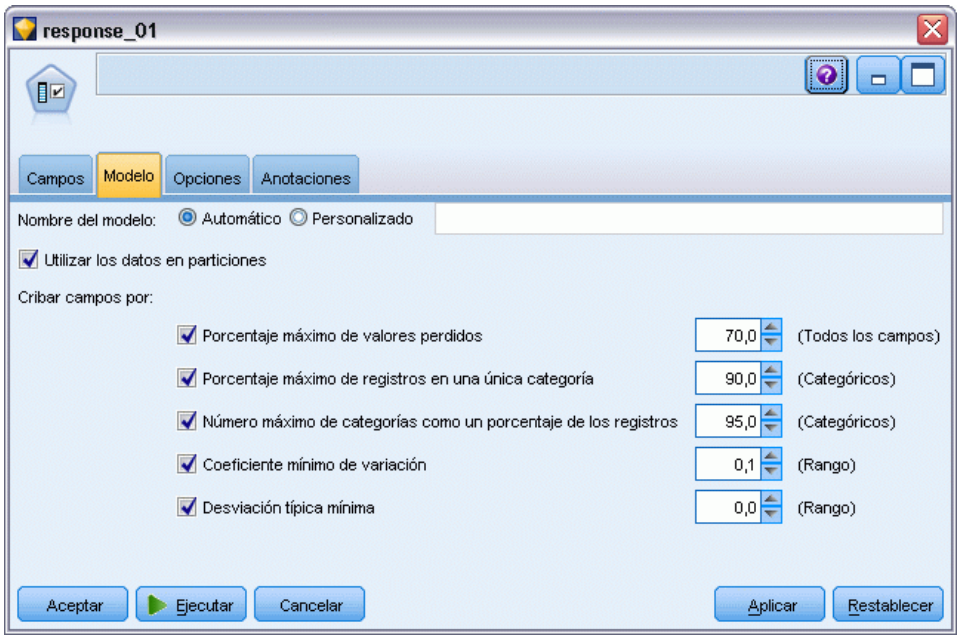
Ejemplo. Una compañía telefónica tiene un almacén de datos con información sobre las respuestas de 5.000 clientes en relación con una promoción especial. Los datos incluyen un gran número de campos que contienen los estadísticos del uso del teléfono, las edades de los clientes, el puesto de trabajo y 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. Si desea obtener más información, consulte el tema Predictores de filtrado (Selección de características) en el capítulo 9 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

Requisitos. Un único campo de objetivo (uno con su papel 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 del modelo estándar, además de configuraciones que permiten ajustar los criterios necesarios para cribar campos de entrada.

Figura 4-1
Pestaña Modelo de la selección de características



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.

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 medición es el índice de la desviación típica 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 típica mínima.** Criba campos con desviación típica menor o igual que el mínimo especificado. Esta opción sólo se aplica a campos continuos (rango numérico).

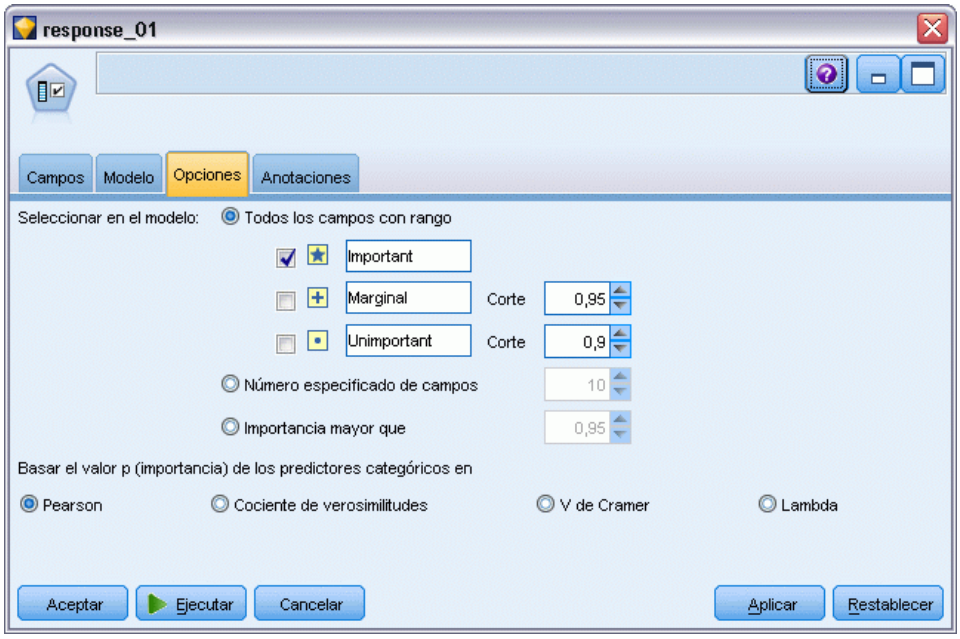
Registros con datos perdidos. Los registros o casos que tienen valores perdidos en el campo objetivo, o bien valores perdidos en todas las entradas, se excluyen automáticamente de todos los cálculos utilizados en las ordenaciones por rangos.

Opciones de la selección de características

La ficha Opciones permite especificar la configuración por defecto 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 por defecto permite aplicar el nugget de modelo sin más cambios, lo que puede ser especialmente útil con fines de creación de procesamientos.

Si desea obtener más información, consulte el tema Resultados del modelo de selección de características el p. 81.

Figura 4-2
Pestaña Opciones de la selección de características



Se encuentran disponibles las siguientes opciones:

Todos los campos con rango. Selecciona los campos según el orden por rangos como *important*, *marginal*, or *unimportant*. Se puede editar la etiqueta de cada rango, 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 ordenación de la importancia por rangos

Todos categóricos. Cuando todas las entradas y el objetivo son categóricos, la importancia se puede ordenar por rangos 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 del cociente de verosimilitudes.** 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 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 pronostica 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 (si bien no todos) son categóricos y el objetivo también lo es, la importancia se puede ordenar por rangos según los chi-cuadrado de Pearson o del cociente de verosimilitudes. (La *V* de Cramer y lambda no estarán disponibles a menos que todas las entradas sean categóricas.)

Categóricos frente a continuos. Cuando se ordena por rangos una entrada categórica al compararlo 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 ordena por rangos una entrada continua al compararlo 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 ordenación por rangos. Si desea obtener más información, consulte el tema Nodo Selección de características el p. 76.

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 por defecto) 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.

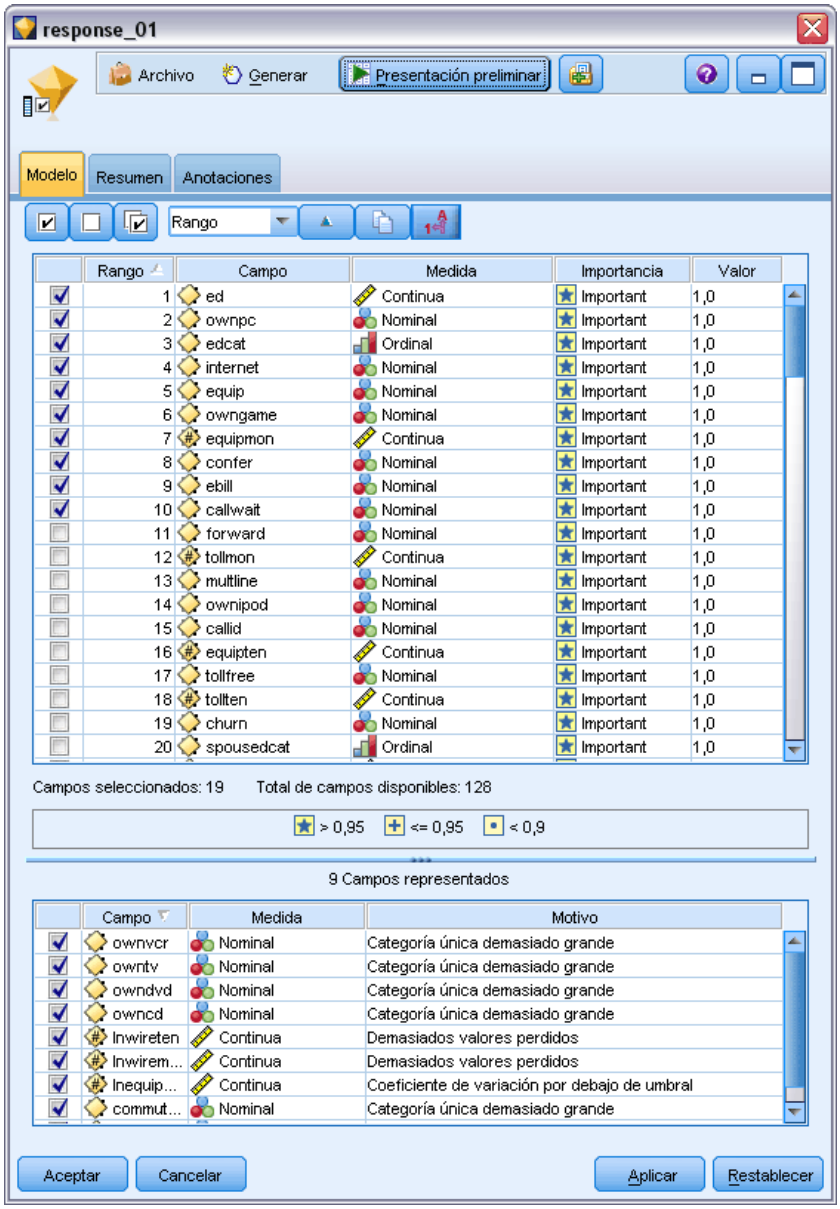
El filtrado se basa únicamente en el nombre del campo; por ejemplo, si selecciona *edad* e *ingresos*, se conservará cualquier campo que coincida con uno de estos nombres. El modelo no actualiza la asignación de rangos a los campos a partir de datos nuevos, sino que simplemente filtra los campos en función 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 se conservan los campos marcados. El resto queda descartado. Las selecciones por defecto 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 ordenaciones por rangos de acuerdo 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. Si desea obtener más información, consulte el tema Configuración del modelo de selección de características el p. 77.

Figura 4-3
Resultados del modelo de selección de características



- 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 anular la selección de cualquier campo 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. Si desea obtener más información, consulte el tema Selección de campos por importancia el p. 83.

- Los valores de umbral para ordenar las entradas por rangos como importantes, marginales o sin importancia se muestran en la leyenda bajo la tabla. Estos valores se especifican en el nodo de modelado. Si desea obtener más información, consulte el tema Opciones de la selección de características el p. 79.

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.

Figura 4-4
Cuadro de diálogo Seleccionar campos

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

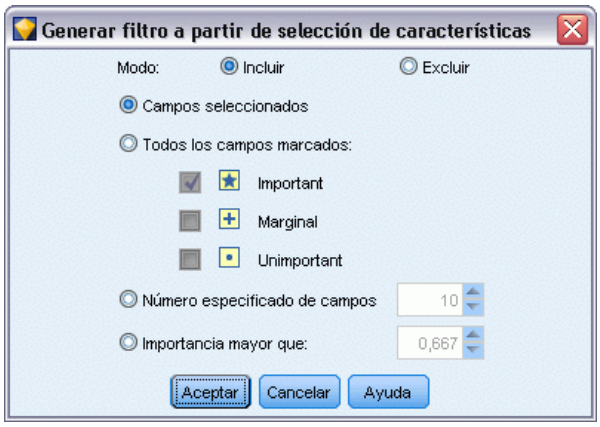
Número especificado de campos. Le permite seleccionar los *n* campos principales en función de su importancia.

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

Con los resultados de un modelo de selección de características se pueden generar uno o varios nodos Filtro que incluyen o excluyen subconjuntos de campos en función de 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.

Figura 4-5
Generación de un nodo Filtro



Incluir/Excluir. Se puede elegir entre incluir campos o excluirlos; por ejemplo, incluir los 10 campos principales o excluir todos los campos que se hayan marcado 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.

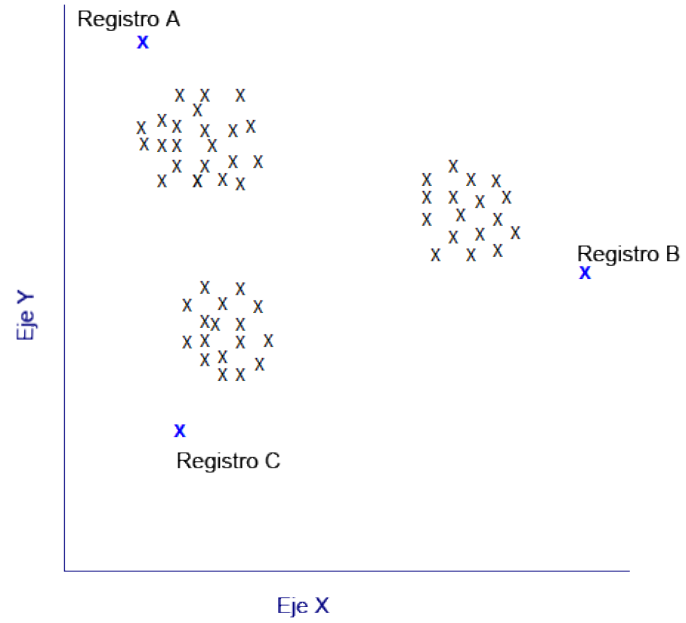
Importancia mayor que. Selecciona todos los campos con una importancia superior al umbral especificado.

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 conglomerados 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 conglomerados diferentes y marcar aquellos que se sitúen lejos del centro de cualquier conglomerado.

Figura 4-6
Uso de conglomerados para identificar anomalías potenciales



Se asigna un índice de anomalías a cada registro, que es el cociente del índice de desviación del grupo sobre su media sobre el conglomerado 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 defecto 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 conglomerados 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 pronosticar. 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 orden de campos por rangos. 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.)

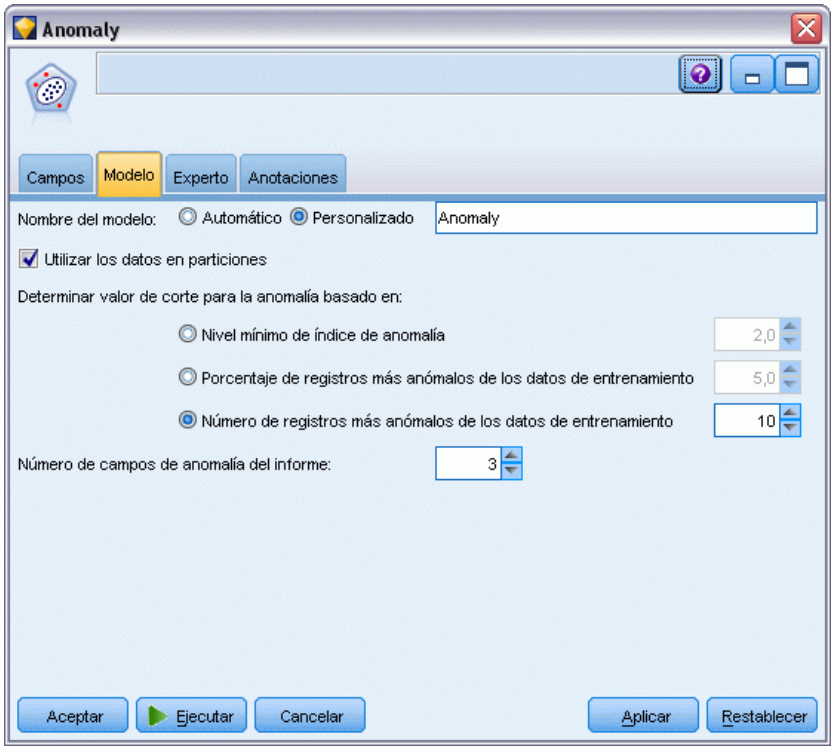
Ejemplo. Al cribar subvenciones para el desarrollo agrícola para posibles casos de fraude, se puede utilizar la detección de anomalías para descubrir las desviaciones de la norma, resaltando aquellos registros que sean anómalos y dignos de una investigación más detallada. En particular, le interesan 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 papel definido como Entrada mediante un nodo de origen o un nodo Tipo. Se omitirán los campos objetivo (con el papel 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

Figura 4-7
Pestaña Modelo de la detección de anomalías



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.

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.

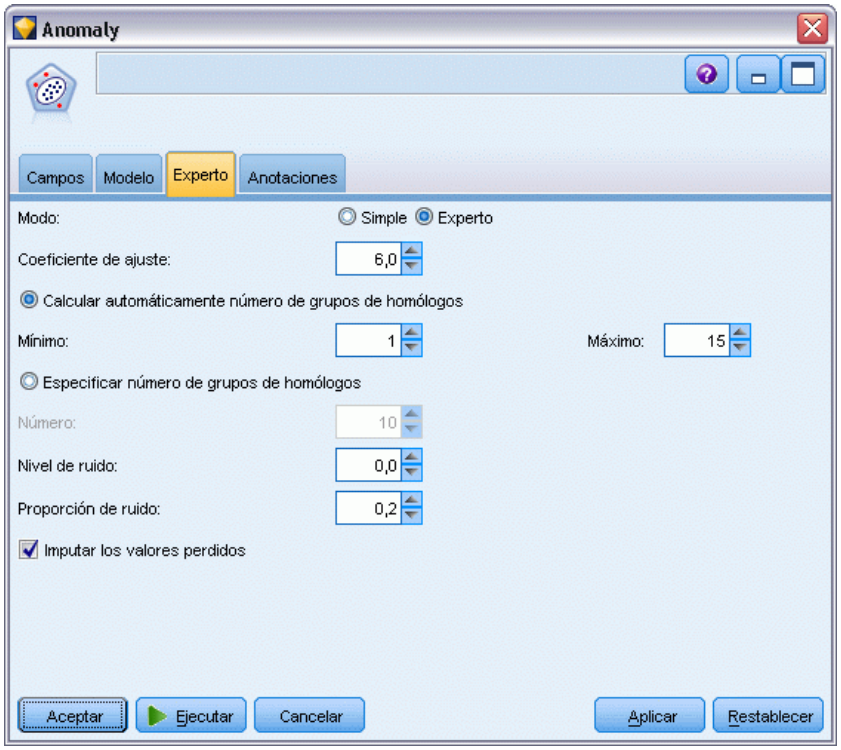
Nota: independientemente de cómo se determine el valor de corte, esto no incidirá en el valor de índice de anomalía subyacente mostrado para cada registro. Simplemente define el umbral para marcar los registros como anómalos al calcular o puntuar el modelo. Si posteriormente desea examinar un número 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 conglomerado 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.

Figura 4-8
Pestaña Experto de la detección de anomalías



Coeficiente de ajuste. Valor utilizado para equilibrar el peso relativo dado 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 busque en un rango más amplio de posibles soluciones, si bien esto supone un tiempo de procesamiento más prolongado.

Especificar número de grupos de homólogos. Si sabe el número de conglomerados 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 conglomerado de dos fases. En la primera fase, se utiliza un árbol de características de conglomerados (CF) para reducir los datos de un gran número de registros individuales a un número de conglomerados 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 filiales. En la segunda fase, comienza el conglomerado jerárquico 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

conglomerado de ruido de la primera lectura de datos se asignan a los conglomerados 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. Tras la conglomeración final, los valores que no se puedan asignar a un conglomerado se considerarán como valores atípicos. El conglomerado de los atípicos tendrá asignado el número de identificación –1 y no se incluirá en el recuento de número de conglomerados; es decir, si especifica *n* conglomerados y tratamiento de ruido, el algoritmo dará como resultado *n* conglomerados y un conglomerado 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 conglomerado 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 conglomerado.

- **Proporción de ruido.** Especifica la parte de memoria asignada al componente que debe 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 conglomerado 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.

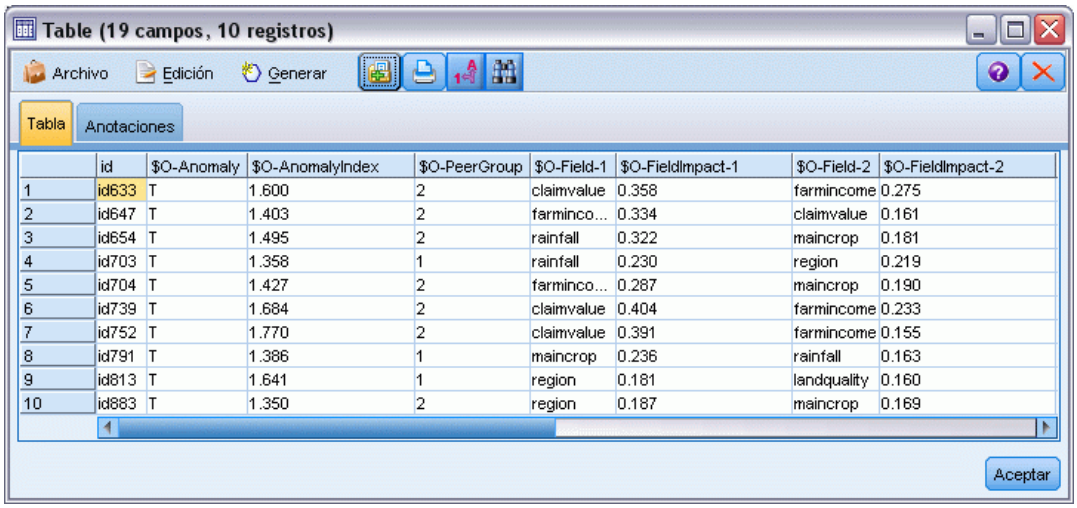
Cuando se ejecuta una ruta que contiene un nugget de modelo de detección de anomalías, se añade un número de campos nuevos a dicha ruta en función de las selecciones realizadas en la pestaña Configuración del nugget de modelo. Si desea obtener más información, consulte el tema Configuración del modelo de detección de anomalías el p. 92.Los nombres de los campos nuevos se basan en el nombre del modelo, precedido de *$O*, tal como se resume en la tabla siguiente:

$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 conglomerado.
$O-FieldImpact-n	Índice de desviación variable del campo. Este valor mide la desviación de la norma de campo relativa al conglomerado 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.

Figura 4-9
Resultados de puntuación con registros no anómalos suprimidos

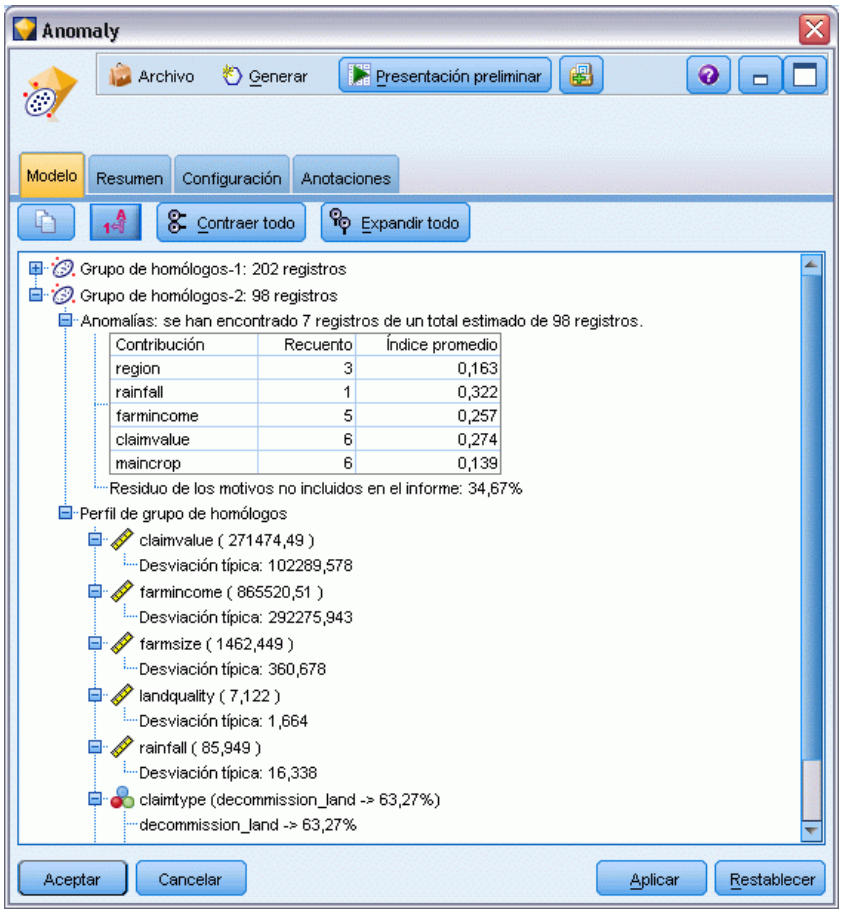


	id	$O-Anomaly	$O-AnomalyIndex	$O-PeerGroup	$O-Field-1	$O-FieldImpact-1	$O-Field-2	$O-FieldImpact-2
1	id633	T	1.600	2	claimvalue	0.358	farmincome	0.275
2	id647	T	1.403	2	farminco...	0.334	claimvalue	0.161
3	id654	T	1.495	2	rainfall	0.322	maincrop	0.181
4	id703	T	1.358	1	rainfall	0.230	region	0.219
5	id704	T	1.427	2	farminco...	0.287	maincrop	0.190
6	id739	T	1.684	2	claimvalue	0.404	farmincome	0.233
7	id752	T	1.770	2	claimvalue	0.391	farmincome	0.155
8	id791	T	1.386	1	maincrop	0.236	rainfall	0.163
9	id813	T	1.641	1	region	0.181	landquality	0.160
10	id883	T	1.350	2	region	0.187	maincrop	0.169

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.

Figura 4-10
Detalles del nugget de modelo de detección de anomalías

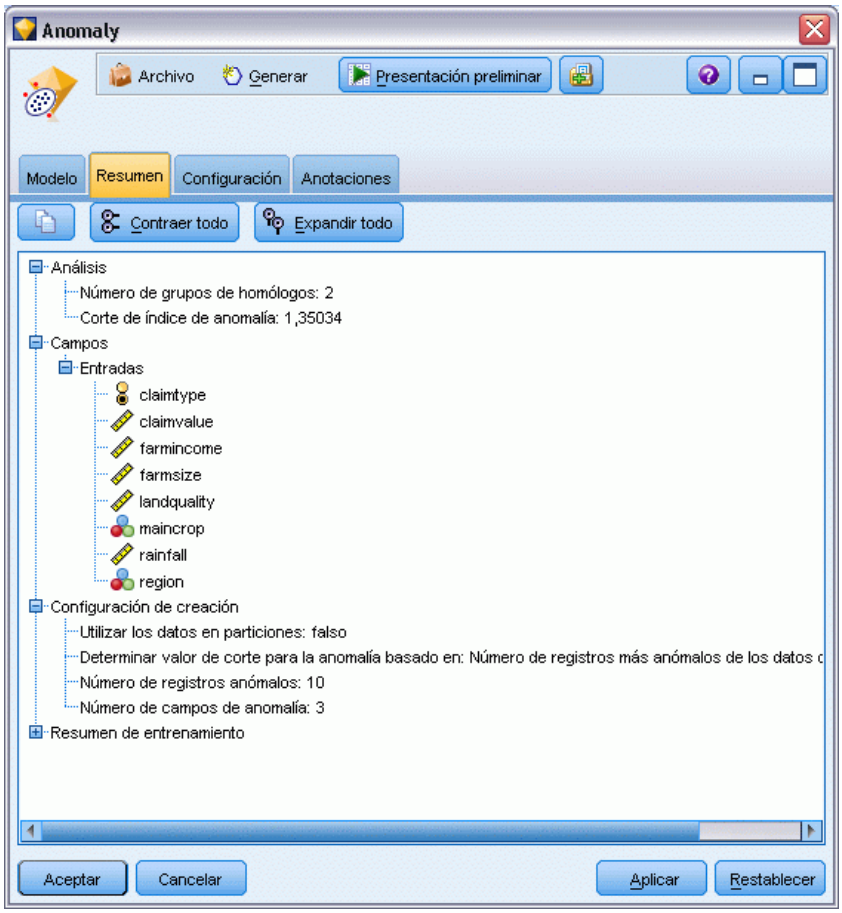


Tenga en cuenta que los tamaños de los grupos de homólogos y los estadísticos de los que se ha informado son cálculos basados en los datos de entrenamiento y, por lo tanto, pueden diferir ligeramente de los resultados de puntuación, aun cuando se ejecuten en los mismos datos.

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.

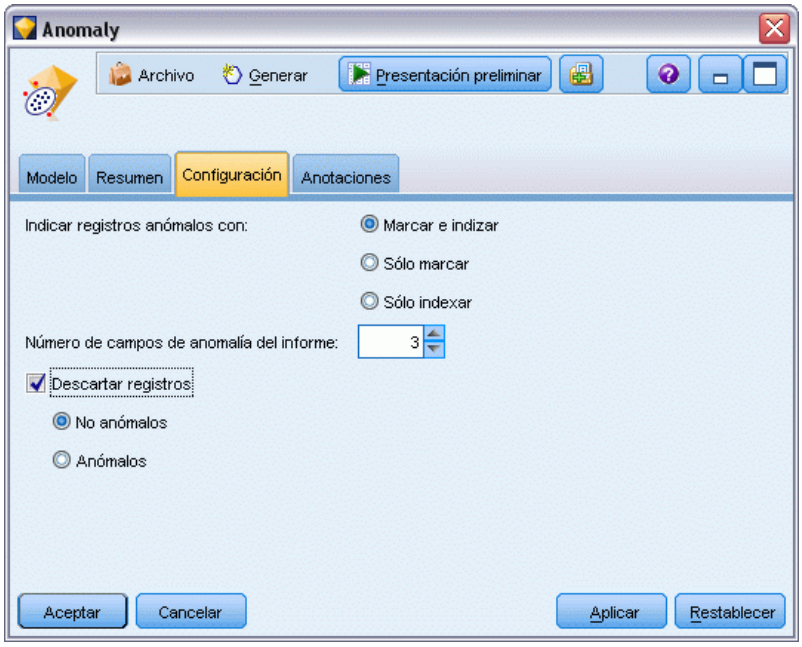
Figura 4-11
Resumen del nugget de modelo de detección de anomalías



Configuración del modelo de detección de anomalías

La pestaña Configuración permite especificar opciones para puntuar el nugget de modelo.

Figura 4-12
Opciones de puntuación para un nugget de modelo de detección de anomalías



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

- **Marcar e indizar.** 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. Si desea obtener más información, consulte el tema Opciones del modelo de detección de anomalías el p. 86.
- **Sólo marcar.** Crea un campo de marcas, pero no se informa del índice de anomalía de cada registro.
- **Sólo indexar.** Informa del índice de anomalía sin crear un campo de marcas.

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 conglomerado 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 por debajo de 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.

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 el modelo de cada 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, pérdida 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. Si desea obtener más información, consulte el tema Nodo Clasificador automático el p. 97.

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. Si desea obtener más información, consulte el tema Nodo Autonumérico el p. 107.

El nodo Autoconglomeración calcula y compara los modelos de conglomerado que identifican grupos de registros con características similares. El nodo funciona de la misma manera que otros nodos de modelado de conglomerado, 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 conglomerado y proporcionar una medida según la importancia de campos concretos. Si desea obtener más información, consulte el tema Nodo Autoconglomeración el p. 113.

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

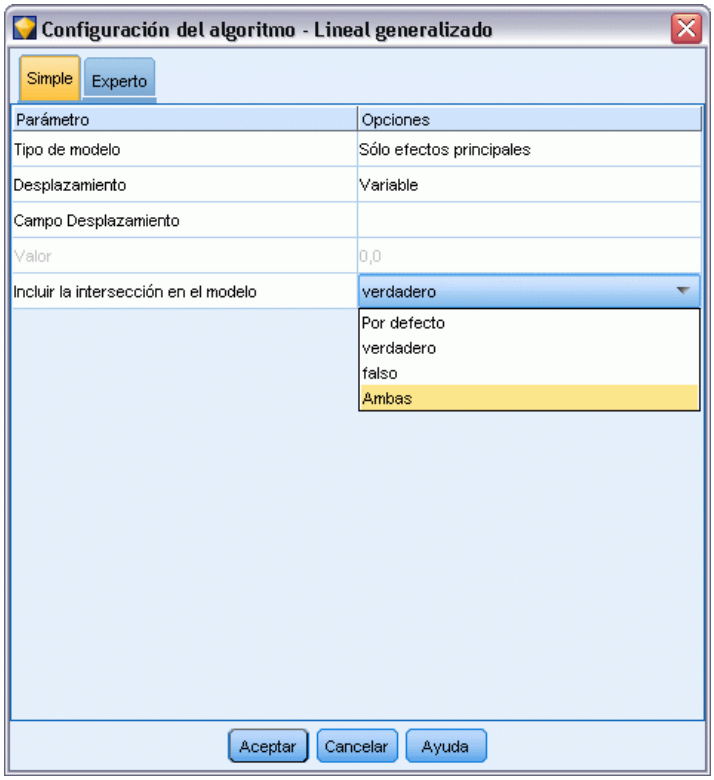
Dependiendo del conjunto de datos y del número de modelos, los nodos de modelado automático pueden tardar horas en ejecutarse. Preste atención al número de modelos que se están generando al seleccionar opciones. Si le resulta práctico, puede programar modelados para que se ejecuten por las noches o durante 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. Consulte Nodo Muestrear o Nodo Partición si desea obtener más información.
- Para reducir el número de campos de entrada, utilice Selección de características. Si desea obtener más información, consulte el tema Nodo Selección de características en el capítulo 4 el p. 76. 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 ordenar modelos.

Ajustes de algoritmo de nodos de modelado automático

Puede usar la configuración por defecto 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.

Figura 5-1
Selección de ajustes de algoritmo para el modelado automático



Para seleccionar las opciones para cada tipo de modelo

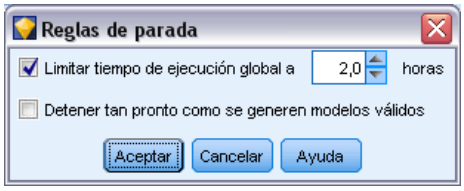
- En el nodo de modelado automatizado, seleccione la pestaña Experto.
- Haga clic en la columna Parámetros de modelo para el tipo de modelo.
- En el menú desplegable, seleccione Especificar.
- En el cuadro de diálogo Configuración de algoritmo, seleccione las opcioens de la columna Opciones.

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.

Figura 5-2
Reglas de parada



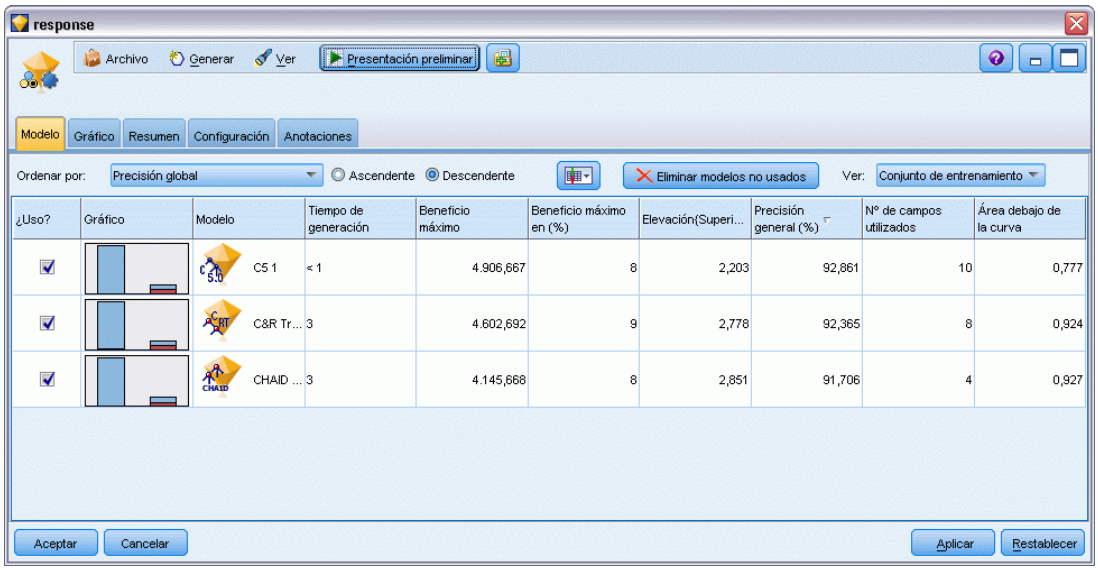
Limitar tiempo de ejecución global a. (Sólo modelos de Red neuronal, K-medias, Kohonen, Bietápico, SVM, KNN, Red bayesiana y C&R Tree) Detiene la ejecución tras un número específico de horas. Se incluirán en el nugget de modelo todos los modelos generados hasta ese momento, pero no se producirán más 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 Autoconglomerado) o la pestaña Modelo (para el nodo Autonumérico). Si desea obtener más información, consulte el tema Opciones para descartar el nodo Clasificador automático el p. 105.Si desea obtener más información, consulte el tema Opciones para descartar del nodo Autoconglomeración el p. 118.

Nodo Clasificador automático

El nodo Clasificador automático calcula y compara los modelos de objetivos nominales (conjuntos) o binarios (yes/no) utilizando métodos diferentes y permitiéndole probar una gran variedad de métodos en una única tirada. Puede seleccionar los algoritmos que se utilizarán y experimentar con múltiples combinaciones de opciones. 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 el modelo de cada candidato en función de la medida especificada y guarda los mejores para su uso en la puntuación o en futuros análisis.Si desea obtener más información, consulte el tema Nodos de modelado automático el p. 94.

Figura 5-3
Resultados del modelado de Clasificador automático



Ejemplo. Una empresa minorista contiene datos históricos en los que se registran las ofertas realizadas a determinados clientes en campañas anteriores. La empresa ahora desea lograr resultados más rentables, realizando la mejor oferta para cada cliente.Demostración

Requisitos. Un campo de objetivo con un nivel de medición de *Nominal* o *Marca* (con el papel establecido a Objetivo), y al menos un campo de entrada (con el papel 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 el tema Uso de campos de frecuencia y ponderación en el capítulo 3 el p. 41.

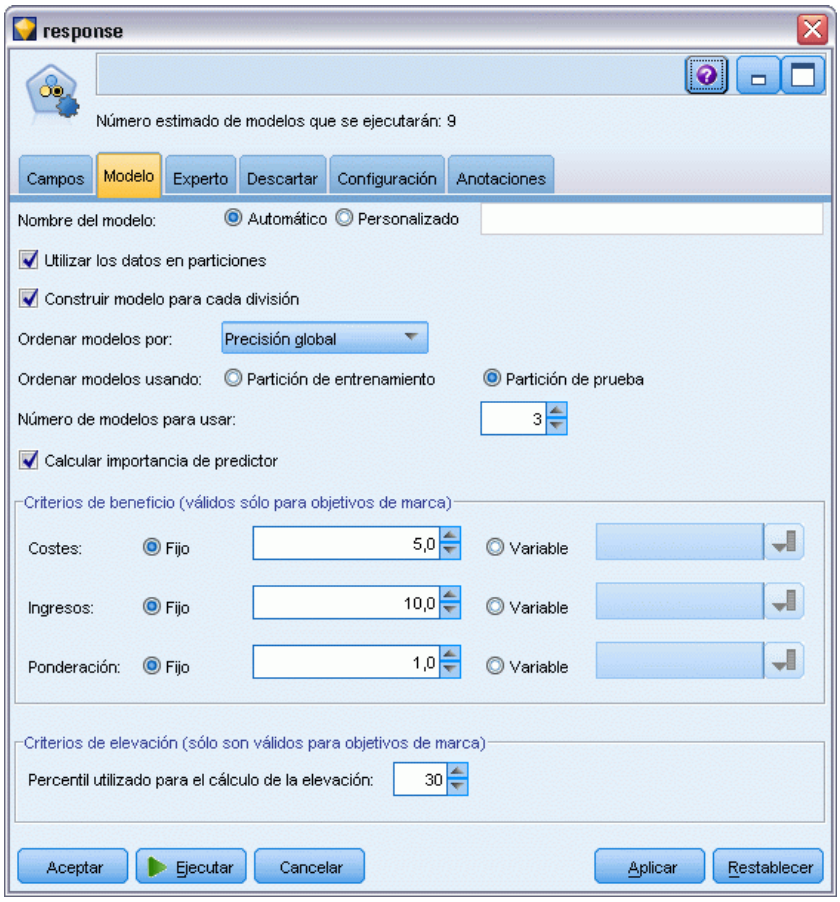
Tipos de modelos admitidos

Los tipos de modelo admitidos incluyen Red neuronal, C&RT, QUEST, CHAID, C5.0, Regresión logística, Lista de decisiones, Red bayesiana, Discriminante, Vecino más cercano y SVM. Si desea obtener más información, consulte el tema Opciones de experto para el nodo Clasificador automático el p. 101.

Opciones de modelo para el nodo Clasificador automático

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

Figura 5-4
Nodo Clasificador automático: Pestaña 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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Crear modelos divididos. Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como modelos divididos. Si desea obtener más información, consulte el tema Generación de modelos divididos en el capítulo 3 el p. 32.

Ordenar 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 ordenación 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.

-
-
-
-
-

Ordenar 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 filtrado 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 lista se enumeran en función del criterio de ordenación 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. Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

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.

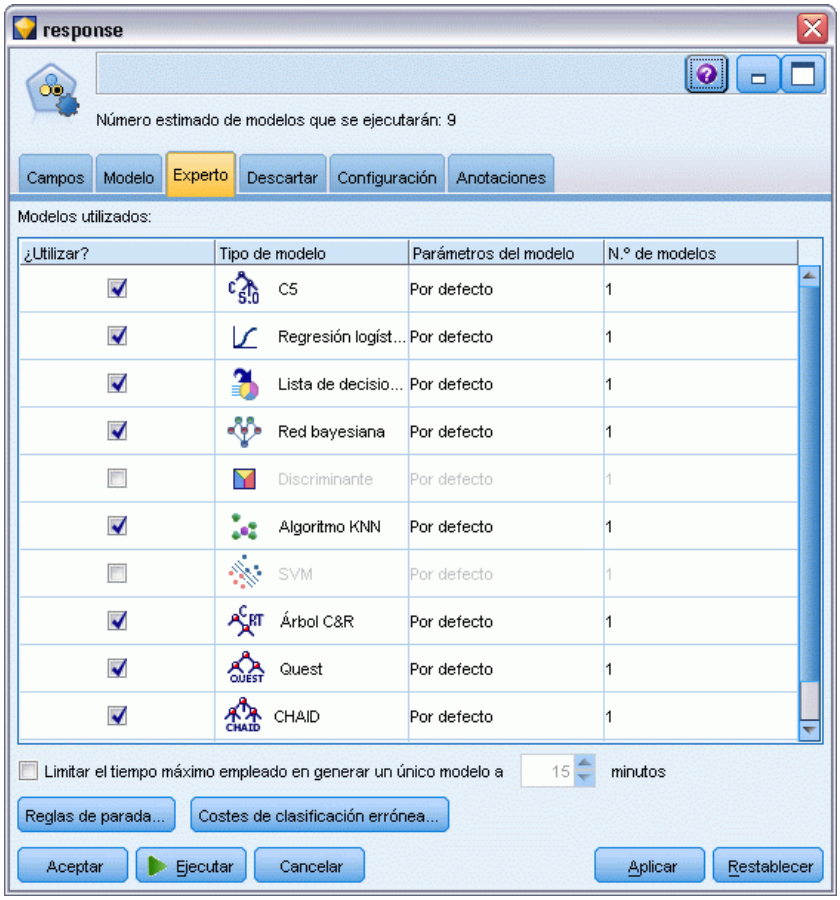
- **Costes.** Permite especificar el coste asociado con cada registro. Puede seleccionar costes fijos o variables. En el caso de los costes fijos, especifique el valor del coste. En el caso de los costes variables, pulse en el selector de campos para elegir un campo de costes.
- **Ingresos.** Permite 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.
- **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.

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. Si desea obtener más información, consulte el tema Nugget de modelo automático el p. 119.

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 va a usar y especificar las reglas de parada.

Figura 5-5
Nodo Clasificador automático: Pestaña Experto



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 por defecto 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 objetivo es un campo nominal (conjunto), la opción Lista de decisiones no está disponible.

Algoritmos admitidos

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 estimadores potentes de funciones generales y requieren un conocimiento matemático o estadístico mínimo para entrenarlas o aplicarlas.

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. Si desea obtener más información, consulte el tema Nodo C5.0 en el capítulo 6 el p. 174.

El nodo de árbol de clasificación y regresión (C&R) genera un árbol de decisión que permite pronosticar 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). Si desea obtener más información, consulte el tema Nodo Árbol C&R en el capítulo 6 el p. 152.

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. Si desea obtener más información, consulte el tema Nodo QUEST en el capítulo 6 el p. 154.

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. Si desea obtener más información, consulte el tema Nodo CHAID en el capítulo 6 el p. 153.

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. Si desea obtener más información, consulte el tema Nodo Logística en el capítulo 10 el p. 278.

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 perder clientes 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. Si desea obtener más información, consulte el tema Lista de decisiones en el capítulo 9 el p. 219.

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 cadena de Markov que se utilizan principalmente para la clasificación. Si desea obtener más información, consulte el tema Nodo Red bayesiana en el capítulo 7 el p. 194.

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. Si desea obtener más información, consulte el tema Nodo Discriminante en el capítulo 10 el p. 307.

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. Si desea obtener más información, consulte el tema Nodo KNN en el capítulo 16 el p. 495.

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. Si desea obtener más información, consulte el tema Nodo SVM en el capítulo 15 el p. 489.

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 pronóstico.

Los costes de clasificación errónea son básicamente ponderaciones aplicadas a resultados específicos. Estas ponderaciones se extraen en el modelo y pueden realmente cambiar el pronóstico (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 ordenan 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*.

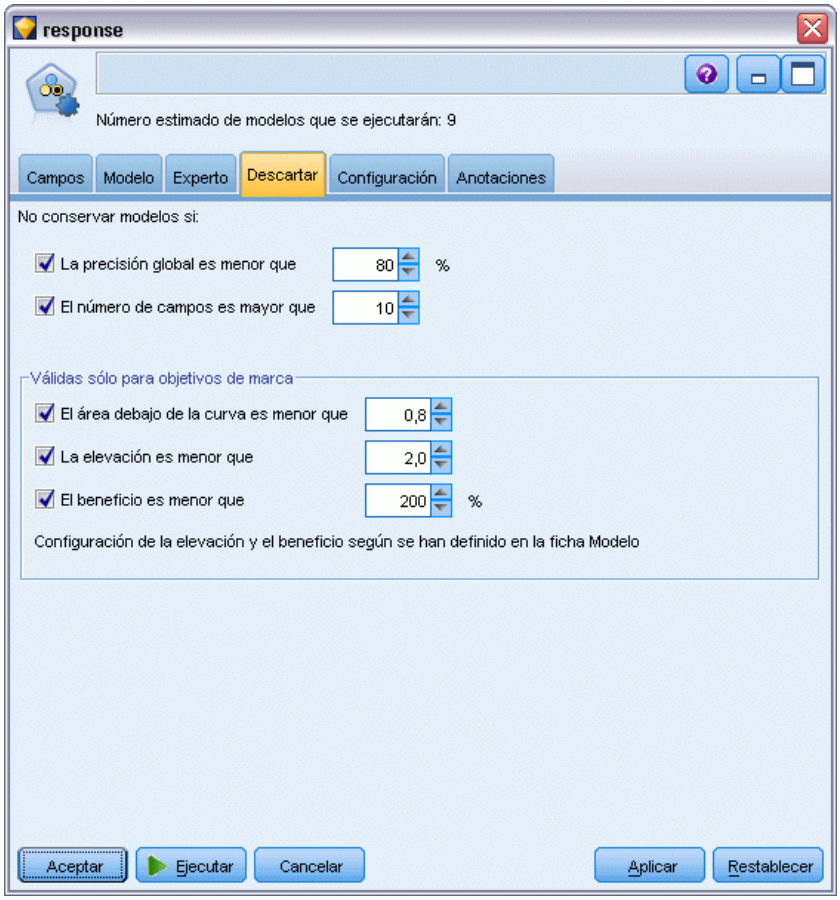
La matriz de costes muestra el coste para cada combinación posible de categoría pronosticada y categoría real. Por defecto, todos los costes de clasificación errónea se establecen en 1,0. Para introducir valores de coste personalizados, seleccione Utilizar costes de clasificación errónea e introduzca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

Para cambiar un coste de clasificación errónea, seleccione la casilla correspondiente a la combinación deseada de valores pronosticados 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 por defecto 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.

Figura 5-6
Nodo Clasificador automático: Pestaña Descartar



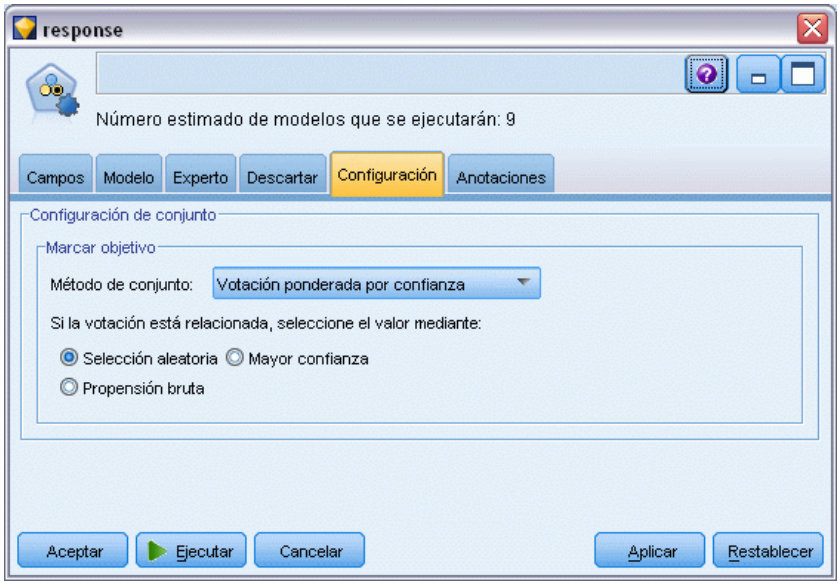
Puede especificar un umbral mínimo para la precisión global y un umbral máximo para el número de variables usadas en el modelo. Además, en el caso de objetivos de marca, puede especificar un umbral mínimo para la elevación, los beneficios y un área debajo de la curva; la elevación y los beneficios se determinan según lo especificado en la pestaña Modelo. Si desea obtener más información, consulte el tema Opciones de modelo para el nodo Clasificador automático el p. 99.

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. Si desea obtener más información, consulte el tema Reglas de parada de nodos de modelado automático el p. 96.

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.

Figura 5-7
Nodo Clasificador automático: Pestaña Configuración



Método de conjunto. En objetivos puede seleccionar uno de los siguientes métodos de conjunto:

- Voting
- Votación ponderada de confianza
- Votación ponderada de propensión bruta (sólo objetivos de marca)
- La mayor confianza gana
- Propensión bruta media (sólo objetivos de marca).

Si desea obtener más información, consulte el tema Configuración de nodo Conjunto en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Si la votación está empatada, seleccione el valor de uso. En los métodos de votación puede especificar cómo se resuelven los empates:

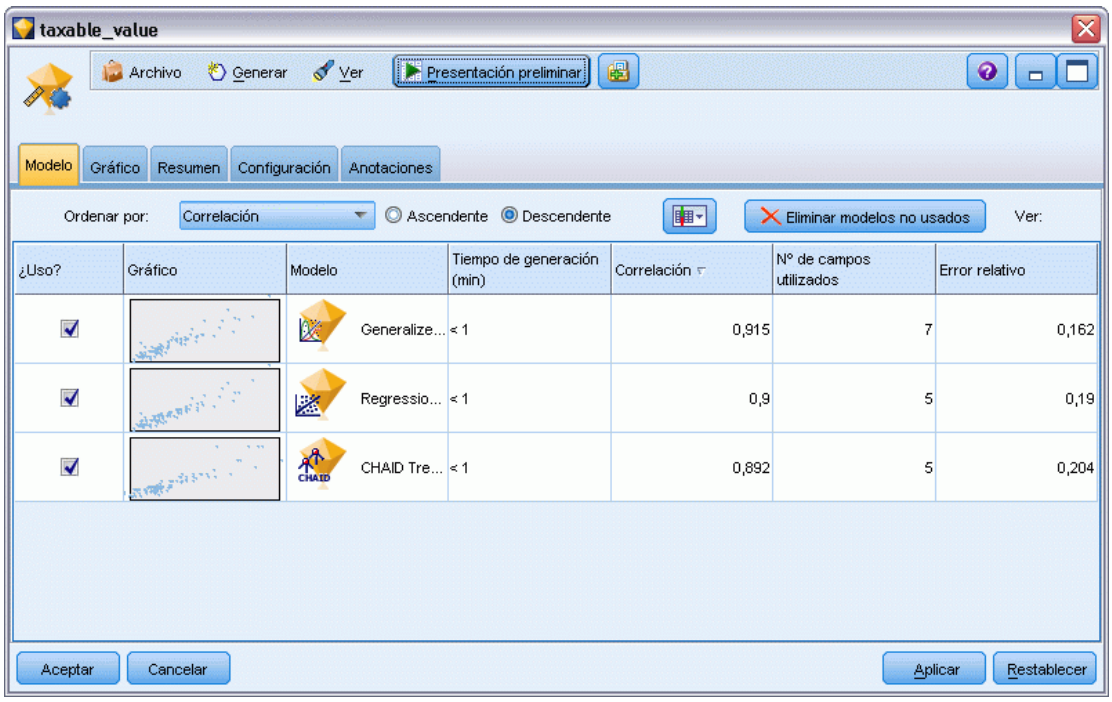
- **Selección aleatoria.** Se seleccionará aleatoriamente uno de los valores empatados.
- **La mayor confianza.** El valor con el mayor índice de confianza gana. Tenga en cuenta que no es necesariamente la misma opción que la del índice de mayor confianza de todos los valores pronosticados.
- **Propensión bruta.** (Sólo objetivos de marca) El valor empatado pronosticado con la mayor propensión absoluta, donde la propensión absoluta se calcula como:

abs(0,5 - propensión) * 2

Nodo Autonumérico

El nodo Autonumérico calcula y compara los modelos de resultados de rango numérico continuo utilizando métodos diferentes y permitiéndole probar una gran variedad de métodos en una única tirada. Puede seleccionar los algoritmos que se utilizarán y experimentar con múltiples combinaciones de opciones. Por ejemplo, puede pronosticar 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 el modelo de cada candidato en función de la medida especificada y guarda los mejores para su uso en la puntuación o en futuros análisis. Si desea obtener más información, consulte el tema Nodos de modelado automático el p. 94.

Figura 5-8
Resultados Autonuméricos



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 pronostican valores de propiedad basándose en el tipo de edificación, vecindario, tamaño y otros factores conocidos. Si desea obtener más información, consulte el tema Valores de propiedad (Autonumérico) en el capítulo 5 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

Requisitos. Un único campo objetivo (con el papel establecido a Objetivo), y al menos un campo de entrada (con el papel 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. Si desea obtener más información, consulte el tema Uso de campos de frecuencia y ponderación en el capítulo 3 el p. 41.

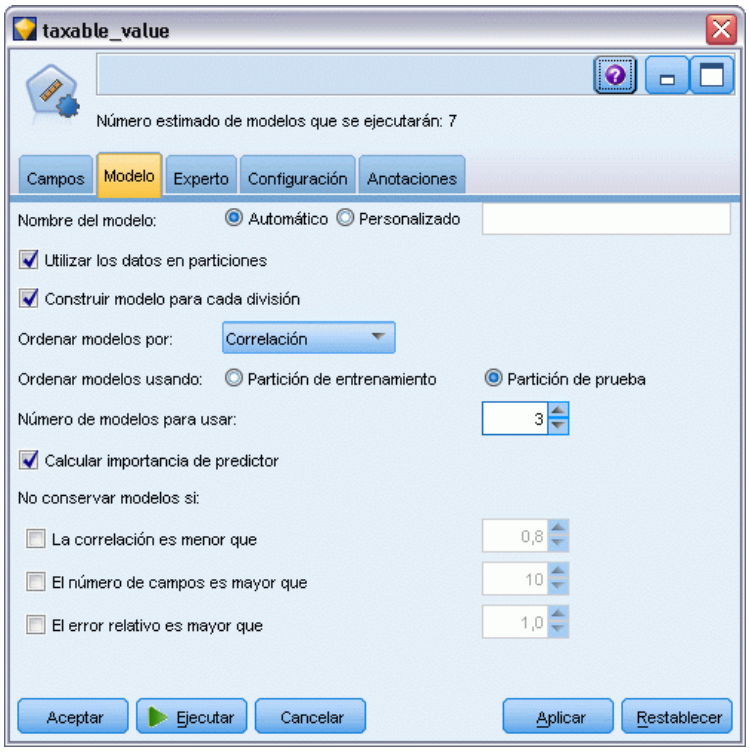
Tipos de modelos admitidos

Los tipos de modelo admitidos incluyen Red neuronal, C&RT, CHAID, Regresión, GenLin, Vecino más cercano y SVM. Si desea obtener más información, consulte el tema Opciones de experto para el nodo Autonumérico el p. 111.

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.

Figura 5-9
Nodo autonumérico: Pestaña 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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Crear modelos divididos. Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como modelos divididos. Si desea obtener más información, consulte el tema Generación de modelos divididos en el capítulo 3 el p. 32.

Ordenar modelos por. Especifique los criterios utilizados para comparar modelos.

- **Correlación.** Correlación de Pearson entre el valor observado para cada registro y el valor pronosticado 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 entre –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 pronosticados 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 intersección** que simplemente devuelve el valor medio del campo objetivo como el pronóstico. 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.

Ordenar 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 filtrado preliminar de modelos, puede mejorar rendimiento en gran medida. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

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 lista se enumeran en función del criterio de ordenación 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. Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

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.

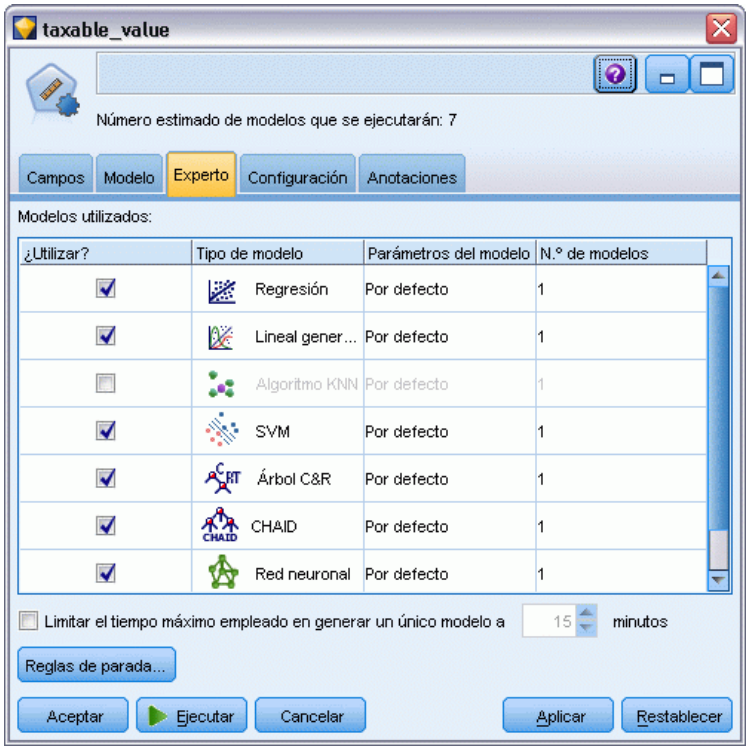
- **Correlación menor que.** Correlación mínima (en cuanto a valor absoluto) para que un modelo se incluya en el informe de resumen.
- **Número de campos utilizados mayor que.**Número máximo de campos que puede utilizar cualquier modelo que vaya a incluirse.
- **Error relativo mayor que.** Error relativo 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. Si desea obtener más información, consulte el tema Reglas de parada de nodos de modelado automático el p. 96.

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.

Figura 5-10
Nodo autonumérico: Pestaña Experto



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 por defecto 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 admitidos

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 estimadores potentes 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 pronosticar 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). Si desea obtener más información, consulte el tema Nodo Árbol C&R en el capítulo 6 el p. 152.

El nodo CHAID genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. A diferencia de los nodos C&R 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. Si desea obtener más información, consulte el tema Nodo CHAID en el capítulo 6 el p. 153.

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

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 que no sea 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. Si desea obtener más información, consulte el tema Nodo GenLin en el capítulo 10 el p. 316.

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. Si desea obtener más información, consulte el tema Nodo KNN en el capítulo 16 el p. 495.

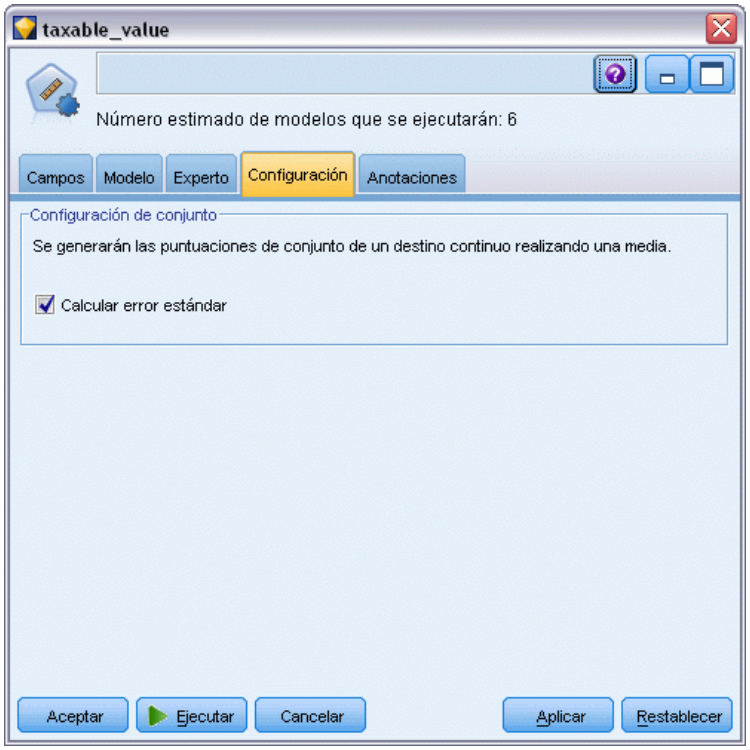
El nodo Máquina de vectores de soporte (SVM) le permite clasificar datos en uno o dos grupos sin que haya un ajuste por exceso. SVM funciona bien con conjuntos de datos grandes, como aquellos con un gran número de campos de entrada. Si desea obtener más información, consulte el tema Nodo SVM en el capítulo 15 el p. 489.

Los modelos de regresión lineal predicen un destino continuo tomando como base las relaciones lineales entre el destino y uno o más predictores. Si desea obtener más información, consulte el tema Modelos lineales en el capítulo 10 el p. 258.

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.

Figura 5-11
Nodo autonumérico: Pestaña Configuración



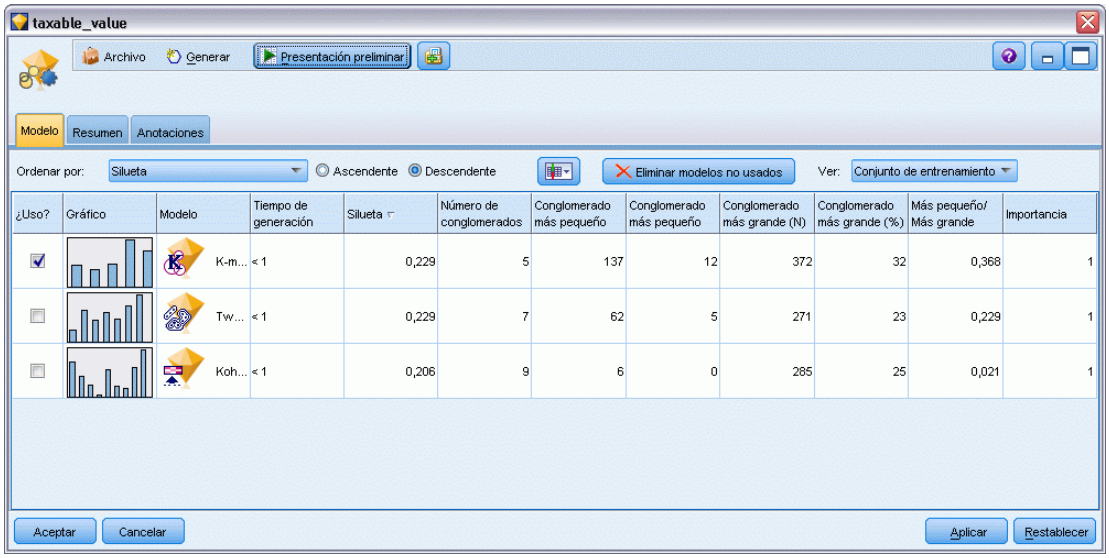
Calcular error estándar. Para un objetivo continuo (rango numérico), se ejecuta un error estándar por defecto para calcular la diferencia entre los valores medidos o estimados y los valores true; y para mostrar si las estimaciones coinciden.

Nodo Autoconglomeración

El nodo Autoconglomeración calcula y compara los modelos de conglomerado que identifican grupos de registros con características similares. El nodo funciona de la misma manera que otros nodos de modelado de conglomerado, 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 conglomerado y proporcionar una medida según la importancia de campos concretos.

Los modelos de conglomerado 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 funciones que debe utilizar para definirlos. Los modelos de conglomerado 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 conglomerado viene determinado por su capacidad de capturar agrupaciones interesantes en los datos y proporcionar descripciones útiles de dichas agrupaciones. Si desea obtener más información, consulte el tema Modelos de conglomerados en el capítulo 11 el p. 370.

Figura 5-12
Resultados de autoconglomerado



Requisitos. Uno o más campos que definen las características de interés. Los modelos de conglomerados 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 conglomerado para predecir si un cliente concreto abandonará o responderá a una oferta. Pero puede utilizar un modelo de conglomerado 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 conglomerado se puede evaluar lo bien (o mal) que los conglomerados diferencian los campos.

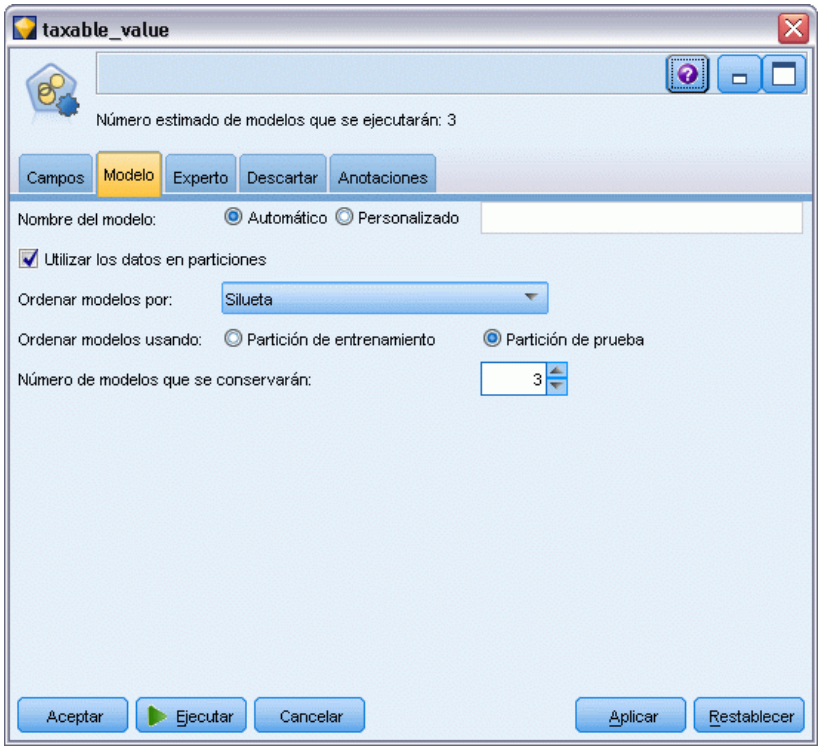
Tipos de modelos admitidos

Entre los tipos de modelos admitidos se incluyen TwoStep, K-Means y Kohonen.

Opciones de modelo para el nodo Autoconglomeración

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

Figura 5-13
Nodo Autoconglomeración: Pestaña 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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Ordenar modelos por. Especifique los criterios utilizados para comparar y clasificar los modelos.

- **Silueta.** Un índice que mide la cohesión y separación del conglomerado. Consulte *Medida de ordenación de siluetas* a continuación para obtener más información.
- **Número de conglomerados.** El número de conglomerados que se utilizan en el modelo.
- **Tamaño del conglomerado más pequeño.** El menor tamaño de conglomerado.
- **Tamaño del conglomerado mayor.** El mayor tamaño de conglomerado.
- **Conglomerado mayor / menor.** La razón del tamaño del conglomerado 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.

Ordenar 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 filtrado 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 lista se enumeran en función del criterio de ordenación especificado. Tenga en cuenta que si aumenta este límite puede ralentizarse el rendimiento. El valor máximo permitido es 100.

Medida de ordenación de siluetas

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

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

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

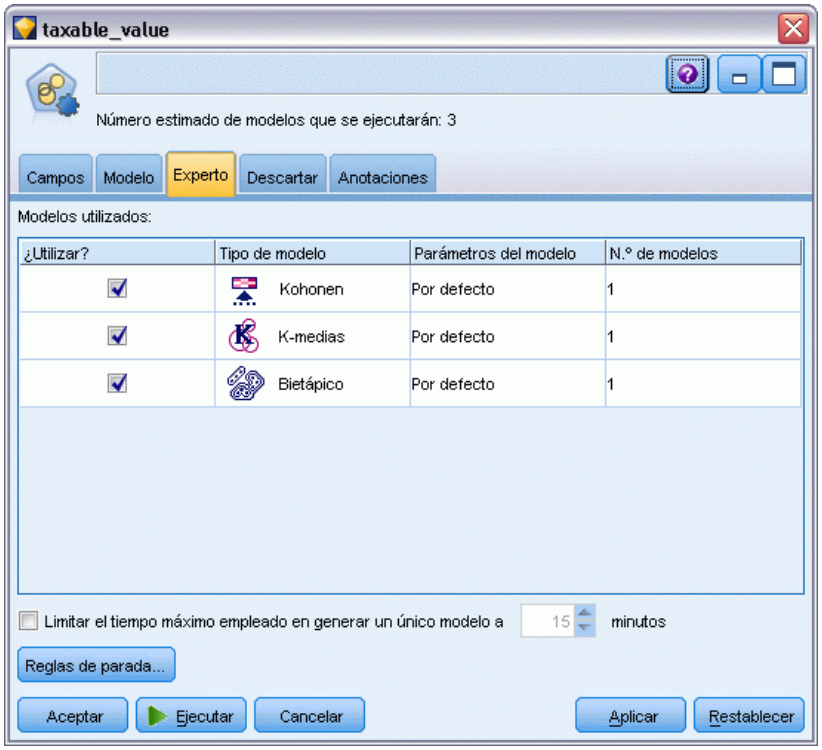
donde A es la distancia desde el caso hasta el centroide del conglomerado al que pertenece el caso; y B es la distancia mínima desde el caso hasta el centroide de cada uno de los otros conglomerados.

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 conglomerados (lo cual produce una silueta de conglomerados). Las distancias pueden calcularse utilizando distancias euclídeas.

Opciones de experto para el nodo Autoconglomeración

La pestaña Experto del nodo Autoconglomeración le permite aplicar una partición (si está disponible), seleccionar los algoritmos que se va a usar y especificar las reglas de parada.

Figura 5-14
Nodo Autoconglomeración: Pestaña Experto



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 por defecto 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 admitidos

El nodo K-medias agrupa conjuntos de datos en grupos distintos (o conglomerados). El método define un número fijo de conglomerados, de forma iterativa asigna registros a los conglomerados y ajusta los centros de los conglomerados hasta que no se pueda mejorar el modelo. En lugar de intentar pronosticar 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. Si desea obtener más información, consulte el tema Nodo K-medias en el capítulo 11 el p. 378.

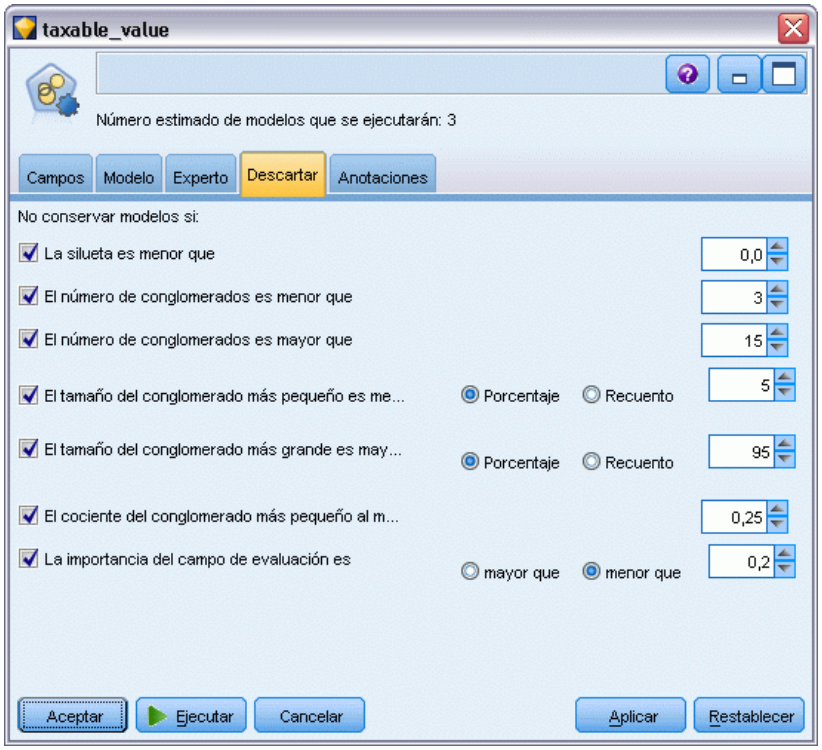
El nodo Kohonen genera un tipo de red neuronal que se puede usar para conglomerar un conjunto de datos en grupos distintos. Cuando la red se termina de entrenar, los registros que son similares se deberían cerrar juntos en el mapa de resultados, mientras que los registros que son diferentes aparecerían 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 conglomerados. Si desea obtener más información, consulte el tema Nodo Kohonen en el capítulo 11 el p. 371.

El nodo Bietápico es un método de conglomerado 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 subconglomerados administrable. El segundo paso utiliza un método de conglomerado jerárquico para fundir progresivamente los subconglomerados en conglomerados cada vez más grandes. El bietápico tiene la ventaja de estimar automáticamente el número óptimo de conglomerados para los datos de entrenamiento. Puede gestionar tipos de campos mixtos y grandes conjuntos de datos eficazmente. Si desea obtener más información, consulte el tema Nodo de conglomerado Bietápico en el capítulo 11 el p. 383.

Opciones para descartar del nodo Autoconglomeración

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

Figura 5-15
Nodo Autoconglomeración: Pestaña Descartar



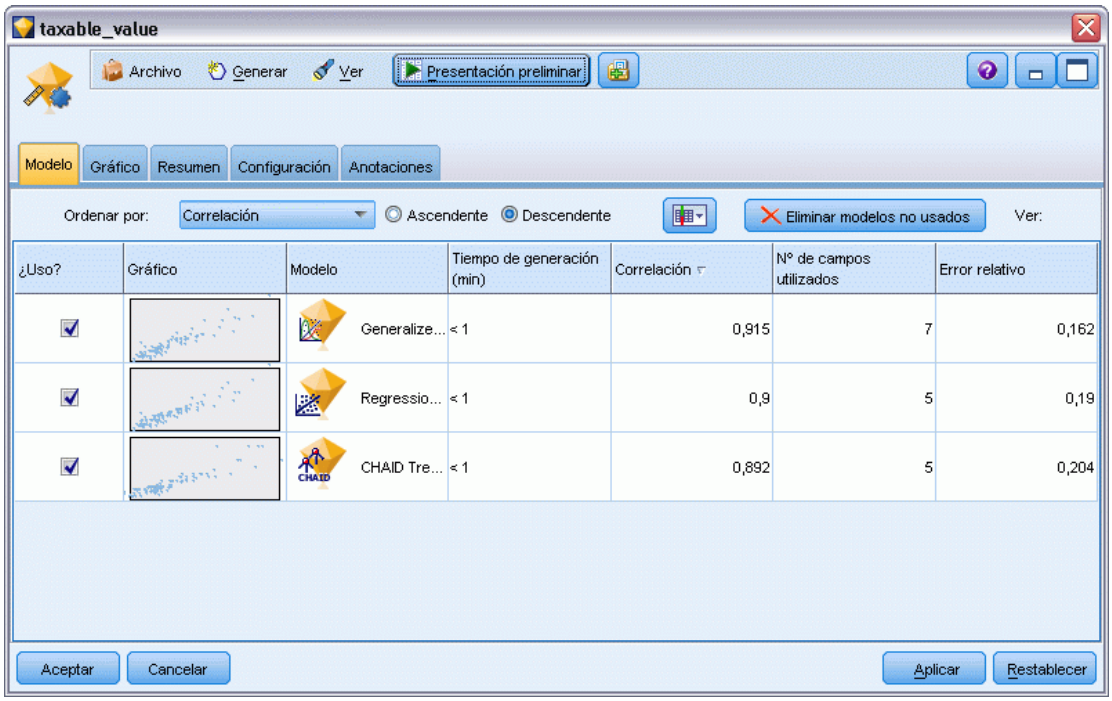
Puede especificar el valor mínimo de silueta, número de conglomerados, tamaños de conglomerados y la importancia del campo de evaluación en el modelo. La silueta y el número y tamaño de clústeres están determinados en la especificación del nodo de modelado. Si desea obtener más información, consulte el tema Opciones de modelo para el nodo Autoconglomeración el p. 115.

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. Si desea obtener más información, consulte el tema Reglas de parada de nodos de modelado automático el p. 96.

Nugget de modelo automático

Cuando se ejecuta el nodo de modelado automático, el nodo estima modelos de candidato 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.

Figura 5-16
Resultados Autonuméricos



- 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 valoración 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 cualquier cambio en los datos anteriores en la ruta (como la actualización de una muestra aleatoria o partición si Establecer semilla aleatoria no está seleccionado) pueda 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. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

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 pronosticados, para ofrecer una rápida indicación visual de cuántos registros se pronosticaron correctamente en cada categoría.
- Los criterios de clasificación coinciden con las opciones del nodo de modelado Clasificador automático. Si desea obtener más información, consulte el tema Opciones de modelo para el nodo Clasificador automático el p. 99.
- 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 pronosticados, para ofrecer una rápida indicación visual de cuántos registros se pronosticaron correctamente en cada categoría.
- Los criterios de clasificación coinciden con las opciones del nodo de modelado Clasificador automático. Si desea obtener más información, consulte el tema Opciones de modelo para el nodo Clasificador automático el p. 99.

Objetivos continuos

- En el caso de modelos continuos (rango numérico), el gráfico representa los valores pronosticados 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 conglomerarse 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. Si desea obtener más información, consulte el tema Opciones de modelo para el nodo Autonumérico el p. 108.

Objetivos de conglomerado

- En el caso de modelos de conglomerado, el gráfico representa los recuentos frente a los conglomerados para cada modelo, lo que ofrece una rápida indicación visual de la distribución de conglomerados.
- Los criterios de clasificación coinciden con las opciones del nodo de autoconglomerado. Si desea obtener más información, consulte el tema Opciones de modelo para el nodo Autoconglomeración el p. 115.

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 conglomerado sólo puede seleccionar un modelo de puntuación cada vez. Por defecto, 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

▶ 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.

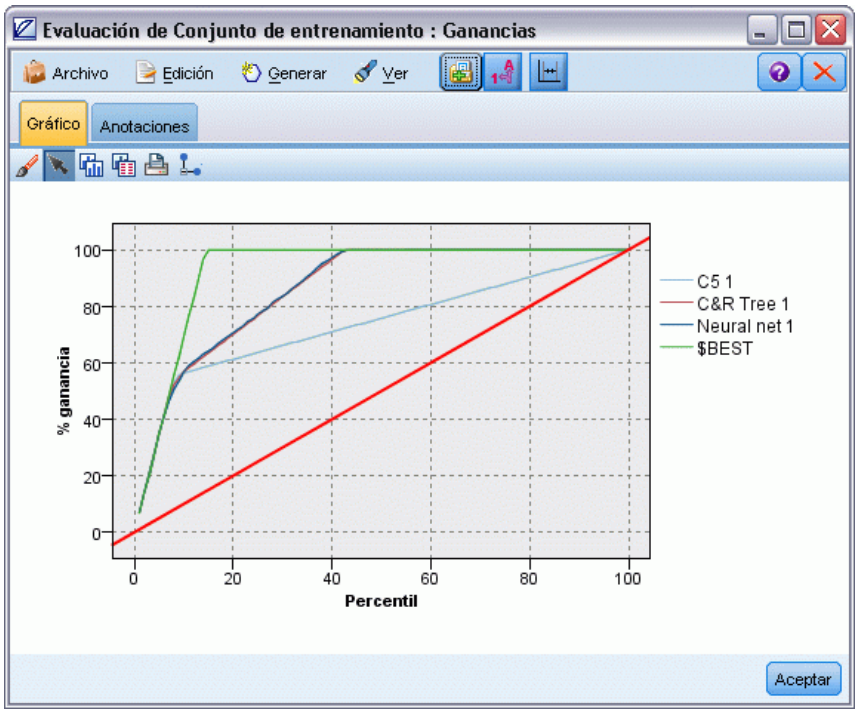
▶ En el menú Generar del nuevo cuadro de diálogo, seleccione Modelo(s) a paleta para añadir el nugget de modelado individual a la paleta de modelos.

▶ 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

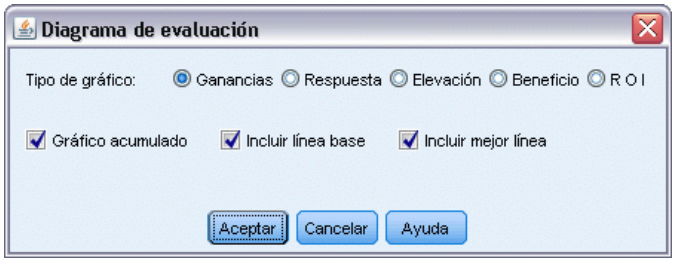
Sólo en el caso de modelos binarios, puede generar gráficos de evaluación que ofrecen un modo visual de valorar y comparar el rendimiento de cada modelo. Los gráficos de evaluación no están disponibles para modelos generados por los nodos Autonumérico o Autoconglomerado. Si desea obtener más información, consulte el tema Nodo Evaluación en el capítulo 5 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Figura 5-17
Gráfico de respuestas (acumulado) con mejor línea y línea base



- En la columna *Uso?* del nugget de modelo automático Clasificador automático, seleccione los modelos que desee evaluar.
- En el menú Generar, seleccione Diagrama(s) de evaluación.

Figura 5-18
Generación de un diagrama de evaluación



- Seleccione el tipo de diagrama y otras opciones que desee. Si desea obtener más información, consulte el tema Pestaña Gráfico de Evaluación en el capítulo 5 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

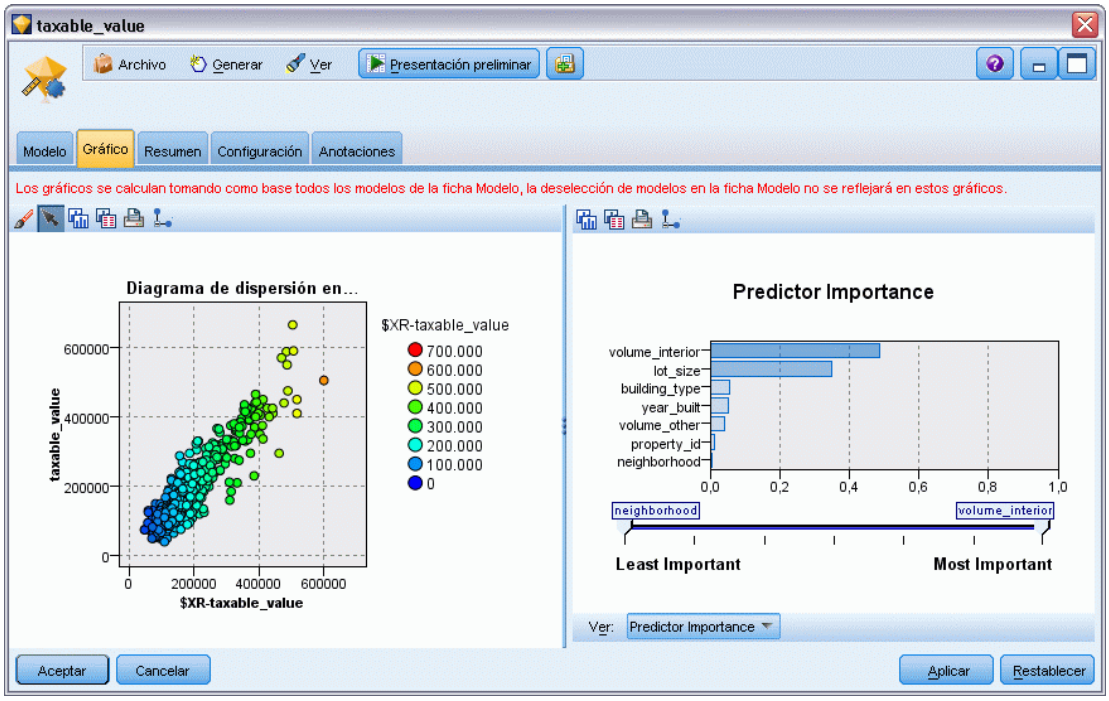
Gráficos de evaluación

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. Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

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).Si desea obtener más información, consulte el tema Características comunes de nodos de gráficos en el capítulo 5 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Figura 5-19
Autonumérico: gráfico múltiple para los modelos de conjunto dentro del nugget de modelo automático



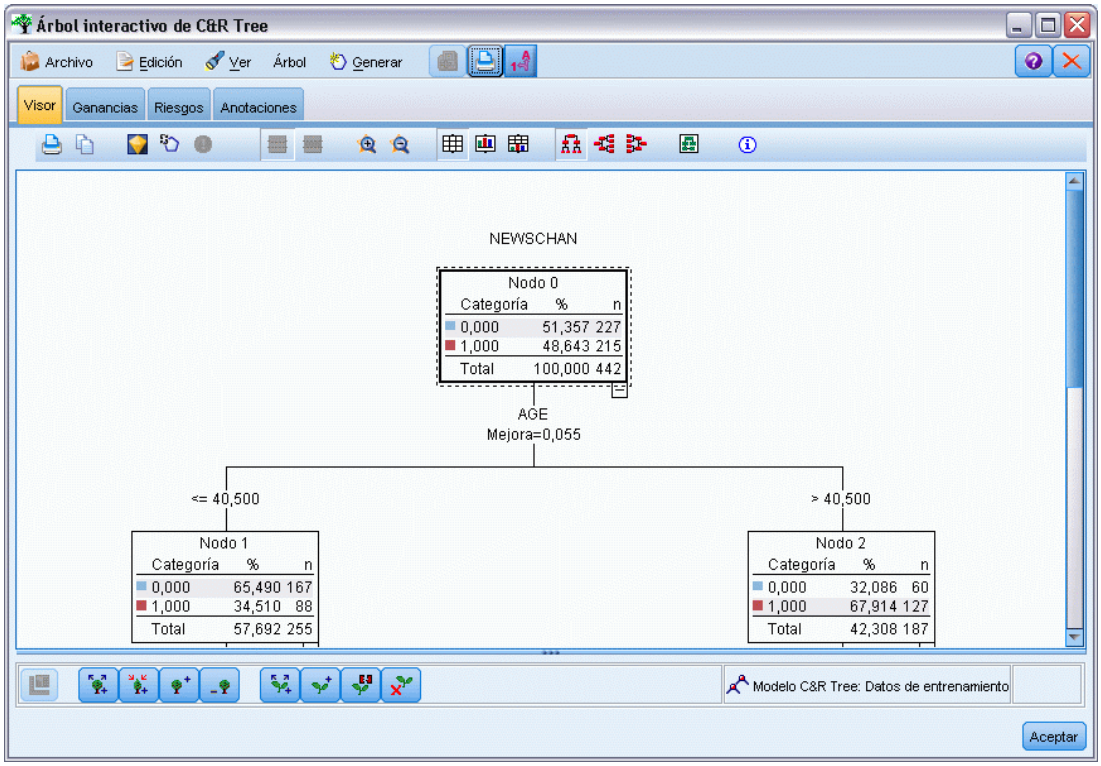
Capítulo 6

Árboles de decisión

Modelos de árboles de decisión

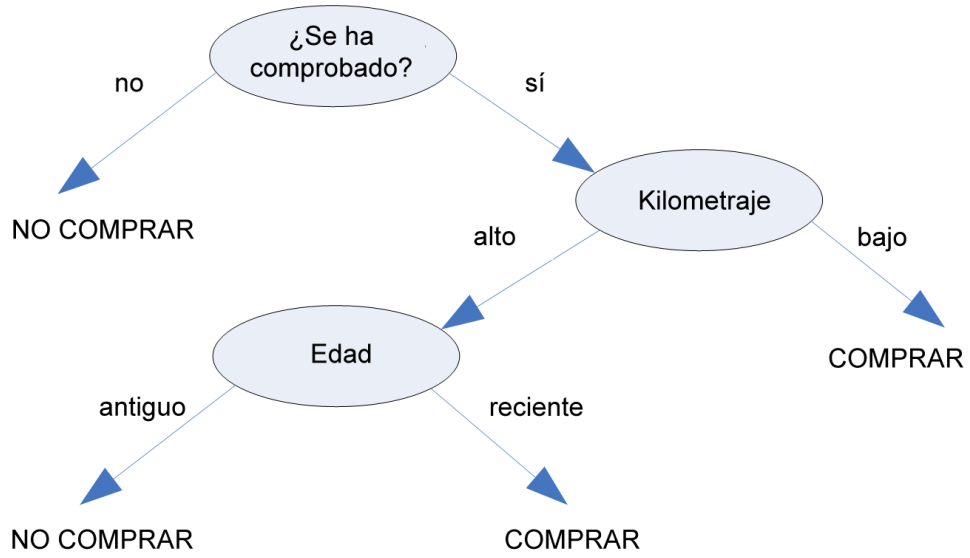
Los modelos de árboles de decisión permiten desarrollar sistemas de clasificación que pronostican o clasifican observaciones según un conjunto de reglas de decisión. Si dispone de datos divididos en clases que le interesan (por ejemplo, préstamos de alto riesgo frente a préstamos de bajo riesgo, suscriptores 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.

Figura 6-1
Ventana Árbol interactivo



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.

Figura 6-2
Árbol de decisión simple para comprar un coche



En segundo lugar, el proceso incluirá 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 árboles de decisión se pueden convertir en una recopilación de reglas si-entonces (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 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 nos proporciona un **perfil** de un grupo de vehículos que merece la pena comprar:

```
IF tested = 'yes'
AND mileage = 'low'
THEN -> 'BUY'.
```

Algoritmos de generación de árboles

Hay cuatro algoritmos disponibles para realizar un análisis de segmentación y clasificación. Todos estos algoritmos son básicamente similares: examinan todos los campos de su conjunto de datos para detectar el que proporciona la mejor clasificación o pronóstico 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 pronosticar 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). Si desea obtener más información, consulte el tema Nodo Árbol C&R el p. 152.

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. Si desea obtener más información, consulte el tema Nodo CHAID el p. 153.

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. Si desea obtener más información, consulte el tema Nodo QUEST el p. 154.

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. Si desea obtener más información, consulte el tema Nodo C5.0 el p. 174.

Usos generales del análisis basado en árboles

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

Segmentación. Identifica personas con probabilidad de pertenecer a una determinada .

Estratificación. Asigna casos en una o varias categorías, como grupos de alto, medio y bajo riesgo.

Pronóstico. Crea reglas y las usa para pronosticar eventos futuros. Los pronósticos también pueden significar intentos de relacionar atributos predictivos con valores de una variable continua.

Reducción de datos y filtrado de variables. Selecciona un subconjunto útil de predictores de un gran conjunto de variables para usarlo en la creación de un modelo paramétrico formal.

Identificación de interacción . Identifica las relaciones que pertenecen sólo a subgrupos determinados y las especifica en un modelo paramétrico formal.

Fusión de categorías y unión de variables continuas. Recodifica categorías de un predictor de grupos y variables continuas con una pérdida mínima de información.

El Generador de árboles interactivos

Puede generar un modelo de árbol automáticamente, que permita al algoritmo seleccionar 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.

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

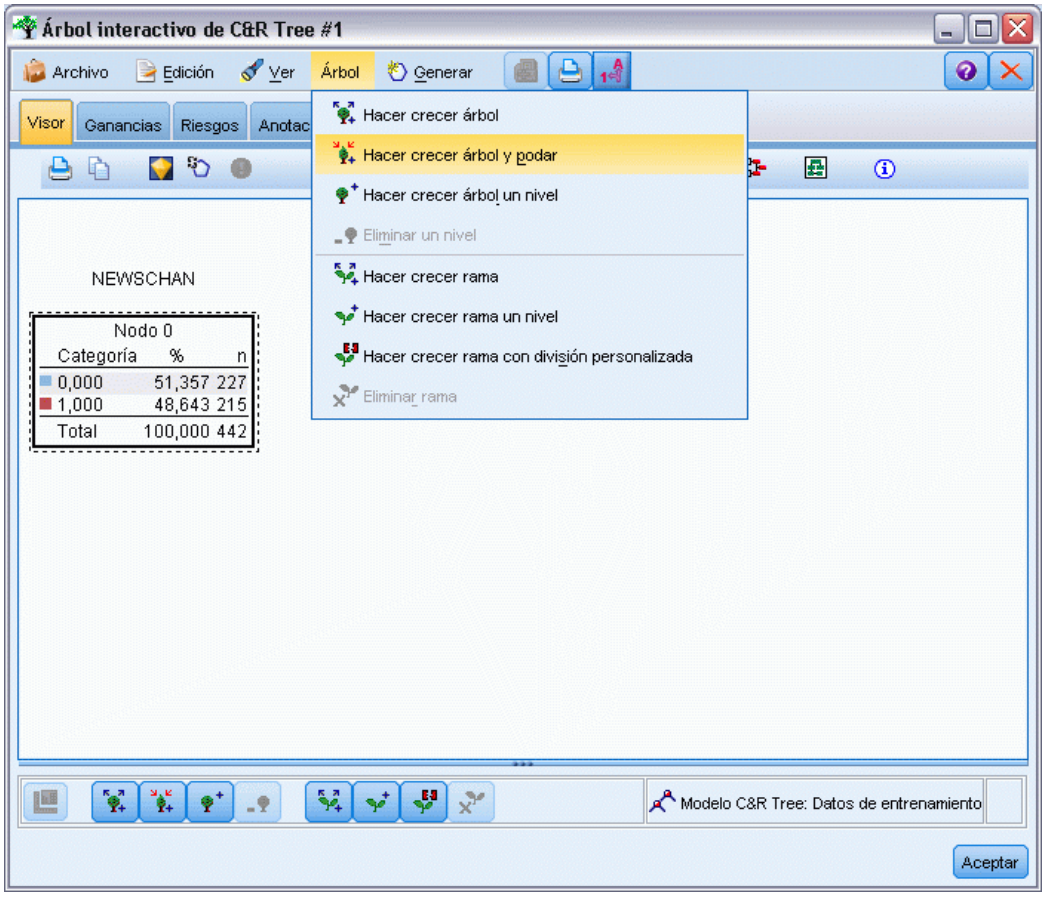
  Nota: los árboles C5.0 no admiten la generación de árboles interactivos.

- 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.

- En el panel Objetivos de la pestaña Opciones de generación, seleccione Iniciar sesión interactiva.

- Pulse en Ejecutar para iniciar el generador de árboles.

Figura 6-3
Ventana del Generador de árboles interactivos



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 desea obtener más información, consulte el tema Nodo Reclasificar en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.
- Si lo desea, puede utilizar un campo de partición para separar los datos de las muestras de comprobación y entrenamiento. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.
- 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. Si desea obtener más información, consulte el tema Creación directa de un modelo de árbol el p. 150.

Desarrollo y poda del árbol

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

▶ 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.

▶ Si lo prefiere, también puede seleccionar Hacer crecer árbol un nivel para añadir un solo nivel.

▶ Para añadir una rama bajo un nodo específico, seleccione el nodo y, a continuación, Hacer crecer rama.

▶ Para seleccionar el predictor que debe utilizarse en una división, seleccione el nodo que desee y, a continuación, Hacer crecer rama con división personalizada. Si desea obtener más información, consulte el tema Definición de divisiones personalizadas el p. 130.

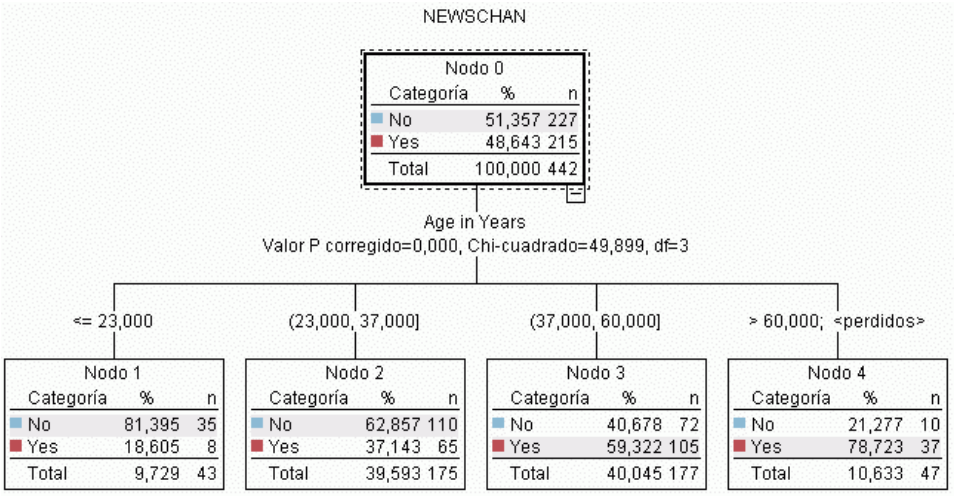
▶ Para podar una rama, seleccione un nodo y seleccione Eliminar rama para borrar el nodo seleccionado.

▶ Para eliminar el nivel inferior del árbol, seleccione Eliminar un nivel.

▶ 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. Si desea obtener más información, consulte el tema Nodo Árbol C&R el p. 152.

Lectura de reglas divididas en la pestaña Visor

Figura 6-4
Reglas divididas mostradas 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 6-5
Botón Detener ejecución

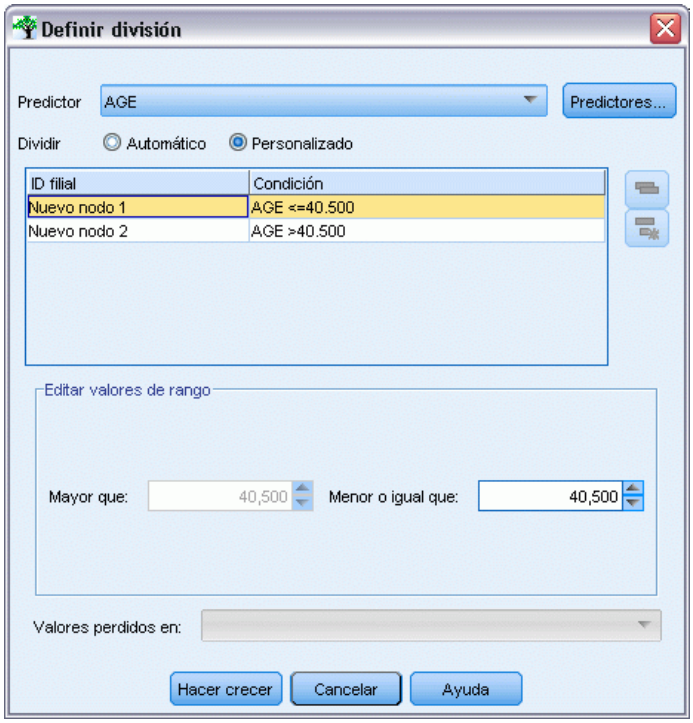
El botón solamente se activa durante la generación del árbol. Detiene la operación de desarrollo en curso en el punto en que se encuentra: se conservan los nodos añadidos, no se guardan cambios, se cierra la ventana, etc. El generador de árboles permanece abierto y permite generar un modelo, actualizar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

Definición de divisiones personalizadas

En el cuadro de diálogo Definir división podrá seleccionar el predictor y especificar las condiciones de las distintas divisiones.

- En el generador de árboles, seleccione un nodo en la pestaña Visor y, en los menús, seleccione:
  Árbol > Hacer crecer rama con división personalizada

Figura 6-6
Cuadro de diálogo Definir división



- Seleccione el predictor que desee en la lista desplegable, o bien, pulse en el botón Predictores para ver detalles acerca de los distintos predictores. Si desea obtener más información, consulte el tema Visualización de detalles de predictores el p. 132.

- Puede aceptar las condiciones por defecto 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 asignen a cada nuevo nodo.

- 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 filial específico. (Con Árbol C&RT y QUEST, los valores perdidos se gestionan mediante sustitutos, tal como se define en el algoritmo.) Si desea obtener más información, consulte el tema Sustitutos y detalles de la división el p. 132.)

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.

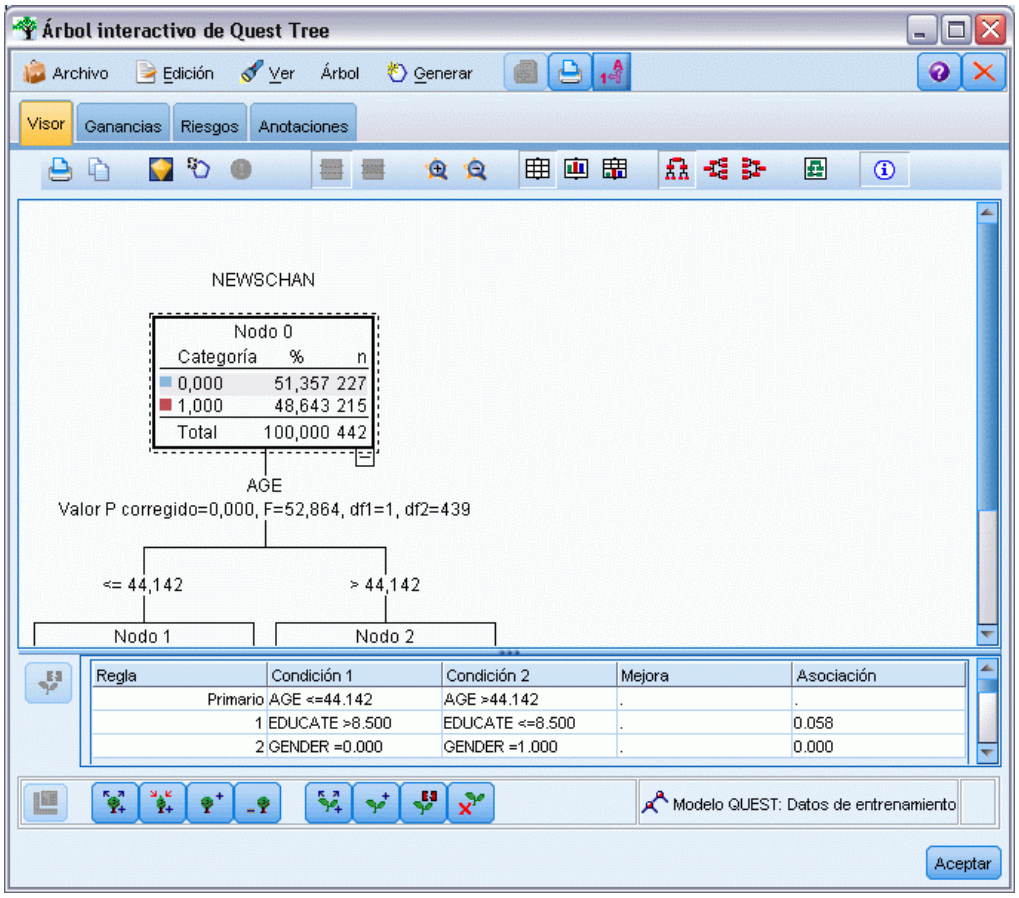
Figura 6-7
Cuadro de diálogo Seleccionar predictor.

- 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. Un estadístico de chi-cuadrado elevado normalmente indica una baja probabilidad, 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.
- En los nodos Árbol 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 parental y filial 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. En los nodos Árbol C&RT, se muestran la mejora y la asociación. La asociación es una medida de la correspondencia entre un sustituto y el campo de división principal, siendo normalmente el "mejor" sustituto aquel que mejor imita al campo de división. En los nodos Árbol C&RT y QUEST, también se mostrarán todos los sustitutos que se hayan utilizado en lugar del predictor principal.

Figura 6-8
Ventana del Generador de árboles interactivos en que se muestra la información de división



- 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 método abreviado, 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 ficha está vacía.

Nota: para que se incluyan 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 filial que tenga un mayor número de registros. Si se esperan valores perdidos durante la comprobación o puntuación, asegúrese de que los

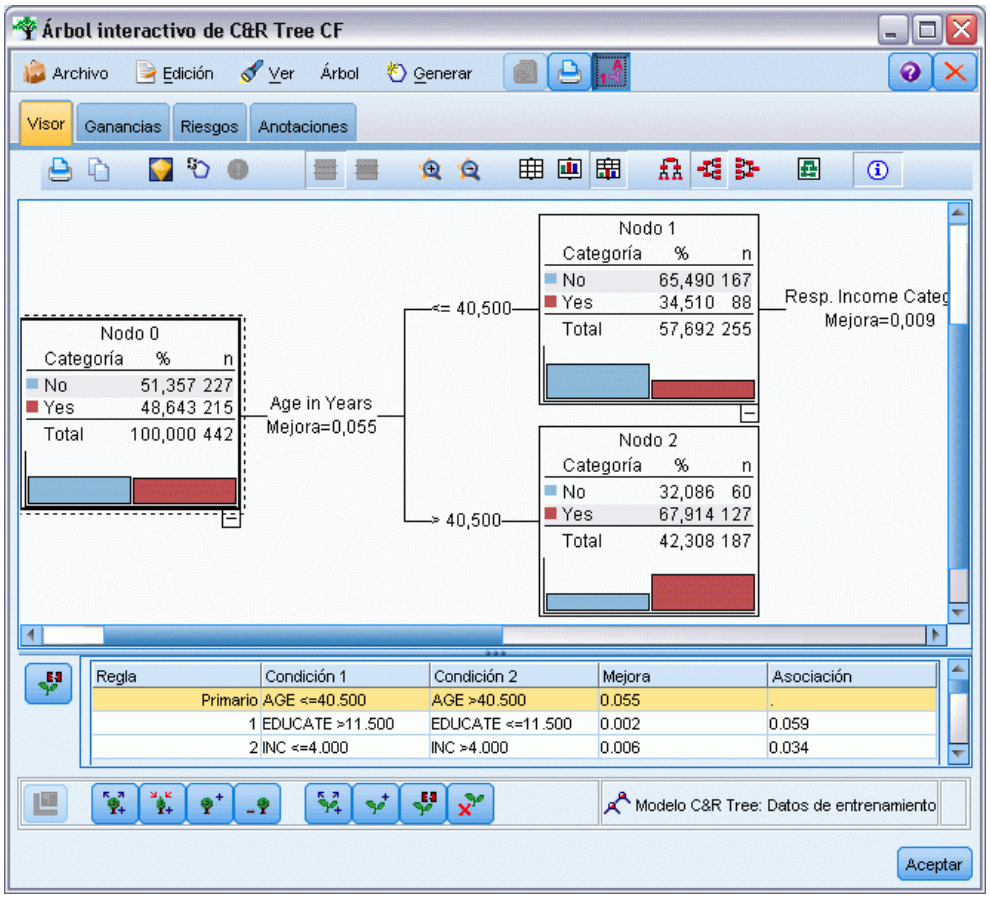
valores no están presentes en la muestra de entrenamiento. No hay sustitutos disponibles para los árboles CHAID.

Aunque con los árboles CHAID no se utilizan sustitutos, podrá asignarlos a un nodo filial específico al definir una división personalizada. Si desea obtener más información, consulte el tema Definición de divisiones personalizadas el p. 130.

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 por defecto, sin embargo, puede expandir y contraer las ramas y personalizar la configuración restante según considere necesario.

Figura 6-9
Vista de izquierda a derecha con detalles de división de regla, etiquetas y gráficos de nodo visibles



- Pulse en el signo menos (–) situado en la esquina inferior izquierda de un nodo parental para ocultar todos sus nodos filiales. Pulse en el signo más (+) situado en la esquina inferior izquierda de un nodo parental para mostrar sus nodos filiales.
- 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. Si desea obtener más información, consulte el tema Sustitutos y detalles de la división el p. 132.
- 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 típica, el número de registros y el valor pronosticado 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

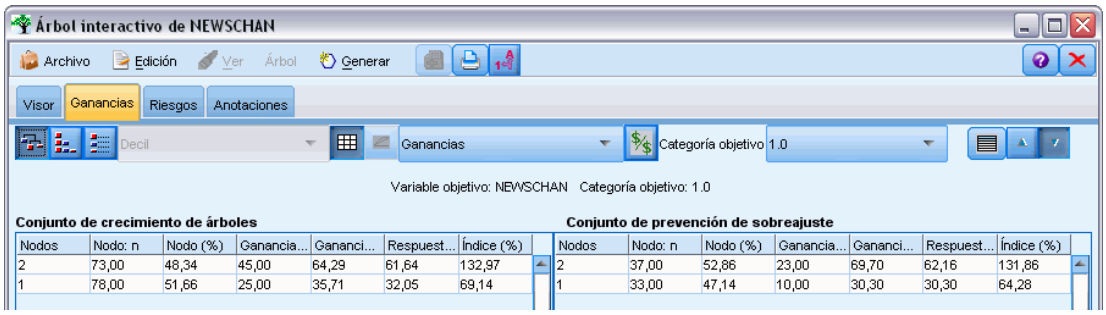
La pestaña Ganancias muestra los estadísticos de todos los nodos terminales de un árbol. Las ganancias proporcionan una medida para considerar la distancia de la media o proporción de un nodo determinado respecto a la media global. Por lo general, cuanto mayor es esta diferencia, más útil resulta el 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.

Figura 6-10
Pestaña Ganancias



La pestaña Ganancias 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 (solamente 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

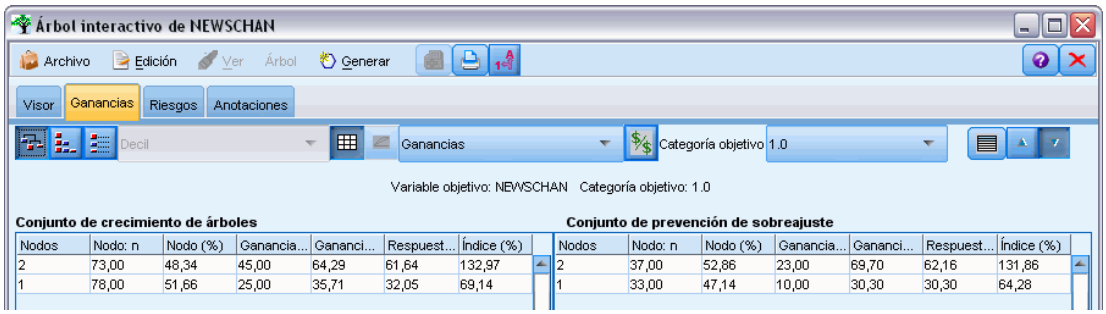
En esta vista, la tabla muestra una fila por cada nodo terminal. Por ejemplo, si la respuesta global a una campaña publicitaria por correo ha sido del 10%, pero el 20% de los registros correspondientes al nodo X ha emitido una respuesta positiva, el índice porcentual del nodo sería del 200%, lo que indica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

Figura 6-11
Estadísticos de ganancias nodo por nodo



Nodos	Nodo: n	Nodo (%)	Ganancia...	Gananci...	Respuest...	Índice (%)
2	73,00	48,34	45,00	64,29	61,64	132,97
1	78,00	51,66	25,00	35,71	32,05	69,14

Nodos	Nodo: n	Nodo (%)	Ganancia...	Gananci...	Respuest...	Índice (%)
2	37,00	52,86	23,00	69,70	62,16	131,86
1	33,00	47,14	10,00	30,30	30,30	64,28

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

Nodo: n. El número total de registros de dicho 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.

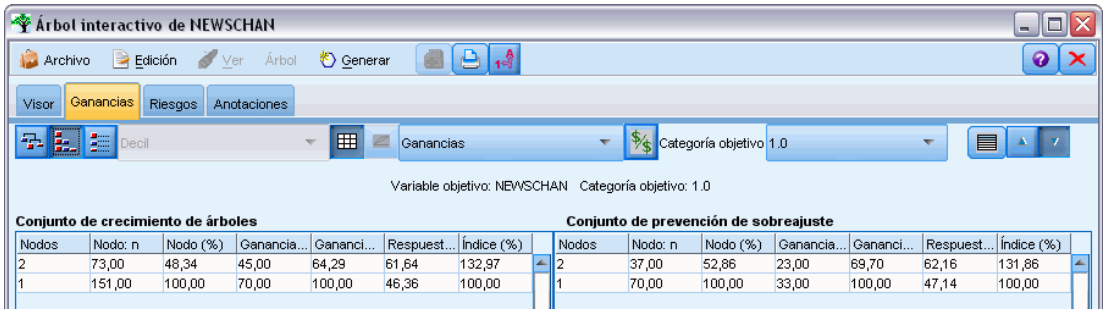
Respuesta (%). El porcentaje de registros del nodo actual correspondiente a la categoría objetivo. En ocasiones, a las respuestas se les denomina “aciertos” en este contexto.

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.

Figura 6-12
Ganancias acumuladas clasificadas en orden descendente por porcentaje de índice



Nodos	Nodo: n	Nodo (%)	Ganancia...	Gananci...	Respuest...	Índice (%)
2	73,00	48,34	45,00	64,29	61,64	132,97
1	151,00	100,00	70,00	100,00	46,36	100,00

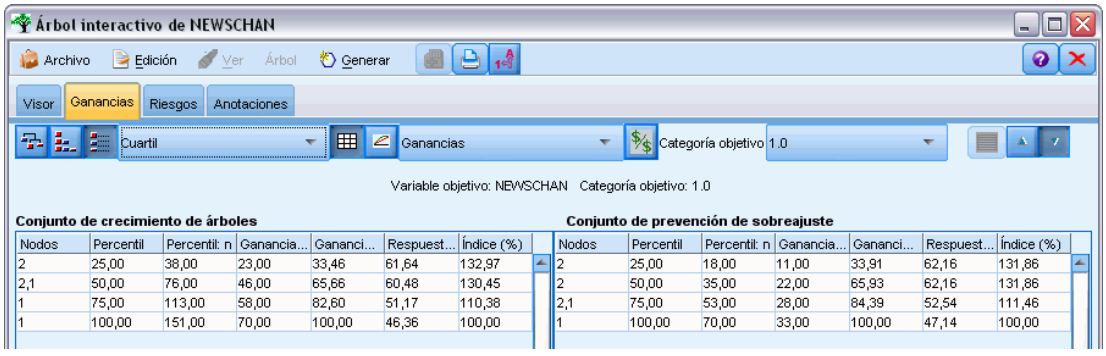
Nodos	Nodo: n	Nodo (%)	Ganancia...	Gananci...	Respuest...	Índice (%)
2	37,00	52,86	23,00	69,70	62,16	131,86
1	70,00	100,00	33,00	100,00	47,14	100,00

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.

Figura 6-13
Ganancias por cuartil dispuestas en orden descendente por porcentaje de índice



Nodos	Percentil	Percentil: n	Ganancia...	Gananci...	Respuest...	Índice (%)
2	25,00	38,00	23,00	33,46	61,64	132,97
2,1	50,00	76,00	46,00	65,66	60,48	130,45
1	75,00	113,00	58,00	82,60	51,17	110,38
1	100,00	151,00	70,00	100,00	46,36	100,00

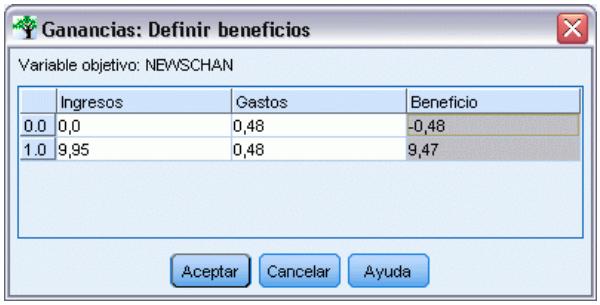
Nodos	Percentil	Percentil: n	Ganancia...	Gananci...	Respuest...	Índice (%)
2	25,00	18,00	11,00	33,91	62,16	131,86
2	50,00	35,00	22,00	65,93	62,16	131,86
2,1	75,00	53,00	28,00	84,39	52,54	111,46
1	100,00	70,00	33,00	100,00	47,14	100,00

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 podrá especificar los ingresos y los gastos de las distintas categorías.

- En la pestaña Ganancias, pulse en el botón Beneficio (con la etiqueta $/$) de la barra de herramientas para acceder al cuadro de diálogo.

Figura 6-14
Cuadro de diálogo Definir beneficios



- Introduzca los valores de beneficios y gastos a las distintas categorías del campo objetivo.

  Por ejemplo, si el envío por correo de una oferta a cada uno de sus clientes tiene un coste de 48 céntimos por cliente y los beneficios de una respuesta positiva son 9,95 dólares por cada suscripción trimestral, entonces cada oferta que se realice *sin respuesta* tendrá un coste de 48 céntimos, mientras que con cada *respuesta afirmativa* se obtendrá un beneficio de 9,47 dólares (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 solamente 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.

- 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.)
- Seleccione el icono de gráficos.

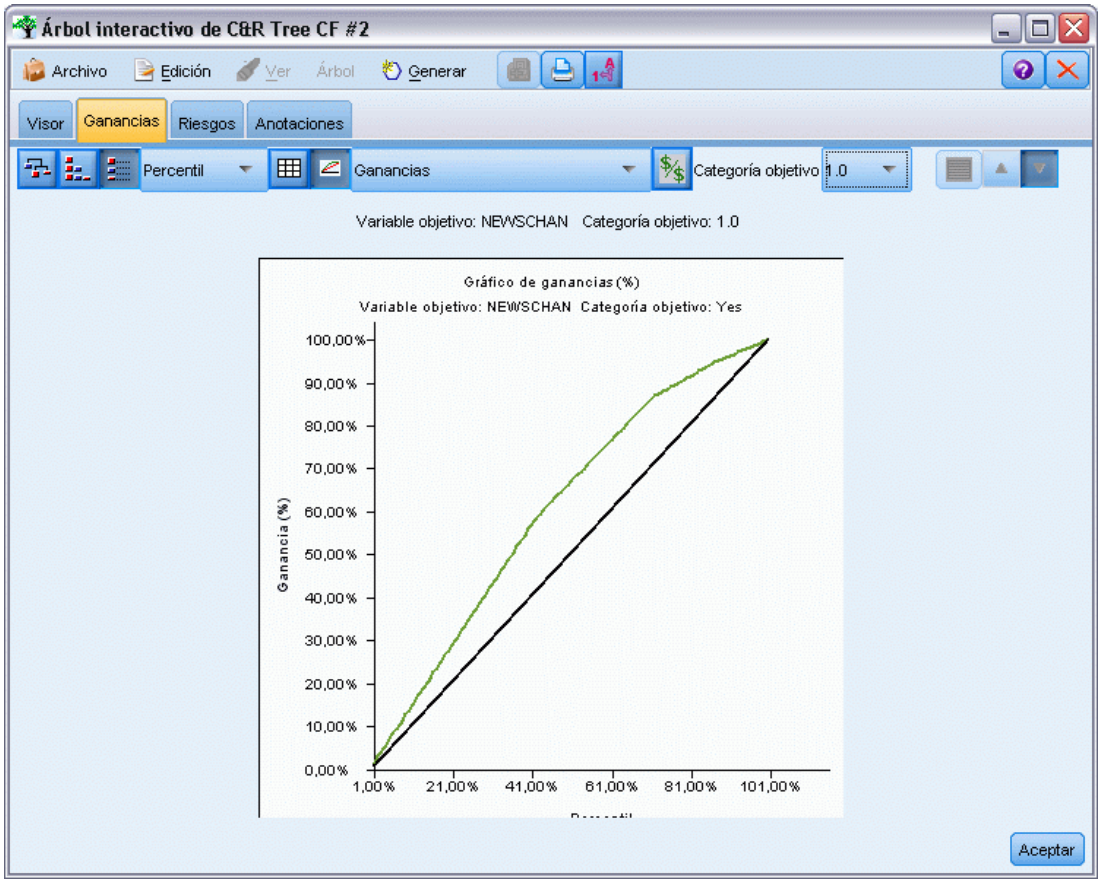
- Seleccione las unidades que desea (percentiles, deciles, etc.) en la lista desplegable.
- Seleccione Ganancias, Respuesta o Elevación para cambiar la medida que se muestra.

Gráfico de 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%

Figura 6-15
Gráfico de ganancias



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)

Figura 6-16
Gráfico de elevación

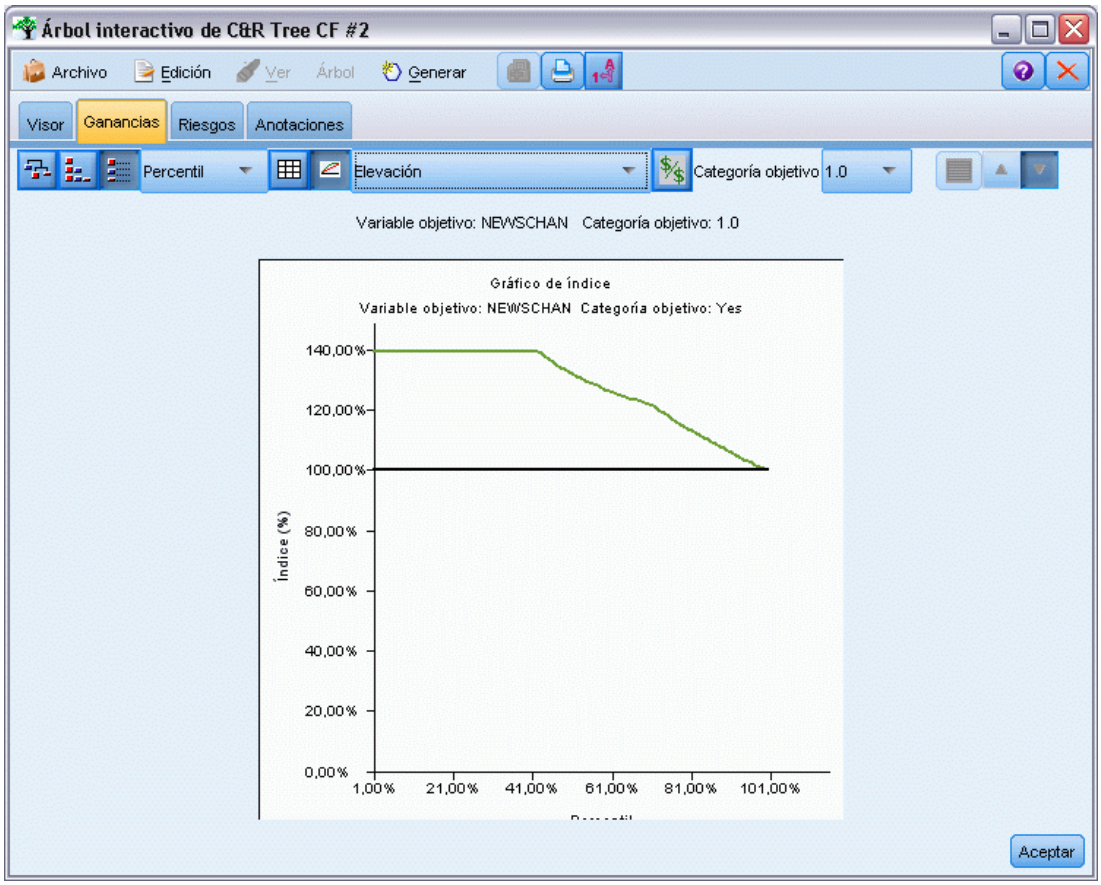
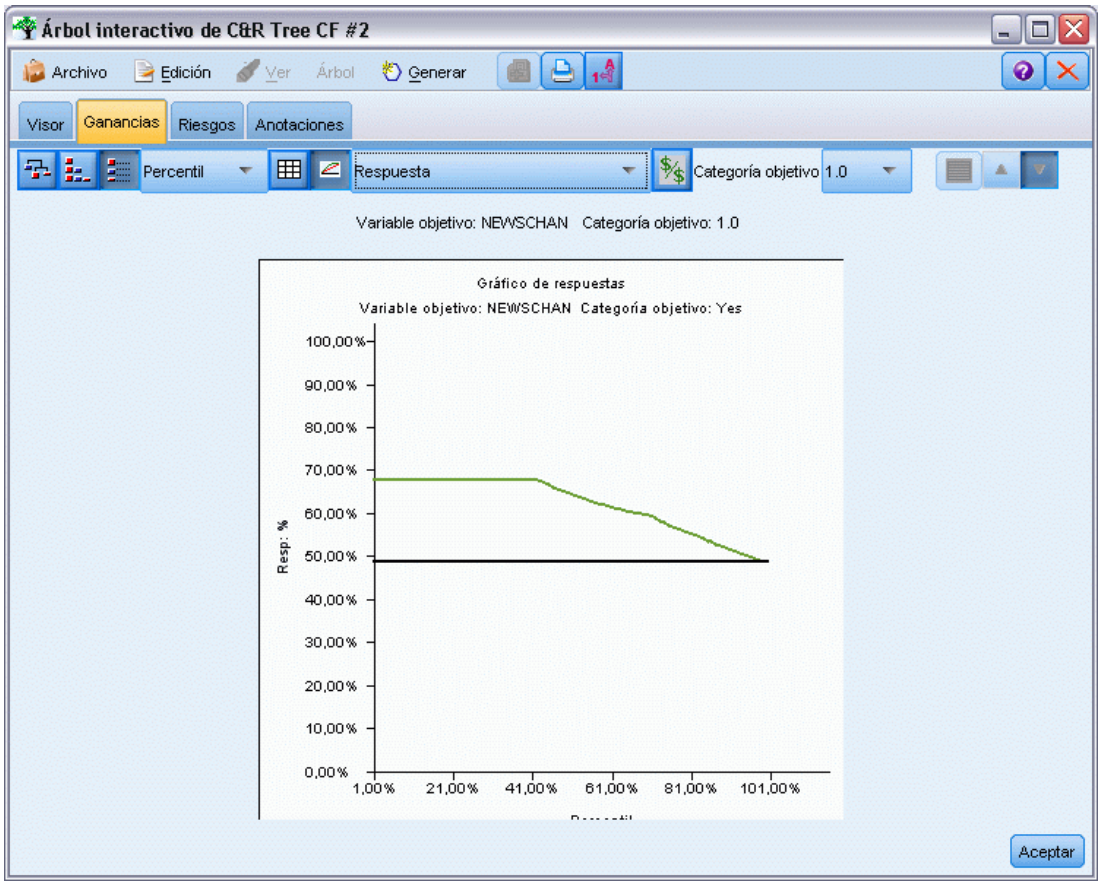


Gráfico de respuestas

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%

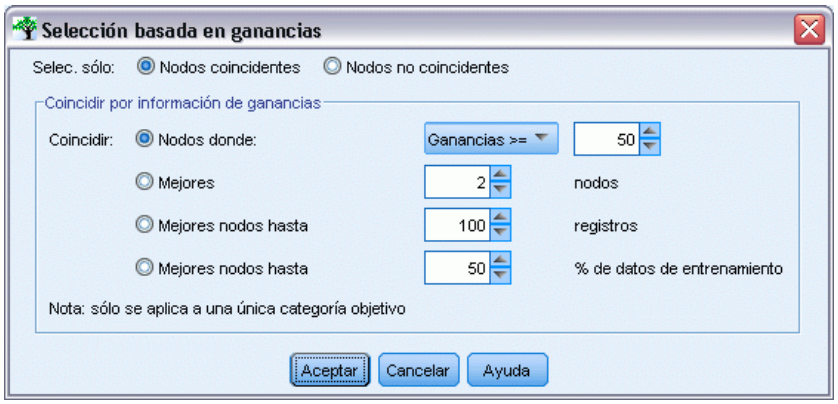
Figura 6-17
Gráfico de respuestas



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.

Figura 6-18
Cuadro de diálogo Selección basada en ganancias



- 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.)
- En la pestaña Ganancias, seleccione en los menús:
  Editar > Seleccionar nodos terminales > Selección basada en ganancias

  Seleccionar solamente. Puede seleccionar nodos coincidentes *o* 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:

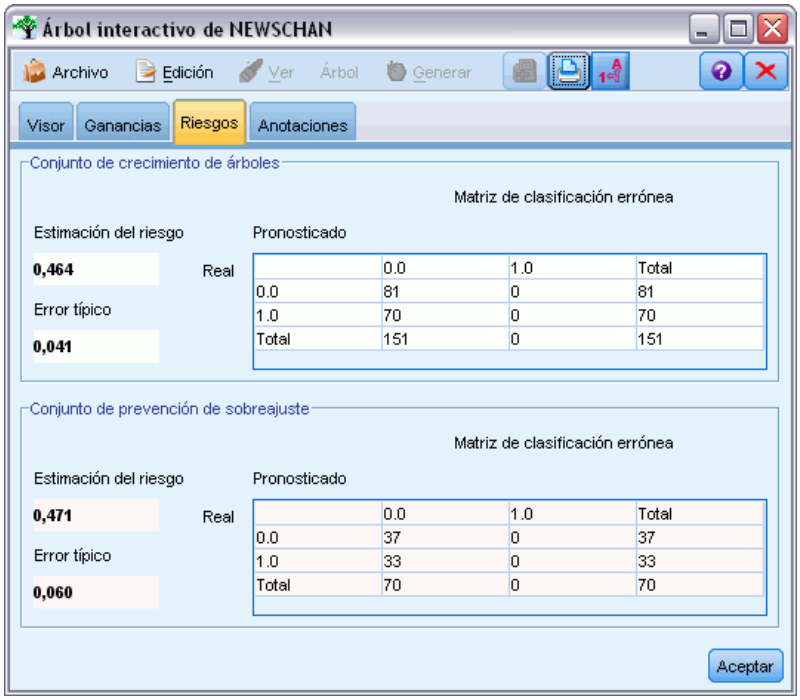
  - Nodos en los que la ganancia, la respuesta o la elevación (índice) coinciden con un umbral especificado, por ejemplo, respuesta mayor o igual que el 50 %.
  - Los *n* mejores nodos basados en la ganancia establecida para la categoría objetivo.
  - Los mejores nodos hasta un número especificado de registros.
  - Los mejores nodos hasta un porcentaje especificado de datos de entrenamiento.
- Pulse en Aceptar para actualizar la selección de la pestaña Visor.
- Para crear un nuevo nodo Seleccionar en función de la selección actual de la pestaña Visor, seleccione Nodo Seleccionar en el menú Generar. Si desea obtener más información, consulte el tema Generación nodos Seleccionar y Filtro el p. 148.

  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.

Figura 6-19
Tabla de errores de clasificación para un objetivo categórico



- En el caso de los pronósticos numéricos, el riesgo es una estimación combinada de la varianza de cada uno de los nodos terminales.
- En el caso de los pronósticos categóricos, 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. Si desea obtener más información, consulte el tema Exportación de la información de riesgo, ganancias y modelo el p. 148.

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

- Generar un filtro o seleccionar un nodo basado en el árbol actual. Si desea obtener más información, consulte el tema Generación nodos Seleccionar y Filtro el p. 148.
- 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. Si desea obtener más información, consulte el tema Generación de un conjunto de reglas desde un árbol de decisión el p. 149.

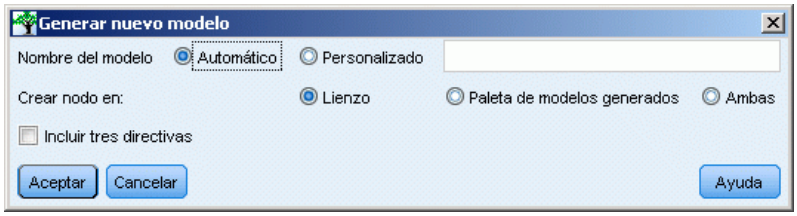
- Además, en el caso de los nugget de árbol generado, puede exportar el modelo en formato PMML. Si desea obtener más información, consulte el tema La paleta de modelos en el capítulo 3 el p. 50. Si el modelo incluye divisiones personalizadas, esta información no se conserva en el PMML exportado. (La 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. Si desea obtener más información, consulte el tema Generación de gráficos el p. 187.

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

Figura 6-20
Generación de un modelo de árbol de decisión



Puede elegir entre las siguientes opciones:

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. Si desea obtener más información, consulte el tema Directivas de desarrollo de árboles el p. 145.

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. Si desea obtener más información, consulte el tema Actualización de directivas de árbol el p. 148. También puede editar las directivas de forma manual, aunque se debe 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,

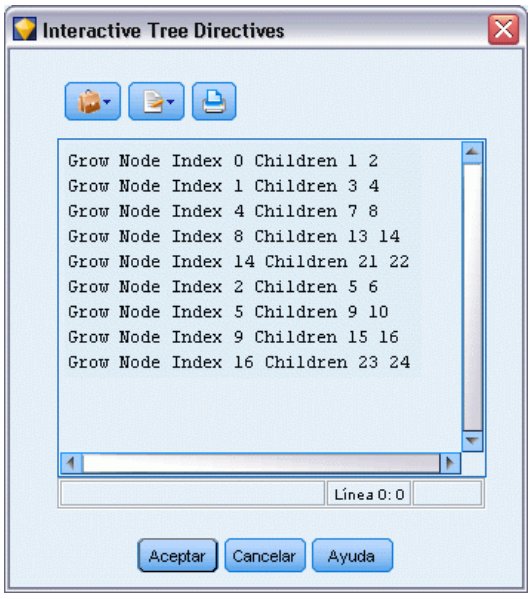
si el algoritmo CHAID cambia una división de dos factores a otra de tres basándose en los datos actualizados, podrían generarse errores en cualquiera de las directivas basadas en la división de dos factores anterior.

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

Edición de directivas

- 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.
- Seleccione Iniciar sesión interactiva para activar los controles, Utilizar directivas de árbol y, a continuación, Directivas.

Figura 6-21
Directivas de desarrollo de árboles



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 (Index 0) y que es necesario especificar los valores de índice para ambos filiales (en este caso, 1 y 2). No es válido especificar Grow Node Index 2 Children 3 4 a menos que en primer lugar haga crecer la raíz que ha creado el nodo 2.

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.) distinguen entre mayúsculas y 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 Árbol C&R 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 filiales:

```
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 filiales que podrían generar errores en la segunda línea de la sintaxis.

Uso de directivas en procesos

Las directivas también pueden incrustarse en procesos mediante comillas triples. Si desea obtener más información, consulte el tema Bloques de texto literal en el capítulo 3 en *Manual de procesos y automatización de IBM SPSS Modeler 15*.

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:
  File > 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. Si desea obtener más información, consulte el tema Directivas de desarrollo de árboles el p. 145.

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.

- En la ventana del generador de árboles, seleccione la pestaña o la vista que desea exportar.
- Seleccione en los menús:
  File > Exportar
- 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 cuantil, acumulados 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 solamente 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 Filtro

  o

  > Nodo Seleccionar

Nodo Filtro 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 papel *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

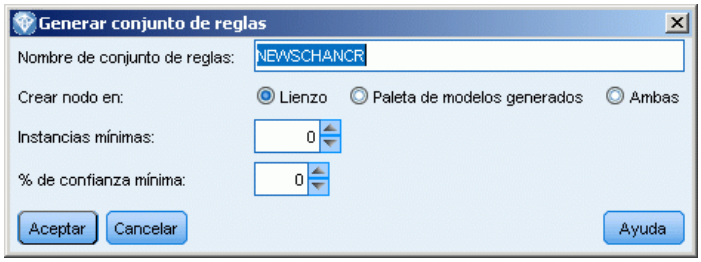
Puede generar un nugget de modelo Conjunto de reglas que representa la estructura del árbol como un conjunto de reglas y define las ramas terminales del árbol. Por lo general, los conjuntos de reglas pueden retener la mayor parte de la información significativa de un árbol de decisión completo, aunque 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 pronostican un resultado *negativo*, seguidas de todas aquellas que pronostican un resultado *afirmativo*. Al aplicar varias reglas, cada una de ellas obtiene un "voto" ponderado basado en la confianza que se asocia a dicha regla. El pronóstico 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 un pronóstico por defecto.

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

Figura 6-22
Cuadro de diálogo Generar conjunto de reglas



Nombre de conjunto de reglas. Permite especificar 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.

Ocurrencias 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.

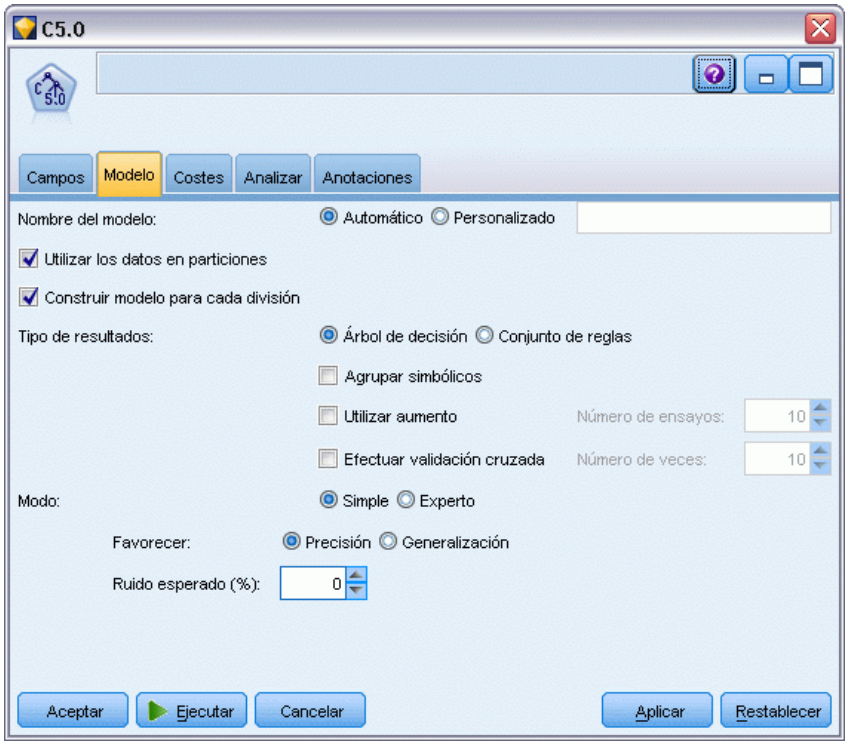
Confianza mínima. Especifique la confianza mínima para las reglas que desea conservar en el nugget de modelo Conjunto de reglas. El nuevo conjunto de reglas no incluirá las reglas con un valor de confianza inferior al especificado.

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. Este es el único método que puede utilizarse en el caso de los modelos de árboles C5.0, que no se admiten en el Generador de árboles interactivos.

- Cree una ruta y añada uno de los nodos de árboles de decisión Árbol C&R, CHAID, QUEST o C5.0.

Figura 6-23
Creación directa de un árbol C5.0



- En Árbol C&R, QUEST o CHAID, en el panel Objetivo de la pestaña Opciones de generación, seleccione uno de los objetivos principales. Si selecciona Crear un árbol único, asegúrese de que el modo está definido a Generar modelo.

  En C5.0, en la pestaña Modelo, defina Tipo de resultado a Árbol de decisión.

- 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.

- Ejecute la ruta para generar el modelo.

Comentarios

- 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 algoritmos de generación de árboles introducidos anteriormente:

- Árbol C&R
- QUEST
- CHAID
- C5.0

Si desea obtener más información, consulte el tema Modelos de árboles de decisión el p. 125.

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 de cualquier de los siguientes tipos (niveles de medición): continuos, categóricos, de marca, nominales u ordinales. Si desea obtener más información, consulte el tema Niveles de medida en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Campos objetivo. Solamente es posible especificar un campo objetivo. Para el Árbol C&R y CHAID, 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.Los nodos Árbol C&R y QUEST solamente admiten divisiones binarias (es decir, cada nodo del árbol se puede dividir en no más dos ramas). Por contra, CHAID y C5.0 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 (por defecto, el coeficiente Gini, aunque se puede modificar). En el caso de objetivos continuos, se utiliza el método de desviación cuadrática mínima. CHAID utiliza 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 diferente y permite utilizarlos en la generación del 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. Permite crear un nivel de árbol cada vez, editar las divisiones y podar el árbol antes de crear el modelo. C5.0 no tiene una opción interactiva.

Probabilidades previas. Árbol C&R y QUEST admiten la especificación de probabilidades previas de categorías al pronosticar 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 y C5.0 no admiten la especificación antes de las probabilidades previas.

Conjuntos de reglas. 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. Si desea obtener más información, consulte el tema Generación de un conjunto de reglas desde un árbol de decisión el p. 149.

Nodo Árbol C&R

El nodo Árbol de clasificación y regresión (C&R) es un método de pronóstico 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&R 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 pronosticar 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. Si desea obtener más información, consulte el tema Nodo Reclasificar en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

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. Si desea obtener más información, consulte el tema Nodo Reclasificar en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

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.

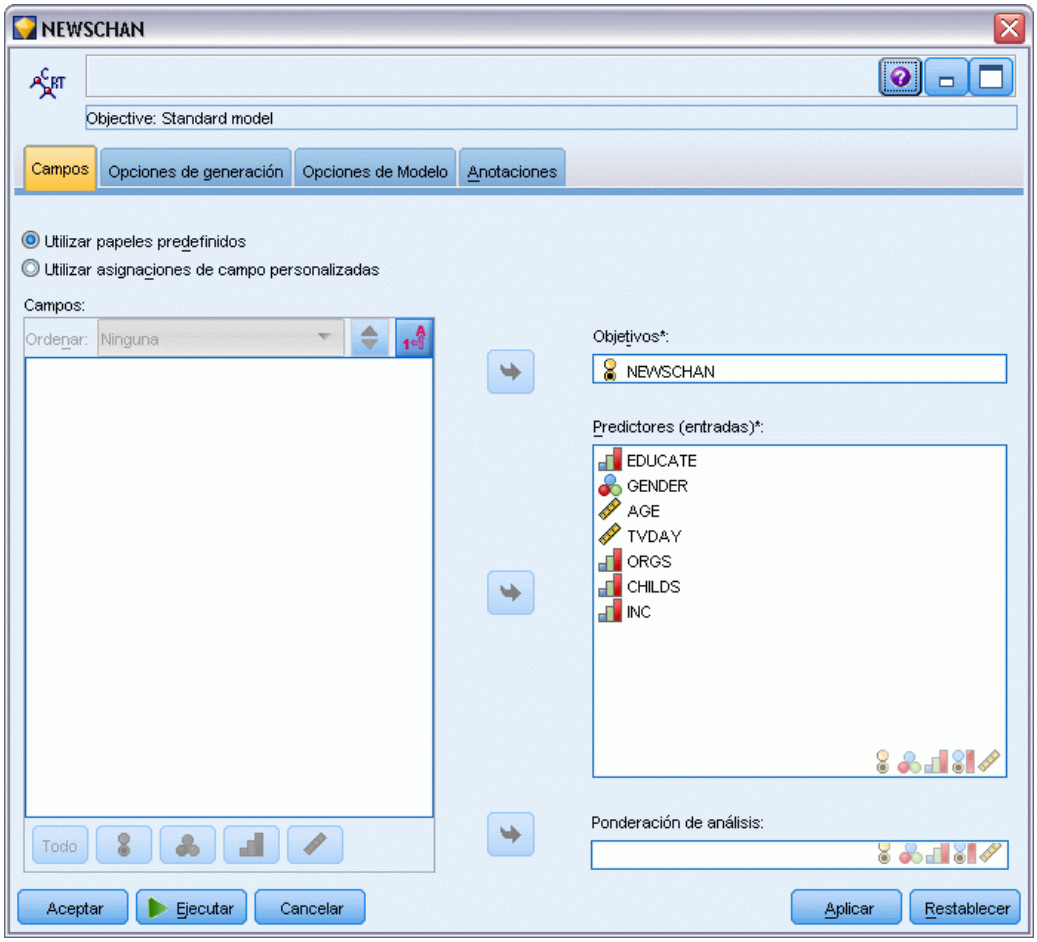
Requisitos. Los campos de entrada pueden ser continuos (rango numérico), sin embargo el campo objetivo debe ser categórico. Todas las divisiones son binarias. No podrá utilizar los campos de ponderación. Todos los campos ordinales (conjunto ordenado) utilizados en el modelo deben disponer de almacenamiento numérico (no en cadenas). Si lo considera necesario, utilice a continuación el nodo Reclasificar para realizar las conversiones. Si desea obtener más información, consulte el tema Nodo Reclasificar en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

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 el campo definición de papeles ya definidos en nodos anteriores o realizando las asignaciones de campos manualmente.

Figura 6-24
Nodo Árbol C&R, pestaña Campos



Utilizar papeles predefinidos. Esta opción utiliza las definiciones de papeles (objetivos, predictores, etcétera) desde un nodo Tipo anterior (o la pestaña Tipo de un nodo de origen anterior). Si desea obtener más información, consulte el tema Definición del papel de campos en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Utilizar asignaciones de campos personalizadas. Seleccione esta opción si desea asignar objetivos, predictores y otros papeles manualmente en esta pantalla.

Campos. Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los diferentes campos de papeles en la parte derecha de la pantalla. Los iconos indican los niveles de medidas diferentes de cada campo.

Pulse en el botón Todos para seleccionar todos los campos de la lista o pulse en un botón de nivel de medida individual para seleccionar todos los campos con ese nivel de medida.

Destino. Seleccione un campo como el destino de la predicción.

Predictores (Entradas). Seleccione uno o más campos como entradas de la predicción.

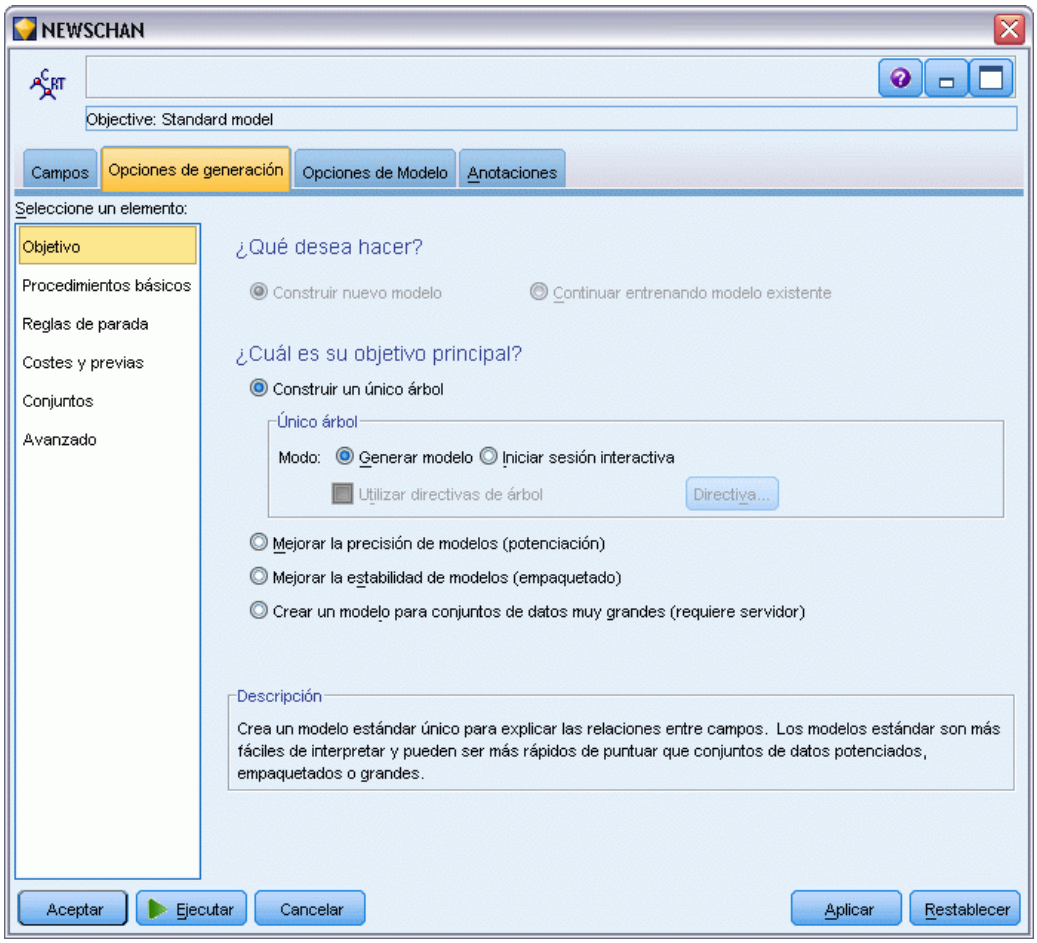
Ponderación de análisis. (CHAID y C&RT únicamente) Para utilizar un campo como ponderación de caso, especifique el campo aquí. Las ponderaciones de caso se utilizan para contabilizar las diferencias existentes en la varianza entre los niveles del campo de salida. Si desea obtener más información, consulte el tema Uso de campos de frecuencia y ponderación en el capítulo 3 el p. 41.

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 por defecto, pero normalmente querrá personalizar la generación de sus tareas.

Puede seleccionar si desea generar un nuevo modelo o actualizar uno existente. También puede definir el objetivo principal del nodo: Para generar un modelo estándar, para crear uno con una precisión o estabilidad mejorada o para generar uno para utilizarlo con conjuntos de datos de grandes dimensiones.

Figura 6-25
Nodo Árbol C&R, pestaña Opciones de generación



¿Qué desea hacer?

Crear modelo nuevo. (por defecto) Crea un nuevo modelo completamente nuevo cada vez que ejecute una ruta con este nodo de modelado.

Continuar entrenando modelo existente. Por defecto, 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 solamente se activa si selecciona como objetivo Crear un modelo para conjuntos de datos muy grandes.

¿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.

  Moda. 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 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. Si desea obtener más información, consulte el tema Actualización de directivas de árbol el p. 148.

- **Mejorar la precisión del modelo (aumento).** Seleccione esta opción si desea utilizar un método especial, conocido como **aumento**, para mejorar el índice 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 los distintos pronósticos en un pronóstico global. El aumento puede mejorar significativamente la precisión del modelo de árbol de decisión, aunque también precisa de un entrenamiento más largo.

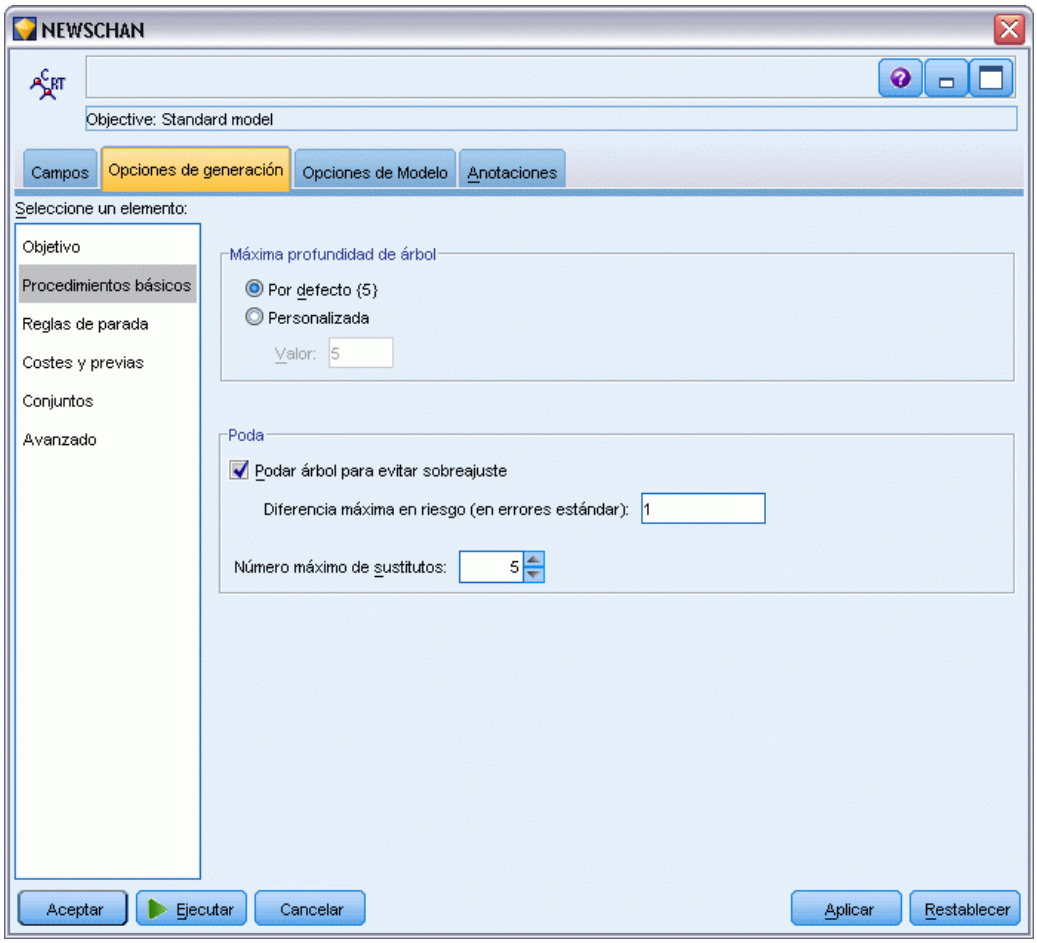
- **Mejorar la estabilidad del modelo (agregación autodocimante).** Seleccione esta opción si desea utilizar un método especial, conocido como **agregación autodocimante**, para mejorar la estabilidad del modelo para evitar sobreajustes. Esta opción crea múltiples modelos y los combina, con objeto de obtener pronósticos 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. Si desea obtener más información, consulte el tema Conexión con IBM SPSS Modeler Server en el capítulo 3 en *Manual de usuario de IBM SPSS Modeler 15*.

Nodos de árbol de decisión: conceptos básicos

Aquí es donde especifica las opciones básicas sobre cómo se crea el árbol de decisiones.

Figura 6-26
Opciones básicas para los árboles de decisión



Algoritmo de desarrollo de árboles. (CHAID únicamente) 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.

Máxima profundidad de árbol. Especifique el número máximo de niveles bajo el nodo raíz (el número de veces que la muestra se dividirá repetidamente). El valor por defecto es 5; seleccione Personalizado e introduzca un valor para especificar un número diferente de niveles.

Poda del árbol (C&RT y QUEST únicamente)

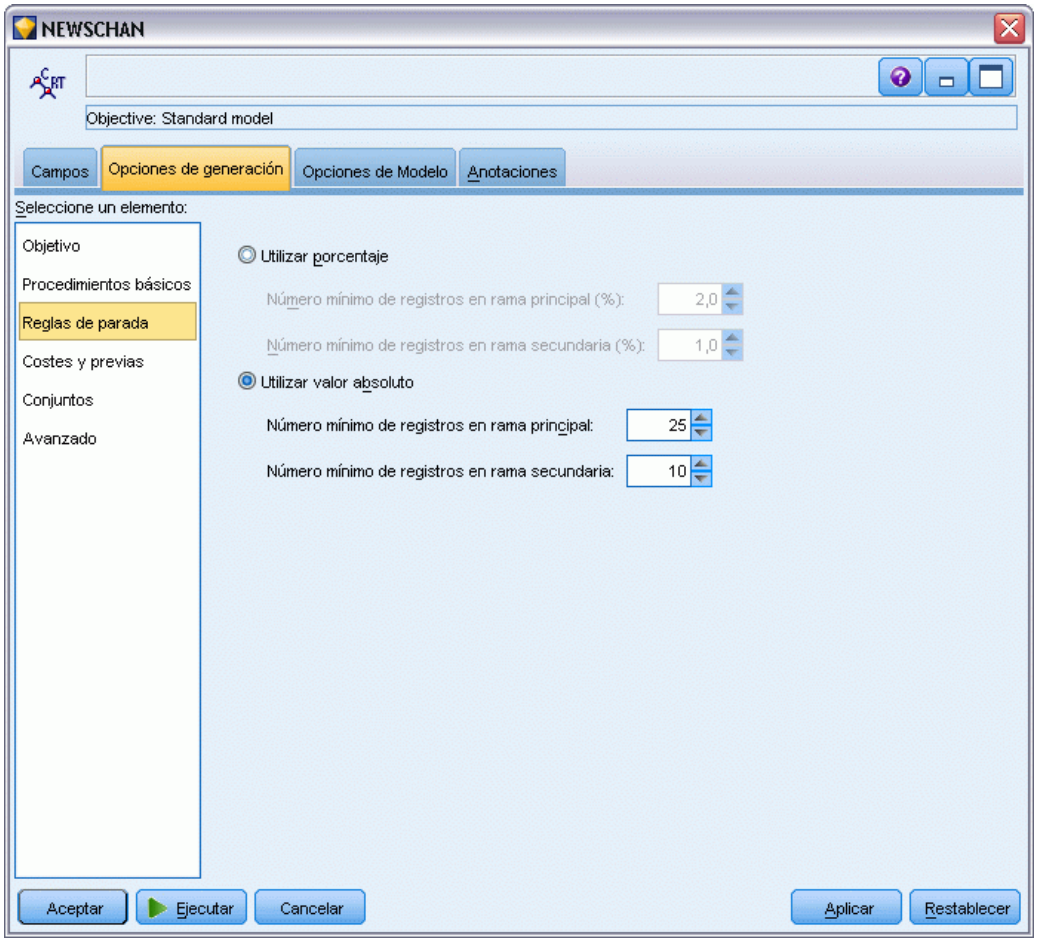
Poda del árbol para evitar sobreajustes. La poda consiste en eliminar las divisiones de nivel inferior que no aportan demasiado a la precisión del árbol. La poda puede ayudar a simplificar un árbol, que resultará más fácil de interpretar y, en determinados casos, mejora la generalización. Deje esta opción sin seleccionar para conservar un árbol completo sin podar.

- **Máxima diferencia en riesgos (en errores típicos).** Permite especificar una regla de poda más general. La regla de error típico permite al algoritmo seleccionar el árbol más simple con una estimación del riesgo próxima (y 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 con estimación del riesgo superior (2 × error típico) a la del árbol completo.

Número 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

Figura 6-27
Opciones para las reglas de parada

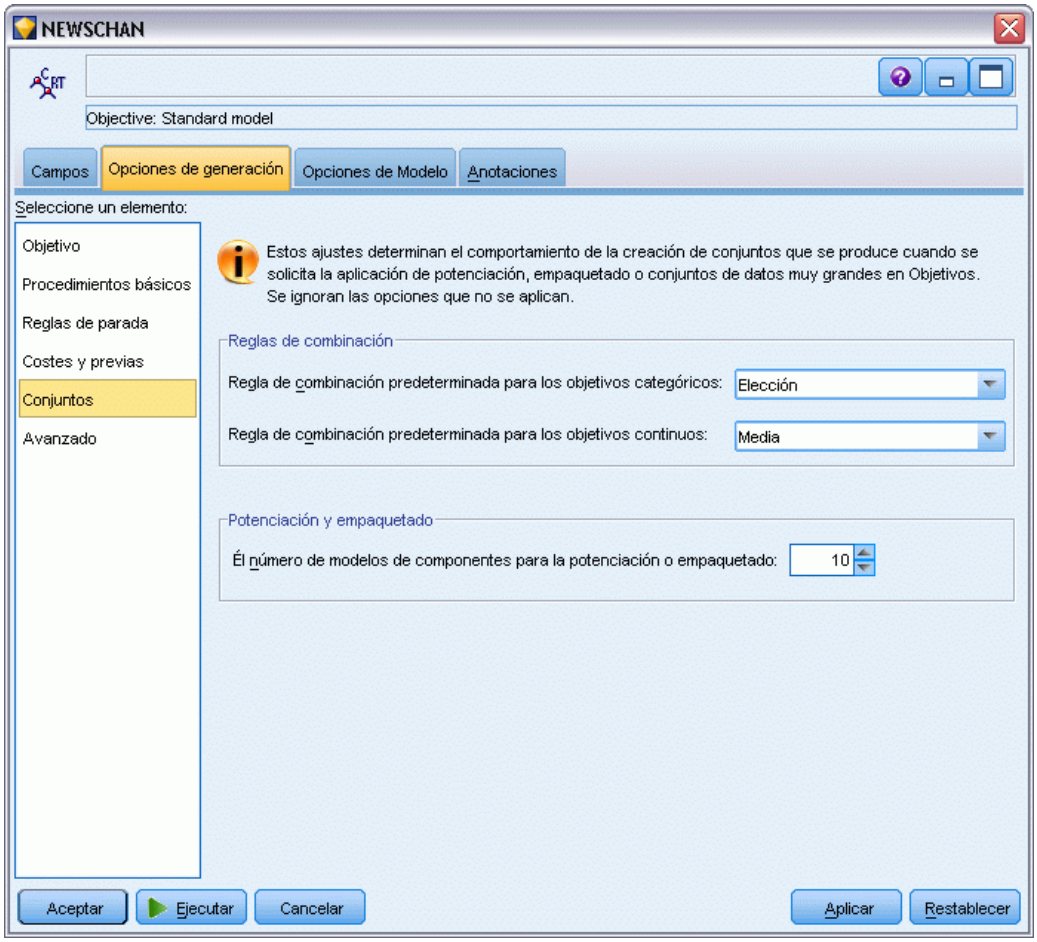


Estas opciones controlan la construcción del árbol. Las reglas de parada determinan cuándo debe detenerse la división de ramas específicas del árbol. Establezca los tamaños de rama mínimos para evitar las divisiones a partir de las cuales se crearían subgrupos muy pequeños. Número mínimo de registros en rama parental impedirá una división cuando el número de registros del nodo que ha de dividirse (**parental**) sea inferior al valor especificado. Número mínimo de registros en rama filial impedirá una división cuando el número de registros de cualquiera de las ramas creadas por la división (**filiales**) resulte inferior al valor especificado.

- **Utilizar porcentaje.** Permite especificar tamaños en términos de un porcentaje de datos de entrenamiento globales.
- **Utilizar valor absoluto.** Permite especificar tamaños como números de registros absolutos.

Nodos de árbol de decisión: Conjuntos

Figura 6-28
Opciones para 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.

Bagging y conjuntos de datos muy grandes. Al puntuar un conjunto, ésta es la regla utilizada para combinar los valores pronosticados a partir de los modelos básicos para calcular el valor de puntuación del conjunto.

- **Regla de combinación predeterminada para objetivos categóricos.** Los valores pronosticados 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.

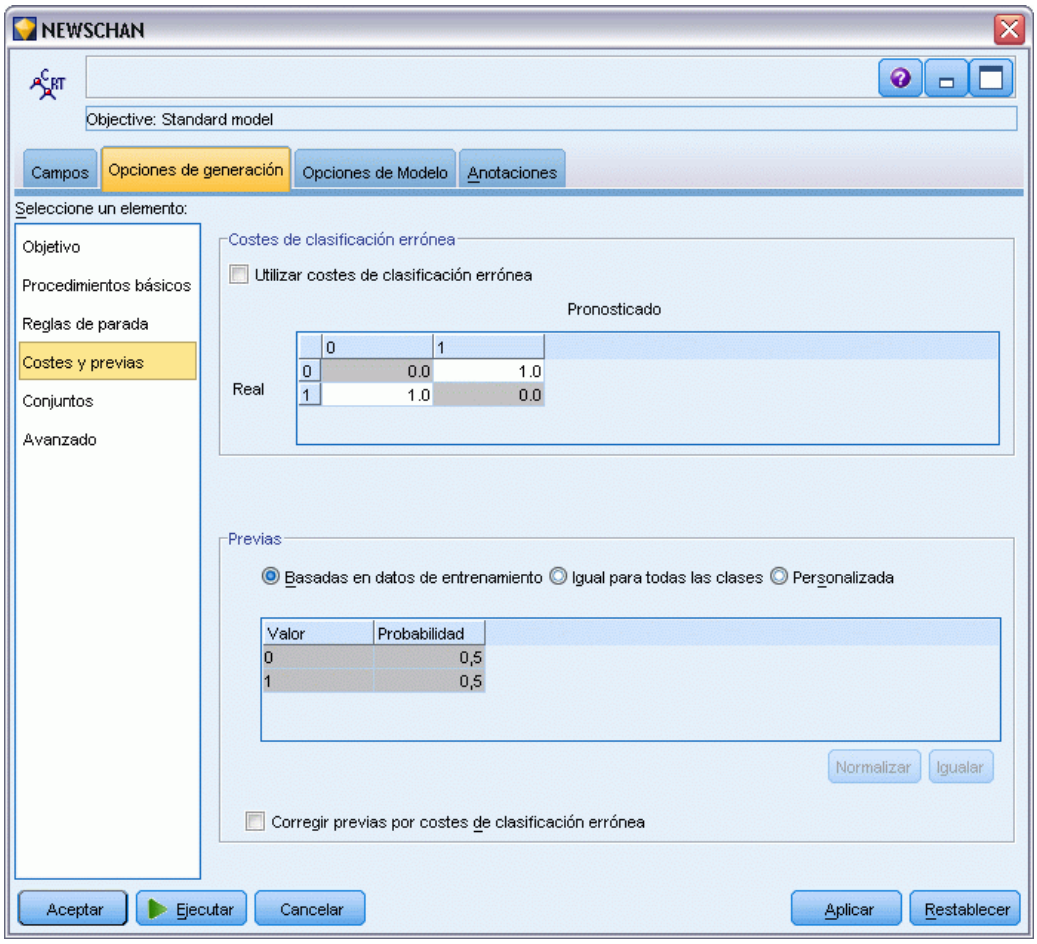
- **Regla de combinación predeterminada para objetivos continuos.** Los valores pronosticados de conjunto para objetivos continuos pueden combinarse mediante la media o mediana de los valores pronosticados a partir de los modelos básicos.

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 autodocimantes. Debe ser un número entero positivo.

Nodos Árbol C&R y QUEST - Costes y Previas

Figura 6-29
Configuración de costes de clasificación errónea y probabilidades 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 pronóstico.

Los costes de clasificación errónea son básicamente ponderaciones aplicadas a resultados específicos. Estas ponderaciones se extraen en el modelo y pueden realmente cambiar el pronóstico (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 ordenan 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 pronosticada y categoría real. Por defecto, 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 pronosticados 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 por defecto 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:

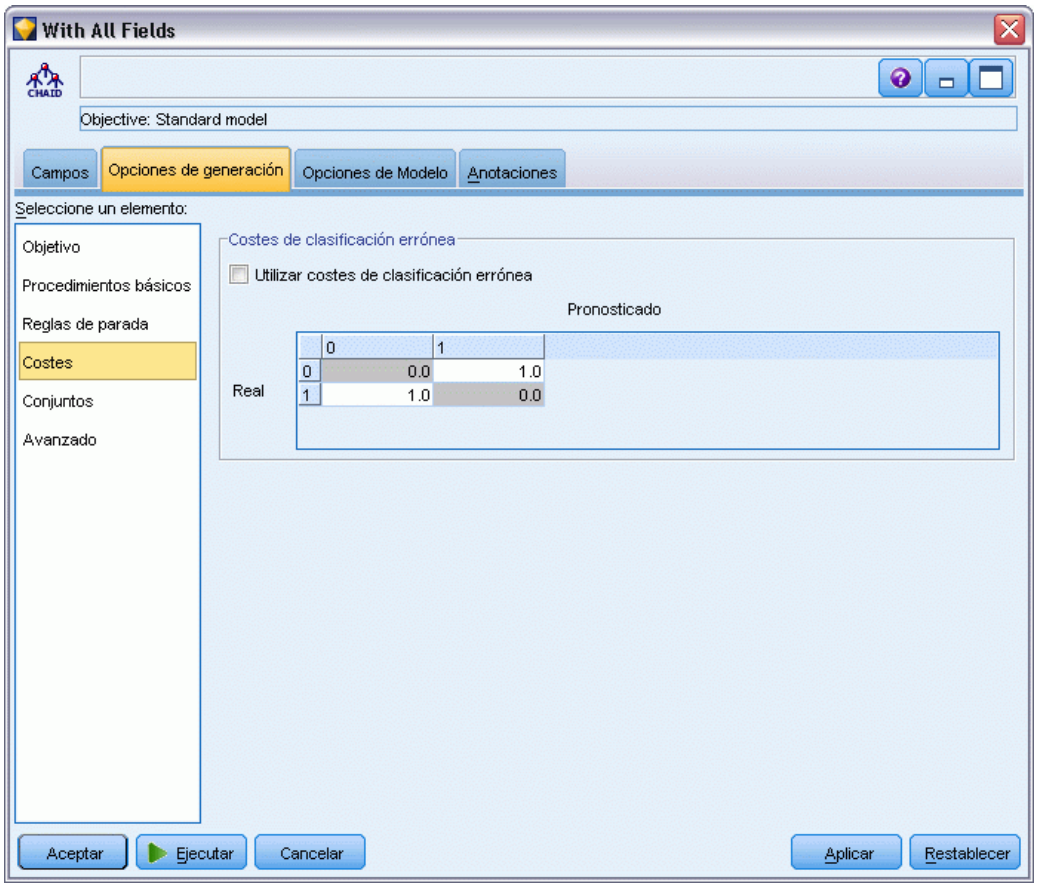
- **Basadas en datos de entrenamiento.** Ésta es la opción por defecto. Las probabilidades previas se basan en las frecuencias relativas de las categorías en los datos de entrenamiento.
- **Igual para todas las clases.** Las probabilidades previas de todas las categorías se definen como $1/k$, donde 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.

Corregir las previas mediante los costes de clasificación errónea. Esta opción 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

Figura 6-30
Costes de clasificación errónea en el nodo CHAID



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 pronóstico.

Los costes de clasificación errónea son básicamente ponderaciones aplicadas a resultados específicos. Estas ponderaciones se extraen en el modelo y pueden realmente cambiar el pronóstico (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 ordenan 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*.

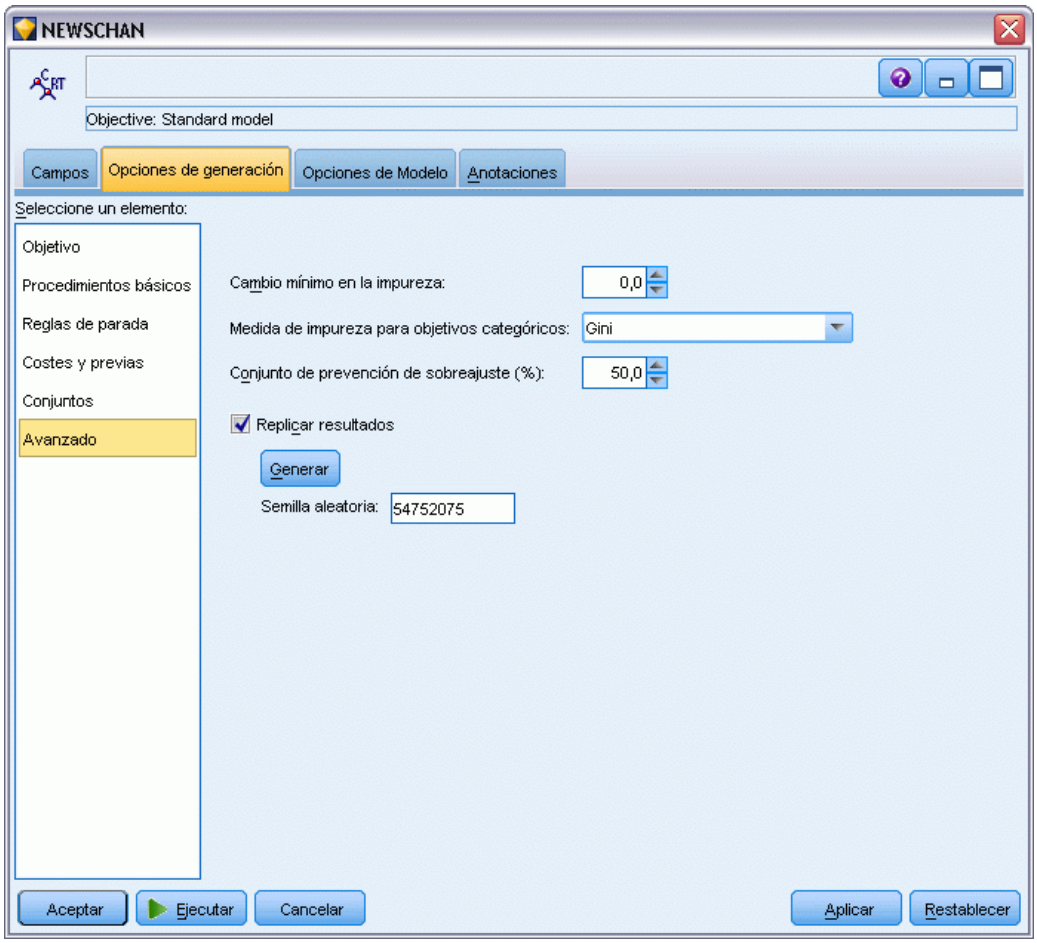
La matriz de costes muestra el coste para cada combinación posible de categoría pronosticada y categoría real. Por defecto, todos los costes de clasificación errónea se establecen en 1,0. Para introducir valores de coste personalizados, seleccione Utilizar costes de clasificación errónea e introduzca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

Para cambiar un coste de clasificación errónea, seleccione la casilla correspondiente a la combinación deseada de valores pronosticados 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 por defecto 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.

Figura 6-31
Configuración de opciones avanzadas para el nodo Árbol C&R



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 por defecto la medida binaria estándar.

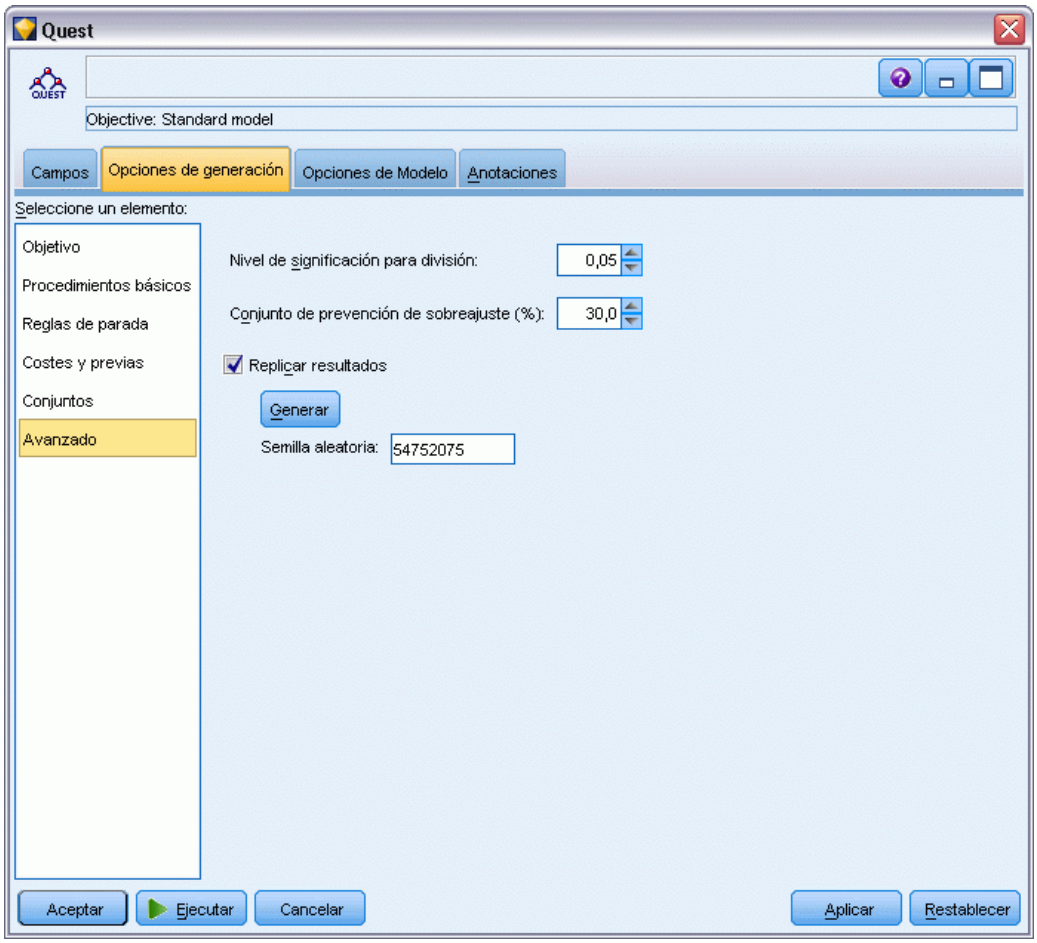
Conjunto de prevención sobreajustado. El algoritmo divide los registros de manera interna en un conjunto de creación de modelos y un conjunto de prevención sobreajustado, el cual es un conjunto independiente de registros de datos utilizado para realizar un seguimiento de errores durante la formación para evitar que el método modele una variación atribuible al azar en los datos. Especifique un porcentaje de registros. El valor predeterminado 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.

Nodo QUEST: Avanzado

Las opciones avanzadas permiten ajustar el proceso de generación de árboles.

Figura 6-32
Configuración de opciones avanzadas para el nodo QUEST



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.

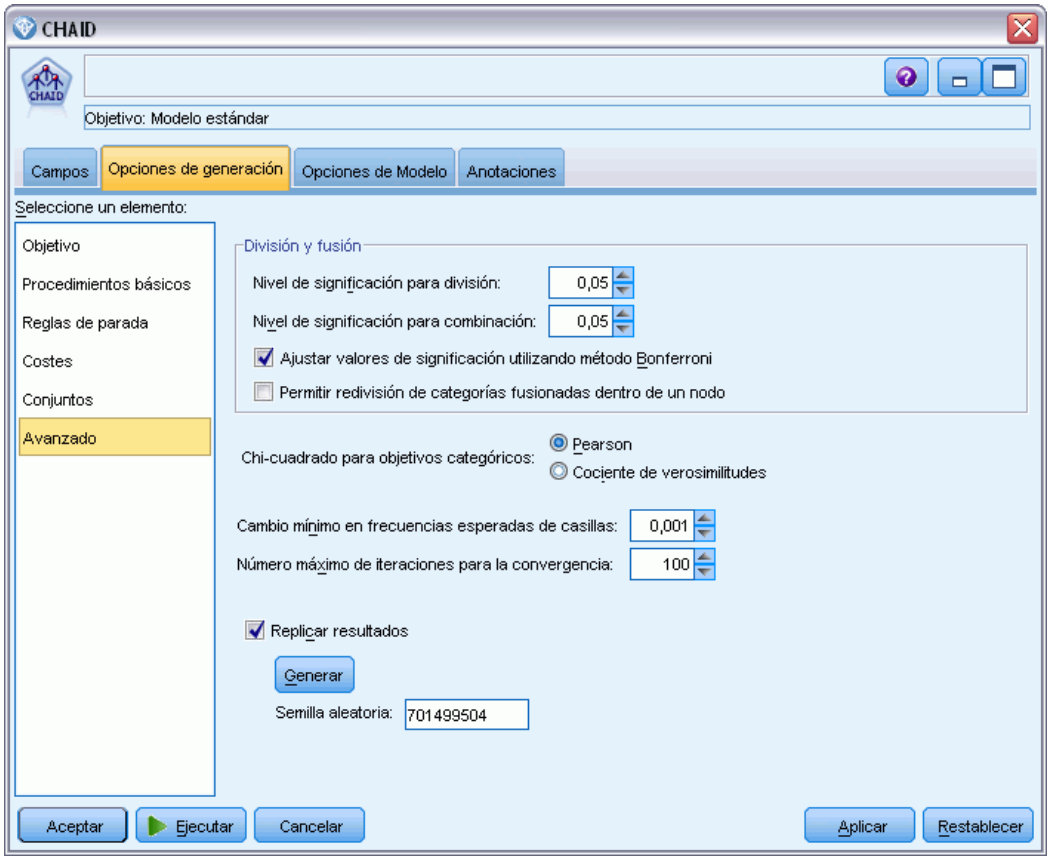
Conjunto de prevención sobreajustado. El algoritmo divide los registros de manera interna en un conjunto de creación de modelos y un conjunto de prevención sobreajustado, el cual es un conjunto independiente de registros de datos utilizado para realizar un seguimiento de errores durante la formación 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 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.

Figura 6-33
Configuración de opciones avanzadas para el nodo CHAID



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 medida 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. Seleccione esta opción para permitir que las categorías fusionadas vuelvan a dividirse si puede considerarse una solución adecuada.

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.

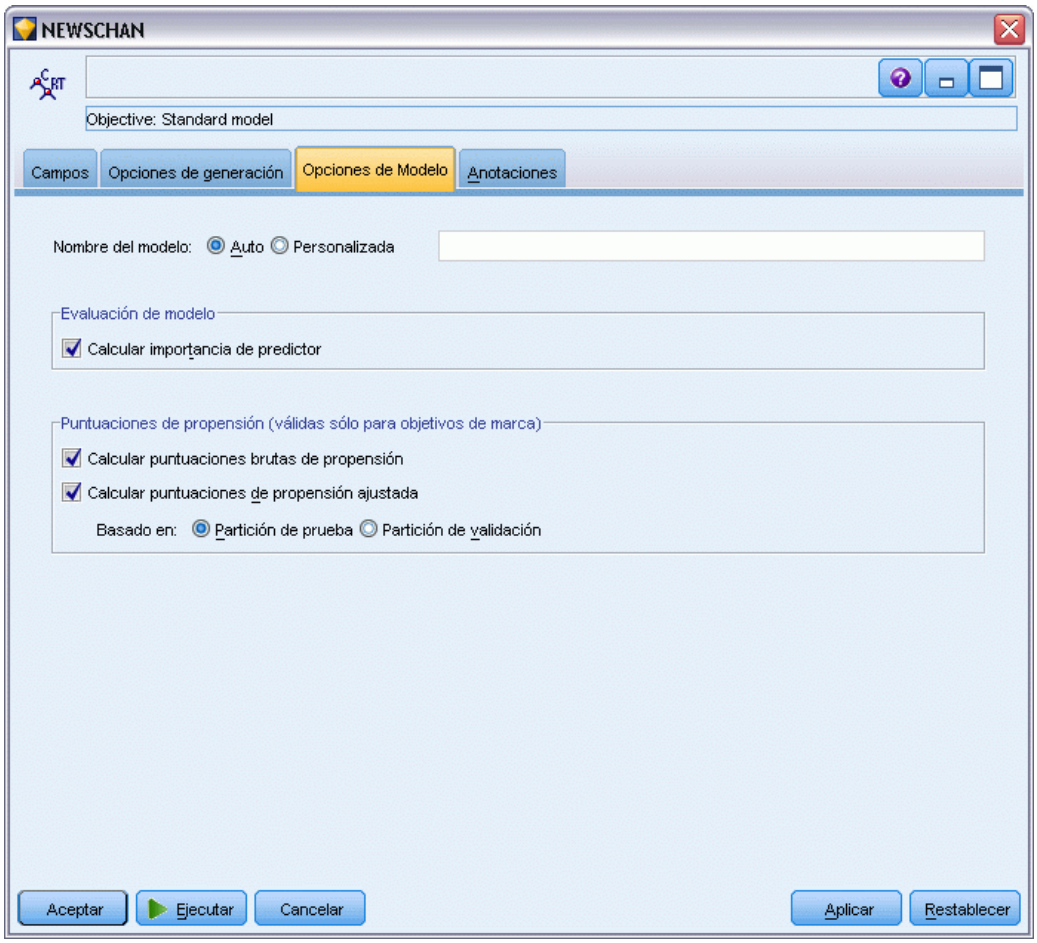
Número máximo de iteraciones para la convergencia. Especifica el número máximo de iteraciones que deben producirse antes de la parada, haya o no tenido lugar la convergencia.

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.

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.

Figura 6-34
Configuración de las opciones de modelo para un nodo de árbol de decisión



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.

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 manera predeterminada. La importancia del predictor no está disponible para modelos de listas de decisiones. Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

Puntuaciones de propensión

Las puntuaciones de propensión pueden activarse en el nodo de modelado y en la ficha Configuración del nugget de modelo. Esta funcionalidad sólo está disponible cuando el objetivo seleccionado es un campo de marca. Si desea obtener más información, consulte el tema Puntuaciones de propensión en el capítulo 3 el p. 45.

Calcular puntuaciones brutas de propensión. Las puntuaciones brutas de propensión 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 del pronóstico. Si el modelo predice el valor false, la propensión se calculará como (1 – P).

- Si selecciona esta opción al crear el modelo, las puntuaciones de propensión se activarán en el nugget de modelo por defecto. Sin embargo, siempre puede activar las puntuaciones brutas de propensión en el nugget de modelo independientemente de si las selecciona o no en el nodo de modelado.
- Al puntuar el modelo, se añadirán puntuaciones brutas de propensión a un campo con las letras *RP* unidas al prefijo estándar. Por ejemplo, si los pronósticos 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. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.
- 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 los pronósticos 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. Si desea obtener más información, consulte el tema Opciones de configuración del nodo Equilibrar en el capítulo 3 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*. 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. Si desea obtener más información, consulte el tema Modelos C5.0 aumentados el p. 186.

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. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Nodo C5.0

Nota: Esta función 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 un pronóstico 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 pronósticos 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. El pronóstico 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 un pronóstico por defecto.

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. Si desea obtener más información, consulte el tema Tratamientos con medicamentos (Gráficos exploratorios/C5.0) en el capítulo 8 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

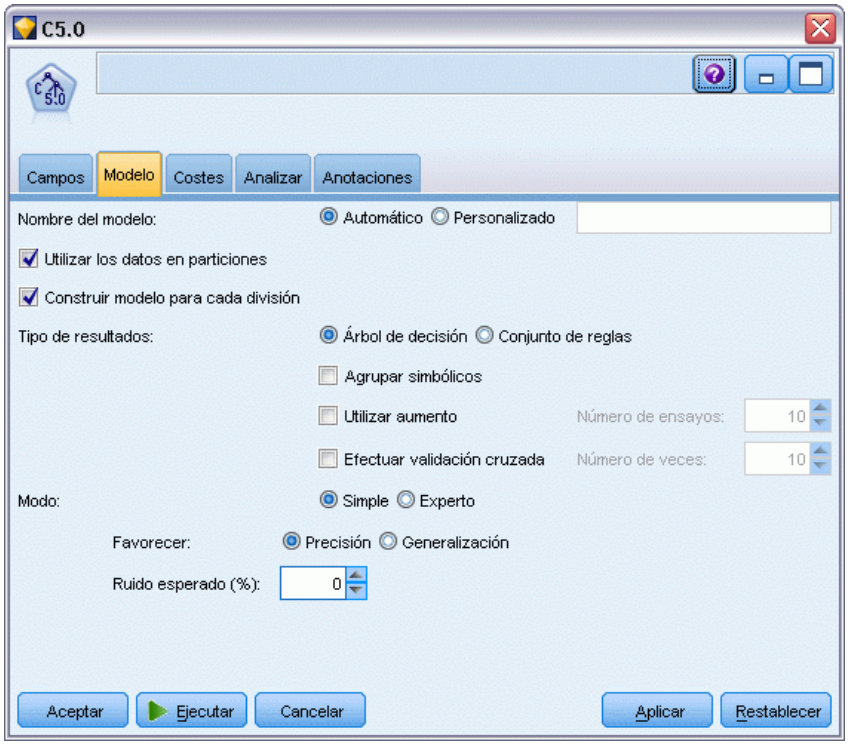
Requisitos. Para entrenar un modelo C5.0, debe existir un campo categórico (por ejemplo, nominal u ordinal) *Objetivo* y uno o más campos *Entrada* de cualquier tipo. Se ignorarán los campos establecidos en *Ambos* o *Ninguno*. Los tipos de los campos utilizados en el modelo deben estar completamente 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. Si desea obtener más información, consulte el tema Configuración de opciones de optimización de las rutas en el capítulo 5 en *Manual de usuario de IBM SPSS Modeler 15*.

Opciones de modelo para el nodo C5.0

Figura 6-35
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 por defecto.
- **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 utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Crear modelos divididos. Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como modelos divididos. Si desea obtener más información, consulte el tema Generación de modelos divididos en el capítulo 3 el p. 32.

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 filial para cada uno de los valores del campo simbólico utilizado para dividir el nodo parental. Por ejemplo, si C5.0 realiza divisiones en un campo *COLOR* (con valores *ROJO*, *VERDE* y *AZUL*), se creará por defecto 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 los distintos pronósticos en un pronóstico 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 función 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.

Moda. 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. C5.0 intentará producir por defecto el árbol lo 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 filial. 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 por defecto es 2. Auméntelo para impedir el **sobreentrenamiento** con los datos con ruido.

Utilizar poda global. Los árboles se podan en dos fases: La primera es una fase de poda local, que examina los subárboles y contraen 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. Por defecto 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. Si desea obtener más información, consulte el tema Configuración de opciones de optimización de las rutas en el capítulo 5 en *Manual de usuario de IBM SPSS Modeler 15*.

Nugget de modelo de árboles de decisión

Los nugget de modelo de árboles de decisión representan las estructuras de árbol para el pronóstico de un determinado campo de salida descubierto por uno de los nodos de modelado de árboles de decisión (C&RT, CHAID, QUEST, C5.0). Se pueden generar tres modelos directamente desde el nodo de generación de árboles o indirectamente desde el Generador de árboles interactivos. Si desea obtener más información, consulte el tema El Generador de árboles interactivos el p. 128.

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 pronosticado y la confianza para cada registro. El pronóstico 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*, el pronóstico para todos los registros asignados a dicho nodo es sí.
- En los árboles de regresión, solamente los valores pronosticados 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 *$R-* para el campo de pronóstico, *$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 pronóstico y *$CC-* para el campo de confianza. Si hay presentes varios nodos del modelo de árbol, los nuevos nombres de campos incluirán números en el *prefijo* para distinguirlos si fuera necesario; por ejemplo, *$R1-* y *$RC1-* y *$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 desde la ventana del generador de árboles.

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

- Generar un filtro o seleccionar un nodo basado en el árbol actual. Si desea obtener más información, consulte el tema Generación nodos Seleccionar y Filtro el p. 148.
- 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. Si desea obtener más información, consulte el tema Generación de un conjunto de reglas desde un árbol de decisión el p. 149.
- Además, en el caso de los nugget de árbol generado, puede exportar el modelo en formato PMML. Si desea obtener más información, consulte el tema La paleta de modelos en el capítulo 3 el p. 50. 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. Si desea obtener más información, consulte el tema Generación de gráficos el p. 187.
- 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. Si desea obtener más información, consulte el tema Modelos C5.0 aumentados el p. 186.

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.

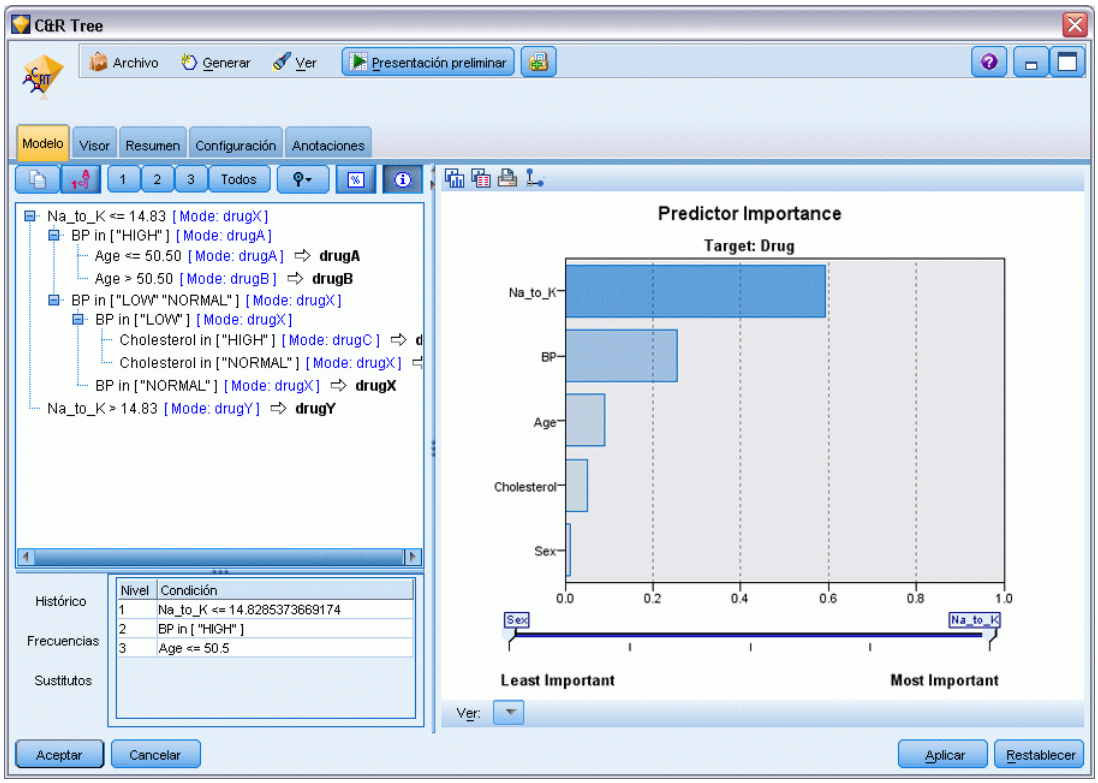
Pestaña	Descripción	Más información
Modelo	Muestra las reglas que definen el modelo.	Si desea obtener más información, consulte el tema Reglas de modelo de árboles de decisión el p. 179.
Visor	Muestra la vista de árbol del modelo.	Si desea obtener más información, consulte el tema Visor de modelo de árboles de decisión el p. 183.
Resumen	Muestra información sobre los campos, los ajustes de versión y el proceso de estimación del modelo.	Si desea obtener más información, consulte el tema Información / Resumen de nugget de modelo en el capítulo 3 el p. 54.
Configuración	Le permite especificar las opciones de confianzas y generación SQL durante la puntuación del modelo.	Si desea obtener más información, consulte el tema Configuración del nugget de modelo de conjunto de reglas/árboles de decisión el p. 184.
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.	Si desea obtener más información, consulte el tema Anotaciones en el capítulo 5 en *Manual de usuario de IBM SPSS Modeler 15*.

Reglas de modelo de árboles 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.

Figura 6-36
Nugget de modelo de árboles de decisión



Reglas de árbol

El panel izquierdo muestra una lista de condiciones que definen la partición de los datos que ha descubierto el algoritmo; dicha lista se compone esencialmente de una serie de reglas que se pueden utilizar para asignar registros individuales a los nodos filiales 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 filiales**, basados en el valor de un determinado campo de entrada. Cada rama filial 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**).

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. Si desea obtener información general sobre el explorador de modelos, consulte Exploración de nugget de modelo.

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 un pronóstico agregando una flecha y el valor pronosticado al final de la condición de la regla. Por ejemplo, una hoja definida por *ingresos* > 100 que pronostica 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 parental. 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:

```
revenue > 100
 region = "Norte"
 region in ["Sur" "Este" "Oeste"]
  revenue <= 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

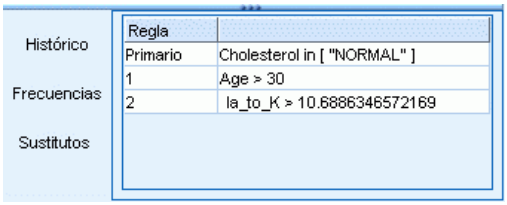
Opcionalmente, en la pestaña Modelo también se puede mostrar un gráfico que indique la importancia relativa de cada predictor cuando se calcule el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los

que importan menos. Tenga en cuenta que este gráfico sólo está disponible si se ha seleccionado Calcular importancia de predictoren la pestaña Analizar antes de generar el modelo. Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

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.

Figura 6-37
Sustitutos en el panel de información



Histórico. 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 ficha está vacía.

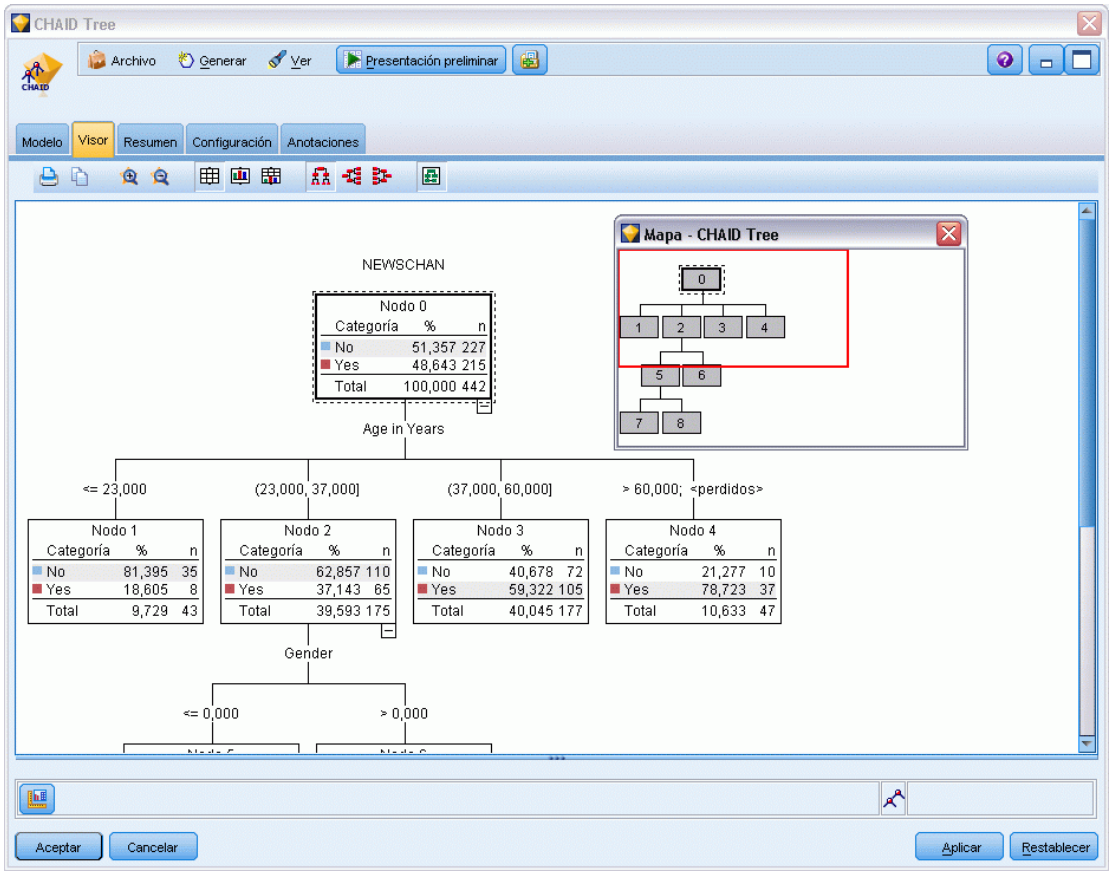
Nota: para que se incluyan en el modelo, los sustitutos se deben identificar durante la fase de entrenamiento. Si la muestra de entrenamiento no tiene valores perdidos, no se identificarán sustitutos. Los registros con valores perdidos que se encuentren durante la comprobación o puntuación pasarán automáticamente al nodo filial 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 árboles de decisión

La pestaña Visor para un nugget de modelo de árboles 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. Si desea obtener más información, consulte el tema Personalización de la vista del árbol el p. 134.

Nota: la pestaña Visor no se muestra en nuggets de modelo de CHAID si selecciona la opción Crear un modelo para conjuntos de datos de grandes dimensiones en el panel Objetivo de la pestaña Opciones de generación.

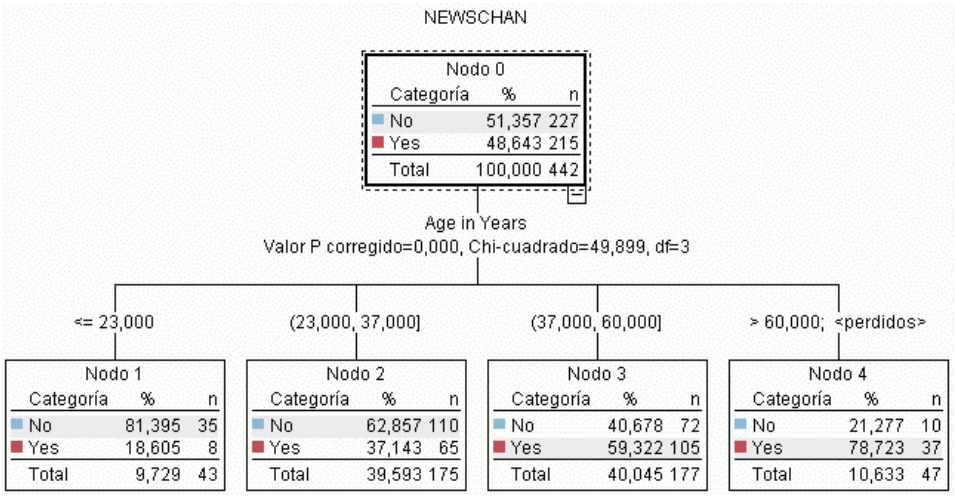
Figura 6-38
Pestaña Visor del árbol de decisión con la ventana del mapa de árboles



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

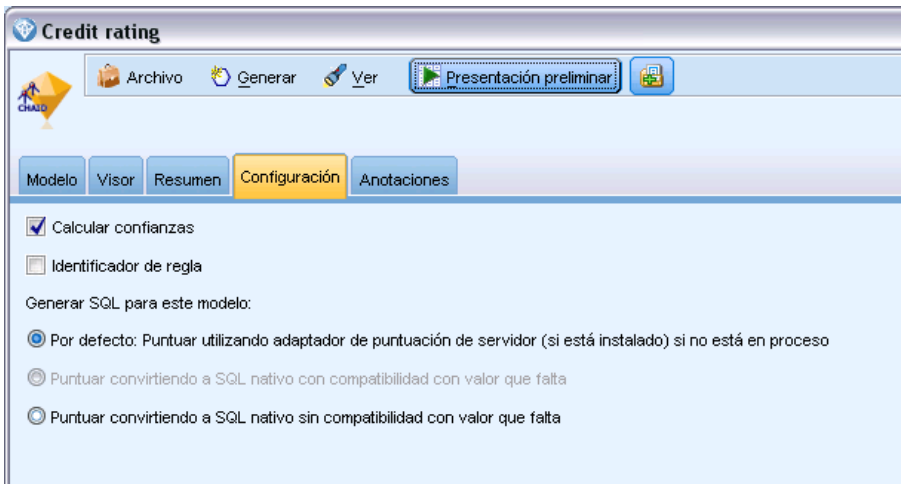
Figura 6-39
Reglas divididas mostradas en la pestaña Visor



Configuración del nugget de modelo de conjunto de reglas/árboles de decisión

La pestaña Configuración de un nugget de modelo de árbol de decisión o conjunto de reglas permite especificar la configuración de confianzas y para la generación de SQL durante la puntuación del modelo. Esta pestaña está solamente disponible después de que el nugget de modelo se haya añadido a una ruta.

Figura 6-40
Ejemplo de la configuración del nugget de modelo de árboles de decisión



Calcular confianzas. Seleccione incluir confianzas al puntuar operaciones. Al puntuar modelos en una 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 brutas de propensión. En el caso de modelos con un objetivo de marca (que devuelve un pronóstico 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 pronóstico y confianza que pueden generarse durante la puntuación.

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 con un objetivo categórico de marcas.

Calcular puntuaciones de propensión ajustada. Las puntuaciones brutas de propensión se basan solamente 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 la pestaña Analizar antes de generar el modelo.

Nota: las puntuaciones ajustadas de propensión no están disponibles para modelos de árbol aumentado y de conjuntos de reglas. Si desea obtener más información, consulte el tema Modelos C5.0 aumentados el p. 186.

Identificador de regla. En los modelos CHAID, QUEST y C&RT, esta opción añade un campo en el resultado de puntuación que indica el ID para el nodo 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. Si desea obtener más información, consulte el tema Optimización de SQL en el capítulo 6 en *Guía de rendimiento y administración de IBM SPSS Modeler Server 15*.

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

- **Por defecto: Puntuar mediante el adaptador de puntuación del servidor (si está instalado) en curso.** Si se conecta a una base de datos con un adaptador de puntuación instalado, se genera SQL con el adaptador de puntuación; de lo contrario, se genera SQL de forma nativa dentro de SPSS Modeler.
- **Generar sin compatibilidad de valores perdidos.** Seleccione esta opción para activar la generación de SQL evitando la gestión de valores perdidos. Esta opción simplemente establece el pronóstico como nulo ($null$) cuando se encuentra un valor perdido durante la puntuación de un caso.

Nota: esta opción no se encuentra disponible para modelos CHAID. Para otros tipos de modelo, solamente está disponible para árboles de decisión (no conjuntos de reglas).

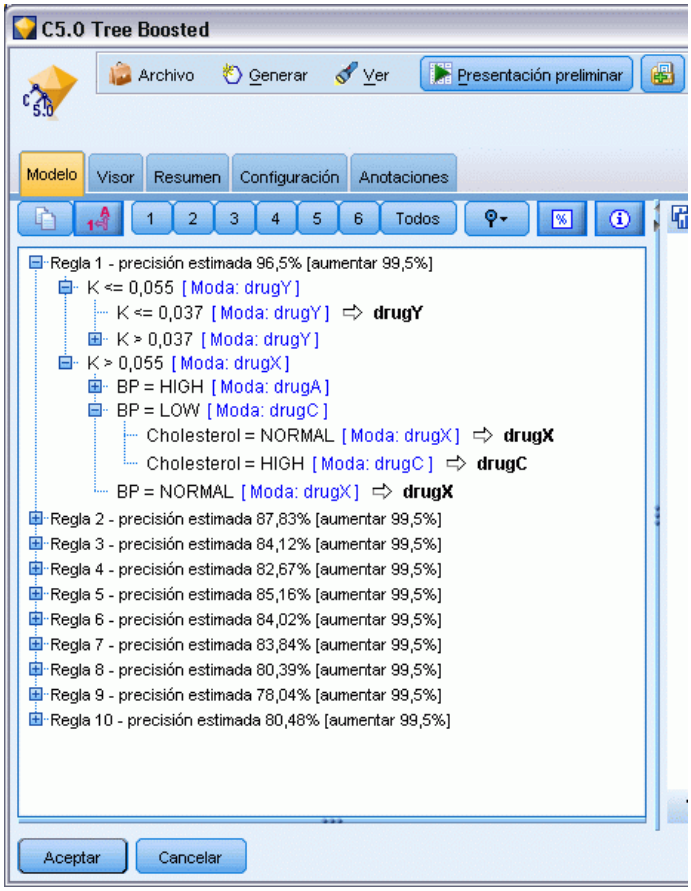
- **Generar con compatibilidad de valores perdidos.** En el caso de los modelos CHAID, QUEST y C&RT, puede activar la generación de SQL con una compatibilidad completa de valores perdidos. Esto significa que se genera SQL de manera que los valores perdidos se gestionan tal y como se especificó en el modelo. Por ejemplo, los árboles C&RT utilizan reglas basadas en sustitutos garantizadas por el descendiente mayor.

  Nota: para modelos C5.0, esta opción solamente está disponible para conjuntos de reglas (no árboles de decisión).

Modelos C5.0 aumentados

Nota: Esta función está disponible en SPSS Modeler Professional y SPSS Modeler Premium.

Figura 6-41
Nugget de modelo C5.0 aumentados, pestaña Modelo



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 comerciales. 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:

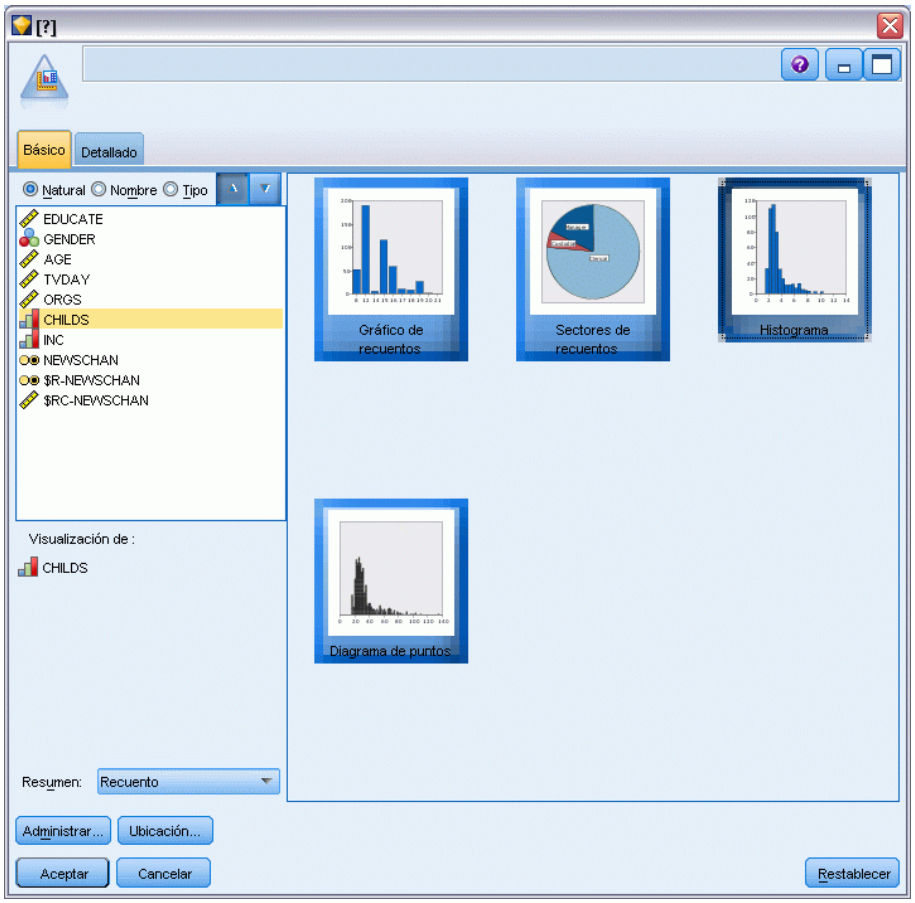
- En la pestaña Modelo de un nugget, expanda la lista de condiciones y reglas en el panel izquierdo y seleccione la lista que le interese.
- En la pestaña Visor de un nugget, expanda la lista de las ramas y seleccione el nodo que le interese.
- En la pestaña Visor de un árbol interactivo, expanda la 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:

▶ 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.

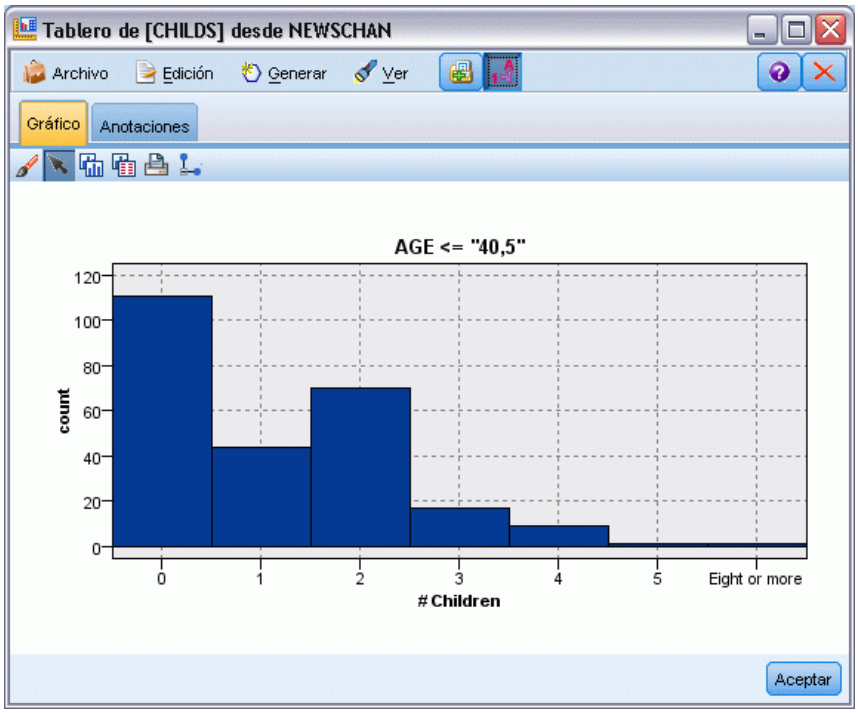
Figura 6-42
Cuadro de diálogo del nodo Tablero, pestaña Básico



Nota: cuando abre la pestaña Tablero de esta forma, las únicas pestañas disponibles son Básico y Detallado. Si desea obtener más información, consulte el tema Nodo Tablero en el capítulo 5 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

- Si utiliza la configuración de la pestaña Básico o Detallado, especifique los detalles que se mostrarán en el gráfico.
- Pulse en Aceptar para generar el gráfico.

Figura 6-43
Histograma generado a partir de la pestaña Tablero básico



El encabezado del gráfico identifica los nodos o reglas seleccionadas que se incluirán.

Nuggets de modelo para aumento, agregación autodocimante y conjuntos de datos muy amplios

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. Si desea obtener más información, consulte el tema Modelos de conjuntos en el capítulo 3 el p. 58.

El nugget de modelo resultante contiene las siguientes pestañas. La pestaña Modelo ofrece una serie de vistas del modelo.

Pestaña	Ver	Descripción	Más información
Modelo	Resumen de 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.	Si desea obtener más información, consulte el tema Resumen del modelo en el capítulo 3 el p. 60.
	Importancia del predictor	Muestra un gráfico que indica la importancia relativa de cada predictor (campo de	Si desea obtener más información, consulte el tema Importancia del

Pestaña	Ver	Descripción	Más información
		entrada) cuando se calcula el modelo.	predictor en el capítulo 3 el p. 61.
	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.	Si desea obtener más información, consulte el tema Frecuencia de predictor en el capítulo 3 el p. 62.
	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.	Si desea obtener más información, consulte el tema Detalles de modelo de componente en el capítulo 3 el p. 65.
	Información	Muestra información sobre los campos, los ajustes de versión y el proceso de estimación del modelo.	Si desea obtener más información, consulte el tema Información / Resumen de nugget de modelo en el capítulo 3 el p. 54.
Configuración		Le permite incluir confianzas en operaciones de puntuación.	Si desea obtener más información, consulte el tema Configuración del nugget de modelo de conjunto de reglas/árboles de decisión el p. 184.
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.	Si desea obtener más información, consulte el tema Anotaciones en el capítulo 5 en *Manual de usuario de IBM SPSS Modeler 15*.

Nuggets de modelo del conjunto de reglas

Un nugget de modelo Conjunto de reglas representa las reglas para pronosticar un determinado campo de salida descubierto por los nodos de modelado de reglas de asociación (Apriori) o por uno de los nodos de generación de árboles (Árbol C&R, 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, 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 pronósticos.

Cuando se ejecuta una ruta que contiene el nugget Conjunto de reglas, se añaden a la ruta dos nuevos campos que contienen el valor pronosticado 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 *$A-* para el campo de pronóstico y *$AC-* para el campo de confianza. Para los conjuntos de reglas C5.0, los prefijos son *$C-* para el campo de

pronóstico 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 pronóstico y *$RC-* para el campo de confianza. En una ruta con varios nugget Conjunto de reglas en una serie que pronostican los mismos campos de salida, los nuevos nombres de campos contendrán números en el *prefijo* para que se puedan distinguir entre sí. El primer nugget Conjunto de reglas de asociación de la ruta utilizará los nombres comunes, el segundo usará los nombres que comiencen por *$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 varios pronósticos y puede que no coincidan. Existen dos métodos para generar pronósticos 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 los pronósticos 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 un pronóstico 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 pronóstico final. La confianza para el pronóstico 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 el pronóstico.

El método utilizado puede controlarse en las opciones de ruta. Si desea obtener más información, consulte el tema Configuración de opciones generales de las rutas en el capítulo 5 en *Manual de usuario de IBM SPSS Modeler 15*.

Generación de nodos. El menú Generar permite crear nuevos nodos basados en el conjunto de reglas.

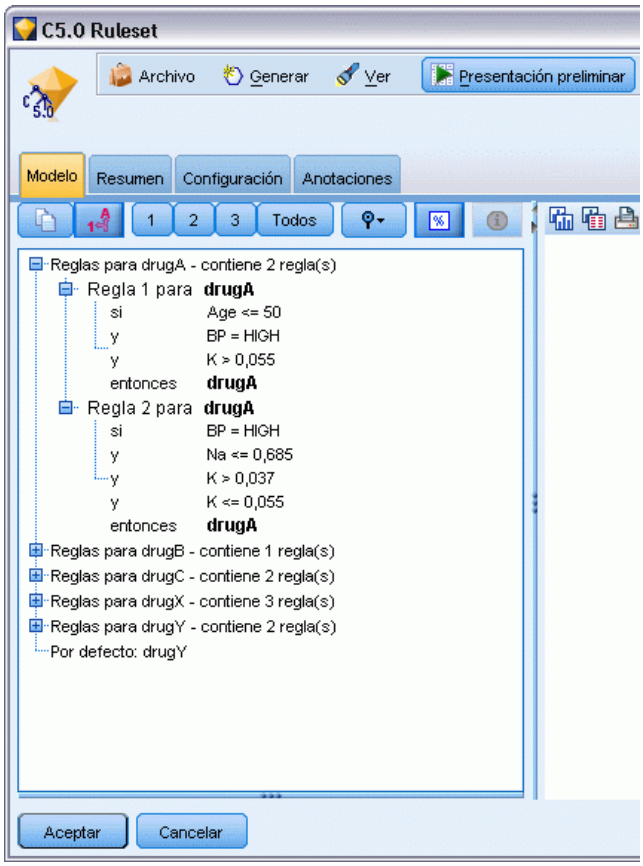
- **Nodo Filtro** Crea un nuevo nodo Filtro para filtrar los campos que las reglas no utilizan en el conjunto de reglas.
- **Nodo Seleccionar** Crea un nuevo nodo Seleccionar para seleccionar los registros a los que se aplica la regla seleccionada. El nodo generado seleccionará los registros para los 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 el pronóstico 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 elección, éste es una cadena más compleja que muestra la entrada al mecanismo de elección.
- **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. Solamente está disponible para los modelos C5.0 **aumentados**. Si desea obtener más información, consulte el tema Modelos C5.0 aumentados el p. 186.
- **Modelo a paleta** Devuelve el modelo a la paleta de modelos. Esto resulta útil en situaciones en las que un compañero puede haber enviado una transmisión que contenga el modelo, pero no el modelo propiamente dicho.

Nota: las pestañas Configuración y Resumen del nugget Conjunto de reglas son idénticas a las que utilizan los modelos de árboles 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.

Figura 6-44
Nugget de modelo Conjunto de reglas, pestaña Modelo



El consecuente descompone las reglas (categoría pronosticada), que se presentan en el siguiente formato:

si *antecedente_1*
y *antecedente_2*
...
y *antecedente_n*
entonces *valor pronosticado*

donde consequent y antecedent_1 medianteantecedent_n son condiciones. La regla se interpreta como "para los registros donde antecedent_1 mediante antecedent_n son verdaderos, consequent es posible que también sea 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:

- Seleccione en los menús de SPSS Modeler:
  File > Abrir ruta
- En la lista desplegable Archivos de tipo, seleccione Archivos de proyecto AT (*.atp, *.ats).

  Los proyectos importados se convierten en rutas de SPSS Modeler con los siguientes nodos:
  - Un nodo de fuente que define la fuente de datos utilizada (por ejemplo, una fuente 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 papel (campo predictor o de entrada frente a campo pronosticado 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).
- Para ver los árboles generados, ejecute la ruta.

Comentarios

- Por lo general, no es posible exportar a AnswerTree los árboles de decisión generados en SPSS Modeler; la importación de AnswerTree a SPSS Modeler es una acción monodireccional.
- Los beneficios definidos en AnswerTree no se conservan una vez importado el proyecto a 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 cadena de Markov que se utilizan principalmente para la clasificación.

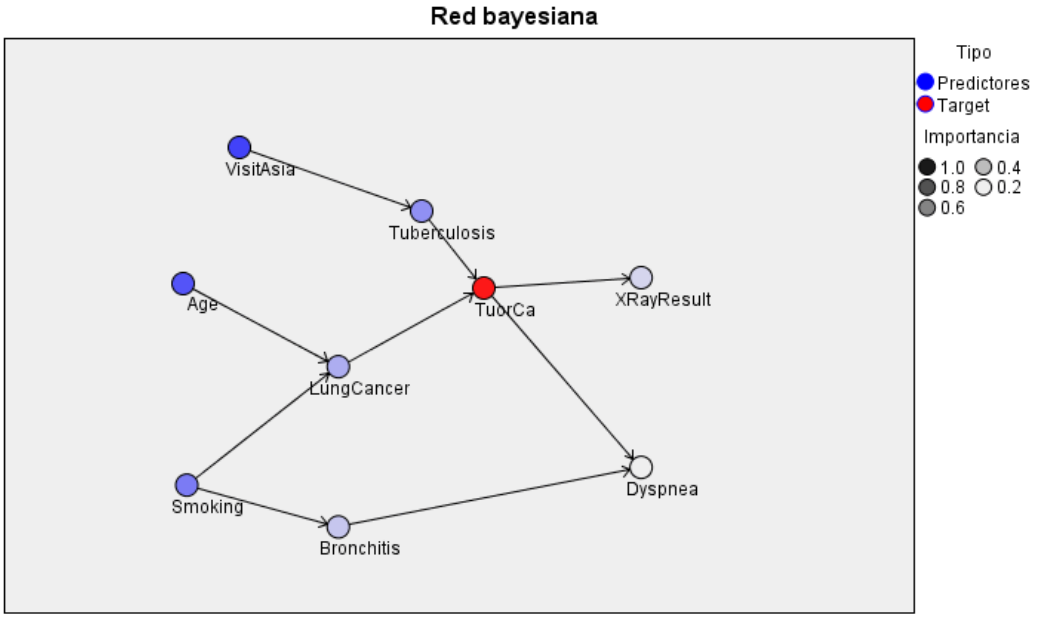
Las redes bayesianas se utilizan para realizar pronósticos en diferentes situaciones; algunos ejemplos son los siguientes:

- 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 los mejores pronósticos 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 de detectadas en una radiografía de los pulmones pueden estar causadas por tuberculosis o cáncer, mientras que las posibilidades de que un paciente tenga dificultades respiratorias (disnea) aumentan si también padece bronquitis o cáncer de pulmón.

Figura 7-1
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 pronosticar 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. Si desea obtener más información, consulte el tema Nodo Intervalos en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

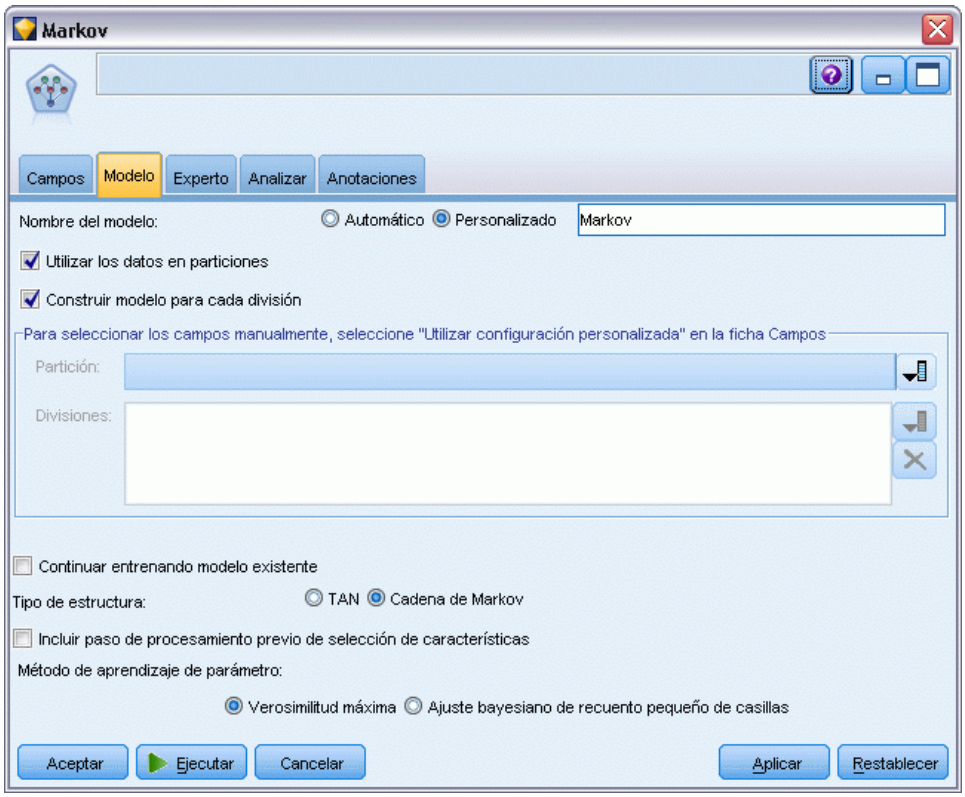
Ejemplo. Un analista de un banco quiere poder pronosticar clientes o clientes potenciales, con posibilidades de impago de sus créditos. Puede utilizar un modelo de red bayesiana para identificar las características de los clientes con más posibilidades de impago y generar varios tipos diferentes de modelo para establecer el mejor método para pronosticar los clientes con más posibilidades de impago. Si desea obtener más información, consulte el tema Predicción de moras en préstamos (red bayesiana) en el capítulo 17 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

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. Si desea obtener más información, consulte el tema Reentrenamiento de un modelo mensualmente (red bayesiana) en el capítulo 18 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

Opciones de modelo de nodo de red bayesiana

Figura 7-2
Nodo Red bayesiana: Pestaña 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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Construir modelo para cada división. Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como modelos divididos. Si desea obtener más información, consulte el tema Generación de modelos divididos en el capítulo 3 el p. 32.

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 sobre la adecuación 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 sólo hay una partición, se usará automáticamente siempre que se active la partición.) Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*. 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 dividido. Se trata de una acción similar a establecer el papel del campo en *Segmentar* en un nodo Tipo. Sólo puede designar campos con un nivel de medida de Marca, Nominal, Ordinal o Continuo como campos divididos. Los campos seleccionados como campos divididos no se pueden utilizar como campos de destino, entrada, partición, frecuencia o ponderación. Si desea obtener más información, consulte el tema Generación de modelos divididos en el capítulo 3 el p. 32.

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 del 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.
- **Cadena de Markov.** Selecciona el conjunto de nodos del conjunto de datos que contiene las variables objetivo principales, sus filiales y sus filiales principales. Esencialmente, una cadena de Markov identifica todas las variables de la red que son necesarias para pronosticar 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.

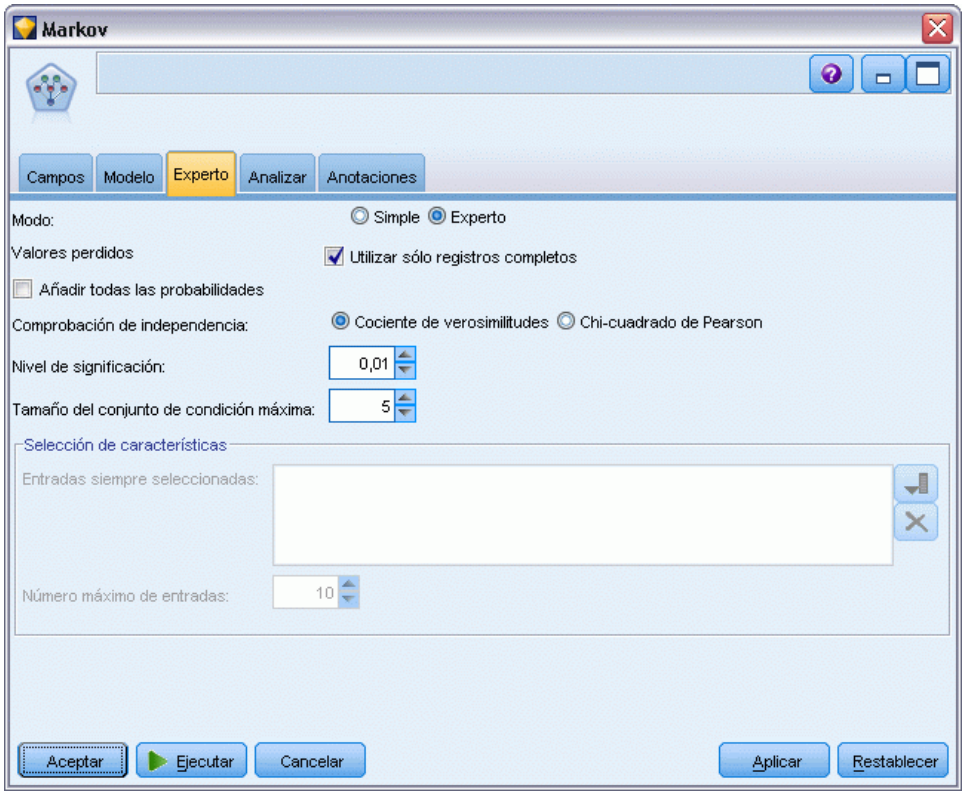
Incluir paso de procesamiento previo de selección de características. Si selecciona esta casilla podrá utilizar las opciones de selección de características de la pestaña Experto.

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 principales. 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 las principales:

- **Verosimilitud máxima.** Seleccione esta casilla si utiliza un conjunto de datos grande. Ésta es la opción seleccionada por defecto.
- **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

Figura 7-3
Nodo Red bayesiana: Pestaña Experto



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. Por defecto, 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

tal caso, puede cancelar la selección de Utilizar sólo registros completos. SPSS Modeler intentará utilizar tanta información como sea posible para calcular el modelo, incluyendo 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 pronosticada.

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 comprobaciones de independencia sólo están disponibles si selecciona Incluir paso de procesamiento previo de selección de características o un Tipo de estructura de una cadena de Markov 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 por defecto es 0,01.

Nota: Esta opción sólo está disponible si selecciona Incluir paso de procesamiento previo de selección de características o un Tipo de estructura de una cadena de Markov en la pestaña Modelo.

Tamaño del conjunto de condición máxima. El algoritmo para crear una estructura de cadena 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. La configuración por defecto es de 5 líneas.

Nota: Esta opción sólo está disponible si selecciona Incluir paso de procesamiento previo de selección de características o un Tipo de estructura de una cadena de Markov 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 una cadena de Markov, debido a un posible

número elevado de entradas potenciales; permite seleccionar las entradas significativas en función de su variable objetivo.

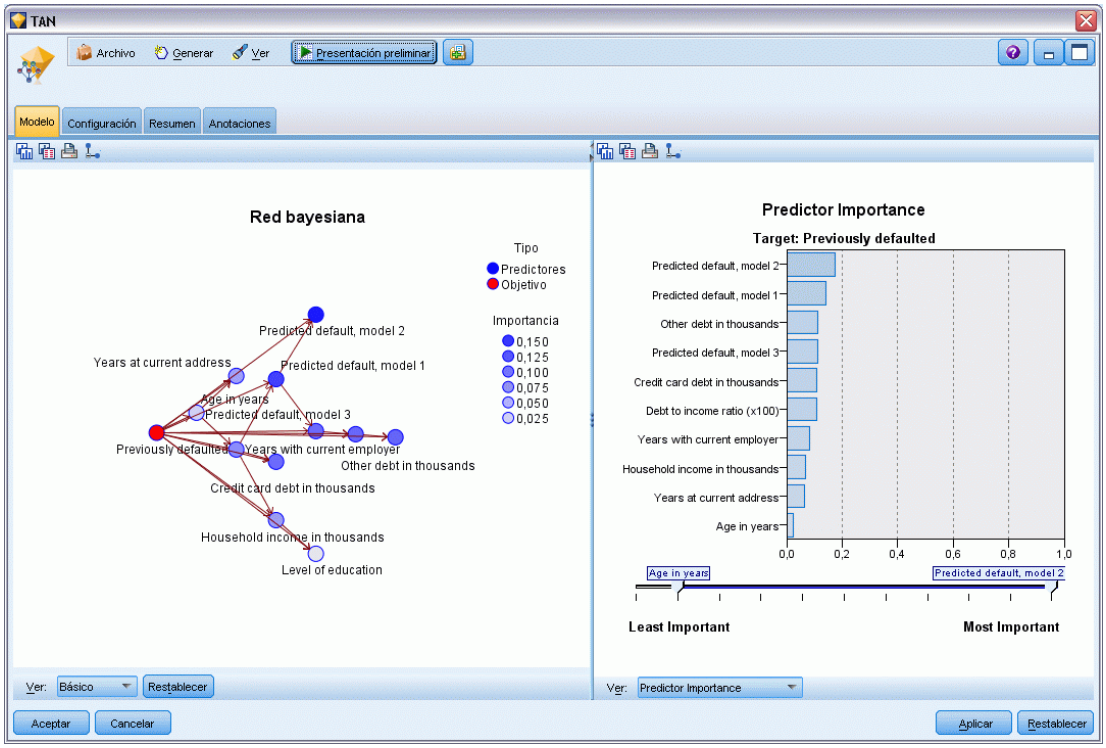
Nota: Las opciones de selección sólo están disponibles si selecciona Incluir paso de procesamiento previo de selección de características en la pestaña Modelo.

- **Entradas siempre seleccionadas** Si utiliza el botón Selector de campos (a la derecha del campo de texto), seleccione los campos del conjunto de datos que siempre se utilizarán cuando cree el modelo de red bayesiana. Tenga en cuenta que el campo objetivo siempre está seleccionado.
- **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 seleccionado en Entradas siempre seleccionadas supera el valor de Número máximo de entradas, aparecerá un mensaje de error.

Nugget de modelo de red bayesiana

Figura 7-4
Detalles de modelo de red bayesiana e Importancia de predictor asociada



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, así como las relaciones 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 grupos de valores de nodos que representan un rango se muestran en una ventana emergente cuando pasa el puntero sobre el nodo.

Puede utilizar las herramientas de gráfico 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.

Sugerencia: Si la red contiene numerosos nodos, puede pulsar en un nodo 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 mostrar sus valores en una ventana emergente.

Panel derecho

Importancia del predictor. Muestra un gráfico que indica la importancia relativa de cada predictor cuando se calcula el modelo. Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

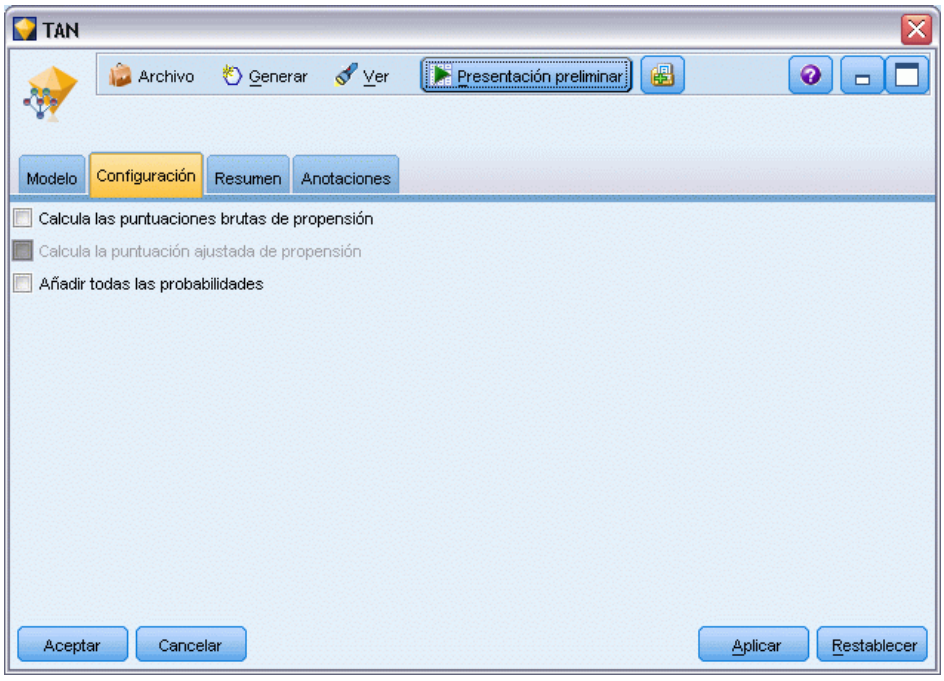
Probabilidades condicionales. Si selecciona un nodo o un minigráfico de distribución en la columna izquierda, se muestra la tabla de probabilidades condicionales asociadas en el panel derecho. Esta tabla contiene el valor de probabilidad condicional de cada valor de nodo y combinación de valores en sus nodos principales. Además, incluye el número de registros observados para cada valor de registro y cada combinación de valores en los nodos principales.

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.

Figura 7-5
Pestaña Configuración de un modelo de red bayesiana



Calcular puntuaciones brutas de propensión. En el caso de modelos con un objetivo de marca (que devuelve un pronóstico 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 pronóstico y confianza que pueden generarse durante la puntuación.

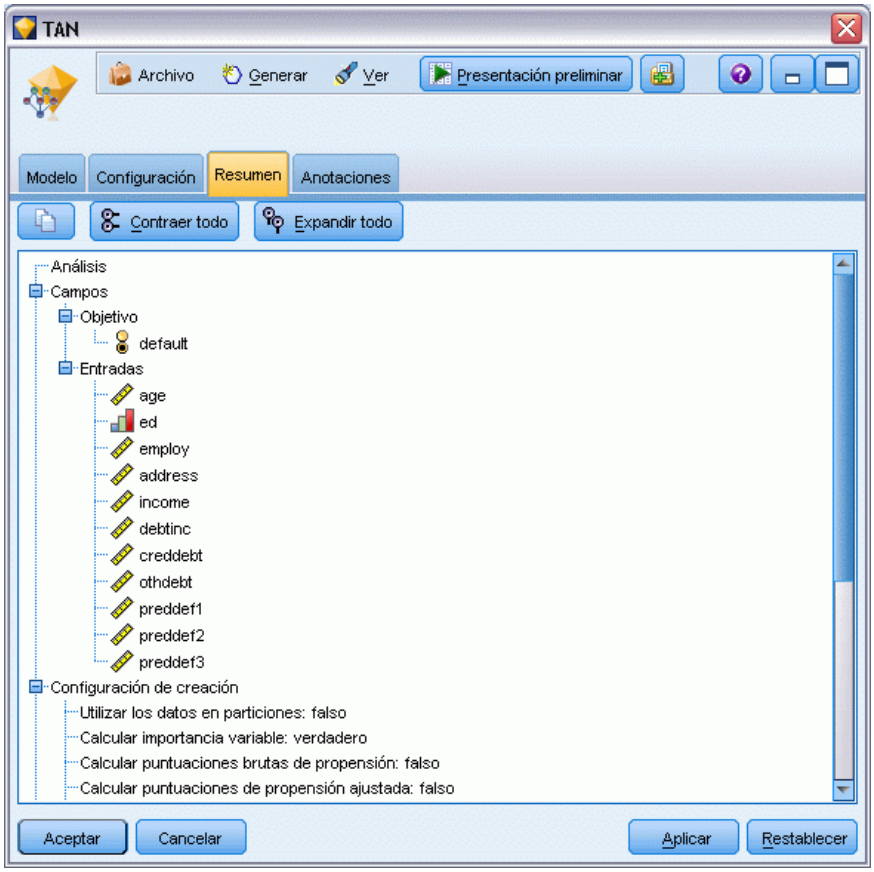
Calcular puntuaciones de propensión ajustada. Las puntuaciones brutas de propensión se basan sólo en los datos de entrenamiento y pueden ser demasiado optimistas debido a la tendencia de muchos modelos a sobreajustar estos datos. Las propensiones ajustadas intentan compensar evaluando el rendimiento 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 la ficha Analizar 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 pronosticada.

El valor por defecto de esta casilla de verificación está determinado por la casilla de verificación correspondiente en la pestaña Experto del nodo de modelado. Si desea obtener más información, consulte el tema Opciones de experto del nodo de red bayesiana el p. 198.

Resumen de modelo de red bayesiana

Figura 7-6
Pestaña Resumen de un modelo de red bayesiana



La ficha 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 ficha 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.

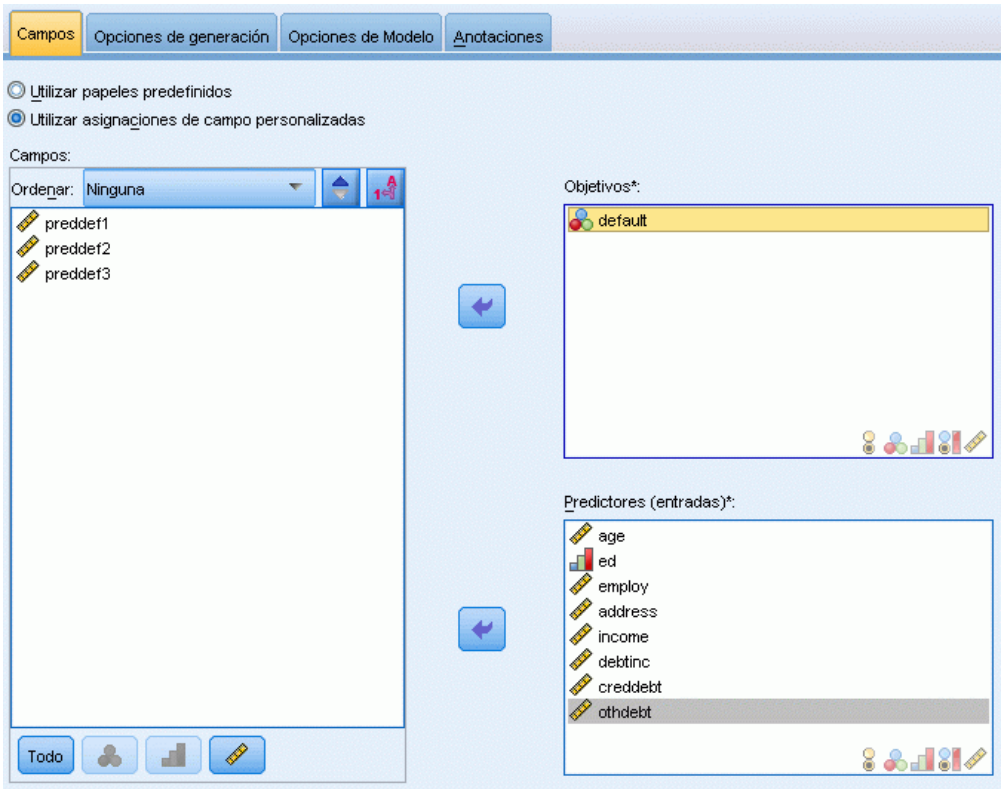
Capítulo
8

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 8-1
Pestaña campos

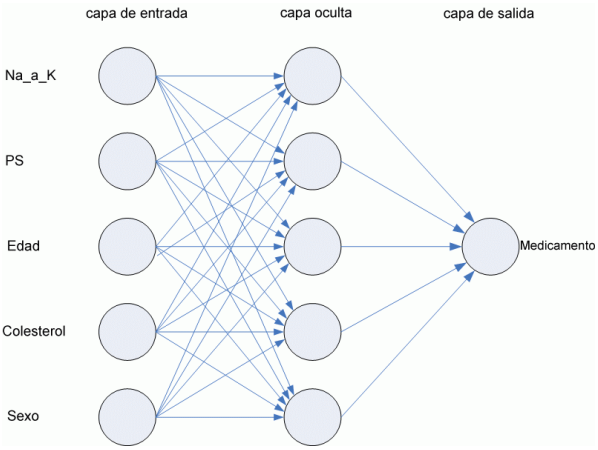


Requisitos del campo. Debe haber al menos un objetivo y una entrada. Se ignorarán los campos establecidos en Ambos o Ninguno. No hay restricciones de nivel de medida en los objetivos o en los predictores (entradas). Si desea obtener más información, consulte el tema Opciones de los campos del nodo de modelado en el capítulo 3 el p. 38.

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 8-2
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. Existen, generalmente, tres capas 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 unidades que representan los campos objetivo. 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 un pronóstico para cada registro y realizando ajustes a las ponderaciones cuando realiza un pronóstico incorrecto. Este proceso se repite muchas veces y la red sigue mejorando sus pronósticos 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.

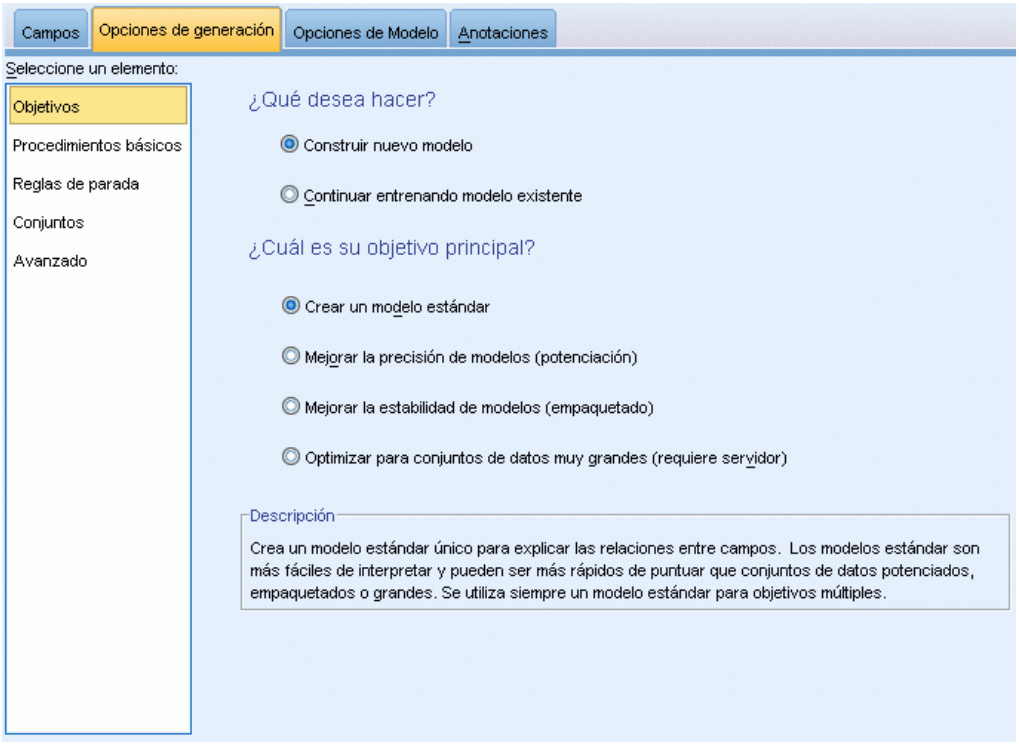
Uso de redes neuronales con rutas heredadas

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) en la versión 13 o posterior, debería marcar los valores desconocidos como valores perdidos (por ejemplo, por medio del nodo Tipo).

Objetivos

Figura 8-3
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 se activa esta opción, se desactivan el resto de los controles de las pestañas Campos y Opciones de creación.

¿Cuál es su objetivo principal?

- **Crear un modelo estándar.** El método genera un modelo simple para pronosticar el destino mediante los predictores. Por lo general, los modelos estándar son más fáciles de interpretar y pueden puntuarse más rápido que conjuntos de datos de gran tamaño o de aumento o agregación autodocimante.
- **Mejorar la precisión del modelo (aumento).** El método genera un modelo de conjunto mediante el aumento, que genera una secuencia de modelos para obtener pronósticos más precisos. Se puede tardar más tiempo en generar y puntuar conjuntos que un modelo estándar.

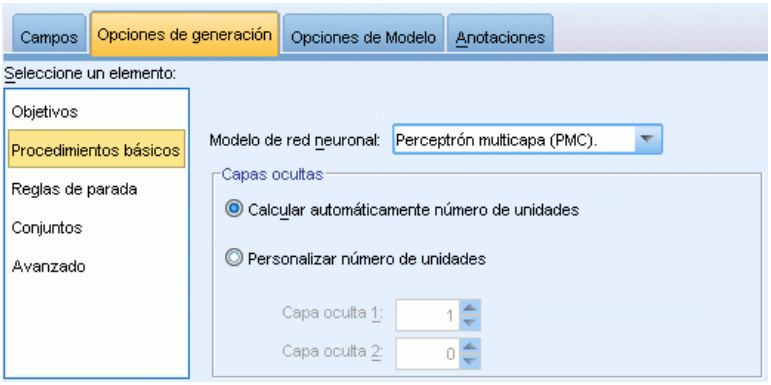
  El aumento produce una sucesión de "modelos de componente", cada uno de ellos basados en el conjunto de datos completo. Antes de crear cada modelo de componente sucesivo, los registros se ponderan en función de los residuos del modelo del componente anterior. Los casos con residuos de grandes dimensiones tienen ponderaciones de análisis relativamente superiores para que el siguiente modelo de componente se centre en pronosticar 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 pronósticos 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 autodocimantes 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.
- **Cree un modelo para conjuntos de datos muy grandes (necesita IBM® SPSS® Modeler Server).** 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 generar uno 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. Esta opción requiere conexión al servidor SPSS Modeler Server.

Cuando existen objetivos múltiples, este método sólo creará un modelo estándar, sin importar el objetivo seleccionado.

Conceptos básicos

Figura 8-4
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.

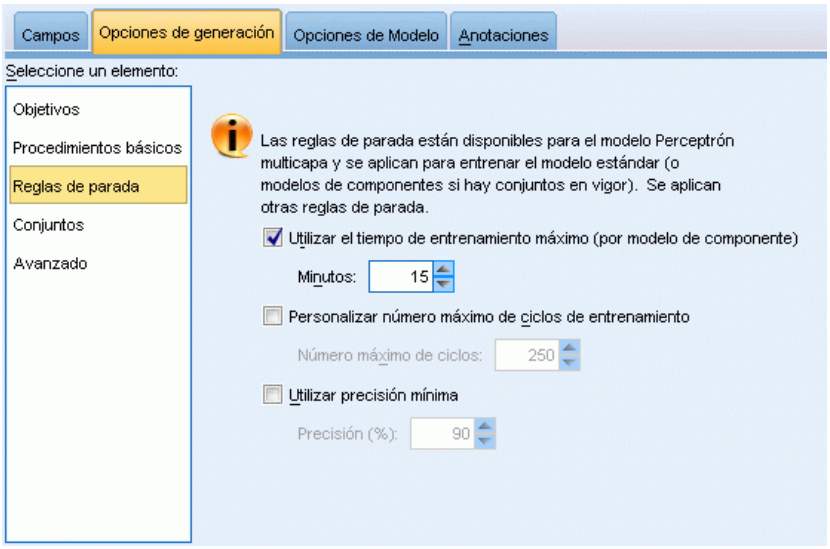
Capas ocultas. Las capas ocultas de una red neuronal contienen unidades no observables. El valor de cada unidad oculta es alguna función de los predictores; la forma exacta de la función depende en parte del tipo de red. Los perceptrones multicapa pueden tener una o dos capas ocultas; la red de función 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.

  Nota: 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 8-5
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 (pase de datos) y puede detenerse luego según los siguientes criterios.

Emplear el tiempo de 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.

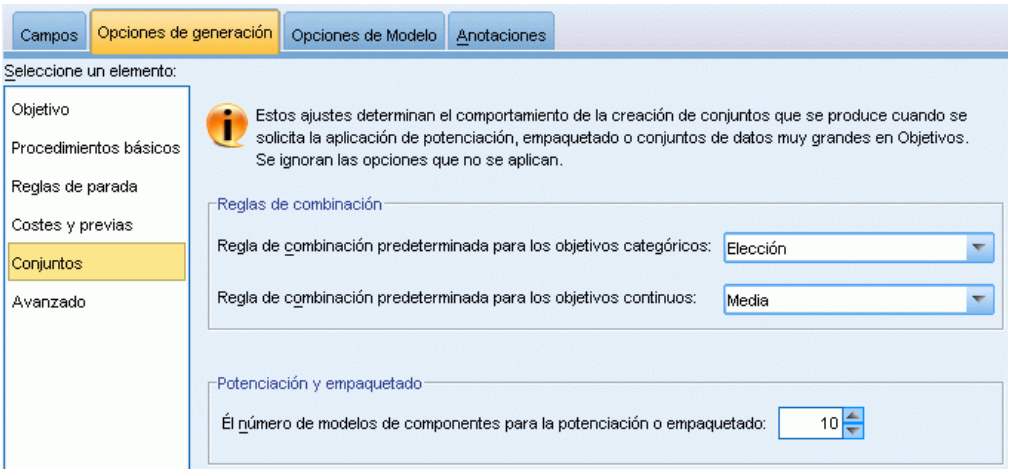
Personalizar el número máximo de ciclos de entrenamiento. El número máximo de ciclos de entrenamiento permitidos. Si se supera el número máximo de ciclos, el entrenamiento se detiene. Especifique un entero mayor que 0.

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 8-6
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.

Bagging y conjuntos de datos muy grandes. Al puntuar un conjunto, ésta es la regla utilizada para combinar los valores pronosticados a partir de los modelos básicos para calcular el valor de puntuación del conjunto.

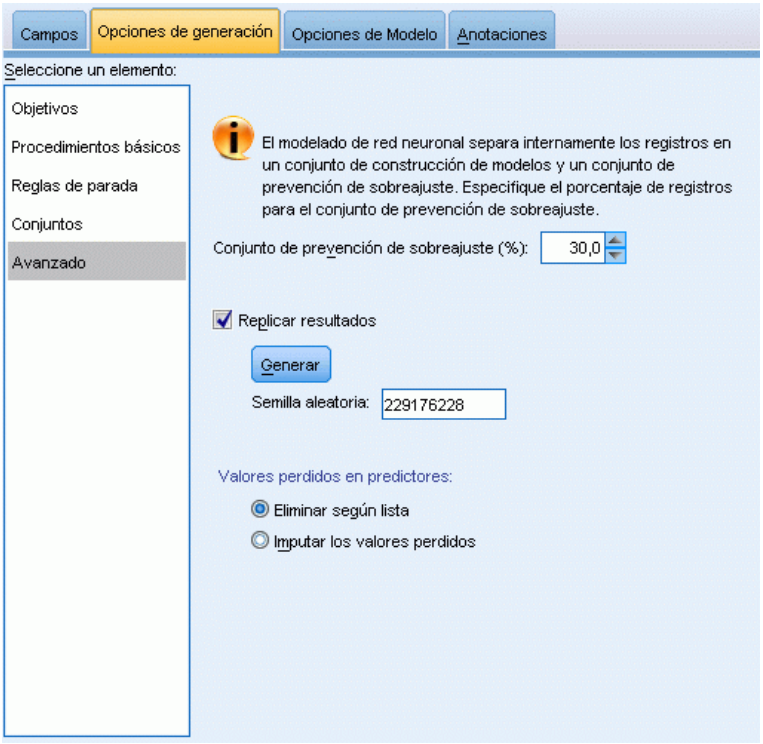
- **Regla de combinación predeterminada para objetivos categóricos.** Los valores pronosticados de conjunto para objetivos categóricos pueden combinarse mediante votación, la mayor probabilidad o la mayor probabilidad media. **Votación** selecciona la categoría que tenga la mayor probabilidad más frecuentemente 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 pronosticados de conjunto para objetivos continuos pueden combinarse mediante la media o mediana de los valores pronosticados 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 autodocimantes. Debe ser un número entero positivo.

Avanzados

Figura 8-7
Configuración avanzada



La configuración avanzada controla las opciones que no se ajustan bien en otros grupos de configuraciones.

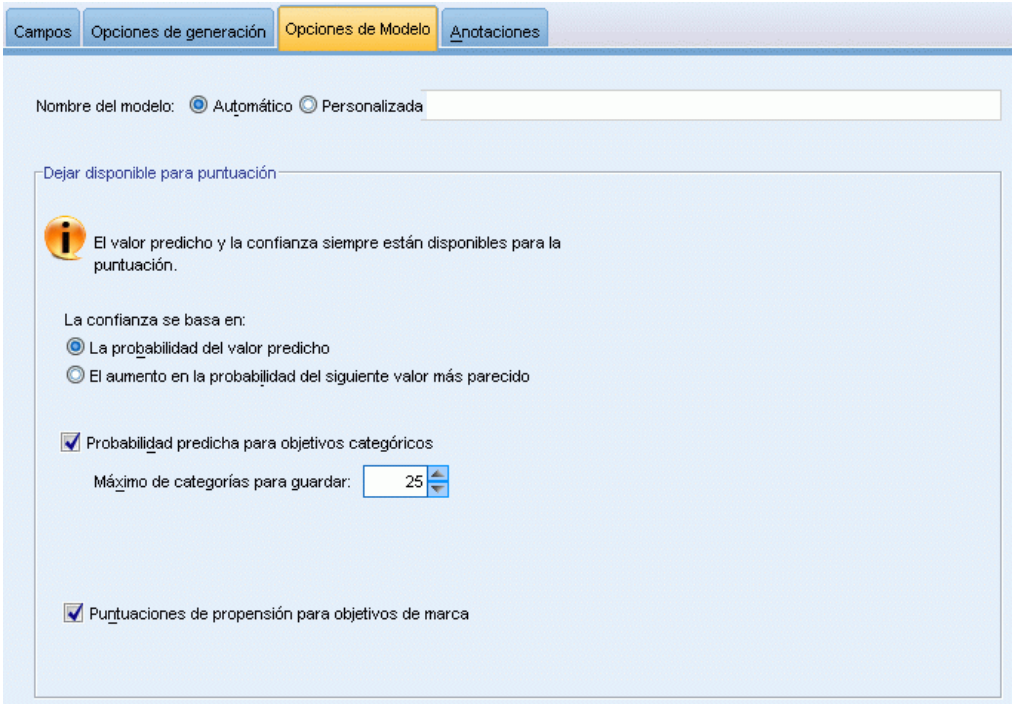
Conjunto de prevención sobreajustado. El método de red neuronal divide los registros de manera interna en un conjunto de creación de modelos y un conjunto de prevención sobreajustado, el cual es un conjunto independiente de registros de datos utilizado para realizar un seguimiento de errores durante la formación para evitar que el método modele una variación atribuible al azar en los datos. Especifique un porcentaje de registros. El valor predeterminado 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. Por defecto, 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 8-8
Pestaña Opciones de modelo



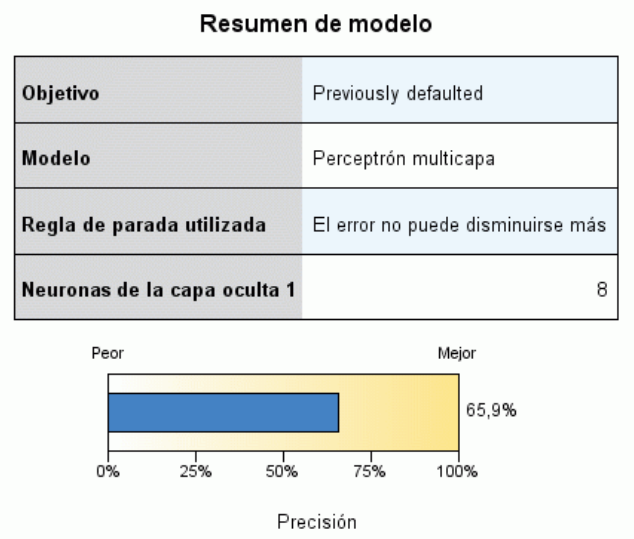
Nombre del modelo. Puede generar el nombre del modelo automáticamente tomando como base los campos objetivo o especificar un nombre personalizado. El nombre generado automáticamente es el nombre del campo objetivo. Si existen objetivos múltiples, el nombre del modelo se forma con los nombres de campos en orden, conectados por símbolos &. Por ejemplo, si *campo1 campo2 campo3* son objetivos, el nombre de modelo es: *campo1 & campo2 & campo3*.

Dejar disponible para puntuar. 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 un pronóstico 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 brutas de propensión; si hay particiones activas, el modelo también producirá puntuaciones de propensión ajustadas en función de la partición de prueba. Si desea obtener más información, consulte el tema Puntuaciones de propensión en el capítulo 3 el p. 45.

Resumen del modelo

Figura 8-9
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.

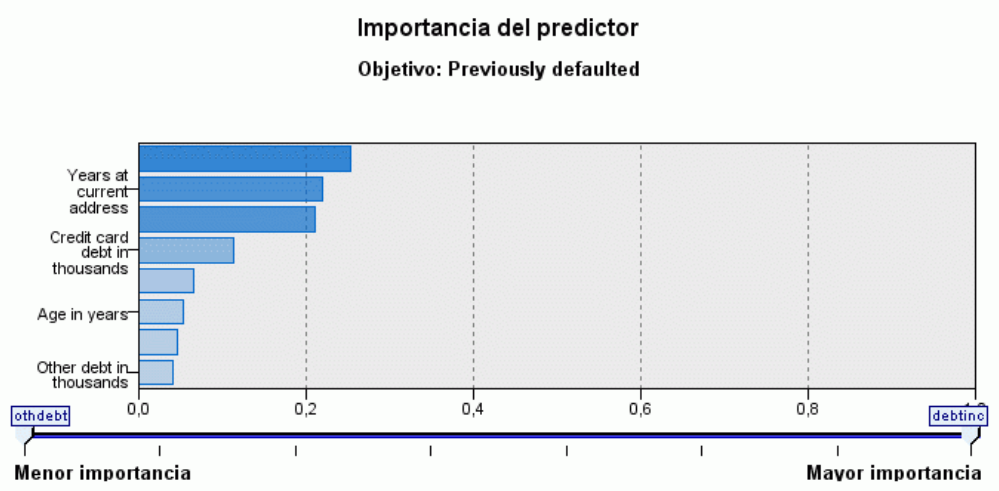
Resumen del modelo. La tabla identifica el objetivo, el tipo de red neuronal entrenada, la regla de parada que detuvo el entrenamiento (se muestra si se entrenó una red de perceptrones multicapa) y el número de neuronas en cada capa oculta de la red.

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, este 1 menos el índice del error absoluto de la media en la predicción (la media de los valores absolutos de los valores predichos menos los valores observados) hasta el intervalo de los valores predichos (el valor predicho máximo menos el valor predicho mínimo).

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 8-10
Vista de importancia del predictor

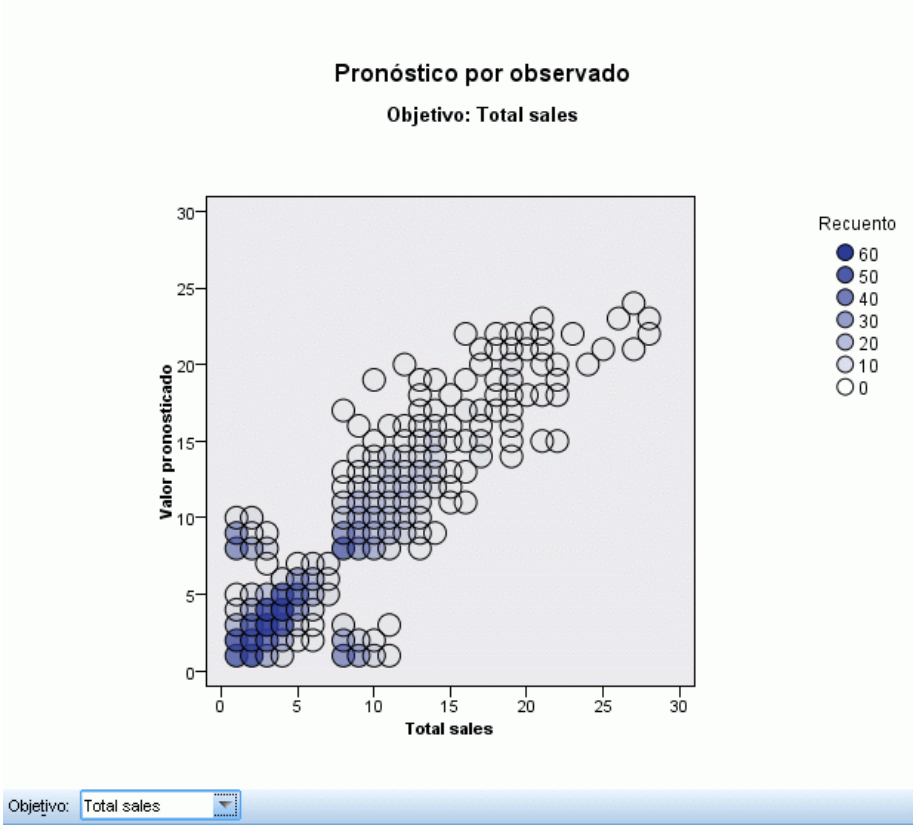


Es normal centrar los esfuerzos de modelado en los campos predictores más importantes y valorar la omisión de aquellos con menor relevancia. El gráfico de importancia de los predictores le ayuda a hacerlo indicando la importancia relativa de cada predictor en la estimación del modelo. Como los valores 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 8-11
Vista Predicho por observado

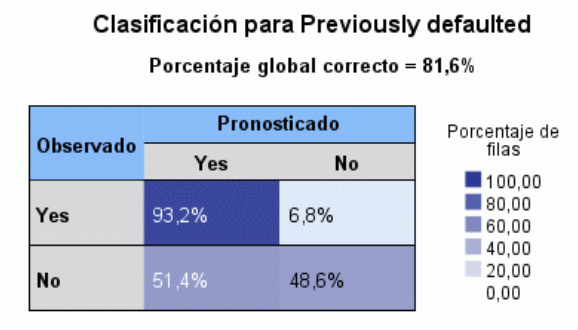


Para objetivos continuos, muestra un diagrama de dispersión en intervalos de los valores predichos en el eje vertical por los valores observados en el eje horizontal.

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 8-12
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 por defecto.
- **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 encabezados de filas o columnas, ni valores en las casillas. Puede ser útil cuando el objetivo tiene muchas categorías.

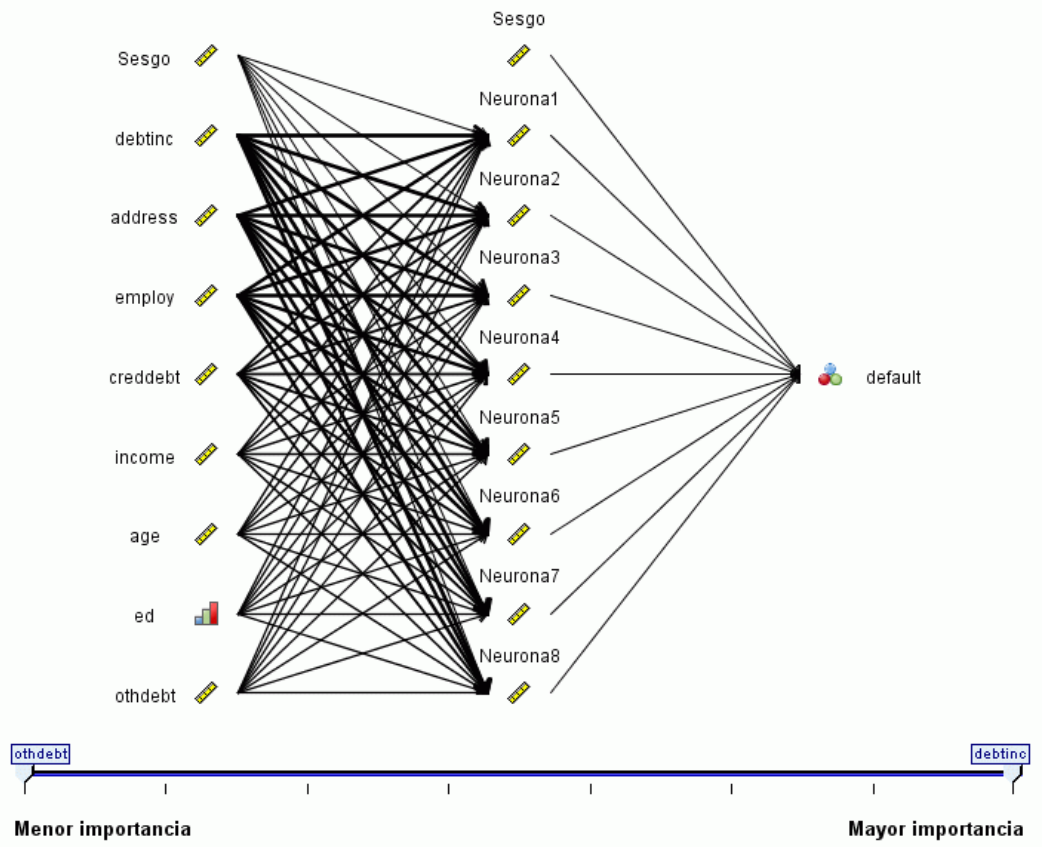
Perdidos. Si cualquier registro tiene valores perdidos en el objetivo, se muestran en una fila (Perdidos) bajo todas las filas válidas. Los registros con valores perdidos no contribuyen al porcentaje global correcto.

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 8-13
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 por defecto.
- **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. Por defecto, 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.

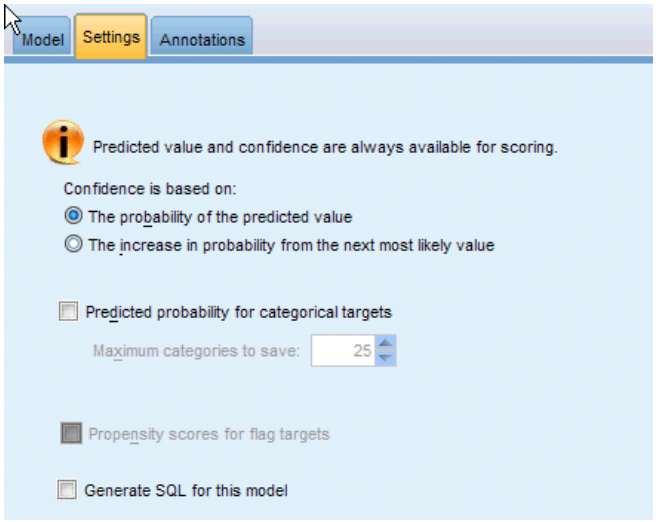
Importancia del predictor. Las líneas de conexión del diagrama se ponderan tomando como base la importancia de predictores, con un grosor de línea mayor correspondiente a una importancia mayor. Existe un control deslizante Importancia del predictor en la barra de herramientas

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 8-14
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 un pronóstico 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 brutas de propensión; si hay particiones activas, el modelo también producirá puntuaciones de propensión ajustadas en función de la partición de prueba. Si desea obtener más información, consulte el tema Puntuaciones de propensión en el capítulo 3 el p. 45.

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. Si desea obtener más información, consulte el tema Optimización de SQL en el capítulo 6 en *Guía de rendimiento y administración de IBM SPSS Modeler Server 15*.

Puntuar convirtiendo a SQL nativo. Si selecciona esta opción, se genera SQL para puntuar el modelo de manera nativa dentro de la aplicación.

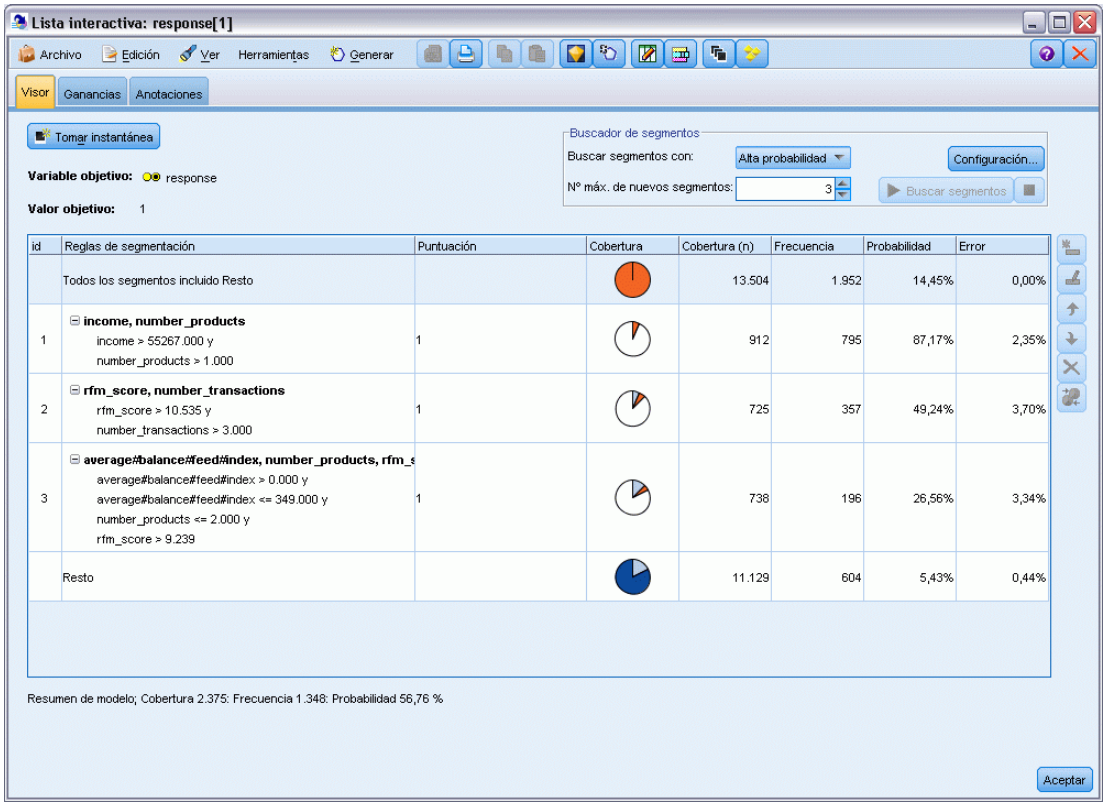
Capítulo

9

Lista de decisiones

Los modelos de Decision List 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 pérdida o con más posibilidades de decir sí a una campaña u oferta determinada. El Decision List Viewer proporciona control total sobre el modelo, ya que le permite editar segmentos, añadir sus propias reglas de negocios, 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.

Figura 9-1
Modelo de lista de decisiones



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:

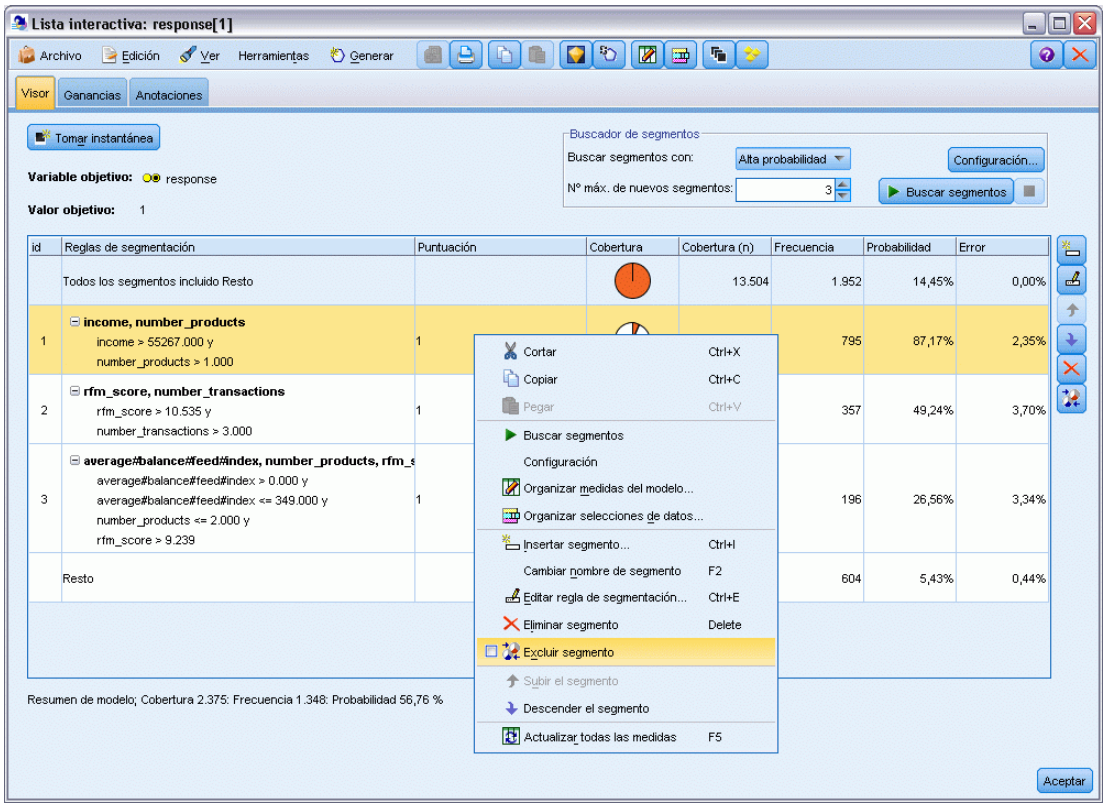
RFM_SCORE > 10 y
MONTHS_CURRENT <= 9

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 Decision List Viewer 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 Decision List devuelven *Sí* para los segmentos incluidos y *$null$* para todo los demás, incluido el resto. Este control directo sobre la puntuación hace que los modelos Decision List 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.

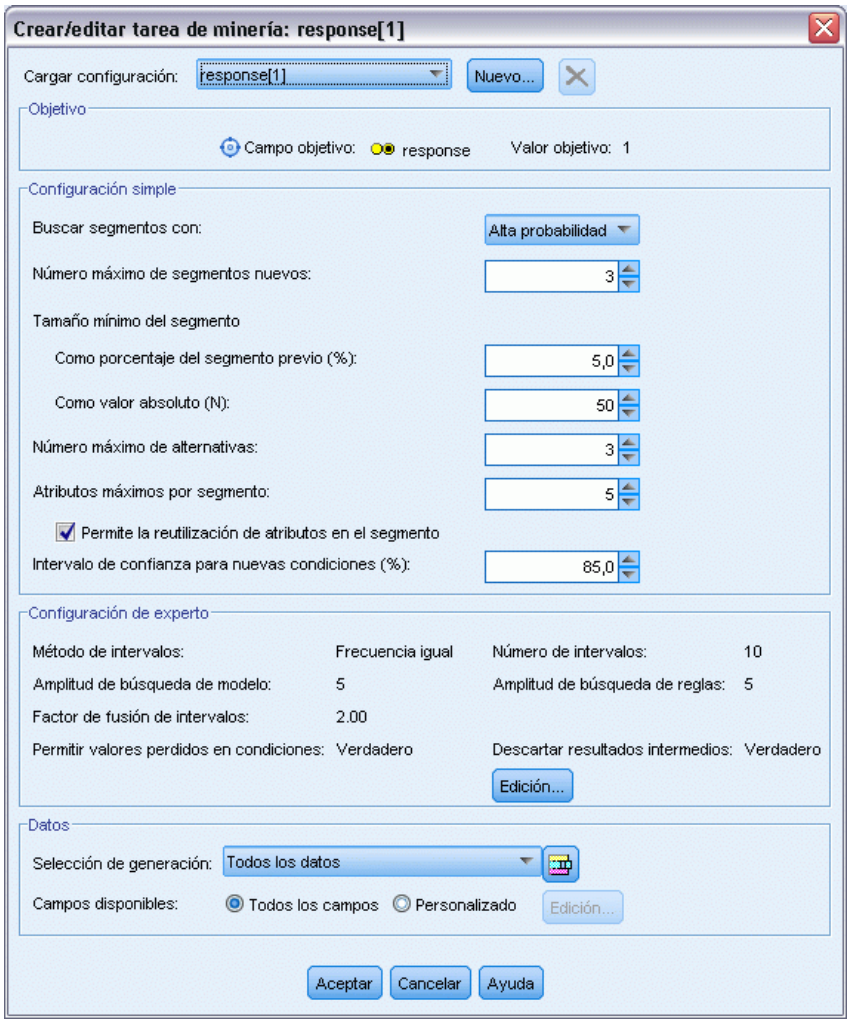
Figura 9-2
Modelo de lista de decisiones



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 por defecto se basa en las especificaciones iniciales del nodo Decision List, 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.

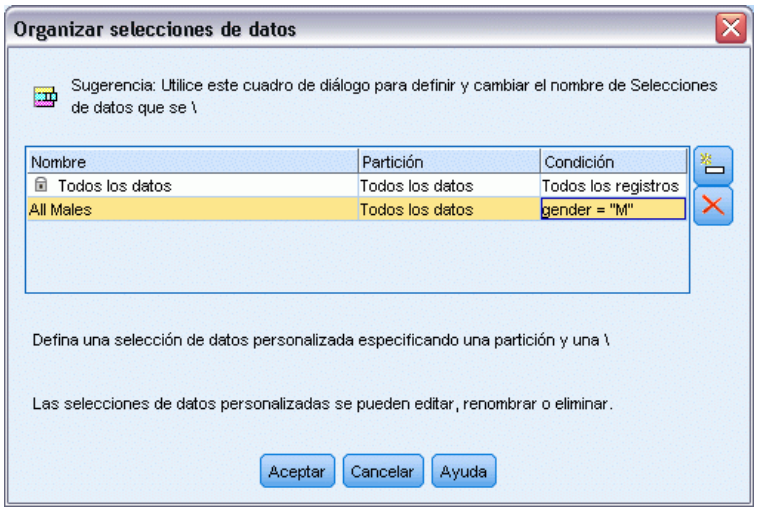
Figura 9-3
Creación de una tarea de minería



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.

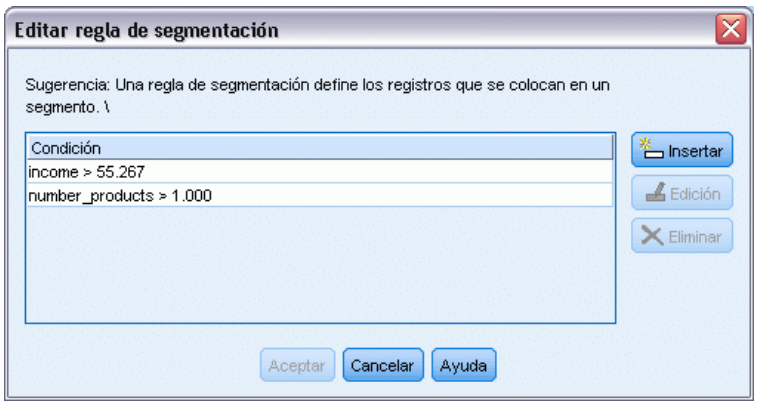
Figura 9-4
Creación de una selección de datos



Incorporación de conocimiento empresarial

Al ajustar y ampliar segmentos identificados por el algoritmo, el Decision List Viewer 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.

Figura 9-5
Especificación de una regla



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 aparecen en el Decision List Viewer mientras se genera el modelo.

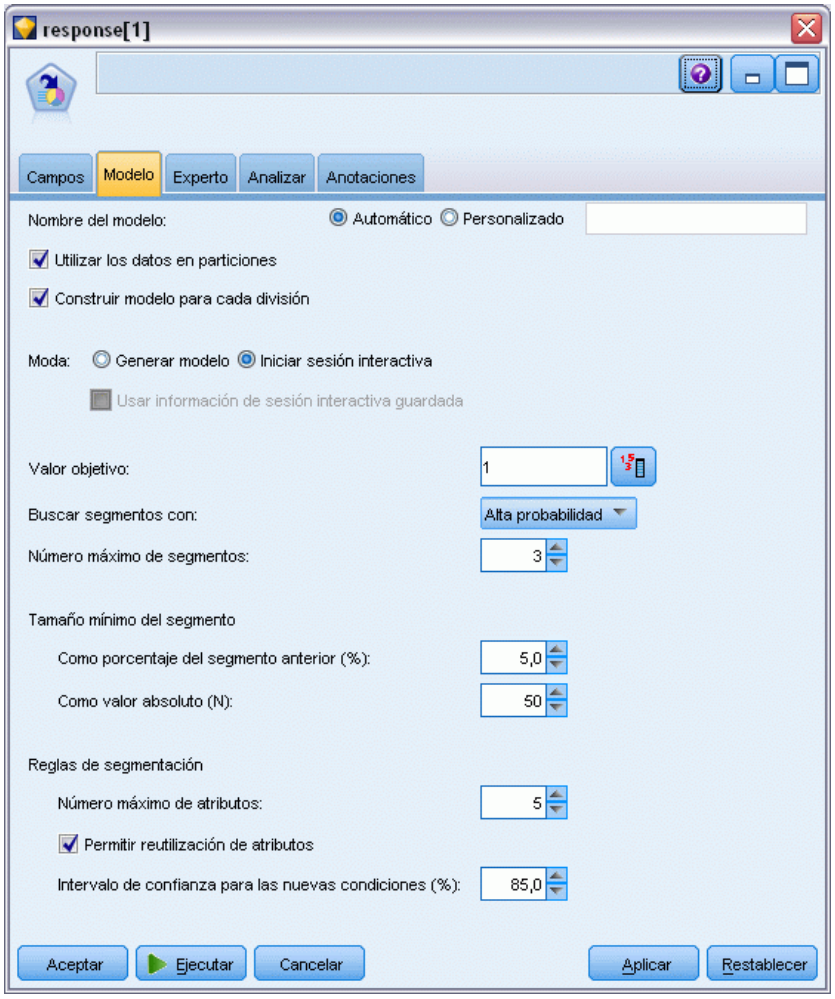
Ejemplo. El departamento de marketing de una entidad financiera desea obtener resultados más rentables en las futuras campañas adaptando la oferta adecuada a cada cliente. Puede utilizar un modelo de lista de decisiones para identificar las características de los clientes que es más probable que respondan favorablemente teniendo en cuenta las promociones anteriores y generar

una lista de mailing a partir de estos resultados. Si desea obtener más información, consulte el tema Modelado de respuesta de clientes (Lista de decisiones) en el capítulo 11 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

Requisitos. Un único campo objetivo categórico con un nivel de medición del tipo *Marca* o *Nominal* que indica el resultado binario que desea pronosticar (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

Figura 9-6
Nodo Lista de decisiones: Pestaña 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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Crear modelos divididos. Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como modelos divididos. Si desea obtener más información, consulte el tema Generación de modelos divididos en el capítulo 3 el p. 32.

Moda. 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) interactivo de Decision List Viewer, que le permite elegir entre varias alternativas y aplicar repetidamente el algoritmo con diferentes configuraciones para hacer crecer o modificar progresivamente el modelo. Si desea obtener más información, consulte el tema Decision List Viewer el p. 229.
- **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 Decision List Viewer 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 la pérdida del campo objetivo se codifica 0 = no y 1 = yes, 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 parámetro mínimo permitido es 1, no hay parámetro 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 parámetro mínimo permitido es 0, el máximo es 99,9.
- **Como valor absoluto (N).** Especifica el tamaño mínimo de grupo como número absoluto de registros. El parámetro mínimo permitido es 1, no hay parámetro máximo.

Reglas de segmentación.

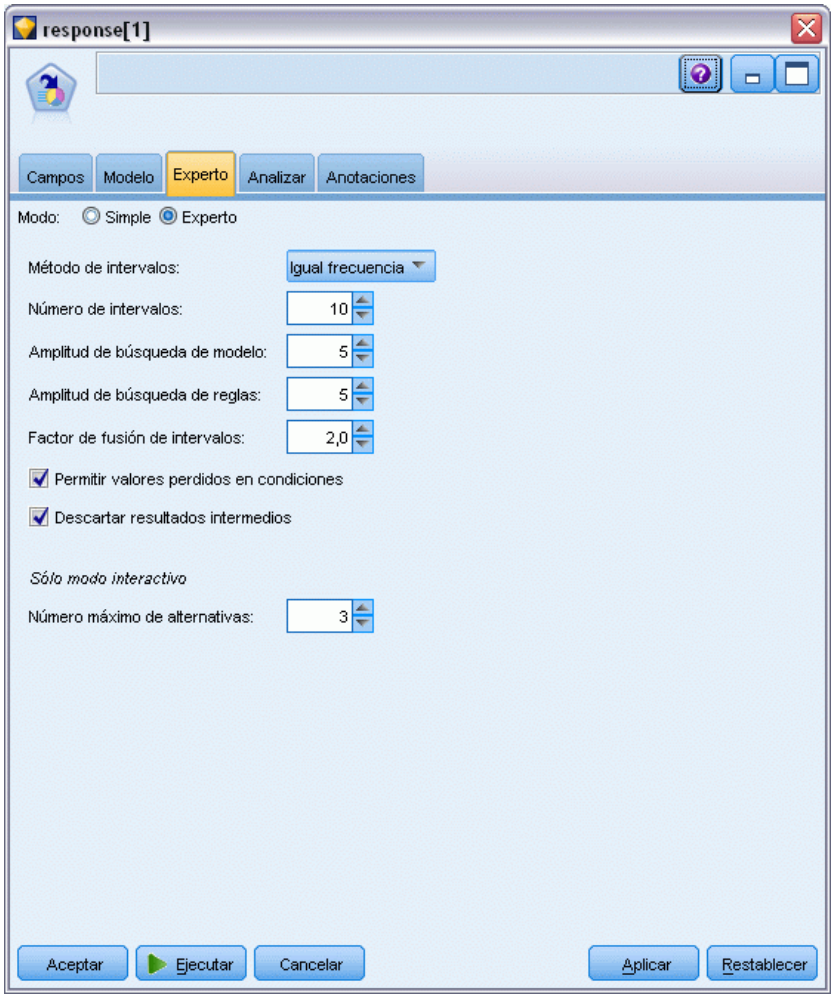
Número máximo de atributos. Especifica el número máximo de condiciones por regla de segmentación. El parámetro mínimo permitido es 1, no hay parámetro máximo.

- **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 papel 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

Figura 9-7
Nodo Lista de decisiones: Pestaña Experto



Las opciones de experto permiten ajustar 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 parámetro mínimo permitido es 2, no hay parámetro máximo.

Amplitud de búsqueda de modelo. Número máximo de resultados de modelo por ciclo que se puede utilizar para el siguiente ciclo. El parámetro mínimo permitido es 1, no hay parámetro máximo.

Amplitud de búsqueda de reglas. Número máximo de resultados de regla por ciclo que se pueden utilizar para el siguiente ciclo. El parámetro mínimo permitido es 1, no hay parámetro 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.

- **Permiten valores perdidos en las 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 parámetro mínimo permitido es 1, no hay parámetro 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.

Figura 9-8
Nugget del modelo de la lista de decisiones

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 *$null* (para los campos excluidos), la probabilidad (tasa de aciertos) del segmento al que corresponde el registro y el número de ID del segmento. Los nombres de los nuevos campos se derivan del nombre del campo de salida que se está pronosticado, con el prefijo *$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 los segmentos con valores menores al promedio de las tasas de acierto, estos segmentos "inferior a" 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 listas 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 brutas de propensión. En el caso de modelos con un objetivo de marca (que devuelve un pronóstico 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 pronóstico y confianza que pueden generarse durante la puntuación.

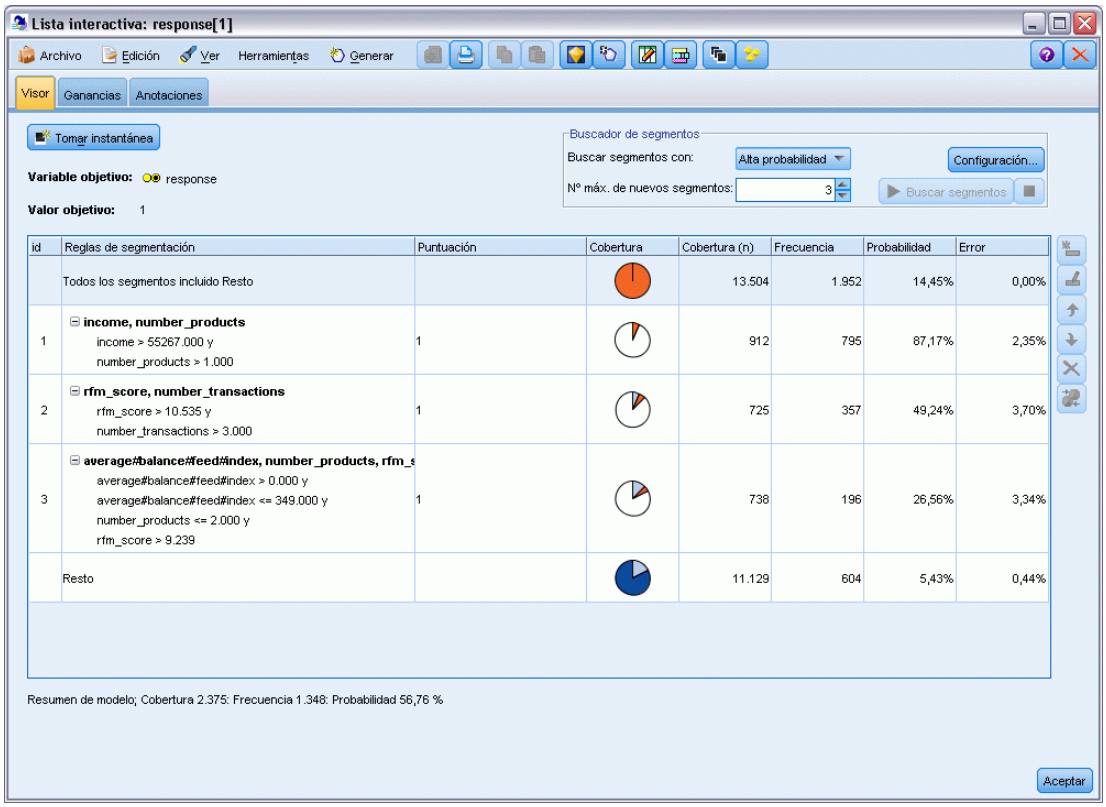
Calcular puntuaciones de propensión ajustada. Las puntuaciones brutas de propensión se basan sólo en los datos de entrenamiento y pueden ser demasiado optimistas debido a la tendencia de muchos modelos a sobreajustar estos datos. Las propensiones ajustadas intentan compensar evaluando el rendimiento 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 la ficha Analizar antes de generar el modelo.

Puntuar convirtiendo a SQL nativo. Si selecciona esta opción, se genera SQL para puntuar el modelo de manera nativa dentro de la aplicación.

Decision List Viewer

La interfaz gráfica de Decision List Viewer, 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 la configuración de los objetivos, la selección de los grupos objetivo, el análisis de los resultados y la elección del modelo óptimo.

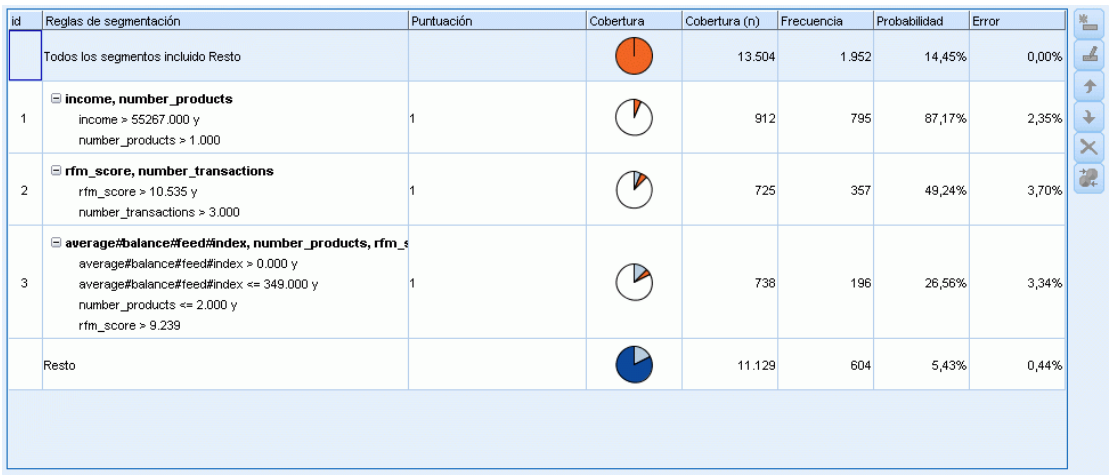
Figura 9-9
Visor interactivo de listas de decisiones



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.

Figura 9-10
Panel de 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. Por defecto, 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 pronosticar, 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.

Cubierto. El gráfico de sectores muestra visualmente la cobertura de cada segmento respecto a la cobertura total.

Cubierto (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 9-1
Botones de la barra de herramientas del modelo de trabajo

	Inicia el cuadro de diálogo Generar nuevo modelo que incluye opciones que permiten crear un nuevo nugget del modelo.
	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. Si desea obtener más información, consulte el tema Organización de las medidas del modelo el p. 248.
	Muestra el cuadro de diálogo Organizar selecciones de datos. Si desea obtener más información, consulte el tema Organización de selecciones de datos el p. 240.

	Muestra la pestaña Instantáneas. Si desea obtener más información, consulte el tema Pestaña Instantáneas el p. 234.
	Muestra la pestaña Alternativas. Si desea obtener más información, consulte el tema Pestaña Alternativas el p. 232.
	Toma una instantánea de la estructura del modelo actual. Las instantáneas aparecen en la pestaña Instantáneas y suelen utilizarse para comparar modelos.
	Inicia el cuadro de diálogo Inserción de segmentos que incluye opciones que permiten crear nuevos segmentos del modelo.
	Inicia el cuadro de diálogo Edición de reglas de segmentación que incluye opciones que permiten añadir condiciones a los segmentos del modelo o cambiar las condiciones de segmento del modelo anteriormente definidas.
	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.

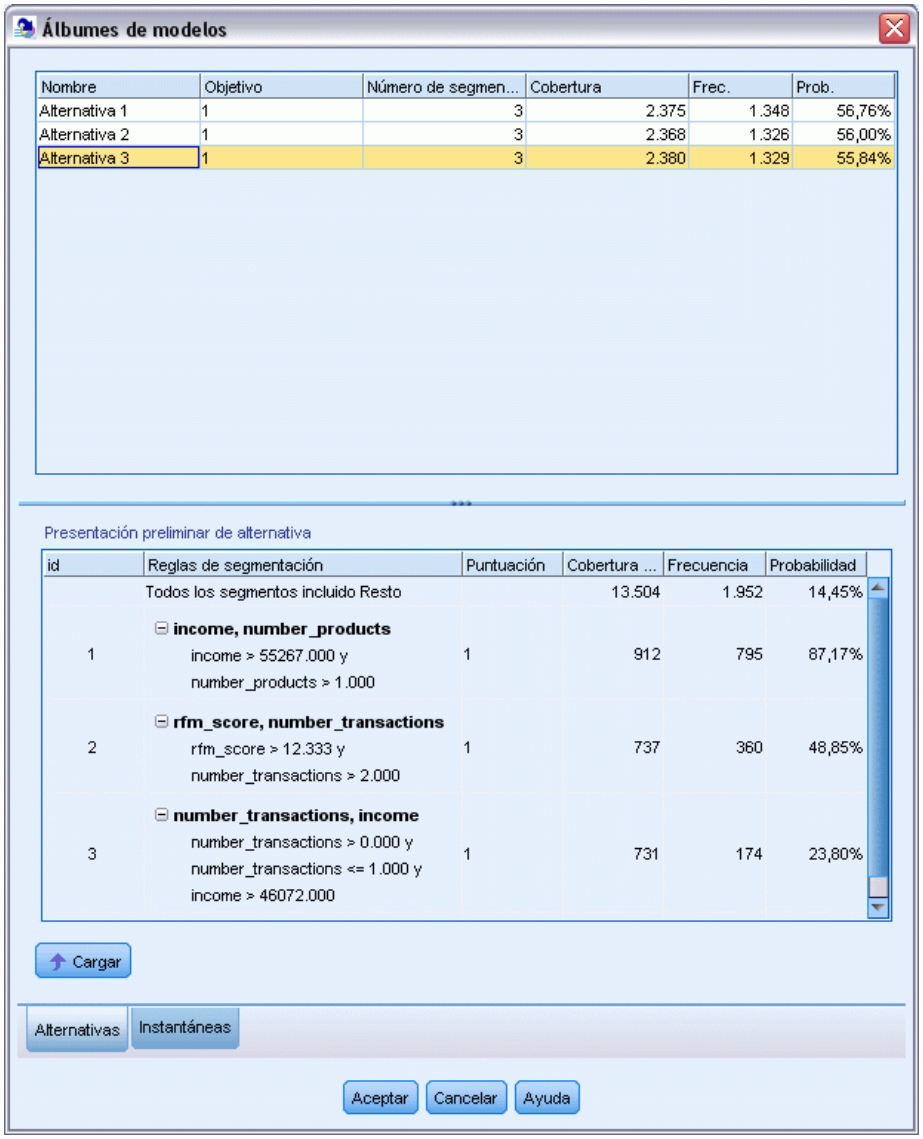
Pestaña Alternativas

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.

Figura 9-11
Pestaña Alternativas



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.

Cubierto. 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

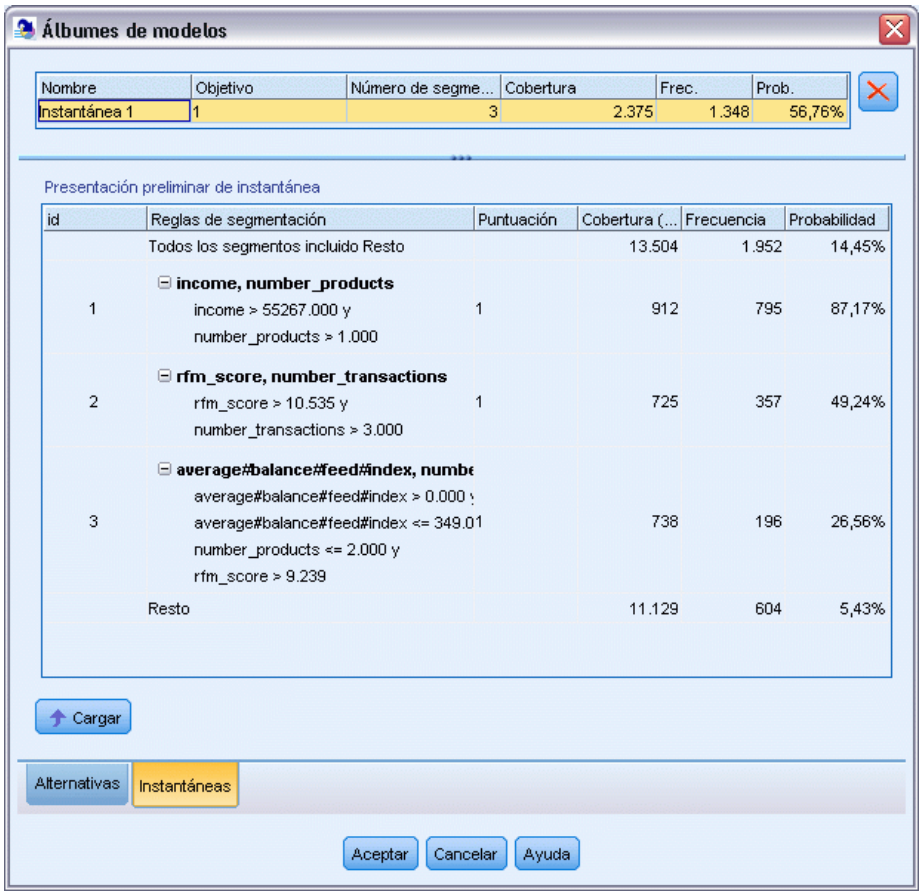
Una instantánea es una vista de un modelo en un momento determinado. Por ejemplo, puede tomar una instantánea de modelo cuando desee cargar otro modelo alternativo en el panel Modelo de trabajo, pero no quiera perder el trabajo realizado con el modelo actual. La pestaña Instantáneas muestra todas las 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

- Seleccione el modelo o la alternativa que desea que aparezca en el panel Modelo de trabajo.
- Realice todos los cambios necesarios al modelo de trabajo.
- Pulse en Tomar instantánea. Aparecerá una nueva instantánea en la pestaña Instantáneas.

Figura 9-12
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.

Cubierto. La cobertura del modelo.

Frecuencia. Muestra el número de aciertos respecto a la cobertura.

P. Indica es el porcentaje de probabilidad del modelo.

- 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.
- Puede eliminar una instantánea pulsando en Eliminar o pulsando con el botón derecho del ratón en la instantánea y seleccionando Eliminar en el menú.

Trabajo con Decision List Viewer

La generación de un modelo que pronostique de manera óptima la respuesta y el comportamiento de los clientes se realiza en varias fases. Al iniciar Decision List Viewer, 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
- Organización de selecciones de datos

Ejecución de tareas de minería

Decision List Viewer le permite añadir manualmente reglas de segmento a un modelo ejecutando tareas de minería o copiando y pegando reglas 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 pronosticar 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:

▶ 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:

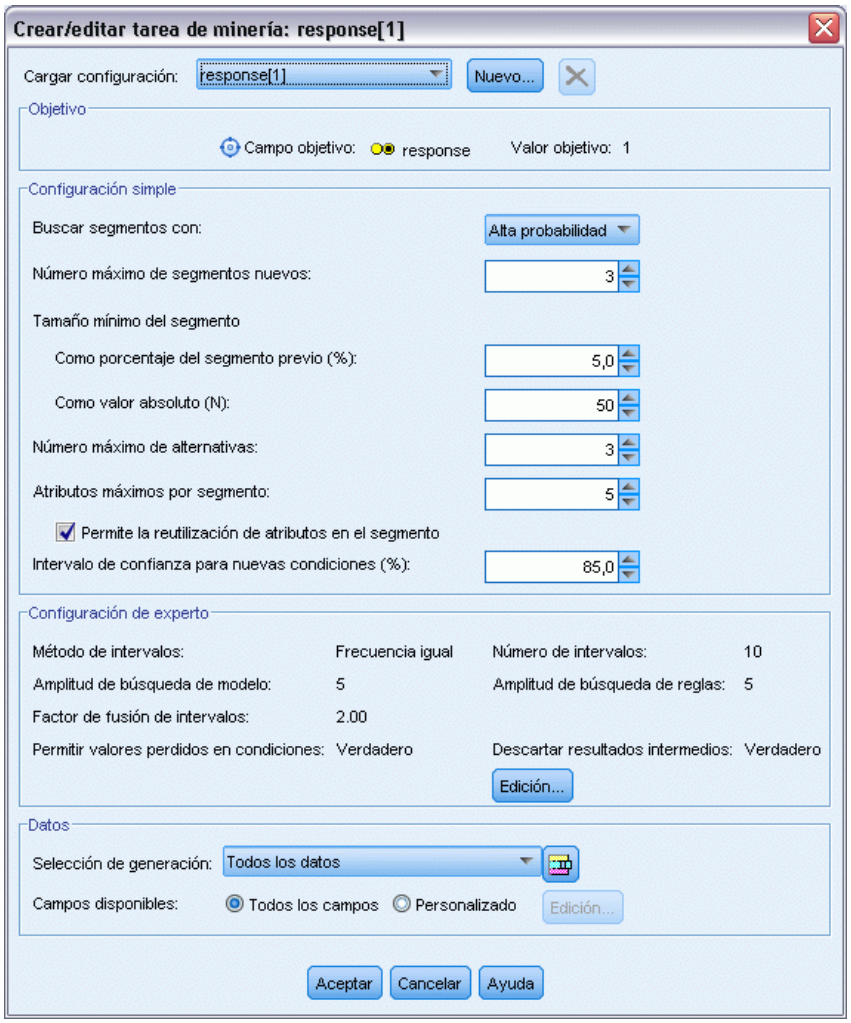
- Seleccione el segmento a partir del que desea buscar condiciones de segmento adicionales.
- Pulse en Configuración. Aparecerá el cuadro de diálogo Crear/editar tarea de minería. El cuadro de diálogo ofrece las opciones para definir la tarea de minería.
- Realice todos los cambios necesarios y pulse en Aceptar para volver al cuadro de diálogo Organizar tareas de minería. Decision List Viewer utiliza la configuración como ajustes por defecto para ejecutar todas las tareas hasta que se selecciona una tarea o configuración alternativa.
- 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. Si desea obtener más información, consulte el tema Opciones del modelo de la lista de decisiones el p. 224.

Figura 9-13
Cuadro de diálogo Crear/editar tarea de minería



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.

Destino

Campo objetivo: Representa el campo que se desea pronosticar, 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 la pérdida del campo objetivo se codifica 0 = no y 1 = yes, 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 parámetro mínimo permitido es 1, no hay parámetro máximo.

Configuración de experto

Editar... Abre el cuadro de diálogo Editar parámetros avanzados que permite definir la configuración avanzada. Si desea obtener más información, consulte el tema Editar parámetros avanzados el p. 239.

Data

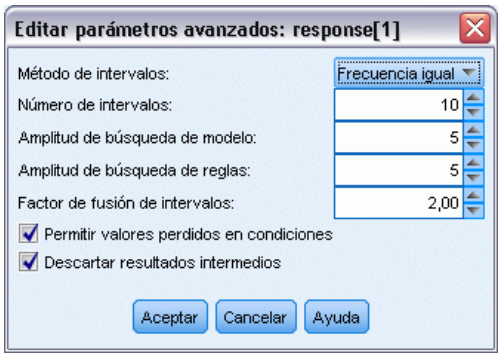
Selección de generación. Incluye opciones que permiten especificar la medida de evaluación que Decision List Viewer analizará para buscar nuevas reglas. Las medidas de evaluación de la lista se crean y editan en el cuadro de diálogo Organizar selecciones de datos.

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 abrirá el cuadro de diálogo Personalizar campos disponibles que le permite seleccionar los campos disponibles como atributos de segmentos que ha encontrado la tarea de minería. Si desea obtener más información, consulte el tema Personalizar campos disponibles el p. 240.

Editar parámetros avanzados

Figura 9-14
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 parámetro mínimo permitido es 2, no hay parámetro máximo.

Amplitud de búsqueda de modelo. Número máximo de resultados de modelo por ciclo que se puede utilizar para el siguiente ciclo. El parámetro mínimo permitido es 1, no hay parámetro máximo.

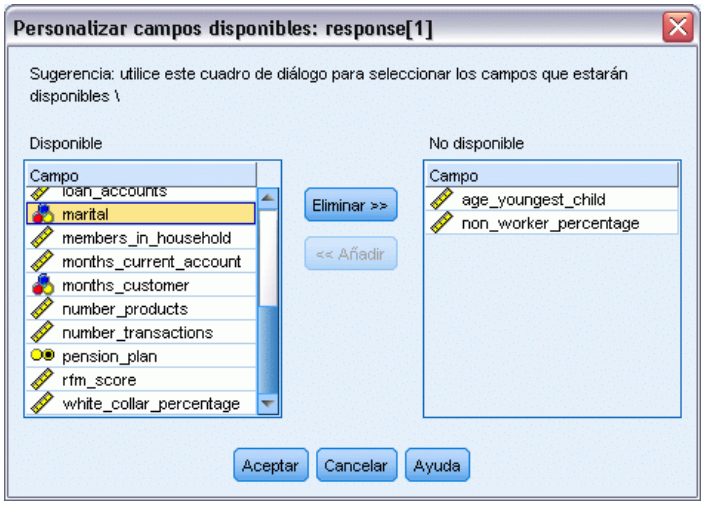
Amplitud de búsqueda de reglas. Número máximo de resultados de regla por ciclo que se pueden utilizar para el siguiente ciclo. El parámetro mínimo permitido es 1, no hay parámetro máximo.

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.

- **Permiten valores perdidos en las 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

Figura 9-15
Cuadro de diálogo Personalizar campos disponibles



El cuadro de 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.

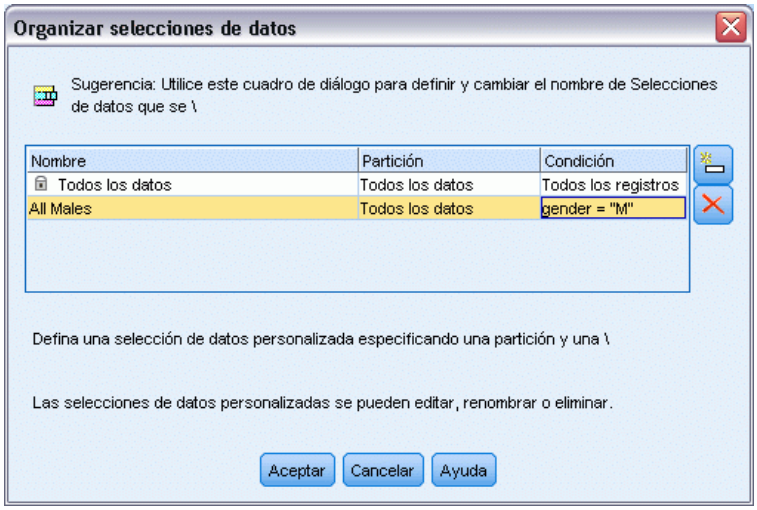
Organización de selecciones de datos

Mediante la organización de las selecciones de datos (el conjunto de datos de minería), puede especificar las medidas de evaluación que Decision List Viewer 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:

- 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.

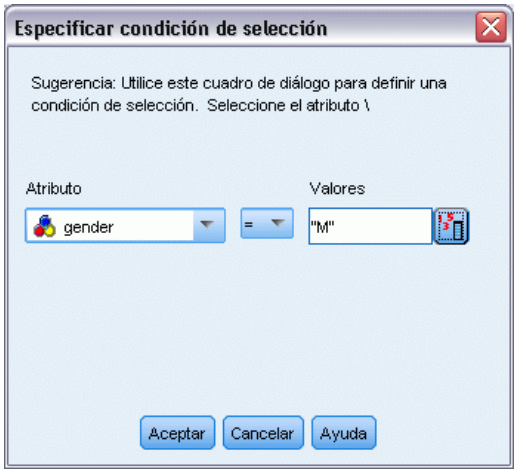
Figura 9-16
Cuadro de diálogo Organizar selecciones de datos



Nota: el cuadro de diálogo Organizar selecciones de datos también le permite editar o eliminar selecciones de datos existentes.

- Pulse en el botón Añadir nueva selección de datos. Se añadirá una nueva entrada de selección de datos a la tabla existente.
- Pulse en Nombre y escriba el nombre de la selección.
- Pulse en Partición y seleccione el tipo de partición.
- 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.

Figura 9-17
Cuadro de diálogo Especificar condición de selección



▶ 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. Esta 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:

▶ 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.

- 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.

- Pulse en Insertar. El cuadro de diálogo Insertar condición se abrirá, permitiéndole definir los atributos para la nueva condición de regla.

- 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 region = 'TOWN' aparece en rojo, indica que TOWN se excluye del conjunto de resultados.

- Introduzca uno o más valores o pulse en el icono Insertar valor para acceder al cuadro de diálogo Insertar valor. Este cuadro de diálogo permite elegir un valor definido para el campo seleccionado. Por ejemplo, el campo casado ofrecerá los valores sí y no.

- 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:

- Seleccione el segmento del modelo que desea editar.

- 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.

- 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.

- 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 region = 'TOWN' aparece en rojo, indica que TOWN se excluye del conjunto de resultados.

- Introduzca uno o más valores o pulse en el botón Insertar valor para acceder al cuadro de diálogo Insertar valor. Este cuadro de diálogo permite elegir un valor definido para el campo seleccionado. Por ejemplo, el campo casado ofrecerá los valores sí y no.
- 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:

- Seleccione el segmento del modelo que contiene las condiciones de reglas que desea eliminar.
- 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 cuadro de diálogo Editar regla de segmentación, que le permite eliminar una o más condiciones de reglas de segmentación.

- Seleccione la condición de regla adecuada y pulse en Eliminar.
- 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

Decision List Viewer 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 Decision List Viewer, 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:

- Seleccione el segmento del modelo que desea utilizar en otro modelo.
- 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**.
- Abra el modelo en el que desea pegar el segmento del modelo.

- Seleccione uno de los segmentos del modelo y pulse en **Pegar**.

  Nota: en lugar de los comandos **Cortar**, **Copiar** y **Pegar**, también puede utilizar las siguientes 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:

- 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.
- 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.

Decision List Viewer 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. Por defecto, los segmentos del modelo aparecen por orden de prioridad, siendo el primer segmento el que tiene la mayor prioridad. Cuando se asigna una prioridad 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 una prioridad mayor o menor.

Para asignar una prioridad a los segmentos del modelo:

- Seleccione el segmento del modelo al que desea asignar una prioridad diferente.
- 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:

- Seleccione un segmento del modelo.
- 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 distribuir 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:

- Seleccione un segmento en el panel del modelo de trabajo.
- 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 cuadro de diálogo Cambiar valor objetivo le permite cambiar el valor 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 único 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 Decision List Viewer, aparecerá la configuración realizada en un principio utilizando el Visor.

Si desea obtener más información, consulte el tema Regeneración de un nodo de modelado en el capítulo 3 el p. 70.

Valoración del modelo

Para crear con éxito un modelo, es necesario evaluar cuidadosamente el modelo antes de implementarlo en el entorno de producción. Decision List Viewer 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 de la distribución.

Puede evaluar el modelo de las siguientes formas:

- Utilizando las medidas estadísticas y empresariales predefinidas que ofrece Decision List Viewer (probabilidad, frecuencia).
- Evaluando las medidas importadas de Microsoft Excel.
- Visualizando el modelo mediante un gráfico de ganancias.

Organización de las medidas del modelo

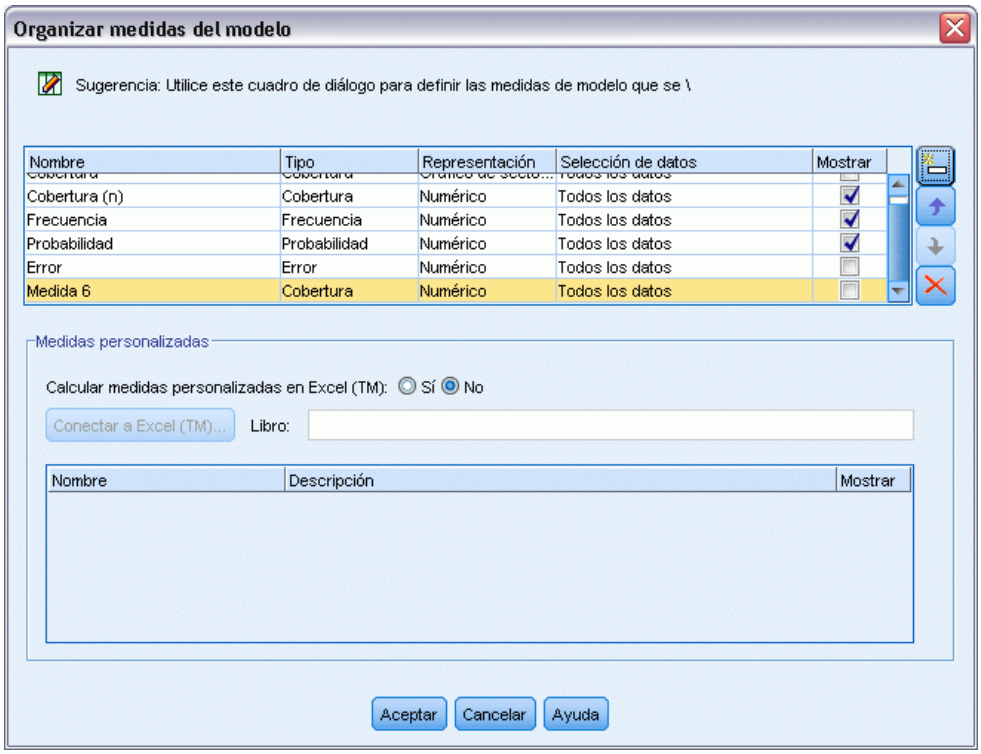
Decision List Viewer incluye opciones que permiten definir las medidas que se calculan y muestran como columnas. Cada segmento puede incluir la cobertura por defecto, la frecuencia, la probabilidad y las medidas de error representadas como columnas. También puede crear nuevas medidas que aparecerán como columnas.

Definición de las medidas del modelo

Para añadir una medida al modelo o definir una medida existente:

▶ 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.

Figura 9-18
Organizar medidas del modelo



▶ Pulse en el botón Añadir nueva medida de modelo (a la derecha de la columna Mostrar). Se mostrará una nueva medida en la tabla.

- 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.

- Pulse en Aceptar para volver al espacio de trabajo de Decision List Viewer. Si se activó la columna Mostrar para la nueva medida, aparecerá la nueva medida para el modelo de trabajo.

Métricas personalizadas en Excel

Si desea obtener más información, consulte el tema Evaluación en Excel el p. 249.

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

Decision List Viewer puede integrarse con Microsoft Excel, lo que le permite utilizar sus propios cálculos de los valores y las fórmulas de beneficios directamente en el proceso de generación del modelo, con el fin de simular escenarios de coste/beneficio. El enlace con Excel permite exportar datos a Excel, donde se pueden utilizar para crear gráficos para presentaciones y calcular medidas personalizadas (por ejemplo, medidas de beneficio completo y rendimiento de la inversión), así como verlos en Decision List Viewer durante la generación del modelo.

Si desea obtener más información, consulte el tema Cálculo de las medidas personalizadas con Excel en el capítulo 11 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

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 Decision List Viewer con Microsoft Excel. La configuración figura en un archivo de hoja de cálculo de Excel y especifica la información que se transfiere de Decision List Viewer 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:

- Abra el modelo, ejecute una sesión interactiva y seleccione Organizar medidas del modelo desde el menú Herramientas.
- Seleccione Sí para la opción Calcular medidas personalizadas en Excel. Se activará el campo Libro, que permite seleccionar una plantilla de libro de Excel preconfigurada.
- Pulse en el botón Conectar a Excel. Aparecerá el cuadro de diálogo Abrir, que permite buscar la ubicación de la plantilla preconfigurada en el sistema de archivos de red o local.
- 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.
- Seleccione la correspondencia adecuada 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 Decision List Viewer.

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. Si desea obtener más información, consulte el tema Pestaña Instantáneas el p. 234. Las métricas de Excel no aparecen en las instantáneas hasta que se vuelve a establecer una conexión con la plantilla de Excel.

Configuración de integración de MS Excel

La integración entre Decision List Viewer 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 Decision List Viewer 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 Decision List Viewer importadas y las medidas de Excel personalizadas.

Configuración. Incluye opciones que permiten especificar las medidas que se importarán de Decision List Viewer y definir las medidas de Excel personalizadas.

ADVERTENCIA: 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 Decision List Viewer que se utilizarán en los cálculos.
- **Métricas al modelo.** Indica las métricas generadas en Excel que se devolverán a Decision List Viewer. Las métricas generadas por Excel aparecen en Decision List Viewer 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

- Abra la plantilla y seleccione la hoja de trabajo Configuración.
- 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 y se mostrará en el modelo, permitiendo al usuario seleccionar la medida apropiada para el mapa.

Importar una medida estándar adicional desde el modelo

- Abra la plantilla y seleccione la hoja de trabajo Configuración.
- Seleccione en los menús:
  Herramientas > Protección > Hoja no protegida
- Seleccione la casilla A5, que se encuentra sombreada y contiene la palabra End.
- Seleccione en los menús:
  Insertar > Filas
- Introduzca el Nombre y Descripción de la nueva medida. Por ejemplo, Error se convierte Error asociado con segmento.
- En la casilla C5, introduzca la fórmula =COLUMN('Medidas del modelo'!N3).
- 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.
- Seleccione en los menús:
  Herramientas > Protección > Hoja protegida
- Pulse en Aceptar.
- En la hoja de trabajo Medidas del modelo, asegúrese de que la casilla N3 tiene Error como título de la nueva columna.
- Seleccione toda la columna N.

- Seleccione en los menús:
  Formato > Casillas
- Por defecto, 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.
- Pulse en Aceptar.
- 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

- Abra la plantilla a la que ha añadido la columna Error en el ejemplo anterior; seleccione la hoja de trabajo Configuración.
- Seleccione en los menús:
  Herramientas > Protección > Hoja no protegida
- Seleccione la casilla A14, que se encuentra sombreada y contiene la palabra End.
- Seleccione en los menús:
  Insertar > Filas
- Introduzca el Nombre y Descripción de la nueva medida. Por ejemplo, Error escalado y Escala aplicada a un error de Excel.
- En la casilla C14, introduzca la fórmula =COLUMN('Medidas del modelo'!O3).
- 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.
- Seleccione la hoja de cálculo Parámetros.
- En la casilla A17, introduzca la descripción '- Error escalado.
- En la casilla B17, introduzca el factor de escala 10.
- En la hoja de trabajo Medidas del modelo, introduzca la descripción Error escalado en la casilla O3 como título de la nueva columna.
- En la casilla O4, introduzca la fórmula =N4*Settings!B17.
- Seleccione la esquina de la casilla O4 y arrástrela a la casilla O22 para copiar la fórmula en cada casilla.
- Seleccione en los menús:
  Herramientas > Protección > Hoja protegida
- Pulse en Aceptar.

- 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 Gráfico de ganancias muestra las ventajas que ofrecen los modelos respecto a la toma aleatoria de decisiones y permite comparar directamente varios gráficos cuando hay modelos alternativos.

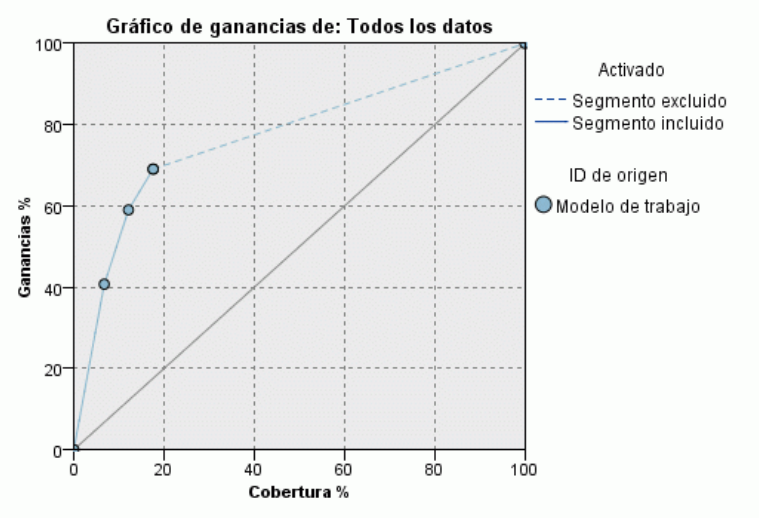
Gráfico de ganancias

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.

Figura 9-19
Pestaña Ganancias



Para ver un gráfico de ganancias:

- Abra una ruta que contenga un nodo Lista de decisiones e inicie una sesión interactiva desde dicho nodo.
- 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).

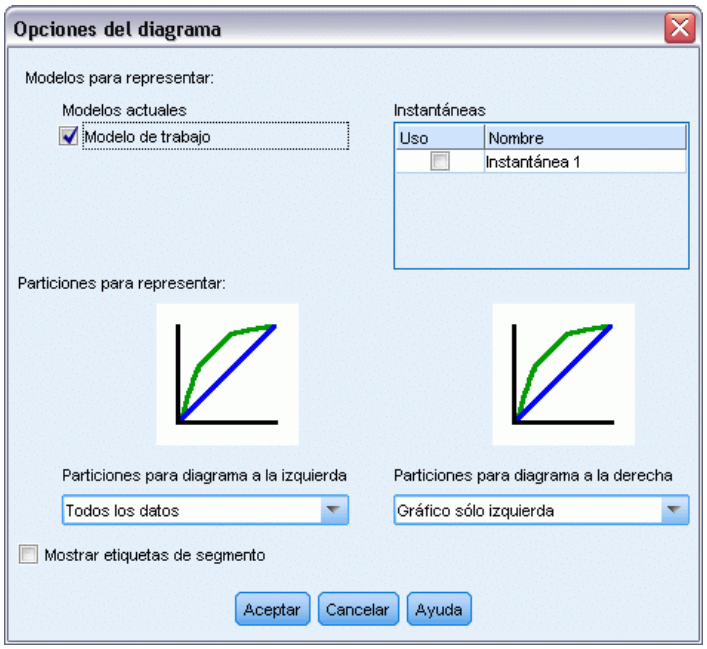
  Por defecto, los gráficos aparecen como segmentos. Si desea que los gráficos aparezcan como cuantiles, seleccione Cuantiles y, a continuación, seleccione el método de cuantiles en el menú desplegable.

  Nota: consulte Edición de visualizaciones si desea obtener información sobre cómo trabajar con gráficos.

Opciones de gráfico

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.

Figura 9-20
Cuadro de diálogo Opciones del diagrama



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 destino continuo tomando como base las relaciones lineales entre el destino y uno o más predictores. Si desea obtener más información, consulte el tema Modelos lineales el p. 258.

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. Si desea obtener más información, consulte el tema Nodo Logística el p. 278.

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. Si desea obtener más información, consulte el tema Nodo PCA/Factorial el p. 299.

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. Si desea obtener más información, consulte el tema Nodo Discriminante el p. 307.

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 que no sea 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. Si desea obtener más información, consulte el tema Nodo GenLin el p. 316.

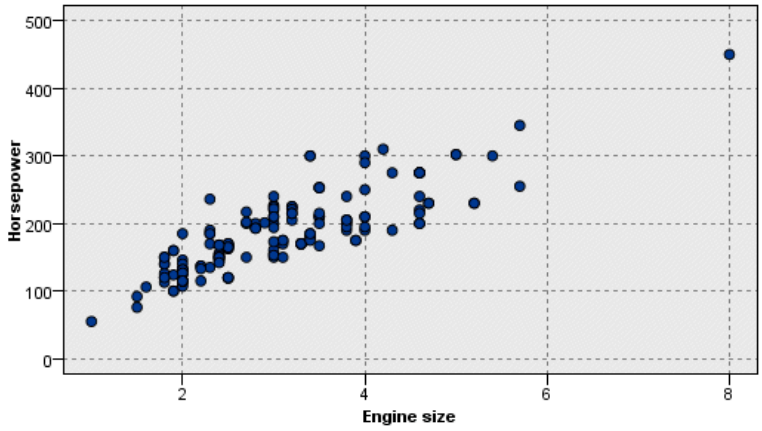
Un modelo lineal mixto generalizado (GLMM) amplía el modelo lineal de modo que el objetivo pueda tener una distribución no normal, esté linealmente relacionado con los factores y covariables mediante una función de enlace especificada y las observaciones se puedan correlacionar. Los modelos lineales mixtos generalizados cubren una amplia variedad de modelos, desde modelos de regresión lineal simple hasta modelos multinivel complejos para datos longitudinales no normales. Si desea obtener más información, consulte el tema Nodo GLMM el p. 331.

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 pronostica la probabilidad de que el evento de interés se haya producido en el momento dado (*t*) para valores determinados de las variables de entrada. Si desea obtener más información, consulte el tema Nodo Cox el p. 359.

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 pronosticados y reales.

Figura 10-1
Gráfico de regresión lineal simple



Requisitos. Sólo se pueden utilizar campos numéricos en un modelo de regresión lineal. Debe tener exactamente un campo objetivo (con el papel definido a *Objetivo*) y uno o más predictores (con el papel definido a *Entrada*). Los campos con un papel *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. Si desea obtener más información, consulte el tema Nueva codificación de valores con el nodo Derivar en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.)

Puntos fuertes. Los modelos de regresión lineal son relativamente simples y proporcionan una fórmula matemática fácil de interpretar para la creación de pronósticos. 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 *perder*/*no perder*, 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. Si desea obtener más información, consulte el tema Nodo Logística el p. 278.

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 comprenden bien y se pueden crear rápidamente en comparación con el resto de tipos de modelos (como redes neuronales o árboles de decisión) en el 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 implementar este modelo en centros de servicios, los representantes pueden introducir información sobre reclamaciones mientras atienden por teléfono al cliente y obtienen inmediatamente el coste "pronosticado" de la reclamación en función de los datos pasados.

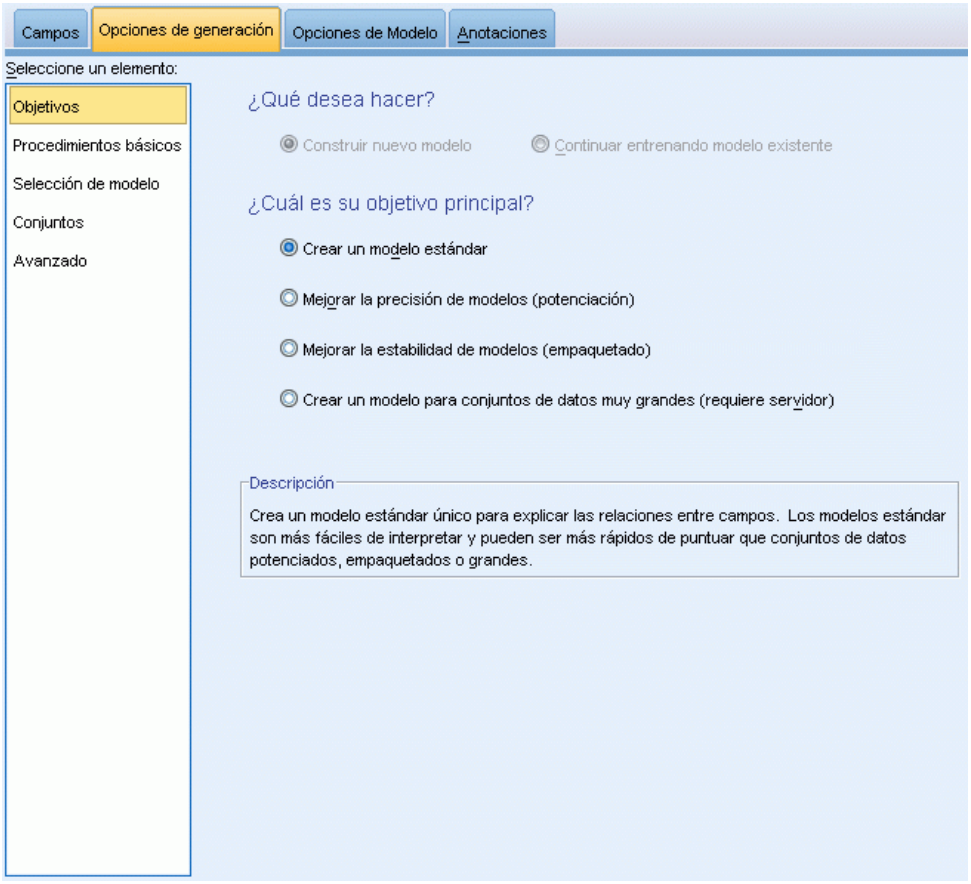
Figura 10-2
Pestaña Campos

Requisitos del campo. Debe haber un objetivo y al menos una entrada. Por defecto los campos con los papeles 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.Si desea obtener más información, consulte el tema Opciones de los campos del nodo de modelado en el capítulo 3 el p. 38.

Objetivos

Figura 10-4
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 se activa esta opción, se desactivan el resto de los controles de las pestañas Campos y Opciones de creación.

¿Cuál es su objetivo principal?

- **Crear un modelo estándar.** El método genera un modelo simple para pronosticar el destino mediante los predictores. Por lo general, los modelos estándar son más fáciles de interpretar y pueden puntuarse más rápido que conjuntos de datos de gran tamaño o de aumento o agregación autodocimante.
- **Mejorar la precisión del modelo (aumento).** El método genera un modelo de conjunto mediante el aumento, que genera una secuencia de modelos para obtener pronósticos más precisos. Se puede tardar más tiempo en generar y puntuar conjuntos que un modelo estándar.

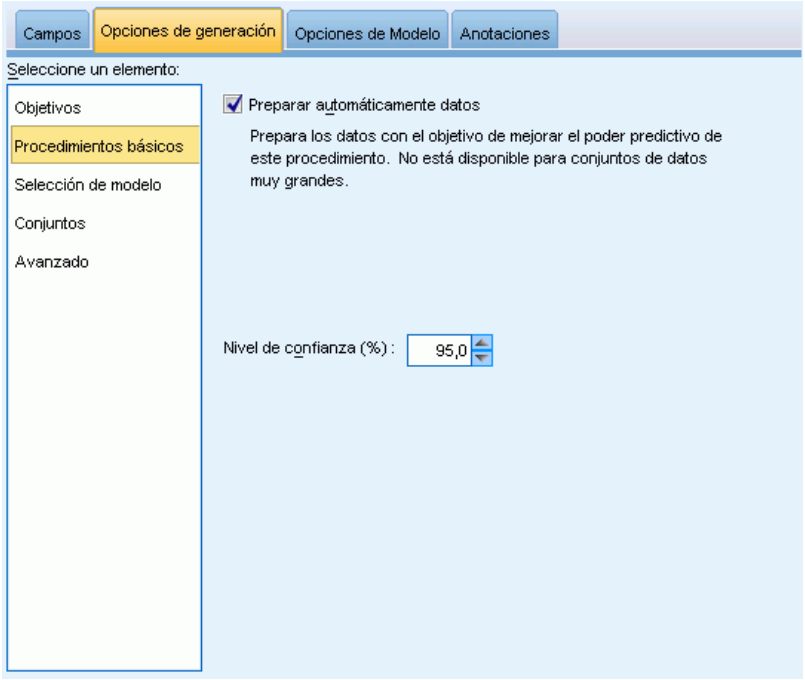
  El aumento produce una sucesión de "modelos de componente", cada uno de ellos basados en el conjunto de datos completo. Antes de crear cada modelo de componente sucesivo, los registros se ponderan en función de los residuos del modelo del componente anterior. Los casos con residuos de grandes dimensiones tienen ponderaciones de análisis relativamente superiores para que el siguiente modelo de componente se centre en pronosticar 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 pronósticos 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 autodocimantes 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.
- **Cree un modelo para conjuntos de datos muy grandes (necesita IBM® SPSS® Modeler Server).** 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 generar uno 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. Esta opción requiere conexión al servidor SPSS Modeler Server.

Consulte Conjuntos el p. 264 para configuraciones relacionadas con opciones de aumento, agregación autodocimante y conjuntos de datos de gran tamaño.

Conceptos básicos

Figura 10-5
Configuración básica



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. Por defecto, se realiza la siguiente preparación automática de datos.

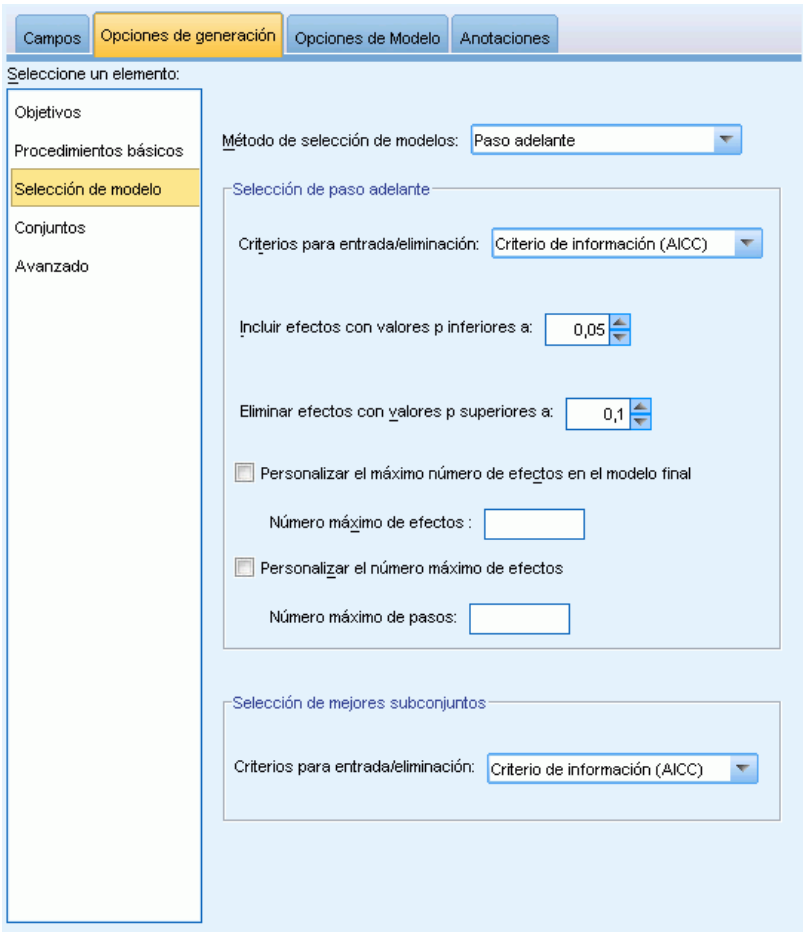
- **Fecha y hora.** Cada predictor de fecha se transforma en un nuevo predictor continuo que contiene el tiempo transcurrido desde una fecha de referencia (01-01-1970). Cada predictor de hora se transforma en un nuevo predictor continuo que contiene el tiempo transcurrido desde una hora de referencia (00:00:00).
- **Ajustar nivel de medida.** 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 típicas 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 por defecto es 95.

Selección de modelos

Figura 10-6
Configuración de 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. Por defecto, 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.** Por defecto, 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. Por defecto, es 3 veces el número de efectos disponibles. Del mismo modo, especifique un entero positivo para el número máximo de pasos.

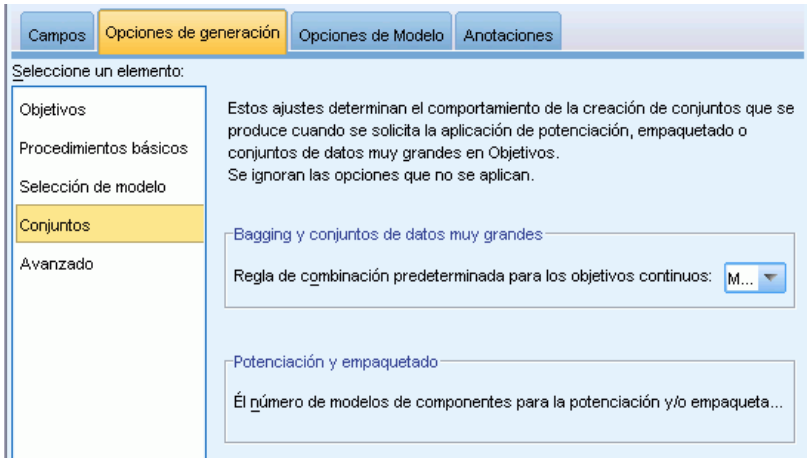
Selección de mejores subconjuntos. Comprueba "todos los modelos posibles", o al menos el subconjunto más grande de los modelos posibles que los pasos sucesivos hacia adelante, para seleccionar el mejor según el criterio de mejores subconjuntos. Criterio de información (AICC) se basa en la similitud del 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

Figura 10-7
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 pronosticados a partir de los modelos básicos para calcular el valor de puntuación del conjunto.

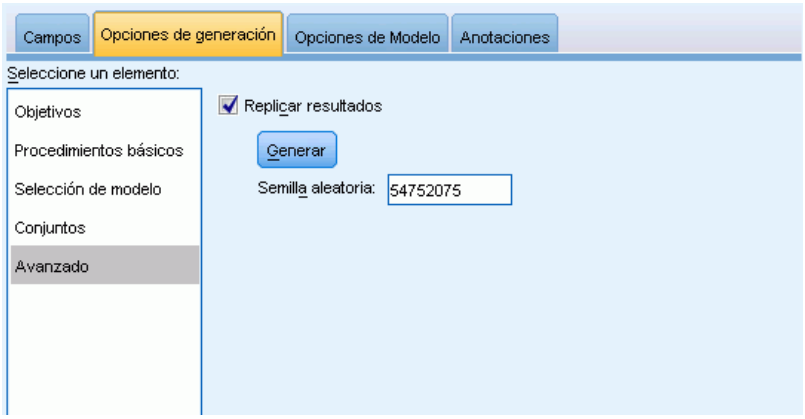
- **Regla de combinación predeterminada para objetivos continuos.** Los valores pronosticados de conjunto para objetivos continuos pueden combinarse mediante la media o mediana de los valores pronosticados a partir de los modelos básicos.

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 autodocimantes. Debe ser un número entero positivo.

Avanzado

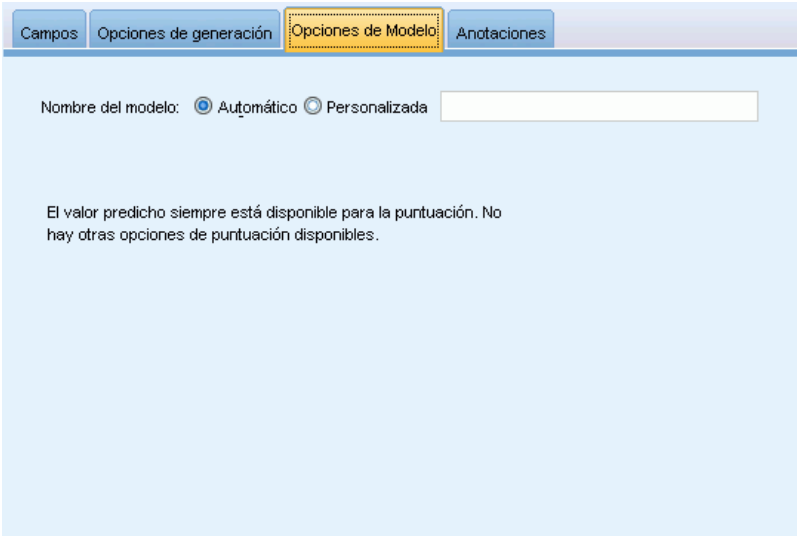
Figura 10-8
Configuración avanzada



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

Figura 10-9
Pestaña Opciones de modelo

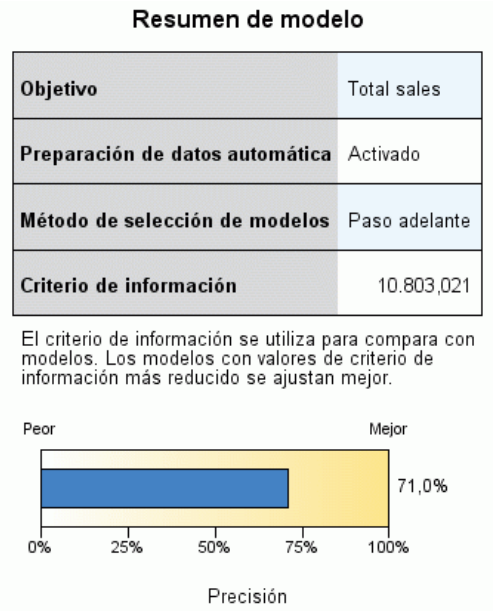


Nombre del modelo. Puede generar el nombre del modelo automáticamente tomando como base los campos objetivo o especificar un nombre personalizado. El nombre generado automáticamente es el nombre del campo objetivo.

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

Figura 10-10
Vista Resumen del modelo



La vista Resumen del modelo es una instantánea, un resumen visual del modelo y su ajuste.

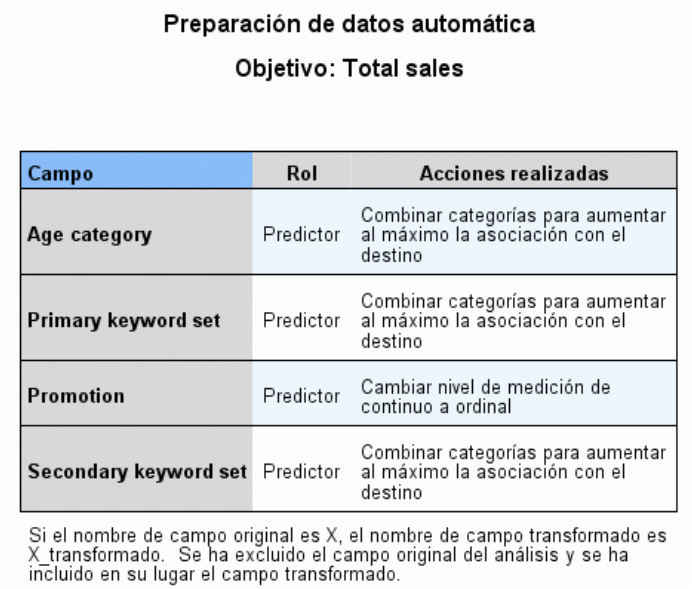
Tabla. La tabla identifica algunos parámetros de modelo superior, incluyendo:

- El nombre del destino especificado en la pestaña Campos,
- Si se ha ejecutado la preparación de datos automática tal y como se ha especificado en los ajustes de Básico,
- El método y el criterio de selección de modelo especificado en los ajustes de Selección de modelos. 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

Figura 10-11
Vista Preparación de datos automática



Preparación de datos automática

Objetivo: Total sales

Campo	Rol	Acciones realizadas
Age category	Predictor	Combinar categorías para aumentar al máximo la asociación con el destino
Primary keyword set	Predictor	Combinar categorías para aumentar al máximo la asociación con el destino
Promotion	Predictor	Cambiar nivel de medición de continuo a ordinal
Secondary keyword set	Predictor	Combinar categorías para aumentar al máximo la asociación con el destino

Si el nombre de campo original es X, el nombre de campo transformado es X_transformado. Se ha excluido el campo original del análisis y se ha incluido en su lugar el campo transformado.

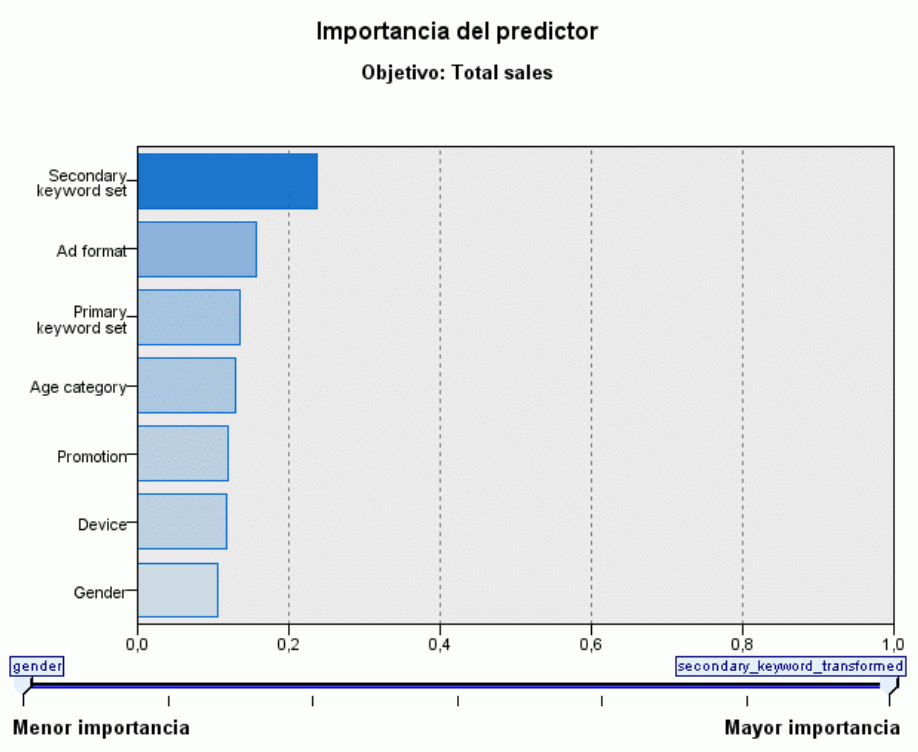
Esta vista muestra información a cerca de qué campos se excluyen y cómo los campos transformados se derivaron en el paso de preparación automática de datos (ADP). Para cada campo que fue transformado o excluido, la tabla enumera el nombre del campo, su papel en el análisis y la acción tomada por el paso 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 únicos como campos ordinales.
- Cambiar nivel de medición de ordinal a continuo reestructura los campos continuos con más de 10 valores únicos como campos continuos.
- Recortar valores atípicos define los valores de los predictores continuos que recaen más allá de un valor de corte (3 desviaciones típicas 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 valor único, posiblemente después de tomar otras acciones ADP.

Importancia de predictor

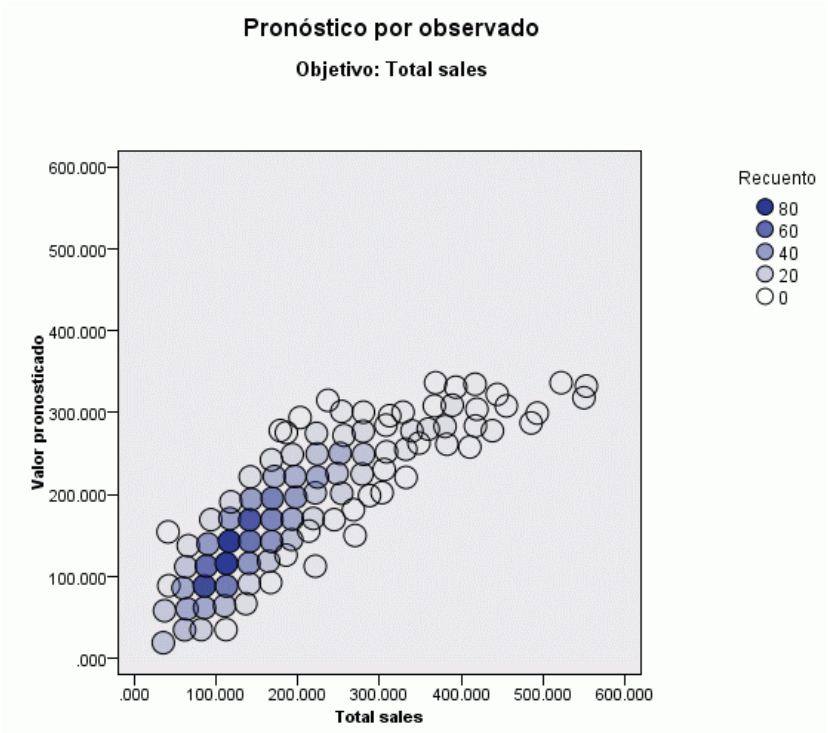
Figura 10-12
Vista de importancia del predictor



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.

Predicho por observado

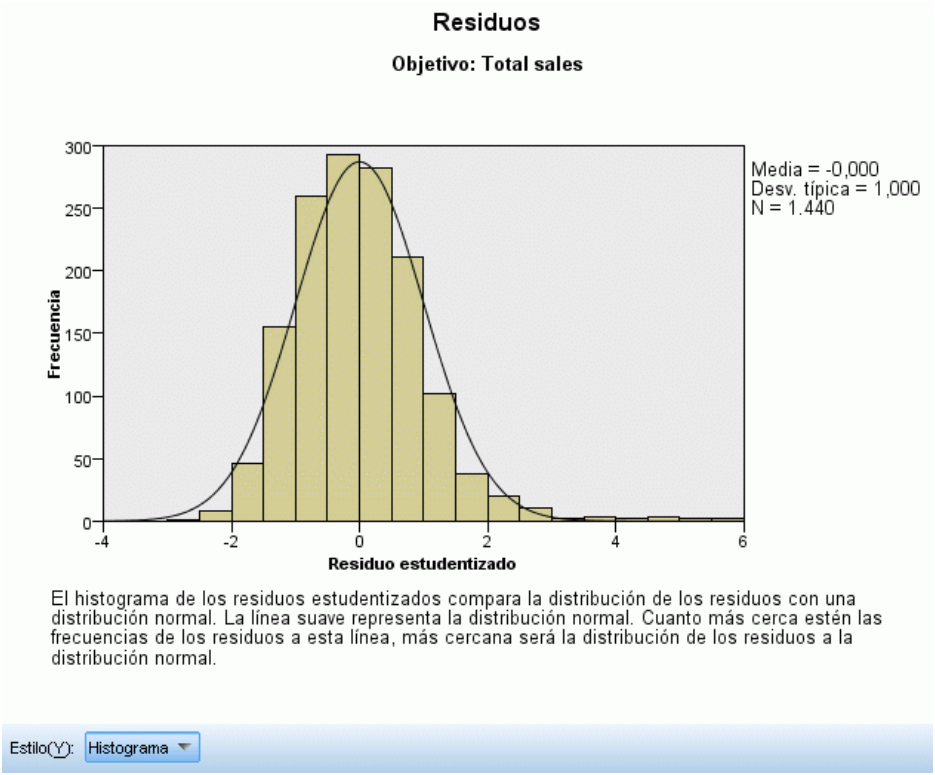
Figura 10-13
Vista Predicho por observado



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 pronosticado de manera incorrecta en el modelo.

Residuos

Figura 10-14
Vista de residuos, estilo histograma



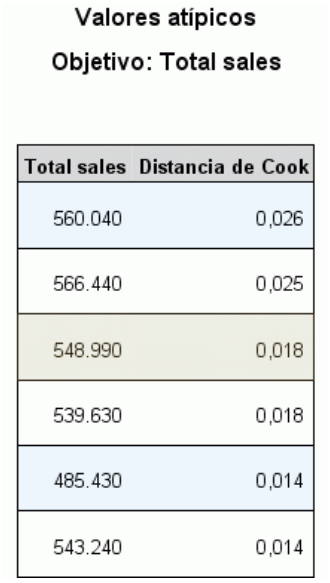
Muestra un gráfico de diagnosis de los residuos del modelo.

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

Figura 10-15
Vista Valores atípicos



Valores atípicos

Objetivo: Total sales

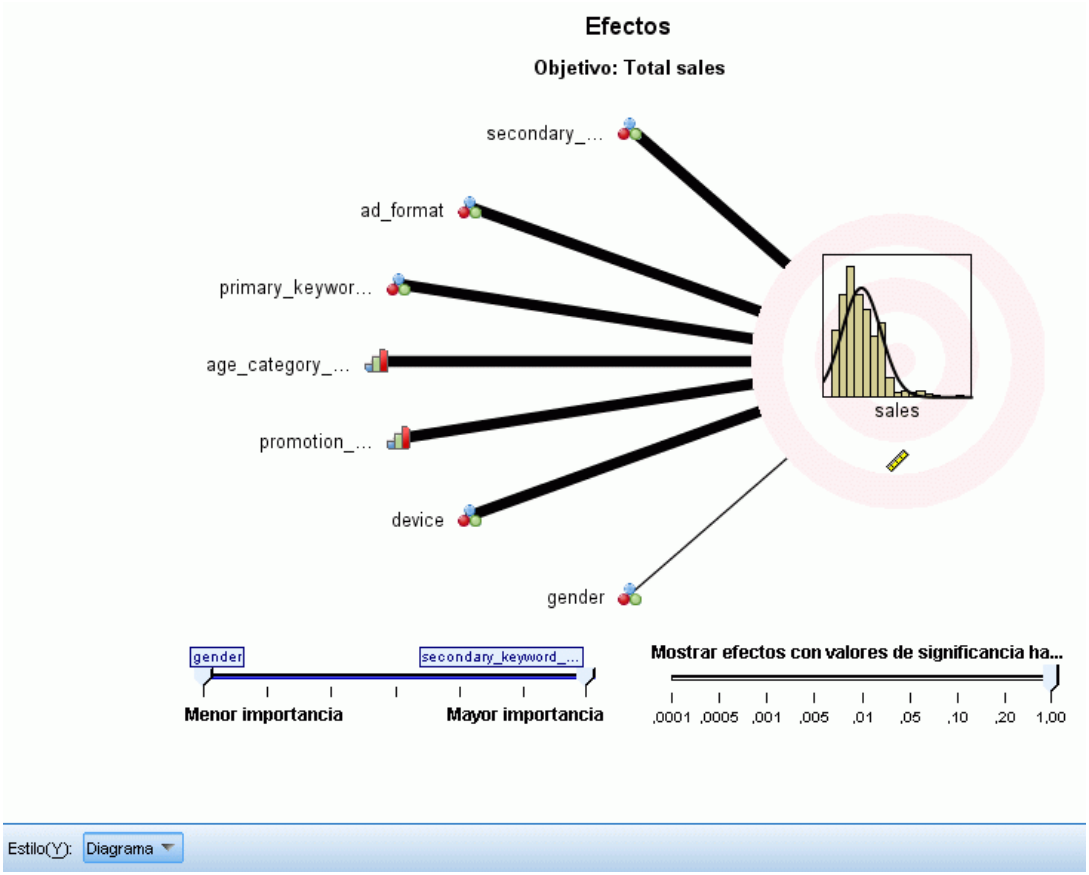
Total sales	Distancia de Cook
560.040	0,026
566.440	0,025
548.990	0,018
539.630	0,018
485.430	0,014
543.240	0,014

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

Figura 10-16
Vista de efectos, estilo de diagrama



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.

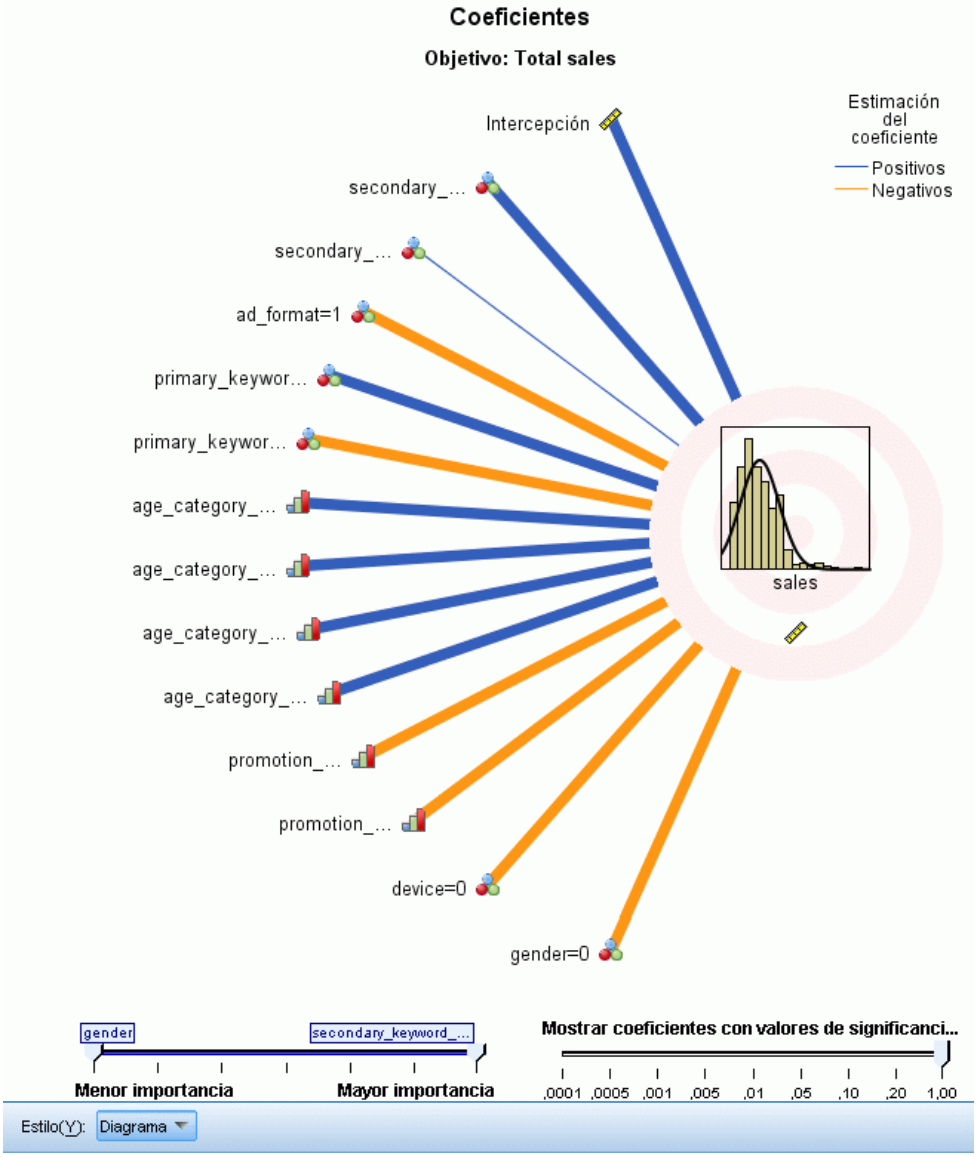
- **Diagrama.** Es un gráfico en el que los efectos se clasifican desde arriba hacia abajo con una importancia de predictores descendente. Las líneas de conexión del diagrama se ponderan tomando como base la significación del efecto, con un grosor de línea mayor correspondiente a efectos con mayor significación (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 por defecto.
- **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, por defecto, 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. Por defecto, 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 por defecto es 1.00, de modo que no se filtran efectos tomando como base la significación.

Coeficientes

Figura 10-17
Vista de coeficientes, estilo de diagrama



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.

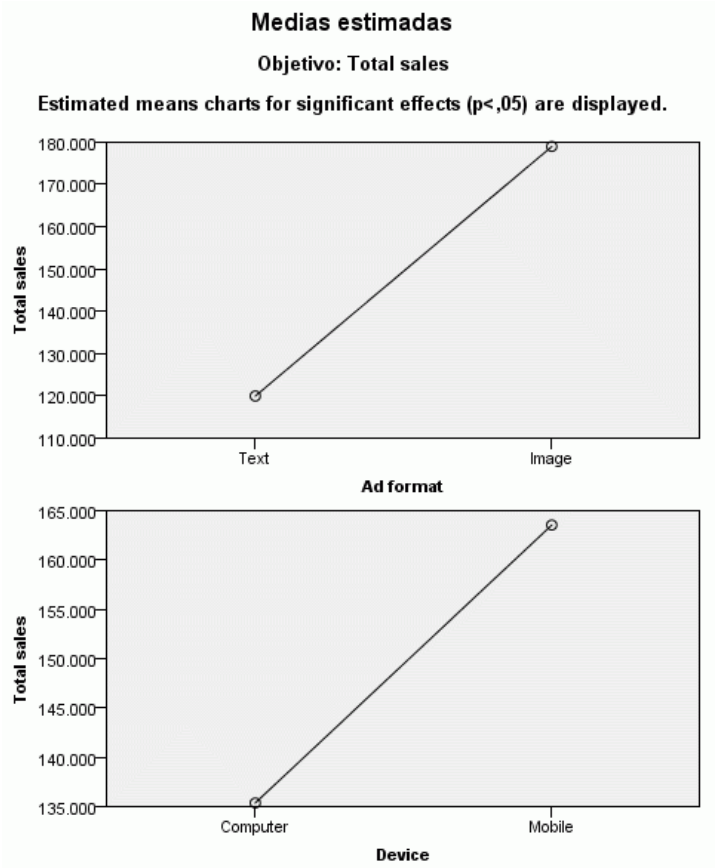
- **Diagrama.** Es un gráfico que muestra la intercepción primero, y luego clasifica los efectos desde arriba hacia abajo con una importancia de predictores descendente. Dentro de los efectos que contienen factores, los coeficientes se clasifican en orden ascendente de valores de datos. Las líneas de 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 por defecto.
- **Tabla.** Muestra los valores, las pruebas de significación y los intervalos de confianza para los coeficientes de modelos individuales. Tras la intercepción, 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, por defecto, la tabla se contrae para mostrar únicamente el coeficiente, la significación y la importancia de cada parámetro del modelo. Para ver el error típico, 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. Por defecto, 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 por defecto es 1.00, de modo que no se filtran coeficientes tomando como base la significación.

Medias estimadas

Figura 10-18
Vista 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.

Nota: si no hay predictores significativos, no se generan medias estimadas.

Resumen de creación de modelos

Figura 10-19
Vista de Resumen de creación de modelos, algoritmo de selección por pasos hacia adelante

Resumen de construcción de modelo

Objetivo: Total sales

		Paso						
		1	2	3	4	5	6	7
Criterio de información		11.949,413	11.597,758	11.347,000	11.118,878	10.965,287	10.816,338	10.803,021
Efecto	secondary_keyword_transformed	✓	✓	✓	✓	✓	✓	✓
	ad_format		✓	✓	✓	✓	✓	✓
	primary_keyword_transformed			✓	✓	✓	✓	✓
	age_category_transformed				✓	✓	✓	✓
	promotion_transformed					✓	✓	✓
	device						✓	✓
	gender							✓

El método de construcción de modelo es Pasos sucesivos hacia delante utilizando el criterio de información.
Una marca de verificación indica que el efecto está en el modelo en este paso.

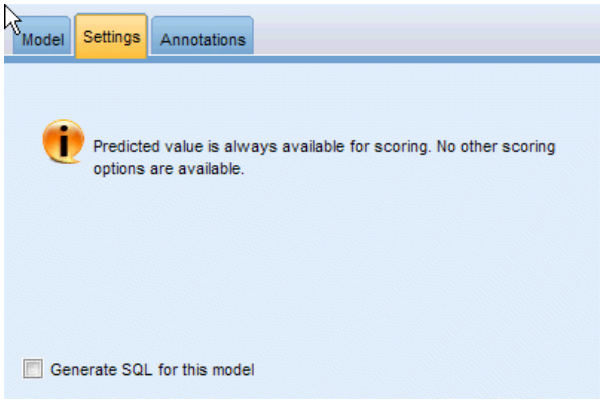
Cuando se selecciona un algoritmo de selección de modelos que no sea Ninguno, proporciona algunos detalles del proceso de creación del modelo.

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

Figura 10-20
Pestaña 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. Si desea obtener más información, consulte el tema Optimización de SQL en el capítulo 6 en *Guía de rendimiento y administración de IBM SPSS Modeler Server 15*.

Puntuar convirtiendo a SQL nativo. Si selecciona esta opción, se genera SQL para puntuar el modelo de manera nativa dentro de la aplicación.

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 pronosticado 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 pronosticar 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 diferentes (probabilidad de pasar a la competencia o no). Si desea obtener más información, consulte el tema Pérdida de clientes

de telecomunicaciones (Regresión logística binomial) en el capítulo 13 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

Nota: solo en el caso de los modelos binomiales, los campos de cadena deben estar limitados a ocho caracteres. Si es necesario, las cadenas más largas se pueden recodificar mediante un nodo Reclasificar. Si desea obtener más información, consulte el tema Nodo Reclasificar en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.Si desea obtener más información, consulte el tema Reducción de la longitud de cadena de datos de entrada (Reclasificar) en el capítulo 10 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

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 pronosticar 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. Si desea obtener más información, consulte el tema Clasificación de clientes de telecomunicaciones (Regresión logística multinomial) en el capítulo 12 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

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 pronosticadas para todas las categorías objetivo, de forma que el "segundo mejor pronóstico" 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 seleccionar campos de manera automática, aunque otros métodos, como los modelos de árboles o de selección de características, pueden hacerlo más rápido en conjuntos de datos de mayor tamaño). 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. Si desea obtener más información, consulte el tema Salida avanzada de regresión logística el p. 288.

Opciones de modelo para el nodo Logística

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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Crear modelos divididos. Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como modelos divididos. Si desea obtener más información, consulte el tema Generación de modelos divididos en el capítulo 3 el p. 32.

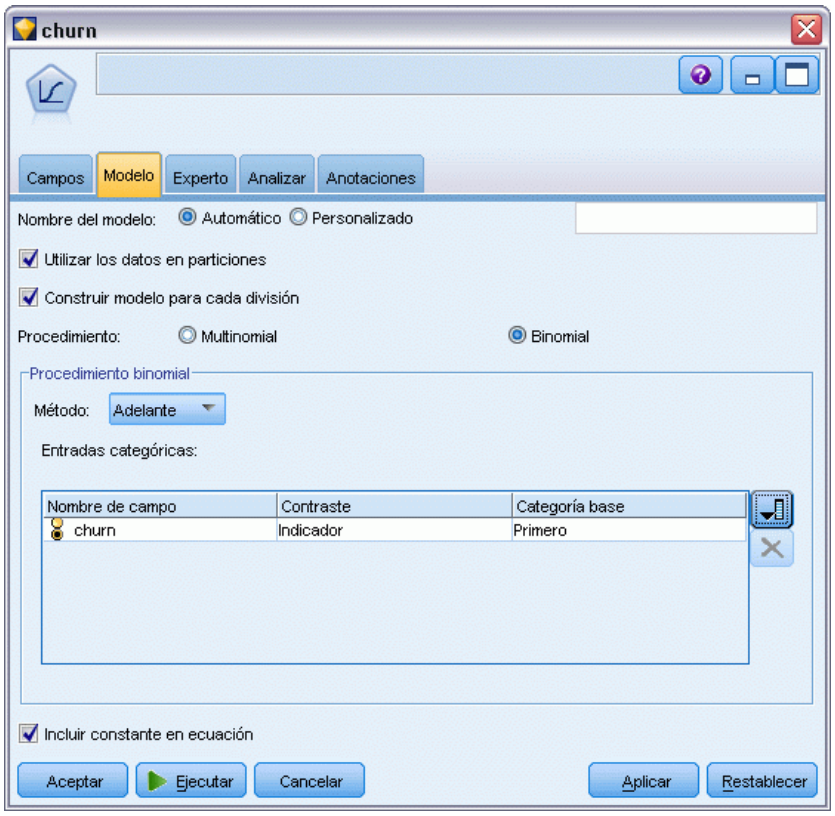
Procedimiento. Especifica si se ha creado un modelo binomial o multinomial. Las opciones disponibles en el cuadro de diálogo varían en función del tipo de procedimiento de modelado seleccionado.

- **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

Figura 10-21
Opciones de modelo binomial para el nodo Logística



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 por defecto que introduce directamente todos los términos en la ecuación. No se realiza ninguna selección de campos en la creación del modelo.
- **Adelante.** El método de selección de campos Adelante genera el modelo moviéndose hacia delante, paso a paso. Con este método, el modelo inicial es el más simple y sólo se pueden añadir al modelo los términos y la constante. 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.

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 de base para cada campo categórico.

- **Nombre de campo.** Esta columna contiene los nombres de campo de las entradas categóricas y está rellenada previamente con todos los valores de marcas y nominales de los datos. 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. Este es el método por defecto.

  Simple. 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 equidistantemente. 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 del campo de entrada, clasificado alfabéticamente, o seleccione Última para utilizar la última categoría. El valor por defecto es Primera.

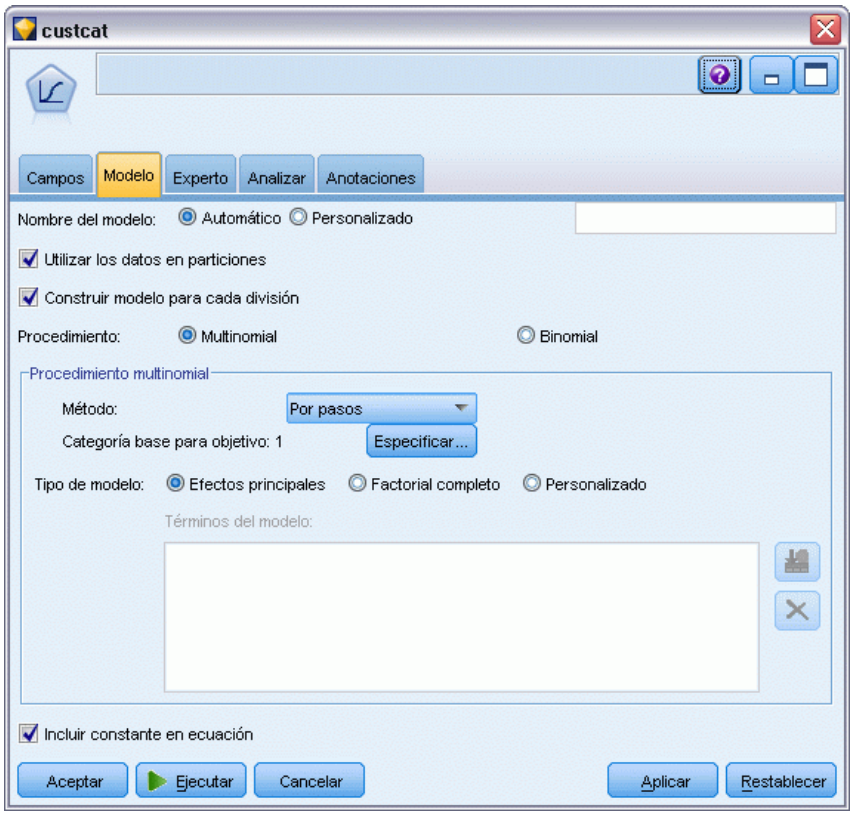
  Nota: este campo no está disponible cuando el contraste se establece en Diferencia, Helmert, Repetido o Polinómico.

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 de 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 de base. Si desea obtener más información, consulte el tema Detalles del nugget de modelo logístico el p. 292.

Modelos multinomiales

Figura 10-22
Opciones de modelo multinomial para el nodo Logística



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 por defecto 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, 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. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Categoría de 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—, clasificado 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. Si desea obtener más información, consulte el tema Utilización del cuadro de diálogo de valores en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Normalmente se especifica la categoría en la que se está menos interesado como categoría de base, por ejemplo, un producto líder con pérdidas. A continuación, se relaciona con la categoría de 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 de 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 de base. Si desea obtener más información, consulte el tema Detalles del nugget de modelo logístico el p. 292.

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 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 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 nuevo cuadro de diálogo en el que podrá especificar los términos.

Figura 10-23
Cuadro de diálogo de nuevos términos de regresión logística

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.

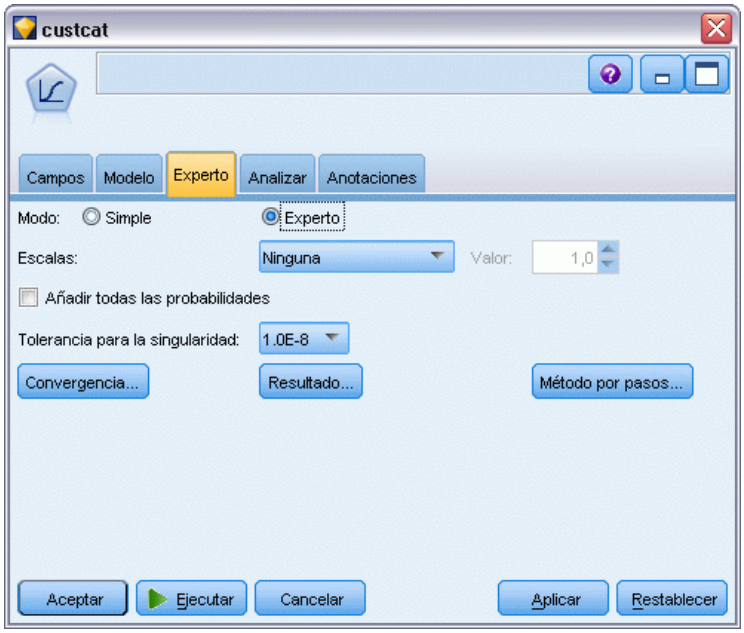
Presentación preliminar. Muestra los términos que se añadirán al modelo si pulsa en Insertar, según los campos seleccionados y el tipo de término.

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 sobre regresión logística. Para acceder a las opciones de experto, en la pestaña Experto, establezca Modo como Experto.

Figura 10-24
Pestaña Experto de Regresión logística



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 desviación (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 pronosticada.

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 pronostique 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. Si desea obtener más información, consulte el tema Nugget de modelo logístico el p. 291.

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). Si desea obtener más información, consulte el tema Opciones de convergencia de regresión logística el p. 287.

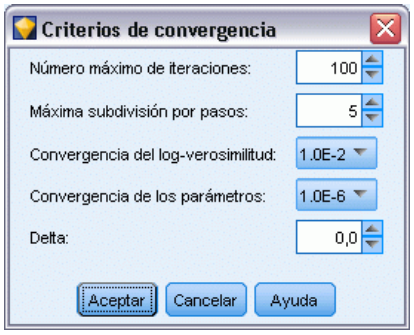
Resultados. Estas opciones le permiten solicitar estadísticos adicionales que aparecerán en el resultado avanzado del nugget de modelo construido por el nodo. Si desea obtener más información, consulte el tema Salida avanzada de regresión logística el p. 288.

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.) Si desea obtener más información, consulte el tema Opciones del método por pasos de regresión logística el p. 290.

Opciones de convergencia de regresión logística

Puede establecer los parámetros de convergencia para la estimación del modelo de regresión logística.

Figura 10-25
Opciones de Convergencia 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 por defecto.

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.

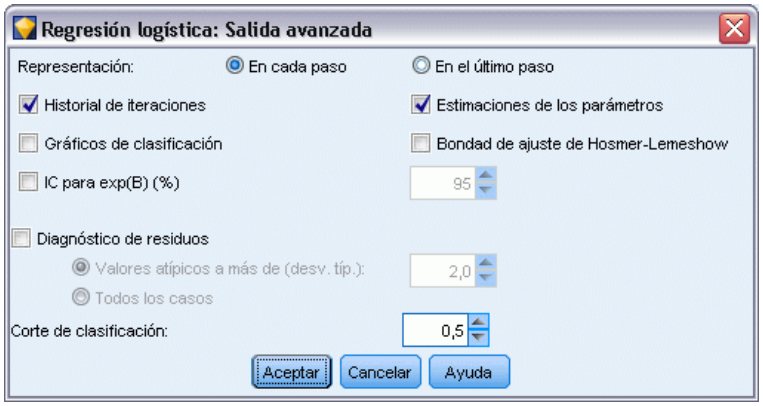
Delta (sólo modelos multinomiales). Puede especificar un valor entre 0 y 1 para añadirlo a cada casilla vacía (combinación de valores de campos de entrada y de salida). Esto puede ayudar al algoritmo de estimación a gestionar los datos en los que hay muchas combinaciones posibles de valores de los campos respecto al número de registros existente en los datos. El valor predeterminado 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. Si desea obtener más información, consulte el tema Resultado avanzado del nugget de modelo logístico el p. 297.

Opciones binomiales

Figura 10-26
Opciones de resultados binomiales de regresión logística



Seleccione los tipos de resultados que se generarán para el modelo. Si desea obtener más información, consulte el tema Resultado avanzado del nugget de modelo logístico el p. 297.

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 por defecto es el 95%).

Diagnóstico de residuos. Solicita una tabla de diagnósticos por caso de los residuos.

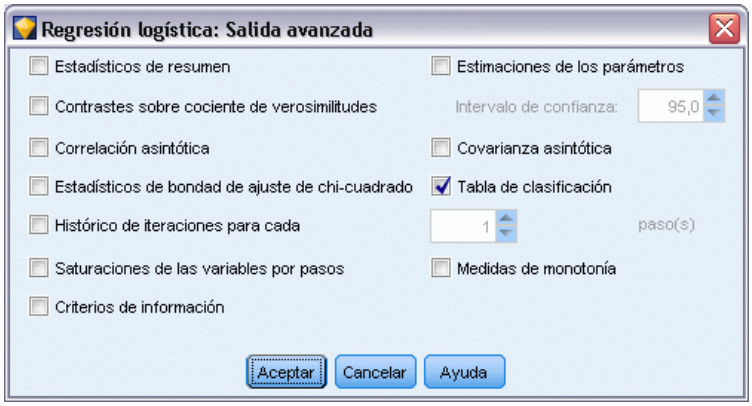
- **Valores atípicos a más de (desv. típ.).** Muestra sólo los casos residuales para los que el valor absoluto estandarizado de la variable de la lista es como mínimo tan grande como el valor especificado. El valor por defecto es 2.
- **Todos los casos.** Incluye todos los casos en la tabla de diagnósticos 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. Permite determinar el punto de corte para clasificar casos. Los casos con valores pronosticados que han sobrepasado el punto de corte para la clasificación se clasifican como positivos, mientras que aquéllos con valores pronosticados menores que el punto de corte se clasifican como negativos. Para cambiar los valores por defecto, introduzca un valor comprendido entre 0,01 y 0,99.

Opciones multinomiales

Figura 10-27
Opciones de resultados multinomiales de regresión logística



Seleccione los tipos de resultados que se generarán para el modelo. Si desea obtener más información, consulte el tema Resultado avanzado del nugget de modelo logístico el p. 297.

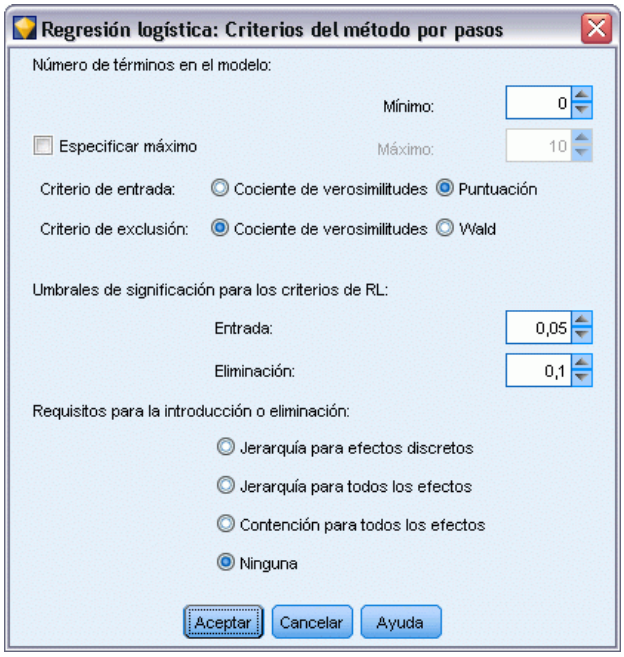
Nota: si se selecciona la opción Contrastes de razón de verosimilitud, aumenta en gran medida el tiempo de procesamiento necesario para generar un modelo de regresión logística. Si la generación del modelo está tardando demasiado, puede desactivar esta opción o utilizar, en su lugar, los estadísticos Wald o de puntuación. Si desea obtener más información, consulte el tema Opciones del método por pasos de regresión logística el p. 290.

Histórico 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 por defecto es el 95%).

Opciones del método por pasos de regresión logística

Figura 10-28
Criterios del método por pasos de Regresión logística



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 por defecto es el uso del estadístico Puntuación.

Criterio de exclusión. Seleccione Cociente de verosimilitudes 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 por defecto es el uso del estadístico cociente de verosimilitudes. 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. Si desea obtener más información, consulte el tema Niveles de medida en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.
- **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 el pronóstico del modelo y la probabilidad asociada. Los nombres de los nuevos campos se derivan del nombre del campo de salida que se está pronosticando, con el prefijo *$L-* para la categoría pronosticada y *$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 Filtro. El menú Generar permite crear un nuevo nodo Filtro 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.

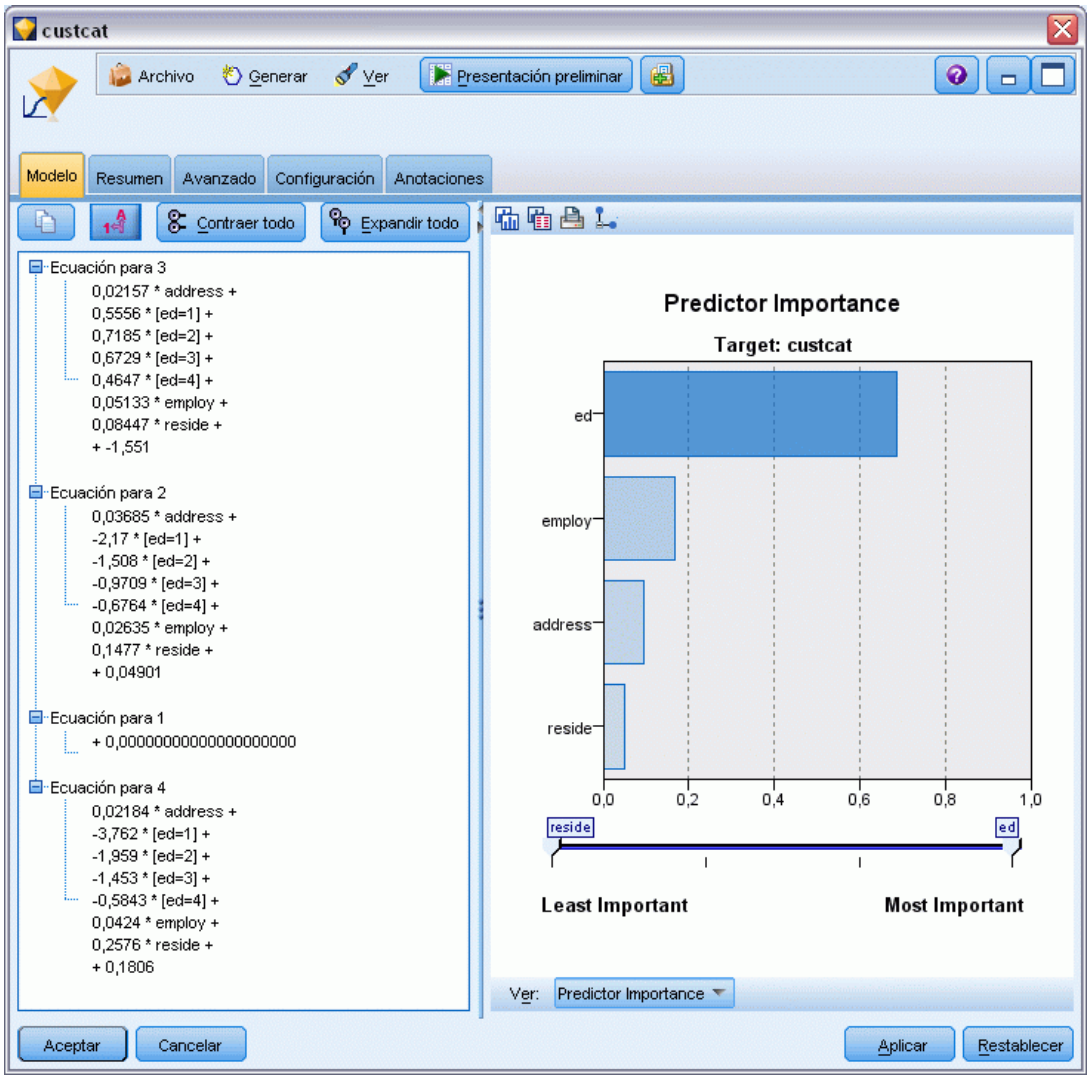
Detalles del nugget de modelo logístico

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. Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

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.

Figura 10-29
Detalles del nugget de modelo logístico con visualización de la importancia del predictor



Ecuación para. Muestra las ecuaciones de regresión utilizadas para derivar las probabilidades de la categoría objetivo, dado un conjunto de valores de pronóstico. La última categoría del campo objetivo se considera la **categoría de línea de base**; las ecuaciones mostradas ofrecen los registros de ventajas para el resto de categorías de destino relativas a la categoría de línea de base para un conjunto de valores de pronóstico concretos. La probabilidad pronosticada para cada categoría del patrón de pronóstico dado se deriva de estos valores de registro de ventajas.

¿Cómo se calculan las probabilidades?

Cada ecuación calcula el registro de ventajas de una categoría objetivo particular, relativa a la categoría de línea de base. El **registro de ventajas**, 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 registro de ventajas 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 registros de ventajas 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 registro de ventajas calculado, i es el índice de categoría y k varía entre 1 y el número de categorías objetivo.

Importancia del predictor

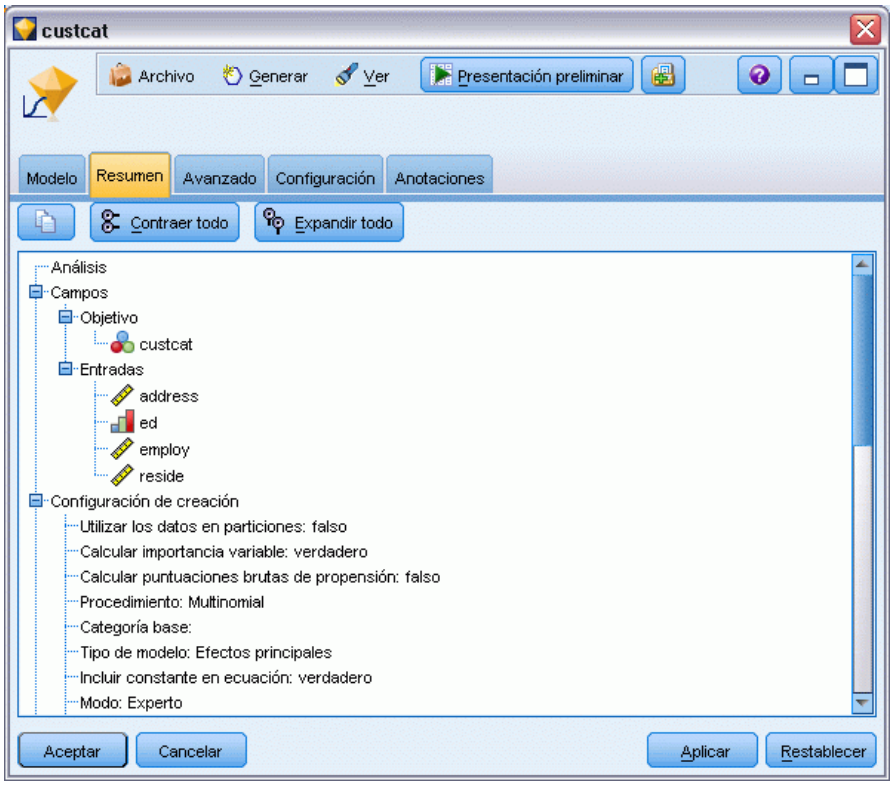
Opcionalmente, en la pestaña Modelo también se puede mostrar un gráfico que indique la importancia relativa de cada predictor cuando se calcule el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Tenga en cuenta que este gráfico sólo está disponible si se ha seleccionado Calcular importancia de predictoren la pestaña Analizar antes de generar el modelo. Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

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 por defecto. 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. Si desea obtener más información, consulte el tema Nodo Análisis en el capítulo 6 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.Si desea obtener información general sobre el explorador de modelos, consulte Exploración de nugget de modelo el p. 53.

Figura 10-30
Pestaña Resumen de nugget de modelo de regresión logística



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 los modelos multinomiales, están disponibles las siguientes opciones:

Calcular confianzas. Especifica si las confianzas se calculan durante la puntuación.

Calcular puntuaciones brutas de propensión (sólo objetivos de marca). En el caso de modelos sólo con objetivos de marca, puede solicitar puntuaciones brutas de propensión que indican la probabilidad del resultado *true* especificado para el campo objetivo. Éstas se añaden a los valores estándar de pronóstico y confianza. Las puntuaciones ajustadas de propensión no están disponibles. Si desea obtener más información, consulte el tema Opciones de análisis del nodo de modelado en el capítulo 3 el p. 42.

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 pronosticada. 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 pronostique. Por ejemplo, si las probabilidades de las categorías *Rojo*, *Verde* y *Azul* son 0,6, 0,3 y 0,1 respectivamente, la categoría pronosticada sería *Rojo*, con una probabilidad de 0,6.

Puntuar convirtiendo a SQL nativo. Si selecciona esta opción, se genera SQL para puntuar el modelo de manera nativa dentro de la aplicación.

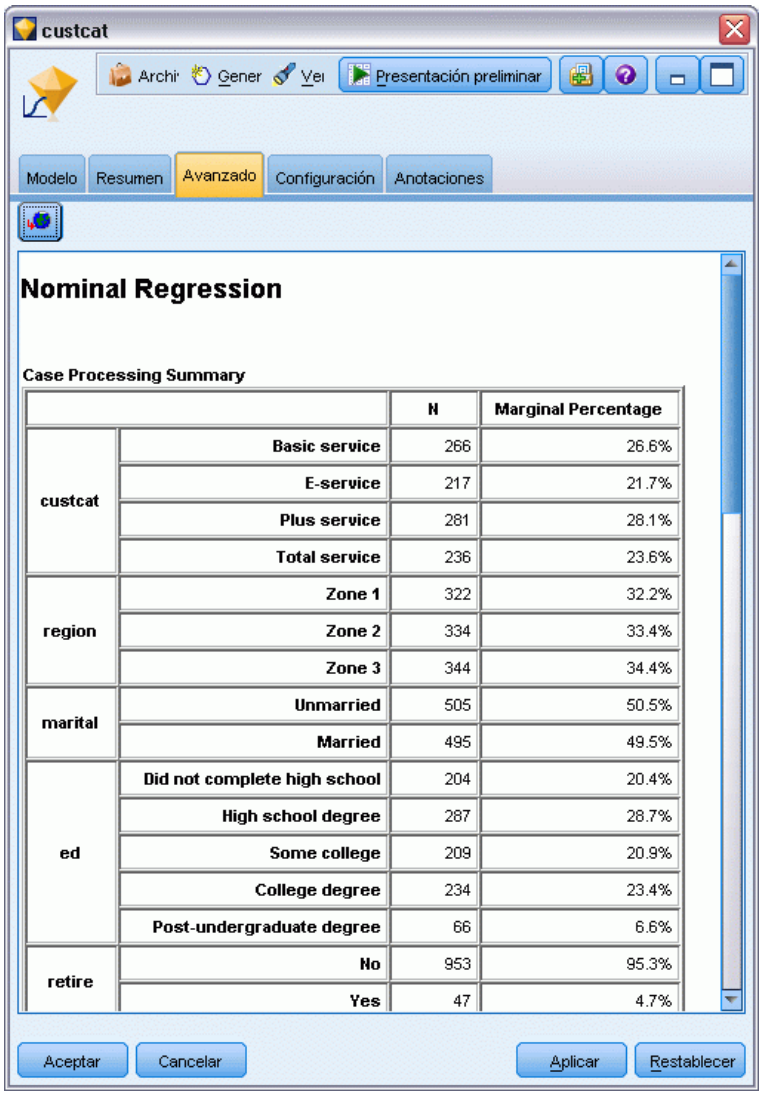
Nota: para modelos multinomiales, la generación de SQL no está disponible si se ha seleccionado Añadir todas las probabilidades o, en el caso de modelos con objetivos nominales, si se ha seleccionado Calcular confianzas. La generación de SQL con cálculos de confianza sólo se admite para modelos multinomiales 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 activadas 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 brutas de propensión. Como se ha indicado anteriormente para los modelos multinomiales, esto sólo es aplicable en modelos con objetivos de marca. Si desea obtener más información, consulte el tema Opciones de análisis del nodo de modelado en el capítulo 3 el p. 42.

Resultado avanzado del nugget de modelo logístico

Figura 10-31
Pestaña de muestra Avanzado del nodo de ecuación de regresión logística



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. Por defecto se imprimen todas las iteraciones (n=1).

Información de ajuste de los modelos (modelos multinomiales). Muestra los contrastes sobre cocientes de verosimilitudes de su modelo (final) en comparación con uno en el que todos los coeficientes de parámetros son 0 (sólo intersección).

Clasificación (opcional). Muestra la matriz de los valores de campo de salida reales y pronosticados con porcentajes.

Estadísticos de bondad de ajuste de chi-cuadrado (opcional). Muestra los estadísticos de Pearson y de chi-cuadrado de cociente de verosimilitudes. 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 cociente de verosimilitudes (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 pronosticadas (opcional). En cada patrón de covariable, muestra las frecuencias observadas y pronosticadas 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 únicas. PCA se centra en todas las varianzas, incluidas las compartidas y las únicas.
- **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 única correspondiente 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.

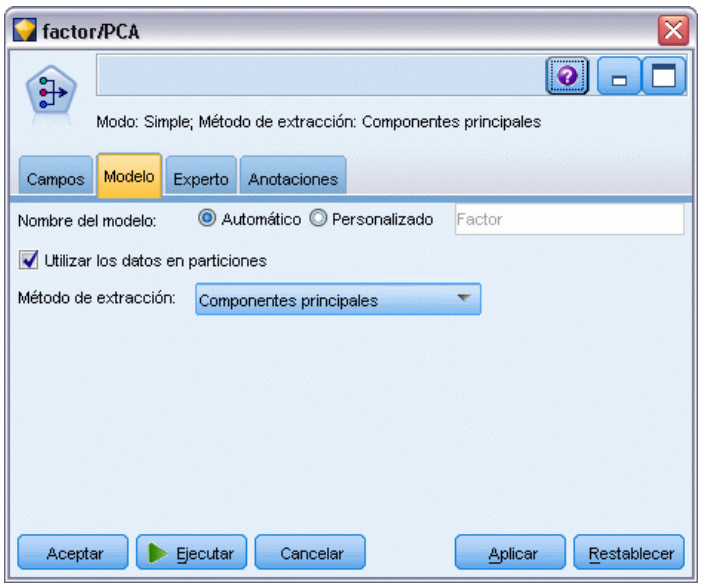
Para los dos métodos, el objetivo es encontrar un número pequeño de campos derivados que resuman de forma eficaz la información del conjunto original de campos.

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 papel definido a campos de *Entrada*. Los campos el papel 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

Figura 10-32
Pestaña Nugget de modelo PCA/Factorial



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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

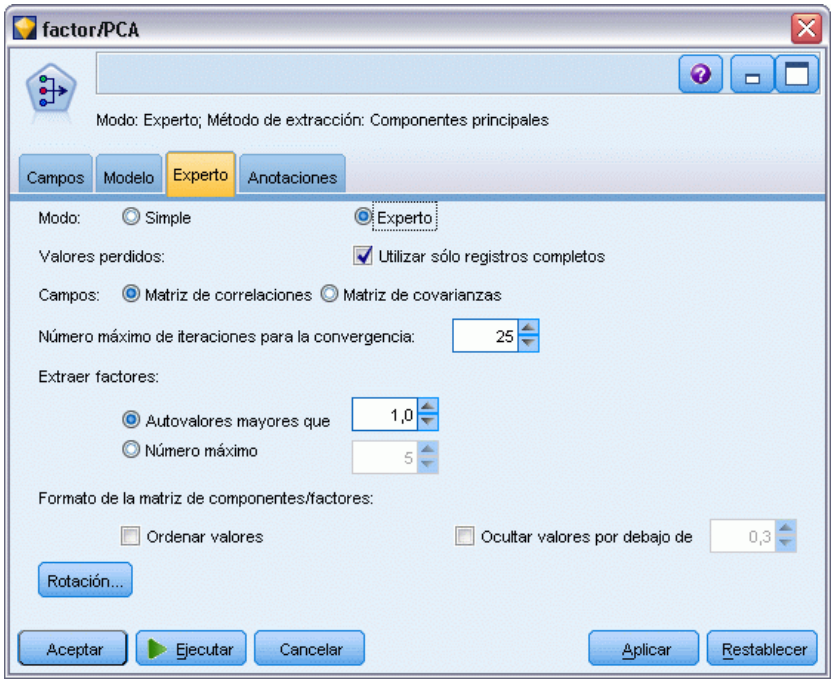
Método de extracción. Especifica el método que se va a utilizar para la reducción de datos.

- **Componentes principales.** Método por defecto 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 única (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 multivariada.
- **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.

Figura 10-33
Pestaña Experto de PCA/Factorial



Valores perdidos. Por defecto, 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 tal caso, puede cancelar la selección de Sólo utilizar registros completos. SPSS Modeler intentará utilizar tanta información como sea posible para calcular el modelo, incluyendo 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 por defecto) 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 por defecto es de cinco factores/componentes.

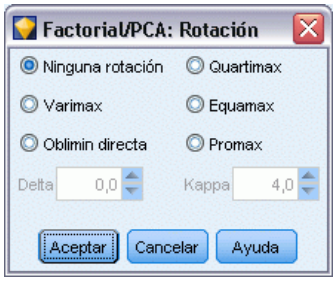
Formato de la matriz de componentes/factores. Estas opciones controlan el formato de la matriz de factores (o matriz de componentes para los modelos PCA).

- **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. Si desea obtener más información, consulte el tema Opciones de rotación para el nodo PCA/Factorial el p. 302.

Opciones de rotación para el nodo PCA/Factorial

Figura 10-34
Opciones de rotación de 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 por defecto. 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 por defecto), las soluciones serán oblicuas. A medida que delta se va haciendo más negativo, los factores son menos oblicuos. Para anular el valor por defecto 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*.

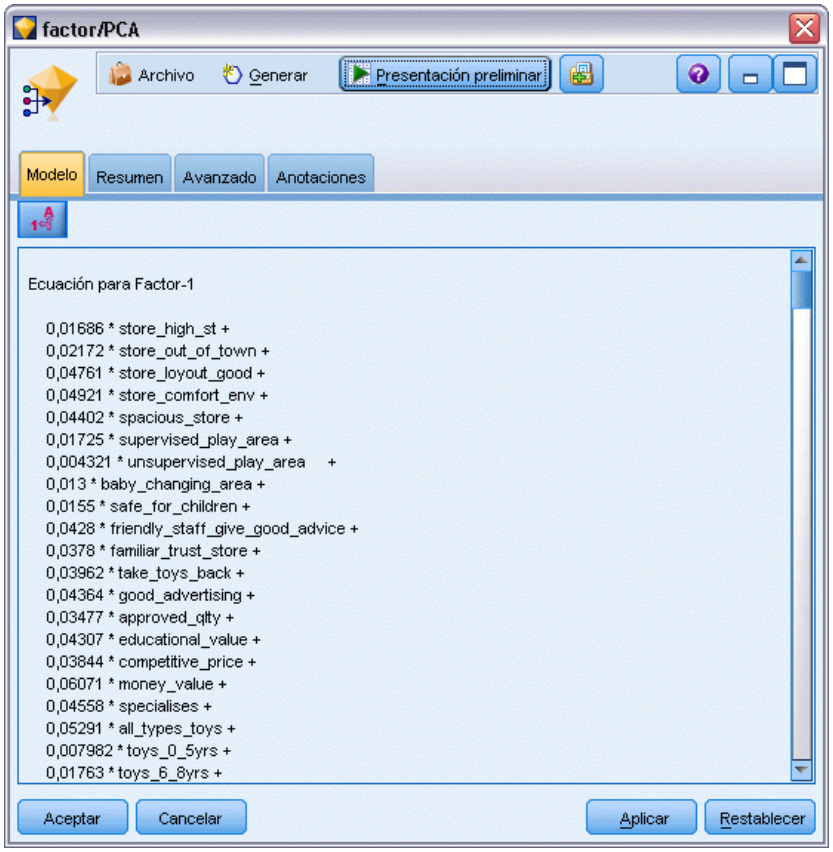
Para obtener una impresión más acertada de qué ha codificado el modelo, puede realizar un análisis hacia abajo más profundo. Una forma útil de consultar el resultado del modelo factorial es ver las correlaciones entre los campos de factores y entradas mediante un nodo Estadísticos. De este modo se detectan 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. Si desea obtener más información, consulte el tema Nodo Estadísticos en el capítulo 6 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

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. Si desea obtener más información, consulte el tema Resultado avanzado del nugget de modelo PCA/Factorial el p. 306.

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.

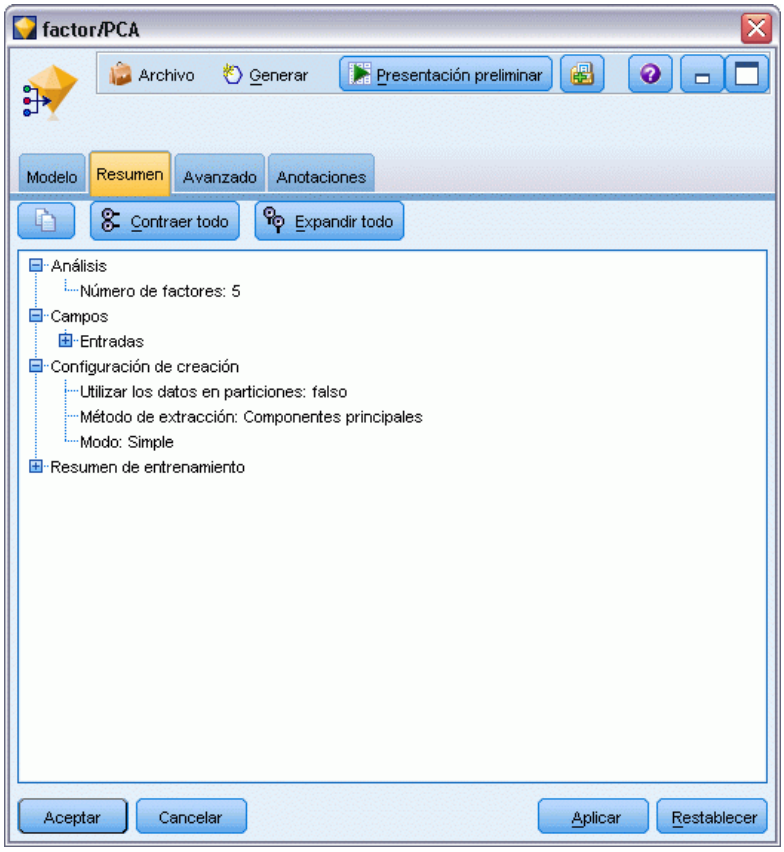
Figura 10-35
Pestaña Nugget de modelo PCA/Factorial



Resumen de nugget de modelo PCA/Factorial

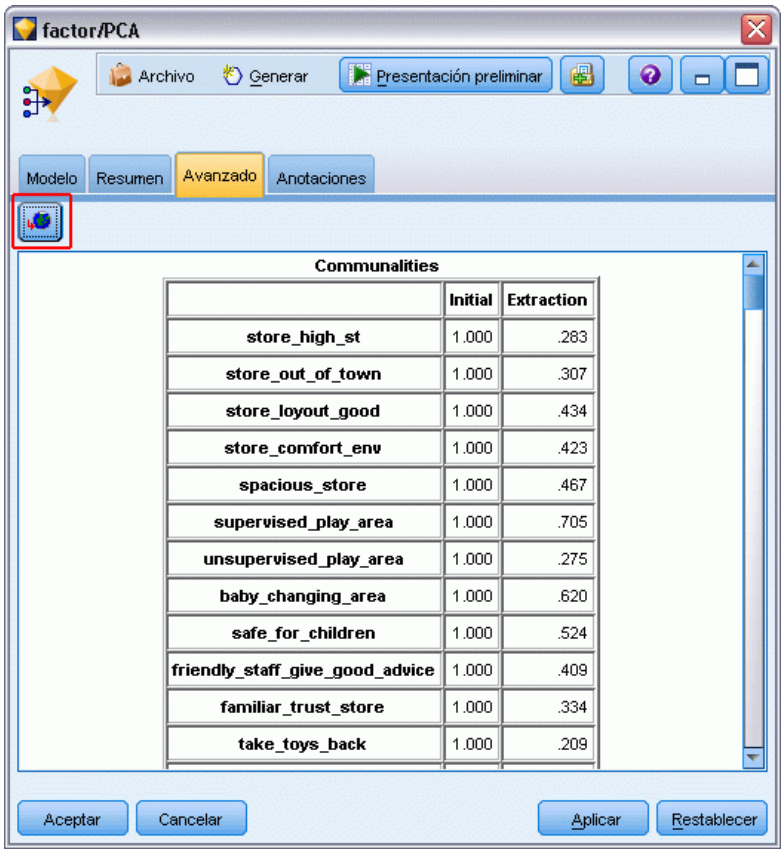
La pestaña Resumen de un modelo factorial muestra el número de factores que se retienen en el modelo PCA/factorial, junto con información adicional de los campos y ajustes utilizados para generar el modelo. Si desea obtener más información, consulte el tema Exploración de nugget de modelo en el capítulo 3 el p. 53.

Figura 10-36
Pestaña de muestra Resumen del nodo de ecuación factorial



Resultado avanzado del nugget de modelo PCA/Factorial

Figura 10-37
Pestaña de muestra Avanzado del nodo de ecuación 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.

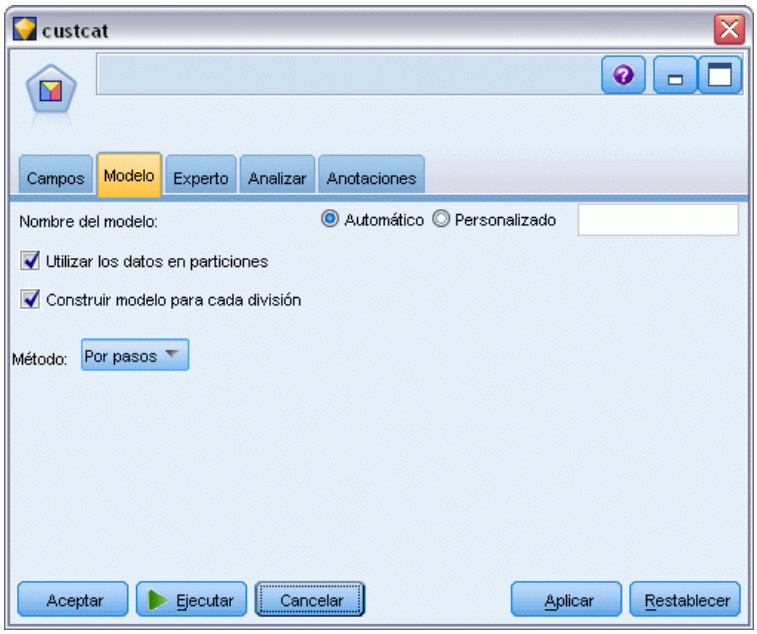
Ejemplo. Una empresa de telecomunicaciones puede usar el análisis discriminante para clasificar a los clientes en grupos basados en datos de uso. Esto permite puntuar a los clientes potenciales y dirigirse a los que tienen más posibilidades de estar incluidos en los grupos más valiosos. Si desea obtener más información, consulte el tema Clasificación de clientes de telecomunicaciones (Análisis discriminante) en el capítulo 21 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

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. Si desea obtener más información, consulte el tema Conversión del almacenamiento mediante el nodo Rellenar en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.) Se ignorarán los campos definidos como *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

Figura 10-38
Cuadro de diálogo del nodo Discriminante, pestaña 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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Crear modelos divididos. Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como modelos divididos. Si desea obtener más información, consulte el tema Generación de modelos divididos en el capítulo 3 el p. 32.

Método. Éstas son las opciones disponibles para introducir predictores en el modelo:

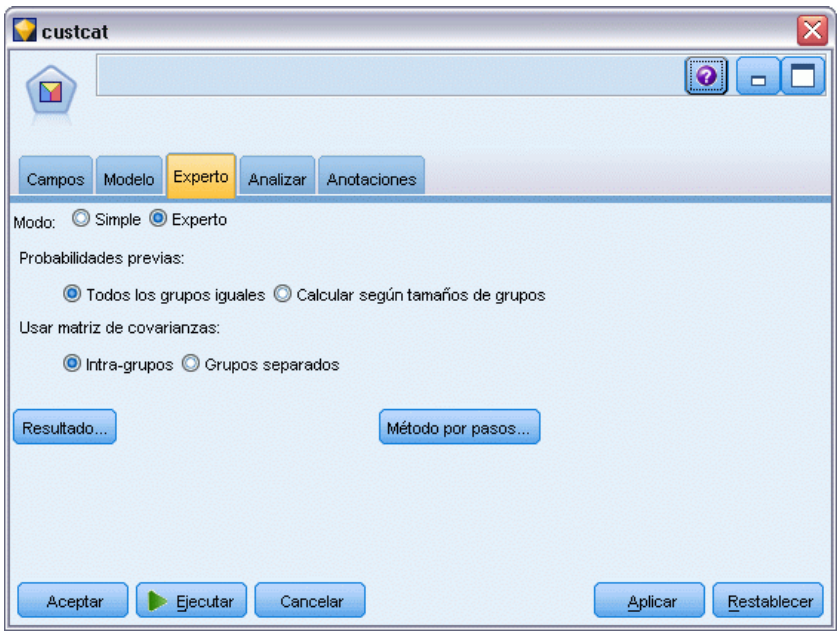
- **Intro.** Éste es el método por defecto que introduce directamente todos los términos en la ecuación. No se añaden los términos que no aumentan de forma significativa el poder predictivo del modelo.
- **Por pasos.** El modelo inicial es el más simple que puede haber, sin ningún término del modelo (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 estas opciones, active el modoExperto en la pestaña Experto.

Figura 10-39
Cuadro de diálogo del nodo Discriminante, pestaña Experto



Probabilidades previas Esta opción determina si los coeficientes de clasificación están ajustados para un conocimiento a priori de pertenencia a un grupo.

- **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.

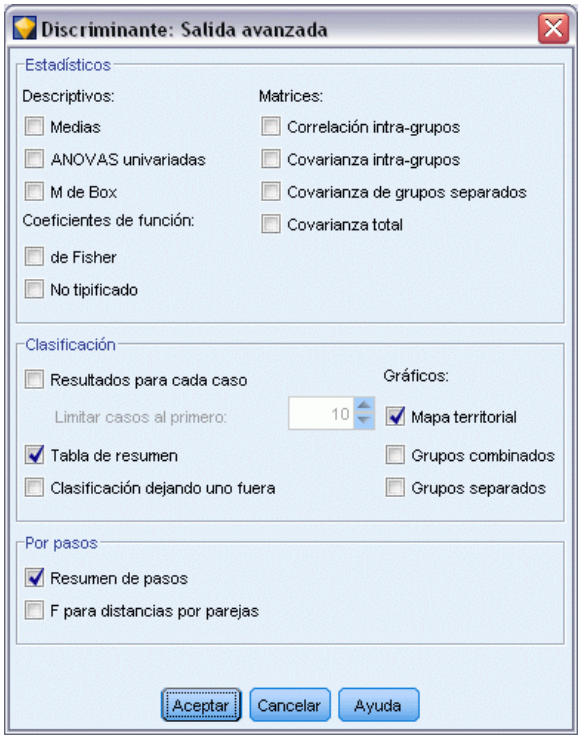
- **Intra-grupos.** Se utiliza la matriz de covarianza intra-grupos combinada para clasificar los casos.
- **Grupos separados.** Para la clasificación se utilizan las matrices de covarianza de los grupos separados. Dado que la clasificación se basa en las funciones discriminantes y no en las variables 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. Si desea obtener más información, consulte el tema Opciones de resultados del nodo Discriminante el p. 310.

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.) Si desea obtener más información, consulte el tema Opciones del método por pasos del nodo Discriminante el p. 312.

Opciones de resultados del nodo Discriminante

Figura 10-40
Opciones de resultados avanzados 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. Si desea obtener más información, consulte el tema Resultados avanzados del nugget de modelo Discriminante el p. 314.

Descriptivos. Las opciones disponibles son medias (incluyendo las desviaciones estándar), ANOVAs univariadas y prueba *M* de Box.

- **Medias.** Muestra la media y desviación típica totales y las medias y desviaciones típicas 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 tipificados.

- **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.

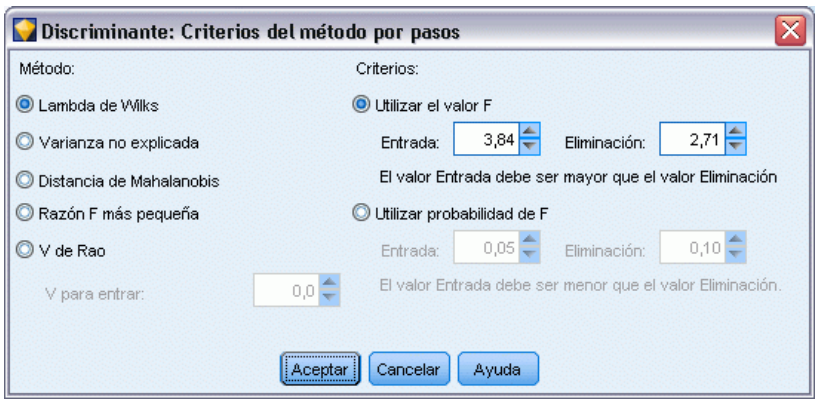
Clasificación. El resultado siguiente pertenece a los resultados de clasificación.

- **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 los pasos muestra los estadísticos para todas las variables después de cada paso; F para distancias por parejas muestra 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

Figura 10-41
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. Especifique valores para introducir 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 salida. La entrada debe ser mayor que la salida 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 salida.

- **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 salida. La entrada debe ser menor que la salida 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 salida.

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 el pronóstico del modelo y la probabilidad asociada. Los nombres de los nuevos campos se derivan del nombre del campo de salida que se está pronosticando, con el prefijo *$D-* para la categoría pronosticada y *$DP-* para la probabilidad asociada. Por ejemplo, para un campo de salida llamado *colorpref*, los nuevos campos se llamarían *$D-colorpref* y *$DP-colorpref*.

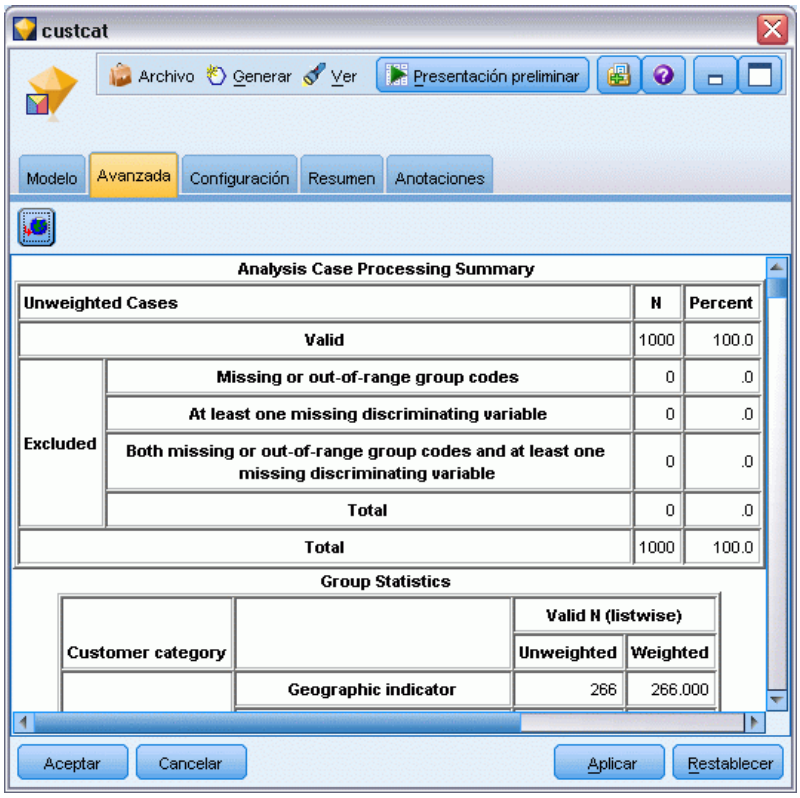
Generación de un nodo Filtro. El menú Generar permite crear un nuevo nodo Filtro para pasar los campos de entrada en función de los resultados del modelo.

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 predictoren la pestaña Analizar antes de generar el modelo. Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

Resultados avanzados del nugget de modelo Discriminante

Figura 10-42
Nugget de modelo Discriminante, pestaña Avanzada

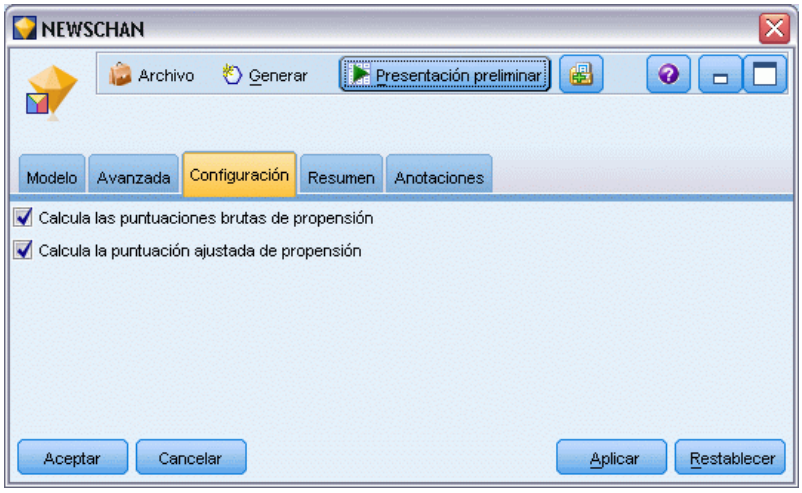


Los resultados avanzados del análisis discriminante ofrecen información detallada sobre el modelo estimado y su rendimiento. La mayoría de la información contenida en los resultados avanzados es bastante técnica y es necesario tener amplios conocimientos sobre análisis discriminante para interpretar correctamente estos resultados. Si desea obtener más información, consulte el tema Opciones de resultados del nodo Discriminante el p. 310.

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.

Figura 10-43
Nugget de modelo Discriminante, ficha Configuración para un objetivo de marca



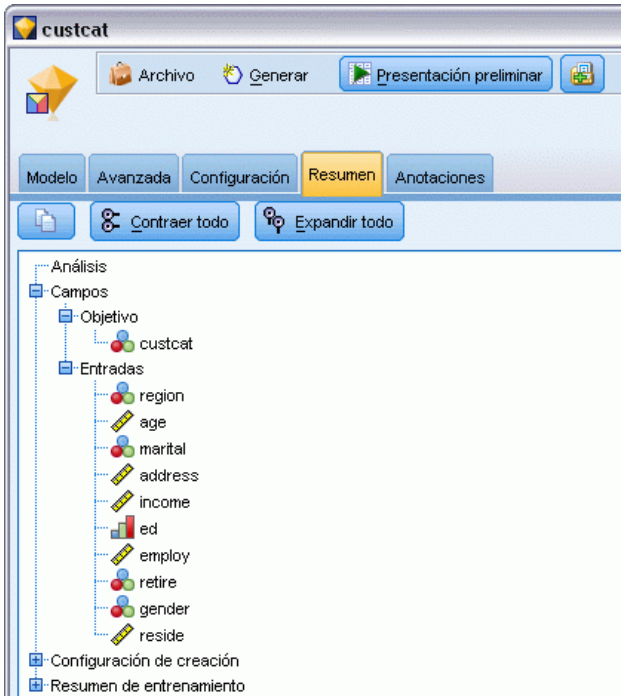
Calcular puntuaciones brutas de propensión. En el caso de modelos con un objetivo de marca (que devuelve un pronóstico de sí o no), puede solicitar puntuaciones de propensión que indican la probabilidad del resultado true especificado para el campo objetivo. Éstas se añaden a otros valores de pronóstico y confianza que pueden generarse durante la puntuación.

Calcular puntuaciones de propensión ajustada. Las puntuaciones brutas de propensión 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 la ficha Analizar antes de generar el modelo.

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. Si desea obtener más información, consulte el tema Nodo Análisis en el capítulo 6 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.Si desea obtener información general sobre el explorador de modelos, consulte Exploración de nugget de modelo el p. 53.

Figura 10-44
Nugget de modelo Discriminante, pestaña Resumen



Nodo GenLin

El modelo lineal generalizado amplía el modelo lineal general para que la variable dependiente esté linealmente relacionada con 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. Si desea obtener más información, consulte el tema Uso de la regresión de Poisson para analizar las tasas de daños sufridos por barcos (modelos lineales generalizados) en el capítulo 23 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

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. Si desea obtener más información, consulte el tema Ajuste de una regresión gamma a reclamaciones de seguros de coches (modelos lineales generalizados) en el capítulo 24 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

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. Si desea obtener más información, consulte el tema Análisis de datos de supervivencia censurados por intervalos (modelos lineales generalizados) en el capítulo 22 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

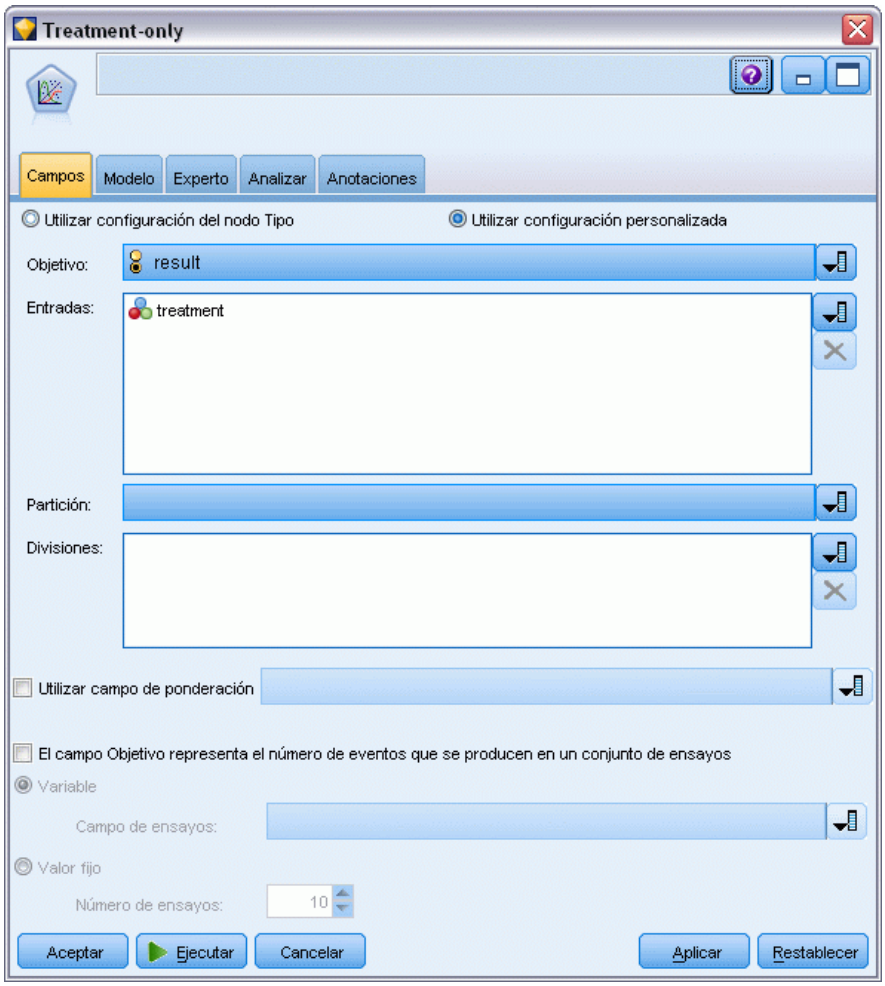
Los modelos lineales generalizados funcionan generando una ecuación que relaciona los valores de los campos de entrada con los valores de los campos de salida. Una vez se ha generado el modelo, se puede utilizar para calcular los valores de datos nuevos. Para cada registro, se calcula una probabilidad de pertenencia a cada categoría posible de salida. La categoría objetivo con la probabilidad más alta se asigna como el valor de salida pronosticado 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

Figura 10-45
Cuadro de diálogo del nodo GenLin, pestaña Campos



Además de las opciones personalizadas de objetivo, entrada y partición que suelen ofrecer las pestañas Campos del nodo de modelado (consulte Opciones de los campos del nodo de modelado el p. 38), el nodo GenLin ofrece la siguiente funcionalidad adicional.

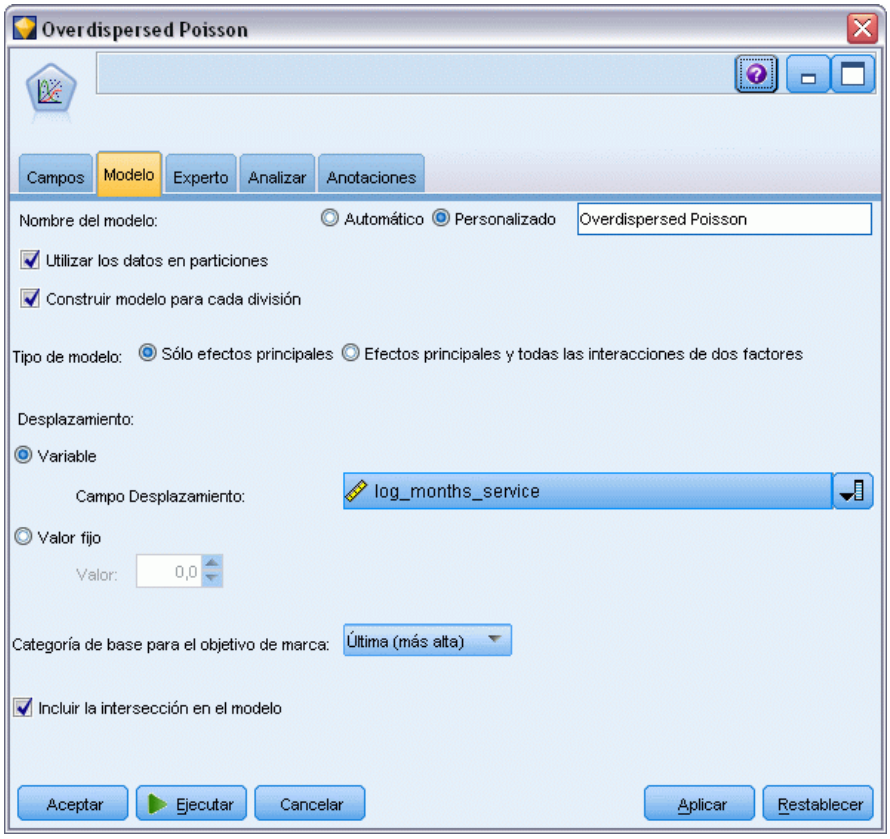
Utilizar campo de ponderación. El parámetro de escala es un parámetro del modelo estimado relacionado con la varianza de la respuesta. Los pesos de escala son valores "conocidos" que pueden variar de una observación a otra. Si se especifica una variable de peso de escala, el parámetro de escala, que está relacionado con la varianza de la respuesta, se divide por él 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

Figura 10-46
Cuadro de diálogo del nodo GenLin, pestaña 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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Crear modelos divididos. Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como modelos divididos. Si desea obtener más información, consulte el tema Generación de modelos divididos en el capítulo 3 el p. 32.

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.

Nota: Si se utiliza un campo de desplazamiento de variable, el campo especificado no debe utilizarse también como una entrada. Defina el papel del campo de desplazamiento como Ninguno en un origen anterior de la ruta o nodo Tipo si es necesario. Si desea obtener más información, consulte el tema Definición del papel de campos en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Categoría de 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.

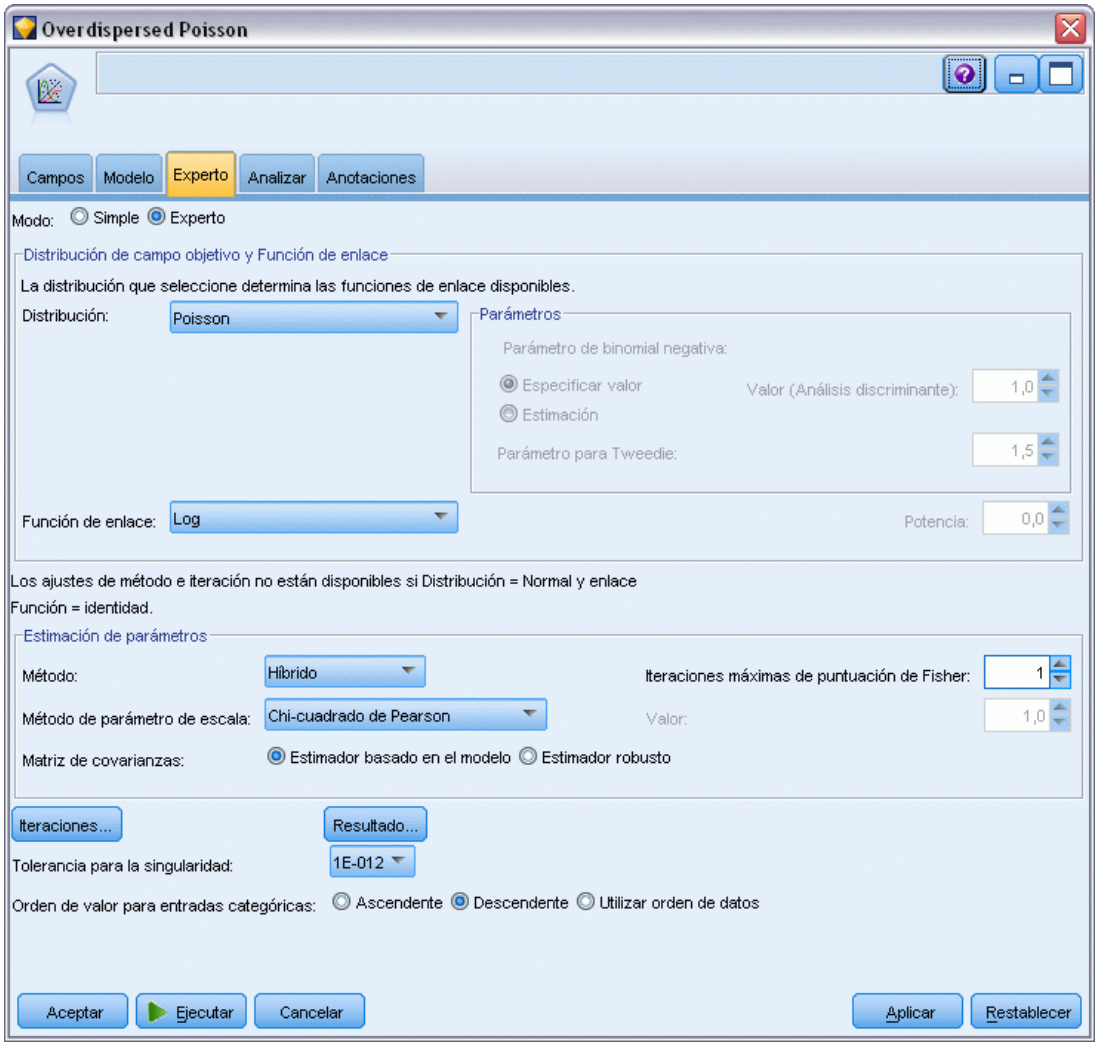
- Por defecto, el procedimiento convierte la última categoría (la de mayor valor), o 1, en categoría de referencia. En esta situación, las probabilidades guardadas por el modelo calculan la posibilidad de que un caso determinado tome el valor 0 y los cálculos del parámetro deberían interpretarse como 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 intersección en el modelo. La intersección se incluye normalmente en el modelo. Si asume que los datos pasan por el origen, puede excluir la intersección.

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 estas opciones, active el modoExperto en la pestaña Experto.

Figura 10-47
Cuadro de diálogo del nodo GenLin, 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 fijo del parámetro auxiliar de la distribución binomial negativa puede ser cualquier número superior o igual a 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. Están disponibles las siguientes funciones:

- **Identidad.** $f(x)=x$. La variable dependiente no se transforma. Este enlace se puede utilizar con cualquier distribución.
- **Log-log complementario.** $f(x)=\log(-\log(1-x))$. Adecuado sólo con la distribución binomial.
- **Cauchit acumulado.** $f(x) = \tan(\pi\ (x - 0{,}5))$, aplicado a la probabilidad acumulada de cada categoría de la respuesta. Adecuado sólo con la distribución multinomial.

- **Log-log complementario acumulado.** $f(x)=\ln(-\ln(1-x))$, aplicado a la probabilidad acumulada de cada categoría de la respuesta. Adecuado sólo con la distribución multinomial.
- **Logit acumulado.** $f(x)=\ln(x / (1-x))$, aplicado a la probabilidad acumulada de cada categoría de la respuesta. Adecuado sólo con la distribución multinomial.
- **Log-log negativo acumulado.** $f(x)=-\ln(-\ln(x))$, aplicado a la probabilidad acumulada de cada categoría de la respuesta. Adecuado sólo con la distribución multinomial.
- **Probit acumulado.** $f(x)=\Phi^{-1}(x)$, aplicado a la probabilidad acumulada de cada categoría de la respuesta, donde Φ^{-1} es la función de distribución acumulada normal estándar inversa. Adecuado sólo con la distribución multinomial.
- **Log.** $f(x)=\log(x)$. Este enlace se puede utilizar con cualquier distribución.
- **Complemento log.** $f(x)=\log(1-x)$. Adecuado sólo con la distribución binomial.
- **Logit.** $f(x)=\log(x / (1-x))$. Adecuado sólo con la distribución binomial.
- **Binomial negativa.** $f(x)=\log(x / (x+k^{-1}))$, donde k es el parámetro auxiliar de la distribución binomial negativa. Adecuado sólo con la distribución binomial negativa.
- **Log-log negativo.** $f(x)=-\log(-\log(x))$. Adecuado sólo 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 $\alpha=0$. α es la especificación de número requerida y debe ser un número real. Adecuado sólo con la distribución binomial.
- **Probit.** $f(x)=\Phi^{-1}(x)$, donde Φ^{-1} es la función de distribución acumulada normal típica inversa. Adecuado sólo con la distribución binomial.
- **Potencia.** $f(x)=x^{\alpha}$, si $\alpha \neq 0$. $f(x)=\log(x)$, si $\alpha=0$. α es la especificación de número requerida 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 la negativa de la inversa generalizada de la matriz hessiana. El estimador robusto (también llamado el estimador de Huber/White/Sandwich) es un estimador basado en el modelo "corregido" que proporciona una estimación coherente de la covarianza, incluso cuando la varianza y las funciones de enlace no se han especificado correctamente.

Iteraciones. Estas opciones le permiten controlar los parámetros de la convergencia del modelo. Si desea obtener más información, consulte el tema Iteraciones de modelos lineales generalizados el p. 324.

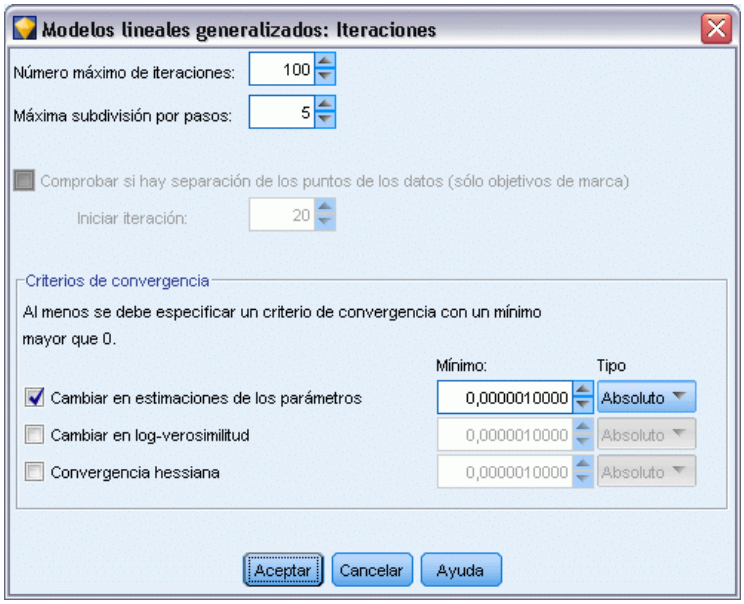
Resultados. Estas opciones le permiten solicitar estadísticos adicionales que aparecerán en el resultado avanzado del nugget de modelo construido por el nodo. Si desea obtener más información, consulte el tema Resultados avanzados de modelos lineales generalizados el p. 326.

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.

Figura 10-48
Opciones de iteraciones de modelos lineales generalizados



Iteraciones.

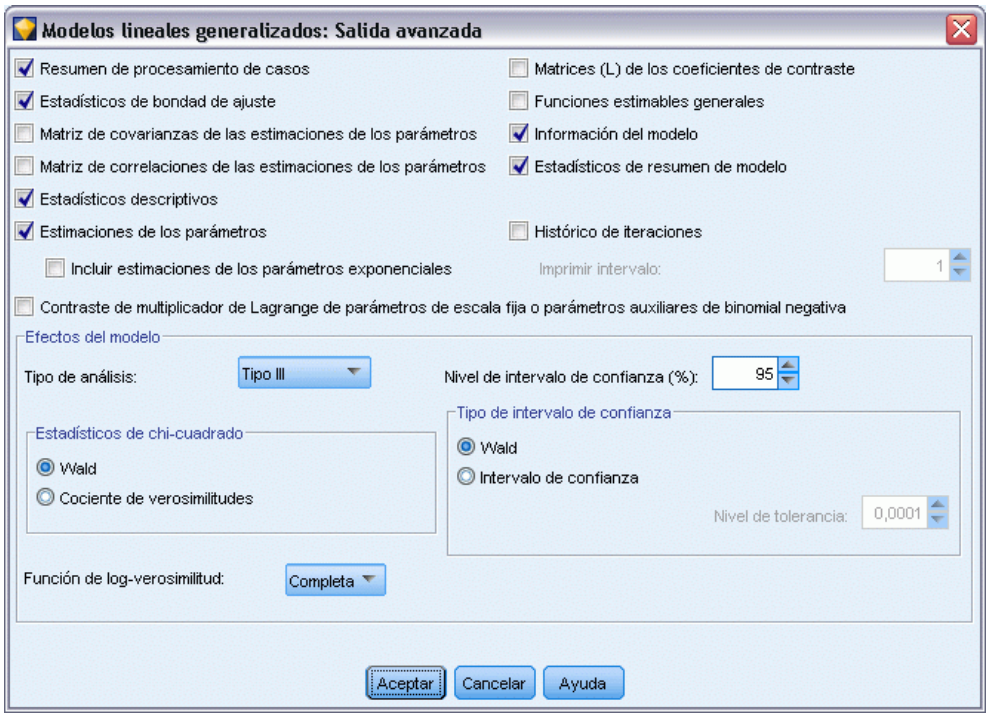
- **Número máximo de iteraciones.** Número máximo de iteraciones que se ejecutará el algoritmo. Especifique un número entero no negativo.
- **Máxima subdivisión por pasos.**En cada iteración, se reduce el tamaño del paso mediante un factor de 0,5 hasta que aumenta el logaritmo de la verosimilitud o se 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 está disponible para respuestas binomiales con formato binario.

Criterios de convergencia.

- **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 log-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

Figura 10-49
Opciones de 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. Si desea obtener más información, consulte el tema Resultado avanzado del nugget de modelo GenLin el p. 329.

Los siguientes resultados están disponibles:

- **Resumen de procesamiento de 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 desviación y la desviación 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 por defecto y para las medias marginales estimadas, si se solicitaron en la ficha Medias marginales estimadas.
- **Funciones estimables generales.** Muestra las matrices para generar las matrices (L) de los coeficientes de contraste.
- **Histórico 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 del histórico de iteraciones se muestran las estimaciones para cada iteración *n*-ésima que comienza con la iteración 0 (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, sirve como contraste del parámetro auxiliar fijo.

Efectos del modelo.

- **Tipo de análisis.** Especifique el tipo de análisis que se va a producir. El análisis de tipo I suele ser adecuado cunado se tienen motivos a priori para realizar predictores en el modelo, mientras que el tipo III se aplica de forma 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 pronosticada 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 pronosticando, con el prefijo *$G-* para la categoría pronosticada y *$GP-* para la probabilidad asociada. Por ejemplo, para un campo de salida llamado *por_defecto*, los nuevos campos se llamarían *$G-por_defecto* y *$GP-por_defecto*. Los nombres de los posteriores dos campos adicionales se asignan en función de los valores del campo de salida, con el prefijo *$GP-*. Por ejemplo, si los valores correctos de *por_defecto* son *Sí* y *No*, los nuevos campos se denominarán *$GP-Sí* y *$GP-No*.
- **Destino continuo.** Añade campos que contienen la medida pronosticada y el error típico.
- **Objetivo continuo, representando el número de eventos en una serie de ensayos.** Añade campos que contienen la medida pronosticada y el error típico.

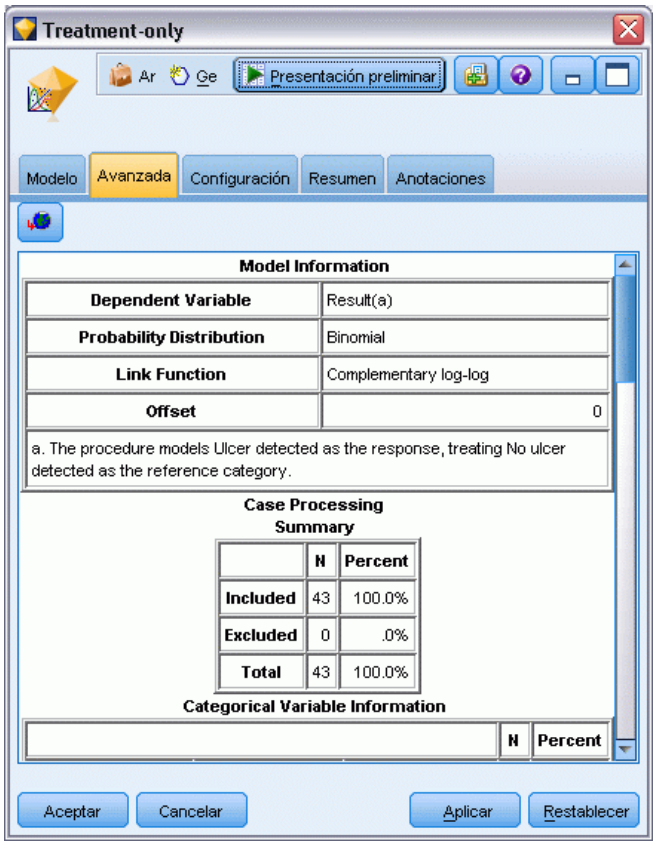
Generación de un nodo Filtro. El menú Generar permite crear un nuevo nodo Filtro para pasar los campos de entrada en función de los resultados del modelo.

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 predictoren la pestaña Analizar antes de generar el modelo. Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

Resultado avanzado del nugget de modelo GenLin

Figura 10-50
Nugget de modelo GenLin, pestaña Avanzado

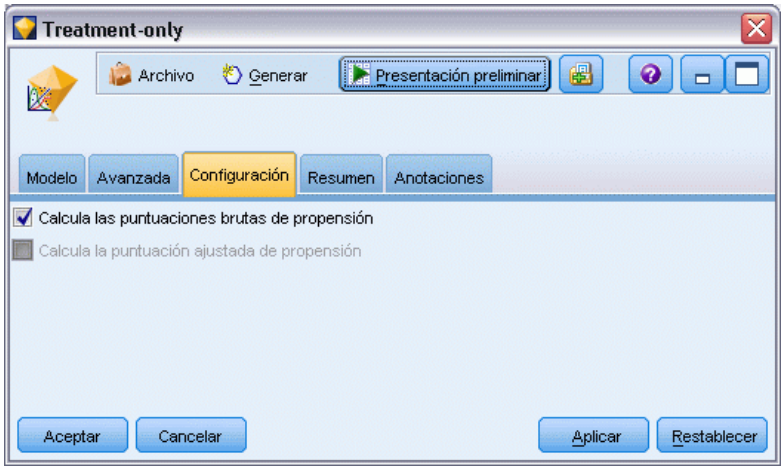


El resultado avanzado del modelo lineal generalizado ofrece información detallada sobre el modelo estimado y su rendimiento. La mayoría de la información contenida en los resultados avanzados es bastante técnica y es necesario tener amplios conocimientos sobre este tipo de análisis para interpretar correctamente estos resultados. Si desea obtener más información, consulte el tema Resultados avanzados de modelos lineales generalizados el p. 326.

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. 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.

Figura 10-51
Nugget de modelo GenLin, pestaña Configuración para objetivos de marca



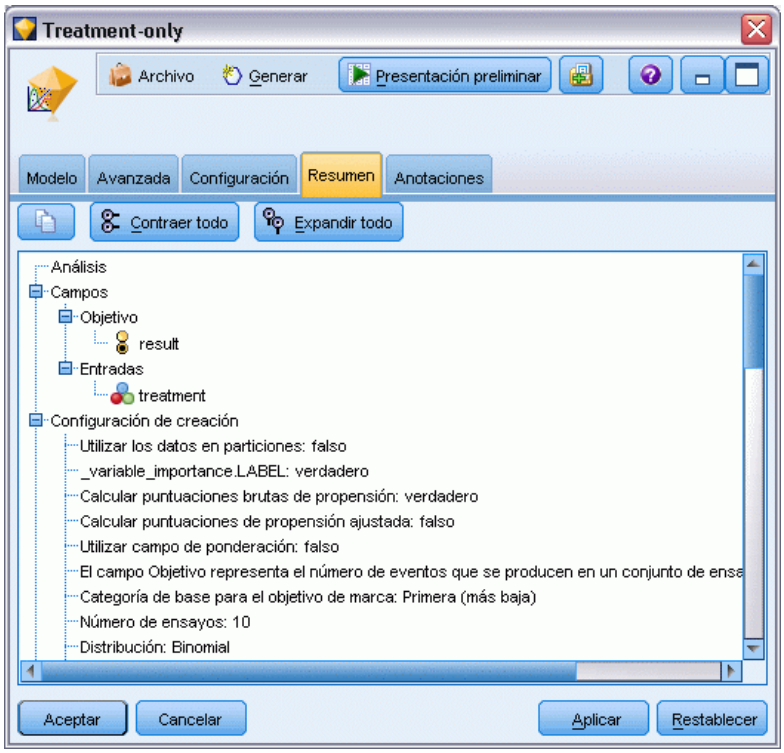
Calcular puntuaciones brutas de propensión. En el caso de modelos con un objetivo de marca (que devuelve un pronóstico de sí o no), puede solicitar puntuaciones de propensión que indican la probabilidad del resultado true especificado para el campo objetivo. Éstas se añaden a otros valores de pronóstico y confianza que pueden generarse durante la puntuación.

Calcular puntuaciones de propensión ajustada. Las puntuaciones brutas de propensión 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 la ficha Analizar antes de generar el modelo.

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. Si desea obtener más información, consulte el tema Nodo Análisis en el capítulo 6 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.Si desea obtener información general sobre el explorador de modelos, consulte Exploración de nugget de modelo el p. 53.

Figura 10-52
Nugget de modelo GenLin, ficha Resumen



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.

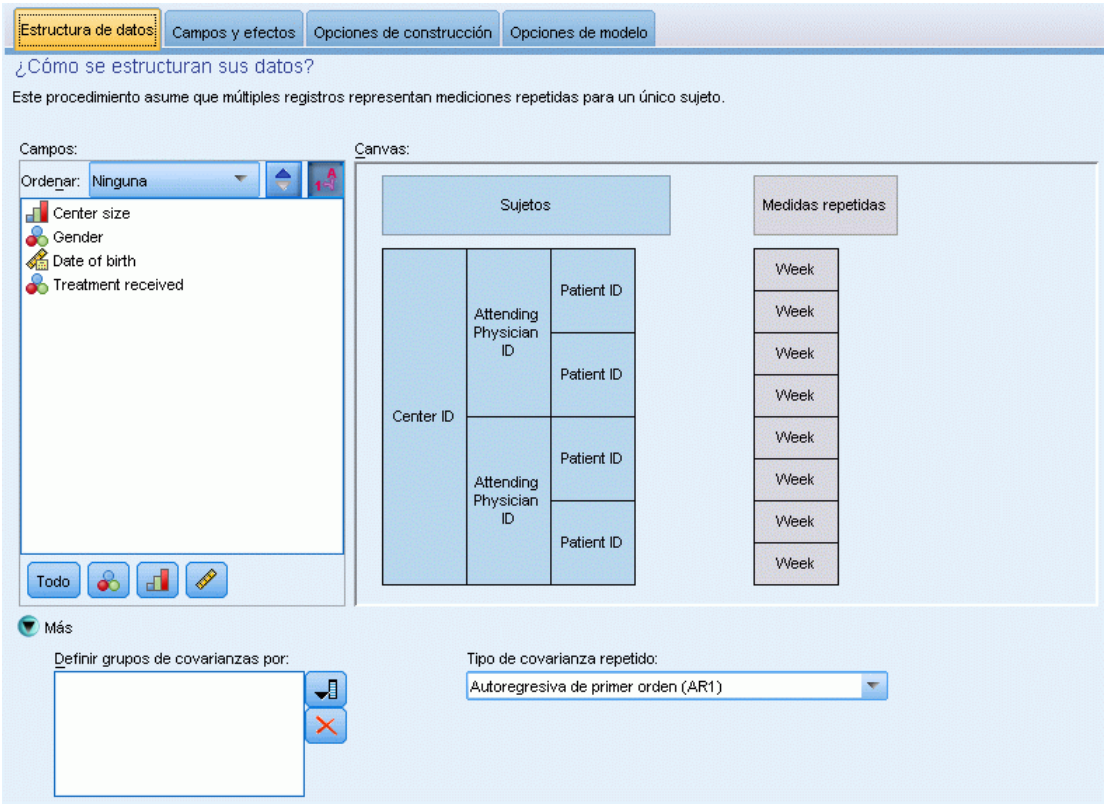
Ejemplos. El consejo escolar del distrito puede utilizar un modelo lineal mixto generalizado para determinar si un método educativo experimental es eficaz para mejorar las notas en matemáticas. Los estudiantes de la misma clase deberían correlacionarse dado que les enseña el mismo maestro. Asimismo, 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 medidas 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 medida nominal, el analista de la empresa utiliza un modelo logit mixto generalizado con una intersección 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.

Figura 10-53
Pestaña Estructura de datos



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 manera única 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 únicos 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 medidas 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 la 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.

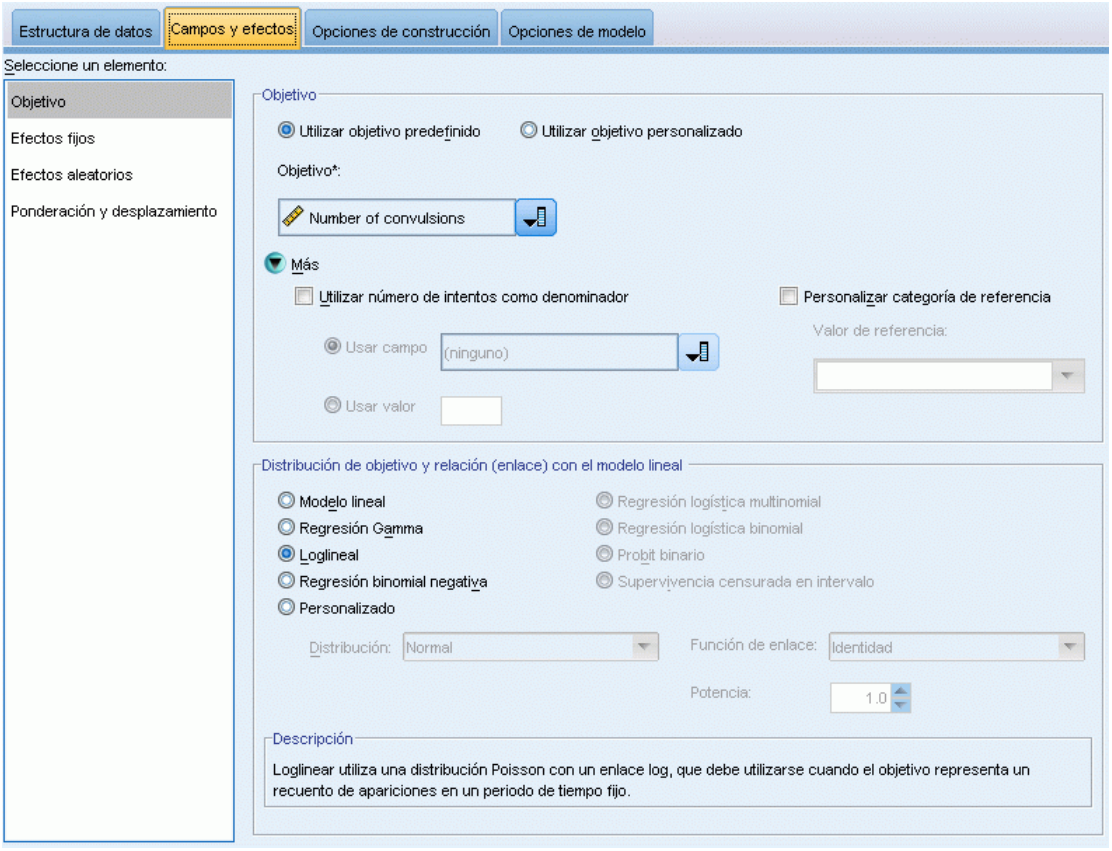
Definir grupos de covarianzas por. Los campos especificados aquí definen conjuntos independientes de parámetros de covarianza de efectos repetidos; uno por cada categoría definida por la clasificación cruzada de los campos de agrupación. Todos los sujetos tienen el mismo tipo de covarianza; los sujetos con la misma agrupación de covarianza tendrán los mismos valores para los parámetros.

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
- Toeplitz
- Sin estructura
- Componentes de la varianza

Objetivo

Figura 10-55
Configuración de 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 medida. Dicho nivel de medida 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, por defecto, 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 métodos abreviados para varios modelos comunes, o también puede seleccionar el ajuste Personalizado si hay una combinación específica de distribución y función de enlace que quiera ajustar y que no esté en la lista breve.

- **Modelo lineal.** Especifica una distribución normal con un enlace de identidad, que resulta de utilidad cuando se puede pronosticar 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 pronosticada 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 medida 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 intersección 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

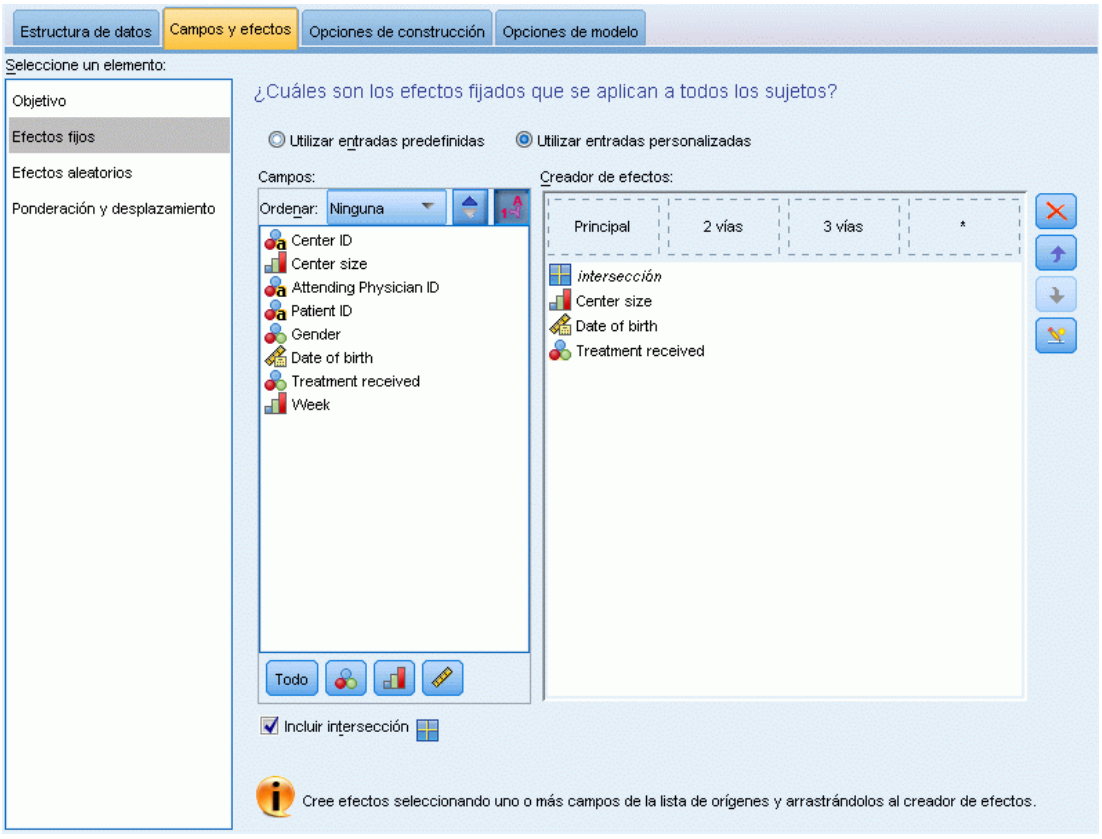
La función de enlace es una transformación del objetivo que permite una estimación del modelo. Se encuentran disponibles las siguientes funciones:

- **Identidad.** $f(x)=x$. El objetivo 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 negativa.** $f(x)=-\log(-\log(x))$. Adecuado solamente con la distribución binomial o multinomial.
- **Probit.** $f(x)=\Phi^{-1}(x)$, donde Φ^{-1} es la función de distribución acumulada normal típica inversa. Adecuado solamente con la distribución binomial o multinomial.
- **Potencia.** $f(x)=x^{\alpha}$, si $\alpha \neq 0$. $f(x)=\log(x)$, if $\alpha=0$. α es la especificación de número requerida y debe ser un número real. Este enlace se puede utilizar con cualquier distribución, excepto la multinomial.

Efectos fijos

Figura 10-56
Configuración de 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. Por defecto, los campos con el papel 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 fuentes 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 que suelte aparece como una única interacción en la parte inferior de la lista de efectos.

Los botones a la derecha del generador de efectos le permiten hacer lo siguiente:

Eliminar términos del modelo de efectos fijos seleccionando los términos que quiera eliminar y pulsando en el botón Eliminar;

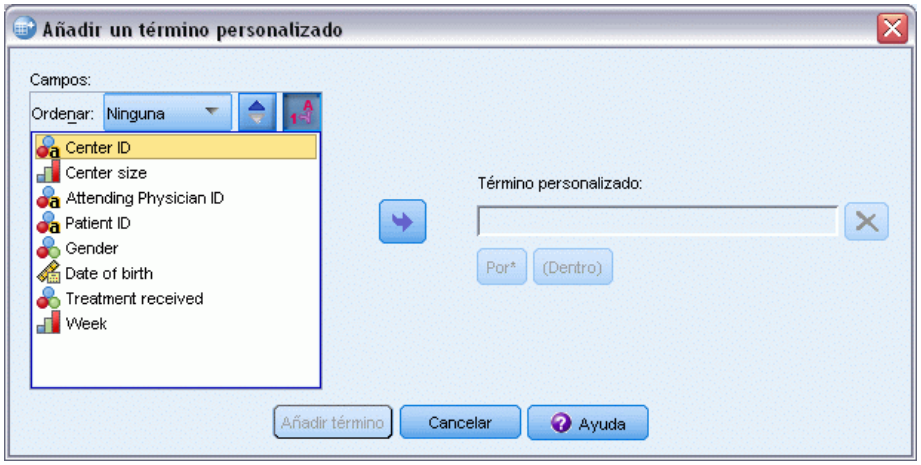
Reordenar los términos dentro del modelo de efectos fijos seleccionando los términos que quiera reordenar y pulsando en la flecha hacia arriba o hacia abajo; y

Añadir términos anidados al modelo mediante el cuadro de diálogo Añadir un término personalizado , pulsando en el botón Añadir un término personalizado.

Incluir intersección. La intersección se incluye normalmente en el modelo. Si asume que los datos pasan por el origen, puede excluir la intersección.

Añadir un término personalizado

Figura 10-57
Cuadro de diálogo 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

▶ Seleccione un factor o covariable que esté anidado en otro factor y, a continuación, pulse en el botón de flecha.

▶ Pulse en (Dentro).

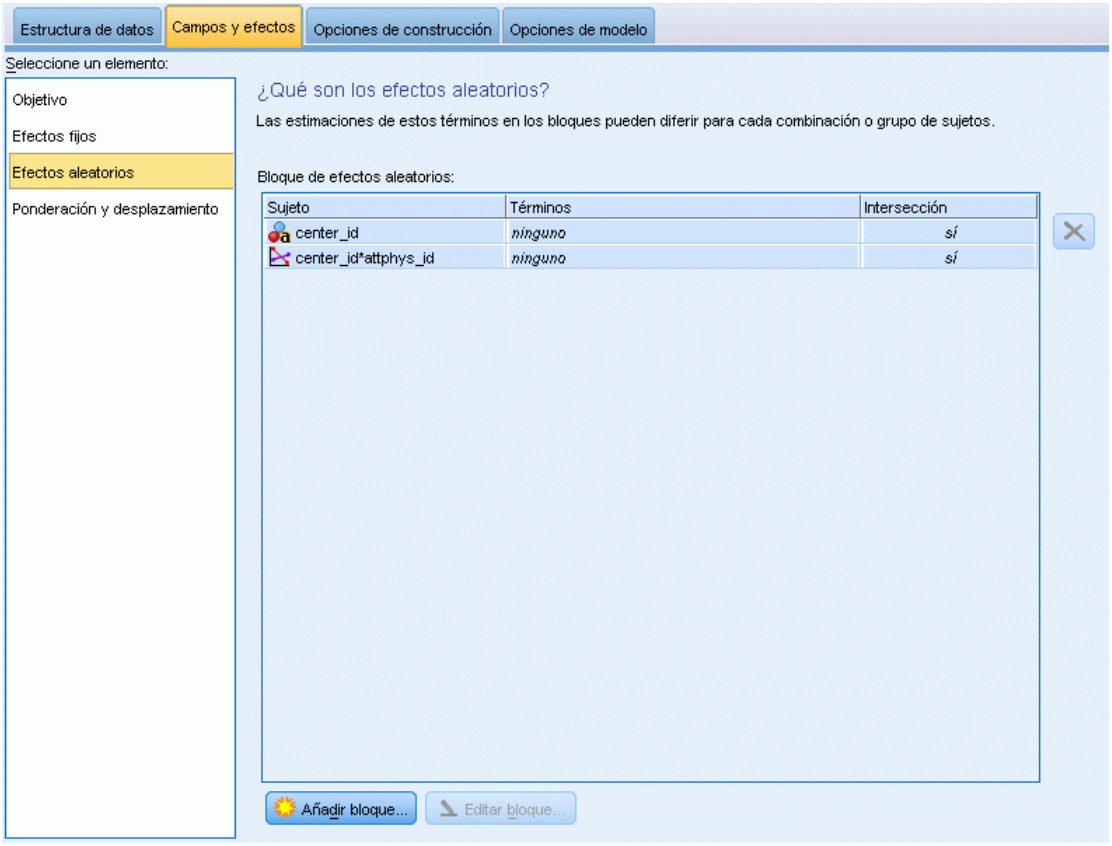
▶ Seleccione el factor dentro del cual el factor o covariable anterior se anida y, a continuación, pulse en el botón de flecha.

▶ Pulse en Añadir término.

Si lo desea, puede incluir efectos de interacción o añadir varios niveles de anidación al término anidado.

Efectos aleatorios

Figura 10-58
Configuración de 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. Por defecto, 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, solamente intersección)
- Efecto aleatorio 2: el sujeto es colegio * clase (sin efectos, solamente intersección)

Puede trabajar con bloques de efectos aleatorios de las maneras siguientes:

▶ Para añadir un nuevo bloque, pulse en Añadir bloque...Esto abrirá el cuadro de diálogo Bloque de efectos aleatorios .

- Para editar un bloque existente, seleccione el bloque que quiera editar y pulse en Editar bloque... Esto abrirá el cuadro de diálogo Bloque de efectos aleatorios .
- Para eliminar uno o más bloques, seleccione los bloques que quiera eliminar y pulse en el botón Eliminar.

Bloque de efectos aleatorios

Figura 10-59
Cuadro de diálogo Bloque de efectos aleatorios

Introduzca efectos en el modelo seleccionando uno o más campos en la lista de fuentes 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 que suelte aparece como una única interacción en la parte inferior de la lista de efectos.

Los botones a la derecha del generador de efectos le permiten hacer lo siguiente:

Eliminar términos del modelo de efectos fijos seleccionando los términos que quiera 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 hacia arriba o hacia abajo; y

Añadir términos anidados al modelo mediante el cuadro de diálogo Añadir un término personalizado , pulsando en el botón Añadir un término personalizado.

Incluir intersección. La inserción no está incluida en el modelo de efectos aleatorios por defecto. Si asume que los datos pasan por el origen, puede excluir la intersección.

Definir grupos de covarianzas por. Los campos especificados aquí definen conjuntos independientes de parámetros de covarianza de efectos aleatorios; uno por 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.

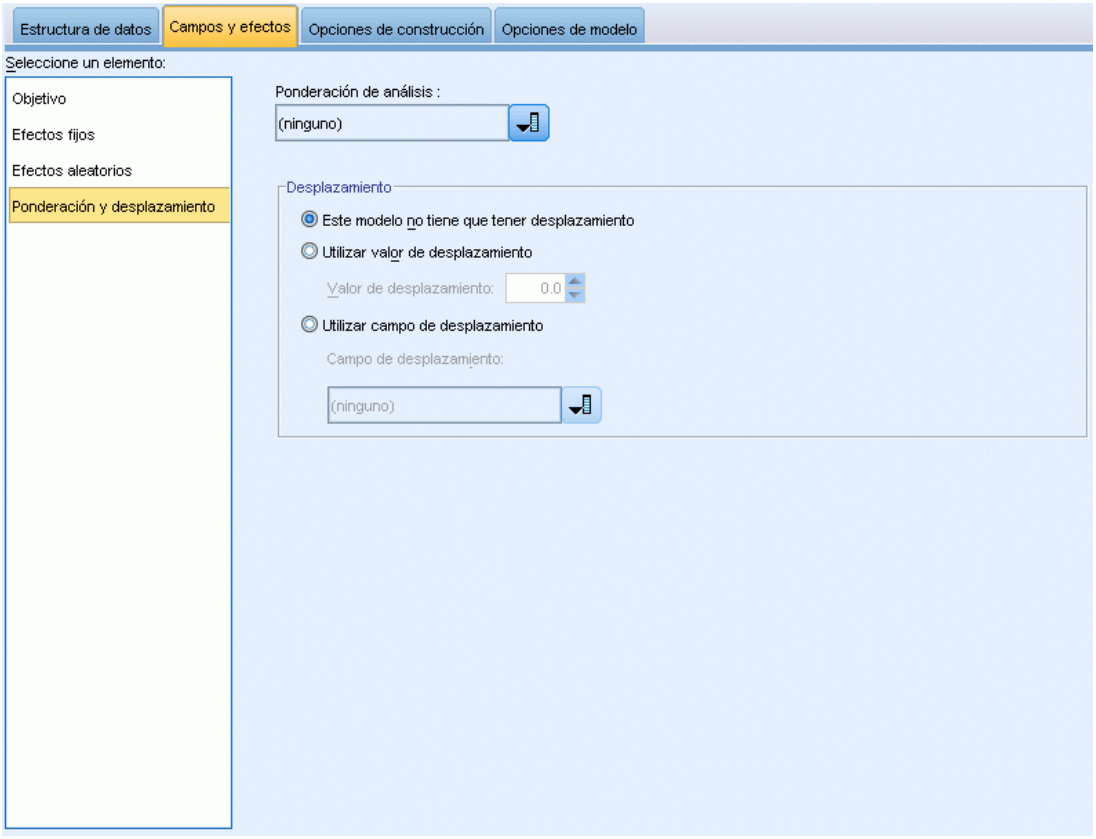
Combinación de sujetos. Esto le permite especificar sujetos de efectos aleatorios a partir de combinaciones de sujetos predefinidas desde la pestaña Estructura de datos. Por ejemplo, si *Colegio*, *Clase* y *Estudiante* se definen como sujetos en la pestaña Estructura de datos, en ese orden, entonces 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

Figura 10-60
Configuración de 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.

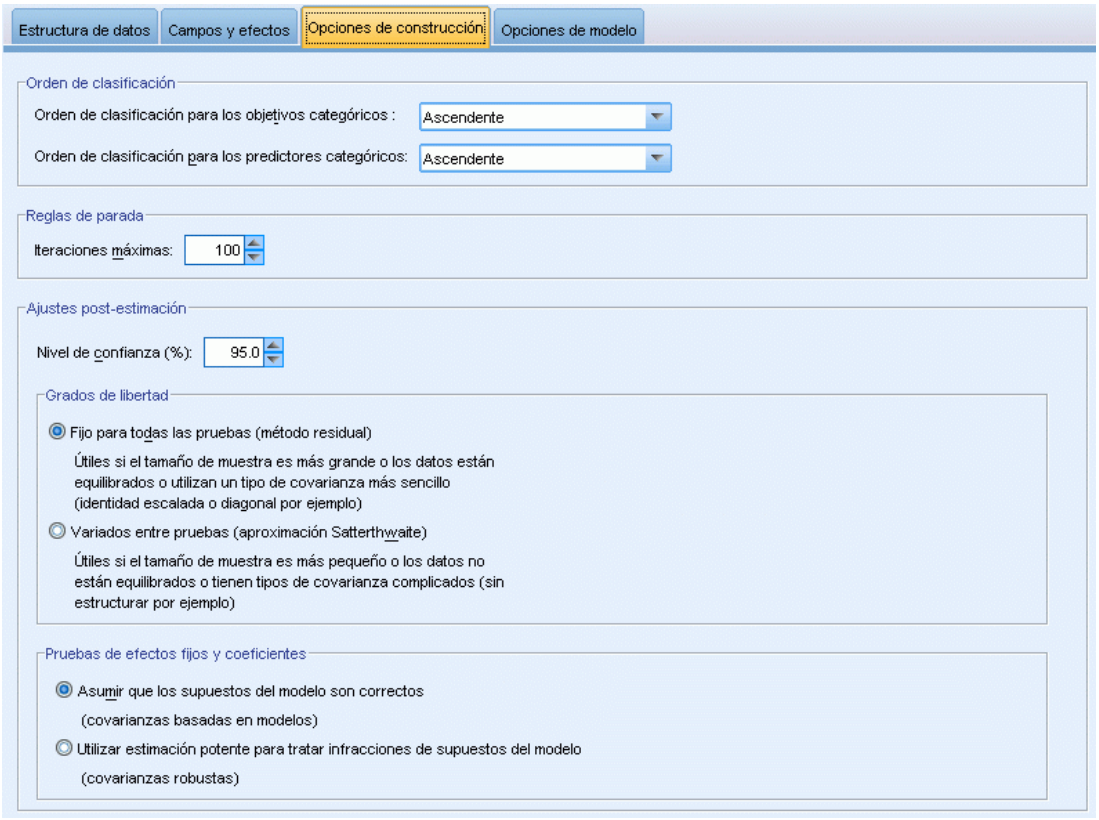
Desplazamiento. El término desplazamiento es un predictor "estructural". Su coeficiente no se estima por el modelo pero se supone que tiene el valor 1. Por tanto, los valores del desplazamiento se suman sencillamente al predictor lineal del destino. Esto resulta especialmente útil en los modelos de regresión 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

Figura 10-61
Configuración de opciones de generación



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 ordenación del objetivo se ignora si el objetivo no es categórico o si una categoría de referencia personalizada se especifica en la configuración de Objetivo .

Reglas de parada. Puede especificar el número máximo de iteraciones que se ejecutará el algoritmo. Especifique un número entero no negativo. El valor predeterminado es 100.

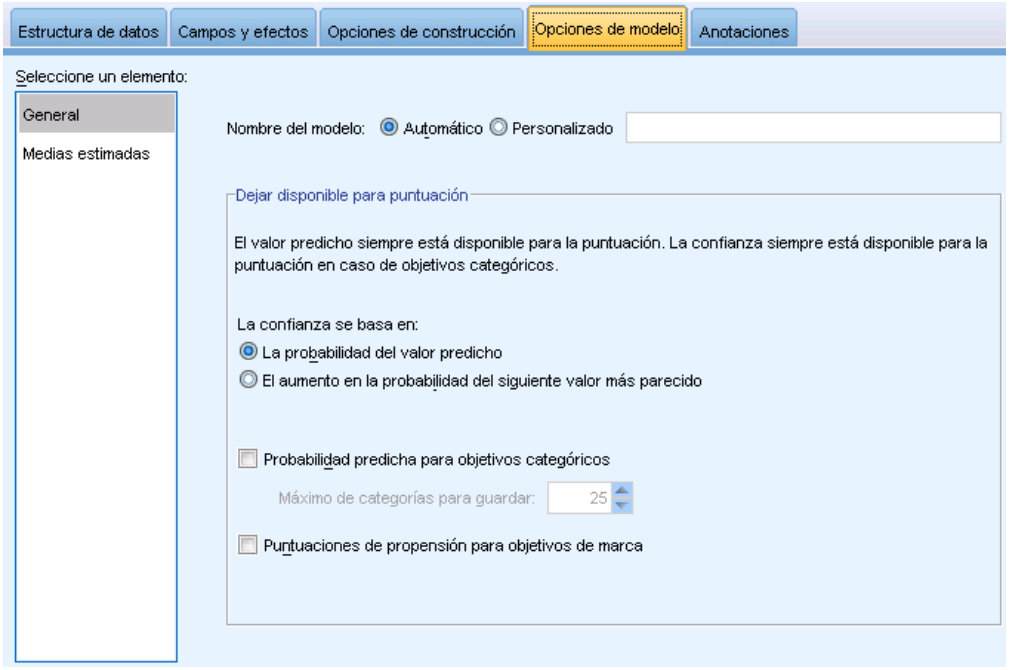
Configuración de estimación posterior. Estos ajustes determinan el modo en que algunos de los resultados de modelo se calculan para su visualización.

- **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 por defecto 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 por defecto. 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.

General

Figura 10-62
Configuración general



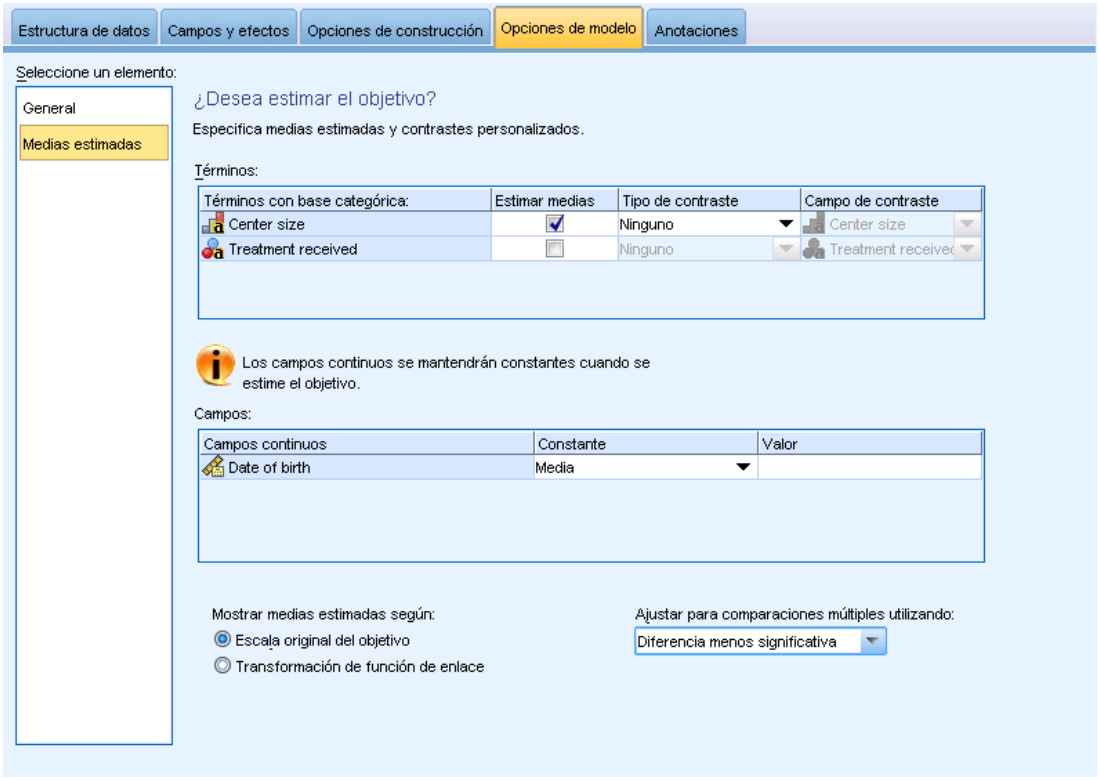
Nombre del modelo. Puede generar el nombre del modelo automáticamente tomando como base los campos objetivo o especificar un nombre personalizado. El nombre generado automáticamente es el nombre del campo objetivo. Si existen objetivos múltiples, el nombre del modelo se forma con los nombres de campos en orden, conectados por símbolos &. Por ejemplo, si *campo1 campo2 campo3* son objetivos, el nombre de modelo es: *campo1 & campo2 & campo3*.

Dejar disponible para puntuar. 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 un pronóstico 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 brutas de propensión; si hay particiones activas, el modelo también producirá puntuaciones de propensión ajustadas en función de la partición de prueba. Si desea obtener más información, consulte el tema Puntuaciones de propensión en el capítulo 3 el p. 45.

Medias estimadas

Figura 10-63
Configuración de 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.** Este es un procedimiento de Bonferroni 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.
- **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

Por defecto, se muestra la vista Resumen del modelo. Para ver otra vista de modelo, selecciónela entre las vistas en miniatura.

Resumen del modelo

Figura 10-64
Vista Resumen del modelo

Resumen del modelo

Objetivo: Number of convulsions

Objetivo		Number of convulsions
Distribución de probabilidad		Poisson
Función de enlace		Log
Criterio de información	**Akaike Corregida**	6,983.683
	Bayesiana	7,008.180

Los criterios de información se basan en Log-pseudo-verosimilitud-2 (6,975.671) y se utilizan para comparar modelos. Los modelos con valores de criterio de información menores se ajustan mejor. Cuando se comparan modelos utilizando valores de Log-pseudo-verosimilitud, se debe proceder con cuidado porque se pueden utilizar transformaciones de datos diferentes en los modelos.

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 -2 log verosimilitud (restringida). Los valores menores indican modelos mejores. El AICC "corrige" el AIC respecto a tamaños muestrales pequeños. A medida que aumenta el tamaño muestral, el AICC converge con el AIC.
- **Bayesiano.** Una medida para seleccionar y comparar modelos basada en la -2 log verosimilitud. Los valores menores indican modelos mejores. El BIC también penaliza los modelos sobreparametrizados, pero de manera 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

Figura 10-65
Vista de Estructura de datos

Estructura de datos

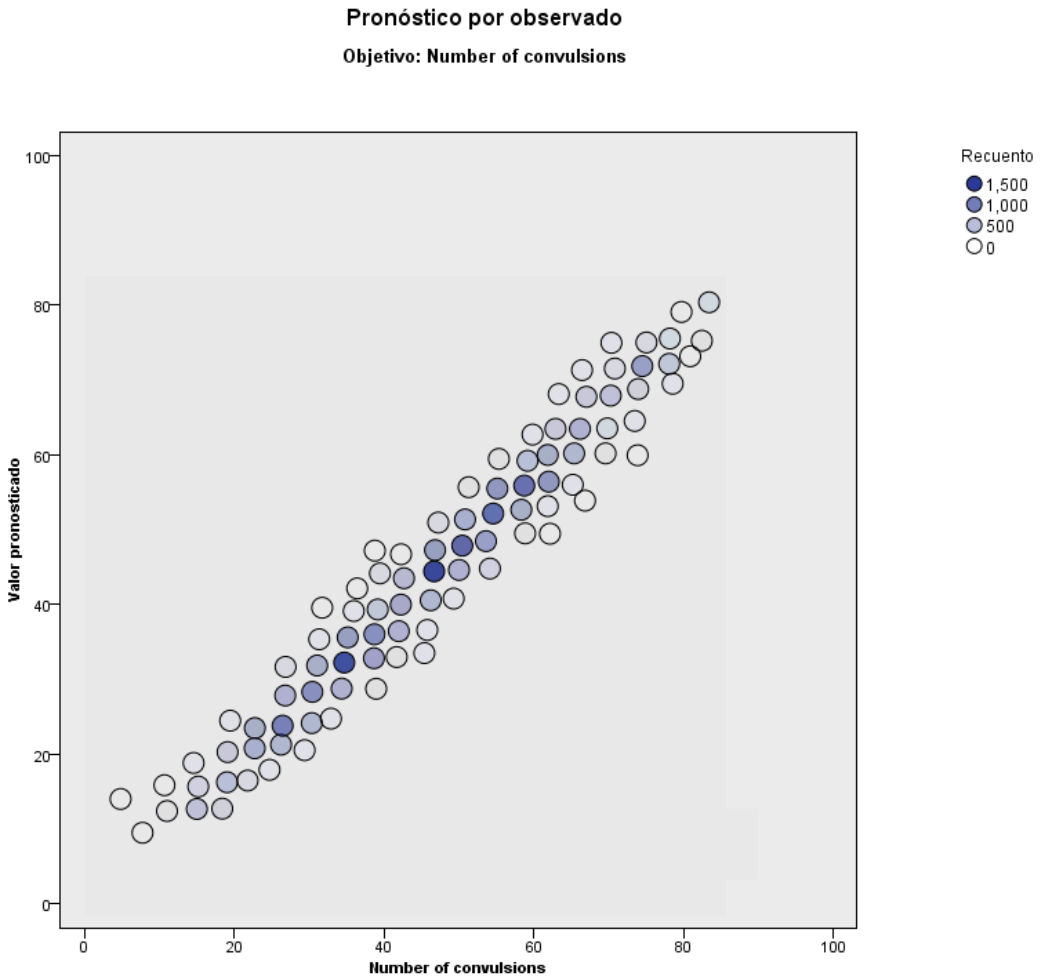
Objetivo: Number of convulsions

	Sujetos			Medidas repetidas	Objetivo
	Center ID	Attending Physician ID	Patient ID	Week	Number of convulsions
Datos del primer asunto	07057	COMX	1FSL	0	2
	07057	COMX	1FSL	1	6
	07057	COMX	1FSL	2	4
	07057	COMX	1FSL	3	4
	07057	COMX	1FSL	4	6
	07057	COMX	1FSL	5	3
Número total de niveles:	8	49	565	6	

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

Figura 10-66
Vista Predicho por observado



Para objetivos continuos, incluidos objetivos especificados como eventos/ensayos, muestra un diagrama de dispersión en intervalos de los valores pronosticados 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 pronosticado de manera incorrecta en el modelo.

Clasificación

Figura 10-67
Vista de Clasificación

Clasificación

Destino:Service usage

Porcentaje correcto global =85.2%

Observado	Pronosticado		
	No service	Other provider	Service with company
No service	60.6%	13.0%	26.4%
Other provider	1.5%	86.8%	11.7%
Service with company	0.8%	6.7%	92.5%

Porcentaje de filas
100.00
80.00
60.00
40.00
20.00
0.00

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 por defecto.
- **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 encabezados 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

Figura 10-68
Vista de Efectos fijos, estilo de diagrama

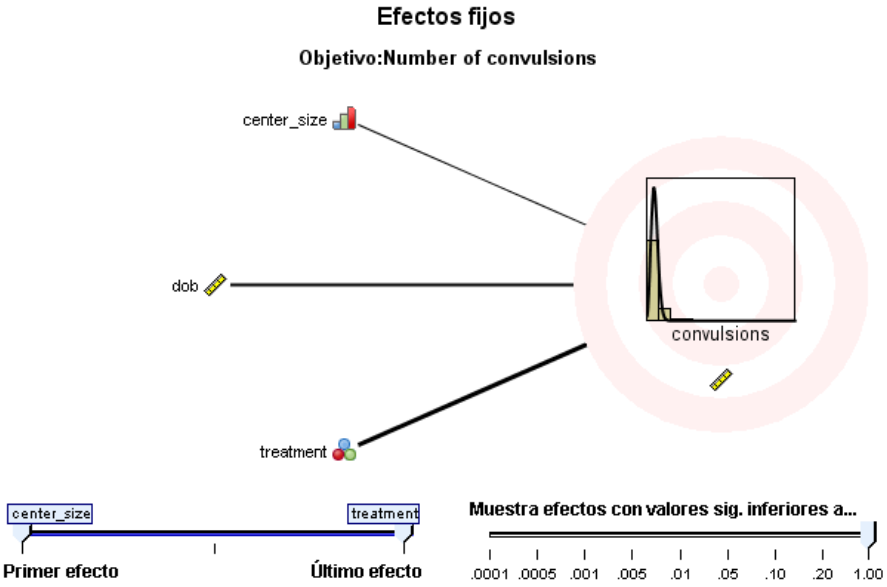


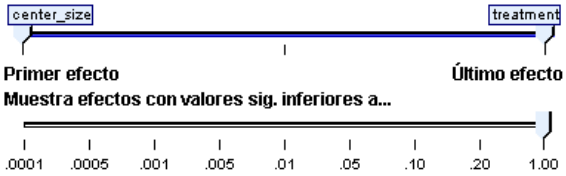
Figura 10-69
Vista de Efectos fijos, estilo de tabla

Efectos fijos

Objetivo:Number of convulsions

Origen	F	df1	df2	Sig.
Modelo corregido ▼	1.993	4	3,385	.093
center_size	1.340	2	3,385	.262
dob	1.759	1	3,385	.185
treatment	3.762	1	3,385	.053

Distribución de probabilidad:Poisson
Función de enlace:Log



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 por defecto.
- **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 por defecto es 1.00, de modo que no se filtran efectos tomando como base la significación.

Coeficientes fijos

Figura 10-70
Vista de Coeficientes fijos, estilo de diagrama

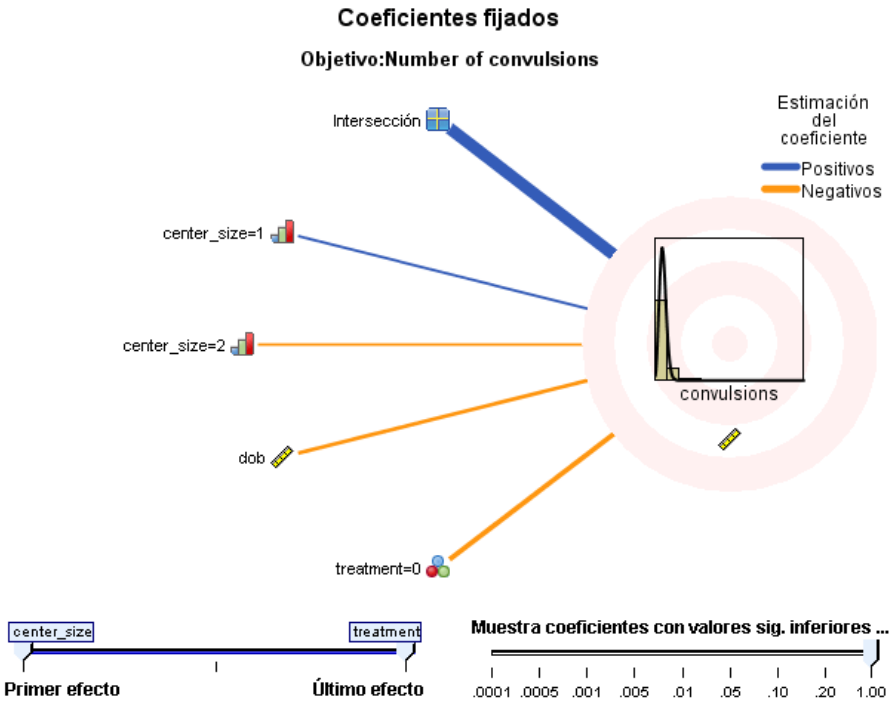


Figura 10-71
Vista de Coeficientes fijos, estilo de tabla

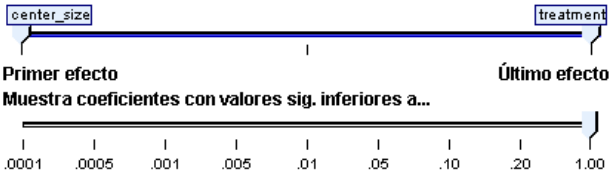
Coeficientes fijados

Objetivo:Number of convulsions

Término del modelo	Coeficiente ▶	Sig.	Exp (Coeficiente)
Intersección	1.745	.000	5.727
center_size=1	0.078	.610	1.081
center_size=2	-0.157	.201	0.855
center_size=3	0[a]		
dob	-0.000	.185	1.000
treatment=0	-0.138	.053	0.871
treatment=1	0[a]		

Distribución de probabilidad:Poisson
Función de enlace:Log

[a]Este coeficiente está establecido en cero porque es redundante.



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 intersección 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 por defecto.

- **Tabla.** Muestra los valores, las pruebas de significación y los intervalos de confianza para los coeficientes de modelos individuales. Después de la intersección, 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 por defecto 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 por defecto.
- **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 los encabezados 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

Figura 10-72
Vista de Parámetros de covarianza

Parámetros de covarianza

Objetivo:Number of convulsions

Parámetros de covarianza	**Efecto residual**	2
	Efectos aleatorios	2
Columnas de matriz de diseño	**Efectos fijos**	7
	Efectos aleatorios	10[a]
Sujetos comunes		8

Los sujetos comunes se basan en las especificaciones del sujeto de los efectos residuales y aleatorios y se utilizan para fragmentar los datos para mejorar el rendimiento.

[a]Es el número de columnas por sujetos comunes.

Efecto residual	Estimación	Error típico	Z	Sig.	Intervalo de confianza 95%	
					Inferior	Superior
AR1 Diagonal	5.379	0.243	22.151	.000	4.923	5.876
AR1 Rho	0.805	0.010	82.823	.000	0.785	0.824

Estructura de covarianza:Autorregresiva de primer orden
Especificación de sujeto:center_id*attphys_id*patient_id

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 típico 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 típico y el intervalo de confianza para cada combinación de niveles de los campos del efecto; el resto de predictores se mantiene constante.

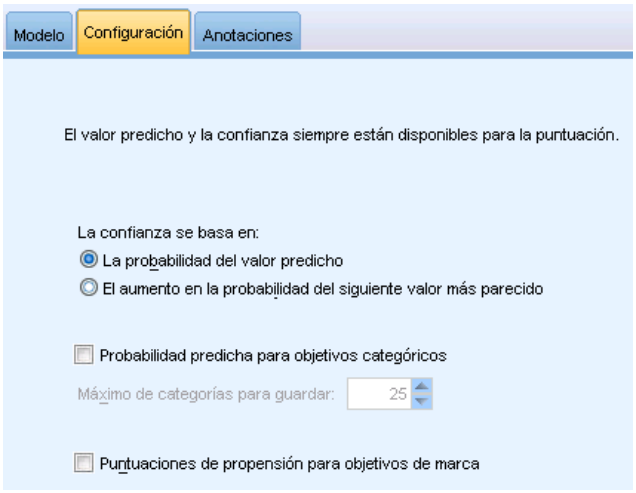
  Si se solicitan contrastes, se muestra otra tabla con la estimación, el error típico, la prueba de significación y el intervalo de confianza para cada contraste; para las interacciones, hay un conjunto de filas independiente para cada combinación de niveles de los efectos distintos del campo Contraste. 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

Figura 10-73
Configuración del modelo



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 un pronóstico 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 brutas de propensión; si hay particiones activas, el modelo también producirá puntuaciones de propensión ajustadas en función de la partición de prueba. Si desea obtener más información, consulte el tema Puntuaciones de propensión en el capítulo 3 el p. 45.

Nodo Cox

La regresión de Cox crea un modelo predictivo para datos de tiempo hasta el evento. El modelo produce una función de supervivencia que pronostica la probabilidad de que el evento de interés se haya producido en el momento dado *t* para valores determinados de las variables del predictor. 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 medidas 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 este 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 Si desea obtener más información, consulte el tema Uso de la regresión de Cox en el modelo de tiempo de abandono de cliente en el capítulo 26 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

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. Si desea obtener más información, consulte el tema Conversión del almacenamiento mediante el nodo Rellenar en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.) Se ignorarán los campos definidos como *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.

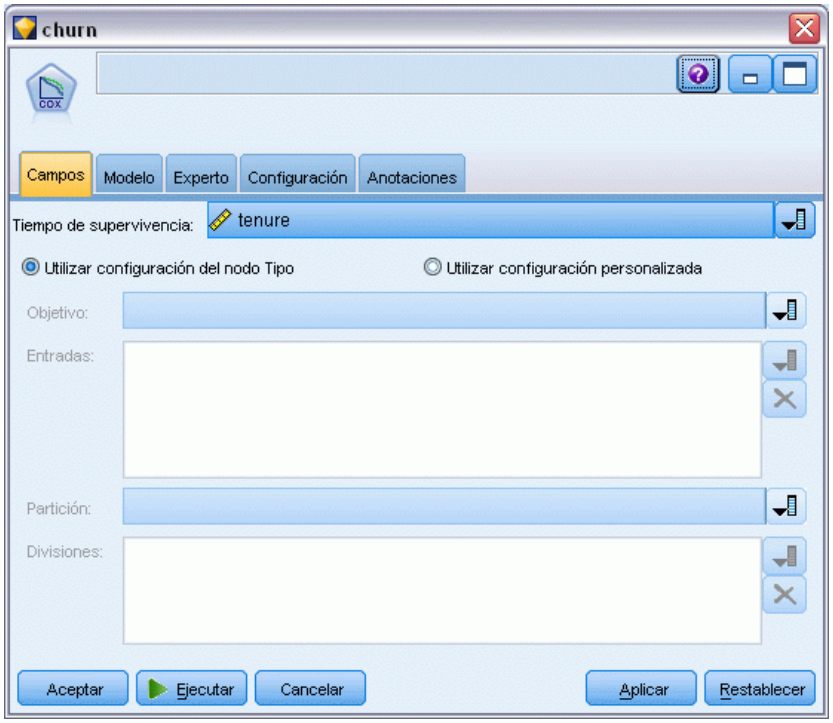
Fechas y Horas. Los campos Fecha y Hora no se pueden utilizar para definir directamente el tiempo de supervivencia; si tiene campos Fecha y Hora debe utilizarlos para crear un campo que contenga tiempos de supervivencia, basados en la diferencia entre la fecha de entrada en el estudio

y la fecha de observación. Si desea obtener más información, consulte el tema Cómo trabajar con fechas y horas en el capítulo 7 en *Manual de usuario de IBM SPSS Modeler 15*.

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

Figura 10-74
Cuadro de diálogo del nodo Cox, pestaña Campos



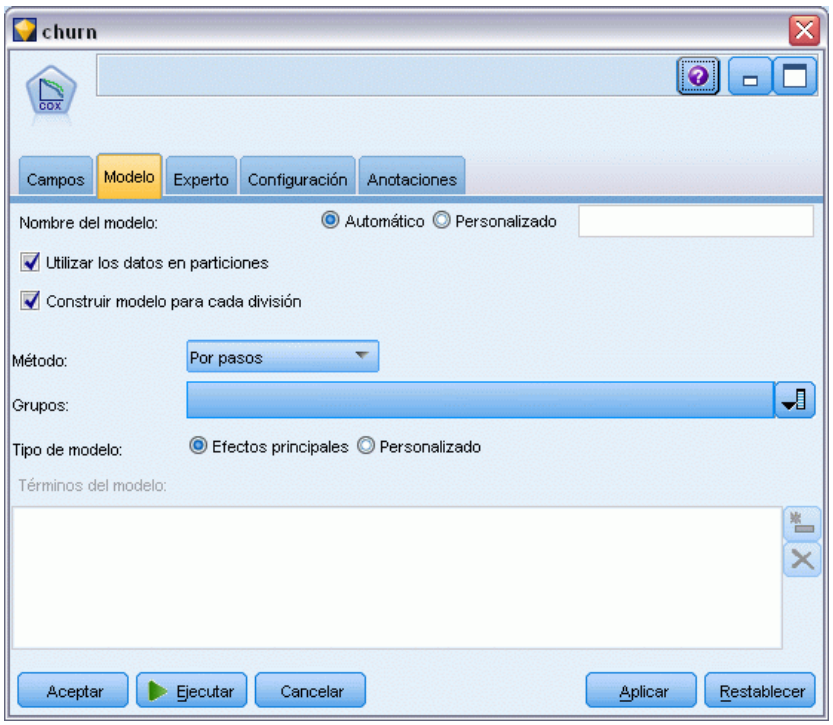
Tiempo de supervivencia. Seleccione un campo numérico (uno con un nivel de medición de *Continuo*) para que el nodo se pueda ejecutar. El tiempo de supervivencia indica la vida útil del registro que se está pronosticando. Por ejemplo, si modela el tiempo de abandono de cliente, éste será el campo que 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 relevante el 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. Si desea obtener más información, consulte el tema Opciones de los campos del nodo de modelado en el capítulo 3 el p. 38.

Opciones de modelo para el nodo Cox

Figura 10-75
Cuadro de diálogo del nodo Cox, pestaña 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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Crear modelos divididos. Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como modelos divididos. Si desea obtener más información, consulte el tema Generación de modelos divididos en el capítulo 3 el p. 32.

Método. Éstas son las opciones disponibles para introducir predictores en el modelo:

- **Intro.** Éste es el método por defecto 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. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

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.

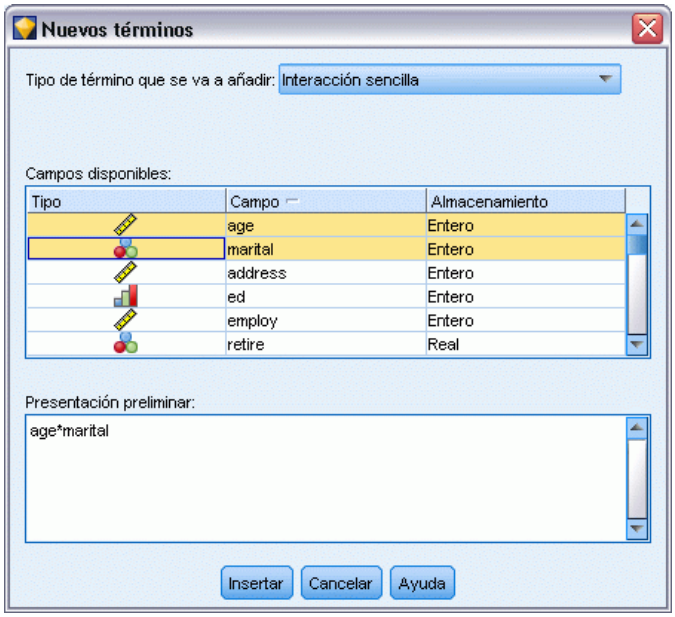
Términos del modelo. Al crear un modelo personalizado, deberá especificar explícitamente los términos del modelo. La lista muestra el conjunto actual de términos para el modelo. Los botones situados en la parte derecha de la lista Términos del modelo le permitirán añadir y eliminar los términos del modelo.

- 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.

Figura 10-76
Cuadro de diálogo Nuevos 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*.

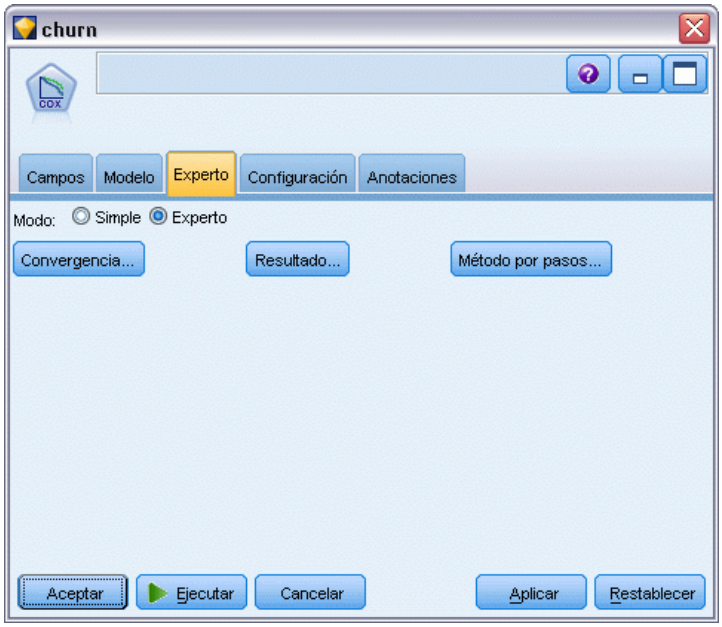
Campos disponibles. Muestra los campos de entrada disponibles que se van a utilizar en la construcción de términos del modelo. Tenga en cuenta que la lista puede incluir campos que no son campos de entrada correctos, por lo que asegúrese de que todos los términos del modelo incluyen sólo campos de entrada.

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

Figura 10-77
Cuadro de diálogo del nodo Cox, pestaña Experto



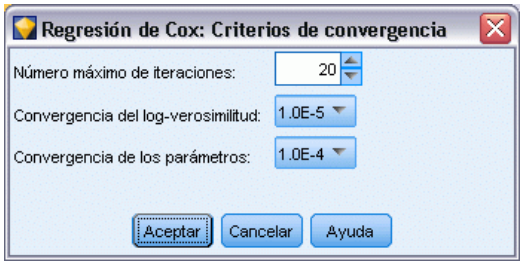
Convergencia. Estas opciones le permiten controlar los parámetros de la convergencia del modelo. Cuando se ejecuta el modelo, la configuración de la convergencia controla cuántas veces se ejecutan los distintos parámetros a través de éste para comprobar si se ajustan. Cuanta más veces se prueben los parámetros, más próximos estarán los resultados (es decir, los resultados convergirán). Si desea obtener más información, consulte el tema Criterios de convergencia del nodo Cox el p. 365.

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. Si desea obtener más información, consulte el tema Opciones de resultados avanzados del nodo Cox el p. 365.

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.) Si desea obtener más información, consulte el tema Criterios del método por pasos del nodo Cox el p. 366.

Criterios de convergencia del nodo Cox

Figura 10-78
Cuadro de diálogo Criterios de convergencia del nodo Cox



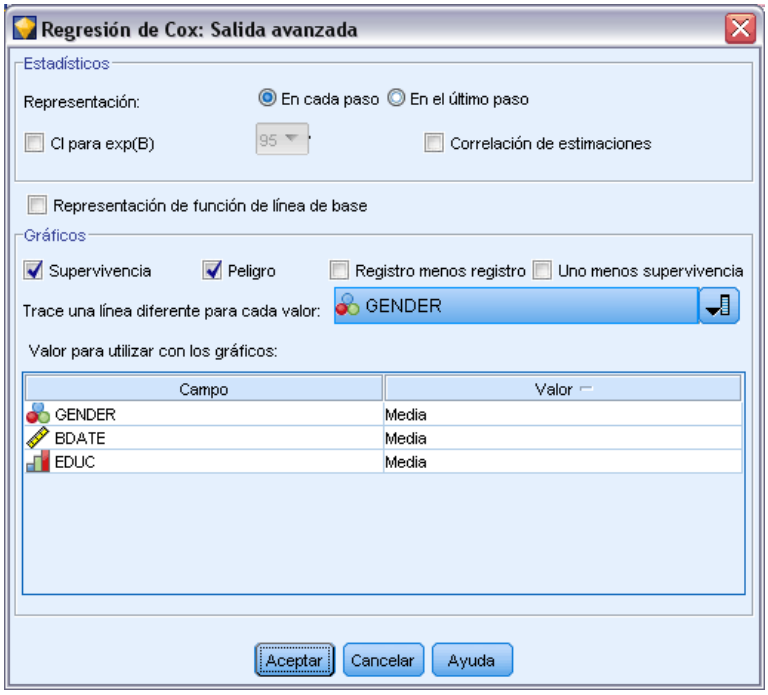
Iteraciones máximas. Permite especificar el número máximo de iteraciones para el modelo, que controla durante cuánto tiempo el procedimiento buscará una solución.

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

Figura 10-79
Cuadro de diálogo Resultado avanzado de regresión de Cox



Estadísticos. Es posible obtener estadísticos para los parámetros del modelo, incluyendo los intervalos de confianza para exp(*B*) y la correlación de las estimaciones. Es posible solicitar estos estadísticos en cada paso o solamente en el último paso.

Mostrar la función de línea base. Permite visualizar la función de impacto basal y la supervivencia acumulada en la media de las covariables.

Gráficos

Los gráficos pueden ayudarle a evaluar el modelo estimado e interpretar los resultados. Es posible representar gráficamente las funciones de supervivencia, de impacto, log-menos-log y uno menos la supervivencia.

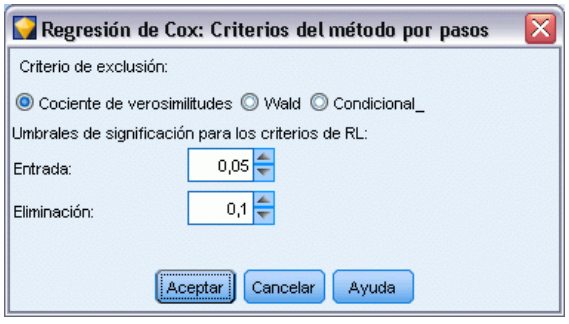
- **Superviv..** Muestra la función de supervivencia acumulada, en una escala lineal.
- **Impacto.** Muestra la función de impacto acumulada, en una escala lineal.
- **Log menos log.** Muestra la estimación de supervivencia acumulada después de aplicar la transformación In(-In) a la estimación.
- **Uno 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 por defecto 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

Figura 10-80
Cuadro de diálogo Criterios del método por pasos de regresión de Cox



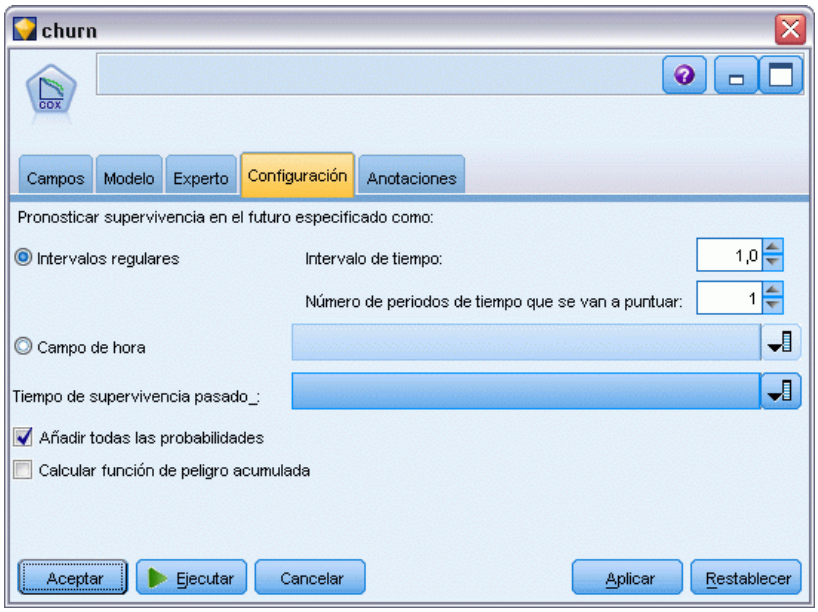
Criterio de exclusión. Seleccione Cociente de verosimilitudes para un modelo más robusto. Si desea reducir el tiempo necesario para generar el modelo, puede intentar seleccionar Wald. Existe la opción adicional Condicional que permite una comprobación de eliminación en función de la

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

Figura 10-81
Cuadro de diálogo del nodo Cox, pestaña Configuración



Pronosticar 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 pronostica para cada registro en cada valor de tiempo, un pronóstico 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 pronosticará 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 pronóstico), 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 futuro y los tiempos de supervivencia pasados deben situarse dentro del 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 pronosticada. Las probabilidades se calculan para cada tiempo futuro.

Calcular función de peligro acumulada. Especifique si el valor del peligro acumulado se añade a cada registro. El peligro 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 el pronóstico del modelo y la probabilidad asociada. Los nombres de los nuevos campos se derivan del nombre del campo de salida que se está pronosticando, con el prefijo *$C-* para la categoría pronosticada y *$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 peligro acumulada en el nodo Cox, se añadirá un campo adicional para cada tiempo futuro, que contenga la función de peligro acumulada 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*.

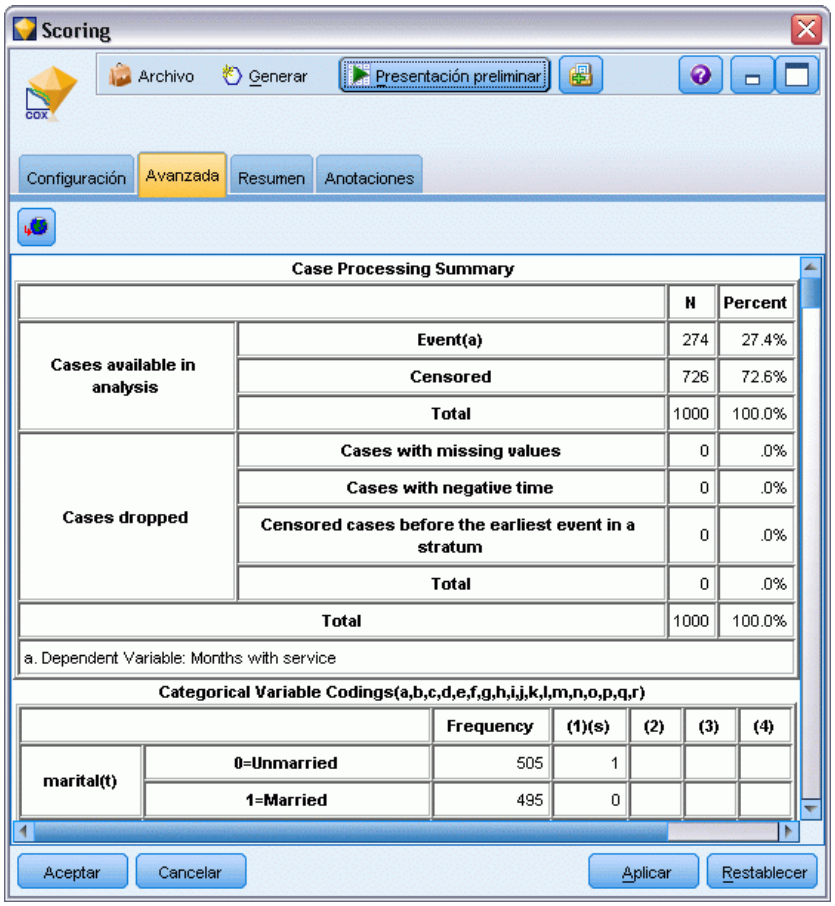
Configuración de resultados de regresión de Cox

La pestaña Configuración del nugget contiene el mismo control que la pestaña Configuración del nodo del modelo. Los valores por defecto de los controles del nugget se determinan por el conjunto de valores definidos en el nodo de modelo. Si desea obtener más información, consulte el tema Opciones de configuración para el nodo Cox el p. 367.

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.

Figura 10-82
Nugget de modelo Cox, pestaña Avanzada



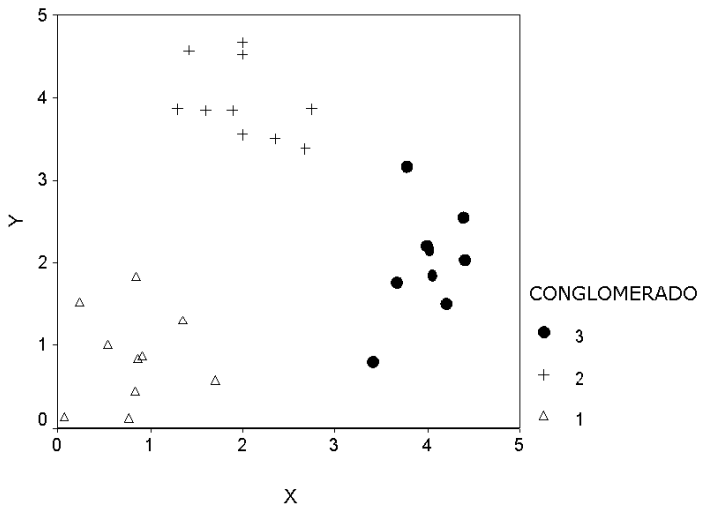
Capítulo
11

Modelos de conglomerados

Los modelos de conglomerados 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 buscar. Esto es lo que diferencia a los modelos de conglomerados de otras técnicas de aprendizaje de las máquinas: no hay campo objetivo o salida predefinidos para el modelo que se va a pronosticar. 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 conglomerado se basan en la medición de distancias entre registros y entre conglomerados. Los registros se asignan a los conglomerados de un modo que tiende a minimizar la distancia entre los registros pertenecientes al mismo conglomerado.

Figura 11-1
Modelo de conglomerado simple



Se ofrecen tres métodos de conglomeración:

El nodo K-medias agrupa conjuntos de datos en grupos distintos (o conglomerados). El método define un número fijo de conglomerados, de forma iterativa asigna registros a los conglomerados y ajusta los centros de los conglomerados hasta que no se pueda mejorar el modelo. En lugar de intentar pronosticar 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. Si desea obtener más información, consulte el tema Nodo K-medias el p. 378.

El nodo Bietápico es un método de conglomerado 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 subconglomerados administrable. El segundo paso utiliza un método de conglomerado jerárquico para fundir progresivamente los subconglomerados en conglomerados cada vez más grandes. El bietápico tiene la ventaja de estimar automáticamente el número óptimo de conglomerados para los datos de entrenamiento. Puede gestionar tipos de campos mixtos y grandes conjuntos de datos eficazmente. Si desea obtener más información, consulte el tema Nodo de conglomerado Bietápico el p. 383.

El nodo Kohonen genera un tipo de red neuronal que se puede usar para conglomerar 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 conglomerados. Si desea obtener más información, consulte el tema Nodo Kohonen el p. 371.

Los modelos de conglomerado se usan a menudo para crear conglomerados 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.

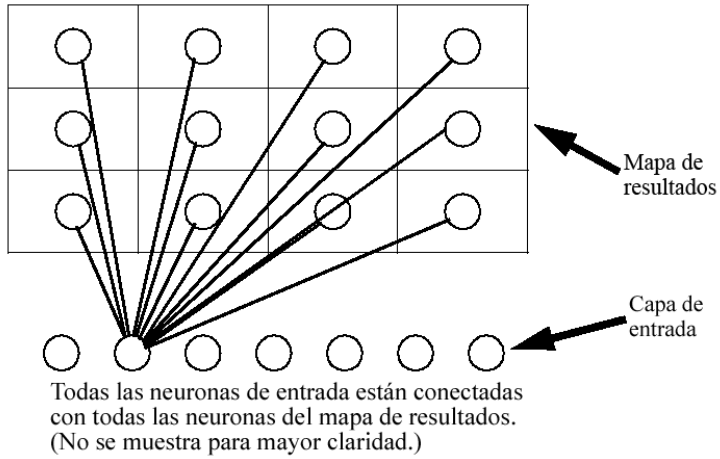
Nodo Kohonen

Las redes de Kohonen son un tipo de red neuronal que realiza conglomerados, también conocidos como **knet** o como un **mapa autoorganizativo**. Este tipo de redes se puede utilizar para conglomerar el conjunto de datos en distintos grupos cuando no se sabe lo que son al principio. Los registros se agrupan de manera que los de un mismo grupo o conglomerado tiendan a ser similares entre ellos y que los de otros grupos sean distintos.

Las unidades básicas son las **neuronas**, que 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. A continuación se detalla un mapa de 3 × 4, aunque los mapas suelen tener un tamaño mayor que éste.

Figura 11-2
Estructura de una red de Kohonen



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 conglomerados (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 conglomerados.

Otro uso de las redes de Kohonen es el de la **reducción de dimensión**. La característica espacial de las cuadrículas bidimensionales permite una correspondencia desde los predictores originales *k* a dos funciones 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 por defecto de la cuadrícula de salida ha cambiado con respecto a versiones anteriores de SPSS Modeler. El nuevo método suele generar capas de salida más pequeñas, que se entrenan más rápidamente y se generalizan mejor. Si obtiene

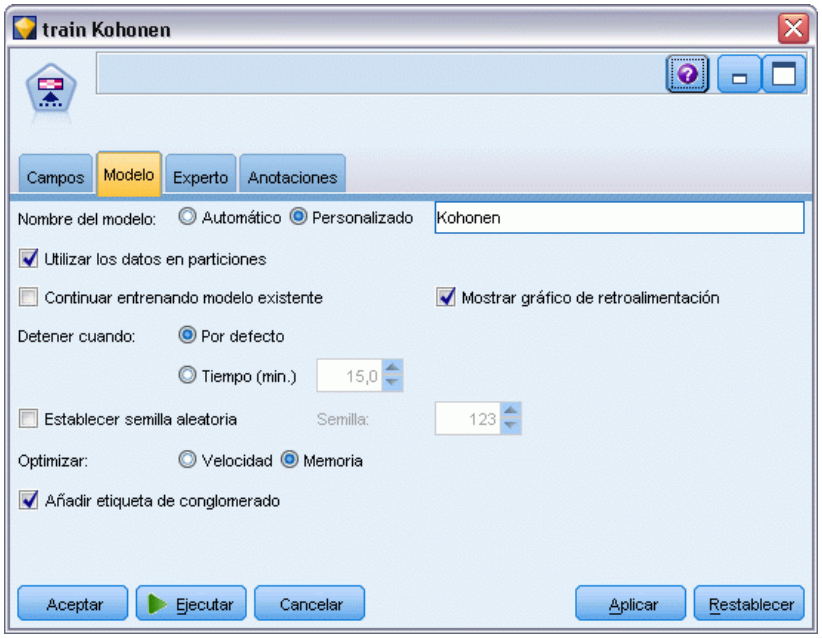
unos resultados no significativos con el tamaño por defecto, intente aumentar el tamaño de la cuadrícula de salida en la pestaña Experto. Si desea obtener más información, consulte el tema Opciones de experto para el nodo Kohonen el p. 375.

Requisitos. Para entrenar una red de Kohonen, necesita uno o más campos con su papel establecido como *Entrada*. Se ignorarán los campos con el papel 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 conglomerados 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 conglomerados.

Opciones de modelo para el nodo Kohonen

Figura 11-3
Opciones de modelo para el nodo Kohonen



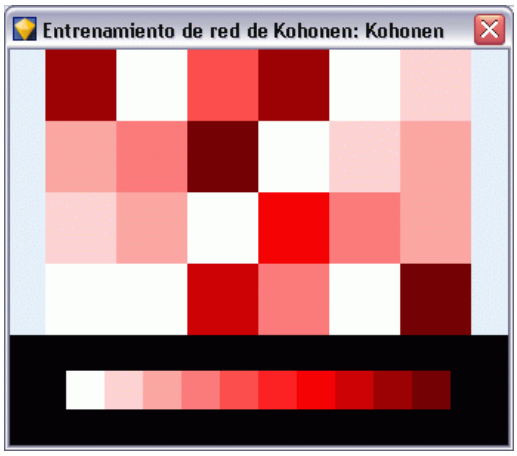
Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Continuar entrenando modelo existente. Por defecto, 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 función puede ralentizar el tiempo de entrenamiento. Para acelerarlo, anule la selección de esta opción.

Figura 11-4
Gráfico de retroalimentación de Kohonen



Detener cuando. Los criterios de parada por defecto detienen el entrenamiento en función de los 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. Si desea obtener más información, consulte el tema Nodo Ordenar en el capítulo 3 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

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 por defecto.

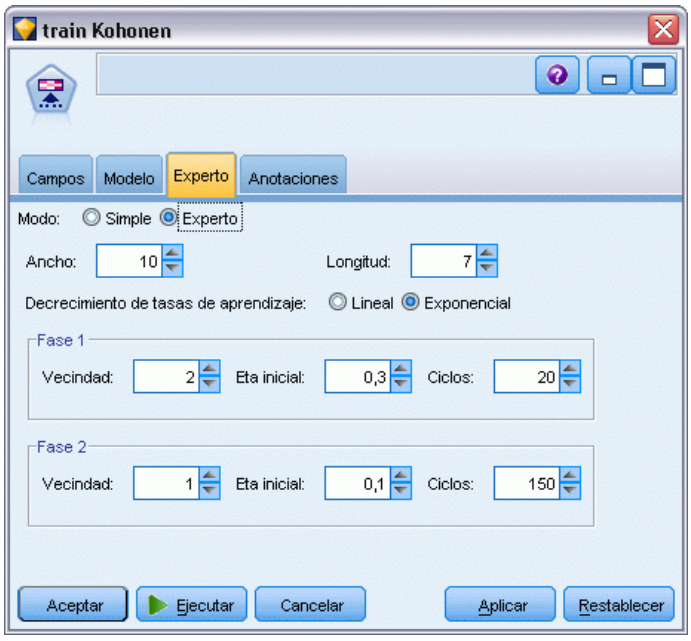
  Nota: cuando se ejecuta en modo distribuido, esta configuración puede quedar anulada por las opciones del administrador especificadas en *options.cfg*. Si desea obtener más información, consulte el tema Uso del archivo options.cfg en el capítulo 4 en *Guía de rendimiento y administración de IBM SPSS Modeler Server 15*.

Añadir etiqueta de conglomerado. Seleccionada por defecto 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 Autoconglomeración cuando se calculan medidas de ordenación de los diferentes tipos de modelos. Si desea obtener más información, consulte el tema Nodo Autoconglomeración en el capítulo 5 el p. 113.

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.

Figura 11-5
Opciones de experto para Kohonen



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 conglomerados, que ofrece una representación gráfica de los conglomerados, campos y niveles de importancia. Si desea obtener más información, consulte el tema Visor de conglomerados - Pestaña Modelo el p. 388.

Si prefiere visualizar los conglomerados 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.

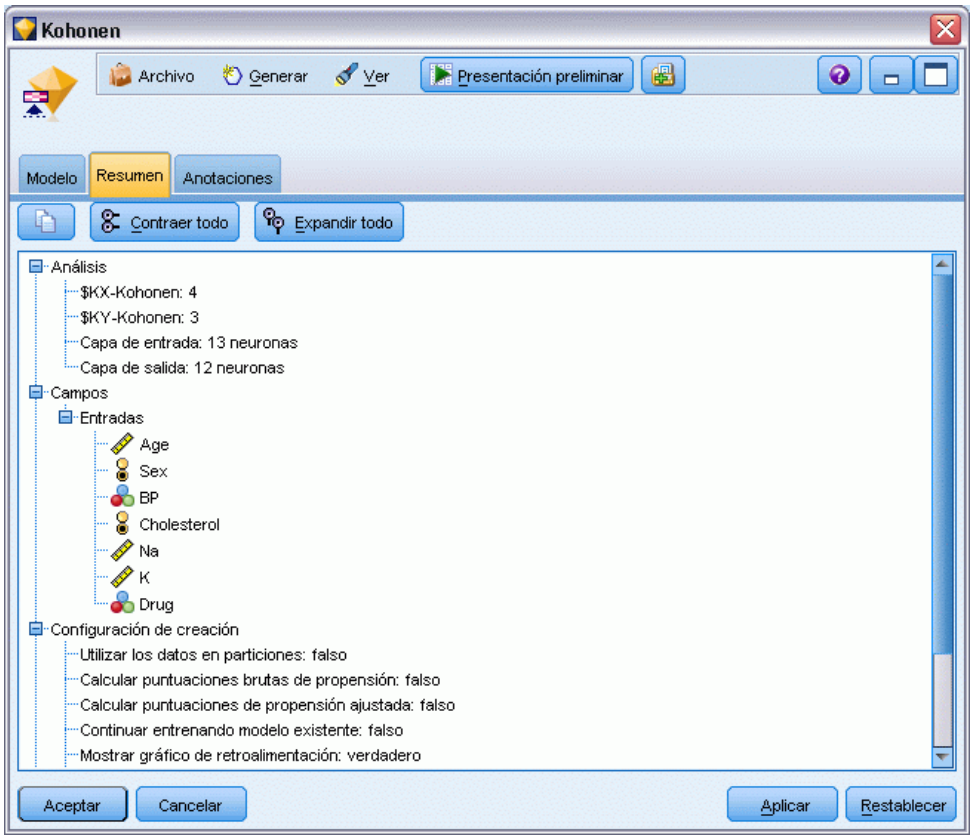
Otra buena forma de comprender la red de Kohonen es utilizar la inducción de reglas para descubrir las características que distinguen los conglomerados que la red ha encontrado. Si desea obtener más información, consulte el tema Nodo C5.0 en el capítulo 6 el p. 174.

Si desea obtener información general sobre el explorador de modelos, consulte Exploración de nugget de modelo

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 Kohonen de funciones bidimensional (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.

Figura 11-6
Pestaña Resumen de nugget de modelo de Kohonen



Nodo K-medias

El nodo K-medias ofrece un método de **análisis de conglomerados**. Se puede utilizar para conglomerar el conjunto de datos en distintos grupos 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 conglomerado tiendan a ser similares entre ellos, y que los de otros grupos sean distintos.

K-medias empieza definiendo un conjunto de centros de conglomerados iniciales derivados de datos. Después asigna cada registro al conglomerado de registros más similares, basándose en los valores de los campos de entrada de registros. Una vez asignados todos los casos, los centros de conglomerados se actualizan para reflejar el nuevo conjunto de registros asignados a cada conglomerado. Los registros se vuelven a comprobar para ver si se deben reasignar a otro conglomerado, y el proceso de iteración de conglomerado/asignación continúa hasta que se alcanza el número máximo de iteraciones o el cambio entre una iteración y otra no sobrepasa el umbral especificado.

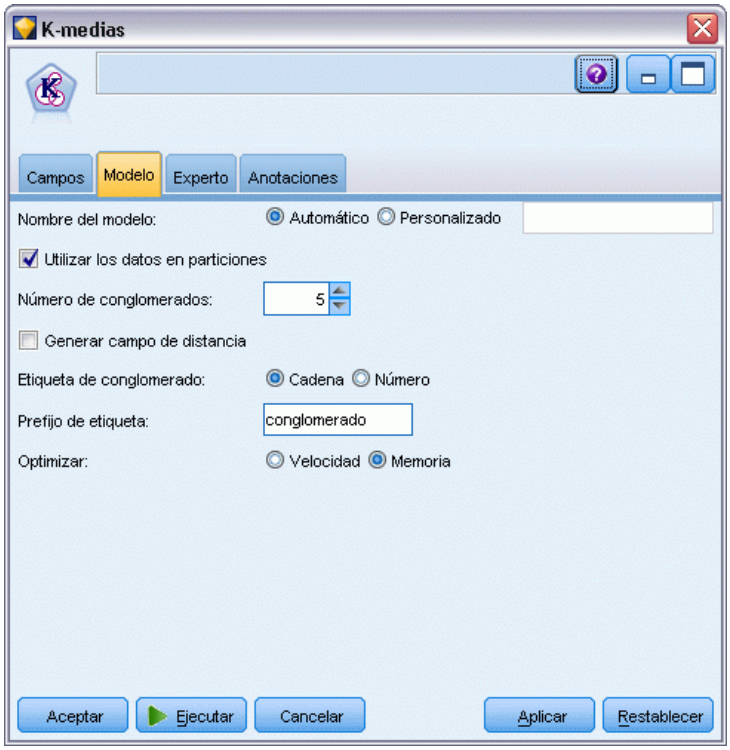
Nota: el modelo resultante depende, hasta cierto punto, del orden de los datos de entrenamiento. Reordenar los datos y regenerar el modelo puede dar como resultado un modelo de conglomerados final distinto.

Requisitos. Para entrenar un modelo K-Means, necesita uno o más campos con su papel establecido como *Entrada*. Se ignorarán los campos con el papel 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 conglomerado para conjuntos de datos grandes.

Opciones de modelo para el nodo K-medias

Figura 11-7
Opciones de modelo para el nodo K-medias



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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Número de conglomerados especificado. Especifique el número de conglomerados que generar. El valor por defecto 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 conglomerado que le ha sido asignado.

Etiqueta de conglomerado. Especifique el formato de los valores del campo de pertenencia al conglomerado generado. La pertenencia a un conglomerado se puede indicar con una Cadena con el Prefijo de etiqueta especificado (por ejemplo "Cluster 1", "Cluster 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.

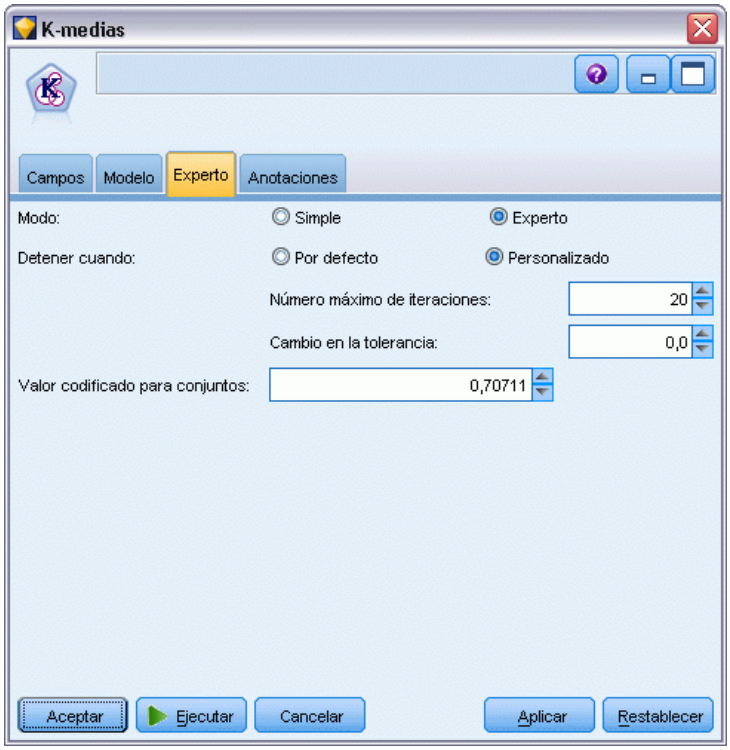
- Seleccione Velocidad para indicar al algoritmo que nunca debe recurrir al volcado en disco para mejorar el rendimiento.
- Seleccione Memoria para indicar al algoritmo que utilice el volcado en disco cuando lo considere oportuno en detrimento de la velocidad. Esta opción está seleccionada por defecto.

  Nota: cuando se ejecuta en modo distribuido, esta configuración puede quedar anulada por las opciones del administrador especificadas en *options.cfg*. Si desea obtener más información, consulte el tema Uso del archivo options.cfg en el capítulo 4 en *Guía de rendimiento y administración de IBM SPSS Modeler Server 15*.

Opciones de experto para el nodo K-medias

Para aquellos con conocimientos avanzados de los conglomerados 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.

Figura 11-8
Opciones de experto para K-medias



Detener cuando. Especifique los criterios de parada que utilizar en el entrenamiento del modelo. El criterio de parada Por defecto es 20 iteraciones o cambiar en < 0.000001 (lo que ocurra primero). Seleccione Personalizado para especificar sus propios criterios de parada.

- **Nº 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 conglomerados 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 por defecto 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 conglomerados, 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 conglomerado y la distancia a partir del centro del conglomerado asignado para ese registro. Del nombre del modelo se derivan los nuevos nombres de campos con el prefijo *$KM-* para la pertenencia del conglomerado y *$KMD-* para la distancia desde el centro del conglomerado. Por ejemplo, si el modelo se llama *Kmeans*, los nuevos campos se llamarían *$KM-Kmeans* y *$KMD-Kmeans*.

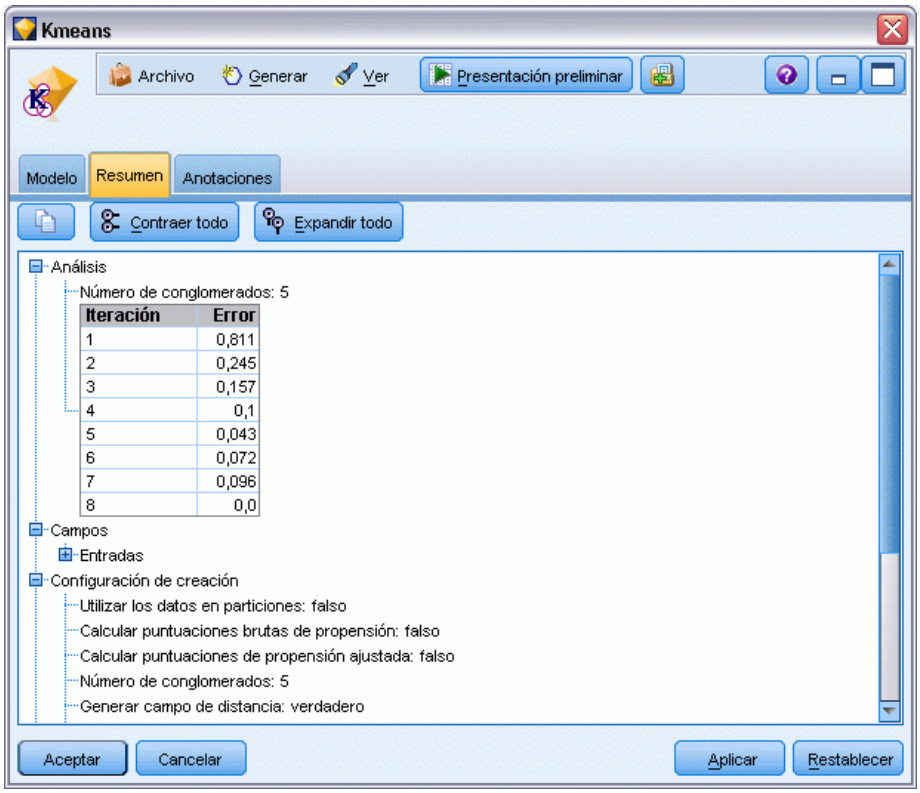
Una buena forma de comprender el modelo K-medias es utilizar la inducción de reglas para descubrir las características que distinguen los conglomerados que el modelo ha encontrado. Si desea obtener más información, consulte el tema Nodo C5.0 en el capítulo 6 el p. 174. Asimismo, puede pulsar en la pestaña Modelo del explorador del nugget de modelo para ver el visor de conglomerados, que ofrece una representación gráfica de conglomerados, campos y niveles de importancia. Si desea obtener más información, consulte el tema Visor de conglomerados - Pestaña Modelo el p. 388.

Si desea obtener información general sobre el explorador de modelos, consulte Exploración de nugget de modelo

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 conglomerados definidos por el modelo. Se muestra el número de conglomerados 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. Si desea obtener más información, consulte el tema Nodo Análisis en el capítulo 6 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Figura 11-9
Pestaña Resumen de nugget de modelo de K-medias



Nodo de conglomerado Bietápico

El nodo de conglomerado Bietápico ofrece un método de **análisis de conglomerados**. Se puede utilizar para conglomerar el conjunto de datos en distintos grupos cuando no se sabe lo que son al principio. Al igual que los nodos Kohonen y K-medias, los modelos de conglomerados bietápicos *no* utilizan un campo objetivo. En lugar de intentar predecir un resultado, el conglomerado 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 conglomerado tiendan a ser similares entre ellos, y que los de otros grupos sean distintos.

El conglomerado Bietápico es un método de conglomerado 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 subconglomerados que se puede administrar. El segundo paso utiliza un método de conglomerado jerárquico para fundir progresivamente los subconglomerados en conglomerados cada vez más grandes, sin necesidad de realizar otra pasada por los datos. El conglomerado jerárquico tiene la ventaja de que no es necesario seleccionar el número de conglomerados por adelantado. Muchos métodos de conglomerado jerárquico comienzan con registros individuales como conglomerados iniciales y los van fundiendo sucesivamente para generar conglomerados más grandes. Aunque estos métodos suelen no funcionar bien con grandes cantidades de datos, el preconglomerado inicial de Bietápico permite que el conglomerado jerárquico sea rápido incluso con grandes conjuntos de datos.

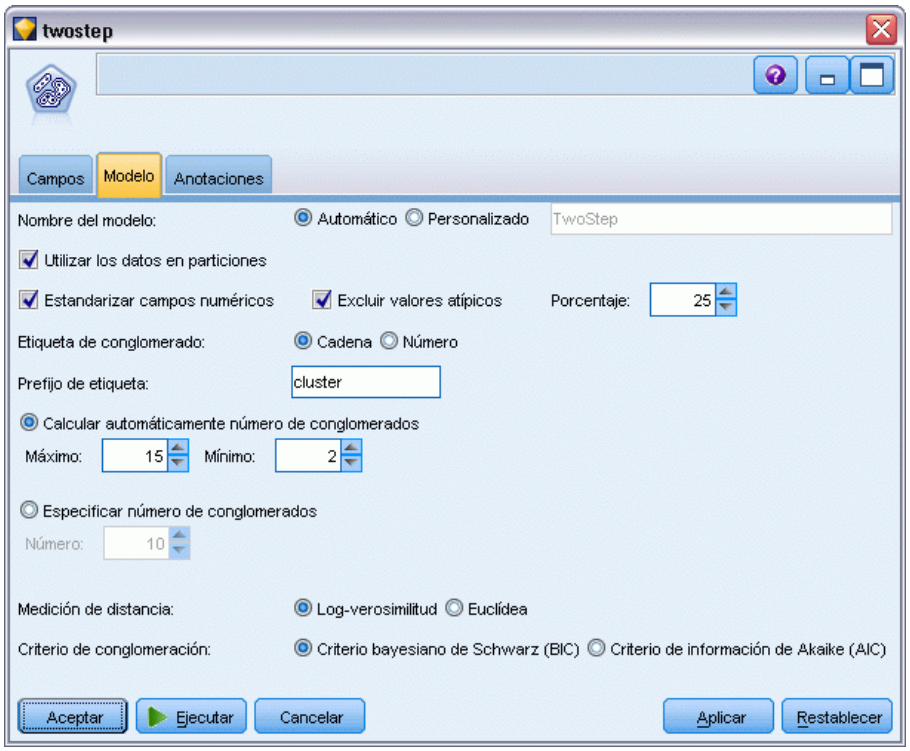
Nota: el modelo resultante depende, hasta cierto punto, del orden de los datos de entrenamiento. Reordenar los datos y regenerar el modelo puede dar como resultado un modelo de conglomerados final distinto.

Requisitos. Para entrenar un modelo de conglomerado Bietápico, necesita uno o más campos con su papel establecido como *Entrada*. Se ignorarán los campos con el papel establecido como *Objetivo*, *Ambos* o *Ninguno*. El algoritmo de conglomerados 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 conglomerado Bietápico puede gestionar distintos tipos de campos mezclados y conjuntos de datos grandes con eficacia. También tiene capacidad para comprobar varias soluciones de conglomerados y seleccionar la mejor, por lo que no tendrá que saber el número de conglomerados que hay que pedir al comienzo. El conglomerado Bietápico se puede configurar para que excluya automáticamente **valores atípicos** o casos muy extraños que puedan contaminar sus resultados.

Opciones de modelo para el nodo de conglomerado Bietápico

Figura 11-10
Opciones de modelo para el nodo de conglomerado Bietápico



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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Estandarizar campos numéricos. Por defecto, 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 conglomerado 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 preconglomerado. Cuando se selecciona esta opción, los subconglomerados con pocos registros relativos a otros subconglomerados se consideran valores atípicos potenciales y se vuelve a crear el árbol de subconglomerados excluyendo esos registros. El tamaño por debajo del cual se considera que los subconglomerados contienen posibles valores atípicos está controlado por la opción Porcentaje. Algunos de esos registros de valores atípicos potenciales pueden añadirse a los subconglomerados creados de nuevo, si son lo suficientemente similares a alguno de los nuevos perfiles de subconglomerados. Los demás valores atípicos potenciales que no se puedan fundir se considerarán valores atípicos, se añadirán a un conglomerado "ruido" y se excluirán del paso de conglomerado jerárquico.

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 conglomerado significativo más cercano, se consideran valores atípicos y se asignan al conglomerado "ruido" con el nombre -1.

Etiqueta de conglomerado. Especifique el formato del campo de pertenencia al conglomerado generado. La pertenencia a un conglomerado se puede indicar con una Cadena con el Prefijo de etiqueta especificado (por ejemplo "Cluster 1", "Cluster 2", etc.) o con un Número.

Calcular automáticamente número de conglomerados. El conglomerado Bietápico puede analizar rápidamente un gran número de soluciones de conglomerado para seleccionar el número óptimo de conglomerados para los datos de entrenamiento. Especifique un rango de soluciones que comprobar estableciendo el número Máximo y Mínimo de conglomerados. Bietápico utiliza un proceso de dos etapas para determinar el número óptimo de conglomerados. En la primera etapa se selecciona un límite superior para el número de conglomerados del modelo en función del cambio del Criterio de información bayesiano (BIC) a medida que se van añadiendo conglomerados. En la segunda etapa se encuentra el cambio de la distancia mínima entre conglomerados para todos los modelos con menor número de conglomerados que el mínimo de la solución de BIC. El mayor cambio de distancia se utiliza para identificar el modelo de conglomerados final.

Especificar número de conglomerados. Si conoce el número de conglomerados 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 conglomerados.

- **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 conglomerados. Sólo se puede utilizar cuando todas las variables son continuas.

Criterio de conglomeración. Esta opción determina cómo el algoritmo de conglomeración determina el número de conglomerados. Se puede especificar tanto el criterio de información bayesiano (BIC) como el criterio de información de Akaike (AIC).

Nugget de modelo de conglomerado Bietápico

Los nugget de modelo de conglomerado Bietápico contienen toda la información capturada por el modelo de conglomerados, 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 conglomerado 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*.

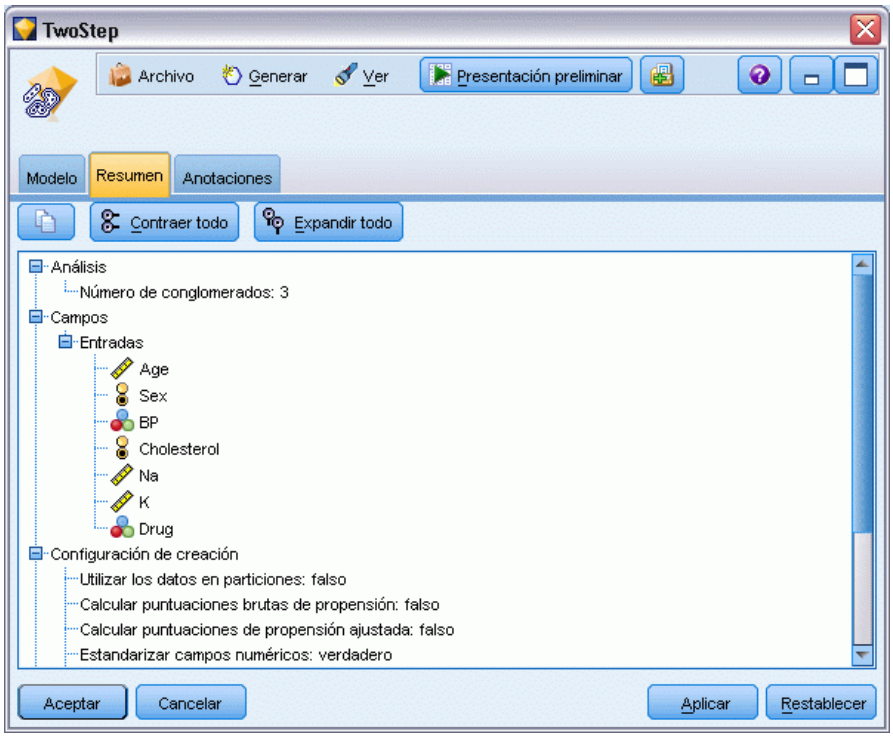
Una buena forma de comprender el modelo Bietápico es utilizar la inducción de reglas para descubrir las características que distinguen los conglomerados que el modelo ha encontrado. Si desea obtener más información, consulte el tema Nodo C5.0 en el capítulo 6 el p. 174. Asimismo, puede pulsar en la pestaña Modelo del explorador del nugget de modelo para ver el visor de conglomerados, que ofrece una representación gráfica de conglomerados, campos y niveles de importancia. Si desea obtener más información, consulte el tema Visor de conglomerados - Pestaña Modelo el p. 388.

Si desea obtener información general sobre el explorador de modelos, consulte Exploración de nugget de modelo

Resumen de modelo bietápico

La pestaña Resumen de un nugget de modelo de conglomerado Bietápico muestra el número de conglomerados encontrados, junto con la información acerca de los datos de entrenamiento, el proceso de estimación y la configuración de creación utilizada.

Figura 11-11
Pestaña Resumen de nugget de modelo de muestra de conglomerado Bietápico



Si desea obtener más información, consulte el tema Exploración de nugget de modelo en el capítulo 3 el p. 53.

Visor de conglomerados

Los modelos de conglomerados suelen utilizarse para encontrar grupos (o conglomerados) de registros similares en función de las variables examinadas, en los que la similitud entre los miembros del mismo grupo es alta, mientras que la similitud entre miembros de diferentes grupos es baja. Los resultados pueden utilizarse para identificar asociaciones que de otra forma no serían evidentes. Por ejemplo, es posible utilizar un análisis de conglomerados 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 conglomerados:

- Examinar los conglomerados para determinar las características únicas de cada uno. *¿Hay algún conglomerado que contenga todos los socios con un alto nivel de ingresos? ¿Contiene este conglomerado más registros que los demás?*
- Examinar los campos de los conglomerados para determinar la forma en que se distribuyen los valores entre los conglomerados. *¿Afecta el nivel de educación a la inclusión en un conglomerado? ¿Sirve una puntuación alta de crédito para distinguir entre la pertenencia a un conglomerado o a otro?*

Puede utilizar las vistas principales y las diferentes vistas vinculadas en el visor de conglomerados 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 conglomerados:

- Nugget de modelo de red Kohonen
- Nugget de modelos de K-medias
- Nugget de modelo de conglomerado en dos fases

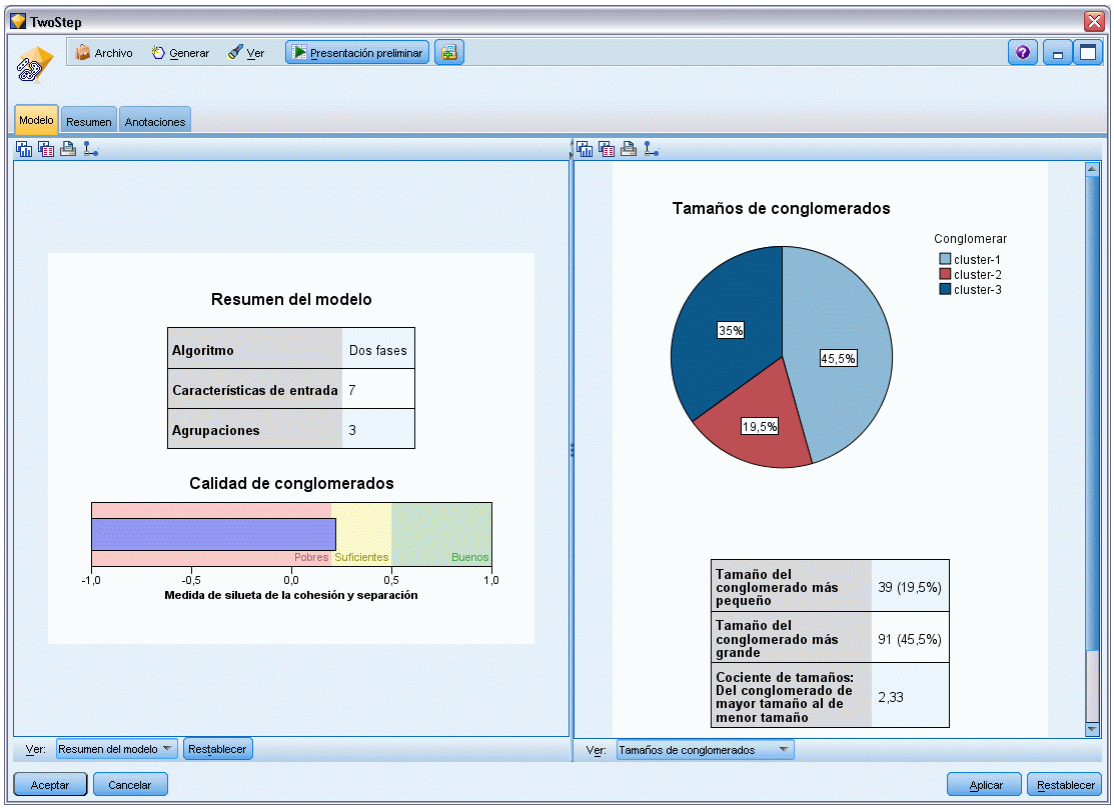
Para ver información sobre los nuggets de modelos de conglomerados, pulse con el botón derecho en el nodo del modelo y seleccione Examinar en el menú contextual (o Modificar en los nodos de una transmisión). Además, si está utilizando el nodo de modelado de conglomerado automático, pulse dos veces en el nugget de conglomerados que proceda, dentro del nugget de modelo Conglomerado automático. Si desea obtener más información, consulte el tema Nodo Autoconglomeración en el capítulo 5 el p. 113.

Visor de conglomerados - Pestaña Modelo

La pestaña Modelos de los modelos de conglomerados muestra una visualización gráfica de estadísticas y distribuciones de resúmenes para campos entre conglomerados, que se conoce como el **Visor de conglomerados**.

Nota: La pestaña Modelo no está disponible para modelos creados en versiones de IBM® SPSS® Modeler anteriores a la 13.

Figura 11-12
Visor de conglomerados con visualización predeterminada



El Visor de conglomerados 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:

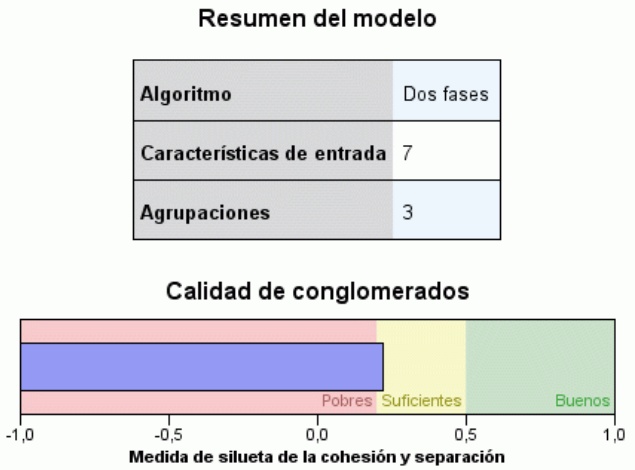
- Resumen del modelo (predeterminado). Si desea obtener más información, consulte el tema Vista Resumen del modelo el p. 390.
- Conglomerados. Si desea obtener más información, consulte el tema Vista de conglomerados el p. 391.

Hay cuatro vistas relacionadas/auxiliares:

- Importancia del predictor. Si desea obtener más información, consulte el tema Vista Importancia del predictor de conglomerados el p. 394.
- Tamaños de conglomerados (predeterminado) Si desea obtener más información, consulte el tema Vista de tamaños de conglomerados el p. 395.
- Distribución de casillas. Si desea obtener más información, consulte el tema Vista Distribución de casillas el p. 396.
- Comparación de conglomerados. Si desea obtener más información, consulte el tema Vista Comparación de conglomerados el p. 397.

Vista Resumen del modelo

Figura 11-13
Vista Resumen del modelo en el panel principal



La vista Resumen del modelo muestra una instantánea o resumen del modelo de conglomerado, incluyendo una medida de silueta de la cohesión y separación de conglomerados 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 conglomerado 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 conglomerados. 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 conglomerados, 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.

Las medias de medida de silueta, en todos los registros, (B−A) / max(A,B), donde A es la distancia del registro al centro de su conglomerado y B es la distancia del registro al centro del conglomerado más cercano al que no pertenece. Un coeficiente de silueta de 1 podría implicar que todos los casos están ubicados directamente en los centros de sus conglomerados. Un valor de −1 significaría que todos los casos se encuentran en los centros de conglomerado de otro conglomerado. Un valor de 0 implica, de media, que los casos están equidistantes entre el centro de su propio conglomerado y el siguiente conglomerado más cercano.

El resumen incluye una tabla que contiene la siguiente información:

- **Algoritmo.** El algoritmo de conglomeración utilizado, por ejemplo, "Dos fases".
- **Características de entrada.** El número de campos, también conocidos como **entradas** o **predictores**.
- **Conglomerados.** Número de conglomerados de la solución.

Vista de conglomerados

Figura 11-14
Vista Centros de conglomerados del panel principal

La vista Conglomerados contiene una cuadrícula de conglomerados por funciones que incluye nombres de conglomerados, tamaños y perfiles para cada conglomerado.

Las columnas de la cuadrícula contienen la siguiente información:

- **Conglomerado.** Números de conglomerados creados por el algoritmo.
- **Etiqueta.** Etiquetas aplicadas a cada conglomerado (está en blanco de forma predeterminada). Pulse dos veces la casilla para introducir una etiqueta que describa el contenido del conglomerado, por ejemplo “Compradores de automóviles de lujo”.
- **Descripción.** Cualquier descripción de los contenidos de los conglomerados (está en blanco de forma predeterminada). Pulse dos veces la casilla para introducir una descripción del conglomerado, por ejemplo “Más de 55 años de edad, profesionales, con ingresos superiores a 100.000 €”.
- **Tamaño.** El tamaño de cada conglomerado como porcentaje de la muestra general del conglomerado. Cada casilla de tamaño de la cuadrícula muestra una barra vertical que muestra el porcentaje de tamaño del conglomerado, un porcentaje de tamaño en formato numérico y los recuentos de casos de conglomerado.
- **Funciones** 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 conglomerado.

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 conglomerados, 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 Conglomerados, puede seleccionar varias formas de mostrar la información de conglomerados:

- Transponer conglomerados y características Si desea obtener más información, consulte el tema Transponer conglomerados y características el p. 392.
- Clasificar características Si desea obtener más información, consulte el tema Clasificar características el p. 393.
- Clasificar conglomerados Si desea obtener más información, consulte el tema Clasificar conglomerados el p. 393.
- Seleccionar contenido de casilla Si desea obtener más información, consulte el tema Contenido de casilla el p. 393.

Transponer conglomerados y características

Por defecto, los conglomerados se muestran como columnas y las características aparecen como filas. Para invertir esta visualización, pulse el botón Transponer conglomerados y características a la izquierda de los botones Clasificar características. Por ejemplo, puede que desea hacer esto cuando se muestren muchos conglomerados para reducir la cantidad de desplazamiento horizontal necesario para visualizar los datos.

Figura 11-15
Conglomerados transpuestos en el panel principal

Conglomerado	Etiqueta	Descripción	Tamaño	
cluster-1			45,0% (91)	BP HIGH (41,8%)
cluster-3			35,0% (70)	BP NORMAL (51,4%)
cluster-2			19,0% (39)	BP HIGH (100,0%)

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 conglomerados. Si hay características que empatan en valores de importancia, éstas se muestran en orden de clasificación ascendente según el nombre.
- **Importancia dentro del conglomerado** Las características se clasifican con respecto de su importancia para cada conglomerado. 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 conglomerados.
- **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 conglomerados

Por defecto, los conglomerados se clasifican en orden de tamaño descendente. Los botones Clasificar conglomerados por le permiten ordenarlos por nombre en orden alfabético o, si ha creado etiquetas de únicas, por orden de etiqueta alfanumérico.

Las características con la misma etiqueta se clasifican por nombre de conglomerado. Si los conglomerados se clasifican por etiqueta y modifica la etiqueta de un conglomerado, 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 conglomerados.** Por defecto, las casillas muestran nombres/etiquetas de características y la tendencia central para cada combinación de conglomerado/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 conglomerado. En el caso de las funciones 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 conglomerado.

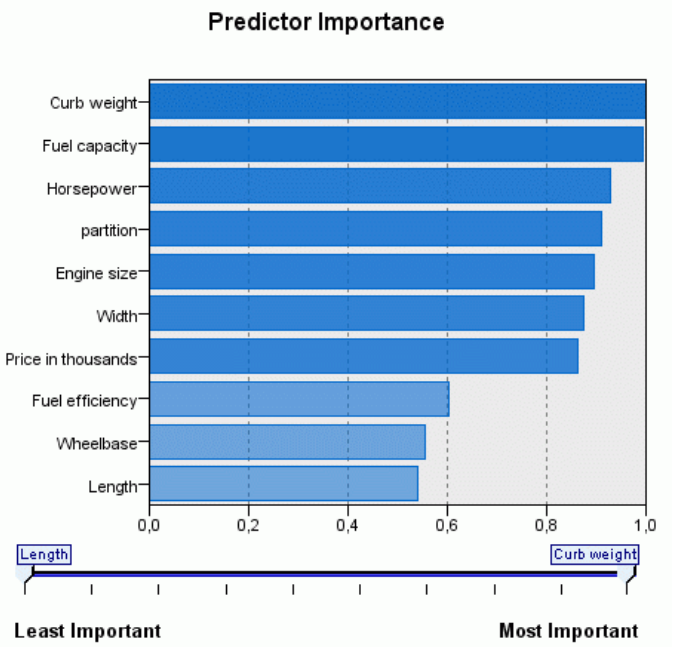
  La visualización en color rojo oscuro muestra la distribución de conglomerados, mientras que la más clara representa los datos generales.
- **Distribuciones relativas** Muestra los nombres/etiquetas de características y las distribuciones relativas en las casillas. En general, las visualizaciones son 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 conglomerados, mientras que la más clara representa los datos generales.

- **Vista básica.** Si hay muchos conglomerados, 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 conglomerados

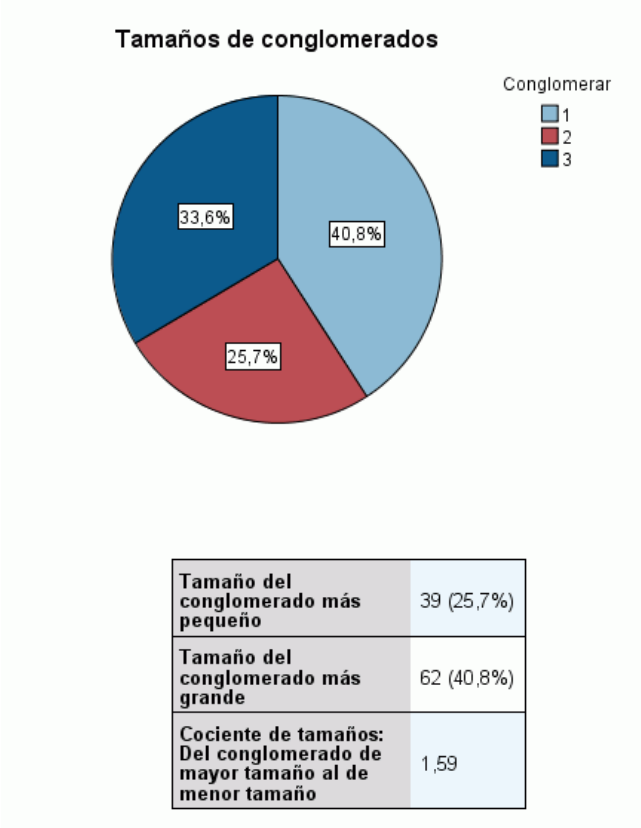
Figura 11-16
Vista Importancia del predictor de conglomerados del panel de vínculos



La vista Importancia del predictor muestra la importancia relativa de cada campo en la estimación del modelo. Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

Vista de tamaños de conglomerados

Figura 11-17
Vista Tamaños de conglomerados del panel de vínculos



Tamaño del conglomerado más pequeño	39 (25,7%)
Tamaño del conglomerado más grande	62 (40,8%)
Cociente de tamaños: Del conglomerado de mayor tamaño al de menor tamaño	1,59

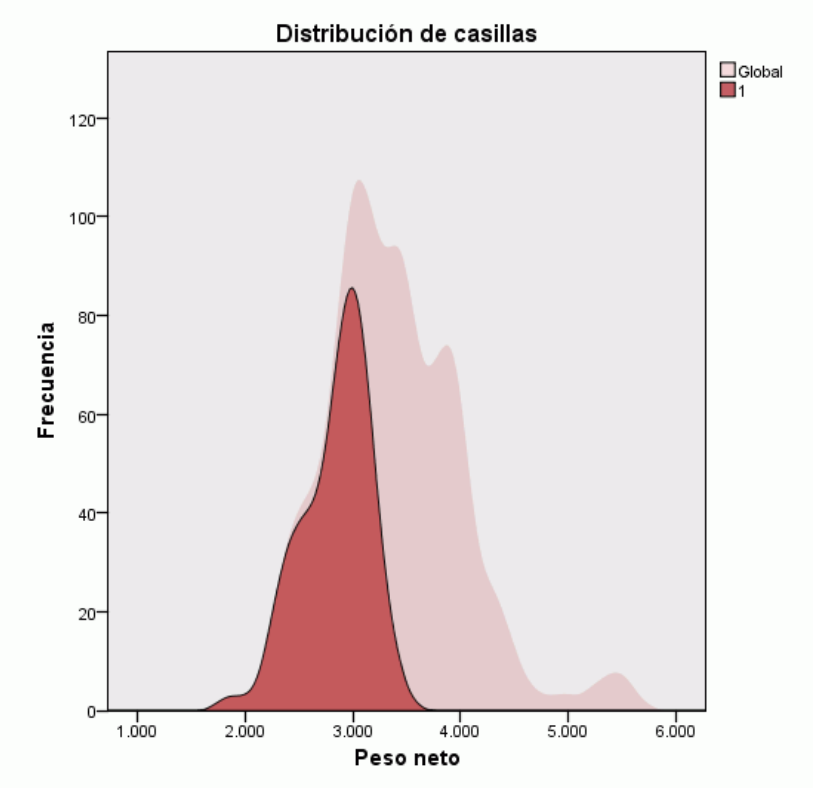
La vista Tamaños de conglomerados muestra el gráfico de sectores que contiene cada conglomerado. El tamaño de porcentaje de cada conglomerado se muestra en cada sector, pase el ratón sobre cada sector para mostrar el recuento de ese sector.

Bajo el gráfico, una tabla enumera la siguiente información de tamaño:

- El tamaño del conglomerado más pequeño (un recuento y porcentaje del conjunto).
- El tamaño del conglomerado mayor (un recuento y porcentaje del conjunto).
- La proporción entre el tamaño del mayor conglomerado y el del menor.

Vista Distribución de casillas

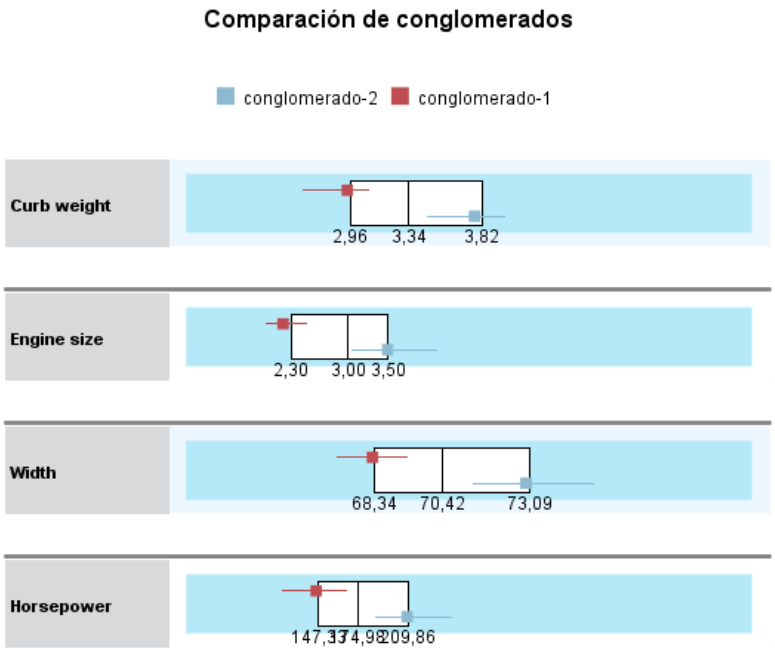
Figura 11-18
Vista Distribución de casillas del panel de vínculos



La vista Distribución de casillas muestra un gráfico expandido y más detallado de la distribución de los datos para cualquier casilla que seleccione en el panel principal Conglomerados.

Vista Comparación de conglomerados

Figura 11-19
Vista Comparación de conglomerados del panel de vínculos



La vista Comparación de conglomerados se compone de un diseño en estilo de cuadrícula, con características en las filas y conglomerados seleccionados en las columnas. Esta vista le ayuda a entender mejor los factores de los que se componen los conglomerados, y le permite ver las diferencias entre los conglomerados no sólo con respecto a los datos generales, sino entre sí.

Para seleccionar conglomerados para su visualización, pulse en la parte superior de la columna del conglomerado en el panel principal Conglomerados. Pulse las teclas Ctrl o Mayús y pulse para seleccionar o cancelar la selección de más de un conglomerado para su comparación.

Nota: Puede seleccionar hasta cinco conglomerados para su visualización.

Los conglomerados 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 conglomerado, 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 conglomerado (por característica).
- Las características continuas se muestran como diagramas de caja, que muestran las medianas globales y las amplitudes intercuartiles.

En estas vistas de fondo aparecen superpuestos diagramas de caja para los conglomerados seleccionados:

- En funciones continuas, los marcadores cuadrados de puntos y las líneas horizontales indican la amplitud de la media e intercuantil de cada conglomerado.
- Cada conglomerado viene representado por un color distinto, que se muestra en la parte superior de la vista.

Navegación en el Visor de conglomerados

El Visor de conglomerados es una visualización interactiva. Tiene la posibilidad de:

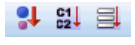
- Seleccionar un campo o conglomerado para ver más detalles.
- Comparar conglomerados para seleccionar elementos de interés.
- Alterar la visualización.
- Transponer ejes.
- Generar derivadas, filtrar y seleccionar nodos mediante el menú Generar.

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 Conglomerados en el panel principal.

Figura 11-20
Barras de herramientas de control de los datos que se muestran en el Visor de conglomerados

Las opciones Clasificar características por, Clasificar conglomerados por, Casillas y Mostrar sólo están disponibles cuando selecciona la vista Conglomerados en el panel principal. Si desea obtener más información, consulte el tema Vista de conglomerados el p. 391.

	Consulte Transponer conglomerados y características el p. 392
	Consulte Clasificar características por el p. 393
	Consulte Clasificar conglomerados por el p. 393
	Consulte Casillas el p. 393

Generación de nodos desde los modelos de conglomerado

El menú Generar le permite crear nuevos nodos basados en el modelo de conglomerados. 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 conglomerados visibles o todos

los seleccionados). Por ejemplo, puede seleccionar una única característica y generar un nodo Filtro 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.

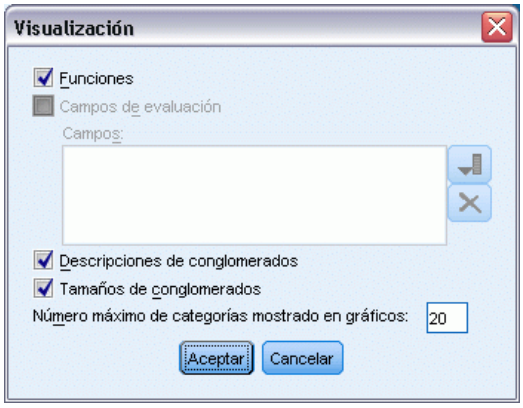
- **Generar nodo de modelado** Crea un nodo de modelado en el lienzo de la transmisión. Esto puede resultar útil, por ejemplo, si tiene una transmisión en la que desea utilizar estos ajustes de modelo, pero ya no tiene el nodo de modelado utilizado para generarlos.
- **Modelo a paleta** Crea un nugget en la paleta Modelos. Esto resulta útil en situaciones en las que un compañero puede haber enviado una transmisión que contenga el modelo, pero no el modelo propiamente dicho.
- **Nodo Filtro** Crea un nuevo nodo Filtro para filtrar los campos que no se utilicen en el modelo de conglomerados o que no sea visible en la visualización Visor de conglomerados actual. Si hay un nodo Tipo anterior a este nodo Conglomerado, cualquier campo con el papel *destino* queda descartado por el nodo Filtro generado.
- **Nodo Filtro (desde la selección)** Crea un nuevo nodo Filtro para filtrar los campos en función de las selecciones del Visor de conglomerados. Seleccione múltiples campos mediante presionando Ctrl y pulsando cada uno de ellos. Los campos seleccionados en el Visor de conglomerados se descartan posteriormente, aunque puede cambiar este comportamiento editando el nodo Filtro antes de su ejecución.
- **Nodo Seleccionar** Crea un nuevo nodo Seleccionar para seleccionar registros en función de su pertenencia a cualquiera de los conglomerados visibles en la visualización actual del Visor de conglomerados. Automáticamente se genera una condición seleccionar.
- **Nodo Seleccionar (desde la selección)** Crea un nuevo nodo Seleccionar para seleccionar registros en función de la pertenencia a conglomerados seleccionados en el Visor de conglomerados. Seleccione múltiples conglomerados presionando Ctrl y pulsando cada uno de ellos.
- **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 conglomerados visibles en el Visor de conglomerados. Automáticamente se genera una condición derivar.
- **Nodo Derivar (desde la selección)** Crea un nuevo nodo Derivar que deriva un campo de marca en función de la pertenencia a conglomerados seleccionados en el Visor de conglomerados. Seleccione múltiples conglomerados presionando Ctrl y pulsando cada uno de ellos.

Además de generar nodos, también puede crear gráficos desde el menú Generar. Si desea obtener más información, consulte el tema Generación de gráficos desde los modelos de conglomerado el p. 400.

Control de la visualización de conglomerados

Para controlar qué se muestra en la vista Conglomerados del panel principal, pulse el botón Mostrar y se abrirá el cuadro de diálogo Mostrar.

Figura 11-21
Visor de conglomerados - Opciones de Mostrar



Funciones Está seleccionado por defecto. 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 conglomerado, sino que se envían al visor de modelos para evaluar los conglomerados) que desea mostrar, ya que ninguno se muestra de forma predeterminada. *Nota*: Esta casilla de verificación no está disponible si no hay ningún campo de evaluación disponible.

Descripciones de conglomerados Está seleccionado por defecto. Para ocultar todas las casillas de descripción de conglomerado, cancele la selección de la casilla de verificación.

Tamaños de conglomerados Está seleccionado por defecto. Para ocultar todas las casillas de tamaño de conglomerado, 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 conglomerado

Los modelos de conglomerado ofrecen mucha información, aunque no siempre en un formato accesible para los usuarios de las empresas. 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 conglomerados puede generar un gráfico para un conglomerado seleccionado, creando así un gráfico únicamente para los casos de ese conglomerado.

Nota: Sólo puede generar un gráfico desde el Visor de conglomerados si el nugget de modelo está vinculado a otros nodos de una transmisión.

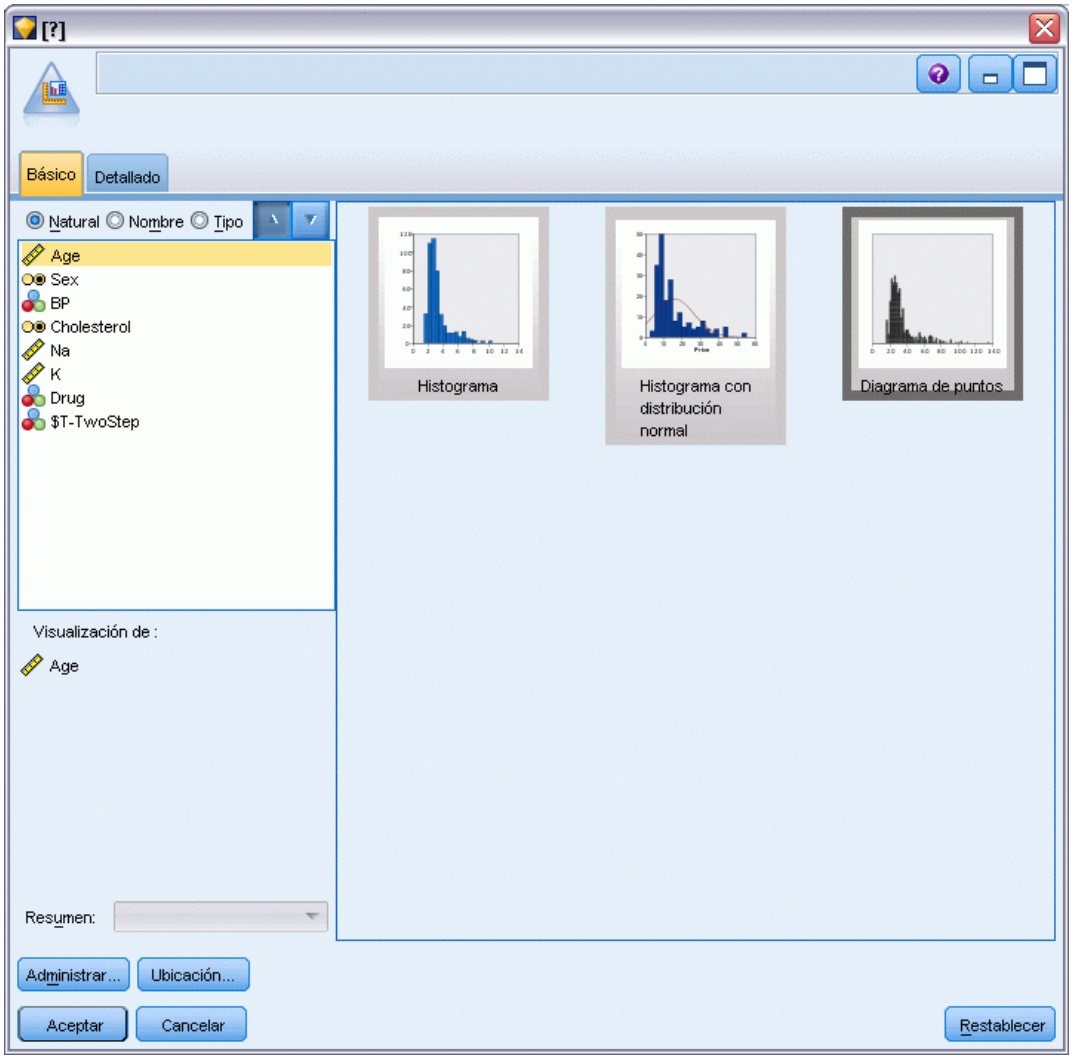
Generación de un gráfico

- Abra el nugget de modelo que contiene el Visor de conglomerados.
- En la pestaña Modelo, seleccione *Conglomerados* en la lista desplegable Vista.

- En la vista principal, seleccione el conglomerado o conglomerados de los que desea crear un gráfico.

- En el menú Generar, seleccione Gráfico (desde selección) y se mostrará la pestaña Tablero básico.

Figura 11-22
Cuadro de diálogo del nodo Tablero, pestaña Básico

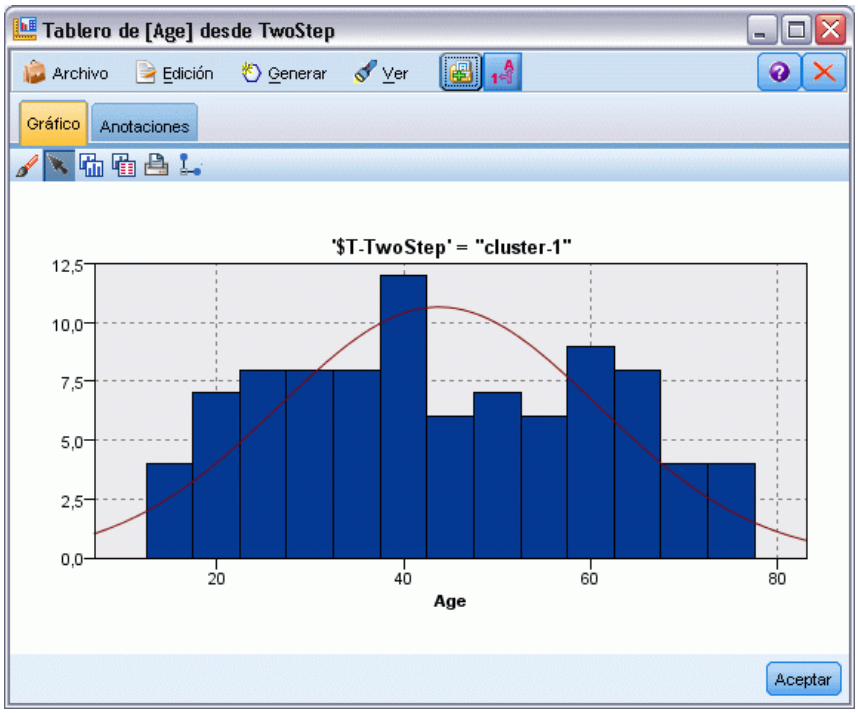


Nota: Cuando abre la pestaña Tablero de esta forma, las únicas pestañas disponibles son Básico y Detallado. Si desea obtener más información, consulte el tema Nodo Tablero en el capítulo 5 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

- Si utiliza la configuración de la pestaña Básico o Detallado, especifique los detalles que se mostrarán en el gráfico.

- Pulse en Aceptar para generar el gráfico.

Figura 11-23
Histograma generado a partir de la pestaña Tablero básico



El encabezado del gráfico identifica el tipo de modelo y el conglomerado o conglomerados que se seleccionaron para su inclusión.

Capítulo
12

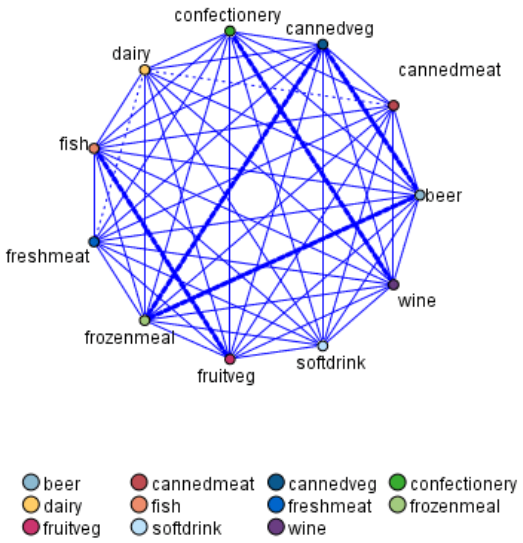
Reglas de asociación

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 12-1
Nodo Malla mostrando asociaciones entre elementos de la cesta del supermercado



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 pronósticos 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 A priori extrae un conjunto de reglas de los datos y destaca aquellas reglas con un mayor contenido de información. A priori 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, A priori 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. A priori 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. Si desea obtener más información, consulte el tema Nodo A priori el p. 405.

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, es probable que un cliente que compra una cuchilla y una loción para después del afeitado compre 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. Si desea obtener más información, consulte el tema Nodo Secuencia el p. 431.

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 A priori, 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, y Secuencia pueden utilizar datos tabulares.

Nodo A priori

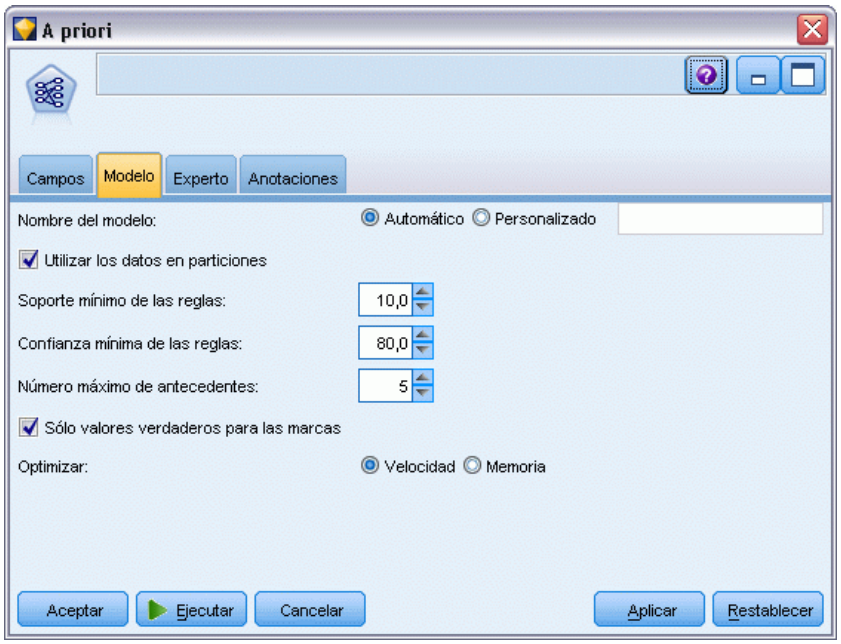
El nodo A priori también encuentra reglas de asociación en los datos. A priori ofrece cinco métodos diferentes para la selección de reglas y utiliza un sofisticado esquema de indización para procesar conjuntos de datos de gran tamaño de forma eficaz.

Requisitos. Para crear un conjunto de reglas de A priori, se precisan uno o varios campos de *Entrada* y uno o varios campos de *Objetivo*. Los campos de entrada y de salida (con dirección *Entrada*, *Objetivo* o *Ambos*) deben ser simbólicos. Los campos con el papel *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. Si desea obtener más información, consulte el tema Datos tabulares frente a datos transaccionales el p. 404.

Puntos fuertes. En los problemas de grandes dimensiones, A priori 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. A priori 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 A priori

Figura 12-2
Opciones de modelo para el nodo A priori



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.

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. Si desea obtener más información, consulte el tema Opciones de modelo para el nodo Secuencia el p. 434.) 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 A priori 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 pronósticos basados en la regla que son correctos. 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.

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. Si desea obtener más información, consulte el tema Datos tabulares frente a datos transaccionales el p. 404.

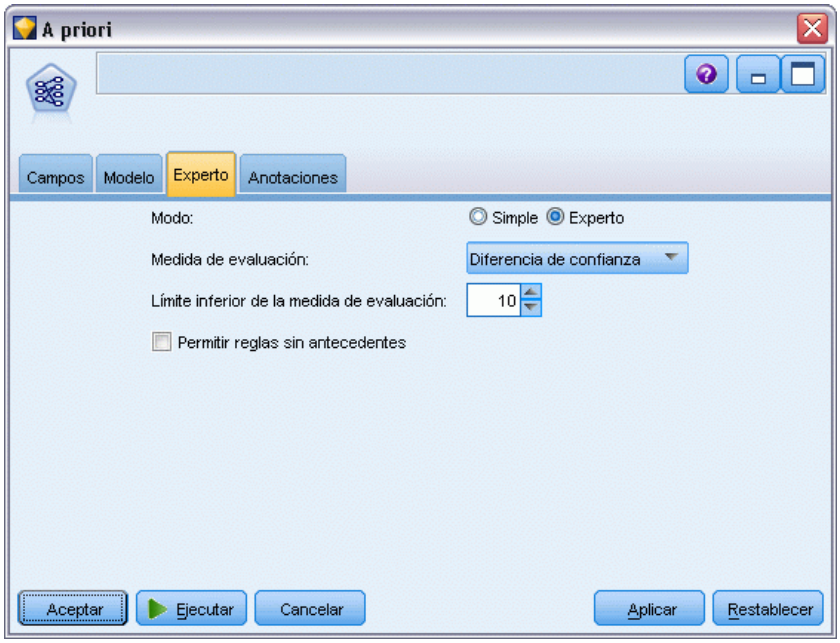
Optimizar. Seleccione opciones diseñadas para aumentar el rendimiento durante la generación de modelos según sus necesidades específicas.

- 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 por defecto. *Nota*: cuando se ejecuta en modo distribuido, esta configuración puede quedar anulada por las opciones del administrador especificadas en *options.cfg*. Si desea obtener más información, consulte el *Manual del administrador de IBM® SPSS® Modeler Server*.

Opciones de experto para el nodo A priori

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.

Figura 12-3
Opciones de experto de A priori



Medida de evaluación. A priori admite cinco métodos de evaluación de reglas potenciales.

- **Confianza de la regla.** El método por defecto 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. Si desea obtener más información, consulte el tema Opciones de modelo para el nodo A priori el p. 406.
- **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 a priori. 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 pronostica 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 pronostican eventos raros. Por ejemplo, supongamos que hay una enfermedad que sólo se da en el 1% de los pacientes. Una regla que puede pronosticar 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: puede que sea necesario experimentar con los distintos límites inferiores para obtener un conjunto de reglas satisfactorio, ya que la escala de esta medida es algo menos intuitiva que otras escalas.

- **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 los pronósticos de mayor confianza. Esta opción esta desactivada por defecto. 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

si *antecedente(s)* **entonces** *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 A priori, 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 A priori 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.

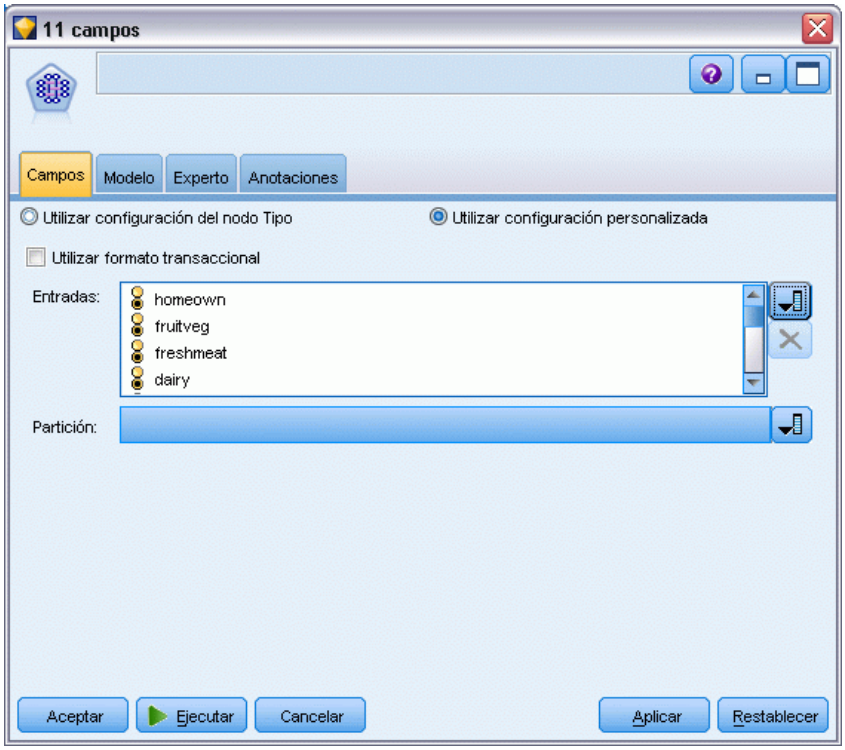
Para crear un conjunto de reglas de CARMA, es necesario especificar un campo de ID y uno o varios campos de contenido. El campo de ID puede tener cualquier papel o nivel de medición. Los campos con el papel *Ninguno* se omiten. Los tipos de campo deben estar completamente instanciados antes de ejecutar el nodo. Al igual que en A priori, los datos pueden estar en formato tabular o transaccional. Si desea obtener más información, consulte el tema Datos tabulares frente a datos transaccionales el p. 404.

Puntos fuertes. El nodo CARMA se basa en el algoritmo de reglas de asociación de CARMA. A diferencia de A priori, 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 A priori, los modelos que genera un nodo CARMA se pueden insertar en una ruta de datos para crear pronósticos. Si desea obtener más información, consulte el tema Nuggets de modelo en el capítulo 3 el p. 46.

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.

Figura 12-4
Opciones de campos para el nodo CARMA



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 por defecto.

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. Si desea obtener más información, consulte el tema Datos tabulares frente a datos transaccionales el p. 404.

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 papel 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 sobre la adecuación 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 ficha Campos en todos los nodos de modelado que usen la partición. (Si sólo hay una partición, se usará automáticamente siempre que se active la partición.) Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*. Debe tener en cuenta que al aplicar la partición seleccionada en su análisis, también debe activar la partición en la ficha 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

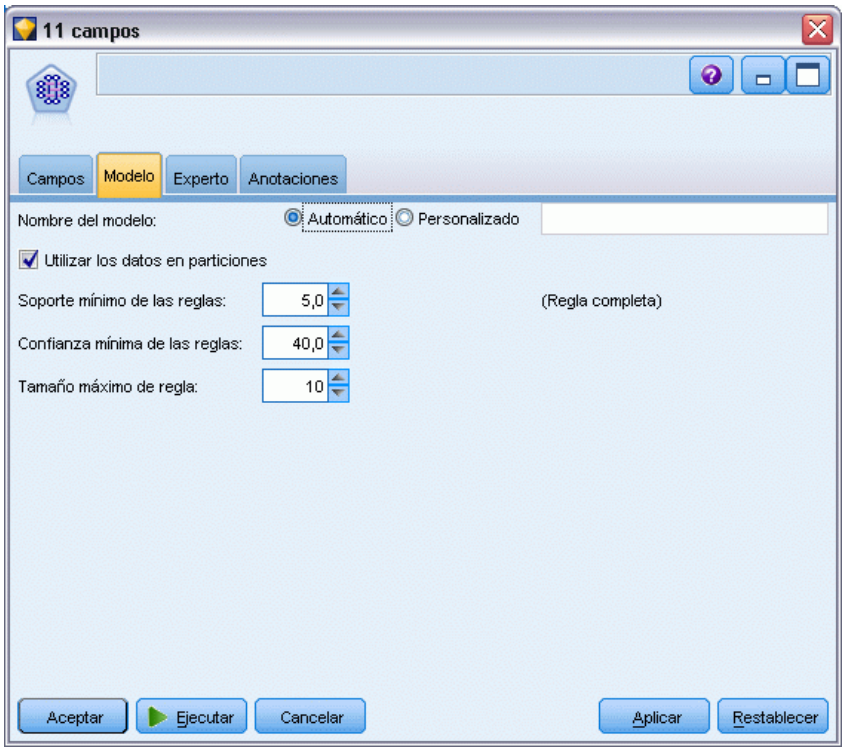
Si selecciona Utilizar formato transaccional, se muestran los siguientes campos.

- **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 único de este campo debe indicar una unidad de análisis específica. Por ejemplo, en una aplicación de la cesta del supermercado, 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

Figura 12-5
Opciones de modelo para el nodo CARMA



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.

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 A priori.) Si desea centrarse en las reglas más comunes, aumente el valor de este parámetro.

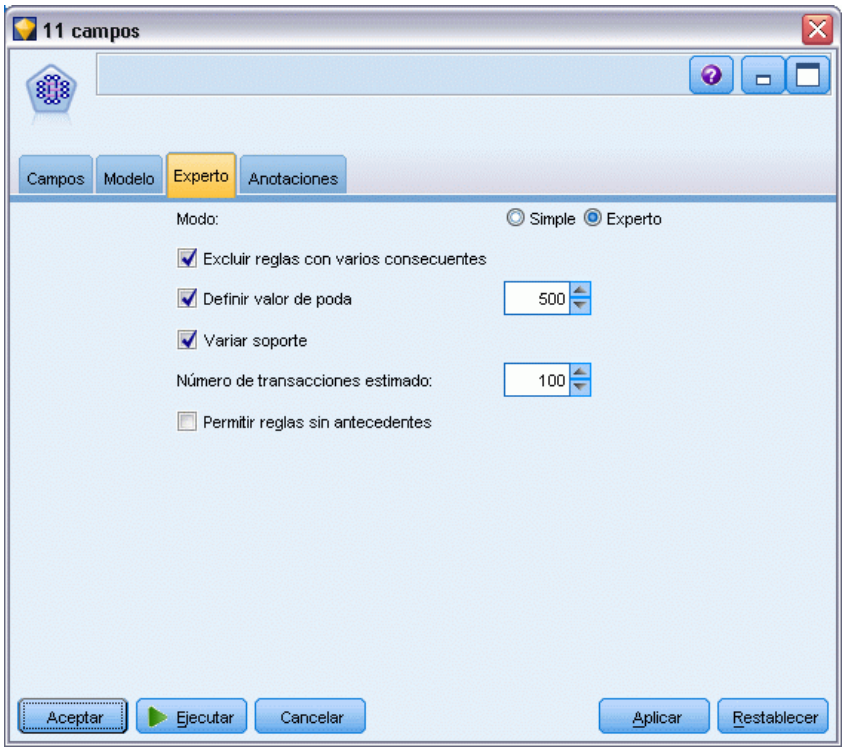
Confianza mínima de las reglas (%). Se puede especificar un criterio de confianza para mantener las reglas en el conjunto de reglas. **La confianza** hace referencia al porcentaje de campos de ID en los que se realiza un pronóstico correcto (de todos los campos de ID para los que la regla realiza un pronóstico). 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.

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.

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.

Figura 12-6
Opciones de experto para el nodo CARMA



Excluir reglas con varios consecuentes. Seleccione esta opción para excluir los "consecuentes "de dos destinos—, es decir, los que contienen dos elementos. Por ejemplo, la regla bread & cheese & fish -> wine&fruit contiene una regla consecuente de dos direcciones, wine&fruit. Estas reglas se incluyen por defecto.

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 defecto 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 los pronósticos de mayor confianza. Esta opción esta desactivada por defecto.

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.

Visualización de los resultados

Para examinar las reglas generadas por los modelos de asociación (A priori 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

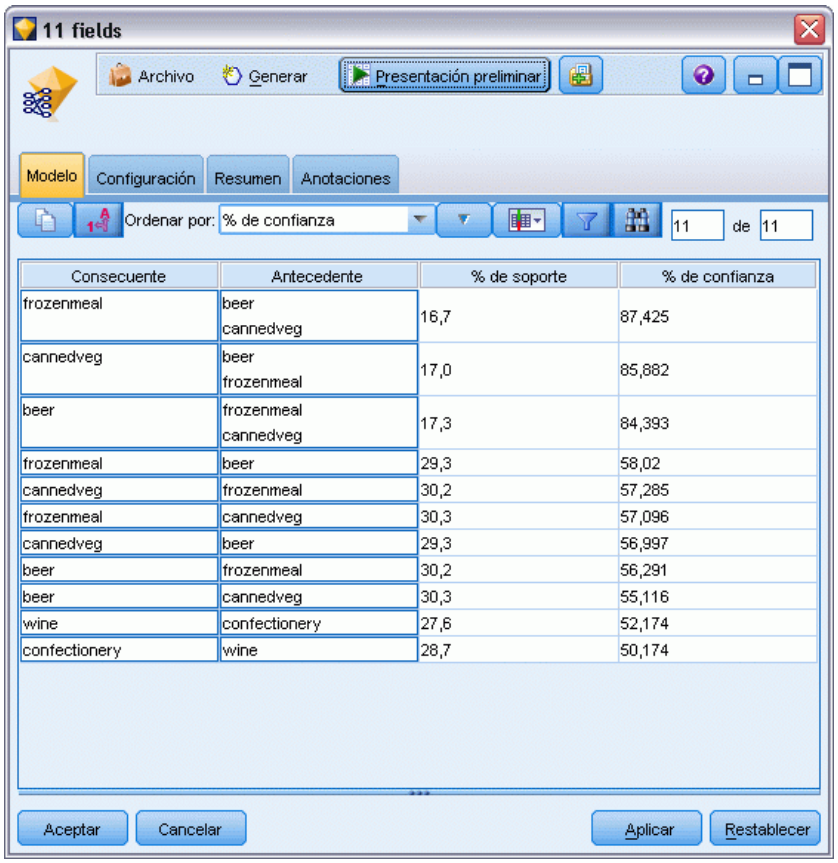
Los nugget de modelo refinado (A Priori, CARMA y Secuencia) se pueden añadir a una ruta y utilizarse para puntuar. Si desea obtener más información, consulte el tema Uso de nugget de modelo en rutas en el capítulo 3 el p. 68. Los nugget de modelo utilizados para puntuar incluyen una pestaña Configuración adicional en sus respectivos cuadros de diálogo. Si desea obtener más información, consulte el tema Configuración del nugget de modelo de reglas de asociación el p. 422.

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. Si desea obtener más información, consulte el tema Generación de un conjunto de reglas desde un nugget de modelo de asociación el p. 425.

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.

Figura 12-7
Pestaña Nugget de modelo de reglas de asociación



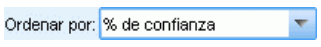
Las reglas de asociación a menudo se muestran en el siguiente formato:

Consecuente	**Antecedente**
Drug = drugY	Sex = F BP = HIGH

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).

Figura 12-8
Opciones de ordenación de la barra de herramientas



Los valores se pueden ordenar por:

- Support
- Confianza
- Soporte de regla
- Consecuente
- Lift
- Deployability

Menú Mostrar/ocultar. El menú Mostrar/ocultar (botón de criterios de la barra de herramientas) controla las opciones de visualización de las reglas.

Figura 12-9
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 un determinado pronóstico. Los ID de regla también permiten, más adelante, fundir información adicional de las reglas, como capacidad de distribución, información del producto o antecedentes.
- **Instancias** muestra información acerca del número de ID únicos a los que se aplica la regla; es decir, para los que se cumple la condición de los antecedentes. Por ejemplo, dada la regla bread -> cheese, se hace referencia al número de registros en los datos de entrenamiento que incluyan el antecedente *pan* como **ocurrencias**.
- **Soporte** muestra el soporte de antecedentes; es decir, la proporción de ID para los que se cumple que los antecedentes son verdad, basándose en los datos de entrenamiento. Por ejemplo, si el 50% de los datos de entrenamiento incluyen la compra de pan, entonces la regla bread -> cheese 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 contienen pan (indicando así el soporte del antecedente), pero sólo el 20% contiene pan y queso (lo que indica el soporte de regla), la confianza de la regla bread -> cheese sería Rule Support / Antecedent Support 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 bread -> cheese 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 pronostica 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 distribuibilidad** 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. La estadística de distribución se define como (Antecedent Support in # of Records - Rule Support in # of Records) / Number of Records) * 100, donde *Soporte de antecedentes* significa el número de registros por el que los antecedentes son verdaderos y *Soporte de regla* significa el número de registros por el que antecedentes y consecuente 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 12-10
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 12-11
Botón Buscar regla

Para ordenar de nuevo los ID de la regla:

▶ Es posible reorganizar los ID de regla en IBM® SPSS® Modeler ordenando, en primer lugar, la tabla de representación de reglas según la medida que desee, como la confianza o la elevación.

▶ A continuación, puede crear un modelo filtrado mediante las opciones del menú Generar.

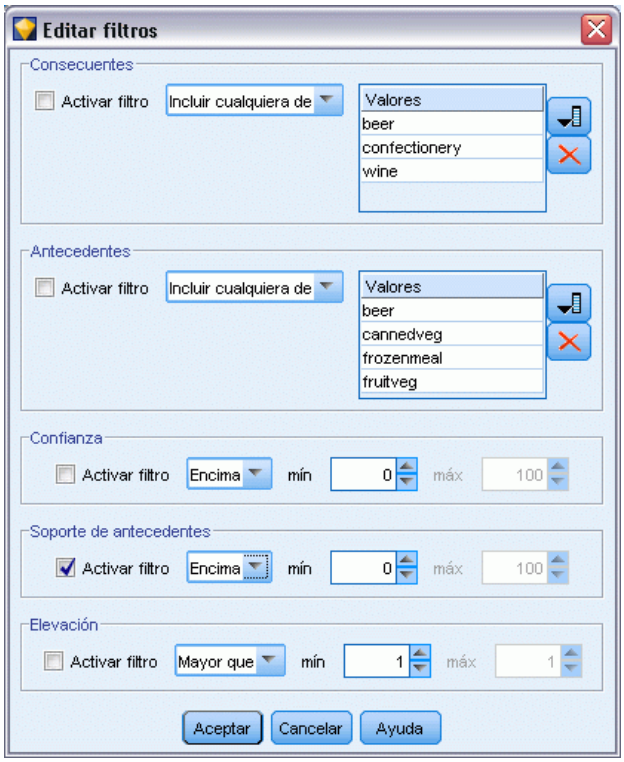
▶ En el cuadro de diálogo Modelo filtrado, seleccione Volver a numerar reglas consecutivamente, comenzando por y especifique un número de inicio.

Si desea obtener más información, consulte el tema Generación de un modelo filtrado el p. 426.

Especificación de filtros para reglas

Por defecto, los algoritmos de regla, como A priori, 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.

Figura 12-12
Cuadro de diálogo del filtro del explorador de reglas



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 medida 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 medida 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.

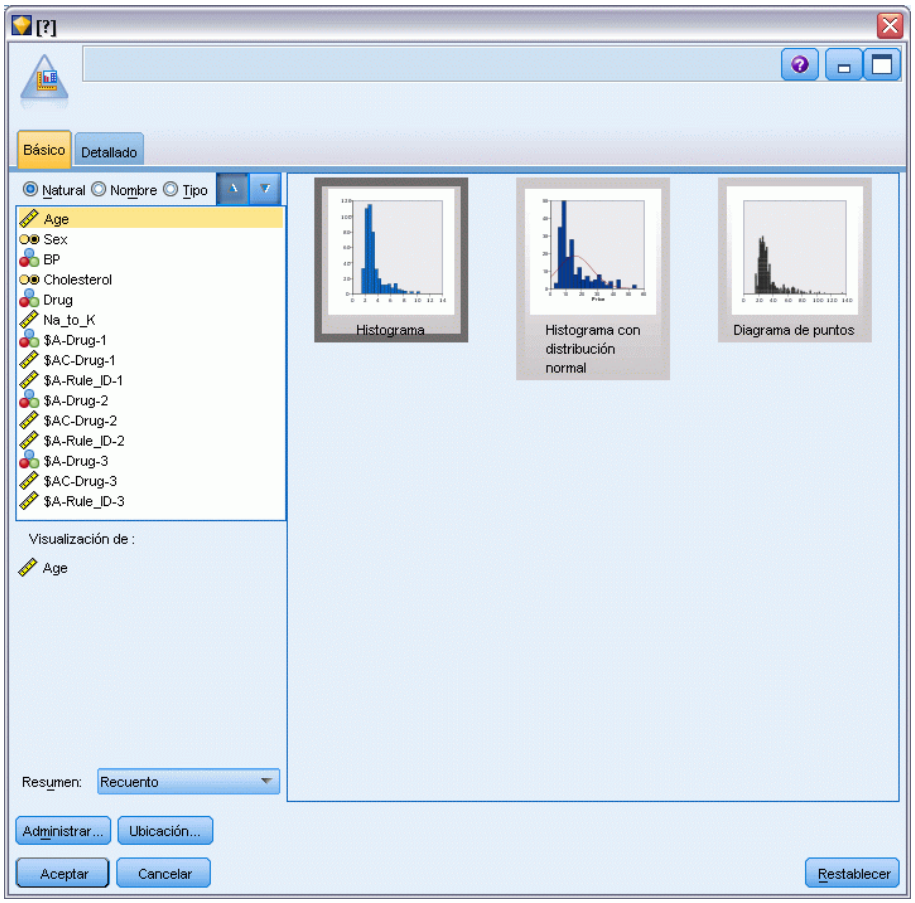
Generación de gráficos para reglas

Los nodos Asociación proporcionan gran cantidad de información, sin embargo, es posible que no estén en un formato fácilmente accesible para usuarios comerciales. 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.

▶ En la pestaña Modelo, selecciona la regla que le interese.

▶ En el menú Generar, seleccione Gráfico (desde selección). Aparecerá la pestaña Tablero básico.

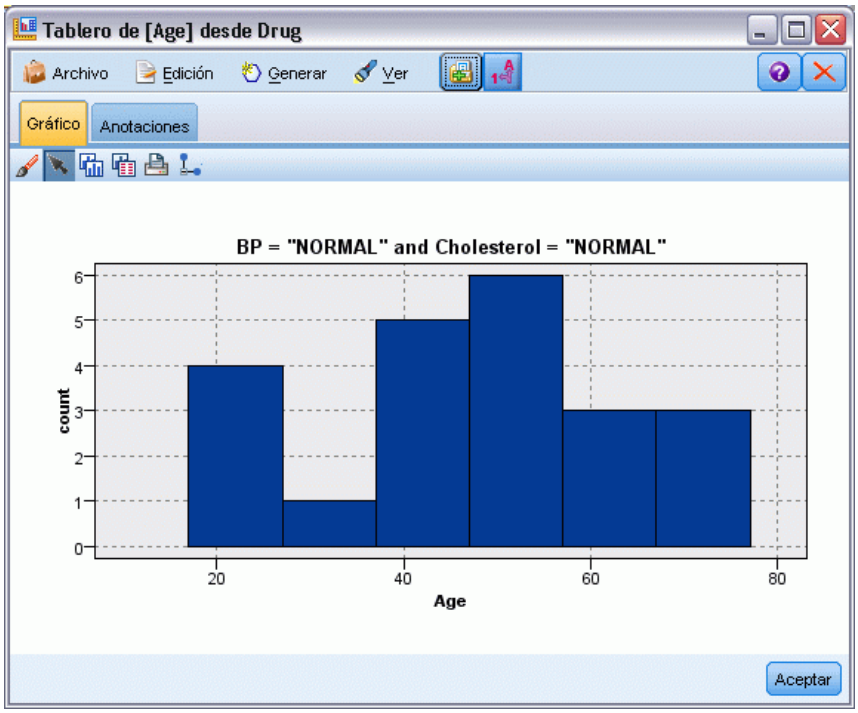
Figura 12-13
Cuadro de diálogo del nodo Tablero, pestaña Básico



Nota: Cuando abre la pestaña Tablero de esta forma, las únicas pestañas disponibles son Básico y Detallado. Si desea obtener más información, consulte el tema Nodo Tablero en el capítulo 5 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

- Si utiliza la configuración de la pestaña Básico o Detallado, especifique los detalles que se mostrarán en el gráfico.
- Pulse en Aceptar para generar el gráfico.

Figura 12-14
Cuadro de diálogo del nodo Tablero, pestaña Básico



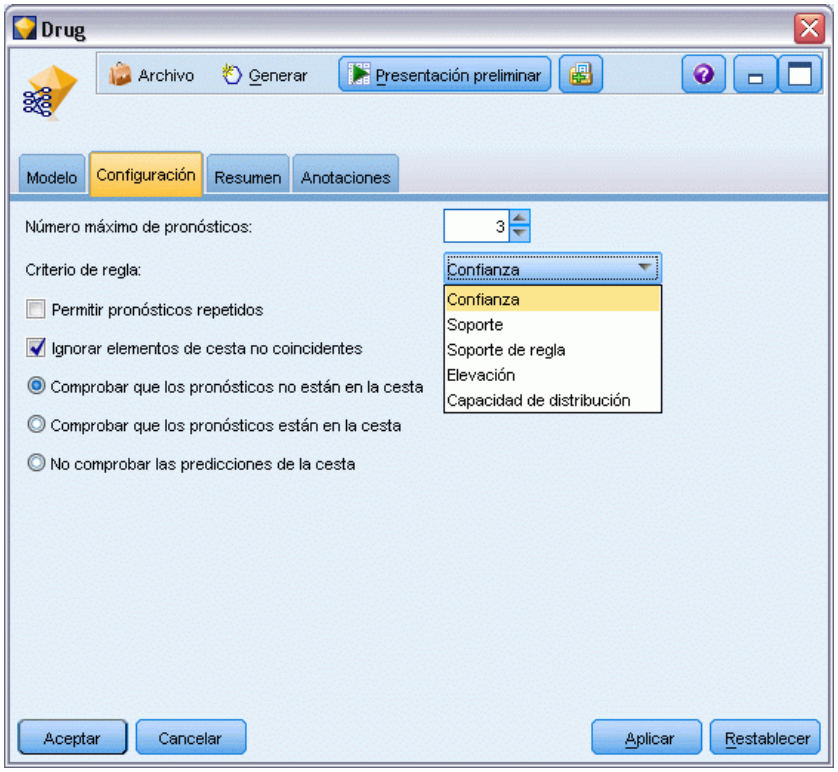
El encabezado del gráfico identifica la regla y los detalles de antecedentes seleccionados para incluir.

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 (A priori 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 sin refinar no incluye la pestaña Configuración porque no se puede puntuar. Para puntuar el , debe, en primer lugar, generar un conjunto de reglas. Si desea obtener más información, consulte el tema Generación de un conjunto de reglas desde un nugget de modelo de asociación el p. 425.

Figura 12-15
Pestaña Configuración del nugget de modelo de reglas de asociación



Número máximo de pronósticos. Especifique el número máximo de pronósticos incluidos para cada conjunto de elementos de la cesta. Se utiliza esta opción junto con el criterio de regla situado debajo para generar los "mejores" pronósticos, donde *mejores* indica el nivel de confianza más alto, el soporte, la 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 los mejores pronósticos para un conjunto de elementos. Algunos criterios disponibles son:

- Confianza
- Support
- Soporte de regla (Soporte * Confianza)
- Lift
- Deployability

Permitir pronósticos repetidos. Seleccione esta opción 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 queso > vino queso y fruta
 >vino
```

Desactive esta opción para excluir los pronósticos repetidos a la hora de puntuar.

Nota: Las reglas con varios consecuentes (bread & cheese & fruit -> wine & pate) se consideran pronósticos repetidos sólo si todos los consecuentes (wine & pate) se han pronosticado antes.

Ignorar elementos de cesta no coincidentes. Seleccione esta opción para omitir la presencia de elementos adicionales en el conjunto de elementos. Por ejemplo, si selecciona esta opción para una cesta que contiene [tent & sleeping bag & kettle], se aplicará la regla tent & sleeping bag -> gas_stove a pesar del elemento extra (kettle) presente 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 el mejor pronóstico. 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. Por defecto, se omiten los elementos no coincidentes.

Comprobar que los pronósticos 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.

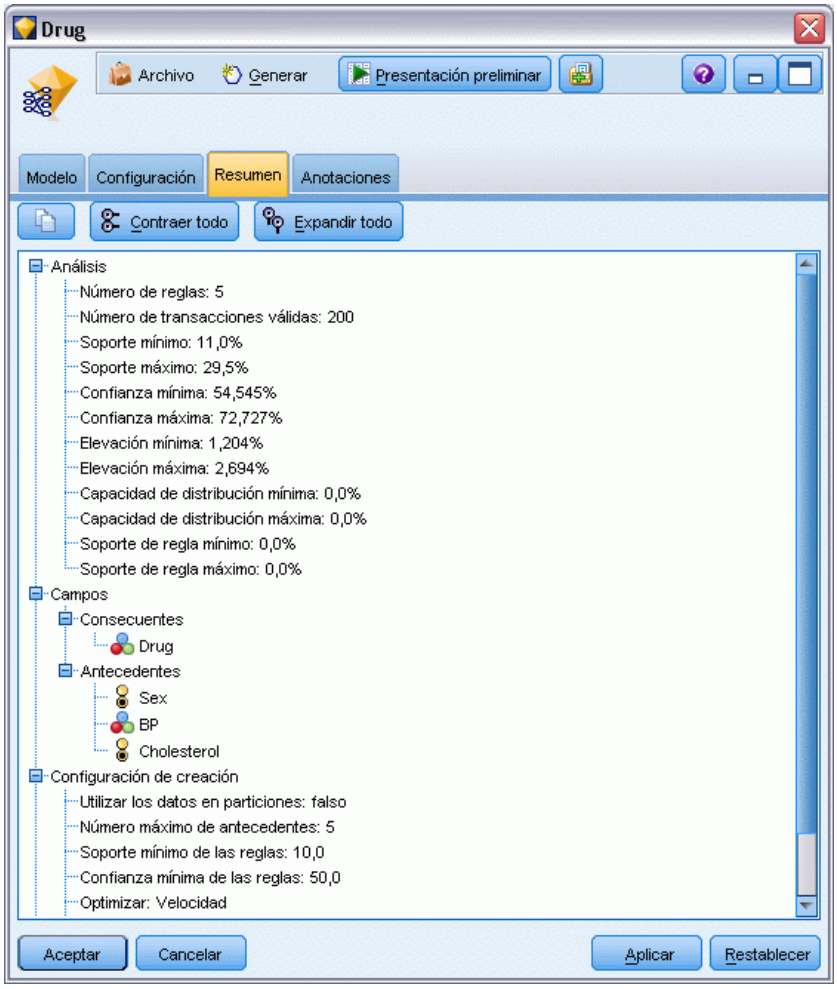
Comprobar que los pronósticos están en la cesta. Seleccione esta opción para asegurarse de que los consecuentes también se encuentran en la cesta. Este método es útil cuando se intenta comprender mejor los clientes o las transacciones existentes. Por ejemplo, es posible que desee identificar las reglas con la mayor elevación y, a continuación, explorar qué clientes se ajustan a estas reglas.

No comprobar los pronósticos 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.

Resumen del nugget de modelo de reglas de asociación

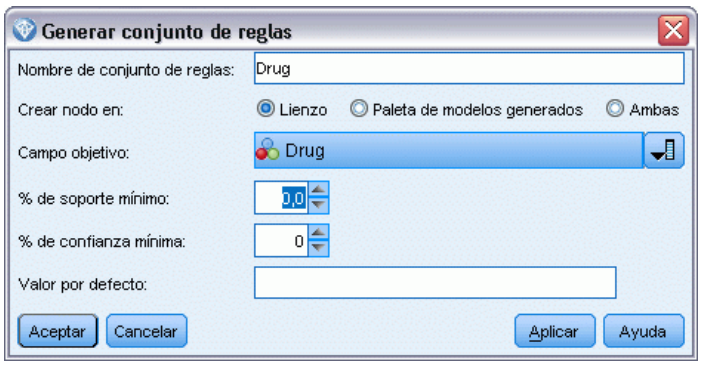
La pestaña Resumen para un nugget de modelo de reglas de una asociación muestra el número de reglas descubiertas y el mínimo y máximo de soporte y distribuibilidad de las reglas en el conjunto de reglas.

Figura 12-16
Pestaña Resumen del nugget de modelo de reglas de asociación



Generación de un conjunto de reglas desde un nugget de modelo de asociación

Figura 12-17
Cuadro de diálogo Generar conjunto de reglas



Los nugget de modelo de asociación, como A priori 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 sin refinar que no puede utilizarse directamente para la puntuación. Si desea obtener más información, consulte el tema Modelos sin refinar en el capítulo 3 el p. 75.

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.

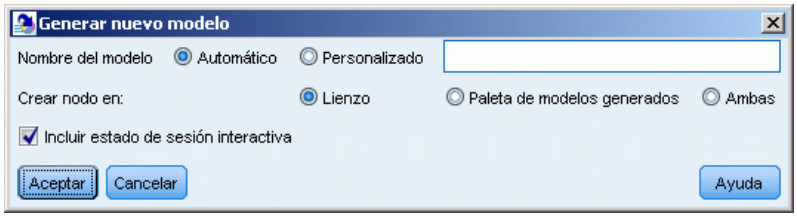
Soporte mínimo. Especifique el soporte mínimo para las reglas que desea conservar en el conjunto de reglas generado. El nuevo conjunto de reglas no incluirá reglas con un soporte inferior al valor especificado.

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 por defecto. Permite especificar un valor por defecto para el campo objetivo que se asigna a los registros puntuados para los que no se activa una regla.

Generación de un modelo filtrado

Figura 12-18
Cuadro de diálogo Generar nuevo modelo



Para generar un modelo filtrado desde un nugget de modelo de asociación, como un nodo A priori, 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. Por defecto, 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 pronóstico: *P* representa el pronóstico, *C* representa la confianza y, por último, *I* representa el ID de regla. 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 el p. 404 para ver conceptos básicos de estos formatos.

Por ejemplo, suponga que está puntuando datos de la cesta de la compra con un modelo que genera pronósticos basados en estas tres reglas:

```
Rule_15  pan y vino -> carne (confianza 54%)
Rule_22  queso -> fruta (confianza 43%)
Rule_5  pan y queso -> verducong (confianza 24%)
```

Datos tabulares. En el caso de los datos tabulares, se devuelven los tres pronósticos (3 es el valor por defecto) en un único registro.

Tabla 12-1
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	verducong	0,24	5

Datos transaccionales. En los datos transaccionales, se genera un registro independiente para cada pronóstico. Los pronósticos se siguen añadiendo en columnas independientes, pero las puntuaciones se devuelven cuando se calculan. Así se producen registros con pronósticos incompletos, como se muestra en el siguiente resultado de muestra. El segundo y el tercer pronóstico (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 los tres pronósticos.

Tabla 12-2
Puntuaciones en formato transaccional

ID	Elemento	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	verducong	0,24	5

Para incluir sólo los pronósticos completos con propósito de informar o distribuir, 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 de la siguiente manera:

Campo nuevo	Nombre del campo de ejemplo
Pronóstico	*$A-TRANSACCIÓN_NÚMERO-1*
Confianza (u otro criterio)	*$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 esas reglas como “de dos direcciones”, se devuelven los pronósticos en el formato mostrado en la siguiente tabla:

Tabla 12-3
Puntuación de resultados incluido un pronóstico con varios consecuentes

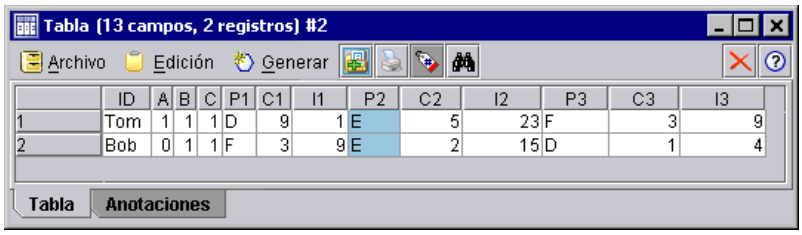
ID	Pan	Vino	Queso	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	carne&verdu	0,54	16	fruta	0,43	22	verducong	0,24	5

En ciertos casos, es posible que sea necesario dividir dichas puntuaciones antes de realizar la distribución. Para dividir un pronóstico con varios consecuentes, deberá analizar el campo mediante las funciones de cadena de CLEM. Si desea obtener más información, consulte el tema Funciones de cadena en el capítulo 8 en *Manual de usuario de IBM SPSS Modeler 15*.

Distribución de modelos de asociación

Al puntuar modelos de asociación, los pronósticos y las confianzas se muestran en columnas independientes (donde *P* representa el pronóstico, *C* representa la confianza y *I* representa el ID de regla). En este caso los datos de entrada pueden ser tabulares o transaccionales. Si desea obtener más información, consulte el tema Reglas de asociación de la puntuación el p. 427.

Figura 12-19
Puntuaciones tabulares con pronósticos en columnas



	ID	A	B	C	P1	C1	I1	P2	C2	I2	P3	C3	I3
1	Tom	1	1	1	D	9	1	E	5	23	F	3	9
2	Bob	0	1	1	F	3	9	E	2	15	D	1	4

Al preparar puntuaciones para distribución, puede que la aplicación requiera transponer los datos de salida a un formato con pronósticos en filas en lugar de columnas (un pronóstico por fila, que en ocasiones se conoce como formato “anidado”).

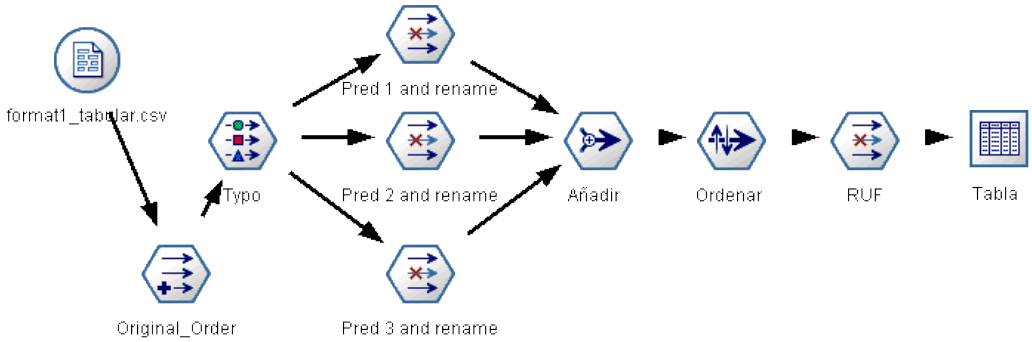
Figura 12-20
Puntuaciones transpuestas con pronósticos en filas

	ID	A	B	C	Pred	Crit	Rule_ID
1	Tom	1	1	1	D	9	1
2	Tom	1	1	1	E	5	23
3	Tom	1	1	1	F	3	9
4	Bob	0	1	1	F	3	9
5	Bob	0	1	1	.	$null$	$null$
6	Bob	0	1	1	.	$null$	$null$

Transposición de puntuaciones tabulares

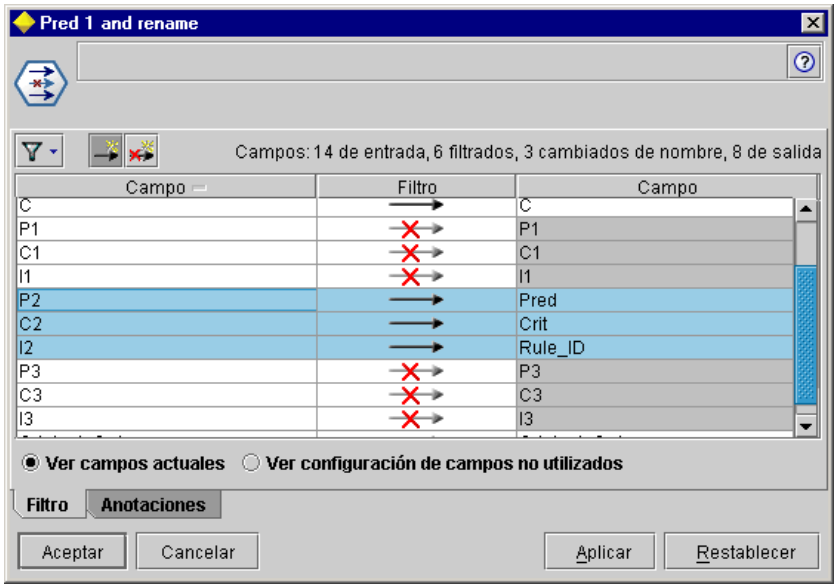
Puede transponer puntuaciones tabulares de columnas a filas utilizando una combinación de pasos en IBM® SPSS® Modeler, como se describe en los siguientes pasos.

Figura 12-21
La ruta de ejemplo utilizada para transponer los datos tabulares en formato anidado



- Utilice la función @INDEX en un nodo Derivar para comprobar el orden actual de los pronósticos y guardar este indicador en un nuevo campo, como *Orden_original*.
- Añada un nodo Tipo para asegurarse de que todos los campos están instanciados.
- Utilice un nodo Filtro para cambiar el nombre por defecto del pronóstico, la confianza y los campos de ID (*P1, C1, I1*) en campos comunes, como *Pron*, *Crit* e *ID_de_regla*, que se utilizará para añadir registros más tarde. Necesitará un nodo Filtro para cada pronóstico generado.

Figura 12-22
Filtrado de campos para los pronósticos 1 y 3 a la vez que se cambia el nombre de los campos para el pronóstico 2.



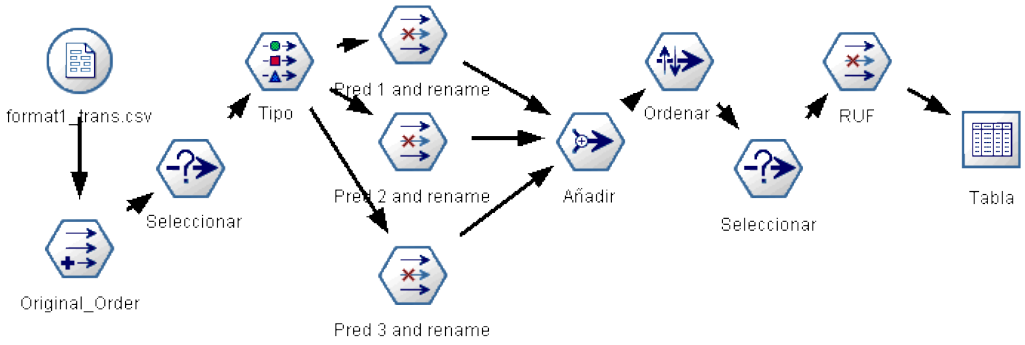
- Utilice un nodo Agregar para añadir valores para los datos compartidos *Pron*, *Crit* e *ID_de_regla*.
- Conecte un nodo Ordenar para clasificar registros en orden ascendente para el campo *Orden_original* y en orden descendente para *Crit*, que es el campo utilizado para clasificar los pronósticos por criterios como confianza, elevación y soporte.
- Utilice otro nodo Filtro para filtrar el campo *Orden_original* desde el resultado.

En este punto, los datos ya están listos para la distribución.

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 un único pronóstico en cada fila según sea necesario para la distribución.

Figura 12-23
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 únicos 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 Rule_ID 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 la distribución, contacte el servicio de asistencia técnica.

Nodo Secuencia

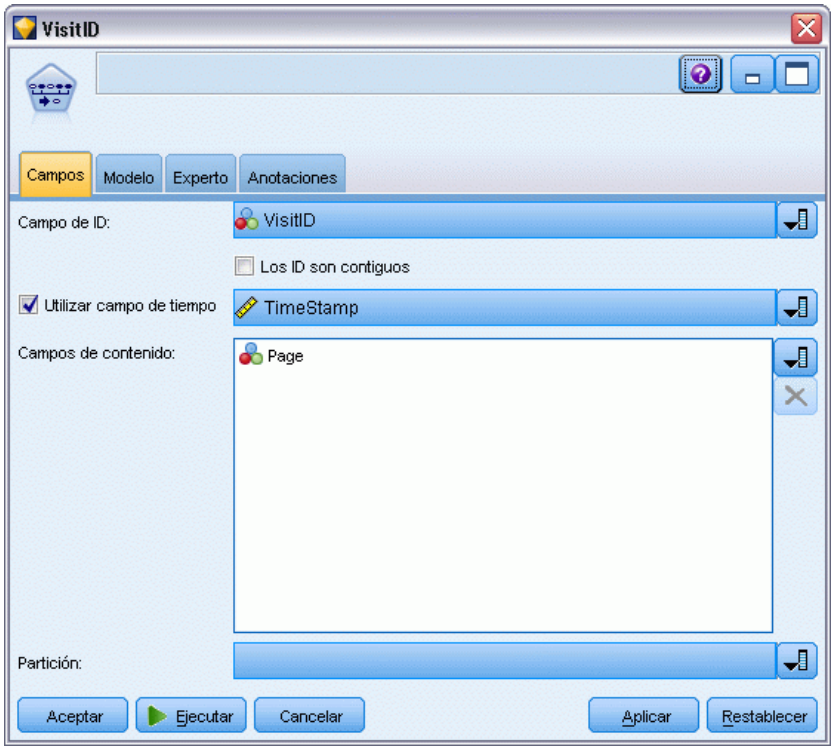
El nodo Secuencia descubre patrones, en datos secuenciales u ordenados en el tiempo, con el formato pan > queso. Los elementos de una secuencia son **conjuntos de elementos** que constituyen una única transacción. Por ejemplo, si una persona va a la tienda y compra pan y leche y, varios días después, 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 pronósticos.

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 papel o nivel de medición. Si se especifica un campo de tiempo, éste puede tener cualquier papel, 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 papel, 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 pronósticos. El nodo de modelo generado también puede generar Supernodos para detectar y contar las secuencias específicas y realizar pronósticos basados en secuencias específicas.

Opciones de campos para el nodo Secuencia

Figura 12-24
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 único de este campo debe indicar una unidad de análisis específica. Por ejemplo, en una aplicación de la cesta del supermercado, 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).

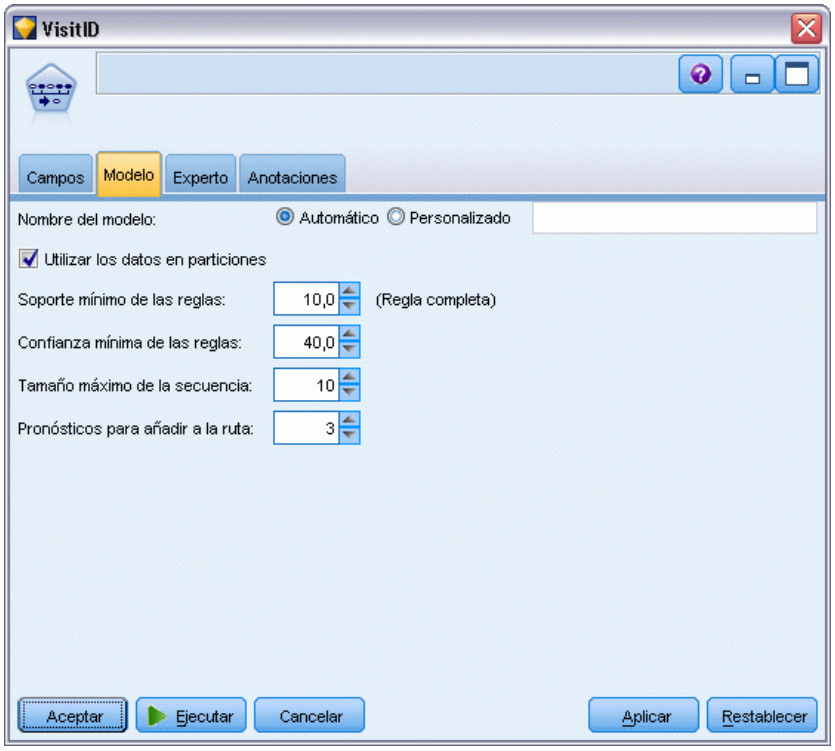
Campos de contenido. Especifique los campos de contenido del modelo. Estos campos contienen los eventos de interés del modelado de secuencia.

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. Si desea obtener más información, consulte el tema Datos tabulares frente a datos transaccionales el p. 404.

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 sobre la adecuación 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 sólo hay una partición, se usará automáticamente siempre que se active la partición.) Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*. 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

Figura 12-25
Opciones de modelo para el nodo Secuencia



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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Soporte mínimo de las 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 mínima de las reglas (%). 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 un pronóstico correcto a partir de todos los campos de ID para los que la regla realiza un pronóstico. 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.

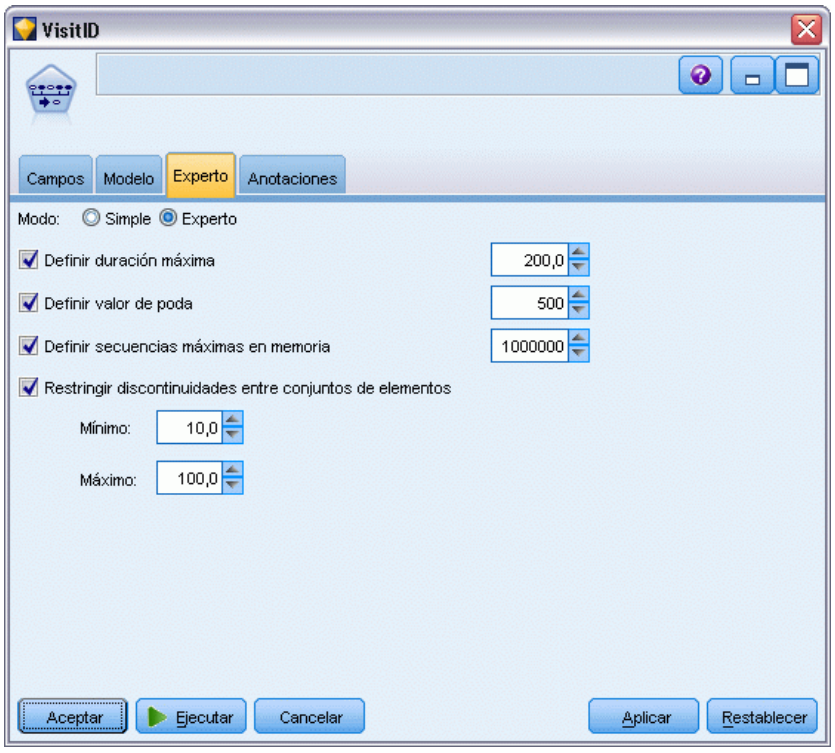
Tamaño máximo de la secuencia. Se puede configurar el número máximo de *conjuntos de elementos* (a diferencia de los *elementos*) distintos en una secuencia. Si las secuencias de interés resultantes son pocas, se puede disminuir el valor del parámetro para que el conjunto de secuencias se genere más rápido.

Pronósticos para añadir a la ruta. Especifique el número de pronósticos que desea que el nodo Modelo resultante añada a la ruta. Si desea obtener más información, consulte el tema Nugget del modelo de secuencia el p. 437.

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.

Figura 12-26
Opciones de experto para el nodo Secuencia



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 esta lista de transacciones:

ID	Time	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 apples -> bread porque la discontinuidad entre apples y bread es inferior a la discontinuidad mínima. Sucedería igual si los datos fueran los siguientes:

ID	Time	Contenido
1001	1	manzanas
1001	2	pan
1001	5	queso
1001	20	ropa

y la discontinuidad máxima se hubiese establecido en 10, no aparecería ninguna secuencia con dressing, porque la discontinuidad entre cheese y dressing es demasiado amplia para que se consideren parte de la misma secuencia.

Nugget del modelo de secuencia

Los nuggets de modelo de secuencia representan las secuencias que se encuentran para un campo de salida determinado descubierto por el nodo Secuencia y pueden añadirse a rutas para generar pronósticos.

Al ejecutar una ruta que contenga un nodo de secuencia, dicho nodo añade en los datos un par de campos con los pronósticos y los valores de confianza asociados para cada pronóstico desde el modelo de secuencia. Por defecto, se añaden tres pares de campos con los tres mejores pronósticos (y los valores de confianza asociados). Puede cambiar el número de pronósticos generados 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. Si desea obtener más información, consulte el tema Configuración del nugget de modelo de secuencia el p. 441.

Los nuevos nombres de campos se derivan del nombre del modelo. Los nombres de campos son *$S-secuencia-n* para el campo de pronóstico (donde *n* indica el *enésimo* pronóstico) y *$SC-secuencia-n* para el campo de la confianza. En una ruta con varios nodos de reglas de secuencia en una serie, los nuevos nombres de campos incluirán números en el prefijo para distinguirse entre sí. El primer nodo de conjuntos de secuencia 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. Los pronósticos se muestran en orden de confianza, por lo que *$S-secuencia-1* contiene el pronóstico con la mayor confianza, mientras que*$S-secuencia-2* contiene la siguiente mayor confianza, y así sucesivamente. Para los registros en los que el número de pronósticos disponibles es menor que el número de pronósticos solicitados, los pronósticos restantes contienen el valor $null$. Por ejemplo, si solo se pueden realizar dos pronósticos 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 los pronósticos *k* para el registro, donde *k* es el número de pronósticos especificados en la pestaña Configuración después de añadir el modelo a la ruta. (Si varias reglas pronostican el mismo resultado para el conjunto de transacciones, sólo se

utilizará la regla con la mayor confianza.) Si desea obtener más información, consulte el tema Configuración del nugget de modelo de secuencia el p. 441.

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. Si desea obtener más información, consulte el tema Reglas de asociación de la puntuación el p. 427.

Nota: al puntuar los datos mediante un nodo de conjuntos de secuencia generado en una ruta, cualquier configuración de tolerancia o discontinuidad seleccionada en la creación del modelo se omite para efectuar la puntuación.

Pronósticos 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, los pronósticos 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*, dada la historia de transacciones para el ID actual hasta el registro actual.

Tenga en cuenta que los pronósticos 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 pronóstico útil a la secuencia, el pronóstico 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 la pasa a los siguientes registros:

ID	Compra	Pronóstico
001	jam	bread
001	milk	bread

Observe que el primer registro genera un pronóstico de *pan*, como se esperaba. El segundo registro también contiene un pronóstico de *pan* porque no existe ninguna regla para la *mermelada* seguida de *leche*; por lo tanto, la transacción de *leche* no añade información útil y la regla Jam -> Bread se sigue aplicando.

Generación de nuevos nodos

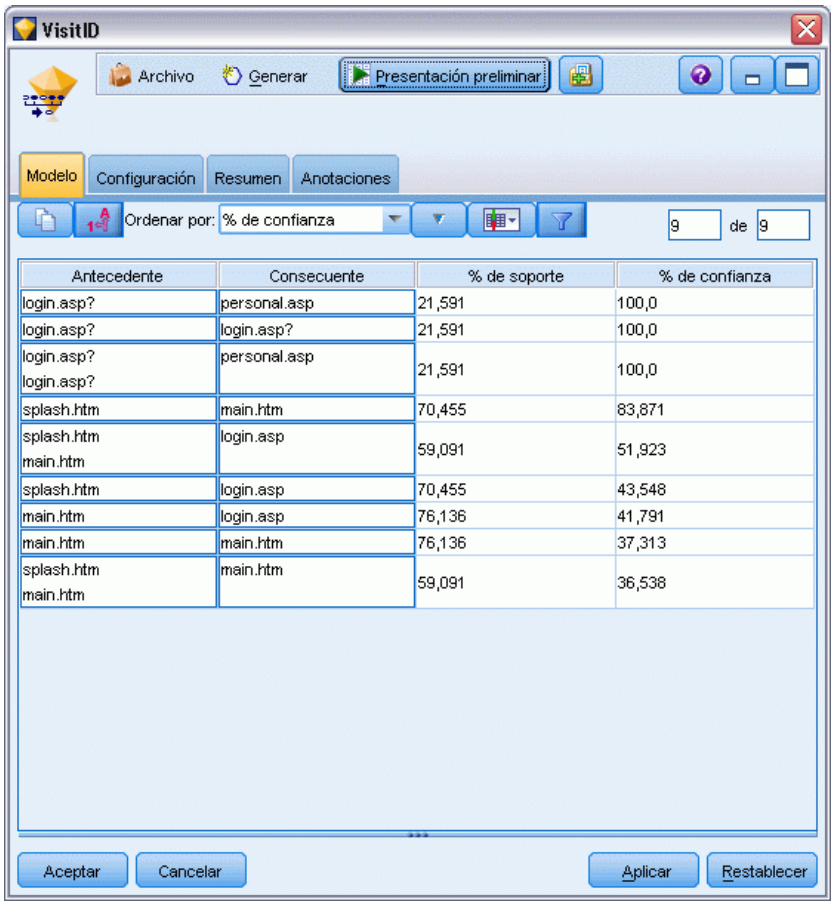
El menú Generar permite crear nuevos Supernodos basados en el modelo de secuencia.

- **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. Si desea obtener más información, consulte el tema Generación de un Supernodo Regla a partir de un nugget de modelo de secuencia el p. 444.
- **Modelo a paleta** Devuelve el modelo a la paleta de modelos. Esto resulta útil en situaciones en las que un compañero puede haber enviado una transmisión que contenga el modelo, pero no el modelo propiamente dicho.

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.

Figura 12-27
Pestaña Modelo del nugget Secuencia



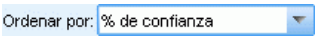
Cada regla se muestra en el siguiente formato:

Antecedente	Consecuente
beer and cannedveg	beer
fish fish	fish

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 instancia 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 instancia 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).

Figura 12-28
Opciones de ordenación de la barra de herramientas



Los valores se pueden ordenar por:

- 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.

Antecedente	Consecuente
beer and cannedveg and frozenmeal	frozenmeal
beer and cannedveg	beer
fish fish	fish
softdrink	softdrink

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 únicos para los que aparece la *secuencia completa*, tanto el antecedente como el consecuente. (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 bread -> cheese, 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 bread -> cheese 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 un pronóstico correcto a partir de todos los campos de ID para los que la regla realiza un pronóstico. 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 cannedveg (indicando así soporte del antecedente), pero sólo el 20% contiene tanto cannedveg como frozenmeal, la confianza de la regla cannedveg -> frozenmeal sería Rule Support / Antecedent Support 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 contiene tanto *pan* como *queso*, el soporte de regla para la regla bread -> cheese es 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. Se descartan las transacciones no válidas (las que no disponen ni de elementos ni de valores verdaderos) 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.

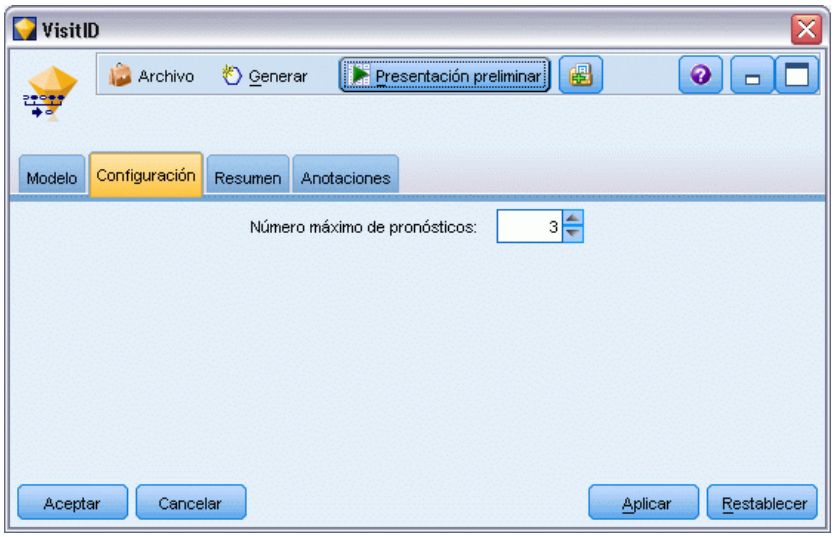
Figura 12-29
Botón de filtrado

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.

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.

Figura 12-30
Pestaña Configuración del nugget Secuencia

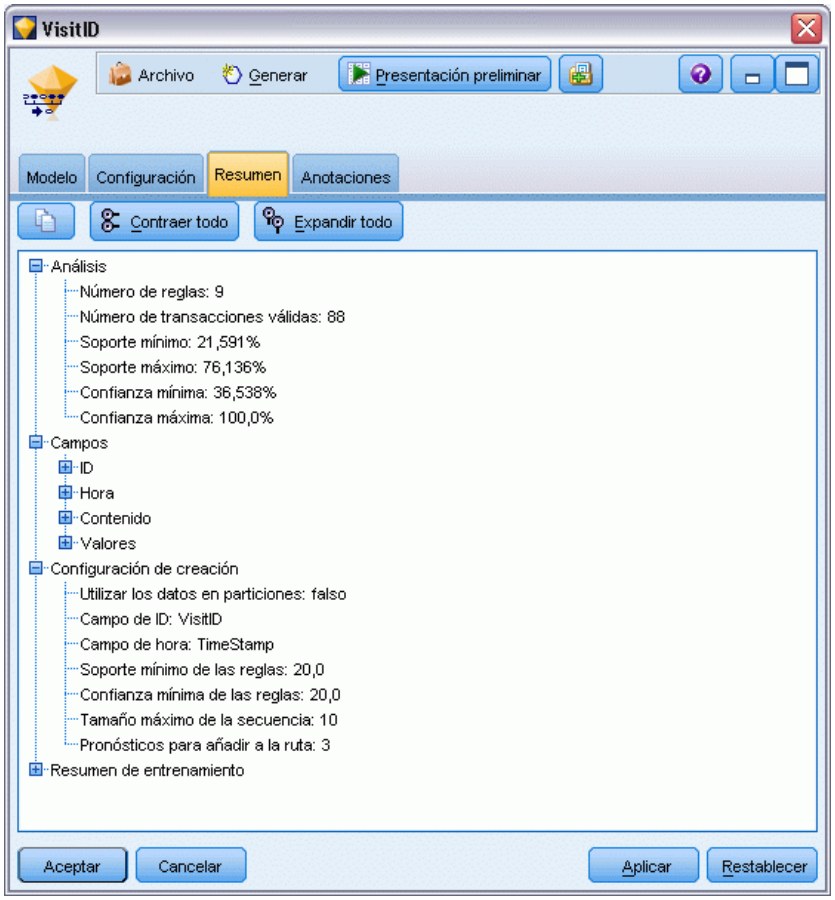


Número máximo de pronósticos. Especifique el número máximo de pronósticos incluidos para cada conjunto de elementos de la cesta. Las reglas con los valores de confianza más altos que se aplican a este conjunto de transacciones se utilizan para generar predicciones para el registro hasta el límite especificado.

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. Si desea obtener más información, consulte el tema Nodo Análisis en el capítulo 6 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

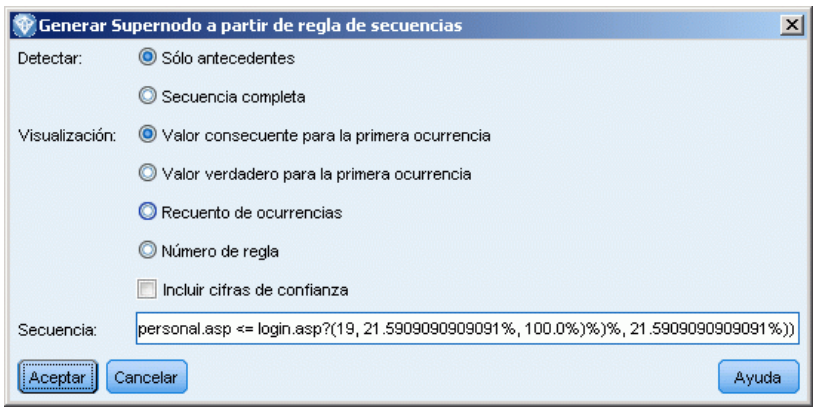
Figura 12-31
Pestaña Resumen de modelo del nugget Secuencia



Si desea obtener más información, consulte el tema Exploración de nugget de modelo en el capítulo 3 el p. 53.

Generación de un Supernodo Regla a partir de un nugget de modelo de secuencia

Figura 12-32
Cuadro de diálogo de generación del Supernodo Regla



Para generar un Supernodo Regla basado en una regla de secuencia:

- 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.
- 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 el consecuente de la regla seleccionada en el orden correcto dentro de un conjunto de registros que tengan el mismo ID. Este proceso no tiene en cuenta la tolerancia de marca de tiempo o la configuración de la restricción de la discontinuidad de los elementos 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 pronosticado 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*.

Capítulo
13

Modelos de series temporales

¿Por qué es importante pronosticar?

Pronosticar 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 predicciones 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 predicciones más precisas en las que podrá basar sus decisiones.

Datos de series temporales

Una **serie temporal** es una colección ordenada de medidas tomadas en intervalos regulares; por ejemplo, los precios diarios de las acciones o los datos de ventas semanales. Las medidas pueden estar relacionadas con cualquier cosa que le interese, y cada serie se suele clasificar en una de las siguientes categorías:

- **Dependiente.** Serie que desea pronosticar.
- **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 como, por ejemplo, las 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 medidas. Además, si algún intervalo no tiene ninguna medida, debe definirse en el valor perdido. De esta forma, el número de intervalos con medidas (incluidos los que tienen valores perdidos) define la duración del período histórico de los datos.

Características de las series temporales

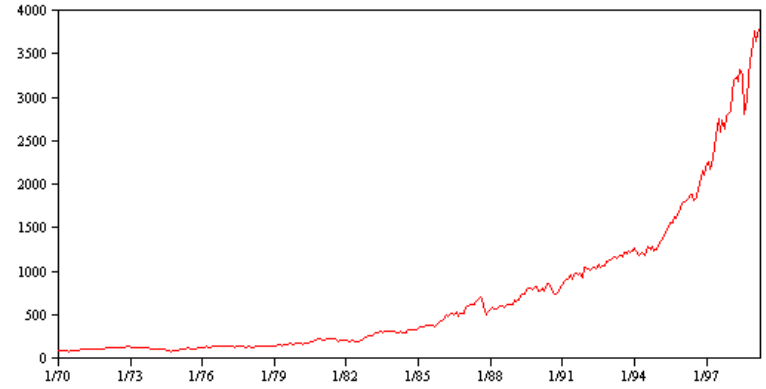
Estudiar el comportamiento pasado de una serie le ayudará a identificar los patrones y realizar mejores pronósticos. Cuando se representan, muchas series temporales muestran una o varias de estas características:

- 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 13-1
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.

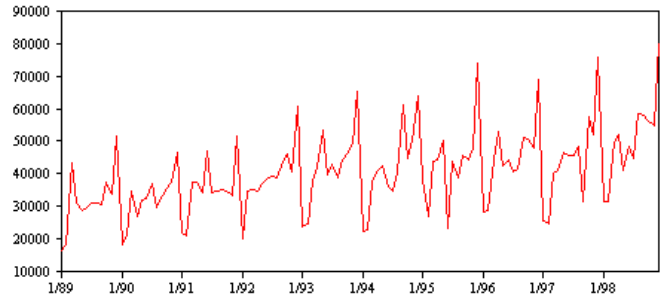
Las tendencias también pueden ser **lineales** o **no lineales**. Las tendencias lineales son incrementos aditivos positivos o negativos en el nivel de la serie, comparables al efecto del interés simple sobre el principal. Las tendencias no lineales suelen ser multiplicativas, con incrementos proporcionales a los valores de series anteriores.

Las tendencias lineales globales se ajustan y pronostican correctamente mediante modelos ARIMA y 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 13-2
Ciclo estacional



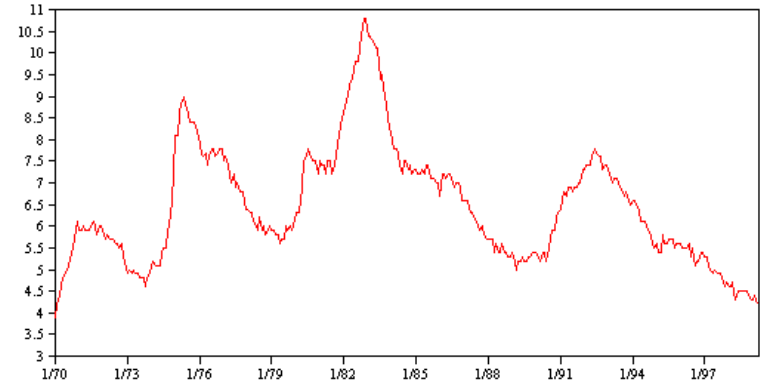
Los ciclos estacionales están ligados al intervalo de la serie. Por ejemplo, los datos mensuales suelen mostrar un comportamiento cíclico a lo largo de trimestres y años. Una serie mensual puede mostrar un ciclo trimestral significativo con un mínimo en el primer trimestre o un ciclo anual con un pico 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 predicciones. 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 13-3
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.

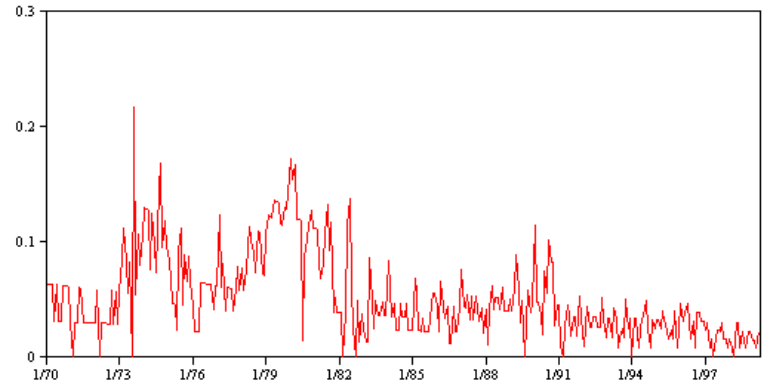
Los patrones cíclicos no estacionales son difíciles de modelar y suelen aumentar la incertidumbre de los pronósticos. El mercado de valores, por ejemplo, proporciona numerosos ejemplos de series que han desafiado el trabajo de los pronosticadores. No obstante, los patrones no estacionales se deben tener 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 de los pronósticos.

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 13-4
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 predicciones 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.)

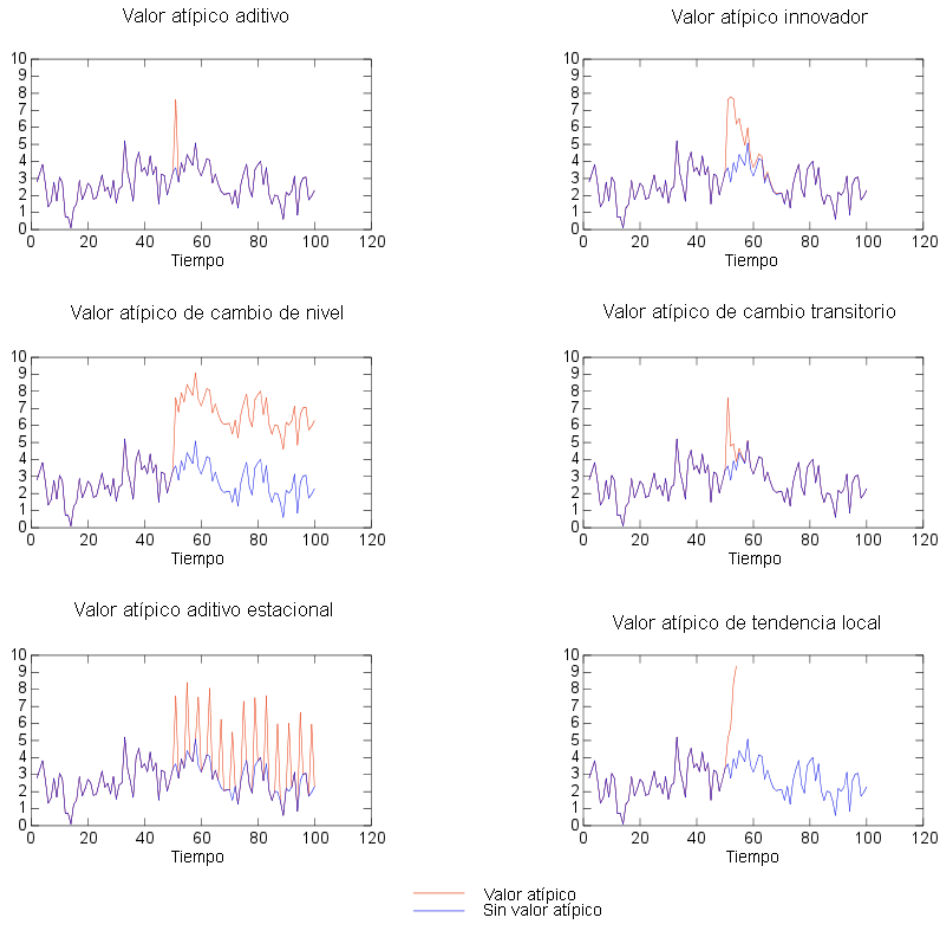
Si puede explicarse una alteración, se puede modelar mediante una **intervención** o un **evento**. Por ejemplo, en agosto de 1973, la Organización de Países Exportadores de Petróleo (OPEP) impuso un embargo sobre el petróleo que cambió drásticamente la tasa de inflación, aunque recuperó sus niveles 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 predicciones 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 pronóstico 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 13-5
Tipos de valores atípicos



- **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 siguientes valores atípicos aditivos 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.

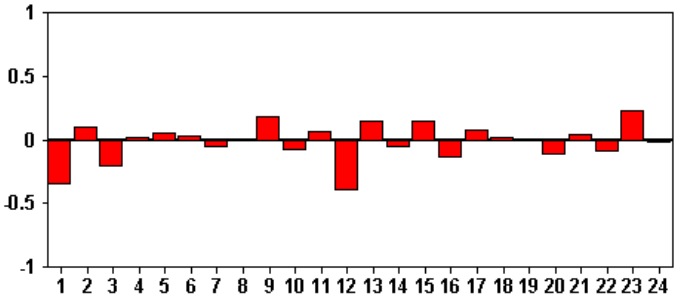
La detección de valores atípicos en una serie temporal implica determinar la ubicación, tipo y magnitud de todos los valores atípicos presentes. Tsay (1998) propuso un procedimiento iterativo para detectar el cambio del nivel de la media con el fin de identificar los valores atípicos deterministas. Este 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 13-6
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 que se va a pronosticar. Por ejemplo, un minorista de venta por catálogo o por Internet podría predecir 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 pronosticar 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 un mejor pronóstico, por lo que una mayor complejidad puede no valer la pena. Lo ideal sería identificar el modelo más simple que mejores pronósticos realice.

Como norma general, se recomienda que el número de predictores sea inferior al tamaño muestral 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 predicción. 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 Modelos de series temporales

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 predicciones a partir de los datos de series temporales.

El **suavizado exponencial** es un método de predicción que utiliza los valores ponderados de las observaciones anteriores de la serie para pronosticar los valores futuros. Como tal, el suavizado exponencial no se basa en una comprensión teórica de los datos. Pronostica un punto cada vez, corrigiendo las predicciones a medida que entran nuevos datos. Esta técnica es útil para pronosticar las series que muestran una tendencia, estacionalidad o ambas. Puede elegir entre una amplia variedad de modelos de suavizado exponencial que difieren en el 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 puedan ayudar a explicar el comportamiento de la serie que se está pronosticando, como el número de catálogos enviados por correo o el número de visitas de 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 ha mostrado picos cada 12 meses, es probable que lo siga haciendo aunque se desconozca el motivo.

Existe también un **modelizador experto** que identifica y estima automáticamente el modelo ARIMA o de suavizado exponencial que mejor se ajusta para una o más variables objetivo, lo que elimina la necesidad de identificar un modelo adecuado mediante ensayo y error. En todos los casos, el modelizador experto elige el mejor modelo para cada variable objetivo especificada. En caso de duda, utilice el modelizador experto.

Si se especifican variables predictoras, el modelizador experto selecciona, para su inclusión en los modelos ARIMA, las variables que tienen una relación estadísticamente significativa con la serie dependiente. Las variables del modelo se transforman cuando es necesario mediante una diferenciación y/o una raíz cuadrada o una transformación logarítmica natural. Por defecto, 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.

Ejemplo. Un analista que trabaja para un proveedor de banda ancha a nivel nacional debe generar predicciones de las suscripciones de usuarios para pronosticar la utilización de la banda ancha. Las predicciones se deben realizar para cada uno de los mercados locales que conforman la base nacional de suscriptores. Puede utilizar el modelado de series temporales para generar predicciones de los tres meses siguientes para varios mercados locales. Si desea obtener más información, consulte el tema Predicciones con el nodo Serie temporal en el capítulo 14 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

Requisitos

El nodo Serie temporal se distingue de otros nodos de IBM® SPSS® Modeler en que no se puede insertar simplemente en una ruta y ejecutar dicha ruta. El nodo Serie temporal debe ir siempre precedido por un nodo Intervalos de tiempo que especifique información como el intervalo de tiempo que se va a utilizar (años, trimestres, meses, etc.), los datos que se van a utilizar para la estimación y lo que se va a prolongar un pronóstico en el futuro, si se utiliza.

Figura 13-7
Un nodo Serie temporal siempre debe ir precedido de un nodo Intervalos de tiempo

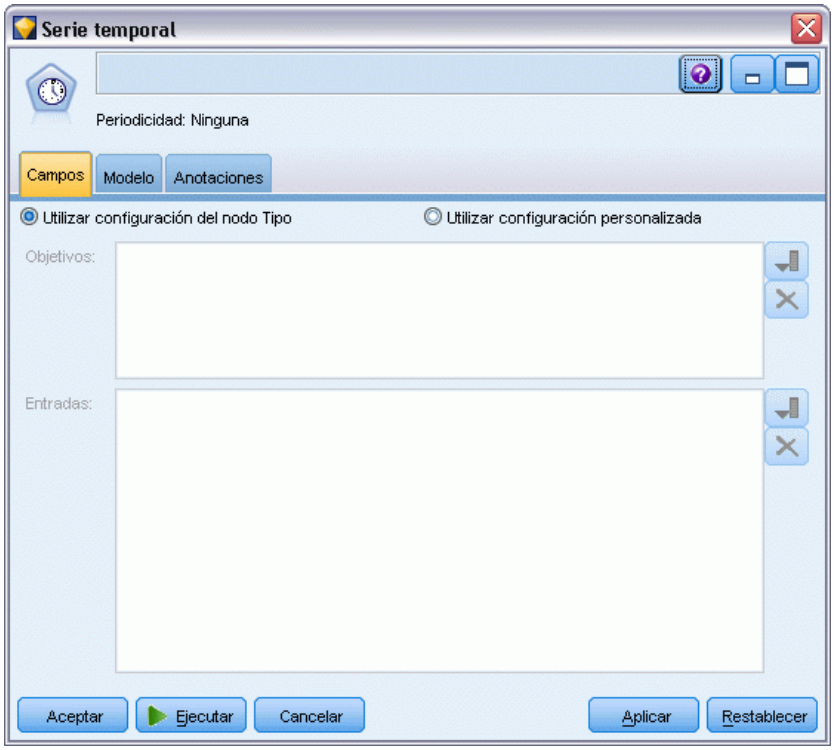
Los datos de serie temporal deben estar espaciados de manera uniforme. Los métodos para modelar datos de series temporales precisan de un intervalo uniforme entre cada medida. Los valores perdidos deben indicarse mediante filas vacías. Si sus datos aún no cumplen este requisito, el nodo Intervalos de tiempo puede transformar los valores según sea necesario. Si desea obtener más información, consulte el tema Nodo de intervalos de tiempo en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Deben tenerse en cuenta estas otras cuestiones en relación con el nodo Serie temporal:

- Los campos deben ser numéricos
- Los campos de fecha no se pueden utilizar como entradas
- Las particiones se ignorarán

Opciones de los campos

Figura 13-8
Cuadro de diálogo del nodo Serie temporal, pestaña Campos



En la pestaña Campos se especifican los campos que se van a utilizar para generar el modelo. Para generar un modelo, primero se deben especificar los campos que se desea usar como objetivos y como entradas. El nodo Serie temporal suele utilizar información de campo de un nodo Tipo situado en un punto anterior de la ruta. Si utiliza un nodo Tipo para seleccionar campos de entrada y objetivo, no es necesario cambiar nada en esta pestaña.

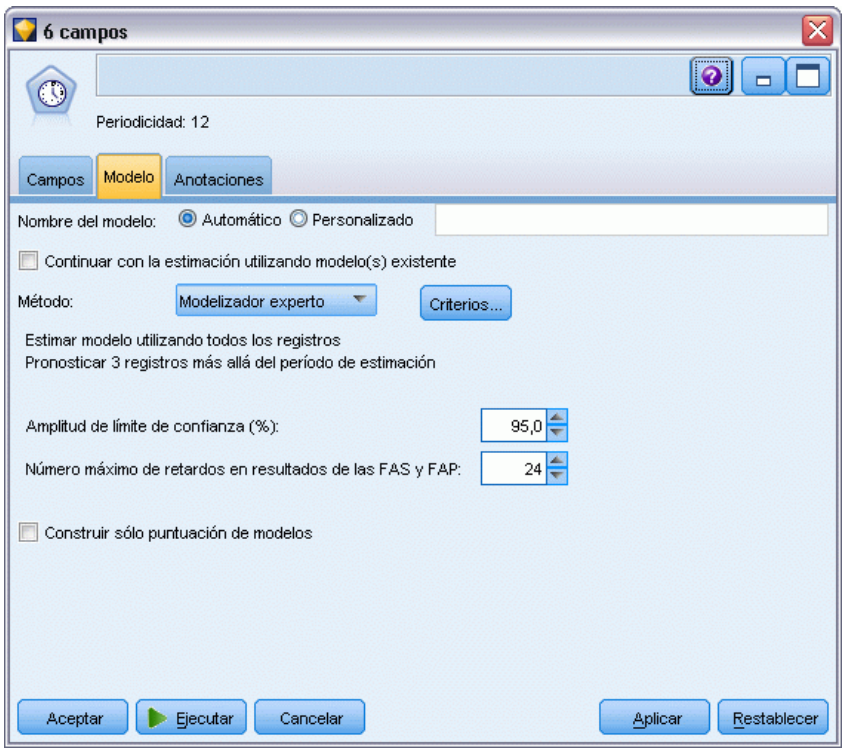
Utilizar configuración del nodo Tipo. Esta opción permite indicar al nodo que use la información de campo de un nodo Tipo situado en un punto anterior de la ruta. Este es el método por defecto.

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. Tenga en cuenta que los campos almacenados como fechas no se aceptan como campos objetivo o de entrada.

- **Objetivos.** Seleccione uno o varios campos objetivo. Se trata de una acción similar a establecer un papel de un campo a *Objetivo* en un nodo Tipo. Los campos objetivos de un modelo de series temporales debe tener un nivel de medición de *Continuo*. Se crea un modelo distinto para cada campo objetivo. Un campo objetivo tiene en cuenta todos los campos *Entrada* especificados (excepto dicho campo) como posibles entradas. Por lo tanto, el mismo campo puede incluirse en ambas listas; ese campo se utilizará como posible entrada para todos los modelos, excepto en el que actúa como objetivo.
- **Entradas.** Seleccione los campos de entrada. Se trata de una acción similar a establecer un papel de un campo a *Entrada* en un nodo Tipo. Los campos de entrada de un modelo de serie temporal deben ser numéricos.

Opciones del modelo de serie temporal

Figura 13-9
Cuadro de diálogo del nodo Serie temporal, pestaña Modelo



Nombre del modelo. Especifica el nombre asignado al modelo generado al ejecutar el nodo.

- **Automático.** Genera el nombre del modelo de forma automática basándose en los nombres de los campos objetivo o de ID, o en el nombre del tipo de modelo en los casos en los que no se especifique ningún campo objetivo (como en los modelos de conglomerado).
- **Personalizado.** Permite especificar un nombre personalizado para el nugget de modelo.

Continuar con la estimación utilizando modelo(s) existente. Si ya ha generado un modelo de serie temporal, seleccione esta opción para reutilizar la configuración de criterios especificada para ese modelo y generar un nuevo nodo de modelo en la paleta de modelos, en lugar de crear uno a partir de cero. De esta manera podrá ahorrar tiempo si vuelve a estimar y generar un nuevo pronóstico basado en la misma configuración del modelo de antes, pero con 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 un pronóstico con ellos; el sistema no volverá a intentar buscar el mejor tipo de modelo para los nuevos datos. Si selecciona esta opción desactivará los controles Método y Criterios. Si desea obtener más información, consulte el tema Nueva estimación y predicción el p. 468.

Método. Puede elegir Modelizador experto, Suavizado exponencial o ARIMA. Si desea obtener más información, consulte el tema Nodo Modelos de series temporales el p. 453. Seleccione Criterios para especificar las opciones del método seleccionado.

- **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.

Información de intervalo de tiempo

Esta sección del cuadro de diálogo contiene información sobre las especificaciones para las estimaciones y las predicciones realizadas en el nodo Intervalos de tiempo. Tenga en cuenta que esta sección no aparecerá si elige la opción Continuar con la estimación utilizando modelo(s) existente.

En la primera línea se indica si hay algún registro que se excluya del modelo o se utilice como caso reservado. Si desea obtener más información, consulte el tema Período de estimación en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

En la segunda línea aparece información sobre todos los períodos de predicción especificados en el nodo Intervalos de tiempo. Si desea obtener más información, consulte el tema Predicciones en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Si en la primera línea aparece No se ha definido ningún intervalo de tiempo, esto significa que no hay ningún nodo Intervalos de tiempo conectado. Esta situación provocará un error al intentar ejecutar la ruta, por lo que deberá incluir un nodo Intervalos de tiempo en un punto de la ruta anterior al nodo Serie temporal.

Información variada

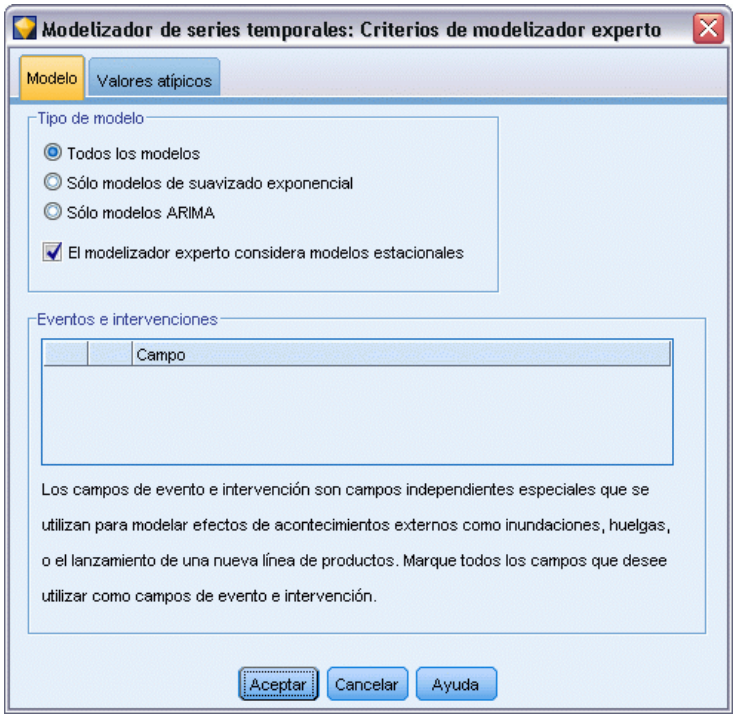
Amplitud de límite de confianza (%). Los intervalos de confianza se calculan para las predicciones del modelo y las autocorrelaciones residuales. Puede especificar cualquier valor positivo inferior a 100. Por defecto, se utiliza un intervalo de confianza del 95 %.

Número máximo de retardos en resultados de las FAS y FAP. Puede establecer el número máximo de retardos que se muestran en las tablas y en los gráficos de autocorrelaciones y autocorrelaciones parciales.

Construir únicamente modelo de puntuación. Marque esta casilla para reducir la cantidad de datos que están almacenados en el modelo. Al hacerlo mejorará el rendimiento al construir modelos con cifras muy grandes de series temporales (cientos de miles). Si selecciona esta opción, las pestañas Modelo, Parámetros y Residuos no se muestran en el nugget de modelo Series temporales, pero puede puntuar los datos de la manera habitual.

Criterios del modelizador experto de series temporales

Figura 13-10
Cuadro de diálogo Criterios de modelizador experto, pestaña Modelo



Tipo de modelo. Se encuentran disponibles las siguientes opciones:

- **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 sólo tiene en cuenta los modelos de suavizado exponencial.
- **Sólo modelos ARIMA.** El modelizador experto sólo tiene en cuenta los modelos ARIMA.

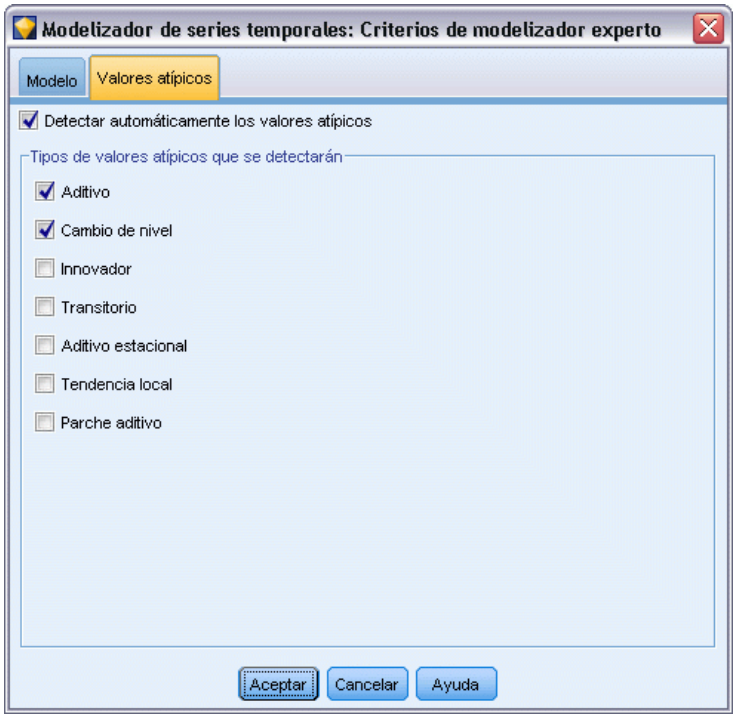
El modelizador experto considera modelos estacionales. Esta opción sólo está activada si se ha definido una periodicidad para el conjunto de datos activo. Si esta opción está seleccionada, el modelizador experto tiene en cuenta los modelos tanto estacionales como no estacionales. Si esta opción no está seleccionada, el modelizador experto sólo tiene en cuenta los modelos no estacionales.

Eventos e intervenciones. Permite designar determinados campos de entrada como campos de eventos o intervención. De esta forma, se identifica el campo que contiene datos de series temporales afectados por eventos (situaciones recurrentes predecibles, como promociones de ventas) o intervenciones (incidentes puntuales, como apagones o huelgas). El modelizador experto sólo tendrá en cuenta la regresión simple y no las funciones de transferencia arbitrarias para entradas identificadas como campos de evento o intervención.

Los campos de entrada deben tener un nivel de medición de *Marca*, *Nominal* u *Ordinal* y deben ser numéricos (por ejemplo, 1/0, no True/False, para un campo marca), antes de que aparezcan en esta lista. Si desea obtener más información, consulte el tema Pulsos y pasos el p. 449.

Valores atípicos

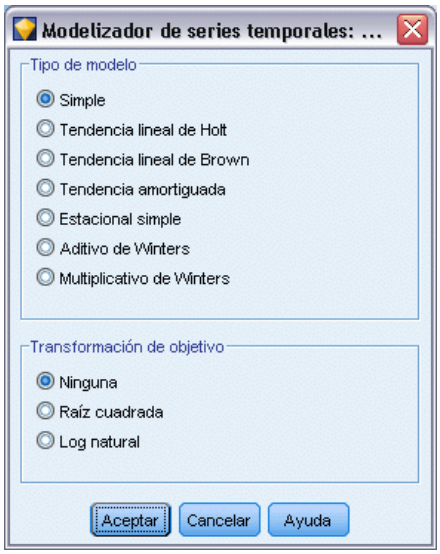
Figura 13-11
Cuadro de diálogo Criterios de modelizador experto, pestaña Valores atípicos



Detectar automáticamente los valores atípicos. Por defecto, 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. Si desea obtener más información, consulte el tema Valores atípicos el p. 450.

Criterios de suavizado exponencial de series temporales

Figura 13-12
Cuadro de diálogo Criterios de suavizado exponencial



Tipo de modelo. Los modelos de suavizado exponencial se clasifican como estacionales o no estacionales. Los modelos estacionales sólo están disponibles si se ha definido una periodicidad estacional mediante el nodo Intervalos de tiempo. Las periodicidades estacionales son las siguientes: períodos cíclicos, años, trimestres, meses, días por semana, horas por día, minutos por día y segundos por día. Si desea obtener más información, consulte el tema Nodo de intervalos de tiempo en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

- **Simple.** 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 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.
- **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.

- **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, $p = 12$.
- **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.

Transformación de objetivo. Puede especificar una transformación para que se lleve a cabo en cada variable dependiente antes de su modelado. Si desea obtener más información, consulte el tema Transformaciones de series el p. 452.

- **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.

Criterios ARIMA de series temporales

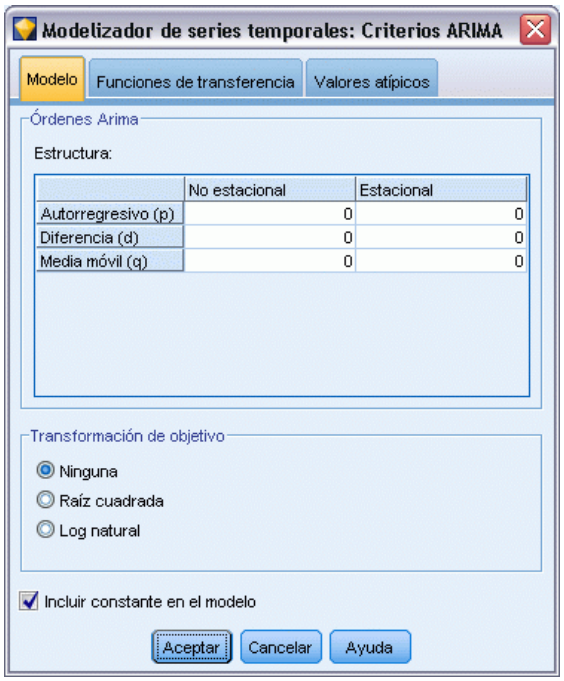
El nodo Serie temporal le permite generar modelos ARIMA estacionales y no estacionales personalizados (también conocidos como modelos Box-Jenkins) con o sin un conjunto fijo de variables de entrada (predictoras) . Puede definir funciones de transferencia para algunas o todas las variables de entrada y especificar la detección automática de valores atípicos o especificar un conjunto explícito de valores atípicos.

Todas las variables de entrada especificadas se incluyen en el modelo de manera explícita, a diferencia de lo que ocurre al utilizar el modelizador experto, donde las variables de entrada sólo se incluyen si tienen una relación estadísticamente significativa con la variable objetivo.

Modelo

La pestaña Modelo le permite especificar la estructura de un modelo ARIMA personalizado.

Figura 13-13
Cuadro de diálogo Criterios ARIMA, pestaña Modelo



Órdenes ARIMA. Escriba valores para los distintos componentes ARIMA del modelo en las casillas correspondientes de la cuadrícula Estructura. 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 sólo se activan 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 igual a 2 especifica que se van a utilizar los valores de la serie correspondientes a dos períodos de tiempo del 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 se corresponde con el grado de la tendencia de la serie (la diferenciación de primer orden representa las tendencias lineales, la diferenciación de segundo orden representa las tendencias cuadráticas, etc.).
- **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 de 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.

Órdenes estacionales. Los componentes estacionales autorregresivos, de media móvil y de diferenciación tienen la misma función que los componentes no estacionales correspondientes. No obstante, en el caso de los órdenes estacionales, los valores de la serie actual se ven afectados por los valores de la serie anterior separados por uno o más períodos 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.

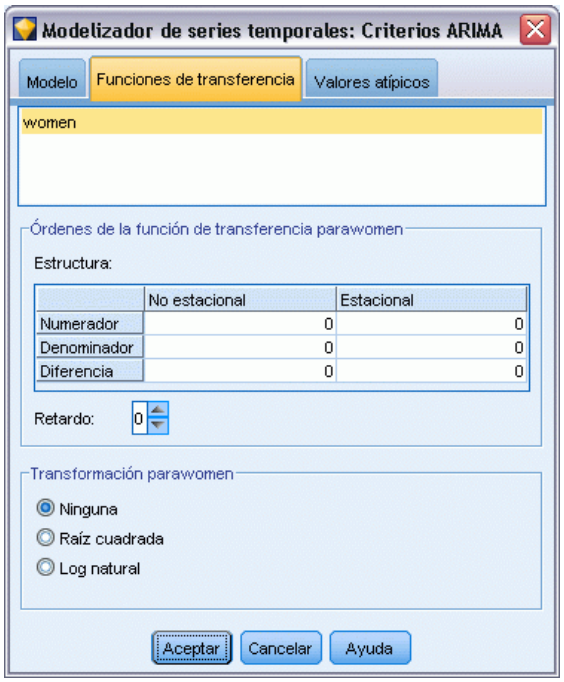
Transformación de objetivo. Puede especificar una transformación para que se lleve a cabo en cada variable objetivo antes de su modelado. Si desea obtener más información, consulte el tema Transformaciones de series el p. 452.

- **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.

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.

Funciones de transferencia

Figura 13-14
Cuadro de diálogo Criterios ARIMA, pestaña Funciones de transferencia



La pestaña Funciones de transferencia le permite definir funciones de transferencia para algunos o todos los campos de entrada. Las funciones de transferencia le permiten especificar el modo en que se utilizan los valores anteriores de estos campos para predecir valores futuros de la serie objetivo.

La pestaña se muestra únicamente si los campos de entrada (con el papel definido como *Entrada*) se especifican, ya sea en el nodo Tipo o en la pestaña Campos del nodo Series temporales (seleccione Utilizar configuración personalizada—Entradas). Si desea obtener más información, consulte el tema Definición del papel de campos en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

En la lista superior aparecen todos los campos de entrada. El resto de la información que aparece en este cuadro de diálogo es específica del campo de entrada seleccionado en la lista.

Órdenes de la función de transferencia. Escriba valores para los distintos componentes de la función de transferencia en las casillas correspondientes de la cuadrícula 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 sólo se activan 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 una serie independiente de un período anterior (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, un orden de denominador de 1 especifica que las desviaciones del valor medio de una serie independiente para un período de tiempo anterior se tienen en cuenta 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.

Órdenes estacionales. Los componentes estacionales de numerador, denominador y diferenciación tienen la misma función que los componentes no estacionales correspondientes. No obstante, en el caso de los órdenes estacionales, los valores de la serie actual se ven afectados por los valores de la serie anterior separados por uno o más períodos 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.

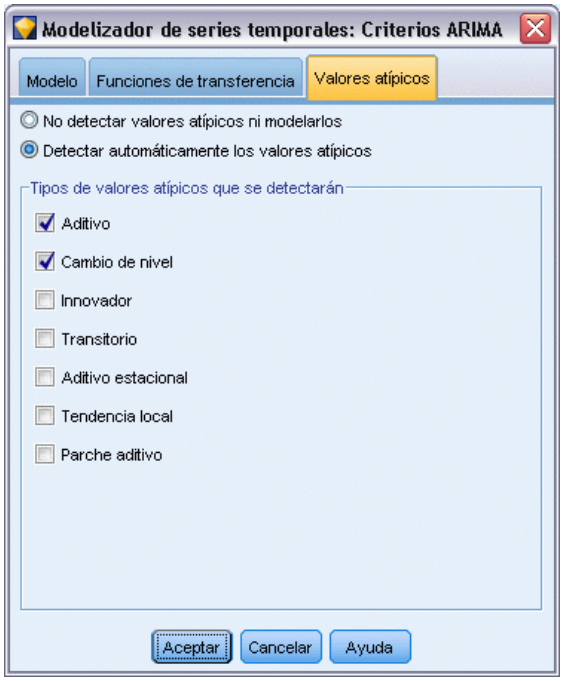
Retardo. Establecer un retardo provoca que la influencia del campo de entrada se retrase según el número de intervalos especificados. Por ejemplo, si el retardo se establece en 5, el valor de la variable de entrada en el tiempo t no afecta a las predicciones hasta que han transcurrido cinco períodos ($t + 5$).

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.

Tratamiento de valores atípicos

Figura 13-15
Cuadro de diálogo Criterios ARIMA, pestaña Valores atípicos



La pestaña Valores atípicos ofrece varias opciones para tratar los valores atípicos de los datos .

No detectar valores atípicos ni modelarlos. Por defecto, los valores atípicos no se detectan ni modelan. Seleccione esta opción para desactivar la detección o el modelado de valores atípicos.

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 valores atípicos que se muestran.

Tipos de valores atípicos que se detectarán. Seleccione los tipos de valores atípicos que desea detectar. Los tipos admitidos son:

- Aditivo (por defecto)
- Cambio de nivel (por defecto)
- Innovador
- Transitorio
- Aditivo estacional
- Tendencia local
- Parche aditivo

Si desea obtener más información, consulte el tema Valores atípicos el p. 450.

Generación de modelos de series temporales

Esta sección proporciona información general acerca de algunos aspectos de generación de modelos de series temporales:

- Generación de varios modelos
- Uso de los modelos de series temporales en predicciones
- Nueva estimación y predicción

El nugget del modelo generado se describe en un tema diferente. Si desea obtener más información, consulte el tema Nugget de modelo Serie temporal el p. 468.

Generación de varios modelos

El modelado de series temporales en IBM® SPSS® Modeler genera un único modelo (ARIMA o de suavizado exponencial) para cada campo objetivo. De esta manera, si tiene varios campos objetivo, SPSS Modeler genera varios modelos en una única operación, lo que le permitirá ahorrar tiempo y comparar la configuración de cada modelo.

Si desea comparar un modelo ARIMA y un modelo de suavizado exponencial para el mismo campo objetivo, puede realizar diferentes ejecuciones del nodo Serie temporal especificando de un modelo distinto cada vez.

Uso de los modelos de series temporales en predicciones

Una operación de generación de series temporales usa una serie concreta de casos ordenados, conocida como amplitud de estimación, para generar un modelo que pueda utilizarse para predecir valores futuros de las series. Este modelo contiene información acerca del período de tiempo usado, incluido el intervalo. Para poder realizar predicciones con este modelo, debe utilizarse la misma información de período temporal e intervalo con la misma serie para la variable objetivo y las variables predictoras.

Por ejemplo, supongamos que a principios de enero desea pronosticar las ventas mensuales del Producto 1 durante el primer trimestre del año. Para ello, genera un modelo utilizando los datos reales de ventas mensuales para el Producto 1 de enero a diciembre del año anterior (que denominaremos Año 1), configurando Intervalo de tiempo como “Meses”. A continuación, se puede utilizar el modelo para predecir las ventas del Producto 1 durante el primer trimestre del Año 2.

De hecho, puede pronosticar cualquier número de meses del futuro pero, como es lógico, cuanto más lejos en el tiempo se encuentren las predicciones, menos eficaz será el modelo. No obstante, no será posible predecir las tres primeras semanas del Año 2, ya que el intervalo usado para generar el modelo era “Meses”. Tampoco tendría sentido usar este modelo para predecir las ventas del Producto 2, ya que un modelo de series temporales sólo es relevante para los datos que se usaron para definirlo. Si desea obtener más información, consulte el tema Predicciones con el nodo Serie temporal en el capítulo 14 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

Nueva estimación y predicción

El período de estimación está codificado internamente en el modelo que se ha generado. Es decir, se ignorarán los valores situados fuera del período de estimación si aplica el modelo actual a datos nuevos. De esta forma, un modelo de serie temporal se debe volver a estimar cada vez que haya nuevos datos disponibles, en contraste con otros modelos de IBM® SPSS® Modeler, que se pueden volver a aplicar para obtener puntuaciones sin necesidad de modificarlos.

Para seguir con el ejemplo anterior, supongamos que, para principios de abril en el Año 2, tiene los datos de ventas mensuales reales para el período comprendido entre enero y marzo. No obstante, si vuelve a aplicar el modelo que generó a principios de enero, volverá a predecir de enero a marzo e ignorará los datos de ventas conocidos para ese período.

La solución es generar un nuevo modelo basado en los datos reales actualizados. Suponiendo que no cambia los parámetros de predicción, el nuevo modelo se puede utilizar para predecir los tres meses siguientes (de abril a junio). Si aún puede acceder a la ruta que se utilizó para generar el modelo original, puede simplemente sustituir la referencia al archivo fuente de esa ruta con una referencia al archivo que contiene los datos actualizados y volver a ejecutar la ruta para generar el nuevo modelo. No obstante, si sólo tiene el modelo original guardado en un archivo, puede seguir utilizándolo para generar un nodo Serie temporal que podrá añadir después a una nueva ruta con una referencia al archivo fuente actualizado. La ejecución de esta nueva ruta generará, a continuación, el nuevo modelo necesario, siempre que en esta nueva ruta antes del nodo Serie temporal haya un nodo Intervalos de tiempo en el que el intervalo esté configurado como “Meses”. Si desea obtener más información, consulte el tema Nueva aplicación de modelos de series temporales en el capítulo 14 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

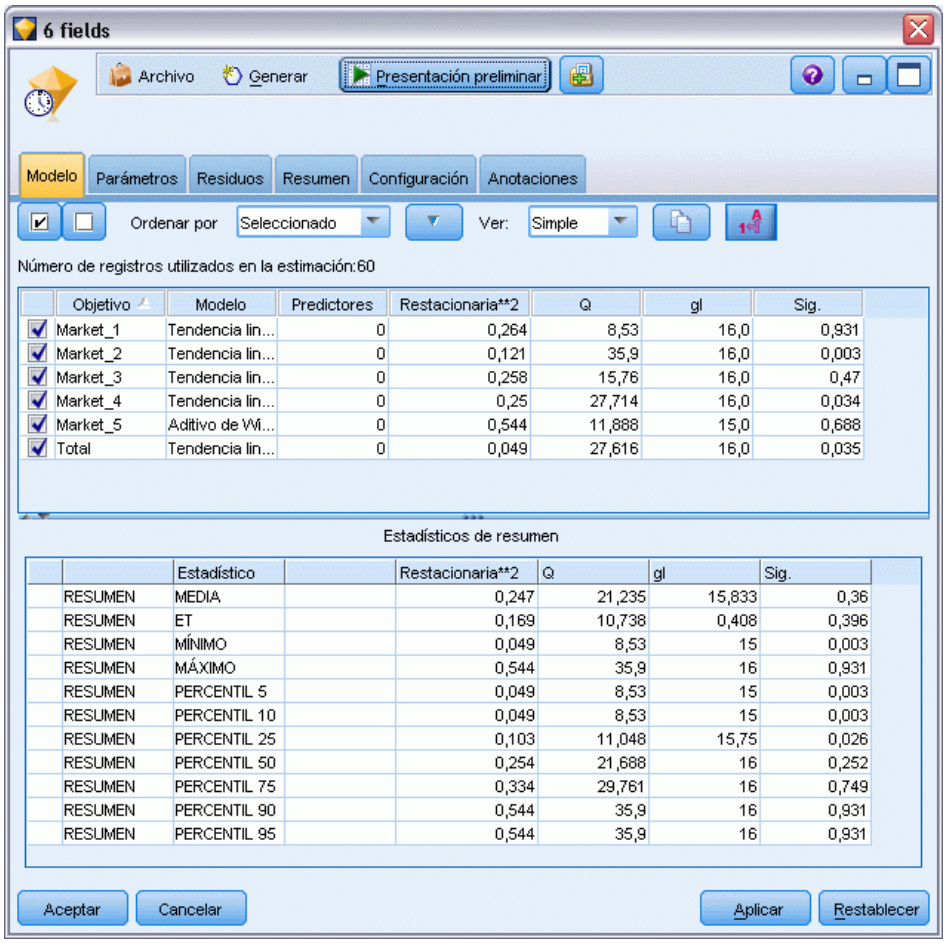
Nugget de modelo Serie temporal

La operación de modelado de series temporales crea varios campos nuevos con el prefijo $TS- como se indica a continuación:

$TS-*nombrecol*	Valor pronosticado por el modelo para cada serie objetivo.
$TSLCI-*nombrecol*	Los intervalos de confianza más bajos para cada serie pronosticada.*
$TSUCI-*nombrecol*	Los intervalos de confianza más altos para cada serie pronosticada.*
$TSNR-*nombrecol*	Valor de residuo de ruido para cada columna de datos del modelo generado.*
$TS-Total	Total de los valores de $TS-*nombrecol* de esta fila.
$TSLCI-Total	Total de los valores de $TSLCI-*nombrecol* de esta fila.*
$TSUCI-Total	Total de los valores de $TSUCI-*nombrecol* de esta fila.*
$TSNR-Total	Total de los valores de $TSNR-*nombrecol* de esta fila.*

* La visibilidad de estos campos (por ejemplo, en los resultados del nodo Tabla conectado) depende de las opciones de la pestaña Configuración del nugget de modelo Serie temporal. Si desea obtener más información, consulte el tema Configuración del modelo de serie temporal el p. 475.

Figura 13-16
Nugget de modelo Serie temporal, pestaña Modelo



El nugget de modelo Serie temporal muestra detalles de diferentes modelos seleccionados para cada entrada de serie en el nodo de generación Serie temporal. Se pueden introducir varias series (como datos relacionados con las gamas de productos, regiones o almacenes), y se genera un modelo independiente para cada serie objetivo. Por ejemplo, si se descubre que los ingresos de la región oriental se ajustan a un modelo ARIMA, pero la región occidental se ajusta sólo a una simple media móvil, cada región se puntuará con el modelo adecuado.

Para cada modelo generado, los resultados por defecto muestran el tipo de modelo, el número de predictores especificados y la medida de bondad de ajuste (*R* cuadrado estacionaria es la medida por defecto). Si ha especificado métodos de valores atípicos, hay una columna donde aparece el número de valores atípicos detectados. Los resultados por defecto también incluyen columnas para Q de Ljung-Box, grados de libertad y valores de significación.

También puede seleccionar el resultado avanzado, que mostrará estas otras columnas:

- R cuadrado
- RMSE (raíz del error cuadrático promedio)
- MAPE (error absoluto porcentual promedio)

- MAE (error absoluto promedio)
- MaxAPE (error absoluto máximo porcentual)
- MaxAE (error absoluto máximo)
- Norm. BIC (criterio de información bayesiano normalizado)

Generar. Le permite volver a generar un nodo de modelado Serie temporal en la ruta o un nugget de modelo en la paleta.

- **Generar nodo de modelado** Coloca un nodo de modelado Serie temporal en una ruta con la configuración usada para crear este conjunto de modelos. Esto resultará útil, por ejemplo, si tiene una ruta en la que desee usar esta configuración de modelo y ya no tiene el nodo de modelado que se utilizó para generarla.
- **Modelo a paleta** Sitúa un nugget de modelo con todos los objetivos en el administrador de modelos.

Modelo

Figura 13-17
Botones Seleccionar todos y Anular la selección de todos

Casillas de verificación. Seleccione los modelos que desea utilizar en la puntuación. Todas las casillas están activadas por defecto. Los botones Seleccionar todos y Anular la selección de todos afectan a todas las casillas en una única operación.

Ordenar por. Le permite ordenar las filas de resultados en orden ascendente o descendente de una columna especificada de la presentación. La opción "Seleccionado" ordena el resultado según una o más filas seleccionadas por casillas de verificación. Esto resultaría útil, por ejemplo, para forzar que los campos objetivo denominados "Mercado_1" a "Mercado_9" aparezcan antes de "Mercado_10", ya que el orden de clasificación por defecto muestra "Mercado_10" justo después de "Mercado_1."

Ver. La vista por defecto (simple) muestra el conjunto básico de columnas de resultados. La opción Avanzado muestra columnas adicionales para medidas de bondad de ajuste.

Número de registros utilizados en la estimación. Número de filas existentes en el archivo de datos fuente original.

Destino. Campo o campos identificados como campos objetivo (aquellos con un papel de *Objetivo*) en el nodo Tipo.

Modelo. Tipo de modelo usado para este campo objetivo.

Predictores. El número de predictores (aquellos con un papel de *Entrada*) utilizados para este campo objetivo.

Valores atípicos. Esta columna sólo aparece si ha solicitado (en el modelizador experto o los criterios ARIMA) la detección automática de valores atípicos. El valor mostrado es el número de valores atípicos detectados.

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. R-cuadrado puede ser negativa con un rango desde menos infinito 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.

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.

RMSE. Raíz del error cuadrático promedio. 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.

MAPE. Error absoluto porcentual promedio. 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.

MAE. Error absoluto promedio. Mide la desviación de la serie del nivel pronosticado por el modelo. El MAE se informa en las unidades originales de la serie.

MaxAPE. Error absoluto máximo porcentual. El mayor error pronosticado, expresado como porcentaje. Esta medida es útil para imaginar el peor escenario de un caso en las predicciones.

MaxAE. Error absoluto máximo. El mayor error pronosticado, 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 la predicción. El error absoluto máximo y el error absoluto máximo porcentual 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 máximo porcentual corresponderá al valor de la serie menor.

BIC normalizado. Criterio de información Bayesiano normalizado. 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.

Q. Estadístico Q de Ljung-Box. Prueba de la aleatoriedad de los errores residuales de este modelo.

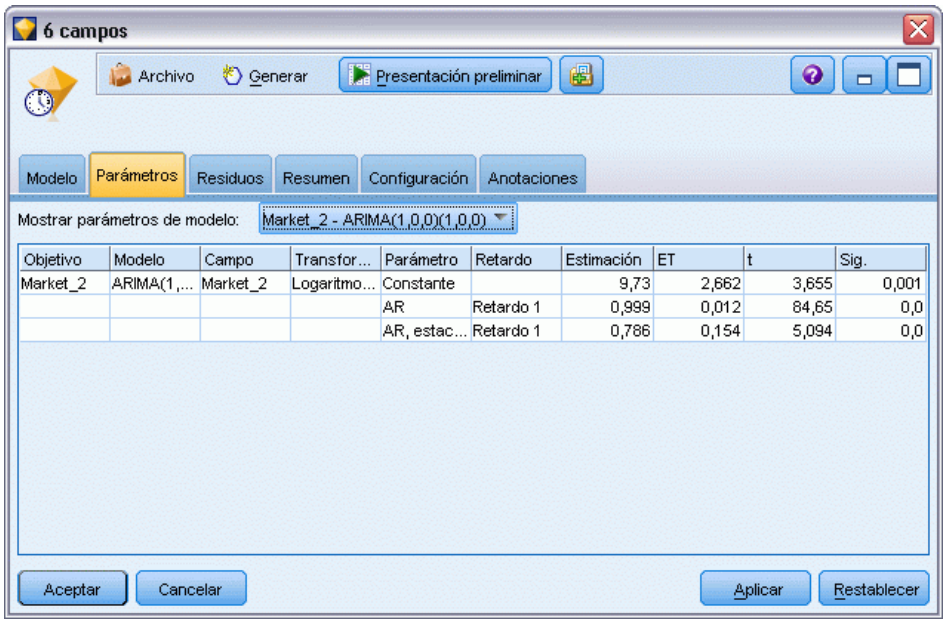
gl. Grados de libertad. Número de parámetros del modelo que pueden variar al calcular un objetivo específico.

Sig. Valor de significación del estadístico de Ljung-Box. Un valor de significación inferior a 0,05 indica que los errores residuales no son aleatorios.

Estadísticos de resumen. Esta sección contiene diferentes estadísticos de resumen para las distintas columnas, incluidos los valores de la media, mínimo, máximo y los percentiles.

Parámetros del modelo Serie temporal

Figura 13-18
Modelo Serie temporal, pestaña Parámetros



La pestaña Parámetros enumera los detalles de los distintos parámetros que se utilizaron para crear un modelo seleccionado.

Mostrar parámetros de modelo. Seleccione el modelo para el que desea mostrar los detalles de parámetros.

Destino. Nombre del campo objetivo (con el papel *Objetivo*) pronosticado por este modelo.

Modelo. Tipo de modelo usado para este campo objetivo.

Campo (sólo modelos ARIMA). Contiene una entrada para cada una de las variables utilizadas en el modelo, con el objetivo en primer lugar, seguido por los predictores, si los hubiera.

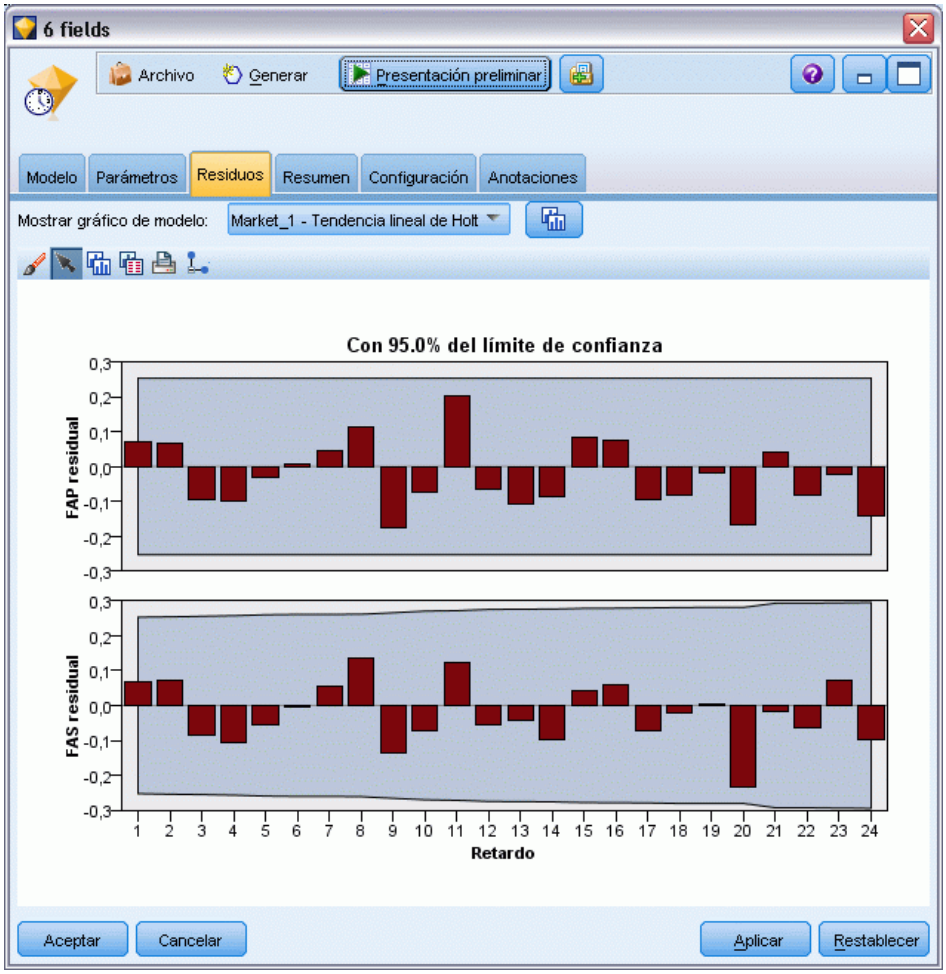
Transformación. Indica qué tipo de transformación se ha especificado para este campo antes de generar el modelo, si hay alguna transformación.

Parámetro. Parámetro del modelo para el que se muestran los siguientes detalles:

- **Retardo (sólo modelos ARIMA).** Indica los retardos, en caso de que los haya, que se tienen en cuenta para este parámetro en el modelo.
- **Estimación.** Estimación del parámetro. Este valor se utiliza para calcular el valor pronosticado y los intervalos de confianza para el campo objetivo.
- **SE.** Error estándar de la estimación del parámetro.
- **t.** Valor de la estimación del parámetro dividido entre el error estándar.
- **Sig.** Nivel de significación para la estimación del parámetro. Los valores por encima de 0,05 no se consideran estadísticamente significativos.

Residuos del modelo de serie temporal

Figura 13-19
Modelo Serie temporal, pestaña Residuos, representación de FAS y FAP

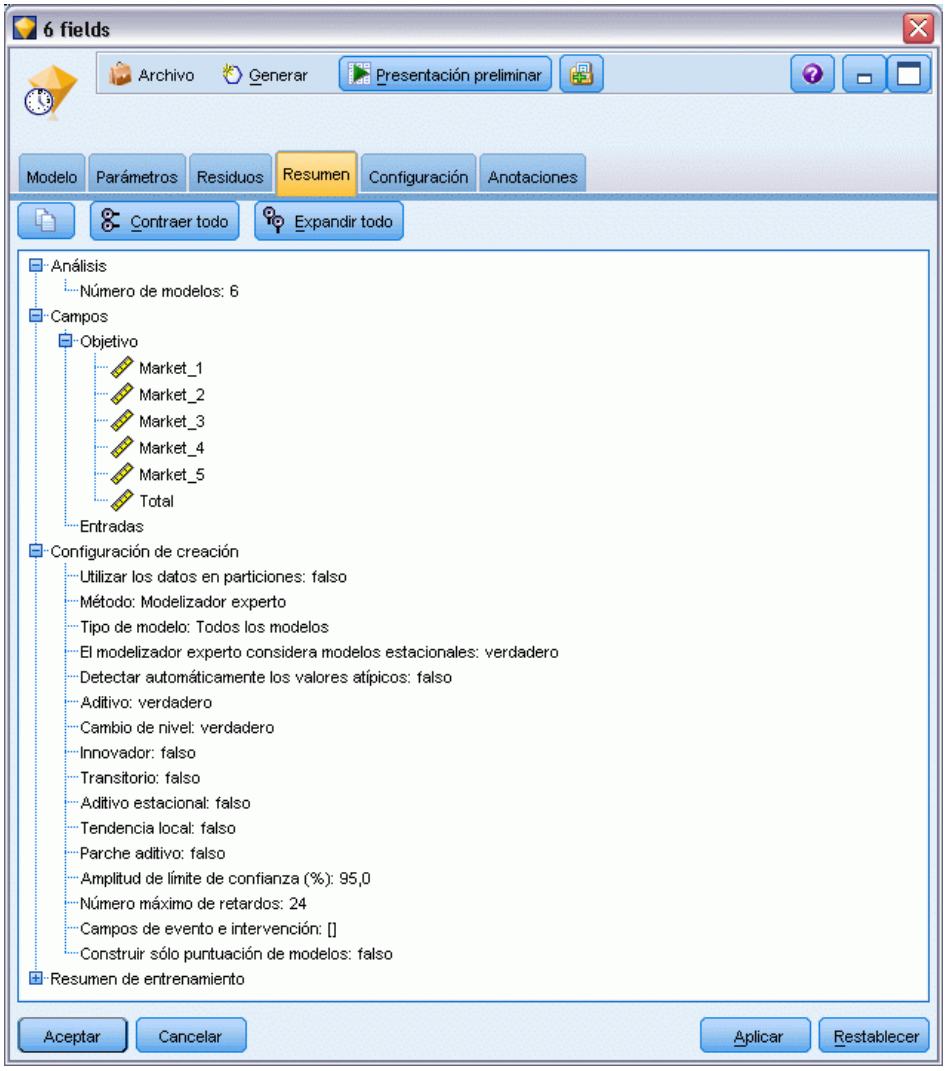


La pestaña Residuos muestra la función de autocorrelación (FAS) y la función de autocorrelación parcial (FAP) de los residuos (diferencia entre los valores esperados y los reales) de cada modelo generado. Si desea obtener más información, consulte el tema Funciones de autocorrelación y autocorrelación parcial el p. 451.

Mostrar gráfico de modelo. Seleccione el modelo cuyas FAS y FAP de residuos desea mostrar.

Resumen del modelo de serie temporal

Figura 13-20
Modelo Serie temporal, pestaña Resumen



La ficha 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 ficha 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.

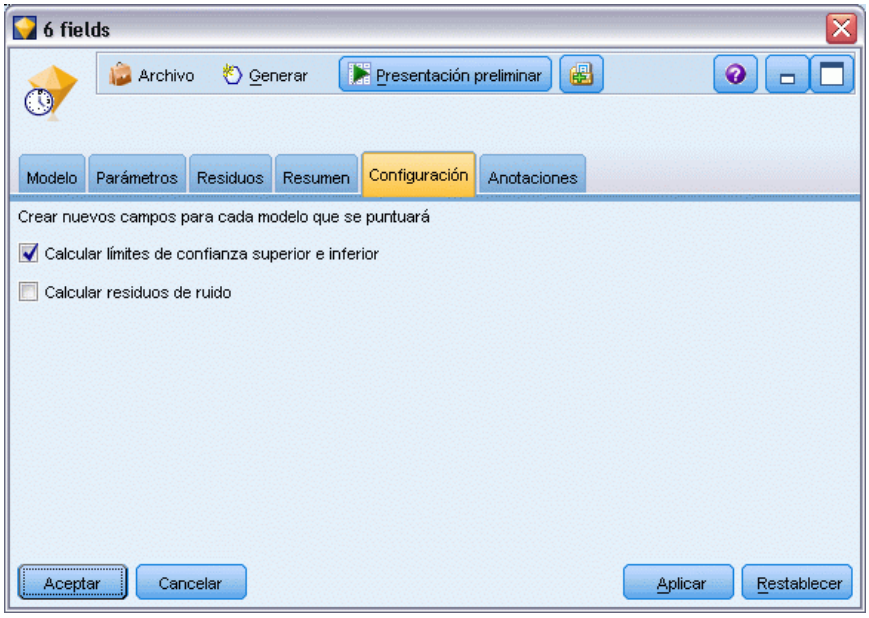
Campos. Enumera los campos utilizados como objetivo y entradas en la generación del modelo.

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.

Configuración del modelo de serie temporal

Figura 13-21
Modelo Serie temporal, pestaña Configuración



La pestaña Configuración le permite especificar los campos adicionales que se van a crear mediante la operación de modelado.

Crear nuevos campos para cada modelo que se puntuará. Permite especificar los nuevos campos que se crearán para cada modelo que se puntuará.

- **Calcular límites de confianza superior e inferior.** Si está seleccionada, se crearán nuevos campos (con los prefijos por defecto $TSLCI- y $TSUCI-) para los intervalos de confianza superior e inferior, respectivamente, de cada campo objetivo, además de los totales de estos valores.
- **Calcular residuos de ruido.** Si está activada, creará un nuevo campo (con el prefijo por defecto $TSNR-) para los residuos del modelo de cada campo objetivo, junto con un total de estos valores.

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.

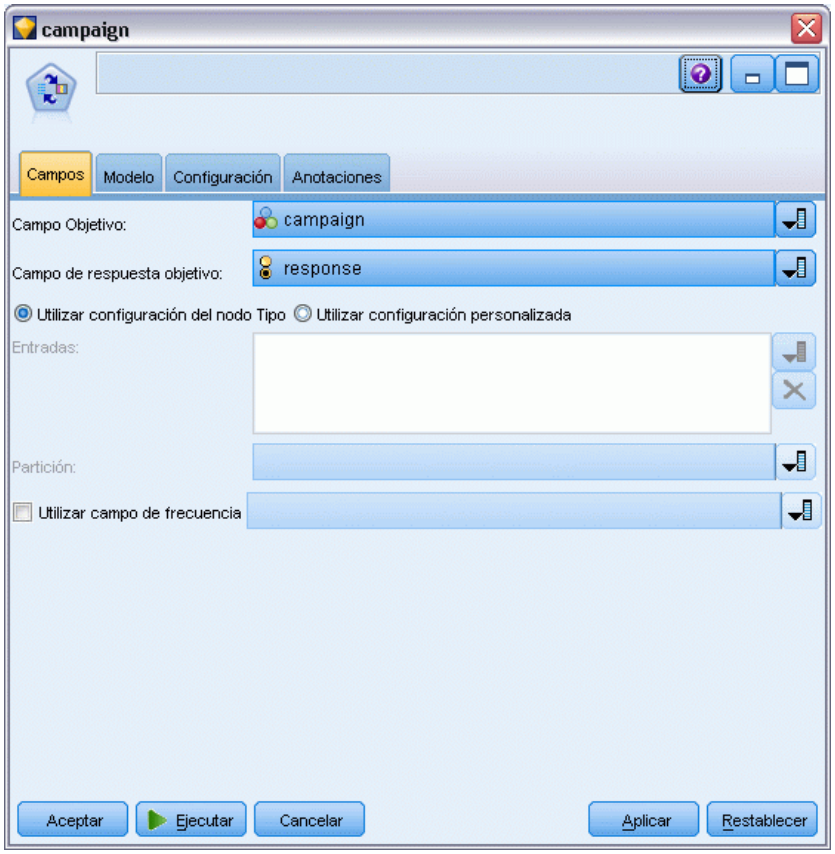
El modelo se puede generar inicialmente con un conjunto de datos pequeño con ofertas realizadas aleatoriamente y las respuestas a éstas. A medida que crece el conjunto de datos, el modelo se puede actualizar y, de ese modo, aumenta su capacidad para predecir las ofertas más adecuadas para los clientes 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 analista 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. Si desea obtener más información, consulte el tema Realización de ofertas a clientes (Autoaprendizaje) en el capítulo 16 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

Opciones de los campos del nodo SLRM

Figura 14-1
Cuadro de diálogo del nodo SLRM, pestaña Campos



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.

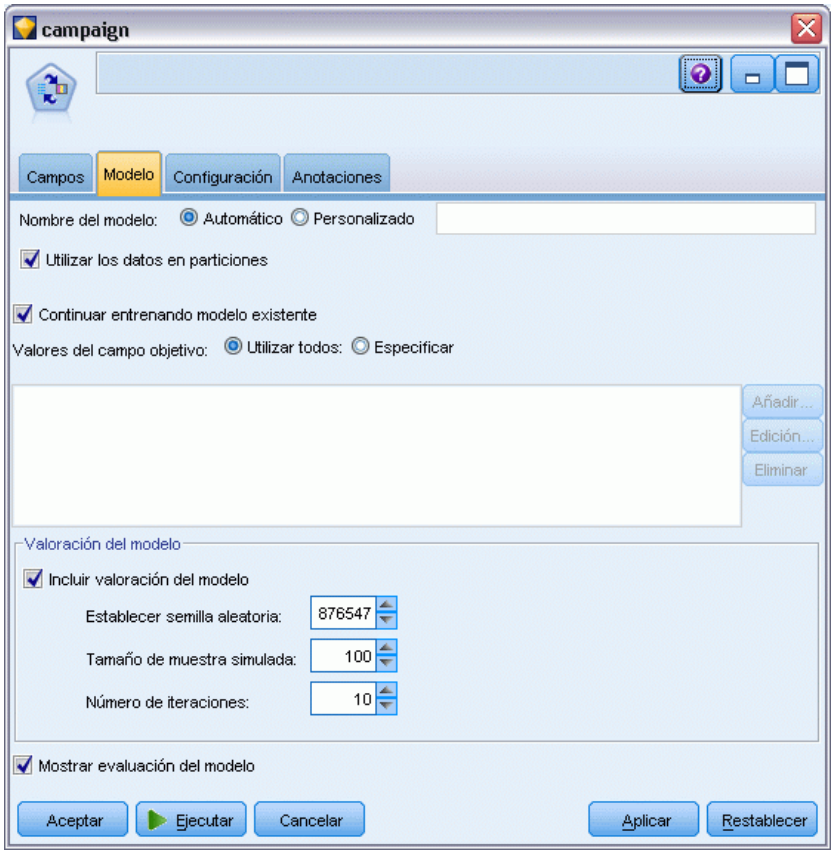
Nota: este campo debe ser de marcas. El valor para verdadero de la marca indica la aceptación de la oferta y el valor para falso el rechazo de la oferta.

Los campos restantes de este cuadro de diálogo son los que se utilizan normalmente en IBM® SPSS® Modeler. Si desea obtener más información, consulte el tema Opciones de los campos del nodo de modelado en el capítulo 3 el p. 38.

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

Figura 14-2
Cuadro de diálogo del nodo SLRM, pestaña 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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Continuar entrenando modelo existente. Por defecto, 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 Por defecto, 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 pronosticará 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.

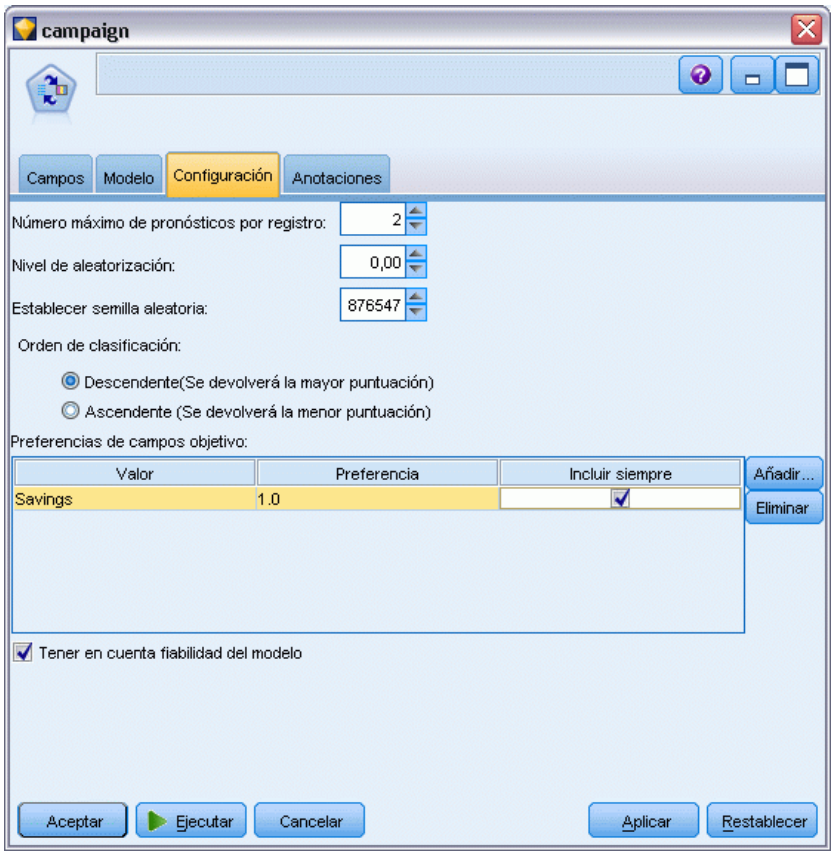
- **Incluir valoración del modelo.** Seleccione esta casilla para crear gráficos que muestren la precisión pronosticada del modelo para cada oferta seleccionada.
- **Establecer semilla aleatoria.** Cuando se estima la precisión de un modelo a partir de un porcentaje aleatorio, esta opción permite duplicar 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 por defecto es 100.
- **Número de iteraciones.** Permite detener la generación de la valoración del modelo tras un número de iteraciones especificado. Especifique el número máximo de iteraciones, el valor por defecto 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

Figura 14-3
Cuadro de diálogo del nodo SLRM, pestaña Configuración



Las opciones de configuración del nodo permiten ajustar el proceso de generación de modelos.

Número máximo de pronósticos por registro. Esta opción le permite limitar el número de predicciones realizadas para cada registro del conjunto de datos. El valor por defecto 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 por defecto 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. Si desea obtener más información, consulte el tema Nodo Ordenar en el capítulo 3 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

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 por defecto es 0.0.
- **Incluir siempre.** Para asegurarse de que se incluye siempre una oferta específica en las predicciones, active esta casilla.

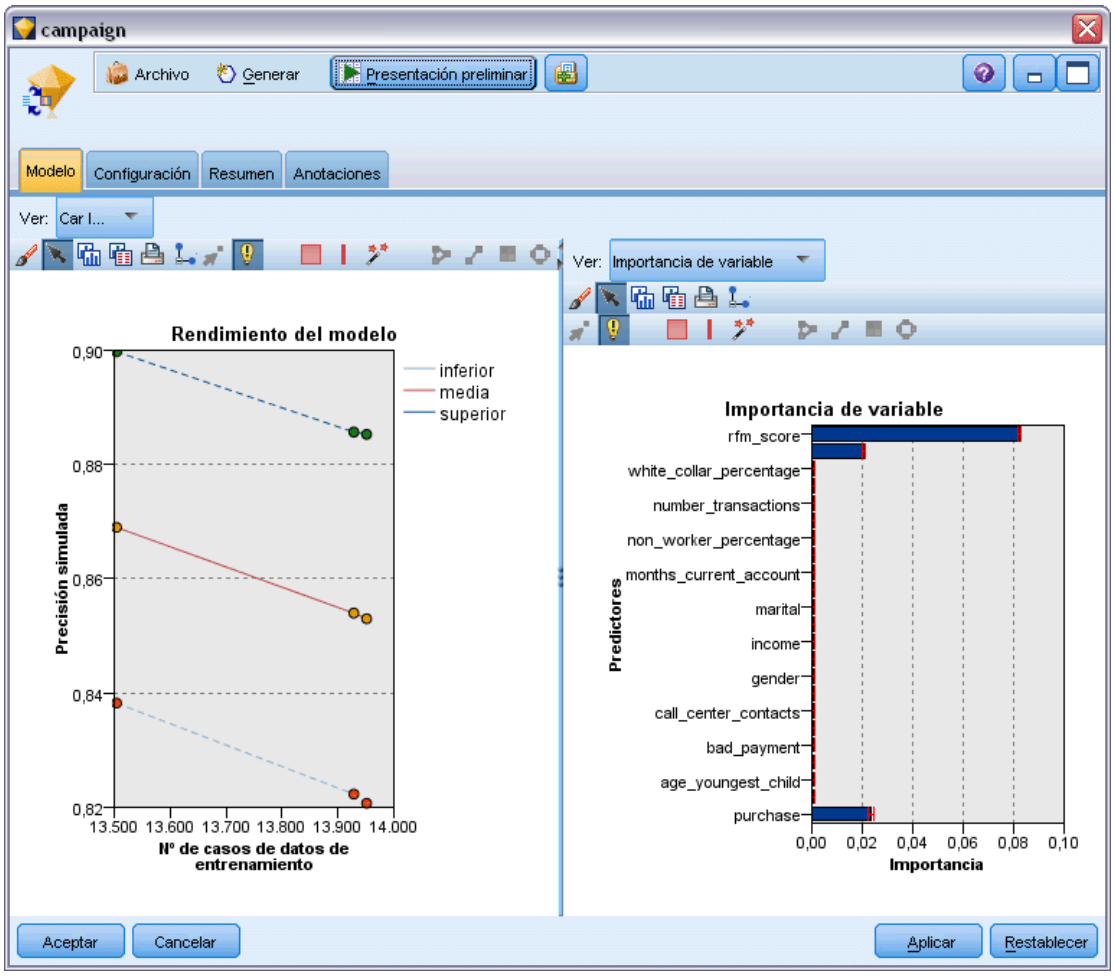
  Nota: si la Preferencia se establece en 0,0, se ignorará la configuración de Incluir siempre.

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.

Figura 14-4
Visualización del nugget de modelo SLRM



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.

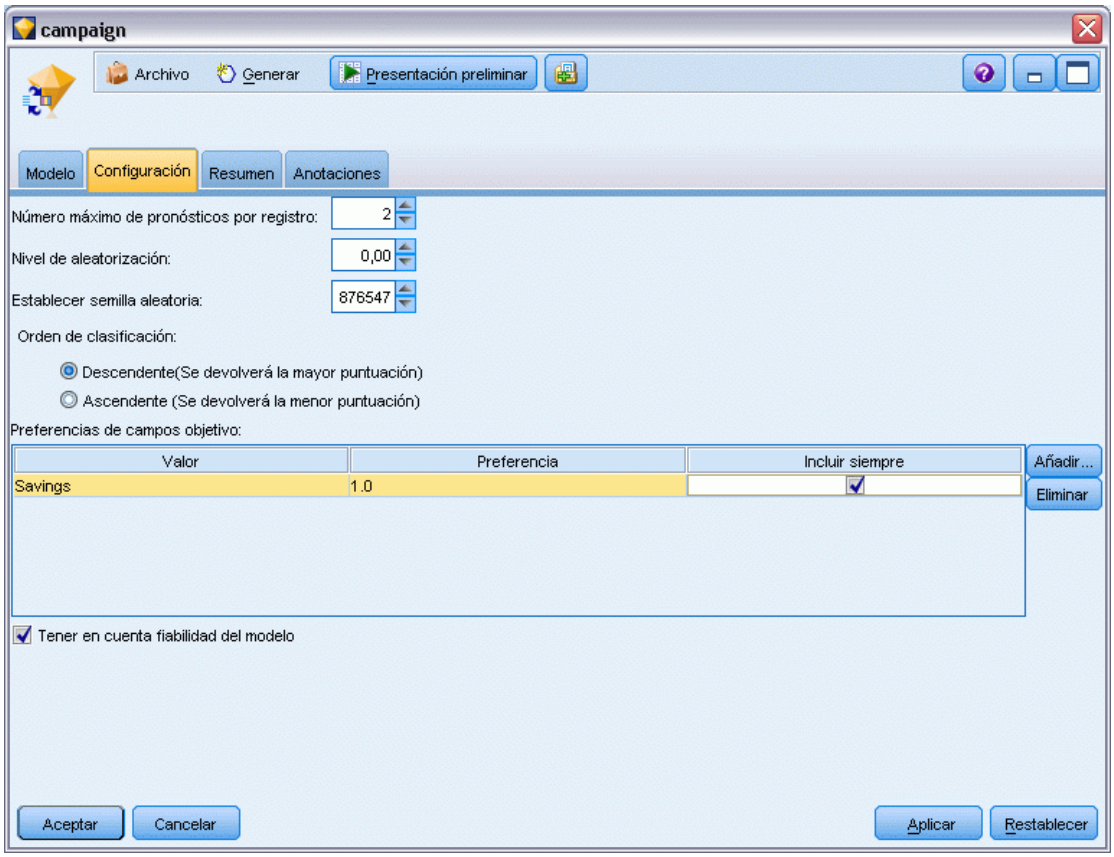
- **Ver.** Seleccione si desea visualizar los detalles de Asociación con respuesta o Importancia de variable.
- **Asociación con respuesta.** Muestra la asociación (correlación) de cada predictor con la variable objetivo.
- **Importancia del predictor.** Indica la importancia relativa de cada predictor 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. Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

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.

Figura 14-5
Cuadro de diálogo Nugget de modelo SLRM, pestaña Configuración



Número máximo de pronósticos por registro. Esta opción le permite limitar el número de predicciones realizadas para cada registro del conjunto de datos. El valor por defecto 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 por defecto 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. Si desea obtener más información, consulte el tema Nodo Ordenar en el capítulo 3 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

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 por defecto 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.

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 los pronósticos 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).

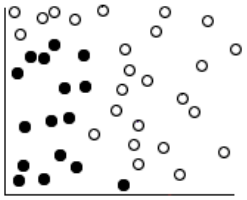
SVM tiene aplicaciones en multitud de disciplinas, incluyendo la gestión de relaciones con los clientes (CRM), el reconocimiento facial y de otras imágenes, bioinformática, extracción de conceptos de minería de texto, detección de intrusiones, pronóstico de estructura de proteínas y reconocimiento de la voz.

Funcionamiento de SVM

SVM funciona asignando datos a un espacio de función 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 pronosticar el grupo al que pertenece el nuevo registro.

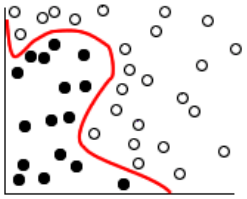
Por ejemplo, imagine la siguiente figura, en la que los puntos de datos corresponden a dos categorías diferentes:

Figura 15-1
Conjunto de datos original



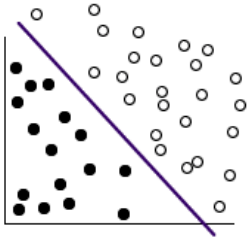
Las dos categorías se pueden separar con una curva:

Figura 15-2
Datos con un separador añadido



Tras la transformación, el límite entre las dos categorías se puede definir por un hiperplano:

Figura 15-3
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:

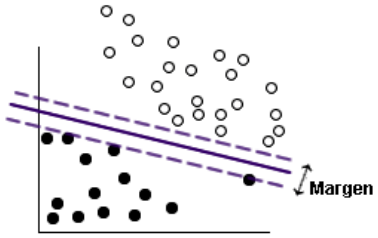
- Lineal
- Polinómico
- Función de base radial (RBF)
- Sigmoide

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, una clasificación del modelo SVM también encuentra líneas marginales que definen el espacio entre las dos categorías:

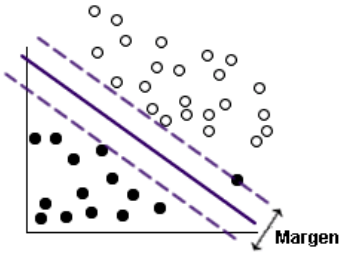
Figura 15-4
Datos con un modelo preliminar



Los puntos de datos que están en los márgenes se conocen como **vectores de soporte**.

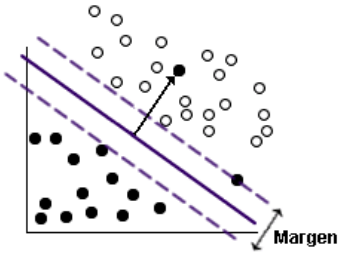
Cuanto más amplio sea el margen entre las dos categorías, mejor será el modelo para pronosticar la categoría de nuevos registros. En el ejemplo anterior, el margen no es muy amplio y el modelo se conoce como **ajuste en exceso**. Se puede aceptar una pequeña cantidad de clasificación errónea para ampliar el margen, por ejemplo:

Figura 15-5
Datos con un modelo mejorado



En algunos casos, la separación lineal es más difícil, por ejemplo:

Figura 15-6
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 por defecto 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.

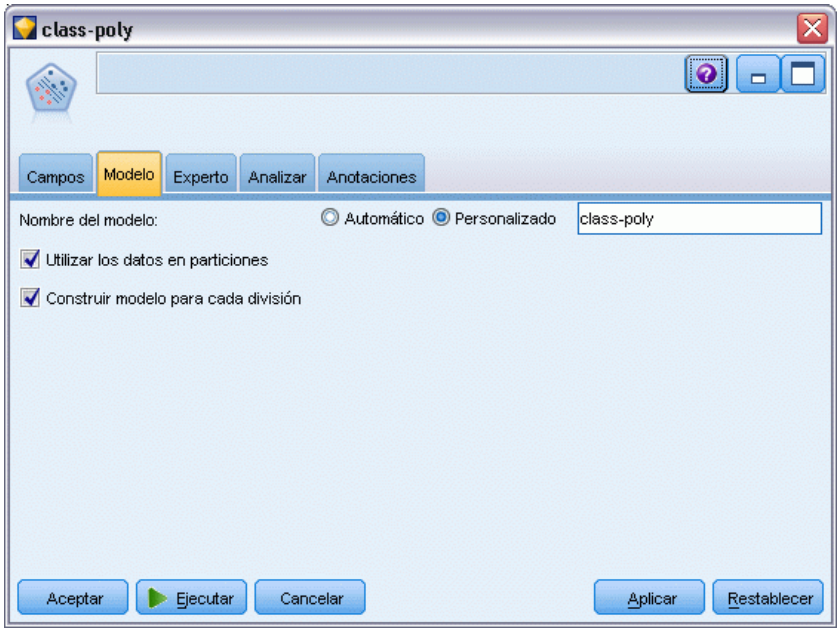
Cuando haya creado el modelo, podrá:

- 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. Si desea obtener más información, consulte el tema Clasificación de muestras de células (SVM) en el capítulo 25 en *Guía de aplicaciones de IBM SPSS Modeler 15*.

Opciones de modelo del nodo SVM

Figura 15-7
Opciones del modelo del nodo SVM



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.

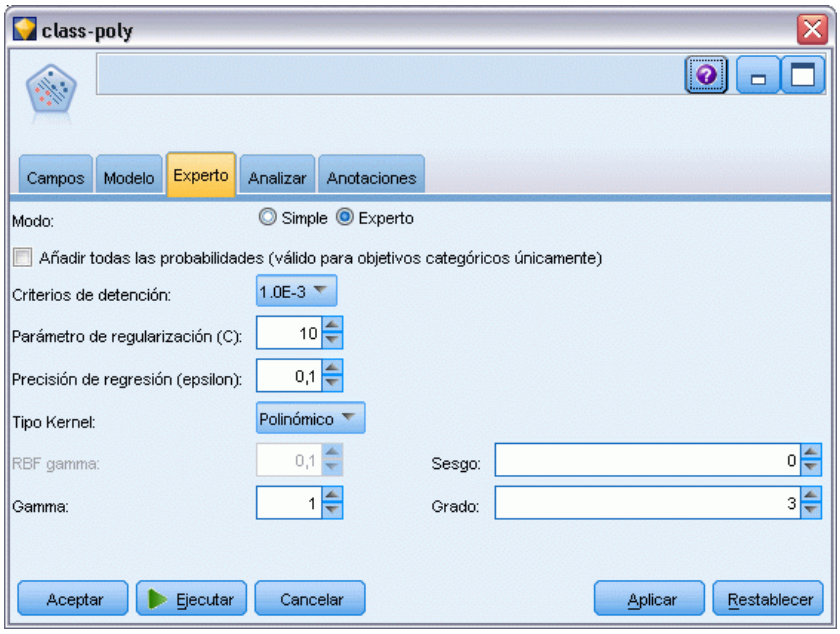
Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Crear modelos divididos. Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como modelos divididos. Si desea obtener más información, consulte el tema Generación de modelos divididos en el capítulo 3 el p. 32.

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.

Figura 15-8
Opciones de experto de nodo SVM



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 pronosticado de campos objetivo nominal o marca. El ajuste de esta casilla de verificación determina el estado por defecto 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 oscilan entre 1.0E–1 a 1.0E–6; el valor por defecto es 1.0E–3. Si reduce el valor obtendrá un modelo más preciso, pero deberá realizar un entrenamiento más largo.

Parámetro de regularización (C). Controla el equilibrio entre la maximización del margen y la minimización del término de error de entrenamiento. El valor debe oscilar entre 1 y 10; el valor por defecto es 10. Si aumenta el valor aumentará la precisión de la clasificación (o se reduce el error de regresión) de los datos de entrenamiento, aunque 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 por defecto 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 por defecto 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 asignado. 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 15-1
El nivel de medición del campo objetivo es "Nominal" o "Marca".

Nombre del campo nuevo	Descripción
$S-*nombrede-campo*	Valor pronosticado del campo objetivo.
$SP-*nombrede-campo*	Probabilidad del valor pronosticado.
$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 brutas (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. Si desea obtener más información, consulte el tema Opciones de análisis del nodo de modelado en el capítulo 3 el p. 42.
$SAP-*valor*	

Tabla 15-2
El nivel de medición del campo objetivo es "Continuo"

Nombre del campo nuevo	Descripción
$S-*nombrede-campo*	Valor pronosticado del campo objetivo.

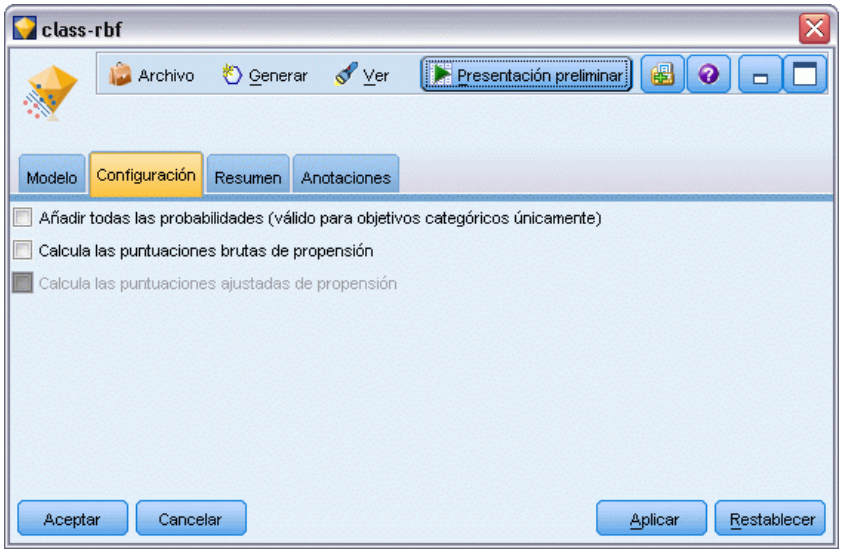
Importancia del predictor

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 predictoren la pestaña Analizar antes de generar el modelo. Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

Nota: La importancia del predictor SVM puede requerir más tiempo de cálculo que otros tipos de modelos y no está seleccionada en la pestaña Analizar por defecto. Si selecciona esta opción se reducirá el rendimiento, especialmente con conjuntos de datos más grandes.

Configuración de modelo SVM

Figura 15-9
Modelo SVM, pestaña Configuración



La ficha 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 pronosticado y su probabilidad para campos objetivo nominal o marca.

El valor por defecto de esta casilla de verificación está determinado por la casilla de verificación correspondiente en el nodo de modelado.

Calcular puntuaciones brutas de propensión. En el caso de modelos con un objetivo de marca (que devuelve un pronóstico 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 pronóstico y confianza que pueden generarse durante la puntuación.

Calcular puntuaciones de propensión ajustada. Las puntuaciones brutas de propensión 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 la ficha Analizar antes de generar el modelo.

Capítulo
16

Modelos de vecinos más próximos

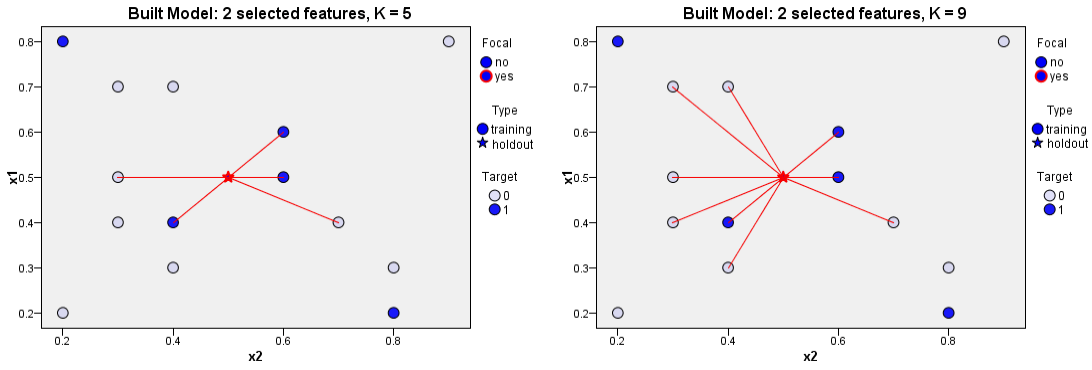
Nodo KNN

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 similares están cercanos entre sí y los casos no similares están distantes entre sí. Además, la distancia entre dos casos es una medida de sus diferencias.

Los casos muy cercanos a otros se denominan "vecinos". Cuando se presenta un nuevo caso (reserva), se calcula su distancia desde cada caso del modelo. Las clasificaciones de la mayoría de casos similares (los vecinos más próximos) se anotan y el nuevo caso se coloca en la categoría que contiene el mayor número de vecinos más próximos.

Puede especificar el número de vecinos más próximos que se van a examinar; este valor se denomina *k*. Las imágenes muestran cómo se clasifica un nuevo caso utilizando dos valores diferentes de *k*. Si *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, si *k* = 9, el nuevo caso se coloca en la categoría *0* porque una mayoría de los vecinos más próximos pertenecen a esa categoría *0*.

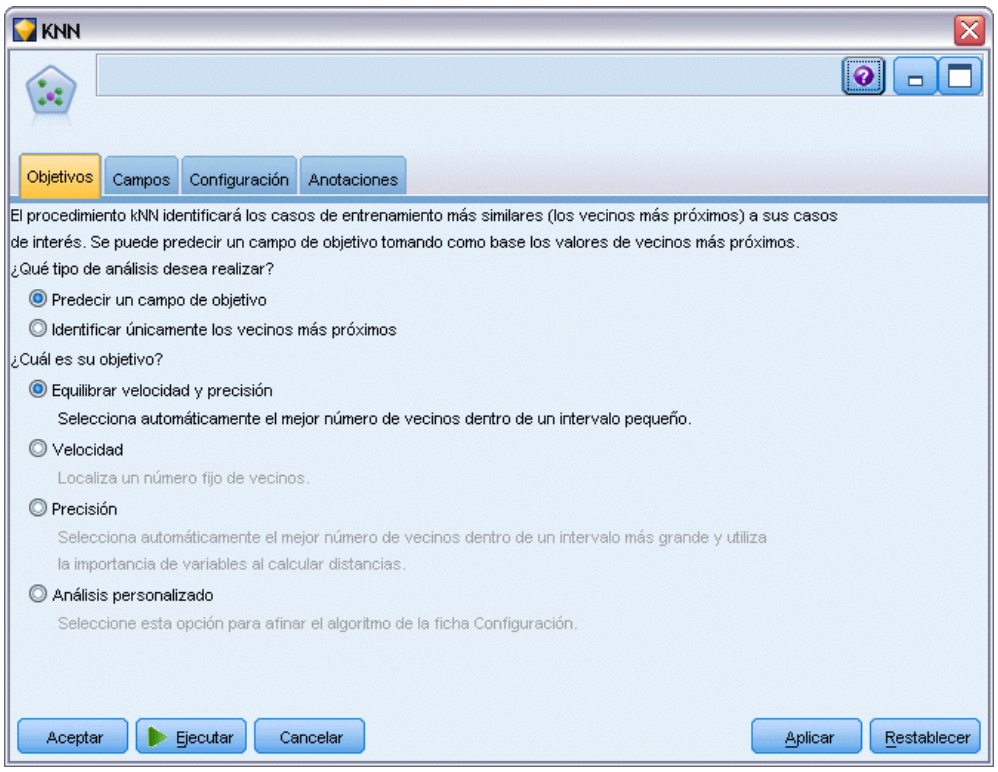
Figura 16-1
Los efectos de modificar k en la clasificación



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 pronosticado del nuevo caso.

Opciones de objetivos del nodo KNN

Figura 16-2
Opciones de objetivos del nodo KNN



En la pestaña Objetivos podrá seleccionar si desea generar un modelo que pronostique 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 ejecutar?

Pronosticar un campo de destino. Seleccione esta opción si desea pronosticar 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 pronósticos.

¿Cuál es su 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 pronosticar 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 de forma lineal 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

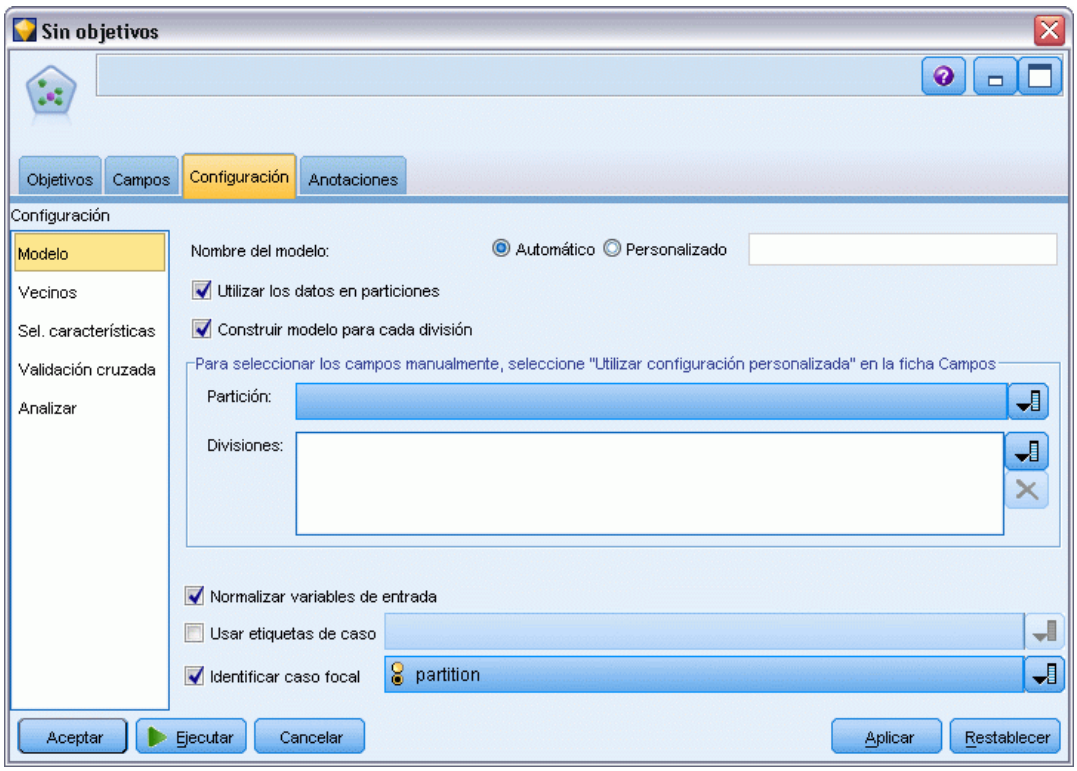
e introduzca el tamaño nuevo en el campo Memoria máxima. Los cambios realizados en el cuadro de diálogo Opciones de sistema no surtirán efecto hasta que no reinicie SPSS Modeler.

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

Figura 16-3
Opciones de modelo de nodo KNN



El panel Modelo proporcionan las opciones que controlan cómo se va a generar el modelo, por ejemplo, si se utilizarán modelos de partición o de división, 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.

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.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Crear modelos divididos. Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como modelos divididos. Si desea obtener más información, consulte el tema Generación de modelos divididos en el capítulo 3 el p. 32.

Para seleccionar campos manualmente... Por defecto, el nodo utiliza la partición y divide la configuración del campo (si tuviera) desde el nodo Tipo, pero puede sustituir esa configuración aquí. Para activar los campos Partición y Divisiones, seleccione la ficha 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 sobre la adecuación 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 ficha Campos en todos los nodos de modelado que usen la partición. (Si sólo hay una partición, se usará automáticamente siempre que se active la partición.) Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*. Debe tener en cuenta que al aplicar la partición seleccionada en su análisis, también debe activar la partición en la ficha 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 dividido. Se trata de una acción similar a establecer el papel del campo en *Segmentar* en un nodo Tipo. Sólo puede diseñar campos de tipo Marca, Nominal u Ordinal como campos de división. Los campos seleccionados como campos divididos no se pueden utilizar como campos de destino, entrada, partición, frecuencia o ponderación. Si desea obtener más información, consulte el tema Generación de modelos divididos en el capítulo 3 el p. 32.

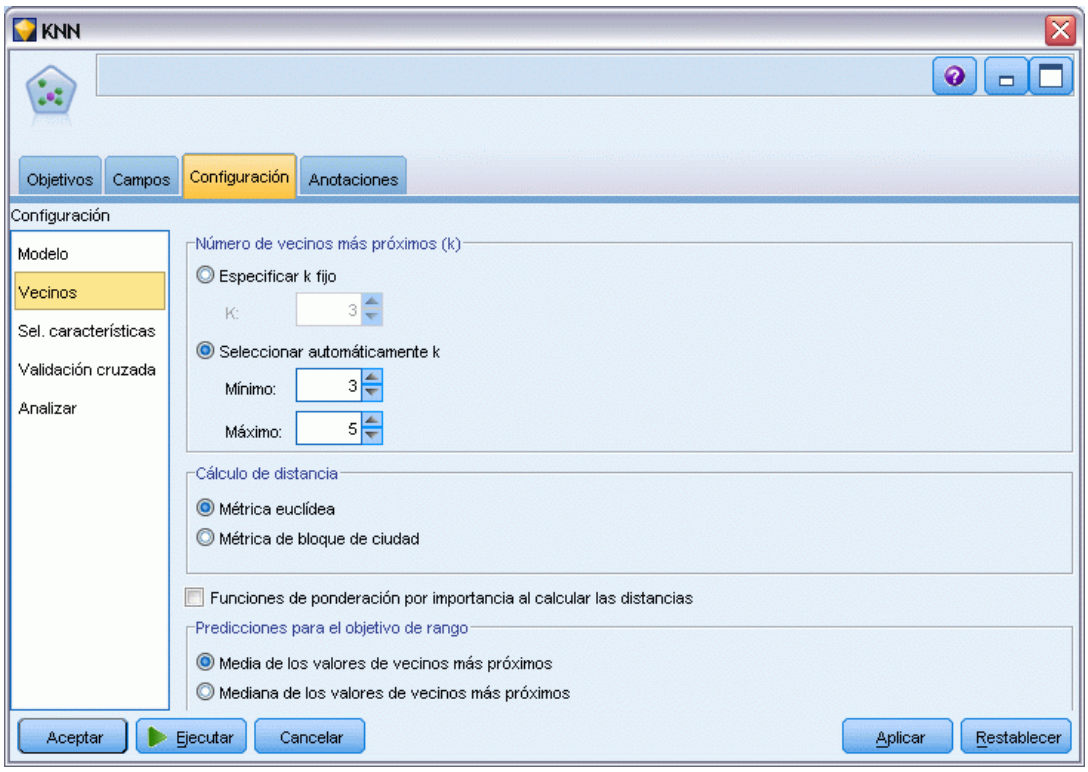
Normalizar entradas de rango. Seleccione esta casilla para normalizar los valores de campos de entrada continuos. Las funciones normalizadas tienen el mismo rango de valores, que pueden mejorar el rendimiento del algoritmo de cálculo. Se utilizará la normalización ajustada [2*(x−mín.)/(máx.−mín.)]−1. Los valores normalizados ajustados quedan comprendidos 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

Figura 16-4
Opciones de vecinos de nodo KNN



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 pronosticar 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 funciones se solicita en el panel Selección de características:

  Si la selección de funciones está activada, ésta se realizará para cada valor de *k* en el rango solicitado, y se seleccionará la *k* y el conjunto de funciones compañero con la menor tasa de error (o el menor error cuadrático si el destino es continuo).

  Si la selección de funciones 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.

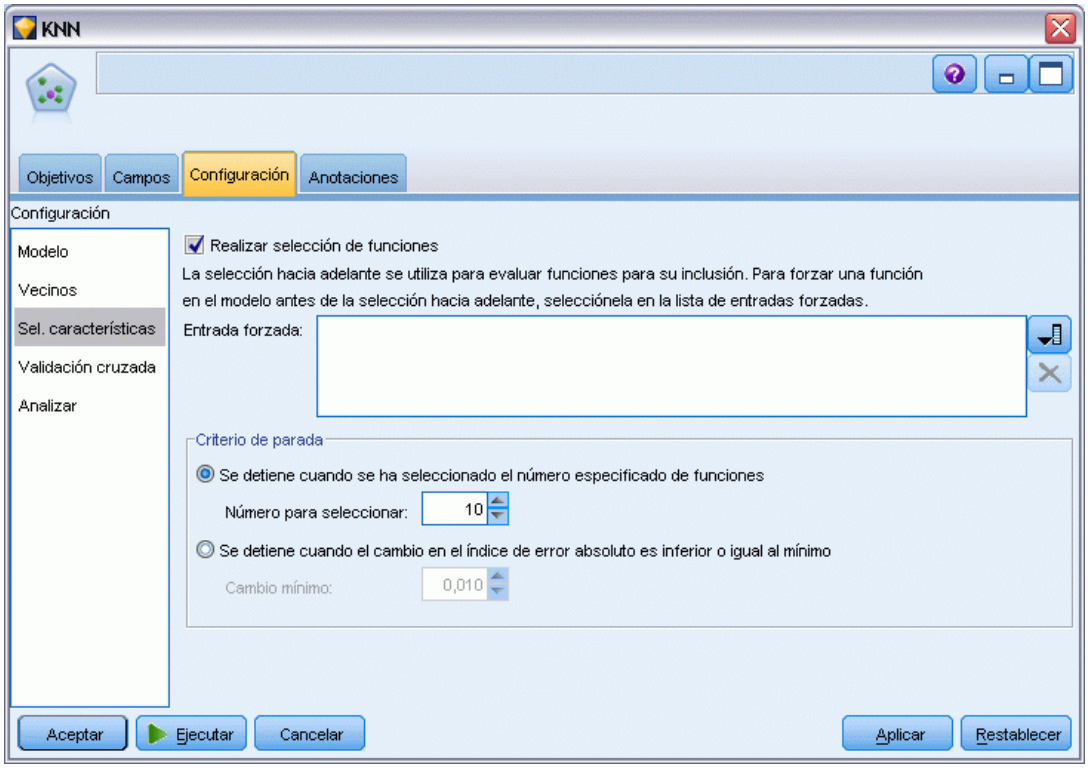
Opcionalmente, si el objetivo es pronosticar un objetivo, podrá seleccionar ponderar funciones según la importancia normalizada al calcular las distancias. La importancia de funciones de un predictor se calcula por la relación de la tasa de errores, o el error de la suma de los cuadrados del modelo sin 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 función para que sumen 1.

Ponderar funciones por importancia al calcular distancias. (Se muestra sólo si el objetivo es pronosticar 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). Si desea obtener más información, consulte el tema Importancia del predictor en el capítulo 3 el p. 55.

Predicciones del destino de rango. (Se muestra sólo si el objetivo es pronosticar un objetivo.) Si se especifica un objetivo continuo (rango numérico), defines si el valor pronosticado se calcula en función de la mediana o el valor de la mediana de los vecinos más próximos.

Selección de funciones

Figura 16-5
Opciones de selección de funciones de nodo KNN



Este panel se activa sólo si el objetivo es pronosticar un objetivo. Permite solicitar y especificar opciones de selección de funciones. Por defecto, todas las funciones se tienen en cuenta para la selección de funciones, pero es posible seleccionar un subconjunto de funciones para forzarlas en el modelo.

Realizar selección de características. Seleccione esta casilla para activar las opciones de selección de funciones.

- **Entrada forzada.** Pulse el botón de selección de campos junto a esta casilla y seleccione una o más funciones que se forzarán en el modelo.

Criterio de parada. En cada paso, la función 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 destino continuo) se tiene en cuenta para su inclusión en el conjunto de modelos. La selección continúa hasta que se cumple la condición especificada.

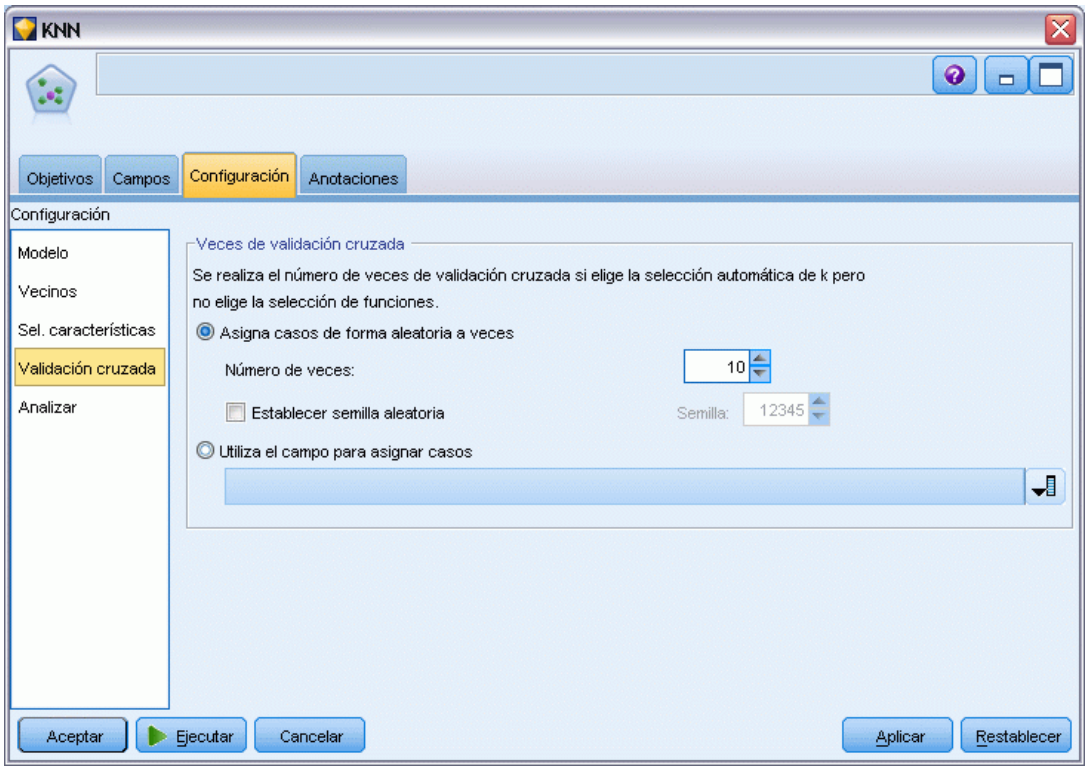
- **Parar cuando se ha seleccionado el número especificado de funciones.**El algoritmo añade un número fijo de funciones además de las forzadas en el modelo. Especifique un número entero positivo. Si se disminuyen los valores de número que se puede seleccionar se obtiene un modelo más reducido, lo que supone el riesgo de perder importantes funciones. Si se

aumentan los valores de número que se puede seleccionar se incluirán todas las funciones importantes, pero se corre el riesgo de añadir funciones 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 funciones. Especifique un número positivo. Los valores decrecientes del cambio mínimo tienden a incluir más funciones, con el riesgo de incluir funciones que no añaden demasiado valor al modelo. Si se aumentan los valores del cambio mínimo se excluirán más funciones, pero puede que se pierdan funciones importantes para el modelo. El valor "óptimo" de cambio mínimo dependerá de sus datos y de la aplicación. Consulte el Registro de errores de selección de funciones en los resultados para poder evaluar qué funciones son más importantes. Si desea obtener más información, consulte el tema Registro de errores de selección de predictores el p. 514.

Validación cruzada

Figura 16-6
Opciones de validación cruzada de nodo KNN



Este panel se activa sólo si el objetivo es pronosticar 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

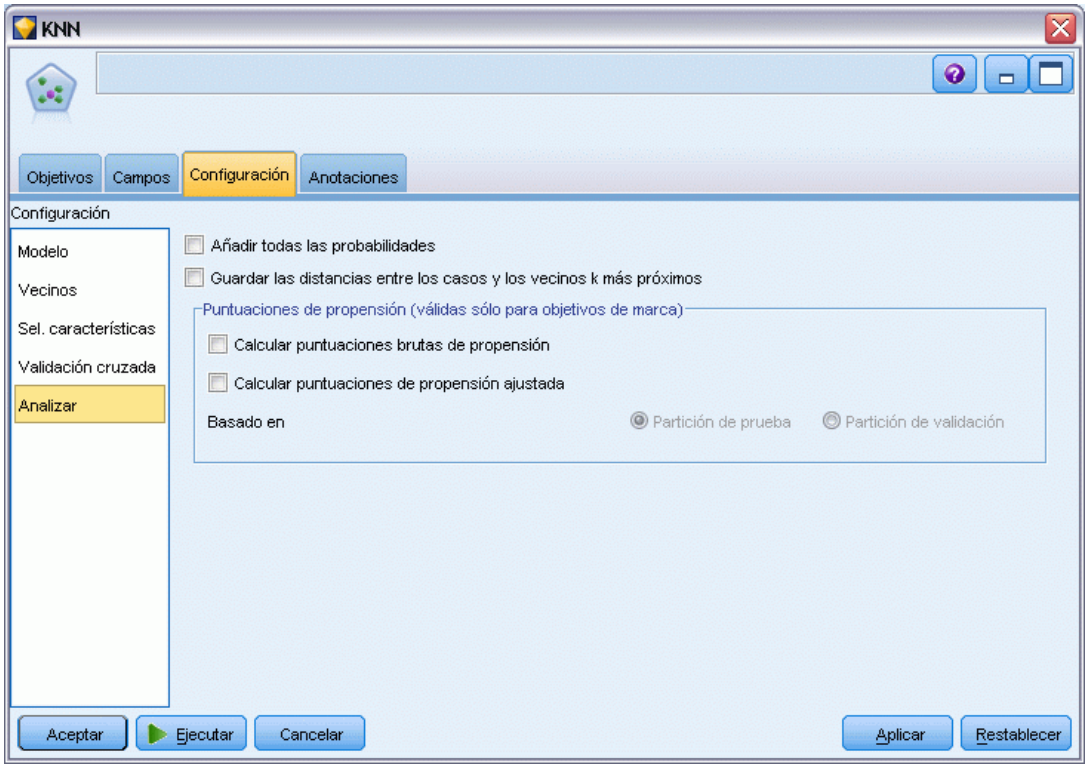
muestra y así sucesivamente. Para cada modelo se calcula el error aplicando el modelo a la submuestra que se excluyó al generarse este. El "mejor" número de vecinos más próximos será el que produzca el menor error entre los pliegues.

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 funciones.

- **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 toma el valor entre 1 a *V*. Si falta cualquiera de los valores de este rango, y en los campos divididos de si los modelos divididos están activados, se generará un error.

Analizar

Figura 16-7
Opciones de análisis de modelo de nodo KNN



El panel Analizar se activa sólo si el objetivo es pronosticar 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 brutas de propensión (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 pronosticado 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 los vecinos más próximos' *k* de cada 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 ficha Configuración del nugget de modelo. Esta funcionalidad sólo está disponible cuando el objetivo seleccionado es un campo de marca. Si desea obtener más información, consulte el tema Puntuaciones de propensión en el capítulo 3 el p. 45.

Calcular puntuaciones brutas de propensión. Las puntuaciones brutas de propensión 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 del pronóstico. Si el modelo predice el valor false, la propensión se calculará como (1 – P).

- Si selecciona esta opción al crear el modelo, las puntuaciones de propensión se activarán en el nugget de modelo por defecto. Sin embargo, siempre puede activar las puntuaciones brutas de propensión en el nugget de modelo independientemente de si las selecciona o no en el nodo de modelado.
- Al puntuar el modelo, se añadirán puntuaciones brutas de propensión a un campo con las letras *RP* unidas al prefijo estándar. Por ejemplo, si los pronósticos 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. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.
- 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 los pronósticos 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. Si desea obtener más información, consulte el tema Opciones de configuración del nodo Equilibrar en el capítulo 3 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*. 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. Si desea obtener más información, consulte el tema Modelos C5.0 aumentados en el capítulo 6 el p. 186.

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. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 15*.

Nugget de modelo KNN

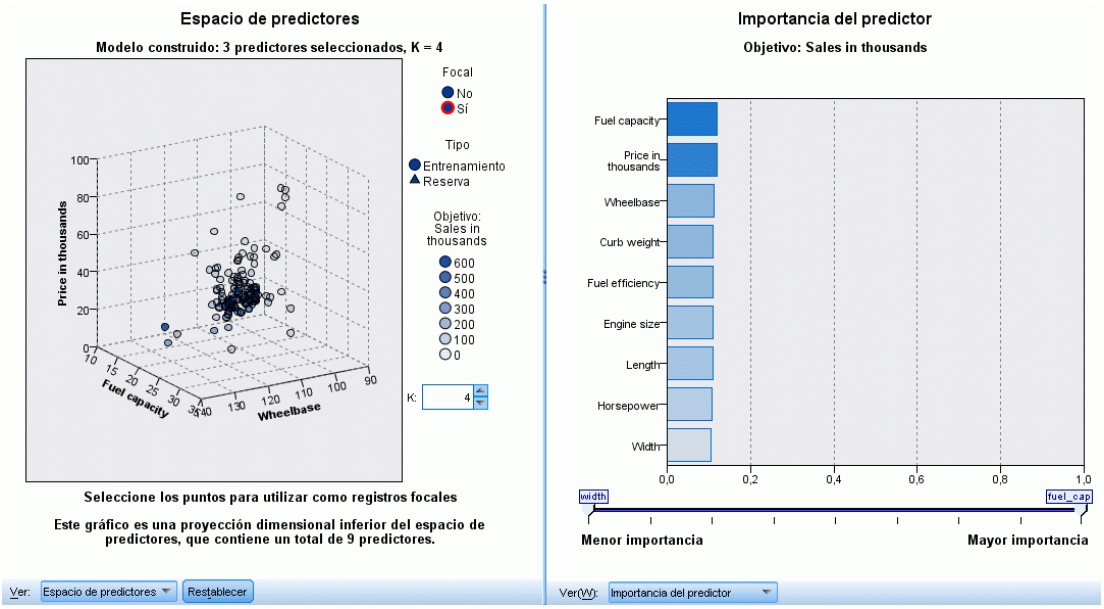
El modelo KNN crea un número de campos nuevos, tal y como se muestra en la tabla siguiente. Para ver estos campos y sus valores, añada un nodo Tabla al nugget de modelo KNN y ejecute el nodo Tabla, o pulse en el botón Presentación preliminar en el nugget.

Tabla 16-1
Campos de modelo KNN

Nombre del campo nuevo	Descripción
$KNN-*nombredecampo*	Valor pronosticado del campo objetivo.
$KNNP-*nombredecampo*	Probabilidad del valor pronosticado.
$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

Figura 16-8
Análisis de vecinos más próximos: Vista de modelos



La vista de modelos tiene una ventana con dos paneles:

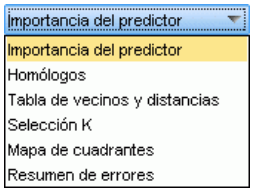
- El primer panel muestra una descripción general del modelo denominado vista principal.
- El segundo panel muestra uno de los dos tipos de vistas:

  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 función del modelo cuando el usuario desglosa parte de la vista principal.

De manera 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 los predictores no está disponible, es decir, si no ha seleccionado Ponderar funciones por importanciadel panel Vecinos de la pestaña Configuración, se muestra la primer vista disponible en la lista desplegable Ver.

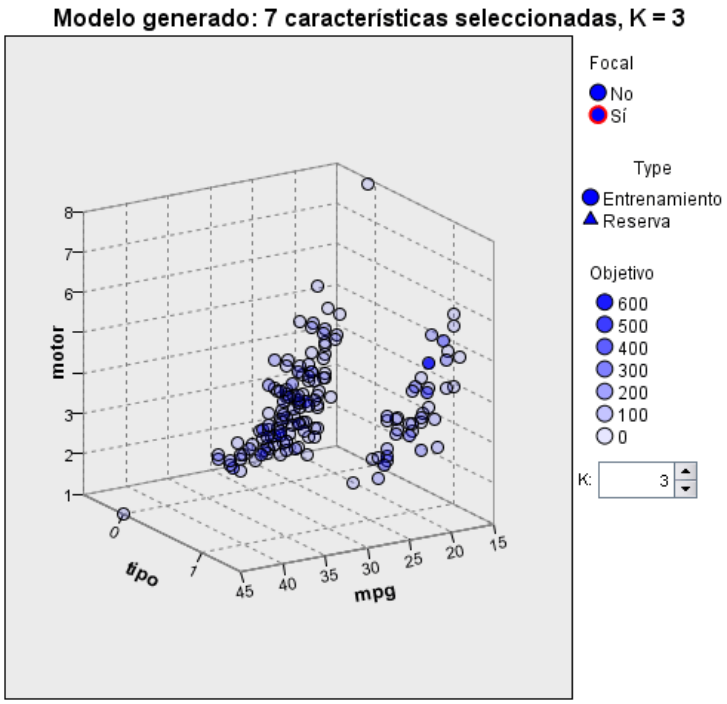
Figura 16-9
Análisis de vecinos más próximos: lista desplegable Ver



Cuando una vista no tiene información disponible, se omite del cuadro desplegable Ver.

Espacio predictor

Figura 16-10
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 diferente representa las categorías de un destino categórico y las sombras indican el rango de valores de un destino 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 pronosticado. 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. A la selección de puntos se aplican los controles “normales”, es decir, si pulsa en un punto éste se selecciona y se cancela la selección de todos los demás y si pulsa Control y el ratón sobre un punto éste se añadirá 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 pronosticados.
- 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 funciones se muestran en los ejes del gráfico Espacio predictor.

Para cambiar la configuración de los ejes:

▶ Pulse en el botón Modo edición (icono de pincel) en el panel de la izquierda para seleccionar el Modo edición de Espacio predictor.

▶ Cambie la vista (a cualquier cosa) en el panel derecho. Aparecerá el panel Mostrar zonas entre los dos paneles principales.

▶ Pulse en la casilla de verificación Mostrar zonas.

▶ Pulse en cualquier punto de datos en Espacio predictor.

▶ 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.

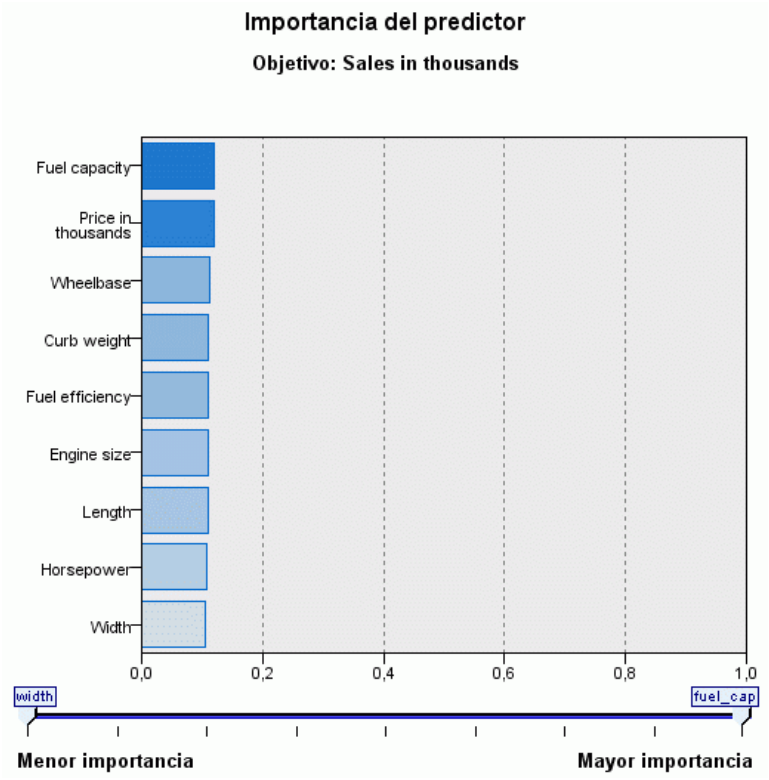
▶ Para sustituir un eje por un predictor de un tipo de datos diferente:

- 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.

▶ 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

Figura 16-11
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

Figura 16-12
Distancias de vecinos más próximos

Vecinos más próximos k y distancias

Mostrado para los registros focales iniciales

Registros focal	Vecinos más próximos				Distancias más próximas			
	1	2	3	4	1	2	3	4
112	15	13	77	33	0,082	0,103	0,121	0,126

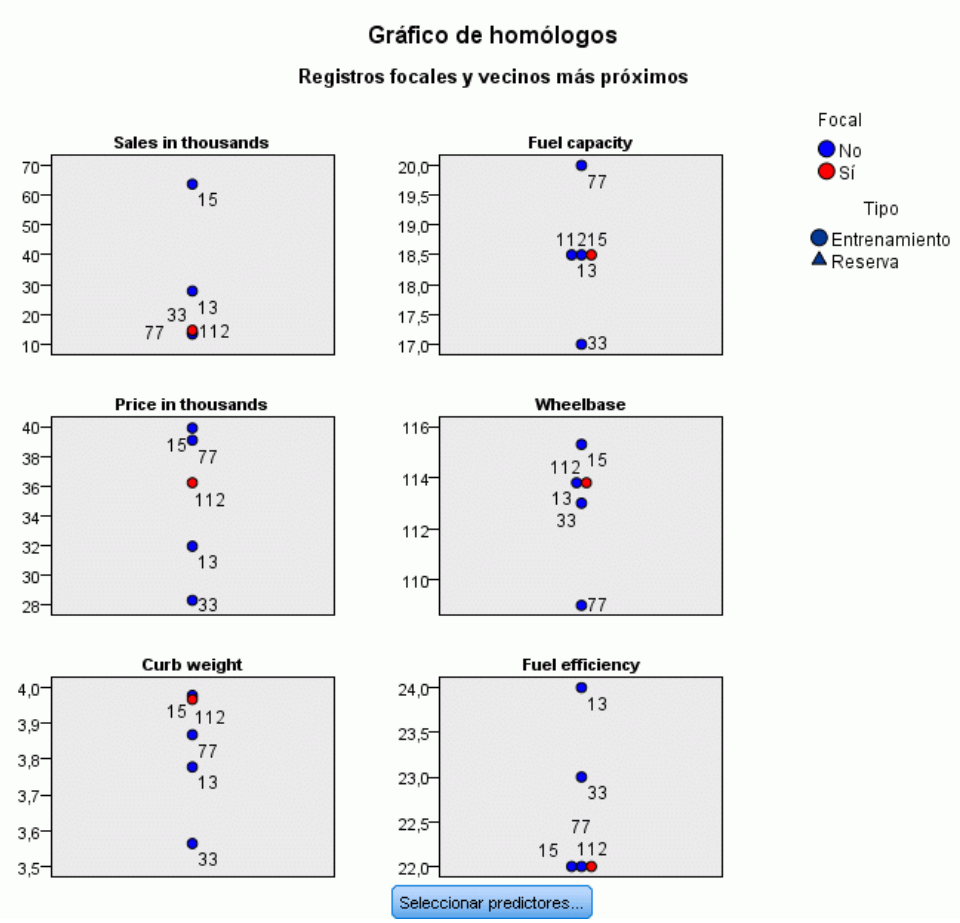
Esta tabla muestra los *k* vecinos más próximos y las distancias de registros focales únicamente. Está disponible si se especifica un identificador de registro focal en la nodo de modelado y sólo muestra los registros focales identificados por esta variable.

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

Figura 16-13
Gráfico 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.

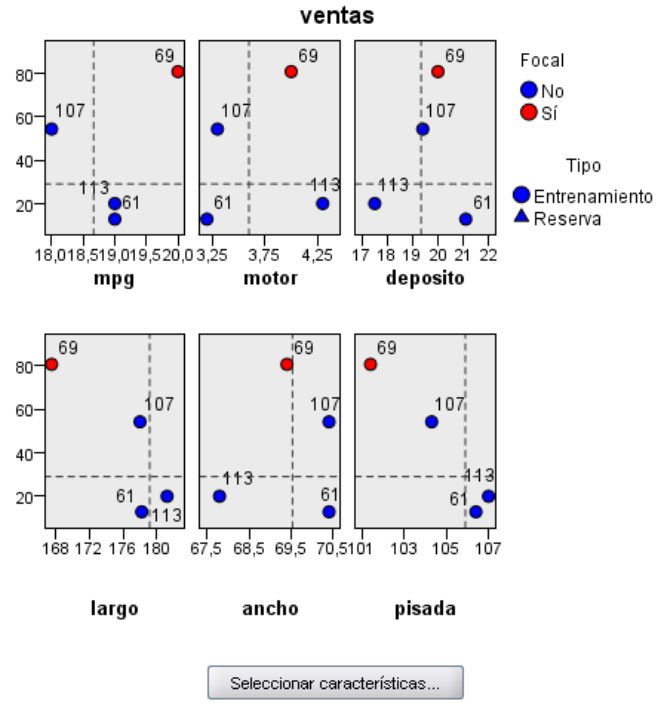
El gráfico Homólogos se vincula a Espacio predictor de dos maneras.

- 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

Figura 16-14
Mapa de cuadrantes



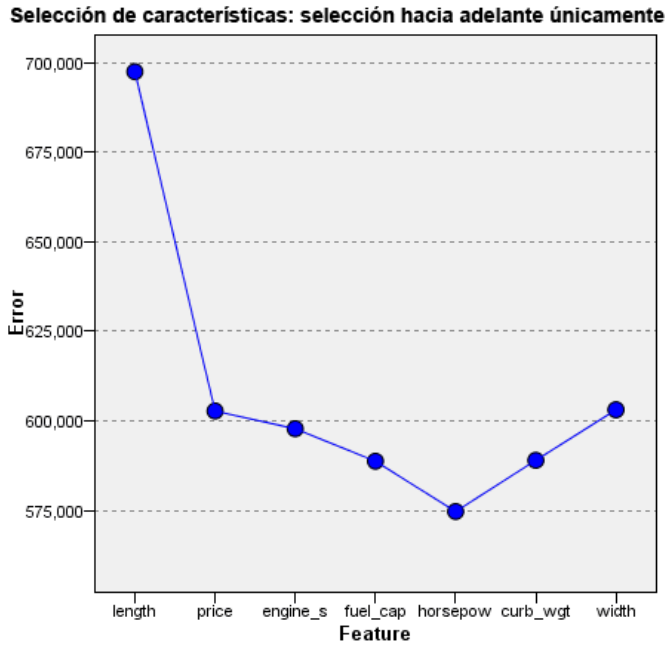
Este gráfico muestra los casos focales y sus *k* vecinos más próximos en un diagrama de dispersión (o gráfico de puntos, dependiendo del nivel de medición del destino) con el destino en el eje *y* y un predictor de escala en el eje *x*, panelado por predictores. Está disponible si hay un destino y se 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

Figura 16-15
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 funciones a la izquierda del eje *x*). Este gráfico está disponible si hay un destino y la selección de funciones está activada.

Tabla de clasificación

Figura 16-18
Tabla de clasificación

Partición		Pronósticos		
		0	**1**	**Porcentaje correcto**
Entrenamiento	**0**	111	1	99.11%
	1	7	33	82.50%
	Porcentaje global	77.64%	22.37%	94.74%

Esta tabla muestra la clasificación cruzada de los valores observados en comparación con los valores pronosticados 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: Valores de Porcentaje global, pero no a los valores de Porcentaje correcto.

Resumen de error

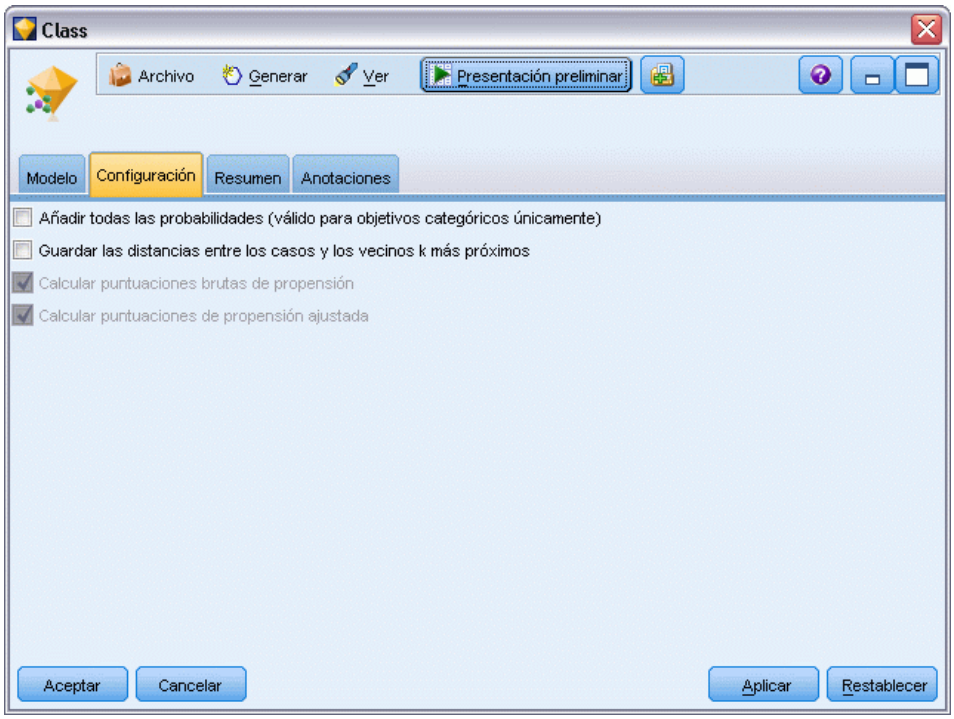
Figura 16-19
Resumen de error

Partición	Sum-of-Squares Error
Entrenamiento	622043

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

Figura 16-20
Ajustes de nugget de modelo KNN



La ficha 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 pronosticado y su probabilidad para campos objetivo nominal o marca.

El valor por defecto de esta casilla de verificación está determinado por la casilla de verificación correspondiente en el nodo de modelado.

Calcular puntuaciones brutas de propensión. En el caso de modelos con un objetivo de marca (que devuelve un pronóstico 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 pronóstico y confianza que pueden generarse durante la puntuación.

Calcular puntuaciones de propensión ajustada. Las puntuaciones brutas de propensión 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 la ficha Analizar 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.

Apéndice

A

Avisos

Esta información se ha desarrollado para los productos y servicios ofrecidos en todo el mundo.

Puede que IBM no ofrezca los productos, los servicios o las características de los que se habla en este documento en otros países. Consulte a su representante local de IBM para obtener información acerca de los productos y servicios que está disponibles actualmente en su zona. Toda referencia que se haga de un producto, programa o servicio de IBM no implica que sólo se deba utilizar ese producto, programa o servicio de IBM. En su lugar, puede utilizarse todo producto, programa o servicio con funcionalidades equivalentes que no infrinjan 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 aplicaciones de patentes pendientes que cubren el asunto descrito 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, Armonk, NY 10504-1785, Estados Unidos

Si tiene alguna pregunta sobre la licencia relacionada con la información del juego de caracteres de doble byte (DBCS), póngase en contacto con el departamento de propiedad intelectual de IBM de su país o envíe sus preguntas por escrito a:

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan.

El párrafo siguiente no se aplica a los Reino Unido o cualquier otro país donde tales disposiciones son incompatibles con la legislación local: INTERNATIONAL BUSINESS MACHINES PROPORCIONA ESTA PUBLICACIÓN “TAL CUAL” SIN GARANTÍA DE NINGÚN TIPO, YA SEA EXPRESA O IMPLÍCITA, INCLUYENDO, PERO NO LIMITADA A, LAS GARANTÍAS IMPLÍCITAS DE NO INFRACCIÓN, COMERCIALIZACIÓN O IDONEIDAD PARA UN PROPÓSITO PARTICULAR. Algunos estados no permiten la renuncia a expresar o a garantías implícitas en determinadas transacciones , por lo tanto , esta declaración no se aplique a usted.

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 mejoras y/o cambios en los productos y/o los programas descritos en esta publicación en cualquier momento sin previo aviso.

Cualquier referencia a sitios Web que no sean de IBM en esta información sólo es ofrecida por comodidad y de ningún modo sirve como aprobación de esos sitios Web. Los materiales en esos sitios Web no forman parte del material de este producto de IBM y el uso de estos sitios Web es bajo su propio riesgo.

IBM puede utilizar cualquier información que le suministre en cualquier forma que considere adecuada, sin incurrir en ninguna obligación para usted.

Los licenciatarios de este programa que deseen tener información sobre el mismo con el objetivo de habilitar: (i) el intercambio de información entre programas creados independientemente y otros programas (incluyendo este) y (ii) el uso común de la información que se ha intercambiado, deben ponerse en contacto con:

IBM Software Group, a la atención de: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

Esta información estará disponible, bajo las condiciones adecuadas, incluyendo en algunos casos el pago de una cuota.

IBM proporciona el programa bajo licencia que se describe en este documento y todo el material bajo licencia disponible para el mismo bajo los términos de IBM Customer Agreement (Acuerdo de cliente de IBM), IBM International Program License Agreement (Acuerdo de licencia de programa internacional de IBM) o cualquier acuerdo equivalente entre las partes.

Cualquier dato de rendimiento mencionado aquí ha sido determinado en un entorno controlado. Por lo tanto, los resultados obtenidos en otros entornos operativos pueden variar de forma significativa. Es posible que algunas medidas se hayan realizado en sistemas en desarrollo y no existe ninguna garantía de que estas medidas sean las mismas en los sistemas comerciales. Además, es posible que algunas medidas hayan sido estimadas a través de extrapolación. Los resultados reales pueden variar. Los usuarios de este documento deben consultar los datos que corresponden a su entorno específico.

Se ha obtenido información acerca de productos que no son de IBM de los proveedores de esos productos, de sus publicaciones anunciadas o de otras fuentes disponibles públicamente. IBM no ha probado estos productos y no puede confirmar la precisión de su rendimiento, su compatibilidad o cualquier otra reclamación relacionada con productos que no sean de IBM. Las preguntas acerca de las aptitudes de productos que no sean de IBM deben dirigirse a los proveedores de dichos productos.

Todas las declaraciones sobre el futuro del rumbo y la 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 esos nombres son ficticios y cualquier parecido con los nombres y direcciones utilizados por una empresa real es pura coincidencia.

Si está viendo esta información en copia electrónica, es posible que las fotografías y las ilustraciones en color no aparezcan.

Marcas comerciales

IBM, el logotipo de IBM, ibm.com y SPSS son marcas comerciales de IBM Corporation, registradas en muchas jurisdicciones de todo el mundo. Existe una lista actualizada de marcas comerciales de IBM en Internet en *http://www.ibm.com/legal/copytrade.shtml*.

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 de Sun Microsystems, Inc. en Estados Unidos, otros países o ambos.

Otros productos y nombres de servicio pueden ser marcas comerciales de IBM u otras empresas.

Índice