IBM SPSS Decision Trees 22

Nota
Antes de utilizar esta información y el producto al que da soporte, lea la información de "Avisos" en la página 25.

Product Information

Esta edición se aplica a la versión 22, release 0, modificación 0 de IBM SPSS Statistics y a todos los releases y modificaciones posteriores hasta que se indique lo contrario en ediciones nuevas.

Contenido

Capítulo 1. Creación de árboles de decisión

El procedimiento Árbol de decisión crea un modelo de clasificación basado en árboles y clasifica casos en grupos o pronostica valores de una variable (criterio) dependiente basada en valores de variables independientes (predictores). El procedimiento proporciona herramientas de validación para análisis de clasificación exploratorios y confirmatorios.

El procedimiento se puede utilizar para:

Segmentación. Identifica las personas que pueden ser miembros de un grupo específico.

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

Predicción. Crea reglas y las utiliza para predecir eventos futuros, como la verosimilitud de que una persona cause mora en un crédito o el valor de reventa potencial de un vehículo o una casa.

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

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

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

Ejemplo. Un banco desea categorizar a los solicitantes de créditos en función de si representan o no un riesgo crediticio razonable. Basándose en varios factores, incluyendo las valoraciones del crédito conocidas de clientes anteriores, se puede generar un modelo para pronosticar si es probable que los clientes futuros causen mora en sus créditos.

Un análisis basado en árboles ofrece algunas características atractivas:

- Permite identificar grupos homogéneos con alto o bajo riesgo.
- Facilita la creación de reglas para realizar predicciones sobre casos individuales.

Consideraciones de los datos

Datos. Las variables dependientes e independientes pueden ser:

- *Nominal*. Una variable puede ser tratada como nominal cuando sus valores representan categorías que no obedecen a una clasificación intrínseca. Por ejemplo, el departamento de la compañía en el que trabaja un empleado. Algunos ejemplos de variables nominales son: región, código postal o confesión religiosa.
- *Ordinal*. Una variable puede ser tratada como ordinal cuando sus valores representan categorías con alguna clasificación intrínseca. Por ejemplo, los niveles de satisfacción con un servicio, que abarquen desde muy insatisfecho hasta muy satisfecho. Entre los ejemplos de variables ordinales se incluyen escalas de actitud que representan el grado de satisfacción o confianza y las puntuaciones de evaluación de las preferencias.
- *Escalas*. Una variable puede tratarse como escala (continua) cuando sus valores representan categorías ordenadas con una métrica con significado, por lo que son adecuadas las comparaciones de distancia entre valores. Son ejemplos de variables de escala: la edad en años y los ingresos en dólares.

Ponderaciones de frecuencia Si se encuentra activada la ponderación, las ponderaciones fraccionarias se redondearán al número entero más cercano; de esta manera, a los casos con un valor de ponderación menor que 0,5 se les asignará una ponderación de 0 y, por consiguiente, se verán excluidos del análisis.

Supuestos. Este procedimiento supone que se ha asignado el nivel de medición adecuado a todas las variables del análisis; además, algunas características suponen que todos los valores de la variable dependiente incluidos en el análisis tienen etiquetas de valor definidas.

- **Nivel de medición.** El nivel de medición afecta a los tres cálculos; por lo tanto, todas las variables deben tener asignado el nivel de medición adecuado. De forma predeterminada, se supone que las variables numéricas son de escala y que las variables de cadena son nominales, lo cual podría no reflejar con exactitud el verdadero nivel de medición. Un icono junto a cada variable en la lista de variables identifica el tipo de variable.

Tabla 1. Iconos de nivel de medición.

Icono	Nivel de medición
	Escala
	Nominal
	Ordinal

Puede cambiar de forma temporal el nivel de medición de una variable; para ello, pulse con el botón derecho del ratón en la variable en la lista de variables de origen y seleccione un nivel de medición del menú emergente.

- **Etiquetas de valor.** La interfaz del cuadro de diálogo para este procedimiento supone que o todos los valores no perdidos de una variable dependiente categórica (nominal, ordinal) tienen etiquetas de valor definidas o ninguno de ellos las tiene. Algunas características no estarán disponibles a menos que haya como mínimo dos valores no perdidos de la variable dependiente categórica que tengan etiquetas de valor. Si al menos dos valores no perdidos tienen etiquetas de valor definidas, todos los demás casos con otros valores que no tengan etiquetas de valor se excluirán del análisis.

Para obtener árboles de decisión

1. Seleccione en los menús:
   Analizar > **Clasificar** > **Árbol...**
2. Seleccione una variable dependiente.
3. Seleccionar una o más variables independientes.
4. Seleccione un método de crecimiento.

Si lo desea, puede:

- Cambiar el nivel de medición para cualquier variable de la lista de origen.
- Forzar que la primera variable en la lista de variables independientes en el modelo sea la primera variable de segmentación.
- Seleccionar una variable de influencia que defina cuánta influencia tiene un caso en el proceso de crecimiento de un árbol. Los casos con valores de influencia inferiores tendrán menos influencia, mientras que los casos con valores superiores tendrán más. Los valores de la variable de influencia deben ser valores positivos.
- Validar el árbol.
- Personalizar los criterios de crecimiento del árbol.

- Guardar los números de nodos terminales, valores pronosticados y probabilidades pronosticadas como variables.
- Guardar el modelo en formato XML (PMML).

Campos con un nivel de medición desconocido

La alerta de nivel de medición se muestra si el nivel de medición de una o más variables (campos) del conjunto de datos es desconocido. Como el nivel de medición afecta al cálculo de los resultados de este procedimiento, todas las variables deben tener un nivel de medición definido.

Explorar datos. Lee los datos del conjunto de datos activo y asigna el nivel de medición predefinido en cualquier campo con un nivel de medición desconocido. Si el conjunto de datos es grande, puede llevar algún tiempo.

Asignar manualmente. Abre un cuadro de diálogo que contiene todos los campos con un nivel de medición desconocido. Puede utilizar este cuadro de diálogo para asignar el nivel de medición a esos campos. También puede asignar un nivel de medición en la Vista de variables del Editor de datos.

Como el nivel de medición es importante para este procedimiento, no puede acceder al cuadro de diálogo para ejecutar este procedimiento hasta que se hayan definido todos los campos en el nivel de medición.

Cambio del nivel de medición

1. En la lista de origen, pulse con el botón derecho del ratón en la variable.
2. Seleccione un nivel de medición del menú emergente.

Esto modifica de forma temporal el nivel de medición para su uso en el procedimiento Árbol de decisión.

Métodos de crecimiento

Los métodos de crecimiento disponibles son:

CHAID. Detección automática de interacciones mediante chi-cuadrado (CHi-square Automatic Interaction Detection). En cada paso, CHAID elige la variable independiente (predictora) que presenta la interacción más fuerte con la variable dependiente. Las categorías de cada predictor se funden si no son significativamente distintas respecto a la variable dependiente.

CHAID exhaustivo. Una modificación del CHAID que examina todas las divisiones posibles de cada predictor.

CRT. Árboles de clasificación y regresión (Classification and Regression Trees). CRT divide los datos en segmentos para que sean lo más homogéneos que sea posible respecto a la variable dependiente. Un nodo terminal en el que todos los casos toman el mismo valor en la variable dependiente es un nodo homogéneo y "puro".

QUEST. Árbol estadístico rápido, insesgado y eficiente (Quick, Unbiased, Efficient Statistical Tree). Método rápido y que evita el sesgo que presentan otros métodos al favorecer los predictores con muchas categorías. Sólo puede especificarse QUEST si la variable dependiente es nominal.

Cada método presenta ventajas y limitaciones, entre las que se incluyen:

Tabla 2. Características del método de crecimiento.

Característica	**CHAID***	**CRT**	**QUEST**
Basado en chi-cuadrado**	X		

Tabla 2. Características del método de crecimiento (continuación).

Característica	CHAID*	CRT	QUEST
Variables (predictoras) independientes sustitutas		X	X
Poda de árboles		X	X
División de nodos multinivel	X		
División de nodos binarios		X	X
Variables de influencia	X	X	
Probabilidades previas		X	X
Costes de clasificación errónea	X	X	X
Cálculo rápido	X		X

*Incluye CHAID exhaustivo.

**QUEST también utiliza una medida de chi-cuadrado para variables independientes nominales.

Selección de categorías

Para variables dependientes categóricas (nominales, ordinales), puede:

- Controlar qué categorías se incluirán en el análisis.
- Identificar las categorías objetivo de interés.

Inclusión y exclusión de categorías

Puede limitar el análisis a categorías específicas de la variable dependiente.

- Aquellos casos que tengan valores de la variable dependiente en la lista de exclusión no se incluirán en el análisis.
- Para variables dependientes nominales, también puede incluir en el análisis categorías perdidas del usuario. (De forma predeterminada, las categorías perdidas del usuario se muestran en la lista de exclusión.)

Categorías objetivo

Las categorías seleccionadas (marcadas) se tratarán durante el análisis como las categorías de interés fundamental. Por ejemplo, si persigue identificar a las personas que es más probable que causen mora en un crédito, podría seleccionar como categoría objetivo la categoría "negativa" de valoración del crédito.

- No hay ninguna categoría objetivo predeterminada. Si no se selecciona ninguna categoría, algunas opciones de las reglas de clasificación y algunos resultados relacionados con las ganancias no estarán disponibles.
- Si hay varias categorías seleccionadas, se generarán gráficos y tablas de ganancias independientes para cada una de las categorías objetivo.
- La designación de una o más categorías como categorías objetivo no tiene ningún efecto sobre los resultados de clasificación errónea, modelo de árbol o estimación del riesgo.

Categorías y etiquetas de valor

Este cuadro de diálogo requiere etiquetas de valor definidas para la variable dependiente. No estará disponible a menos que dos valores como mínimo de la variable dependiente categórica tengan etiquetas de valor definidas.

Para incluir/excluir categorías y seleccionar categorías objetivo

1. En el cuadro de diálogo principal Árbol de decisión, seleccione una variable dependiente categórica (nominal, ordinal) con dos o más etiquetas de valor definidas.
2. Pulse **Categorías**.

Validación

La validación permite evaluar la bondad de la estructura de árbol cuando se generaliza para una mayor población. Hay dos métodos de validación disponibles: validación cruzada y validación por división muestral.

Validación cruzada

La validación cruzada divide la muestra en un número de **submuestras**. A continuación, se generan los modelos de árbol, que no incluyen los datos de cada submuestra. El primer árbol se basa en todos los casos excepto los correspondientes al primer pliegue de la muestra; el segundo árbol se basa en todos los casos excepto los del segundo pliegue de la muestra y así sucesivamente. Para cada árbol se calcula el riesgo de clasificación errónea aplicando el árbol a la submuestra que se excluyó al generarse este.

- Se puede especificar un máximo de 25 pliegues de la muestra. Cuanto mayor sea el valor, menor será el número de casos excluidos de cada modelo de árbol.
- La validación cruzada genera un modelo de árbol único y final. La estimación de riesgo mediante validación cruzada para el árbol final se calcula como promedio de los riesgos de todos los árboles.

Validación por división muestral

Con la validación por división muestral, el modelo se genera utilizando una muestra de entrenamiento y después pone a prueba ese modelo con una muestra reservada.

- Puede especificar un tamaño de la muestra de entrenamiento, expresado como un porcentaje del tamaño total de la muestra, o una variable que divida la muestra en muestras de entrenamiento y de comprobación.
- Si utiliza una variable para definir las muestras de entrenamiento y de comprobación, los casos con un valor igual a 1 para la variable se asignarán a la muestra de entrenamiento y todos los demás casos se asignarán a la muestra de comprobación. Dicha variable no puede ser ni la variable dependiente, ni la de ponderación, ni la de influencia ni una variable independiente forzada.
- Los resultados se pueden mostrar tanto para la muestra de entrenamiento como para la de comprobación, o sólo para esta última.
- La validación por división muestral se debe utilizar con precaución en archivos de datos pequeños (archivos de datos con un número pequeño de casos). Si se utilizan muestras de entrenamiento de pequeño tamaño, pueden generarse modelos que no sean significativos, ya que es posible que no haya suficientes casos en algunas categorías para lograr un adecuado crecimiento del árbol.

Para validar un árbol de decisión

1. En el cuadro de diálogo Árboles de decisión principal, pulse en **Validación**.
2. Seleccione **Validación cruzada** o **Validación por división muestral**.

Nota: ambos métodos de validación asignan casos a los grupos muestrales de forma aleatoria. Si desea poder reproducir exactamente los mismos resultados en un análisis subsiguiente, deberá definir la semilla de aleatorización (menú Transformar, Generadores de números aleatorios) antes de ejecutar el análisis la primera vez y, a continuación, restablecer la semilla a dicho valor para el subsiguiente análisis.

Criterios de crecimiento del árbol

Los criterios de crecimiento disponibles pueden depender del método de crecimiento, del nivel de medición de la variable dependiente o de una combinación de ambos.

Límites de crecimiento

La pestaña Límites de crecimiento permite limitar el número de niveles del árbol y controlar el número de casos mínimo para nodos padre y para nodos hijo.

Máxima profundidad de árbol. Controla el número máximo de niveles de crecimiento por debajo del nodo raíz. El ajuste **Automática** limita el árbol a tres niveles por debajo del nodo raíz para los métodos CHAID y CHAID exhaustivo y a cinco niveles para los métodos CRT y QUEST.

Número de casos mínimo. Controla el número de casos mínimo para los nodos. Los nodos que no cumplen estos criterios no se dividen.

- El aumento de los valores mínimos tiende a generar árboles con menos nodos.
- La disminución de dichos valores mínimos generará árboles con más nodos.

Para archivos de datos con un número pequeño de casos, es posible que, en ocasiones, los valores predeterminados de 100 casos para nodos padre y de 50 casos para nodos hijo den como resultado árboles sin ningún nodo por debajo del nodo raíz; en este caso, la disminución de los valores mínimos podría generar resultados más útiles.

Para especificar los límites del crecimiento

1. En el cuadro de diálogo Árbol de decisión principal, pulse en **Criterios**.
2. Pulse en la pestaña **Límites de crecimiento**.

Criterios para CHAID

Para los métodos CHAID y CHAID exhaustivo, puede controlar:

Nivel de significación. Puede controlar el valor de significación para la división de nodos y la fusión de categorías. Para ambos criterios, el nivel de significación predeterminado es igual a 0,05.

- La división de nodos requiere un valor mayor que 0 y menor que 1. Los valores inferiores tienden a generar árboles con menos nodos.
- La fusión de categorías requiere que el valor sea mayor que 0 y menor o igual que 1. Si desea impedir la fusión de categorías, especifique un valor igual a 1. Para una variable independiente de escala, esto significa que el número de categorías para la variable en el árbol final será el número especificado de intervalos (el valor predeterminado es 10). Consulte el tema “Intervalos de escala para el análisis CHAID” en la página 7 para obtener más información.

Estadístico de Chi-cuadrado. Para variables dependientes ordinales, el valor de chi-cuadrado para determinar la división de nodos y la fusión de categorías se calcula mediante el método de la razón de verosimilitud. Para variables dependientes nominales, puede seleccionar el método:

- **Pearson.** Aunque este método ofrece cálculos más rápidos, debe utilizarse con precaución en muestras pequeñas. Éste es el método predeterminado.
- **Razón de verosimilitud.** Este método es más robusto que el de Pearson pero tarda más en realizar los cálculos. Es el método preferido para las muestras pequeñas

Estimación del modelo. Para variables dependientes ordinales y nominales, puede especificar:

- **Número máximo de iteraciones.** El valor predeterminado es 100. Si el árbol detiene su crecimiento porque se ha alcanzado el número máximo de iteraciones, puede que desee aumentar el número máximo o modificar alguno de los demás criterios que controlan el crecimiento del árbol.

- **Cambio mínimo en las frecuencias esperadas de las casillas.** El valor debe ser mayor que 0 y menor que 1. El valor predeterminado es 0,05. Los valores inferiores tienden a generar árboles con menos nodos.

Corregir los valores de significación mediante el método de Bonferroni. Para comparaciones múltiples, los valores de significación para los criterios de división y fusión se corrigen utilizando el método de Bonferroni. Este es el método predeterminado.

Permitir nueva división de las categorías fusionadas dentro de un nodo. A menos que se impida de forma explícita la fusión de categorías, el procedimiento intentará la fusión de las categorías de variables (predictoras) independientes entre sí para generar el árbol más simple que describa el modelo. Esta opción permite al procedimiento volver a dividir las categorías fusionadas si con ello se puede obtener una solución mejor.

Para especificar criterios para CHAID

1. En el cuadro de diálogo Árbol de decisión principal, seleccione **CHAID** o **CHAID exhaustivo** como método de crecimiento.
2. Pulse en **Criterios**.
3. Pulse en la pestaña **CHAID**.

Intervalos de escala para el análisis CHAID

En el análisis CHAID, las variables (predictoras) independientes de escala siempre se categorizan en grupos discretos (por ejemplo, 0–10, 11–20, 21–30, etc.) antes del análisis. Puede controlar el número inicial/máximo de grupos (aunque el procedimiento puede fundir grupos contiguos después de la división inicial):

- **Número fijo.** Todas las variables independientes de escala se categorizan inicialmente en el mismo número de grupos. El valor predeterminado es 10.
- **Personalizado.** Todas las variables independientes de escala se categorizan inicialmente en el número de grupos especificado para esta variable.

Para especificar intervalos para variables independientes de escala

1. En el cuadro de diálogo principal Árbol de decisión, seleccione una o más variables independientes de escala.
2. Para el método de crecimiento, seleccione **CHAID** o **CHAID exhaustivo**.
3. Pulse en **Criterios**.
4. Pulse en la pestaña **Intervalos**.

En los análisis CRT y QUEST, todas las divisiones son binarias y las variables independientes de escala y ordinales se tratan de la misma manera; por lo tanto, no se puede especificar un número de intervalos para variables independientes de escala.

Criterios para CRT

El método de crecimiento CRT procura maximizar la homogeneidad interna de los nodos. El grado en el que un nodo no representa un subconjunto homogéneo de casos es una indicación de **impureza**. Por ejemplo, un nodo terminal en el que todos los casos tienen el mismo valor para la variable dependiente es un nodo homogéneo que no requiere ninguna división más ya que es "puro".

Puede seleccionar el método utilizado para medir la impureza así como la reducción mínima de la impureza necesaria para dividir nodos.

Medida de la impureza. Para variables dependientes de escala, se utilizará la medida de impureza de desviación cuadrática mínima (LSD). Este valor se calcula como la varianza dentro del nodo, corregida para todas las ponderaciones de frecuencia o valores de influencia.

Para variables dependientes categóricas (nominales, ordinales), puede seleccionar la medida de la impureza:

- **Gini.** Se obtienen divisiones que maximizan la homogeneidad de los nodos hijo con respecto al valor de la variable dependiente. Gini se basa en el cuadrado de las probabilidades de pertenencia de cada categoría de la variable dependiente. El valor mínimo (cero) se alcanza cuando todos los casos de un nodo corresponden a una sola categoría. Esta es la medida predeterminada.
- **Binaria.** Las categorías de la variable dependiente se agrupan en dos subclases. Se encuentran las divisiones que separan mejor los dos grupos.
- **Binaria ordinal.** Similar a la regla binaria con la única diferencia de que sólo se pueden agrupar las categorías adyacentes. Esta medida sólo se encuentra disponible para variables dependientes ordinales.

Cambio mínimo en la mejora. Esta es la reducción mínima de la impureza necesaria para dividir un nodo. El valor predeterminado es 0,0001. Los valores superiores tienden a generar árboles con menos nodos.

Para especificar criterios para CRT

1. Para el método de crecimiento, seleccione **CRT**.
2. Pulse en **Criterios**.
3. Pulse en la pestaña **CRT**.

Criterios para QUEST

Para el método QUEST, puede especificar el nivel de significación para la división de nodos. No se puede utilizar una variable independiente para dividir nodos a menos que el nivel de significación sea menor o igual que el valor especificado. El valor debe ser mayor que 0 y menor que 1. El valor predeterminado es 0,05. Los valores más pequeños tenderán a excluir más variables independientes del modelo final.

Para especificar criterios para QUEST

1. En el cuadro de diálogo principal Árbol de decisión, seleccione una variable dependiente nominal.
2. Para el método de crecimiento, seleccione **QUEST**.
3. Pulse en **Criterios**.
4. Pulse en la pestaña **QUEST**.

Poda de árboles

Con los métodos CRT y QUEST, puede evitar el sobreajuste del modelo mediante la **poda** del árbol: el árbol crece hasta que se cumplen los criterios de parada y, a continuación, se recorta de forma automática hasta obtener el subárbol más pequeño basado en la máxima diferencia en el riesgo especificada. El valor del riesgo se expresa en errores estándar. El valor predeterminado es 1. El valor debe ser no negativo. Para obtener el subárbol con el mínimo riesgo, especifique 0.

Para podar un árbol:

1. En el cuadro de diálogo principal Árbol de decisión, para el método de crecimiento, seleccione **CRT** o **QUEST**.
2. Pulse en **Criterios**.
3. Pulse en la pestaña **Poda del árbol**.

La poda del árbol frente a la ocultación de nodos

Cuando se crea un árbol podado, ninguno de los nodos podados del árbol estarán disponibles en el árbol final. Es posible ocultar y mostrar de forma interactiva los nodos hijo en el árbol final, pero no se pueden mostrar los nodos podados durante el proceso de creación del árbol. Consulte el tema Capítulo 2, “Editor del árbol”, en la página 19 para obtener más información.

Sustitutos

CRT y QUEST pueden utilizar **sustitutos** para variables (predictoras) independientes. Para los casos en que el valor de esa variable falte, se utilizarán otras variables independientes con asociaciones muy cercanas a la variable original para la clasificación. A estas variables predictoras alternativas se les denomina sustitutos. Se puede especificar el número máximo de sustitutos que utilizar en el modelo.

- De forma predeterminada, el número máximo de sustitutos es igual al número de variables independientes menos uno. Es decir, para cada variable independiente, se pueden utilizar todas las demás variables independientes como sustitutos.
- Si no desea que el modelo utilice sustitutos, especifique 0 para el número de sustitutos.

Para especificar sustitutos

1. En el cuadro de diálogo principal Árbol de decisión, para el método de crecimiento, seleccione **CRT** o **QUEST**.
2. Pulse en **Criterios**.
3. Pulse en la pestaña **Sustitutos**.

Opciones

Las opciones disponibles pueden depender del método de crecimiento, del nivel de medición de la variable dependiente y de la existencia de etiquetas de valor definidas para los valores de la variable dependiente.

Costes de clasificación errónea

Para las variables dependientes categóricas (nominales, ordinales), los costes de clasificación errónea permiten incluir información referente a las penalizaciones relativas asociadas a una clasificación incorrecta. Por ejemplo:

- El coste de negar crédito a un cliente solvente será diferente al coste de otorgar crédito a un cliente que posteriormente incurra en un incumplimiento.
- El coste de clasificación errónea de una persona con un alto riesgo de dolencias cardíacas como de bajo riesgo es, probablemente, mucho mayor que el coste de clasificar erróneamente a una persona de bajo riesgo como de alto riesgo.
- El coste de realizar un mailing a alguien con poca propensión a responder es probablemente muy bajo, mientras que el coste de no enviar dicho mailing a personas con propensión a responder es relativamente más alto (en términos de pérdida de beneficios).

Costes de clasificación errónea y etiquetas de valor

Este cuadro de diálogo no estará disponible a menos que dos valores como mínimo de la variable dependiente categórica tengan etiquetas de valor definidas.

Para especificar los costes de clasificación errónea

1. En el cuadro de diálogo principal Árbol de decisión, seleccione una variable dependiente categórica (nominal, ordinal) con dos o más etiquetas de valor definidas.
2. Pulse en **Opciones**.
3. Pulse en la pestaña **Costes de clasificación errónea**.
4. Pulse en **Personalizados**.
5. Introduzca uno o más costes de clasificación errónea en la cuadrícula. Los valores deben ser no negativos. (Las clasificaciones correctas, representadas en la diagonal, son siempre 0.)

Rellenar matriz. Es posible que en muchos casos se desee que los costes sean simétricos, es decir, que el coste de clasificar erróneamente A como B sea el mismo que el coste de clasificar erróneamente B como A. Las siguientes opciones le ayudarán a especificar una matriz de costes simétrica:

- **Duplicar triángulo inferior.** Copia los valores del triángulo inferior de la matriz (bajo la diagonal) en las casillas correspondientes del triángulo superior.
- **Duplicar triángulo superior.** Copia los valores del triángulo superior de la matriz (sobre la diagonal) en las casillas correspondientes del triángulo inferior.
- **Usar valores promedio de casillas** Para cada casilla de cada mitad de la matriz, se calcula el promedio de los dos valores (triángulo superior e inferior) y dicho promedio reemplaza ambos valores. Por ejemplo, si el coste de clasificación errónea de A como B es 1, y el coste de clasificación errónea de B como A es 3, esta opción reemplaza ambos valores por el promedio obtenido: (1+3)/2 = 2.

Beneficios

Para las variables dependientes categóricas, puede asignar valores de ingresos y gastos a niveles de la variable dependiente.

- El beneficio se calcula como la diferencia entre ingresos y gastos.
- Los valores de beneficio afectan a los valores del beneficio promedio y ROI (retorno de la inversión) en las tablas de ganancias. No afectan, sin embargo, a la estructura básica del modelo del árbol.
- Los valores de ingresos y gastos deben ser numéricos y se deben estar especificados para todas las categorías de la variable dependiente que aparezcan en la cuadrícula.

Beneficios y etiquetas de valor

Este cuadro de diálogo requiere etiquetas de valor definidas para la variable dependiente. No estará disponible a menos que dos valores como mínimo de la variable dependiente categórica tengan etiquetas de valor definidas.

Para especificar los beneficios

1. En el cuadro de diálogo principal Árbol de decisión, seleccione una variable dependiente categórica (nominal, ordinal) con dos o más etiquetas de valor definidas.
2. Pulse en **Opciones**.
3. Pulse en la pestaña **Beneficios**.
4. Pulse en **Personalizados**.
5. Introduzca los valores de ingresos y gastos para todas las categorías de la variable dependiente que aparecen en la cuadrícula.

Probabilidades previas

Para los árboles CRT y QUEST con variables dependientes categóricas, puede especificar probabilidades previas de pertenencia al grupo. Las **probabilidades previas** son estimaciones de la frecuencia relativa global de cada categoría de la variable dependiente, previas a cualquier conocimiento sobre los valores de las variables (predictoras) independientes. La utilización de las probabilidades previas ayuda a corregir cualquier crecimiento del árbol causado por datos de la muestra que no sean representativos de la totalidad de la población.

Obtener de la muestra de entrenamiento (previas empíricas). Utilice este ajuste si la distribución de los valores de la variable dependiente en el archivo de datos es representativa de la distribución de población. Si se usa validación por división muestral, se utilizará la distribución de los casos en la muestra de entrenamiento.

Nota: como en la validación por división muestral se asignan los casos de forma aleatoria a la muestra de entrenamiento, no podrá conocer de antemano la distribución real de los casos en la muestra de entrenamiento. Consulte el tema “Validación” en la página 5 para obtener más información.

Iguales para todas las categorías. Utilice este ajuste si las categorías de la variable dependiente tienen la misma representación dentro de la población. Por ejemplo, si hay cuatro categorías con aproximadamente el 25% de los casos en cada una de ellas.

Personalizado. Introduzca un valor no negativo para cada categoría de la variable dependiente que aparezca en la cuadrícula. Los valores pueden ser proporciones, porcentajes, frecuencias o cualquier otro valor que represente la distribución de valores entre categorías.

Corregir previas por costes de clasificación errónea. Si define costes de clasificación errónea personalizados, podrá corregir las probabilidades previas basándose en dichos costes. Consulte el tema "Costes de clasificación errónea" en la página 9 para obtener más información.

Beneficios y etiquetas de valor

Este cuadro de diálogo requiere etiquetas de valor definidas para la variable dependiente. No estará disponible a menos que dos valores como mínimo de la variable dependiente categórica tengan etiquetas de valor definidas.

Para especificar probabilidades previas

1. En el cuadro de diálogo principal Árbol de decisión, seleccione una variable dependiente categórica (nominal, ordinal) con dos o más etiquetas de valor definidas.
2. Para el método de crecimiento, seleccione **CRT** o **QUEST**.
3. Pulse en **Opciones**.
4. Pulse en la pestaña **Probabilidades previas**.

Puntuaciones

Para CHAID y CHAID exhaustivo con una variable dependiente ordinal, puede asignar puntuaciones personalizadas a cada categoría de la variable dependiente. Las puntuaciones definen el orden y la distancia entre las categorías de la variable dependiente. Puede utilizar las puntuaciones para aumentar o disminuir la distancia relativa entre valores ordinales o para cambiar el orden de los valores.

- **Utilizar para cada categoría su rango ordinal.** A la categoría inferior de la variable dependiente se le asigna una puntuación de 1, a la siguiente categoría superior se le asigna una puntuación de 2, etc. Este es el método predeterminado.
- **Personalizado.** Introduzca una puntuación numérica para cada categoría de la variable dependiente que aparezca en la cuadrícula.

Ejemplo

Tabla 3. Valores de puntuación personalizados.

Etiqueta de valor	Valor original	Puntuación
No especializado	1	1
Obrero especializado	2	4
Administrativo	3	4,5
Professional	4	7
Directivo	5	6

- Las puntuaciones aumentan la distancia relativa entre *No especializado* y *Obrero especializado* y disminuyen la distancia relativa entre *Obrero especializado* y *Administrativo*.
- Las puntuaciones invierten el orden entre *Directivo* y *Profesional*.

Puntuaciones y etiquetas de valor

Este cuadro de diálogo requiere etiquetas de valor definidas para la variable dependiente. No estará disponible a menos que dos valores como mínimo de la variable dependiente categórica tengan etiquetas de valor definidas.

Para especificar puntuaciones

1. En el cuadro de diálogo principal Árbol de decisión, seleccione una variable dependiente ordinal con dos o más etiquetas de valor definidas.
2. Para el método de crecimiento, seleccione **CHAID** o **CHAID exhaustivo**.
3. Pulse en **Opciones**.
4. Pulse en la pestaña **Puntuaciones**.

Valores perdidos

La pestaña Valores perdidos controla el tratamiento de los valores perdidos del usuario de las variables (predictoras) independientes nominales.

- El tratamiento de los valores perdidos del usuario de las variables independientes ordinales y de escala varía en función del método de crecimiento.
- En el cuadro de diálogo Categorías, se especifica el tratamiento de las variables dependientes nominales. Consulte el tema "Selección de categorías" en la página 4 para obtener más información.
- Para las variables dependientes ordinales y de escala, siempre se excluyen los casos con valores de variables dependientes perdidos del sistema o del usuario.

Tratar como valores perdidos. Los valores perdidos del usuario reciben el mismo tratamiento que los valores perdidos del sistema. El tratamiento de los valores perdidos del sistema varía según el método de crecimiento.

Tratar como valores válidos. Los valores perdidos del usuario de las variables independientes nominales se tratan como valores ordinarios en la clasificación y crecimiento del árbol.

Reglas dependientes del método

Si algunos, pero no todos, los valores de las variables independientes son valores perdidos del sistema o del usuario:

- Para CHAID y CHAID exhaustivo, los valores de las variables independientes perdidos del sistema y del usuario se incluyen en el análisis como una única categoría combinada. Para las variables independientes ordinales y de escala, los algoritmos primero generan categorías utilizando valores válidos y, a continuación, deciden si fundir la categoría de valores perdidos con la categoría (válida) que más se le parece o se mantiene como una categoría separada.
- Para CRT y QUEST, los casos con valores perdidos en variables independientes se excluyen del proceso de crecimiento del árbol pero se clasifican utilizando sustitutos si estos están incluidos en el método. Si los valores perdidos del usuario nominales se tratan como perdidos, también se procesarán de la misma manera. Consulte el tema "Sustitutos" en la página 9 para obtener más información.

Para especificar el tratamiento de los valores perdidos del usuario de variables independientes nominales

1. En el cuadro de diálogo principal Árbol de decisión, seleccione al menos una variable independiente nominal.
2. Pulse en **Opciones**.
3. Pulse en la pestaña **Valores perdidos**.

Almacenamiento de información del modelo

Puede guardar la información sobre el modelo como variables en el archivo de datos de trabajo y, asimismo, puede guardar todo el modelo en formato XML (PMML) en un archivo externo.

Variables guardadas

Número del nodo terminal. Identifica el nodo terminal al que se asigna cada caso. El valor es el número de nodo del árbol.

Valor predicho. La clase (grupo) o valor de la variable dependiente pronosticada por el modelo.

Probabilidades pronosticadas. La probabilidad asociada con la predicción del modelo. Se guarda una variable por cada categoría de la variable dependiente. No disponible para variables dependientes de escala.

Asignación muestral (entrenamiento/comprobación). Para la validación por división muestral, esta variable indica si se ha utilizado un caso en la muestra de entrenamiento o de comprobación. El valor es 1 si la muestra es de entrenamiento y 0 si es de comprobación. No disponible a menos que se haya seleccionado la validación por división muestral. Consulte el tema "Validación" en la página 5 para obtener más información.

Exportar modelo de árbol como XML

Puede guardar todo el modelo del árbol en formato XML (PMML). Puede utilizar este archivo de modelo para aplicar la información del modelo a otros archivos de datos para puntuarlo.

Muestra de entrenamiento. Escribe el modelo en el archivo especificado. Para árboles validados por división muestral, este es el modelo para la muestra de entrenamiento.

Muestra de comprobación. Escribe el modelo para la muestra de comprobación en el archivo especificado. No disponible a menos que se haya seleccionado la validación por división muestral.

Resultados

Las opciones de resultados disponibles dependen del método de crecimiento, del nivel de medición de la variable dependiente y de otros valores de configuración.

Presentación del árbol

Permite controlar el aspecto inicial del árbol o suprimir completamente la presentación del árbol.

Árbol. De forma predeterminada, el diagrama del árbol se incluye en los resultados que se muestran en el Visor. Desactive la selección (quite la marca) de esta opción para excluir el diagrama de árbol de los resultados.

Representación. Estas opciones controlan el aspecto inicial del diagrama de árbol en el Visor. Todos estos atributos también se pueden modificar editando el árbol generado.

- **Orientación.** El árbol se puede mostrar de arriba a abajo con el nodo raíz situado en la parte superior, de izquierda a derecha, o de derecha a izquierda.
- **Contenidos de los nodos.** Los nodos pueden mostrar tablas, gráficos o ambos. Para variables dependientes categóricas, las tablas muestran recuentos y porcentajes de frecuencia, y los gráficos son diagramas de barras. Para variables dependientes de escala, las tablas muestran medias, desviaciones estándar, número de casos y valores pronosticados, y los gráficos son histogramas.
- **Escalas.** De forma predeterminada, los árboles grandes se reducen de forma automática para intentar ajustar el árbol a la página. Puede especificar un porcentaje de escala personalizado de hasta el 200%.

- **Estadísticos de las variables independientes.** Para CHAID y CHAID exhaustivo, los estadísticos incluyen el valor *F* (para variables dependientes de escala) o el valor chi-cuadrado (para variables dependientes categóricas) así como el valor de significación y los grados de libertad. Para CRT, se muestra el valor de mejora. Para QUEST, se muestra el valor *F*, el valor de significación y los grados de libertad para las variables independientes ordinales y de escala; para las variables independientes nominales, se muestra el valor chi-cuadrado, el valor de significación y los grados de libertad.
- **Definiciones de los nodos.** Las definiciones de nodos muestran el valor o valores de la variable independiente utilizados en cada división de nodos.

Árbol en formato de tabla. Información de resumen para cada nodo del árbol, incluyendo el número del nodo padre, los estadísticos de las variables independientes, el valor o valores de las variables independientes para el nodo, la media y la desviación estándar para variables dependientes de escala, o los recuentos y porcentajes para variables dependientes categóricas.

Para controlar la presentación inicial del árbol

1. En el cuadro de diálogo Árbol de decisión principal, pulse en **Resultados**.
2. Pulse en la pestaña **Árbol**.

Estadísticas

Las tablas de estadísticos disponibles dependen del nivel de medición de la variable dependiente, del método de crecimiento y de otros valores de configuración.

Modelo

Resumen. El resumen incluye el método utilizado, las variables incluidas en el modelo y las variables especificadas pero no incluidas en el modelo.

Riesgo. Estimación del riesgo y su error estándar. Una medida de la precisión predictiva del árbol.

- Para variables dependientes categóricas, la estimación de riesgo es la proporción de casos clasificados incorrectamente después de corregidos respecto a las probabilidades previas y los costes de clasificación errónea.
- Para variables dependientes de escala, la estimación de riesgo corresponde a la varianza dentro del nodo.

Tabla de clasificación. Para variables dependientes categóricas (nominales, ordinales), esta tabla muestra el número de casos clasificados correcta e incorrectamente para cada categoría de la variable dependiente. No disponible para variables dependientes de escala.

Valores de costes, probabilidades previas, puntuaciones y beneficios. Para variables dependientes categóricas, esta tabla muestra los valores de costes, probabilidades previas, puntuaciones y beneficios utilizados en el análisis. No disponible para variables dependientes de escala.

Variables independientes

Importancia en el modelo. Para el método de crecimiento CRT, esta opción asigna rangos a cada variable (predictora) independiente de acuerdo con su importancia para el modelo. No disponible para los métodos QUEST o CHAID.

Sustitutos por división. Para los métodos de crecimiento CRT y QUEST, si el modelo incluye sustitutos, se enumeran estos para cada división en el árbol. No disponible para los métodos CHAID. Consulte el tema “Sustitutos” en la página 9 para obtener más información.

Comportamiento del nodo

Resumen. En el caso de variables dependientes de escala, la tabla incluye el número de nodo, el número de casos y el valor de la media de la variable dependiente. En el caso de variables dependientes categóricas con beneficios definidos, la tabla incluye el número de nodo, el número de casos, el beneficio promedio y los valores de ROI (retorno de la inversión). No disponible para variables dependientes categóricas para las que no se hayan definido beneficios. Consulte el tema "Beneficios" en la página 10 para obtener más información.

Por categoría objetivo. Para variables dependientes categóricas con categorías objetivo definidas, la tabla incluye el porcentaje de ganancia, el porcentaje de respuestas y el índice porcentual (elevación) por nodo o grupo de percentiles. Se genera una tabla separada para cada categoría objetivo. No disponible para variables dependientes de escala o categóricas para las que no se hayan definido categorías objetivo. Consulte el tema "Selección de categorías" en la página 4 para obtener más información.

Filas. Las tablas de comportamiento de los nodos pueden mostrar resultados por nodos terminales, por percentiles o por ambos. Si selecciona ambos, se generan dos tablas por cada categoría objetivo. Las tablas de percentiles muestran valores acumulados para cada percentil, basados en el orden.

Incremento del percentil. Para las tablas de percentiles, puede seleccionar el incremento del percentil: 1, 2, 5, 10, 20, ó 25.

Mostrar estadísticos acumulados. Para las tablas de nodos terminales, muestra columnas adicionales en cada tabla con resultados acumulados.

Para seleccionar los resultados de los estadísticos

1. En el cuadro de diálogo Árbol de decisión principal, pulse en **Resultados**.
2. Pulse en la pestaña **Estadísticos**.

Gráficos

Los gráficos disponibles dependen del nivel de medición de la variable dependiente, del método de crecimiento y de otros valores de configuración.

Importancia de la variable independiente en el modelo. Diagrama de barras de la importancia del modelo por variable (predictora) independiente. Disponible sólo con el método de crecimiento CRT.

Comportamiento del nodo

Ganancia. La ganancia es el porcentaje de los casos totales en la categoría objetivo en cada nodo, calculada como: (*n* criterio de nodo / *n* total de criterios) x 100. El gráfico de ganancias es un gráfico de líneas de las ganancias por percentiles acumulados, calculadas como: The gains chart is a line chart of cumulative percentile gains, computed as: (*n* de percentil de criterios acumulados / *n* total de criterios) x 100. Se genera un gráfico de líneas separado para cada categoría objetivo. Disponible sólo para variables dependientes categóricas con categorías objetivo definidas. Consulte el tema "Selección de categorías" en la página 4 para obtener más información.

El gráfico de ganancias representa los mismos valores que se muestran en la columna *Porcentaje de ganancia* en la tabla de ganancias para los percentiles, que también informa de los valores acumulados.

Índice. El índice es la proporción del porcentaje de respuestas en la categoría criterio del nodo en comparación con el porcentaje global de respuestas en la categoría criterio para toda la muestra. El gráfico de índices es un gráfico de líneas que representa los valores de los índices de percentiles acumulados. Disponible sólo para variables dependientes categóricas. El índice de percentiles acumulados se calcula como: (porcentaje de respuestas de percentiles acumulados / porcentaje de respuestas total) x 100. Se genera un gráfico separado para cada categoría objetivo, y las categorías objetivo deben estar definidas.

El gráfico de índices representa los mismos valores que se muestran en la columna *Índice* en la tabla de ganancias para los percentiles.

Respuesta. Porcentaje de casos pertenecientes al nodo que pertenecen a la categoría objetivo especificada. El gráfico de respuestas es un gráfico de líneas de las respuestas por percentiles acumulados, calculado como: (*n* de percentil de criterios acumulados / *n* total de percentiles acumulados) x 100. Disponible sólo para variables dependientes categóricas con categorías objetivo definidas.

El gráfico de respuestas representa los mismos valores que se muestran en la columna *Responde* en la tabla de ganancias para los percentiles.

Media. Gráfico de líneas de los valores de las medias de percentiles acumulados para la variable dependiente. Disponible sólo para variables dependientes de escala.

Beneficio promedio. Gráfico de líneas del beneficio promedio acumulado. Disponible sólo para variables dependientes categóricas con beneficios definidos. Consulte el tema “Beneficios” en la página 10 para obtener más información.

El gráfico de los beneficios promedios representa los mismos valores que se muestran en la columna *Beneficio* en la tabla de resumen de ganancias para los percentiles.

Retorno de la inversión (ROI). Gráfico de líneas de ROI (retorno de la inversión) acumulado. ROI se calcula como la relación entre los beneficios y los gastos. Disponible sólo para variables dependientes categóricas con beneficios definidos.

El gráfico de ROI representa los mismos valores que se muestran en la columna *ROI* en la tabla de resumen de ganancias para los percentiles.

Incremento del percentil. Para todos los gráficos de percentiles, este ajuste controla los incrementos de los percentiles que se muestran en el gráfico: 1, 2, 5, 10, 20, ó 25.

Para seleccionar los resultados de los gráficos

1. En el cuadro de diálogo Árbol de decisión principal, pulse en **Resultados**.
2. Pulse en la pestaña **Gráficos**.

Reglas de selección y puntuación

La pestaña Reglas ofrece la capacidad de generar reglas de selección o clasificación/predicción en forma de sintaxis de comandos, SQL o sólo texto (inglés sin formato). Estas reglas se pueden visualizar en el Visor y/o guardar en un archivo externo.

Sintaxis. Controla la forma de las reglas de selección en los resultados que se muestran en el Visor y de las reglas de selección almacenadas en un archivo externo.

- **IBM® SPSS Statistics.** Lenguaje de sintaxis de comandos. Las reglas se expresan como un conjunto de comandos que definen una condición de filtrado que permite la selección de subconjuntos de casos o como sentencias `COMPUTE` que se pueden utilizar para asignar puntuaciones a los casos.
- **SQL.** Las reglas SQL estándar se generan para seleccionar o extraer registros de una base de datos, o para asignar valores a dichos registros. Las reglas SQL generadas no incluyen nombres de tablas ni ninguna otra información sobre orígenes de datos.
- **Sólo texto.** Pseudocódigo en inglés sin formato. Las reglas se expresan como un conjunto de sentencias lógicas "if...then" que describen las clasificaciones o predicciones del modelo para cada nodo. Las reglas expresadas en esta forma pueden utilizar etiquetas de variable y de valor definidas o nombres de variables y valores de datos.

Tipo. Para reglas SQL y IBM SPSS Statistics, controla el tipo de reglas generadas: reglas de puntuación o de selección.

- **Asignar valores a los casos.** Las reglas se pueden utilizar para asignar las predicciones del modelo a los casos que cumplan los criterios de pertenencia al nodo. Se genera una regla independiente para cada nodo que cumple los criterios de pertenencia.
- **Seleccionar casos.** Las reglas se pueden utilizar para seleccionar aquellos casos que cumplan los criterios de pertenencia al nodo. Para las reglas de IBM SPSS Statistics y de SQL, se genera una única regla para seleccionar todos los casos que cumplan los criterios de selección.

Incluir sustitutos en las reglas SQL y de IBM SPSS Statistics. Para CRT y QUEST, puede incluir predictores sustitutos del modelo en las reglas. Es conveniente tener en cuenta que las reglas que incluyen sustitutos pueden ser bastante complejas. En general, si sólo desea derivar información conceptual sobre el árbol, excluya a los sustitutos. Si algunos casos tienen datos de variables (predictoras) independientes incompletas y desea reglas que imiten a su árbol, entonces deberá incluir a los sustitutos. Consulte el tema "Sustitutos" en la página 9 para obtener más información.

Nodos. Controla el ámbito de las reglas generadas. Se genera una regla distinta para cada nodo incluido en el ámbito.

- **Todos los nodos terminales.** Genera reglas para cada nodo terminal.
- **Mejores nodos terminales.** Genera reglas para los *n* nodos terminales superiores según los valores de índice. Si la cifra supera el número de nodos terminales del árbol, se generan reglas para todos los nodos terminales. (Consulte la siguiente nota.)
- **Mejores nodos terminales hasta un porcentaje de casos especificado.** Genera reglas para nodos terminales para el porcentaje *n* de casos superiores según los valores de índice. (Consulte la siguiente nota.)
- **Nodos terminales cuyo valor del índice alcanza o excede un valor de corte.** Genera reglas para todos los nodos terminales con un valor de índice mayor o igual que el valor especificado. Un valor de índice mayor que 100 significa que el porcentaje de casos en la categoría objetivo en dicho nodo supera el porcentaje del nodo raíz. (Consulte la siguiente nota.)
- **Todos los nodos.** Genera reglas para todos los nodos.

Nota: la selección de nodos basada en los valores de índice sólo está disponible para las variables dependientes categóricas con categorías objetivo definidas. Si ha especificado varias categorías objetivo, se generará un conjunto separado de reglas para cada una de las categorías objetivo.

Nota 2: en el caso de reglas SQL y de IBM SPSS Statistics para la selección de casos (no reglas para la asignación de valores), **Todos los nodos** y **Todos los nodos terminales** generarán de forma eficaz una regla que seleccione todos los casos utilizados en el análisis.

Exportar reglas a un archivo. Guarda las reglas en un archivo de texto externo.

También se pueden generar y guardar, de forma interactiva, reglas de selección o puntuación, basadas en los nodos seleccionados en el modelo del árbol final. Consulte el tema "Reglas de selección de casos y puntuación" en la página 22 para obtener más información.

Nota: si aplica reglas con el formato de sintaxis de comandos a otro archivo de datos, dicho archivo deberá contener variables con los mismos nombres que las variables independientes incluidas en el modelo final, medidas con la misma métrica y con los mismos valores perdidos del usuario (si hubiera).

Para especificar reglas de selección o puntuación

1. En el cuadro de diálogo Árbol de decisión principal, pulse en **Resultados**.
2. Pulse en la pestaña **Reglas**.

Capítulo 2. Editor del árbol

Con el Editor del árbol es posible:

- Ocultar y mostrar ramas seleccionadas del árbol.
- Controlar la presentación del contenido de los nodos, los estadísticos que se muestran en las divisiones de los nodos y otra información.
- Cambiar los colores de los nodos, fondos, bordes, gráficos y fuentes.
- Cambiar el estilo y el tamaño de la fuente.
- Cambiar la alineación de los árboles.
- Seleccionar subconjuntos de casos para realizar análisis más detallados basados en los nodos seleccionados.
- Crear y guardar reglas para la selección y puntuación de casos basadas en los nodos seleccionados.

Para editar un modelo de árbol:

1. Pulse dos veces en el modelo del árbol en la ventana del Visor.

   o
2. En el menú Edición o el menú emergente que aparece al pulsar el botón derecho, seleccione:

   Editar contenido > **En otra ventana**

Ocultación y presentación de nodos

Para ocultar, contraer, todos los nodos hijo en una rama por debajo de un nodo padre:

1. Pulse en el signo menos (-) de la pequeña casilla situada debajo de la esquina derecha inferior del nodo padre.

   Se ocultarán todos los nodos de esa rama situados por debajo del nodo padre.

   Para mostrar, expandir, los nodos hijo en una rama por debajo de un nodo padre:
2. Pulse en el signo más (+) de la pequeña casilla situada debajo de la esquina derecha inferior del nodo padre.

Nota: ocultar los nodos hijo que hay en una rama no es lo mismo que podar un árbol. Si desea un árbol podado, deberá solicitar la poda antes de crear el árbol y las ramas podadas no se incluirán en el árbol final. Consulte el tema “Poda de árboles” en la página 8 para obtener más información.

Selección de varios nodos

Utilizando como base los nodos seleccionados actualmente, es posible seleccionar casos, generar reglas de puntuación y de selección, así como realizar otras acciones. Para seleccionar varios nodos:

1. Pulse en un nodo que desee seleccionar.
2. Mientras mantiene pulsada Ctrl pulse con el ratón en los demás nodos que desee añadir a la selección.

Puede realizar una selección múltiple de nodos hermanos y/o de nodos padre en una rama, y de nodos hijo en otra rama. Sin embargo, no podrá utilizar la selección múltiple en un nodo padre y en un nodo hijo/descendiente de la misma rama del nodo.

Trabajo con árboles grandes

En ocasiones, los modelos de árbol pueden contener tantos nodos y ramas que resulta difícil o imposible ver todo el árbol a tamaño completo. Para ello existen ciertas características que le serán de utilidad a la hora de trabajar con árboles grandes:

- **Mapa del árbol.** Puede utilizar el mapa del árbol, que es una versión más pequeña y simplificada del árbol, para desplazarse por él y seleccionar nodos. Consulte el tema "Mapa del árbol" para obtener más información.
- **Escalamiento.** Puede acercarse o alejarse cambiando el porcentaje de escala para la presentación del árbol. Consulte el tema "Escalamiento de la presentación del árbol" para obtener más información.
- **Presentación de nodos y ramas.** Puede hacer que la presentación de un árbol sea más compacta mostrando sólo tablas o sólo gráficos en los nodos, o desactivando la visualización de las etiquetas de los nodos o la información de las variables independientes. Consulte el tema "Control de la información que se muestra en el árbol" en la página 21 para obtener más información.

Mapa del árbol

El mapa del árbol proporciona una vista compacta y simplificada del árbol que puede utilizar para desplazarse por el árbol y seleccionar nodos.

Para utilizar la ventana del mapa del árbol:

1. En los menús del Editor del árbol, seleccione:

   Ver > **Mapa del árbol**

- El nodo seleccionado actualmente aparece resaltado tanto en el Editor del modelo del árbol como en la ventana del mapa del árbol.
- La parte del árbol que se ve actualmente en el área de presentación del Editor del modelo del árbol aparece indicada con un rectángulo rojo en el mapa del árbol. Pulse con el botón derecho en el rectángulo y arrástrelo para cambiar la sección del árbol que se muestra en el área de presentación.
- Si selecciona un nodo en el mapa del árbol que no aparece actualmente en el área de presentación del Editor del árbol, la vista cambiará para incluir el nodo seleccionado.
- La selección de varios nodos en el mapa del árbol funciona de la misma manera que en el Editor del árbol: mantenga pulsada la tecla Ctrl al mismo tiempo que pulsa el botón del ratón para seleccionar varios nodos. No podrá utilizar la selección múltiple en un nodo padre y en un nodo hijo/descendiente de la misma rama del nodo.

Escalamiento de la presentación del árbol

De forma predeterminada, los árboles se escalan de forma automática para ajustarse a la ventana del Visor, lo que puede dar como resultado que, inicialmente, algunos árboles sean difíciles de leer. Puede seleccionar un ajuste de escala predefinida o introducir su propio valor de escala entre el 5% y el 200%.

Para cambiar la escala del árbol:

1. Seleccione un porcentaje de escala de la lista desplegable situada en la barra de herramientas o introduzca un valor de porcentaje personalizado.

   o

2. En los menús del Editor del árbol, seleccione:

   Ver > **Escala...**

También puede especificar un valor de escala antes de crear el modelo del árbol. Consulte el tema "Resultados" en la página 13 para obtener más información.

Ventana de resumen de nodos

La ventana de resumen de nodos proporciona una vista de mayor tamaño de los nodos seleccionados. También puede utilizar la ventana de resumen para ver, aplicar o guardar las reglas de selección o de puntuación basadas en los nodos seleccionados.

- Utilice el menú Ver de la ventana de resumen de nodos para cambiar entre las vistas de tabla, gráfico o reglas de resumen.
- Utilice el menú Reglas de la ventana de resumen de nodos para seleccionar el tipo de reglas que desea ver. Consulte el tema "Reglas de selección de casos y puntuación" en la página 22 para obtener más información.
- Todas las vistas de la ventana de resumen de nodos reflejan un resumen combinado para todos los nodos seleccionados.

Para utilizar la ventana de resumen de nodos:

1. Seleccione los nodos en el Editor del árbol. Mantenga pulsada la tecla Ctrl al mismo tiempo que pulsa el botón del ratón para seleccionar varios nodos.
2. Seleccione en los menús:

   Ver > **Resumen**

Control de la información que se muestra en el árbol

El menú Opciones del Editor del árbol le permite controlar la presentación del contenido de los nodos, estadísticos y nombres de las variables (predictoras) independientes, definiciones de nodos y otros valores de configuración. Muchos de estos ajustes también se pueden controlar desde la barra de herramientas.

Modificación de las fuentes de texto y los colores del árbol

En los árboles, se pueden modificar los siguientes colores:

- Color del borde, del fondo y del texto de los nodos
- Color de las ramas y del texto de las ramas
- Color del fondo del árbol
- Color de resalte de las categorías pronosticadas (variables dependientes categóricas)
- Colores de los gráficos de los nodos

Asimismo, se puede modificar el tipo, estilo y tamaño de las fuentes de todo el texto del árbol.

Nota: no se puede cambiar el color o los atributos de fuente para nodos o ramas individuales. Los cambios de color se aplican a todos los elementos del mismo tipo, y los cambios de fuente (que no sean el cambio de color) se aplican a todos los elementos del gráfico.

Para modificar los colores y los atributos de la fuente de texto

1. Utilice la barra de herramientas para cambiar los atributos de fuente para todo el árbol o los colores para los distintos elementos de dicho árbol. (Las ayudas contextuales describen todos los controles de la barra de herramientas cuando se sitúa el puntero del ratón sobre ellos.)

   o
2. Pulse dos veces en cualquier lugar del Editor del árbol para abrir la ventana Propiedades, o, en los menús, seleccione:

   Ver > **Propiedades**
3. Para el borde, rama, fondo de los nodos, categoría pronosticada y fondo del árbol, pulse en la pestaña **Color**.
4. Para los colores y atributos de fuente, pulse en la pestaña **Texto**.
5. Para los colores de los gráficos de los nodos, pulse en la pestaña **Gráficos de nodos**.

Reglas de selección de casos y puntuación

Puede utilizar el Editor del árbol para:

- Seleccionar subconjuntos de casos basados en los nodos seleccionados. Consulte el tema "Filtrado de casos" para obtener más información.
- Generar reglas de selección de casos o reglas de puntuación en sintaxis de comandos de IBM SPSS Statistics o formato SQL. Consulte el tema "Almacenamiento de las reglas de selección y puntuación" para obtener más información.

También puede guardar de forma automática reglas basadas en distintos criterios cuando ejecute el procedimiento Árbol de decisión para crear el modelo del árbol. Consulte el tema "Reglas de selección y puntuación" en la página 16 para obtener más información.

Filtrado de casos

Si desea obtener más información sobre los casos de un determinado nodo o de un grupo de nodos, puede seleccionar un subconjunto de casos para realizar un análisis más detallado en los nodos seleccionados.

1. Seleccione los nodos en el Editor del árbol. Mantenga pulsada la tecla Ctrl al mismo tiempo que pulsa el botón del ratón para seleccionar varios nodos.
2. Seleccione en los menús:

   Reglas > **Filtrar casos...**
3. Introduzca un nombre de variable de filtro. Los casos de los nodos seleccionados recibirán un valor igual a 1 para esta variable. Todos los demás casos recibirán un valor igual a 0 y se excluirán del análisis subsiguiente hasta que se modifique el estado del filtro.
4. Pulse en **Aceptar**.

Almacenamiento de las reglas de selección y puntuación

Puede guardar las reglas de selección de casos y puntuación en un archivo externo y, a continuación, aplicar dichas reglas a otro origen de datos. Las reglas están basadas en los nodos seleccionados en el Editor del árbol.

Sintaxis. Controla la forma de las reglas de selección en los resultados que se muestran en el Visor y de las reglas de selección almacenadas en un archivo externo.

- **IBM SPSS Statistics.** Lenguaje de sintaxis de comandos. Las reglas se expresan como un conjunto de comandos que definen una condición de filtrado que permite la selección de subconjuntos de casos o como sentencias `COMPUTE` que se pueden utilizar para asignar puntuaciones a los casos.
- **SQL.** Las reglas SQL estándar se generan para seleccionar o extraer registros de una base de datos, o para asignar valores a dichos registros. Las reglas SQL generadas no incluyen nombres de tablas ni ninguna otra información sobre orígenes de datos.

Tipo. Puede crear reglas de selección o de puntuación.

- **Seleccionar casos.** Las reglas se pueden utilizar para seleccionar aquellos casos que cumplan los criterios de pertenencia al nodo. Para las reglas de IBM SPSS Statistics y de SQL, se genera una única regla para seleccionar todos los casos que cumplan los criterios de selección.
- **Asignar valores a los casos.** Las reglas se pueden utilizar para asignar las predicciones del modelo a los casos que cumplan los criterios de pertenencia al nodo. Se genera una regla independiente para cada nodo que cumple los criterios de pertenencia.

Incluir sustitutos. Para CRT y QUEST, puede incluir predictores sustitutos del modelo en las reglas. Es conveniente tener en cuenta que las reglas que incluyen sustitutos pueden ser bastante complejas. En general, si sólo desea derivar información conceptual sobre el árbol, excluya a los sustitutos. Si algunos

casos tienen datos de variables (predictoras) independientes incompletas y desea reglas que imiten a su árbol, entonces deberá incluir a los sustitutos. Consulte el tema “Sustitutos” en la página 9 para obtener más información.

Para guardar reglas de selección de casos o puntuación:

1. Seleccione los nodos en el Editor del árbol. Mantenga pulsada la tecla Ctrl al mismo tiempo que pulsa el botón del ratón para seleccionar varios nodos.
2. Seleccione en los menús:

   Reglas > **Exportar...**
3. Seleccione el tipo de reglas que desea e introduzca un nombre de archivo.

Nota: si aplica reglas con el formato de sintaxis de comandos a otro archivo de datos, dicho archivo deberá contener variables con los mismos nombres que las variables independientes incluidas en el modelo final, medidas con la misma métrica y con los mismos valores perdidos del usuario (si hubiera).

Avisos

Esta información se ha desarrollado para productos y servicios ofrecidos en EE.UU.

Es posible que IBM no ofrezca los productos, servicios o las características que se describen este documento en otros países. Consulte al representante local de IBM para obtener información sobre los productos y servicios disponibles actualmente en su zona. Las referencias a un programa, producto o servicio de IBM no pretenden afirmar ni implicar que solo se pueda utilizar el producto, programa o servicio de IBM. En su lugar, se puede utilizar cualquier producto, programa o servicio funcionalmente equivalente que no infrinja los derechos de propiedad intelectual de IBM. Sin embargo, es responsabilidad del usuario evaluar y comprobar el funcionamiento de todo producto, programa o servicio que no sea de IBM.

Puede que IBM tenga patentes o solicitudes de patente pendientes que cubran la materia descrita en este documento. Este documento no le otorga ninguna licencia para estas patentes. Puede enviar preguntas acerca de las licencias, por escrito, a:

IBM Director of Licensing
IBM, S.A.
North Castle Drive
Armonk, NY 10504-1785
EE.UU.

Para consultas sobre licencias relacionadas con información de doble byte (DBCS), póngase en contacto con el departamento de propiedad intelectual de IBM en su país o envíe la consulta por escrito a:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japón

El párrafo siguiente no se aplica al Reino Unido ni a ningún otro país donde estas disposiciones sean incompatibles con la legislación vigente: INTERNATIONAL BUSINESS MACHINES CORPORATION PROPORCIONA ESTA PUBLICACIÓN "TAL CUAL" SIN GARANTÍAS DE NINGÚN TIPO, NI EXPLÍCITAS NI IMPLÍCITAS, INCLUIDAS, PERO NO LIMITÁNDOSE A ELLAS, LAS GARANTÍAS IMPLÍCITAS DE NO VULNERACIÓN, COMERCIALIZACIÓN O IDONEIDAD PARA UN FIN CONCRETO. 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 en cualquier momento mejoras o cambios en los productos o programas descritos en esta publicación 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. El material de esos sitios web no forma parte del material de este producto de IBM y el uso de dichos sitios web es responsabilidad del usuario.

IBM puede utilizar o distribuir cualquier información que proporcione en la forma que considere adecuada sin incurrir en ninguna obligación con el usuario.

Los usuarios con licencia de este programa que deseen obtener información sobre éste con el propósito de habilitar: (i) el intercambio de información entre programas creados independientemente y otros programas (incluido este) y (ii) el uso mutuo de la información que se ha intercambiado, deben ponerse en contacto con:

Tel. 901 100 400
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
España

Esta información estará disponible, bajo las condiciones adecuadas, incluyendo en algunos casos el pago de una cuota.

El programa bajo licencia que se describe en este documento y todo el material bajo licencia disponible los proporciona IBM bajo los términos de las Condiciones Generales de IBM, Acuerdo Internacional de Programas Bajo Licencia de IBM o cualquier acuerdo equivalente entre las partes.

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 mediciones se hayan realizado en sistemas en desarrollo y no existe ninguna garantía de que estas mediciones sean las mismas en los sistemas comerciales. Además, es posible que algunas mediciones 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 otros orígenes disponibles públicamente. IBM no ha comprobado estos productos y no puede confirmar la precisión de su rendimiento, compatibilidad ni contemplar ninguna otra reclamación relacionada con los productos que no son 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.

Cada una de las copias, totales o parciales, de estos programas de ejemplo o cualquier trabajo derivado de ellos, debe incluir el siguiente aviso de copyright:

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.

Cada una de las copias, totales o parciales, de estos programas de ejemplo o cualquier trabajo derivado de ellos, debe incluir el siguiente aviso de copyright:

© nombre de la empresa (año). Algunas partes de este código proceden de IBM Corp. Sample Programs.

Marcas comerciales

IBM, el logotipo de IBM e ibm.com son marcas registradas o marcas comerciales registradas de International Business Machines Corp., registrada en muchas jurisdicciones en todo el mundo. Otros nombres de servicios y productos podrían ser marcas registradas de IBM u otras empresas. Hay disponible una lista actual de marcas registradas de IBM en la Web en “Información de marca registrada y copyright en ” at www.ibm.com/legal/copytrade.shtml.

Adobe, el logotipo Adobe, PostScript y el logotipo PostScript son marcas registradas o marcas comerciales de Adobe Systems Incorporated en Estados Unidos y/o otros países.

Intel, el logotipo de Intel, Intel Inside, el logotipo de Intel Inside, Intel Centrino, el logotipo de Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium y Pentium son marcas comerciales o marcas registradas de Intel Corporation o sus filiales en Estados Unidos y otros países.

Linux es una marca registrada de Linus Torvalds en Estados Unidos, otros países o ambos.

Microsoft, Windows, Windows NT, y el logotipo de Windows son marcas comerciales de Microsoft Corporation en Estados Unidos, otros países o ambos.

UNIX es una marca registrada de The Open Group en Estados Unidos y otros países.

Java y todas las marcas registradas y logotipos basados en Java son marcas registradas de Oracle y/o sus filiales.

Índice

Impreso en España