IBM SPSS - Preparación de datos 26

Nota
Antes de utilizar esta información y el producto al que da soporte, lea la información del apartado "Avisos" en la página 7.

Información de producto

Esta edición se aplica a la versión 26, 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

Preparación de los datos

Se han incluido las características de preparación adicional de datos siguientes en SPSS Statistics Professional Edition o la opción Preparación de los datos.

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.

Las del módulo adicional de preparación de datos le permiten 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 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.

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 sobre 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.

Identificación de 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.
4. Pulse **Aplicar**.

Campos con nivel de medición desconocido

La alerta de nivel de medición se muestra cuando 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
Lista todos los campos con un nivel de medición desconocido. Puede asignar un nivel de medición a estos campos. También puede asignar un nivel de medición en el panel de Lista de variables del editor de datos.

Dado que el nivel de medición es importante para este procedimiento, no puede ejecutar este procedimiento hasta que se hayan definido todos los campos en el nivel de medición.

Identificar casos atípicos: Salida

El diálogo Salida proporciona opciones para generar salida tabular.

Lista de casos atípicos y motivos por los que se consideran atípicos
Cuando se selecciona, 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 ID de los casos se visualizan si se especifica la variable de identificador de caso en el diálogo Variables.

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

El diálogo Guardar proporciona opciones de guardar de variables y modelos.

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.

Sustituir variables existentes que tengan el mismo nombre o nombre raíz

Cuando se seleccionen, se sustituirán las variables existentes cuyos nombres entran en conflicto con las variables que se van a guardar.

Exportar archivo de modelo

Le permite guardar el modelo a un archivo XML externo.

Identificar casos atípicos: Valores perdidos

El diálogo Valores perdidos se utilizan para controlar el tratamiento de valores perdidos de usuario y de 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

El diálogo Opciones incluye valores para criterios de casos atípicos y la definición de un rango para el número de grupos de homólogos.

Criterios para identificar casos atípicos

Estos valores siguientes 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 busca 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.

Avisos

Esta información se ha desarrollado para productos y servicios ofrecidos en los EE.UU. Este material puede estar disponible en IBM en otros idiomas. Sin embargo, es posible que deba ser propietario de una copia del producto o de la versión del producto en dicho idioma para acceder a él.

Es posible que IBM no ofrezca los productos, servicios o características que se tratan en este documento en otros países. El representante local de IBM le puede informar sobre los productos y servicios que están actualmente disponibles en su localidad. 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. En su lugar, se puede utilizar cualquier producto, programa o servicio funcionalmente equivalente que no infrinja los derechos de propiedad intelectual de IBM. Sin embargo, es responsabilidad del usuario evaluar y comprobar el funcionamiento de todo producto, programa o servicio que no sea de IBM.

IBM puede tener patentes o solicitudes de patente en tramitación 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, MD-NC119
Armonk, NY 10504-1785
EE.UU.

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

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

INTERNATIONAL BUSINESS MACHINES CORPORATION PROPORCIONA ESTA PUBLICACIÓN "TAL CUAL", SIN GARANTÍAS DE NINGUNA CLASE, NI EXPLÍCITAS NI IMPLÍCITAS, INCLUYENDO, PERO SIN LIMITARSE A, LAS GARANTÍAS IMPLÍCITAS DE NO VULNERACIÓN, COMERCIALIZACIÓN O ADECUACIÓN A UN PROPÓSITO DETERMINADO. Algunas jurisdicciones no permiten la renuncia a las garantías explícitas o implícitas en determinadas transacciones; por lo tanto, es posible que esta declaración no sea aplicable a su caso.

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.

Las referencias hechas en esta publicación a sitios web que no son de IBM se proporcionan sólo para la comodidad del usuario y no constituyen de modo alguno un aval de esos sitios web. La información de esos sitios web no forma parte de la información de este producto de IBM y la utilización de esos sitios web se realiza bajo la responsabilidad del usuario.

IBM puede utilizar o distribuir la información que se le proporcione del modo que considere adecuado sin incurrir por ello en ninguna obligación con el remitente.

Los titulares de licencias de este programa que deseen tener información sobre el mismo con el fin de permitir: (i) el intercambio de información entre programas creados independientemente y otros programas (incluido este) y (ii) el uso mutuo de la información que se ha intercambiado, deberán ponerse en contacto con:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
EE.UU.

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.

Los ejemplos de datos de rendimiento y de clientes citados se presentan solamente a efectos ilustrativos. Los resultados reales de rendimiento pueden variar en función de las configuraciones específicas y condiciones de operación.

La información relacionada con productos no IBM se ha obtenido de los proveedores de esos productos, de sus anuncios publicados o de otras fuentes disponibles públicamente. IBM no ha probado esos productos y no puede confirmar la exactitud del rendimiento, la compatibilidad ni ninguna otra afirmación relacionada con productos no IBM. Las preguntas sobre las posibilidades de productos que no son de IBM deben dirigirse a los proveedores de esos productos.

Las declaraciones sobre el futuro rumbo o 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 estos nombres son ficticios y cualquier parecido con personas o empresas comerciales reales es pura coincidencia.

LICENCIA DE DERECHOS DE AUTOR:

Esta información contiene programas de aplicación de muestra escritos en lenguaje fuente, los cuales muestran técnicas de programación en diversas plataformas operativas. Puede copiar, modificar y distribuir estos programas de muestra de cualquier modo sin realizar ningún pago a IBM, con el fin de desarrollar, utilizar, comercializar o distribuir programas de aplicación que se ajusten a la interfaz de programación de aplicaciones para la plataforma operativa para la que se han escrito los programas de muestra. Estos ejemplos no se han probado exhaustivamente en todas las condiciones. Por lo tanto, IBM no puede garantizar ni dar por supuesta la fiabilidad, la capacidad de servicio ni la funcionalidad de estos programas. Los programas de muestra se proporcionan "TAL CUAL" sin garantía de ningún tipo. IBM no será responsable de ningún daño derivado del uso de los programas de muestra.

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:

Marcas comerciales

IBM, el logotipo de IBM e ibm.com son marcas registradas o marcas comerciales de International Business Machines Corp., registradas en muchas jurisdicciones en todo el mundo. Otros nombres de productos y servicios podrían ser marcas registradas de IBM u otras compañías. En Internet hay disponible una lista actualizada de las marcas registradas de IBM, en "Copyright and trademark information", 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 comerciales y los logotipos basados en Java son marcas comerciales o registradas de Oracle y/o sus afiliados.

Índice

G

I

M

V

Impreso en España