IBM SPSS Decision Trees 24

Comunicado

Antes de usar estas informações e o produto suportado por elas, leia as informações nos "Avisos" na página 25.

Informações sobre o produto

Esta edição aplica-se à versão 24, liberação 0, modificação 0 do IBM SPSS Statistics e a todas as liberações e modificações subsequentes até que seja indicado de outra forma em novas edições.

Índice

Capítulo 1. Criando árvores de decisão

O procedimento de Árvore de decisão cria um modelo de classificação baseado em árvore. Ele classifica casos em grupos ou prevê valores de uma variável dependente (de destino) com base em valores de variáveis independentes (preditoras). O procedimento fornece ferramentas de validação para análise de classificação exploratória e confirmatória.

O procedimento pode ser usado para:

Segmentação. Identifique pessoas que provavelmente sejam membros de um determinado grupo.

Estratificação. Designe casos em uma de várias categorias, como grupos de risco alto, médio e baixo.

Predição. Crie regras e use-as para prever eventos futuros, como a probabilidade de que alguém dará calote em um empréstimo ou o possível valor de revenda de um veículo ou imóvel.

Redução de dados e triagem de variáveis. Selecione um subconjunto útil de preditores de um grande conjunto de variáveis para uso na construção de um modelo paramétrico formal.

Identificação de interação. Identifique relacionamentos referentes apenas a subgrupos específicos e especifique-os em um modelo paramétrico formal.

Mesclando e distinguindo variáveis contínuas de categoria. Recodifique categorias do preditor do grupo e variáveis contínuas com uma perda mínima de informações.

Exemplo. Um banco quer categorizar requerentes de crédito de acordo com se eles representam ou não um risco de crédito razoável. Com base em vários fatores, incluindo as classificações de crédito conhecidas de clientes antigos, é possível construir um modelo para prever se os futuros clientes têm probabilidade de darem calote em seus empréstimos.

Uma análise baseada em árvore fornece alguns recursos atrativos:
- Permite identificar grupos homogêneos com alto ou baixo risco.
- Facilita a construção de regras para fazer predições sobre casos individuais.

Considerações de dados

Dados. As variáveis dependentes e independentes podem ser:
- *Nominal*. Uma variável pode ser tratada como nominal quando seus valores representarem categorias sem ranqueamento intrínseco (por exemplo, o departamento da empresa na qual um funcionário trabalha). Exemplos de variáveis nominais incluem região, código de endereçamento postal e filiação religiosa.
- *Ordinal*. Uma variável pode ser tratada como ordinal quando seus valores representarem categorias com algum ranqueamento intrínseco (por exemplo, níveis de satisfação de serviço de muito insatisfeito para muito satisfeito). Exemplos de variáveis ordinais incluem escores de atitude que representam o grau de satisfação ou de confiança e os escores de classificação de preferência.
- *Escala*. Uma variável pode ser tratada como escala (contínua) quando os seus valores representarem categorias ordenadas com uma métrica significativa, de forma que as comparações de distância entre os valores sejam apropriadas. Exemplos de variáveis de escala incluem idade em anos e rendimento em milhares de dólares.

Ponderações de frequência Se a ponderação estiver em vigor, as ponderações fracionárias serão arredondadas para o número inteiro mais próximo; portanto, os casos com um valor de ponderação menor que 0,5 serão designados a uma ponderação de 0 e, portanto, serão excluídos da análise.

Suposições. Esse procedimento supõe que o nível de medição apropriado tenha sido designado a todas as variáveis de análise, e alguns recursos consideram que todos os valores da variável dependente incluídos na análise tenham rótulos de valor definidos.

- **Nível de medição.** O nível de medição afeta os cálculos de árvore; portanto, todas as variáveis devem ser designadas ao nível de medição apropriado. Por padrão, as variáveis numéricas são consideradas como de escala e as variáveis de sequência de caracteres são consideradas como nominais, o que pode não refletir com exatidão o nível de medição verdadeiro. Um ícone próximo a cada variável na lista de variáveis identifica o tipo de variável.

Tabela 1. Ícones do nível de medição.

Ícone	Nível de medição
	Escala
	Nominal
	Ordinal

É possível mudar temporariamente o nível de medição para uma variável, clicando com o botão direito na variável na lista de variáveis de origem e selecionando um nível de medição no menu pop-up.

- **Rótulos de valor.** A interface da caixa de diálogo para esse procedimento considera que todos os valores não omissos de uma variável dependente categórica (nominal, ordinal) tenham rótulos de valor definidos ou que nenhum deles tem. Alguns recursos não estarão disponíveis, a não ser que pelo menos dois valores não omissos da variável dependente categórica tenham rótulos de valor. Se pelo menos dois valores não omissos tiverem rótulos de valor definidos, todos os casos com outros valores que não possuem rótulos de valor serão excluídos da análise.

Para obter árvores de decisão

1. Nos menus, escolha:
   Analisar > **Classificar** > **Árvore...**
2. Selecione uma variável dependente.
3. Selecionar uma ou mais variáveis independentes.
4. Selecione um método de desenvolvimento.

Como opção, você pode:

- Mudar o nível de medição para qualquer variável na lista de origem.
- Forçar a primeira variável na lista de variáveis independentes no modelo como a primeira variável de divisão.
- Selecionar uma variável de influência que defina o grau de influência que um caso tem no processo de crescimento da árvore. Os casos com valores de influência mais baixos têm menos influência; os casos com valores mais altos têm mais. Os valores da variável de influência devem ser positivos.
- Validar a árvore.
- Customizar os critérios crescentes da árvore.
- Salvar números de nó terminal, valores preditos e probabilidades preditas como variáveis.
- Salvar o modelo no formato XML (PMML).

Campos com nível de medição desconhecido

O alerta de Nível de Medição é exibido quando o nível de medição para uma ou mais variáveis (campos) no conjunto de dados é desconhecido. Como o nível de medição afeta o cálculo de resultados para este procedimento, todas as variáveis devem ter um nível de medição definido.

Dados de varredura. Lê os dados no conjunto de dados ativo e designa o nível de medição padrão para quaisquer campos com um nível de medição desconhecido atualmente. Se o conjunto de dados for grande, isso poderá demorar algum tempo.

Designar Manualmente. Abre um diálogo que lista todos os campos com um nível de medição desconhecido. É possível utilizar este diálogo para designar o nível de medição para esses campos. Também é possível designar o nível de medição na Visualização de Variável do Editor de Dados.

Como o nível de medição é importante para este procedimento, não é possível acessar o diálogo para executar este procedimento até que todos os campos possuam um nível de medição definido.

Mudando o nível de medição

1. Clique com o botão direito na variável na lista de origem.
2. Selecione um nível de medição no menu pop-up.

Isso muda temporariamente o nível de medição para uso no procedimento Árvore de decisão.

Métodos crescentes

Os métodos crescentes disponíveis são:

CHAID. Chi-squared Automatic Interaction Detection. Em cada passo, o CHAID escolhe a variável (preditora) independente que possuir a interação mais forte com a variável dependente. As categorias de cada preditor serão mescladas se elas não forem significativamente diferentes com relação à variável dependente.

CHAID exaustivo. Uma modificação do CHAID que examina todas as possíveis divisões para cada preditor.

CRT. Árvores de Classificação e Regressão. O CRT divide os dados em segmentos que são tão homogêneos quanto possíveis com relação à variável dependente. Um nó terminal no qual todos os casos têm o mesmo valor para a variável dependente é um nó homogêneo "puro".

QUEST. Árvore Estatística Eficiente, Não Tendenciosa e Rápida. Um método que é rápido e evita outros vieses de métodos em favor de preditores com muitas categorias. O QUEST poderá ser especificado apenas se a variável dependente for nominal.

Existem benefícios e limitações com cada método, incluindo:

Tabela 2. Recursos do método crescente.

Recursos	**CHAID***	**CRT**	**QUEST**
Baseado em qui-quadrado**	P		
Variáveis independentes substitutas (preditoras)		P	P
Poda de árvore		P	P
Divisão de nó multiponto	P		
Divisão de nó binário		P	P
Variáveis de influência	P	P	

Tabela 2. Recursos do método crescente (continuação).

Recursos	CHAID*	CRT	QUEST
Probabilidades anteriores		P	P
Custos de classificação errada	P	P	P
Cálculo rápido	P		P

*Inclui CHAID exaustivo.

**QUEST também usa uma medida qui-quadrado para variáveis independentes nominais.

Selecionando categorias

Para variáveis dependentes categóricas (nominais, ordinais), é possível:

- Controlar quais categorias estão incluídas na análise.
- Identificar as categorias de destino de interesse.

Incluindo/Excluindo categorias

É possível limitar a análise a categorias específicas da variável dependente.

- Os casos com valores da variável dependente na lista Exclusão não são incluídos na análise.
- Para variáveis dependentes nominais, também é possível incluir categorias com usuário desconhecido na análise. (Por padrão, as categorias com usuário desconhecido são exibidas na lista Exclusão.)

Categorias de destino

As categorias selecionadas (marcadas) são tratadas como as categorias de interesse principal na análise. Por exemplo, se você estiver interessado principalmente em identificar as pessoas com maior probabilidade de dar calote em um empréstimo, poderá selecionar a categoria de classificação de crédito "ruim" como a categoria de destino.

- Não há nenhuma categoria de destino padrão. Se nenhuma categoria estiver selecionada, algumas opções de regra de classificação e saída relacionada a ganhos não estarão disponíveis.
- Se várias categorias forem selecionadas, tabelas e gráficos de ganhos separados serão produzidos para cada categoria de destino.
- A designação de uma ou mais categorias como categorias de destino não tem nenhum efeito no modelo de árvore, na estimativa de risco ou nos resultados de classificação errada.

Categorias e rótulos de valor

Essa caixa de diálogo requer rótulos de valor definidos para a variável dependente. Ela não está disponível, a não ser que pelo menos dois valores da variável dependente categórica tenham rótulos de valor definidos.

Para incluir/excluir categorias e selecionar categorias de destino

1. Na caixa de diálogo principal Árvore de decisão, selecione uma variável dependente categórica (nominal, ordinal) com dois ou mais rótulos de valor definidos.
2. Clique em **Categorias**.

Validação

A validação permite avaliar como a estrutura em árvore generaliza para uma população maior. Dois métodos de validação estão disponíveis: validação cruzada e validação de amostra de divisão.

Validação cruzada

A validação cruzada divide a amostra em várias subamostras ou **dobras**. Os modelos de árvore são então gerados, excluindo os dados de cada subamostra, sucessivamente. A primeira árvore é baseada em todos os casos, exceto os que estão na primeira dobra de amostra, a segunda árvore é baseada em todos os casos, exceto os que estão na segunda dobra de amostra, e assim por diante. Para cada árvore, o risco de classificação errada é estimado aplicando a árvore à subamostra excluída em sua geração.

- É possível especificar um máximo de 25 dobras de amostra. Quanto maior o valor, menor o número de casos excluídos para cada modelo de árvore.
- A validação cruzada produz um único modelo de árvore final. A estimativa de risco de validação cruzada para a árvore final é calculada como a média dos riscos para todas as árvores.

Validação de amostra de divisão

Com a validação de amostra de divisão, o modelo é gerado usando uma amostra de treinamento e testado em uma amostra de validação.

- É possível especificar um tamanho de amostra de treinamento, expresso como uma porcentagem do tamanho total da amostra, ou uma variável que divide a amostra em amostras de treinamento e de teste.
- Se você usar uma variável para definir amostras de treinamento e de teste, os casos com um valor 1 para a variável serão designados à amostra de treinamento e todos os outros casos serão designados à amostra de teste. A variável não pode ser a variável dependente, uma variável de ponderação, uma variável de influência ou uma variável independente forçada.
- É possível exibir resultados para as amostras de treinamento e de teste ou apenas para a amostra de teste.
- A validação de amostra de divisão deve ser usada com cuidado em arquivos de dados pequenos (arquivos de dados com um pequeno número de casos). Os tamanhos pequenos da amostra de treinamento podem resultar em modelos simples, pois pode não haver casos suficientes em algumas categorias para o crescimento adequado da árvore.

Para validar uma árvore de decisão

1. No diálogo principal Árvores de decisão, clique em **Validação**.
2. Selecione **Validação cruzada** ou **Validação de amostra de divisão**.

Nota: Os dois métodos de validação designam casos aleatoriamente a grupos de amostra. Se desejar ser capaz de reproduzir exatamente os mesmos resultados em uma análise subsequente, você deve configurar a valor semente de número aleatório (menu Transformar, Geradores de números aleatórios) antes de executar a análise pela primeira vez e, em seguida, reconfigure a semente como esse valor para a análise subsequente.

Critérios crescentes da árvore

Os critérios crescentes disponíveis podem depender do método crescente, do nível de medição da variável dependente ou uma combinação dos dois.

Limites de crescimento

A guia Limites de crescimento permite limitar o número de níveis na árvore e controlar o número mínimo de casos para nós pais e filhos.

Profundidade máxima da árvore. Controla o número máximo de níveis de crescimento abaixo do nó raiz. A configuração **Automático** limita a árvore a três níveis abaixo do nó raiz para os métodos CHAID e CHAID Exaustivo e a cinco níveis para os métodos CRT e QUEST.

Número mínimo de casos. Controla o número mínimo de casos para os nós. Os nós que não satisfazem esses critérios não serão divididos.

- Aumentar os valores mínimos tende a produzir árvores com menos nós.
- Diminuir os valores mínimos produz árvores com mais nós.

Para arquivos de dados com um pequeno número de casos, os valores padrão de 100 casos para nós pai e 50 casos para nós filhos podem, às vezes, resultar em árvores sem nenhum nó abaixo do nó raiz; nesse caso, reduzir os valores mínimos pode produzir resultados mais úteis.

Para especificar limites de crescimento

1. No diálogo principal Árvore de decisão, clique em **Critérios**.
2. Clique na guia **Limites de crescimento**.

Critérios de CHAID

Para os métodos CHAID e CHAID Exaustivo, é possível controlar:

Nível de significância. É possível controlar o valor de significância para dividir nós e mesclar categorias. Para os dois critérios, o nível de significância padrão é 0,05.

- Para dividir nós, o valor deve ser maior que 0 e menor que 1. Valores mais baixos tendem a produzir árvores com menos nós.
- Para mesclar categorias, o valor deve ser maior que 0 e menor ou igual a 1. Para evitar a mesclagem de categorias, especifique um valor 1. Para uma variável independente de escala, isso significa que o número de categorias para a variável na árvore final é o número especificado de intervalos (o padrão é 10). Consulte o tópico “Intervalos de escala para análise CHAID” na página 7 para obter mais informações

Estatística qui-quadrado. Para variáveis dependentes ordinais, o qui-quadrado para determinar a divisão do nó e a mesclagem de categoria é calculado usando o método de razão de verossimilhança. Para variáveis dependentes nominais, é possível selecionar o método:

- **Pearson.** Este método fornece cálculos mais rápidos, mas deve ser utilizado com cuidado em amostras pequenas. Este é o método padrão.
- **Razão de verossimilhança.** Esse método é mais robusto que o de Pearson, mas leva mais tempo para ser calculado. Para pequenas amostras, este é o método preferencial.

Estimação de modelo. Para variáveis dependentes nominais e ordinais, é possível especificar:

- **Número máximo de iterações.** O padrão é 100. Se a árvore parar de crescer porque o número máximo de iterações foi atingido, talvez você queira aumentar o máximo ou mudar um ou mais dos outros critérios que controlam o crescimento da árvore.
- **Mudança mínima nas frequências de célula esperadas.** O valor deve ser maior que 0 e menor que 1. O padrão é 0,05. Valores mais baixos tendem a produzir árvores com menos nós.

Ajustar valores de significância usando o método de Bonferroni. Para comparações múltiplas, os valores de significância para critérios de mesclagem e divisão são ajustados usando o método de Bonferroni. Esse é o padrão.

Permitir nova divisão de categorias mescladas em um nó. A menos que você evite explicitamente a mesclagem de categoria, o procedimento tentará mesclar as categorias de variáveis independentes (preditoras) juntas para produzir a árvore mais simples que descreve o modelo. Essa opção permite que o procedimento divida novamente as categorias mescladas se for fornecida uma solução melhor.

Para especificar critérios de CHAID

1. No diálogo principal Árvore de decisão, selecione **CHAID** ou **CHAID Exaustivo** como o método crescente.
2. Clique em **Critérios**.
3. Clique na guia **CHAID**.

Intervalos de escala para análise CHAID

Na análise CHAID, as variáveis independentes de escala (preditoras) são sempre divididas em grupos distintos (por exemplo, 0–10, 11–20, 21–30, etc.) antes da análise. É possível controlar o número inicial/máximo de grupos (embora o procedimento possa mesclar grupos contínuos após a divisão inicial):

- **Número fixo.** Todas as variáveis independentes de escala são inicialmente divididas no mesmo número de grupos. O padrão é 10.
- **Customizado.** Cada variável independente de escala é inicialmente dividida no número de grupos especificados para essa variável.

Para especificar intervalos para variáveis independentes de escala

1. Na caixa de diálogo principal Árvore de decisão, selecione uma ou mais variáveis independentes de escala.
2. Para o método crescente, selecione **CHAID** ou **CHAID exaustivo**.
3. Clique em **Critérios**.
4. Clique na guia **Intervalos**.

Na análise CRT e QUEST, todas as divisões são binárias e as variáveis independentes de escala e ordinais são tratadas da mesma forma; portanto, não é possível especificar diversos intervalos para variáveis independentes de escala.

Critérios de CRT

O método crescente CRT tenta maximizar a homogeneidade no nó. A extensão até onde um nó não representa um subconjunto homogêneo de casos é uma indicação de **impureza**. Por exemplo, um nó terminal no qual todos os casos possuem o mesmo valor para a variável dependente é um nó homogêneo que não requer nenhuma divisão adicional porque é "puro."

É possível selecionar o método usado para medir a impureza e a diminuição mínima na impureza necessária para dividir nós.

Medida de Impureza. Para variáveis dependentes de escala, a medida de desvio de quadrado mínimo (LSD) de impureza é usada. Ela é calculada como a variância no nó, ajustada para quaisquer ponderações de frequência ou valores de influência.

Para variáveis dependentes categóricas (nominais, ordinais), é possível selecionar a medida de impureza:

- **Gini.** Foram encontradas divisões que maximizam a homogeneidade de nós filhos com relação ao valor da variável dependente. Gini é baseado nas probabilidades quadradas de associação para cada categoria da variável dependente. Ele atinge seu mínimo (zero) quando todos os casos em um nó estão incluídos em uma única categoria. Esta é a medida padrão.
- **Twoing.** Categorias da variável dependente são agrupadas em duas subclasses. São localizadas divisões que melhor separam os dois grupos.
- **Twoing ordenado.** Semelhante ao twoing, exceto que apenas as categorias adjacentes podem ser agrupadas. Esta medida está disponível apenas para variáveis dependentes ordinais.

Mudança mínima na melhoria. Essa é a diminuição mínima de impureza necessária para dividir um nó. O padrão é 0,0001. Valores mais altos tendem a produzir árvores com menos nós.

Para especificar critérios de CRT

1. Para o método crescente, selecione **CRT**.
2. Clique em **Critérios**.
3. Clique na guia **CRT**.

Critérios de QUEST

Para o método QUEST, é possível especificar o nível de significância para divisão de nós. Uma variável independente não pode ser usada para dividir nós, a menos que o nível de significância seja menor ou igual ao valor especificado. O valor deve ser maior que 0 e menor que 1. O padrão é 0,05. Valores menores tendem a excluir mais variáveis independentes do modelo final.

Para especificar critérios de QUEST

1. Na caixa de diálogo principal Árvore de decisão, selecione uma variável dependente nominal.
2. Para o método crescente, selecione **QUEST**.
3. Clique em **Critérios**.
4. Clique na guia **QUEST**.

Podando árvores

Com os métodos CRT e QUEST, é possível evitar o superajuste do modelo **podando** a árvore: a árvore crescerá até que os critérios de parada sejam atendidos e, em seguida, será aparada automaticamente até a menor subárvore, com base na diferença máxima especificada em risco. O valor de risco é expresso em erros padrão. O padrão é 1. O valor deve ser não negativo. Para obter a subárvore com o risco mínimo, especifique 0.

Para podar uma árvore

1. Na caixa de diálogo principal Árvore de decisão, para o método crescente, selecione **CRT** ou **QUEST**.
2. Clique em **Critérios**.
3. Clique na guia **Poda**.

Poda versus ocultar nós

Ao criar uma árvore podada, os nós podados da árvore não estão disponíveis na árvore final. É possível ocultar e mostrar interativamente os nós filhos selecionados na árvore final, mas não é possível mostrar nós que foram podados no processo de criação da árvore. Consulte o tópico Capítulo 2, “Editor de árvore”, na página 19 para obter mais informações

Substitutos

CRT e QUEST podem usar **substitutos** para variáveis independentes (preditores). Para casos em que o valor para essa variável estiver omisso, outras variáveis independentes que possuem altas associações com a variável original serão usadas para classificação. Esses preditores alternativos são chamados substitutos. É possível especificar o número máximo de substitutos a serem usados no modelo.

- Por padrão, o número máximo de substitutos é um menos o número de variáveis independentes. Em outras palavras, para cada variável independente, todas as outras variáveis independentes podem ser usadas como substitutas.
- Se não desejar que o modelo use substitutos, especifique 0 para o número de substitutos.

Para especificar substitutos

1. Na caixa de diálogo principal Árvore de decisão, para o método crescente, selecione **CRT** ou **QUEST**.
2. Clique em **Critérios**.
3. Clique na guia **Substitutos**.

Opções

As opções disponíveis podem depender do método crescente, do nível de medição da variável dependente e/ou da existência de rótulos do valor definidos para valores da variável dependente.

Custos de classificação errada

Para variáveis dependentes categóricas (nominais, ordinais), os custos de classificação errada permitem incluir informações sobre a penalidade relativa associada à classificação incorreta. Por exemplo:

- O custo de negar crédito a um cliente fidedigno provavelmente deve ser diferente do custo de estender crédito a um cliente que dá calote no empréstimo.
- O custo de classificação errada de uma pessoa com um alto risco de doença cardíaca como baixo risco provavelmente é muito mais alto do que o custo de classificação errada de uma pessoa de baixo risco como alto risco.
- O custo de envio de correspondência em massa para alguém que provavelmente não responderá provavelmente é muito baixo, enquanto o custo de não enviar a correspondência para alguém que provavelmente responderá é relativamente mais alto (em termos de renda perdida).

Custos de classificação errada e rótulos de valor

Essa caixa de diálogo não estará disponível, a não ser que pelo menos dois valores da variável dependente categórica tenham rótulos de valor definidos.

Para especificar custos de classificação errada

1. Na caixa de diálogo principal Árvore de decisão, selecione uma variável dependente categórica (nominal, ordinal) com dois ou mais rótulos de valor definidos.
2. Clique em **Opções**.
3. Clique na guia **Custos de classificação errada**.
4. Clique em **Customizado**.
5. Insira um ou mais custos de classificação errada na grade. Os valores devem ser não negativos. (As classificações corretas, representadas na diagonal, são sempre 0.)

Matriz de preenchimento. Em muitos casos, talvez você queira que os custos sejam simétricos — ou seja, o custo de classificação errada de A como B seja igual ao custo de classificação errada de B como A. Os seguintes controles podem facilitar a especificação de uma matriz de custos simétricos:

- **Duplicar triângulo inferior.** Copia valores no triângulo inferior da matriz (abaixo da diagonal) para as células do triângulo superior correspondente.
- **Duplicar triângulo superior.** Copia valores no triângulo superior da matriz (acima da diagonal) para as células do triângulo inferior correspondente.
- **Usar valores médios de célula.** Para cada célula em cada metade da matriz, os dois valores (triangular superior e inferior) têm sua média calculada e a média substitui ambos os valores. Por exemplo, se o custo de classificação errada de A como B for 1 e o custo de classificação errada de B como A for 3, esse controle substituirá esses dois valores pela média (1+3)/2 = 2.

Lucros

Para variáveis dependentes categóricas, é possível designar valores de renda e despesa a níveis da variável dependente.

- O lucro é calculado como renda menos despesa.
- Os valores de lucro afetam o lucro médio e o ROI (retorno sobre investimento) em tabelas de ganhos. Eles não afetam a estrutura básica do modelo de árvore.
- Os valores de renda e de despesa devem ser numéricos e devem ser especificados para todas as categorias da variável dependente exibida na grade.

Lucros e rótulos de valor

Essa caixa de diálogo requer rótulos de valor definidos para a variável dependente. Ela não está disponível, a não ser que pelo menos dois valores da variável dependente categórica tenham rótulos de valor definidos.

Para especificar lucros

1. Na caixa de diálogo principal Árvore de decisão, selecione uma variável dependente categórica (nominal, ordinal) com dois ou mais rótulos de valor definidos.
2. Clique em **Opções**.
3. Clique na guia **Lucros**.
4. Clique em **Customizado**.
5. Insira valores de renda e despesa para todas as categorias de variável dependente listadas na grade.

Probabilidades anteriores

Para árvores CRT e QUEST com variáveis dependentes categóricas, é possível especificar probabilidades anteriores de associação ao grupo. **Probabilidades anteriores** são estimativas da frequência relativa geral para cada categoria da variável dependente antes de saber a respeito dos valores das variáveis independentes (preditoras). Usar probabilidades anteriores ajuda a corrigir qualquer crescimento de árvore causado por dados na amostra que não são representantes da população inteira.

Obter de amostra de treinamento (informações a priori empíricas). Use essa configuração se a distribuição de valores de variável dependente no arquivo de dados for representante da distribuição da população. Se estiver usando a validação de amostra de divisão, a distribuição de casos na amostra de treinamento será usada.

Nota: Como os casos são designados aleatoriamente à amostra de treinamento na validação de amostra de divisão, você não saberá com antecedência a distribuição real de casos na amostra de treinamento. Consulte o tópico "Validação" na página 5 para obter mais informações

Igual entre categorias. Use essa configuração se as categorias das variável dependente forem representadas igualmente na população. Por exemplo, se houver quatro categorias, aproximadamente 25% dos casos estarão em cada categoria.

Customizado. Insira um valor não negativo para cada categoria da variável dependente listada na grade. Os valores podem ser proporções, porcentagens, contagens de frequência, ou quaisquer outros valores que representam a distribuição de valores entre categorias.

Ajustar informações a priori usando custos de classificação errada. Se você definir custos de classificação errada customizados, será possível ajustar probabilidades anteriores com base nesses custos. Consulte o tópico "Custos de classificação errada" na página 9 para obter mais informações

Lucros e rótulos de valor

Essa caixa de diálogo requer rótulos de valor definidos para a variável dependente. Ela não está disponível, a não ser que pelo menos dois valores da variável dependente categórica tenham rótulos de valor definidos.

Para especificar probabilidades anteriores

1. Na caixa de diálogo principal Árvore de decisão, selecione uma variável dependente categórica (nominal, ordinal) com dois ou mais rótulos de valor definidos.
2. Para o método crescente, selecione **CRT** ou **QUEST**.
3. Clique em **Opções**.

4. Clique na guia **Probabilidades anteriores**.

Pontuações

Para CHAID e CHAID Exaustivo com uma variável dependente ordinal, é possível designar escores customizados a cada categoria da variável dependente. Os escores definem a ordem e a distância entre as categorias da variável dependente. É possível usar escores para aumentar ou diminuir a distância relativa entre valores ordinais ou mudar a ordem dos valores.

- **Usar ranqueamento ordinal para cada categoria.** A categoria mais baixa da variável dependente é designada a um escore de 1, a próxima categoria mais alta é designada a um escore de 2, e assim por diante. Este é o padrão.
- **Customizado.** Insira um valor de escore numérico para cada categoria da variável dependente listada na grade.

Exemplo

Tabela 3. Valores de escore customizado.

Rótulo de valor	Valor original	Pontuação
Não qualificado	1	1
Qualificação manual	2	4
Escrita	3	4.5
Professional	4	7
Gerenciamento	5	6

- Os escores aumentam a distância relativa entre *Não qualificado* e *Qualificação manual* e diminuem a distância relativa entre *Não qualificado* e *Administrativo*.
- Os escores invertem a ordem de *Gerenciamento* e *Profissional*.

Escores e rótulos de valor

Essa caixa de diálogo requer rótulos de valor definidos para a variável dependente. Ela não está disponível, a não ser que pelo menos dois valores da variável dependente categórica tenham rótulos de valor definidos.

Para especificar escores

1. Na caixa de diálogo principal Árvore de decisão, selecione uma variável dependente ordinal com dois ou mais rótulos de valor definidos.
2. Para o método crescente, selecione **CHAID** ou **CHAID exaustivo**.
3. Clique em **Opções**.
4. Clique na guia **Escores**.

Valores ausentes

A guia Valores omissos controla o tratamento de valores de variáveis independentes nominais, omissos de usuário (preditoras).

- O tratamento de valores de variáveis independentes omissos de usuário ordinais e de escala varia entre os métodos crescentes.
- O tratamento de variáveis dependentes nominais é especificado na caixa de diálogo Categorias. Consulte o tópico "Selecionando categorias" na página 4 para obter mais informações
- Para variáveis dependentes ordinais e de escala, os casos com valores de variáveis dependentes omissos do sistema ou omissos de usuário são sempre excluídos.

Tratar como valores omissos. Os valores omissos de usuário são tratados como valores omissos do sistema. O tratamento de valores omissos do sistema varia entre métodos crescentes.

Tratar como valores válidos. Os valores omissos de usuário de variáveis independentes nominais são tratados como valores ordinários no crescimento e classificação da árvore.

Regras dependentes de método

Se alguns, mas não todos os valores de variáveis independentes forem omissos do sistema ou de usuário:

- Para CHAID e CHAID Exaustivo, os valores de variáveis independentes omissos do sistema e de usuário estão incluídos na análise como uma categoria única, combinada. Para variáveis independentes de escala e ordinais, os algoritmos primeiro geram categorias usando valores válidos e, em seguida, decidem se mesclarão a categoria omissa com sua categoria mais similar (válida) ou se a manterão como uma categoria separada.
- Para CRT e QUEST, os casos com valores de variáveis independentes omissos são excluídos do processo de crescimento da árvore, mas são classificados usando substitutos, se os substitutos estiverem incluídos no método. Se os valores omissos de usuário nominais forem tratados como omissos, eles também serão tratados dessa maneira. Consulte o tópico "Substitutos" na página 8 para obter mais informações

Para especificar tratamento omisso de usuário nominal, independente

1. Na caixa de diálogo principal Árvore de decisão, selecione pelo menos uma variável independente nominal.
2. Clique em **Opções**.
3. Clique na guia **Valores omissos**.

Salvando informações de modelo

É possível salvar informações do modelo como variáveis no arquivo de dados de trabalho e também é possível salvar o modelo inteiro em formato XML (PMML) em um arquivo externo.

Variáveis salvas

Número de nó terminal. O nó terminal ao qual cada caso é designado. O valor é o número de nó da árvore.

Valor predito. A classe (grupo) ou valor para a variável dependente predita pelo modelo.

Probabilidades preditas. A probabilidade associada à predição do modelo. Uma variável é salva para cada categoria da variável dependente. Não disponível para variáveis dependentes de escala.

Designação de amostra (treinamento/teste). Para validação de amostra de divisão, essa variável indica se um caso foi usado na amostra de treinamento ou de teste. O valor é 1 para a amostra de treinamento e 0 para a amostra de teste. Não disponível, a menos que você tenha selecionado a validação de amostra de divisão. Consulte o tópico "Validação" na página 5 para obter mais informações

Exportar modelo de árvore como XML

É possível salvar o modelo de árvore inteiro em formato XML (PMML). É possível usar esse arquivo de modelo para aplicar as informações de modelo a outros arquivos de dados para propósitos de escoragem.

Amostra de treinamento. Grava o modelo no arquivo especificado. Para árvores validadas de amostra de divisão, esse é o modelo para a amostra de treinamento.

Amostra de teste. Grava o modelo para a amostra de teste no arquivo especificado. Não disponível, a menos que você tenha selecionado a validação de amostra de divisão.

Saída

As opções de saída disponíveis dependem do método crescente, do nível de medição da variável dependente e de outras configurações.

Exibição em Árvore

É possível controlar a aparência inicial da árvore ou suprimir completamente a exibição em árvore.

Árvore. Por padrão, o diagrama de árvore está incluído na saída exibida no Visualizador. Cancele a seleção dessa opção para excluir o diagrama de árvore da saída.

Exibição. Essas opções controlam a aparência inicial do diagrama de árvore no Visualizador. Todos esses atributos também podem ser modificados editando a árvore gerada.

- **Orientação.** A árvore pode ser exibida de cima para baixo com o nó raiz na parte superior, da esquerda para a direita ou da direita para a esquerda.
- **Conteúdos do nó.** Os nós podem exibir tabelas, gráficos ou ambos. Para variáveis dependentes categóricas, as tabelas exibem as contagens de frequência e porcentagens, e os gráficos são gráficos de barras. Para variáveis dependentes de escala, as tabelas exibem médias, desvios padrão, número de casos e valores preditos, e os gráficos são histogramas.
- **Escala.** Por padrão, as árvores grandes são automaticamente diminuídas em uma tentativa de ajustar a árvore na página. É possível especificar uma porcentagem de escala customizada de até 200%.
- **Estatísticas de variável independente.** Para CHAID e CHAID Exaustivo, as estatísticas incluem o valor *F* (para variáveis dependentes de escala) ou valor qui-quadrado (para variáveis dependentes categóricas), bem como o valor de significância e graus de liberdade. Para CRT, o valor de melhoria é mostrado. Para QUEST, *F*, valor de significância e graus de liberdade são mostrados para variáveis independentes de escala e ordinais; para variáveis independentes nominais, qui-quadrado, valor de significância e graus de liberdade são mostrados.
- **Definições de nó.** As definições de nó exibem os valores da variável independente usada em cada divisão de nó.

Árvore em formato de tabela. Informações de sumarização para cada nó na árvore, incluindo o número de nó pai, estatísticas de variável independente, valores de variável independente para o nó, média e desvio padrão para variáveis dependentes de escala ou contagens e porcentagens para variáveis dependentes categóricas.

Para controlar a exibição em árvore inicial

1. No diálogo principal Árvore de decisão, clique em **Saída**.
2. Clique na guia **Árvore**.

Estatísticas

As tabelas de estatísticas disponíveis dependem do nível de medição da variável dependente, do método crescente e de outras configurações.

Modelo

Sumarização. A sumarização inclui o método usado, as variáveis incluídas no modelo e as variáveis especificadas mas não incluídas no modelo.

Risco. Estimativa de risco e seu erro padrão. Uma medida de precisão preditiva da árvore.

- Para variáveis dependentes categóricas, a estimativa de risco é a proporção de casos classificados incorretamente após o ajustamento de probabilidades anteriores e custos de classificação errada.
- Para variáveis dependentes de escala, a estimativa de risco está na variância do nó.

Tabela de classificação. Para variáveis dependentes categóricas (nominais, ordinais), esta tabela mostra o número de casos classificados corretamente e incorretamente para cada categoria da variável dependente. Não disponível para variáveis dependentes de escala.

Custo, probabilidade anterior, escore e valores de lucro. Para variáveis dependentes categóricas, essa tabela mostra o custo, a probabilidade anterior, o escore e valores de lucro usados na análise. Não disponível para variáveis dependentes de escala.

Variáveis independentes

Importância para o modelo. Para o método crescente CRT, ranqueia cada variável independente (preditora) de acordo com sua importância para o modelo. Não disponível para os métodos QUEST ou CHAID.

Substitutos por divisão. Para os métodos crescentes CRT e QUEST, se o modelo incluir substitutos, lista substitutos para cada divisão na árvore. Não disponível para métodos CHAID. Consulte o tópico “Substitutos” na página 8 para obter mais informações

Desempenho do nó

Sumarização. Para variáveis dependentes de escala, a tabela inclui o número de nó, o número de casos e o valor médio da variável dependente. Para variáveis dependentes categóricas com lucros definidos, a tabela inclui o número de nó, o número de casos, o lucro médio e valores do ROI (retorno sobre investimento). Não disponível para variáveis dependentes categóricas sem lucros definidos. Consulte o tópico “Lucros” na página 9 para obter mais informações

Por categoria de destino. Para variáveis dependentes categóricas com categorias de destino definidas, a tabela inclui o ganho de porcentagem, a porcentagem de resposta e porcentagem de índice (ganho) por grupo de nós ou de percentis. Uma tabela separada é produzida para cada categoria de destino. Não disponível para variáveis dependentes de escala ou variáveis dependentes categóricas sem categorias de destino definidas. Consulte o tópico “Selecionando categorias” na página 4 para obter mais informações

Linhas. As tabelas de desempenho do nó podem exibir resultados por nós terminais, percentis ou ambos. Se você selecionar ambos, duas tabelas serão produzidas para cada categoria de destino. As tabelas de percentis exibem valores acumulativos para cada percentil, com base na ordenação.

Aumento de percentil. Para tabelas de percentis, é possível selecionar o aumento do percentil: 1, 2, 5, 10, 20 ou 25.

Exibir estatísticas acumulativas. Para tabelas de nós terminais, exibe colunas adicionais em cada tabela com resultados acumulativos.

Para selecionar saída de estatísticas

1. No diálogo principal Árvore de decisão, clique em **Saída**.
2. Clique na guia **Estatísticas**.

Gráficos

Os gráficos disponíveis dependem do nível de medição da variável dependente, do método crescente e outras configurações.

Importância da variável independente para o modelo. Gráfico de barras de importância do modelo por variável independente (preditor). Disponível apenas com o método crescente CRT.

Desempenho do nó

Ganho. Ganho é a porcentagem do total de casos na categoria de destino em cada nó, calculada como: (destino do nó *n* / destino do total *n*) x 100. O gráfico de ganhos é um gráfico de linha de ganhos de percentis acumulativos, calculados como: (destino de percentil acumulativo *n* / destino total *n*) x 100. Um gráfico de linha separado é produzido para cada categoria de destino. Disponível somente para variáveis dependentes categóricas com categorias de destino definidas. Consulte o tópico "Selecionando categorias" na página 4 para obter mais informações

O gráfico de ganhos plota os mesmos valores que você veria na coluna *Percentual de ganho* na tabela de ganhos para percentis, que também relata valores acumulativos.

Index. O índice é a razão da porcentagem de resposta do nó para a categoria de destino comparada com a porcentagem de resposta da categoria de destino geral para a amostra inteira. O gráfico de índice é um gráfico de linha de valores de índice de percentil acumulativo. Disponível apenas para variáveis dependentes categóricas. O índice de percentil acumulativo é calculado como: (percentual de respostas de percentil acumulativo / percentual do total de respostas) x 100. Um gráfico separado é produzido para cada categoria de destino e categorias de destino devem ser definidas.

O gráfico de índice plota os mesmos valores que você veria na coluna *Índice* na tabela de ganhos para percentis.

Resposta. A porcentagem de processos no nó na categoria de destino especificada. O gráfico de resposta é um gráfico de linha de resposta de percentil acumulativo, calculado como: (destino de percentil acumulativo *n* / total de percentis acumulativos *n*) x 100. Disponível somente para variáveis dependentes categóricas com categorias de destino definidas.

O gráfico de resposta plota os mesmos valores que você veria na coluna *Resposta* na tabela de ganhos para percentis.

Média. Gráfico de linha de valores médios de percentil acumulativo para a variável dependente. Disponível apenas para variáveis dependente de escala.

Média de lucro. Gráfico de linha de média de lucro acumulativo. Disponível somente para variáveis dependentes categóricas com lucros definidos. Consulte o tópico "Lucros" na página 9 para obter mais informações

O gráfico de média de lucro plota os mesmos valores que você veria na coluna *Lucro* na tabela de sumarização de ganhos para percentis.

Retorno sobre investimento (ROI). Gráfico de linha de ROI acumulativo (retorno sobre investimento). O ROI é calculado como a razão de lucros para despesas. Disponível somente para variáveis dependentes categóricas com lucros definidos.

O gráfico ROI plota os mesmos valores que você veria na coluna *ROI* na tabela sumarização de ganhos para percentis.

Aumento de percentil. Para todos os gráficos de percentil, essa configuração controla os incrementos de percentil exibidos no gráfico: 1, 2, 5, 10, 20 ou 25.

Para selecionar a saída de gráfico

1. No diálogo principal Árvore de decisão, clique em **Saída**.

2. Clique na guia **Gráficos**.

Regras de seleção e escoragem

A guia Regras permite gerar regras de seleção ou classificação/predição no formato de sintaxe de comando, SQL ou texto simples (inglês simples). É possível exibir essas regras no Visualizador e/ou salvar as regras em um arquivo externo.

Sintaxe. Controla o formato das regras de seleção na saída exibida no Visualizador e das regras de seleção salvas em um arquivo externo.

- **IBM® SPSS Statistics.** Linguagem da sintaxe de comando. As regras são expressas como um conjunto de comandos que definem uma condição do filtro que pode ser usada para selecionar subconjuntos de casos ou como instruções `COMPUTE` que podem ser usadas para escorar casos.
- **SQL.** As regras SQL padrão são geradas para selecionar ou extrair registros de um banco de dados ou designar valores a esses registros. As regras SQL geradas não incluem nenhum nome de tabela ou outras informações de origem de dados.
- **Texto simples.** Pseudocódigo de inglês simples. As regras são expressas como um conjunto de instruções lógicas "if...then" que descrevem as classificações ou predições do modelo para cada nó. As regras nesse formato podem usar rótulos definidos de variável e de valor ou nomes de variáveis e valores de dados.

Tipo. Para regras do IBM SPSS Statistics e de SQL, controla o tipo de regras geradas: regras de seleção ou de escoragem.

- **Designar valores a casos.** As regras podem ser usadas para designar as predições do modelo a casos que atendem aos critérios de associação do nó. É gerada uma regra separada para cada nó que atende aos critérios de associação do nó.
- **Selecionar casos.** As regras podem ser usadas para selecionar casos que atendem aos critérios de associação do nó. Para regras do IBM SPSS Statistics e de SQL, é gerada uma única regra para selecionar todos os casos que atendem aos critérios de seleção.

Incluir substitutos em regras do IBM SPSS Statistics e de SQL. Para CRT e QUEST, é possível incluir preditores substitutos do modelo nas regras. As regras que incluem substitutos podem ser bastante complexas. Em geral, se você apenas deseja derivar informações conceituais sobre sua árvore, exclua os substitutos. Se alguns casos tiverem dados da variável independente (preditora) incompletos e você desejar regras que imitam sua incluir, inclua substitutos. Consulte o tópico “Substitutos” na página 8 para obter mais informações

Nós. Controla o escopo das regras geradas. É gerada uma regra separada para cada nó incluído no escopo.

- **Todos os nós terminais.** Gera regras para cada nó terminal.
- **Melhores nós terminais.** Gera regras para os principais *n* nós terminais com base em valores de índice. Se o número exceder o número de nós terminais na árvore, as regras serão geradas para todos os nós terminais. (Consulte a nota abaixo.)
- **Melhores nós terminais até uma porcentagem especificada de casos.** Gera regras para nós terminais para a porcentagem *n* principal de casos com base em valores de índice. (Consulte a nota abaixo.)
- **Nós terminais cujo valor de índice atende ou excede um valor de corte.** Gera regras para todos os nós terminais com um valor de índice maior ou igual ao valor especificado. Um valor de índice maior que 100 significa que a porcentagem de casos na categoria de destino nesse nó excede a porcentagem no nó raiz. (Consulte a nota abaixo.)
- **Todos os nós.** Gera regras para todos os nós.

Nota 1: A seleção de nó baseada em valores de índice está disponível apenas para variáveis dependentes categóricas com categorias de destino definidas. Se você tiver especificado várias categorias de destino, será gerado um conjunto separado de regras para cada categoria de destino.

Nota 2: Para regras do IBM SPSS Statistics e de SQL para selecionar casos (não regras para designar valores), **Todos os nós** e **Todos os nós terminais** gerarão efetivamente uma regra que seleciona todos os casos usados na análise.

Exportar regras para um arquivo. Salva as regras em um arquivo de texto externo.

Também é possível gerar e salvar regras de seleção ou de escoragem interativamente, com base em nós selecionados no modelo de árvore final. Consulte o tópico "Seleção de caso e regras de escoragem" na página 21 para obter mais informações

Nota: Se você aplicar regras no formato de sintaxe de comando a outro arquivo de dados, esse arquivo de dados deverá conter variáveis com os mesmos nomes que as variáveis independentes incluídas no modelo final, medidas na mesma métrica, com os mesmos valores omissos definidos pelo usuário (se houver).

Para especificar regras de seleção ou de escoragem

1. No diálogo principal Árvore de decisão, clique em **Saída**.
2. Clique na guia **Regras**.

Capítulo 2. Editor de árvore

Com o Editor de árvore, é possível:

- Ocultar e mostrar ramificações de árvore selecionadas.
- Controlar a exibição de conteúdo do nó, estatísticas exibidas em divisões do nó e outras informações.
- Mudar as cores do nó, do plano de fundo, da borda, do gráfico e da fonte.
- Mudar o estilo e tamanho da fonte.
- Mudar o alinhamento da árvore.
- Selecionar subconjuntos de casos para análise adicional com base em nós selecionados.
- Criar e salvar regras para casos de seleção e escoragem com base em nós selecionados.

Para editar um modelo de árvore:

1. Clique duas vezes no modelo de árvore na janela Visualizador.
   ou
2. No menu Editar ou no menu pop-up de clique com o botão direito, escolha:
   Editar conteúdo > **Em janela separada**

Ocultando e mostrando nós

Para ocultar (reduzir) todos os nós filhos em uma ramificação abaixo de um nó pai:

1. Clique no sinal de menos (-) na caixa pequena abaixo do canto inferior direito do nó pai.
   Todos os nós abaixo do nó pai nessa ramificação ficarão ocultos.
   Para mostrar (expandir) os nós filhos em uma ramificação abaixo de um nó pai:
2. Clique no sinal de mais (+) na caixa pequena abaixo do canto inferior direito do nó pai.

Nota: Ocultar os nós filhos em uma ramificação não é igual a podar uma árvore. Se desejar podar uma árvore, deve-se solicitar a poda antes da criação da árvore, e as ramificações podadas não serão incluídas na árvore final. Consulte o tópico “Podando árvores” na página 8 para obter mais informações

Selecionando vários nós

É possível selecionar casos, gerar regras de escoragem e seleção e executar outras ações com base nos nós selecionados atualmente. Para selecionar vários nós:

1. Clique em um nó que você deseja selecionar.
2. Ctrl-clique nos outros nós que você deseja selecionar.

É possível selecionar vários nós irmãos e/ou nós pais em uma ramificação e nós filhos em outra ramificação. No entanto, não é possível usar a seleção múltipla em um nó pai e em um filho/descendente da mesma ramificação do nó.

Trabalhando com árvores grandes

Os modelos de árvore às vezes podem conter muitos nós e ramificações que fica difícil ou impossível visualizar a árvore inteira em tamanho normal. Existem vários recursos que você pode achar úteis ao trabalhar com árvores grandes:

- **Mapa de árvore.** É possível usar o mapa de árvore, uma versão muito menor e simplificada da árvore, para navegar na árvore e selecionar nós. Consulte o tópico “Mapa de árvore” na página 20 para obter mais informações

- **Escala.** É possível diminuir o zoom e aumentar o zoom mudando a porcentagem de escala para a exibição em árvore. Consulte o tópico "Escalando a exibição de árvore" para obter mais informações
- **Exibição de nó e de ramificação.** É possível tornar uma árvore mais compacta exibindo somente tabelas ou somente gráficos nos nós e/ou suprimindo a exibição de rótulos do nó ou informações de variável independente. Consulte o tópico "Controlando informações exibidas na árvore" na página 21 para obter mais informações

Mapa de árvore

O mapa de árvore fornece uma visualização compactada e simplificada da árvore que pode ser usada para navegar pela árvore e selecionar nós.

Para usar a janela do mapa de árvore:

1. Nos menus do Editor de árvore, escolha:

   Visualizar > **Mapa de árvore**

- O nó selecionado atualmente é destacado no Editor de modelo de árvore e na janela do mapa de árvore.
- A parte da árvore que está atualmente na área de visualização do Editor de modelo de árvore é indicada com um retângulo vermelho no mapa de árvore. Clique com o botão direito e arraste o retângulo para mudar a seção da árvore exibida na área de visualização.
- Se você selecionar um nó no mapa de árvore que não está atualmente na área de visualização do Editor de árvore, a visualização mudará para incluir o nó selecionado.
- A seleção de múltiplos nós funciona da mesma forma no mapa de árvore e no Editor de árvore: Ctrl-clique para selecionar múltiplos nós. Não é possível usar a seleção múltipla em um nó pai e um filho/descendente na mesma ramificação de nó.

Escalando a exibição de árvore

Por padrão, as árvores são automaticamente escaladas para se ajustar na janela Visualizador, o que pode resultar inicialmente em uma grande dificuldade de leitura de algumas árvores. É possível selecionar uma configuração de escala pré-configurada ou inserir seu próprio valor de escala customizado entre 5% e 200%.

Para mudar a escala da árvore:

1. Selecione uma porcentagem de escala da lista suspensa na barra de ferramentas ou insira um valor de porcentagem customizado.

   ou

2. Nos menus do Editor de árvore, escolha:

   Visualizar > **Escala...**

Também é possível especificar um valor de escala antes de criar o modelo de árvore. Consulte o tópico "Saída" na página 13 para obter mais informações

Janela de sumarização de nó

A janela de sumarização de nó fornece uma visualização maior dos nós selecionados. Também é possível usar a janela de sumarização para visualizar, aplicar ou salvar regras de seleção ou de escoragem com base nos nós selecionados.

- Use o menu Visualizar na janela de sumarização de nó para alternar entre visualizações de uma tabela de sumarização, gráfico ou regras.
- Use o menu Regras na janela de sumarização de nó para selecionar o tipo de regras que você deseja ver. Consulte o tópico "Seleção de caso e regras de escoragem" na página 21 para obter mais informações

- Todas as visualizações na janela de sumarização de nó refletem uma sumarização combinada para todos os nós selecionados.

Para usar a janela de sumarização de nó:

1. Selecione os nós no Editor de árvore. Para selecionar vários nós, use Ctrl-clique.
2. Nos menus, escolha:

   Visualizar > **Sumarização**

Controlando informações exibidas na árvore

O menu Opções no Editor de árvore permite controlar a exibição de conteúdos do nó, nomes e estatísticas de variáveis independentes (preditoras), definições de nó e outras configurações. Muitas dessas configurações também podem ser controladas a partir da barra de ferramentas.

Mudando as cores da árvore e fontes de texto

É possível mudar as seguintes cores na árvore:

- Cor da borda do nó, do plano de fundo e do texto
- Cor da ramificação e cor do texto da ramificação
- Cor do plano de fundo da árvore
- Cor de destaque da categoria predita (variáveis dependentes categóricas)
- Cores do gráfico do nó

Também é possível mudar o tipo, estilo e tamanho de todo o texto na árvore.

Nota: Não é possível mudar a cor ou atributos de fonte para nós ou ramificações individuais. As mudanças de cor se aplicam a todos os elementos do mesmo tipo e as mudanças de fonte (diferente de cor) se aplicam a todos os elementos do gráfico.

Para mudar cores e atributos de fonte do texto:

1. Use a barra de ferramentas para mudar os atributos de fonte para a árvore inteira ou cores para diferentes elementos da árvore. (As Dicas de ferramenta descrevem cada controle na barra de ferramentas quando você coloca o cursor do mouse no controle.)

   ou

2. Clique duas vezes em qualquer lugar no Editor de árvore para abrir a janela Propriedades, ou nos menus, escolha:

   Visualizar > **Propriedades**
3. Para borda, ramificação, plano de fundo do nó, categoria predita e plano de fundo da árvore, clique na guia **Cor**.
4. Para cores e atributos da fonte, clique na guia **Texto**.
5. Para cores do gráfico do nó, clique na guia **Gráficos do nó**.

Seleção de caso e regras de escoragem

É possível usar o Editor de árvore para:

- Selecionar subconjuntos de casos com base nos nós selecionados. Consulte o tópico “Filtrando casos” na página 22 para obter mais informações
- Gerar regras de seleção de caso ou regras de escoragem na sintaxe de comando ou formato SQL do IBM SPSS Statistics. Consulte o tópico “Salvando regras de seleção e de escoragem” na página 22 para obter mais informações

Também é possível salvar automaticamente regras com base em vários critérios ao executar o procedimento Árvore de decisão para criar o modelo de árvore. Consulte o tópico “Regras de seleção e escoragem” na página 16 para obter mais informações

Filtrando casos

Se desejar saber mais sobre os casos em um nó específico ou grupo de nós, é possível selecionar um subconjunto de casos para análise adicional com base nos nós selecionados.

1. Selecione os nós no Editor de árvore. Para selecionar vários nós, use Ctrl-clique.
2. Nos menus, escolha:

   Regras > **Filtrar casos...**
3. Insira um nome de variável de filtro. Os casos dos nós selecionados receberão um valor 1 para essa variável. Todos os outros casos receberão um valor 0 e serão excluídos da análise subsequente até que o status do filtro seja mudado.
4. Clique em **OK**.

Salvando regras de seleção e de escoragem

É possível salvar a seleção de caso ou regras de escoragem em um arquivo externo e, em seguida, aplicar essas regras a uma origem de dados diferente. As regras são baseadas nos nós selecionados no Editor de árvore.

Sintaxe. Controla o formato das regras de seleção na saída exibida no Visualizador e das regras de seleção salvas em um arquivo externo.

- **IBM SPSS Statistics.** Linguagem da sintaxe de comando. As regras são expressas como um conjunto de comandos que definem uma condição do filtro que pode ser usada para selecionar subconjuntos de casos ou como instruções `COMPUTE` que podem ser usadas para escorar casos.
- **SQL.** As regras SQL padrão são geradas para selecionar/extrair registros de um banco de dados ou designar valores a esses registros. As regras SQL geradas não incluem nenhum nome de tabela ou outras informações de origem de dados.

Tipo. É possível criar regras de seleção ou de escoragem.

- **Selecionar casos.** As regras podem ser usadas para selecionar casos que atendem aos critérios de associação do nó. Para regras do IBM SPSS Statistics e de SQL, é gerada uma única regra para selecionar todos os casos que atendem aos critérios de seleção.
- **Designar valores a casos.** As regras podem ser usadas para designar as predições do modelo a casos que atendem aos critérios de associação do nó. É gerada uma regra separada para cada nó que atende aos critérios de associação do nó.

Incluir substitutos. Para CRT e QUEST, é possível incluir preditores substitutos do modelo nas regras. As regras que incluem substitutos podem ser bastante complexas. Em geral, se você apenas deseja derivar informações conceituais sobre sua árvore, exclua os substitutos. Se alguns casos tiverem dados da variável independente (preditora) incompletos e você desejar regras que imitam sua incluir, inclua substitutos. Consulte o tópico “Substitutos” na página 8 para obter mais informações

Para salvar a seleção de caso ou regras de escoragem:

1. Selecione os nós no Editor de árvore. Para selecionar vários nós, use Ctrl-clique.
2. Nos menus, escolha:

   Regras > **Exportar...**
3. Selecione o tipo de regras desejado e insira um nome de arquivo.

Nota: Se você aplicar regras no formato de sintaxe de comando a outro arquivo de dados, esse arquivo de dados deverá conter variáveis com os mesmos nomes que as variáveis independentes incluídas no modelo final, medidas na mesma métrica, com os mesmos valores omissos definidos pelo usuário (se

houver).

Avisos

Essas informações foram desenvolvidas para produtos e serviços oferecidos nos Estados Unidos. Esse material pode estar disponível a partir da IBM em outros idiomas. No entanto, pode ser necessário possuir uma cópia do produto ou da versão do produto nesse idioma para acessá-lo.

É possível que a IBM não ofereça produtos, serviços ou recursos discutidos neste documento em outros países. Consulte um representante IBM local para obter informações sobre produtos e serviços disponíveis atualmente em sua área. Qualquer referência a produtos, programas ou serviços IBM não significa que apenas produtos, programas ou serviços IBM possam ser utilizados. Qualquer produto, programa ou serviço funcionalmente equivalente, que não infrinja nenhum direito de propriedade intelectual da IBM poderá ser utilizado em substituição a este produto, programa ou serviço. Entretanto, a avaliação e verificação da operação de qualquer produto, programa ou serviço não IBM são de responsabilidade do Cliente.

A IBM pode ter patentes ou solicitações de patentes pendentes relativas a assuntos tratados nesta publicação. O fornecimento desta publicação não lhe garante direito algum sobre tais patentes. Pedidos de licença podem ser enviados, por escrito, para:

Gerência de Relações Comerciais e Industriais da IBM Brasil
Av. Pasteur, 138-146
CEP 22290-240
Rio de Janeiro, RJ
Brasil

Para pedidos de licença relacionados a informações de DBCS (Conjunto de Caracteres de Byte Duplo), entre em contato com o Departamento de Propriedade Intelectual da IBM em seu país ou envie pedidos de licença, por escrito, para:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

A INTERNATIONAL BUSINESS MACHINES CORPORATION FORNECE ESTA PUBLICAÇÃO "NO ESTADO EM QUE SE ENCONTRA", SEM GARANTIA DE NENHUM TIPO, SEJA EXPRESSA OU IMPLÍCITA, INCLUINDO, MAS NÃO SE LIMITANDO ÀS GARANTIAS IMPLÍCITAS DE NÃO-VIOLAÇÃO, COMERCIALIZAÇÃO OU ADEQUAÇÃO A UM DETERMINADO PROPÓSITO. Alguns países não permitem a exclusão de garantias explícitas ou implícitas em certas transações; portanto, esta instrução pode não se aplicar ao Cliente.

Essas informações podem conter imprecisões técnicas ou erros tipográficos. São feitas alterações periódicas nas informações aqui contidas; tais alterações serão incorporadas em futuras edições desta publicação. A IBM pode, a qualquer momento, aperfeiçoar e/ou alterar o(s) produto(s) e/ou programa(s) descritos nesta publicação, sem aviso prévio.

Qualquer referência nestas informações a websites não IBM são fornecidas apenas por conveniência e não representam de forma alguma um endosso a esses websites. Os materiais contidos nesses websites não fazem parte dos materiais para esse produto IBM e o uso desses websites é de inteira responsabilidade do Cliente.

A IBM por usar ou distribuir as informações fornecidas da forma que julgar apropriada sem incorrer em qualquer obrigação para com o Cliente.

Licenciados deste programa que desejam obter informações sobre o mesmo com o objetivo de permitir: (i) a troca de informações entre programas criados independentemente e outros programas (incluindo este) e (ii) o uso mútuo de informações trocadas, devem entrar em contato com:

Gerência de Relações Comerciais e Industriais da IBM Brasil
Av. Pasteur, 138-146
CEP 22290-240
Rio de Janeiro, RJ
Brasil

Tais informações podem estar disponíveis, sujeitas a termos e condições apropriadas, incluindo em alguns casos o pagamento de uma taxa.

O programa licenciado descrito nesta publicação e todo o material licenciado disponível são fornecidos pela IBM sob os termos do Contrato com o Cliente IBM, do Contrato Internacional de Licença do Programa IBM ou de qualquer outro contrato equivalente.

Os exemplos de dados de desempenho e do Cliente citados são apresentados apenas para propósitos ilustrativos. Resultados de desempenho reais podem variar dependendo das configurações específicas e das condições operacionais.

Informações relativas a produtos não IBM foram obtidas junto aos fornecedores dos respectivos produtos, de seus anúncios publicados ou de outras fontes disponíveis publicamente. A IBM não testou esses produtos e não pode confirmar a precisão de desempenho, compatibilidade nem qualquer outra reivindicação relacionada a produtos não IBM. Perguntas sobre os recursos de produtos não IBM devem ser endereçadas aos fornecedores desses produtos.

Instruções relativas à direção futura ou intento da IBM estão sujeitas a mudança ou retirada sem aviso e representam metas e objetivos apenas.

Estas informações contêm exemplos de dados e relatórios utilizados nas operações diárias de negócios. Para ilustrá-los da forma mais completa possível, os exemplos podem incluir nomes de assuntos, empresas, marcas e produtos. Todos esses nomes são fictícios e qualquer semelhança com pessoas ou empresas reais é mera coincidência.

LICENÇA DE COPYRIGHT:

Estas informações contêm programas de aplicativos de amostra na linguagem fonte, ilustrando as técnicas de programação em diversas plataformas operacionais. O Cliente pode copiar, modificar e distribuir estes programas de amostra sem a necessidade de pagar à IBM, com objetivos de desenvolvimento, utilização, marketing ou distribuição de programas aplicativos em conformidade com a interface de programação de aplicativo para a plataforma operacional para a qual os programas de amostra são criados. Esses exemplos não foram testados completamente em todas as condições. Portanto, a IBM não pode garantir ou implicar a confiabilidade, manutenção ou função destes programas. Os programas de amostra são fornecidos "NO ESTADO EM QUE SE ENCONTRAM", sem garantia de qualquer tipo. A IBM não será responsabilizada por quaisquer danos decorrentes do uso dos programas de amostra.

Cada cópia ou parte destes programas de amostra ou qualquer trabalho derivado deve incluir um aviso de copyright com os dizeres:

© nome de sua empresa) (ano). Partes deste código são derivadas dos Programas de Amostra da IBM Corp.

Marcas comerciais

IBM, o logotipo IBM e ibm.com são marcas comerciais ou marcas registradas da International Business Machines Corp., registradas em muitos países no mundo todo. Outros nomes de produtos e serviços podem ser marcas comerciais da IBM ou de outras empresas. ma lista atual de marcas comerciais da IBM está disponível na web em "Copyright and trademark information" em www.ibm.com/legal/copytrade.shtml.

Adobe, o logotipo Adobe, PostScript e o logotipo PostScript são marcas registradas ou marcas comerciais da Adobe Systems Incorporated nos Estados Unidos e/ou em outros países.

Intel, o logotipo Intel, Intel Inside, o logotipo Intel Inside, Intel Centrino, o logotipo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium são marcas comerciais ou marcas registradas da Intel Corporation ou de suas subsidiárias nos Estados Unidos e em outros países.

Linux é uma marca registrada da Linus Torvalds nos Estados Unidos, e/ou em outros países.

Microsoft, Windows, Windows NT e o logotipo Windows são marcas comerciais da Microsoft Corporation nos Estados Unidos e/ou em outros países.

UNIX é uma marca registrada da The Open Group nos Estados Unidos e em outros países.

Java e todas as marcas comerciais e logotipos baseados em Java são marcas comerciais ou marcas registradas da Oracle e/ou suas afiliadas.

Índice Remissivo

Impresso no Brasil