# IBM SPSS Decision Trees 19

*Note*: Before using this information and the product it supports, read the general information under Notices el p. 114.

# Prefacio

IBM® SPSS® Statistics es un sistema global para el análisis de datos. El módulo adicional opcional Árboles de decisión proporciona las técnicas de análisis adicionales que se describen en este manual. El módulo adicional Árboles de decisión se debe utilizar con el sistema básico de SPSS Statistics y está completamente integrado en dicho sistema.

## Acerca de SPSS Inc., an IBM Company

SPSS Inc., an IBM Company, es uno de los principales proveedores globales de software y soluciones de análisis predictivo. La gama completa de productos de la empresa (recopilación de datos, análisis estadístico, modelado y distribución) capta las actitudes y opiniones de las personas, predice los resultados de las interacciones futuras con los clientes y, a continuación, actúa basándose en esta información incorporando el análisis en los procesos comerciales. Las soluciones de SPSS Inc. tratan los objetivos comerciales interconectados en toda una organización centrándose en la convergencia del análisis, la arquitectura de TI y los procesos comerciales. Los clientes comerciales, gubernamentales y académicos de todo el mundo confían en la tecnología de SPSS Inc. como ventaja ante la competencia para atraer, retener y hacer crecer los clientes, reduciendo al mismo tiempo el fraude y mitigando los riesgos. SPSS Inc. fue adquirida por IBM en octubre de 2009. Para obtener más información, visite *http://www.spss.com*.

## Asistencia técnica

El servicio de asistencia técnica está a disposición de todos los clientes de mantenimiento. Los clientes podrán ponerse en contacto con este servicio de asistencia técnica si desean recibir ayuda sobre la utilización de los productos de SPSS Inc. o sobre la instalación en alguno de los entornos de hardware admitidos. Para ponerse en contacto con el servicio de asistencia técnica, consulte el sitio web de SPSS Inc. en *http://support.spss.com* o encuentre a su representante local a través del sitio web *http://support.spss.com/default.asp?refpage=contactus.asp*. Tenga a mano su identificación, la de su organización y su contrato de asistencia cuando solicite ayuda.

## Servicio de atención al cliente

Si tiene cualquier duda referente a la forma de envío o pago, póngase en contacto con su oficina local, que encontrará en el sitio Web en *http://www.spss.com/worldwide*. Recuerde tener preparado su número de serie para identificarse.

## Cursos de preparación

SPSS Inc. ofrece cursos de preparación, tanto públicos como in situ. Todos los cursos incluyen talleres prácticos. Los cursos tendrán lugar periódicamente en las principales ciudades. Si desea obtener más información sobre estos cursos, póngase en contacto con su oficina local que encontrará en el sitio Web en *http://www.spss.com/worldwide*.

## Publicaciones adicionales

Los documentos *SPSS Statistics: Guide to Data Analysis*, *SPSS Statistics: Statistical Procedures Companion* y *SPSS Statistics: Advanced Statistical Procedures Companion*, escritos por Marija Norušis y publicados por Prentice Hall, están disponibles y se recomiendan como material adicional. Estas publicaciones cubren los procedimientos estadísticos del módulo SPSS Statistics Base, el módulo Advanced Statistics y el módulo Regression. Tanto si da sus primeros pasos en el análisis de datos como si ya está preparado para las aplicaciones más avanzadas, estos libros le ayudarán a aprovechar al máximo las funciones ofrecidas por IBM® SPSS® Statistics. Si desea información adicional sobre el contenido de la publicación o muestras de capítulos, consulte el sitio web de la autora: *http://www.norusis.com*

# Contenido

## Parte I: Manual del usuario

## 1 Creación de árboles de decisión 1



## 2 Editor del árbol 39

## *Parte II: Ejemplos*

## *3 Requisitos y supuestos de los datos 51*



## *4 Utilización de árboles de decisión para evaluar riesgos de crédito 59*

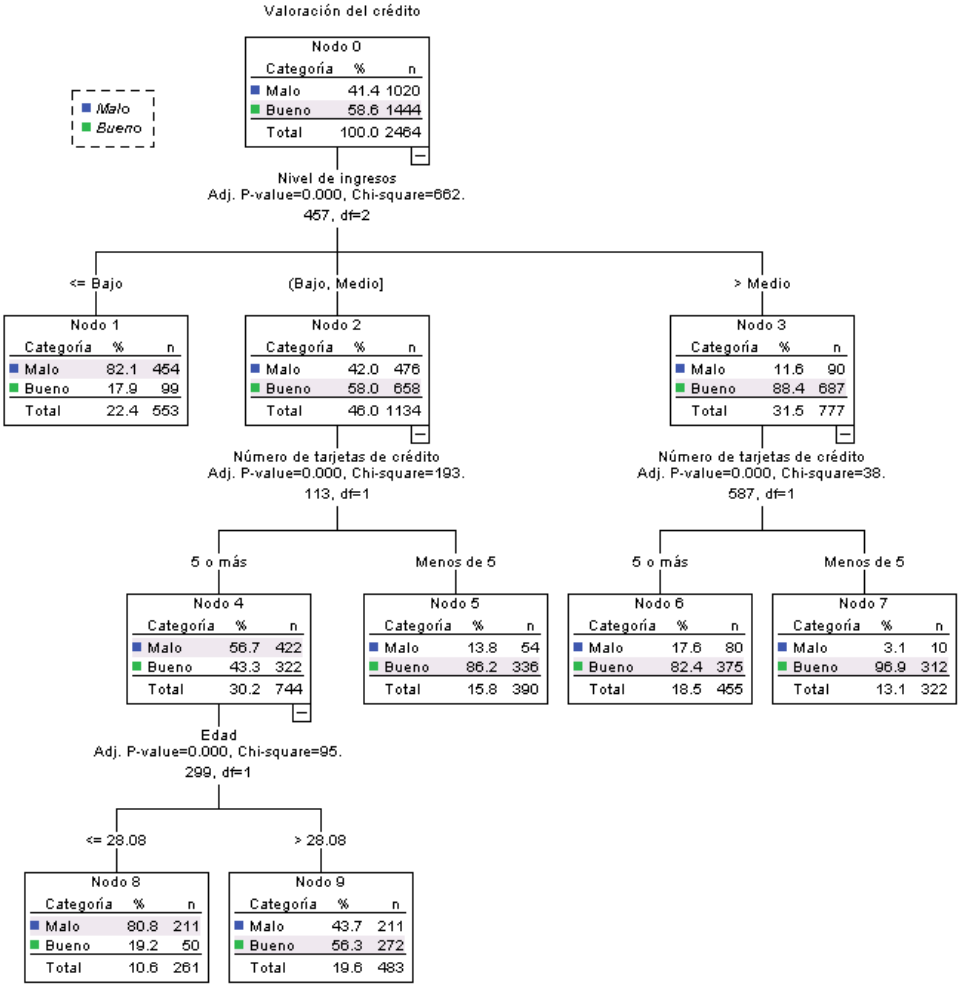
# *Parte I:*
# *Manual del usuario*

Capítulo

# 1

# Creación de árboles de decisión

Figura 1-1
*Árbol 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 pronósticos sobre casos individuales.

### *Consideraciones de los datos*

**Datos.** Las variables dependientes e independientes pueden ser:

- **Nominal.** Una variable se puede tratar como nominal si sus valores representan categorías que no obedecen a una ordenación intrínseca (por ejemplo, el departamento de la empresa en el que trabaja un empleado). Algunos ejemplos de variables nominales son: región, código postal o confesión religiosa.
- **Ordinal.** Una variable puede tratarse como ordinal cuando sus valores representan categorías con alguna ordenación intrínseca (por ejemplo, los niveles de satisfacción con un servicio, que vayan 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.
- **Escala.** 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 medida adecuado a todas las variables del análisis; además, algunas funciones suponen que todos los valores de la variable dependiente incluidos en el análisis tienen etiquetas de valor definidas.

- **Nivel de medida.** El nivel de medida afecta a los tres cálculos; por lo tanto, todas las variables deben tener asignado el nivel de medida adecuado. Por defecto, 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 medida. Un icono junto a cada variable en la lista de variables identifica el tipo de variable.


Escalas


Nominal


Ordinal

Puede cambiar de forma temporal el nivel de medida 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 medida del menú contextual.

- **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 funciones 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***

▶ Seleccione en los menús:
Analizar > Clasificar > Árbol...

Figura 1-2
*Cuadro de diálogo Árbol de decisión*


- Seleccione una variable dependiente.
- Seleccionar una o más variables independientes.
- Seleccione un método de crecimiento.

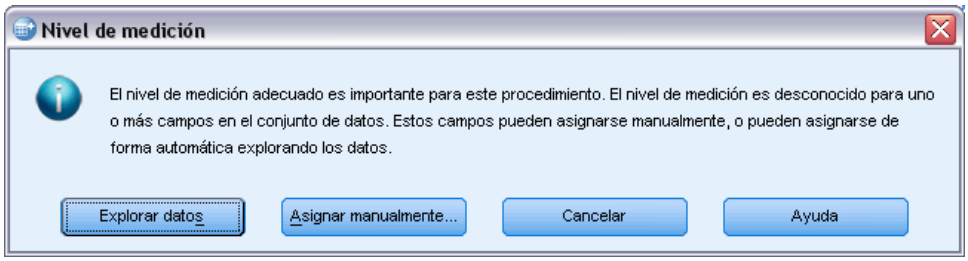
Si lo desea, puede:

- Cambiar el nivel de medida 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 división.
- 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.

Figura 1-3
*Alerta de nivel de medición*




- **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 medida*

▶ En la lista de origen, pulse con el botón derecho del ratón en la variable.

▶ Seleccione un nivel de medida del menú contextual emergente.

Esto modifica de forma temporal el nivel de medida 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. 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:

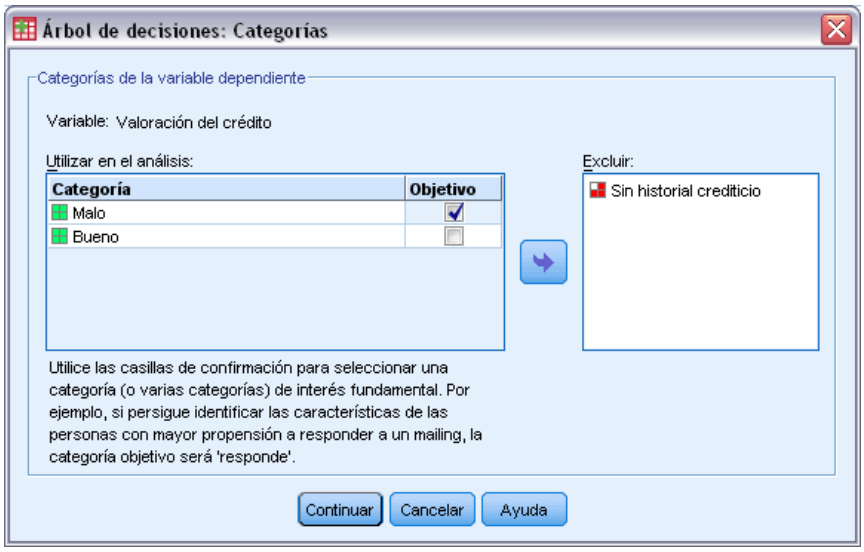
| | CHAID* | CRT | QUEST |
|---|---|---|---|
| Basado en chi-cuadrado** | X | | |
| Variables (predictoras) independientes sustitutas | | X | X |
| Poda de árboles | | X | X |
| División de nodos multinivel | X | | |
| División de nodos binarios | | X | X |
| Variables de influencia | X | X | |
| Probabilidades 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*

Figura 1-4
*Cuadro de diálogo 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 definidas como perdidas por el usuario. (Por defecto, las categorías definidas como perdidas por el 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 por defecto. 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.

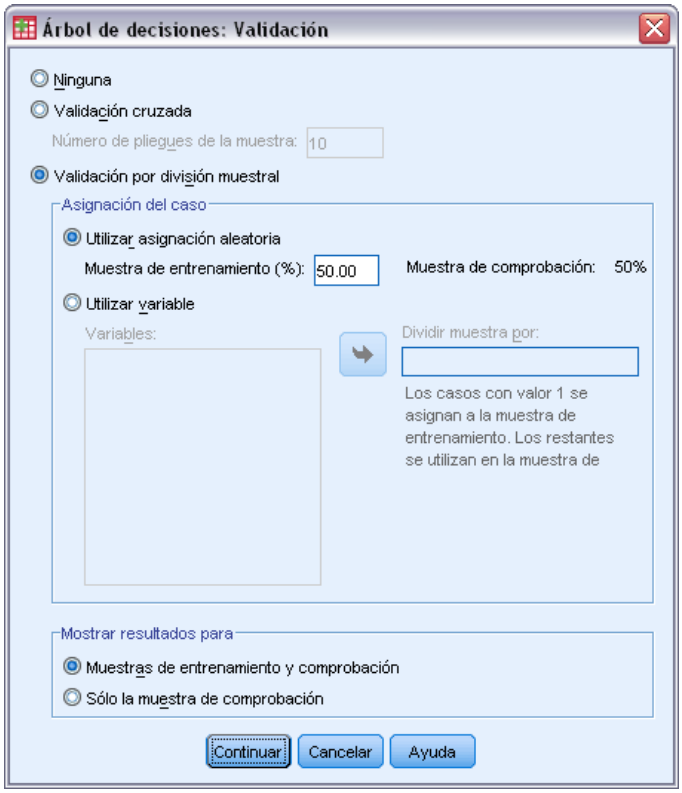
#### *Para incluir/excluir categorías y seleccionar categorías objetivo*

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

▶ Pulse Categorías.

# Validación

Figura 1-5
*Cuadro de diálogo 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 de reserva.

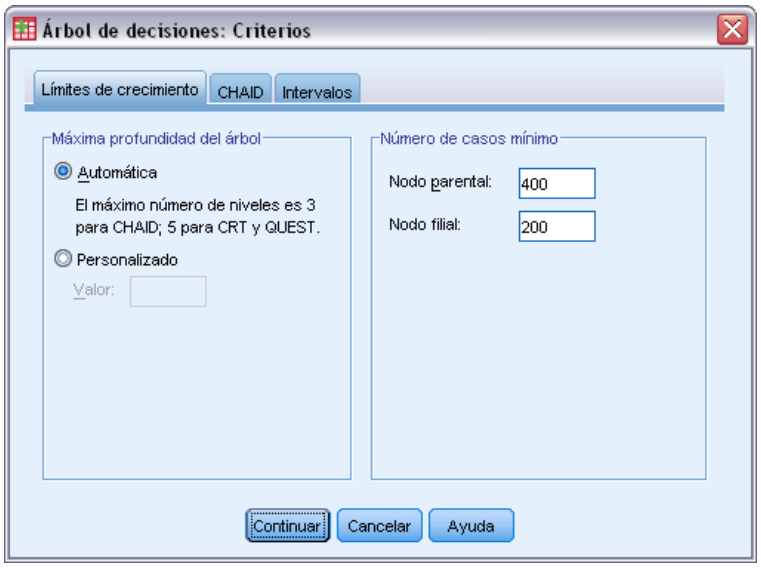
- Puede especificar un tamaño de la muestra de entrenamiento, expresado como un porcentaje del tamaño muestral total, o una variable que divida la muestra en muestras de entrenamiento y de comprobación.
- Si utiliza una variable para definir las muestras de entrenamiento y de comprobación, los casos 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.

## *Criterios de crecimiento del árbol*

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

### *Límites de crecimiento*

Figura 1-6
*Cuadro de diálogo Criterios, pestaña 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 parentales y filiales.

**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.

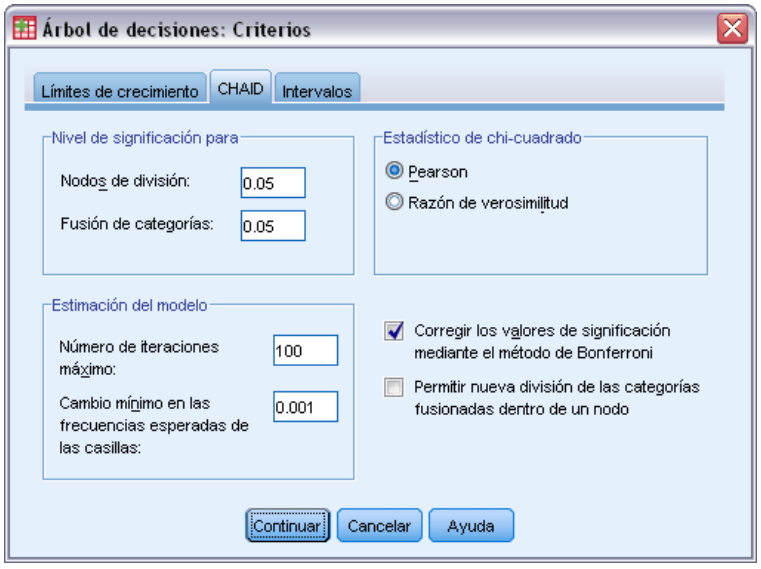
**Número de casos mínimo.** Controla el número de casos mínimo para los nodos. Los nodos que no cumplen estos criterios no se dividen.

- 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 por defecto de 100 casos para nodos parentales y de 50 casos para nodos filiales 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.

## *Criterios para CHAID*

Figura 1-7
*Cuadro de diálogo Criterios, pestaña 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 por defecto 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 por defecto es 10). Si desea obtener más información, consulte el tema Intervalos de escala para el análisis CHAID el p. 12.

**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 por defecto.
- **Cociente de verosimilitudes.** 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:

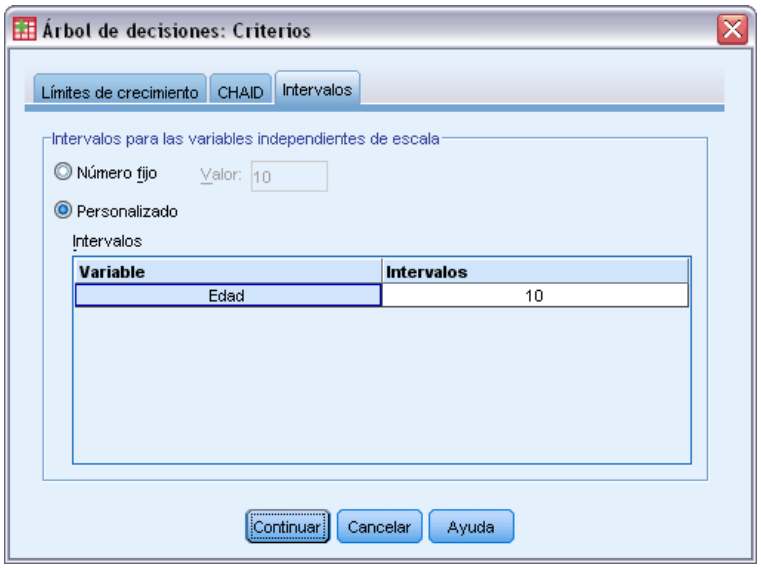
- **Número máximo de iteraciones.** El valor por defecto es 100. Si el árbol detiene su crecimiento porque se ha alcanzado el número máximo de iteraciones, puede que desee aumentar el número máximo o modificar alguno de los demás criterios que controlan el crecimiento del árbol.
- **Cambio mínimo en las frecuencias esperadas de las casillas.** El valor debe ser mayor que 0 y menor que 1. El valor por defecto 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. Ésta es la opción por defecto.

**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.

### *Intervalos de escala para el análisis CHAID*

Figura 1-8
*Cuadro de diálogo Criterios, pestaña Intervalos*




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 por defecto es 10.
- **Personalizado.** Todas las variables independientes de escala se categorizan inicialmente en el número de grupos especificado para esta variable.

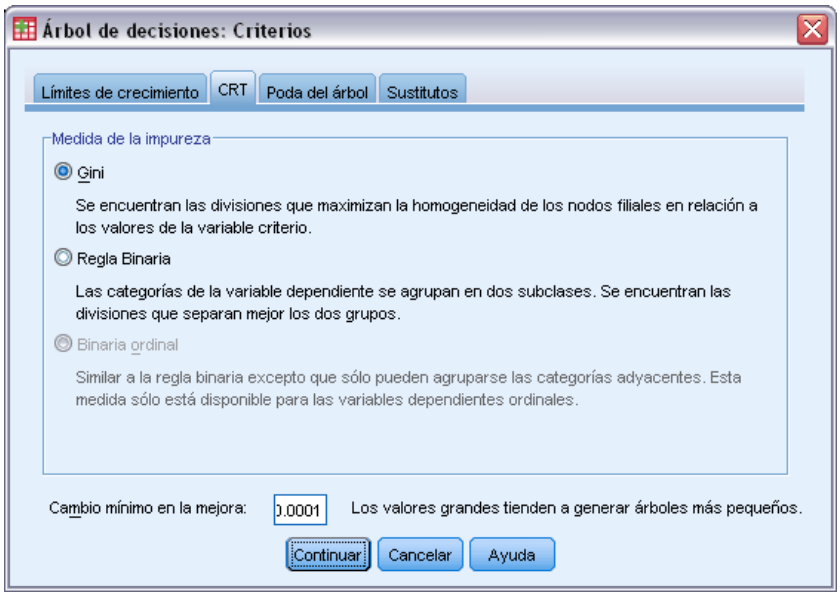
#### *Para especificar intervalos para variables independientes de escala*

- En el cuadro de diálogo principal Árbol de decisión, seleccione una o más variables independientes de escala.
- Para el método de crecimiento, seleccione CHAID o CHAID exhaustivo.
- Pulse en Criterios.
- 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*

Figura 1-9
*Cuadro de diálogo Criterios, pestaña 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.

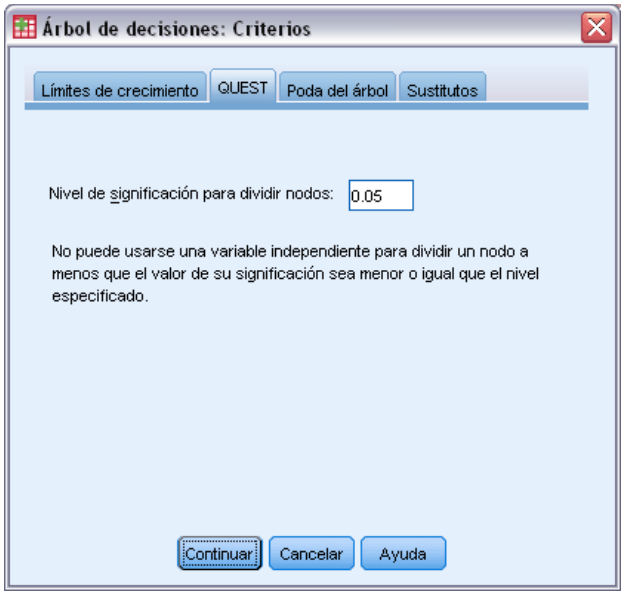
Para variables dependientes categóricas (nominales, ordinales), puede seleccionar la medida de la impureza:

- **Gini.** Se obtienen divisiones que maximizan la homogeneidad de los nodos filiales 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 por defecto.
- **Binaria.** Las categorías de la variable dependiente se agrupan en dos subclases. Se obtienen las divisiones que mejor separan 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 por defecto es 0,0001. Los valores superiores tienden a generar árboles con menos nodos.

## Criterios para QUEST

Figura 1-10
*Cuadro de diálogo Criterios, pestaña QUEST*



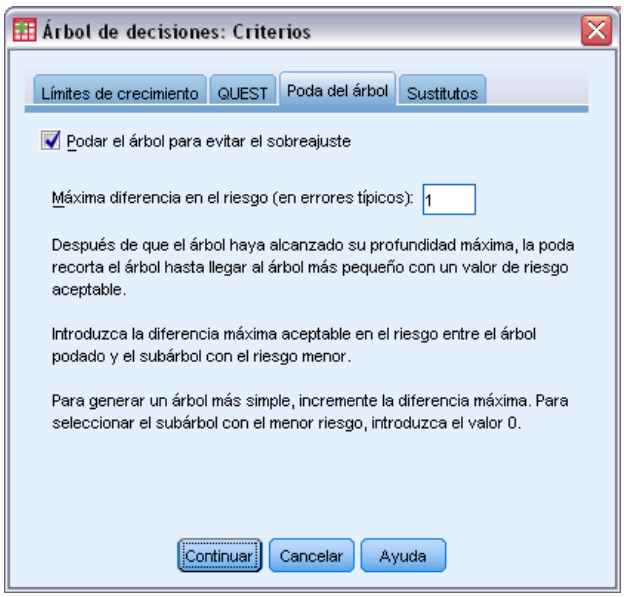

Para el método QUEST, puede especificar el nivel de significación para la división de nodos. No se puede utilizar una variable independiente para dividir nodos a menos que el nivel de significación sea menor o igual que el valor especificado. El valor debe ser mayor que 0 y menor que 1. El valor por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

### *Para especificar criterios para QUEST*

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

## Poda de árboles

Figura 1-11
*Cuadro de diálogo Criterios, pestaña Poda del árbol*



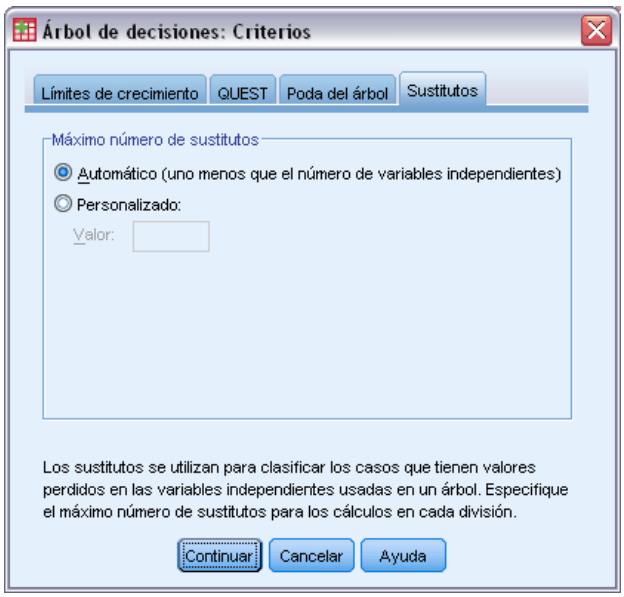

Con los métodos CRT y QUEST, puede evitar el sobreajuste del modelo mediante la **poda** del árbol: el árbol crece hasta que se cumplen los criterios de parada y, a continuación, se recorta de forma automática hasta obtener el subárbol más pequeño basado en la máxima diferencia en el riesgo especificada.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

### *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 filiales en el árbol final, pero no se pueden mostrar los nodos podados durante el proceso de creación del árbol. Si desea obtener más información, consulte el tema Editor del árbol en el capítulo 2 el p. 39.

## *Sustitutos*

Figura 1-12
*Cuadro de diálogo Criterios, pestaña 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.

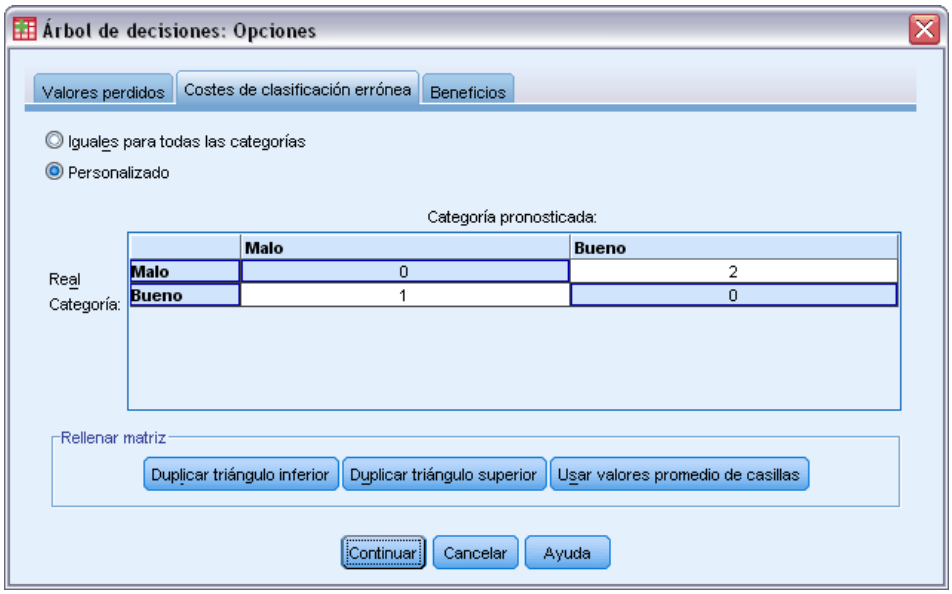
- Por defecto, el número máximo de sustitutos es igual al número de variables independientes menos uno. Es decir, para cada variable independiente, se pueden utilizar todas las demás variables independientes como sustitutos.
- Si no desea que el modelo utilice sustitutos, especifique 0 para el número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

# *Opciones*

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

## Costes de clasificación errónea

Figura 1-13
*Cuadro de diálogo Opciones, pestaña 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*

- ▶ En el cuadro de diálogo principal Árbol de decisión, seleccione una variable dependiente categórica (nominal, ordinal) con dos o más etiquetas de valor definidas.
- ▶ Pulse en Opciones.
- ▶ Pulse en la pestaña Costes de clasificación errónea.

- ▶ Pulse en Personalizados.

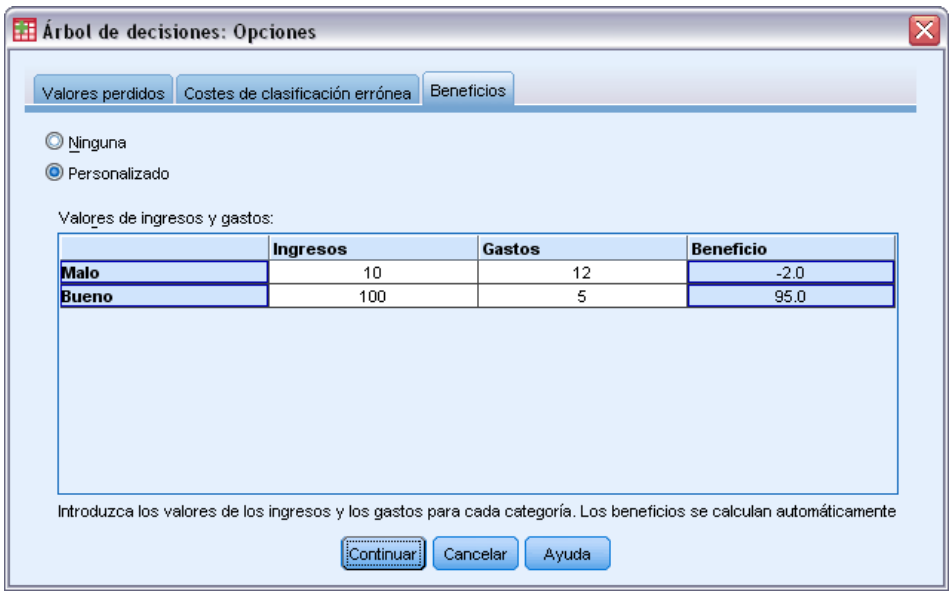
- ▶ Introduzca uno o más costes de clasificación errónea en la cuadrícula. Los valores deben ser no negativos. (Las clasificaciones correctas, representadas en la diagonal, son siempre 0.)

**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*

Figura 1-14
*Cuadro de diálogo Opciones, pestaña 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*

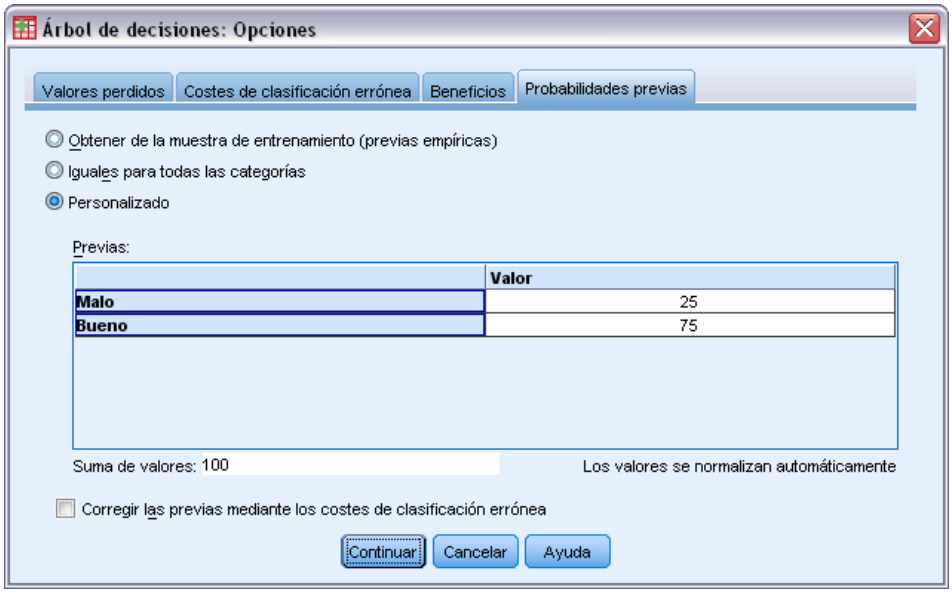
Este cuadro de diálogo requiere etiquetas de valor definidas para la variable dependiente. No estará disponible a menos que dos valores como mínimo de la variable dependiente categórica tengan etiquetas de valor definidas.

#### *Para especificar los beneficios*

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

## *Probabilidades previas*

Figura 1-15
*Cuadro de diálogo Opciones, pestaña 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. Si desea obtener más información, consulte el tema Validación el p. 8.

**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. Si desea obtener más información, consulte el tema Costes de clasificación errónea el p. 17.

### *Beneficios y etiquetas de valor*

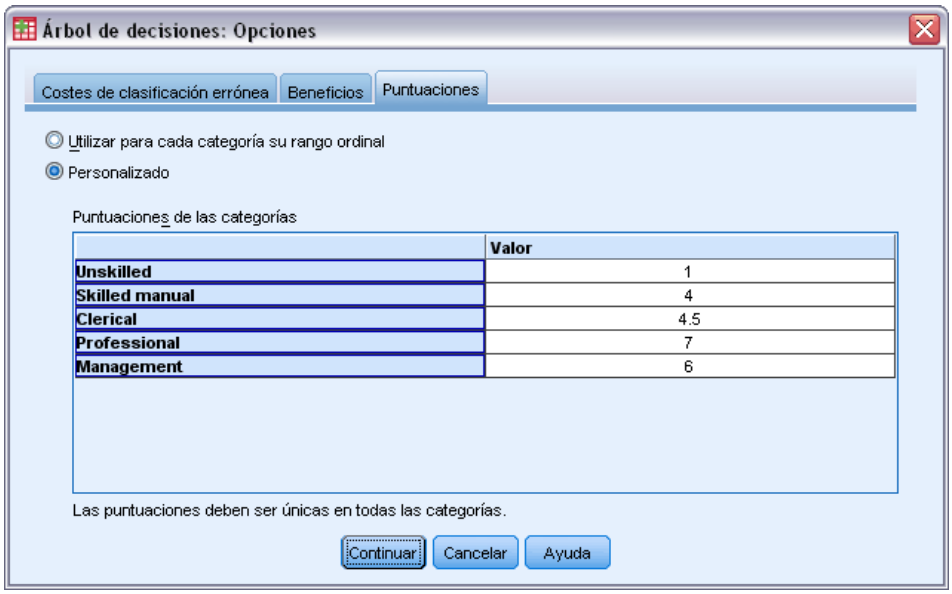
Este cuadro de diálogo requiere etiquetas de valor definidas para la variable dependiente. No estará disponible a menos que dos valores como mínimo de la variable dependiente categórica tengan etiquetas de valor definidas.

### *Para especificar probabilidades previas*

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

## Puntuaciones

Figura 1-16
*Cuadro de diálogo Opciones, pestaña 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. Ésta es la opción por defecto.
- **Personalizado.** Introduzca una puntuación numérica para cada categoría de la variable dependiente que aparezca en la cuadrícula.

### Ejemplo

| 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*

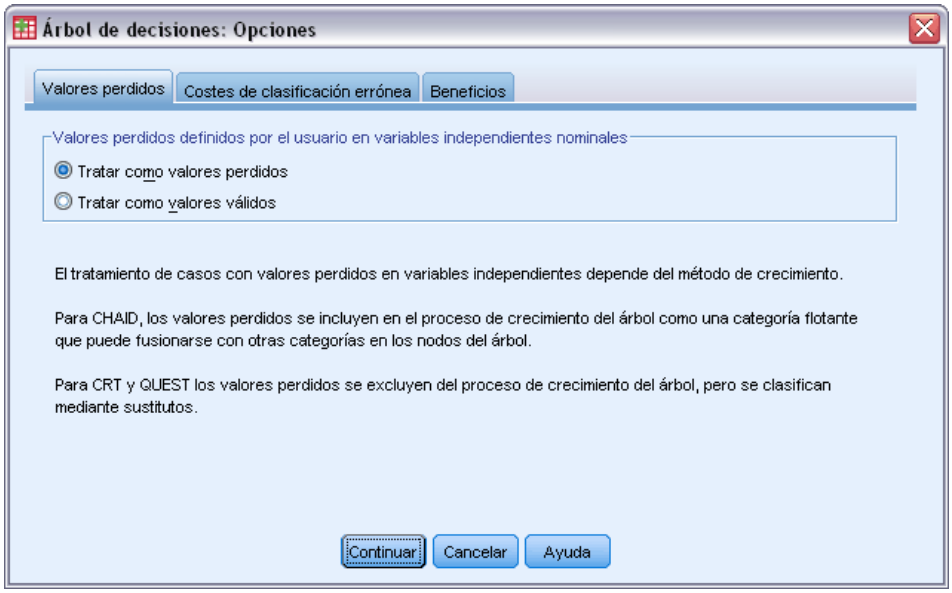
Este cuadro de diálogo requiere etiquetas de valor definidas para la variable dependiente. No estará disponible a menos que dos valores como mínimo de la variable dependiente categórica tengan etiquetas de valor definidas.

### *Para especificar puntuaciones*

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

## *Valores perdidos*

Figura 1-17
*Cuadro de diálogo Opciones, pestaña Valores perdidos*




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

- El tratamiento de los valores definidos como perdidos por el 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. Si desea obtener más información, consulte el tema Selección de categorías el p. 6.
- Para las variables dependientes ordinales y de escala, siempre se excluyen los casos con valores de variables dependientes perdidos del sistema o definidos como tales por el usuario.

**Tratar como valores perdidos.** Los valores definidos como perdidos por el usuario reciben el mismo tratamiento que los valores perdidos del sistema. El tratamiento de estos varía en función del método de crecimiento.

**Tratar como valores válidos.** Los valores definidos como perdidos por el 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 definidos como tales por el usuario:

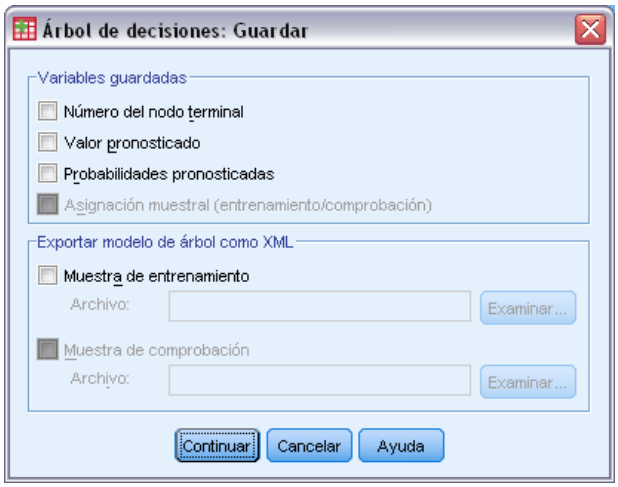
- Para CHAID y CHAID exhaustivo, los valores de las variables independientes perdidos del sistema o definidos como perdidos por el usuario se incluyen en el análisis como una única categoría combinada. Para las variables independientes ordinales y de escala, los algoritmos primero generan categorías 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 definidos como perdidos por el usuario nominales se tratan como perdidos, también se procesarán de la misma manera. Si desea obtener más información, consulte el tema Sustitutos el p. 16.

### *Para especificar el tratamiento de los valores definidos como perdidos por el usuario de variables independientes nominales*

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

# Almacenamiento de información del modelo

Figura 1-18
*Cuadro de diálogo Guardar*




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 pronosticado.** 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. Si desea obtener más información, consulte el tema Validación el p. 8.

### *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 puntuarlos.

**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.

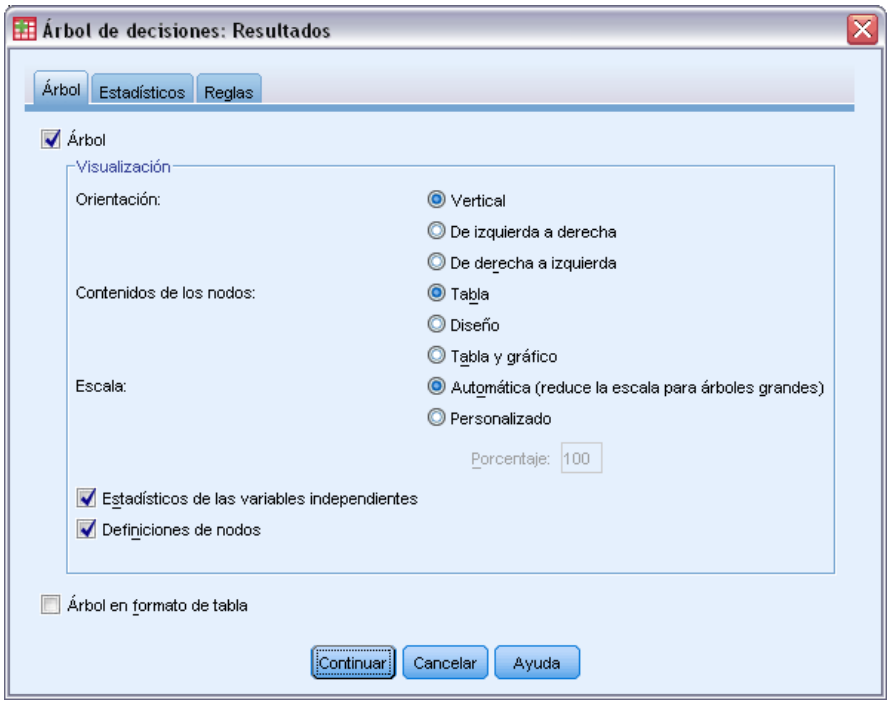
**Muestra de comprobación.** Escribe el modelo para la muestra de comprobación en el archivo especificado. No disponible a menos que se haya seleccionado la validación por división muestral.

# Resultados

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

## Presentación del árbol

Figura 1-19
*Cuadro de diálogo Resultados, pestaña Árbol*




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

**Árbol.** Por defecto, 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 frecuencias y porcentajes, y los gráficos son diagramas de barras. Para variables dependientes de escala, las tablas muestran medias, desviaciones típicas, número de casos y valores pronosticados, y los gráficos son histogramas.
- **Escalas.** Por defecto, 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.

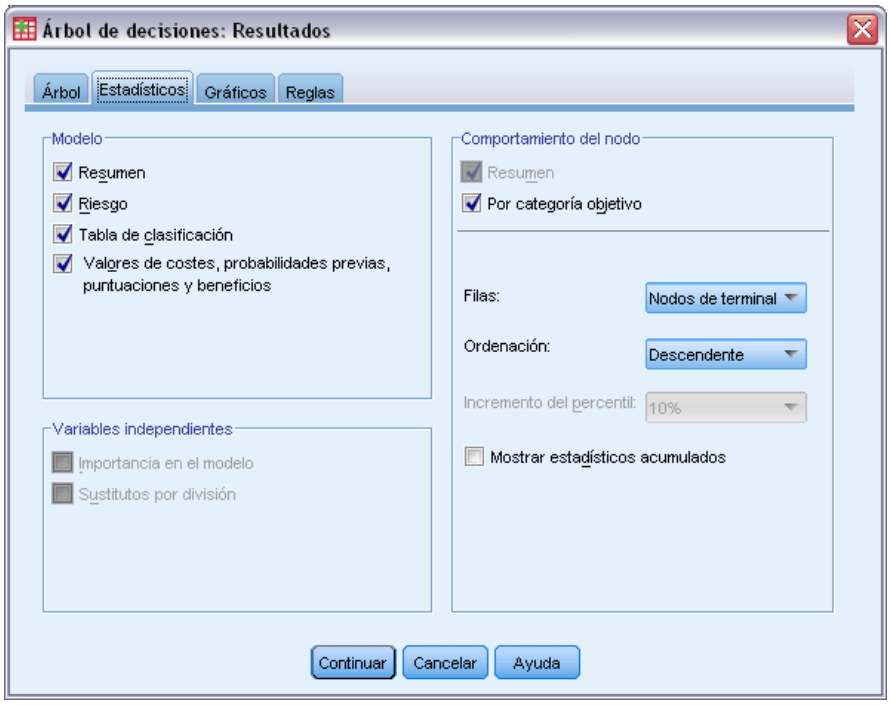
**Árbol en formato de tabla.** Información de resumen para cada nodo del árbol, incluyendo el número del nodo parental, los estadísticos de las variables independientes, el valor o valores de las variables independientes para el nodo, la media y la desviación típica para variables dependientes de escala,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

Figura 1-20
*Árbol en formato de tabla*

| | | Malo | | Bueno | | Total | | | | Variable independiente primaria | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Muestra | Nodo | N | Porcentaje | N | Porcentaje | N | Porcentaje | Categoría pronosticada | Nodo principal | Variable | Mejora | Segmentar valores |
| Entrenamiento | 0 | 538 | 42,5% | 727 | 57,5% | *** | 100,0% | Bueno | | | | |
| | 1 | 238 | 82,4% | 51 | 17,6% | 289 | 16,3% | Malo | 0 | Nivel de ingresos | ,071 | <= Bajo |
| | 2 | 300 | 30,7% | 676 | 69,3% | 976 | 83,7% | Bueno | 0 | Nivel de ingresos | ,071 | > Bajo |
| | 3 | 133 | 62,7% | 79 | 37,3% | 212 | 14,3% | Bueno | 2 | Edad | ,024 | <= 27,074877023612267 |
| | 4 | 167 | 21,9% | 597 | 78,1% | 764 | 69,3% | Bueno | 2 | Edad | ,024 | > 27,074877023612267 |
| | 5 | 119 | 76,3% | 37 | 23,7% | 156 | 9,3% | Malo | 3 | Número de tarjetas de crédito | ,014 | 5 o más |
| | 6 | 14 | 25,0% | 42 | 75,0% | 56 | 5,0% | Bueno | 3 | Número de tarjetas de crédito | ,014 | Menos de 5 |
| | 7 | 149 | 33,2% | 300 | 66,8% | 449 | 37,9% | Bueno | 4 | Número de tarjetas de crédito | ,008 | 5 o más |
| | 8 | 18 | 5,7% | 297 | 94,3% | 315 | 31,5% | Bueno | 4 | Número de tarjetas de crédito | ,008 | Menos de 5 |
| | 9 | 116 | 44,1% | 147 | 55,9% | 263 | 20,6% | Bueno | 7 | Nivel de ingresos | ,006 | <= Medio |

## *Estadísticas*

Figura 1-21
*Cuadro de diálogo Resultados, pestaña Estadísticos*




Las tablas de estadísticos disponibles dependen del nivel de medida 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.

Figura 1-22
*Tabla de resumen del modelo*

| | | |
|---|---|---|
| Especificaciones | Método de crecimiento | CHAID |
| | Variable dependiente | Valoración del crédito |
| | Variables independientes | Edad, Nivel de ingresos, Número de tarjetas de crédito, Educación, Créditos de coche |
| | Validación | SPLITSAMPLE |
| | Máxima profundidad de árbol | 3 |
| | Casos mínimos en nodo principal | 400 |
| | Casos mínimos en nodo secundario | 200 |
| Resultados | Variables independientes incluidas | Nivel de ingresos, Edad, Número de tarjetas de crédito, Créditos de coche |
| | Número de nodos | 10 |
| | Número de nodos terminales | 6 |
| | Profundidad | 3 |

**Riesgo.** Estimación del riesgo y su error típico. 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.

Figura 1-23
*Tablas de riesgos y de clasificación*

**Riesgo**

| Estimación | Desviación Error |
|---|---|
| ,205 | ,008 |

Método de crecimiento: CHAID
Variable dependiente: Valoración del crédito

**Clasificación**

| Observado | Pronosticado | | |
|---|---|---|---|
| | Malo | Bueno | Porcentaje correcto |
| Malo | 665 | 355 | 65,2% |
| Bueno | 149 | 1295 | 89,7% |
| Porcentaje global | 33,0% | 67,0% | 79,5% |

Método de crecimiento: CHAID
Variable dependiente: Valoración del crédito

**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. Si desea obtener más información, consulte el tema Sustitutos el p. 16.

### *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. Si desea obtener más información, consulte el tema Beneficios el p. 18.

Figura 1-24
*Tablas de resumen de ganancias para nodos y percentiles*

**Resumen de ganancia para los nodos**

| Nodo | N | Porcentaje | Beneficios | ROI |
|---|---|---|---|---|
| 7 | 322 | 13,1% | 77,826 | 377,4% |
| 5 | 390 | 15,8% | 70,308 | 308,8% |
| 6 | 455 | 18,5% | 67,692 | 287,9% |
| 9 | 483 | 19,6% | 49,420 | 172,0% |
| 8 | 261 | 10,6% | 23,410 | 64,7% |
| 1 | 553 | 22,4% | 22,532 | 61,9% |

**Resumen de ganancia para los percentiles**

| Percentil | Nodo | N | Beneficios | ROI |
|---|---|---|---|---|
| 10 | 7 | 246 | 77,826 | 377,4% |
| 20 | 7 ; 5 | 493 | 75,218 | 352,0% |
| 30 | 5 ; 6 | 739 | 73,488 | 336,2% |
| 40 | 6 | 986 | 72,036 | 323,4% |
| 50 | 6 ; 9 | 1232 | 70,205 | 307,9% |
| 60 | 9 | 1478 | 66,745 | 280,6% |
| 70 | 9 ; 8 | 1725 | 63,134 | 254,4% |
| 80 | 8 ; 1 | 1971 | 58,149 | 221,6% |
| 90 | 1 | 2218 | 54,183 | 197,9% |
| 100 | 1 | 2464 | 51,023 | 180,4% |

**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. Si desea obtener más información, consulte el tema Selección de categorías el p. 6.

Figura 1-25
*Ganancias de categorías objetivo para nodos y percentiles*

**Target Category: Malo**

**Ganancias para los nodos**

| | Nodo | | Ganancia | | | |
|---|---|---|---|---|---|---|
| Nodo | N | Porcentaje | N | Porcentaje | Respuesta | Índice |
| 1 | 553 | 22,4% | 454 | 44,5% | 82,1% | 198,3% |
| 8 | 261 | 10,6% | 211 | 20,7% | 80,8% | 195,3% |
| 9 | 483 | 19,6% | 211 | 20,7% | 43,7% | 105,5% |
| 6 | 455 | 18,5% | 80 | 7,8% | 17,6% | 42,5% |
| 5 | 390 | 15,8% | 54 | 5,3% | 13,8% | 33,4% |
| 7 | 322 | 13,1% | 10 | 1,0% | 3,1% | 7,5% |

**Ganancias para los percentiles**

| | | | Ganancia | | | |
|---|---|---|---|---|---|---|
| Percentil | Nodo | N | N | Porcentaje | Respuesta | Índice |
| 10 | 1 | 246 | 202 | 19,8% | 82,1% | 198,3% |
| 20 | 1 | 493 | 405 | 39,7% | 82,1% | 198,3% |
| 30 | 1 ; 8 | 739 | 604 | 59,3% | 81,8% | 197,6% |
| 40 | 8 ; 9 | 986 | 740 | 72,6% | 75,1% | 181,3% |
| 50 | 9 | 1232 | 848 | 83,1% | 68,8% | 166,2% |
| 60 | 9 ; 6 | 1478 | 908 | 89,0% | 61,4% | 148,4% |
| 70 | 6 | 1725 | 951 | 93,3% | 55,1% | 133,2% |
| 80 | 6 ; 5 | 1971 | 986 | 96,7% | 50,0% | 120,9% |
| 90 | 5 ; 7 | 2218 | 1012 | 99,3% | 45,6% | 110,3% |
| 100 | 7 | 2464 | 1020 | 100,0% | 41,4% | 100,0% |

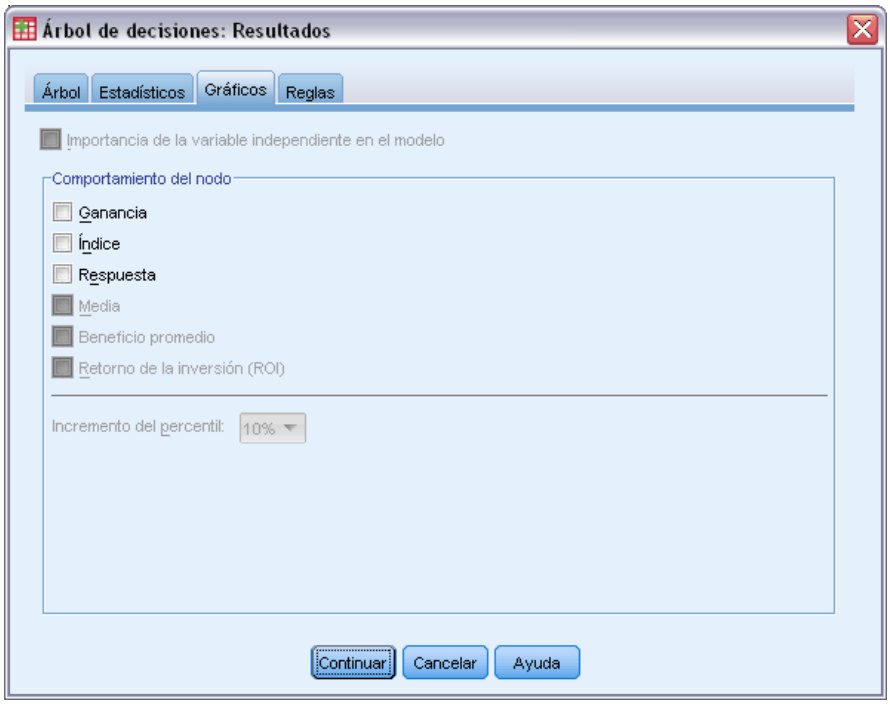
**Filas.** Las tablas de comportamiento de los nodos pueden mostrar resultados por nodos terminales, por percentiles o por ambos. Si selecciona ambos, se generan dos tablas por cada categoría objetivo. Las tablas de percentiles muestran valores acumulados para cada percentil, basados en el orden.

**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.

## *Gráficos*

Figura 1-26
*Cuadro de diálogo Resultados, pestaña Gráficos*




Los gráficos disponibles dependen del nivel de medida 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*

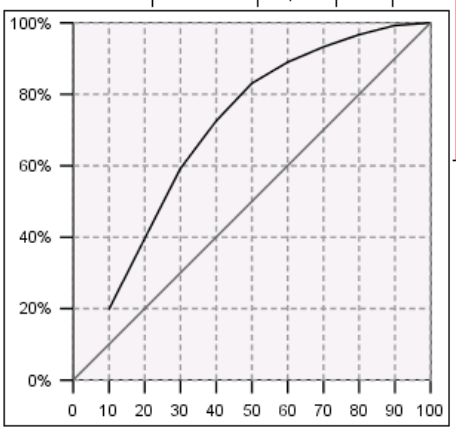
**Ganancia.** La ganancia es el porcentaje de los casos totales en la categoría objetivo en cada nodo, calculada como: (*n* criterio de nodo / *n* total de criterios) x 100. El gráfico de ganancias es un gráfico de líneas de las ganancias por percentiles acumulados, calculadas como: (*n* de percentil de criterios acumulados / *n* total de criterios) x 100. Se generará un gráfico de líneas distinto para cada categoría objetivo. Disponible sólo para variables dependientes categóricas con categorías objetivo definidas. Si desea obtener más información, consulte el tema Selección de categorías el p. 6.

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.

Figura 1-27
*Tabla de ganancias para los percentiles y gráfico de ganancias*

**Ganancias para los percentiles**

| Percentil | Nodo | N | Ganancia N | Ganancia Porcentaje | Respuesta | Índice |
|---|---|---|---|---|---|---|
| 10 | 1 | 246 | 202 | 19,8% | 82,1% | 198,3% |
| 20 | 1 | 493 | 405 | 39,7% | 82,1% | 198,3% |
| 30 | 1 ; 8 | 739 | 604 | 59,3% | 81,8% | 197,6% |
| 40 | 8 ; 9 | 986 | 740 | 72,6% | 75,1% | 181,3% |
| | | | | 83,1% | 68,8% | 166,2% |
| | | | | 89,0% | 61,4% | 148,4% |
| | | | | 93,3% | 55,1% | 133,2% |
| | | | | 96,7% | 50,0% | 120,9% |
| | | | | 99,3% | 45,6% | 110,3% |
| | | | | 100,0% | 41,4% | 100,0% |



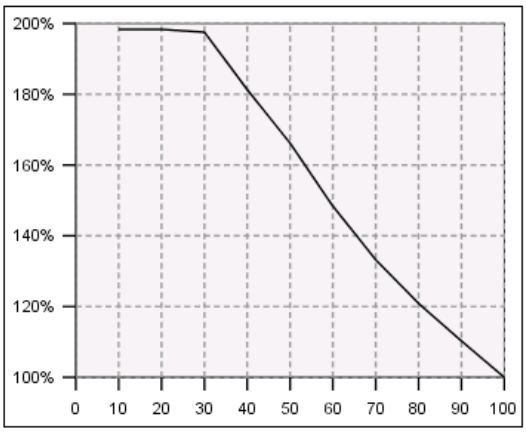

**Índice.** El índice es la proporción del porcentaje de respuestas en la categoría criterio del nodo en comparación con el porcentaje global de respuestas en la categoría criterio para toda la muestra. El gráfico de índices es un gráfico de líneas que representa los valores de los índices de percentiles 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.

Figura 1-28
*Tabla de ganancias para los percentiles y gráfico de índices*

**Ganancias para los percentiles**

| Percentil | Nodo | N | Ganancia N | Ganancia Porcentaje | Respuesta | Índice |
|---|---|---|---|---|---|---|
| 10 | 1 | 246 | 202 | 19,8% | 82,1% | 198,3% |
| 20 | 1 | 493 | 405 | 39,7% | 82,1% | 198,3% |
| 30 | 1 ; 8 | 739 | 604 | 59,3% | 81,8% | 197,6% |
| 40 | 8 ; 9 | 986 | 740 | 72,6% | 75,1% | 181,3% |
| | | | | | | 166,2% |
| | | | | | | 148,4% |
| | | | | | | 133,2% |
| | | | | | | 120,9% |
| | | | | | | 110,3% |
| | | | | | | 100,0% |



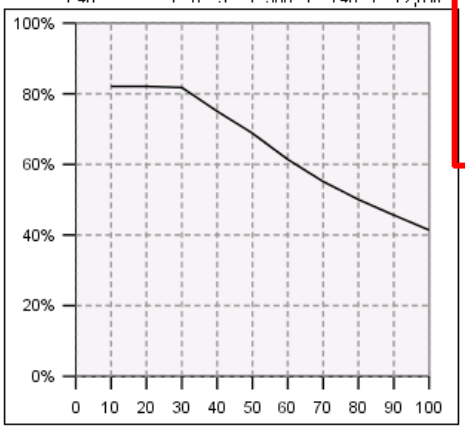

**Respuestas.** Porcentaje de casos pertenecientes al nodo que pertenecen a la categoría objetivo especificada. El gráfico de respuestas es un gráfico de líneas de las respuestas por percentiles acumulados, calculado como: (*n* de percentil de criterios acumulados / *n* total de percentiles acumulados) 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.

Figura 1-29
*Tabla de ganancias para los percentiles y gráfico de respuestas*

**Ganancias para los percentiles**

| Percentil | Nodo | N | Ganancia N | Ganancia Porcentaje | Respuesta | Índice |
|---|---|---|---|---|---|---|
| 10 | 1 | 246 | 202 | 19,8% | 82,1% | 198,3% |
| 20 | 1 | 493 | 405 | 39,7% | 82,1% | 198,3% |
| 30 | 1 ; 8 | 739 | 604 | 59,3% | 81,8% | 197,6% |
| 40 | 8 · 9 | 986 | 740 | 72,6% | 75,1% | 181,3% |
| | | | | | 68,8% | 166,2% |
| | | | | | 61,4% | 148,4% |
| | | | | | 55,1% | 133,2% |
| | | | | | 50,0% | 120,9% |
| | | | | | 45,6% | 110,3% |
| | | | | | 41,4% | 100,0% |




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

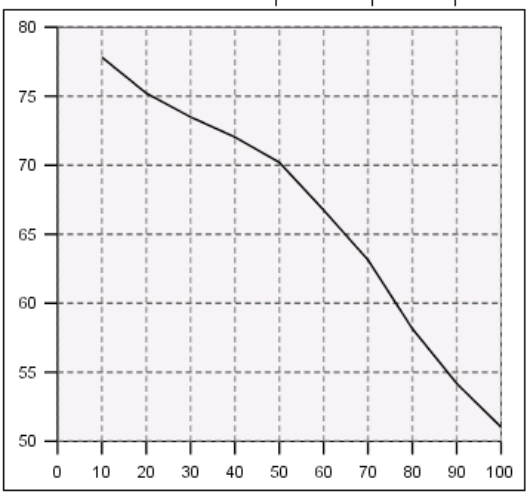
**Beneficio promedio.** Gráfico de líneas del beneficio promedio acumulado. Disponible sólo para variables dependientes categóricas con beneficios definidos. Si desea obtener más información, consulte el tema Beneficios el p. 18.

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.

Figura 1-30
*Tabla de resumen de ganancias para los percentiles y gráfico de beneficio medio*

**Resumen de ganancia para los percentiles**

| Percentil | Nodo | N | Beneficios | ROI |
|---|---|---|---|---|
| 10 | 7 | 246 | 77,826 | 377,4% |
| 20 | 7 ; 5 | 493 | 75,218 | 352,0% |
| 30 | 5 ; 6 | 739 | 73,488 | 336,2% |
| 40 | 6 | 986 | 72,036 | 323,4% |
| | | | 70,205 | 307,9% |
| | | | 66,745 | 280,6% |
| | | | 63,134 | 254,4% |
| | | | 58,149 | 221,6% |
| | | | 54,183 | 197,9% |
| | | | 51,023 | 180,4% |



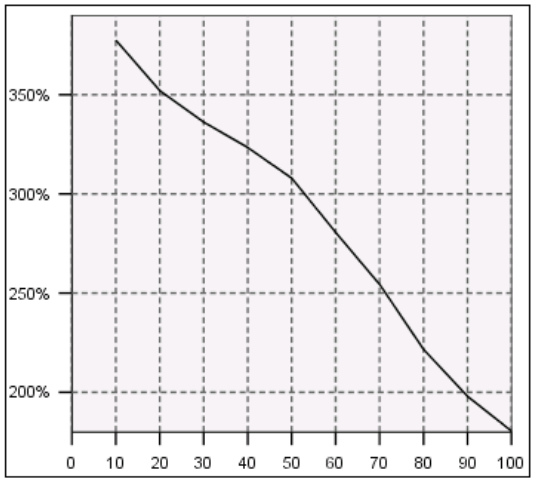

**Retorno de la inversión (ROI).** Gráfico de líneas de ROI (retorno de la inversión) acumulado. ROI se calcula como la relación entre los beneficios y los gastos. Disponible sólo para variables dependientes categóricas con beneficios definidos.

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.

Figura 1-31
*Tabla de resumen de ganancias para los percentiles y gráfico de ROI*

**Resumen de ganancia para los percentiles**

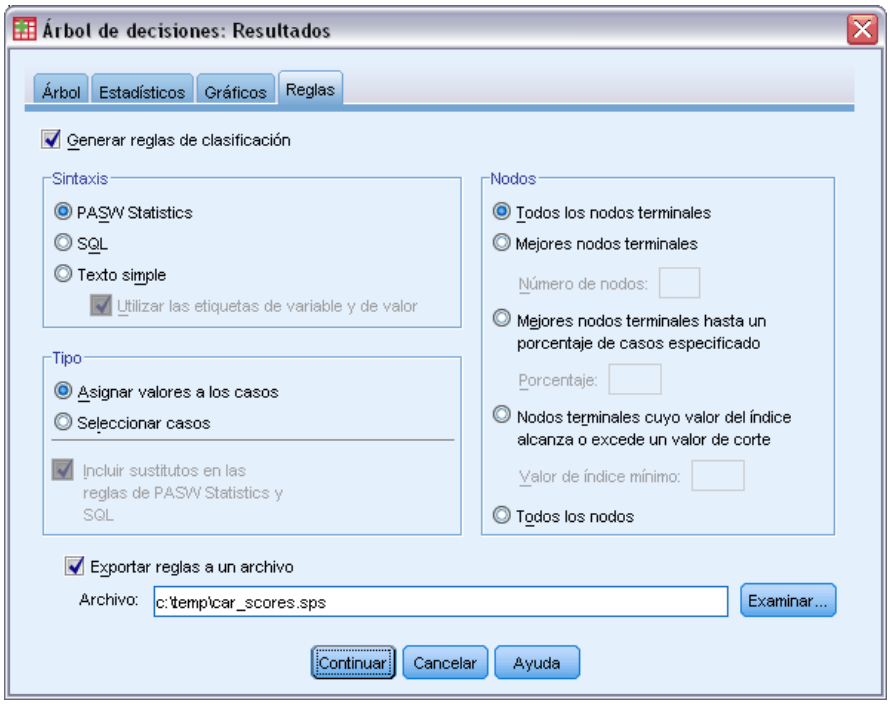
| Percentil | Nodo | N | Beneficios | ROI |
|---|---|---|---|---|
| 10 | 7 | 246 | 77,826 | 377,4% |
| 20 | 7 ; 5 | 493 | 75,218 | 352,0% |
| 30 | 5 ; 6 | 739 | 73,488 | 336,2% |
| 40 | 6 | 986 | 72,036 | 323,4% |
| | | | | 307,9% |
| | | | | 280,6% |
| | | | | 254,4% |
| | | | | 221,6% |
| | | | | 197,9% |
| | | | | 180,4% |




**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.

## *Reglas de selección y puntuación*

Figura 1-32
*Cuadro de diálogo Resultados, pestaña Reglas*




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 instrucciones `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 instrucciones 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 SPSS Statistics y las reglas de SQL, controla el tipo de reglas generadas: reglas de selección o puntuación.

- **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 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 de SPSS Statistics y de SQL.** 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. Si desea obtener más información, consulte el tema Sustitutos el p. 16.

**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 1*: 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 de SPSS Statistics y de SQL 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. Si desea obtener más información, consulte el tema Reglas de selección de casos y puntuación en el capítulo 2 el p. 47.

*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 definidos como perdidos por el usuario (si hubiera).

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:

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

*o*

▶ En el menú Edición o el menú contextual que aparece al pulsar el botón derecho, seleccione:
Editar contenido > En otra ventana

### *Ocultación y presentación de nodos*

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

▶ Pulse en el signo menos (–) de la pequeña casilla situada debajo de la esquina derecha inferior del nodo parental.

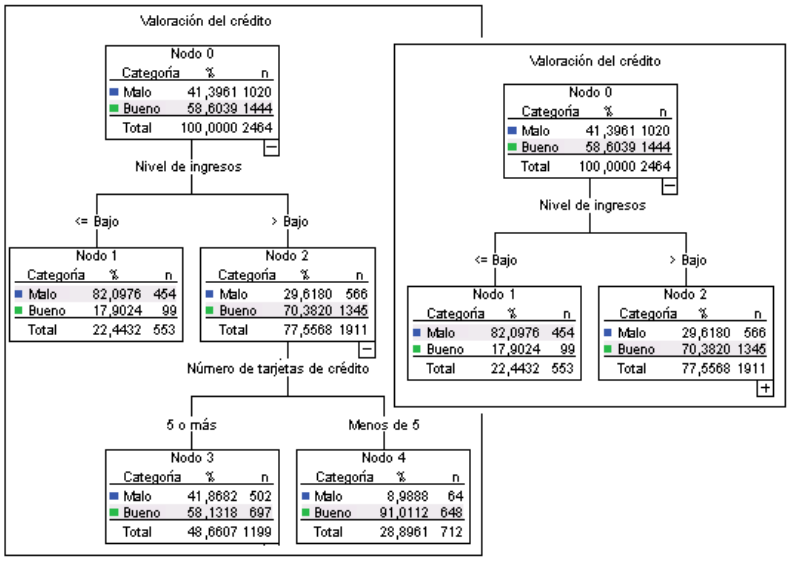
Se ocultarán todos los nodos de esa rama situados por debajo del nodo parental.

Para mostrar, expandir, los nodos filiales en una rama por debajo de un nodo parental:

▶ Pulse en el signo más (+) de la pequeña casilla situada debajo de la esquina derecha inferior del nodo parental.

*Nota*: ocultar los nodos filiales 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. Si desea obtener más información, consulte el tema Poda de árboles en el capítulo 1 el p. 15.

Figura 2-1
*Árbol expandido y contraído*




***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:

- ▶ Pulse en un nodo que desee seleccionar.
- ▶ 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 parentales en una rama, y de nodos filiales en otra rama. Sin embargo, no podrá utilizar la selección múltiple en un nodo parental y en un nodo filial/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 funciones 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. Si desea obtener más información, consulte el tema Mapa del árbol el p. 41.

- **Escalamiento.** Puede acercarse o alejarse cambiando el porcentaje de escala para la presentación del árbol. Si desea obtener más información, consulte el tema Escalamiento de la presentación del árbol el p. 42.
- **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. Si desea obtener más información, consulte el tema Control de la información que se muestra en el árbol el p. 43.

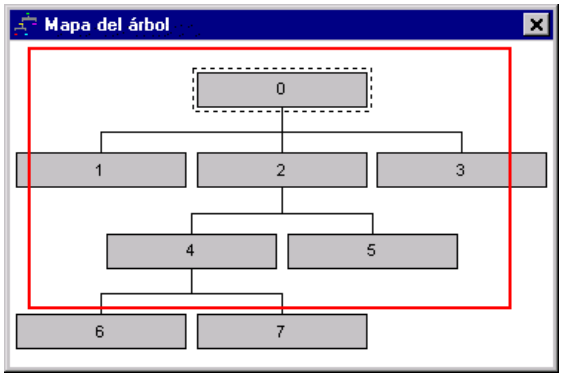
## *Mapa del árbol*

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:

▶ En los menús del Editor del árbol, seleccione:
Ver > Mapa del árbol

Figura 2-2
*Ventana del 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 parental y en un nodo filial/descendiente de la misma rama del nodo.

## Escalamiento de la presentación del árbol

Por defecto, 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%.

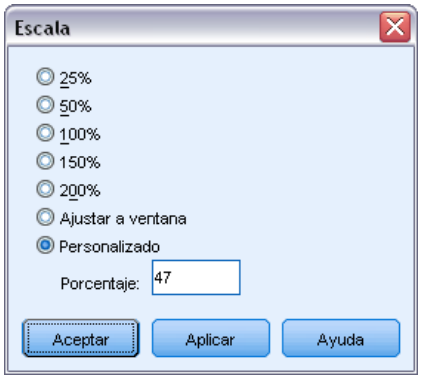
Para cambiar la escala del árbol:

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

*o*

- En los menús del Editor del árbol, seleccione:
  Ver > Escala...

Figura 2-3
*Cuadro de diálogo Escala*




También puede especificar un valor de escala antes de crear el modelo del árbol. Si desea obtener más información, consulte el tema Resultados en el capítulo 1 el p. 25.

## Ventana de resumen de nodos

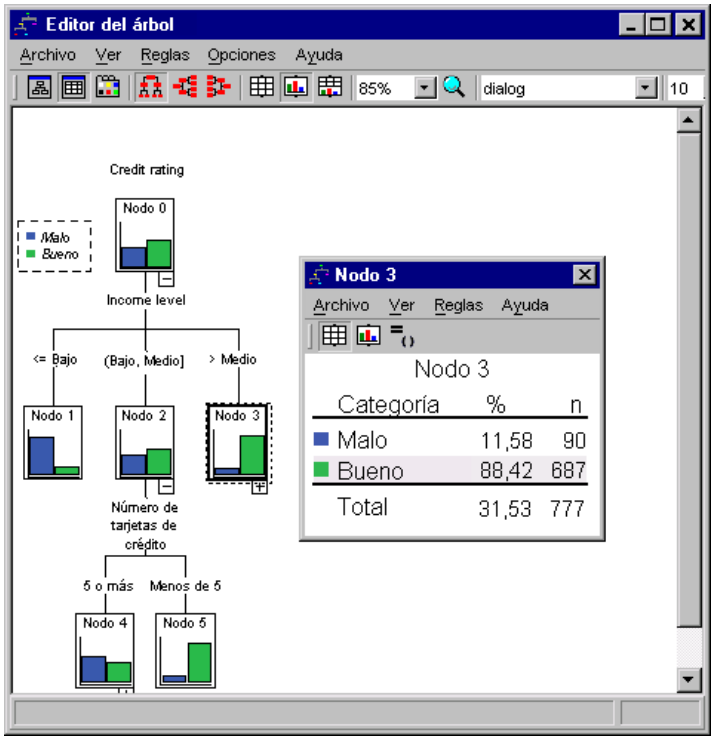
La ventana de resumen de nodos proporciona una vista de mayor tamaño de los nodos seleccionados. También puede utilizar la ventana de resumen para ver, aplicar o guardar las reglas de selección o de puntuación basadas en los nodos seleccionados.

- 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. Si desea obtener más información, consulte el tema Reglas de selección de casos y puntuación el p. 47.
- 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:

- 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.
- Elija en los menús:
  Ver > Resumen

Figura 2-4
*Ventana de resumen*



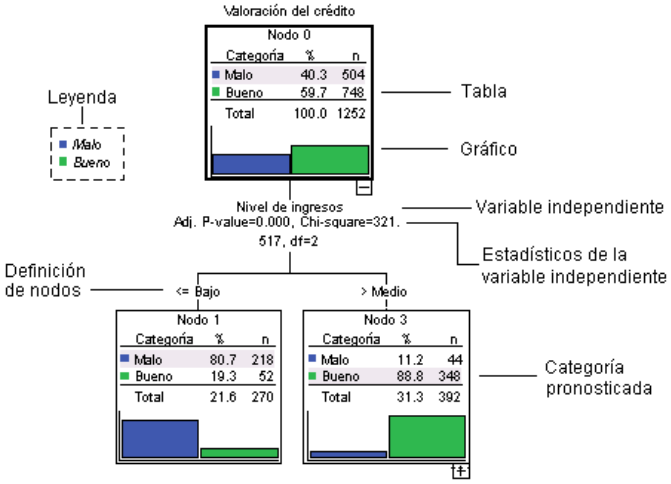

## Control de la información que se muestra en el árbol

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.

| Configuración | Selección en el menú Opciones |
|---|---|
| Resaltar categoría pronosticada (variable dependiente categórica) | Resaltar pronosticada |
| Tablas y/o gráficos en el nodo | Contenidos de los nodos |
| Valores de la prueba de significación y valores *p* | Estadísticos de las variables independientes |
| Nombres de las variables (predictoras) independientes | Variables independientes |
| Valor(es) independientes (predictores) para nodos | Definiciones de los nodos |

| Configuración | Selección en el menú Opciones |
|---|---|
| Alineación (arriba-abajo, izquierda-derecha, derecha-izquierda) | Orientación |
| Leyenda del gráfico | Leyenda |

Figura 2-5
*Elementos del árbol*




## 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

- 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 pistas para las herramientas describen todos los controles de la barra de herramientas cuando se sitúa el puntero del ratón sobre ellos.)

  *o*

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

- Para el borde, rama, fondo de los nodos, categoría pronosticada y fondo del árbol, pulse en la pestaña Color.

- Para los colores y atributos de fuente, pulse en la pestaña Texto.

- Para los colores de los gráficos de los nodos, pulse en la pestaña Gráficos de nodos.

Figura 2-6
*Ventana Propiedades, pestaña Color*

Figura 2-7
*Ventana Propiedades, pestaña Texto*

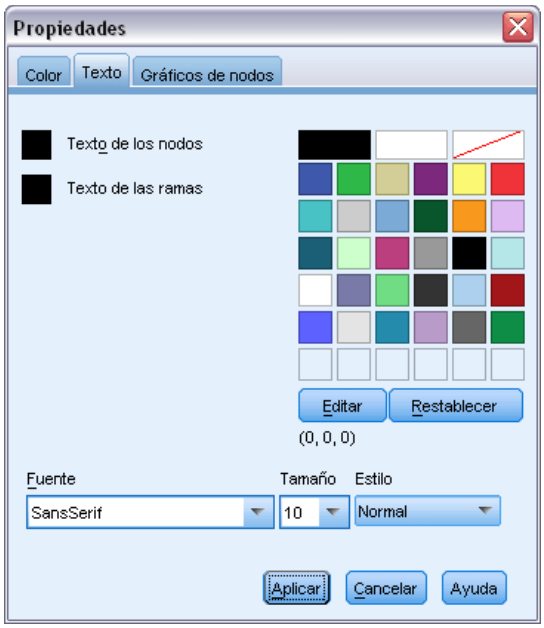
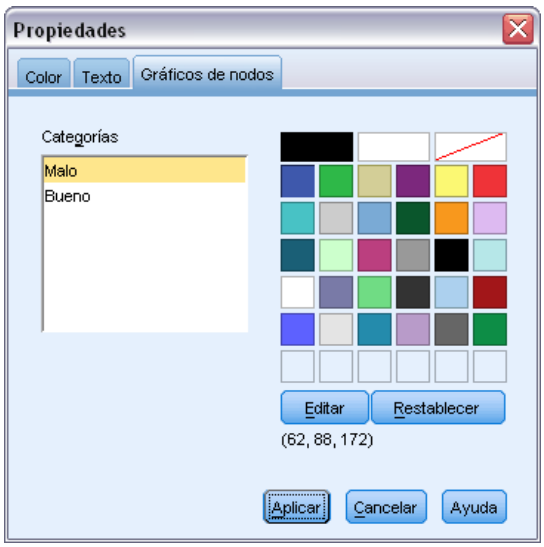



Figura 2-8
*Ventana Propiedades, pestaña Gráficos de nodos*



# Reglas de selección de casos y puntuación

Puede utilizar el Editor del árbol para:

- Seleccionar subconjuntos de casos basados en los nodos seleccionados. Si desea obtener más información, consulte el tema Filtrado de casos el p. 47.
- Generar reglas de selección de casos o reglas de puntuación en sintaxis de comandos de IBM® SPSS® Statistics o formato SQL. Si desea obtener más información, consulte el tema Almacenamiento de las reglas de selección y puntuación el p. 47.

También puede guardar de forma automática reglas basadas en distintos criterios cuando ejecute el procedimiento Árbol de decisión para crear el modelo del árbol. Si desea obtener más información, consulte el tema Reglas de selección y puntuación en el capítulo 1 el p. 37.

## 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.

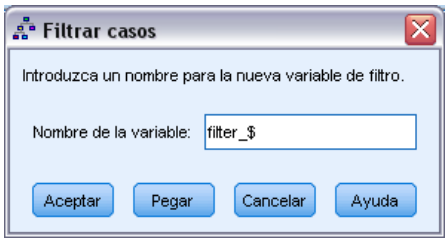
▶ Seleccione los nodos en el Editor del árbol. Mantenga pulsada la tecla Ctrl al mismo tiempo que pulsa el botón del ratón para seleccionar varios nodos.

▶ Elija en los menús:
Reglas > Filtrar casos...

▶ 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.

▶ Pulse en Aceptar.

Figura 2-9
*Cuadro de diálogo Filtrar casos*




## 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 instrucciones `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 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. Si desea obtener más información, consulte el tema Sustitutos en el capítulo 1 el p. 16.

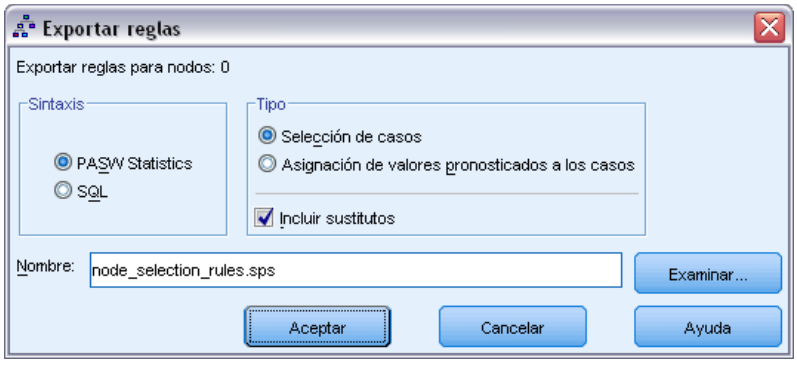
Para guardar reglas de selección de casos o puntuación:

▶ 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.

▶ Elija en los menús:
Reglas > Exportar...

▶ Seleccione el tipo de reglas que desea e introduzca un nombre de archivo.

Figura 2-10
*Cuadro de diálogo Exportar reglas*



*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 definidos como perdidos por el usuario (si hubiera).

# *Parte II: Ejemplos*

Capítulo
3

# Requisitos y supuestos de los datos

El procedimiento Árbol de decisión supone que:

- Se ha asignado el nivel de medida adecuado a todas las variables del análisis.
- En el caso de variables dependientes categóricas (**nominales**, **ordinales**), se han definido etiquetas de valor para todas las categorías que se deben incluir en el análisis.

Utilizaremos el archivo *tree_textdata.sav* para ilustrar la importancia de estos dos requisitos. Este archivo de datos refleja el estado por defecto de los datos leídos o introducidos antes de definir ningún atributo, como el nivel de medida o las etiquetas de valor. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Decision Trees 19*.

## Efectos del nivel de medida en los modelos de árbol

Las dos variables de este archivo de datos son numéricas y ambas tienen asignadas el nivel de medición **escala**. Pero, como veremos más adelante, ambas variables son en realidad variables categóricas que utilizan códigos numéricos para indicar valores de categoría.

▶ Para ejecutar un análisis de Árbol de decisiones, elija en los menús:
Analizar > Clasificar > Árbol...

Los iconos situados junto a las dos variables en la lista de variables de origen indican que se ambas se tratarán como variables de escala.

Figura 3-1
*Cuadro de diálogo principal Árbol de decisión con dos variables de escala*


- Seleccione *dependiente* como la variable dependiente.
- Seleccione *independiente* como la variable independiente.
- Pulse en Aceptar para iniciar el procedimiento.
- Vuelva a abrir el cuadro de diálogo Árbol de decisión y pulse en Restablecer.
- Pulse con el botón derecho en *dependiente* en la lista de origen y, en el menú contextual, seleccione Nominal.
- Realice los mismos pasos para la variable *independiente* en la lista de origen.

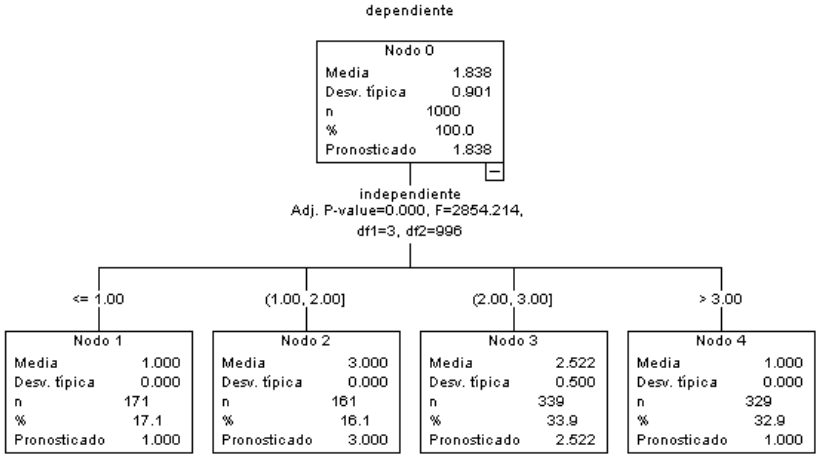
Ahora los iconos situados junto a cada variable indican que serán tratadas como variables nominales.

Figura 3-2
*Iconos nominales en la lista de origen*


- Seleccione *dependiente* como variable dependiente e *independiente* como variable independiente y pulse en Aceptar para ejecutar el procedimiento.

Comparemos los dos árboles. Primero estudiaremos el árbol en el que las dos variables numéricas se han tratado como variables de escala.

Figura 3-3
*Árbol con las dos variables tratadas como variables de escala*

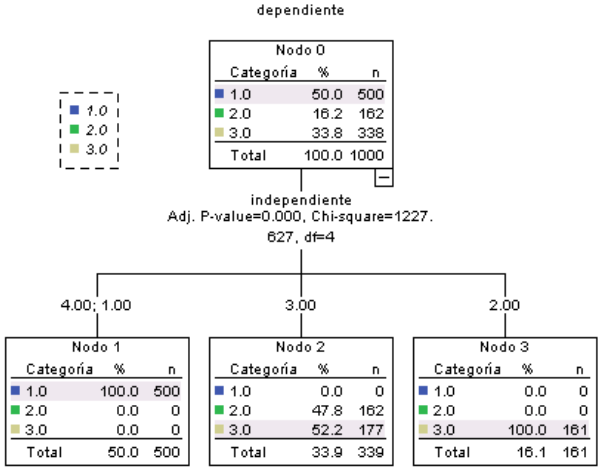


- Cada nodo del árbol muestra el valor “pronosticado”, que es el valor de la media de la variable dependiente en dicho nodo. Para una variable que es en realidad categórica, puede que la media no sea un estadístico significativo.
- El árbol tiene cuatro nodos filiales, uno para cada valor de la variable independiente.

Los modelos de árbol fundirán a menudo nodos similares, pero para una variable de escala, sólo se pueden fundir valores contiguos. En este ejemplo, no hay valores contiguos que se hayan considerado lo suficientemente similares como para fundir nodos entre sí.

El árbol en el que se ha tratado a las dos variables como nominales es algo distinto en varios aspectos.

Figura 3-4
*Árbol con las dos variables tratadas como nominales*




- En lugar de un valor pronosticado, cada nodo contiene una tabla de frecuencias que muestra el número de casos (frecuencia y porcentaje) para cada categoría de la variable dependiente.
- La categoría “pronosticada”, que es la categoría con el mayor valor de frecuencia en cada nodo, aparece resaltada. Por ejemplo, la categoría pronosticada para el nodo 2 es la categoría 3.
- En lugar de cuatro nodos filiales, sólo hay tres, con dos valores de la variable independiente fundidos en un único nodo.

Los dos valores independientes fundidos en el mismo nodo son el 1 y el 4. Ya que, por definición, no hay ningún orden inherente a los valores nominales, se permite la fusión de valores aunque estos no sean contiguos.

## Asignación permanente del nivel de medida

Cuando se modifica el nivel de medida para una variable en el cuadro de diálogo Árbol de decisión, el cambio es sólo temporal; y no se almacenará con el archivo de datos. Es más, es posible que no siempre sepa cuál es el nivel de medida correcto para todas las variables.

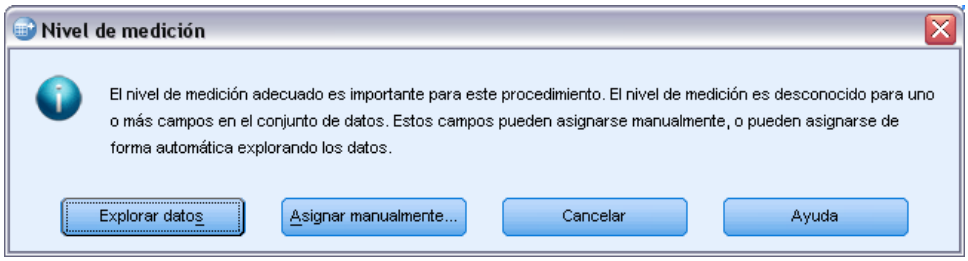
La opción Definir propiedades de variables puede ayudarle a determinar el nivel de medida correcto para cada variable y modificar, de forma permanente, el nivel de medida asignado. Para utilizar la opción Definir propiedades de variables:

- Seleccione en los menús:
  Datos > Definir propiedades de variables...

### *Variables 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.

Figura 3-5
*Alerta de nivel de medición*




- **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.

## *Efectos de las etiquetas de valor en los modelos de árbol*

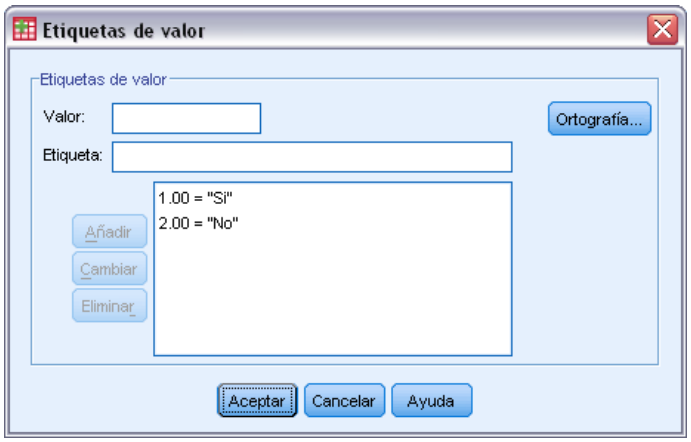
La interfaz del cuadro de diálogo Árbol de decisión supone que o *todos* los valores no perdidos de una variable dependiente categórica (nominal, ordinal) tienen etiquetas de valor definidas o *ninguno* de ellos las tienen. Algunas características no estarán disponibles a menos que dos valores como mínimo de la variable dependiente categórica 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.

El archivo de datos original de este ejemplo no contiene ninguna etiqueta de valor definida y, cuando la variable dependiente se trata como nominal, el modelo de árbol utiliza todos los valores no perdidos en el análisis. En este ejemplo, dichos valores son 1, 2 y 3.

Pero, ¿qué sucede si definimos etiquetas de valor para algunos, aunque no todos, valores de la variable dependiente?

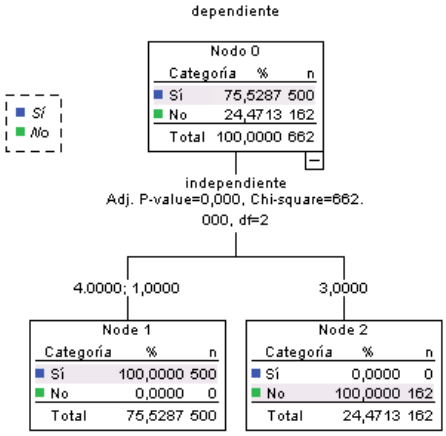
- En la ventana del Editor de datos, pulse en la pestaña Vista de variables.
- Pulse en la casilla Valores para la variable *dependiente*.

Figura 3-6
*Definición de etiquetas de valor para la variable dependiente*




- Primero, introduzca 1 para Valor y Sí para Etiqueta de valor y, a continuación, pulse en Añadir.
- A continuación, introduzca 2 para Valor y No para Etiqueta de valor y, a continuación, vuelva a pulsar en Añadir.
- A continuación, pulse en Aceptar.
- Vuelva a abrir el cuadro de diálogo Árbol de decisión. En el cuadro de diálogo aún debe aparecer seleccionada *dependiente* como la variable dependiente, con un nivel de medida nominal.
- Pulse en Aceptar para volver a ejecutar el procedimiento.

Figura 3-7
*Árbol para la variable dependiente nominal con etiquetas de valor parciales*




Ahora sólo se incluirán en el modelo de árbol los dos valores de la variable dependiente con etiquetas de valor definidas. Se han excluido todos los casos con un valor igual a 3 para la variable dependiente, lo que podría no apreciarse con facilidad si no se está familiarizado con los datos.

## *Asignación de etiquetas de valor a todos los valores*

Para evitar la omisión accidental del análisis de valores categóricos válidos, utilice la opción Definir propiedades de variables para asignar etiquetas de valor a todos los valores de la variable dependiente encontrados en los datos.

Cuando aparezca la información del diccionario de datos para la variable *nombre* en el cuadro de diálogo Definir propiedades de variables, se observa que aunque hay unos 300 casos con valor igual a 3 para dicha variable, no se ha definido ninguna etiqueta de valor para dicho valor.

Figura 3-8

*Variable con etiquetas de valor parciales en el cuadro de diálogo Definir propiedades de variables*

Capítulo
4

# Utilización de árboles de decisión para evaluar riesgos de crédito

Los bancos mantienen una base de datos con información histórica sobre los clientes a los que el banco ha concedido préstamos, incluido si han o no reintegrado o causado mora en el pago de dichos préstamos. Es posible utilizar árboles de decisión para analizar las características de los dos grupos de clientes y generar modelos para pronosticar la verosimilitud de que los solicitantes de préstamos causen mora en el pago de los mismos.

Los datos de los créditos se almacenan en *tree_credit.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Decision Trees 19*.

## Creación del modelo

El procedimiento Árbol de decisión ofrece varios métodos diferentes para crear modelos de árboles. Para este ejemplo, utilizaremos el método por defecto:

**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.

### Creación del modelo de árbol CHAID

- Para ejecutar un análisis de Árbol de decisiones, elija en los menús:
  Analizar > Clasificar > Árbol...

Figura 4-1
*Cuadro de diálogo Árbol de decisión*


- Seleccione *Valoración de crédito* como la variable dependiente.
- Seleccione las restantes variables como variables independientes. (El procedimiento excluirá de forma automática cualquier variable cuya contribución al modelo final no sea significativa.)

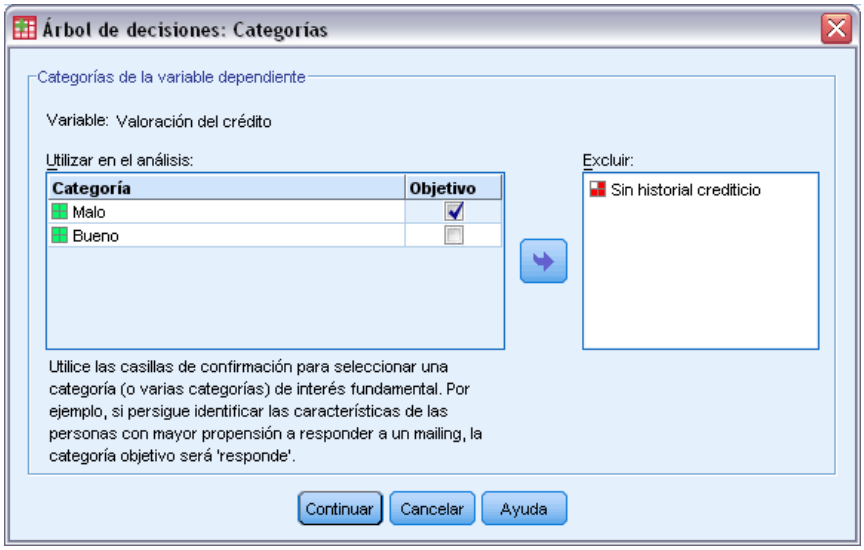
  En este momento ya se puede ejecutar el procedimiento y generar un modelo de árbol básico, pero vamos a seleccionar algunos resultados adicionales y realizar algunos pequeños ajustes a los criterios utilizados para generar el modelo.

## *Selección de categorías objetivo*

- Pulse en el botón Categorías situado debajo de la variable dependiente seleccionada.

Se abrirá el cuadro de diálogo Categorías, en el que se pueden especificar las categorías objetivo de interés de la variable dependiente. Hay que tener en cuenta que si bien las categorías objetivo no afectan al modelo del árbol propiamente dicho, algunos resultados y opciones sólo estarán disponibles si se han seleccionado categorías objetivo.

Figura 4-2
*Cuadro de diálogo Categorías*



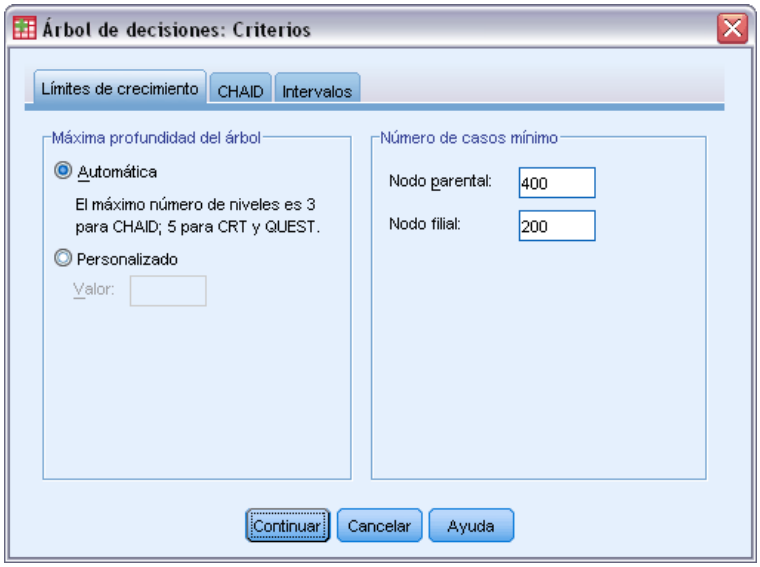

- Seleccione (marque) las casillas de verificación Objetivo para la categoría *Negativa*. Los clientes con una valoración del crédito negativa (que han causado mora en un préstamo) se tratarán como la categoría objetivo de interés.
- Pulse en Continuar.

## Especificación de los criterios de crecimiento del árbol

Para este ejemplo, deseamos que el árbol sea lo más sencillo posible, así que limitaremos el crecimiento del árbol elevando el número de casos mínimo para nodos parentales y filiales.

- En el cuadro de diálogo principal Árbol de decisión, pulse en Criterios.

Figura 4-3
*Cuadro de diálogo Criterios, pestaña Límites de crecimiento*



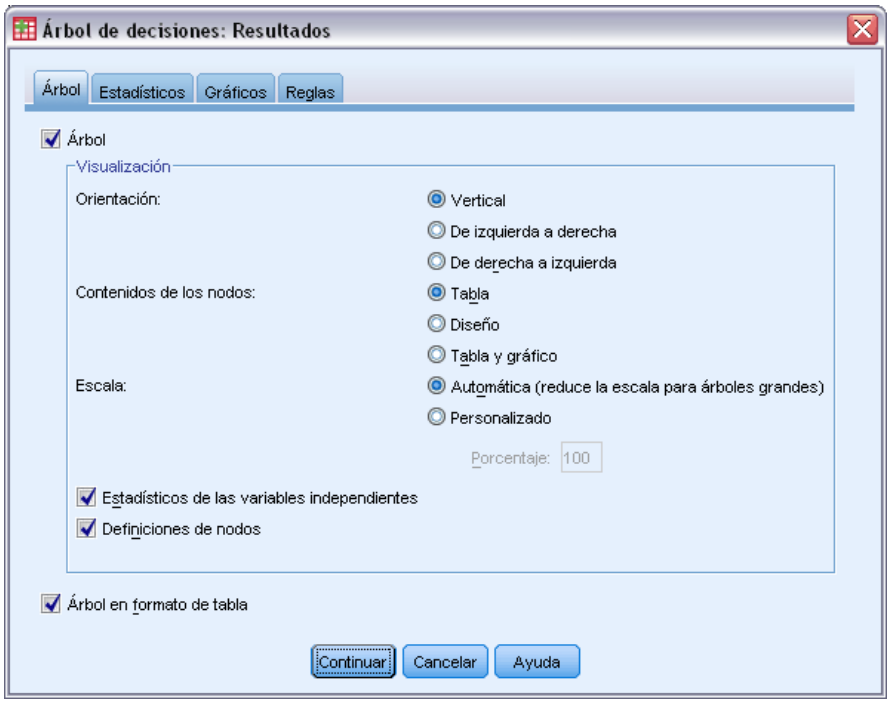

- En el grupo Número de casos mínimo, escriba 400 para Nodo parental y 200 para Nodo filial.
- Pulse en Continuar.

## *Selección de resultados adicionales*

- En el cuadro de diálogo Árbol de decisión, pulse en Resultados.

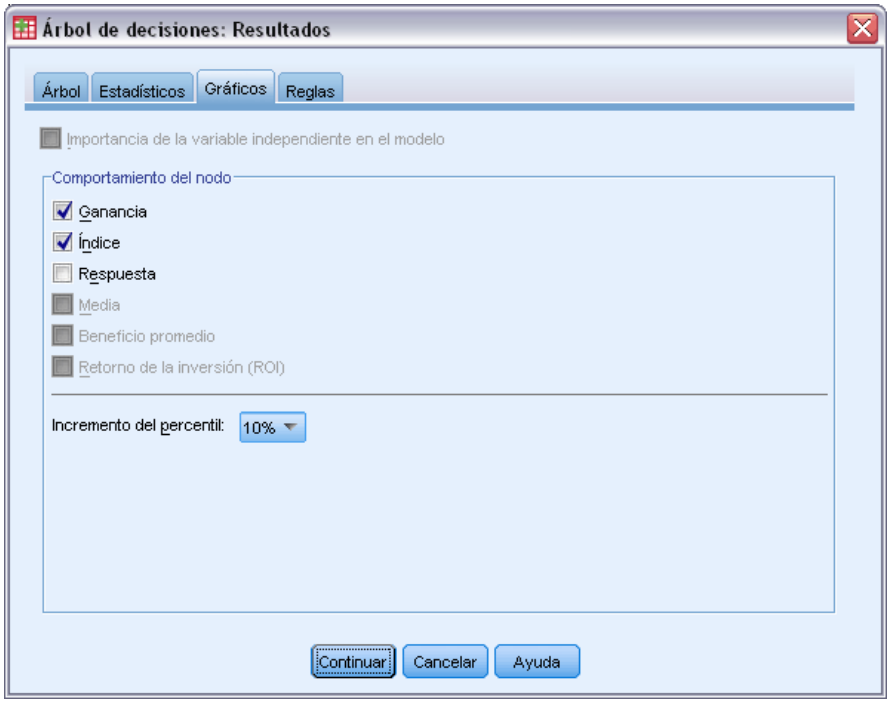
Se abrirá un cuadro de diálogo con pestañas, en el que podrá seleccionar distintos tipos de resultados adicionales.

Figura 4-4
*Cuadro de diálogo Resultados, pestaña Árbol*




- En la pestaña Árbol, seleccione (marque) Árbol en formato de tabla.
- A continuación, pulse en la pestaña Gráficos.

Figura 4-5
*Cuadro de diálogo Resultados, pestaña Gráficos*




▶ Seleccione (marque) Ganancia e Índice.

*Nota*: estos gráficos requieren una categoría objetivo para la variable dependiente. En este ejemplo sólo se podrá acceder a la pestaña Gráficos cuando se hayan seleccionado una o más categorías objetivo.

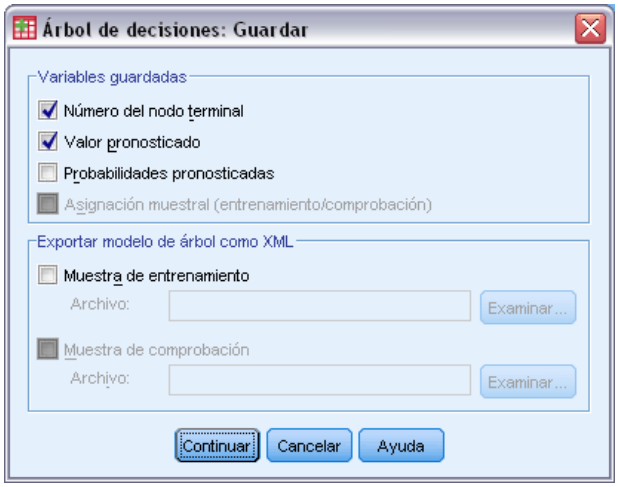
▶ Pulse en Continuar.

## *Almacenamiento de los valores pronosticados*

Es posible guardar las variables que contienen información sobre los pronósticos del modelo. Por ejemplo, puede guardar la valoración de crédito pronosticada para cada caso y, a continuación, comparar dichos pronósticos con las valoraciones de crédito reales.

▶ En el cuadro de diálogo principal Árbol de decisión, pulse en Guardar.

Figura 4-6
*Cuadro de diálogo Guardar*




- Seleccione (marque) Número del nodo terminal, Valor pronosticado y Probabilidades pronosticadas.
- Pulse en Continuar.
- En el cuadro de diálogo principal Árbol de decisión, pulse en Aceptar para ejecutar el procedimiento.

## Evaluación del modelo

Para este ejemplo, los resultados del modelo incluyen:

- Tablas que proporcionan información acerca del modelo.
- Diagrama del árbol.
- Gráficos que ofrecen una indicación sobre el rendimiento del modelo.
- Las variables de predicción del modelo añadidas al conjunto de datos activo.

## *Tabla de resumen del modelo*

Figura 4-7
*Resumen del modelo*

| | | |
|---|---|---|
| Especificaciones | Método de crecimiento | CHAID |
| | Variable dependiente | Valoración del crédito |
| | Variables independientes | Edad, Nivel de ingresos, Número de tarjetas de crédito, Educación, Créditos de coche |
| | Validación | SPLITSAMPLE |
| | Máxima profundidad de árbol | 3 |
| | Casos mínimos en nodo principal | 400 |
| | Casos mínimos en nodo secundario | 200 |
| Resultados | Variables independientes incluidas | Nivel de ingresos, Edad, Número de tarjetas de crédito, Créditos de coche |
| | Número de nodos | 10 |
| | Número de nodos terminales | 6 |
| | Profundidad | 3 |

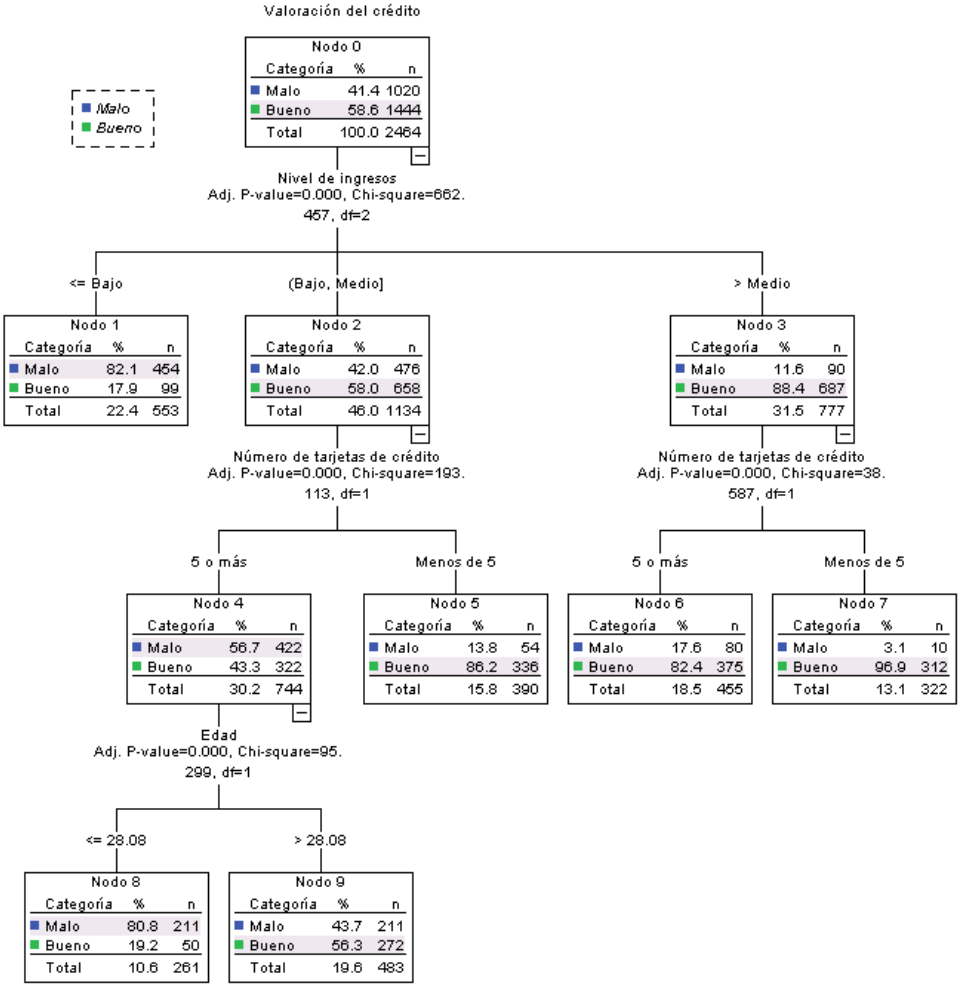
La tabla de resumen del modelo proporciona cierta información muy general sobre las especificaciones utilizadas para crear el modelo y sobre el modelo resultante.

- La sección Especificaciones ofrece información sobre los valores de configuración utilizados para generar el modelo de árbol, incluidas las variables utilizadas en el análisis.
- La sección Resultados muestra información sobre el número de nodos totales y terminales, la profundidad del árbol (número de niveles por debajo del nodo raíz) y las variables independientes incluidas en el modelo final.

Se han especificado cinco variables independientes, pero sólo se han incluido tres en el modelo final. Las variables para *estudios* y número actual de *préstamos para coches* no contribuyen de forma significativa al modelo, por lo que se eliminarán automáticamente del modelo final.

## Diagrama del árbol

Figura 4-8
*Diagrama del árbol para el modelo de valoración de créditos*




El diagrama del árbol es una representación gráfica del modelo del árbol. Este diagrama del árbol muestra que:

- Si se utiliza el método CHAID, *nivel de ingresos* es el mejor predictor para *valoración de crédito*.
- Para la categoría de ingresos bajos, *nivel de ingresos* es el único predictor significativo para *valoración de crédito*. De todos los clientes del banco que pertenecen a esta categoría, el 82% ha causado mora en los créditos. Como no hay ningún nodo filial por debajo de él, se considera un nodo **terminal**.
- Para las categorías de ingresos medios y altos, el siguiente mejor predictor es *número de tarjetas de crédito*.

- Para clientes con ingresos medios con cinco o más tarjetas de crédito, el modelo incluye un predictor más: *edad*. Cerca del 80% de dichos clientes con 28 o menos años tienen una valoración de crédito negativa, mientras que poco menos de la mitad de los clientes con más de 28 años tienen ese tipo de valoración.

Se puede utilizar el Editor del árbol para ocultar o mostrar ramas seleccionadas, cambiar el color y las fuentes, y seleccionar subconjuntos de casos basados en nodos seleccionados. Si desea obtener más información, consulte el tema Selección de casos en nodos el p. 74.

## Tabla del árbol

Figura 4-9
*Tabla del árbol para la valoración de créditos*

| | Malo | | Bueno | | Total | | | |
|---|---|---|---|---|---|---|---|---|
| Nodo | N | Porcentaje | N | Porcentaje | N | Porcentaje | Categoría pronosticada | Nodo principal |
| 0 | 1020 | 41,4% | 1444 | 58,6% | 2464 | 100,0% | Bueno | |
| 1 | 454 | 82,1% | 99 | 17,9% | 553 | 22,4% | Malo | 0 |
| 2 | 476 | 42,0% | 658 | 58,0% | 1134 | 46,0% | Bueno | 0 |
| 3 | 90 | 11,6% | 687 | 88,4% | 777 | 31,5% | Bueno | 0 |
| 4 | 422 | 56,7% | 322 | 43,3% | 744 | 30,2% | Malo | 2 |
| 5 | 54 | 13,8% | 336 | 86,2% | 390 | 15,8% | Bueno | 2 |
| 6 | 80 | 17,6% | 375 | 82,4% | 455 | 18,5% | Bueno | 3 |
| 7 | 10 | 3,1% | 312 | 96,9% | 322 | 13,1% | Bueno | 3 |
| 8 | 211 | 80,8% | 50 | 19,2% | 261 | 10,6% | Malo | 4 |
| 9 | 211 | 43,7% | 272 | 56,3% | 483 | 19,6% | Bueno | 4 |

La tabla del árbol, como su nombre indica, proporciona la mayor parte de la información esencial sobre el diagrama del árbol en forma de tabla. Para cada nodo, la tabla muestra:

- El número y porcentaje de casos dentro de cada categoría de la variable dependiente.
- La categoría pronosticada para la variable dependiente. En este ejemplo, la categoría pronosticada es la categoría *valoración del crédito*, con más del 50% de los casos en ese nodo, ya que sólo hay dos valoraciones de crédito posibles.
- El nodo parental para cada nodo del árbol. Observe que el nodo 1, el nodo de nivel de ingresos bajos, no es el nodo parental de ningún nodo. Como es un nodo terminal, no tiene ningún nodo filial.

Figura 4-10
*Tabla del árbol para la valoración de créditos (continuación)*

| Variable independiente primaria | | | | |
|---|---|---|---|---|
| Variable | Sig.[a] | Chi-cuadrado | gl | Segmentar valores |
| Nivel de ingresos | ,000 | 662,457 | 2 | <= Bajo |
| Nivel de ingresos | ,000 | 662,457 | 2 | (Bajo, Medio] |
| Nivel de ingresos | ,000 | 662,457 | 2 | > Medio |
| Número de tarjetas de crédito | ,000 | 193,113 | 1 | 5 o más |
| Número de tarjetas de crédito | ,000 | 193,113 | 1 | Menos de 5 |
| Número de tarjetas de crédito | ,000 | 38,587 | 1 | 5 o más |
| Número de tarjetas de crédito | ,000 | 38,587 | 1 | Menos de 5 |
| Edad | ,000 | 95,299 | 1 | <= 28,079205818990676 |
| Edad | ,000 | 95,299 | 1 | > 28,079205818990676 |

- Variable independiente utilizada para dividir el nodo.
- El valor de chi-cuadrado (ya que el árbol se generó utilizando el método CHAID), grados de libertad (*gl*) y nivel de significación (*Sig.*) para la división. Para propósitos más prácticos, es probable que sólo esté interesado en el nivel de significación, que es de menos de 0,0001 para todas las divisiones de este modelo.
- El valor o valores de la variable independiente para dicho nodo.

*Nota*: para variables independientes ordinales y de escala, puede que vea rangos en el árbol y en la tabla del árbol expresados con el formato general *(valor1, valor2]*, que básicamente significa "mayor que valor1 y menor o igual que valor2". En este ejemplo, el nivel de ingresos sólo tiene tres valores posibles, *Bajos*, *Medios* y *Altos*, y *(Bajos, Medios]* simplemente significa *Medios*. De manera similar, *>Medios* significa *Altos*.

## *Ganancias para nodos*

Figura 4-11
*Ganancias para nodos*

| | Nodo | | Ganancia | | | |
|---|---|---|---|---|---|---|
| Nodo | N | Porcentaje | N | Porcentaje | Respuesta | Índice |
| 1 | 553 | 22,4% | 454 | 44,5% | 82,1% | 198,3% |
| 8 | 261 | 10,6% | 211 | 20,7% | 80,8% | 195,3% |
| 9 | 483 | 19,6% | 211 | 20,7% | 43,7% | 105,5% |
| 6 | 455 | 18,5% | 80 | 7,8% | 17,6% | 42,5% |
| 5 | 390 | 15,8% | 54 | 5,3% | 13,8% | 33,4% |
| 7 | 322 | 13,1% | 10 | 1,0% | 3,1% | 7,5% |

Método de crecimiento: CHAID
Variable dependiente: Valoración del crédito

La tabla de ganancias para nodos ofrece un resumen de información sobre los nodos terminales del modelo.

- En esta tabla sólo se muestran los nodos terminales, aquellos en los que se detiene el crecimiento del árbol. Con frecuencia, el único interés lo suscitan los nodos terminales, ya que representan los mejores pronósticos de clasificación para el modelo.

- Como los valores de ganancia proporcionan información sobre las categorías objetivo, esta tabla sólo estará disponible si se especifican una o más categorías objetivo. En este ejemplo, sólo hay una categoría objetivo, por lo que sólo habrá una tabla de ganancias para nodos.
- *N del Nodo* indica el número de casos en cada nodo terminal y *Porcentaje del Nodo* indica el porcentaje del número total de casos en cada nodo.
- *N de Ganancia* indica el número de casos en cada nodo terminal en la categoría objetivo y *Porcentaje de la Ganancia* indica el porcentaje de casos en la categoría objetivo con respecto al número global de casos en la categoría objetivo; en este ejemplo, muestran el número y el porcentaje de casos con una valoración de crédito negativa.
- En el caso de variables dependientes categóricas, *Responde* indica el porcentaje de casos en el nodo en la categoría objetivo especificada. En este ejemplo, son los mismos porcentajes que se muestran en la categoría *Negativa* en el diagrama del árbol.
- En el caso de variables dependientes categóricas, *Índice* indica la razón del porcentaje de respuestas para la categoría objetivo en comparación con el porcentaje de respuestas de toda la muestra.

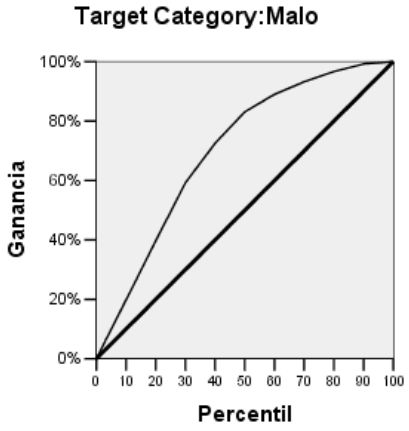
#### *Valores de índice*

El valor del índice es básicamente una indicación de cuánto difiere el porcentaje *observado* de la categoría objetivo para dicho nodo del porcentaje *esperado* para dicha categoría objetivo. El porcentaje de la categoría objetivo en el nodo raíz representa el porcentaje esperado antes de considerar los efectos de cualquiera de las variables independientes.

Un valor de índice superior al 100% significa que hay más casos en la categoría objetivo que el porcentaje global de dicha categoría objetivo. Por el contrario, un valor de índice inferior al 100% significa que hay menos casos en la categoría objetivo que el porcentaje global.

## *Gráfico de ganancias*

Figura 4-12
*Gráfico de ganancias para una categoría objetivo de valoración de crédito negativa*

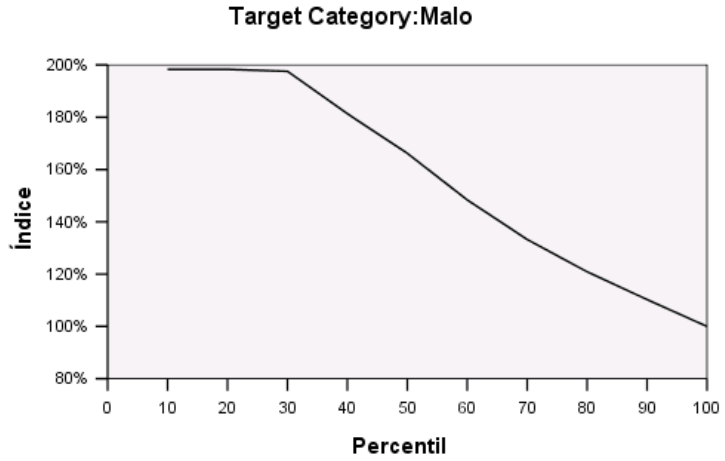



Este gráfico de ganancias indica que el modelo es bastante bueno.

Los gráficos de ganancias acumuladas siempre comienzan en el 0% y finalizan en el 100% al ir de un extremo a otro. Si el modelo es bueno, el gráfico de ganancias irá subiendo vertiginosamente hacia el 100% y, a continuación, se estabilizará. Un modelo que no proporciona ninguna información seguirá la línea diagonal de referencia.

## *Gráfico de índice*

Figura 4-13
*Gráfico de índice para una categoría objetivo de valoración de crédito negativa*




Este gráfico de índice indica que el modelo es bueno. Los gráficos de índices acumulados suelen comenzar por encima del 100% y descienden gradualmente hasta que alcanzan el 100%.

En un buen modelo, el valor de índice debe comenzar muy por encima del 100%, permanecer en una meseta elevada a medida que se avanza y, a continuación, descender bruscamente hasta el 100%. Un modelo que no proporciona ninguna información la línea rondará el 100% durante todo el gráfico.

## *Estimación de riesgo y clasificación*

Figura 4-14
*Tablas de riesgos y de clasificación*

**Riesgo**

| Estimación | Desviación Error |
|---|---|
| ,205 | ,008 |

Método de crecimiento: CHAID
Variable dependiente: Valoración del crédito

**Clasificación**

| | Pronosticado | | |
|---|---|---|---|
| Observado | Malo | Bueno | Porcentaje correcto |
| Malo | 665 | 355 | 65,2% |
| Bueno | 149 | 1295 | 89,7% |
| Porcentaje global | 33,0% | 67,0% | 79,5% |

Método de crecimiento: CHAID
Variable dependiente: Valoración del crédito

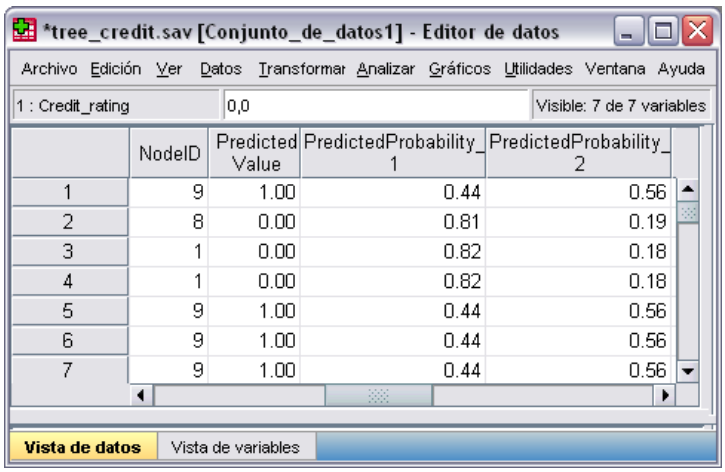
Las tablas de riesgos y de clasificación proporcionan una rápida evaluación de la bondad del funcionamiento del modelo.

- Una estimación de riesgo de 0,205 indica que la categoría pronosticada por el modelo (valoración de crédito positiva o negativa) es errónea para el 20,5% de los casos. Por lo tanto, el "riesgo" de clasificar erróneamente a un cliente es de aproximadamente el 21%.
- Los resultados en la tabla de clasificación son coherentes con la estimación de riesgo. La tabla muestra que el modelo clasifica de forma correcta, aproximadamente, al 79,5% de los clientes.

No obstante, la tabla de clasificación revela un problema potencial con este modelo: Para aquellos clientes con una valoración de crédito negativa, pronostica una valoración negativa para sólo el 65% de ellos, lo que significa que el 35% de los clientes con una valoración de crédito negativa aparecen inapropiadamente clasificados como clientes "buenos".

## Valores pronosticados

Figura 4-15
*Variables nuevas para valores pronosticados y probabilidades*




| | NodeID | PredictedValue | PredictedProbability_1 | PredictedProbability_2 |
|---|---|---|---|---|
| 1 | 9 | 1.00 | 0.44 | 0.56 |
| 2 | 8 | 0.00 | 0.81 | 0.19 |
| 3 | 1 | 0.00 | 0.82 | 0.18 |
| 4 | 1 | 0.00 | 0.82 | 0.18 |
| 5 | 9 | 1.00 | 0.44 | 0.56 |
| 6 | 9 | 1.00 | 0.44 | 0.56 |
| 7 | 9 | 1.00 | 0.44 | 0.56 |

Se han creado cuatro variables nuevas en el conjunto de datos activo:

**IDNodo.** Número del nodo terminal para cada caso.

**ValorPronosticado.** Valor pronosticado de la variable dependiente para cada caso. Como la variable dependiente está codificada como 0 = *Negativa* y 1 = *Positiva*, un valor pronosticado igual a 0 significa que el pronóstico del caso es una valoración de crédito negativa.

**ProbabilidadPronosticada.** Probabilidad de que el caso pertenezca a cada categoría de la variable dependiente. Como sólo hay dos valores posibles para la variable dependiente, se crean dos variables:

- **ProbabilidadPronosticada_1.** Probabilidad de que el caso pertenezca a la categoría de valoración de crédito negativa.
- **ProbabilidadPronosticada_2.** Probabilidad de que el caso pertenezca a la categoría de valoración de crédito positiva.

La probabilidad pronosticada es simplemente la proporción de casos en cada categoría de la variable dependiente para el nodo terminal que contiene cada caso. Por ejemplo, en el nodo 1, el 82% de los casos están en la categoría negativa y el 18% están en la categoría positiva, dando como resultado probabilidades pronosticadas de 0,82 y 0,18, respectivamente.

En caso de una variable dependiente categórica, el valor pronosticado es la categoría con la mayor proporción de casos en el nodo terminal para cada caso. Por ejemplo, para el primer caso, el valor pronosticado es 1 (valoración de crédito positiva) porque aproximadamente el 56% de los casos contenidos en su nodo terminal tienen una valoración de crédito positiva. Por el contrario, para el segundo caso, el valor pronosticado es 0 (valoración de crédito negativa) porque aproximadamente el 81% de los casos contenidos en su nodo terminal tienen una valoración de crédito negativa.

No obstante, si hay costes definidos, la relación entre la categoría pronosticada y las probabilidades pronosticadas puede que no sea tan directa. Si desea obtener más información, consulte el tema Asignación de costes a resultados el p. 78.

# Ajuste del modelo

En general, el modelo tiene una tasa de clasificación correcta situada justo por debajo del 80%. Esto se ve reflejado en la mayoría de los nodos terminales, en los que la categoría pronosticada, que aparece resaltada en el nodo, es la misma que la categoría real para el 80% o más de los casos.

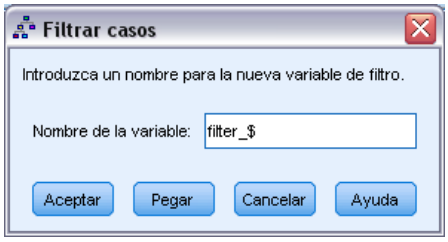
No obstante, hay un nodo terminal en el que los casos están uniformemente divididos entre valoraciones de crédito positivas y negativas. En el nodo 9, la valoración del crédito pronosticada es "positiva", pero sólo el 56% de los casos del nodo tienen realmente una valoración positiva. Esto significa que casi la mitad de los casos del nodo (44%) tendrán la categoría pronosticada errónea. Y considerando que el principal objetivo es la identificación de riesgos crediticios negativos, este nodo no realiza su función correctamente.

## Selección de casos en nodos

Estudiemos los casos del nodo 9 para ver si los datos revelan alguna información adicional de utilidad.

- Pulse dos veces en el árbol en la ventana del Visor para abrir el Editor del árbol.
- Pulse en el nodo 9 para seleccionarlo. (Si desea seleccionar varios nodos, mantenga pulsada la tecla Ctrl al mismo tiempo que pulsa el botón del ratón).
- En los menús del Editor del árbol, seleccione:
  Reglas > Filtrar casos...

Figura 4-16
*Cuadro de diálogo Filtrar casos*



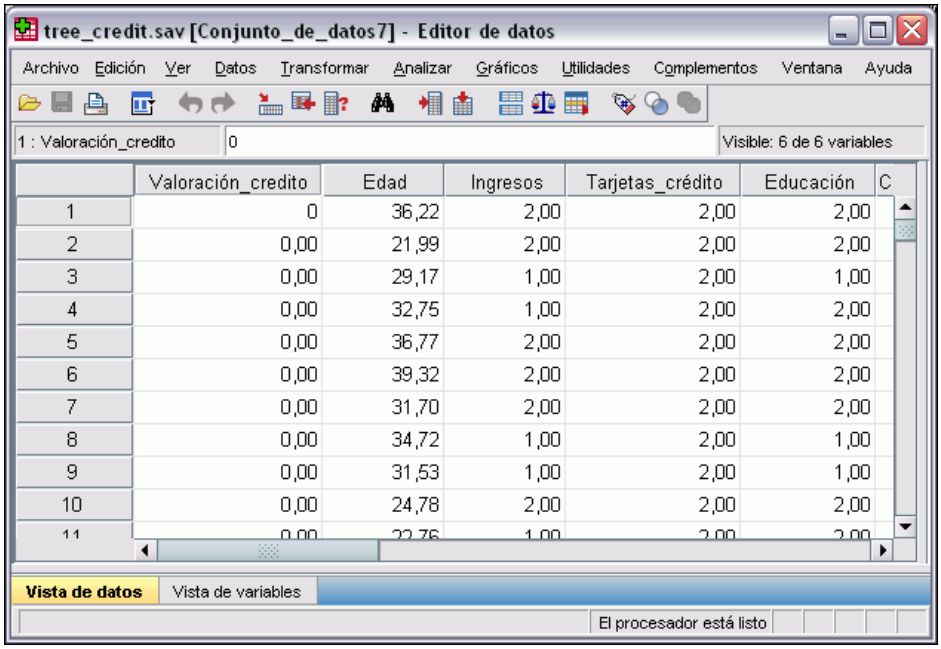

El cuadro de diálogo Filtrar casos creará una variable de filtro y aplicará un ajuste de filtrado basado en los valores de dicha variable. El nombre por defecto de una variable de filtro es *filter_$*.

- 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 de los análisis subsiguientes hasta que se modifique el estado del filtro.

En este ejemplo, esto significa que se filtrarán (pero no se eliminarán) los casos que no estén en el nodo 9.

- Pulse en Aceptar para crear la variable de filtro y aplicar la condición de filtrado.

Figura 4-17
*Casos filtrados en el Editor de datos*




En el Editor de datos, los casos que se han filtrado se indican con una barra transversal sobre el número de fila. Se filtrarán todos los casos que no estén en el nodo 9. Y viceversa, no se filtrarán aquellos casos que estén en el nodo 9; por consiguiente los subsiguientes análisis incluirán sólo los casos del nodo 9.

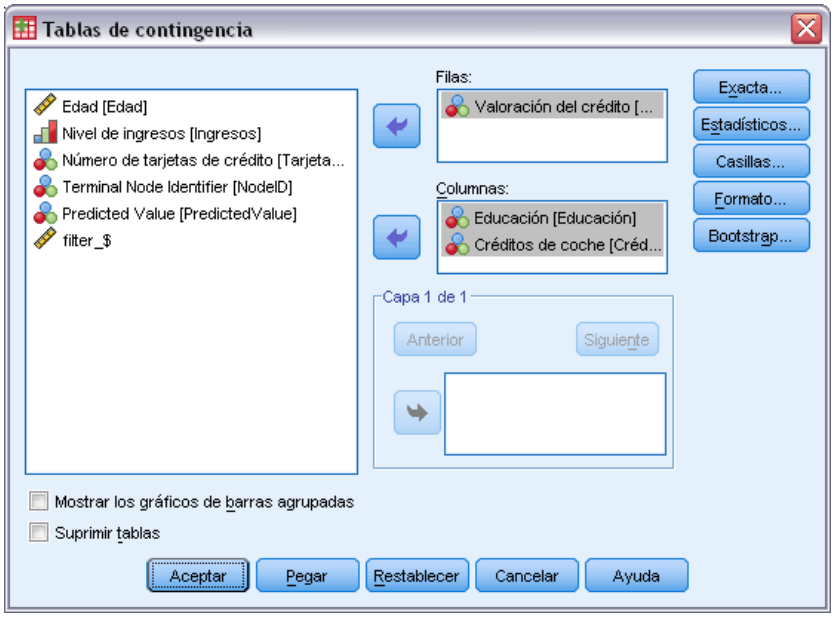
## *Examen de los casos seleccionados*

Como primer paso para el examen de los casos del nodo 9, podría ser interesante observar las variables que no se utilizan en este modelo. En este ejemplo, todas las variables del archivo de datos se han incluido en el análisis, pero dos de ellas no se han incluido en el modelo final: *estudios* y *préstamos para coches*. Como seguramente existe un buen motivo para que el procedimiento las haya excluido del modelo final, es probable que no nos ofrezcan mucha información. A pesar de ello, vamos a observarlas.

▶ Elija en los menús:

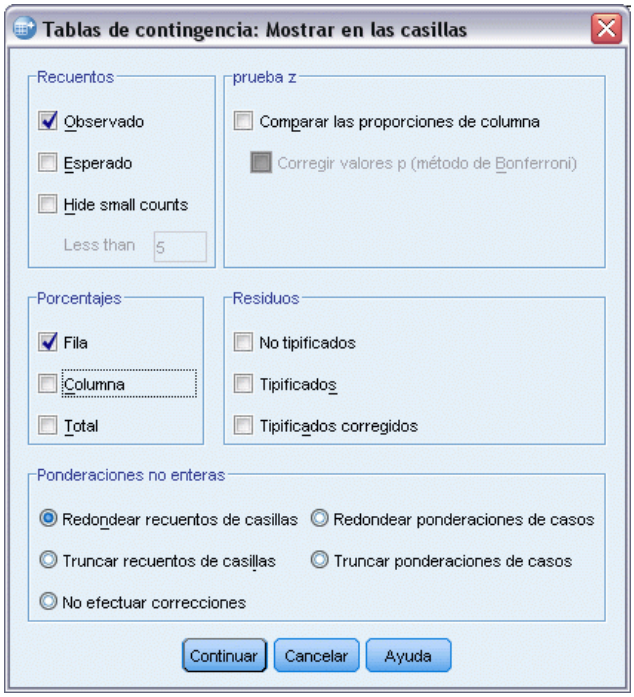
Analizar > Estadísticos descriptivos > Tablas de contingencia...

Figura 4-18
*Cuadro de diálogo Tablas de contingencia*




- Seleccione *Valoración de crédito* como la variable de fila.
- Seleccione *Estudios* y *Préstamos para coches* como las variables de columna.
- Pulse en Casillas.

Figura 4-19
*Cuadro de diálogo Tablas de contingencia: Mostrar en las casillas*




- Seleccione (marque) Fila en el grupo Porcentajes.
- A continuación, pulse en Continuar y, en el cuadro de diálogo principal Tablas de contingencia, pulse en Aceptar para ejecutar el procedimiento.

Al examinar las tablas de contingencia, se observa que no existe una gran diferencia entre casos en las categorías de valoración de crédito positiva y negativa para las dos variables que no se han incluido en el modelo.

Figura 4-20
*Tablas de contingencia para los casos del nodo seleccionado*

**Tabla de contingencia Valoración del crédito * Educación**

| | | | Educación | | |
|---|---|---|---|---|---|
| | | | Universitario | Bachillerato | Total |
| Valoración del crédito | Malo | Recuento | 110 | 101 | 211 |
| | | % de Valoración del crédito | 52,1% | 47,9% | 100,0% |
| | Bueno | Recuento | 128 | 144 | 272 |
| | | % de Valoración del crédito | 47,1% | 52,9% | 100,0% |
| Total | | Recuento | 238 | 245 | 483 |
| | | % de Valoración del crédito | 49,3% | 50,7% | 100,0% |

**Tabla de contingencia Valoración del crédito * Créditos de coche**

| | | | Créditos de coche | | |
|---|---|---|---|---|---|
| | | | 1 o ninguno | 2 o más | Total |
| Valoración del crédito | Malo | Recuento | 18 | 193 | 211 |
| | | % de Valoración del crédito | 8,5% | 91,5% | 100,0% |
| | Bueno | Recuento | 39 | 233 | 272 |
| | | % de Valoración del crédito | 14,3% | 85,7% | 100,0% |
| Total | | Recuento | 57 | 426 | 483 |
| | | % de Valoración del crédito | 11,8% | 88,2% | 100,0% |

- Para la variable *estudios*, un poco más de la mitad de los casos con una valoración de crédito negativa sólo tienen estudios secundarios, mientras que un poco más de la mitad de los casos con una valoración de crédito positiva tienen estudios universitarios; si bien esta diferencia no es estadísticamente significativa.
- Para la variable *préstamos para coches*, el porcentaje de casos de créditos positivos con sólo uno o ningún préstamo para coche es superior al porcentaje correspondiente a los casos de créditos negativos, pero la amplia mayoría de casos en ambos grupos tiene uno o más préstamos para coches.

Por lo tanto, aunque ahora ya esté claro por qué no se incluyeron estas variables en el modelo final, desafortunadamente no hemos obtenido ninguna información sobre cómo mejorar la predicción para el nodo 9. Si hubiera otras variables no especificadas para el análisis, puede que desee examinar algunas antes de continuar.

## Asignación de costes a resultados

Tal y como se ha comentado anteriormente, aparte del hecho de que casi la mitad de los casos del nodo 9 pertenecen a cada una de las categorías de valoración de crédito, la cuestión de que la categoría pronosticada sea “positiva” es problemática si el objetivo principal es generar un modelo que identifique correctamente los riesgos de crédito negativos. Pese a que es posible que no se pueda mejorar el rendimiento del nodo 9, aún se puede ajustar el modelo para mejorar la tasa de

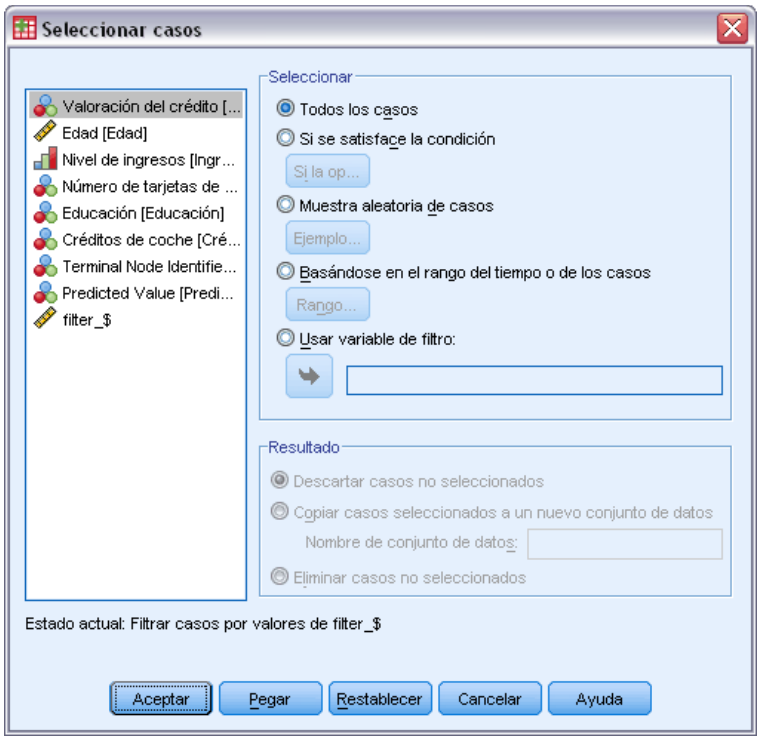
clasificación correcta de los casos de valoración de crédito negativa; aunque esto dará como resultado una mayor tasa de clasificación errónea para los casos de valoración de crédito positiva.

Primero es necesario desactivar el filtrado de casos de manera que todos los casos se vuelvan a utilizar en el análisis.

- Elija en los menús:
  Datos > Seleccionar casos...

- En el cuadro de diálogo Seleccionar casos, seleccione Todos los casos y, a continuación, pulse en Aceptar.

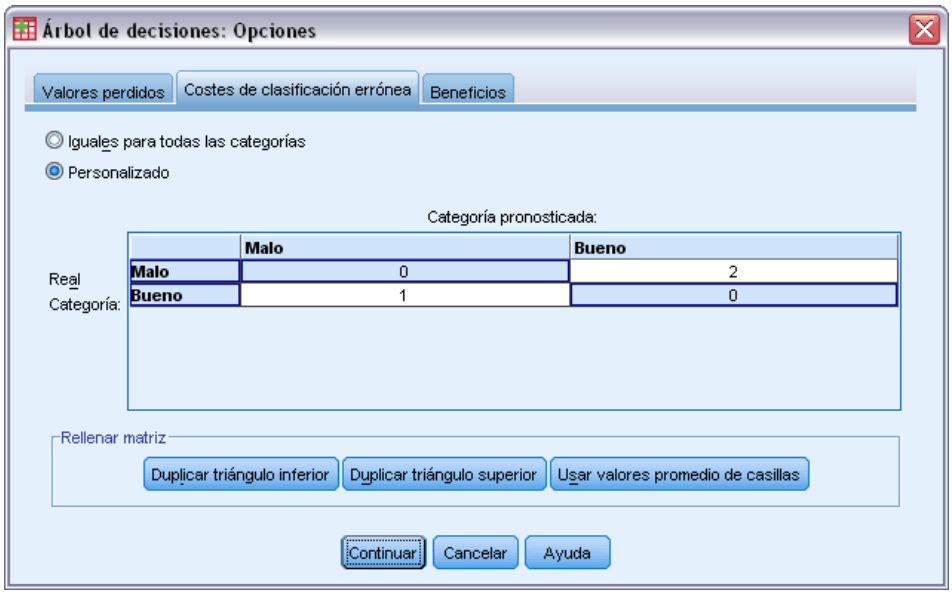
Figura 4-21
*Cuadro de diálogo Seleccionar casos*




- Abra el cuadro de diálogo Árbol de decisión y pulse en Opciones.

▶ Pulse en la pestaña Costes de clasificación errónea.

Figura 4-22
*Cuadro de diálogo Opciones, pestaña Costes de clasificación errónea*

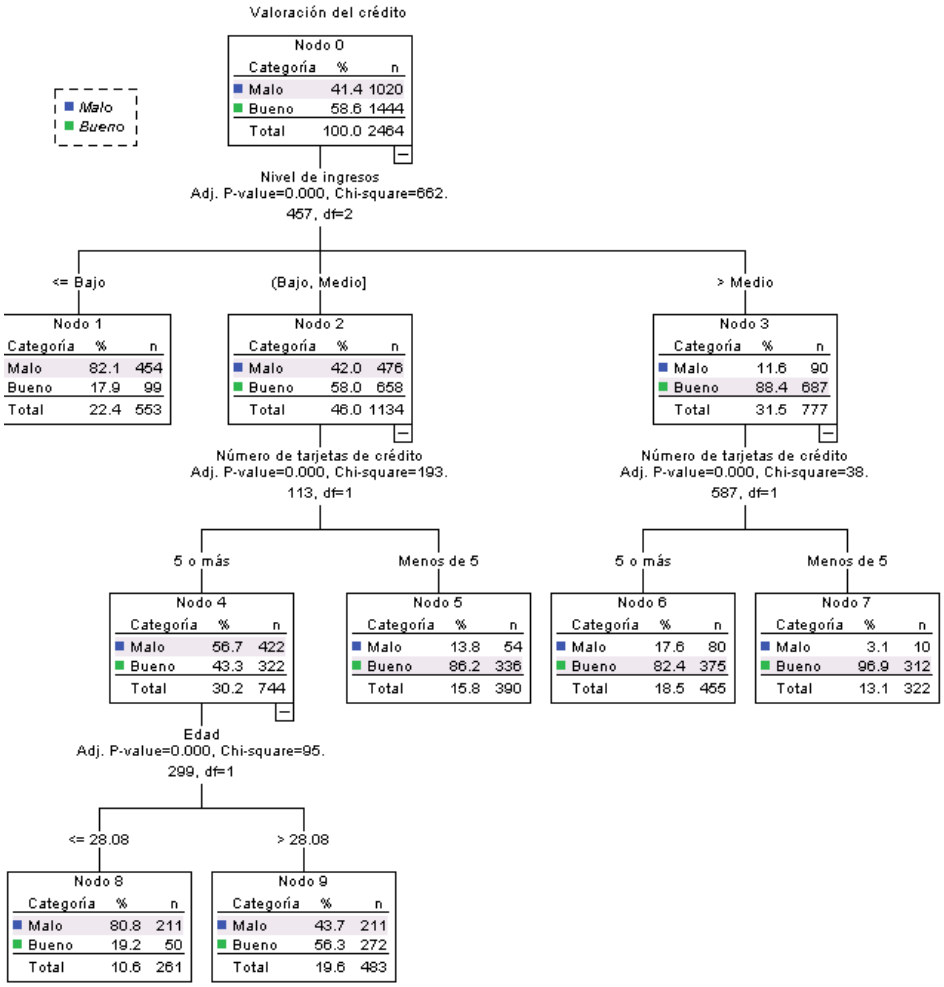



▶ Seleccione Personalizar y, para la Categoría real *Negativa* / Categoría pronosticada *Positiva*, introduzca un valor de 2.

Esto indica al procedimiento que el "coste" de clasificar erróneamente un riesgo de crédito negativo como positivo es el doble de alto que el "coste" de clasificar erróneamente un riesgo de crédito positivo como negativo.

▶ Pulse en Continuar y, a continuación, pulse en Aceptar en el cuadro de diálogo principal para ejecutar el procedimiento.

Figura 4-23
*Modelo del árbol con los valores de costes corregidos*




A primera vista, el árbol generado por el procedimiento parece esencialmente el mismo que el árbol original. Sin embargo, una inspección más detallada revela que si bien la distribución de los casos en cada nodo no ha variado, algunas categorías pronosticadas sí lo han hecho.

En el caso de los nodos terminales, la categoría pronosticada sigue siendo la misma en todos los nodos excepto en uno: el nodo 9. La categoría pronosticada es ahora *Negativa*, a pesar de que más de la mitad de los casos están en la categoría *Positiva*.

Como hemos indicado al procedimiento que la clasificación errónea de los riesgos de crédito negativos como positivos tenía un coste superior a la clasificación errónea de los riesgos de crédito positivos como negativos, cualquier nodo en el que los casos estén distribuidos de una forma bastante uniforme entre las dos categorías, ahora tendrá una categoría pronosticada *Negativa*, a pesar de que una ligera mayoría de casos esté en la categoría *Positiva*.

Este cambio en la categoría pronosticada se refleja en la tabla de clasificación.

Figura 4-24
*Tablas de riesgos y de clasificación basadas en costes corregidos*

**Riesgo**

| Estimación | Desviación Error |
|---|---|
| ,288 | ,011 |

Método de crecimiento: CHAID
Variable dependiente: Valoración del crédito

**Clasificación**

| | Pronosticado | | |
|---|---|---|---|
| Observado | Malo | Bueno | Porcentaje correcto |
| Malo | 876 | 144 | 85,9% |
| Bueno | 421 | 1023 | 70,8% |
| Porcentaje global | 52,6% | 47,4% | 77,1% |

Método de crecimiento: CHAID
Variable dependiente: Valoración del crédito

- Casi el 86% de los riesgos de crédito negativos aparecen ahora correctamente clasificados, comparado con el anterior 65%.
- Por otra parte, la correcta clasificación de los riesgos de crédito positivos ha disminuido del 90% al 71% y la clasificación correcta global ha descendido del 79,5% al 77,1%.

Se observa también que la estimación de riesgo y la tasa de clasificación correcta global ya no son coherentes la una con la otra. Si la tasa de clasificación correcta global es del 77,1%, se esperaría una estimación de riesgo de 0,229. En este ejemplo, al aumentar el coste de clasificación errónea para los casos de créditos negativos, se ha inflado el valor de riesgo, haciendo que su interpretación sea más compleja.

## Resumen

Se pueden utilizar los modelos de árbol para clasificar casos en grupos identificados por ciertas características, como son las características asociadas con los clientes de los bancos con registros de créditos positivos y negativos. Si un determinado resultado pronosticado es más importante que los demás posibles resultados, se puede ajustar el modelo para asociar un mayor coste de clasificación errónea a dicho resultado; sin embargo, la reducción de las tasas de clasificación errónea para un resultado aumentará las tasas de clasificación errónea para otros resultados.

Capítulo 5

# Creación de un modelo de puntuación

Una de las características más potentes y útiles del procedimiento Árbol de decisión es la capacidad de crear modelos que después se pueden aplicar a otros archivos de datos para pronosticar resultados. Por ejemplo, basándonos en un archivo de datos que contenga tanto información demográfica como información sobre precios de compra de vehículos, podemos generar un modelo que se pueda utilizar para pronosticar cuánto se gastarían en la compra de un nuevo coche personas con características demográficas similares; y, a continuación, aplicar dicho modelo a otros archivos de datos que contengan información demográfica pero no dispongan de información sobre adquisiciones previas de vehículos.

Para este ejemplo, utilizaremos el archivo de datos *tree_car.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Decision Trees 19*.

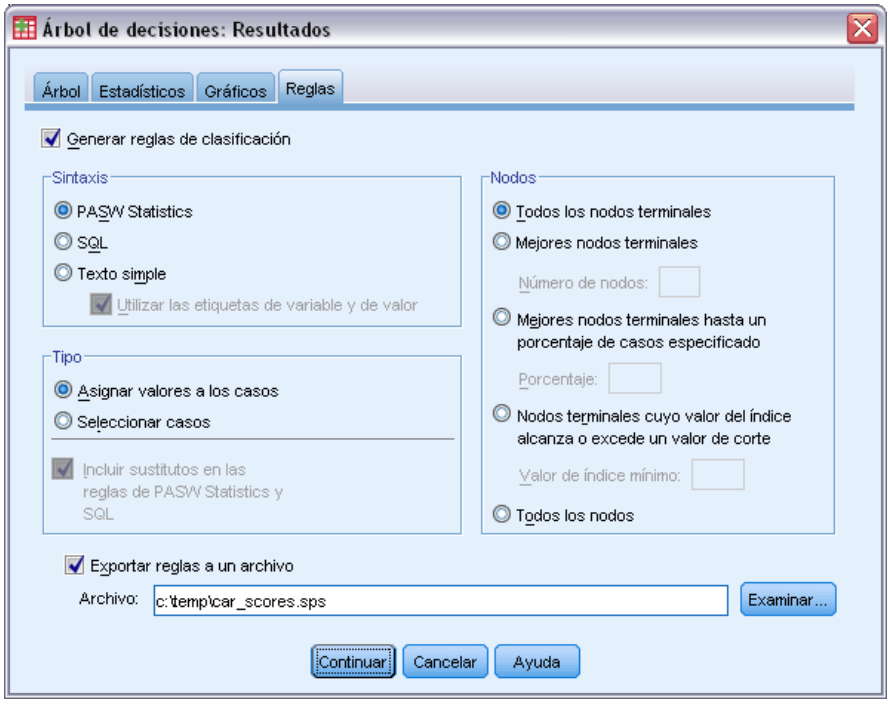
## Creación del modelo

▶ Para ejecutar un análisis de Árbol de decisiones, elija en los menús:
Analizar > Clasificar > Árbol...

Figura 5-1
*Cuadro de diálogo Árbol de decisión*


- Seleccione *Precio del vehículo principal* como la variable dependiente.
- Seleccione las restantes variables como variables independientes. (El procedimiento excluirá de forma automática cualquier variable cuya contribución al modelo final no sea significativa.)
- Para el método de crecimiento, seleccione CRT.
- Pulse en Resultados.

Figura 5-2
*Cuadro de diálogo Resultados, pestaña Reglas*




- Pulse en la pestaña Reglas.
- Seleccione (marque) Generar reglas de clasificación.
- Para Sintaxis, seleccione IBM® SPSS® Statistics.
- Para Tipo, seleccione Asignar valores a los casos.
- Seleccione (marque) Exportar reglas a un archivo e introduzca un nombre de archivo y la ubicación del directorio.

  Recuerde el nombre de archivo y la ubicación o anótelos porque necesitará esta información más adelante. Si no incluye una ruta de directorio, puede que no sepa dónde se ha guardado el archivo. Puede utilizar el botón Examinar para desplazarse hasta una ubicación de directorio específica (y válida).
- Pulse en Continuar y, a continuación, pulse en Aceptar para ejecutar el procedimiento y crear el modelo de árbol.

## Evaluación del modelo

Antes de aplicar el modelo a otros archivos de datos, probablemente deseará asegurarse de que el modelo funciona razonablemente bien con los datos originales utilizados para crearlo.

## *Resumen del modelo*

Figura 5-3
*Tabla de resumen del modelo*

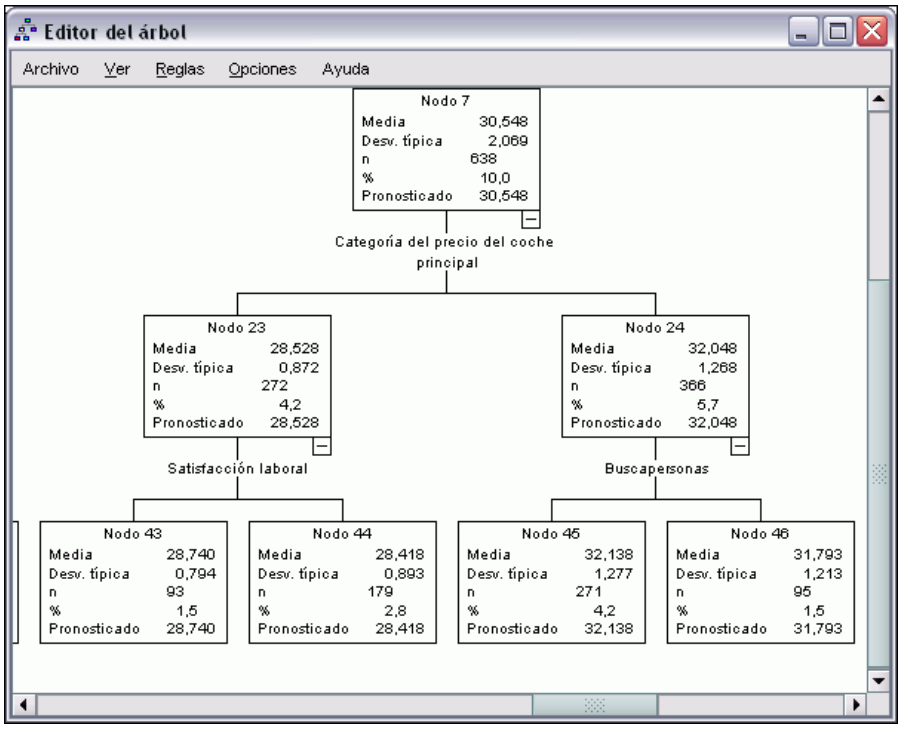
| | | |
|---|---|---|
| Especificaciones | Método de crecimiento | CRT |
| | Variable dependiente | Precio del vehículo principal |
| | Variables independientes | Edad en años, Sexo, Categoría de ingresos en miles, Nivel de estudios, Estado civil |
| | Validación | NONE |
| | Máxima profundidad de árbol | 5 |
| | Casos mínimos en nodo principal | 100 |
| | Casos mínimos en nodo secundario | 50 |
| Resultados | Variables independientes incluidas | Categoría de ingresos en miles, Edad en años, Nivel de estudios |
| | Número de nodos | 29 |
| | Número de nodos terminales | 15 |
| | Profundidad | 5 |

La tabla de resumen del modelo indica que sólo tres de las variables independientes seleccionadas han tenido una contribución lo suficientemente significativa como para ser incluidas en el modelo final: *ingresos*, *edad* y *estudios*. Esta información es importante si desea aplicar este modelo a otros archivos de datos, ya que las variables independientes utilizadas en la creación del modelo deberán estar presentes en todos los archivos de datos a los que se desee aplicar el modelo.

La tabla de resumen también indica que el propio modelo de árbol no es en particular un modelo simple ya que lo forman 29 nodos y 15 nodos terminales. Puede que este hecho no sea un problema si se desea un modelo fiable y que se pueda aplicar en la práctica en lugar de un modelo sencillo que sea fácil de describir o explicar. Por supuesto, para efectos prácticos, probablemente también desee un modelo que no dependa de demasiadas variables (predictoras) independientes. En este caso esto no es un problema ya que sólo se han incluido tres variables independientes en el modelo final.

## *Diagrama del modelo de árbol*

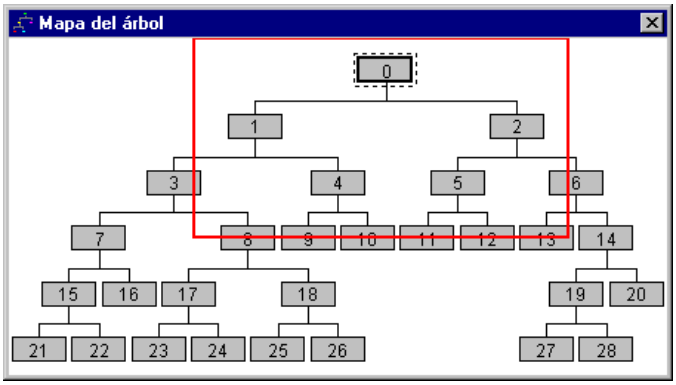
Figura 5-4
*Diagrama del modelo de árbol en el Editor del árbol*




El diagrama de modelo de árbol tiene tantos nodos que puede ser difícil ver el modelo en toda su extensión con un tamaño en el que la información contenida en el nodo aún sea legible. Puede utilizar el mapa del árbol para verlo completo:

- Pulse dos veces en el árbol en la ventana del Visor para abrir el Editor del árbol.
- En los menús del Editor del árbol, seleccione:
  Ver > Mapa del árbol

Figura 5-5
*Mapa del árbol*




- El mapa del árbol muestra todo el árbol. Se puede modificar el tamaño de la ventana del mapa del árbol, y se ampliará o reducirá la presentación del mapa del árbol para que se ajuste al tamaño de la ventana.
- El área resaltada en el mapa del árbol es el área del árbol que se muestra actualmente en el Editor del árbol.
- El mapa del árbol se puede utilizar para desplazarse por el árbol y seleccionar nodos.

Si desea obtener más información, consulte el tema Mapa del árbol en el capítulo 2 el p. 41.

En el caso de variables dependientes de escala, cada nodo muestra la media y la desviación típica de la variable dependiente. El nodo 0 muestra una media global del precio de compra de los vehículos de cerca de 29,9 (en miles), con una desviación típica de cerca de 21,6.

- El nodo 1, que representa los casos con unos ingresos por debajo de los 75 (también en miles), tiene una media del precio de los vehículos de sólo 18,7.
- En contraste, el nodo 2, que representa los casos con unos ingresos de 75 o más, tiene una media del precio de los vehículos de 60,9.

Un estudio en detalle del árbol mostraría que la *edad* y los *estudios* también presentan una relación con el precio de compra de los vehículos, pero en este momento estamos más interesados en la aplicación práctica del modelo que en un examen detallado de sus componentes.

## Estimación de riesgo

Figura 5-6
*Tabla de riesgo*

**Riesgo**

| Estimación | Desviación Error |
|---|---|
| 68,485 | 2,985 |

Método de crecimiento: CRT
Variable dependiente: Precio del vehículo principal

Ninguno de los resultados examinados hasta ahora nos indica si este es un modelo particularmente bueno. Un indicador del rendimiento del modelo es la estimación de riesgo. En el caso de una variable dependiente de escala, la estimación de riesgo es una medida de la varianza dentro del nodo, que por sí misma no aporta mucha información. Una menor varianza indica un mejor modelo, pero la varianza está relacionada con la unidad de medida. Si, por ejemplo, se hubiera registrado el precio en unidades en vez de en miles, la estimación de riesgo sería miles de veces más grande.

Para obtener una interpretación significativa de la estimación de riesgo con una variable dependiente de escala, es necesario realizar algunos pasos adicionales:

- La varianza total es igual a la varianza dentro del nodo (error) más la varianza entre los nodos (explicada).
- La varianza dentro del nodo es el valor de la estimación de riesgo: 68.485.
- La varianza total es la varianza para las variables dependientes antes de tener en consideración a las variables independientes o, lo que es lo mismo, la varianza en el nodo raíz.
- La desviación típica que se muestra en el nodo raíz es de 21,576; por lo que la varianza total es ese valor al cuadrado: 465.524.
- La proporción de la varianza debida al error (varianza no explicada) es 68,485/465,524 = 0,147.
- La proporción de la varianza explicada por el modelo es 1–0,147 = 0,853 ó 85,3%, lo que indica que es un modelo bastante bueno. (La interpretación de estos valores es similar a la de la tasa de clasificación correcta global para una variable dependiente categórica.)

## Aplicación del modelo a otro archivo de datos

Una vez que se ha determinado que el modelo es razonablemente bueno, se puede aplicar dicho modelo a otros archivos de datos que contengan variables de *edad*, *ingresos* y *estudios* similares y generar una variable nueva que represente el precio de compra de vehículos pronosticado para cada caso del archivo. A menudo, se hace referencia a este proceso como **puntuación**.

En el momento de generar el modelo, se especificó que las “reglas” para la asignación de valores a los casos se guardaran en un archivo de texto, con el formato de sintaxis de comandos. A continuación, se utilizarán los comandos almacenados en dicho archivo para generar puntuaciones en otro archivo de datos.

▶ Abra el archivo *tree_score_car.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Decision Trees 19*.

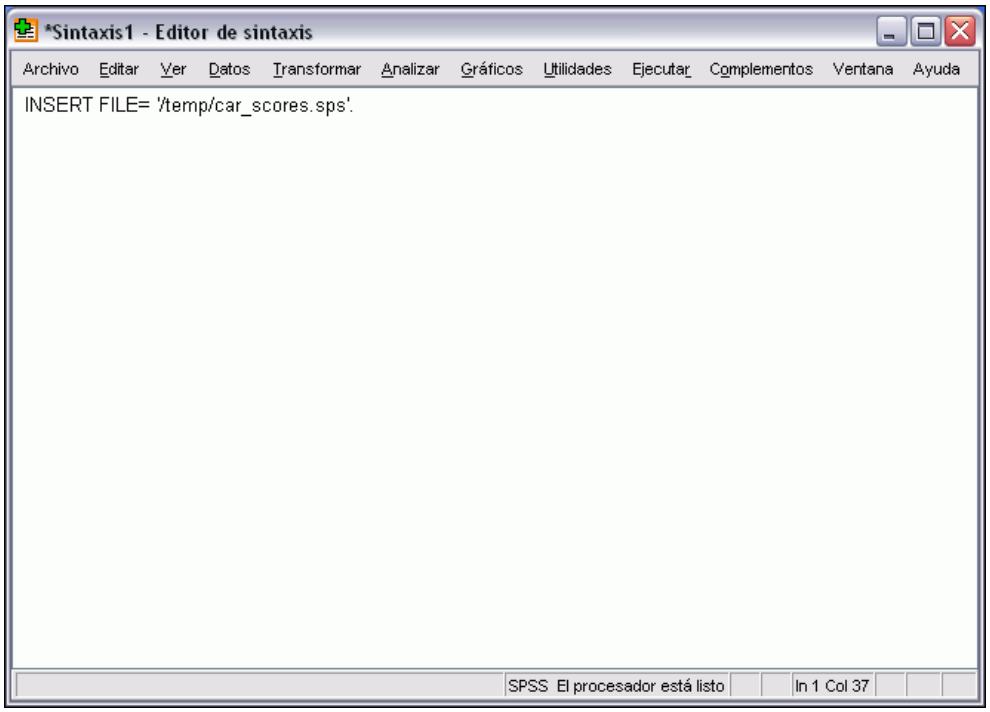
▶ A continuación, en los menús, elija:
Archivo > Nuevo > Sintaxis

▶ En la ventana de sintaxis de comandos, escriba:

```
INSERT FILE=
 '/temp/car_scores.sps'.
```

Si utilizó otro nombre de archivo o ubicación, realice las oportunas modificaciones.

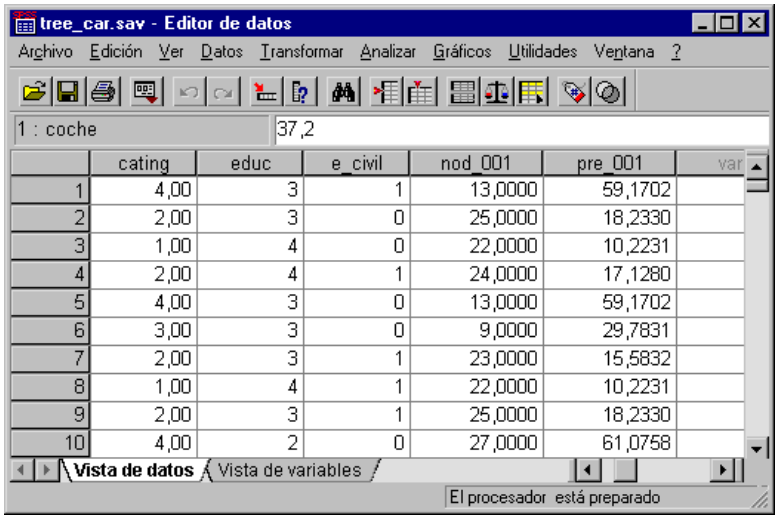
Figura 5-7
*Ventana de sintaxis con el comando INSERT para ejecutar un archivo de comandos*




El comando `INSERT` ejecutará los comandos almacenados en el archivo especificado, que es el archivo de "reglas" generado durante la creación del modelo.

- En los menús de la ventana de sintaxis de comandos, seleccione:
  Ejecutar > Todo

Figura 5-8
*Valores pronosticados añadidos al archivo de datos*




| | cating | educ | e_civil | nod_001 | pre_001 | var |
|---|---|---|---|---|---|---|
| 1 | 4,00 | 3 | 1 | 13,0000 | 59,1702 | |
| 2 | 2,00 | 3 | 0 | 25,0000 | 18,2330 | |
| 3 | 1,00 | 4 | 0 | 22,0000 | 10,2231 | |
| 4 | 2,00 | 4 | 1 | 24,0000 | 17,1280 | |
| 5 | 4,00 | 3 | 0 | 13,0000 | 59,1702 | |
| 6 | 3,00 | 3 | 0 | 9,0000 | 29,7831 | |
| 7 | 2,00 | 3 | 1 | 23,0000 | 15,5832 | |
| 8 | 1,00 | 4 | 1 | 22,0000 | 10,2231 | |
| 9 | 2,00 | 3 | 1 | 25,0000 | 18,2330 | |
| 10 | 4,00 | 2 | 0 | 27,0000 | 61,0758 | |

Este proceso añade dos nuevas variables al archivo de datos:

- *nod_001* contiene el número del nodo terminal pronosticado por el modelo para cada caso.
- *pre_001* contiene el valor pronosticado para el precio de compra de vehículos para cada caso.

Como hemos solicitado reglas para la asignación de valores para nodos terminales, el número de valores pronosticados posibles será el mismo que el número de nodos terminales, que en este caso es de 15. Por ejemplo, cada caso con un número de nodo pronosticado de 10 tendrá el mismo precio de compra de vehículos pronosticado: 30.56. Este es, y no por casualidad, el valor de la media indicado para el nodo terminal 10 en el modelo original.

Aunque normalmente el modelo se aplica a datos para los que no se conoce el valor de la variable dependiente, en este ejemplo, el archivo de datos al que se aplica el modelo contiene realmente dicha información; por lo que se pueden comparar las predicciones del modelo con los valores reales.

▶ Elija en los menús:

Analizar > Correlaciones > Bivariadas...

▶ Seleccione *Precio del vehículo principal* y *pre_001*.

Figura 5-9
*Cuadro de diálogo Correlaciones bivariadas*


▶ Pulse en Aceptar para ejecutar el procedimiento.

Figura 5-10
*Correlación entre el precio de los vehículos real y el precio pronosticado*

| | | Precio del vehículo principal | pre_001 |
|---|---|---|---|
| Precio del vehículo principal | Correlación de Pearson | 1 | ,923** |
| | Sig. (bilateral) | | ,000 |
| | N | 3110 | 3110 |
| pre_001 | Correlación de Pearson | ,923** | 1 |
| | Sig. (bilateral) | ,000 | |
| | N | 3110 | 3110 |

**. La correlación es significativa al nivel 0,01 (bilateral).

La correlación de 0,92 indica una correlación positiva muy alta entre el precio de los vehículos real y el precio pronosticado, lo que indica que el modelo funciona correctamente.

## Resumen

Se puede utilizar el procedimiento Árbol de decisión para crear modelos que después se pueden aplicar a otros archivos de datos para pronosticar resultados. El archivo de datos de destino 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 definidos como perdidos por el usuario (si hubiera). No obstante, no será necesario que en el archivo de datos de destino estén presentes ni la variable dependiente ni las variables independientes excluidas del modelo final.

Capítulo 6

# Valores perdidos en modelos de árbol

Los diferentes métodos de crecimiento tratan los valores perdidos para variables (predictoras) independientes de distintas maneras:

- CHAID y CHAID exhaustivo tratan los valores perdidos del sistema o definidos como perdidos por el usuario para cada variable independiente como una única categoría. En el caso de variables independientes ordinales y de escala, se podrá fundir dicha categoría a continuación con otras categorías de la variable independiente, dependiendo de los criterios de crecimiento.
- 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.

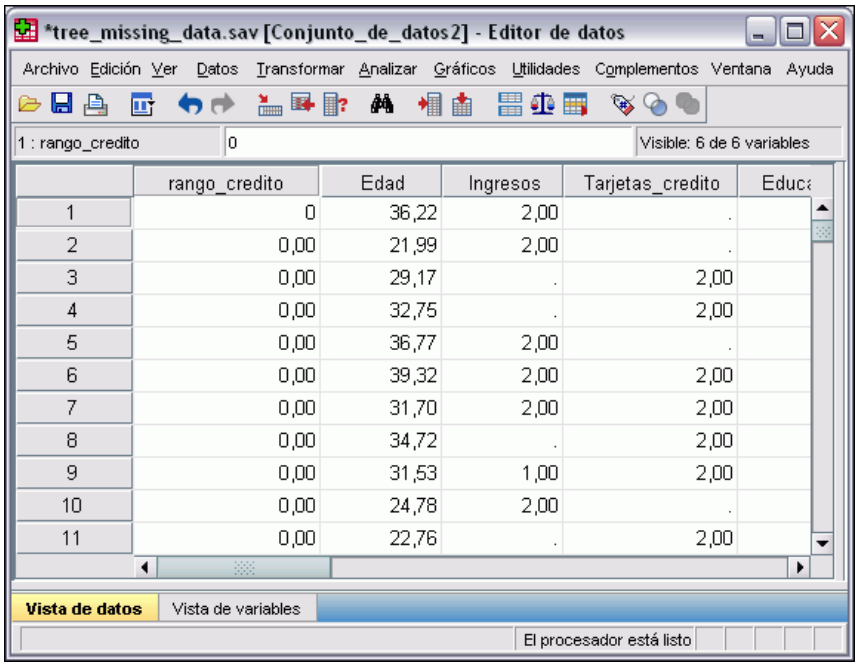
Este ejemplo muestra la diferencia entre CHAID y CRT cuando hay valores perdidos para variables independientes utilizadas en el modelo.

Para este ejemplo, utilizaremos el archivo de datos *tree_missing_data.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Decision Trees 19*.

*Nota*: en el caso de variables independientes nominales y de variables dependientes nominales, se puede elegir tratar los valores **definidos como perdidos por el usuario** como valores válidos, en cuyo caso dichos valores se tratarán como cualquier otro valor no perdido. Si desea obtener más información, consulte el tema Valores perdidos en el capítulo 1 el p. 22.

# Valores perdidos con CHAID

Figura 6-1
*Datos de crédito con valores perdidos*




| | rango_credito | Edad | Ingresos | Tarjetas_credito | Educa |
|---|---|---|---|---|---|
| 1 | 0 | 36,22 | 2,00 | . | |
| 2 | 0,00 | 21,99 | 2,00 | . | |
| 3 | 0,00 | 29,17 | . | 2,00 | |
| 4 | 0,00 | 32,75 | . | 2,00 | |
| 5 | 0,00 | 36,77 | 2,00 | . | |
| 6 | 0,00 | 39,32 | 2,00 | 2,00 | |
| 7 | 0,00 | 31,70 | 2,00 | 2,00 | |
| 8 | 0,00 | 34,72 | . | 2,00 | |
| 9 | 0,00 | 31,53 | 1,00 | 2,00 | |
| 10 | 0,00 | 24,78 | 2,00 | . | |
| 11 | 0,00 | 22,76 | . | 2,00 | |

De la misma manera que en el ejemplo del riesgo de crédito (para obtener más información, consulte el capítulo 4), en este ejemplo se intentará generar un modelo para clasificar los riesgos de crédito positivos y negativos. La principal diferencia es que este archivo de datos contiene valores perdidos para algunas variables independientes utilizadas en el modelo.

- Para ejecutar un análisis de Árbol de decisiones, elija en los menús:

  Analizar > Clasificar > Árbol...

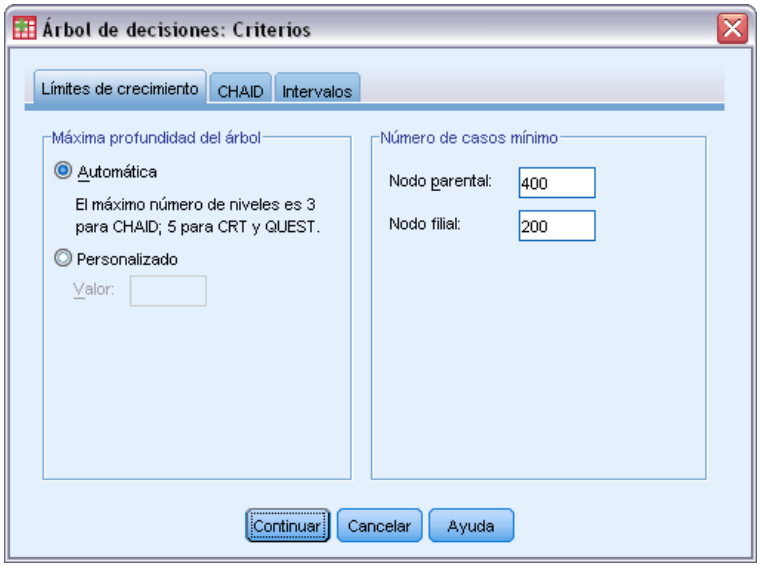
Figura 6-2
*Cuadro de diálogo Árbol de decisión*


- Seleccione *Valoración de crédito* como la variable dependiente.
- Seleccione todas las demás variables como variables independientes. (El procedimiento excluirá de forma automática cualquier variable cuya contribución al modelo final no sea significativa.)
- Para el método de crecimiento, seleccione CHAID.

  Para este ejemplo, deseamos que el árbol sea lo más sencillo posible, así que limitaremos el crecimiento del árbol elevando el número de casos mínimo para nodos parentales y filiales.

- En el cuadro de diálogo principal Árbol de decisión, pulse en Criterios.

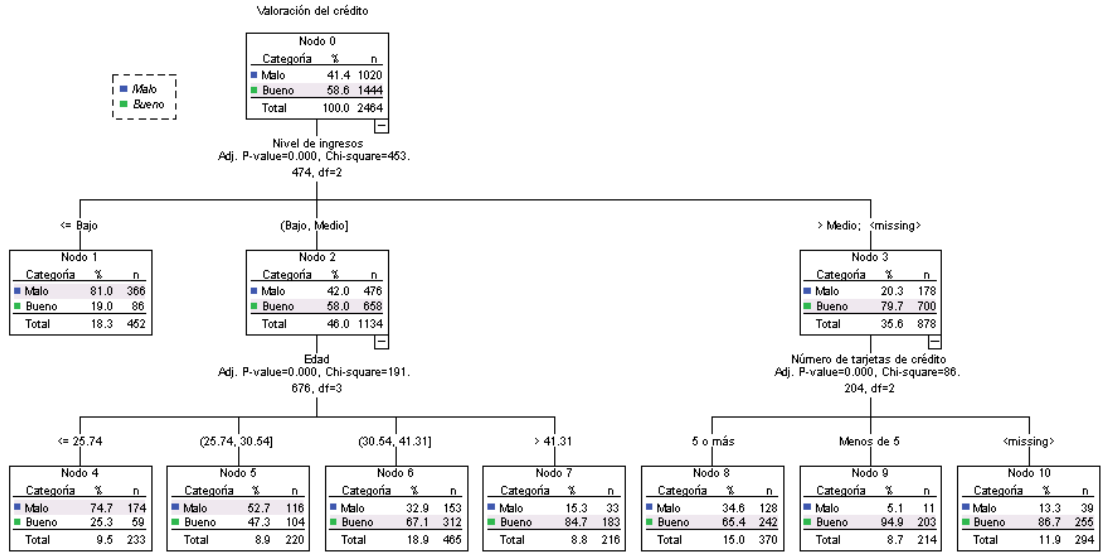
Figura 6-3
*Cuadro de diálogo Criterios, pestaña Límites de crecimiento*




- En el grupo Número de casos mínimo, escriba 400 para Nodo parental y 200 para Nodo filial.
- Pulse en Continuar y, a continuación, pulse en Aceptar para ejecutar el procedimiento.

## *Resultados de CHAID*

Figura 6-4
*Árbol CHAID con valores de variables independientes perdidos*




Para el nodo 3, el valor de *nivel de ingresos* aparece como *>Medio;<perdido>*. Esto significa que el nodo contiene casos en la categoría de ingresos altos además de todos los casos con valores perdidos para *nivel de ingresos*.

El nodo terminal 10 contiene casos con valores perdidos para *número de tarjetas de crédito*. Si está interesado en identificar riesgos de crédito positivos, éste es en realidad el segundo mejor nodo terminal, lo que puede ser problemático si se desea utilizar este modelo para pronosticar riesgos de crédito positivos. Probablemente, no es lo más deseable generar un modelo que pronostica una valoración de crédito positiva sencillamente porque no se tiene ninguna información sobre el número de tarjetas de crédito que tienen los casos y, además, es posible que alguno de dichos casos tengan información perdida sobre los niveles de ingresos.

Figura 6-5
*Tablas de riesgos y de clasificación para el modelo CHAID*

**Riesgo**

| Estimación | Desviación Error |
|---|---|
| ,249 | ,009 |

Método de crecimiento: CHAID
Variable dependiente: Valoración del crédito

**Clasificación**

| | Pronosticado | | |
|---|---|---|---|
| Observado | Malo | Bueno | Porcentaje correcto |
| Malo | 656 | 364 | 64,3% |
| Bueno | 249 | 1195 | 82,8% |
| Porcentaje global | 36,7% | 63,3% | 75,1% |

Método de crecimiento: CHAID
Variable dependiente: Valoración del crédito

Las tablas de riesgos y de clasificación indican que el modelo CHAID clasifica correctamente cerca del 75% de los casos. No es un mal porcentaje, pero tampoco es fantástico. Además, tenemos razones para sospechar que la tasa de clasificación correcta para los casos con valoración de crédito positiva sea excesivamente optimista, ya que se basa en parte en el supuesto de que la falta de información sobre dos variables independientes (*nivel de ingresos* y *número de tarjetas de crédito*) es una indicación de una valoración de crédito positiva.

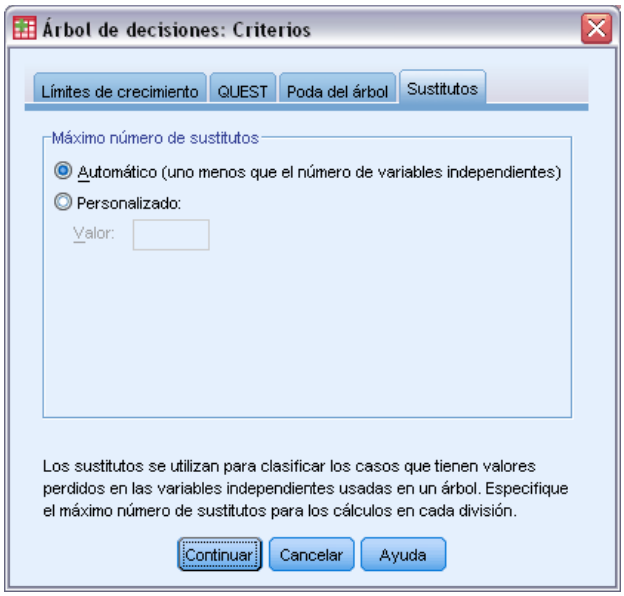
## Valores perdidos con CRT

A continuación probaremos los mismos análisis básicos, excepto que en esta ocasión utilizaremos CRT como método de crecimiento.

- En el cuadro de diálogo principal Árbol de decisión, para el método de crecimiento, seleccione CRT.
- Pulse en Criterios.
- Asegúrese de que el número de casos mínimo sigue establecido en 400 para los nodos parentales y en 200 para los nodos filiales.
- Pulse en la pestaña Sustitutos.

  *Nota*: la pestaña Sustitutos no será visible a menos que haya seleccionado CRT o QUEST como método de crecimiento.

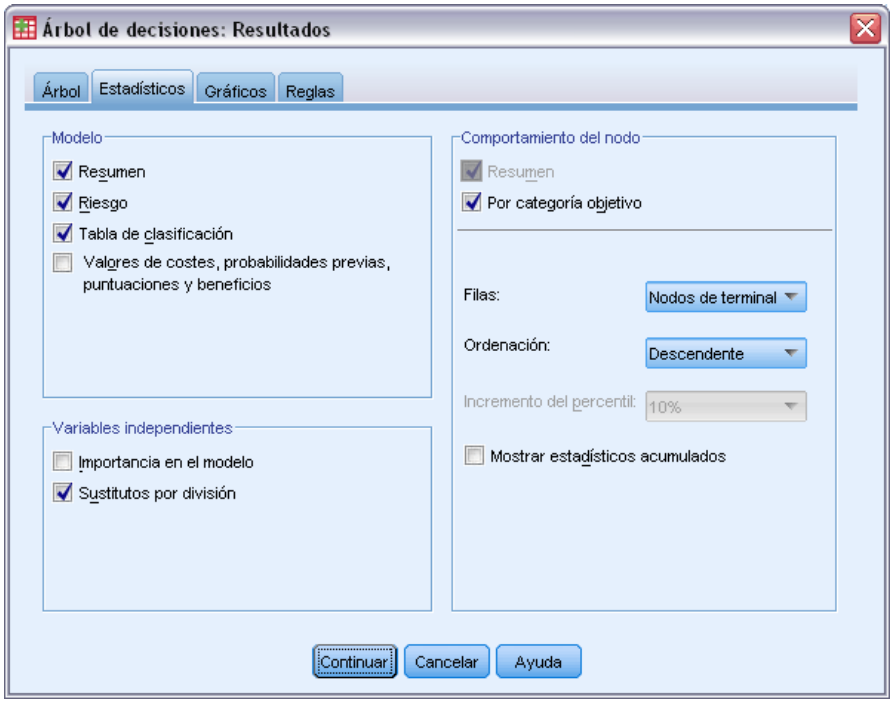
Figura 6-6
*Cuadro de diálogo Criterios, pestaña Sustitutos*




Para cada una de las divisiones de los nodos de las variables independientes, el ajuste Automático considerará todas las demás variables independientes del modelo como posibles sustitutos. Como en este ejemplo no hay muchas variables independientes, el ajuste Automático es adecuado.

- Pulse en Continuar.
- En el cuadro de diálogo Árbol de decisión, pulse en Resultados.

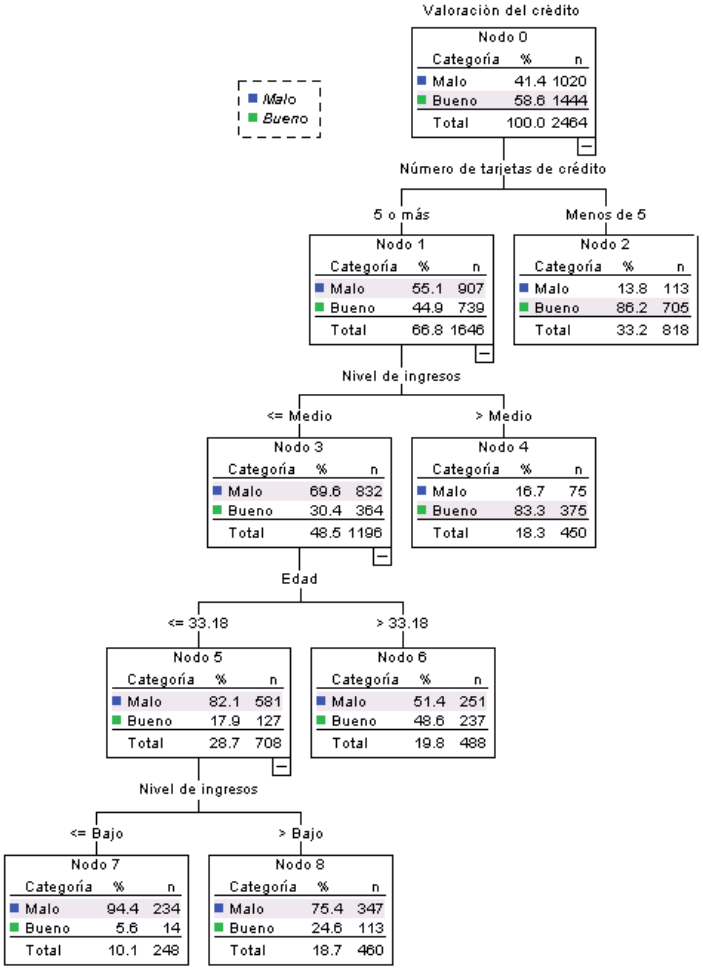
Figura 6-7
*Cuadro de diálogo Resultados, pestaña Estadísticos*




- Pulse en la pestaña Estadísticos.
- Seleccione Sustitutos por división.
- Pulse en Continuar y, a continuación, pulse en Aceptar para ejecutar el procedimiento.

## Resultados de CRT

Figura 6-8
*Árbol CRT con valores de variables independientes perdidos*




A primera vista ya se observa que este árbol no se parece mucho al árbol CHAID. De por sí, este hecho no tiene necesariamente que ser significativo. En un modelo de árbol CRT, todas las divisiones son binarias; es decir, cada nodo parental se divide únicamente en dos nodos filiales. En un modelo CHAID, los nodos parentales se pueden dividir en muchos nodos filiales. Por lo tanto, los árboles tienen un aspecto distinto aunque ambos representen el mismo modelo subyacente.

Sin embargo, existen varias diferencias importantes:

- La variable (predictora) independiente más importante del modelo CRT es *número de tarjetas de crédito*, mientras que en el modelo CHAID, el predictor más importante era *nivel de ingresos*.
- Para los casos con menos de cinco tarjetas de crédito, *número de tarjetas de crédito* es el único predictor significativo de la valoración de crédito y el nodo 2 es un nodo terminal.

- Igual que con el modelo CHAID, *nivel de ingresos* y *edad* también están incluidas en el modelo, aunque *nivel de ingresos* es ahora el segundo predictor en lugar del primero.
- No hay nodos que contengan una categoría <*perdido*>, porque CRT utiliza en el modelo predictores sustitutos en vez de valores perdidos.

Figura 6-9
*Tablas de riesgos y de clasificación para el modelo CRT*

**Riesgo**

| Estimación | Desviación Error |
|---|---|
| ,224 | ,008 |

Método de crecimiento: CRT
Variable dependiente: Valoración del crédito

**Clasificación**

| | Pronosticado | | |
|---|---|---|---|
| Observado | Malo | Bueno | Porcentaje correcto |
| Malo | 832 | 188 | 81,6% |
| Bueno | 364 | 1080 | 74,8% |
| Porcentaje global | 48,5% | 51,5% | 77,6% |

Método de crecimiento: CRT
Variable dependiente: Valoración del crédito

- Las tablas de riesgos y de clasificación muestran una tasa de clasificación correcta de casi un 78%, un ligero aumento frente al modelo CHAID (75%).
- La tasa de clasificación correcta para los casos con valoración de crédito negativa es mucho mayor para el modelo CRT: 81,6% frente a sólo un 64,3% del modelo CHAID.
- Sin embargo, la tasa de clasificación correcta para los casos con valoración de crédito positiva ha descendido del 82,8% del modelo CHAID al 74,8% del modelo CRT.

### *Sustitutos*

Las diferencias entre los modelos CHAID y CRT se deben, en parte, a la utilización de sustitutos en el modelo CRT. La tabla de sustitutos indica cómo se utilizaron los sustitutos en el modelo.

Figura 6-10
*Tabla Sustitutos*

| Nodo principal | Variable independiente | | Mejora | Association |
|---|---|---|---|---|
| 0 | Primario | Número de tarjetas de crédito | ,090 | |
| | Surrogate | Créditos de coche | ,052 | ,643 |
| | | Edad | ,001 | ,004 |
| 1 | Primario | Nivel de ingresos | ,071 | |
| | Surrogate | Edad | ,001 | ,004 |
| 3 | Primario | Edad | ,022 | |
| 5 | Primario | Nivel de ingresos | ,006 | |
| | Surrogate | Edad | 3,93E-005 | ,009 |

Growing Method: CRT
Dependent Variable: Valoración del crédito

- En el nodo raíz (nodo 0), la mejor variable (predictora) independiente es *número de tarjetas de crédito*.

- En todos los casos con valores perdidos para *número de tarjetas de crédito*, se utiliza *préstamos para coches* como el predictor sustituto, ya que esta variable tiene una asociación bastante alta (0,643) con *número de tarjetas de crédito*.
- Si un caso también tiene un valor perdido para *préstamos para coches*, entonces se utiliza *edad* como el sustituto (aunque tenga un valor de asociación bastante bajo de sólo 0,004).
- También se utiliza *edad* como sustituto para *nivel de ingresos* en los nodos 1 y 5.

## Resumen

Los distintos métodos de crecimiento tratan los datos perdidos de diferentes maneras. Si los datos que se han utilizado para crear el modelo contienen muchos valores perdidos (o si se desea aplicar un modelo a otros archivos de datos que contienen muchos valores perdidos), debe evaluar el efecto de los valores perdidos en los distintos modelos. Si desea utilizar sustitutos en el modelo para compensar el impacto los valores perdidos, utilice los métodos CRT o QUEST.

Apéndice
A

# Archivos muestrales

Los archivos muestrales instalados con el producto se encuentran en el subdirectorio *Samples* del directorio de instalación. Hay una carpeta independiente dentro del subdirectorio Samples para cada uno de los siguientes idiomas: Inglés, francés, alemán, italiano, japonés, coreano, polaco, ruso, chino simplificado, español y chino tradicional.

No todos los archivos muestrales están disponibles en todos los idiomas. Si un archivo muestral no está disponible en un idioma, esa carpeta de idioma contendrá una versión en inglés del archivo muestral.

### *Descripciones*

A continuación, se describen brevemente los archivos muestrales usados en varios ejemplos que aparecen a lo largo de la documentación.

- **accidents.sav.** Archivo de datos hipotéticos sobre una compañía de seguros que estudia los factores de riesgo de edad y género que influyen en los accidentes de automóviles de una región determinada. Cada caso corresponde a una clasificación cruzada de categoría de edad y género.
- **adl.sav.** Archivo de datos hipotéticos relativo a los esfuerzos para determinar las ventajas de un tipo propuesto de tratamiento para pacientes que han sufrido un derrame cerebral. Los médicos dividieron de manera aleatoria a pacientes (mujeres) que habían sufrido un derrame cerebral en dos grupos. El primer grupo recibió el tratamiento físico estándar y el segundo recibió un tratamiento emocional adicional. Tres meses después de los tratamientos, se puntuaron las capacidades de cada paciente para realizar actividades cotidianas como variables ordinales.
- **advert.sav.** Archivo de datos hipotéticos sobre las iniciativas de un minorista para examinar la relación entre el dinero invertido en publicidad y las ventas resultantes. Para ello, se recopilaron las cifras de ventas anteriores y los costes de publicidad asociados.
- **aflatoxin.sav.** Archivo de datos hipotéticos sobre las pruebas realizadas en las cosechas de maíz con relación a la aflatoxina, un veneno cuya concentración varía ampliamente en los rendimientos de cultivo y entre los mismos. Un procesador de grano ha recibido 16 muestras de cada uno de los 8 rendimientos de cultivo y ha medido los niveles de aflatoxinas en partes por millón (PPM).
- **aflatoxin20.sav.** Este archivo de datos contiene las medidas de aflatoxina de cada una de las 16 muestras de los rendimientos 4 y 8 procedentes del archivo de datos *aflatoxin.sav*.
- **anorectic.sav.** Mientras trabajaban en una sintomatología estandarizada del comportamiento anoréxico/bulímico, los investigadores realizaron un estudio de 55 adolescentes con trastornos de la alimentación conocidos. Cada paciente fue examinado cuatro veces durante cuatro años, lo que representa un total de 220 observaciones. En cada observación, se puntuó a los

pacientes por cada uno de los 16 síntomas. Faltan las puntuaciones de los síntomas para el paciente 71 en el tiempo 2, el paciente 76 en el tiempo 2 y el paciente 47 en el tiempo 3, lo que nos deja 217 observaciones válidas.

- **autoaccidents.sav.** Archivo de datos hipotéticos sobre las iniciativas de un analista de seguros para elaborar un modelo del número de accidentes de automóvil por conductor teniendo en cuenta la edad y el género del conductor. Cada caso representa un conductor diferente y registra el sexo, la edad en años y el número de accidentes de automóvil del conductor en los últimos cinco años.
- **band.sav** Este archivo de datos contiene las cifras de ventas semanales hipotéticas de CD de música de una banda. También se incluyen datos para tres variables predictoras posibles.
- **bankloan.sav.**Archivo de datos hipotéticos sobre las iniciativas de un banco para reducir la tasa de moras de créditos. El archivo contiene información financiera y demográfica de 850 clientes anteriores y posibles clientes. Los primeros 700 casos son clientes a los que anteriormente se les ha concedido un préstamo. Al menos 150 casos son posibles clientes cuyos riesgos de crédito el banco necesita clasificar como positivos o negativos.
- **bankloan_binning.sav.** Archivo de datos hipotéticos que contiene información financiera y demográfica sobre 5.000 clientes anteriores.
- **behavior.sav.** En un ejemplo clásico , se pidió a 52 estudiantes que valoraran las combinaciones de 15 situaciones y 15 comportamientos en una escala de 10 puntos que oscilaba entre 0 =“extremadamente apropiado” y 9=“extremadamente inapropiado”. Los valores promediados respecto a los individuos se toman como disimilaridades.
- **behavior_ini.sav.** Este archivo de datos contiene una configuración inicial para una solución bidimensional de *behavior.sav*.
- **brakes.sav.** Archivo de datos hipotéticos sobre el control de calidad de una fábrica que produce frenos de disco para automóviles de alto rendimiento. El archivo de datos contiene las medidas del diámetro de 16 discos de cada una de las 8 máquinas de producción. El diámetro objetivo para los frenos es de 322 milímetros.
- **breakfast.sav.**En un estudio clásico , se pidió a 21 estudiantes de administración de empresas de la Wharton School y sus cónyuges que ordenaran 15 elementos de desayuno por orden de preferencia, de 1=“más preferido” a 15=“menos preferido”. Sus preferencias se registraron en seis escenarios distintos, de “Preferencia global” a “Aperitivo, con bebida sólo”.
- **breakfast-overall.sav.** Este archivo de datos sólo contiene las preferencias de elementos de desayuno para el primer escenario, “Preferencia global”.
- **broadband_1.sav** Archivo de datos hipotéticos que contiene el número de suscriptores, por región, a un servicio de banda ancha nacional. El archivo de datos contiene números de suscriptores mensuales para 85 regiones durante un período de cuatro años.
- **broadband_2.sav** Este archivo de datos es idéntico a *broadband_1.sav* pero contiene datos para tres meses adicionales.
- **car_insurance_claims.sav.** Un conjunto de datos presentados y analizados en otro lugar estudia las reclamaciones por daños en vehículos. La cantidad de reclamaciones media se puede modelar como si tuviera una distribución Gamma, mediante una función de enlace inversa para relacionar la media de la variable dependiente con una combinación lineal de la edad del asegurado, el tipo de vehículo y la antigüedad del vehículo. El número de reclamaciones presentadas se puede utilizar como una ponderación de escalamiento.

- **car_sales.sav.** Este archivo de datos contiene estimaciones de ventas, precios de lista y especificaciones físicas hipotéticas de varias marcas y modelos de vehículos. Los precios de lista y las especificaciones físicas se han obtenido de *edmunds.com* y de sitios de fabricantes.
- **car_sales_uprepared.sav.**Ésta es una versión modificada de *car_sales.sav* que no incluye ninguna versión transformada de los campos.
- **carpet.sav** En un ejemplo muy conocido , una compañía interesada en sacar al mercado un nuevo limpiador de alfombras desea examinar la influencia de cinco factores sobre la preferencia del consumidor: diseño del producto, marca comercial, precio, sello de *buen producto para el hogar* y garantía de devolución del importe. Hay tres niveles de factores para el diseño del producto, cada uno con una diferente colocación del cepillo del aplicador; tres nombres comerciales (*K2R*, *Glory* y *Bissell*); tres niveles de precios; y dos niveles (no o sí) para los dos últimos factores. Diez consumidores clasificaron 22 perfiles definidos por estos factores. La variable *Preferencia* contiene el rango de las clasificaciones medias de cada perfil. Las clasificaciones inferiores corresponden a preferencias elevadas. Esta variable refleja una medida global de la preferencia de cada perfil.
- **carpet_prefs.sav** Este archivo de datos se basa en el mismo ejemplo que el descrito para *carpet.sav*, pero contiene las clasificaciones reales recogidas de cada uno de los 10 consumidores. Se pidió a los consumidores que clasificaran los 22 perfiles de los productos empezando por el menos preferido. Las variables desde *PREF1* hasta *PREF22* contienen los ID de los perfiles asociados, como se definen en *carpet_plan.sav*.
- **catalog.sav**Este archivo de datos contiene cifras de ventas mensuales hipotéticas de tres productos vendidos por una compañía de venta por catálogo. También se incluyen datos para cinco variables predictoras posibles.
- **catalog_seasfac.sav**Este archivo de datos es igual que *catalog.sav*, con la excepción de que incluye un conjunto de factores estacionales calculados a partir del procedimiento Descomposición estacional junto con las variables de fecha que lo acompañan.
- **cellular.sav.** Archivo de datos hipotéticos sobre las iniciativas de una compañía de telefonía móvil para reducir el abandono de clientes. Las puntuaciones de propensión al abandono de clientes se aplican a las cuentas, oscilando de 0 a 100. Las cuentas con una puntuación de 50 o superior pueden estar buscando otros proveedores.
- **ceramics.sav.**Archivo de datos hipotéticos sobre las iniciativas de un fabricante para determinar si una nueva aleación de calidad tiene una mayor resistencia al calor que una aleación estándar. Cada caso representa una prueba independiente de una de las aleaciones; la temperatura a la que registró el fallo del rodamiento.
- **cereal.sav.** Archivo de datos hipotéticos sobre una encuesta realizada a 880 personas sobre sus preferencias en el desayuno, teniendo también en cuenta su edad, sexo, estado civil y si tienen un estilo de vida activo o no (en función de si practican ejercicio al menos dos veces a la semana). Cada caso representa un encuestado diferente.
- **clothing_defects.sav.** Archivo de datos hipotéticos sobre el proceso de control de calidad en una fábrica de prendas. Los inspectores toman una muestra de prendas de cada lote producido en la fábrica, y cuentan el número de prendas que no son aceptables.
- **coffee.sav.** Este archivo de datos pertenece a las imágenes percibidas de seis marcas de café helado . Para cada uno de los 23 atributos de imagen de café helado, los encuestados seleccionaron todas las marcas que quedaban descritas por el atributo. Las seis marcas se denotan AA, BB, CC, DD, EE y FF para mantener la confidencialidad.

- **contacts.sav.**Archivo de datos hipotéticos sobre las listas de contactos de un grupo de representantes de ventas de ordenadores de empresa. Cada uno de los contactos está categorizado por el departamento de la compañía en el que trabaja y su categoría en la compañía. Además, también se registran los importes de la última venta realizada, el tiempo transcurrido desde la última venta y el tamaño de la compañía del contacto.
- **creditpromo.sav.** Archivo de datos hipotéticos sobre las iniciativas de unos almacenes para evaluar la eficacia de una promoción de tarjetas de crédito reciente. Para este fin, se seleccionaron aleatoriamente 500 titulares. La mitad recibieron un anuncio promocionando una tasa de interés reducida sobre las ventas realizadas en los siguientes tres meses. La otra mitad recibió un anuncio estacional estándar.
- **customer_dbase.sav.** Archivo de datos hipotéticos sobre las iniciativas de una compañía para usar la información de su almacén de datos para realizar ofertas especiales a los clientes con más probabilidades de responder. Se seleccionó un subconjunto de la base de clientes aleatoriamente a quienes se ofrecieron las ofertas especiales y sus respuestas se registraron.
- **customer_information.sav.** Archivo de datos hipotéticos que contiene la información de correo del cliente, como el nombre y la dirección.
- **customer_subset.sav.** Un subconjunto de 80 casos de *customer_dbase.sav*.
- **customers_model.sav.** Este archivo contiene datos hipotéticos sobre los individuos a los que va dirigida una campaña de marketing. Estos datos incluyen información demográfica, un resumen del historial de compras y si cada individuo respondió a la campaña. Cada caso representa un individuo diferente.
- **customers_new.sav.** Este archivo contiene datos hipotéticos sobre los individuos que son candidatos potenciales para una campaña de marketing. Estos datos incluyen información demográfica y un resumen del historial de compras de cada individuo. Cada caso representa un individuo diferente.
- **debate.sav.** Archivos de datos hipotéticos sobre las respuestas emparejadas de una encuesta realizada a los asistentes a un debate político antes y después del debate. Cada caso corresponde a un encuestado diferente.
- **debate_aggregate.sav.** Archivo de datos hipotéticos que agrega las respuestas de *debate.sav*. Cada caso corresponde a una clasificación cruzada de preferencias antes y después del debate.
- **demo.sav.** Archivos de datos hipotéticos sobre una base de datos de clientes adquirida con el fin de enviar por correo ofertas mensuales. Se registra si el cliente respondió a la oferta, junto con información demográfica diversa.
- **demo_cs_1.sav.**Archivo de datos hipotéticos sobre el primer paso de las iniciativas de una compañía para recopilar una base de datos de información de encuestas. Cada caso corresponde a una ciudad diferente, y se registra la identificación de la ciudad, la región, la provincia y el distrito.
- **demo_cs_2.sav.**Archivo de datos hipotéticos sobre el segundo paso de las iniciativas de una compañía para recopilar una base de datos de información de encuestas. Cada caso corresponde a una unidad familiar diferente de las ciudades seleccionadas en el primer paso, y se registra la identificación de la unidad, la subdivisión, la ciudad, el distrito, la provincia y la región. También se incluye la información de muestreo de las primeras dos etapas del diseño.

- **demo_cs.sav.**Archivo de datos hipotéticos que contiene información de encuestas recopilada mediante un diseño de muestreo complejo. Cada caso corresponde a una unidad familiar distinta, y se recopila información demográfica y de muestreo diversa.
- **dmdata.sav.** Éste es un archivo de datos hipotéticos que contiene información demográfica y de compras para una empresa de marketing directo. *dmdata2.sav* contiene información para un subconjunto de contactos que recibió un envío de prueba, y *dmdata3.sav* contiene información sobre el resto de contactos que no recibieron el envío de prueba.
- **dietstudy.sav.**Este archivo de datos hipotéticos contiene los resultados de un estudio sobre la "dieta Stillman" . Cada caso corresponde a un sujeto distinto y registra sus pesos antes y después de la dieta en libras y niveles de triglicéridos en mg/100 ml.
- **dvdplayer.sav.** Archivo de datos hipotéticos sobre el desarrollo de un nuevo reproductor de DVD. El equipo de marketing ha recopilado datos de grupo de enfoque mediante un prototipo. Cada caso corresponde a un usuario encuestado diferente y registra información demográfica sobre los encuestados y sus respuestas a preguntas acerca del prototipo.
- **german_credit.sav.**Este archivo de datos se toma del conjunto de datos "German credit" de las Repository of Machine Learning Databases de la Universidad de California, Irvine.
- **grocery_1month.sav.** Este archivo de datos hipotéticos es el archivo de datos *grocery_coupons.sav* con las compras semanales "acumuladas" para que cada caso corresponda a un cliente diferente. Algunas de las variables que cambiaban semanalmente desaparecen de los resultados, y la cantidad gastada registrada se convierte ahora en la suma de las cantidades gastadas durante las cuatro semanas del estudio.
- **grocery_coupons.sav.** Archivo de datos hipotéticos que contiene datos de encuestas recopilados por una cadena de tiendas de alimentación interesada en los hábitos de compra de sus clientes. Se sigue a cada cliente durante cuatro semanas, y cada caso corresponde a un cliente-semana distinto y registra información sobre dónde y cómo compran los clientes, incluida la cantidad que invierten en comestibles durante esa semana.
- **guttman.sav.**Bell presentó una tabla para ilustrar posibles grupos sociales. Guttman utilizó parte de esta tabla, en la que se cruzaron cinco variables que describían elementos como la interacción social, sentimientos de pertenencia a un grupo, proximidad física de los miembros y grado de formalización de la relación con siete grupos sociales teóricos, incluidos multitudes (por ejemplo, las personas que acuden a un partido de fútbol), espectadores (por ejemplo, las personas que acuden a un teatro o de una conferencia), públicos (por ejemplo, los lectores de periódicos o los espectadores de televisión), muchedumbres (como una multitud pero con una interacción mucho más intensa), grupos primarios (íntimos), grupos secundarios (voluntarios) y la comunidad moderna (confederación débil que resulta de la proximidad cercana física y de la necesidad de servicios especializados).
- **health_funding.sav.** Archivo de datos hipotéticos que contiene datos sobre inversión en sanidad (cantidad por 100 personas), tasas de enfermedad (índice por 10.000 personas) y visitas a centros de salud (índice por 10.000 personas). Cada caso representa una ciudad diferente.
- **hivassay.sav.** Archivo de datos hipotéticos sobre las iniciativas de un laboratorio farmacéutico para desarrollar un ensayo rápido para detectar la infección por VIH. Los resultados del ensayo son ocho tonos de rojo con diferentes intensidades, donde los tonos más oscuros indican una mayor probabilidad de infección. Se llevó a cabo una prueba de laboratorio de 2.000 muestras de sangre, de las cuales una mitad estaba infectada con el VIH y la otra estaba limpia.

- **hourlywagedata.sav.** Archivo de datos hipotéticos sobre los salarios por horas de enfermeras de puestos de oficina y hospitales y con niveles distintos de experiencia.
- **insurance_claims.sav.** Éste es un archivo de datos hipotéticos sobre una compañía de seguros que desee generar un modelo para etiquetar las reclamaciones sospechosas y potencialmente fraudulentas. Cada caso representa una reclamación diferente.
- **insure.sav.** Archivo de datos hipotéticos sobre una compañía de seguros que estudia los factores de riesgo que indican si un cliente tendrá que hacer una reclamación a lo largo de un contrato de seguro de vida de 10 años. Cada caso del archivo de datos representa un par de contratos (de los que uno registró una reclamación y el otro no), agrupados por edad y sexo.
- **judges.sav.** Archivo de datos hipotéticos sobre las puntuaciones concedidas por jueces cualificados (y un aficionado) a 300 actuaciones gimnásticas. Cada fila representa una actuación diferente; los jueces vieron las mismas actuaciones.
- **kinship_dat.sav.** Rosenberg y Kim comenzaron a analizar 15 términos de parentesco [tía, hermano, primos, hija, padre, nieta, abuelo, abuela, nieto, madre, sobrino, sobrina, hermana, hijo, tío]. Le pidieron a cuatro grupos de estudiantes universitarios (dos masculinos y dos femeninos) que ordenaran estos grupos según las similitudes. A dos grupos (uno masculino y otro femenino) se les pidió que realizaran la ordenación dos veces, pero que la segunda ordenación la hicieran según criterios distintos a los de la primera. Así, se obtuvo un total de seis "fuentes". Cada fuente se corresponde con una matriz de proximidades de $15 \times 15$ cuyas casillas son iguales al número de personas de una fuente menos el número de veces que se particionaron los objetos en esa fuente.
- **kinship_ini.sav.** Este archivo de datos contiene una configuración inicial para una solución tridimensional de *kinship_dat.sav*.
- **kinship_var.sav.** Este archivo de datos contiene variables independientes *sexo*, *gener*(ación), y *grado* (de separación) que se pueden usar para interpretar las dimensiones de una solución para *kinship_dat.sav*. Concretamente, se pueden usar para restringir el espacio de la solución a una combinación lineal de estas variables.
- **marketvalues.sav.** Archivo de datos sobre las ventas de casas en una nueva urbanización de Algonquin, Ill., durante los años 1999 y 2000. Los datos de estas ventas son públicos.
- **nhis2000_subset.sav.** La National Health Interview Survey (NHIS, encuesta del Centro Nacional de Estadísticas de Salud de EE.UU.) es una encuesta detallada realizada entre la población civil de Estados Unidos. Las encuestas se realizaron en persona a una muestra representativa de las unidades familiares del país. Se recogió tanto la información demográfica como las observaciones acerca del estado y los hábitos de salud de los integrantes de cada unidad familiar. Este archivo de datos contiene un subconjunto de información de la encuesta de 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Archivo de datos y documentación de uso público. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Fecha de acceso: 2003.
- **ozono.sav.** Los datos incluyen 330 observaciones de seis variables meteorológicas para pronosticar la concentración de ozono a partir del resto de variables. Los investigadores anteriores, han encontrado que no hay linealidad entre estas variables, lo que dificulta los métodos de regresión típica.

- **pain_medication.sav.** Este archivo de datos hipotéticos contiene los resultados de una prueba clínica sobre medicación antiinflamatoria para tratar el dolor artrítico crónico. Resulta de particular interés el tiempo que tarda el fármaco en hacer efecto y cómo se compara con una medicación existente.
- **patient_los.sav.** Este archivo de datos hipotéticos contiene los registros de tratamiento de pacientes que fueron admitidos en el hospital ante la posibilidad de sufrir un infarto de miocardio (IM o "ataque al corazón"). Cada caso corresponde a un paciente distinto y registra diversas variables relacionadas con su estancia hospitalaria.
- **patlos_sample.sav.** Este archivo de datos hipotéticos contiene los registros de tratamiento de una muestra de pacientes que recibieron trombolíticos durante el tratamiento del infarto de miocardio (IM o "ataque al corazón"). Cada caso corresponde a un paciente distinto y registra diversas variables relacionadas con su estancia hospitalaria.
- **polishing.sav.** Archivo de datos "Nambeware Polishing Times" (Tiempo de pulido de metal) de la biblioteca de datos e historiales. Contiene datos sobre las iniciativas de un fabricante de cuberterías de metal (Nambe Mills, Santa Fe, N. M.) para planificar su programa de producción. Cada caso representa un artículo distinto de la línea de productos. Se registra el diámetro, el tiempo de pulido, el precio y el tipo de producto de cada artículo.
- **poll_cs.sav.** Archivo de datos hipotéticos sobre las iniciativas de los encuestadores para determinar el nivel de apoyo público a una ley antes de una asamblea legislativa. Los casos corresponden a votantes registrados. Cada caso registra el condado, la población y el vecindario en el que vive el votante.
- **poll_cs_sample.sav.** Este archivo de datos hipotéticos contiene una muestra de los votantes enumerados en *poll_cs.sav*. La muestra se tomó según el diseño especificado en el archivo de plan *poll.csplan* y este archivo de datos registra las probabilidades de inclusión y las ponderaciones muestrales. Sin embargo, tenga en cuenta que debido a que el plan muestral hace uso de un método de probabilidad proporcional al tamaño (PPS), también existe un archivo que contiene las probabilidades de selección conjunta (*poll_jointprob.sav*). Las variables adicionales que corresponden a los datos demográficos de los votantes y sus opiniones sobre la propuesta de ley se recopilaron y añadieron al archivo de datos después de tomar la muestra.
- **property_assess.sav.** Archivo de datos hipotéticos sobre las iniciativas de un asesor del condado para mantener actualizada la evaluación de los valores de las propiedades utilizando recursos limitados. Los casos corresponden a las propiedades vendidas en el condado el año anterior. Cada caso del archivo de datos registra la población en que se encuentra la propiedad, el último asesor que visitó la propiedad, el tiempo transcurrido desde la última evaluación, la valoración realizada en ese momento y el valor de venta de la propiedad.
- **property_assess_cs.sav.** Archivo de datos hipotéticos sobre las iniciativas de un asesor de un estado para mantener actualizada la evaluación de los valores de las propiedades utilizando recursos limitados. Los casos corresponden a propiedades del estado. Cada caso del archivo de datos registra el condado, la población y el vecindario en el que se encuentra la propiedad, el tiempo transcurrido desde la última evaluación y la valoración realizada en ese momento.
- **property_assess_cs_sample.sav**Este archivo de datos hipotéticos contiene una muestra de las propiedades recogidas en *property_assess_cs.sav*. La muestra se tomó en función del diseño especificado en el archivo de plan *property_assess.csplan*, y este archivo de datos registra las probabilidades de inclusión y las ponderaciones muestrales. La variable adicional *Valor actual* se recopiló y añadió al archivo de datos después de tomar la muestra.

- **recidivism.sav.** Archivo de datos hipotéticos sobre las iniciativas de una agencia de orden público para comprender los índices de reincidencia en su área de jurisdicción. Cada caso corresponde a un infractor anterior y registra su información demográfica, algunos detalles de su primer delito y, a continuación, el tiempo transcurrido desde su segundo arresto, si ocurrió en los dos años posteriores al primer arresto.
- **recidivism_cs_sample.sav.** Archivo de datos hipotéticos sobre las iniciativas de una agencia de orden público para comprender los índices de reincidencia en su área de jurisdicción. Cada caso corresponde a un delincuente anterior, puesto en libertad tras su primer arresto durante el mes de junio de 2003 y registra su información demográfica, algunos detalles de su primer delito y los datos de su segundo arresto, si se produjo antes de finales de junio de 2006. Los delincuentes se seleccionaron de una muestra de departamentos según el plan de muestreo especificado en *recidivism_cs.csplan*. Como este plan utiliza un método de probabilidad proporcional al tamaño (PPS), también existe un archivo que contiene las probabilidades de selección conjunta (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** Archivo de datos hipotéticos que contiene datos de transacciones de compra, incluida la fecha de compra, los artículos adquiridos y el importe de cada transacción.
- **salesperformance.sav.** Archivo de datos hipotéticos sobre la evaluación de dos nuevos cursos de formación de ventas. Sesenta empleados, divididos en tres grupos, reciben formación estándar. Además, el grupo 2 recibe formación técnica; el grupo 3, un tutorial práctico. Cada empleado se sometió a un examen al final del curso de formación y se registró su puntuación. Cada caso del archivo de datos representa a un alumno distinto y registra el grupo al que fue asignado y la puntuación que obtuvo en el examen.
- **satisf.sav.** Archivo de datos hipotéticos sobre una encuesta de satisfacción llevada a cabo por una empresa minorista en cuatro tiendas. Se encuestó a 582 clientes en total y cada caso representa las respuestas de un único cliente.
- **screws.sav** Este archivo de datos contiene información acerca de las características de tornillos, pernos, clavos y tacos .
- **shampoo_ph.sav.**Archivo de datos hipotéticos sobre el control de calidad en una fábrica de productos para el cabello. Se midieron seis lotes de resultados distintos en intervalos regulares y se registró su pH. El intervalo objetivo es de 4,5 a 5,5.
- **ships.sav.** Un conjunto de datos presentados y analizados en otro lugar sobre los daños en los cargueros producidos por las olas. Los recuentos de incidentes se pueden modelar como si ocurrieran con una tasa de Poisson dado el tipo de barco, el período de construcción y el período de servicio. Los meses de servicio agregados para cada casilla de la tabla formados por la clasificación cruzada de factores proporcionan valores para la exposición al riesgo.
- **site.sav.**Archivo de datos hipotéticos sobre las iniciativas de una compañía para seleccionar sitios nuevos para sus negocios en expansión. Se ha contratado a dos consultores para evaluar los sitios de forma independiente, quienes, además de un informe completo, han resumido cada sitio como una posibilidad “buena”, “media” o “baja”.
- **smokers.sav.**Este archivo de datos es un resumen de la encuesta sobre toxicomanía 1998 National Household Survey of Drug Abuse y es una muestra de probabilidad de unidades familiares americanas. (*http://dx.doi.org/10.3886/ICPSR02934*) Así, el primer paso de un análisis de este archivo de datos debe ser ponderar los datos para reflejar las tendencias de población.

- **stroke_clean.sav.** Este archivo de datos hipotéticos contiene el estado de una base de datos médica después de haberla limpiado mediante los procedimientos de la opción Preparación de datos.
- **stroke_invalid.sav.** Este archivo de datos hipotéticos contiene el estado inicial de una base de datos médica que incluye contiene varios errores de entrada de datos.
- **stroke_survival.** Este archivo de datos hipotéticos registra los tiempos de supervivencia de los pacientes que finalizan un programa de rehabilitación tras un ataque isquémico. Tras el ataque, la ocurrencia de infarto de miocardio, ataque isquémico o ataque hemorrágico se anotan junto con el momento en el que se produce el evento registrado. La muestra está truncada a la izquierda ya que únicamente incluye a los pacientes que han sobrevivido al final del programa de rehabilitación administrado tras el ataque.
- **stroke_valid.sav.** Este archivo de datos hipotéticos contiene el estado de una base de datos médica después de haber comprobado los valores mediante el procedimiento Validar datos. Sigue conteniendo casos potencialmente anómalos.
- **survey_sample.sav.** Este archivo de datos contiene datos de encuestas, incluyendo datos demográficos y diferentes medidas de actitud. Se basa en un subconjunto de variables de NORC General Social Survey de 1998, aunque algunos valores de datos se han modificado y que existen variables ficticias adicionales se han añadido para demostraciones.
- **telco.sav.** Archivo de datos hipotéticos sobre las iniciativas de una compañía de telecomunicaciones para reducir el abandono de clientes en su base de clientes. Cada caso corresponde a un cliente distinto y registra diversa información demográfica y de uso del servicio.
- **telco_extra.sav.** Este archivo de datos es similar al archivo de datos *telco.sav*, pero las variables de meses con servicio y gasto de clientes transformadas logarítmicamente se han eliminado y sustituido por variables de gasto del cliente transformadas logarítmicamente tipificadas.
- **telco_missing.sav.** Este archivo de datos es un subconjunto del archivo de datos *telco.sav*, pero algunos valores de datos demográficos se han sustituido con valores perdidos.
- **testmarket.sav.** Archivo de datos hipotéticos sobre los planes de una cadena de comida rápida para añadir un nuevo artículo a su menú. Hay tres campañas posibles para promocionar el nuevo producto, por lo que el artículo se presenta en ubicaciones de varios mercados seleccionados aleatoriamente. Se utiliza una promoción diferente en cada ubicación y se registran las ventas semanales del nuevo artículo durante las primeras cuatro semanas. Cada caso corresponde a una ubicación semanal diferente.
- **testmarket_1month.sav.** Este archivo de datos hipotéticos es el archivo de datos *testmarket.sav* con las ventas semanales "acumuladas" para que cada caso corresponda a una ubicación diferente. Como resultado, algunas de las variables que cambiaban semanalmente desaparecen y las ventas registradas se convierten en la suma de las ventas realizadas durante las cuatro semanas del estudio.
- **tree_car.sav.** Archivo de datos hipotéticos que contiene datos demográficos y de precios de compra de vehículos.
- **tree_credit.sav** Archivo de datos hipotéticos que contiene datos demográficos y de historial de créditos bancarios.
- **tree_missing_data.sav** Archivo de datos hipotéticos que contiene datos demográficos y de historial de créditos bancarios con un elevado número de valores perdidos.

- **tree_score_car.sav.** Archivo de datos hipotéticos que contiene datos demográficos y de precios de compra de vehículos.
- **tree_textdata.sav.** Archivo de datos sencillos con dos variables diseñadas principalmente para mostrar el estado por defecto de las variables antes de realizar la asignación de nivel de medida y etiquetas de valor.
- **tv-survey.sav.** Archivo de datos hipotéticos sobre una encuesta dirigida por un estudio de TV que está considerando la posibilidad de ampliar la emisión de un programa de éxito. Se preguntó a 906 encuestados si verían el programa en distintas condiciones. Cada fila representa un encuestado diferente; cada columna es una condición diferente.
- **ulcer_recurrence.sav.** Este archivo contiene información parcial de un estudio diseñado para comparar la eficacia de dos tratamientos para prevenir la reaparición de úlceras. Constituye un buen ejemplo de datos censurados por intervalos y se ha presentado y analizado en otro lugar .
- **ulcer_recurrence_recoded.sav.** Este archivo reorganiza la información de *ulcer_recurrence.sav* para permitir modelar la probabilidad de eventos de cada intervalo del estudio en lugar de sólo la probabilidad de eventos al final del estudio. Se ha presentado y analizado en otro lugar .
- **verd1985.sav.** Archivo de datos sobre una encuesta . Se han registrado las respuestas de 15 sujetos a 8 variables. Se han dividido las variables de interés en tres grupos. El conjunto 1 incluye *edad* y *ecivil*, el conjunto 2 incluye *mascota* y *noticia*, mientras que el conjunto 3 incluye *música* y *vivir*. Se escala *mascota* como nominal múltiple y *edad* como ordinal; el resto de variables se escalan como nominal simple.
- **virus.sav.**Archivo de datos hipotéticos sobre las iniciativas de un proveedor de servicios de Internet (ISP) para determinar los efectos de un virus en sus redes. Se ha realizado un seguimiento (aproximado) del porcentaje de tráfico de correos electrónicos infectados en sus redes a lo largo del tiempo, desde el momento en que se descubre hasta que la amenaza se contiene.
- **wheeze_steubenville.sav.** Subconjunto de un estudio longitudinal de los efectos sobre la salud de la polución del aire en los niños . Los datos contienen medidas binarias repetidas del estado de las sibilancias en niños de Steubenville, Ohio, con edades de 7, 8, 9 y 10 años, junto con un registro fijo de si la madre era fumadora durante el primer año del estudio.
- **workprog.sav.** Archivo de datos hipotéticos sobre un programa de obras del gobierno que intenta colocar a personas desfavorecidas en mejores trabajos. Se siguió una muestra de participantes potenciales del programa, algunos de los cuales se seleccionaron aleatoriamente para entrar en el programa, mientras que otros no siguieron esta selección aleatoria. Cada caso representa un participante del programa diferente.

Apéndice
B

# Notices



Patent No. 7,023,453

**The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:** SPSS INC., AN IBM COMPANY, PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. SPSS Inc. may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-SPSS and non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this SPSS Inc. product and use of those Web sites is at your own risk.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Information concerning non-SPSS products was obtained from the suppliers of those products, their published announcements or other publicly available sources. SPSS has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-SPSS products. Questions on the capabilities of non-SPSS products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing,

using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

### *Trademarks*

IBM, the IBM logo, and ibm.com are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at *http://www.ibm.com/legal/copytrade.shmtl*.

SPSS is a trademark of SPSS Inc., an IBM Company, registered in many jurisdictions worldwide.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Other product and service names might be trademarks of IBM, SPSS, or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

# Índice