IBM SPSS Data Preparation 19

Note: Before using this information and the product it supports, read the general information under Notices el p. 148.

Prefacio

IBM® SPSS® Statistics es un sistema global para el análisis de datos. El módulo adicional opcional Preparación de los datos proporciona las técnicas de análisis adicionales que se describen en este manual. El módulo adicional Preparación de los datos 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

Capítulo

1

Introducción a la preparación de datos

A medida que la potencia de los sistemas informáticos se incrementa, la necesidad de información crece proporcionalmente, llevando a un crecimiento cada vez mayor de la recopilación de datos: más casos, más variables y más errores en la entrada de datos. Estos errores son la pesadilla de las predicciones del modelo predictivo, que son el objetivo final del almacenamiento de datos, por lo que existe la necesidad de mantener los datos "limpios". Sin embargo, la cantidad de datos almacenados ha superado de tal forma a la capacidad de comprobar los casos manualmente que resulta vital implementar procesos automatizados para validar los datos.

El módulo adicional Preparación de datos permite identificar casos, variables y valores de datos atípicos y no válidos en el conjunto de datos activo, así como preparar los datos para el modelado.

Uso de los procedimientos de preparación de datos

El uso de los procedimientos de preparación de datos depende de las necesidades específicas. Una ruta típica tras la carga de datos es:

- **Preparación de metadatos.** Revisar las variables del archivo de datos y determinar los valores válidos, las etiquetas y los niveles de medida. Identificar las combinaciones de valores de variable que son imposibles pero suelen estar mal codificadas. Definir las reglas de validación en función de esta información. Esta tarea puede resultar pesada, pero el esfuerzo compensa si debe validar archivos de datos que tengan atributos similares con regularidad.
- **Validación de datos.** Ejecutar comprobaciones básicas y comprobaciones de reglas de validación definidas para identificar casos no válidos, variables y valores de datos. Cuando se encuentran datos no válidos, investigar y corregir la causa. Puede que sea necesario realizar otro paso con la preparación de metadatos.
- **Preparación de modelos.** Utilice la preparación automática de datos para obtener transformaciones de los campos originales que mejorarán la generación de modelos. Identifique valores atípicos estadísticos potenciales que puedan provocar problemas para muchos modelos predictivos. Algunos valores atípicos son el resultado de valores de variable no válidos que no se han identificado. Puede que sea necesario realizar otro paso con la preparación de metadatos.

Una vez que el archivo de datos está "limpio", se pueden generar modelos de otros módulos adicionales.

Capítulo
2

Reglas de validación

Las reglas se utilizan para determinar si un caso es válido. Existen dos tipos de reglas de validación:

- **Reglas de variable única.** Las reglas de variable única constan de un conjunto fijo de comprobaciones que se aplican a una única variable, como las comprobaciones de los valores que están fuera de rango. En el caso de las reglas de variable única, los valores válidos pueden expresarse como un rango de valores o una lista de valores aceptables.
- **Reglas inter-variables.** Las reglas inter-variables son reglas definidas por el usuario que se pueden aplicar a una única variable o a una combinación de variables. Las reglas inter-variables están definidas por una expresión lógica que marca valores no válidos.

Las reglas de validación se guardan en el diccionario de datos del archivo de datos. Esto permite especificar una regla una vez y volver a utilizarla más adelante.

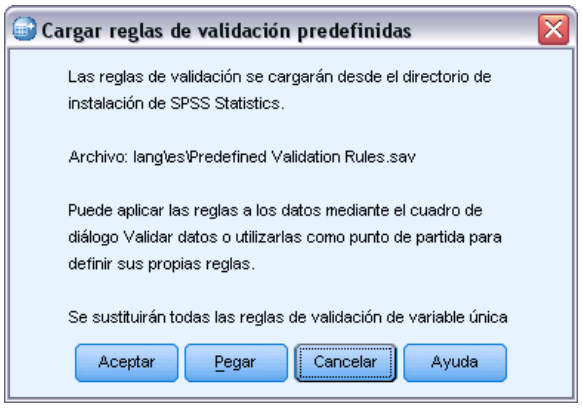
Cargar reglas de validación predefinidas

Puede obtener de manera rápida un conjunto de reglas de validación listas para usar cargando reglas predefinidas a partir de un archivo de datos externo que se incluye en la instalación.

Para cargar reglas de validación predefinidas

▶ En los menús, seleccione:
Datos > Validación > Cargar reglas predefinidas...

Figura 2-1
Cargar reglas de validación predefinidas



Tenga en cuenta que este proceso eliminará cualquier regla de variable única del conjunto de datos activo.

Si lo desea, puede utilizar el Asistente para la copia de propiedades de datos para cargar reglas desde cualquier archivo de datos.

Definir reglas de validación

El cuadro de diálogo Definir reglas de validación permite crear y ver reglas de validación inter-variables y de variable única.

Para crear y ver reglas de validación

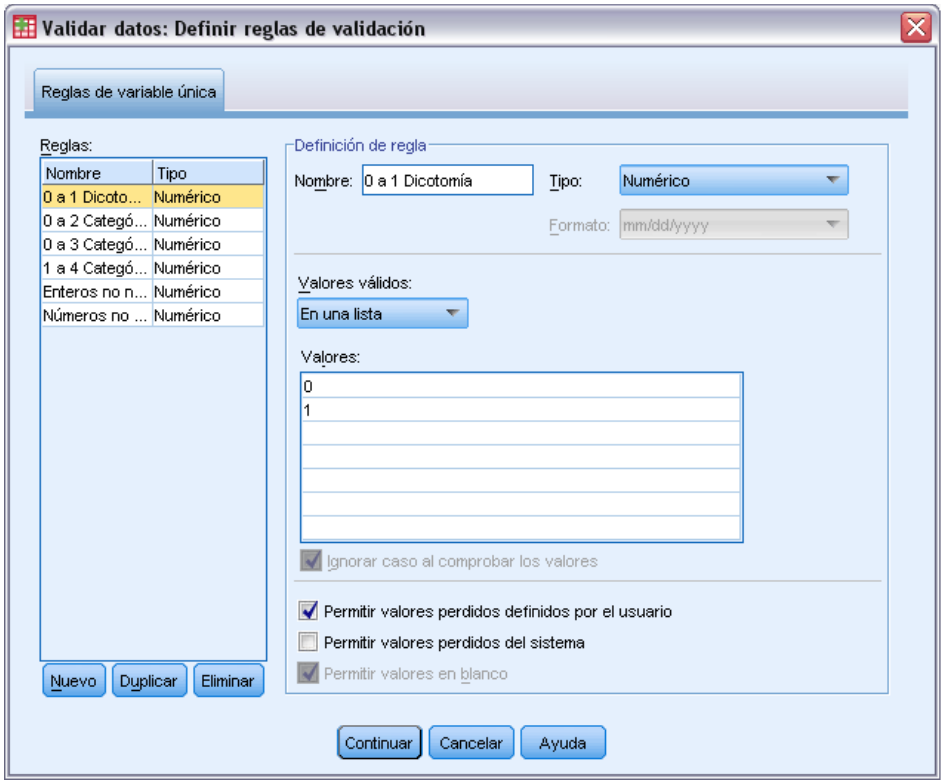
- En los menús, seleccione:
  Datos > Validación > Definir reglas...

  El cuadro de diálogo contiene reglas de validación inter-variables y de variable única que se leen desde el diccionario de datos. Cuando no hay reglas, se crea automáticamente una regla de marcador de posición nueva que se puede modificar para ajustarse a sus necesidades.

- Seleccione las reglas individuales en las pestañas Reglas de variable única y Reglas inter-variables para ver y modificar sus propiedades.

Definir reglas de variable única

Figura 2-2
Cuadro de diálogo Definir reglas de validación, pestaña Reglas de variable única



La pestaña Reglas de variable única permiten crear, ver y modificar reglas de validación de variable única.

Reglas. La lista las muestra reglas de validación de variable única por nombre y el tipo de variable a la que se puede aplicar la regla. Cuando el cuadro de diálogo está abierto, muestra las reglas definidas en el diccionario de datos o, si no hay ninguna regla definida en ese momento, se muestra una regla de marcador de posición denominada "ReglaVarÚnica 1". Los siguientes botones aparecen debajo de la lista Reglas:

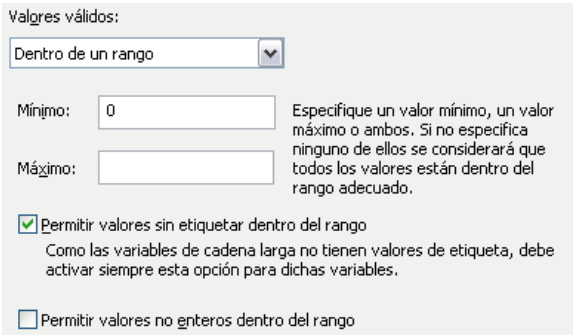
- **Nuevo.** Añade una nueva entrada en la parte inferior de la lista Reglas. La regla se selecciona y se le asigna el nombre "ReglaVarÚnica *n*", donde *n* es un número entero de forma que el nombre de la nueva regla es único en las reglas de variable única y las reglas inter-variables.
- **Duplicar.** Añade una copia de la regla seleccionada en la parte inferior de la lista Reglas. El nombre de la regla se ajusta de forma que sea única entre las reglas de variable única y las reglas inter-variables. Por ejemplo, si duplica "ReglaVarÚnica 1", el nombre de la primera regla duplicada sería "Copia de ReglaVarÚnica 1", la segunda sería "Copia (2) de ReglaVarÚnica 1", y así sucesivamente.
- **Eliminar.** Elimina la regla seleccionada.

Definición de regla. Estos controles permiten ver y establecer propiedades para una regla seleccionada.

- **Nombre.** El nombre de la regla debe ser único para las reglas de variable única y las reglas inter-variables.
- **Tipo.** Éste es el tipo de variable a la que se puede aplicar la regla. Seleccione desde Numérico, Cadena y Fecha.
- **Formato.** Permite seleccionar el formato de fecha para las reglas que se puedan aplicar a las variables de fecha.
- **Valores válidos.** Puede especificar los valores válidos como un rango o como una lista de valores.

Los controles de Definición de rango permiten especificar un rango válido. Los valores que se encuentran fuera del rango aparecen marcados como no válidos.

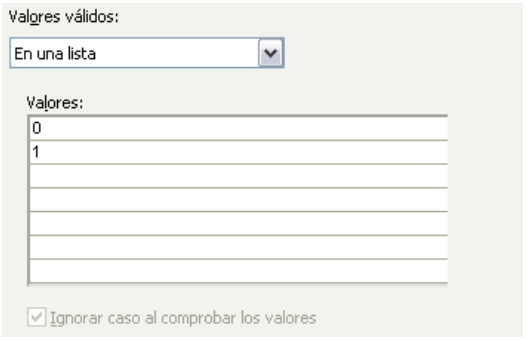
Figura 2-3
Reglas de variable única: Definición de rango



Para especificar un rango, escriba el valor mínimo, el valor máximo o ambos. Los controles de la casilla de verificación permiten marcar valores sin etiqueta y no enteros que se encuentran dentro del rango.

Los controles de definición de lista permiten definir una lista de valores válidos. Los valores que no están incluidos en la lista aparecen marcados como no válidos.

Figura 2-4
Reglas de variable única: Definición de lista

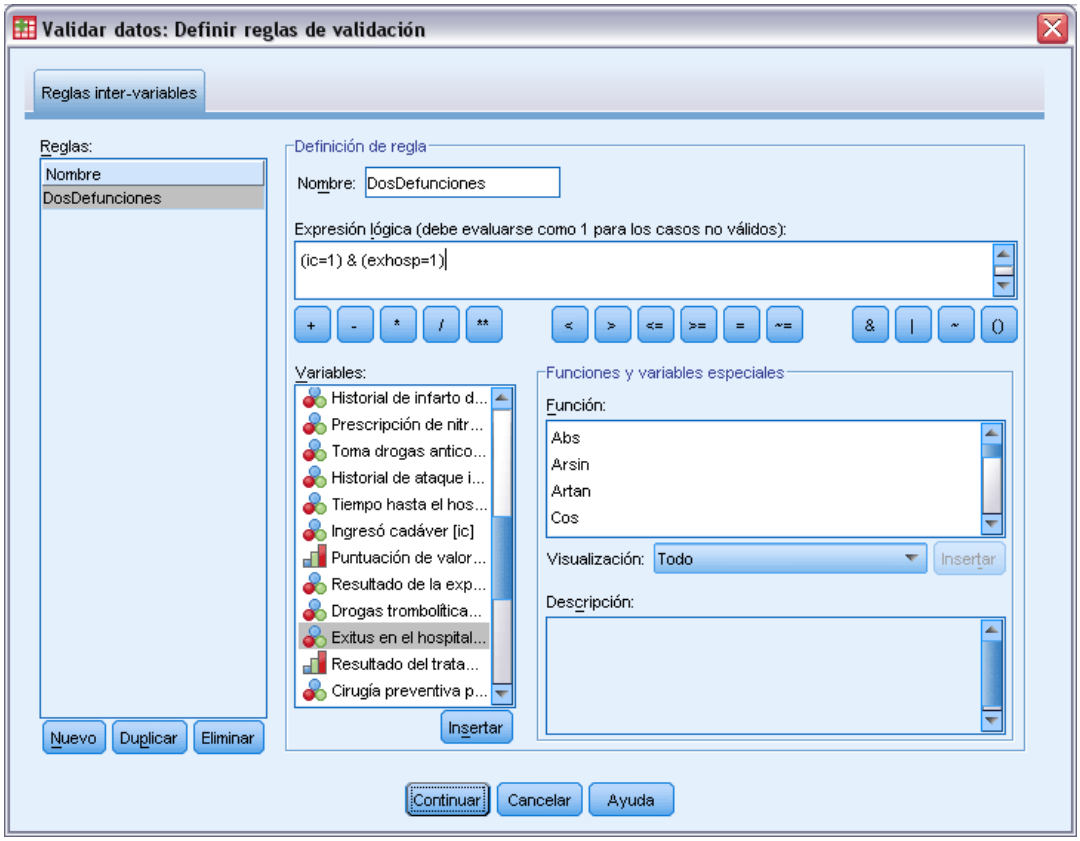


Introduce valores de lista en la cuadrícula. La casilla de verificación determina si el caso tiene importancia cuando los valores de datos de cadena se contrastan con la lista de valores aceptables.

- **Permitir valores perdidos definidos por el usuario.** Controla si los valores perdidos definidos por el usuario están marcados como no válidos.
- **Permitir valores perdidos del sistema.** Controla si los valores perdidos del sistema están marcados como no válidos. Esto no se aplica a tipos de reglas de cadena.
- **Permitir valores en blanco.** Controla si los valores en blanco de cadena (es decir, completamente vacíos) están marcados como no válidos. Esto no se aplica a los tipos de reglas que no son de cadena.

Definir reglas inter-variables

Figura 2-5
Cuadro de diálogo Definir reglas de validación, pestaña Reglas inter-variables



La pestaña Reglas inter-variables permite crear, ver y modificar reglas de validación inter-variables.

Reglas. La lista muestra reglas de validación inter-variables por nombre. Cuando se abre el cuadro de diálogo, muestra una regla de marcador de posición denominada “ReglaInterVar 1”. Los siguientes botones aparecen debajo de la lista Reglas:

- **Nuevo.** Añade una nueva entrada en la parte inferior de la lista Reglas. La regla se selecciona y se le asigna el nombre “ReglaInterVar *n*”, donde *n* es un número entero, de forma que el nombre de la nueva regla es único en las reglas de variable única y la regla inter-variables.

- **Duplicar.** Añade una copia de la regla seleccionada en la parte inferior de la lista Reglas. El nombre de la regla se ajusta de forma que sea única entre las reglas de variable única y las reglas inter-variables. Por ejemplo, si duplica "ReglaInterVar 1", el nombre de la primera regla duplicada sería "Copia de ReglaInterVar 1", la segunda sería "Copia (2) de ReglaInterVar 1", y así sucesivamente.
- **Eliminar.** Elimina la regla seleccionada.

Definición de regla. Estos controles permiten ver y establecer propiedades para una regla seleccionada.

- **Nombre.** El nombre de la regla debe ser único para las reglas de variable única y las reglas inter-variables.
- **Expresión lógica.** Es, en esencia, la definición de la regla. Debe codificar la expresión para que los casos no válidos se evalúen en 1.

Expresiones de generación

▶ Para crear una expresión, puede pegar los componentes en el campo Expresión o escribir directamente en dicho campo.

- Puede pegar las funciones o las variables de sistema utilizadas habitualmente seleccionando un grupo de la lista Grupo de funciones y pulsando dos veces en la función o variable de las listas de funciones y variables especiales (o seleccionando la función o variable y pulsando en Insertar). Rellene los parámetros indicados mediante interrogaciones (aplicable sólo a las funciones). El grupo de funciones con la etiqueta Todo contiene una lista de todas las funciones y variables de sistema disponibles. En un área reservada del cuadro de diálogo se muestra una breve descripción de la función o variable actualmente seleccionada.
- Las constantes de cadena deben ir entre comillas o apóstrofos.
- Si los valores contienen decimales, debe utilizarse una coma(,) como indicador decimal.

Capítulo
3

Validar datos

El cuadro de diálogo Validar datos permite identificar casos, variables y valores de datos no válidos o sospechosos en el conjunto de datos activo.

Ejemplo. Una analista de datos debe proporcionar un informe mensual de satisfacción de usuarios mensual para su cliente. Debe comprobar los datos que recibe cada mes para detectar identificadores de usuarios que estén incompletos, valores de las variables que estén fuera de rango y combinaciones de valores de las variables que se suelen escribir por error. El cuadro de diálogo Validar datos permite a la analista especificar las variables que identifican a los usuarios de forma exclusiva, definir reglas de variable única para los rangos válidos de las variables y definir reglas inter-variables para detectar combinaciones imposibles. El procedimiento devuelve un informe de las variables y los casos problemáticos. Además, los datos contienen los mismos elementos de datos cada mes, de forma que la analista podrá aplicar las reglas al archivo de datos nuevo el mes siguiente.

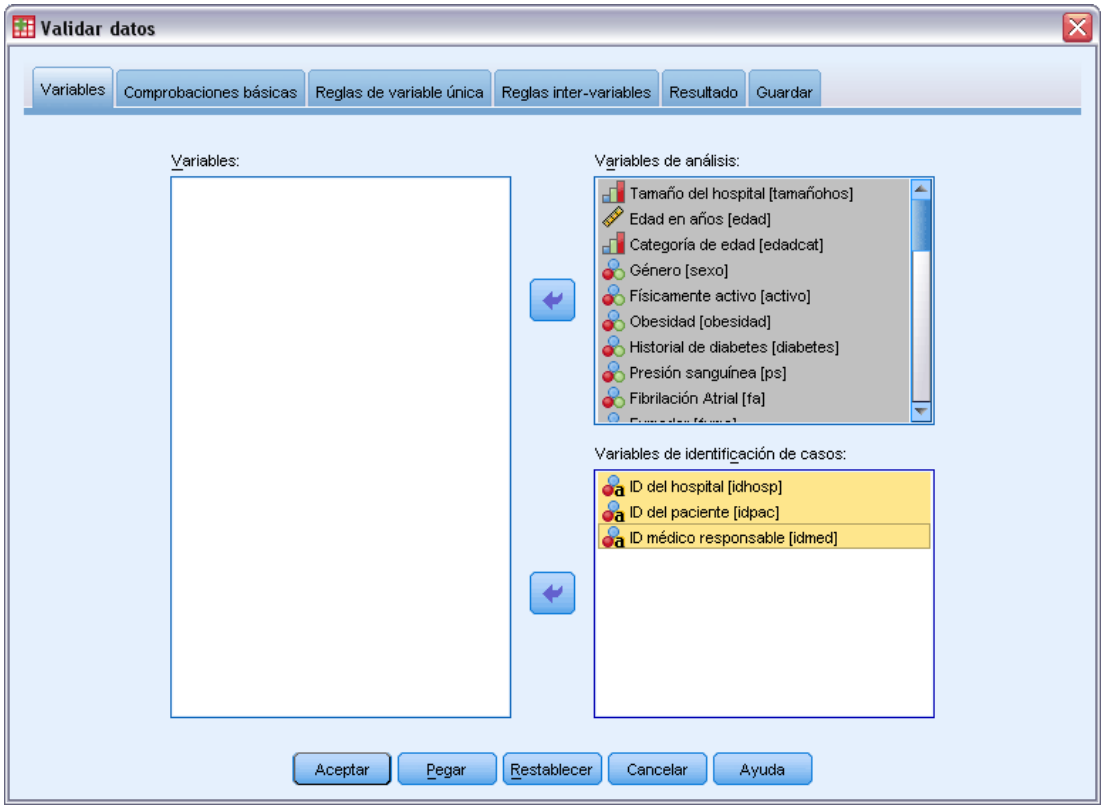
Estadísticos. El procedimiento genera listas de las variables, los casos y los valores de datos que no superan las diversas comprobaciones, recuentos de los incumplimientos de las reglas de variable única y de las reglas inter-variables, así como resúmenes descriptivos sencillos de las variables de análisis.

Ponderaciones. El procedimiento ignora la especificación de la variable de ponderación y, en su lugar, ésta recibe el mismo trato que cualquier otra variable de análisis.

Para validar datos

- Seleccione en los menús:
  Datos > Validación > Validar datos...

Figura 3-1
Cuadro de diálogo Validar datos, pestaña Variables



- Seleccione una o más variables de análisis para validarlas mediante comprobaciones de variables básicas o mediante reglas de validación de variable única.

Si lo desea, puede:

- Pulsar en la pestaña Reglas inter-variables y aplicar una o más reglas inter-variables.

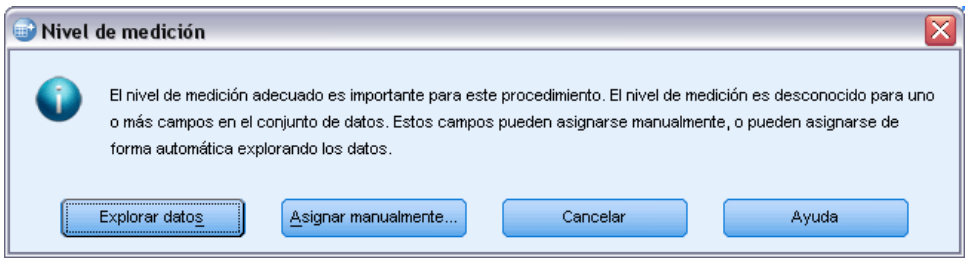
Si lo desea, puede:

- Seleccionar una o más variables de identificación de casos para comprobar si existen ID incompletos o duplicados. Las variables de ID de caso también se utilizan para etiquetar los resultados por casos. Si se especifican dos o más variables de ID de caso, la combinación de sus valores se trata como un identificador de caso.

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 3-2
Alerta de nivel de medición

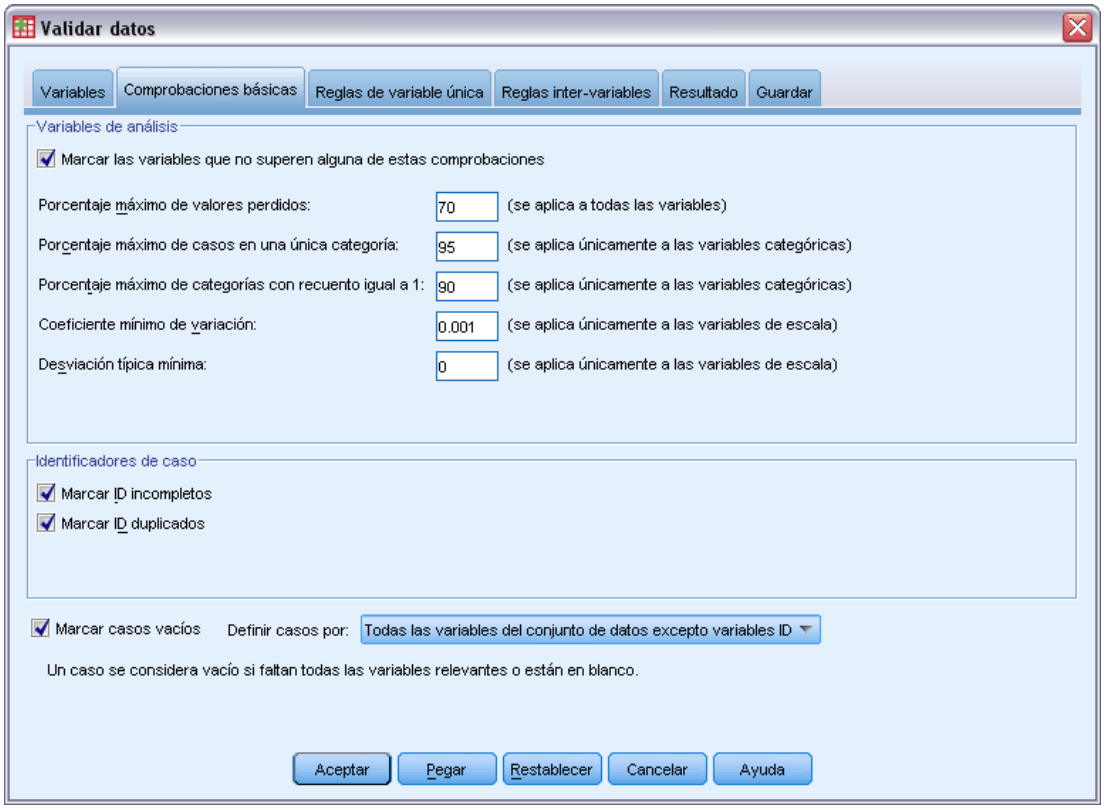


- **Explorar datos.** Lee los datos del conjunto de datos activo y asigna el nivel de medición predefinido en cualquier campo con un nivel de medición desconocido. Si el conjunto de datos es grande, puede llevar algún tiempo.
- **Asignar manualmente.** Abre un cuadro de diálogo que contiene todos los 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.

Validar datos: Comprobaciones básicas

Figura 3-3
Cuadro de diálogo Validar datos, pestaña Comprobaciones básicas



La pestaña Comprobaciones básicas permite seleccionar comprobaciones básicas para variables de análisis, identificadores de caso y casos completos.

Variables de análisis. Si ha seleccionado alguna variable de análisis en la pestaña Variables, podrá seleccionar cualquiera de las siguientes comprobaciones de su validez. La casilla de verificación permite activar o desactivar las comprobaciones.

- **Porcentaje máximo de valores perdidos.** Informa sobre las variables de análisis con un porcentaje de valores perdidos mayor que el valor especificado. El valor especificado debe ser un número positivo menor o igual que 100.
- **Porcentaje máximo de casos en una única categoría.** Si alguna variable de análisis es categórica, esta opción informa sobre las variables de análisis categóricas con un porcentaje de casos que representa una categoría de valores no perdidos mayor que el valor especificado. El valor especificado debe ser un número positivo menor o igual que 100. El porcentaje está basado en casos con valores no perdidos de la variable.
- **Porcentaje máximo de categorías con recuento igual a 1.** Si alguna variable de análisis es categórica, esta opción informa sobre las variables de análisis categóricas en las que el porcentaje de las categorías de variable que sólo contienen un caso es mayor que el valor especificado. El valor especificado debe ser un número positivo menor o igual que 100.

- **Coeficiente mínimo de variación.** Si cualquier variable de análisis es de escala, esta opción informa sobre las variables de análisis de escala en las que el valor absoluto del coeficiente de variación es menor que el valor especificado. Esta opción sólo se aplica a las variables en las que la media no es cero. El valor especificado debe ser un número no negativo. La comprobación del coeficiente de variación se desactiva si se especifica 0.
- **Desviación típica mínima.** Si alguna variable de análisis es de escala, esta opción informa sobre variables de análisis de escala cuya desviación típica es menor que el valor especificado. El valor especificado debe ser un número no negativo. La comprobación de desviación típica se desactiva si se especifica 0.

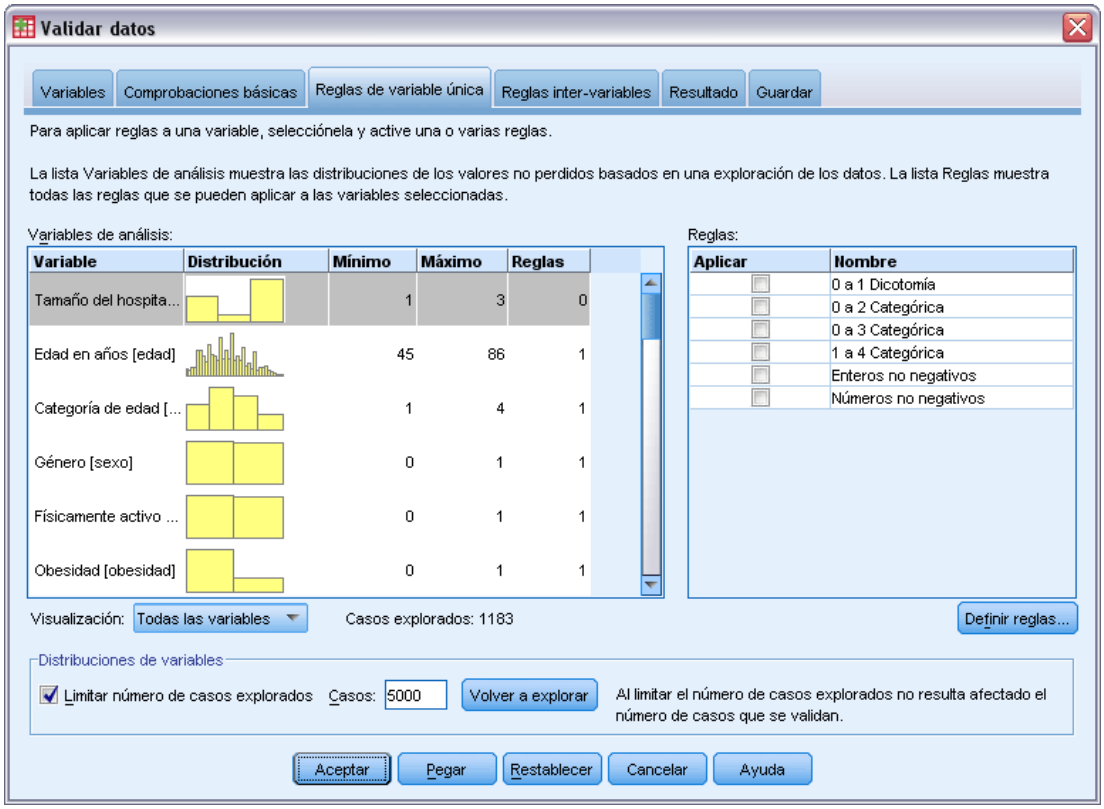
Identificadores de caso. Si ha seleccionado alguna variable de identificador de caso en la pestaña Variables, podrá seleccionar cualquiera de las siguientes comprobaciones de su validez.

- **Marcar ID incompletos.** Esta opción informa sobre casos que tienen identificadores de caso incompletos. Para un caso determinado, un identificador se considera incompleto si el valor de cualquier variable de identificación está en blanco o perdido.
- **Marcar ID duplicados.** Esta opción informa sobre casos que tienen identificadores de caso duplicados. Los identificadores incompletos se excluyen del conjunto de posibles duplicados.

Marcar casos vacíos. Esta opción informa sobre los casos en los que todas las variables están vacías o en blanco. Con el fin de identificar los casos vacíos, puede utilizar todas las variables del archivo (excepto las variables de ID) o sólo las variables de análisis definidas en la pestaña Variables.

Validar datos: Reglas de variable única

Figura 3-4
Cuadro de diálogo Validar datos, pestaña Reglas de variable única



La pestaña Reglas de variable única muestra las reglas de validación de variable única disponibles y permite aplicarlas a las variables de análisis. Para definir reglas de variable única adicionales, pulse en Definir reglas. Si desea obtener más información, consulte el tema Definir reglas de variable única en el capítulo 2 el p. 4.

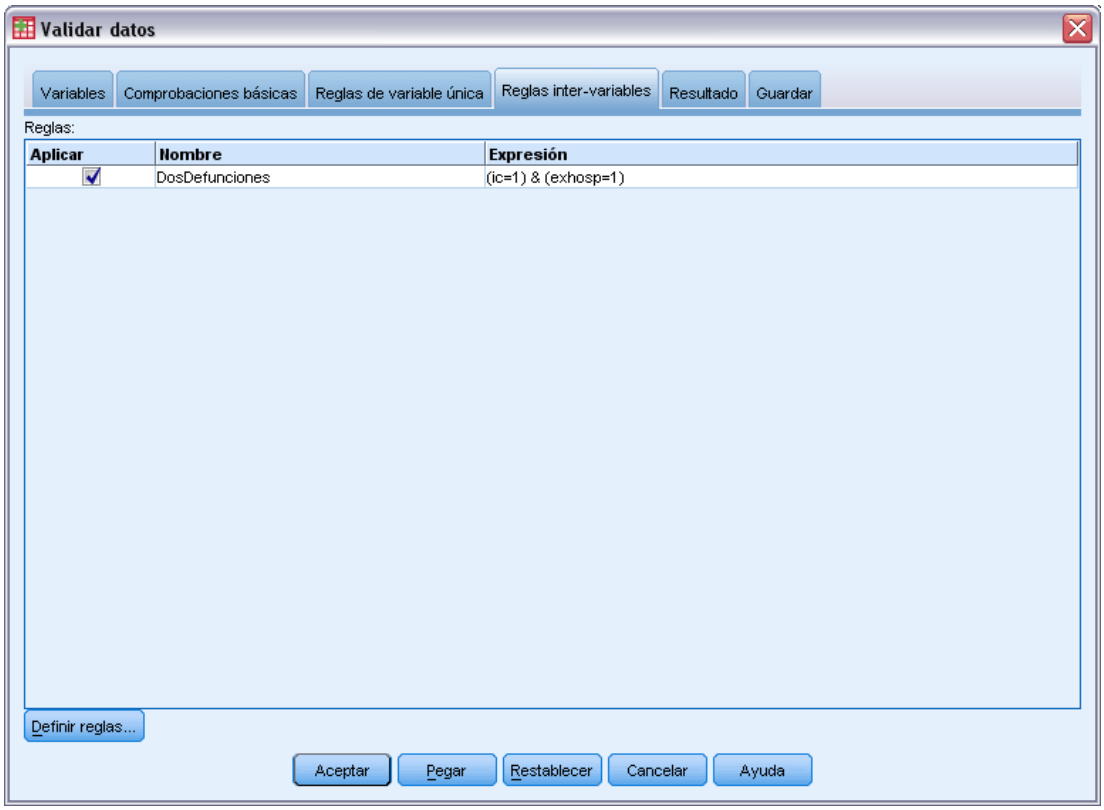
Variables de análisis. La lista muestra variables de análisis, resume sus distribuciones y muestra el número de reglas aplicadas a cada variable. Tenga en cuenta que los valores perdidos del sistema y los valores perdidos definidos por el usuario no están incluidos en los resúmenes. La lista desplegable Visualización controla las variables que se muestran; puede elegir entre Todas las variables, Variables numéricas, Variables de cadena y Variables de fecha.

Reglas. Para aplicar reglas a las variables de análisis, seleccione una o más variables y compruebe todas las reglas que desea aplicar en la lista Reglas. La lista Reglas muestra sólo reglas que son adecuadas para las variables de análisis seleccionadas. Por ejemplo, si se seleccionan variables de análisis numéricas, sólo se mostrarán reglas numéricas; si se selecciona una variable de cadena, sólo se mostrarán reglas de cadena. Si no se selecciona ninguna variable de análisis o si dichas variables tienen tipos de datos mixtos, no se muestra ninguna regla.

Distribuciones de variables. Los resúmenes de distribución que se muestran en la lista Variables de análisis pueden basarse en todos los casos o en una exploración de los primeros *n* casos, como se especifica en el cuadro de texto Casos. Puede actualizar los resúmenes de distribución al pulsar en Volver a explorar.

Validar datos: Reglas inter-variables

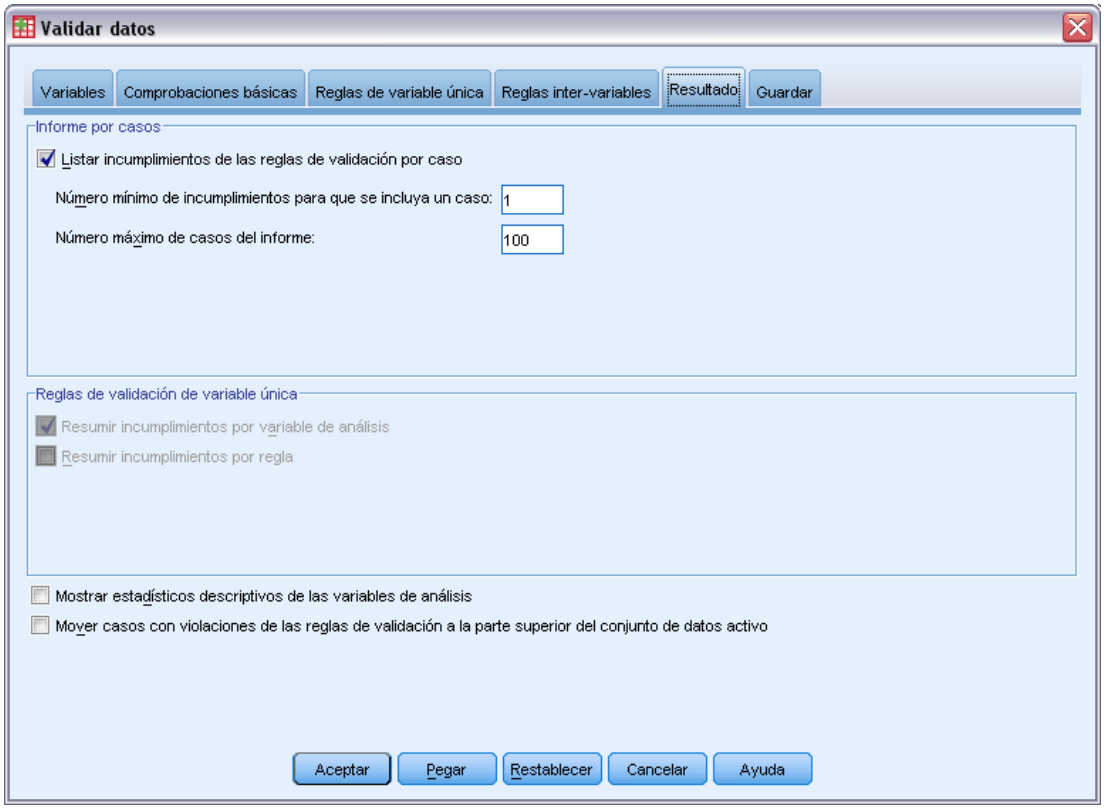
Figura 3-5
Cuadro de diálogo Validar datos, pestaña Reglas reglas inter-variables



La pestaña Reglas inter-variables muestra reglas inter-variables disponibles y permite aplicarlas a los datos. Para definir reglas inter-variables adicionales, pulse en Definir reglas. Si desea obtener más información, consulte el tema Definir reglas inter-variables en el capítulo 2 el p. 6.

Validar datos: Resultados

Figura 3-6
Cuadro de diálogo Validar datos, pestaña Resultado



Informe por casos. Si ha aplicado alguna regla de validación de variable única o inter-variables, puede solicitar un informe que contenga los incumplimientos de las reglas de validación de casos individuales.

- **Número mínimo de incumplimientos.** Esta opción especifica el número mínimo de incumplimientos de reglas requeridos para que un caso se incluya en el informe. Especifique un número entero positivo.
- **Número máximo de casos.** Esta opción especifica el número máximo de casos incluidos en el informe de casos. Especifique un número entero positivo menor o igual que 1000.

Reglas de validación de variable única. Si ha aplicado alguna regla de validación de variable única, puede elegir cómo mostrar los resultados o si se van a mostrar.

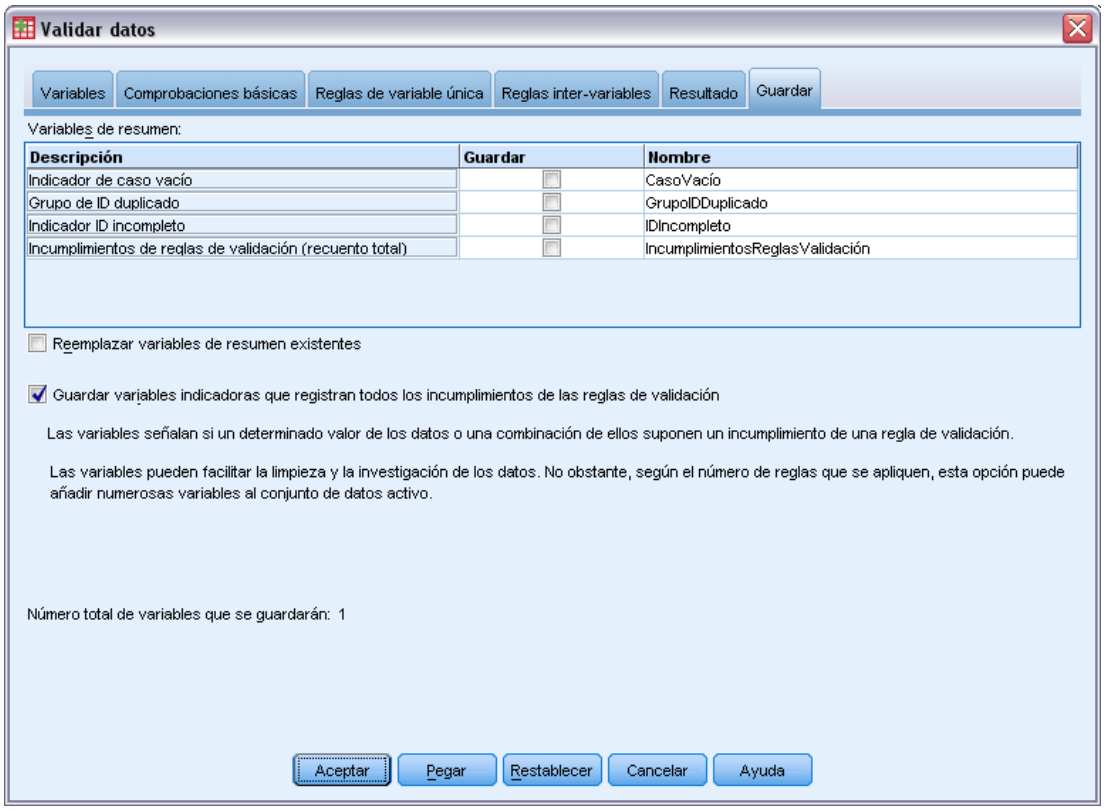
- **Resumir incumplimientos por variable de análisis.** Para cada variable de análisis, esta opción muestra todas las reglas de validación de variable única que se incumplieron y el número de valores que incumplió cada regla. También informa sobre el número total de incumplimientos de regla de variable única de cada variable.
- **Resumir incumplimientos por regla.** Para cada regla de validación de variable única, esta opción informa sobre las variables que incumplieron la regla y el número de valores no válidos por variable. También informa sobre el número total de valores que incumplieron cada regla entre las variables.

Mostrar estadísticos descriptivos. Esta opción permite solicitar estadísticos descriptivos para las variables de análisis. Se genera una tabla de frecuencias para cada variable categórica. Se genera una tabla de resumen de estadísticos que incluye la media, la desviación típica, el mínimo y el máximo para las variables de escala.

Mover casos con incumplimientos de las reglas de validación. Esta opción mueve los casos con incumplimientos de las reglas inter-variables y de variable única a la parte superior del conjunto de datos activo para facilitar su examen.

Validar datos: Guardar

Figura 3-7
Cuadro de diálogo Validar datos, pestaña Guardar



La pestaña Guardar permite guardar variables que registran los incumplimientos de las reglas en el conjunto de datos activo.

Variables de resumen. Variables individuales que se pueden guardar. Marque un cuadro para guardar la variable. Los nombres por defecto de las variables se proporcionan y se pueden editar.

- **Indicador de caso vacío.** El valor 1 se asigna a los casos vacíos. El resto de casos se codifican como 0. Los valores de la variable reflejan el ámbito especificado en la pestaña Comprobaciones básicas.

- **Grupo de ID duplicado** Se asigna el mismo número de grupo a los casos que comparten el mismo identificador de caso (diferentes de los que tienen identificadores incompletos). Los casos con identificadores únicos o incompletos se codifican como 0.
- **Indicador ID incompleto.** Se asigna el valor 1 a los casos con identificadores de casos vacíos o incompletos. El resto de casos se codifica como 0.
- **Incumplimientos de reglas de validación.** Recuento total por caso de los incumplimientos de reglas de validación de variable única e inter-variables.

Reemplazar variables de resumen existentes. Las variables que se guardan en el archivo de datos deben tener nombres únicos o sustituir a las variables con el mismo nombre.

Guardar variables indicadoras. Esta opción permite guardar un registro completo de incumplimientos de reglas de validación. Cada variable corresponde a una aplicación de una regla de validación y tiene un valor de 1 si el caso incumple la regla y un valor de 0 si no lo hace.

Capítulo
4

Preparación automática de datos

La preparación de los datos para su análisis es uno de los pasos más importantes en cualquier proyecto y, tradicionalmente, uno de los que más tiempo requieren. Preparación automática de datos (ADP) controla las tareas automáticamente, analizando los datos e identificando problemas, filtrando campos problemáticos o sin posibilidades de ser útiles, derivando nuevos atributos cuando sea necesario y mejorando el rendimiento mediante técnicas de filtrado inteligente. Puede utilizar el algoritmo de una forma totalmente **automática**, permitiendo seleccionar y aplicar soluciones; o de forma **interactiva**, previendo los cambios antes de que se realicen y aceptarlos o rechazarlos según sea necesario.

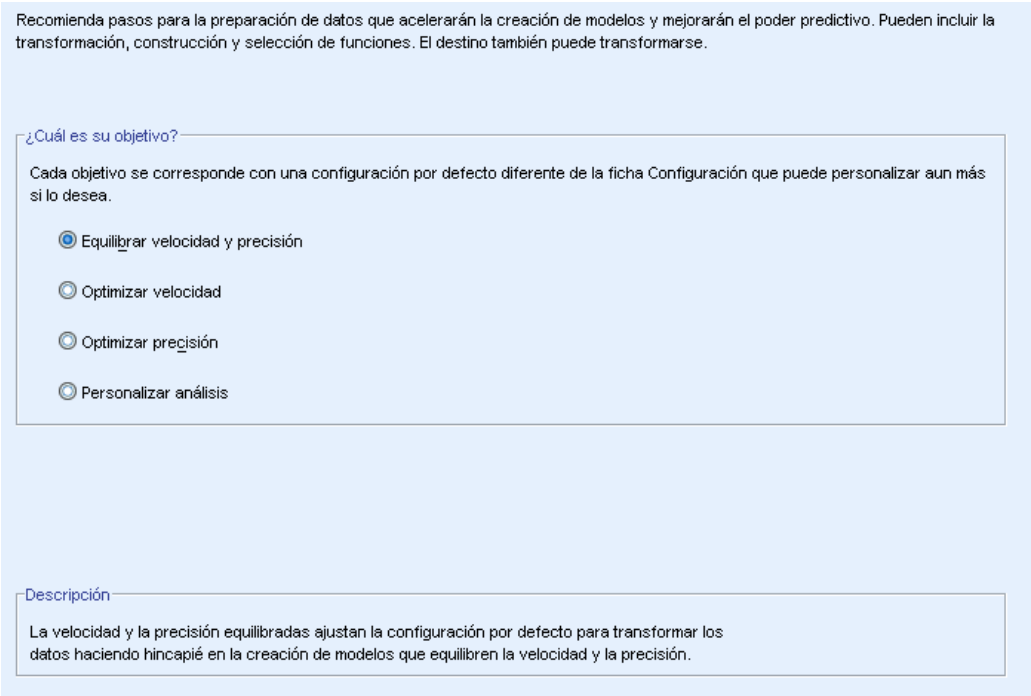
ADP permite hacer que sus datos estén listos para la generación de modelos de forma rápida y fácil, sin necesidad de tener conocimientos previos de los conceptos previos implicados. Los modelos tienden a crearse y puntuarse con mayor rapidez; además, el uso de ADP mejora la solidez de los procesos de modelado automatizados.

Nota: cuando el ADP prepara un campo para su análisis, crea un nuevo campo con los ajustes o transformaciones, en vez de reemplazar los valores y propiedades existentes del campo anterior. El campo anterior no se usa en más análisis, su papel se define como Ninguno. Tenga también en cuenta que cualquier información sobre los valores perdidos definidos por el usuario no se transfiere a estos campos recién creados y cualquier valor perdido en el nuevo campo se considera valores perdidos del sistema.

Ejemplo. Una correduría de seguros con recursos limitados para investigar las reclamaciones de seguros de los asegurados desea crear un modelo para etiquetar las reclamaciones sospechosas y potencialmente fraudulentas. Antes de construir el modelo, leerán los datos para el modelado mediante la preparación automática de datos. Como desean revisar las transformaciones propuestas antes de que se apliquen las transformaciones, utilizarán la preparación automática de datos en modo interactivo. Si desea obtener más información, consulte el tema Uso interactivo de la preparación automática de datos en el capítulo 8 el p. 85.

Un grupo del sector del automóvil desea realizar un seguimiento de las ventas de diversos vehículos a motor. Para poder identificar los modelos como mejor y peor rendimiento, desean establecer una relación entre las ventas de vehículos y las características de los vehículos. Utilizarán la preparación automática de datos para preparar los datos para el análisis y crearán modelos utilizando la preparación “anterior” y “posterior” de datos para ver cómo difieren los resultados. Si desea obtener más información, consulte el tema Uso automático de la preparación automática de datos en el capítulo 8 el p. 96.

Figura 4-1
Pestaña Objetivo de Preparación automática de datos



¿Cuál es su objetivo? Preparación automática de datos recomienda ejecutar pasos para la preparación de datos que afectan a la velocidad con la que el resto de algoritmos pueden generar modelos y mejorar el potencial predictivo de esos modelos. Pueden incluir la transformación, construcción y selección de funciones. El destino también puede transformarse. Puede especificar las prioridades de generación de modelos en las que se deben centrar el proceso de preparación de datos.

- **Equilibrar velocidad y precisión.** Esta opción prepara los datos para dar igual prioridad a la velocidad con la que se procesan los datos por algoritmos de creación de modelos y la precisión de los pronósticos.
- **Optimizar velocidad.** Esta opción prepara los datos para dar prioridad a la velocidad con la que se procesan los datos por los algoritmos de construcción de modelos. Si trabaja con conjuntos de datos muy grandes o busca una respuesta rápida, seleccione esta opción.
- **Optimizar precisión.** Esta opción prepara los datos para dar prioridad a la precisión de los pronósticos producidos por los algoritmos de construcción de modelos.
- **Análisis personalizado.** Seleccione esta opción si desea cambiar manualmente el algoritmo de la pestaña Configuración. Tenga en cuenta que esta configuración se selecciona automáticamente si realiza cambios posteriores a muchas opciones de la pestaña Configuración que sean incompatibles con los de otros objetivos.

Para obtener preparación de datos automática

Seleccione en los menús:

Transformar > Preparar datos para modelado > Automática...

▶ Pulse en Ejecutar.

Si lo desea, puede:

- Especifique un objetivo en la pestaña Objetivos.
- Especifique asignaciones de campo en la pestaña Campos.
- Especifique la configuración de experto en la pestaña Configuración.

Para obtener preparación de datos interactiva

Seleccione en los menús:
Transformar > Preparar datos para modelado > Interactiva...

▶ Pulse en Analizar en la barra de herramientas en la parte superior del cuadro de diálogo.

▶ Pulse en la pestaña análisis y consulte los pasos de preparación de datos sugeridos.

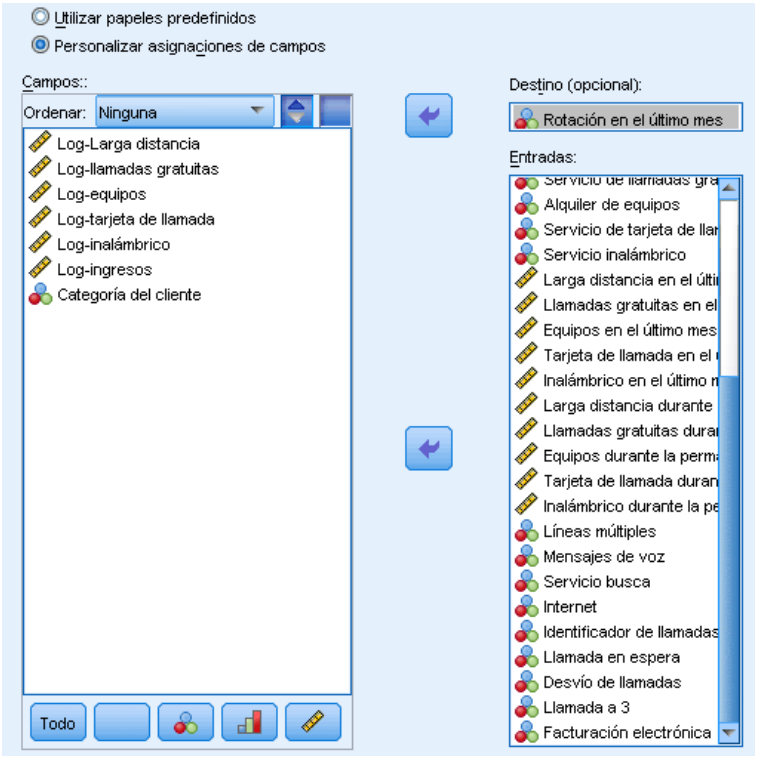
▶ Si está satisfecho, pulse en Ejecutar. En caso contrario, pulse en Borrar análisis, cambie los ajustes que sea necesario y pulse en Analizar.

Si lo desea, puede:

- Especifique un objetivo en la pestaña Objetivos.
- Especifique asignaciones de campo en la pestaña Campos.
- Especifique la configuración de experto en la pestaña Configuración.
- Guardar los pasos recomendados de preparación de datos en un archivo XML pulsando en Guardar XML.

Pestaña Campos

Figura 4-2
Pestaña Campos de Preparación automática de datos



La pestaña Campos especifica los campos que se deben preparar para futuros análisis.

Utilizar papeles predefinidos. Esta opción utiliza información de campos existentes. Si hay un solo campo con una función como Destino, se utilizará como el destino; de lo contrario no habrá ningún objetivo. Todos los campos con un papel predefinido como Entrada se utilizarán como entradas. Al menos un campo de entrada es necesario.

Utilizar asignaciones de campos personalizadas. Cuando sobrescribe los papeles de campos moviendo los campos desde sus listas predeterminadas, el cuadro de diálogo cambia automáticamente a esta opción. Cuando realice asignaciones de campos personalizadas, especifique los siguientes campos:

- **Destino (opcional).** Si planea crear modelos que requieren un destino, seleccione el campo de destino. Es similar a definir el papel del campo a Destino.
- **Entradas.** Seleccione uno o más campos de entrada. Es similar a definir el papel del campo a Entrada.

Pestaña Configuración

La pestaña Configuración contiene diferentes grupos de ajustes que puede modificar para ajustar con precisión la forma en que el algoritmo procesa sus datos. Si realiza algún cambio en la configuración por defecto que sea incompatible con el resto de objetivos, la pestaña Objetivo se actualiza automáticamente para seleccionar la opciónPersonalizar análisis.

Preparar fechas y horas

Figura 4-3
Preparación automática de datos: Configuración de fecha y hora

Muchos algoritmos no pueden tratar directamente los detalles de fecha y hora; estas configuraciones permiten derivar nuevos datos de duración que pueden utilizarse como entradas de modelo de fechas y horas de sus datos existentes. Los campos que contienen las fechas y las horas se deben predefinir con los tipos de almacenamiento de fecha u hora. Los campos de fecha y hora originales no se recomiendan como entradas de modelo posteriores a la preparación automática de datos.

Preparar fechas y horas para el modelado. Si cancela la selección de esta opción se desactivan todos los demás controles de Preparar fechas y horas mientras se mantienen las selecciones.

Calcular tiempo transcurrido hasta fecha de referencia. Esto produce el número de años/meses/días desde una fecha de referencia para cada variable que contenga fechas.

- **Fecha de referencia.** Especifique la fecha desde la que se calculará la duración en lo relativo a la información de fecha de los datos de entrada. Si selecciona Fecha de hoy, la fecha actual del sistema se utilizará siempre que se ejecute el nodo ADP. Para utilizar una fecha específica, seleccione Fecha fija e introduzca la fecha obligatoria.
- **Unidades de duración de fecha.** Especifique si el nodo debería decidir automáticamente sobre la unidad de duraciones de fecha o establezca Unidades fijas como Años, Meses o Días.

Calcular tiempo transcurrido hasta hora de referencia. Esto produce el número de horas/minutos/segundos desde una hora de referencia para cada variable que contenga horas.

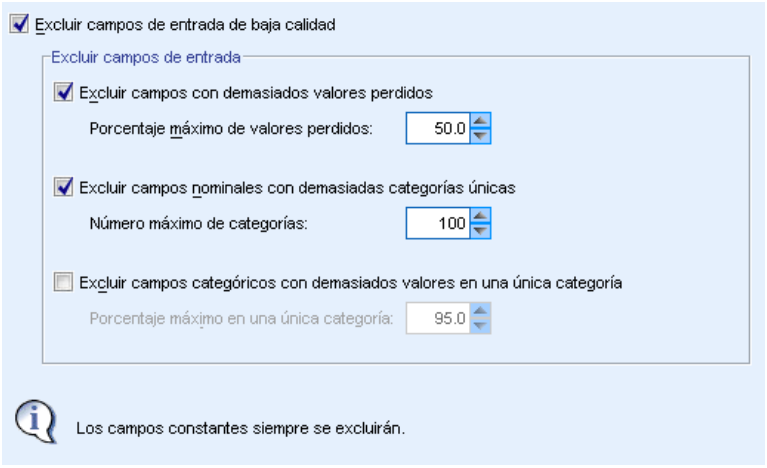
- **Hora de referencia.** Especifique la hora desde la que se calculará la duración en lo relativo a la información de hora de los datos de entrada. Si selecciona Hora actual, la hora actual del sistema se utilizará siempre que se ejecute el nodo ADP. Para utilizar una hora específica, seleccione Hora fija e introduzca los detalles obligatorios.
- **Unidades de duración de tiempo.** Especifique si el nodo debería decidir automáticamente sobre la unidad de duraciones de hora o establezca Unidades fijas como Horas, Minutos o Segundos.

Extraer elementos temporales cíclicos. Utilice esta configuración para dividir un único campo de fecha o de hora en uno o más campos. Por ejemplo, si selecciona las tres casillas de verificación de fecha, el campo de fecha de entrada "1954-05-23" se dividirá en tres campos: 1954, 5 y 23, cada uno con el sufijo definido en el panel Nombres de campos y el campo de fecha original se ignorará.

- **Extraer de fechas.** Para cualquier entrada de fecha, especifique si desea extraer años, meses, días o cualquier combinación.
- **Extraer de horas.** Para cualquier entrada de hora, especifique si desea extraer horas, minutos, segundos o cualquier combinación.

Excluir campos

Figura 4-4
Configuración de Excluir campos de preparación automática de datos



Los datos de mala calidad pueden afectar a la precisión de sus predicciones; por lo tanto, puede especificar el nivel de calidad aceptable de las características de entrada. Todos los campos que no sean constantes o les falte el 100% de los valores se excluirán automáticamente.

Excluir campos de entrada de baja calidad. Si cancela la selección de esta opción se desactivan todos los demás controles de Excluir campos mientras se mantienen las selecciones.

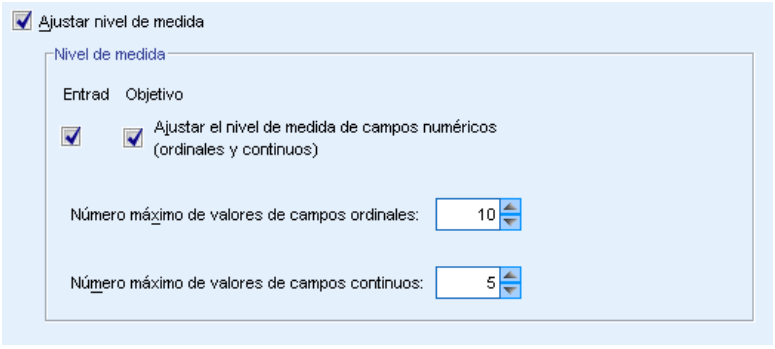
Excluir campos con demasiados valores perdidos. Los campos con un porcentaje de valores perdidos mayor que el porcentaje especificado se eliminan de análisis posteriores. Especifique un valor superior o igual a 0 (que equivale a cancelar la selección de esta opción) y menor o igual a 100, aunque los campos que tienen valores que faltan se excluyen automáticamente. El valor por defecto es 50.

Excluir campos nominales con demasiadas categorías únicas. Los campos nominales con un número de categorías superior al especificado se eliminarán de análisis posteriores. Especifique un número entero positivo. El valor predeterminado es 100. Esto resulta útil para eliminar automáticamente campos que contengan información única de registros para el modelado, como ID, dirección o nombre.

Excluir campos categóricos con demasiados valores en una única categoría. Los campos nominales y ordinales con una categoría con un porcentaje de registros superior al especificado se eliminarán de análisis posteriores. Especifique un valor superior o igual a 0 (que equivale a cancelar la selección de esta opción) y menor o igual a 100, aunque los campos constantes se excluyen automáticamente. El valor por defecto es 95.

Ajustar medida

Figura 4-5
Configuración de Ajustar medida de preparación automática de datos



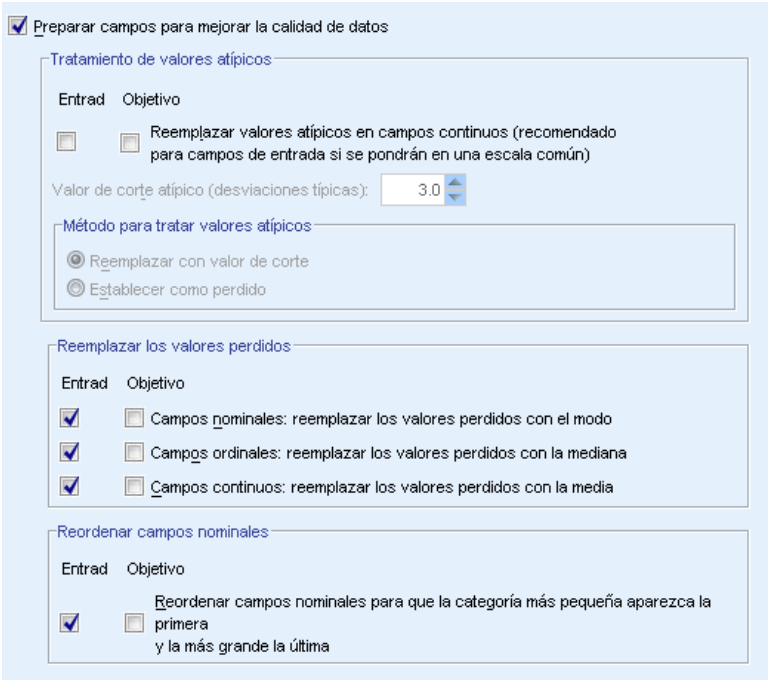
Ajustar nivel de medida. Si cancela la selección de esta opción se desactivan todos los demás controles de Ajustar medida mientras se mantienen las selecciones.

Nivel de medida. Especifique si el nivel de medida de campos continuos con “demasiados pocos” valores se pueden ajustar a ordinales. Los campos ordinales con “demasiados” valores se pueden ajustar a continuos.

- **Número máximo de valores de campos ordinales.** Los campos ordinales con un número de categorías superior al especificado se reestructuran como campos continuos. Especifique un número entero positivo. El valor por defecto es 10. Este valor debe ser mayor o igual al número mínimo de valores de campos continuos.
- **Número mínimo de valores de campos continuos.** Los campos continuos con un número de valores únicos inferior al especificado se reestructuran como campos ordinales. Especifique un número entero positivo. El valor por defecto es 5. Este valor debe ser menor o igual al número máximo de valores de campos ordinales.

Mejorar la calidad de datos

Figura 4-6
Configuración de Mejorar la calidad de datos de preparación automática de datos



Preparar campos para mejorar la calidad de datos. Si cancela la selección de esta opción se desactivan todos los demás controles de Mejorar la calidad de datos mientras se mantienen las selecciones.

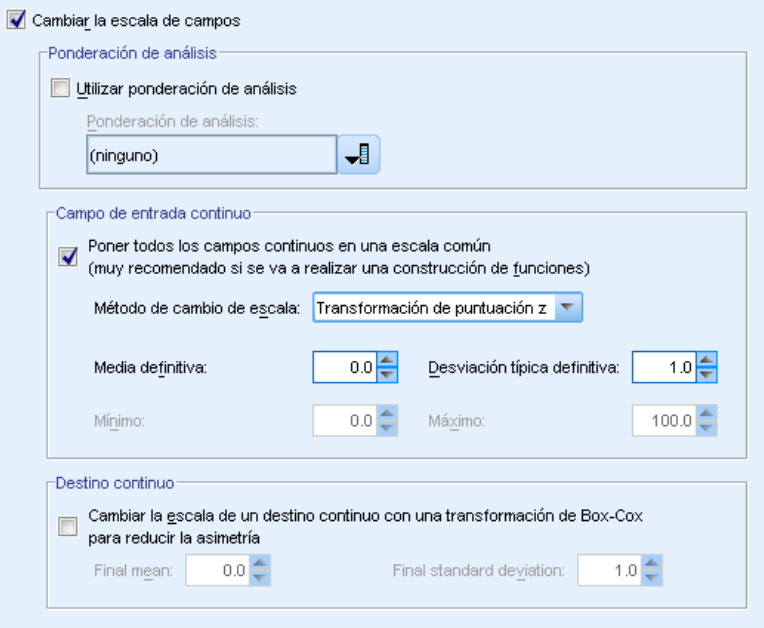
Tratamiento de valores atípicos. Especifique si sustituirá los atípicos por entradas y destino; si es así, especifique un criterio de corte atípico, medido en desviaciones típicas y un método para sustituir atípicos. Los atípicos se pueden sustituir por recorte (ajuste del corte de valor) o configurándolos como valores perdidos. Todos los valores atípicos establecidos como valores ausentes siguen la configuración de gestión de valores ausentes seleccionada a continuación.

Reemplazar valores perdidos. Especifique si desea sustituir los valores perdidos de campos continuos, nominales u ordinales.

Reordenar campos nominales. Seleccione esta opción para recodificar los valores de campos nominales (conjunto) de menor (menos frecuencia) a mayor (mayor frecuencia) según su categoría. Los valores de nuevo campo comienzan por 0, como la categoría menos frecuente. Tenga en cuenta que el nuevo campo será numérico aunque el original sea una cadena. Por ejemplo, si los valores de los datos de un campo nominal son "A", "A", "A", "B", "C", "C", la preparación automática de datos recodificará "B" a 0, "C" a 1 y "A" a 2.

Cambiar la escala de campos

Figura 4-7
Configuración de Cambiar la escala de campos de preparación automática de datos



Cambiar la escala de campos. Si cancela la selección de esta opción se desactivan todos los demás controles de Cambiar la escala de campos mientras se mantienen las selecciones.

Ponderación de análisis. Esta variable contiene ponderaciones de análisis (regresión o muestra). Las ponderaciones de análisis se utilizan para contabilizar las diferencias existentes en la varianza entre los niveles del campo de salida. Seleccione un campo continuo.

Campos de entrada continuos. Se normalizarán los campos de entrada continuos utilizando una transformación de puntuaciones z o transformación mínima/máxima. Las entradas de cambio de escala son especialmente útiles si selecciona Realizar creación de características en la configuración de selección y creación.

- **Transformación de puntuación z.** Si utiliza la media observada y una desviación típica como estimaciones de parámetros de población, los campos se tipifican y las puntuaciones *z* se asignan a los valores correspondientes de una distribución normal con la Media final y Desviación típica final especificadas. Especifique un número para Media final y un número

positivo para Desviación típica final. Los valores por defecto son 0 y 1, respectivamente, correspondientes al cambio de escala tipificado.

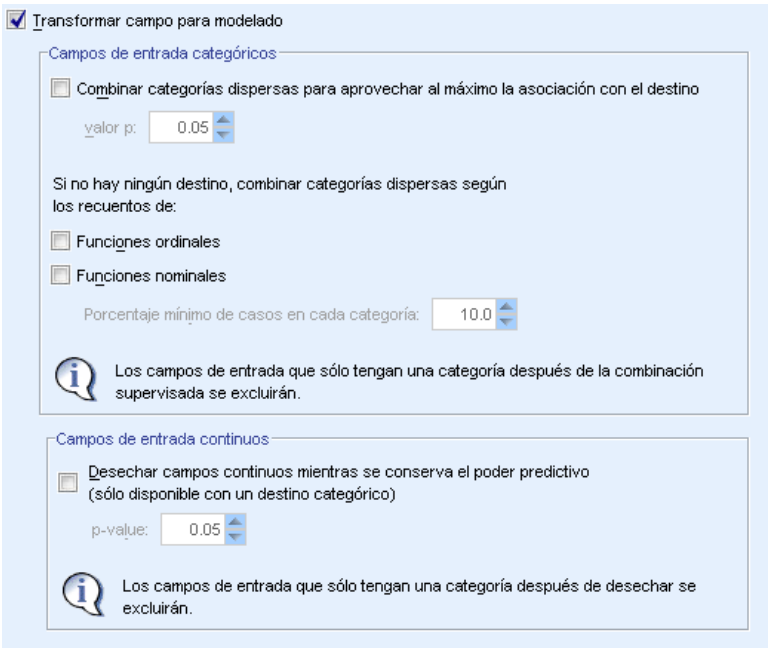
- **Transformación mín. y máx.** Si utiliza los valores mínimo y máximo observados como estimaciones de parámetros de población, los campos se asignan a los valores correspondientes de una distribución uniforme con los valores mínimo y máximo especificados. Especifique números con un valor máximo superior al mínimo.

Destino continuo. Transforma un destino continuo utilizando la Transformación de Box-Cox en un campo con una distribución normal aproximada con Media final y Desviación típica final especificada. Especifique un número para Media final y un número positivo para Desviación típica final. Los valores por defecto son 0 y 1, respectivamente.

Nota: Si ADP transforma un destino, los siguientes modelos generados utilizando el destino transformado puntúan las unidades transformadas. Para interpretar y utilizar los resultados, debe convertir el valor pronosticado a la escala original. Si desea obtener más información, consulte el tema Puntuaciones de transformación retrospectiva el p. 46.

Transformar campos

Figura 4-8
Configuración de transformar campos de preparación automática de datos



Para mejorar el poder predictivo de sus datos, puede transformar los campos de entrada.

Transformar campo para modelado. Si cancela la selección de esta opción se desactivan todos los demás controles de Transformar campos mientras se mantienen las selecciones.

Campos de entrada categóricos

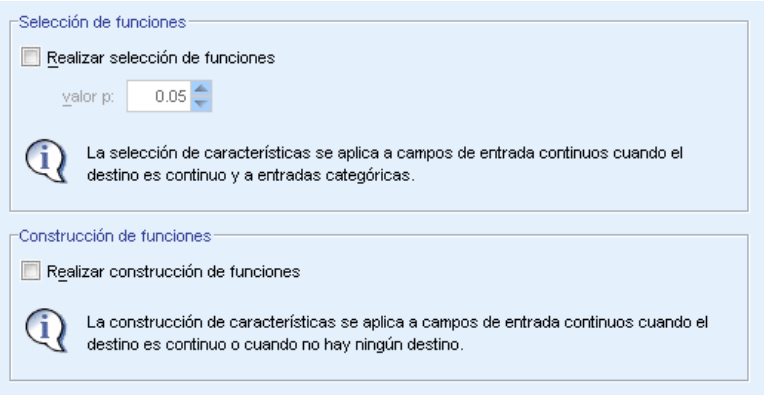
- **Combinar categorías dispersas para aprovechar al máximo la asociación con el destino.** Seleccione esta opción para realizar un modelo más parsimonioso reduciendo el número de campos que deben procesarse junto con el destino. Las categorías similares se identifican en función de la relación entre la entrada y destino. Las categorías que no son significativamente diferentes; es decir, que tienen un valor *p* superior al valor especificado, se fusionan. Especifique un valor mayor o igual que 0 y menor o igual que 1. Si todas las categorías se combinan en una, las versiones original y derivada del campo se excluyen de futuros análisis porque no tienen ningún valor como predictor.
- **Si no hay ningún destino, combine las categorías dispersas según los recuentos.** Si el conjunto de datos no tiene destino, puede fusionar las categorías dispersas de campos ordinales y nominales. El método de frecuencias iguales se utiliza para fusionar categorías con un porcentaje mínimo especificado inferior al número de registros. Especifique un valor mayor o igual que 0 y menor o igual que 100. El valor por defecto es 10. La fusión se detiene si no hay categorías con un porcentaje mínimo especificado menor que el porcentaje de casos o si sólo quedan dos categorías.

Campos de entrada continuos. Si el conjunto de datos incluye un destino categórico, puede crear un intervalo para entradas continuas con asociaciones fuertes para mejorar el rendimiento del procesamiento. Los intervalos se crean en función de las propiedades de "subconjuntos homogéneos", que se identifican por el método Scheffe que utiliza el valor *p* especificado como el valor alfa del valor crítico para determinar subconjuntos homogéneos. Especifique un valor mayor que 0 y menor o igual que 1. El valor por defecto es 0,05. Si la operación de creación de intervalos da como resultado un único intervalo para un campo específico, las versiones original y con intervalos del campo se excluyen porque no tienen ningún valor como predictor.

Nota: Los intervalos en ADP son diferentes de intervalos óptimos. Intervalos óptimos utiliza entropía de información para convertir un campo continuo en un campo categórico; necesita ordenar los datos y almacenarlo todo en memoria. ADP utiliza subconjuntos homogéneos para agrupar un campo continuo, lo que significa que el intervalo ADP no necesita ordenar los datos ni almacenar todos los datos en memoria. El uso del método de subconjunto homogéneo para agrupar un campo continuo significa que el número de categorías después de la agrupación es siempre menor o igual que el número de categorías del destino.

Seleccionar y construir

Figura 4-9
Configuración de Seleccionar y construir de preparación automática de datos



Para mejorar el poder predictivo de sus datos, puede crear nuevos campos basados en los campos existentes.

Realizar selección de características. Una entrada continua se elimina del análisis si el valor de *p* de su correlación con el destino es mayor que el valor *p* especificado.

Realizar construcción de características. Seleccione esta opción para derivar nuevas características de una combinación de varias características existentes. Las características antiguas no se emplean en otros análisis. Esta opción sólo es aplicable a características de entrada continuas en las que el destino es continuo o en las que no hay destino.

Nombres de campos

Figura 4-10
Configuración de Nombrar campos de preparación automática de datos

Para identificar fácilmente las características nuevas y transformadas, ADP crea y aplica nombres, prefijos o sufijos básicos nuevos. Puede modificar estos nombres para que sean más relevantes para sus propias necesidades y datos.

Campos transformados y construidos. Especifique las extensiones de nombre que se aplicarán a campos de entrada y de destino transformado.

Además, especifique el nombre de prefijo que se aplicará a todas las características que se creen mediante la configuración de Crear y seleccionar. El nuevo nombre se crea adjuntando un sufijo numérico a este nombre de raíz de prefijo. El formato del número depende de cuántas nuevas características se deriven, por ejemplo:

- 1-9 características creadas se denominarán: característica1 a característica9.
- 10-99 características creadas se denominarán: característica01 a característica99.
- 100-999 características creadas se denominarán: característica001 a característica999, etcétera.

De esta forma se garantiza que las características creadas se ordenen de forma adecuada independientemente de cuántas sean.

Duraciones calculadas de fechas y horas. Especifique las extensiones de nombre que se aplicarán a duraciones calculadas a partir de fechas y horas.

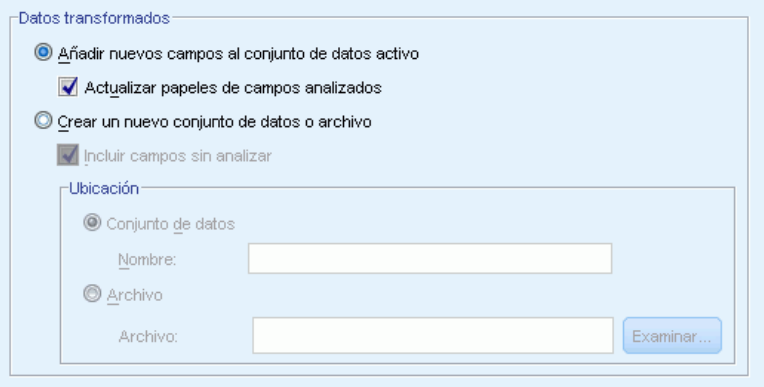
Elementos cíclicos extraídos de fechas y horas. Especifique las extensiones de nombre que se aplicarán a elementos cíclicos extraídos de fechas y horas.

Aplicación y almacenamiento de transformaciones

Dependiendo de si utiliza los cuadros de diálogo de preparación automática de datos o interactiva, los ajuste de aplicación y almacenamiento de transformaciones son ligeramente diferentes.

Configuración de Aplicar transformaciones de preparación automática de datos

Figura 4-11
Configuración de Aplicar transformaciones de preparación automática de datos

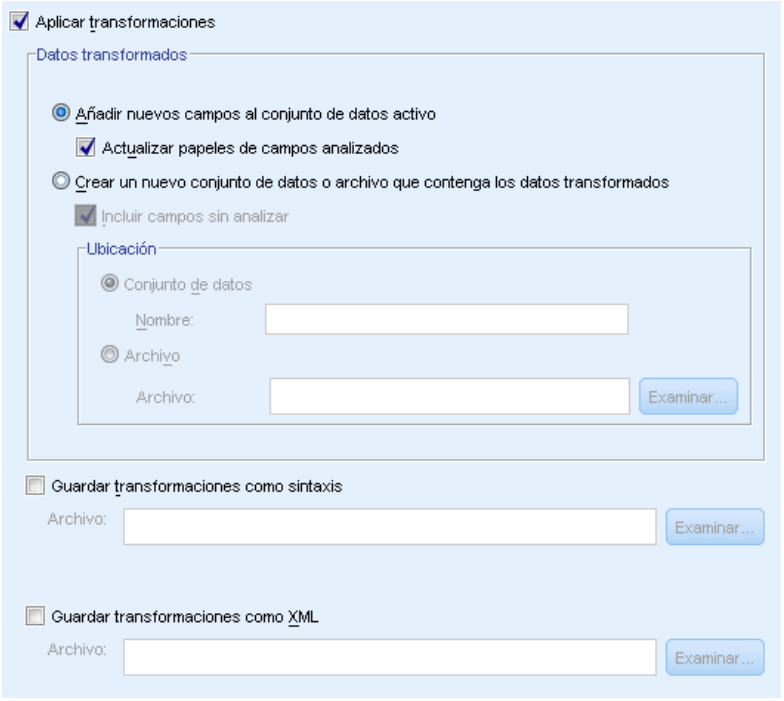


Datos transformados. Esta configuración especifica dónde se guardarán los datos transformados.

- **Añadir nuevos campos al conjunto de datos activo.** Los campos creados con preparación automática de datos se añaden al conjunto de datos activos como campos nuevos. Actualizar papeles de campos analizados definirá el papel a Ninguno para todos los campos excluidos de futuros análisis por preparación automática de datos.
- **Cree un nuevo conjunto de datos o el archivo con los datos transformados.**Los campos recomendados por la preparación automática de datos se añaden a un conjunto de datos o archivo nuevos. Incluir campos sin analizar añade campos en el conjunto de datos original que no se han especificado en la pestaña Campos al nuevo conjunto de datos. Esto resulta útil para transferir campos que contenga información que no se utilice en el modelado, como ID, dirección o nombre, al nuevo conjunto de datos.

Configuración de Aplicar y guardar de preparación automática de datos

Figura 4-12
Configuración de Aplicar y guardar de preparación automática de datos



El grupo Datos transformados es el mismo que en la preparación interactiva de datos. En la preparación automática de datos hay disponibles las siguientes opciones adicionales:

Aplicar transformaciones. En los cuadros de diálogo de preparación automática de datos, si cancela la selección de esta opción se desactivan todos los demás controles de Aplicar y Guardar mientras se mantienen las selecciones.

Guardar transformaciones como sintaxis. Guarda las transformaciones recomendadas como sintaxis de comandos en un archivo externo. El cuadro de diálogo de preparación de datos interactiva no tiene este control porque pegará las transformaciones como sintaxis de comandos en la ventana de sintaxis si pulsa en Pegar.

Guardar transformaciones como XML. Guarda las transformaciones recomendadas como XML en un archivo externo, que se puede fusionar con PMML de modelo utilizando `TMS MERGE` o aplicado a otros conjuntos de datos utilizando `TMS IMPORT`. El cuadro de diálogo de preparación de datos interactiva no tiene este control porque guarda las transformaciones como XML si pulsa en Guardar XML en la barra de herramientas en la parte superior del cuadro de diálogo.

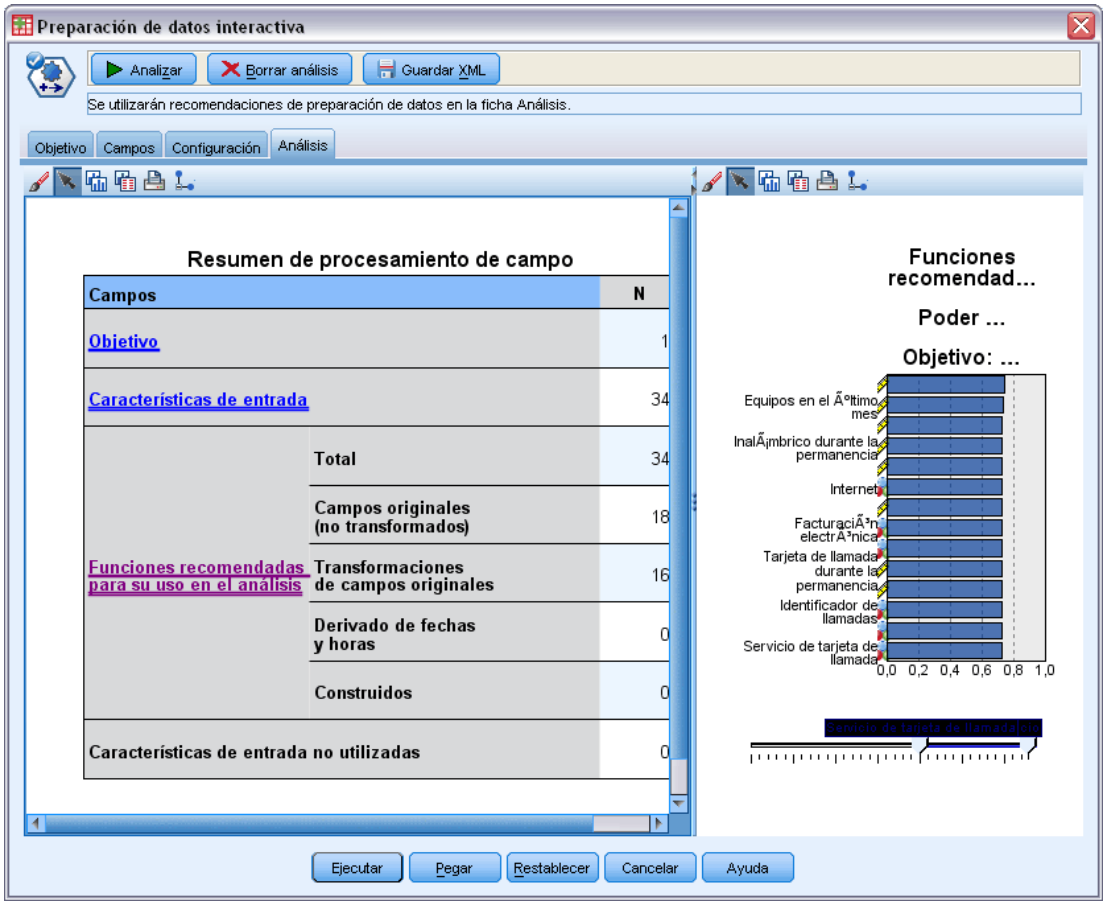
Pestaña análisis

Nota: La pestaña Análisis se utiliza en el cuadro de diálogo de preparación de datos interactiva le permite revisar las transformaciones recomendadas. El cuadro de diálogo de diálogo de preparación automática de datos no incluye este paso.

- Cuando haya terminado con la configuración del nodo ADP, incluyendo las modificaciones realizadas en las pestañas Objetivos, Campos y Configuración, pulse Analizar datos; el algoritmo aplica la configuración a las entradas de datos y muestra los resultados en la pestaña Análisis.

La pestaña Análisis contiene resultados tabulares y gráficos que resumen el procesamiento de sus datos y muestra recomendaciones acerca de cómo los datos se pueden modificar o mejorar para establecer la puntuación. Puede revisar y aceptar o rechazar esas recomendaciones.

Figura 4-13
Pestaña análisis de preparación automática de datos



La pestaña Análisis se compone de dos paneles, la vista principal en la parte izquierda y la vista relacionada o auxiliar de la derecha. Hay tres vistas principales:

- Resumen de procesamiento de campos (la configuración por defecto). Si desea obtener más información, consulte el tema Resumen de procesamiento de campo el p. 35.

- Campos. Si desea obtener más información, consulte el tema Campos el p. 36.
- Resumen de acciones. Si desea obtener más información, consulte el tema Resumen de acciones el p. 38.

Hay cuatro vistas relacionadas/auxiliares:

- Poder predictivo (la configuración por defecto). Si desea obtener más información, consulte el tema Poder predictivo el p. 39.
- Tabla de campos. Si desea obtener más información, consulte el tema Tabla de campos el p. 40.
- Detalles de campo. Si desea obtener más información, consulte el tema Detalles de campo el p. 41.
- Detalles de acción. Si desea obtener más información, consulte el tema Detalles de acción el p. 43.

Enlaces entre vistas

En la vista principal, el texto subrayado de las tablas controla la visualización en la vista vinculada. Si pulsa el texto podrá obtener detalles de un campo concreto, conjunto de campos o paso de procesamiento. El enlace que ha seleccionado aparece en color más oscuro; de esta forma podrá identificar la conexión entre el contenido de los dos paneles de vista.

Restablecimiento de las vistas

Para volver a mostrar las recomendaciones de análisis originales y abandonar los cambios que haya realizado en las vistas de análisis, pulse Restablecer en la parte inferior del panel de vista principal.

Resumen de procesamiento de campo

Figura 4-14
Resumen de procesamiento de campo

Resumen de procesamiento de campos

Campos		N
Objetivo		1
Características de entrada		9
Características recomendadas para su uso en análisis	Total	8
	Campos originales (sin transformar)	1
	Transformaciones de campos originales	7
	Derivados de fechas y horas	0
	Construido	0
Características de entrada no utilizadas		1

La tabla Resumen de procesamiento de campos proporciona una instantánea del impacto total previsto de procesamiento, incluyendo los cambios en el estado y el número de características creadas.

Tenga en cuenta que no se crea un modelo realmente, por lo que no existe una medida ni un gráfico del cambio con el poder predictivo total antes y después de la preparación de los datos. Por contra, puede visualizar los gráficos de poder predictivo de los predictores individuales recomendados.

La tabla muestra la siguiente información:

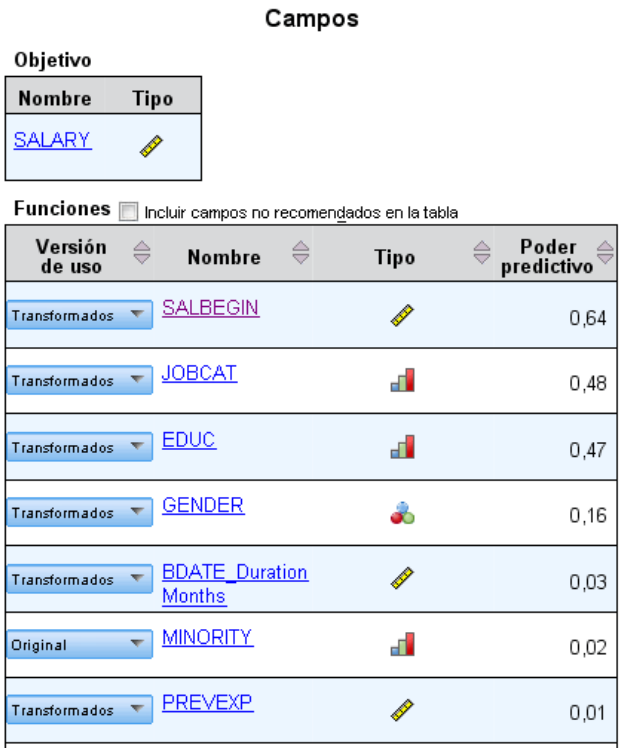
- El número de campos de destino.
- El número de predictores (de entrada) originales.
- Los predictores recomendados para su uso en el análisis y modelado. Incluye el número total de campos recomendados; el número de campos originales sin transformar; campos recomendados; el número de campos transformados recomendados (excluyendo las versiones intermedias de campos, campos derivados de los predictores de fecha y hora y predictores creados); el número de campos recomendados de los campos de fecha/hora; y el número de predictores creados recomendados.
- El número de predictores de entrada no recomendados para su uso en cualquier formulario, ya sea en su formato original, como campo derivado o como entrada en un predictor construido.

Si cualquiera de la información de los Campos está subrayada, pulse para visualizar más detalles en una vista vinculada. Los detalles de Destino, Características de entrada y Características de entrada no utilizadas se muestran en la vista vinculada Tabla de campos. Si desea obtener más

información, consulte el tema Tabla de campos el p. 40. Las características recomendadas para su uso en el análisis se muestran en la vista vinculada Poder predictivo. Si desea obtener más información, consulte el tema Poder predictivo el p. 39.

Campos

Figura 4-15
Campos



La vista principal Campos muestra los campos procesados y si el modo ADP recomienda su uso en modelos posteriores. Puede omitir la recomendación de cualquier campo; por ejemplo, para excluir las características creadas o incluir características que el nodo ADP recomienda excluir. Si un campo se ha transformado, puede decidir si acepta la transformación sugerida o utiliza la versión original.

La vista Campos tiene dos tablas, una para el destino y otra para los predictores procesados o creados.

Tabla Destino

La tabla Destino sólo se muestra si se ha definido un destino en los datos.

La tabla contiene dos columnas:

- **Nombre.** Es el nombre de la etiqueta o del campo de destino; el nombre del original se utiliza siempre, incluso si el campo se ha transformado.
- **Nivel de medida.** Muestra el icono que representa el nivel de medición; pase el ratón por encima del icono para mostrar una etiqueta (continuo, ordinal, nominal, etcétera) que describe los datos.

  Si el destino se ha transformado, la columna Nivel de medición refleja la versión final transformada. *Nota*: no puede desactivar las transformaciones del destino.

Tabla Predictores

La tabla Predictores se muestra siempre. Cada fila de la tabla representa un campo. Por defecto, las filas se clasifican en orden descendente de potencia predictiva.

En características ordinarias, el nombre original siempre se utiliza como el nombre de la fila. Las versiones original y derivada de los campos de fecha/hora aparecen en la tabla (en filas separadas); la tabla también incluye los predictores creados.

Tenga en cuenta que las versiones transformadas de los campos que aparecen en la tabla siempre representan las versiones finales.

Por defecto sólo se muestran los campos recomendados en la tabla Predictores. Para mostrar el resto de campos, seleccione el cuadro Incluir campos no recomendados en la tabla encima de la tabla; estos campos se mostrarán en la parte inferior de la tabla.

La tabla muestra las siguientes columnas:

- **Versión de uso.** Muestra una lista desplegable que controla si un campo se utilizará posteriormente y si se utilizarán las transformaciones sugeridas. Por defecto, la lista desplegable refleja las recomendaciones.

  Para los predictores ordinarios que se han transformado, la lista desplegable tiene tres opciones: Transformada, Original y No utilizar.

  Para los predictores ordinarios sin transformar, las opciones son: Original y No utilizar.

  Para campos derivados de fecha/hora y predictores creados, las opciones son: Transformada y No utilizar.

  Para los campos de fecha originales, la lista desplegable está desactivada y definida a No utilizar.

  Nota: Para predictores con versiones originales y transformados, si cambia entre las versiones Original y Transformadas, se actualiza automáticamente la configuración de Tipo y Poder predictivo de esas características.
- **Nombre.** Cada nombre de campo es un enlace. Pulse en un nombre para ver más información acerca del campo en la vista vinculada. Si desea obtener más información, consulte el tema Detalles de campo el p. 41.

- **Nivel de medida.** Muestra el icono que representa el tipo de datos; pase el ratón por encima del icono para mostrar una etiqueta (continuo, ordinal, nominal, etcétera) que describe los datos.
- **Poder predictivo.** El poder predictivo sólo se muestra en los campos que ADP recomienda Esta columna no se muestra si no hay un destino definido. La potencia predictiva varía de 0 a 1, siendo los mayores valores los que indican "mejores" predictores. En general, la potencia predictiva resulta útil para comparar predictores con un análisis ADP, aunque los valores de potencia predictiva no deben compararse en distintos análisis.

Resumen de acciones

Figura 4-16
Resumen de acción

Resumen de acción

Acción
Campos de texto
Características de fecha y hora
Inspección de características
Tipo de comprobación
Valores atípicos
Valores perdidos
Objetivo
Características categóricas
Características continuas

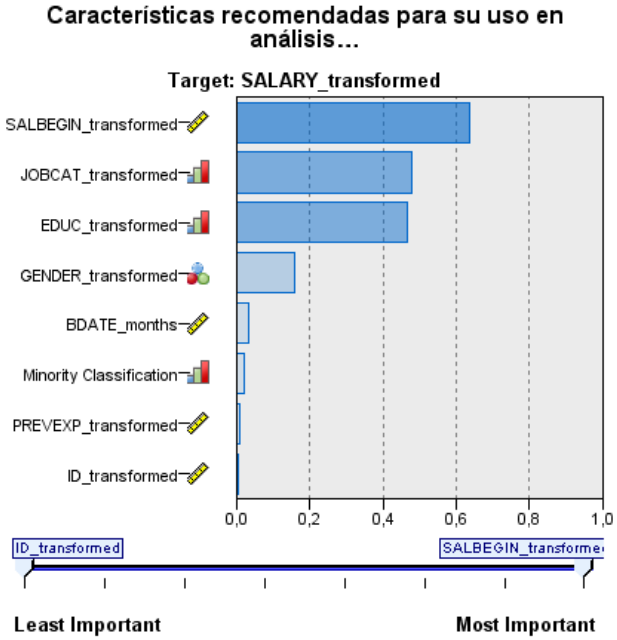
En cada acción realizar por la preparación automática de datos, los predictores de entrada se transforman y/o se filtran; los campos que sobreviven una acción se utilizarán en la acción siguiente. Los campos que sobreviven hasta el último paso se recomiendan para su uso en modelado, mientras que los predictores creados y transformados se filtran.

El Resumen de acciones es una sencilla tabla que enumera las acciones de procesamiento realizadas por ADP. Si alguna Acción está subrayada, pulse para ver más detalles en una vista vinculada sobre las acciones que se realizan. Si desea obtener más información, consulte el tema Detalles de acción el p. 43.

Nota: Sólo se muestran las versiones transformadas originales y finales de cada campo, no las versiones intermedias utilizadas durante el análisis.

Poder predictivo

Figura 4-17
Poder predictivo



Se muestra por defecto cuando el análisis se ejecuta por primera vez o cuando selecciona Los predictores recomendados para su uso en el análisis en la vista principal Resumen del procesamiento de campos, el gráfico muestra el poder predictivo de los predictores recomendados. Los campos se ordenan según el poder predictivo, con el campo de mayor valor en la parte superior.

En versiones transformadas de predictores ordinarios, el nombre del campo refleja su elección del sufijo en el panel Nombres de campos de la pestaña Configuración, por ejemplo: *_transformadas*.

Los iconos de nivel de medición tipo se muestran después de los nombres de campo individuales.

El poder predictivo de cada predictor recomendado, se calcula a partir de una regresión lineal o un modelo de naïve Bayes, dependiendo de si el destino es continuo o categórico.

Tabla de campos

Figura 4-18
Tabla de campos

Características de entrada

Nombre	Tipo
ID	Continuo
GENDER	Establecer
BDATE	Continuo
EDUC	Conjunto ordenado
JOBCAT	Conjunto ordenado
SALBEGIN	Continuo
JOBTIME	Continuo
PREVEXP	Continuo
MINORITY	Conjunto ordenado

Se muestra cuando pulsa en Destino, Predictores o Predictores no utilizados en la vista principal Resumen del procesamiento de campos, la vista Tabla de campos muestra una tabla simple con las características relevantes.

La tabla contiene dos columnas:

- **Nombre.** El nombre del predictor.

  En destinos se utiliza el nombre original o la etiqueta del campo, incluso si el destino se ha transformado.

  En versiones transformadas de predictores ordinarios, el nombre refleja su elección del sufijo en el panel Nombres de campos de la pestaña Configuración, por ejemplo: *_transformadas*.

  En los campos derivados de las fechas y horas se utiliza el nombre de la versión final transformada; por ejemplo: *fnacimiento_años*.

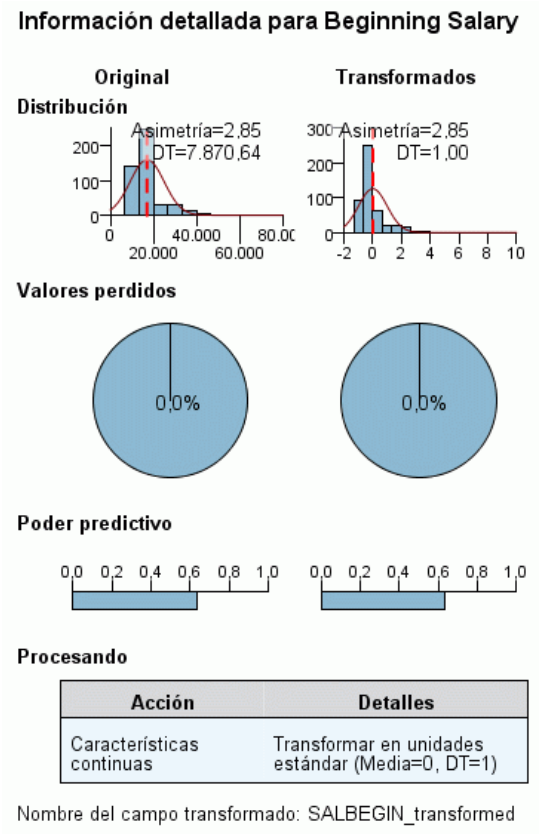
  En los predictores creados, se utiliza el nombre del predictor creado, por ejemplo: *Predictor1*.

- **Nivel de medida.** Muestra el icono que representa el tipo de datos.

  En Destino, el Nivel de medición siempre refleja la versión transformada (si el destino se ha transformado); por ejemplo, si se ha cambiado de ordinal (conjunto ordenado) a continuo (rango, escala) o viceversa.

Detalles de campo

Figura 4-19
Detalles de campo



Se muestra cuando pulsa cualquier Nombre en la vista principal Campos, la vista Detalles de campo contiene distribución, valores perdidos y gráficos de poder predictivo (si procede) del campo seleccionado. Además, también se muestran el historial de procesamiento del campo y el nombre del campo transformado (si procede).

En cada conjunto de gráficos, las dos versiones se muestran juntas para comparar el campo con y sin las transformaciones aplicadas; si no existe una versión transformada del campo, se muestra un gráfico de la versión original únicamente. En los campos de fecha u hora derivados y en los predictores creados, los gráficos sólo se muestran para el nuevo predictor.

Nota: Si se excluye un campo porque tiene demasiadas categorías, solo se muestra el historial de procesamiento.

Gráfico Distribución

La distribución de campos continuos se muestra como una curva normal superpuesta y una línea de referencia vertical para el valor principal; los campos categóricos se muestran como un gráfico de barras.

Los histogramas se etiquetan y muestran la desviación y asimetría típica; sin embargo, la asimetría no se muestra si el número de valores es 2 o menos si la varianza del campo original es inferior a 10-20.

Pase el ratón por encima del gráfico para mostrar la media de los histogramas o el número y el porcentaje del número total de registros para las categorías en gráficos de barras.

Gráfico de valor perdido

Los gráficos de sectores comparan el porcentaje de valores perdidos con y sin transformaciones aplicadas; las etiquetas de gráficos muestran el porcentaje.

Si el nodo ADP ha ejecutado la gestión de valores perdidos, el gráfico de sectores posterior a la transformación también incluye el valor de sustitución como una etiqueta, es decir, el valor que se utiliza en lugar de los valores perdidos.

Pase el ratón por encima del gráfico para mostrar el valor perdido y el porcentaje del número total de registros.

Gráfico de poder predictivo

En los campos recomendados, los gráficos de barras muestran el poder predictivo antes y después de la transformación. Si el destino se ha transformado, el poder predictivo se calcula con respecto al destino transformado.

Nota: Los gráficos de poder predictivo no se muestran si se define el destino o si el pulsa el destino en el panel de vista principal.

Pase el ratón por encima del gráfico para mostrar el valor del poder predictivo.

Tabla Historial de procesamiento

La tabla muestra cómo se ha derivado la versión transformada de un campo. Las acciones que realiza ADP aparecen en el orden en que se ejecutan; sin embargo, en algunos pasos se han realizado varias acciones en un campo concreto.

Nota: Esta tabla no se muestra para los campos que no se han transformado.

La información de la tabla se divide en dos o tres columnas:

- **Acción.** El nombre de la acción. Por ejemplo, Predictores continuos. Si desea obtener más información, consulte el tema Detalles de acción el p. 43.
- **Detalles.** La lista de procesos ejecutados. Por ejemplo, Transformar a unidades estándar.
- **Función.** Sólo se muestra para predictores creados y se muestra la combinación lineal de campos de entrada, por ejemplo, 0,06*edad+ 1,21*altura.

Detalles de acción

Figura 4-20
Análisis ADP: Detalles de acción

Paso 9: Características continuas

Transformación	Número de características	Criterios	
		Media	SD
Transformar en unidades estándar	5	0	1

Construcción de espacio de características	N
Características construidas	0
Características excluidas debido a una asociación baja con el destino	1
Características excluidas por ser constantes tras la agrupación	0

Se muestra cuando selecciona cualquier Acción subrayada en la vista principal Resumen de acciones. La vista vinculada Detalles de acción muestra los datos comunes y específicos de cada paso de procesamiento realizado; los detalles específicos de la acción se muestran primero.

En cada acción, se utiliza la descripción como título en la parte superior de la vista vinculada. Los detalles específicos de la acción se muestran bajo el título y pueden incluir detalles sobre el número de predictores derivados, reestructuración de campo, transformaciones de destinos, categorías fusionadas o reordenadas y predictores creados o excluidos.

A medida que se procesa cada acción, puede cambiar el número de predictores utilizados en el procesamiento, por ejemplo a medida que se excluyen o fusionan los predictores.

Nota: Si se ha desactivado una acción o si no se ha especificado un destino, aparece un mensaje de error en lugar de los detalles de la acción cuando pulsa la acción en la vista principal Resumen de acciones.

Existen nueve acciones disponibles; sin embargo, no todas están necesariamente activas para cada análisis.

Tabla Campos de texto

La tabla muestra el número de:

- Predictores excluidos del análisis.

Tabla Predictores de fecha y hora

La tabla muestra el número de:

- Duraciones de los predictores de fecha y hora.
- Elementos de fecha y hora.
- Predictores derivados de fecha y hora, en total.

La fecha u hora de referencia se muestra como nota al pie si se han calculado algunas de las duraciones de fecha.

Tabla Filtrado de predictores

La tabla muestra el número de los siguientes predictores excluidos del procesamiento:

- Constantes.
- Predictores con demasiados valores perdidos.
- Predictores con demasiados casos en una única categoría.
- Campos nominales (conjuntos) con demasiadas categorías.
- Predictores cribados, en total.

Tabla Comprobar nivel de medición

La tabla muestra los números de reestructuración de campos, que se dividen en:

- Reestructuración de campos ordinales (conjuntos ordenados) como campos continuos.
- Los campos continuos se redistribuyen como ordinales
- Reestructuración de números totales.

Si los campos de entrada (destinos o predictores) no eran conjuntos continuos u ordeninales, se muestra como nota al pie.

Tabla Valores atípicos

La tabla muestra cómo se han tratado los valores atípicos.

- El número de campos continuos donde se han encontrado y suprimido valores atípicos o el número de campos continuos donde se han encontrado valores atípicos y se han definido como perdidos, dependiendo de su configuración en el panel Preparar entradas y destino en la pestaña Configuración.
- Se ha excluido el número de campos continuos porque eran constantes, después del tratamiento de los valores atípicos.

Una nota al pie muestra el valor de corte atípico; mientras que se muestra otra nota al pie si no hay campos de entrada continuos (destino o predictores).

Tabla Valores perdidos

La tabla muestra el número de campos con valores perdidos sustituidos y desglosados en:

- Objetivo. Esta fila no se muestra si no se han especificado destinos.
- Predictores. Pueden desglosarse por el número de nominales (conjunto), ordinales (conjunto ordenado) y continuas.
- El número total de valores perdidos sustituidos.

Tabla Destino

La tabla muestra si se ha transformado el destino, que se muestra como:

- Transformación de Box-Cox a normalidad. Se desglosa a su vez en columnas que muestran los criterios especificados (media y la desviación típica) y Lambda.
- Categorías de destino reordenadas para mejorar la estabilidad.

Tabla Predictores categóricos

La tabla muestra el número de predictores categóricos:

- Las categorías se reordenan de menor a mayor para mejorar la estabilidad.
- Características cuyas categorías se han fusionado para aumentar al máximo su asociación con el destino.
- Características cuyas categorías se han fusionado para tratar categorías dispersas.
- Características cuyas categorías se han excluido por su asociación baja con el destino.
- Características cuyas categorías se han excluido porque eran constantes después de la fusión.

Se muestra una nota al pie si no se han introducido predictores categóricos.

Tabla Predictores continuos

Hay dos tablas. La primera muestra uno de los siguientes números de transformaciones:

- Valores de predictores transformados a unidades estándar. Además, muestra el número de predictores transformados, la media especificada y la desviación estándar.
- Valores de predictores asignados a un rango común. Además, muestra el número de predictores transformados utilizando una transformación mínima-máxima, así como los valores mínimo y máximo especificados.
- Valores de predictores en intervalos y el número de predictores en intervalos.

La segunda tabla muestra los detalles de creación de predictores, mostrados como el número de predictores:

- Construido.
- Características cuyas categorías se han excluido por su asociación baja con el destino.

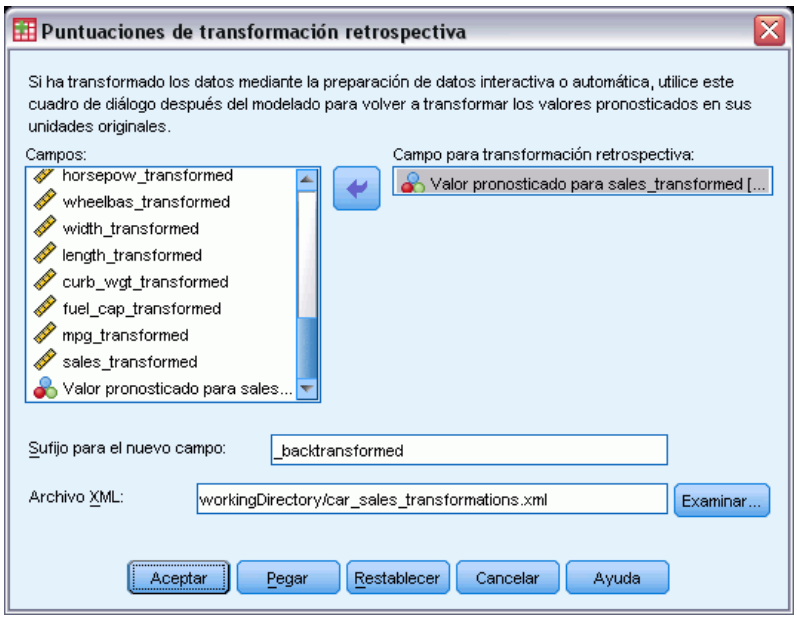
- Características cuyas categorías se han excluido porque eran constantes después de la agrupación.
- Excluido por ser constante tras la construcción.

Se muestra una nota al pie si no se han introducido predictores continuos.

Puntuaciones de transformación retrospectiva

Si ADP transforma un destino, los siguientes modelos generados utilizando el destino transformado puntúan las unidades transformadas. Para interpretar y utilizar los resultados, debe convertir el valor pronosticado a la escala original.

Figura 4-21
Puntuaciones de transformación retrospectiva



Para aplicar transformación retrospectiva a las puntuaciones, seleccione en los menús:
Transformar > Preparar datos para modelado > Puntuaciones de transformación retrospectiva...

▶ Seleccione un campo para aplicar la transformación retrospectiva. Este campo debe contener valores pronosticados por el modelo del destino transformado.

▶ Especifique un sufijo para el nuevo campo. Este nuevo campo contendrá valores pronosticados por el modelo en la escala original del destino sin transformar.

▶ Especifique la ubicación del archivo XML que contiene las transformaciones ADP. Debe ser un archivo guardado en los cuadros de diálogo de preparación automática de datos o interactiva. Si desea obtener más información, consulte el tema Aplicación y almacenamiento de transformaciones el p. 31.

Capítulo
5

Identificar casos atípicos

El procedimiento de detección de anomalías busca casos atípicos basados en desviaciones de las normas de sus agrupaciones. El procedimiento está diseñado para detectar rápidamente casos atípicos con fines de auditoría de datos en el paso del análisis exploratorio de datos, antes de llevar a cabo cualquier análisis de datos inferencial. Este algoritmo está diseñado para la detección de anomalías genéricas; es decir, la definición de un caso anómalo no es específica de ninguna aplicación particular, como la detección de patrones de pago atípicos en la industria sanitaria ni la detección de blanqueo de dinero en la industria financiera, donde la definición de una anomalía puede estar bien definida.

Ejemplo. Un analista de datos contratado para generar modelos predictivos para los resultados de los tratamientos de derrames cerebrales se preocupa por la calidad de los datos ya que tales modelos pueden ser sensibles a observaciones atípicas. Algunas de estas observaciones atípicas representan casos verdaderamente únicos y, por lo tanto, no son adecuadas para la predicción, mientras que otras observaciones están provocadas por errores de entrada de datos donde los valores son técnicamente "correctos" y no pueden ser detectados por los procedimientos de validación de datos. El procedimiento Identificar casos atípicos busca y realiza un informe de estos valores atípicos de forma que el analista pueda decidir cómo tratarlos.

Estadísticos. El procedimiento genera grupos de homólogos, normas de grupos de homólogos para las variables continuas y categóricas, índices de anomalías basados en las desviaciones de las normas de los grupos de homólogos y valores del impacto de las variables para las variables que contribuyen en mayor medida a que el caso se considere atípico.

Consideraciones de los datos

Datos. Este procedimiento trabaja tanto con variables continuas como categóricas. Cada fila representa una observación distinta y cada columna representa una variable distinta en la que se basan los grupos de homólogos. Puede haber una variable de identificación de casos disponible en el archivo de datos para marcar los resultados, pero no se utilizará para el análisis. Los valores perdidos están disponibles. Si se especifica la variable de ponderación, se ignorará.

El modelo de detección puede aplicarse a un archivo de datos de prueba nuevo. Los elementos de los datos de prueba deben ser los mismos que los elementos de los datos de entrenamiento. Además, dependiendo de la configuración del algoritmo, el tratamiento de los valores perdidos que se utiliza para crear el modelo puede aplicarse al archivo de datos de prueba antes de la puntuación.

Orden de casos. Tenga en cuenta que la solución puede depender del orden de los casos. Para minimizar los efectos del orden, ordene los casos aleatoriamente. Para comprobar la estabilidad de una solución dada, puede obtener varias soluciones distintas con los casos ordenados en distintos órdenes aleatorios. En situaciones con tamaños de archivo extremadamente grandes,

se pueden llevar a cabo varias ejecuciones con una muestra de casos ordenados con distintos órdenes aleatorios.

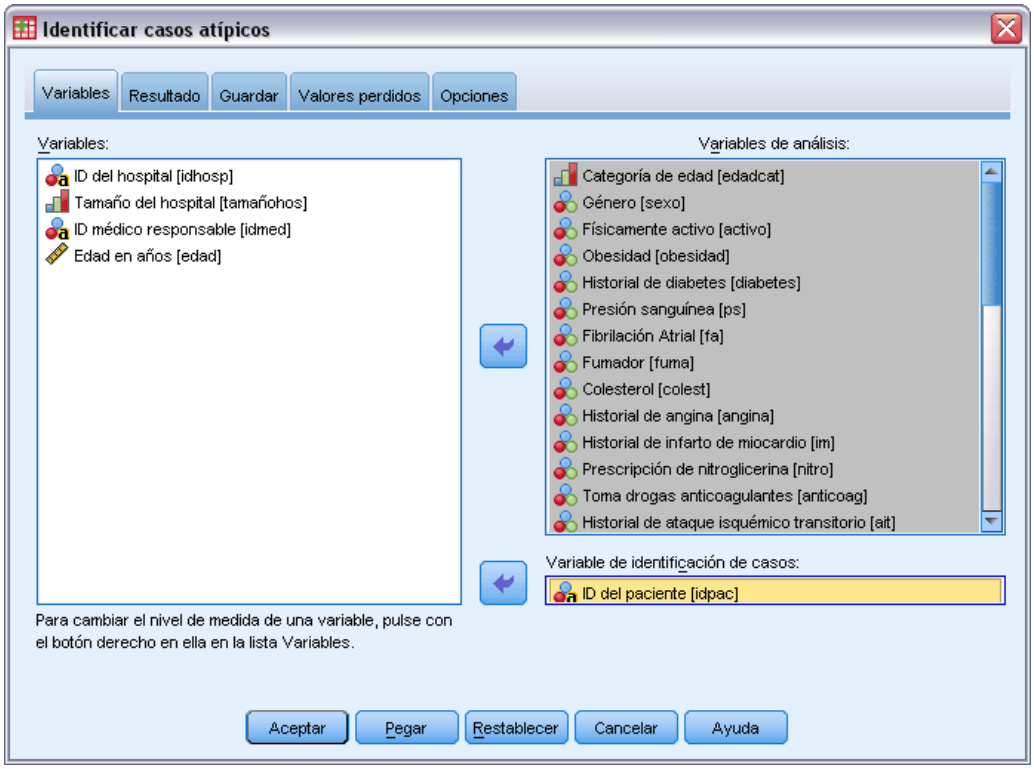
Supuestos. El algoritmo presupone que todas las variables son no constantes e independientes y que ningún caso tiene valores perdidos para ninguna de las variables de entrada. Se supone que cada variable continua tiene una distribución normal (de Gauss) y que cada variable categórica tiene una distribución multinomial. Las comprobaciones empíricas internas indican que este procedimiento es bastante robusto frente a las violaciones tanto del supuesto de independencia como de las distribuciones, pero se debe tener en cuenta hasta qué punto se cumplen estos supuestos.

Para identificar casos atípicos

- Seleccione en los menús:

  Datos > Identificar casos atípicos...

Figura 5-1
Cuadro de diálogo Identificar casos atípicos, pestaña Variables

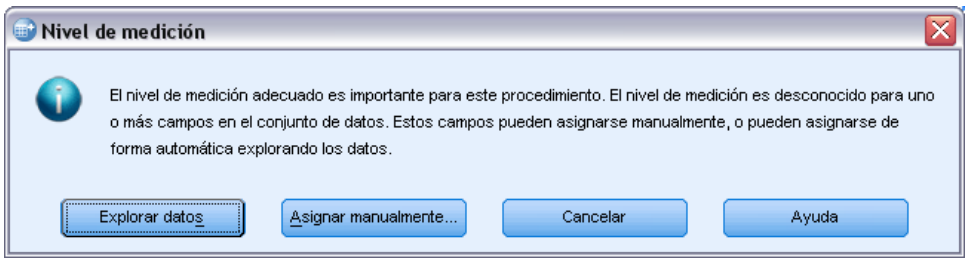


- Seleccione al menos una variable de análisis.
- Si lo desea, seleccione una variable identificadora de caso para utilizarla para etiquetar los resultados.

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 5-2
Alerta de nivel de medición

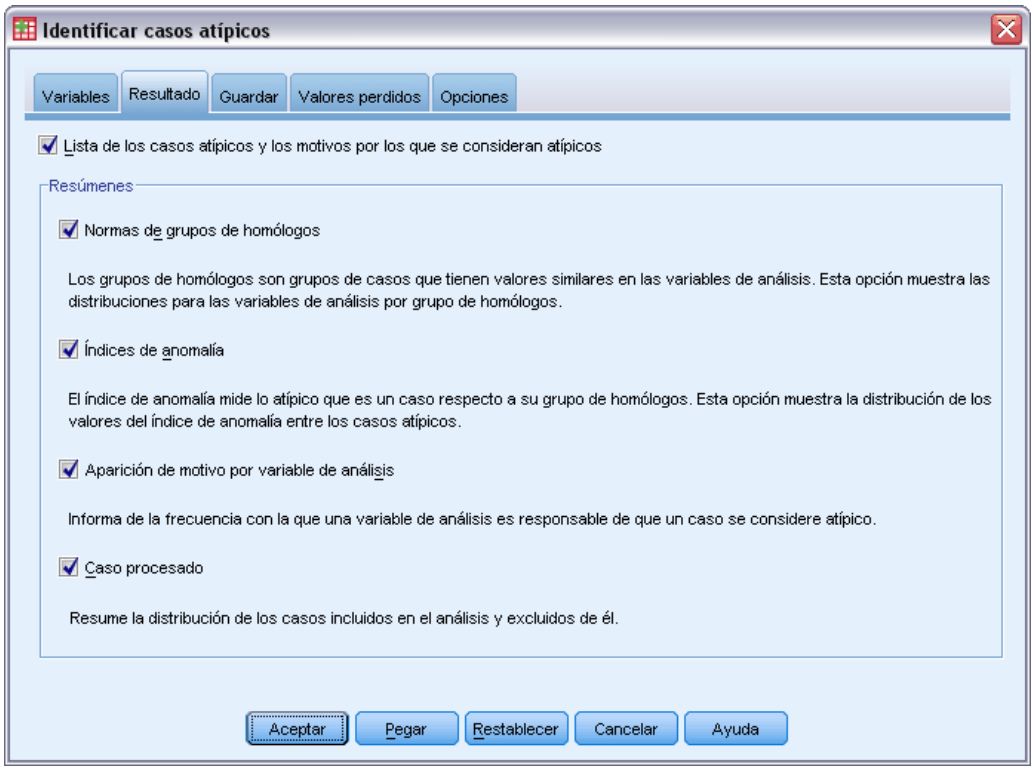


- **Explorar datos.** Lee los datos del conjunto de datos activo y asigna el nivel de medición predefinido en cualquier campo con un nivel de medición desconocido. Si el conjunto de datos es grande, puede llevar algún tiempo.
- **Asignar manualmente.** Abre un cuadro de diálogo que contiene todos los 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.

Identificar casos atípicos: Resultados

Figura 5-3
Cuadro de diálogo Identificar casos atípicos, pestaña Resultado



Lista de casos atípicos y motivos por los que se consideran atípicos. Esta opción produce tres tablas:

- La lista de índice de los casos con anomalías muestra los casos que se identifican como atípicos así como sus valores correspondientes del índice de anomalía.
- La lista de identificadores de los homólogos de los casos con anomalías muestra los casos atípicos e información sobre sus grupos de homólogos correspondientes.
- La lista de motivos de anomalías muestra el número de caso, la variable motivo, el valor de impacto de la variable, el valor de la variable y la norma de la variable de cada motivo.

Todas las tablas se ordenan por índice de anomalía en orden descendente. Además, los identificadores de los casos se muestran si la variable de identificación de caso está especificada en la pestaña Variable.

Resúmenes. Los controles de este grupo generan resúmenes de distribución.

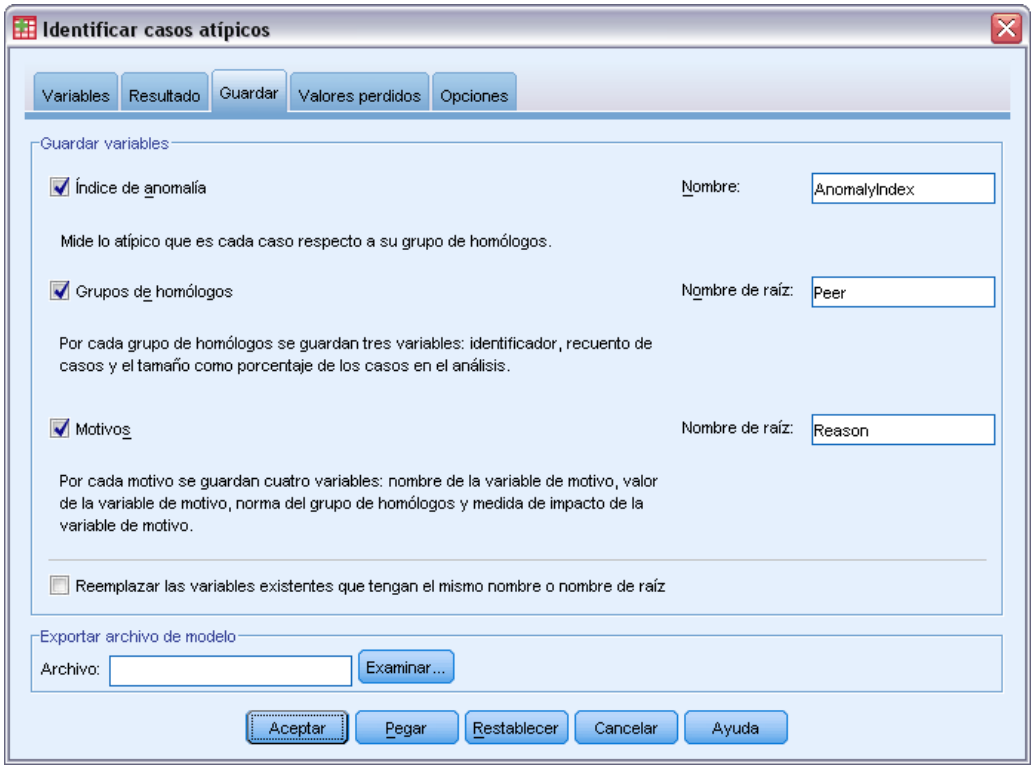
- **Normas de grupos de homólogos.** Esta opción muestra la tabla de normas de las variables continuas (si se utiliza alguna variable continua en el análisis) y la tabla de normas de las variables categóricas (si se utiliza alguna variable categórica en el análisis). La tabla de normas de las variables continuas muestra la media y la desviación típica de cada variable continua para cada grupo de homólogos. La tabla de normas de las variables categóricas muestra la moda (categoría más popular), su frecuencia y el porcentaje de frecuencia de cada

variable categórica para cada grupo de homólogos. En el análisis se utilizan como los valores de norma la media cuando una variable continua y la moda cuando una variable categórica.

- **Índices de anomalía.** El resumen de índice de anomalía muestra estadísticos descriptivos para el índice de anomalía de los casos que se identifican como los más atípicos.
- **Aparición de motivo por variable de análisis.** Para cada motivo, la tabla muestra la frecuencia y el porcentaje de frecuencia de cada aparición de la variable como un motivo. La tabla también informa sobre los estadísticos descriptivos del impacto de cada variable. Si el número máximo de motivos está establecido en 0 en la pestaña Opciones, esta opción no estará disponible.
- **Casos procesados.** El resumen de procesamiento de casos muestra los recuentos y los porcentajes de recuento de todos los casos del conjunto de datos activo, los casos incluidos y excluidos del análisis, y los casos de cada grupo de homólogos.

Identificar casos atípicos: Guardar

Figura 5-4
Cuadro de diálogo Identificar casos atípicos, pestaña Guardar



Guardar variables. Los controles de este grupo permiten guardar las variables del modelo en el conjunto de datos activo. También puede sustituir las variables existentes cuyos nombres entran en conflicto con las variables que se van a guardar.

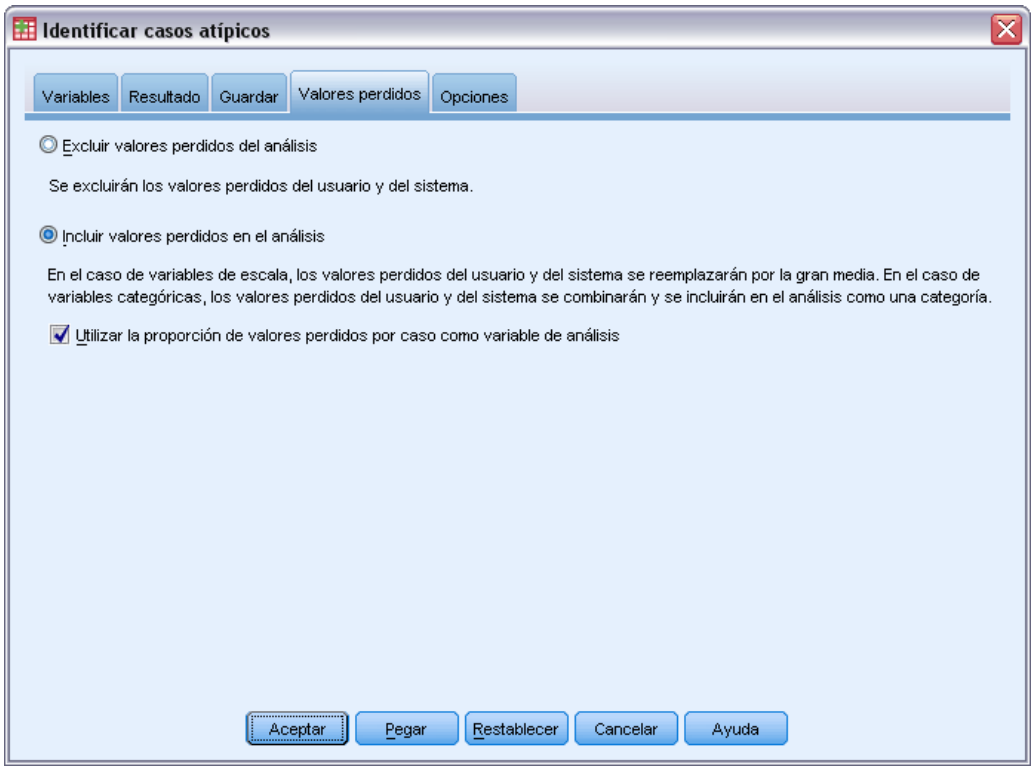
- **Índice de anomalía.** Guarda el valor del índice de anomalía de cada caso en una variable con el nombre especificado.

- **Grupos de homólogos.** Guarda el ID, el recuento de casos y el tamaño del grupo de homólogos como porcentaje de cada caso en las variables con el nombre raíz especificado. Por ejemplo, si se especifica el nombre raíz *Homólogo*, se generarán las variables *HomólogoID*, *HomólogoTam* y *HomólogoPcTam*. *HomólogoID* es el ID del grupo de homólogos del caso, *HomólogoTam* es el tamaño del grupo y *HomólogoPcTam* es el tamaño del grupo como porcentaje.
- **Motivos.** Guarda conjuntos de variables de motivos con el nombre raíz especificado. Un conjunto de variables de motivos consta del nombre de la variable como el motivo, la medida del impacto de la variable, su propio valor y el valor de la norma. El número de conjuntos depende del número de motivos solicitados en la pestaña Opciones. Por ejemplo, si se especifica el nombre de raíz *Reason*, se generarán las variables *ReasonVar_k*, *ReasonMeasure_k*, *ReasonValue_k* y *ReasonNorm_k*, donde *k* es el motivo *k*ésimo. Esta opción no está disponible si el número de motivos está establecido en 0.

Exportar archivo de modelo. Permite guardar el modelo en formato XML.

Identificar casos atípicos: Valores perdidos

Figura 5-5
Cuadro de diálogo Identificar casos atípicos, pestaña Valores perdidos

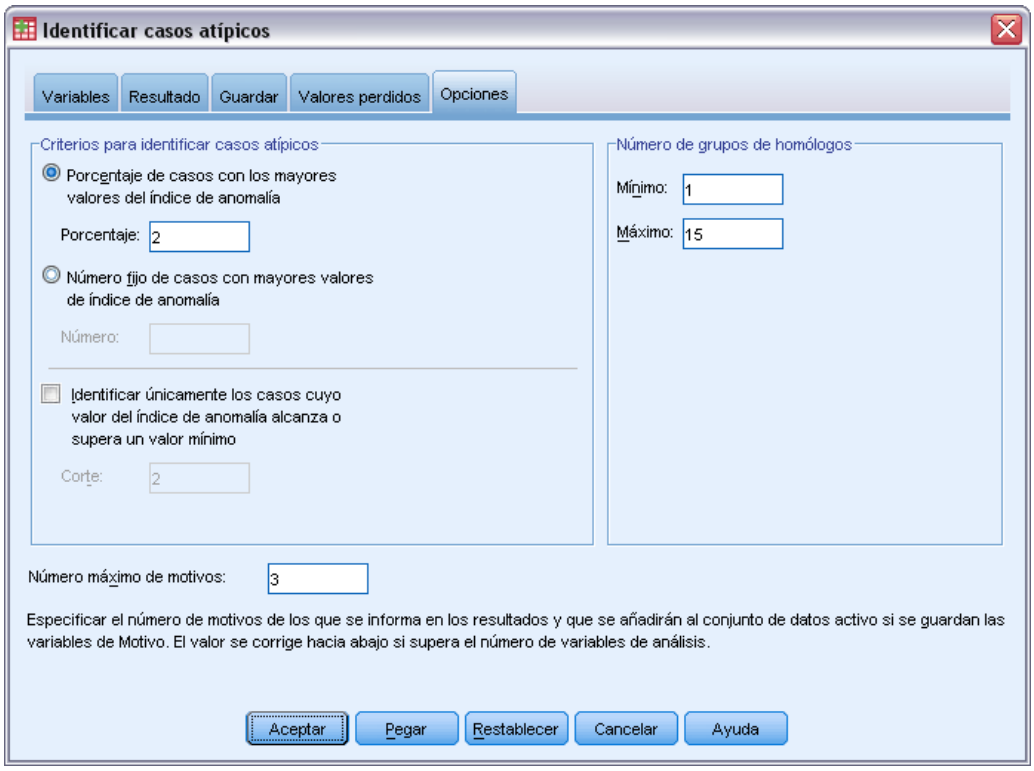


La pestaña Valores perdidos se utiliza para controlar el tratamiento de los valores definidos como perdidos por el usuario y los valores perdidos del sistema.

- **Excluir valores perdidos del análisis.** Los casos con valores perdidos se excluyen del análisis.
- **Incluir valores perdidos en el análisis.** Los valores perdidos de variables continuas se sustituyen por sus medias globales correspondientes y las categorías perdidas de las variables categóricas se agrupan y tratan como una categoría válida. A partir de ese momento, las variables que se han procesado se utilizan en el análisis. Si lo desea, puede solicitar la creación de una variable adicional que represente la proporción de variables perdidas en cada caso y utilizar esa variable en el análisis.

Identificar casos atípicos: Opciones

Figura 5-6
Cuadro de diálogo Identificar casos atípicos, pestaña Opciones



Criterios para identificar casos atípicos. Estas selecciones determinan cuántos casos se incluyen en la lista de anomalías.

- **Porcentaje de casos con los mayores valores del índice de anomalía.** Especifique un número positivo menor o igual que 100.
- **Número de casos fijo con los mayores valores de índice de anomalía.** Especifique un número entero positivo que sea menor o igual que el número total de casos del conjunto de datos activo que se ha utilizado en el análisis.
- **Identificar únicamente los casos cuyo valor del índice de anomalía alcanza o supera un valor mínimo.** Especifique un número que no sea negativo. Un caso se considera anómalo si su valor de índice de anomalía es mayor o igual que el punto de corte especificado. Esta opción se utiliza junto con las opciones Porcentaje de casos y Número fijo de casos. Por

ejemplo, si especifica un número de 50 casos y un valor de punto de corte de 2, la lista de anomalías constará de un máximo de 50 casos, cada uno con un valor del índice de anomalía mayor o igual que 2.

Número de grupos de homólogos. El procedimiento buscará el mejor número de grupos de homólogos entre los valores mínimo y máximo especificados. Los valores deben ser números enteros positivos y el mínimo no debe superar al máximo. Cuando los valores especificados son iguales, el procedimiento presupone un número fijo de grupos de homólogos.

Nota: Dependiendo de la cantidad de variación de los datos, puede haber situaciones en las que el número de grupos de homólogos que los datos pueden admitir sea menor que el número especificado como mínimo. En tal situación, el procedimiento puede generar un número menor de grupos de homólogos.

Número máximo de motivos. Un motivo consta de la medida del impacto de la variable, el nombre de la variable para este motivo, el valor de la variable y el valor del grupo de homólogos correspondiente. Especifique un número entero no negativo; si este valor supera o es igual que el número de variables que se han procesado y se han utilizado en el análisis, se mostrarán todas las variables.

Funciones adicionales del comando DETECTANOMALY

La sintaxis de comandos también le permite:

- Omitir algunas variables del conjunto de datos activo del análisis sin especificar explícitamente todas las variables del análisis (mediante el subcomando `EXCEPT`).
- Especificar una corrección para equilibrar la influencia de las variables continuas y categóricas (mediante la palabra clave `MLWEIGHT` del subcomando `CRITERIA`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo
6

Intervalos óptimos

El procedimiento Intervalos óptimos discretiza una o más variables de escala (a las que denominaremos en lo sucesivo **variables de entrada que se van a agrupar**) mediante la distribución de los valores de cada variable en intervalos. La formación de intervalos es óptima en relación con una variable guía categórica que "supervisa" el proceso de agrupación. Los intervalos se pueden utilizar en lugar de los valores de datos originales para posteriores análisis.

Ejemplos. La reducción del número de valores distintos que puede tomar una variable tiene varios usos, entre los que se incluyen:

- Requisitos de los datos de otros procedimientos. Las variables discretizadas pueden tratarse como categóricas y utilizarse en procedimientos que requieren variables categóricas. Por ejemplo, el procedimiento Tablas de contingencia requiere que todas las variables sean categóricas.
- Privacidad de los datos. Utilizar en los informes los valores agrupados en vez de los valores reales puede ayudar a proteger la privacidad de los orígenes de los datos. El procedimiento Intervalos óptimos puede ayudarle a elegir los intervalos adecuados.
- Agilización del rendimiento. Algunos procedimientos son más eficientes cuando trabajan con un número reducido de valores distintos. Por ejemplo, la velocidad de la regresión logística multinomial puede incrementarse utilizando variables discretizadas.
- Detección de la separación completa o quasi-completa de los datos.

Intervalos óptimos frente al agrupador visual Los cuadros de diálogo de Agrupación visual ofrecen varios métodos automáticos para crear intervalos sin utilizar una variable como guía. Estas reglas "no supervisadas" son útiles para generar estadísticos descriptivos, como tablas de frecuencia, pero Intervalos óptimos es superior cuando el objetivo final es generar un modelo predictivo.

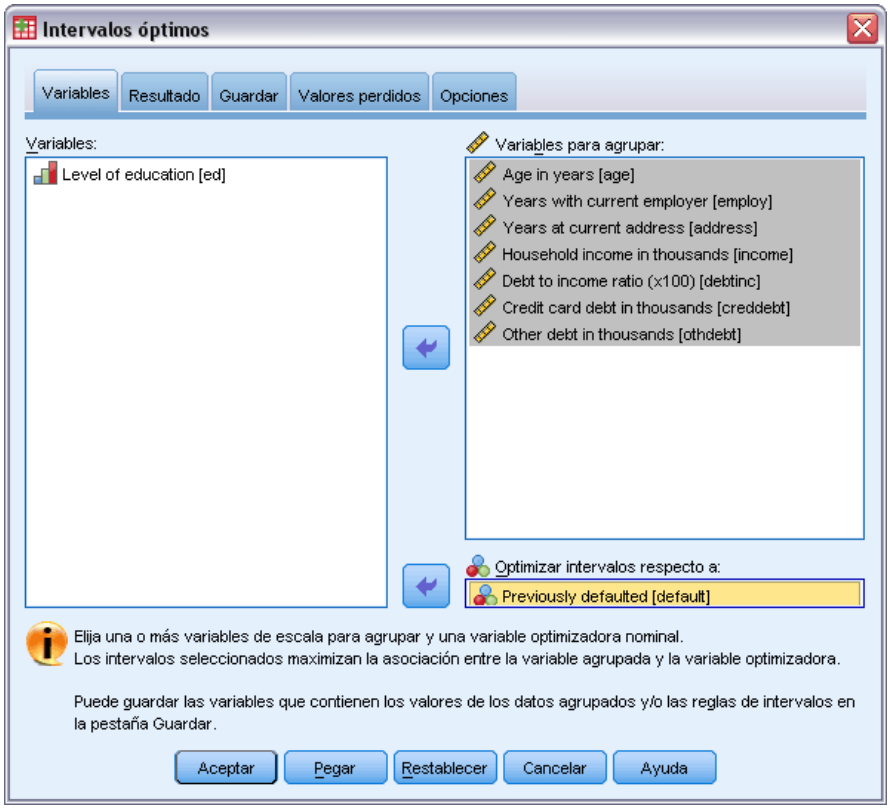
Resultados. El procedimiento genera tablas de puntos de corte para los intervalos y los estadísticos descriptivos de cada una de las variables de entrada que se van a agrupar. Además, puede guardar nuevas variables en el conjunto de datos activo que contengan los valores agrupados de las variables de entrada que se han agrupado, así como guardar las reglas de agrupación como sintaxis de comandos para utilizarlas al discretizar nuevos datos.

Datos. Este procedimiento espera que las variables de entrada que se van a agrupar sean variables numéricas de escala. La variable guía debe ser categórica y puede ser de cadena o numérica.

Para obtener intervalos óptimos

En los menús, seleccione:
Transformar > Intervalos óptimos...

Figura 6-1
Cuadro de diálogo Intervalos óptimos, pestaña Variables

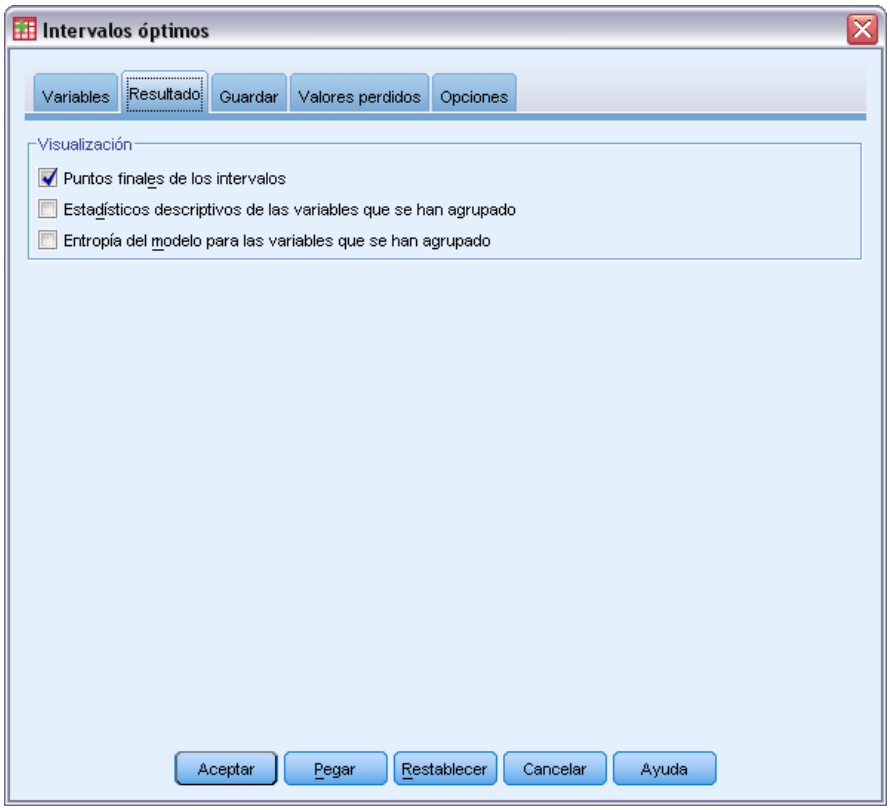


- Seleccione una o más variables de entrada para agruparlas.
- Seleccione una variable guía.

Las variables que contienen los valores de los datos agrupados no se generan por defecto. Utilice la pestaña Guardar para guardar estas variables.

Intervalos óptimos: Resultado

Figura 6-2
Cuadro de diálogo Intervalos óptimos, pestaña Resultado

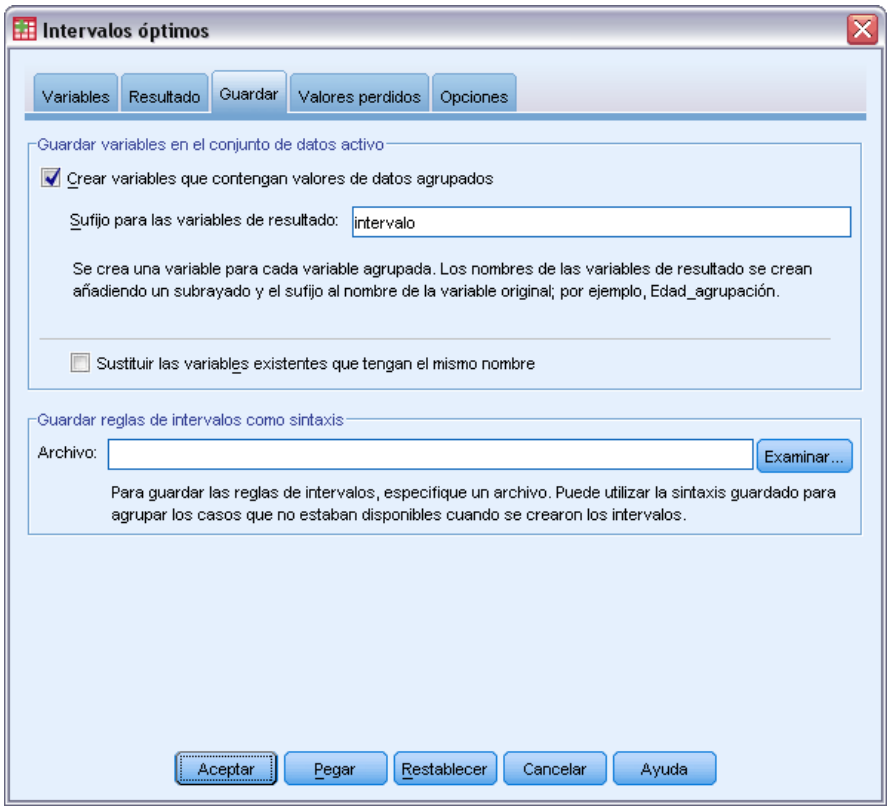


La pestaña Resultados controla la presentación de los resultados.

- **Puntos finales de los intervalos.** Muestra el conjunto de puntos finales de cada variable de entrada que se va a agrupar.
- **Estadísticos descriptivos de las variables que se han agrupado.** Para cada variable de entrada que se ha agrupado, esta opción muestra el número de casos con valores válidos, el número de casos con valores perdidos, el número de valores válidos distintos y los valores mínimo y máximo. Para la variable guía, esta opción muestra la distribución de clase para cada variable de entrada relacionada que se ha agrupado.
- **Entropía del modelo para las variables que se han agrupado.** Para cada variable de entrada que se ha agrupado, esta opción muestra una medida de la precisión predictiva de la variable respecto a la variable guía.

Intervalos óptimos: Guardar

Figura 6-3
Cuadro de diálogo Intervalos óptimos, pestaña Guardar

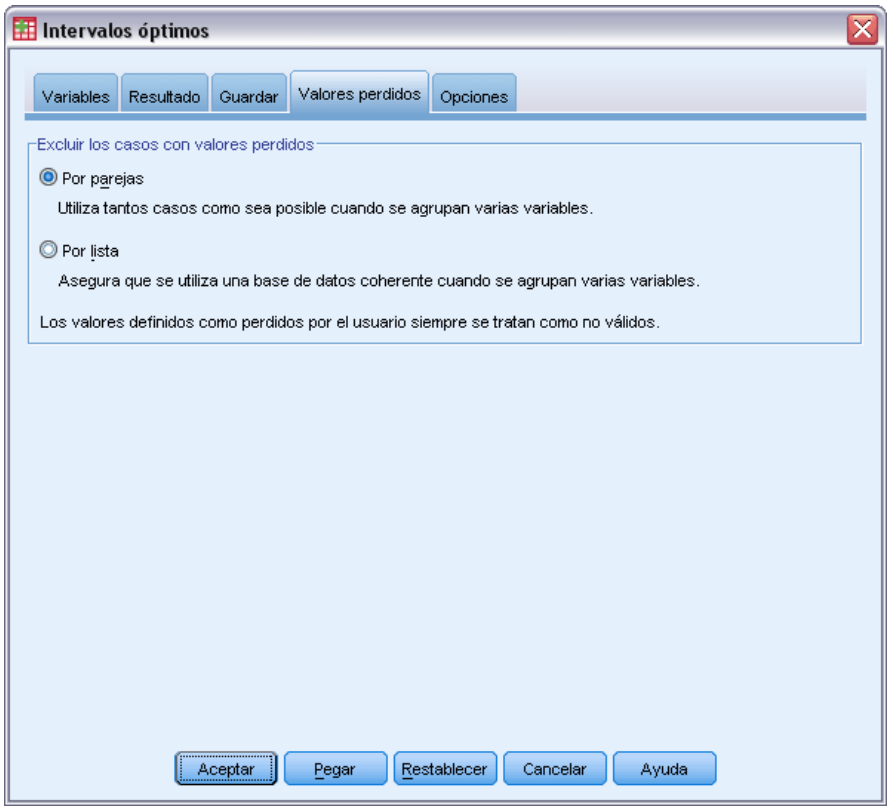


Guardar variables en el conjunto de datos activo. Las variables que contienen los valores de los datos que se han agrupado se pueden utilizar en lugar de las variables originales en análisis posteriores.

Guardar reglas de intervalos como sintaxis de . Genera una sintaxis de comandos que se puede utilizar para agrupar otros conjuntos de datos. Las reglas de recodificación se basan en los puntos de corte determinados por el algoritmo de agrupación.

Intervalos óptimos: Valores perdidos

Figura 6-4
Cuadro de diálogo Intervalos óptimos, pestaña Valores perdidos

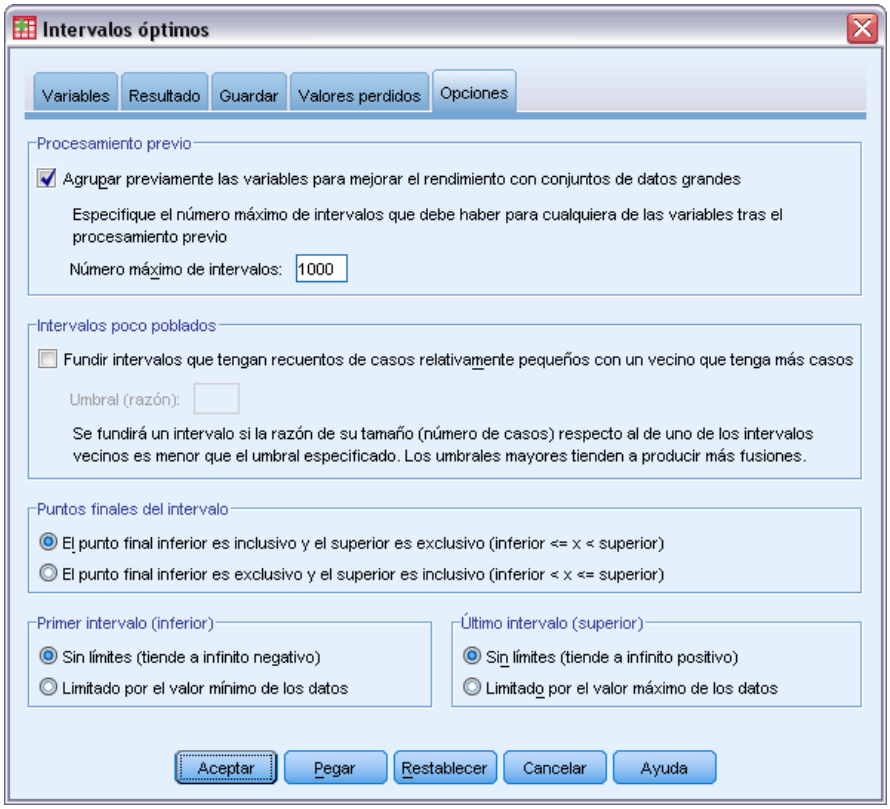


La pestaña Valores perdidos especifica si los valores perdidos se tratarán utilizando eliminación por lista o por parejas. Los valores definidos como perdidos por el usuario siempre se tratan como no válidos. Al recodificar los valores de la variable original en una nueva variable, los valores definidos como perdidos por el usuario se convierten en valores perdidos del sistema.

- **Por parejas.** Esta opción actúa sobre cada par de variables de entrada que se va a agrupar y variable guía. El procedimiento utilizará todos los casos con valores que no sean perdidos en la variable guía y la variable de entrada que se va a agrupar.
- **Por lista** Esta opción actúa sobre todas las variables especificadas en la pestaña Variables. Si algún caso tiene un valor perdido para una variable, se excluirá el caso completo.

Intervalos óptimos: opciones

Figura 6-5
Cuadro de diálogo Intervalos óptimos, pestaña Opciones



Procesamiento previo. La "agrupación previa" de las variables de entrada que se van a agrupar con numerosos valores distintos puede reducir el tiempo de procesamiento sin reducir demasiado la calidad de los intervalos finales. El número máximo de intervalos constituye un límite superior del número de intervalos que se han creado. Por tanto, si especifica 1000 como máximo pero una variable de entrada que se va a agrupar tiene menos de 1000 valores distintos, el número de intervalos preprocesados creados para la variable de entrada que se va a agrupar será igual al número de valores distintos de la variable de entrada que se va a agrupar.

Intervalos poco poblados. En ocasiones, el procedimiento puede generar intervalos con muy pocos casos. La siguiente estrategia elimina estos pseudo puntos de corte:

- Para una determinada variable, supongamos que el algoritmo ha encontrado n_{final} puntos de corte y, por consiguiente, $n_{\text{final}}+1$ intervalos. Para los intervalos $i = 2, ..., n_{\text{final}}$ (desde el segundo intervalo con valores inferiores hasta el segundo intervalo con valores superiores), se calcula

$$\frac{sizeof(b_i)}{mn(sizeof(b_{i-1}),sizeof(b_{i+1}))}$$

donde *tamañode(b)* es el número de casos del intervalo.

- Cuando este valor es menor que el umbral de fusión especificado, b_i se considera poco poblado y se funde con b_{i-1} o b_{i+1}, cualquiera que tenga la entropía de información de clase inferior.

El procedimiento realiza una única pasada a través de los intervalos.

Puntos finales del intervalo. Esta opción especifica cómo se define el límite inferior de un intervalo. Como el procedimiento determina automáticamente los valores de los puntos de corte, es básicamente una cuestión de gustos.

Primer intervalo (inferior) / Último intervalo (superior). Estas opciones especifican cómo se definen los puntos de corte mínimo y máximo para cada variable de entrada que se va a agrupar. En general, el procedimiento supone que las variables de entrada que se van a agrupar pueden tomar cualquier valor de la línea de números reales, pero si tiene algún motivo práctico o teórico para acotar el intervalo, puede limitarlo especificando los valores mínimo y máximo.

Funciones adicionales del comando OPTIMAL BINNING

Con el lenguaje de sintaxis de comandos también podrá:

- Realizar la agrupación no supervisada mediante el método de frecuencias iguales (utilizando el subcomando `CRITERIA`).

Si desea información detallada sobre la sintaxis, consulte la referencia de sintaxis de comandos (*Command Syntax Reference*).

Parte II: Ejemplos

Capítulo
7

Validar datos

El procedimiento Validar datos permite identificar casos, variables y valores de datos no válidos y sospechosos.

Validación de una base de datos médica

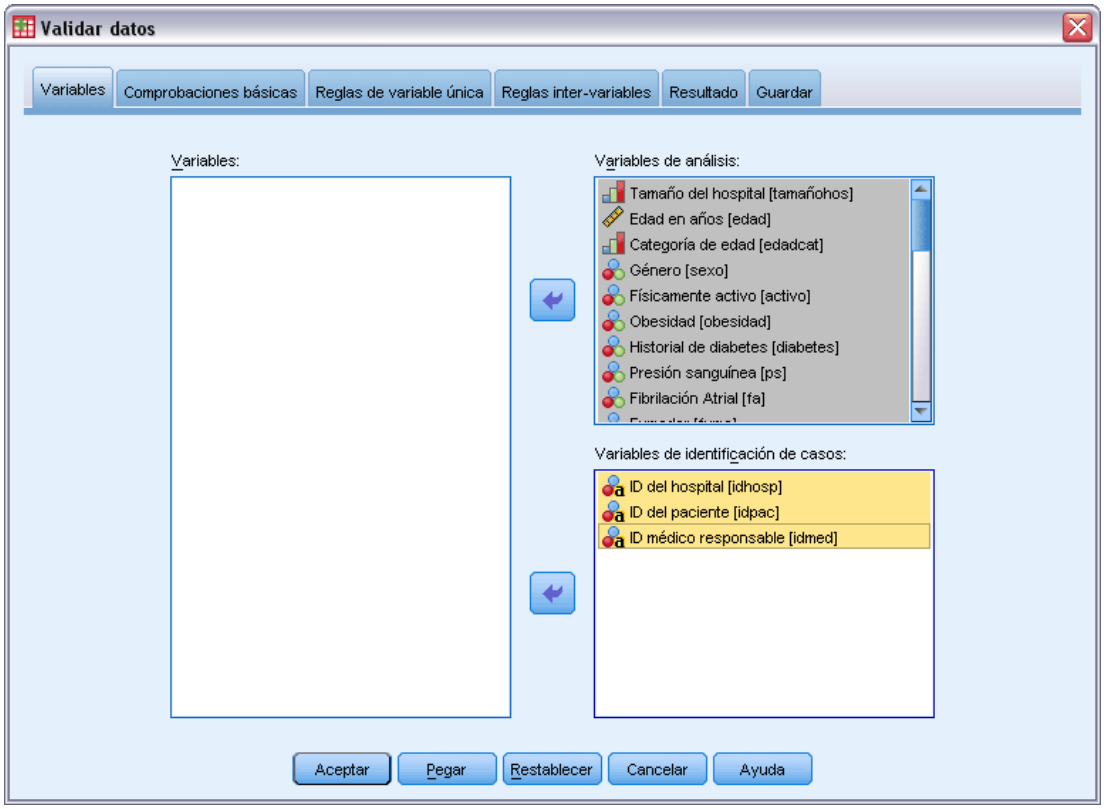
Un analista contratado por un grupo médico está encargado de mantener la calidad de la información del sistema. Este proceso implica comprobar los valores y variables, así como preparar un informe para el administrador del equipo de introducción de datos.

El estado más reciente de la base de datos está recopilado en *stroke_invalid.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A el p. 138. Utilice el procedimiento Validar datos para obtener la información necesaria para generar el informe. Puede encontrar la sintaxis para reproducir estos análisis en *validatedata_stroke.sps*.

Comprobaciones básicas

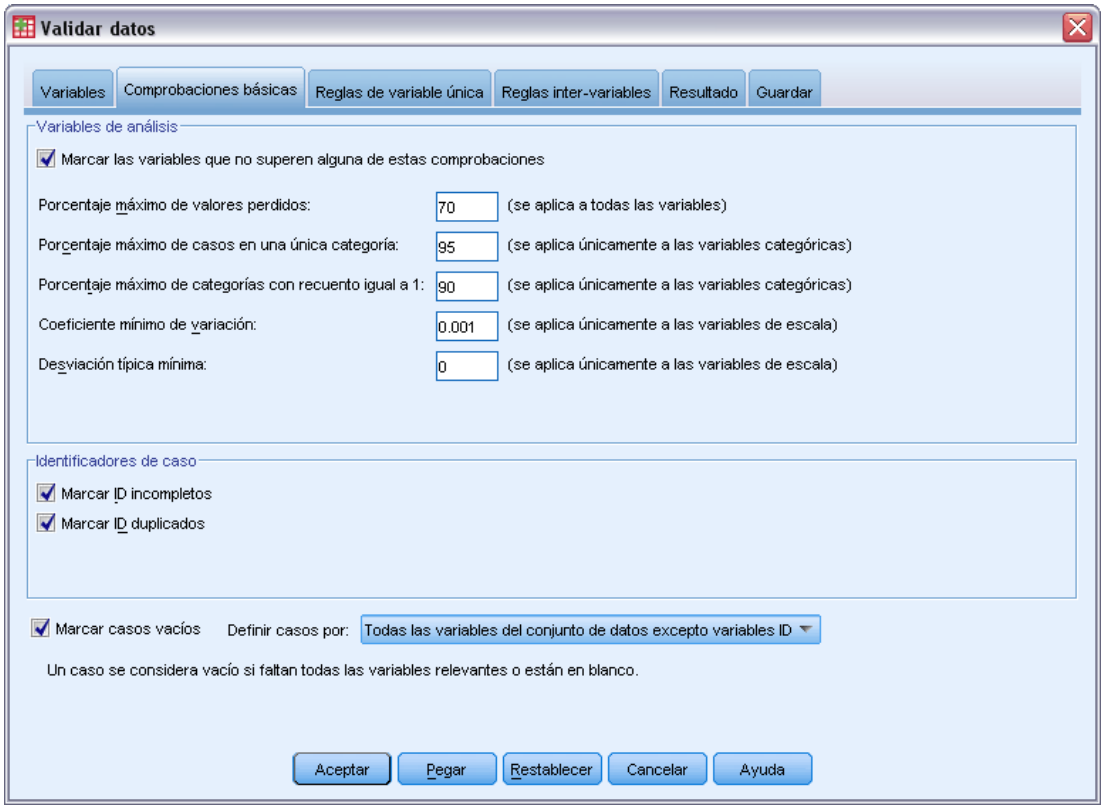
- Para ejecutar un análisis de Validar datos, elija en los menús:
  Datos > Validación > Validar datos...

Figura 7-1
Cuadro de diálogo Validar datos, pestaña Variables



- Seleccione *Tamaño del hospital* y *Edad en años* hasta *Índice de Barthel recodificado al mes 6* como variables de análisis.
- Seleccione *ID del hospital*, *ID del paciente* e *ID médico responsable* como variables de identificación de casos.
- Pulse en la pestaña Comprobaciones básicas.

Figura 7-2
Cuadro de diálogo Validar datos, pestaña Comprobaciones básicas



La configuración por defecto es la configuración que se desea ejecutar.

▶ Pulse en Aceptar.

Advertencias

Figura 7-3
Advertencias

Algunos o todos los resultados solicitados no aparecen porque todos los valores de datos, casos o variables han pasado las comprobaciones solicitadas.

Las variables de análisis superaron las comprobaciones básicas y no hay casos vacíos, por lo que aparece una advertencia que explica por qué no hay ningún resultado que corresponda a esas comprobaciones.

Identificadores incompletos

Figura 7-4
Identificadores de casos incompletos

	Identificador		
Caso	idhosp	idpac	idmed
288	OZN		125304
573		6137798782	790697
774		2322241867	176466

Si hay valores perdidos en las variables de identificación de casos, el caso no se puede identificar correctamente. En este archivo de datos, al caso 288 le falta el *ID de paciente*, y a los casos 573 y 774 les falta el *ID de hospital*.

Identificadores duplicados

Figura 7-5
Identificadores de casos duplicados (se muestran los 11 primeros)

Grupo de identificadores duplicados	Número de duplicados	Casos con identificadores duplicados	Identificador		
			idhosp	idpac	idmed
1	2	10, 11	PBW	1406462419	355184
2	2	14, 15	PBW	2191527525	355184
3	2	21, 22	PBW	7237535360	616528
4	2	28, 29	NHV	4592215163	942982
5	2	30, 31	NHV	7628592330	371884
6	2	64, 65	NHV	0300750006	371884
7	2	83, 84	QWS	4590625286	215041
8	2	86, 87	QWS	6272818258	817329
9	2	96, 97	QWS	1959349605	215041
10	3	100, 101, 102	QWS	5856145337	817329
11	3	104, 105, 106	QWS	1543897849	817329

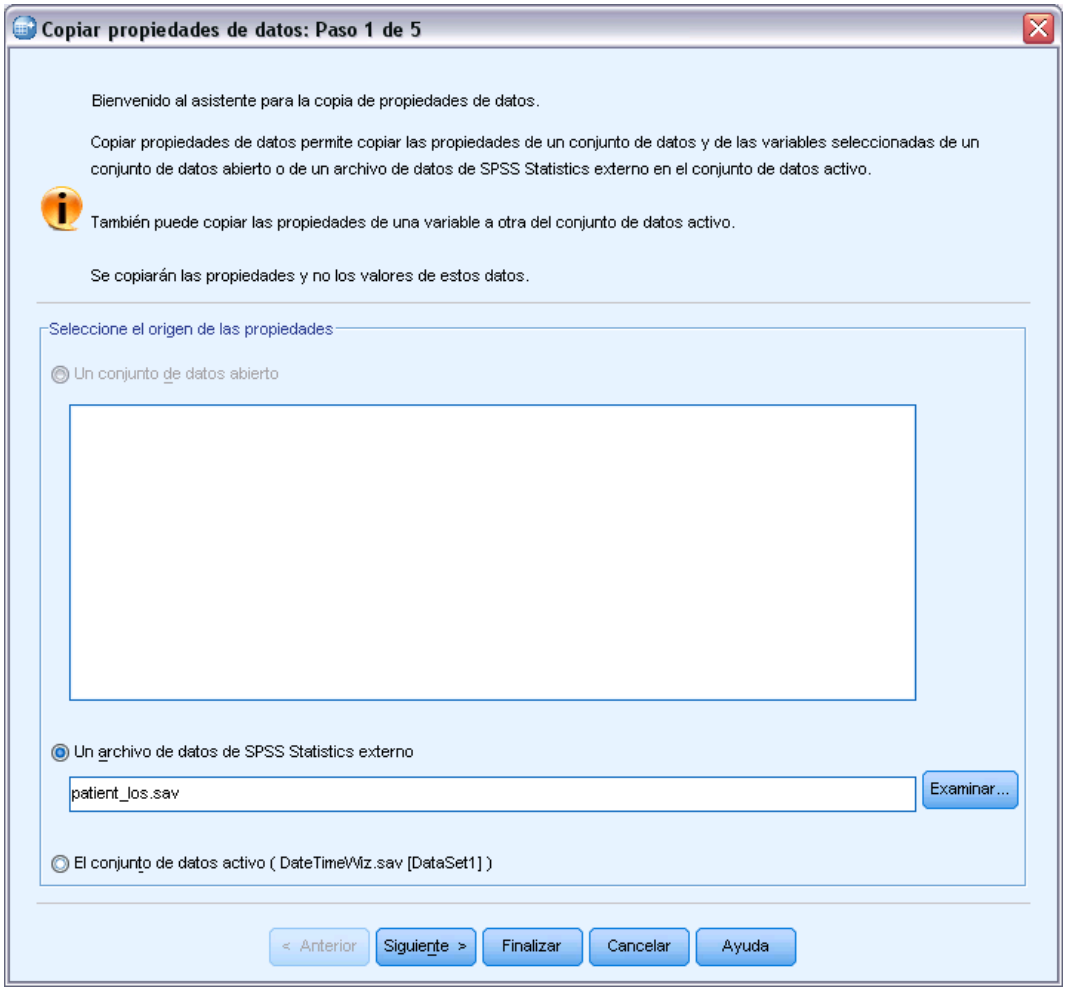
Un caso debe estar identificado de forma única por la combinación de valores de las variables de identificación. A continuación, se muestran las 11 primeras entradas de la tabla de identificadores duplicados. Estos duplicados son pacientes con varios eventos, que se han introducido como casos independientes para cada evento. Como esta información se puede recopilar en una única fila, se deberían limpiar estos casos.

Copia y utilización de reglas desde otro archivo

El analista se da cuenta de que las variables de este archivo de datos son similares a las variables de otro proyecto. Las reglas de validación definidas para ese proyecto se almacenan como propiedades del archivo de datos asociado y se pueden aplicar a este archivo de datos copiando las propiedades de los datos del archivo.

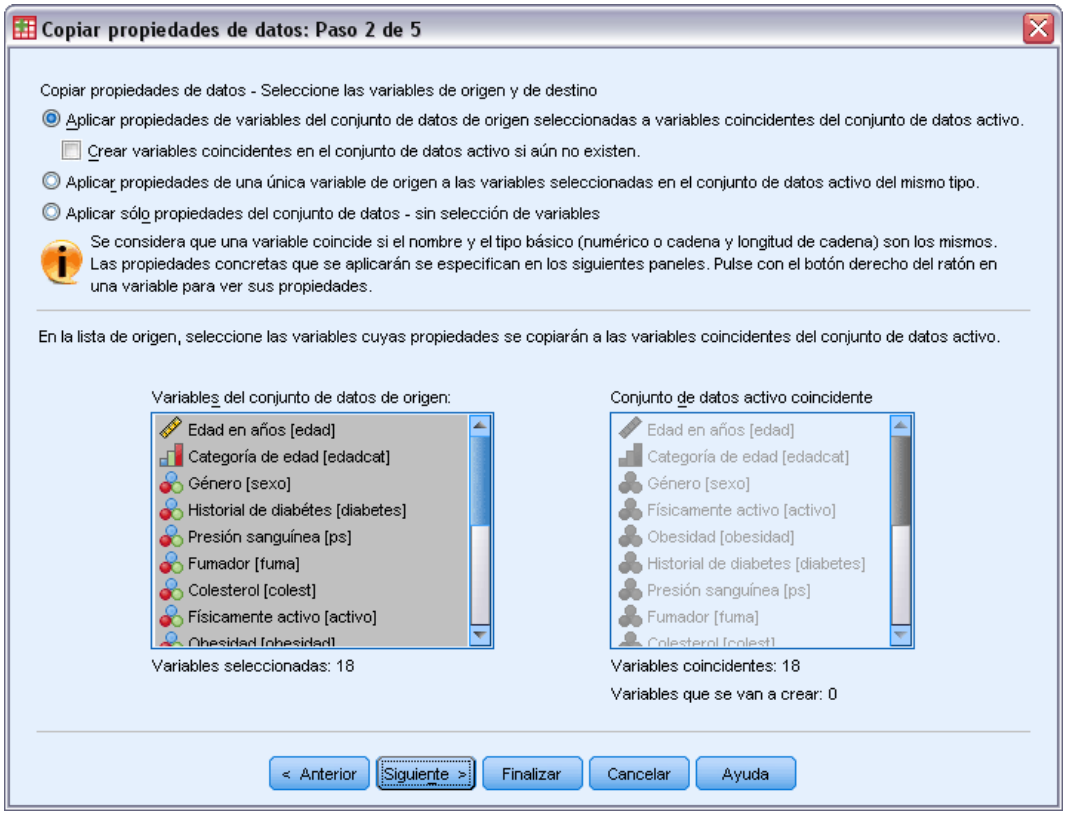
► Para copiar reglas desde otro archivo, elija en los menús:
Datos > Copiar propiedades de datos...

Figura 7-6
Copiar propiedades de datos, Paso 1 (bienvenida)



- Elija copiar las propiedades desde un archivo de datos IBM® SPSS® Statistics externo, *patient_los.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A el p. 138.
- Pulse en Siguiente.

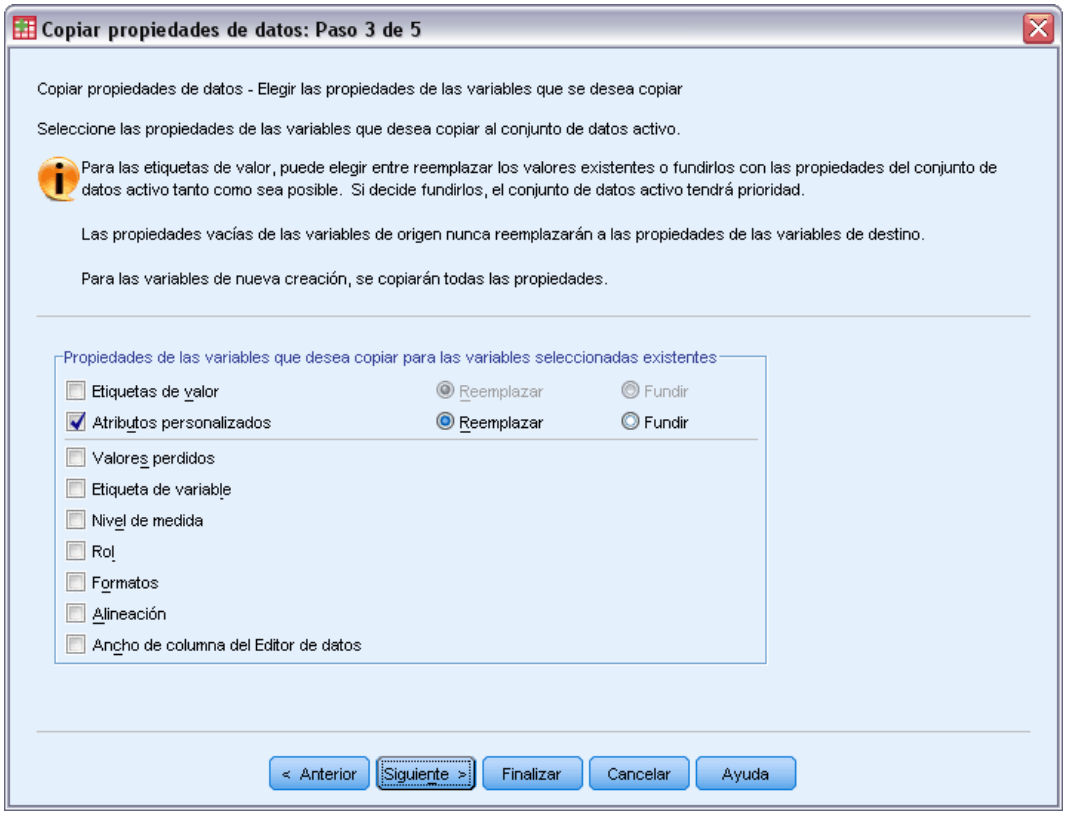
Figura 7-7
Copiar propiedades de datos, Paso 2 (seleccionar variables)



Estas son las variables cuyas propiedades desea copiar desde *patient_los.sav* a las correspondientes variables en *stroke_invalid.sav*.

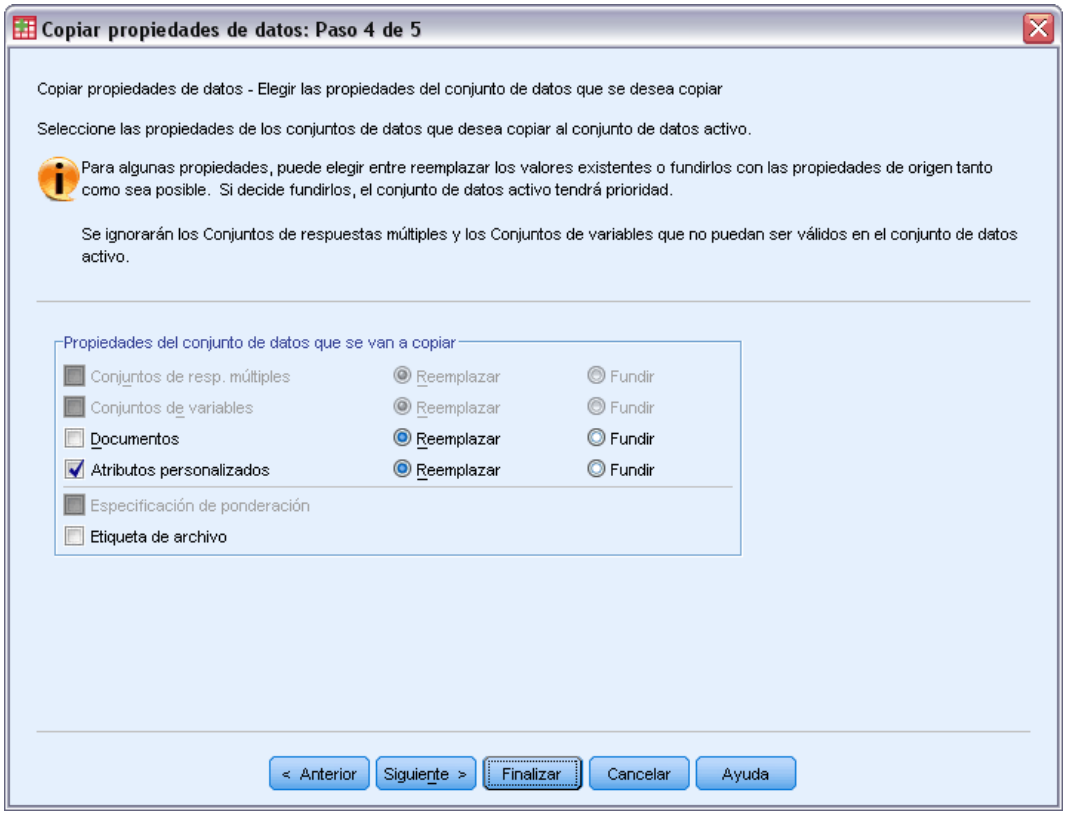
- Pulse en Siguiente.

Figura 7-8
Copiar propiedades de datos, Paso 3 (seleccionar propiedades de variables)



- Anule la selección de todas las propiedades excepto Atributos personalizados.
- Pulse en Siguiente.

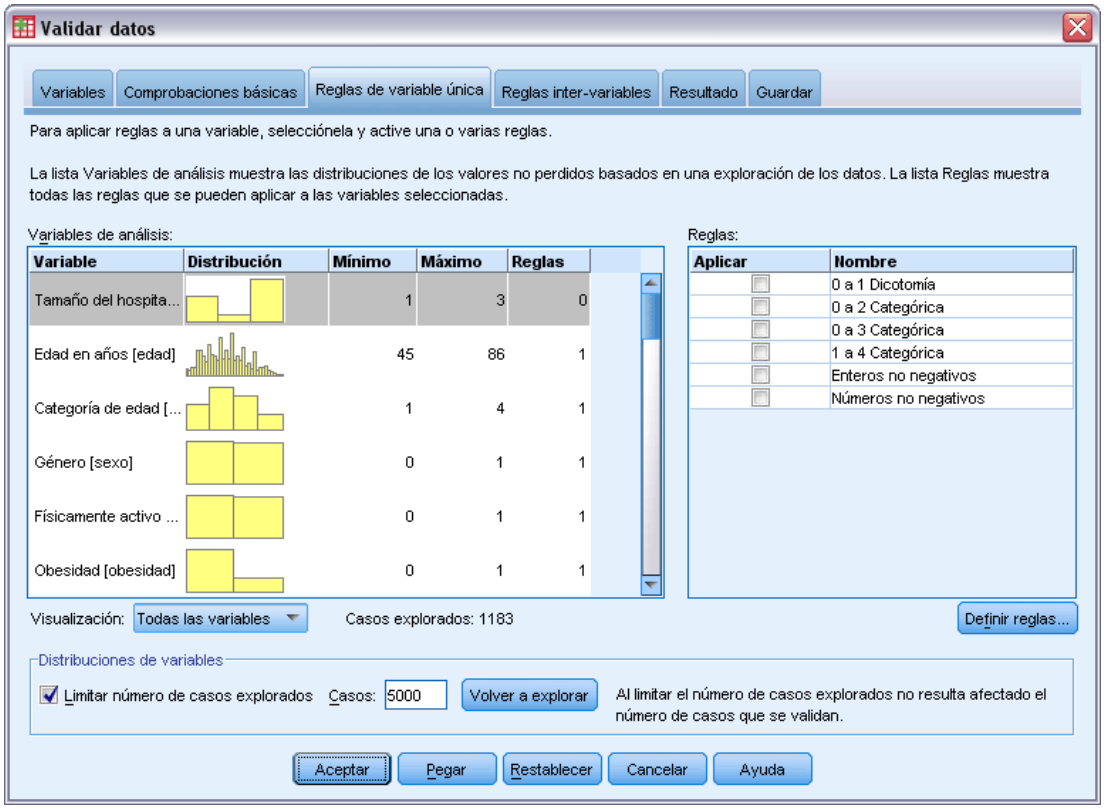
Figura 7-9
Copiar propiedades de datos, Paso 4 (seleccionar propiedades de conjunto de datos)



- Seleccione Atributos personalizados.
- Pulse en Finalizar.

Ya está preparado para volver a utilizar las reglas de validación.

Figura 7-10
Cuadro de diálogo Validar datos, pestaña Reglas de variable única

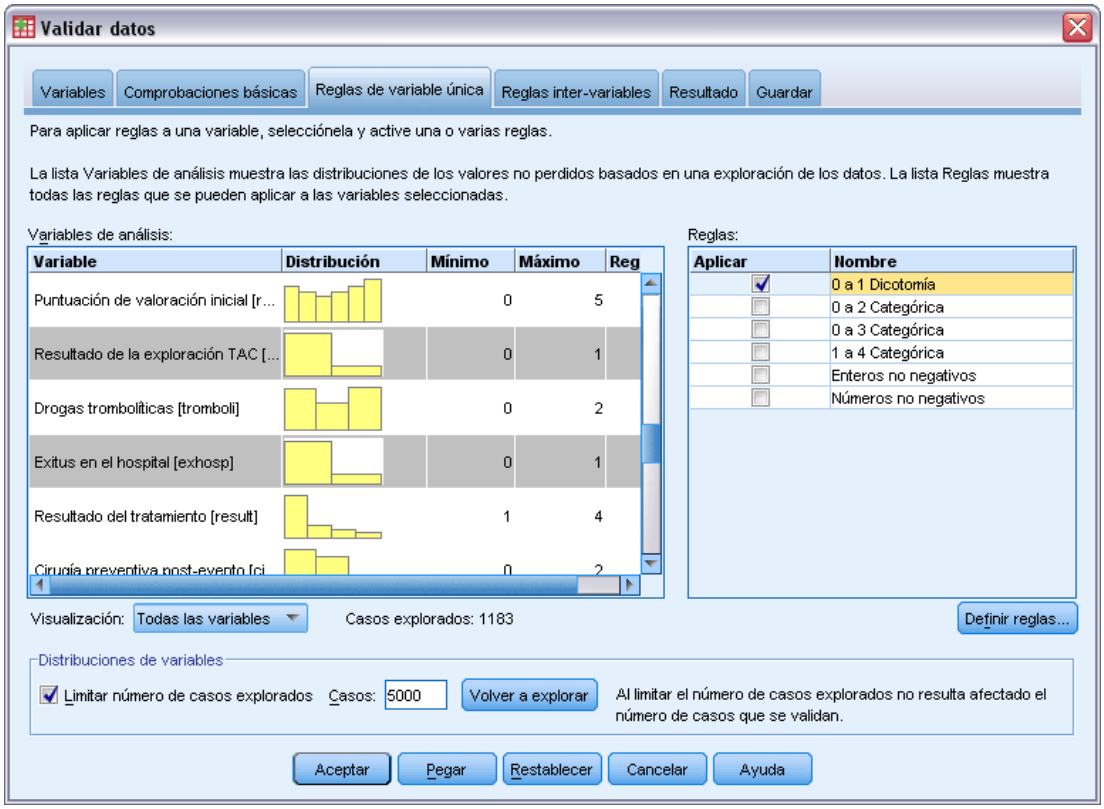


- Para validar los datos de *stroke_invalid.sav* mediante las reglas copiadas, pulse en el botón de la barra de herramientas Recuperar cuadros de diálogo y seleccione Validar datos.
- Pulse en la pestaña Reglas de variable única.

La lista de Variables de análisis muestra las variables seleccionadas en la pestaña Variables, información de resumen sobre sus distribuciones y el número de reglas vinculadas a cada variable. Las variables cuyas propiedades se copiaron de *patient_los.sav* tienen reglas vinculadas.

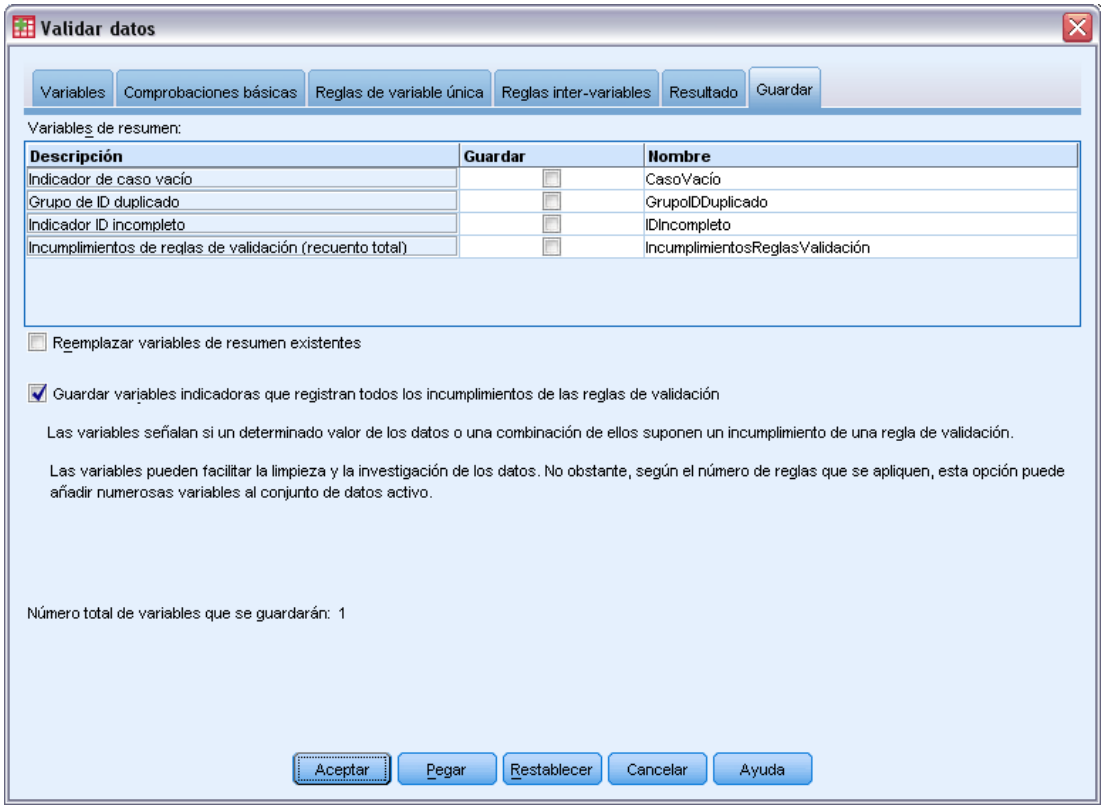
La lista Reglas muestra las reglas de validación de variable única disponibles en el archivo de datos. Todas esas reglas se copiaron del archivo *patient_los.sav*. Observe que algunas de dichas reglas son aplicables a variables que no tienen una análoga exacta en el otro archivo de datos.

Figura 7-11
Cuadro de diálogo Validar datos, pestaña Reglas de variable única



- Seleccione *Fibrilación Atrial*, *Historial de ataque isquémico transitorio*, *Resultado de la exploración TAC* y *Exitus en el hospital*, y aplique la regla 0 to 1 Dichotomy.
- Aplique 0 to 3 Categorical a *Rehabilitación post-evento*.
- Aplique 0 to 2 Categorical a *Cirugía preventiva post-evento*.
- Aplique Nonnegative integer a *Duración de la estancia de rehabilitación*.
- Aplique 1 to 4 Categorical desde el *Índice de Barthel recodificado al mes 1* hasta el *Índice de Barthel recodificado al mes 6*.
- Pulse en la pestaña Guardar.

Figura 7-12
Cuadro de diálogo Validar datos, pestaña Guardar



- Seleccione Guardar variables indicadoras que registran todos los incumplimientos de las reglas de validación. Este proceso simplificará la conexión del caso y la variable que provoca los incumplimientos de la regla de variable única.
- Pulse en Aceptar.

Descripciones de reglas

Figura 7-13
Descripciones de reglas

Regla	Descripción
Enteros no negativos	Type: Numeric Domain: Range Flag user-missing values: No Flag system-missing values: Yes Minimum: 0 Flag unlabeled values within range: No Flag noninteger values within range: Yes $VD.SRule[5]: Rule
0 a 1 Dicotomía	Type: Numeric Domain: List Flag user-missing values: No Flag system-missing values: Yes List: 0; 1 $VD.SRule[1]: Rule
1 a 4 Categórica	Type: Numeric Domain: List Flag user-missing values: No Flag system-missing values: Yes List: 1; 2; 3; 4 $VD.SRule[4]: Rule

La tabla de descripción de reglas ofrece explicaciones de las reglas que se han incumplido. Esta característica es muy útil cuando se realiza el seguimiento de muchas reglas de validación.

Resumen de variables

Figura 7-14
Resumen de variables

	Regla	Número de violaciones
edadcat	1 a 4 Categórica	1
	Total	1
sexo	0 a 1 Dicotomía	1
	Total	1
angina	0 a 1 Dicotomía	1
	Total	1
tiempo	Enteros no negativos	2
	Total	2
ic	0 a 1 Dicotomía	1
	Total	1

La tabla de resumen de variables enumera las variables que han incumplido al menos una regla de validación, las reglas incumplidas y el número de incumplimientos que se han producido por regla y por variable.

Informe de casos

Figura 7-15
Informe de casos

Caso	Violaciones de las reglas	Identificador		
	Variable única[a]	idhosp	idpac	idmed
175	0 a 1 Dicotomía (1)	OZN	0333204686	883285
274	0 a 1 Dicotomía (1)	OZN	1038840465	103254
310	Enteros no negativos (1)	OZN	2090290204	883285
437	0 a 1 Dicotomía (1)	WPA	2349729006	723384
752	Enteros no negativos (1)	GFG	4993307441	828754
1173	1 a 4 Categórica (1)	ALK	8737661990	185787

a. El número de variables que han violado la regla aparece a continuación de dicha regla.

La tabla de informe de casos enumera los casos (tanto por número de caso como por identificador de caso) que han incumplido al menos una regla de validación, las reglas incumplidas y el número de veces que el caso incumplió la regla. Los valores no válidos aparecerán en el Editor de datos.

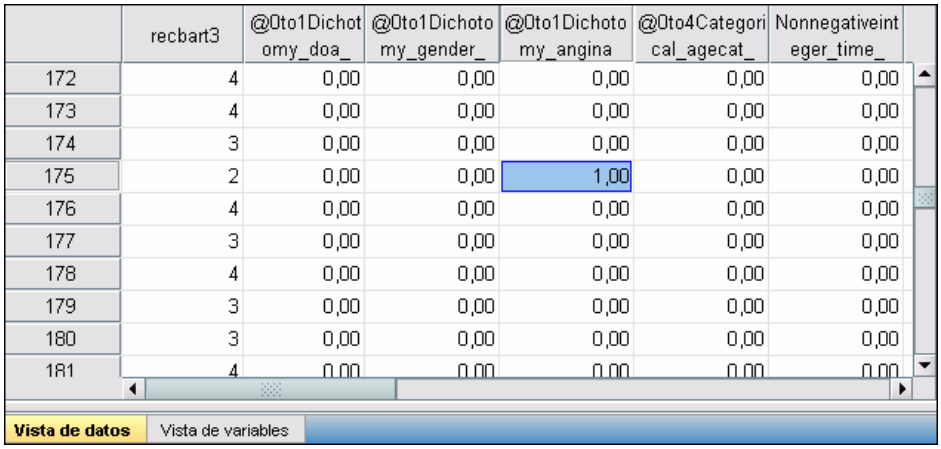
Figura 7-16
El Editor de datos con los indicadores guardados de los incumplimientos de reglas

	recbart3	@0to3Categorical_clotsolv_	@0to3Categorical_rehab_	@0to1Dichotomy_obesity_	@0to1Dichotomy_dhosp_	@0to1Dichotomy_tia_	@0toh
1	4	0,00	0,00	0,00	0,00	0,00	
2	4	0,00	0,00	0,00	0,00	0,00	
3	1	0,00	0,00	0,00	0,00	0,00	
4	4	0,00	0,00	0,00	0,00	0,00	
5	3	0,00	0,00	0,00	0,00	0,00	
6	4	0,00	0,00	0,00	0,00	0,00	
7	4	0,00	0,00	0,00	0,00	0,00	
8	4	0,00	0,00	0,00	0,00	0,00	
9	4	0,00	0,00	0,00	0,00	0,00	
10	2	0,00	0,00	0,00	0,00	0,00	
11	2	0,00	0,00	0,00	0,00	0,00	

Vista de datos | Vista de variables

Se produce una variable indicadora distinta para cada aplicación de una regla de validación. Por lo tanto, *@0to3Categorical_tromboli_* es la aplicación de la regla de validación de variable única "0 to 3 Categorical" a la variable *Toma drogas anticoagulantes*. Para un determinado caso, la forma más fácil de descubrir cuál de los valores de la variable no es válido consiste simplemente en explorar los valores de los indicadores. Un valor de 1 significa que el valor de la variable asociada no es válido.

Figura 7-17
El Editor de datos con indicador de incumplimiento de regla para el caso 175

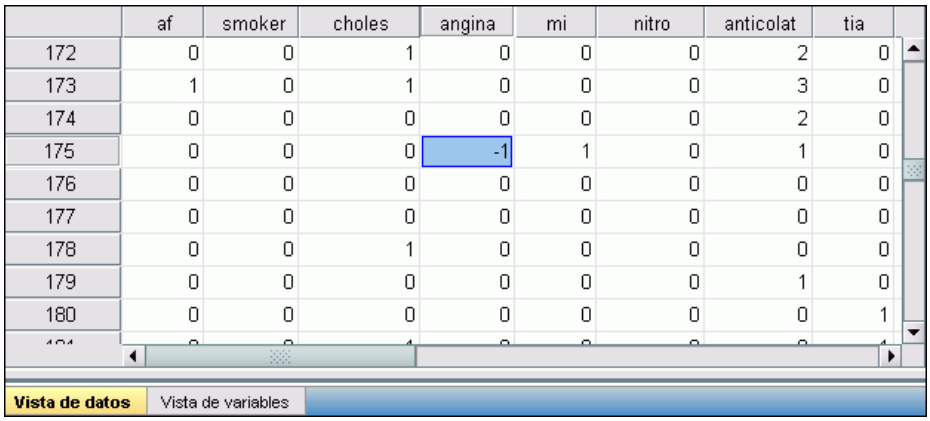


	recbart3	@0to1Dichotomy_doa_	@0to1Dichotomy_gender_	@0to1Dichotomy_angina	@0to4Categorical_agecat_	Nonnegativeinteger_time_
172	4	0,00	0,00	0,00	0,00	0,00
173	4	0,00	0,00	0,00	0,00	0,00
174	3	0,00	0,00	0,00	0,00	0,00
175	2	0,00	0,00	1,00	0,00	0,00
176	4	0,00	0,00	0,00	0,00	0,00
177	3	0,00	0,00	0,00	0,00	0,00
178	4	0,00	0,00	0,00	0,00	0,00
179	3	0,00	0,00	0,00	0,00	0,00
180	3	0,00	0,00	0,00	0,00	0,00
181	4	0,00	0,00	0,00	0,00	0,00

Vista de datos | Vista de variables

Vaya al caso 175, el primer caso con un incumplimiento de reglas. Para acelerar la búsqueda, observe los indicadores que están asociados con variables en la tabla de resumen de variables. Se ve rápidamente que *Historial de angina* tiene el valor no válido.

Figura 7-18
El Editor de datos con el valor no válido para Historial de angina



	af	smoker	choles	angina	mi	nitro	anticolat	tia
172	0	0	1	0	0	0	2	0
173	1	0	1	0	0	0	3	0
174	0	0	0	0	0	0	2	0
175	0	0	0	-1	1	0	1	0
176	0	0	0	0	0	0	0	0
177	0	0	0	0	0	0	0	0
178	0	0	1	0	0	0	0	0
179	0	0	0	0	0	0	1	0
180	0	0	0	0	0	0	0	1

Vista de datos | Vista de variables

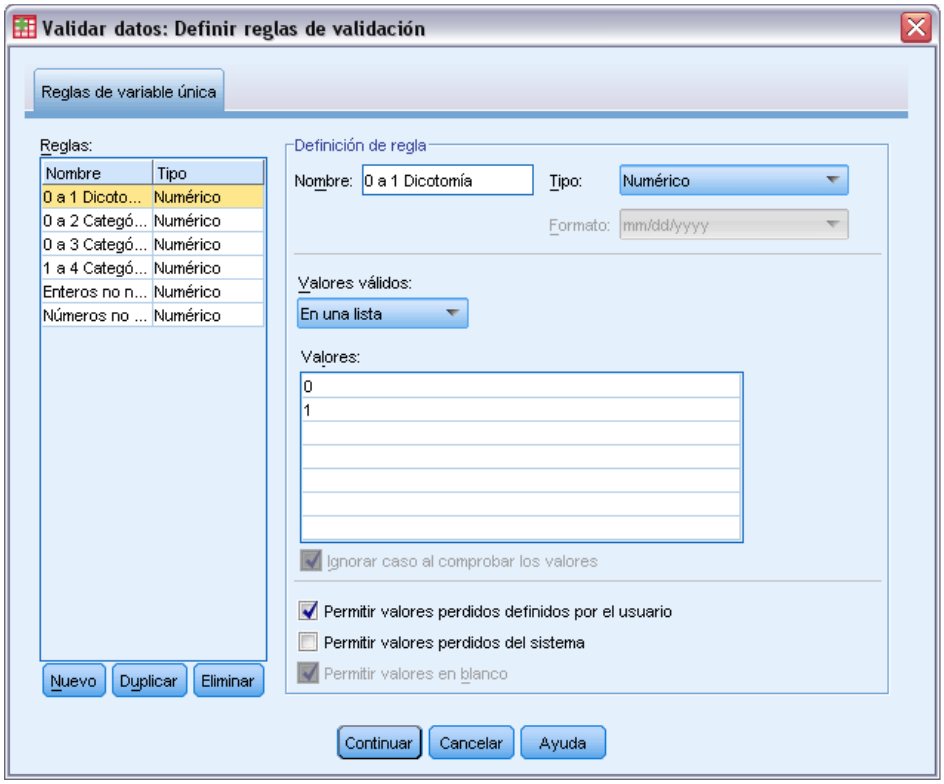
Historial de angina tiene un valor de –1. Aunque este valor es un valor perdido válido para las variables de tratamiento y de resultados en el archivo de datos, aquí no es válido porque los valores del historial de los pacientes no tienen actualmente valores definidos como perdidos por el usuario.

Definición de reglas propias

Las reglas de validación que se copiaron de *patient_los.sav* han sido de gran utilidad, pero deberá definir algunas reglas más para acabar la tarea. Además, en ocasiones algunos pacientes que ingresaron cadáver se anotaron, de forma accidental, como fallecidos en el hospital. Las reglas de validación de variable única no pueden detectar esta situación, por lo que, para ello, deberá definir una regla inter-variables.

- Pulse en el botón de la barra de herramientas Recuperar cuadros de diálogo y seleccione Validar datos.
- Pulse en la pestaña Reglas de variable única. (Deberá definir reglas para *Tamaño del hospital*, las variables que miden las puntuaciones de valoración y las variables que correspondan a los índices de Barthel recodificados.)
- Pulse en Definir reglas.

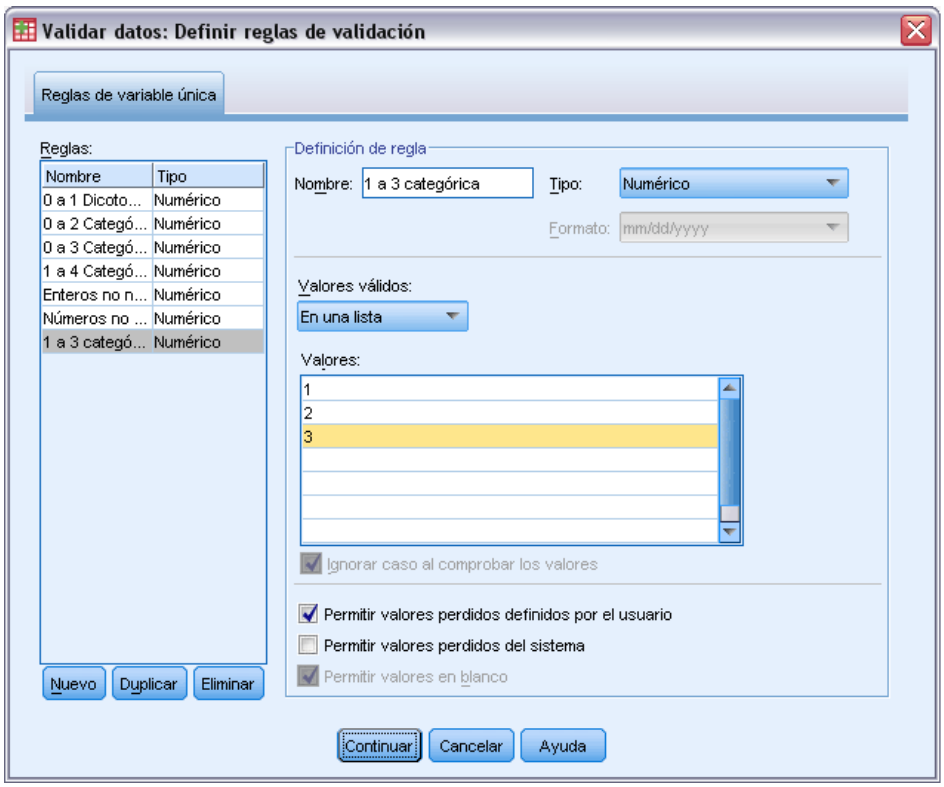
Figura 7-19
Cuadro de diálogo Definir reglas de validación, pestaña Reglas de variable única



Aparecen las reglas definidas actualmente, con la regla 0 to 1 Dichotomy seleccionada en la lista de Reglas y se muestran las propiedades de la regla en el grupo Definición de regla.

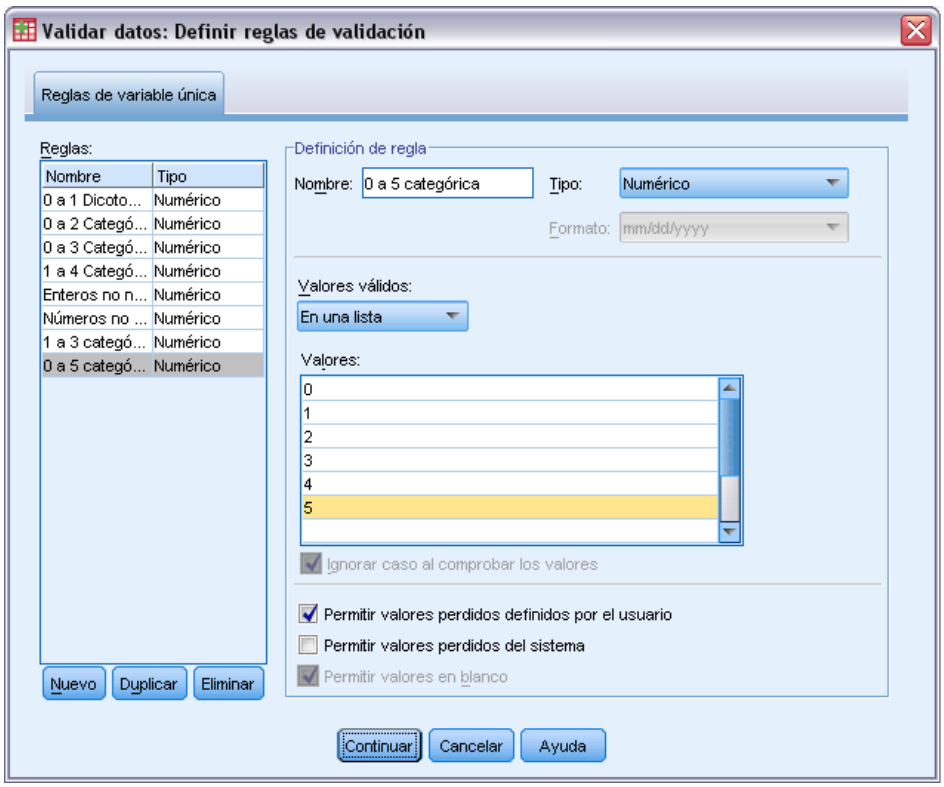
- Para definir una regla, pulse en Nuevo.

Figura 7-20
Cuadro de diálogo Definir reglas de validación, pestaña Reglas de variable única (definida 1 a 3 categórica)



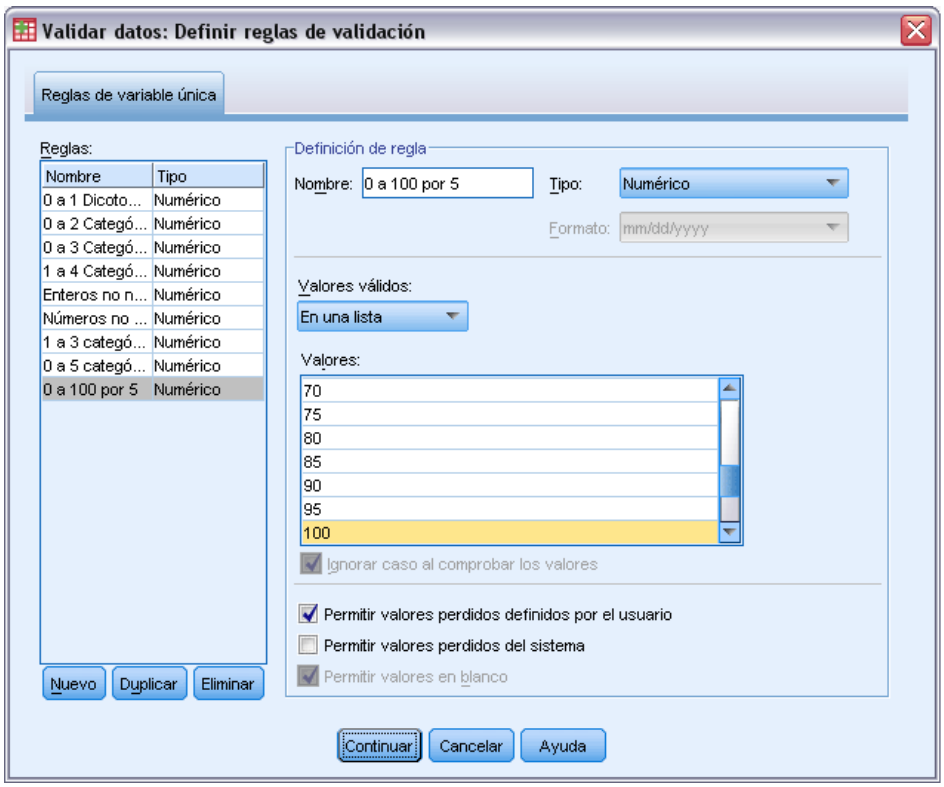
- Escriba 1 a 3 categórica como nombre de la regla.
- Para Valores válidos, elija En una lista.
- Escriba 1, 2 y 3 como los valores.
- Anule la selección de Permitir valores perdidos del sistema.
- Para definir la regla para las puntuaciones de valoración, pulse en Nuevo.

Figura 7-21
Cuadro de diálogo Definir reglas de validación, pestaña Reglas de variable única (definida 0 a 5 categórica)



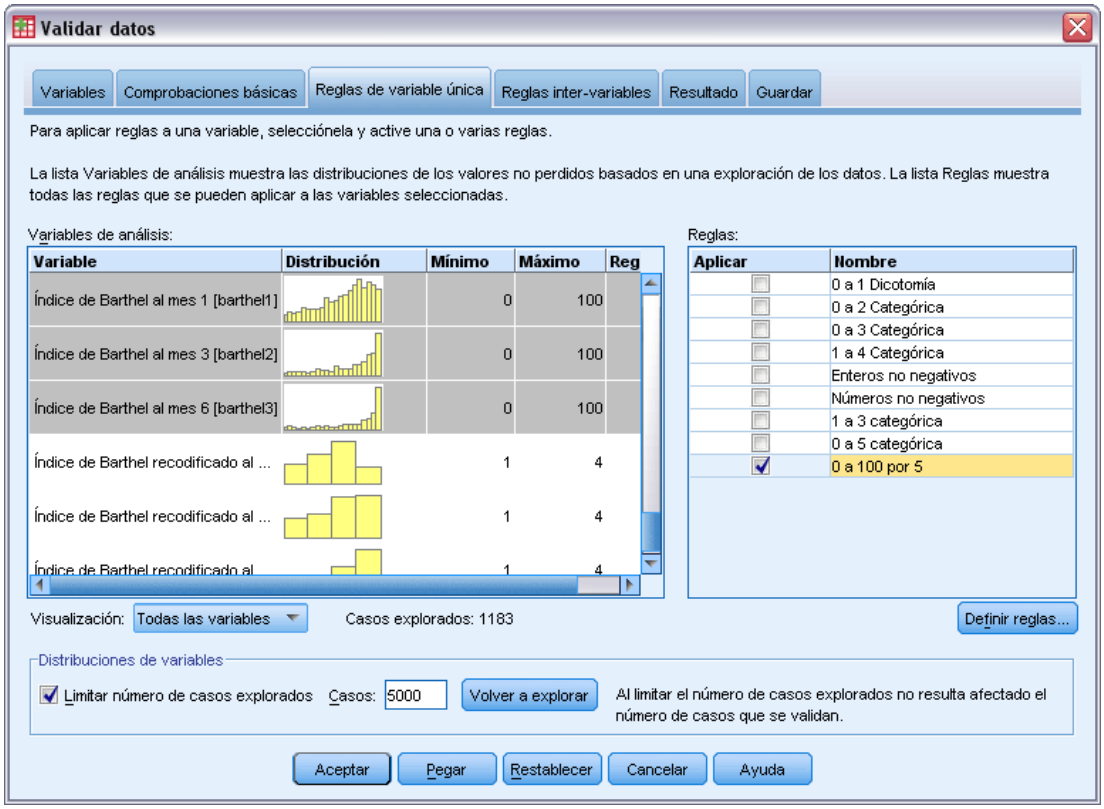
- Escriba 0 a 5 categórica como nombre de la regla.
- Para Valores válidos, elija En una lista.
- Escriba 0, 1, 2, 3, 4 y 5 como los valores.
- Anule la selección de Permitir valores perdidos del sistema.
- Para definir la regla para los índices de Barthel, pulse en Nuevo.

Figura 7-22
Cuadro de diálogo Definir reglas de validación, pestaña Reglas de variable única (definida 0 a 100 por 5 categórica)



- Escriba 0 a 100 por 5 como nombre de la regla.
- Para Valores válidos, elija En una lista.
- Escriba 0, 5, ..., y 100 como los valores.
- Anule la selección de Permitir valores perdidos del sistema.
- Pulse en Continuar.

Figura 7-23
Cuadro de diálogo Validar datos, pestaña Reglas de variable única (definida 0 a 100 por 5)



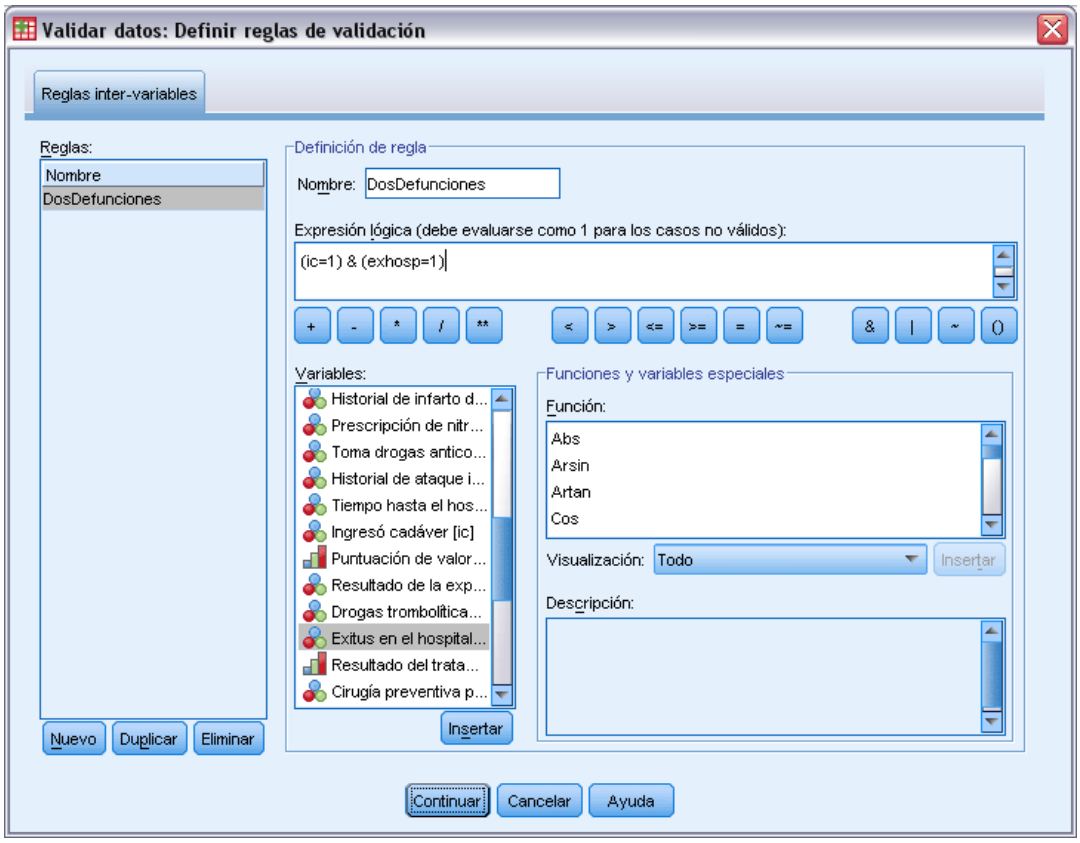
Ahora, es el momento de aplicar las reglas definidas a las variables de análisis.

- Aplique 1 a 3 categórica a *Tamaño del hospital*.
- Aplique 0 a 5 categórica a *Puntuación de valoración inicial* y de *Puntuación de valoración al mes 1* hasta *Puntuación de valoración al mes 6*.
- Aplique 0 a 100 por 5 desde *Índice de Barthel al mes 1* hasta el *Índice de Barthel al mes 6*.
- Pulse en la pestaña Reglas inter-variables.

No hay ninguna regla definida actualmente.

- Pulse en Definir reglas.

Figura 7-24
Cuadro de diálogo Definir reglas de validación, pestaña Reglas inter-variables



Cuando no hay ninguna regla, se crea automáticamente una nueva regla de marcador de posición.

- Escriba DosDefunciones como nombre de la regla.
- Escriba (ic=1) & (exhosp=1) como expresión lógica. Esto devolverá un valor 1 si el paciente aparece registrado como que ingresó cadáver y como fallecido en el hospital.
- Pulse en Continuar.

  La regla recién definida aparece automáticamente seleccionada en la pestaña Reglas inter-variables.
- Pulse en Aceptar.

Reglas inter-variables

Figura 7-25
Reglas inter-variables

Regla	Número de violaciones	Expresión de la regla
DosDefunciones	27	(ic=1) & (exhosp=1)

El resumen de las reglas inter-variables enumera las reglas inter-variables que se han incumplido al menos una vez, el número de incumplimientos que se ha producido y una descripción de cada regla incumplida.

Informe de casos

Figura 7-26
Informe de casos

	Violaciones de las reglas de validación		Identificador		
Caso	Variable única[a]	Variable cruzada	idhosp	idpac	idmed
20		DosDefunciones	PBW	1192970826	355184
49		DosDefunciones	NHV	8717862852	237418
129		DosDefunciones	QWS	6901932085	215041
138		DosDefunciones	RLD	1205005069	695521
162		DosDefunciones	OZN	5546809538	125304
175	0 a 1 Dicotomía (1)		OZN	0333204686	883285
274	0 a 1 Dicotomía (1)		OZN	1038840465	103254
310	Enteros no negativos (1)		OZN	2090290204	883285
414		DosDefunciones	WPA	3351107142	462020
437	0 a 1 Dicotomía (1)		WPA	2349729006	723384
447		DosDefunciones	WPA	7163481282	519548
458		DosDefunciones	WPA	9159094175	652070
462		DosDefunciones	WPA	2137520354	723384
537		DosDefunciones	SLB	5246122506	928076
544		DosDefunciones	SLB	1605957462	506108
620		DosDefunciones	GFG	8141858966	828754
629		DosDefunciones	GFG	3397891610	539412
630		DosDefunciones	GFG	3397891610	539412
639		DosDefunciones	GFG	3962622031	327422
644		DosDefunciones	GFG	4271782383	749432
649		DosDefunciones	GFG	0950686750	618069
653		DosDefunciones	GFG	0663642766	001448
722		DosDefunciones	GFG	0418125590	877354
748		DosDefunciones	GFG	8744721380	539412
752	Enteros no negativos (1) 0 a 1 Dicotomía (1)		GFG	4993307441	828754
868		DosDefunciones	WWL	9714672452	237547
881		DosDefunciones	WWL	6613279456	574275
915		DosDefunciones	EFX	2575793702	501318
933		DosDefunciones	IZO	2807437472	680253
1010		DosDefunciones	BLA	5284009939	657638
1028		DosDefunciones	BLA	8021997463	185703
1054		DosDefunciones	ALK	0950897644	267830
1173	1 a 4 Categórica (1)		ALK	8737661990	185787

a. El número de variables que han violado la regla aparece a continuación de dicha regla.

El informe de casos incluye ahora los casos que incumplieron la regla inter-variables, así como los casos detectados anteriormente que incumplieron las reglas de variable única. Se deberá informar de todos estos casos al equipo de introducción de datos para su corrección.

Resumen

El analista tiene la información necesaria para crear un informe preliminar que enviar al administrador del equipo de introducción de datos.

Procedimientos relacionados

El procedimiento Validar datos es una herramienta muy útil para controlar la calidad de los datos.

- El procedimiento Identificar casos atípicos analiza patrones en los datos e identifica casos con algunos valores significativos que varían del tipo.

Capítulo
8

Preparación automática de datos

La preparación de los datos para su análisis es uno de los pasos más importantes en cualquier proyecto y, tradicionalmente, uno de los que más tiempo requieren. Preparación automática de datos (ADP) controla las tareas automáticamente, analizando los datos e identificando problemas, filtrando campos problemáticos o sin posibilidades de ser útiles, derivando nuevos atributos cuando sea necesario y mejorando el rendimiento mediante técnicas de filtrado inteligente. Puede utilizar el algoritmo de una forma totalmente **automática**, permitiendo seleccionar y aplicar soluciones; o de forma **interactiva**, previendo los cambios antes de que se realicen y aceptarlos o rechazarlos según sea necesario.

ADP permite hacer que sus datos estén listos para la generación de modelos de forma rápida y fácil, sin necesidad de tener conocimientos previos de los conceptos previos implicados. Los modelos tienden a crearse y puntuarse con mayor rapidez; además, el uso de ADP mejora la solidez de los procesos de modelado automatizados.

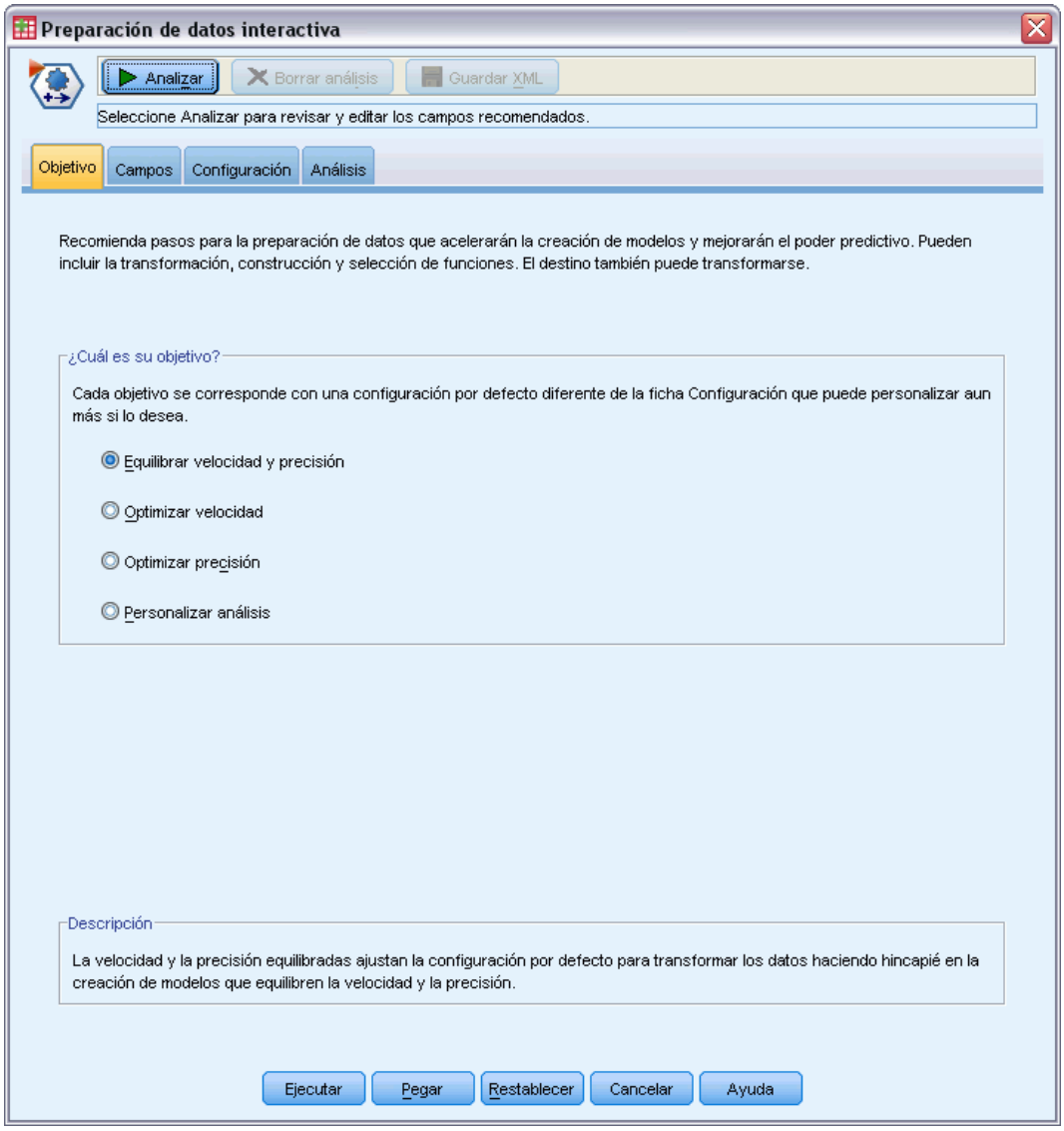
Uso interactivo de la preparación automática de datos

Una correduría de seguros con recursos limitados para investigar las reclamaciones de seguros de los asegurados desea crear un modelo para etiquetar las reclamaciones sospechosas y potencialmente fraudulentas. Tienen una muestra de información de reclamaciones anteriores recopiladas en *insurance_claims.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A el p. 138. Antes de construir el modelo, leerán los datos para el modelado mediante la preparación automática de datos. Como desean revisar las transformaciones propuestas antes de que se apliquen las transformaciones, utilizarán la preparación automática de datos en modo interactivo.

Selección entre objetivos

- Para ejecutar la Preparación automática de datos de forma interactiva, seleccione en los menús:
  Transformar > Preparar datos para modelado > Interactiva...

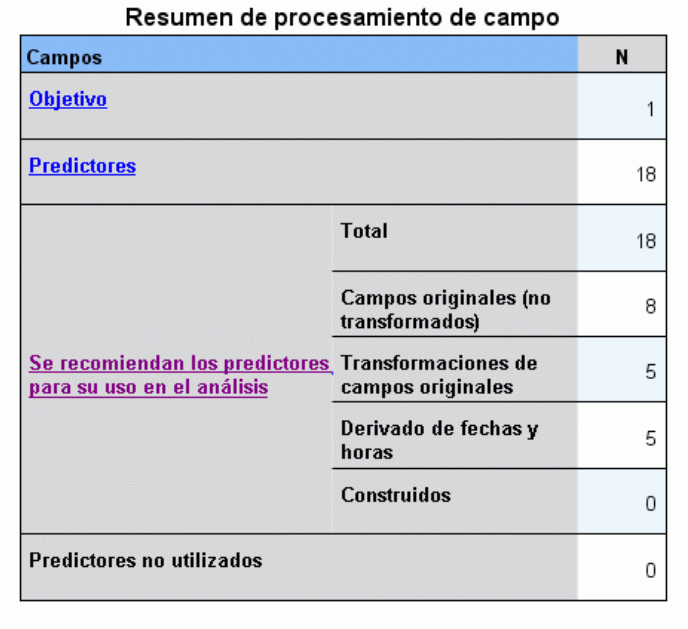
Figura 8-1
Pestaña Objetivo



La primera pestaña pide un objetivo que controla los ajustes predeterminados, pero ¿cuál es la diferencia práctica entre los objetivos? Al ejecutar el procedimiento con uno de los objetivos, podemos ver cómo difieren los resultados.

- Asegúrese de que la opción Equilibrar velocidad y precisión está seleccionada y pulse en Analizar.

Figura 8-2
Pestaña Análisis, resumen de procesamiento de campos para un objetivo equilibrado

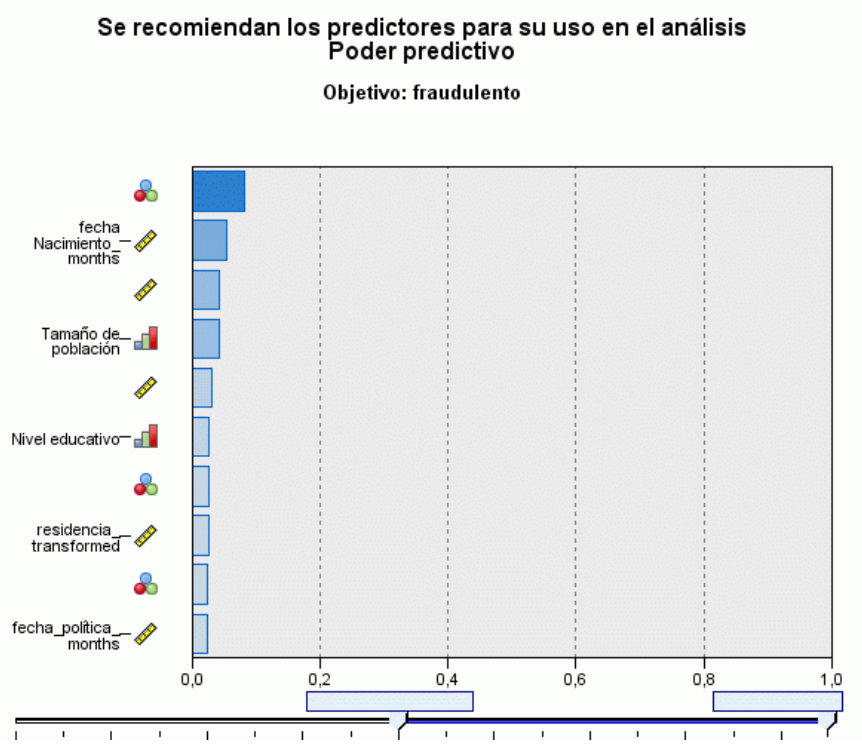


Resumen de procesamiento de campo

Campos		N
Objetivo		1
Predictores		18
Se recomiendan los predictores para su uso en el análisis	Total	18
	Campos originales (no transformados)	8
	Transformaciones de campos originales	5
	Derivado de fechas y horas	5
	Construidos	0
Predictores no utilizados		0

El enfoque cambia automáticamente a la pestaña Análisis, donde el procedimiento procesa los datos. La vista principal predeterminada está en el Resumen de procesamiento de campos, que le ofrece un resumen de cómo la preparación automática de campos procesó los campos. Hay un único destino, 18 entradas y 18 campos recomendados para la construcción de modelos. De los campos recomendados para el modelado, 9 eran campos de entrada originales, 4 eran transformaciones de los campos de entrada originales y 5 eran derivados de campos de fecha y hora.

Figura 8-3
Pestaña Análisis, potencia predictiva para un objetivo equilibrado



La vista auxiliar predeterminada es el Poder predictivo, que ofrece rápidamente una idea de qué campos recomendados serían más útiles para la construcción de modelos. Tenga en cuenta que aunque se recomiendan 18 predictores para el análisis, sólo los 10 primeros se mostrarán por defecto en el gráfico de potencia predictiva. Utilice el control deslizante que aparece bajo el gráfico para mostrar más o menos campos.

Si el objetivo es Equilibrar velocidad y precisión, se identifica que *Tipo de reclamación* es el "mejor" predictor, seguido de *Número de miembros del hogar* y la edad actual del reclamante en meses (duración calculada de la fecha de nacimiento a la fecha actual).

- Pulse en Borrar análisis y después pulse en la pestaña Objetivo.
- Seleccione Optimizar velocidad y pulse en Analizar.

Figura 8-4
Pestaña Análisis, resumen de procesamiento de campos optimizado para mayor velocidad

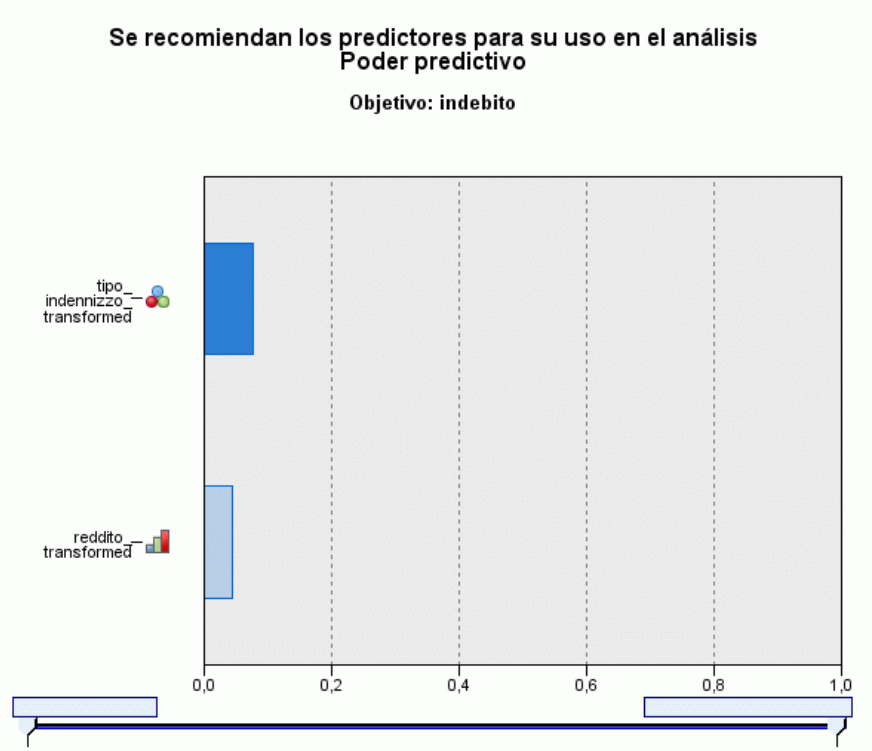
Resumen de procesamiento de campo

Campos		N
Objetivo		1
Predictores		18
Se recomiendan los predictores para su uso en el análisis	Total	2
	Campos originales (no transformados)	0
	Transformaciones de campos originales	2
	Derivado de fechas y horas	0
	Construidos	0
Predictores no utilizados		16

- No se pudieron construir predictores útiles. Las razones más comunes son: demasiados pocos predictores continuos fueron altamente asociados al objetivo o todos los predictores continuos eran independientes.

El enfoque vuelve a cambiar automáticamente a la pestaña Análisis, donde el procedimiento procesa los datos. En este caso, sólo 2 campos están recomendados para la construcción de modelos, y ambos son transformaciones de los campos originales.

Figura 8-5
Pestaña Análisis, potencia predictiva optimizada para mayor velocidad



Si Optimizar velocidad es el objetivo, *tipo_reclamación_transformada* se identifica como el "mejor" predictor, seguido de *ingreso_transformado*.

- Pulse en Borrar análisis y después pulse en la pestaña Objetivo.
- Seleccione Optimizar precisión y pulse en Analizar.

Figura 8-6
Pestaña Análisis, potencia predictiva optimizada para mayor precisión

Resumen de procesamiento de campo

Campos		N
Objetivo		1
Predictores		18
Se recomiendan los predictores para su uso en el análisis	Total	32
	Campos originales (no transformados)	8
	Transformaciones de campos originales	5
	Derivado de fechas y horas	19
	Construidos	0
Predictores no utilizados		0

Con Optimizar precisión como objetivo, se recomiendan 32 campos para la construcción del modelo, ya que hay más campos derivados de las fechas y horas mediante la extracción de días, meses y años a partir de fechas y horas, minutos y segundos a partir de horas.

Figura 8-7
Pestaña Análisis, potencia predictiva optimizada para mayor precisión

Tipo de reclamación se identifica como el "mejor" predictor, seguido del número de días desde que el solicitante inició su trabajo más reciente (la duración calculada desde la fecha de inicio del trabajo a la fecha actual) y el año que el solicitante inició el trabajo actual (extraído desde la fecha de inicio del trabajo).

En resumen:

- Equilibrar velocidad y precisión crea campos que pueden utilizarse para el modelado a partir de fecha, y puede transformar campos continuos como *reside* para distribuirlos más normalmente.
- Optimizar precisión crea algunos campos extra a partir de fechas (también busca valores atípicos y, si el destino es continuo, puede transformarlos para distribuirlos más normalmente).
- Optimizar velocidad no preparar fechas y no cambia la escala de campos continuos, pero fusiona categorías de predictores categóricos y desecha predictores continuos cuando el destino es categórico (y realiza la selección y construcción de características cuando el destino es continuo).

La aseguradora decide profundizar en los resultados de Optimizar para precisión.

▶ En la lista desplegable de la vista principal seleccione Campos.

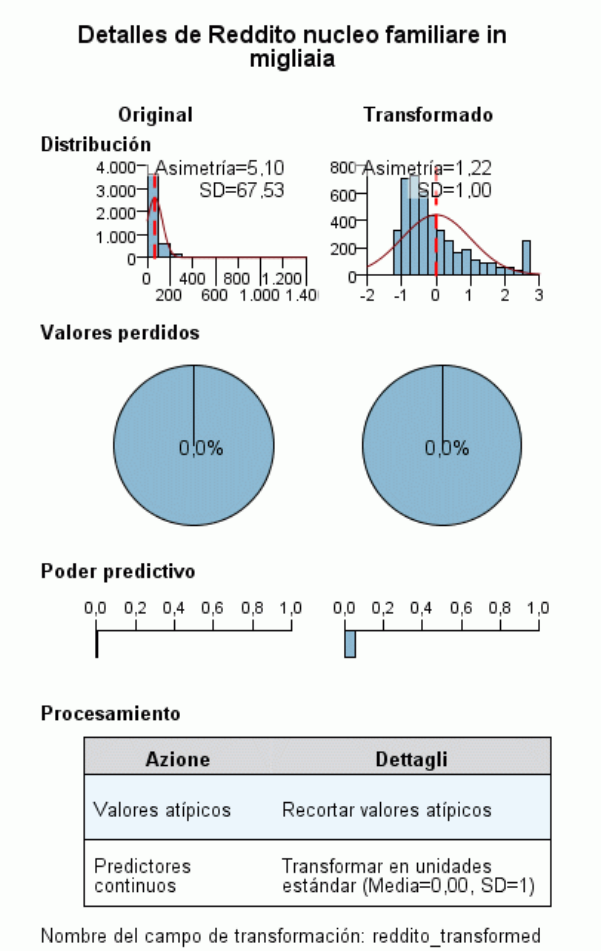
Campos y detalles de campos

Figura 8-8
Campos

La vista principal Campos muestra los campos procesados y si el modo ADP recomienda su uso para la construcción de modelos. Pulse en un nombre de campo para ver más información acerca del campo en la vista vinculada.

- Pulse en ingresos.

Figura 8-9
Detalles de campo para Ingresos del hogar en miles

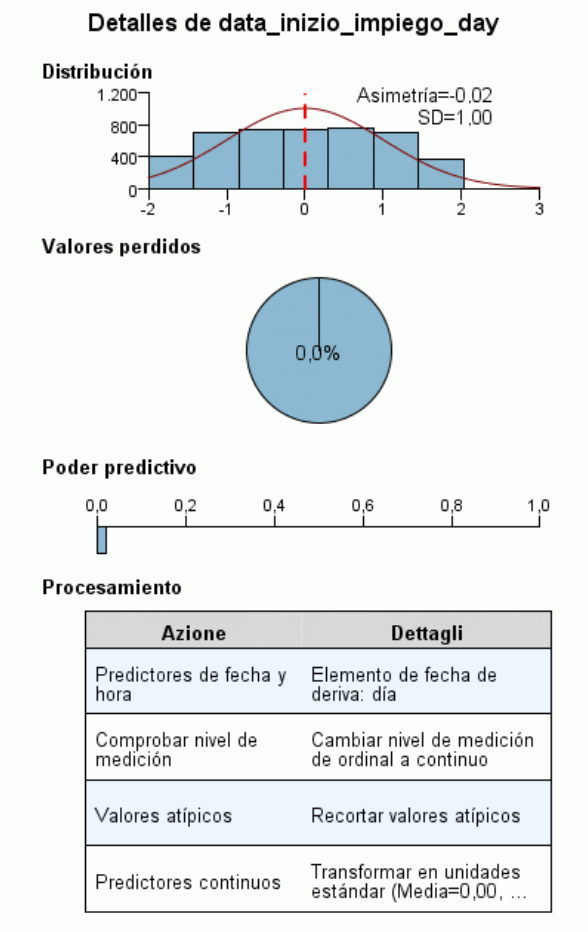


La vista Detalles del campo muestra las distribuciones del *Ingresos del hogar en miles* original y del transformado. De acuerdo con la tabla de procesamiento, se recortan los registros identificados como valores atípicos (definiendo sus valores como iguales al corte para determinar valores atípicos) y el campo se estandarizó para tener una media de 0 y una desviación estándar de 1. La curva del extremo derecho del histograma del campo transformado muestra un número de registros, quizás más de 200, que se identificaron como valores atípicos. Loa ingresos tienen una distribución altamente desviada, por lo que este puede ser un caso en el que corte predeterminado sea demasiado agresivo para determinar los valores atípicos.

Tenga también en cuenta el aumento de la potencia predictiva del campo transformado en el campo original. Parece que se trata de una transformación útil.

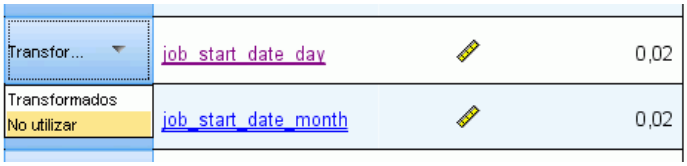
- En la vista Campos, pulse en día_fecha_inicio_trabajo. (Tenga en cuenta que es diferente de *días_fecha_inicio_trabajo*.)

Figura 8-10
Detalles de campo para día_fecha_inicio_trabajo



El campo *día_fecha_inicio_trabajo* es el día extraído desde *Fecha de inicio del empleo [fecha_inicio_trabajo]*. Es muy improbable que este campo tenga ninguna relación con si la reclamación es o no fraudulenta, por lo que la aseguradora desea retirarlo de la construcción del modelo.

Figura 8-11
Detalles de campo para Ingresos del hogar en miles



- En la vista Campos, seleccione No utilizar en la lista desplegable Versión de uso de la fila día_fecha_inicio_trabajo. Realice la misma operación para todos los campos con los prefijos *día_* y *mes_*.
- Para aplicar las transformaciones, pulse en Ejecutar.

El conjunto de datos estará ahora preparado para la construcción de datos, en el sentido de que todos los predictores recomendados (nuevos y viejos) tienen su papel definido como Entrada, mientras que los predictores no recomendados están definidos como Ninguno. Para crear un conjunto de datos sólo con los predictores recomendados, utilice los ajustes de Aplicar transformaciones del cuadro de diálogo.

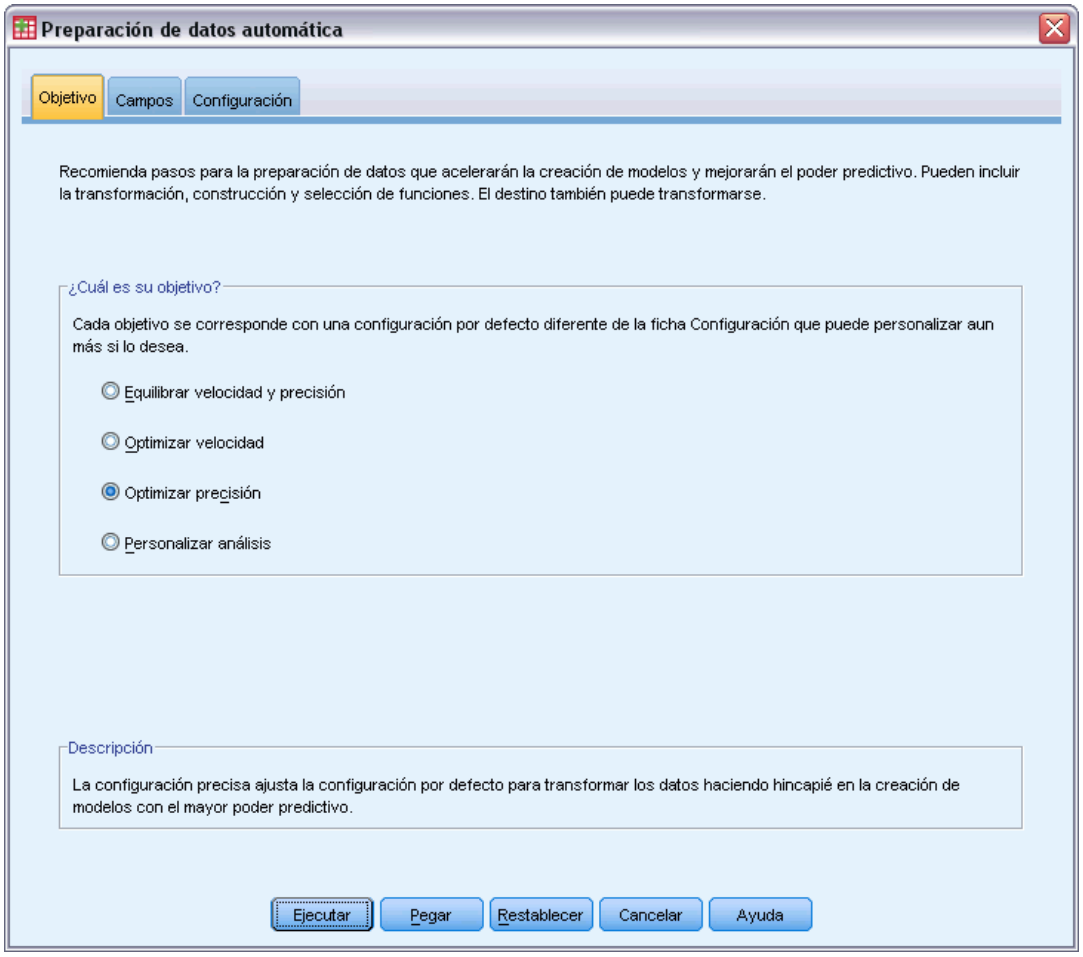
Uso automático de la preparación automática de datos

Un grupo del sector del automóvil desea realizar un seguimiento de las ventas de diversos vehículos a motor. Para poder identificar los modelos como mejor y peor rendimiento, desea establecer una relación entre las ventas de vehículos y las características de los vehículos. Esta información se recoge en el archivo *ventas_automoviles_sinpreparar.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A el p. 138. Utilice la preparación automática de datos para preparar los datos para el análisis. Cree también modelos utilizando la preparación “anterior” y “posterior” de datos para poder comparar los resultados.

Preparación de datos

- Para ejecutar la Preparación automática de datos de forma automática, seleccione en los menús:
  Transformar > Preparar datos para modelado > Automática...

Figura 8-12
Pestaña Objetivo

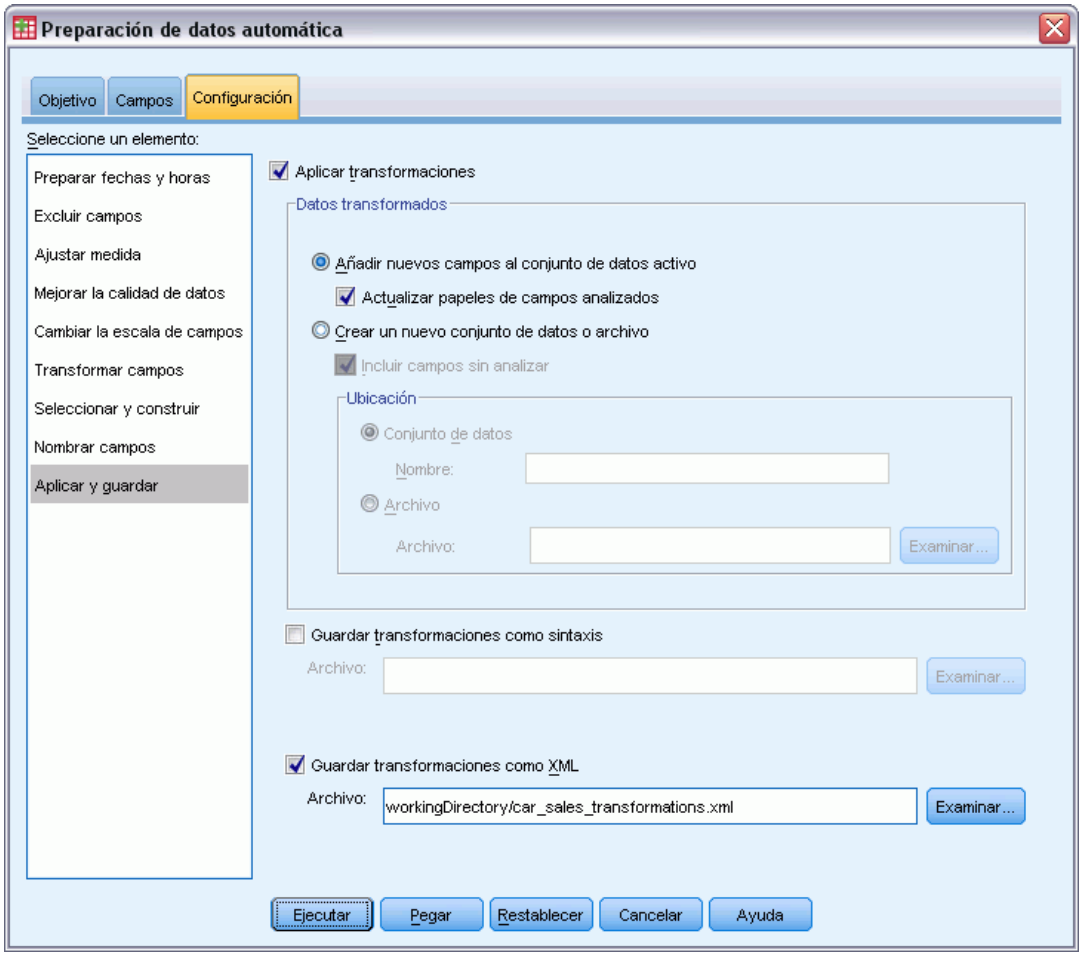


▶ Seleccione Optimizar precisión.

Como el campo de destino *Ventas en miles* es continuo y puede transformarse durante la preparación automática de datos, puede guardar las transformaciones en un archivo XML para poder utilizar el cuadro de diálogo Puntuaciones de transformación retrospectiva para convertir los valores predichos del destino transformado de nuevo a su escala original.

▶ Pulse en la pestaña Configuración y después en los ajustes Aplicar y guardar.

Figura 8-13
Ajustes Aplicar y guardar



- Seleccione Guardar transformaciones como XML y pulse en Examinar para desplazarse a directorioTrabajo/transformaciones_ventas_automoviles.xml, sustituyendo directorioTrabajo por la ruta en la que desea guardar el archivo.
- Pulse en Ejecutar.

Estas selecciones generan la siguiente sintaxis de comandos:

```
*Automatic Data Preparation.
ADP
  /FIELDS TARGET=sales INPUT=resale type price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /PREPDATETIME DATEDURATION=YES(REFERENCE=YMD('2009-06-04') UNIT=AUTO)
    TIMEDURATION=YES(REFERENCE=HMS('08:43:35') UNIT=AUTO) EXTRACTYEAR=YES(SUFFIX='_year')
    EXTRACTMONTH=YES(SUFFIX='_month') EXTRACTDAY=YES(SUFFIX='_day')
    EXTRACTHOUR=YES(SUFFIX='_hour') EXTRACTMINUTE=YES(SUFFIX='_minute')
    EXTRACTSECOND=YES(SUFFIX='_second')
   /SCREENING PCTMISSING=YES(MAXPCT=50) UNIQUECAT=YES(MAXCAT=100) SINGLECAT=NO
  /ADJUSTLEVEL INPUT=YES TARGET=YES MAXVALORDINAL=10 MINVALCONTINUOUS=5
  /OUTLIERHANDLING INPUT=YES TARGET=NO CUTOFF=SD(3) REPLACEWITH=CUTOFFVALUE
  /REPLACEMISSING INPUT=YES TARGET=NO
```

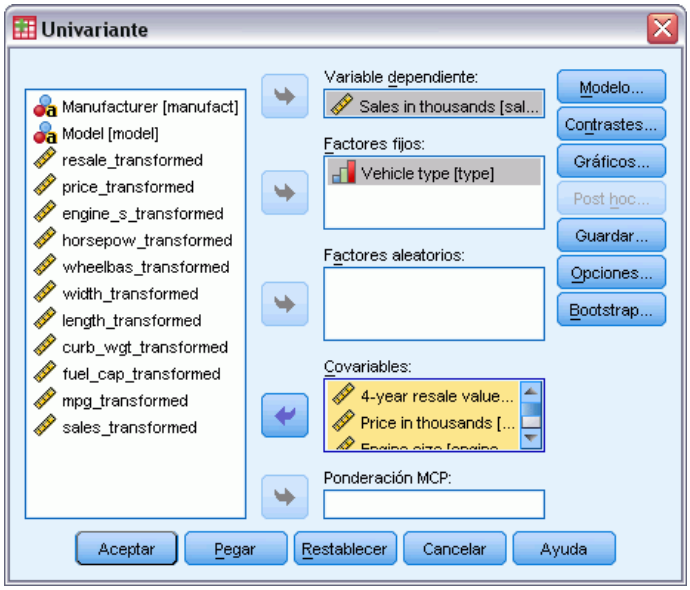
```
  /REORDERNOMINAL INPUT=YES TARGET=NO
  /RESCALE INPUT=ZSCORE(MEAN=0 SD=1) TARGET=BOXCOX(MEAN=0 SD=1)
  /TRANSFORM MERGESUPERVISED=NO MERGEUNSUPERVISED=NO BINNING=NONE SELECTION=NO
    CONSTRUCTION=NO
  /CRITERIA SUFFIX(TARGET='_transformed' INPUT='_transformed')
  /OUTFILE PREPXML='/workingDirectory/car_sales_transformations.xml'.
TMS IMPORT
  /INFILE TRANSFORMATIONS='/workingDirectory/car_sales_transformations.xml'
    MODE=FORWARD (ROLES=UPDATE)
  /SAVE TRANSFORMED=YES.
EXECUTE.
```

- El comando ADP prepara el campo de destino *ventas* y los campos de entrada *reventas* a *mpg*.
- Se especifica el subcomando PREPDATETIME, pero no se utiliza porque ninguno de los campos son de fecha u hora.
- El subcomando ADJUSTLEVEL reclasifica los campos ordinales con más de 10 valores como continuos, y los campos continuos con menos de 5 valores como ordinales.
- El subcomando OUTLIERHANDLING sustituye valores de entradas continuas (no el destino) que presenten más de 3 desviaciones estándar de la media por el valor que suponga 3 desviaciones estándar de la media.
- El subcomando REPLACEMISSING sustituye valores de entradas (no de destino) que falten.
- El subcomando REORDERNOMINAL recodifica los valores de las entradas nominales de la menos frecuente a la más frecuente.
- El subcomando RESCALE estandariza las entradas continuas para que tengan una media de 0 y una desviación estándar 1 mediante una transformación de puntuaciones z, además de estandarizar el destino continuo para que tenga una media 0 y una desviación estándar 1 mediante una transformación Box-Cox.
- El subcomando TRANSFORM desactiva todas las operaciones predeterminadas especificadas por este subcomando.
- El subcomando CRITERIA especifica los sufijos predeterminados para las transformaciones de los destinos y entradas.
- El subcomando OUTFILE especifica que las transformaciones deben guardarse en */directorioTrabajo/transformaciones_ventas_automoviles.xml*, donde */directorioTrabajo* es la ruta en la que guardar *transformaciones_ventas_automoviles.xml*.
- El comando TMS IMPORT lee las transformaciones de *transformaciones_ventas_automoviles.xml* y las aplica al conjunto de datos activo, actualizando los papeles de los campos existentes que se transforman.
- El comando EXECUTE provoca que se procesen las transformaciones. Si lo utiliza como parte de una transmisión de sintaxis más larga, puede borrar el comando EXECUTE para ahorra algo de tiempo de procesamiento.

Creación de un modelo de los datos sin preparar

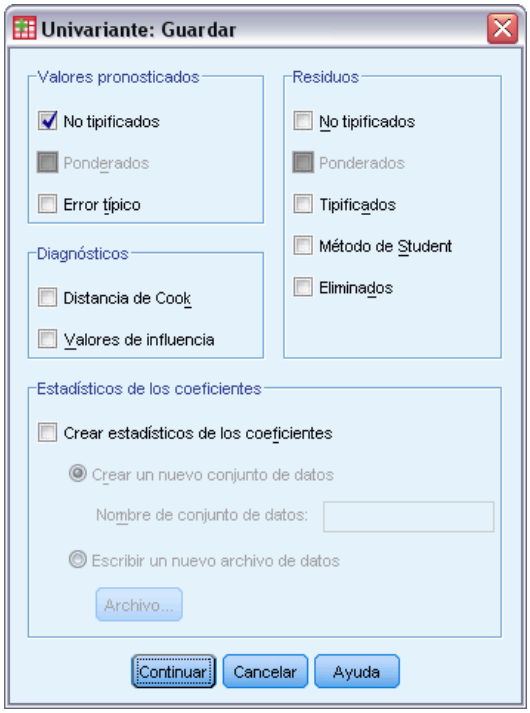
▶ Para ejecutar un modelo de los datos sin preparar, elija en los menús:
Analizar > Modelo lineal general > Univariante...

Figura 8-14
Cuadro de diálogo MLG Univariante



- Seleccione *Ventas en miles [ventas]* como la variable dependiente.
- Seleccione *Tipo de vehículo [tipo]* como factor fijo.
- Seleccione *Valor reventa 4 años [reventa]* a *Consumo de gasolina [mpg]* como covariables.
- Pulse en Guardar.

Figura 8-15
Cuadro de diálogo Guardar



- Seleccione No tipificados en el grupo Valores pronosticados.
- Pulse en Continuar.
- Pulse en Aceptar en el cuadro de diálogo MLG Univariante.

Estas selecciones generan la siguiente sintaxis de comandos:

```
UNIANOVA sales BY type WITH resale price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale price engine_s horsepow wheelbas width length curb_wgt fuel_cap
    mpg type.
```

Figura 8-16
Efectos inter-sujetos para modelos basados en datos sin preparar

Variable dependiente:Sales in thousands

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Modelo corregido	226123.658[a]	11	20556.696	5.050	.000
Intersección	12227.688	1	12227.688	3.004	.086
resale	50.702	1	50.702	.012	.911
price	471.630	1	471.630	.116	.734
engine_s	19872.712	1	19872.712	4.882	.029
horsepow	9644.486	1	9644.486	2.369	.127
wheelbas	29824.272	1	29824.272	7.327	.008
width	263.465	1	263.465	.065	.800
length	1374.525	1	1374.525	.338	.562
curb_wgt	32762.692	1	32762.692	8.049	.005
fuel_cap	1124.237	1	1124.237	.276	.600
mpg	337.585	1	337.585	.083	.774
type	17668.779	1	17668.779	4.341	.040
Error	427402.183	105	4070.497		
Total	1062354.955	117			
Total corregida	653525.841	116			

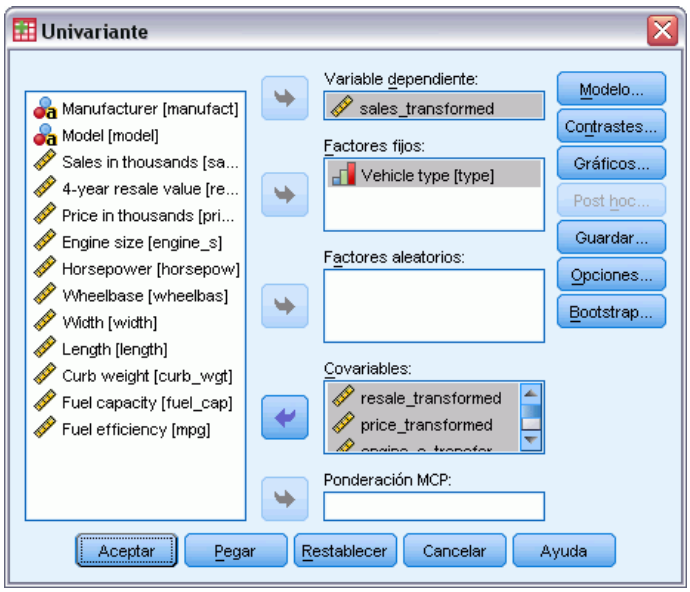
a. R cuadrado = .346 (R cuadrado corregida = .277)

Los resultados de MLG Univariante incluyen los efectos inter-sujetos, que son un análisis de la tabla de varianza. Cada término del modelo, además del modelo en conjunto, se comprueba para conocer su capacidad de afectar a la variación de la variable dependiente. Tenga en cuenta que las etiquetas de variables no se muestran en esta tabla.

Los predictores muestran niveles variables de significancia: aquellos con un valor de significancia inferior a 0,05 suelen considerarse útiles para el modelo.

Creación de un modelo de los datos preparados

Figura 8-17
Cuadro de diálogo MLG Univariante



- Para crear el modelo de los datos preparados, active el cuadro de diálogo MLG Univariante
- Cancele la selección de *Ventas en miles [ventas]* y seleccione *ventas_transformada* como la variable dependiente.
- Cancele la selección desde *Valor reventa 4 años [reventa]* a *Consumo de gasolina [mpg]* y seleccione desde *reventas_transformado* a *mpg_transformado* como covariables.
- Pulse en Aceptar.

Estas selecciones generan la siguiente sintaxis de comandos:

```
UNIANOVA sales_transformed BY type WITH resale_transformed price_transformed
    engine_s_transformed horsepow_transformed wheelbas_transformed width_transformed
    length_transformed curb_wgt_transformed fuel_cap_transformed mpg_transformed
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale_transformed price_transformed engine_s_transformed horsepow_transformed
    wheelbas_transformed width_transformed length_transformed curb_wgt_transformed
    fuel_cap_transformed mpg_transformed type.
```

Figura 8-18
Efectos inter-sujetos para modelos basados en datos preparados

Variable dependiente:sales_transformed

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Modelo corregido	79.327[a]	11	7.212	13.638	.000
Intersección	2.436	1	2.436	4.606	.034
resale_transformed	.954	1	.954	1.804	.181
price_transformed	9.271	1	9.271	17.533	.000
engine_s_transformed	2.885	1	2.885	5.456	.021
horsepow_transformed	.034	1	.034	.064	.801
wheelbas_transformed	1.213	1	1.213	2.293	.132
width_transformed	.037	1	.037	.071	.791
length_transformed	.265	1	.265	.501	.480
curb_wgt_transformed	.103	1	.103	.194	.660
fuel_cap_transformed	.132	1	.132	.249	.618
mpg_transformed	3.390	1	3.390	6.411	.012
type	4.007	1	4.007	7.579	.007
Error	76.673	145	.529		
Total	156.000	157			
Total corregida	156.000	156			

a. R cuadrado = .509 (R cuadrado corregida = .471)

Hay unas diferencias interesantes para tener en cuenta en los efectos inter-sujetos entre el modelo construido con los datos sin preparar y el construido con los datos preparados. En primer lugar, tenga en cuenta que los grados totales de libertad aumentada; esto se debe al hecho de que los valores que faltaban se sustituyeron con valores introducidos durante la preparación automática de datos, de forma que los registros que se eliminaron del primer modelo estaban disponibles en el segundo. Más notable es quizás que la significación de ciertos predictores ha cambiado. Aunque en ambos modelos el tamaño del motor [*tamaño_motor*] y el tipo de vehículo [*tipo*] se utilizan para el modelo, la distancia entre los ejes [*distejes*] y el peso neto del vehículo [*peso_neto*] dejan de ser significativos, mientras que el precio del vehículo [*precio_transformado*] y el consumo de gasolina [*mpg_transformado*] ahora sí lo son.

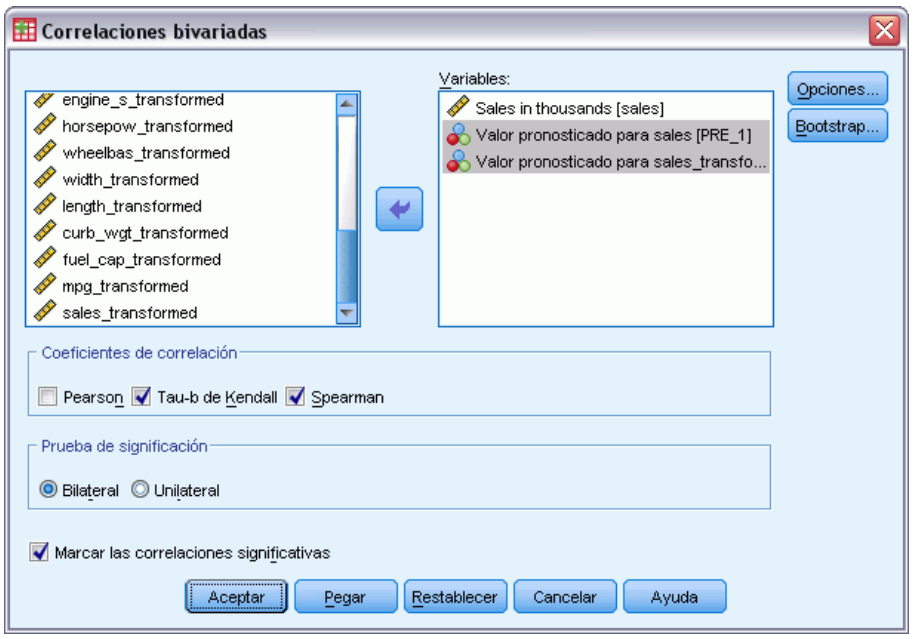
¿Por qué se ha producido este cambio? Las ventas tienen una distribución asimétrica, por lo que puede que la distancia entre ejes y el peso neto del vehículo tengan unos registros influyentes que dejen de serlo cuando se transforma la venta. Otra posibilidad es que los casos adicionales disponibles debido a la sustitución de valores que faltaban cambiaran la significación estadística de estas variables. En cualquier caso, esto exigiría una investigación más detallada que no se realizará en este apartado.

Tenga en cuenta que R cuadrado es mayor para el modelo construido con los datos preparados, pero como las ventas se han transformado, ésta puede no ser la mejor medida para comparar el rendimiento de cada modelo. En su lugar, puede calcular las correlaciones no paramétricas entre los valores observados y los dos conjuntos de valores predichos.

Comparación de los valores predichos

- Para obtener correlaciones de los valores predichos a partir de los dos modelos, elija en los menús: Analizar > Correlaciones > Bivariadas...

Figura 8-19
Cuadro de diálogo Correlaciones bivariadas



- Seleccione *Ventas en miles [ventas]*, *Valor predicho para ventas [PRE_1]* y *Valor predicho para ventas_transformado [PRE_2]* como variables de análisis.
- Cancele la selección de Pearson y seleccione tau-b de Kendall y rho de Spearman en el grupo Coeficientes de correlación.

  Tenga en cuenta que *Valor predicho para ventas_transformado [PRE_2]* puede utilizarse para calcular las correlaciones no paramétricas sin tener que realizar una transformación retrospectiva a la escala original, ya que la transformación retrospectiva no cambia el orden de clasificación de los valores predichos.
- Pulse en Aceptar.

Estas selecciones generan la siguiente sintaxis de comandos:

```
NONPAR CORR
  /VARIABLES=sales PRE_1 PRE_2
  /PRINT=BOTH TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Figura 8-20
Correlaciones no paramétricas

			Sales in thousands	Valor pronosticado para sales	Valor pronosticado para sales_ transformed
Tau_b de Kendall	Sales in thousands	Coeficiente de correlación	1.000	.376**	.484**
		Sig. (bilateral)	.	.000	.000
		N	157	117	157
	Valor pronosticado para sales	Coeficiente de correlación	.376**	1.000	.655**
		Sig. (bilateral)	.000	.	.000
		N	117	117	117
	Valor pronosticado para sales_transformed	Coeficiente de correlación	.484**	.655**	1.000
		Sig. (bilateral)	.000	.000	.
		N	157	117	157
Rho de Spearman	Sales in thousands	Coeficiente de correlación	1.000	.530**	.666**
		Sig. (bilateral)	.	.000	.000
		N	157	117	157
	Valor pronosticado para sales	Coeficiente de correlación	.530**	1.000	.831**
		Sig. (bilateral)	.000	.	.000
		N	117	117	117
	Valor pronosticado para sales_transformed	Coeficiente de correlación	.666**	.831**	1.000
		Sig. (bilateral)	.000	.000	.
		N	157	117	157

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

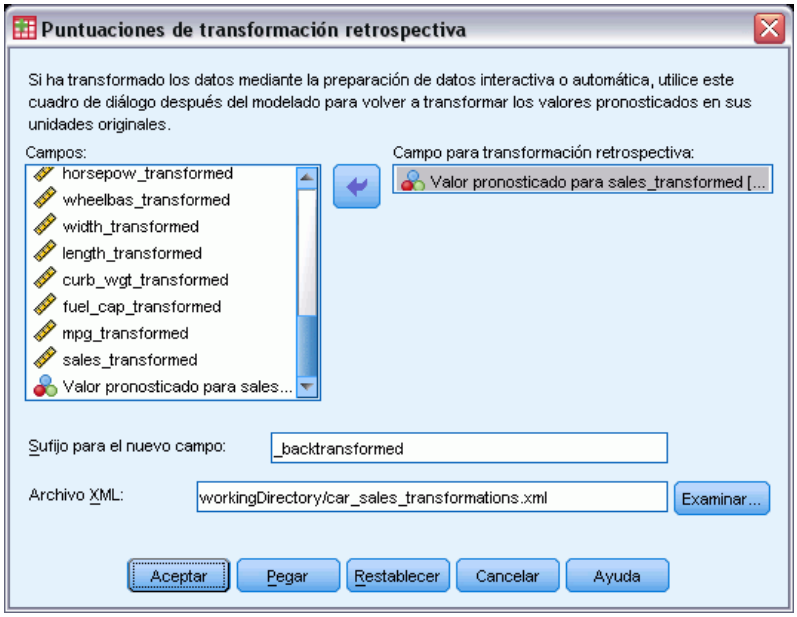
La primera columna muestra que los valores predichos para el modelo creado con los datos preparados están más relacionados con los valores observados por tanto las medidas tau-b de Kendall como rho de Spearman. Esto sugiere que la ejecución de la preparación automática de datos ha mejorado el modelo.

Transformación retrospectiva de los valores predichos

- Los datos preparados incluyen una transformación de las ventas, por lo que los valores predichos a partir de este modelo no pueden utilizarse directamente como puntuaciones. Para transformar los valores predichos a la escala original, seleccione en los menús:

  Transformar > Preparar datos para modelado > Puntuaciones de transformación retrospectiva...

Figura 8-21
Cuadro de diálogo Puntuaciones de transformación retrospectiva



- Seleccione *Valor predicho para ventas_transformado [PRE_2]* como campo al que aplicar la transformación retrospectiva.
- Escriba _transformado_retro como sufijo para el nuevo campo.
- Escriba directorioTrabajo\transformaciones_ventas_automoviles.xml, donde directorioTrabajo es la ubicación del archivo XML que contiene las transformaciones.
- Pulse en Aceptar.

Estas selecciones generan la siguiente sintaxis de comandos:

```
TMS IMPORT
  /INFILE TRANSFORMATIONS='workingDirectory/car_sales_transformations.xml'
    MODE=BACK (PREDICTED=PRE_2 SUFFIX='_backtransformed').
EXECUTE.
```

- El comando `TMS IMPORT` lee las transformaciones de *transformaciones_ventas_automoviles.xml* y aplica la transformación retrospectiva a *PRE_2*.
- El nuevo campo que contienen los valores transformados retrospectivamente se denomina *PRE_2_transformado_retro*.
- El comando `EXECUTE` provoca que se procesen las transformaciones. Si lo utiliza como parte de una transmisión de sintaxis más larga, puede borrar el comando `EXECUTE` para ahorra algo de tiempo de procesamiento.

Resumen

Con la preparación automática de datos puede obtener rápidamente transformaciones de los datos que mejoren su modelo. Si el destino se transforma, puede guardar las transformaciones en un archivo XML y utilizar el cuadro de diálogo Puntuaciones de transformación retrospectiva para convertir los valores predichos para el destino transformado de nuevo a su escala original.

Capítulo
9

Identificar casos atípicos

El procedimiento de detección de anomalías busca casos atípicos basados en desviaciones de las normas de sus agrupaciones. El procedimiento está diseñado para detectar rápidamente casos atípicos con fines de auditoría de datos en el paso del análisis exploratorio de datos, antes de llevar a cabo cualquier análisis de datos inferencial. Este algoritmo está diseñado para la detección de anomalías genéricas; es decir, la definición de un caso anómalo no es específica de ninguna aplicación particular, como la detección de patrones de pago atípicos en la industria sanitaria ni la detección de blanqueo de dinero en la industria financiera, donde la definición de una anomalía puede estar bien definida.

Algoritmo para identificar casos atípicos

Este algoritmo se divide en tres etapas:

Modelado. El procedimiento crea un modelo de agrupación que explica los grupos naturales (o conglomerados) dentro de un conjunto de datos que de otro modo no serían evidentes. La agrupación se basa en un conjunto de variables de entrada. El modelo de agrupación resultante y los estadísticos suficientes para calcular las normas de agrupación se almacenan para su posterior uso.

Puntuación. El modelo se aplica a cada uno de los casos para identificar su grupo y se crean algunos índices para cada caso con el objeto de medir la atipicidad del caso respecto a su propio grupo. Todos los casos se ordenan según los valores de los índices de anomalía. La parte superior de la lista de casos se identifica como el conjunto de anomalías.

Razonamiento. Para cada uno de los casos anómalos, se ordenan las variables por sus correspondientes índices de desviación de las variables. Las variables con los índices más altos, sus valores y los valores de norma correspondientes se presentan como los motivos por los que un caso se identifica como una anomalía.

Identificación de casos atípicos en una base de datos médica

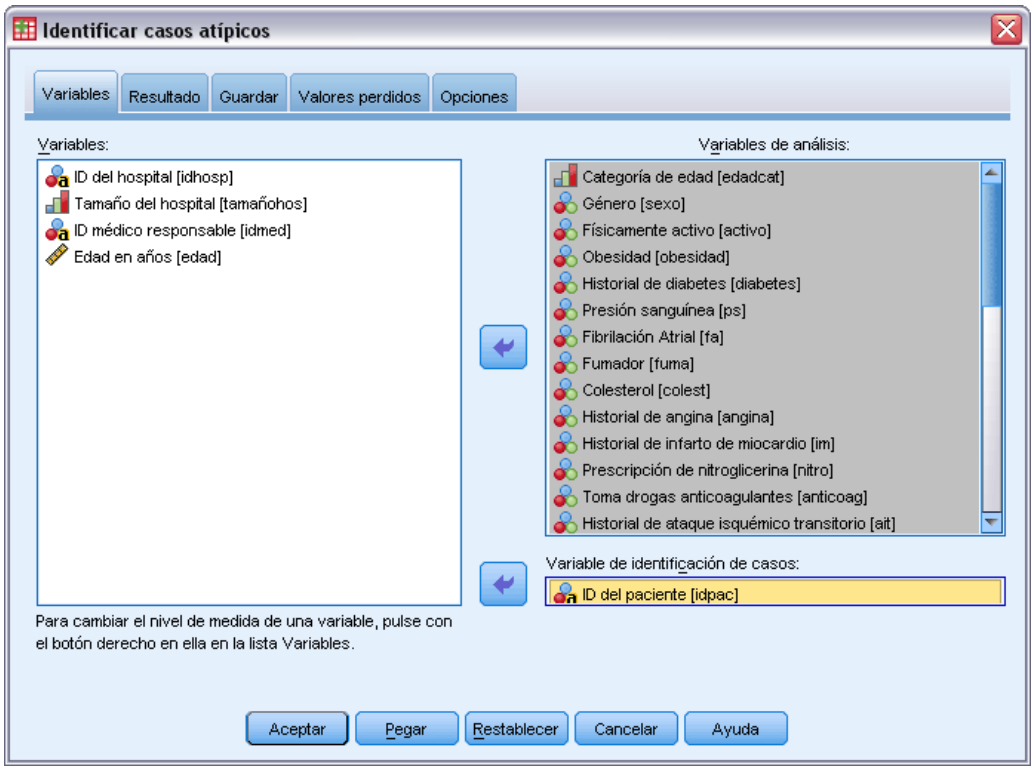
Un analista de datos contratado para generar modelos predictivos para los resultados de los tratamientos de derrames cerebrales se preocupa por la calidad de los datos ya que tales modelos pueden ser sensibles a observaciones atípicas. Algunas de estas observaciones atípicas representan casos verdaderamente únicos y, por lo tanto, no son adecuadas para la predicción, mientras que otras observaciones están provocadas por errores de entrada de datos donde los valores son técnicamente “correctos” y no pueden ser tomados por los procedimientos de validación de datos.

Esta información se recoge en el archivo *stroke_valid.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A el p. 138. Utilice el procedimiento Identificar casos atípicos para limpiar el archivo de datos. Puede encontrar la sintaxis para reproducir estos análisis en *detectanomaly_stroke.sps*.

Ejecución del análisis

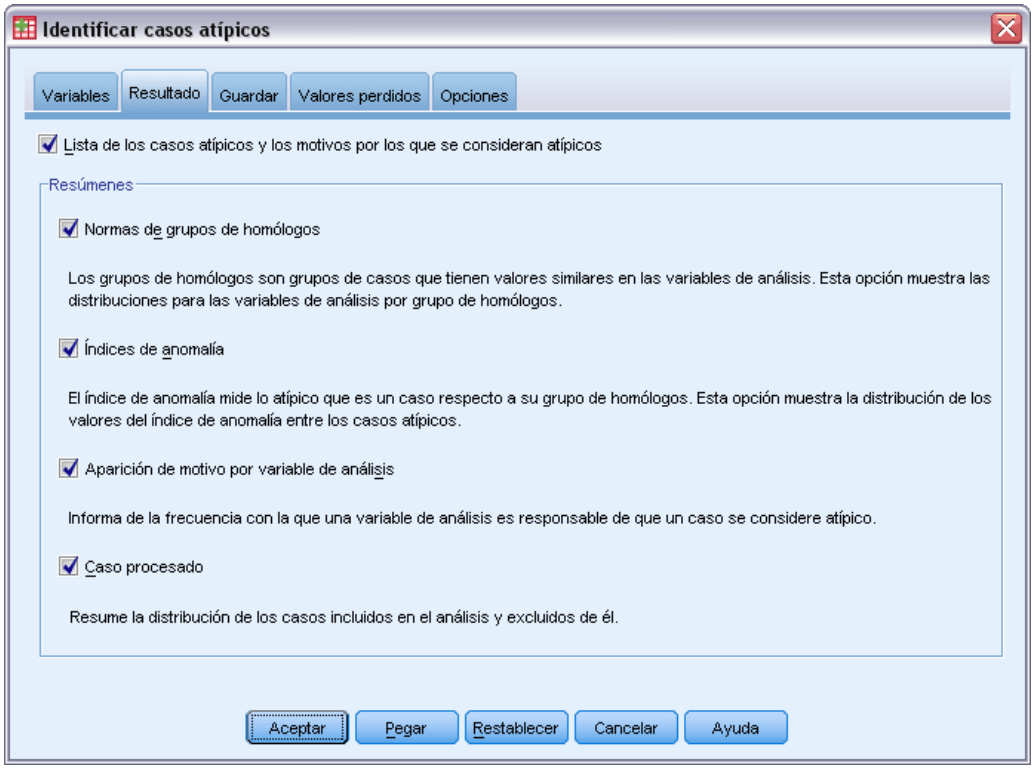
- Para identificar casos atípicos, elija en los menús:
  Datos > Identificar casos atípicos...

Figura 9-1
Cuadro de diálogo Identificar casos atípicos, pestaña Variables



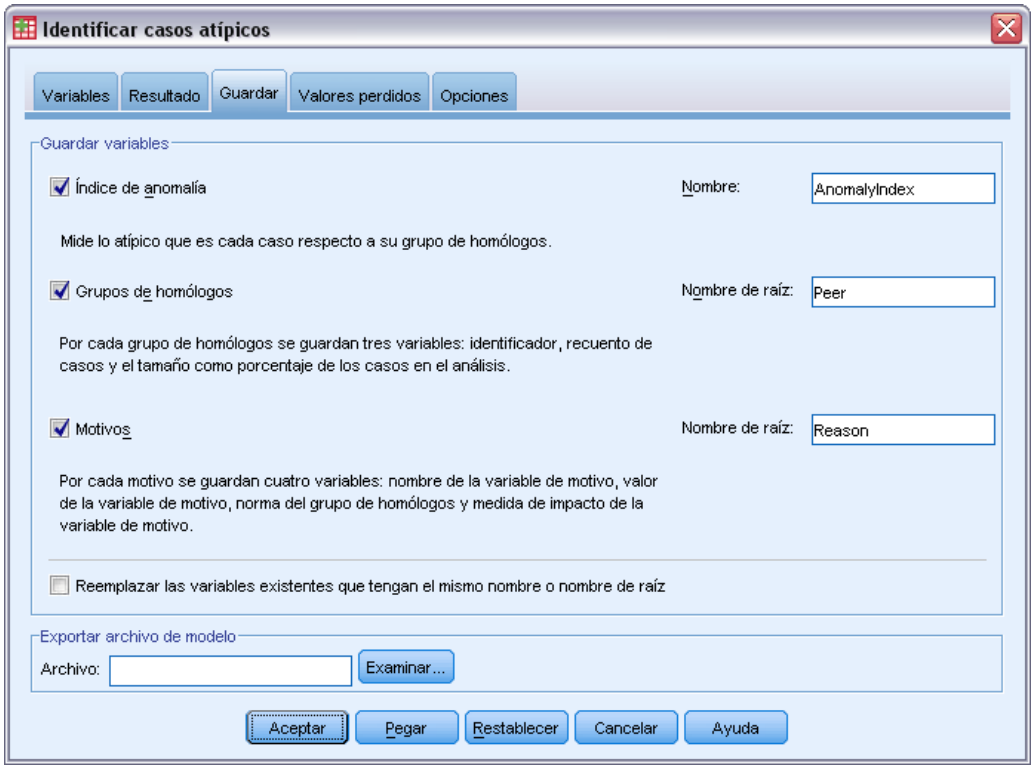
- Seleccione desde *Categoría de edad* hasta *Infarto entre meses 3 y 6* como variables del análisis.
- Seleccione *ID del paciente* como la variable de identificación de casos.
- Pulse en la pestaña Resultados.

Figura 9-2
Cuadro de diálogo Identificar casos atípicos, pestaña Resultado



- Seleccione Normas de grupos de homólogos, Índices de anomalía, Aparición de motivo por variable de análisis y Casos procesados.
- Pulse en la ficha Guardar.

Figura 9-3
Cuadro de diálogo Identificar casos atípicos, pestaña Guardar

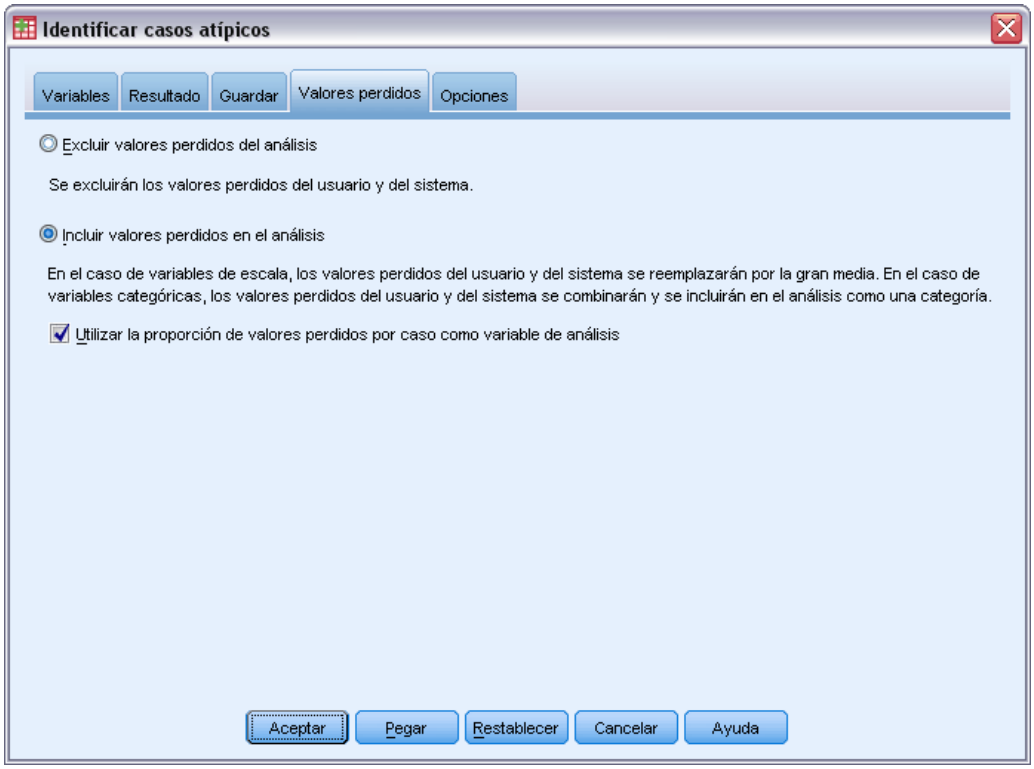


- Seleccione Índice de anomalía, Grupos de homólogos y Motivos.

  Guardar estos resultados permite generar un útil diagrama de dispersión que resuma los resultados.

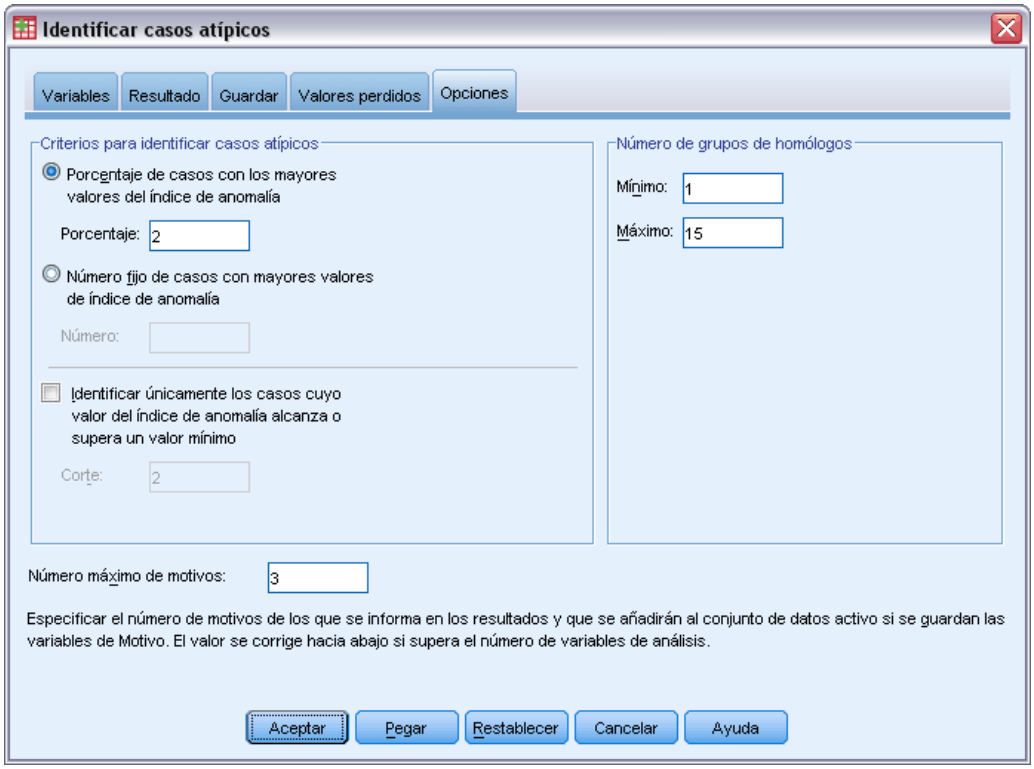
- Pulse en la pestaña Valores perdidos.

Figura 9-4
Cuadro de diálogo Identificar casos atípicos, pestaña Valores perdidos



- Seleccione Incluir valores perdidos en el análisis. Este proceso es necesario porque hay muchos valores definidos como perdidos por el usuario para manejar pacientes que murieron antes o durante el tratamiento. Una variable adicional que mide la proporción de valores perdidos por caso se añade al análisis como una variable de escala.
- Pulse en la ficha Opciones.

Figura 9-5
Cuadro de diálogo Identificar casos atípicos, pestaña Opciones



- Escriba 2 como porcentaje de casos que considerar anómalos.
- Anule la selección de Identificar únicamente los casos cuyo valor del índice de anomalía alcanza o supera un valor mínimo.
- Escriba 3 como el número máximo de motivos.
- Pulse en Aceptar.

Resumen de procesamiento de casos

Figura 9-6
Resumen del procesamiento de los casos

		N	% de los combinados	% del total
Id de homólogos	1	710	67,7%	67,7%
	2	90	8,6%	8,6%
	3	248	23,7%	23,7%
Combinado		1048	100,0%	100,0%
Total		1048		100,0%

Cada caso se categoriza en un grupo de homólogos de casos similares. El resumen de procesamiento de casos muestra el número de grupos de homólogos que se han creado, así como el número y porcentaje de casos que hay en cada grupo de homólogos.

Lista de índices de casos con anomalías

Figura 9-7
Lista de índices de casos con anomalías

Caso	idpac	Índice de anomalías
843	7840326167	2,837
510	0714726620	2,022
623	6553808330	2,014
501	6461046805	2,002
607	1077125669	1,897
884	2260043998	1,889
614	4030164769	1,869
241	1038840465	1,865
13	2191527525	1,826
172	4458028382	1,786
705	1336411777	1,778
651	4103977868	1,767
384	2247641363	1,767
839	0437454972	1,766
861	9746101913	1,757
19	7237535360	1,756
806	4391632997	1,756
871	6961938294	1,739
239	7315965190	1,738
887	6044244232	1,737
245	0816869249	1,736

El índice de anomalía es una medida que refleja la atipicidad del caso respecto al grupo de homólogos. Se muestra el 2 % de los casos con mayores valores de índice de anomalía, junto con sus ID y sus números de caso. En la lista aparecen 21 casos, con valores que van desde 1,736 hasta 2,837. Hay una diferencia relativamente grande en el valor del índice de anomalía entre el primer y el segundo caso de la lista, lo que sugiere que es probable que el caso 843 sea anómalo. Los demás casos se deberán juzgar caso por caso.

Lista de ID de los homólogos de casos con anomalías

Figura 9-8
Lista de ID de los homólogos de casos con anomalías

Caso	idpac	Id de homólogos	Tamaño de homólogos	Tamaño porcentual de homólogos
843	7840326167	3	248	23,7%
510	0714726620	3	248	23,7%
623	6553808330	3	248	23,7%
501	6461046805	3	248	23,7%
607	1077125669	3	248	23,7%
884	2260043998	3	248	23,7%
614	4030164769	3	248	23,7%
241	1038840465	3	248	23,7%
13	2191527525	3	248	23,7%
172	4458028382	3	248	23,7%
705	1336411777	1	710	67,7%
651	4103977868	1	710	67,7%
384	2247641363	3	248	23,7%
839	0437454972	3	248	23,7%
861	9746101913	3	248	23,7%
19	7237535360	1	710	67,7%
806	4391632997	1	710	67,7%
871	6961938294	1	710	67,7%
239	7315965190	3	248	23,7%
887	6044244232	1	710	67,7%
245	0816869249	3	248	23,7%

Se muestran los casos que potencialmente presentan anomalías junto con la información de pertenencia a sus grupos de homólogos. Los primeros 10 casos, y 15 casos en total, pertenecen al grupo de homólogos 3, mientras que el resto pertenece al grupo de homólogos 1.

Lista de motivos de casos con anomalías

Figura 9-9
Lista de motivos de casos con anomalías

Caso	idpac	Variable de motivo	Impacto de la variable	Valor de la variable	Norma de la variable
843	7840326167	coste	,411	200,51	19,83
510	0714726620	coste	,120	96,59	19,83
623	6553808330	coste	,175	114,01	19,83
501	6461046805	barthel1	,084	80	(Valor perdido)
607	1077125669	coste	,126	96,11	19,83
884	2260043998	coste	,138	99,73	19,83
614	4030164769	rango1	,085	3	(Valor perdido)
241	1038840465	barthel1	,115	25	(Valor perdido)
13	2191527525	barthel1	,118	40	(Valor perdido)
172	4458028382	barthel1	,120	100	(Valor perdido)
705	1336411777	coste	,244	198,25	42,47
651	4103977868	barthel1	,064	30	95
384	2247641363	barthel1	,122	20	(Valor perdido)
839	0437454972	barthel1	,109	95	(Valor perdido)
861	9746101913	barthel1	,102	70	(Valor perdido)
19	7237535360	barthel3	,080	5	100
806	4391632997	barthel2	,088	10	100
871	6961938294	barthel1	,094	5	95
239	7315965190	rango1	,092	3	(Valor perdido)
887	6044244232	infart1	,066	1	0
245	0816869249	barthel1	,124	5	(Valor perdido)

Las variables de motivos son las que más contribuyen a que un caso sea clasificado como atípico. Se muestra la variable del motivo principal de cada caso con anomalías, junto con la medida del impacto, el valor para ese caso y la norma de los grupos de homólogos. La norma del grupo de homólogos *(Valor perdido)* para una variable categórica indica que la mayoría de los casos del grupo de homólogos tiene un valor perdido para la variable.

El estadístico de impacto de la variable es la contribución proporcional de la variable de motivo a la desviación del caso respecto a su grupo de homólogos. Con 38 variables en el análisis, incluyendo la variable de proporción de valores perdidos, el impacto esperado de una variable debería ser de 1/38 = 0,026. El impacto de la variable *coste* en el caso 843 es de 0,411, lo que es relativamente grande. El valor de *coste* para el caso 843 es de 200,51, comparado con el valor de la media, 19,83, de los casos del grupo de homólogos 3.

Las selecciones solicitadas del cuadro de diálogo dan como resultado los tres motivos principales.

- Para ver los resultados de los demás motivos, pulse dos veces la tabla para activarla.
- Desplace *Motivo* desde la dimensión de capa a la dimensión de fila.

Figura 9-10
Lista de motivos de casos con anomalías (primeros 8 casos)

Caso	Motivo	idpac	Variable de motivo	Impacto de la variable	Valor de la variable	Norma de la variable
843	1	7840326167	coste	,411	200,51	19,83
	2	7840326167	barthel1	,076	65	(Valor perdido)
	3	7840326167	rango1	,044	2	(Valor perdido)
510	1	0714726620	coste	,120	96,59	19,83
	2	0714726620	barthel1	,083	80	(Valor perdido)
	3	0714726620	rehab	,068	3	(Valor perdido)
623	1	6553808330	coste	,175	114,01	19,83
	2	6553808330	cirugia	,089	2	(Valor perdido)
	3	6553808330	barthel1	,089	70	(Valor perdido)
501	1	6461046805	barthel1	,084	80	(Valor perdido)
	2	6461046805	rehab	,068	3	(Valor perdido)
	3	6461046805	rango1	,063	1	(Valor perdido)
607	1	1077125669	coste	,126	96,11	19,83
	2	1077125669	barthel1	,094	85	(Valor perdido)
	3	1077125669	rehab	,072	3	(Valor perdido)
884	1	2260043998	coste	,138	99,73	19,83
	2	2260043998	barthel1	,114	65	(Valor perdido)
	3	2260043998	rehab	,072	3	(Valor perdido)
614	1	4030164769	rango1	,085	3	(Valor perdido)
	2	4030164769	barthel1	,085	45	(Valor perdido)
	3	4030164769	recbart1	,062	2	(Valor perdido)
241	1	1038840465	barthel1	,115	25	(Valor perdido)
	2	1038840465	rango1	,103	4	(Valor perdido)
	3	1038840465	recbart1	,090	1	(Valor perdido)

Esta configuración facilita la comparación de las contribuciones relativas de los tres principales motivos de cada caso. El caso 843, como se sospecha, se considera anómalo porque el valor de *coste* es atípicamente alto. Por el contrario, no hay ningún motivo que por sí solo contribuya en más de 0,10 a la atipicidad del caso 501.

Normas de variables de escala

Figura 9-11
Normas de variables de escala

		Id de homólogos			
		1	2	3	Combinado
Duración de la estancia de rehabilitación	Media	16,55	16,39	15,91	16,39
	Desviación típica	12,596	,000	6,834	10,887
Coste total de tratamiento y rehabilitación en miles	Media	42,4673	3,5089	19,8273	33,7641
	Desviación típica	26,45401	,50997	20,17309	27,31266
Proporción de perdidos	Media	,006	,541	,354	,134
	Desviación típica	,021	,000	,083	,197

Las normas de variables de escala muestran la media y la desviación típica de cada variable para cada grupo de homólogos y en general. La comparación de los valores ofrece cierta información sobre cuáles son las variables que contribuyen a la formación de los grupos de homólogos.

Por ejemplo, la media de *Duración de la estancia de rehabilitación* es bastante constante en los tres grupos de homólogos, lo que significa que esa variable no contribuye a la formación de los grupos de homólogos. Por el contrario, *Coste total de tratamiento y rehabilitación en miles* y *Proporción de perdidos* ofrecen cierta información sobre la pertenencia a los grupos de homólogos. El grupo de homólogos 1 tiene la mayor media de coste y el menor número de valores perdidos. El grupo de homólogos 2 tiene un coste muy pequeño y muchos valores perdidos. El grupo de homólogos 3 tiene valores intermedios de coste y de valores perdidos.

Esta organización sugiere que el grupo de homólogos 2 está compuesto por pacientes que ingresaron cadáver, por lo que incurrieron en un coste muy pequeño e hicieron que todas las variables de tratamiento y rehabilitación tengan valores perdidos. Muy probablemente, el grupo de homólogos 3 contenga muchos pacientes que murieron durante el tratamiento, por lo que incurrieron en los costes de tratamiento pero no en los de rehabilitación, haciendo que las variables de rehabilitación tengan valores perdidos. Es muy probable que el grupo de homólogos 1 esté compuesto casi completamente por pacientes que sobrevivieron al tratamiento y a la rehabilitación, incurriendo, por lo tanto, en los mayores costes.

Normas de variables categóricas

Figura 9-12
Normas de variables categóricas (primeras 10 variables)

		Id de homólogos			
		1	2	3	Combinado
Categoría de edad	Categoría más popular	2	3	2	2
	Frecuencia	277	25	81	383
	Porcentaje	39,0%	27,8%	32,7%	36,5%
Género	Categoría más popular	0	0	1	0
	Frecuencia	361	46	126	529
	Porcentaje	50,8%	51,1%	50,8%	50,5%
Físicamente activo	Categoría más popular	1	0	0	0
	Frecuencia	373	55	139	531
	Porcentaje	52,5%	61,1%	56,0%	50,7%
Obesidad	Categoría más popular	0	0	0	0
	Frecuencia	555	67	178	800
	Porcentaje	78,2%	74,4%	71,8%	76,3%
Historial de diabetes	Categoría más popular	0	0	0	0
	Frecuencia	665	80	219	964
	Porcentaje	93,7%	88,9%	88,3%	92,0%
Presión sanguínea	Categoría más popular	1	1	1	1
	Frecuencia	445	49	139	633
	Porcentaje	62,7%	54,4%	56,0%	60,4%
Fibrilación Atrial	Categoría más popular	0	0	0	0
	Frecuencia	641	83	216	940
	Porcentaje	90,3%	92,2%	87,1%	89,7%
Fumador	Categoría más popular	0	0	0	0
	Frecuencia	578	69	179	826
	Porcentaje	81,4%	76,7%	72,2%	78,8%
Colesterol	Categoría más popular	0	0	0	0
	Frecuencia	406	52	136	594
	Porcentaje	57,2%	57,8%	54,8%	56,7%
Historial de angina	Categoría más popular	0	0	0	0
	Frecuencia	493	52	167	712
	Porcentaje	69,4%	57,8%	67,3%	67,9%

Las normas de variables categóricas tienen casi el mismo propósito que las variables de escala, aunque las normas de variables categóricas informan de la categoría modal (más popular), así como del número y porcentaje de casos del grupo de homólogos que hay en dicha categoría. Comparar los valores puede ser engañoso; por ejemplo, a primera vista, puede parecer que *Género* contribuye más a la formación de agrupaciones que *Fumador* porque la categoría modal de *Fumador* es la misma para los tres grupos de homólogos, mientras que la categoría modal de *Género* difiere en el grupo de homólogos 3. Sin embargo, como *Género* sólo tiene dos valores, se puede inferir que el 49,2% de los casos del grupo de homólogos 3 tiene un valor igual a 0, que es un porcentaje muy similar al que presentan los demás grupos de homólogos. Por lo contrario, los porcentajes de *Fumador* oscilan entre el 72,2% y el 81,4%.

Figura 9-13
Normas de variables categóricas (variables seleccionadas)

		Id de homólogos			
		1	2	3	Combinado
Ingresó cadáver	Categoría más popular	0	1	0	0
	Frecuencia	710	90	248	958
	Porcentaje	100,0%	100,0%	100,0%	91,4%
Puntuación de valoración inicial	Categoría más popular	0	(Valor perdido)	5	5
	Frecuencia	166	90	104	193
	Porcentaje	23,4%	100,0%	41,9%	18,4%
Resultado de la exploración TAC	Categoría más popular	0	(Valor perdido)	0	0
	Frecuencia	607	90	184	791
	Porcentaje	85,5%	100,0%	74,2%	75,5%
Drogas trombolíticas	Categoría más popular	2	(Valor perdido)	0	2
	Frecuencia	318	90	129	394
	Porcentaje	44,8%	100,0%	52,0%	37,6%
Exitus en el hospital	Categoría más popular	0	(Valor perdido)	1	0
	Frecuencia	710	90	171	787
	Porcentaje	100,0%	100,0%	69,0%	75,1%
Resultado del tratamiento	Categoría más popular	1	(Valor perdido)	1	1
	Frecuencia	524	90	96	620
	Porcentaje	73,8%	100,0%	38,7%	59,2%
Cirugía preventiva post-evento	Categoría más popular	0	(Valor perdido)	(Valor perdido)	0
	Frecuencia	323	90	171	369
	Porcentaje	45,5%	100,0%	69,0%	35,2%
Rehabilitación post-evento	Categoría más popular	0	(Valor perdido)	(Valor perdido)	0
	Frecuencia	278	90	171	314
	Porcentaje	39,2%	100,0%	69,0%	30,0%

Las sospechas que surgieron con las normas de las variables de escala se confirman en la tabla de normas de variables categóricas. El grupo de homólogos 2 está totalmente compuesto por pacientes que ingresaron cadáver, de forma que las variables de tratamiento y rehabilitación tienen valores perdidos. La mayoría de los pacientes del grupo de homólogos 3 (69,0%) murieron durante el tratamiento, por lo que la categoría modal para las variables de rehabilitación es *(Valor perdido)*.

Resumen de índice de anomalía.

Figura 9-14
Resumen de índice de anomalía

	N en la lista de anomalías	Mínimo	Máximo	Media	Desviación típica
Índice de anomalías	21	1,736	2,837	1,872	,240

N en la Lista de anomalías se determina por la especificación: El porcentaje de anomalía es 2%

La tabla proporciona estadísticos de resumen para los valores de los índices de anomalía de los casos incluidos en la lista de anomalías.

Resumen de motivos

Figura 9-15
Resumen de motivos (variables de tratamiento y rehabilitación)

	Aparición como motivo		Estadísticos del impacto de las variables			
	Frecuencia	Porcentaje	Mínimo	Máximo	Media	Desviación típica
Ingresó cadáver	0	,0%	.	.	.	.
Puntuación de valoración inicial	0	,0%	.	.	.	.
Resultado de la exploración TAC	0	,0%	.	.	.	.
Drogas trombolíticas	0	,0%	.	.	.	.
Exitus en el hospital	0	,0%	.	.	.	.
Resultado del tratamiento	0	,0%	.	.	.	.
Cirugía preventiva post-evento	0	,0%	.	.	.	.
Rehabilitación post-evento	0	,0%	.	.	.	.
Puntación de valoración al mes 1	2	9,5%	,085	,092	,089	,005
Puntación de valoración al mes 3	0	,0%	.	.	.	.
Puntación de valoración al mes 6	0	,0%	.	.	.	.
Índice de Barthel al mes 1	10	47,6%	,064	,124	,105	,019
Índice de Barthel al mes 3	1	4,8%	,088	,088	,088	.
Índice de Barthel al mes 6	1	4,8%	,080	,080	,080	.
Índice de Barthel recodificado al mes 1	0	,0%	.	.	.	.
Índice de Barthel recodificado al mes 3	0	,0%	.	.	.	.
Índice de Barthel recodificado al mes 6	0	,0%	.	.	.	.
Infarto entre el alta y mes 1	1	4,8%	,066	,066	,066	.
Infarto entre meses 1 y 3	0	,0%	.	.	.	.
Infarto entre meses 3 y 6	0	,0%	.	.	.	.
Duración de la estancia de rehabilitación	0	,0%	.	.	.	.
Coste total de tratamiento y rehabilitación en miles	6	28,6%	,120	,411	,202	,112
Proporción de perdidos	0	,0%	.	.	.	.
Global	21	100,0%	,064	,411	,127	,076

Para cada variable del análisis, la tabla resume el papel de la variable como un motivo principal. La mayoría de las variables, como las variables desde *Ingresó cadáver* hasta *Rehabilitación post-evento*, no son el motivo principal para que ninguno de los casos esté en la lista de anomalías. *Índice de Barthel al mes 1* es el motivo más frecuente, seguido de *Coste total de tratamiento y rehabilitación en miles*. Los estadísticos que evalúan el impacto de las variables aparecen

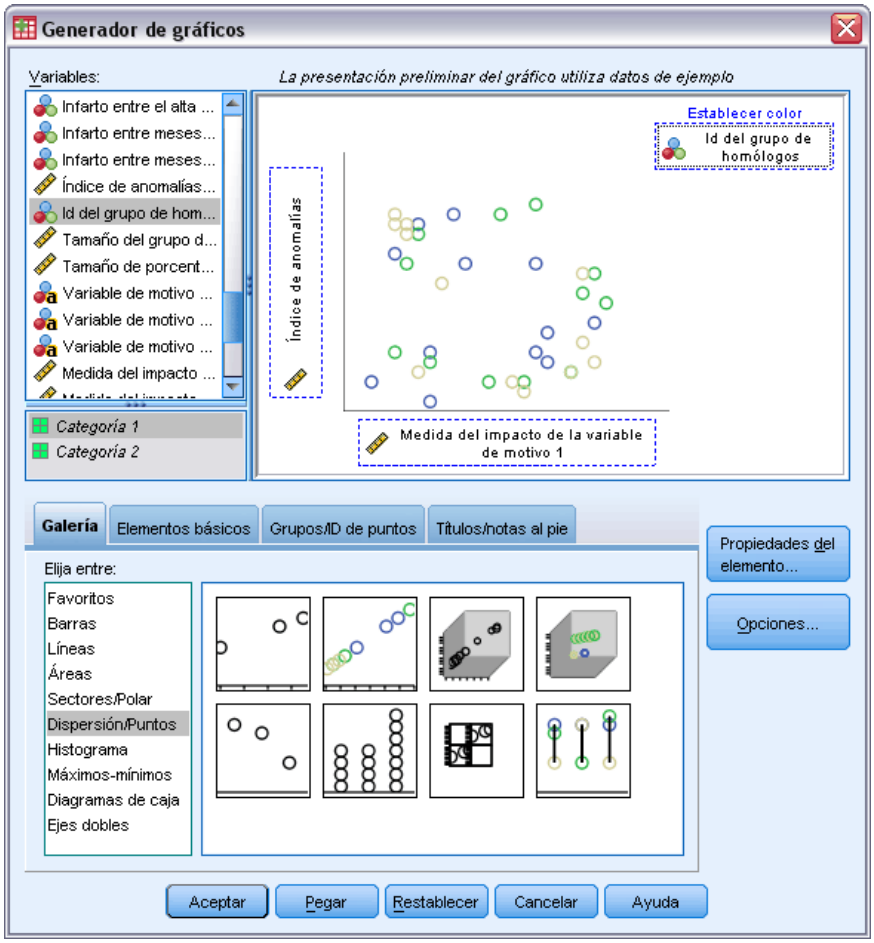
resumidos, con el impacto mínimo, máximo y medio de cada variable, junto con la desviación típica para las variables que sean motivo de más de un caso.

Diagrama de dispersión del índice de anomalía por impacto de las variables

Las tablas contienen gran cantidad de información útil, pero puede ser difícil extraer las relaciones. Utilizando las variables guardadas, se puede construir un gráfico que simplifique este proceso.

- Para generar este diagrama de dispersión, elija en los menús:
  Gráficos > Generador de gráficos...

Figura 9-16
Cuadro de diálogo Generador de gráficos

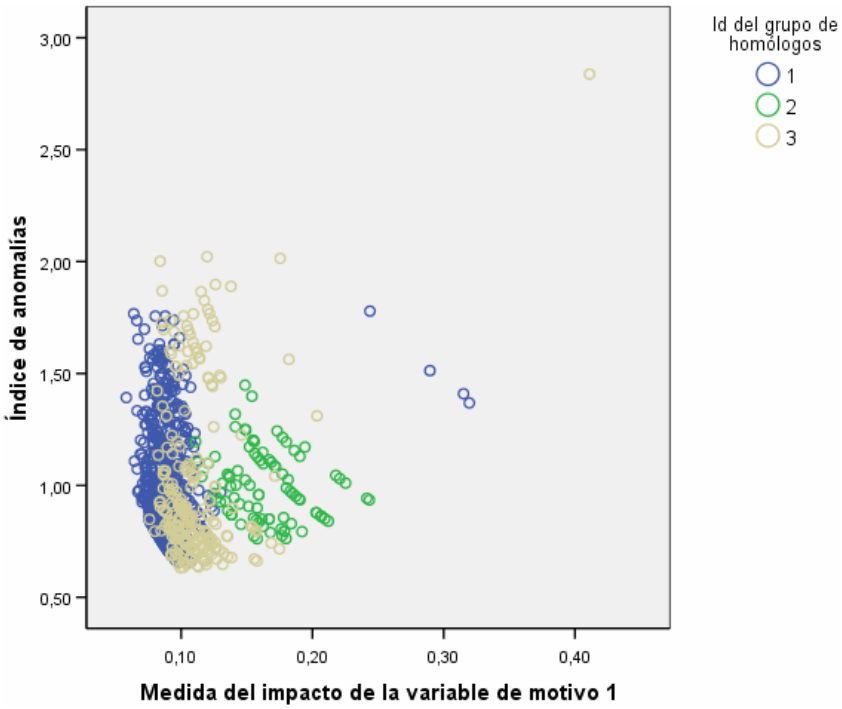


- Seleccione la galería Dispersión/Puntos y arrastre el icono de diagrama de dispersión agrupado al lienzo.
- Seleccione *Índice de anomalía* como variable *y* y *Medida del impacto de la variable de motivo 1* como variable *x*.

- Seleccione *ID de grupo de homólogos* como la variable por la que establecer colores.
- Pulse en Aceptar.

Con estas selecciones se obtiene el diagrama de dispersión.

Figura 9-17
Diagrama de dispersión de índice de anomalía por medida del impacto de la primera variable de motivo



La inspección del gráfico conduce a varias observaciones:

- El caso situado en la esquina superior derecha pertenece al grupo de homólogos 3 y es tanto el caso más anómalo como el caso con la mayor contribución realizada por una única variable.
- Al bajar por el eje *y*, vemos que hay tres casos que pertenecen al grupo de homólogos 3, con valores de índice de anomalía justo por encima de 2,00. Estos casos se deberían investigar con más detalle como anómalos.
- Al recorrer el eje *x*, vemos que hay cuatro casos que pertenecen al grupo de homólogos 1, con medidas de impacto de variables situadas aproximadamente entre 0,23 y 0,33. Estos casos se deberían investigar con mayor profundidad porque esos valores separan a los casos del cuerpo principal de puntos del diagrama.
- El grupo de homólogos 2 parece bastante homogéneo en el sentido de que los valores de índice de anomalía y de impacto de variable no varían mucho de las tendencias centrales.

Resumen

La utilización del procedimiento Identificar casos atípicos ha permitido detectar varios casos que requieren un examen más detallado. Dichos casos no se habrían identificado mediante otros procedimientos de validación, ya que las relaciones entre las variables (no sólo los valores de las propias variables) determinan los casos anómalos.

En cierta forma, es decepcionante que los grupos de homólogos se basen sobre todo en dos variables: *Ingresó cadáver* y *Exitus en el hospital*. En análisis más detallados, se puede estudiar el efecto de forzar la creación de un mayor número de grupos de homólogos o realizar un análisis que incluya sólo pacientes que hayan sobrevivido al tratamiento.

Procedimientos relacionados

El procedimiento Identificar casos atípicos es una herramienta muy útil para detectar casos con anomalías en el archivo de datos.

- El procedimiento Validar datos permite identificar casos, variables y valores de datos no válidos o sospechosos en el conjunto de datos activo.

Capítulo
10

Intervalos óptimos

El procedimiento Intervalos óptimos discretiza una o más variables de escala (a las que se denomina **variables de entrada que se van a agrupar**) mediante la distribución de los valores de cada variable en intervalos. La formación de intervalos es óptima en relación con una variable guía categórica que "supervisa" el proceso de agrupación. Los intervalos se pueden utilizar en lugar de los valores de datos originales en análisis posteriores en procedimientos que requieren o prefieren variables categóricas.

Algoritmo Intervalos óptimos

Los pasos básicos del algoritmo Intervalos óptimos se caracterizan como se indica a continuación:

Procesamiento previo (opcional). La variable de entrada que se va a agrupar en *n* intervalos (donde el usuario especifica el valor de *n*), y cada intervalo contiene el mismo número de casos o una cifra lo más cercana posible a un mismo número de casos.

Identificación de puntos de corte potenciales. Cada valor distinto de la entrada que se va a agrupar que no pertenece a la misma categoría de la variable guía como el siguiente valor distinto superior de la variable de entrada que se va a agrupar es un punto de corte potencial.

Selección de puntos de corte. El punto de corte potencial que produce la mayor ganancia de información se evalúa mediante el criterio de aceptación MDLP. Estos pasos se repiten hasta que no se encuentran más puntos de corte potenciales. Los puntos de corte aceptados definen los límites de los intervalos.

Uso de Intervalos óptimos para discretizar los datos de los solicitantes de créditos

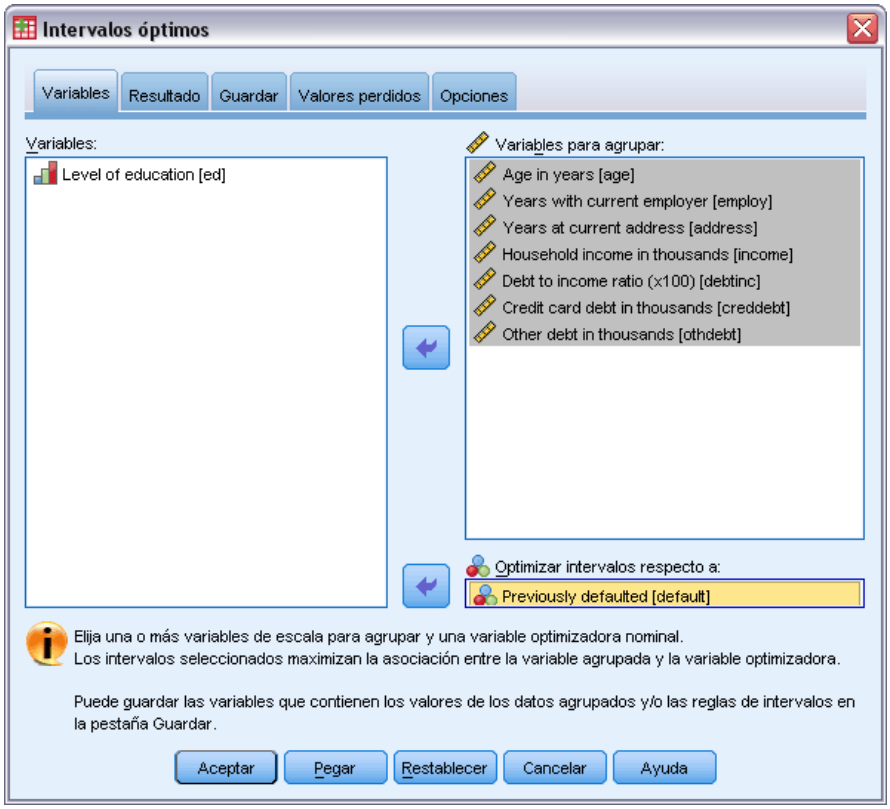
Como parte de la iniciativa del banco para reducir la tasa de moras de créditos, un encargado de créditos ha recopilado información financiera y demográfica sobre los clientes antiguos y actuales con la intención de crear un modelo para pronosticar la probabilidad de causar mora en un crédito. Varios predictores potenciales son de escala, pero el encargado de créditos quiere tener en cuenta modelos que funcionan mejor con predictores categóricos.

La información de los 5000 clientes anteriores está recopilada en *bankloan_binning.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A el p. 138. Utilice el procedimiento Intervalos óptimos para generar reglas de intervalos para los predictores de escala y, a continuación, utilice las reglas generadas para procesar *bankloan.sav*. A continuación, el conjunto de datos procesado puede utilizarse para crear un modelo predictivo.

Ejecución del análisis

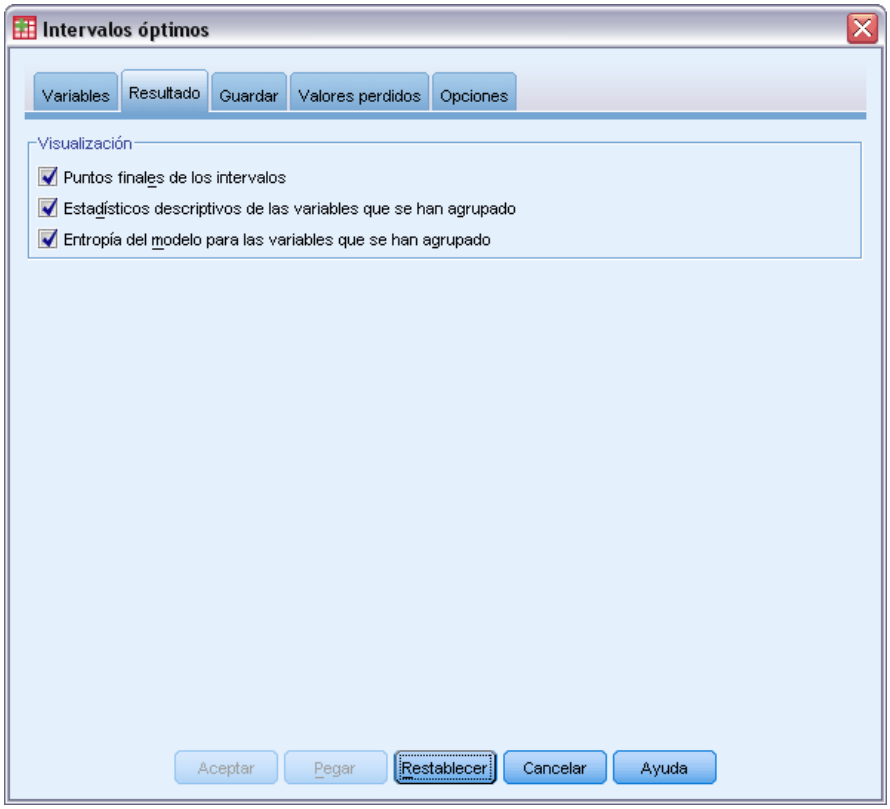
- Para ejecutar un análisis de intervalos óptimos, elija en los menús:
  Transformar > Intervalos óptimos...

Figura 10-1
Cuadro de diálogo Intervalos óptimos, pestaña Variables



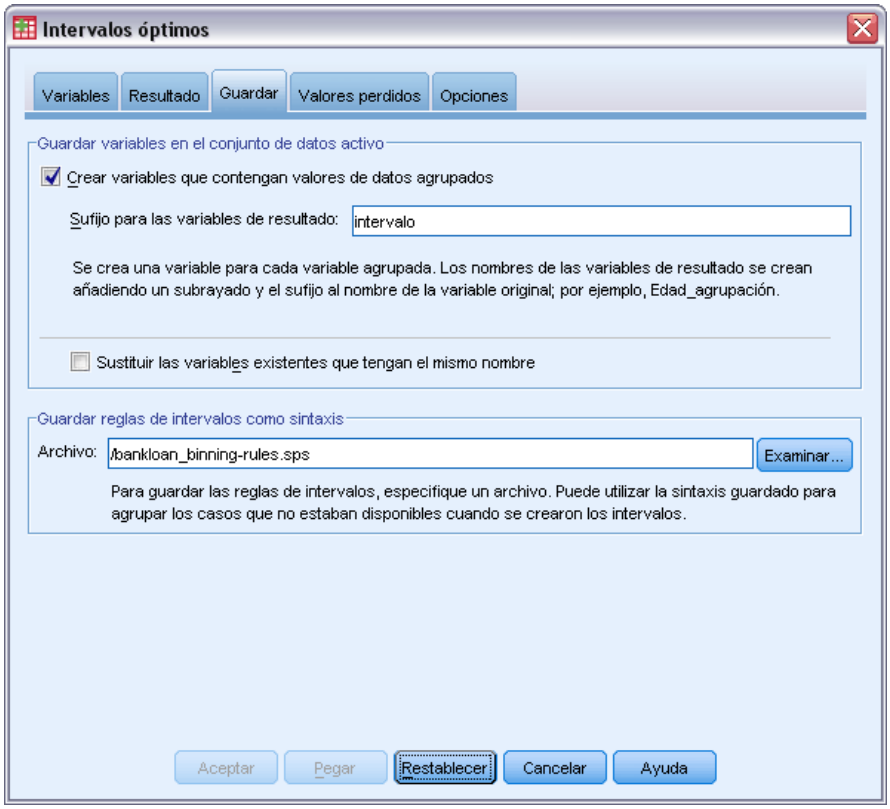
- Seleccione *Edad en años* y *Años con empresa actual* hasta *Otras deudas en miles* como las variables que se van a agrupar.
- Seleccione *Impagos anteriores* como la variable guía.
- Pulse en la pestaña Resultados.

Figura 10-2
Cuadro de diálogo Intervalos óptimos, pestaña Resultado



- Seleccione Estadísticos descriptivos y Entropía del modelo para las variables que se han agrupado.
- Pulse en la pestaña Guardar.

Figura 10-3
Cuadro de diálogo Intervalos óptimos, pestaña Guardar



- Seleccione Crear variables que contengan valores de datos agrupados.
- Escriba una ruta de acceso y un nombre de archivo para el archivo de sintaxis que va a contener las reglas de intervalos generadas. En este ejemplo se ha utilizado */bankloan_binning-rules.sps*.
- Pulse en Aceptar.

Estas selecciones generan la siguiente sintaxis de comandos:

```
* Intervalos óptimos
OPTIMAL BINNING
  /VARIABLES GUIDE=default BIN=edad empleo direccion ingresos deudaing deudacred
  deudaotro SAVE=YES (INTO=edad_bin empleo_bin direccion_bin ingresos_bin deudaing_bin
  deudacred_bin deudacred_bin)
  /CRITERIA METHOD=MDLP
   PREPROCESS=EQUALFREQ (BINS=1000)
   FORCEMERGE=0
   LOWERLIMIT=INCLUSIVE
   LOWEREND=UNBOUNDED
   UPPEREND=UNBOUNDED
  /MISSING SCOPE=PAIRWISE
  /OUTFILE RULES='/bankloan_binning-rules.sps'
  /PRINT ENDPOINTS DESCRIPTIVES ENTROPY.
```

- El procedimiento discretizará las variables de entrada *age*, *employ*, *address*, *income*, *debtinc*, *creddebt* y *othdebt* mediante la agrupación MDLP con la variable guía *default*.

- Los valores discretizados para estas variables se almacenarán en las nuevas variables *age_bin*, *employ_bin*, *address_bin*, *income_bin*, *debtinc_bin*, *creddebt_bin* y *othdebt_bin*.
- Si una variable de entrada que se va a agrupar tiene más de 1000 valores distintos, el método de frecuencias iguales reducirá el número a 1000 antes de llevar a cabo la agrupación MDLP.
- La sintaxis de comandos que representa las reglas de intervalos se guarda en el archivo */bankloan_binning-rules.sps*.
- Los límites de intervalos, los estadísticos descriptivos y los valores de entropía de modelo se solicitan para las variables de entrada que se van a agrupar.
- El resto de criterios de agrupación están establecidos en sus valores por defecto.

Estadísticos descriptivos

Figura 10-4
Estadísticos descriptivos

	N	Mínimo	Máximo	Número de valores distintos	Número de intervalos
Edad en años	5000	20	58	39	2
Años con la empresa actual	5000	0	38	39	4
Años en la dirección actual	5000	0	37	38	3
Ingresos familiares en miles	5000	12,10	2461,70	1100	2
Tasa de deuda sobre ingresos (x100)	5000	,08	44,62	2060	5
Deuda de la tarjeta de crédito en miles	5000	,01	139,58	5000	4
Otras deudas en miles	5000	,01	416,52	4999	2

La tabla de estadísticos descriptivos proporciona información de resumen sobre las variables de entrada que se van a agrupar. Las primeras cuatro columnas se refieren a los valores agrupados previamente.

- N es el número de casos que se utilizan en el análisis. Cuando se utiliza la eliminación por lista de los valores perdidos, este valor debe ser constante entre las variables. Cuando se utiliza el tratamiento de los valores perdidos por parejas, no es necesario que este valor sea constante. Dado que este conjunto de datos no contiene valores perdidos, el valor será sencillamente el número de casos.
- Las columnas Mínimo y Máximo muestran los valores mínimo y máximo (anteriores a la agrupación) del conjunto de datos para cada variable de entrada que se va a agrupar. Además de proporcionar una idea del rango observado de valores para cada variable, pueden resultar útiles para detectar valores que se encuentran fuera del rango esperado.
- El Número de valores distintos indica las variables que se procesaron previamente con el algoritmo de frecuencias iguales. Por defecto, las variables con más de 1000 valores distintos (de *Ingresos familiares en miles* a *Otras deudas en miles*) están previamente agrupados en 1000 intervalos distintos. A continuación, estos intervalos previamente procesados se agrupan

respecto a la variable guía mediante MDLP. Puede controlar la función de procesamiento previo mediante la pestaña Opciones.

- El Número de intervalos es el número final de intervalos generados por el procedimiento y es mucho menor que el número de valores distintos.

Entropía del modelo

Figura 10-5
Entropía del modelo

	Entropía del modelo
Edad en años	,788
Años con la empresa actual	,754
Años en la dirección actual	,781
Ingresos familiares en miles	,803
Tasa de deuda sobre ingresos (x100)	,711
Deuda de la tarjeta de crédito en miles	,776
Otras deudas en miles	,801

Una menor entropía del modelo indica una mayor precisión predictiva de la variable agrupada en la variable guía Impagos anteriores.

La entropía del modelo proporciona una idea de la utilidad que puede tener cada variable en un modelo predictivo para la probabilidad de causar mora.

- El mejor predictor posible es el que contiene casos con el mismo valor que la variable guía para cada intervalo generado; así, la variable guía puede pronosticarse perfectamente. Este tipo de predictores tiene una entropía del modelo no definida. Esto no suele ocurrir en situaciones reales y puede indicar problemas con la calidad de los datos.
- El peor predictor posible es el que no funciona mejor que las suposiciones; el valor de su entropía del modelo depende de los datos. En este conjunto de datos, 1256 (o 0,2512) de los 5000 clientes totales causaron mora y 3744 (o 0,7488) no lo hicieron; así, el peor predictor posible podría tener una entropía del modelo de $-0{,}2512 \times \text{logaritmo}_2(0{,}2512) - 0{,}7488 \times \text{logaritmo}_2(0{,}7488) = 0{,}8132$.

Es difícil desarrollar una sentencia más concluyente que ésta: las variables con valores de entropía del modelo más bajos deberían ser mejores predictores, ya que lo que constituye un buen valor de entropía del modelo depende de la aplicación y los datos. En este caso, parece que las variables con un número de intervalos generados mayor, con relación al número de categorías distintas, tienen valores de entropía del modelo más bajos. Se debería llevar a cabo una evaluación más detenida de estas variables de entrada que se van a agrupar como predictores mediante los procedimientos de creación de modelos predictivos, que ofrecen herramientas más completas para la selección de variables.

Resúmenes de agrupación

El resumen de agrupación indica los límites de los intervalos creados y la frecuencia de recuento de cada intervalo por valores de la variable guía. Se genera una tabla de resumen de agrupación diferente para cada variable de entrada que se ha agrupado.

Figura 10-6
Resumen de agrupación para Edad en años

	Límite		Número de casos por nivel de Impagos anteriores		
Intervalo	Inferior	Superior	No	Sí	Total
1	[a]	32	1129	639	1768
2	32	[a]	2615	617	3232
Total			3744	1256	5000

Cada intervalo se calcula como Inferior <= Edad en años < Superior.

a. Sin límites

El resumen de *Edad en años* muestra que 1768 clientes, todos de 32 años o más jóvenes, se colocan en Intervalo 1, mientras que los 3232 clientes restantes, todos mayores de 32 años, se colocan en Intervalo 2. La proporción de clientes que ha causado mora con anterioridad es mucho mayor en Intervalo 1 (639/1768=0,361) que en Intervalo 2 (617/3232=0,191).

Figura 10-7
Resumen de agrupación de Ingresos familiares en miles

	Límite		Número de casos por nivel de Impagos anteriores		
Intervalo	Inferior	Superior	No	Sí	Total
1	[a]	26,70	1054	513	1567
2	26,70	[a]	2690	743	3433
Total			3744	1256	5000

Cada intervalo se calcula como Inferior <= Ingresos familiares en miles < Superior.

a. Sin límites

El resumen de *Ingresos familiares en miles* muestra un patrón similar, con un único punto de corte en 26,70 y una proporción superior de clientes que han causado mora con anterioridad en Intervalo 1 (513/1567=0,327) que en Intervalo 2 (743/3433=0,216). Como se esperaba a partir de los estadísticos de entropía del modelo, la diferencia en estas proporciones no es tan grande como la de *Edad en años*.

Figura 10-8
Resumen de agrupación de Otras deudas en miles

	Límite		Número de casos por nivel de Impagos anteriores		
Intervalo	Inferior	Superior	No	Sí	Total
1	[a]	2,19	2161	539	2700
2	2,19	[a]	1583	717	2300
Total			3744	1256	5000

Cada intervalo se calcula como Inferior <= Otras deudas en miles < Superior.

a. Sin límites

El resumen de *Otras deudas en miles* muestra un patrón inverso, con un único punto de corte en 2,19 y una proporción inferior de clientes que han causado mora con anterioridad en Intervalo 1 (539/2700=0,200) que en Intervalo 2 (717/2300=0,312). De nuevo, como se esperaba a partir de los estadísticos de entropía del modelo, la diferencia en estas proporciones no es tan grande como la de *Edad en años*.

Figura 10-9
Resumen de agrupación de Años con empresa actual

	Límite		Número de casos por nivel de Impagos anteriores		
Intervalo	Inferior	Superior	No	Sí	Total
1	[a]	3	629	478	1107
2	3	8	1066	461	1527
3	8	18	1471	268	1739
4	18	[a]	578	49	627
Total			3744	1256	5000

Cada intervalo se calcula como Inferior <= Años con la empresa actual < Superior.

a. Sin límites

El resumen de *Años con empresa actual* muestra un patrón de proporciones decrecientes de personas que causan mora a medida que los números del intervalo aumentan.

Intervalo	**Proporción de personas que causan mora**
1	0.432
2	0.302
3	0.154
4	0.078

Figura 10-10
Resumen de agrupación de Años en la dirección actual

	Límite		Número de casos por nivel de Impagos anteriores		
Intervalo	Inferior	Superior	No	Sí	Total
1	[a]	7	1652	829	2481
2	7	14	1184	313	1497
3	14	[a]	908	114	1022
Total			3744	1256	5000

Cada intervalo se calcula como Inferior <= Años en la dirección actual < Superior.

a. Sin límites

El resumen de *Años en la dirección actual* muestra un patrón similar. Como se esperaba a partir de los estadísticos de entropía del modelo, las diferencias entre los intervalos en cuanto a la proporción de personas que causan mora son más acusadas en *Años con empresa actual* que en *Años en la dirección actual*.

Intervalo	**Proporción de personas que causan mora**
1	0.334
2	0.209
3	0.112

Figura 10-11
Resumen de agrupación de Deuda de tarjeta de crédito en miles

	Límite		Número de casos por nivel de Impagos anteriores		
Intervalo	Inferior	Superior	No	Sí	Total
1	[a]	,97	2169	466	2635
2	,97	1,91	848	307	1155
3	1,91	6,05	643	352	995
4	6,05	[a]	84	131	215
Total			3744	1256	5000

Cada intervalo se calcula como Inferior <= Deuda de la tarjeta de crédito en miles < Superior.

a. Sin límites

El resumen de *Deuda de tarjeta de crédito en miles* muestra el patrón inverso, con proporciones crecientes de personas que causan mora a medida que aumentan los números del intervalo. *Años con empresa actual* y *Años en la dirección actual* parecen ser más válidos para identificar personas con una menor probabilidad de causar mora, mientras que *Deuda de tarjeta de crédito en miles* es más útil para identificar personas con mayor probabilidad de causar mora.

Intervalo	**Proporción de personas que causan mora**
1	0.177
2	0.266

Intervalo	Proporción de personas que causan mora
3	0.354
4	0.609

Figura 10-12
Resumen de agrupación de Proporción de la deuda en los ingresos (x100)

	Límite		Número de casos por nivel de Impagos anteriores		
Intervalo	Inferior	Superior	No	Sí	Total
1	[a]	4,39	912	88	1000
2	4,39	12,09	2006	437	2443
3	12,09	18,71	625	386	1011
4	18,71	31,00	198	303	501
5	31,00	[a]	3	42	45
Total			3744	1256	5000

Cada intervalo se calcula como Inferior <= Tasa de deuda sobre ingresos (x100) < Superior.

a. Sin límites

El resumen de *Proporción de la deuda en los ingresos (x100)* muestra un patrón similar a *Deuda de tarjeta de crédito en miles*. Esta variable tiene el valor de entropía del modelo más bajo y, por lo tanto, es el mejor predictor posible para la probabilidad de causar mora. Es más útil para clasificar personas con una alta probabilidad de causar mora que *Deuda de tarjeta de crédito en miles* y casi igual de eficaz para clasificar las personas con una baja probabilidad de causar mora que *Años con empresa actual*.

Intervalo	Proporción de personas que causan mora
1	0.088
2	0.179
3	0.382
4	0.605
5	0.933

Variables agrupadas

Figura 10-13
Variables agrupadas para bankloan_binning.sav en el Editor de datos.

	impago	age_bin	employ_bin	address_bin	income_bin	debtinc_bin	creddebt_bin	othdebt_bin
1	0	2	3	2	2	2	4	2
2	0	1	3	1	2	3	2	2
3	0	2	3	3	2	2	1	1
4	0	2	3	3	2	1	3	1
5	0	1	1	1	2	3	2	2
6	0	2	2	1	1	2	1	1
7	1	2	4	2	2	4	3	2
8	0	2	3	2	2	1	1	1
9	0	1	2	1	1	4	2	2
10	0	2	1	2	1	4	3	1
11	0	1	1	1	1	1	1	1
12	1	1	2	1	1	2	1	1
13	0	2	4	3	2	2	3	2
14	1	2	2	2	2	3	2	2
15	0	2	4	3	2	2	3	2

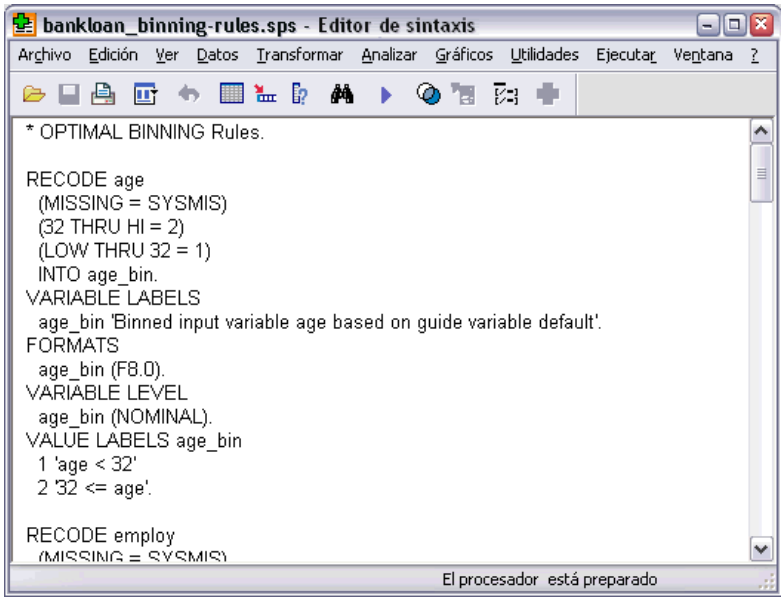
Los resultados del proceso de agrupación en este conjunto de datos pueden observarse claramente en el Editor de datos. Estas variables agrupadas son útiles para generar resúmenes personalizados de los resultados de la agrupación mediante procedimientos descriptivos o de generación de informes, pero no es aconsejable utilizar estos datos para generar un modelo predictivo ya que las reglas de intervalos se generaron con estos casos. Es mejor aplicar las reglas de intervalos a otro conjunto de datos que contenga información sobre otros clientes.

Aplicación de reglas de intervalos de sintaxis

Al ejecutar el procedimiento Intervalos óptimos, solicitó que las reglas de intervalos generadas por el procedimiento se guardaran como una sintaxis de comandos.

- Abra *bankloan_binning-rules.sps*.

Figura 10-14
Archivo de reglas de sintaxis



Para cada variable de entrada que se ha agrupado existe un bloque de sintaxis de comandos que realiza la agrupación, establece la etiqueta de la variable, el formato y el nivel, y define las etiquetas de valor de los intervalos. Estos comandos se pueden aplicar a un conjunto de datos con las mismas variables que *bankloan_binning.sav*.

- Abra *bankloan.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A el p. 138.
- Vuelva a la vista del Editor de sintaxis de *bankloan_binning-rules.sps*.

▶ Para aplicar las reglas de intervalos, seleccione en los menús del Editor de sintaxis:
Ejecutar > Todos...

Figura 10-15
Variables agrupadas para bankloan.sav en el Editor de datos

	morapred3	age_bin	employ_bin	address_bin	income_bin	debtinc_bin	creddebt_bin	othdebt_bin
1	0,21304	2	3	2	2	2	4	2
2	0,43690	1	3	1	2	3	2	2
3	0,14102	2	3	3	2	2	1	1
4	0,10442	2	3	3	2	1	3	1
5	0,43690	1	1	1	2	3	2	2
6	0,23358	2	2	1	1	2	1	1
7	0,81709	2	4	2	2	4	3	2
8	0,11336	2	3	2	2	1	1	1
9	0,66390	1	2	1	1	4	2	2
10	0,51553	2	1	2	1	4	3	1
11	0,09055	1	1	1	1	1	1	1
12	0,13631	1	2	1	1	2	1	1
13	0,22890	2	4	3	2	2	3	2
14	0,40484	2	2	2	2	3	2	2
15	0,20866	2	4	3	2	2	3	2

Las variables de *bankloan.sav* se han agrupado según las reglas generadas al ejecutar el procedimiento Intervalos óptimos en *bankloan_binning.sav*. Este conjunto de datos ya está listo para su uso en la construcción de modelos predictivos que prefieran o requieran variables categóricas.

Resumen

Se ha utilizado el procedimiento Intervalos óptimos para generar reglas de intervalos para variables de escala que son predictores potenciales para la probabilidad de causar mora y para aplicar estas reglas a un conjunto de datos diferente.

Durante el proceso de agrupación, se observa que las agrupaciones *Años con empresa actual* y *Años en la dirección actual* parecen ser más válidas para identificar personas con una menor probabilidad de causar mora, mientras que *Deuda de tarjeta de crédito en miles* es más útil para identificar personas con mayor probabilidad de causar mora. Esta interesante observación ofrece una información extra a la hora de generar modelos predictivos para la probabilidad de causar mora. Si la principal preocupación es evitar las deudas incobrables, *Deuda de tarjeta de crédito en miles* será más relevante que *Años con empresa actual* y *Años en la dirección actual*. Si la prioridad es aumentar la base de clientes, *Años con empresa actual* y *Años en la dirección actual* serán más relevantes.

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 (Van der Ham, Meulman, Van Strien, y Van Engeland, 1997) 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 (Price y Bouffard, 1974), 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 (Green y Rao, 1972), 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 (McCullagh y Nelder, 1989) 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 (Green y Wind, 1973), 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 (Kennedy, Riquier, y Sharp, 1996). 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" (Rickman, Mitchell, Dingman, y Dalen, 1974). 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 (Blake y Merz, 1998) 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 (Bell, 1961) presentó una tabla para ilustrar posibles grupos sociales. Guttman (Guttman, 1968) 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 (Rosenberg y Kim, 1975) 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 15×15 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(Breiman y Friedman, 1985), (Hastie y Tibshirani, 1990) 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 (Hartigan, 1975).
- **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 (McCullagh et al., 1989) 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 (Collett, 2003).
- **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 (Collett et al., 2003).
- **verd1985.sav.** Archivo de datos sobre una encuesta (Verdegaal, 1985). 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 (Ware, Dockery, Spiro III, Speizer, y Ferris Jr., 1984). 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.

Bibliografía

Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. Nueva York: Harper & Row.

Blake, C. L., y C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., y J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, .

Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.

Green, P. E., y V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Green, P. E., y Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, .

Hartigan, J. A. 1975. *Clustering algorithms*. Nueva York: John Wiley and Sons.

Hastie, T., y R. Tibshirani. 1990. *Generalized additive models*. Londres: Chapman and Hall.

Kennedy, R., C. Riquier, y B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, .

McCullagh, P., y J. A. Nelder. 1989. *Modelos lineales generalizados*, 2nd ed. Londres: Chapman & Hall.

Price, R. H., y D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

Rickman, R., N. Mitchell, J. Dingman, y J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, .

Rosenberg, S., y M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, .

Van der Ham, T., J. J. Meulman, D. C. Van Strien, y H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry*, 170, .

Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (en neerlandés)*. Leiden: Department of Data Theory, University of Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, y B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, .

Índice