IBM SPSS Data Preparation 23

Nota

Antes de utilizar esta información y el producto al que da soporte, lea la información contenida en el "Avisos" en la página 33.

Información de producto

Esta edición se aplica a la versión 23, release 0, modificación 0 de IBM® SPSS Statistics y a todos los releases y modificaciones posteriores hasta que se indique lo contrario en ediciones nuevas.

Contenido

Capítulo 1. 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 previsiones 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 medición. 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 de variables de cruce.** Las reglas de variables de cruce son reglas definidas por el usuario que pueden ser aplicadas a una única variable o a una combinación de variables. Las reglas de variables de cruce están definidas por una expresión lógica que señala los 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

1. Seleccione en los menús:

   Datos > **Validación** > **Cargar reglas predefinidas...**

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 de variables de cruce y de variable única.

Para crear y ver reglas de validación

1. Seleccione en los menús:

   Datos > **Validación** > **Definir reglas...**

   El cuadro de diálogo contiene reglas de validación de variables de cruce 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.
2. Seleccione las reglas individuales en las pestañas Reglas de variable única y Reglas de variables de cruce para ver y modificar sus propiedades.

Definir 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 exclusivo en las reglas de variable única y las reglas de variables de cruce.
- **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 exclusiva entre las reglas de variable única y las reglas de variables de cruce. 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 exclusivo para las reglas de variable única y las reglas de variables de cruce.
- **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.

  Definición de rango

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

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

  Definición de lista

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

  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 del usuario.** Controla si los valores perdidos del usuario están señalados como no válidos.
- **Permitir valores perdidos del sistema.** Controla si los valores perdidos del sistema están señalados 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 señalados como no válidos. Esto no se aplica a los tipos de reglas que no son de cadena.

Definir reglas de variables de cruce

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

Reglas. La lista muestra reglas de validación de variables de cruce 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 exclusivo en las reglas de variable única y la regla de variables de cruce.
- **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 exclusiva entre las reglas de variable única y las reglas de variables de cruce. 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 exclusivo para las reglas de variable única y las reglas de variables de cruce.
- **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

1. 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**). Especifique 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 de variables de cruce 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 de variables de cruce, 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

1. Seleccione en los menús:

   Datos > **Validación** > **Validar datos...**
2. 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:
3. Pulsar en la pestaña **Reglas de variables de cruce** y aplicar una o más reglas de variables de cruce.

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.

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

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 estándar mínima.** Si alguna variable de análisis es de escala, esta opción informa sobre variables de análisis de escala cuya desviación estándar es menor que el valor especificado. El valor especificado debe ser un número no negativo. La comprobación de desviación estándar se desactiva si se especifica 0.

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

- **Señalar los 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.
- **Señalar los ID duplicados.** Esta opción informa sobre casos que tienen identificadores de caso duplicados. Los identificadores incompletos se excluyen del conjunto de posibles duplicados.

Señalar los 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

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**. Consulte el tema “Definir reglas de variable única” en la página 3 para obtener más información.

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 del 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 de variables de cruce

La pestaña Reglas de variables de cruce muestra reglas de variables de cruce disponibles y permite aplicarlas a los datos. Para definir reglas de variables de cruce adicionales, pulse en **Definir reglas**. Consulte el tema “Definir reglas de variables de cruce” en la página 4 para obtener más información.

Validar datos: Resultados

Informe por casos. Si ha aplicado alguna regla de validación de variable única o de de variables de cruce, 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 de las variables de análisis. 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 estándar, el mínimo y el máximo para las variables de escala.

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

Validar datos: 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 predeterminados 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 exclusivos 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 y de de variables de cruce.

Reemplazar variables de resumen existentes. Las variables que se guardan en el archivo de datos deben tener nombres exclusivos 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 del 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 señalar las reclamaciones sospechosas y potencialmente fraudulentas. Antes de generar 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.

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.

¿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 características. 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 generación de modelos y la precisión de las predicciones.
- **Optimizar velocidad.** Esta opción prepara los datos para dar prioridad a la velocidad con la que se procesan los datos por los algoritmos de generación 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 las predicciones producidas por los algoritmos de generación 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:

1. Seleccione en los menús:

   Transformar > **Preparar datos para modelado** > **Automática...**
2. 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

1. Seleccione en los menús:

   Transformar > **Preparar datos para modelado** > **Interactiva...**
2. Pulse en **Analizar** en la barra de herramientas en la parte superior del cuadro de diálogo.
3. Pulse la pestaña análisis y consulte los pasos de preparación de datos sugeridos.
4. Si está satisfecho, pulse en **Ejecutar**. En caso contrario, pulse en **Borrar análisis**, cambie los ajustes que desee 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

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 predeterminada que sea incompatible con el resto de objetivos, la pestaña Objetivo se actualiza automáticamente para seleccionar la opción **Personalizar análisis**.

Preparar fechas y horas

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

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 tengan el 100% de valores perdidos 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 perdidos se excluyen automáticamente. El valor predeterminado es 50.

Excluir campos nominales con demasiadas categorías exclusivas. 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 exclusiva 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 predeterminado es 95%.

Ajustar medición

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

Nivel de medición. Especifique si el nivel de medición 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 predeterminado 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 exclusivos inferior al especificado se reestructuran como campos ordinales. Especifique un número entero positivo. El valor predeterminado es 5. Este valor debe ser menor o igual al número máximo de valores de campos ordinales.

Mejorar la calidad 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 estándar 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 perdidos siguen la configuración de manejo de valores perdidos 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

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 estándar como estimaciones de parámetros de población, los campos se tipifican y las puntuaciones *z* se correlacionan con los valores correspondientes de una distribución normal con la **Media final** y **Desviación estándar final** especificadas. Especifique un número para **Media final** y un número positivo para **Desviación estándar final**. Los valores predeterminados 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 correlacionan con 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 estándar final** especificada. Especifique un número para **Media final** y un número positivo para **Desviación estándar final**. Los valores predeterminados 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 predicho a la escala original. Consulte el tema para obtener más información. Consulte el tema "Puntuaciones de transformación retrospectiva" en la página 24 para obtener más información.

Transformar campos

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. Se encuentran disponibles las opciones siguientes:

- **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 predeterminado 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 predeterminado 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

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

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, etc.

De esta forma se garantiza que las características construidas 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

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

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.

1. 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.

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 predeterminada). Consulte el tema “Resumen de procesamiento de campo” en la página 18 para obtener más información.
- Campos. Consulte el tema “Campos” en la página 18 para obtener más información.
- Resumen de acciones. Consulte el tema “Resumen de acciones” en la página 20 para obtener más información.

Hay cuatro vistas relacionadas/auxiliares:

- Poder predictivo (la configuración predeterminada). Consulte el tema "Poder predictivo" en la página 20 para obtener más información.
- Tabla de campos. Consulte el tema "Tabla de campos" en la página 20 para obtener más información.
- Detalles de campo. Consulte el tema "Detalles de campo" en la página 20 para obtener más información.
- Detalles de acción. Consulte el tema "Detalles de acción" en la página 22 para obtener más información.

Enlaces entre vistas

En la vista principal, el texto subrayado de las tablas controla la visualización en la vista enlazada. 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

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 enlazada. Los detalles de **Destino**, **Características de entrada** y **Características de entrada no utilizadas** se muestran en la vista enlazada Tabla de campos. Consulte el tema "Tabla de campos" en la página 20 para obtener más información. Las **Características recomendadas para su uso en análisis** se muestran en la vista enlazada Poder predictivo. Consulte el tema "Poder predictivo" en la página 20 para obtener más información.

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 construidas 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 medición.** 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. De forma predeterminada, 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.

De forma predeterminada 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. De forma predeterminada, la lista desplegable refleja las recomendaciones.

  Para los predictores ordinarios que se han transformado, la lista desplegable tiene tres opciones: **Transformados**, **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: **Transformados** 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 **Nivel de medición** y **Poder predictivo** de esas características.
- **Nombre.** Cada nombre de campo es un enlace. Pulse sobre un nombre para ver más información acerca del campo en la vista enlazada. Consulte el tema “Detalles de campo” en la página 20 para obtener más información.
- **Nivel de medición.** 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

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 enlazada sobre las acciones que se realizan. Consulte el tema “Detalles de acción” en la página 22 para obtener más información.

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

Poder predictivo

Se muestra de forma predeterminada 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

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 medición.** 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

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

Si el nodo ADP ha ejecutado el manejo de valores perdidos, el gráfico circular 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. Consulte el tema "Detalles de acción" en la página 22 para obtener más información.
- **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

Se muestra cuando selecciona cualquier **Acción** subrayada en la vista principal Resumen de acciones. La vista enlazada 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 enlazada. 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 reestructurados como ordinales
- Reestructuración de números totales.

Si los campos de entrada (destinos o predictores) no eran conjuntos continuos u ordinales, 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:

- Destino. 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 estándar) 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 correlacionados 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 predicho a la escala original.

1. Para aplicar transformación retrospectiva a las puntuaciones, seleccione en los menús:

   Transformar > **Preparar datos para modelado** > **Puntuaciones de transformación retrospectiva...**
2. Seleccione un campo para aplicar la transformación retrospectiva. Este campo debe contener valores pronosticados por el modelo del destino transformado.
3. Especifique un sufijo para el nuevo campo. Este nuevo campo contendrá valores pronosticados por el modelo en la escala original del destino sin transformar.
4. 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. Consulte el tema “Aplicación y almacenamiento de transformaciones” en la página 16 para obtener más información.

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 exclusivos 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 manejo 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

1. Seleccione en los menús:

Datos > **Identificar casos atípicos...**

2. Seleccione al menos una variable de análisis.
3. Si lo desea, seleccione una variable de identificación 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.

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

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 estándar 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

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

La pestaña Valores perdidos se utiliza para controlar el tratamiento de los valores perdidos del 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

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.

Características 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 cruzadas 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.

Resultado. 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.

Intervalos óptimos: Consideraciones sobre los 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

1. Seleccione en los menús:

   Transformar > **Intervalos óptimos...**
2. Seleccione una o más variables de entrada para agruparlas.
3. Seleccione una variable guía.

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

Intervalos óptimos: 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

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

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

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, $n_{final}+1$ intervalos. Para los intervalos $i = 2, ..., n_{final}$ (desde el segundo intervalo con valores inferiores hasta el segundo intervalo con valores superiores), se calcula

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

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.

Características adicionales del comando OPTIMAL BINNING

La sintaxis de comandos también le permite:

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

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

Avisos

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

Es posible que IBM no ofrezca los productos, servicios o características que se tratan en este documento en otros países. Consulte al representante local de IBM para obtener información sobre los productos y servicios disponibles actualmente en su zona. Cualquier referencia a un producto, programa o servicio de IBM no pretende afirmar ni implicar que solamente se pueda utilizar ese producto, programa o servicio de IBM. Puede utilizarse en su lugar cualquier otro producto, programa o servicio funcionalmente equivalente que no vulnere ninguno de los derechos de propiedad intelectual de IBM. Sin embargo, es responsabilidad del usuario evaluar y comprobar el funcionamiento de todo producto, programa o servicio que no sea de IBM.

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

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
Estados Unidos

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

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

El párrafo siguiente no se aplica al Reino Unido ni a ningún otro país donde estas disposiciones sean incompatibles con la legislación local: INTERNATIONAL BUSINESS MACHINES CORPORATION PROPORCIONA ESTA PUBLICACIÓN "TAL CUAL" SIN GARANTÍAS DE NINGÚN TIPO, NI EXPLÍCITAS NI IMPLÍCITAS, INCLUIDAS, PERO NO LIMITÁNDOSE A ELLAS, LAS GARANTÍAS IMPLÍCITAS DE NO INFRACCIÓN DE DERECHOS DE TERCEROS, COMERCIALIZACIÓN O IDONEIDAD PARA UN FIN CONCRETO. Algunos estados no permiten la renuncia a expresar o a garantías implícitas en determinadas transacciones , por lo tanto , esta declaración no se aplique a usted.

Esta información puede incluir imprecisiones técnicas o errores tipográficos. Periódicamente, se efectúan cambios en la información aquí y estos cambios se incorporarán en nuevas ediciones de la publicación. IBM puede realizar en cualquier momento mejoras o cambios en los productos o programas descritos en esta publicación sin previo aviso.

Cualquier referencia a sitios Web que no sean de IBM en esta información sólo es ofrecida por comodidad y de ningún modo sirve como aprobación de esos sitios Web. El material de esos sitios web no forma parte del material de este producto de IBM y la utilización de esos sitios web se realizará bajo su total responsabilidad.

IBM puede utilizar o distribuir la información que el usuario le suministre en el modo que considere apropiado sin incurrir en ninguna obligación con el usuario.

Los propietarios de licencia de este programa que deseen tener información sobre el mismo con el fin de: (i) intercambiar información entre programas creados de forma independiente y otros programas (incluido éste) y (ii) utilizar mutuamente la información que se ha intercambiado, deberán ponerse en contacto con:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
Estado Unidos

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

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

Cualquier dato de rendimiento mencionado aquí ha sido determinado en un entorno controlado. Por lo tanto, los resultados obtenidos en otros entornos operativos pueden variar de forma significativa. Es posible que algunas mediciones se hayan realizado en sistemas en desarrollo y no existe ninguna garantía de que estas mediciones sean las mismas en los sistemas comerciales. Además, es posible que algunas mediciones hayan sido estimadas a través de extrapolación. Los resultados reales pueden variar. Los usuarios de este documento deben consultar los datos que corresponden a su entorno específico.

Se ha obtenido información acerca de productos que no son de IBM de los proveedores de esos productos, de sus publicaciones anunciadas o de otros orígenes disponibles públicamente. IBM no ha probado esos productos y no puede confirmar la precisión de su rendimiento, su compatibilidad ni ningún otro aspecto. Las preguntas acerca de las aptitudes de productos que no sean de IBM deben dirigirse a los proveedores de dichos productos.

Todas las declaraciones sobre el futuro del rumbo y la intención de IBM están sujetas a cambio o retirada sin previo aviso y representan únicamente metas y objetivos.

Esta información contiene ejemplos de datos e informes utilizados en operaciones comerciales diarias. Para ilustrarlos lo máximo posible, los ejemplos incluyen los nombres de las personas, empresas, marcas y productos. Todos esos nombres son ficticios y cualquier parecido con los nombres y direcciones utilizados por una empresa real es pura coincidencia.

LICENCIA DE COPYRIGHT:

Esta información contiene programas de aplicaciones de ejemplo en código fuente, que ilustran técnicas de programación en las distintas plataformas operativas. Puede copiar, modificar y distribuir libremente estos programas de muestra, sin pagar por ello a IBM, con la finalidad de desarrollar, utilizar, comercializar o distribuir programas de aplicación conformes a la interfaz de programación de aplicaciones para la plataforma operativa para la que están escritos los programas de muestra. Dichos ejemplos no se han probado exhaustivamente bajo todas las condiciones. Por lo tanto, IBM no puede garantizar ni certificar la fiabilidad, el servicio o el funcionamiento de estos programas. Los programas de ejemplo se proporcionan "TAL CUAL", sin garantía de ninguna clase. IBM no será responsable de los daños que surjan por el uso de los programas de ejemplo.

Cada copia, parcial o completa, de estos programas de ejemplo, o cualquier trabajo obtenido a partir de los mismos, debe incluir el siguiente aviso de copyright:

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

Marcas comerciales

IBM, el logotipo de IBM e ibm.com son marcas registradas o marcas comerciales registradas de International Business Machines Corp., registrada en muchas jurisdicciones en todo el mundo. Otros nombres de producto y servicio podrían ser marcas registradas de IBM u otras compañías. Encontrará una lista actual de marcas registradas de IBM en la Web en "Información de copyright y marca registrada" en www.ibm.com/legal/copytrade.shtml.

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

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

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

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

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

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

Índice

Impreso en España