Guía de aplicaciones de IBM SPSS Modeler 14.2

Nota: Antes de utilizar esta información y el producto, lea la información general en Avisos el p. .

Prefacio

IBM® SPSS® Modeler es el conjunto de programas de minería de datos de IBM Corp. orientado a las empresas. SPSS Modeler ayuda a las organizaciones a mejorar la relación con sus clientes y los ciudadanos a través de la comprensión profunda de los datos. Las organizaciones utilizan la comprensión que les ofrece SPSS Modeler para retener a los clientes más rentables, identificar las oportunidades de venta cruzada, atraer a nuevos clientes, detectar el fraude, reducir el riesgo y mejorar la prestación de servicios del gobierno.

La interfaz visual de SPSS Modeler invita a la pericia empresarial específica de los usuarios, lo que deriva en modelos predictivos más eficaces y la reducción del tiempo necesario para encontrar soluciones. SPSS Modeler ofrece muchas técnicas de modelado tales como pronósticos, clasificaciones, segmentación y algoritmos de detección de asociaciones. Una vez que se crean los modelos, IBM® SPSS® Modeler Solution Publisher permite su distribución en toda la empresa a los encargados de tomar las decisiones o a una base de datos.

Acerca de IBM Business Analytics

El software IBM Business Analytics ofrece información completa, coherente y precisa en la que los órganos de toma de decisiones confían para mejorar el rendimiento comercial. Un conjunto integral de inteligencia empresarial, análisis predictivo,, rendimiento comercial y gestión de estrategias, así como de aplicaciones de análisis le ofrece una información clara, inmediata e interactiva del rendimiento actual y la capacidad para predecir resultados futuros. En combinación con extensas soluciones sectoriales, prácticas probadas y servicios profesionales, las organizaciones de cualquier tamaño pueden conseguir el máximo de productividad y alcanzar mejores resultados.

Como parte de esta familia, el software de análisis predictivo de IBM SPSS ayuda a las organizaciones a predecir eventos futuros y actuar proactivamente según esa información para lograr mejores resultados comerciales. Los clientes comerciales, gubernamentales y académicos de todo el mundo confían en la tecnología de IBM SPSS como ventaja ante la competencia para atraer, retener y hacer crecer los clientes, reduciendo al mismo tiempo el fraude y reduciendo el riesgo. Al incorporar el software de IBM SPSS en sus operaciones diarias, las organizaciones se convierten en empresas predictivas, capaces de dirigir y automatizar decisiones para alcanzar los objetivos comerciales y lograr una ventaja considerable sobre la competencia. Para obtener más información o contactar con un representante, visite *http://www.ibm.com/spss*.

Asistencia técnica

La asistencia técnica está disponible para el mantenimiento de los clientes. Los clientes podrán ponerse en contacto con el servicio de asistencia técnica si desean recibir ayuda sobre la utilización de los productos de IBM Corp. o sobre la instalación en los entornos de hardware admitidos. Para ponerse en contacto con el servicio de asistencia, visite elIBM Corp. sitio Web en *http://www.ibm.com/support*. Prepárese para identificarse, identificar a su organización y su acuerdo de asistencia al solicitar asistencia.

Contenido

Capítulo
1

Acerca de IBM SPSS Modeler

IBM® SPSS® Modeler es un conjunto de herramientas de minería de datos que permite desarrollar rápidamente modelos predictivos mediante técnicas empresariales y utilizarlos en operaciones empresariales para mejorar la toma de decisiones. Con un diseño que sigue el modelo CRISP-DM, estándar del sector, SPSS Modeler admite el proceso completo de minería de datos, desde los propios datos hasta obtener los mejores resultados empresariales.

SPSS Modeler ofrece una gran variedad de métodos de modelado procedentes del aprendizaje automático, la inteligencia artificial y el estadístico. Los métodos disponibles en la paleta de modelado permiten derivar nueva información procedente de los datos y desarrollar modelos predictivos. Cada método tiene ciertos puntos fuertes y es más adecuado para determinados tipos de problemas.

SPSS Modeler puede adquirirse como producto independiente o utilizarse en conjunto con SPSS Modeler Server. También hay disponible cierto número de opciones adicionales que se resumen en las siguientes secciones. Si desea obtener más información, consulte *http://www.ibm.com/software/analytics/spss/products/modeler/*.

IBM SPSS Modeler Server

SPSS Modeler utiliza una arquitectura de cliente/servidor para distribuir peticiones de cliente para operaciones que requieren un uso intensivo de los recursos a un software de servidor de gran potencia, lo que proporciona un rendimiento más rápido con conjuntos de datos de mayor volumen. También puede haber disponibles productos o actualizaciones adicionales que no se incluyan en esta lista. Si desea obtener más información, consulte *http://www.ibm.com/software/analytics/spss/products/modeler/*.

SPSS Modeler. SPSS Modeler es una versión completamente funcional del producto que se instala y ejecuta en el ordenador de escritorio del usuario. Esta versión se puede ejecutar en modo local como un producto independiente o en modo distribuido junto con IBM® SPSS® Modeler Server para mejorar el rendimiento a la hora de trabajar con grandes conjuntos de datos.

SPSS Modeler Server.SPSS Modeler Server se ejecuta ininterrumpidamente en modo de análisis distribuido junto con una o varias instalaciones de IBM® SPSS® Modeler, lo que ofrece un mayor rendimiento cuando se trabaja con grandes conjuntos de datos, ya que las operaciones que requieren un uso intensivo de la memoria se pueden realizar en el servidor sin tener que descargar datos en el equipo cliente. SPSS Modeler Server también ofrece compatibilidad con las capacidades de optimización de SQL y modelado en la base de datos, lo que ofrece ventajas adicionales de rendimiento y automatización. Para ejecutar un análisis debe haber al menos una instalación de SPSS Modeler.

Opciones de IBM SPSS Modeler

Es posible adquirir una licencia de uso de los siguientes componentes y características que pueden utilizarse con SPSS Modeler. Recuerde que también puede haber disponibles productos o actualizaciones adicionales. Si desea obtener más información, consulte *http://www.ibm.com/software/analytics/spss/products/modeler/*.

- Acceso a SPSS Modeler Server, que ofrece una mayor escalabilidad y rendimiento en conjuntos de datos grandes, así como compatibilidad con las capacidades de optimización de SQL y modelado en la base de datos.
- SPSS Modeler Solution Publisher, permite la puntuación automática o en tiempo real fuera del entorno de SPSS Modeler. Si desea obtener más información, consulte el tema IBM SPSS Modeler Solution Publisher en el capítulo 2 en *IBM SPSS Modeler 14.2 Solution Publisher*.
- Adaptadores para permitir la distribución en IBM SPSS Collaboration and Deployment Services o la aplicación IBM SPSS Modeler Advantage de baja intensidad. Si desea obtener más información, consulte el tema Almacenamiento y recuperación de objetos de IBM SPSS Collaboration and Deployment Services Repository en el capítulo 9 en *Manual de usuario de IBM SPSS Modeler 14.2*.

IBM SPSS Text Analytics

IBM® SPSS® Text Analytics es un complemento totalmente integrado en SPSS Modeler que utiliza tecnologías de lingüística avanzada y NLP para procesar con rapidez una gran variedad de datos de texto sin estructurar, extraer y organizar los conceptos clave y agruparlos en categorías. Las categorías y conceptos extraídos se pueden combinar con los datos estructurados existentes, como pueden ser datos demográficos, y se pueden aplicar para modelar utilizando el conjunto completo de herramientas de minería de datos de IBM® SPSS® Modeler para tomar decisiones mejores y más certeras.

- El nodo Text Mining ofrece modelado de conceptos y categorías así como un programa interactivo donde se puede realizar una exploración avanzada de conglomerados y vínculos de texto, crear su propias categorías y refinar las plantillas de recursos lingüísticos.
- Hay diversos formatos de importación compatibles, incluyendo blogs y otros orígenes basados en Web.
- También se incluyen plantillas, bibliotecas y diccionarios personalizados para dominios específicos, como puede ser la terminología CRM y genómica.

Nota: Es necesario disponer de una licencia independiente para acceder a este componente. Si desea obtener más información, consulte *http://www.ibm.com/software/analytics/spss/products/modeler/*.

Documentación de IBM SPSS Modeler

Tiene a su disposición una completa documentación en formato de ayuda en línea desde el menú Ayuda de SPSS Modeler. Se incluye documentación para SPSS Modeler, SPSS Modeler Server y SPSS Modeler Solution Publisher, así como el Manual de aplicaciones y otros materiales de apoyo.

La documentación completa de cada producto en formato PDF está disponible en la carpeta *\Documentation* en cada DVD del producto.

- **Manual del usuario de IBM SPSS Modeler.** Introducción general sobre cómo usar SPSS Modeler, incluyendo cómo crear rutas de datos, tratar valores perdidos, crear expresiones CLEM, trabajar con proyectos e informes y empaquetas rutas para su distribución en IBM SPSS Collaboration and Deployment Services, Predictive Applications o IBM SPSS Modeler Advantage.
- **Nodos Origen, Proceso y Resultado de IBM SPSS Modeler.** Descripciones de todos los nodos utilizados para leer, procesar y dar salida a datos en diferentes formatos. En la práctica, esto implica todos los nodos que no sean nodos de modelado.
- **Nodos de modelado de IBM SPSS Modeler.** Descripciones de todos los nodos utilizados para crear modelos de minería de datos. IBM® SPSS® Modeler ofrece una variedad de métodos de modelado tomados del aprendizaje de las máquinas, la inteligencia artificial y la estadística. Si desea obtener más información, consulte el tema Conceptos básicos sobre nodos de modelado en el capítulo 3 en *Nodos de modelado de IBM SPSS Modeler 14.2*.
- **Manual de algoritmos de IBM SPSS Modeler.** Descripciones de los fundamentos matemáticos de los métodos de modelado que se utilizan en SPSS Modeler.
- **Manual de aplicaciones de IBM SPSS Modeler.** Los ejemplos de esta guía ofrecen introducciones breves y concisas a métodos y técnicas de modelado específicos. También tiene a su disposición una versión en línea de este manual en el menú Ayuda. Si desea obtener más información, consulte el tema Ejemplos de aplicaciones en *Manual de usuario de IBM SPSS Modeler 14.2*.
- **Procesos y automatización de IBM SPSS Modeler.** Información sobre la automatización del sistema a través de procesos, incluidas las propiedades que se pueden utilizar para manipular nodos y rutas.
- **IBM SPSS Modeler Manual de distribución.** Información sobre la ejecución de rutas y escenarios de SPSS Modeler como pasos en trabajos de procesamiento en IBM® SPSS® Collaboration and Deployment Services Deployment Manager.
- **Guía del desarrollador de IBM SPSS Modeler CLEF.**CLEF permite integrar programas de otros fabricantes, como rutinas de procesamiento de datos o algoritmos de modelado como nodos en SPSS Modeler.
- **Manual de minería interna de bases de datos de IBM SPSS Modeler.** Este manual incluye información sobre cómo utilizar la potencia de su base de datos, tanto para mejorar su rendimiento como para ampliar su oferta de capacidades analíticas a través de algoritmos de terceros.
- **Guía de IBM SPSS Modeler Server y su rendimiento.** Información sobre la configuración y administración de IBM® SPSS® Modeler Server.
- **Manual del usuario de IBM SPSS Modeler Administration Console.** Información sobre cómo instalar y utilizar la interfaz de usuario de la consola para supervisar y configurar SPSS Modeler Server. La consola se implementa como complemento de la aplicación Deployment Manager.

- **Manual de IBM SPSS Modeler Solution Publisher.** SPSS Modeler Solution Publisher es un componente complementario que permite a las organizaciones publicar rutas para su uso fuera del entorno estándar de SPSS Modeler.
- **Manual de CRISP-DM de IBM SPSS Modeler.** Manual que explica paso a paso cómo utilizar la metodología de CRISP-DM en la minería de datos con SPSS Modeler.

Ejemplos de aplicaciones

Mientras que las herramientas de minería de datos de SPSS Modeler pueden ayudar a resolver una amplia variedad de problemas organizativos y empresariales, los ejemplos de la aplicación ofrecen introducciones breves y adaptadas de técnicas y métodos de modelado específicos. Los conjuntos de datos utilizados aquí son mucho más pequeños que los enormes almacenes de datos gestionados por algunos analizadores de datos, pero los conceptos y métodos implicados deberían ser escalables a las aplicaciones reales.

Para acceder a los ejemplos pulsando Ejemplos de aplicación en el menú Ayuda de SPSS Modeler. Los archivos de datos y rutas de muestra se instalan en la carpeta *Demos* en el directorio de instalación del producto. Si desea obtener más información, consulte el tema Carpeta Demos en *Manual de usuario de IBM SPSS Modeler 14.2*.

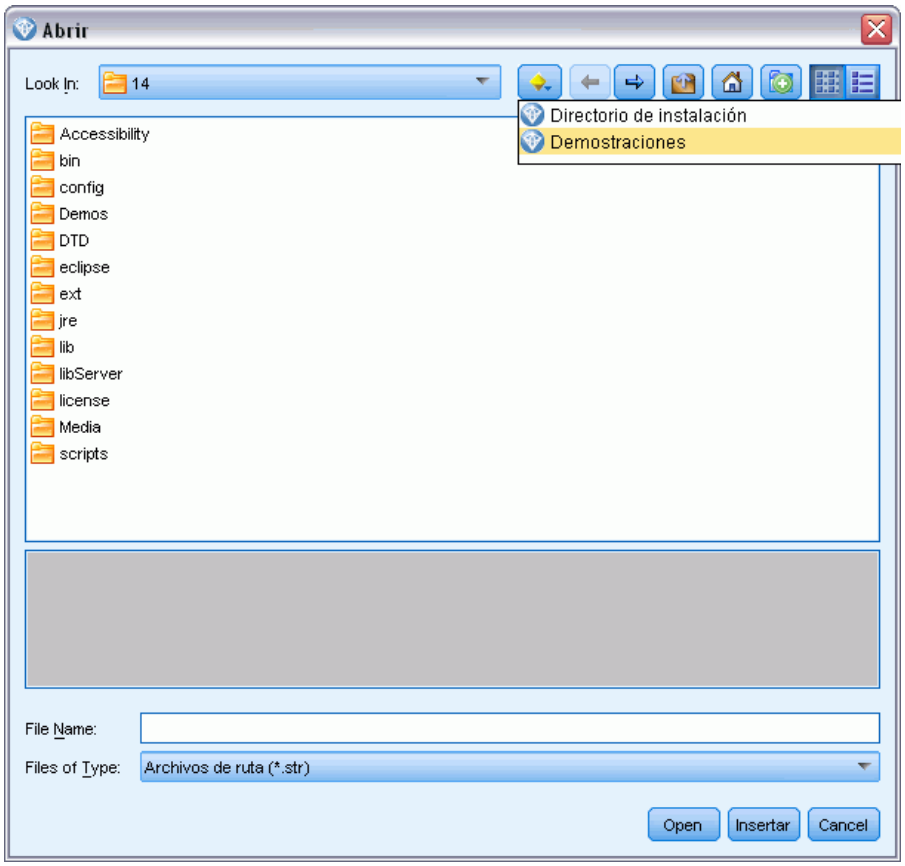
Ejemplos de modelado de base de datos. Consulte los ejemplos que figuran en el Manual de minería interna de bases de datos de *IBM SPSS Modeler*.

Ejemplos de procesos. Consulte los ejemplos que figuran en la Guía de procesos y automatización de *IBM SPSS Modeler*.

Carpeta Demos

Los archivos de datos y rutas de muestra utilizados con los ejemplos de la aplicación se instalan en la carpeta *Demos* en el directorio de instalación del producto. También puede acceder a esta carpeta desde el grupo de programas IBM SPSS Modeler 14.2 del menú Inicio de Windows o pulsando *Demos* de la lista de directorios recientes en el cuadro de diálogo Abrir archivo.

Figura 1-1
Selección de la carpeta Demos desde la lista de directorios utilizados recientemente



Parte I: Introducción

Capítulo
2

Ejemplos de aplicaciones

Mientras que las herramientas de minería de datos de SPSS Modeler pueden ayudar a resolver una amplia variedad de problemas organizativos y empresariales, los ejemplos de la aplicación ofrecen introducciones breves y adaptadas de técnicas y métodos de modelado específicos. Los conjuntos de datos utilizados aquí son mucho más pequeños que los enormes almacenes de datos gestionados por algunos analizadores de datos, pero los conceptos y métodos implicados deberían ser escalables a las aplicaciones reales.

Para acceder a los ejemplos pulsando Ejemplos de aplicación en el menú Ayuda de SPSS Modeler. Los archivos de datos y rutas de muestra se instalan en la carpeta *Demos* en el directorio de instalación del producto. Si desea obtener más información, consulte el tema Carpeta Demos en *Manual de usuario de IBM SPSS Modeler 14.2*.

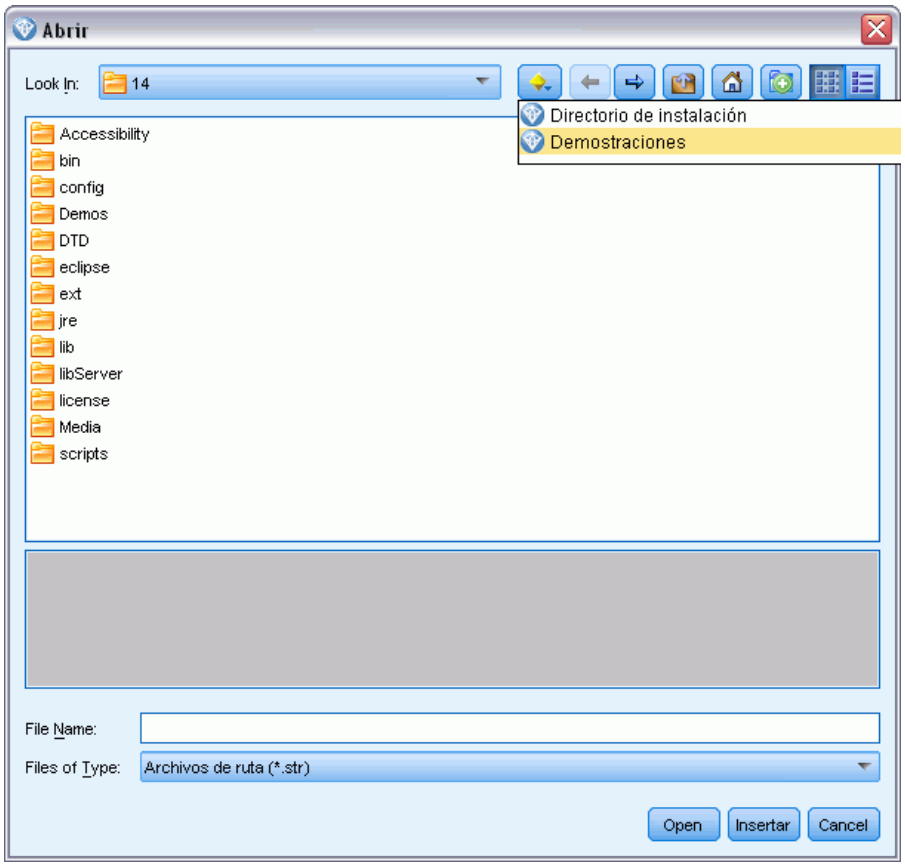
Ejemplos de modelado de base de datos. Consulte los ejemplos que figuran en el Manual de minería interna de bases de datos de *IBM SPSS Modeler*.

Ejemplos de procesos. Consulte los ejemplos que figuran en la Guía de procesos y automatización de *IBM SPSS Modeler*.

Carpeta Demos

Los archivos de datos y rutas de muestra utilizados con los ejemplos de la aplicación se instalan en la carpeta *Demos* en el directorio de instalación del producto. También puede acceder a esta carpeta desde el grupo de programas IBM SPSS Modeler 14.2 del menú Inicio de Windows o pulsando *Demos* de la lista de directorios recientes en el cuadro de diálogo Abrir archivo.

Figura 2-1
Selección de la carpeta Demos desde la lista de directorios utilizados recientemente



Capítulo
3

Conceptos básicos de IBM SPSS Modeler

Primeros pasos

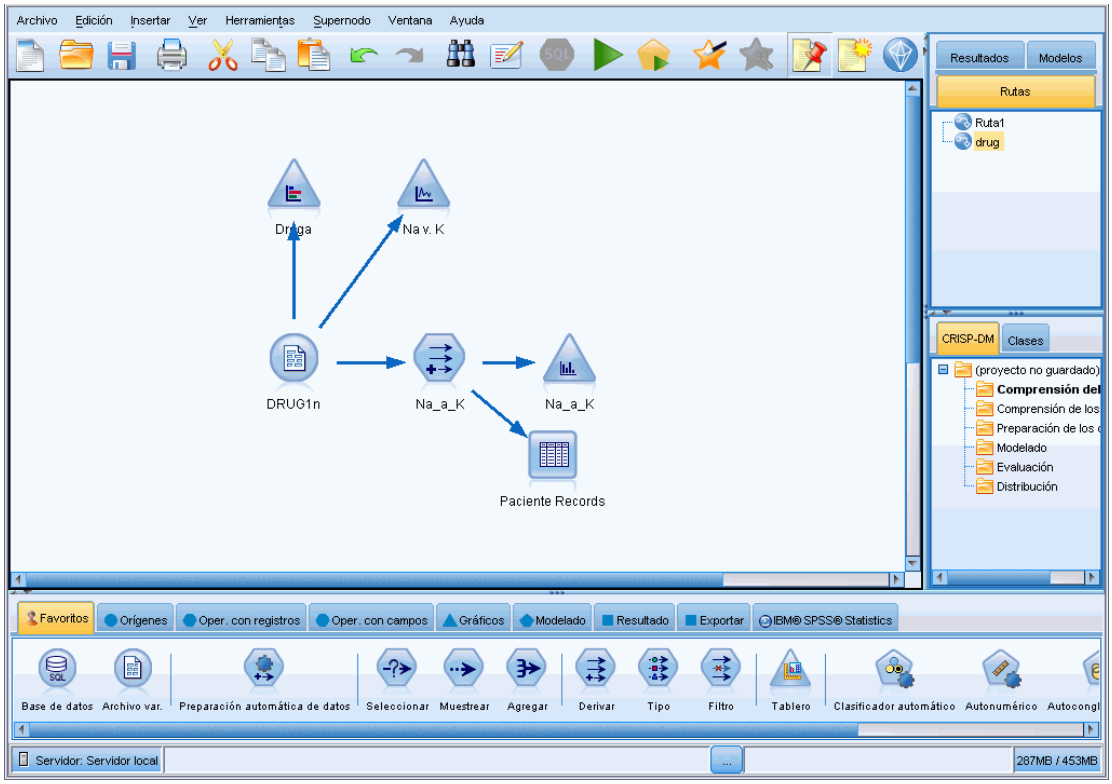
Como aplicación de minería de datos, IBM® SPSS® Modeler ofrece un método estratégico para encontrar relaciones útiles entre grandes conjuntos de datos. Al contrario que los métodos estadísticos más tradicionales, no es necesario saber lo que se está buscando al comenzar. Puede explorar los datos, mediante el ajuste de diferentes modelos y la investigación de diferentes relaciones, hasta que encuentre la información que resulte útil.

Inicie IBM SPSS Modeler

Para iniciar la aplicación, haga clic en:
Inicio > [Todos los] Programas > IBM SPSS Modeler14.2 > IBM SPSS Modeler14.2

La ventana principal se mostrará transcurridos unos segundos.

Figura 3-1
Ventana principal de la aplicación IBM SPSS Modeler



Ejecución desde la línea de comandos

Puede utilizar la línea de comandos del sistema operativo para iniciar IBM® SPSS® Modeler de la siguiente manera:

- En un ordenador en el que se haya instalado IBM® SPSS® Modeler, abra una ventana de DOS o del símbolo del sistema.
- Para iniciar la interfaz de SPSS Modeler en modo interactivo, escriba el comando modelerclient seguido de los argumentos deseados, por ejemplo:

```
modelerclient -stream report.str -execute
```

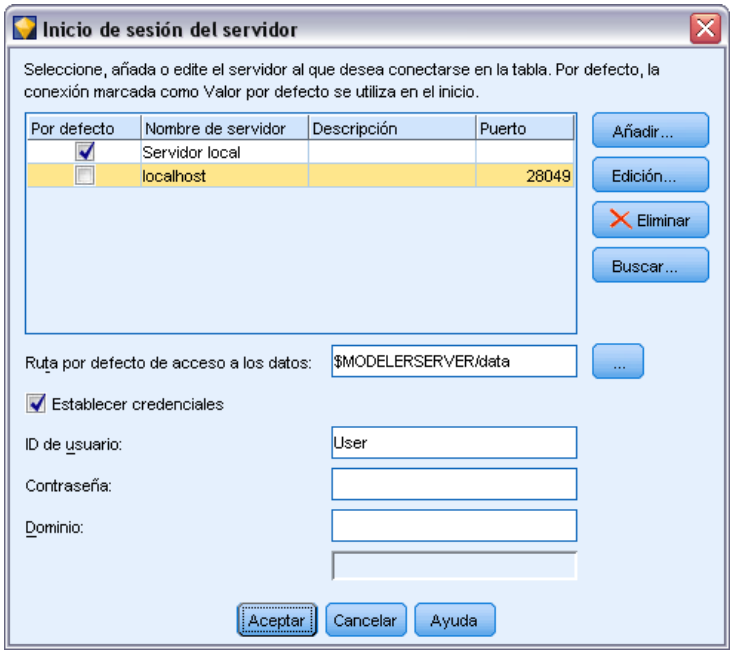
Los argumentos disponibles (modificadores) permiten conectar con un servidor, cargar rutas, ejecutar procesos o especificar otros parámetros, según sea necesario.

Conexión con IBM SPSS Modeler Server

IBM® SPSS® Modeler puede ejecutarse como una aplicación independiente o como un cliente conectado a IBM® SPSS® Modeler Server directamente o a SPSS Modeler Server o un conglomerado de servidores a través del complemento Coordinator of Processes de IBM® SPSS® Collaboration and Deployment Services. El estado de la conexión actual se muestra en la parte inferior izquierda de la ventana de SPSS Modeler.

Siempre que desee conectarse a un servidor, puede introducir manualmente el nombre de servidor al que desee conectarse o seleccione un nombre que haya definido anteriormente. Sin embargo, si tiene IBM SPSS Collaboration and Deployment Services, puede buscar en una lista de servidores o conglomerados de servidores del cuadro de diálogo Inicio de sesión del servidor. La capacidad de buscar entre los servicios de Statistics que se ejecutan en una red está disponible a través de Coordinator of Processes. Si desea obtener más información, consulte el tema Equilibrado de cargas con conglomerados de servidores en el apéndice D en *Guía de administración y rendimiento de IBM SPSS Modeler Server 14.2*.

Figura 3-2
Cuadro de diálogo Acceso al servidor



Para conectar con un servidor

- En el menú Herramientas, pulse en Inicio de sesión del servidor. Se abre el cuadro de diálogo Inicio de sesión del servidor. Si lo prefiere, pulse dos veces con el ratón en el área de estado de la conexión de la ventana de SPSS Modeler.

- En el cuadro de diálogo, especifique las opciones para conectarse al equipo servidor local o seleccione una conexión de la tabla.

- Pulse en Añadir o Edición para añadir o editar una conexión. Si desea obtener más información, consulte el tema Adición y edición de la conexión de IBM SPSS Modeler Server en *Manual de usuario de IBM SPSS Modeler 14.2*.
- Pulse en Buscar para acceder a un servidor o conglomerado de servidores en Coordinator of Processes. Si desea obtener más información, consulte el tema Búsqueda de servidores en IBM SPSS Collaboration and Deployment Services en *Manual de usuario de IBM SPSS Modeler 14.2*.

Tabla Servidor. Esta tabla contiene el conjunto de conexiones de servidor definidas. La tabla muestra la conexión por defecto, el nombre de servidor, la descripción y el número de puerto. Puede añadir manualmente una nueva conexión, así como seleccionar o buscar una conexión existente. Para establecer un servidor específico como la conexión por defecto, seleccione la casilla de verificación en la columna Por defecto de la tabla para la conexión.

Ruta por defecto de acceso a los datos. Especifique la ruta utilizada para los datos del equipo servidor. Pulse en el botón de puntos suspensivos (...) para examinar la ubicación deseada.

Establecer credenciales. Deje esta casilla sin seleccionar para activar la función de **inicio de sesión único**, que tratará de iniciar la sesión del usuario en el servidor con los detalles de nombre de usuario y contraseña del equipo local. Si no es posible el inicio único de sesión o si selecciona esta casilla para desactivar el inicio único de sesión (por ejemplo, para iniciar la sesión en una cuenta de administrador), tendrá activados los siguientes campos para que introduzca las credenciales.

ID de usuario. Introduzca el nombre de usuario con el que se inicia sesión en el servidor.

Contraseña. Introduzca la contraseña asociada al nombre de usuario especificado.

Dominio. Especifique el dominio utilizado para iniciar sesión en el servidor. El nombre de dominio es obligatorio sólo si el equipo servidor está en un dominio de Windows distinto que el del equipo cliente.

▶ Pulse en Aceptar para completar la conexión.

Desconexión de un servidor

▶ En el menú Herramientas, pulse en Inicio de sesión del servidor. Se abre el cuadro de diálogo Inicio de sesión del servidor. Si lo prefiere, pulse dos veces con el ratón en el área de estado de la conexión de la ventana de SPSS Modeler.

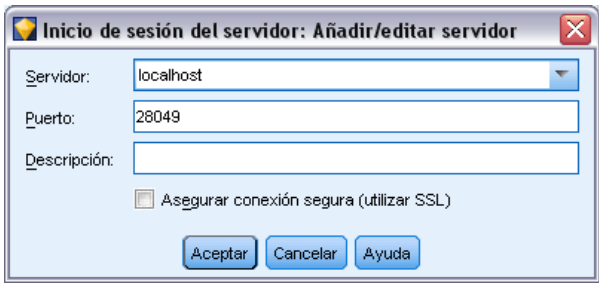
▶ En el cuadro de diálogo, seleccione el Servidor local y pulse en Aceptar.

Adición y edición de la conexión de IBM SPSS Modeler Server

Puede editar o añadir manualmente una conexión de servidor en el cuadro de diálogo Inicio de sesión del servidor. Si pulsa en Añadir, puede acceder al cuadro de diálogo Añadir/editar servidor vacío en el que puede introducir los detalles de conexión de servidor. Al seleccionar una conexión existente y pulsar en Editar en el cuadro de diálogo Inicio de sesión del servidor, se abre el cuadro de diálogo Añadir/editar servidor con los detalles de dicha conexión de modo que puede realizar cualquier cambio.

Nota: No puede editar una conexión de servidor que se haya añadido desde IBM® SPSS® Collaboration and Deployment Services, ya que el nombre, puerto y otros detalles se definen en IBM SPSS Collaboration and Deployment Services.

Figura 3-3
Cuadro de diálogo Inicio de sesión del servidor: Añadir/editar servidor



Adición de conexiones de servidor

- En el menú Herramientas, pulse en Inicio de sesión del servidor. Se abre el cuadro de diálogo Inicio de sesión del servidor.
- En este cuadro de diálogo, pulse en Añadir. Se abre el cuadro de diálogo Inicio de sesión del servidor: Añadir/editar servidor.
- Introduzca los detalles de conexión de servidor y pulse en Aceptar para guardar la conexión y volver al cuadro de diálogo Inicio de sesión del servidor.
  - **Servidor.** Especifique un servidor disponible o seleccione uno de la lista. El equipo servidor se puede identificar por un nombre alfanumérico (por ejemplo, *miservidor*) o por una dirección IP asignada al equipo servidor (por ejemplo, 202.123.456.78).
  - **Puerto.** Especifique el número de puerto en el que el servidor escucha. Si no funciona el número de puerto por defecto, solicite el número de puerto correcto al administrador del sistema.
  - **Descripción.** Introduzca una descripción opcional para esta conexión de servidor.
  - **Asegurar conexión segura (utilizar SSL).** Especifica si se debe usar una conexión SSL (del inglés **Secure Sockets Layer**, capa de sockets seguros). SSL es un protocolo normalmente utilizado para asegurar el conjunto de datos que se envía a través de una red. Para utilizar esta función, SSL debe estar activado en el servidor que aloja IBM® SPSS® Modeler Server. Si es preciso, póngase en contacto con el administrador local para obtener más detalles.

Edición de conexiones de servidor

- En el menú Herramientas, pulse en Inicio de sesión del servidor. Se abre el cuadro de diálogo Inicio de sesión del servidor.
- En este cuadro de diálogo, seleccione la conexión que desee editar y, a continuación, pulse en Editar. Se abre el cuadro de diálogo Inicio de sesión del servidor: Añadir/editar servidor.
- Cambie los detalles de conexión de servidor y pulse en Aceptar para guardar los cambios y volver al cuadro de diálogo Inicio de sesión del servidor.

Búsqueda de servidores en IBM SPSS Collaboration and Deployment Services

En lugar de introducir una conexión de servidor manualmente, puede seleccionar un servidor o conglomerado de servidores disponible en la red a través de Coordinator of Processes, disponible en IBM® SPSS® Collaboration and Deployment Services. Un conglomerado de servidores es un grupo de servidores entre los que Coordinator of Processes determina el servidor más adecuado para responder a una solicitud de procesamiento. Si desea obtener más información, consulte el tema Equilibrado de cargas con conglomerados de servidores en el apéndice D en *Guía de administración y rendimiento de IBM SPSS Modeler Server 14.2*.

Aunque puede añadir servidores manualmente al cuadro de diálogo Inicio de sesión del servidor, la búsqueda de servidores disponibles le permite conectarse a servidores sin que sea necesario que conozca el nombre de servidor y número de puerto correctos. Esta información se proporciona automáticamente. Sin embargo, todavía necesita la información de inicio de sesión correcta, como el nombre de usuario, dominio y contraseña.

Nota: Si no tiene acceso a la capacidad Coordinator of Processes, todavía puede introducir manualmente el nombre de servidor al que desee conectarse o seleccionar un nombre que haya definido anteriormente. Si desea obtener más información, consulte el tema Adición y edición de la conexión de IBM SPSS Modeler Server en *Manual de usuario de IBM SPSS Modeler 14.2*.

Figura 3-4
Cuadro de diálogo Buscar servidores

Búsqueda de servidores y conglomerados

- En el menú Herramientas, pulse en Inicio de sesión del servidor. Se abre el cuadro de diálogo Inicio de sesión del servidor.
- En este cuadro de diálogo, pulse en Buscar para abrir el cuadro de diálogo Buscar servidores. Si no ha iniciado sesión en IBM SPSS Collaboration and Deployment Services cuando intente buscar en Coordinator of Processes, se le pedirá que lo haga. Si desea obtener más información, consulte el tema Conexión con IBM SPSS Collaboration and Deployment Services Repository en el capítulo 9 en *Manual de usuario de IBM SPSS Modeler 14.2*.
- Seleccione el servidor o el conglomerado de servidores de la lista.
- Pulse en Aceptar para cerrar el cuadro de diálogo y añadir esta conexión a la tabla en el cuadro de diálogo Inicio de sesión del servidor.

Modificación del directorio temporal

IBM® SPSS® Modeler Server realiza algunas operaciones que requieren la creación de archivos temporales. Por defecto, IBM® SPSS® Modeler utiliza el directorio temporal del sistema para crear archivos temporales. Se puede modificar la ubicación del directorio temporal con los pasos siguientes.

- Cree un nuevo directorio denominado *spss* y un subdirectorio denominado *servertemp*.
- Edite *options.cfg*, que se encuentra en el directorio */config* del directorio de instalación de SPSS Modeler. Edite el parámetro temp_directory de este archivo para leer: temp_directory, "C:/spss/servertemp".
- A continuación, es necesario reiniciar el servicio SPSS Modeler Server. Esta operación se puede realizar pulsando en la pestaña Servicios del Panel de control de Windows. Es necesario detener el servicio e iniciarlo de nuevo para activar los cambios realizados. Cuando se reinicie el equipo también se reiniciará el servicio.

  Todos los archivos temporales se escribirán a partir de este momento en este directorio.

  Nota: El error más habitual cuando se intenta realizar esta acción es el uso de un tipo de barras incorrecto. Debido al historial de UNIX de SPSS Modeler, se utilizan las barras diagonales.

Inicio de varias sesiones de IBM SPSS Modeler

Si necesita iniciar más de una sesión de IBM® SPSS® Modeler a la vez, deberá realizar algunos cambios en la configuración de IBM® SPSS® Modeler y Windows. Por ejemplo, puede que necesite hacerlo si tiene dos licencias de servidor independientes y desee ejecutar dos rutas frente a dos servidores diferentes del mismo equipo cliente.

Para activar varias sesiones de SPSS Modeler:

- Clic en:
  Inicio > [Todos los] Programas > IBM SPSS Modeler14.2
- En el acceso directo de IBM SPSS Modeler14.2 (el que tiene un icono), pulse con el botón derecho del ratón y seleccione Propiedades.
- En el cuadro de texto Objetivo, añada -noshare al final de la cadena.
- En Windows Explorer, seleccione:
  Herramientas > Opciones de carpeta...
- En la pestaña Tipos de archivo, seleccione la opción Ruta de SPSS Modeler y pulse en Opciones avanzadas.
- En el cuadro de diálogo Editar tipo de archivo, seleccione Abrir con SPSS Modeler y pulse en Editar.
- En el cuadro de texto Aplicación utilizada para realizar la acción, añada -noshare delante del argumento -stream.

Conceptos básicos sobre la interfaz de IBM SPSS Modeler

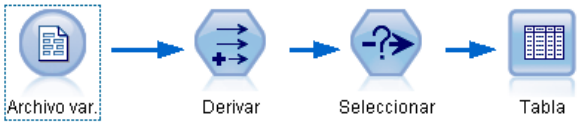
En cada punto del proceso de minería de datos, la interfaz de IBM® SPSS® Modeler fácil de usar implica el uso de técnicas empresariales. Los algoritmos de modelado, tales como predicción, clasificación, segmentación y detección de asociaciones, garantizan la obtención de modelos exactos y potentes. Los resultados del modelo se pueden distribuir y leer fácilmente en bases de datos, IBM® SPSS® Statistics y en una amplia variedad de aplicaciones.

El trabajo con SPSS Modeler es un proceso de tres pasos para trabajar con datos.

- En primer lugar, lee los datos en SPSS Modeler.
- A continuación, ejecuta los datos mediante una serie de manipulaciones.
- Por último, envía los datos a un destino.

Esta secuencia de operaciones se denomina **ruta de datos** porque los datos fluyen registro por registro desde el origen pasando por cada manipulación y, finalmente, llega al destino, que puede ser un modelo o un tipo de datos de resultados.

Figura 3-5
Una ruta simple



Lienzo de rutas de IBM SPSS Modeler

El lienzo de rutas es el área más grande de la ventana de IBM® SPSS® Modeler y en éste se generan y manipulan rutas de datos.

Las rutas se crean dibujando diagramas de operaciones de datos relevantes para su negocio en el lienzo principal de la interfaz. Cada operación se representa con un icono o un **nodo** y los nodos están vinculados entre sí en una **ruta** que representa el flujo de datos en cada operación.

Se puede trabajar con varias rutas al mismo tiempo en SPSS Modeler, en el mismo lienzo de rutas o abriendo uno nuevo. Durante una sesión, las rutas se almacenan en el administrador de rutas, en la parte superior derecha de la ventana de SPSS Modeler.

Paleta de nodos

La mayoría de los datos y las herramientas de modelado de IBM® SPSS® Modeler se encuentran en la **Paleta de nodos**, situadas por la parte inferior de la ventana bajo el lienzo de rutas.

Por ejemplo, la pestaña Paleta Oper. con registros contiene nodos que puede utilizar para realizar operaciones en los **registros** de datos, como la selección, la fusión y la adición.

Para añadir nodos al lienzo, pulse dos veces en los iconos de la Paleta de nodos o arrástrelos y suéltelos en el lienzo. A continuación, conéctelos para crear una **ruta**, que represente el flujo de datos.

Figura 3-6
Pestaña Operaciones con registros de la paleta de nodos

Cada pestaña de paleta contiene una colección de nodos relacionados entre sí que se utilizan en distintas fases de las operaciones de rutas, tales como:

- **Orígenes.** Los nodos introducen datos en SPSS Modeler.
- **Operaciones con registros** Los nodos realizan operaciones en los **registros** de datos como la selección, la fusión y la adición.
- **Operaciones con campos** Los nodos realizan operaciones en los **campos** de datos como el filtrado, la derivación de campos nuevos y la determinación del nivel de medición de campos dados.
- **Gráficos.** Los nodos muestran gráficamente los datos antes y después del modelado. Entre ellos se incluyen gráficos, histogramas, nodos de malla y diagramas de evaluación.
- **Modelado.** Los nodos utilizan los algoritmos de modelado disponibles en SPSS Modeler, tales como las redes neuronales, los árboles de decisión, los algoritmos de conglomerados y las secuencias de datos.
- **Modelado de base de datos.** Los nodos utilizan los algoritmos de modelado disponibles en las bases de datos Microsoft SQL Server, IBM DB2 y Oracle.
- **Resultados.** Los nodos generan una diversidad de resultados para los datos, gráficos y resultados de modelos que pueden visualizarse en SPSS Modeler.
- **Exportar.** Los nodos generan una diversidad de resultados que pueden visualizarse en aplicaciones externas, como IBM® SPSS® Data Collection o Excel.
- **SPSS Statistics.** Los nodos importan datos y exportan datos a IBM® SPSS® Statistics, ejecutando también procedimientos de SPSS Statistics.

Una vez que se familiarice más con SPSS Modeler, podrá personalizar el contenido de la paleta para su propio uso. Si desea obtener más información, consulte el tema Personalización de la paleta de nodos en el capítulo 12 en *Manual de usuario de IBM SPSS Modeler 14.2*.

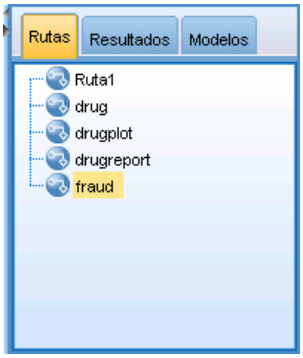
Debajo de la Paleta de nodos, hay un panel de informe que proporciona información sobre el progreso de distintas operaciones, como la lectura de datos en la ruta de datos. Situado también debajo de la Paleta de nodos, hay un panel de estado que proporciona información acerca de la operación que está realizando la aplicación e indica cuándo son necesarios los comentarios del usuario.

Administradores de IBM SPSS Modeler

En la parte superior derecha de la ventana se encuentra el panel de administradores. Este panel cuenta con tres pestañas que se utilizan para administrar rutas, resultados y modelos.

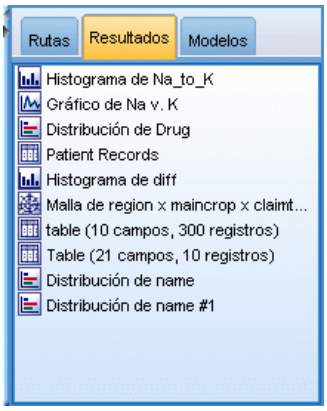
Se puede utilizar la pestaña Rutas para abrir, cambiar nombres, guardar o eliminar las rutas creadas en una sesión.

Figura 3-7
Pestaña Rutas



La pestaña Resultados contiene una serie de archivos, como gráficos y tablas, generados mediante operaciones de rutas en IBM® SPSS® Modeler. Puede mostrar, guardar, cambiar el nombre y cerrar las tablas, gráficos e informes que se enumeran en esta pestaña.

Figura 3-8
Pestaña Resultados



La pestaña Modelos es la pestaña de administradores más potente. Esta pestaña contiene todos los **nugget** de modelo, que son modelos generados en SPSS Modeler, para la sesión actual. Estos modelos se pueden examinar directamente en la pestaña Modelos o añadirlos a la ruta en el lienzo.

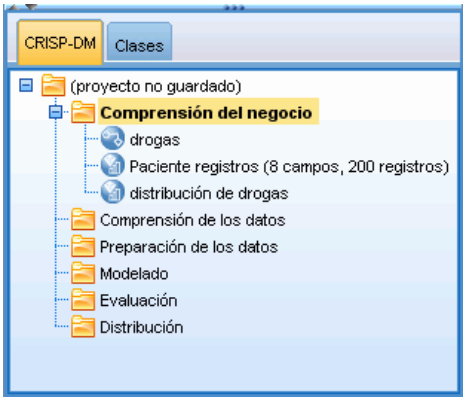
Figura 3-9
Pestaña Modelos que contiene nuggets de modelo

Proyectos de IBM SPSS Modeler

En la parte inferior derecha de la ventana se encuentra el panel de proyectos, que se utiliza para crear y administrar los **proyectos** de minería de datos (grupo de archivos relacionados con una tarea de minería de datos). Existen dos formas de ver los proyectos que se crean en IBM® SPSS® Modeler—: en la vista Clases y la vista CRISP-DM.

La pestaña CRISP-DM permite organizar los proyectos según el proceso CRISP-DM (Cross-Industry Standard Process for Data Mining), una metodología independiente y probada en el sector. Los analizadores de datos con o sin experiencia pueden utilizar la herramienta CRISP-DM para mejorar la organización y la comunicación de los esfuerzos.

Figura 3-10
vista CRISP-DM



La pestaña Clases permite organizar el trabajo en SPSS Modeler de forma categórica, por los tipos de los objetos que se hayan creado. Esta vista resulta útil al realizar un inventario de datos, rutas y modelos.

Figura 3-11
Vista Clases

Barra de herramientas de IBM SPSS Modeler

En la parte superior de la ventana de IBM® SPSS® Modeler hay una barra de herramientas con iconos que proporciona una serie de funciones muy útiles. A continuación se detallan los botones de la barra de herramientas y sus funciones.

	Crear una nueva ruta		Abrir una ruta existente
	Guardar la ruta actual		Imprimir la ruta actual
	Cortar & mover la selección al Portapapeles		Copiar la selección al Portapapeles
	Pegar selección		Deshacer la última acción
	Rehacer		Buscar nodos
	Editar las propiedades de la ruta		Presentación preliminar de generación de SQL
	Ejecutar ruta actual		Ejecutar selección de ruta

Detener ruta (sólo se activa durante la ejecución de la ruta)

Añadir Supernodo

Acercar Supernodo (sólo con Supernodos)

Alejar Supernodo (sólo con Supernodos)

Sin marcas en la ruta

Insertar comentario

Ocultar comentarios de ruta (si los hay)

Mostrar comentarios de ruta ocultos

Abrir una ruta existente en IBM® SPSS® Modeler Advantage

Los comentarios de rutas constan de comentarios, enlaces de modelos e indicaciones de las ramas de puntuación.

Si desea obtener más información sobre los comentarios de rutas, consulte Adición de comentarios y anotaciones a nodos y rutas el p. .

Para obtener más información sobre las indicaciones de las ramas de puntuación, consulte La rama de puntuación el p. .

Los enlaces de modelos se describen en el manual *Nodos de modelado de IBM SPSS*.

Personalización de la barra de herramientas

Puede cambiar varios aspectos de la barra de herramientas, como:

- Si se visualiza
- Si los iconos tienen información sobre herramientas
- Si utiliza iconos grandes o pequeños

Para activar o desactivar la barra de herramientas:

▶ En el menú principal, pulse en:
Ver > Barra de herramientas > Visualización

Para cambiar la información sobre herramientas o la configuración del tamaño de iconos:

▶ En el menú principal, pulse en:
Ver > Barra de herramientas > Personalizar

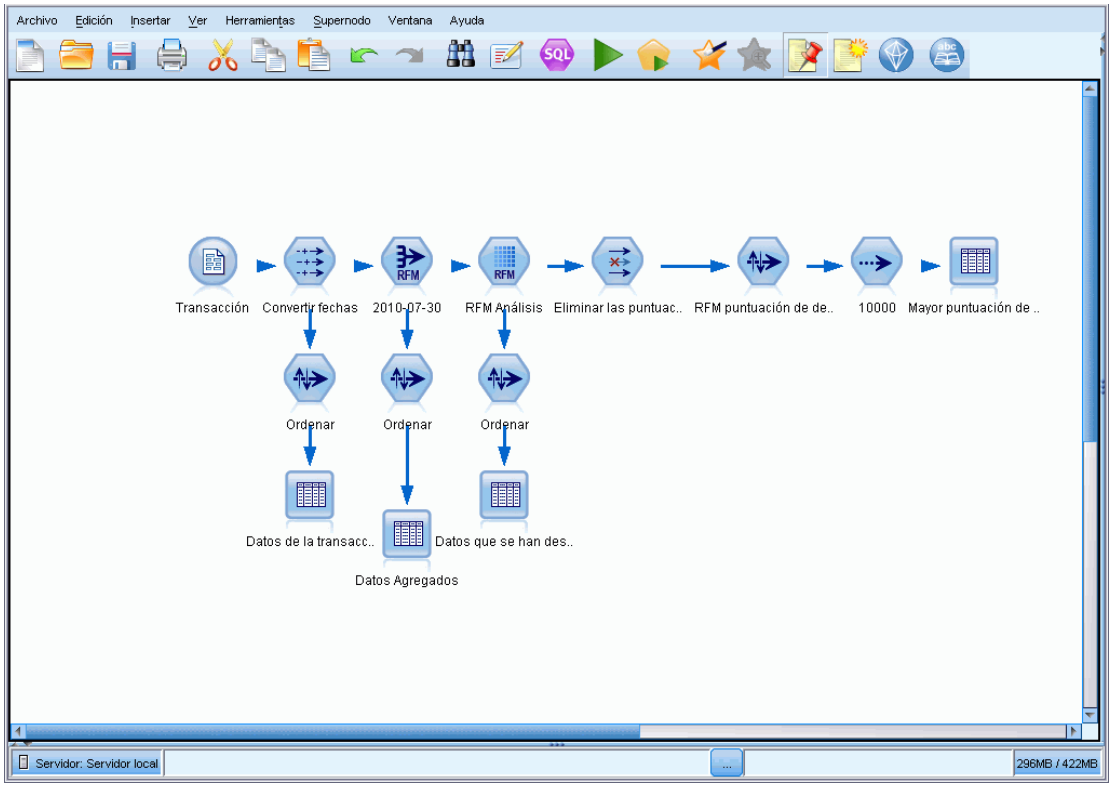
Pulse Mostrar información sobre herramientas o Botones grandes, según sea necesario.

Personalización de la ventana de IBM SPSS Modeler

Se puede cambiar el tamaño de las herramientas o cerrarlas con los separadores de las distintas partes de la interfaz de IBM® SPSS® Modeler. Por ejemplo, si trabaja con una ruta larga, puede utilizar las flechas pequeñas situadas en cada separador para cerrar la paleta de nodos, el panel de administradores y el de proyectos. De esta forma se maximiza el lienzo de rutas y se proporciona espacio de trabajo suficiente para varias rutas o para rutas grandes.

También puede pulsar desde el menú Ver en Paleta de nodos, Administradores o Proyecto para activar o desactivar la visualización de estos elementos.

Figura 3-12
Lienzo de rutas maximizado



En lugar de cerrar la paleta de nodos o los paneles de administradores y de proyectos, también se puede utilizar el lienzo de rutas como una página desplazable moviéndolo vertical y horizontalmente con las barras de desplazamiento situadas en el lateral y en la parte inferior de la ventana de SPSS Modeler.

También puede controlar la visualización de las marcas de pantalla, que consta de los comentarios de rutas, los enlaces de modelos y las indicaciones de las ramas de puntuación. Para activar o desactivar esta visualización, pulse:

Ver > Comentarios de rutas

Utilización del ratón en IBM SPSS Modeler

Los usos más comunes del ratón en IBM® SPSS® Modeler incluyen los siguientes:

- **Pulsar una vez.** Utilice el botón derecho o el izquierdo del ratón para seleccionar las opciones de los menús, abrir menús emergentes y acceder a otros controles y opciones estándar. Pulsar y mantener pulsado el botón para mover y arrastrar nodos.
- **Pulsar dos veces.** Pulse dos veces con el botón izquierdo del ratón para colocar nodos en el lienzo de rutas y editar nodos existentes.
- **Pulsar con el botón central.** Pulse con el botón central del ratón y arrastre el cursor para conectar nodos en el lienzo de rutas. Pulse dos veces con el botón central del ratón para desconectar un nodo. Si el ratón no tiene un botón central, se puede simular esta función pulsando la tecla Alt a la vez que pulsa con el ratón y se arrastra.

Uso de teclas de método abreviado

Muchas operaciones de programación visual de IBM® SPSS® Modeler poseen teclas de acceso rápido asociadas. Por ejemplo, se puede eliminar un nodo pulsando en el nodo y en la tecla Supr del teclado. Del mismo modo, se puede guardar una ruta de forma rápida manteniendo pulsada la tecla Ctrl y pulsando la tecla S. Comandos de control como éste se indican con una combinación de Ctrl con otra tecla; por ejemplo, Ctrl+S.

En las operaciones estándar de Windows se utilizan varias teclas de método abreviado, tales como Ctrl+X para cortar. Estos métodos abreviados son compatibles con SPSS Modeler junto con los siguientes métodos abreviados de aplicaciones específicas.

Nota: En algunos casos, las teclas de método abreviado antiguas de SPSS Modeler entran en conflicto con las de Windows. Estos métodos abreviados antiguos son compatibles si además se pulsa la tecla Alt. Por ejemplo, se puede utilizar Ctrl+Alt+C para activar y desactivar la caché.

Tabla 3-1
Teclas de método abreviado compatibles

Tecla de método abreviado	Función
Ctrl+A	Seleccionar todo
Ctrl+X	Cortar
Ctrl+N	Nueva ruta
Ctrl+O	Abrir una ruta existente
Ctrl+P	Imprimir
Ctrl+C	Copiar
Ctrl+V	Pegar
Ctrl + Z	Deshacer
Ctrl+Q	Selecciona todos los nodos que se encuentren por debajo del nodo seleccionado
Ctrl+W	Anule la selección de todos los nodos posteriores de la ruta (se conmuta con Ctrl+Q)
Ctrl+E	Ejecutar desde el nodo seleccionado
Ctrl+S	Guarda la ruta actual

Tecla de método abreviado	Función
Alt+Teclas de flecha	Mueve los nodos seleccionados en el lienzo de rutas en la dirección de la flecha utilizada.
Mayús+F10	Abre el menú emergente del nodo seleccionado

Tabla 3-2
Métodos abreviados compatibles para teclas de acceso rápido anteriores

Tecla de método abreviado	Función
Ctrl+Alt+D	Duplica el nodo
Ctrl+Alt+L	Carga el nodo
Ctrl+Alt+R	Cambia el nombre del nodo
Ctrl+Alt+U	Crea un nodo Datos Usuario
Ctrl+Alt+C	Conmutar caché activada/desactivada
Ctrl+Alt+F	Vacía la caché
Ctrl+Alt+X	Expande el Supernodo
Ctrl+Alt+Z	Acercar/alejar
Suprimir	Elimina el nodo o la conexión

Impresión

Se pueden imprimir los siguientes objetos en IBM® SPSS® Modeler:

- Diagramas de ruta
- Gráficos
- Tablas
- Informes (del nodo Informe y de los informes de proyectos)
- Procesos (desde los cuadros de diálogo de propiedades de la ruta, Proceso independiente o Proceso de Supernodo)
- Modelos (exploradores de modelos, pestañas de cuadros de diálogo con la vista actual, visores de árboles)
- Anotaciones (mediante la pestaña Anotaciones de resultados)

Para imprimir un objeto:

- Para imprimir sin presentación preliminar, pulse en el botón Imprimir de la barra de herramientas.
- Para configurar la página antes de imprimir, seleccione Configurar página en el menú Archivo.
- Para mostrar la representación preliminar, seleccione Presentación preliminar en el menú Archivo.
- Para que se muestre el cuadro de diálogo de impresión estándar con las opciones para seleccionar las impresoras y especificar las opciones de aspecto, seleccione Imprimir en el menú Archivo.

Automatización de IBM SPSS Modeler

Debido a que la minería de datos avanzada puede ser un proceso complejo y a menudo largo, IBM® SPSS® Modeler incluye varios tipos de soporte de codificación y automatización.

- **Control Language for Expression Manipulation** (CLEM) es un lenguaje para analizar y manipular los datos que fluyen en las rutas de SPSS Modeler. Los analistas de datos suelen utilizar CLEM en las operaciones de rutas para realizar tareas tan simples como derivar beneficios de datos de costes e ingresos, o tan complejas como transformar datos del registro Web en un conjunto de campos y registros con información útil. Si desea obtener más información, consulte el tema Acerca de CLEM en el capítulo 7 en *Manual de usuario de IBM SPSS Modeler 14.2*.
- **El procesamiento en** es una herramienta potente para automatizar procesos en la interfaz de usuario. Los procesos pueden realizar las mismas acciones que los usuarios llevan a cabo con un ratón o un teclado. Se pueden definir opciones para nodos y realizar derivaciones utilizando un subconjunto de CLEM. También pueden especificar los resultados y manipular los modelos generados. Si desea obtener más información, consulte el tema Conceptos básicos del procesamiento en el capítulo 2 en *Guía de procesos y automatización de IBM SPSS Modeler 14.2*.

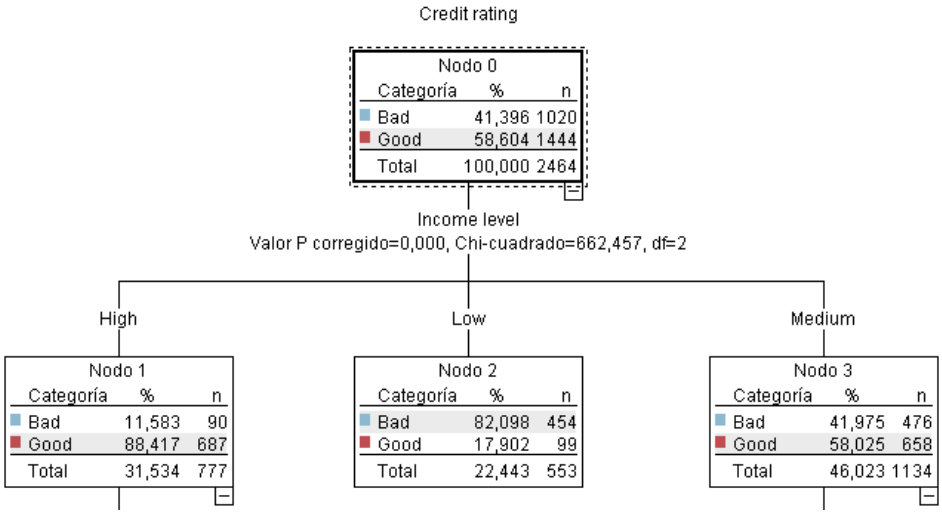
Capítulo
4

Introducción al modelado

Un modelo es un conjunto de reglas, fórmulas o ecuaciones que puede utilizarse para pronosticar un resultado basándose en un conjunto de campos o variables de entrada. Por ejemplo, puede que una institución financiera utilice un modelo para predecir la probabilidad de que los solicitantes de un préstamo sean un riesgo bueno o malo, basándose en información que ya se conoce sobre solicitantes anteriores.

La capacidad de pronosticar un resultado es el objetivo central del análisis predictivo y la comprensión del proceso de modelado es la clave para utilizar IBM® SPSS® Modeler.

Figura 4-1
Modelo de árbol de decisión sencillo



Este ejemplo utiliza un modelo de **árbol de decisión** que clasifica los registros (y pronostica una respuesta) utilizando una serie de reglas de decisión, por ejemplo:

```
IF ingreso = Medio
AND tarjetas <5
THEN -> "Bueno"
```

Aunque este ejemplo utiliza un modelo CHAID (Detección automática de interacciones mediante chi-cuadrado), se presenta como una introducción general y la mayoría de los conceptos se aplica de forma amplia en otros tipos de modelado de SPSS Modeler.

Para comprender cualquier modelo, primero debe comprender los datos que incluye. Los datos de este ejemplo contienen información sobre los clientes de un banco. Se utilizan los siguientes campos:

Nombre de campo	**Descripción**
Valoración_crédito	Valoración de crédito 0=Malo, 1=Bueno, 9=Valores perdidos
Edad	Edad en años
Ingresos	Nivel de ingresos: 1=Bajo, 2=Medio, 3=Alto
Tarjetas_crédito	Número de tarjetas de crédito en propiedad: 1=Menos de cinco, 2=Cinco o más
Educación	Nivel educativo: 1=Instituto, 2=Universidad
Préstamo_coche	Número de préstamos de coche asumidos: 1=Ninguno o uno, 2=Más de dos

El banco mantiene una base de datos con información histórica sobre los clientes a los que el banco ha concedido préstamos, incluido si los han reintegrado o no (Valoración de crédito = Bueno) o causado mora en el pago de dichos préstamos (Valoración de crédito = Malo). Con los datos existentes, el banco quiere generar un modelo que le permita predecir la probabilidad de mora del préstamo de los posibles solicitantes futuros de un préstamo.

Al utilizar un modelo de árbol de decisión, puede analizar las características de los dos grupos de clientes y predecir la probabilidad de mora del préstamo.

Este ejemplo utiliza la ruta denominada *modelingintro.str*, disponible en la carpeta *Demos* bajo la subcarpeta *streams*. El archivo de datos es *tree_credit.sav*. Si desea obtener más información, consulte el tema Carpeta Demos en el capítulo 1 en *Manual de usuario de IBM SPSS Modeler 14.2*.

Veamos la ruta más detenidamente.

- Seleccione lo siguiente en el menú principal:
  File > Abrir ruta

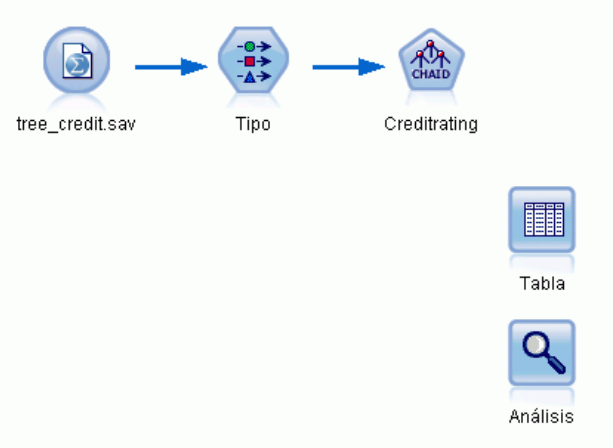
- Pulse en el icono de nugget dorado de la barra de herramientas del cuadro de diálogo Abrir y seleccione la carpeta Demos.

- Pulse dos veces en la carpeta *streams*.

- Pulse dos veces en el archivo llamado *modelingintro.str*.

Generación de la ruta

Figura 4-2
Modelado de la ruta



Para crear una ruta que cree un modelo, necesitamos al menos tres elementos:

- Un nodo de origen que lea los datos de un origen externo, en este caso, un archivo de datos IBM® SPSS® Statistics.
- Un nodo de origen o nodo Tipo que especifique propiedades de campo, como el nivel de medición (el tipo de datos que contiene el campo) y el papel de cada campo como objetivo o entrada en modelado.
- Un nodo de modelado que genera un nugget de modelo cuando se ejecuta la ruta.

En este ejemplo estamos usando un nodo de modelado CHAID. CHAID, o Detección automática de interacciones mediante chi-cuadrado, es un método de clasificación que genera árboles de decisión utilizando un tipo específico de estadísticos denominados estadísticos chi-cuadrado para determinar los mejores lugares para realizar las divisiones en el árbol de decisión.

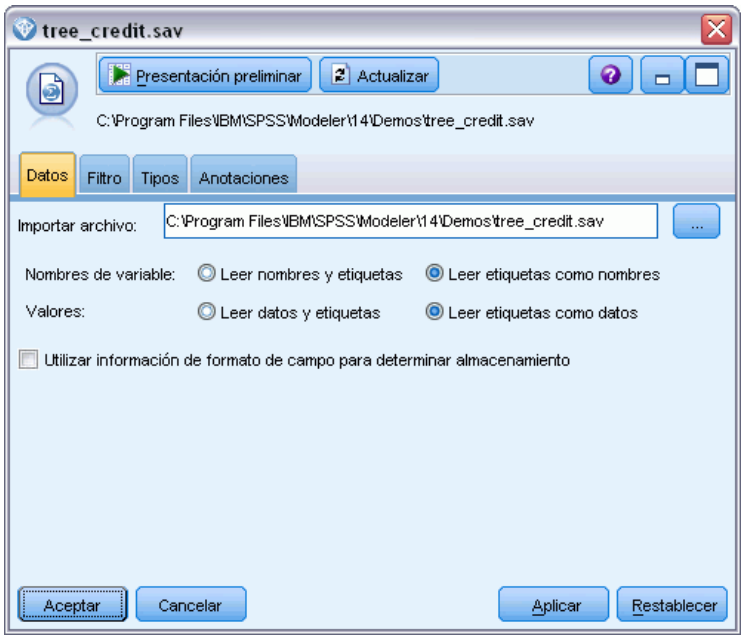
Si se especifican niveles de medición en el nodo de origen, se puede eliminar el nodo Tipo independiente. Funcionalmente, el resultado es el mismo.

Esta ruta también tiene los nodos Tabla y Análisis que se utilizarán para ver los resultados de puntuación después de crear el nugget de modelo y añadirlo a la ruta.

El nodo de origen Archivo Statistics lee los datos en formato SPSS Statistics del archivo de datos *tree_credit.sav*, que está instalado en la carpeta *Demos*. (Una variable especial denominada *$CLEO_DEMOS* se utiliza para hacer referencia a esta carpeta en la instalación actual de IBM®

SPSS® Modeler. Esto garantiza que la ruta será válida independientemente de la carpeta o versión de la instalación actual.)

Figura 4-3
Lectura de datos con un nodo de origen Archivo Statistics

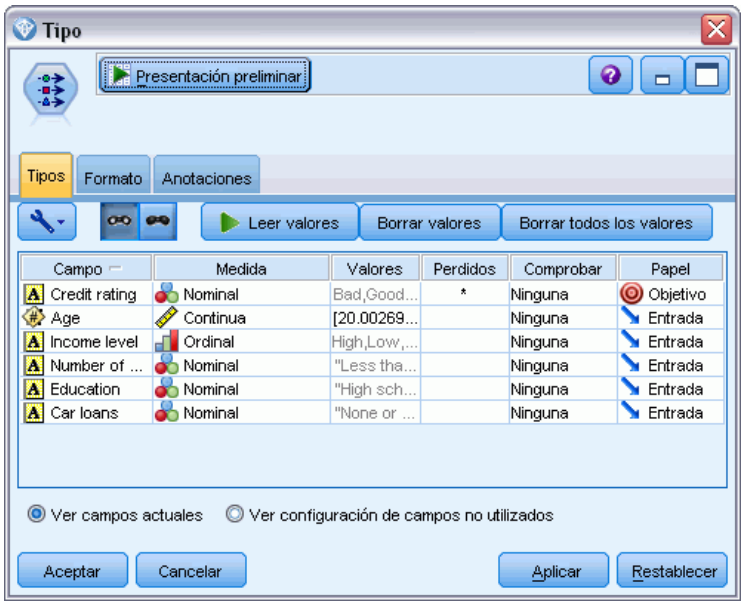


El nodo Tipo especifica el **nivel de medición** de cada campo. El nivel de medición es una categoría que indica el tipo de datos del campo. Nuestro archivo de datos de origen utiliza tres niveles de medición diferentes.

Un campo **Continuo** (como el campo *Edad*) contiene valores numéricos continuos, mientras que un campo **Nominal** (como el campo *Valoración de crédito*) tiene dos o más valores distintos, por ejemplo, *Malo*, *Bueno* o *Sin historial de crédito*. Un campo **Ordinal** (como el campo *Nivel*

de ingresos) describe datos con varios valores distintos que tienen un orden inherente; en este caso, *Bajo*, *Medio* y *Alto*.

Figura 4-4
Configuración de los campos de destino y entrada con el nodo Tipo



Para cada campo, el nodo Tipo también especifica un **papel** para indicar la función que desempeña cada campo en el modelado. El papel se define como *Objetivo* para el campo *Valoración de crédito*, que es el campo que indica si un cliente determinado ha causado mora en el pago del préstamo. Éste es el **objetivo** o campo cuyo valor queremos pronosticar.

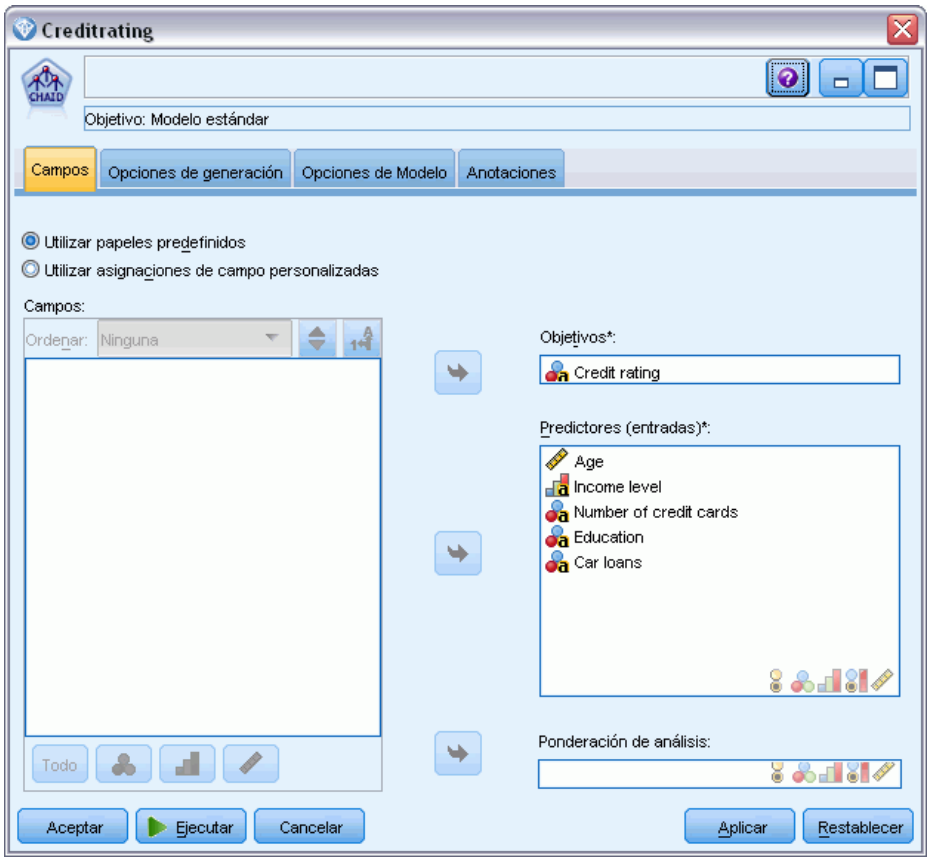
El papel se define a *Entrada* para los otros campos. Los campos de entrada se conocen a menudo como **predictores**, o campos cuyos valores se utilizan en el algoritmo de modelado para predecir el valor del campo objetivo.

El nodo de modelado CHAID genera el modelo.

En la pestaña Campos del nodo de modelado está seleccionada la opción Utilizar las funciones predefinidas, lo que significa que se utilizarán el objetivo y las entradas especificados en el nodo Tipo. En este punto podríamos cambiar las funciones de campo, pero en este ejemplo las usaremos como son.

▶ Pulse en la pestaña Crear opciones.

Figura 4-5
Nodo de modelado CHAID, pestaña Campos



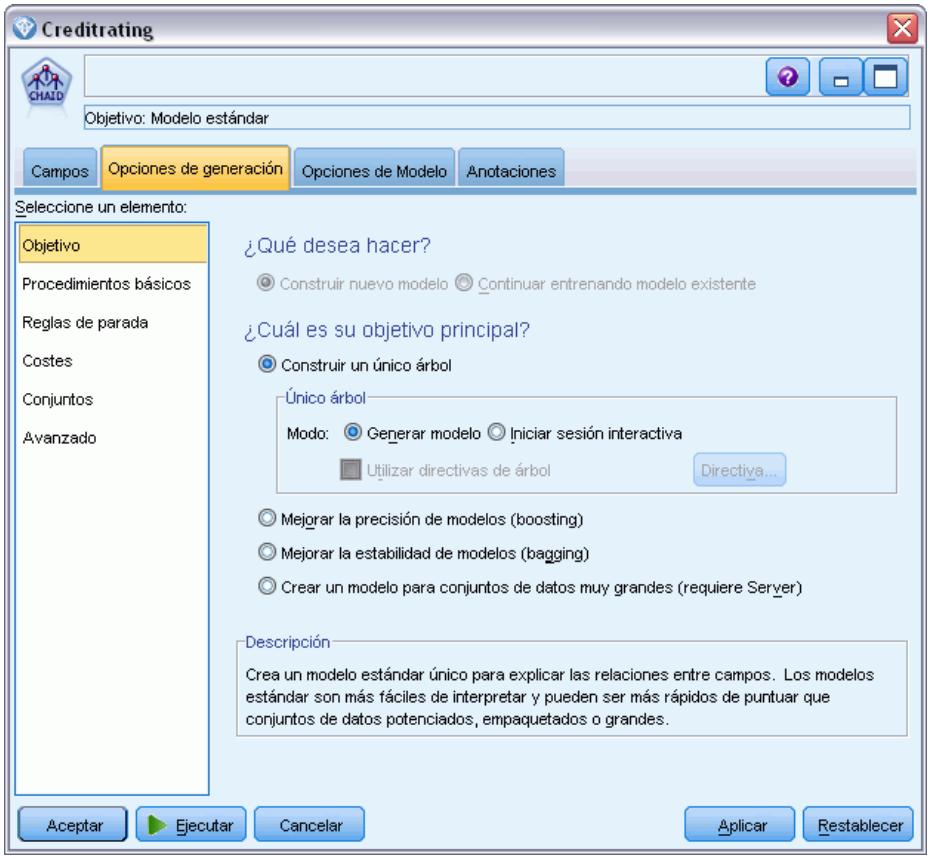
Aquí hay varias opciones en las que podemos especificar el tipo de modelo que queremos generar.

Si queremos un modelo totalmente nuevo usaremos la opción predeterminada Crear modelo nuevo.

También deseamos un único modelo de árbol de decisión estándar sin mejoras, por lo que dejaremos la opción de objetivo predeterminado Crear un árbol único.

Aunque también podemos iniciar una sesión de modelado interactivo que nos permite ajustar con precisión el modelo, este ejemplo simplemente genera un modelo utilizando la configuración de modo por defecto Generar modelo.

Figura 4-6
Nodo de modelado CHAID, pestaña Opciones de generación

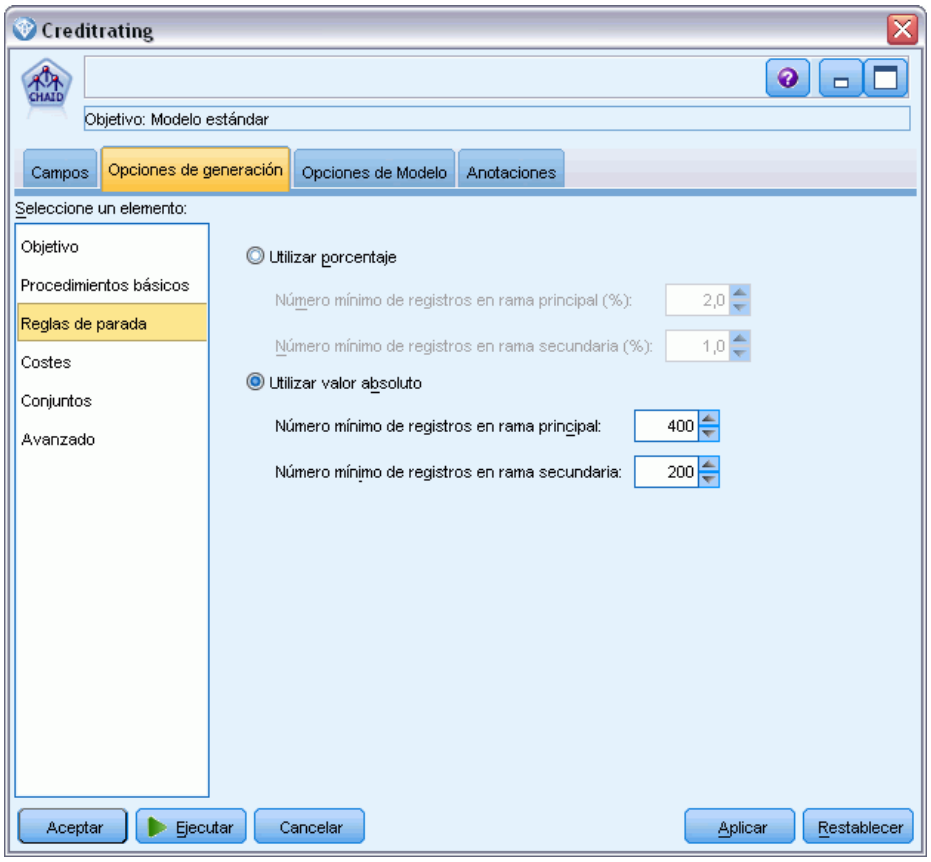


Por ejemplo, queremos que el árbol sea bastante sencillo, así que limitaremos el crecimiento del árbol elevando el número mínimo de casos para los nodos principales y filiales.

- En la pestaña Opciones de generación, seleccione Reglas de parada desde el panel de navegación de la izquierda.
- Seleccione la opción Utilizar valor absoluto.
- Establezca Número mínimo de registros en rama parental como 400.

- Establezca Número mínimo de registros por rama filial como 200.

Figura 4-7
Configuración de los criterios de parada para la generación de árboles de decisión



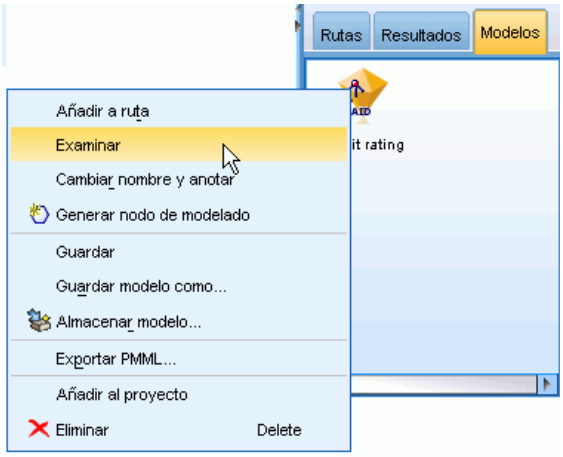
Podemos usar todas las demás opciones predeterminadas para este ejemplo, por lo que pulse en Ejecutar para crear el modelo. (También puede pulsar con el botón derecho del ratón en el nodo y seleccionar Ejecutar del menú contextual o seleccionar el nodo y Ejecutar del menú Herramientas.)

Exploración del modelo

Cuando finaliza la ejecución, se añade el nugget de modelo a la paleta Modelos en la esquina superior derecha de la ventana de aplicación, y también se coloca en el lienzo de rutas con un enlace al nodo de modelado desde el que se creó. Para ver los detalles del modelo, pulse con el

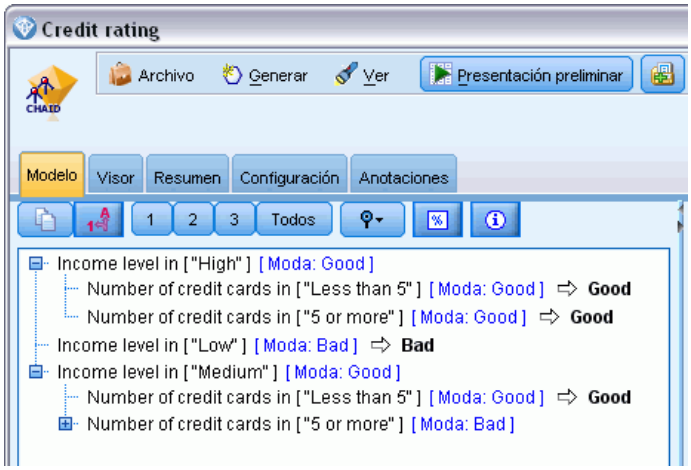
botón derecho del ratón en el nugget y seleccione Examinar (en la paleta de modelos) o Editar (en el lienzo).

Figura 4-8
Paleta de modelos



En el caso del nugget CHAID, la pestaña Modelo muestra los detalles en forma de conjunto de reglas; éste se compone esencialmente de una serie de reglas que se pueden utilizar para asignar registros individuales a los nodos filiales basándose en los valores de distintos campos de entrada.

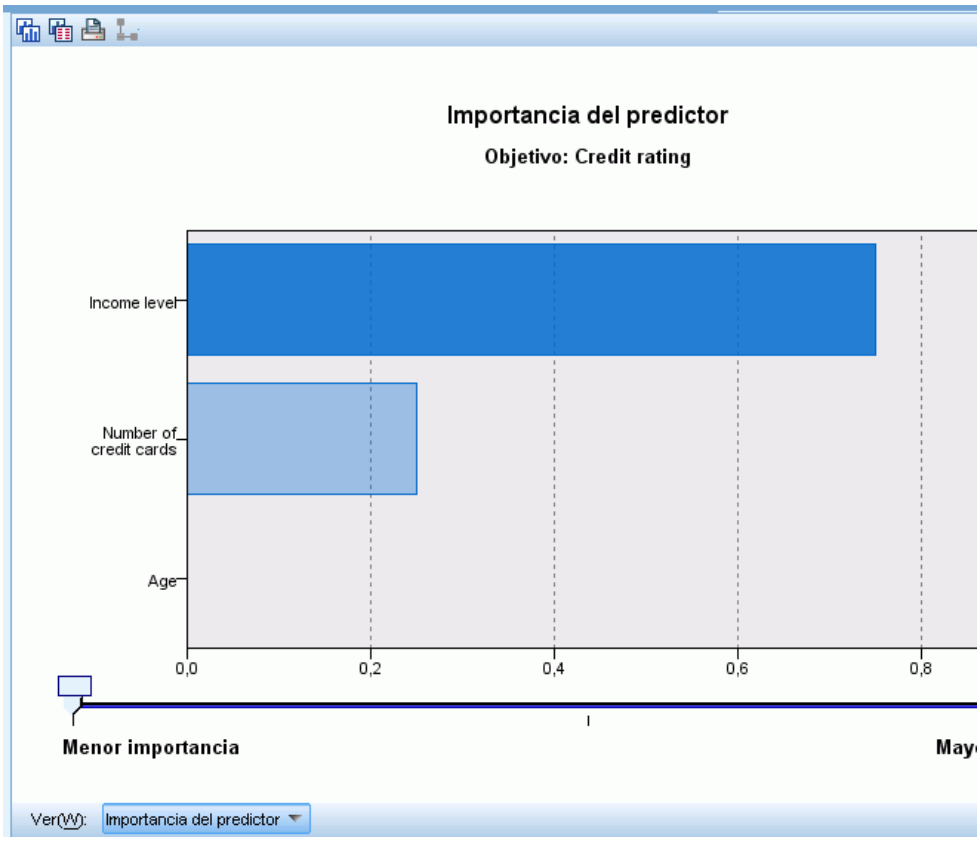
Figura 4-9
Nugget de modelo CHAID, conjunto de reglas



Por cada nodo terminal del árbol de decisión (aquellos nodos que no se dividen más) se devuelve la predicción *Bueno* o *Malo*. En cada caso, el pronóstico está determinado por el **modo** o, la respuesta más común, para registros que se incluyen en dicho nodo.

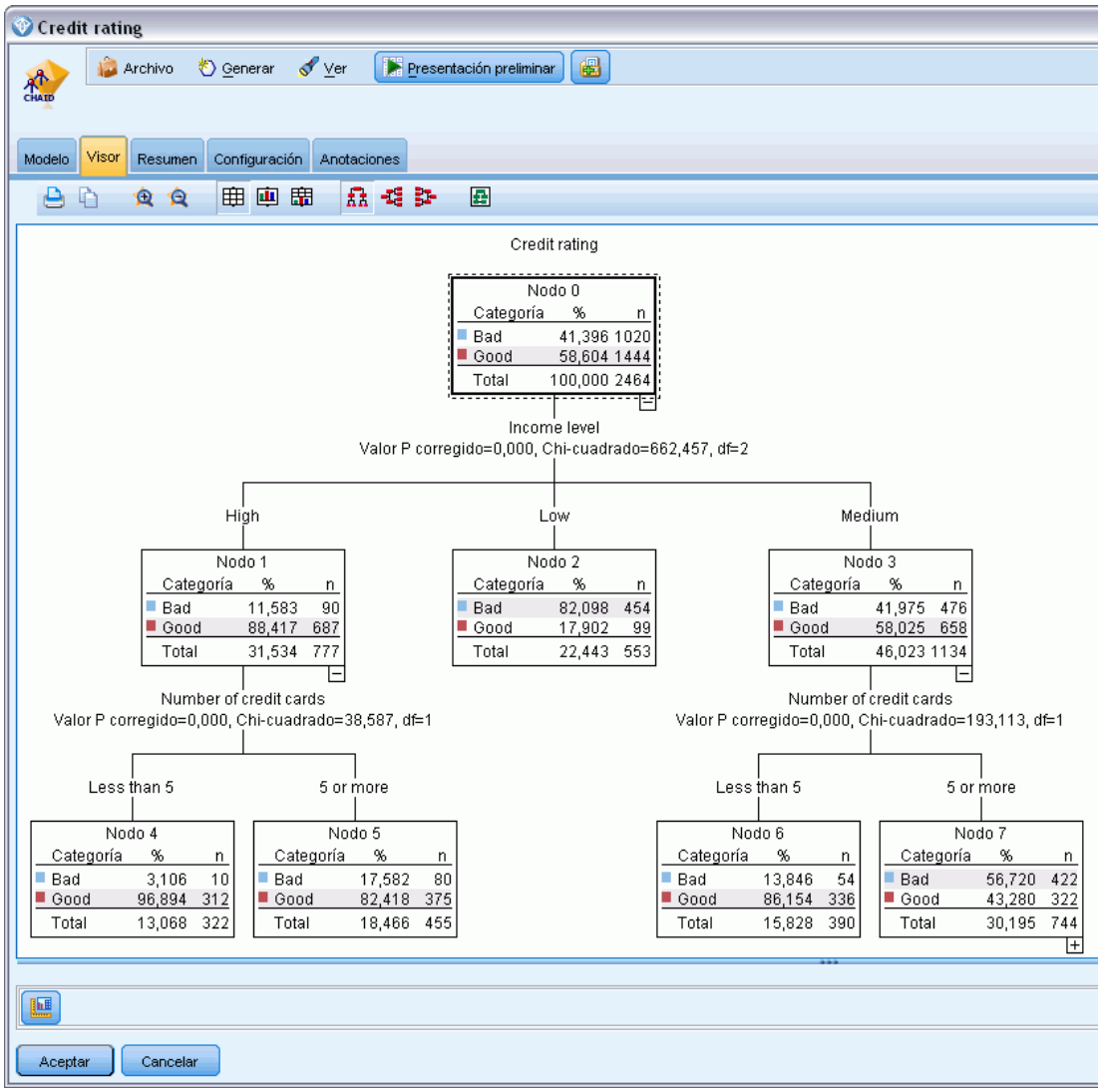
A la derecha del conjunto de reglas, la pestaña Predictor muestra el gráfico Importancia de variable, que muestra la importancia relativa de cada predictor en la estimación del modelo. A partir de aquí podemos determinar que *Nivel de ingresos* es fácilmente lo más significativo de este caso, y que el otro valor significativo es *Número de tarjetas de crédito en propiedad*.

Figura 4-10
Gráfico Importancia del predictor



La pestaña Visor del nugget de modelo muestra el mismo modelo en forma de árbol, con un nodo en cada punto de decisión. Utilice los controles Zoom de la barra de herramientas para acercarse a un nodo específico o alejarse para ver una parte más amplia del árbol.

Figura 4-11
Pestaña Visor del nugget de modelo, con la función alejar seleccionada



Al observar la parte superior del árbol, el primer nodo (Nodo 0) nos ofrece un resumen de todos los registros del conjunto de datos. Algo más del 40% de los casos del conjunto de datos se clasifica como un riesgo malo. Es una proporción bastante alta, de modo que veamos si el árbol puede darnos más pistas sobre qué factores pueden ser los responsables.

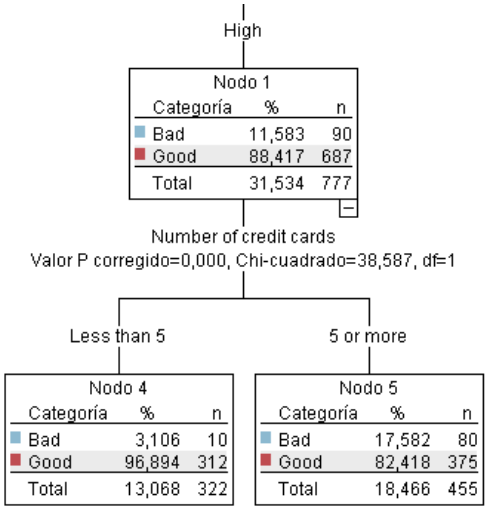
Podemos ver que la primera división es por *Nivel de ingresos*. Los registros cuyo nivel de ingresos están en la categoría *Bajo* se asignan al Nodo 2, por lo que no es sorprendente que esta categoría contenga el mayor porcentaje de morosos de préstamos. Claramente, la concesión de un préstamo a clientes de esta categoría conlleva un alto riesgo.

Sin embargo, el 16% de los clientes de esta categoría no presentó mora en los pagos, por lo que la predicción *no* siempre será correcta. Ningún modelo puede predecir de manera fiable todas las respuestas, pero un buen modelo debe permitirnos predecir la respuesta *más probable* para cada registro basándonos en los datos disponibles.

Del mismo modo, si observamos a los clientes con ingresos elevados (Nodo 1), vemos que la amplia mayoría (89%) es un riesgo bueno. Sin embargo, también más de 1 de 10 de estos clientes ha cometido mora en los pagos. ¿Podemos refinar nuestros criterios de concesión de préstamos para minimizar estos riesgos?

Tenga en cuenta cómo ha dividido el modelo a estos clientes en dos subcategorías (Nodos 4 y 5) basándose en el número de tarjetas de crédito en propiedad. En el caso de clientes con ingresos elevados, si concedemos préstamos sólo a los que tengan menos de 5 tarjetas de crédito, podemos incrementar nuestra tasa de éxito del 89% al 97%, un resultado aun más satisfactorio.

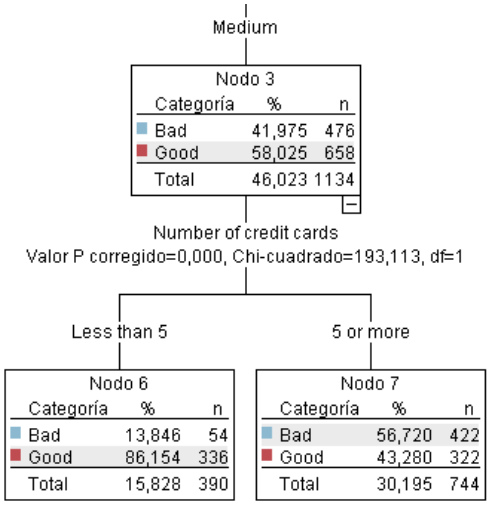
Figura 4-12
Vista de árbol de clientes con ingresos elevados



¿Qué ocurre con los clientes de la categoría de ingresos Medio (Nodo 3)? Están divididos mucho más homogéneamente entre las valoraciones Bueno y Malo.

De nuevo, las subcategorías (Nodos 6 y 7 en este caso) pueden ayudarnos. Esta vez, la concesión de préstamos sólo a los clientes con ingresos medios con menos de 5 tarjetas de crédito aumenta el porcentaje de valoraciones Bueno del 58% al 85%, lo cual es una mejora significativa.

Figura 4-13
Vista de árbol de clientes con ingresos medios



Por lo tanto, hemos aprendido que cada registro que se introduzca en este modelo se asignará a un nodo específico. Asimismo, se le asignará la predicción *Bueno* o *Malo* según la respuesta más común de ese nodo.

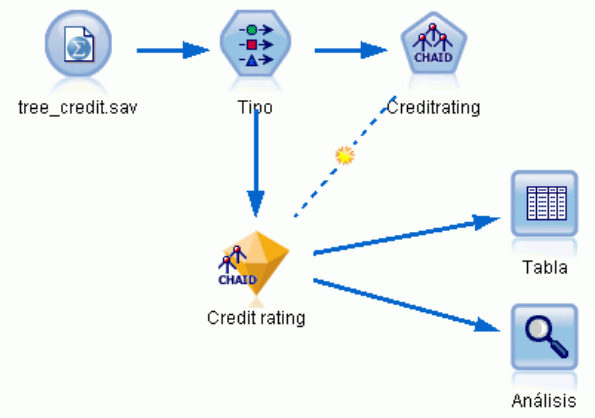
Este proceso de asignar pronósticos a registros individuales se conoce como **puntuación**. Al puntuar los mismos registros utilizados para calcular el modelo, podemos evaluar cuál es el rendimiento preciso en los datos de entrenamiento, es decir, los datos para los que conocemos el resultado. Veamos cómo hacer esto.

Evaluación del modelo

Hemos estado explorando el modelo para comprender cómo funciona la puntuación. Pero para evaluar *con qué precisión* trabaja, debemos puntuar varios registros y comparar las respuestas pronosticadas por el modelo con los resultados reales. Vamos a puntuar los mismos registros que

se utilizaron para estimar el modelo, lo que nos permite comparar las respuestas observadas y predichas.

Figura 4-14
Adición del nugget de modelo a los nodos de salida para la generación del modelo



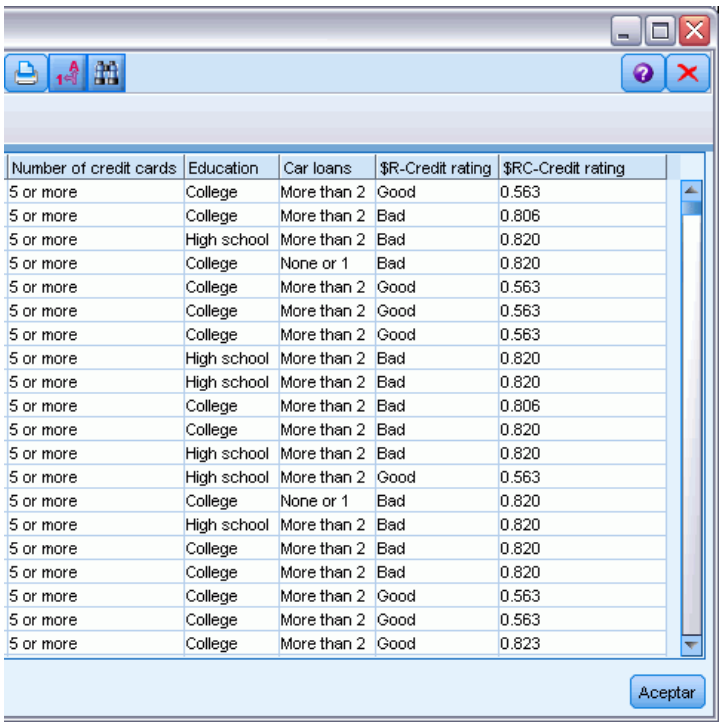
- Para ver las puntuaciones o pronósticos, adjunte el nodo Tabla al nugget de modelo, pulse dos veces en el nodo Tabla y pulse en Ejecutar.

  La tabla muestra las puntuaciones pronosticadas en un campo denominado *$R-Valoración de crédito*, creado por el modelo. Podemos comparar estos valores con el campo *Valoración de crédito* original que contiene las respuestas reales.

  Por convención, los nombres de los campos generados durante la puntuación se basan en el campo objetivo, pero con un prefijo estándar como *$R-* para pronósticos o *$RC-* para valores de confianza. Los distintos tipos de modelo utilizan diferentes conjuntos de prefijos. Un **valor de**

confianza es la estimación del propio modelo, en una escala de 0,0 a 1,0, sobre el grado de precisión de cada valor pronosticado.

Figura 4-15
Tabla que muestra las puntuaciones generadas y los valores de confianza



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Bad	0.806
5 or more	High school	More than 2	Bad	0.820
5 or more	College	None or 1	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.806
5 or more	College	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Good	0.563
5 or more	College	None or 1	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.823

Aceptar

Como se esperaba, el valor pronosticado coincide con las respuestas reales de muchos registros, pero no todos. El motivo es que cada nodo terminal CHAID tiene una mezcla de respuestas. El pronóstico coincide con la *más común*, pero es incorrecto para el resto de dicho nodo. (Recuerde la minoría del 16% de clientes con ingresos bajos que no cometió mora en los pagos.)

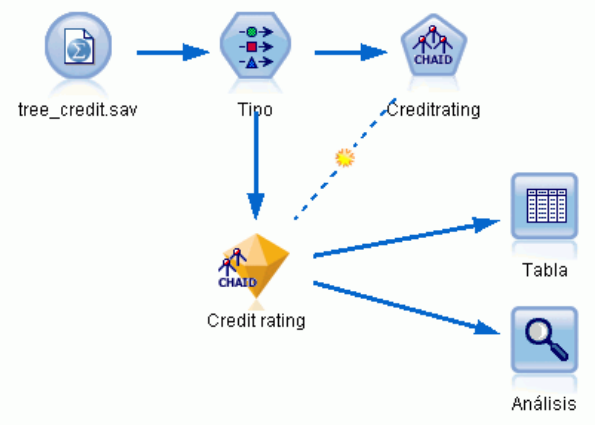
Para evitarlo, podemos seguir dividiendo el árbol en ramas cada vez más pequeñas, hasta que cada nodo sea 100 % puro: todas las respuestas son *Bueno* o *Malo* sin respuestas mezcladas. Pero dicho modelo sería extremadamente complicado y probablemente no se generalizaría bien en otros conjuntos de datos.

Para descubrir exactamente cuántas predicciones son correctas, podríamos observar la tabla y anotar el número de registros en los que el valor del campo pronosticado *$R-Valoración de crédito* coincida con el valor de *Valoración de crédito*. Afortunadamente, hay un modo más sencillo: podemos utilizar un nodo Análisis, que lo hace automáticamente.

- Conecte el nugget de modelo al nodo Análisis.

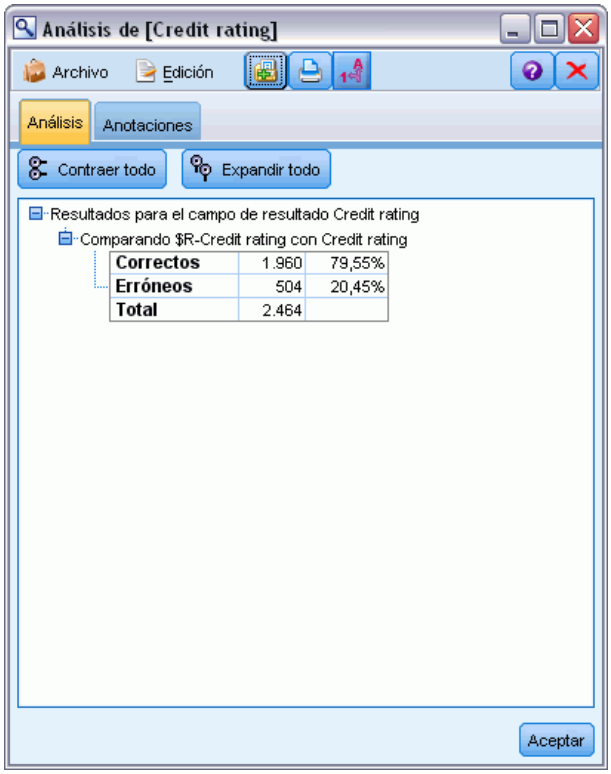
▶ Pulse dos veces en el nodo Análisis y pulse en Ejecutar.

Figura 4-16
Conexión del nodo Análisis



El análisis muestra que para 1899 de 2464 registros (más del 77%), el valor pronosticado por el modelo coincidía con la respuesta real.

Figura 4-17
Resultados de análisis que comparan respuestas observadas y pronosticadas



Este resultado está limitado por el hecho de que los registros que se están puntuando son los mismos utilizados para calcular el modelo. En una situación real, podría utilizar un nodo Partición para dividir los datos en muestras separadas para el entrenamiento y la evaluación.

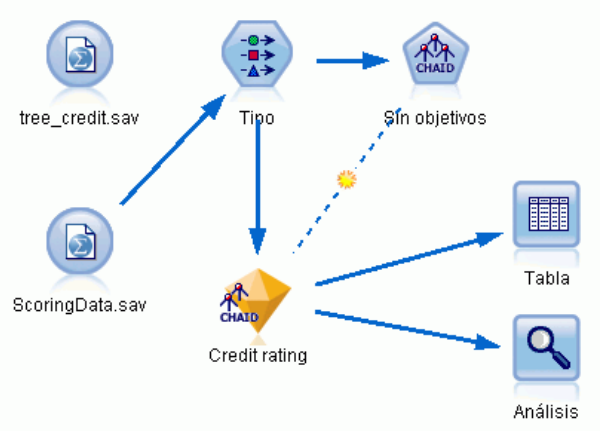
Si utiliza una partición de muestra para generar el modelo y otra muestra para comprobarlo, podrá obtener una indicación mucho mejor de lo bien que se generalizará en otros conjuntos de datos.

El nodo Análisis nos permite comprobar el modelo frente a registros para los que ya conocemos el resultado real. La etapa siguiente muestra cómo podemos utilizar el modelo para puntuar registros cuyos resultados no conocemos. Por ejemplo, esto podría incluir a personas que no son clientes actuales del banco, pero son posibles objetivos de correos promocionales.

Puntuación de registros

Antes hemos puntuado los mismos registros utilizados para calcular el modelo con el fin de evaluar el grado de precisión del modelo. Ahora vamos a ver cómo puntuar un conjunto de registros diferentes de los utilizados para crear el modelo. Ésta es la meta del modelado con un campo objetivo: Estudie los registros de los que conoce los resultados para identificar patrones que le permitirán pronosticar resultados que todavía no conoce.

Figura 4-18
Adición de nuevos datos para su puntuación



Podría actualizar el nodo de origen Archivo Statistics para dirigirse a un archivo de datos diferente o podría añadir un nuevo nodo de origen que lea los datos que desea puntuar. En cualquier caso, el nuevo conjunto de datos debe contener los mismos campos de entrada utilizados por el modelo (*Edad*, *Nivel de ingresos*, *Educación*, etc.) pero no el campo objetivo *Valoración de crédito*.

También podría añadir el nugget de modelo a cualquier ruta que incluya los campos de entrada esperados. El tipo de origen no importa, tanto si se ha leído de un archivo o de una base de datos, siempre que los nombres y tipos de campo coincidan con los utilizados por el modelo.

También podría guardar el nugget de modelo como un archivo independiente, exportar el modelo en formato PMML para su uso con otras aplicaciones que admitan este formato, o almacenar el modelo en un repositorio IBM® SPSS® Collaboration and Deployment Services, que ofrece distribución, puntuación y gestión de modelos en toda la empresa.

Independientemente de la infraestructura utilizada, el propio modelo funciona del mismo modo.

Resumen

Este ejemplo demuestra los pasos básicos para crear, evaluar y puntuar un modelo.

- El nodo de modelado calcula el modelo estudiando registros para los que se conoce el resultado y crea un nugget de modelo. Esto se denomina a veces entrenamiento del modelo.
- El nugget de modelo puede añadirse a cualquier ruta con los campos esperados para puntuar registros. Al puntuar los registros de los que ya conoce el resultado (como los clientes existentes), puede evaluar el grado de rendimiento.
- Una vez quede satisfecho con el rendimiento adecuado del modelo, podrá puntuar nuevos datos (como clientes potenciales) para pronosticar cómo responderán.
- Debe hacerse referencia a los datos utilizados para entrenar o calcular el modelo como los datos analíticos o históricos; también se puede hacer referencia a los datos de puntuación como los datos operativos.

Capítulo
5

Modelado automatizado para un objetivo de marca

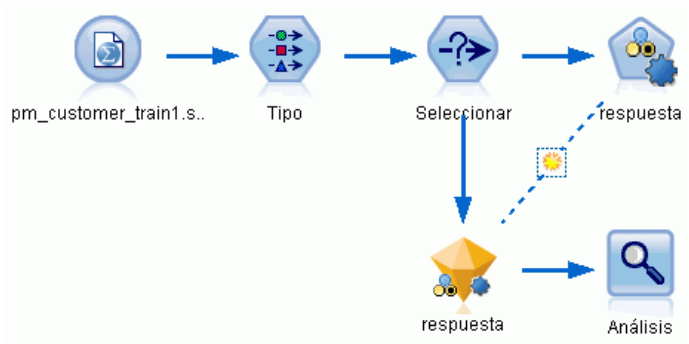
Modelado de respuesta de clientes (clasificador automático)

El nodo Clasificador automático le permite crear y comparar modelos automáticamente un número de modelos para cada marca (como si es probable que un determinado cliente no pueda afrontar el pago de un préstamo o responder a una oferta concreta) u objetivos nominales (conjunto). En este ejemplo buscaremos un resultado de marca (yes o no). Con una ruta relativamente simple, el nodo genera y ordena un conjunto de modelos de candidatos, selecciona los que tienen un mejor rendimiento y los combina en un único modelo agregado (de conjunto). Este método combina la facilidad de la automatización con los beneficios de combinar múltiples modelos, que suelen producir pronósticos más precisos que cualquier otro modelo.

Este ejemplo se basa en una empresa ficticia que desea obtener resultados más rentables adaptando la oferta adecuada a cada cliente.

Este método refuerza las ventajas de la automatización. Para ver un ejemplo similar que utilice un objetivo continuo (rango numérico), consulte el capítulo 6 el p. 56.

Figura 5-1
Ruta de ejemplo de Clasificador automático



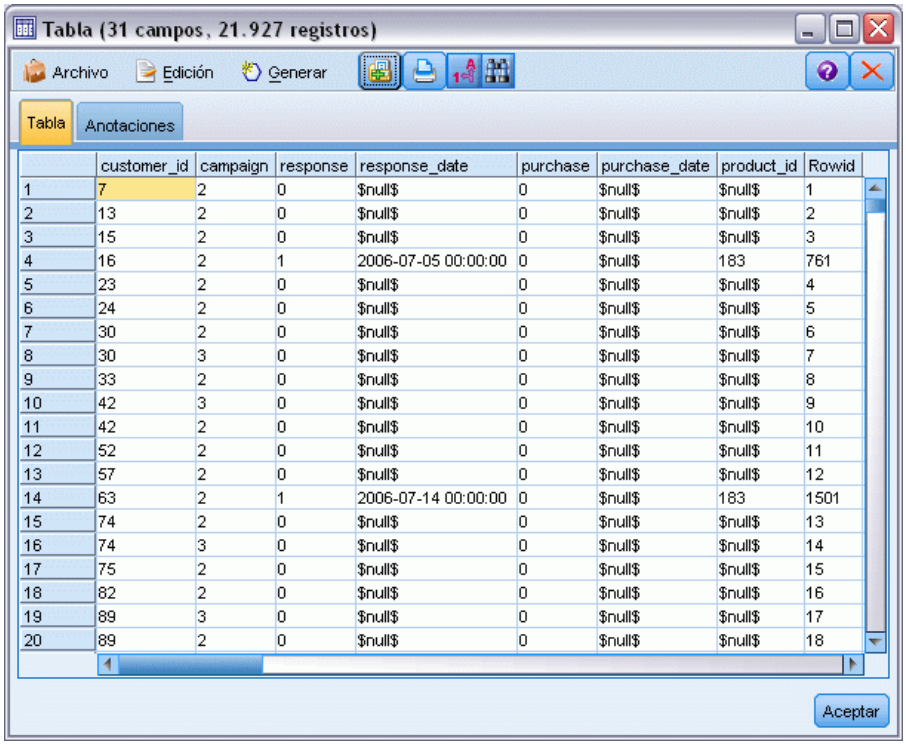
Este ejemplo utiliza la ruta *pm_binaryclassifier.str*, en la carpeta Demo en *streams*. El archivo de datos utilizado es *pm_customer_train1.sav*. Si desea obtener más información, consulte el tema Carpeta Demos en el capítulo 1 el p. 5.

Datos históricos

El archivo *pm_customer_train1.sav* contiene datos históricos en los que se registran las ofertas realizadas a determinados clientes en campañas anteriores, según indica el valor del campo *campaña*. El mayor número de registros corresponden a la campaña *Cuenta principal*.

Los valores del campo *campaña* aparecen codificados como enteros en los datos (por ejemplo, *2 = Cuenta principal*). Posteriormente definirá las etiquetas de estos valores que puede usar para obtener un resultado más significativo.

Figura 5-2
Datos sobre promociones anteriores



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

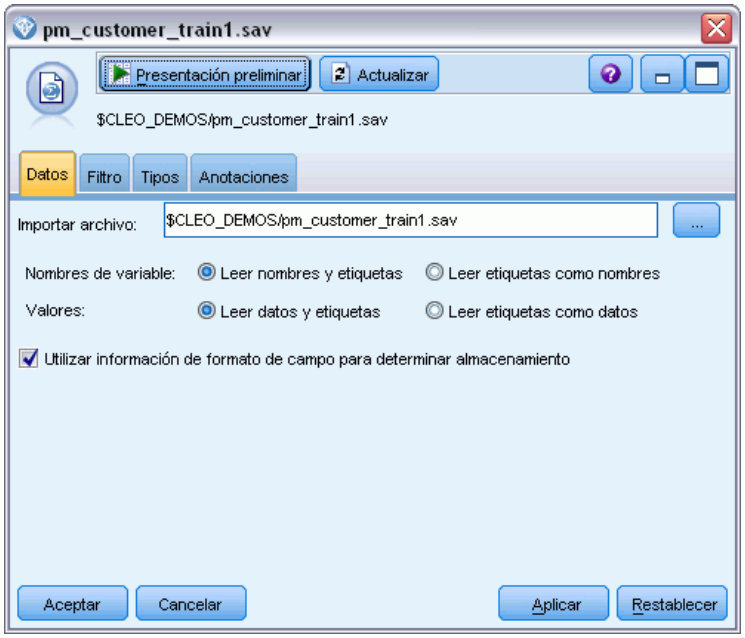
El archivo también incluye un campo *respuesta* que indica si la oferta se ha aceptado (0 = *no*, y 1 = *sí*). Éste es el **campo objetivo** o valor que quiere pronosticar. También se incluyen campos con información demográfica y financiera sobre cada cliente. Se pueden utilizar para genera o "entrenar" un modelo que pronostica índices de respuesta para individuos o grupos basados en características como ingresos, edad o número de transacciones al mes.

Generación de la ruta

- Añada un nodo de origen de Statistics que apunte a *pm_customer_train1.sav*, ubicado en la carpeta *Demos* de la instalación de IBM® SPSS® Modeler. (Puede especificar `$CLEO_DEMOS/` en la ruta

del archivo como acceso directo a referencia de esta carpeta. Tenga en cuenta que se debe usar una barra diagonal en lugar de una barra invertida en la ruta, tal y como se muestra a continuación.)

Figura 5-3
Lectura de datos mezclados



► Añada un nodo Tipo y seleccione *respuesta* como campo objetivo (Papel = Objetivo). Establezca la medición de este campo como Marca.

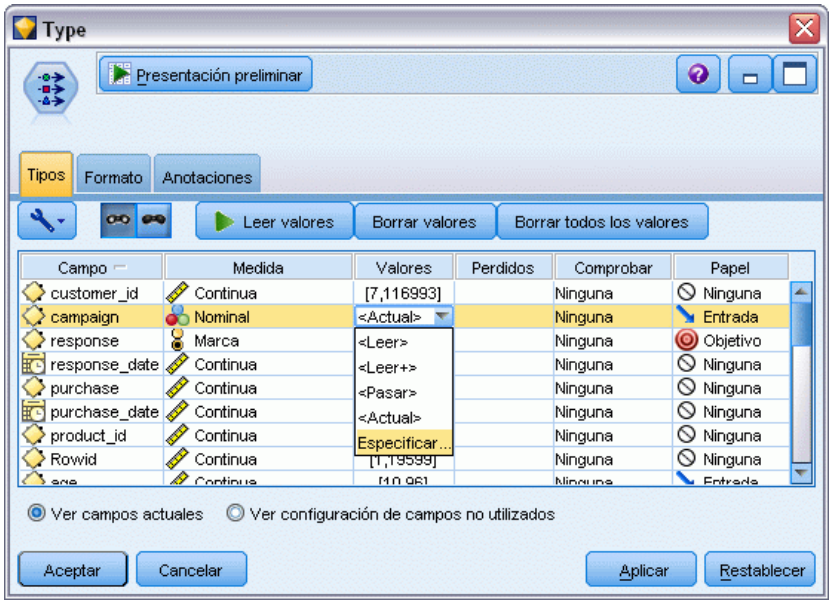
Figura 5-4
Definición del nivel de medición y el papel

- Establezca el papel en Ninguno para los siguientes campos: *id_cliente*, *campaña*, *fecha_respuesta*, *compra*, *fecha_compra*, *id_producto*, *Idfila* y *X_aleatorio*. Estos campos se ignorarán cuando se crea un modelo.

- Pulse en el botón Leer valores del nodo Tipo para asegurarse de que se crea una instancia de los valores.

  Como vimos anteriormente, nuestros datos de origen incluyen información acerca de cuatro diferentes campañas, cada una dirigida a un tipo diferente de cuenta de cliente. Estas campañas están codificadas como enteros en los datos, por lo que para facilitar recordar a qué tipo de cuenta representa cada entero, definamos las etiquetas de cada uno.

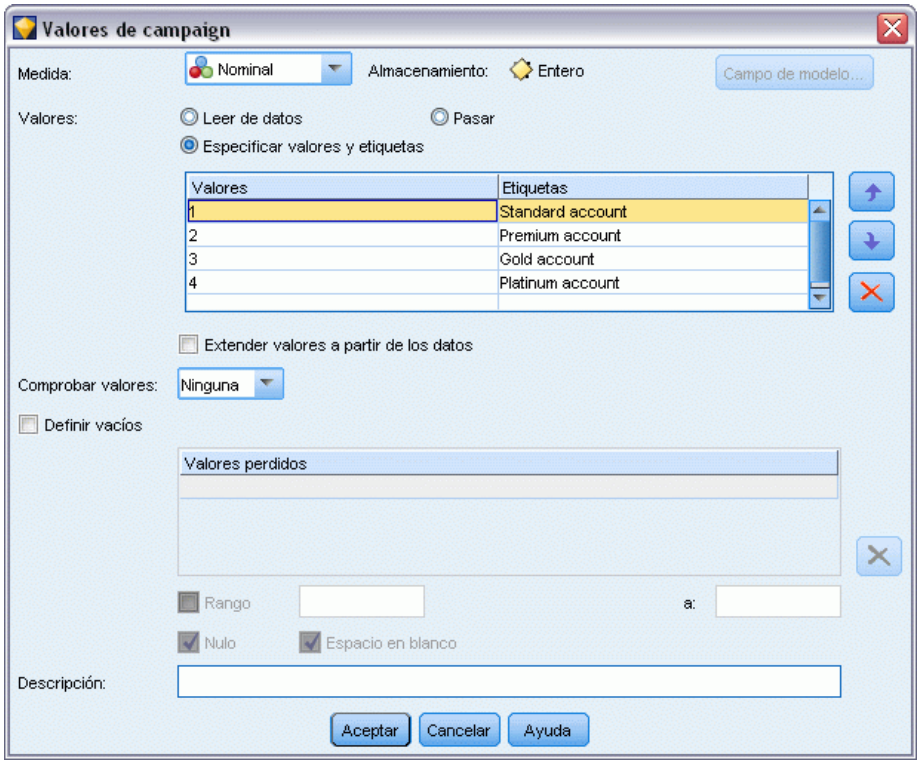
Figura 5-5
Selección de la especificación de valores de un campo



- En la fila del campo campaña, pulse en la columna Valores.

- Seleccione Especificar de la lista desplegable.

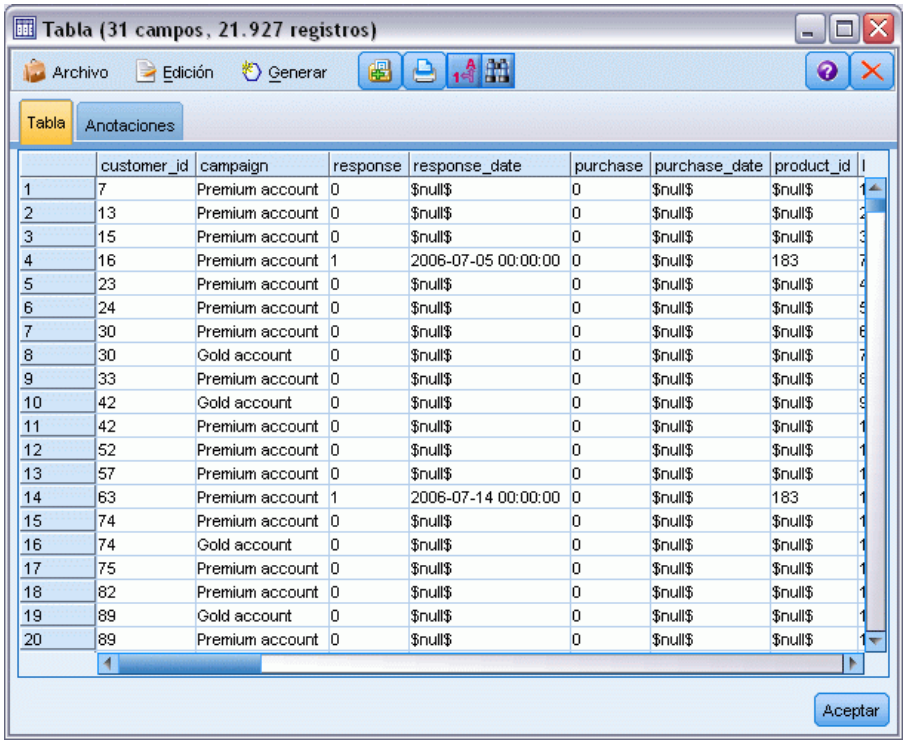
Figura 5-6
Definición de etiquetas de los valores de campos



- En la columna Etiquetas, introduzca las etiquetas como se muestra para cada uno de los cuatro valores del campo campaña.
- Pulse en Aceptar.

Ahora podrá mostrar las etiquetas en las ventanas de salida en lugar de los enteros.

Figura 5-7
Visualización de las etiquetas de valor del campo

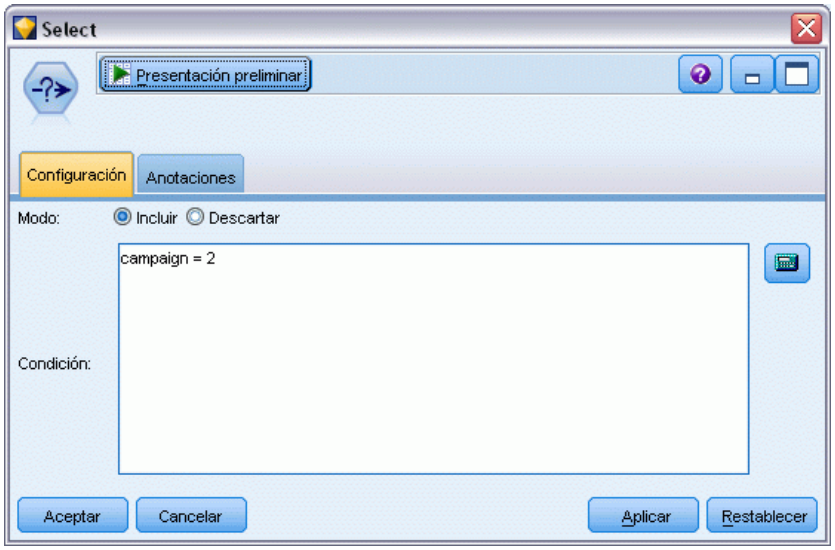


	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	l
1	7	Premium account	0	$null$	0	$null$	$null$	1
2	13	Premium account	0	$null$	0	$null$	$null$	2
3	15	Premium account	0	$null$	0	$null$	$null$	3
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183	7
5	23	Premium account	0	$null$	0	$null$	$null$	4
6	24	Premium account	0	$null$	0	$null$	$null$	5
7	30	Premium account	0	$null$	0	$null$	$null$	6
8	30	Gold account	0	$null$	0	$null$	$null$	7
9	33	Premium account	0	$null$	0	$null$	$null$	8
10	42	Gold account	0	$null$	0	$null$	$null$	9
11	42	Premium account	0	$null$	0	$null$	$null$	1
12	52	Premium account	0	$null$	0	$null$	$null$	1
13	57	Premium account	0	$null$	0	$null$	$null$	1
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183	1
15	74	Premium account	0	$null$	0	$null$	$null$	1
16	74	Gold account	0	$null$	0	$null$	$null$	1
17	75	Premium account	0	$null$	0	$null$	$null$	1
18	82	Premium account	0	$null$	0	$null$	$null$	1
19	89	Gold account	0	$null$	0	$null$	$null$	1
20	89	Premium account	0	$null$	0	$null$	$null$	1

- Conecte un nodo Tabla al nodo Tipo.
- Abra el nodo Tabla y pulse en Ejecutar.
- En la ventana de salida, pulse en el botón Mostrar etiquetas de valor y de campo para mostrar las etiquetas.
- Pulse en Aceptar para cerrar la ventana.

Aunque los datos incluyen información acerca de cuatro campañas diferentes, el análisis lo realizaremos campaña a campaña. Como el mayor número de registros corresponden a la campaña Cuenta principal (codificada como *campaña=2* en los datos), puede utilizar un nodo Seleccionar para incluir únicamente dichos registros en la ruta.

Figura 5-8
Selección de los registros correspondientes a una única campaña

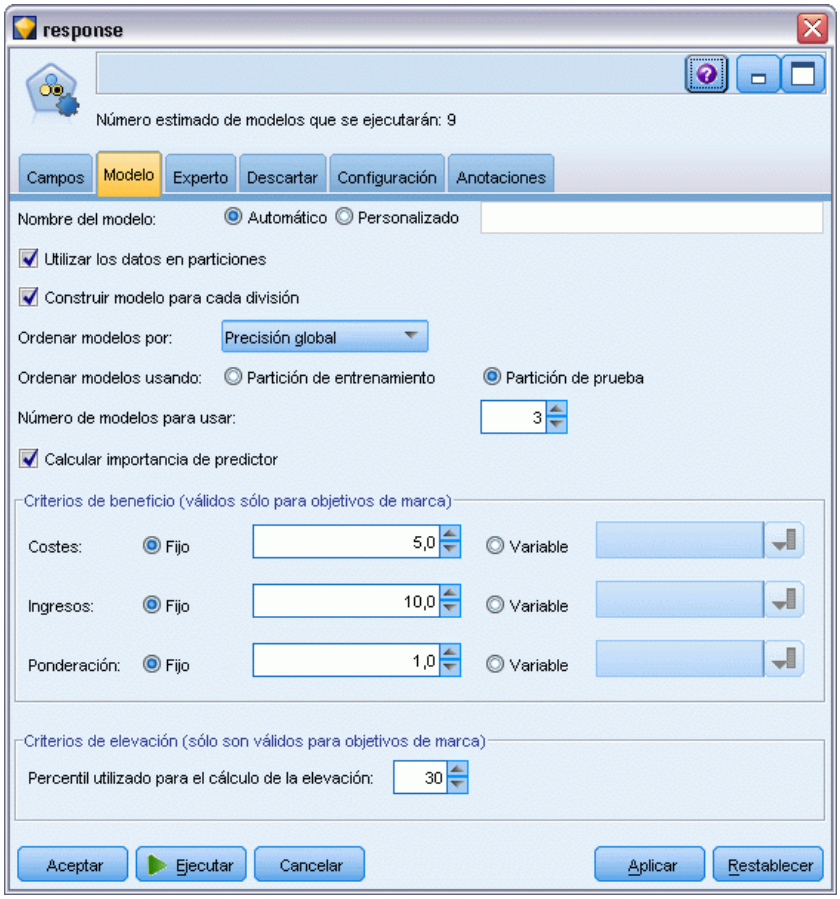


Generación y comparación de modelos

- Conecte un nodo Clasificador automático y seleccione Precisión global como la métrica para ordenar modelos.

- Establezca Número de modelos que se utilizarán como 3. Esto significa que se generarán los tres mejores modelos cuando ejecute el nodo.

Figura 5-9
Pestaña Modelo del nodo Clasificador automático

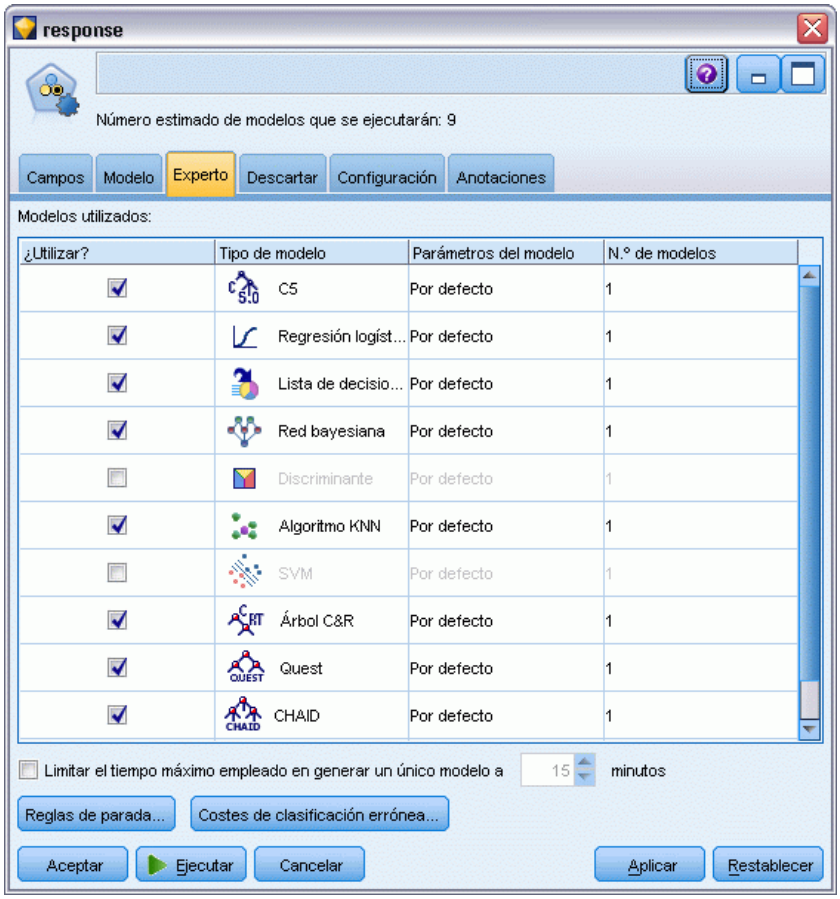


En la pestaña Experto, puede seleccionar entre 11 algoritmos de modelo diferentes.

- Cancele la selección de los tipos de modelo Discriminante y SVM. (Estos modelos tardan más en entrenar los datos, por lo que si cancela su selección, el ejemplo se ejecutará más rápido. Si no le importa esperar, déjelos seleccionados.)

Como ha establecido Número de modelos que se utilizarán como 3 en la pestaña Modelo, el nodo calculará la precisión de los nueve algoritmos restantes y generará un nugget de modelo único con los tres más precisos.

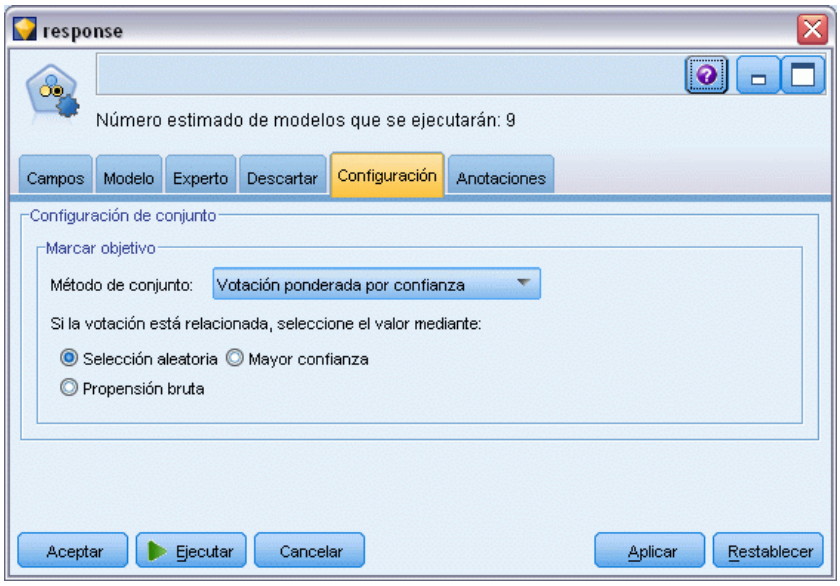
Figura 5-10
Pestaña Experto del nodo Clasificador automático



- En la pestaña Configuración, para el método de conjunto, seleccione Votación ponderada de confianza. Determina cómo se produce una única puntuación agregada para cada registro.

Con una simple votación, si dos o tres modelos pronostican *sí*, *sí* gana por 2 votos a 1. En caso de votación ponderada de confianza, los votos se ponderan en función del valore de confianza de cada predicción. Además, si un modelo pronostica *no* con mayor confianza que los dos pronósticos *sí* combinados, ganará *no*.

Figura 5-11
Nodo Clasificador automático: Pestaña Configuración



▶ Pulse en Ejecutar.

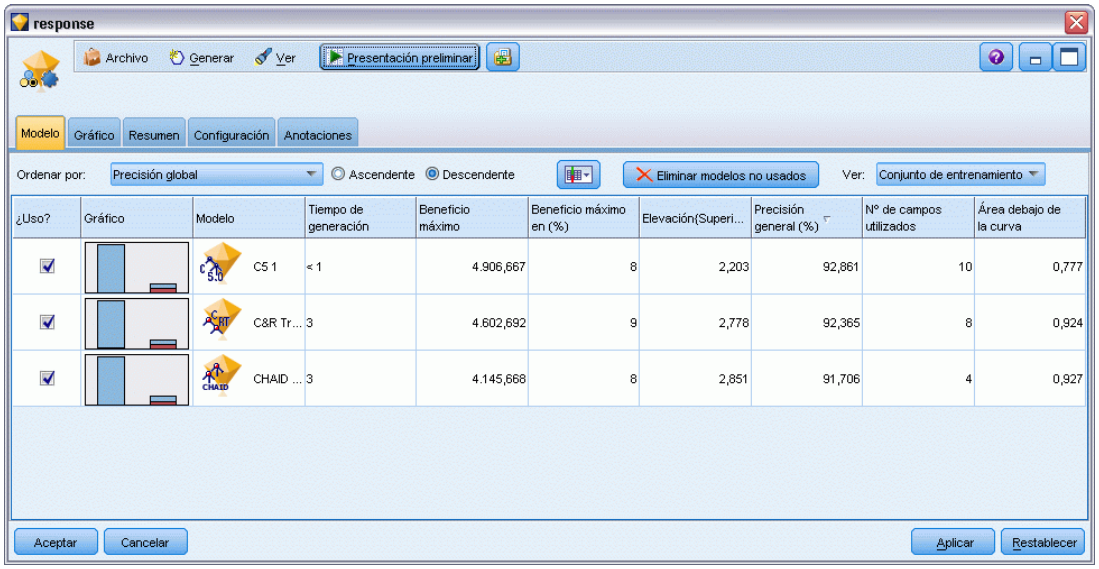
Después de algunos minutos, se crea el nugget del modelo generado y se coloca en el lienzo y, en la paleta Modelos en la esquina superior derecha de la ventana. Puede examinar el nugget de modelo o guardarlo para distribuirlo en diferentes formas.

Abra el nugget de modelo; enumera los detalles de cada uno de los modelos creados durante la ejecución. (En una situación real, en la que se pueden crear cientos de modelos en un conjunto de datos mayor, este proceso puede tardar horas.) Consulte Figura 5-1 el p. 44.

Si desea seguir explorando cualquiera de los modelos individuales, puede pulsar dos veces en el icono del nugget de modelo en la columna Modelo para profundizar y examinar los resultados del modelo individual; desde ahí puede generar nodos de modelado, nugget de modelo o gráficos

de evaluación. En la columna Gráfico, puede pulsar dos veces en una miniatura para generar un gráfico a tamaño completo.

Figura 5-12
Resultados de Clasificador automático



Por defecto, los modelos se clasifican en función de su precisión global, porque es la medida que ha seleccionado en la pestaña Modelo del nodo Clasificador automático. El modelo C51 obtiene una mejor posición con esta medida, pero los modelos C&RT y CHAID son casi igual de precisos.

Puede ordenar una columna diferente pulsando en el encabezado de la columna o seleccionar la medida que desee de la lista desplegable Ordenar por de la barra de herramientas.

Según estos resultados, puede decidir utilizar los tres de estos modelos más precisos. Al combinar predicciones de varios modelos, pueden evitarse las limitaciones en modelos individuales que dan como resultado una precisión global superior.

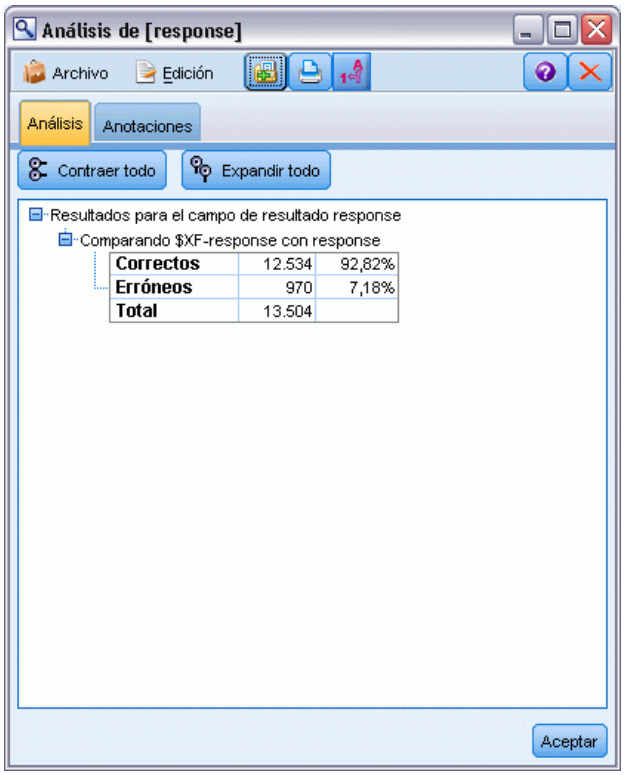
En la columna Uso?, seleccione los modelos C51, C&RT y CHAID.

Añada un nodo Análisis (paleta Resultado) después del nugget de modelo. Pulse con el botón derecho en el nodo Análisis y seleccione Ejecutar para ejecutar la ruta.

La puntuación agregada generada por el modelo de conjunto se muestra en un campo denominado *$XF-response*. Si se comparan con los datos de entrenamiento, el valor pronosticado coincide con la respuesta real (registrada en el campo original *respuesta*) con una precisión global del 92.82%.

Aunque no sea tan precisa como el mejor de los tres modelos individuales en este caso (92,86% de C51), la diferencia es demasiado pequeña para ser significativa. En términos generales, es más probable que un modelo de conjunto se ejecute bien cuando se aplique a conjuntos de datos que no sean los datos de formación.

Figura 5-13
Análisis de los tres modelos de conjunto



Resumen

En resumen, ha utilizado el nodo Clasificador automático para comparar diferentes modelos, ha utilizado los tres modelos más precisos y los ha añadido a la ruta dentro de un nugget de modelo Clasificador automático de conjunto.

- En función de su precisión global, los modelos Árbol C51, C&R y CHAID ejecutan mejor los datos de formación.
- Este modelo de conjunto tiene un rendimiento casi tan bueno como el mejor de los modelos individuales y tendrá un rendimiento aun mejor cuando se aplique a otros conjuntos de datos. Si su objetivo es automatizar el proceso lo máximo posible, este método le permite obtener un modelo robusto en la mayoría de circunstancias, sin tener que entrar demasiado en las características específicas de un modelo.

Capítulo 6

Modelado automatizado para objetivo continuo

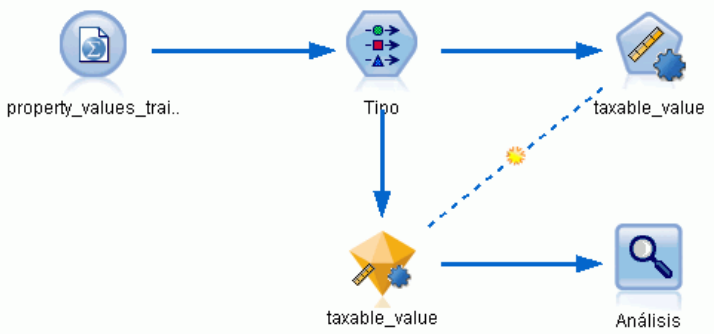
Valores de propiedad (Autonumérico)

El nodo Autonumérico permite crear y comparar de forma automática diferentes modelos de resultados continuo (rango numérico), como pronosticar el valor gravable de una propiedad. Con un nodo único, puede estimar y comparar un conjunto de modelos de candidato y generar un subconjunto de modelos para su análisis posterior. El nodo funciona de la misma manera que el nodo Clasificador automático, pero para continuos en lugar de objetivos marca o nominales.

El nodo combina las mejores opciones de los modelos de candidatos en un único nugget de modelo (agregado). Este método combina la facilidad de la automatización con los beneficios de combinar múltiples modelos, que suelen producir pronósticos más precisos que cualquier otro modelo.

Este ejemplo se centra en una oficina municipal responsable del control y cobro de impuestos sobre bienes inmuebles. Para realizar esta función con mayor precisión, generarán un modelo que pronostica valores en función del tipo de edificio, barrio tamaño y otros factores conocidos.

Figura 6-1
Ruta de ejemplo de Autonumérico



Este ejemplo utiliza la ruta *property_values_numericpredictor.str*, en la carpeta Demo en *streams*. El archivo de datos utilizado es *property_values_train.sav.* Si desea obtener más información, consulte el tema Carpeta Demos en el capítulo 1 el p. 5.

Datos de entrenamiento

El archivo de datos incluye un campo *valor_gravable*, que es el **campo objetivo**, o valor que desea pronosticar. El resto de campos contienen información como el barrio, tipo de edificio y volumen interior y se pueden utilizar como predictores.

Nombre de campo	Label
id_propiedad	ID de la propiedad
barrio	Zona de la ciudad
tipo_edificio	Tipo de edificio
año_construcción	Año de construcción
volumen_interior	Volumen del interior
volumen_otros	Volumen del garaje y de instalaciones extra
tamaño_parcela	Tamaño de la parcela
valor_gravable	Valor gravable

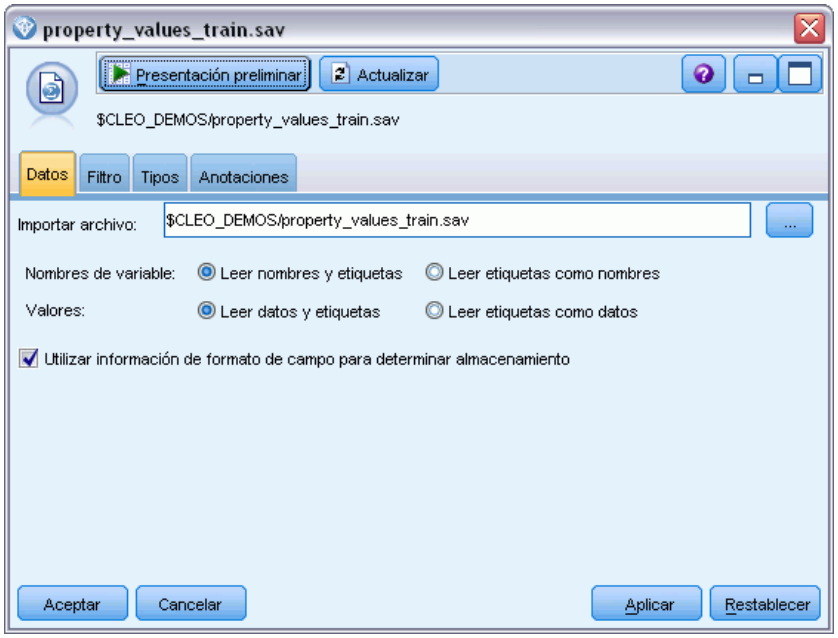
También se incluye un archivo de datos de puntuación en la carpeta Demos, denominado *property_values_score.sav*. Contiene los mismos campos, pero sin el campo *valor_gravable*. Después de entrenar modelos con un conjunto de datos donde se conoce el valor gravable, puede puntuar los registros en los que este valor aún no se conoce.

Generación de la ruta

- Añada un nodo de origen de Statistics que apunte a *property_values_train.sav*, ubicado en la carpeta *Demos* de la instalación de IBM® SPSS® Modeler. (Puede especificar `$CLEO_DEMOS/` en la ruta del archivo como acceso directo a referencia de esta carpeta. Tenga en cuenta que se

debe usar una barra diagonal en lugar de una barra invertida en la ruta, tal y como se muestra a continuación.)

Figura 6-2
Lectura de datos mezclados

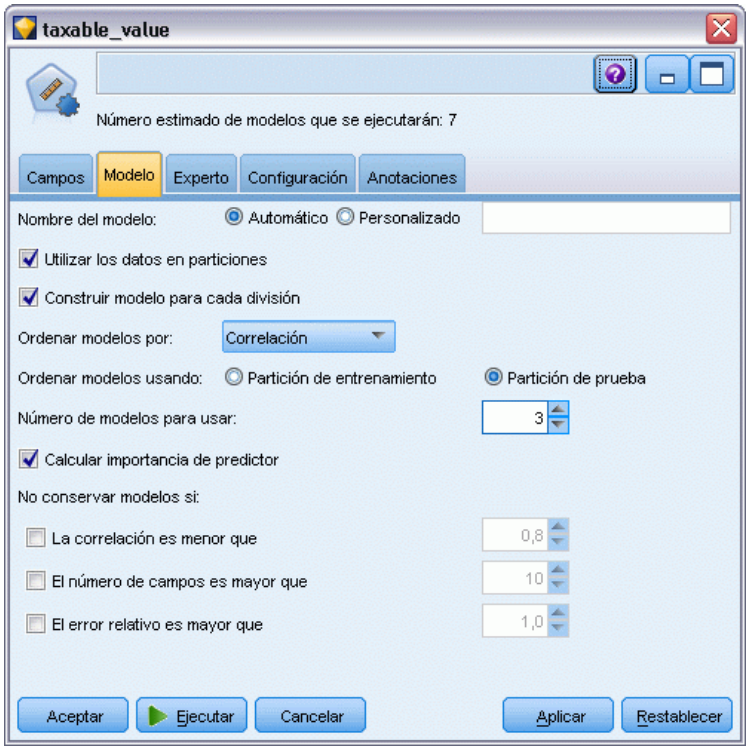


- Añada un nodo Tipo y seleccione *valor_gravable* como campo objetivo (Papel = Objetivo). Debe definirse el papel Entrada para el resto de campos, indicando que se utilizarán como predictores.

Figura 6-3
Configuración del campo objetivo

- Adjunte un nodo Autonumérico y seleccione Correlación como la métrica para ordenar modelos.
- Establezca Número de modelos que se utilizarán como 3. Esto significa que se generarán los tres mejores modelos cuando ejecute el nodo.

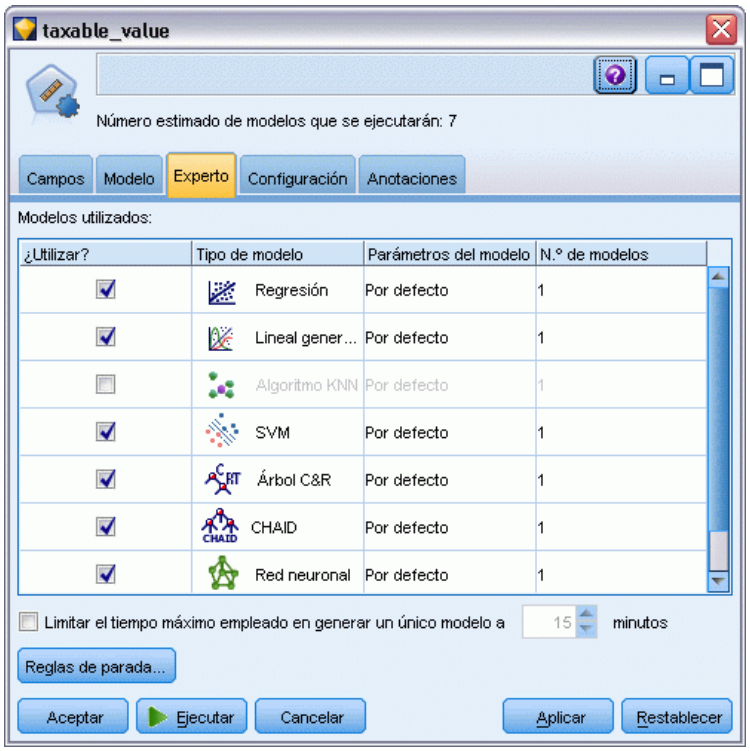
Figura 6-4
Pestaña Modelo del nodo Autonumérico



- En la pestaña Experto, deje la configuración predefinida; el nodo estimará un modelo único para cada algoritmo, para un total de siete modelos. (También puede modificar esta configuración para comparar múltiples variantes para cada tipo de modelo.)

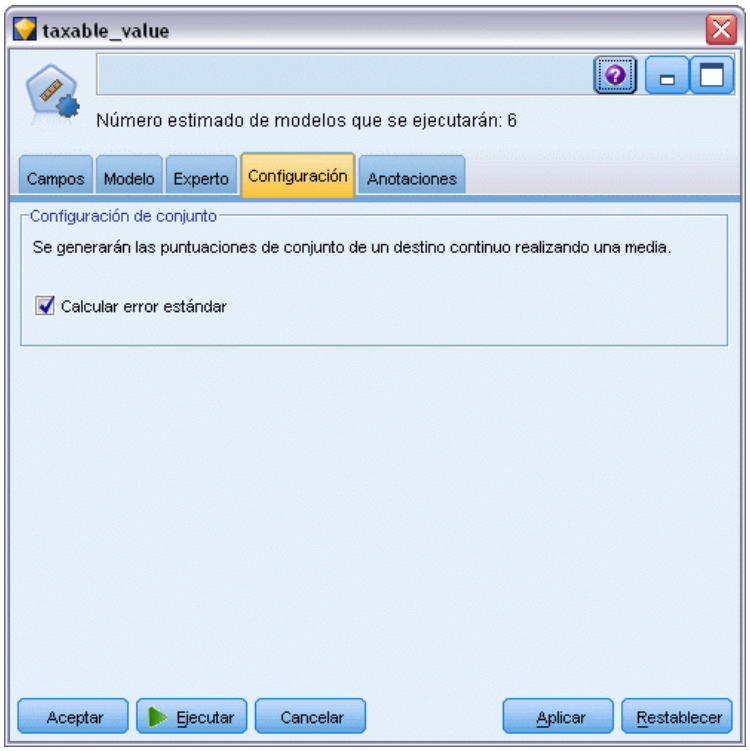
Como ha establecido Número de modelos que se utilizarán como 3 en la pestaña Modelo, el nodo calculará la precisión de los siete algoritmos y generará un nugget de modelo único con los tres más precisos.

Figura 6-5
Pestaña Experto del nodo Autonumérico



- En la pestaña Configuración, deje la configuración predefinida. Como se trata de un objetivo continuo, las puntuaciones se generan promediando las puntuaciones de los modelos individuales.

Figura 6-6
Pestaña Configuración del nodo Autonumérico



Comparación de los modelos

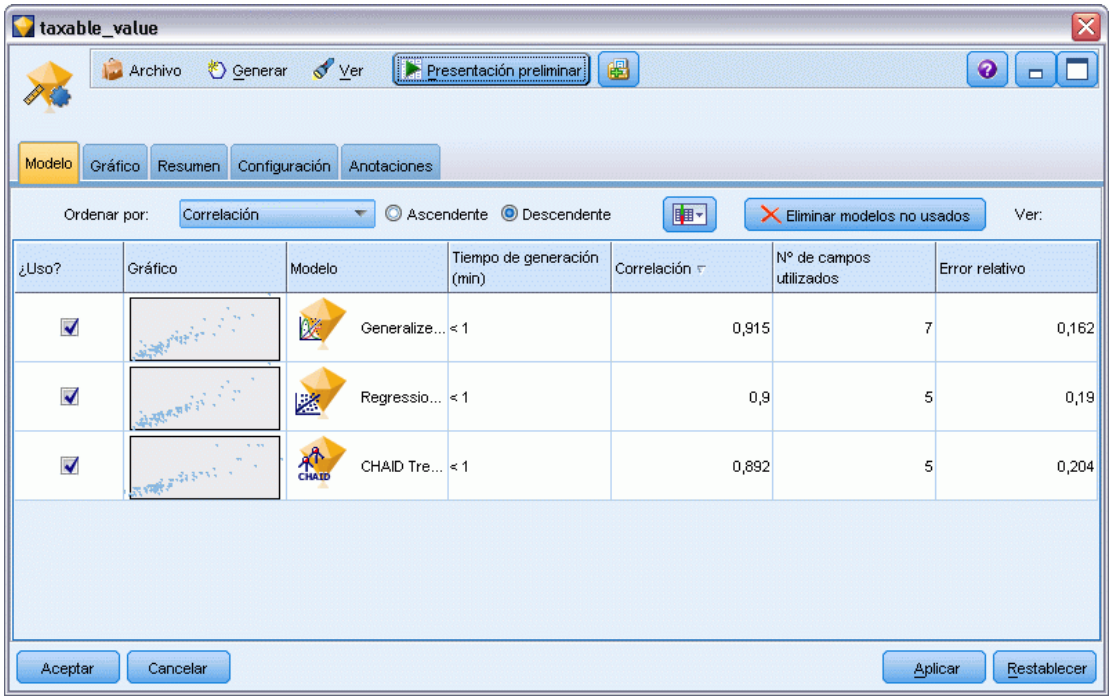
- Pulse en el botón Ejecutar.

Se crea el nugget del modelo y se coloca en el lienzo y, en la paleta Modelos en la esquina superior derecha de la ventana. Puede examinar el nugget o guardarlo para distribuirlo en diferentes formas.

Abra el nugget de modelo; enumera los detalles de cada uno de los modelos creados durante la ejecución. (En una situación real, en la que se estiman cientos de modelos en un conjunto de datos mayor, este proceso puede tardar horas.) Consulte Figura 6-1 el p. 56.

Si desea seguir explorando cualquiera de los modelos individuales, puede pulsar dos veces en el icono del nugget de modelo en la columna Modelo para profundizar y examinar los resultados del modelo individual; desde ahí puede generar nodos de modelado, nugget de modelo o gráficos de evaluación.

Figura 6-7
Resultados Autonuméricos



Por defecto, los modelos se clasifican en función de su correlación, porque es la medida que ha seleccionado en el nodo Autonumérico. Para la clasificación se utiliza el valor absoluto de la correlación, con los valores más cercanos a 1 que indican una relación más estrecha. El modelo Lineal generalizado ordena mejor esta medida, pero hay otros modelos igualmente precisos. El modelo Lineal generalizado también produce el menor error relativo.

Puede ordenar una columna diferente pulsando en el encabezado de la columna o seleccionar la medida que desee de la lista Ordenar por de la barra de herramientas.

Cada gráfico muestra los valores observados en comparación con los valores pronosticados del modelo, lo que ofrece una rápida indicación visual de la correlación entre ellos. En un modelo correcto, los puntos deben estar situados a lo largo de la diagonal, que se cumple para todos los modelos de este ejemplo.

En la columna Gráfico, puede pulsar dos veces en una miniatura para generar un gráfico a tamaño completo.

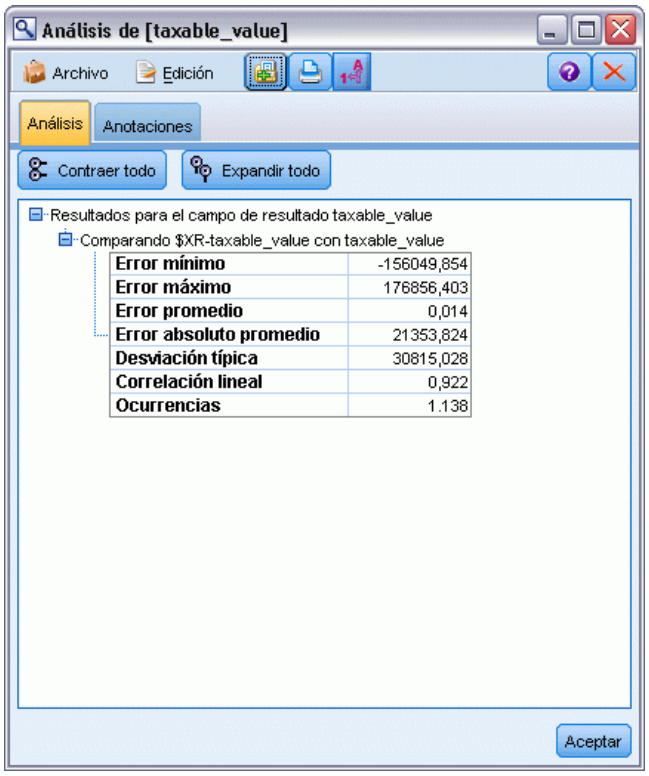
Según estos resultados, puede decidir utilizar los tres de estos modelos más precisos. Al combinar predicciones de varios modelos, pueden evitarse las limitaciones en modelos individuales que dan como resultado una precisión global superior.

En la columna Uso, asegúrese de que ha seleccionado los tres modelos.

Añada un nodo Análisis (paleta Resultado) después del nugget de modelo. Pulse con el botón derecho en el nodo Análisis y seleccione Ejecutar para ejecutar la ruta.

Las puntuaciones promediadas que genera el nodo Conjunto se añaden en un campo denominado *$XR-taxable_value*, con una correlación de 0,922, que tiene un valor superior a los de los tres modelos individuales. Las puntuaciones del conjunto también muestran un error absoluto medio bajo y pueden ejecutarse mejor que cualquier modelo individual cuando se aplica a otros conjuntos de datos.

Figura 6-8
Ruta de ejemplo de Autonumérico



Resumen

En resumen, ha utilizado el nodo Autonumérico para comparar diferentes modelos, ha seleccionado los tres modelos más precisos y los ha añadido a la ruta dentro de un nugget de modelo Autonumérico de conjunto.

- En función de su precisión global, los modelos Lineal generalizado, Regresión y CHAID ejecutan mejor los datos de formación.
- Este conjunto de modelos mostró un rendimiento mejor que el mejor de los dos modelos individuales y se comportarán aún mejor cuando se apliquen a otros conjuntos de datos. Si su objetivo es automatizar el proceso lo máximo posible, este método le permite obtener un modelo robusto en la mayoría de circunstancias, sin tener que entrar demasiado en las características específicas de un modelo.

Parte II:
Ejemplos de preparación de datos

Capítulo

7

Preparación automática de datos (ADP)

La preparación de los datos para el análisis es uno de los pasos más importantes en cualquier proyecto de minería de datos y, tradicionalmente, uno de los que exigen más tiempo. El nodo Preparación automática de datos (ADP) gestiona esta función, analiza los datos e identifica los valores fijos, filtra los campos problemáticos o que no serán útiles, deriva nuevos atributos cuando es necesario y mejora el rendimiento mediante técnicas de filtrado y muestreo inteligente. Puede utilizar el nodo de forma totalmente automática, permitiendo que el nodo seleccione y aplique valores fijos, o bien puede tener una vista previa de los cambios antes de que se apliquen y aceptarlos o rechazarlos.

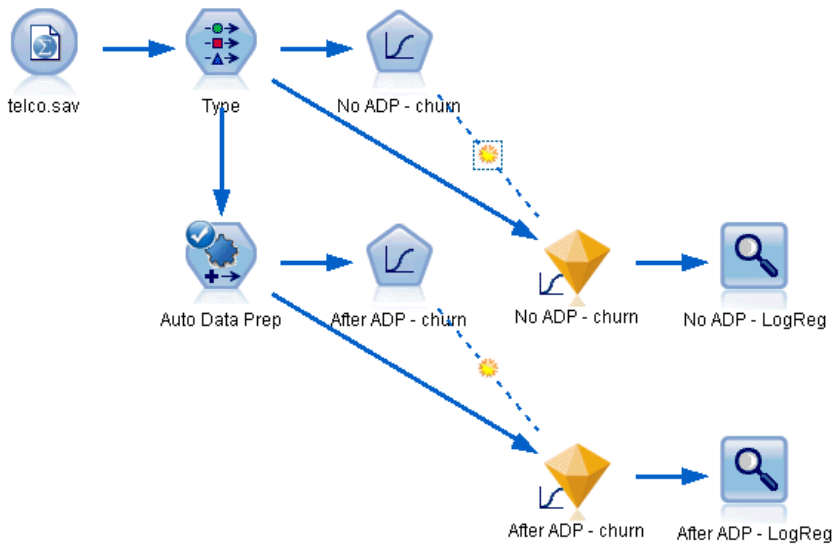
El uso del nodo ADP le permite preparar sus datos de forma rápida y simple para la minería de datos, sin necesidad de tener un conocimiento previo de los conceptos estadísticos necesarios. Si ejecuta el nodo con los valores por defecto, los modelos tenderán a construir y puntuar más rápidamente.

Este ejemplo utiliza la ruta denominada *ADP_basic_demo.str*, que hace referencia al archivo de datos denominado *telco.sav* para demostrar la precisión aumentada que pueden encontrarse utilizando la configuración del nodo ADP por defecto cuando se construyen modelos. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *ADP_basic_demo.str* se encuentra en el directorio *streams*.

Generación de la ruta

- Para generar la ruta, añada un nodo de origen de archivo Statistics que apunte a *telco.sav*, que se encuentra en el directorio *Demos* de la instalación de IBM® SPSS® Modeler.

Figura 7-1
Generación de la ruta

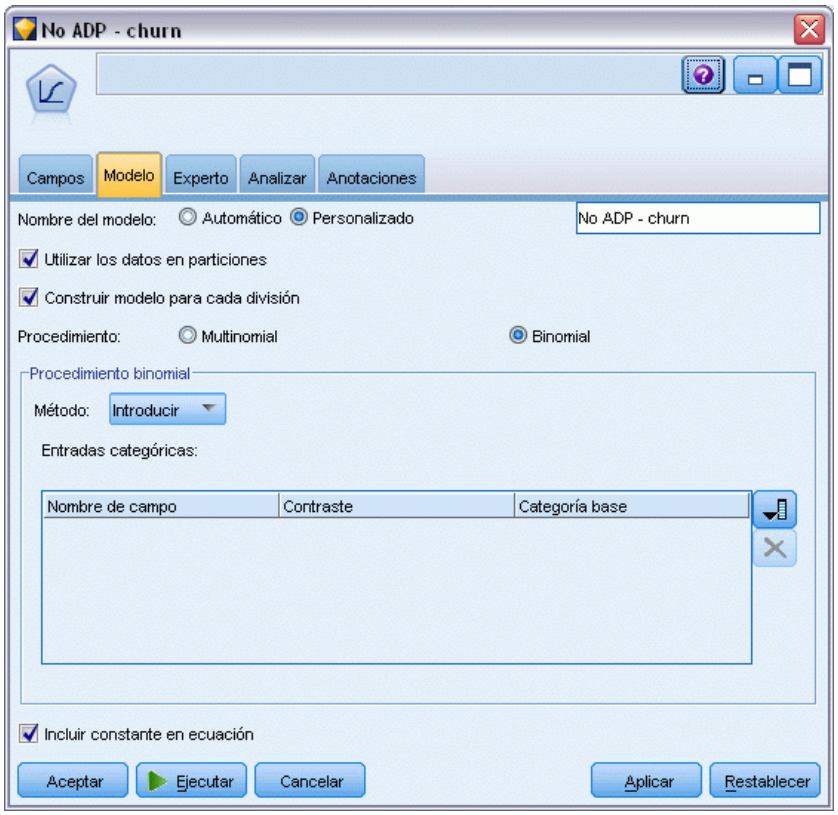


- Conecte un nodo Tipo al nodo de origen, defina el nivel de medición del campo *abandono* a Marca, y defina el papel a Marca. El resto de campos debe tener sus papeles definidas en Entrada.

Figura 7-2
Selección del objetivo

- Conecte un nodo Logística al nodo Tipo.
- En el nodo Logística, pulse en la pestaña Modelo y seleccione el procedimiento Binomial. En el campo *Nombre de modelo*, seleccione Personalizado e introduzca Sin ADP - abandono.

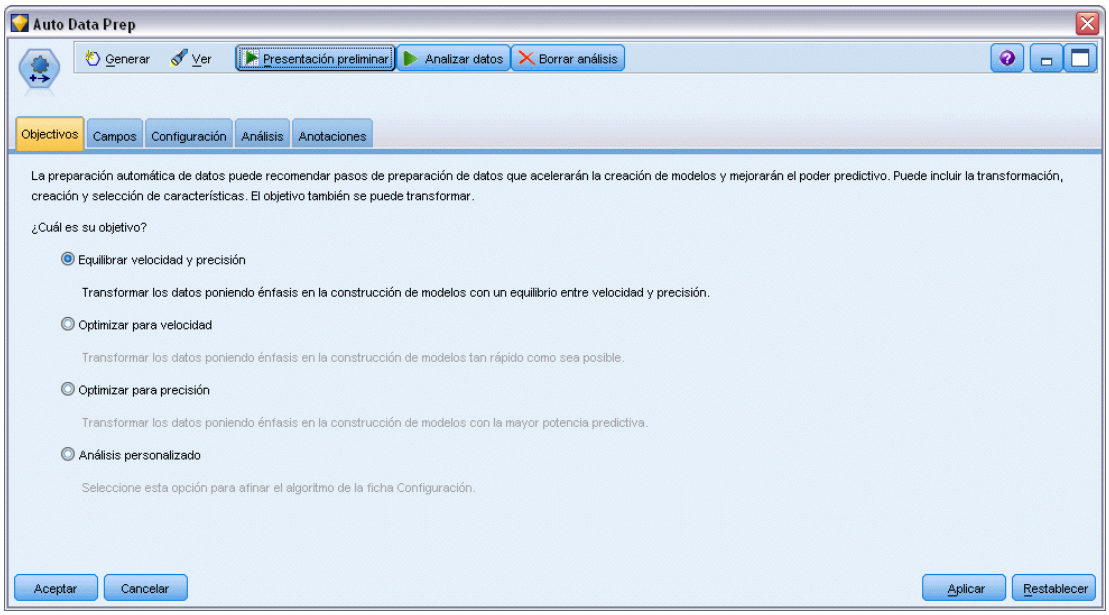
Figura 7-3
Selección de opciones del modelo



- Conecte un nodo ADP al nodo Tipo. En la pestaña Objetivos, deje la configuración por defecto para analizar y preparar sus datos equilibrando la velocidad y la precisión.
- En la parte superior de la pestaña Objetivos, pulse en Analizar datos para analizar y procesar sus datos.

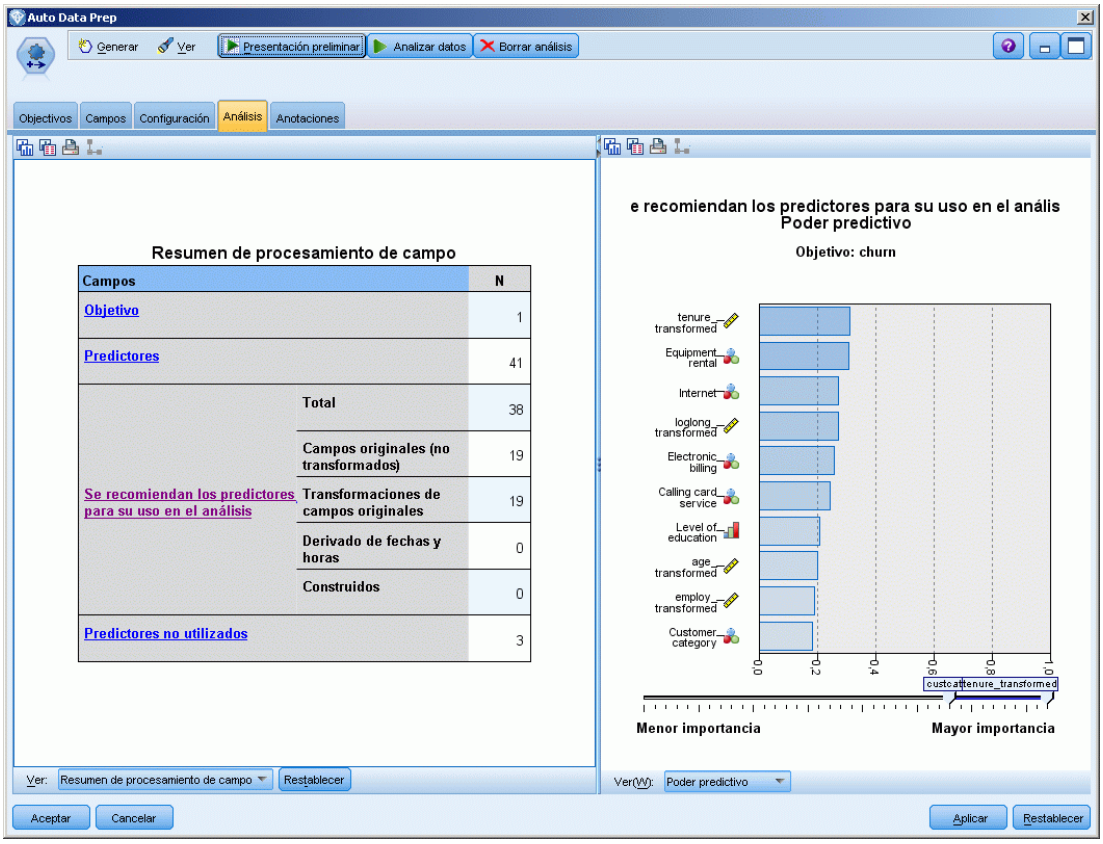
El resto de las opciones del nodo ADP le permiten especificar que desea concentrarse más en la precisión, más en la velocidad de procesamiento o para afinar la cantidad de los pasos de procesamiento de preparación de los datos.

Figura 7-4
Objetivos ADP por defecto



Los resultados del procesamiento de los datos se muestran en la pestaña Análisis. El Resumen del procesamiento de campos muestra que de las 41 funciones de datos que introdujo el nodo ADP, 19 se han transformado para ayudar al procesamiento y que 3 se han descartado como no utilizadas.

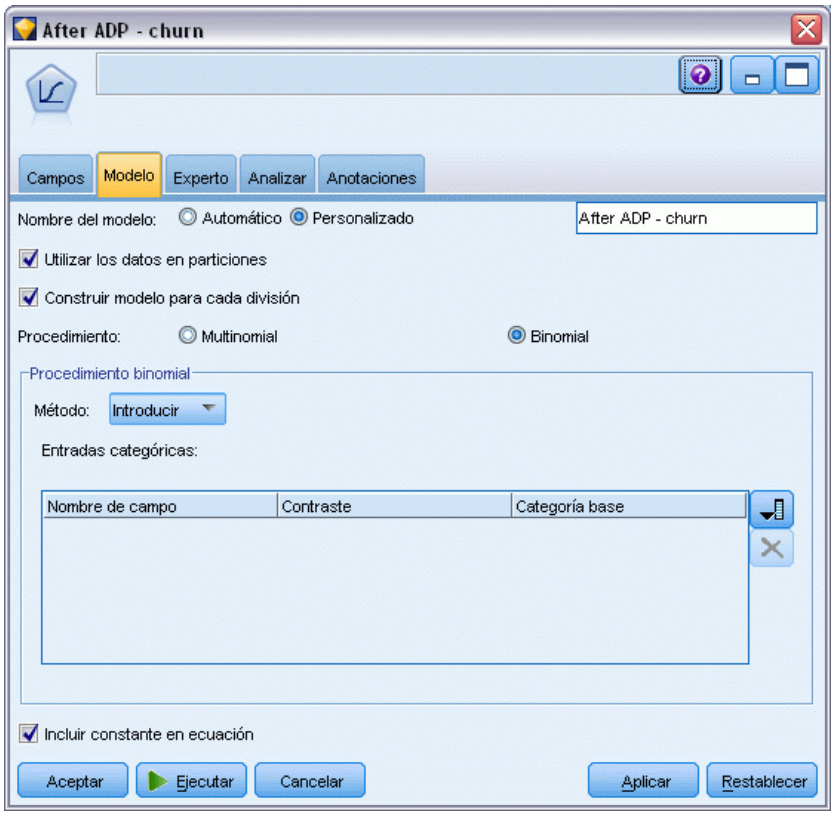
Figura 7-5
Resumen del procesamiento de datos



- Conecte un nodo Logística al nodo ADP.

- En el nodo Logística, pulse en la pestaña Modelo y seleccione el procedimiento Binomial. En el campo *Nombre de modelado*, seleccione Personalizado e introduzca Tras ADP - abandono.

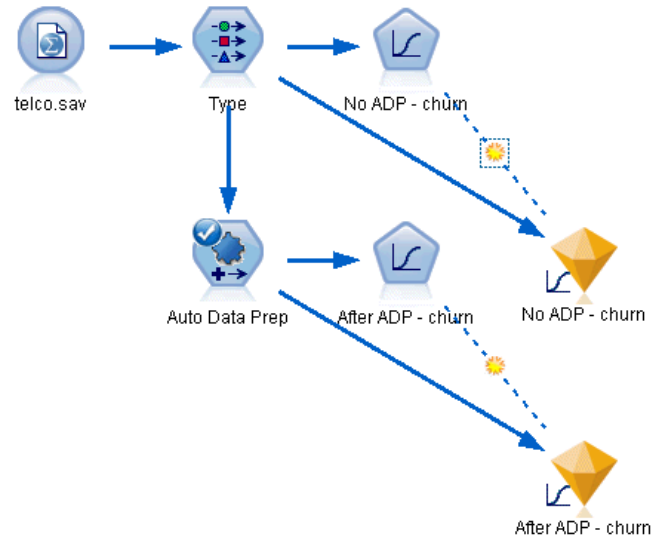
Figura 7-6
Selección de opciones del modelo



Comparación de la precisión de modelos

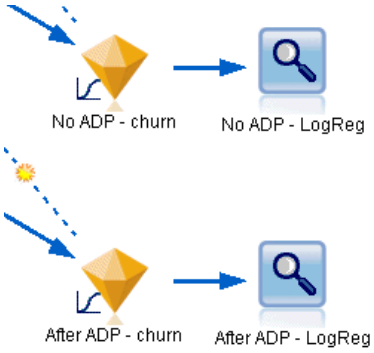
- Ejecute ambos nodos Logística para generar los nuggets de modelos, que se añadirán a la ruta y a la paleta de modelos situada en la esquina superior derecha.

Figura 7-7
Conexión de los nuggets de modelos



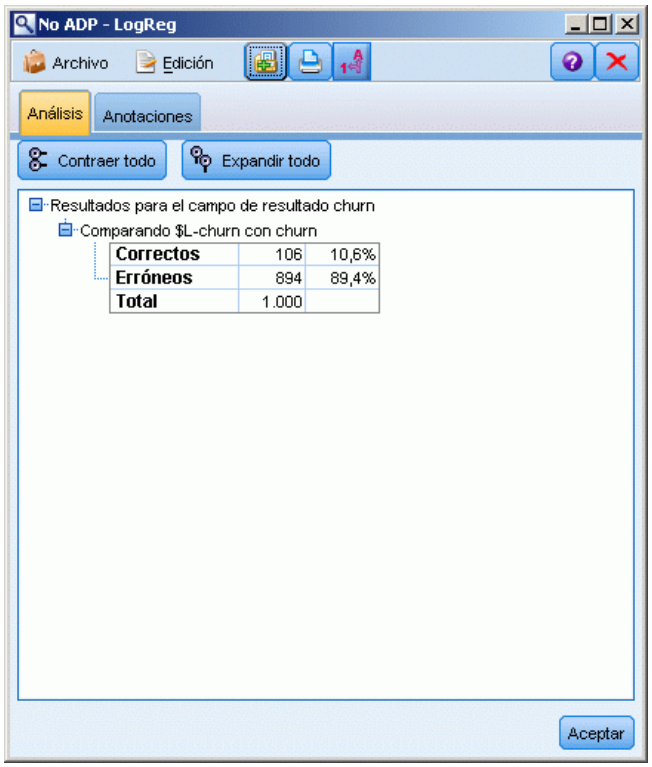
- Conecte los nodos Análisis a los nuggets de modelos y ejecute los nodos Análisis utilizando su configuración por defecto.

Figura 7-8
Conexión de los nodos Análisis



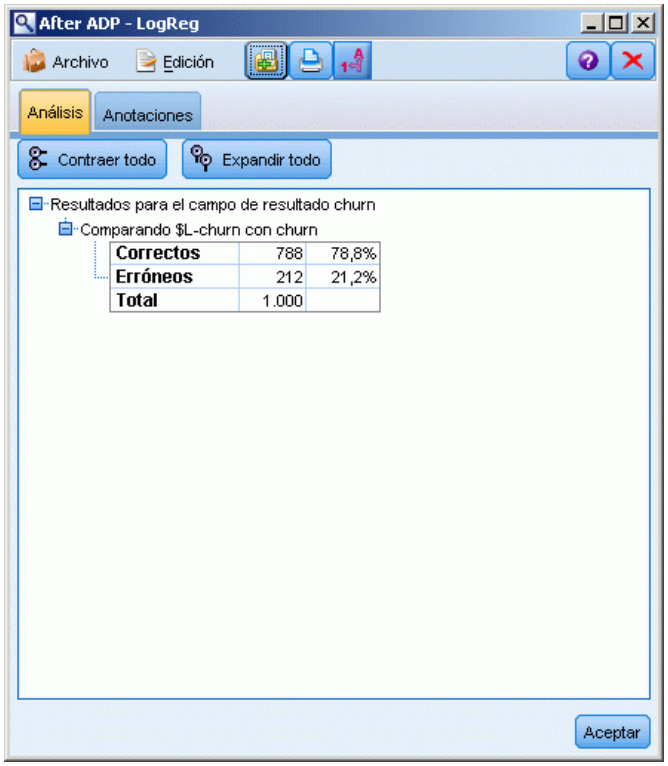
El análisis del modelo derivado no ADP muestra que sólo ejecutando los datos a través del nodo Regresión logística con su configuración por defecto ofrece un modelo con una precisión muy baja de sólo el 10,6%.

Figura 7-9
Resultados de modelos derivados no ADP



El análisis del modelo derivado ADP muestra que la ejecución de los datos con la configuración ADP por defecto ha construido un modelo mucho más preciso que tienen un 78,8% de corrección.

Figura 7-10
Resultados de modelos derivados ADP



En resumen, sólo ejecutando el nodo ADP para afinar el procesamiento de los datos, podrá construir un modelo mucho más preciso con muy poca manipulación directa de los datos.

Obviamente, si está interesado en probar o desaprobar una teoría en particular, o si desea construir modelos específicos, es posible que encuentre beneficioso trabajar directamente con la configuración de modelos; sin embargo, para los usuarios con poco tiempo disponible, o con una gran cantidad de datos para preparar, el nodo ADP puede darle ventaja.

Puede encontrar explicaciones de los fundamentos matemáticos de los métodos de modelado utilizados en IBM® SPSS® Modeler en el *Manual de algoritmos de SPSS Modeler*, disponible en el directorio *\Documentation* del disco de instalación.

Recuerde que estos resultados están basados sólo en los datos de entrenamiento. Para evaluar qué tal se extiende el modelo a otros datos de casos reales, se utilizaría un nodo de partición para reservar un subconjunto de registros para comprobación y validación. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 14.2*.

Capítulo

8

Preparación de los datos para análisis (Auditar datos)

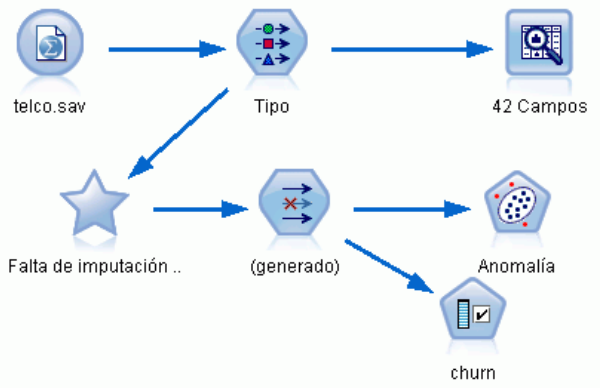
El nodo Auditar datos ofrece un primer vistazo exhaustivo a los datos introducidos en IBM® SPSS® Modeler. Normalmente utilizado durante la exploración de datos iniciales, el informe de auditoría de datos muestra estadísticos de resumen, así como histogramas y gráficos de distribución para cada campo de datos, y permite especificar el tratamiento de valores perdidos, atípicos y extremos.

Este ejemplo utiliza la ruta denominada *telco_dataaudit.str*, que hace referencia al archivo de datos denominado *telco.sav*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas SPSS Modeler en el menú Inicio de Windows. El archivo *telco_dataaudit.str* está ubicado en el directorio *streams*.

Generación de la ruta

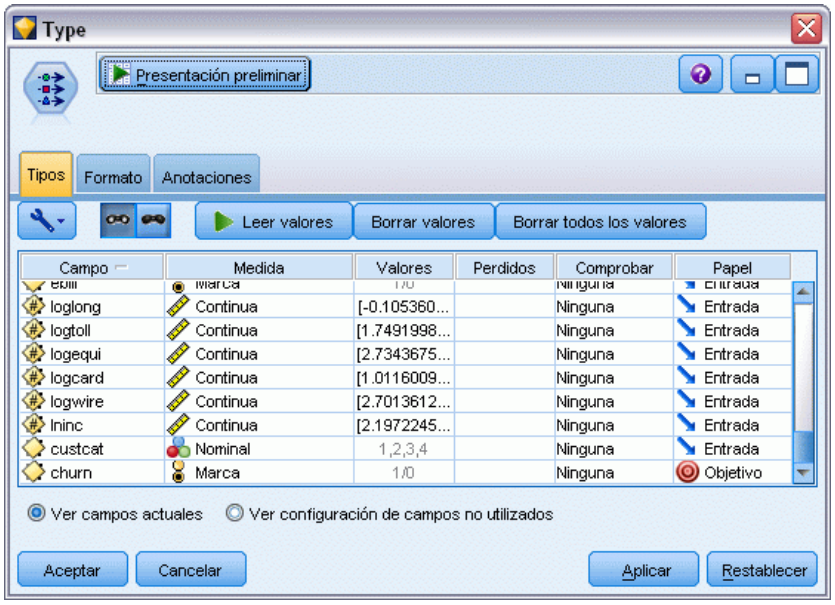
- Para generar la ruta, añada un nodo de origen de archivo Statistics que apunte a *telco.sav*, que se encuentra en el directorio *Demos* de la instalación de IBM® SPSS® Modeler.

Figura 8-1
Generación de la ruta



- Añada un nodo Tipo para definir campos y especifique *churn* como campo objetivo (Papel = Objetivo). Se debe definir el papel como Entrada en el resto de los campos para que éste sea el único objetivo.

Figura 8-2
Definición del objetivo



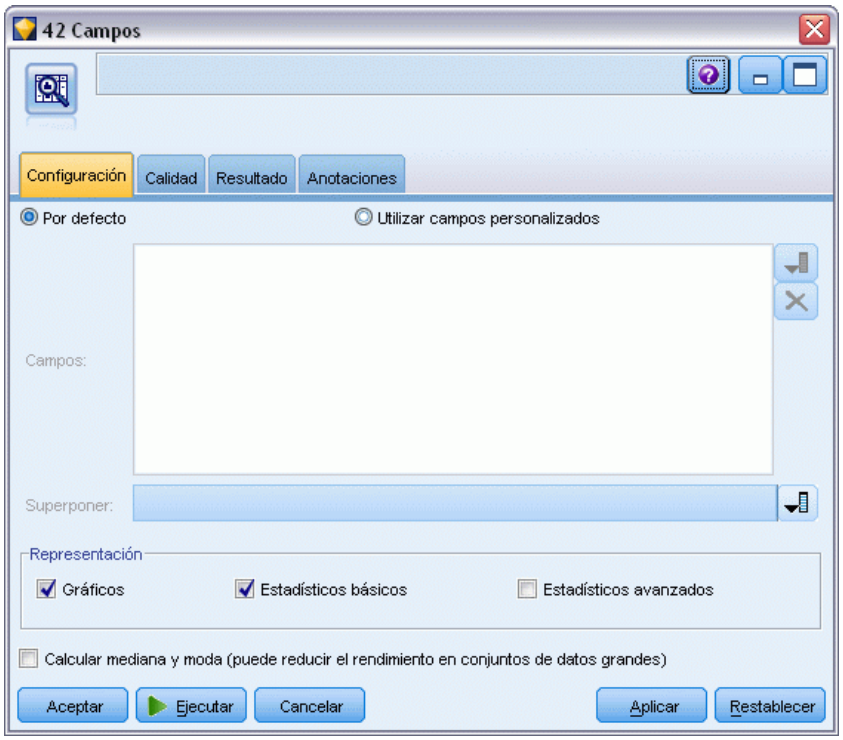
- Confirme que los niveles de medición de campos están definidos correctamente. Por ejemplo, la mayoría de los campos con valores 0 y 1 se pueden considerar como marcas, pero algunos campos, como Sexo, se ven con más precisión como un campo nominal con dos valores.

Figura 8-3
Definición de los niveles de medición

Sugerencia: Para cambiar propiedades de varios campos con valores similares (como 0 y 1), pulse en la cabecera de la columna *Valores* para ordenar campos por dicha columna, y utilice la tecla Mayús para seleccionar todos los campos que quiera cambiar. Después, pulse con el botón derecho en la selección para cambiar el nivel de medición u otros atributos de todos los campos seleccionados.

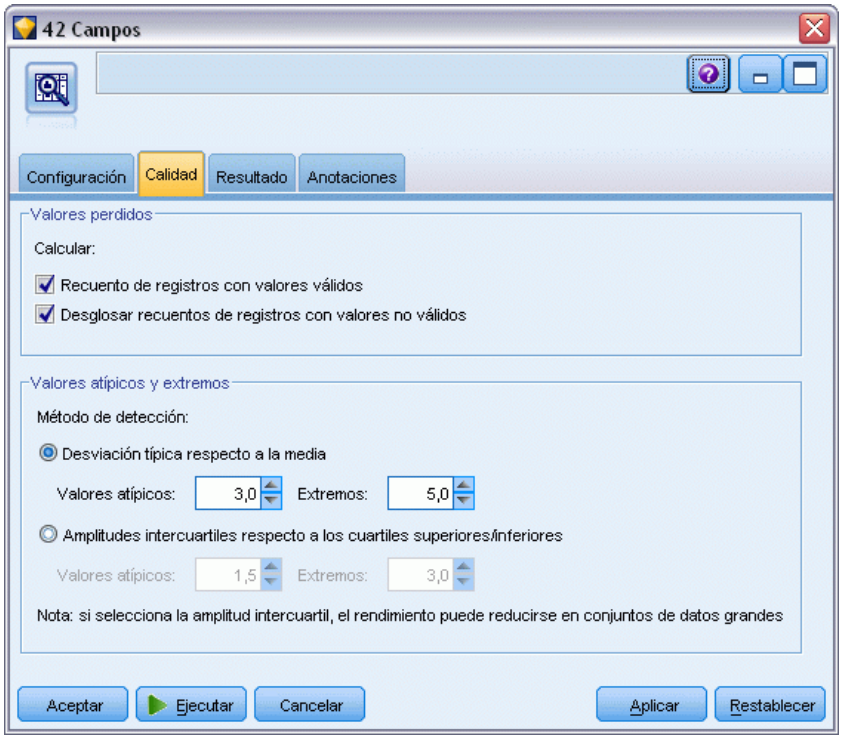
- Conecte a la ruta un nodo Auditar datos. En la pestaña Configuración, deje la configuración por defecto para incluir todos los campos del informe. Puesto que *churn* es el único campo objetivo definido en el nodo Tipo, se utilizará automáticamente como superposición.

Figura 8-4
Pestaña Configuración del nodo Auditar datos



En la pestaña Calidad, deje la configuración por defecto para detectar valores perdidos, atípicos y extremos, y pulse en Ejecutar.

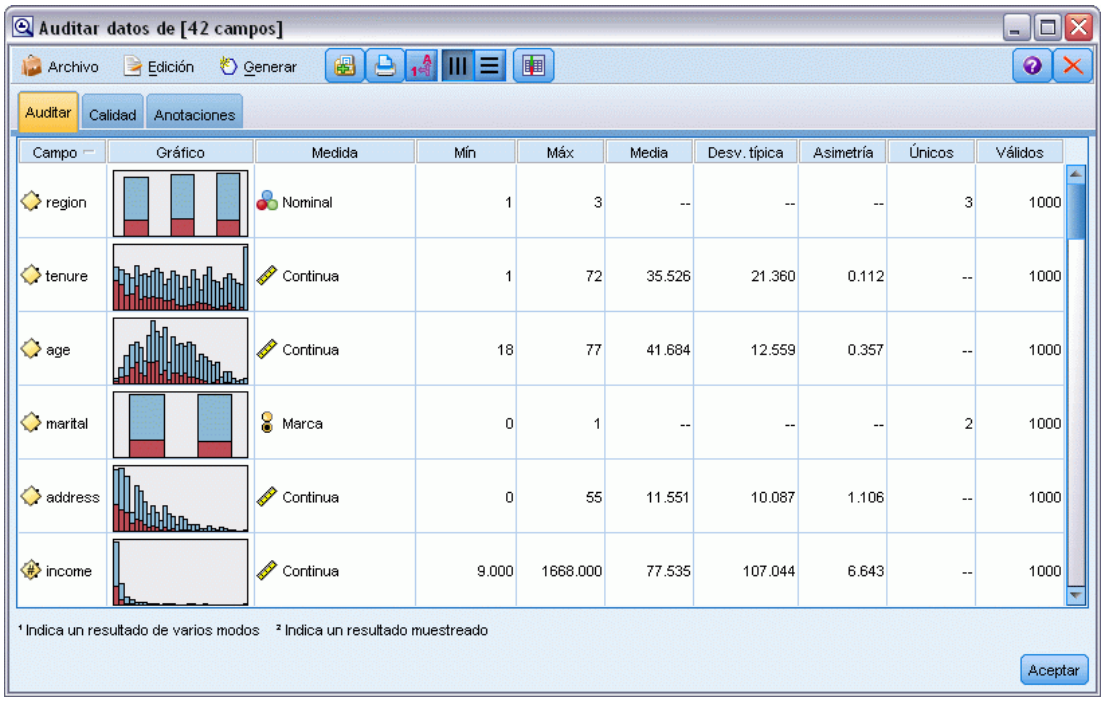
Figura 8-5
Pestaña Calidad del nodo Auditar datos



Exploración de estadísticas y gráficos

Se muestra el explorador de auditoría de datos, con gráficos en miniatura y estadísticos descriptivos para todos los campos.

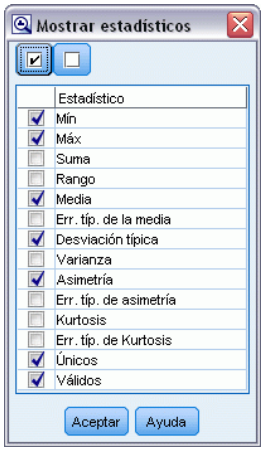
Figura 8-6
explorador de auditoría de datos



Utilice la barra de herramientas para mostrar etiquetas de valor y de campo y para conmutar la alineación de gráficas de horizontal a vertical (sólo para campos categóricos).

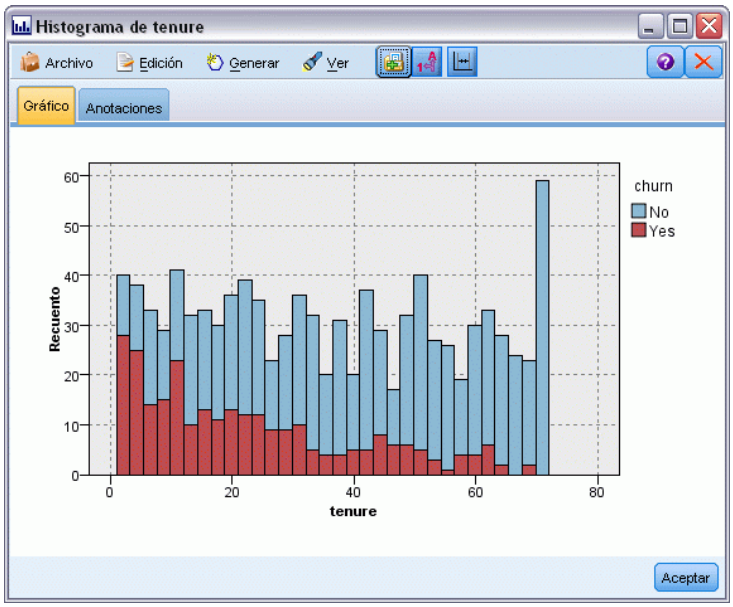
- También puede utilizar la barra de herramientas o el menú Edición para seleccionar los estadísticos que desea mostrar.

Figura 8-7
Mostrar estadísticos



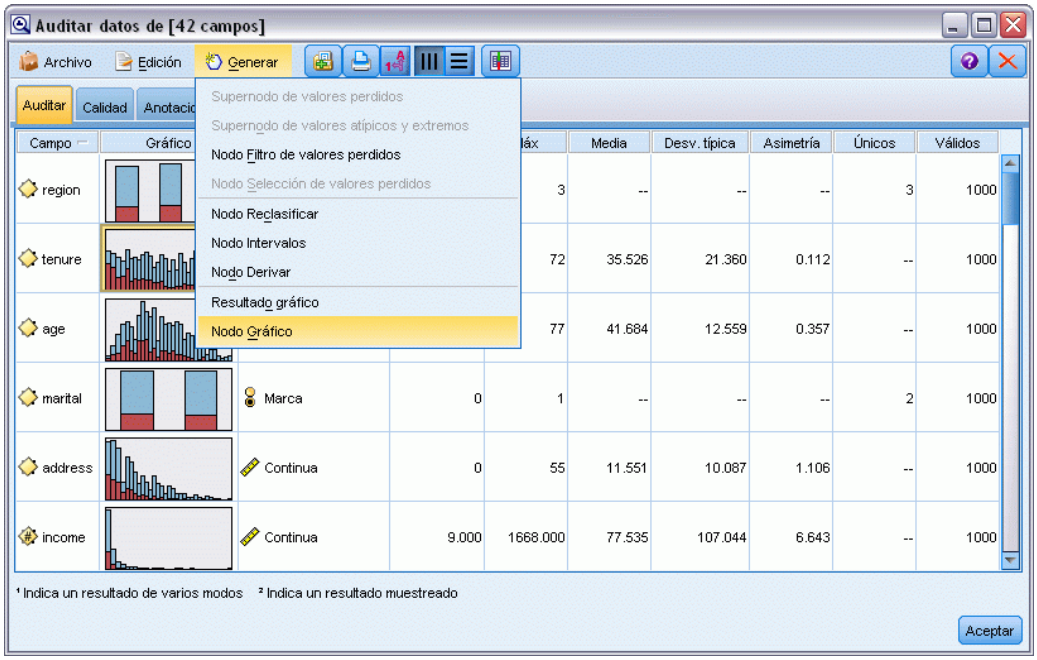
Pulse dos veces en cualquier gráfico en miniatura del informe de auditoría para ver una versión a tamaño completo de dicho gráfico. Puesto que *churn* es el único campo objetivo de la ruta, se utiliza automáticamente como superposición. Si desea cambiar la visualización de las etiquetas de valor y de campo, puede utilizar la barra de herramientas de la ventana del gráfico, o bien pulsar en el botón de modo de edición para personalizar el gráfico.

Figura 8-8
Histograma de cargo



Si lo prefiere, puede seleccionar uno o varios gráficos en miniatura y generar un nodo Gráfico para cada uno. Los nodos generados se colocan en el lienzo de rutas y se pueden añadir a la ruta para volver a crear ese gráfico en concreto.

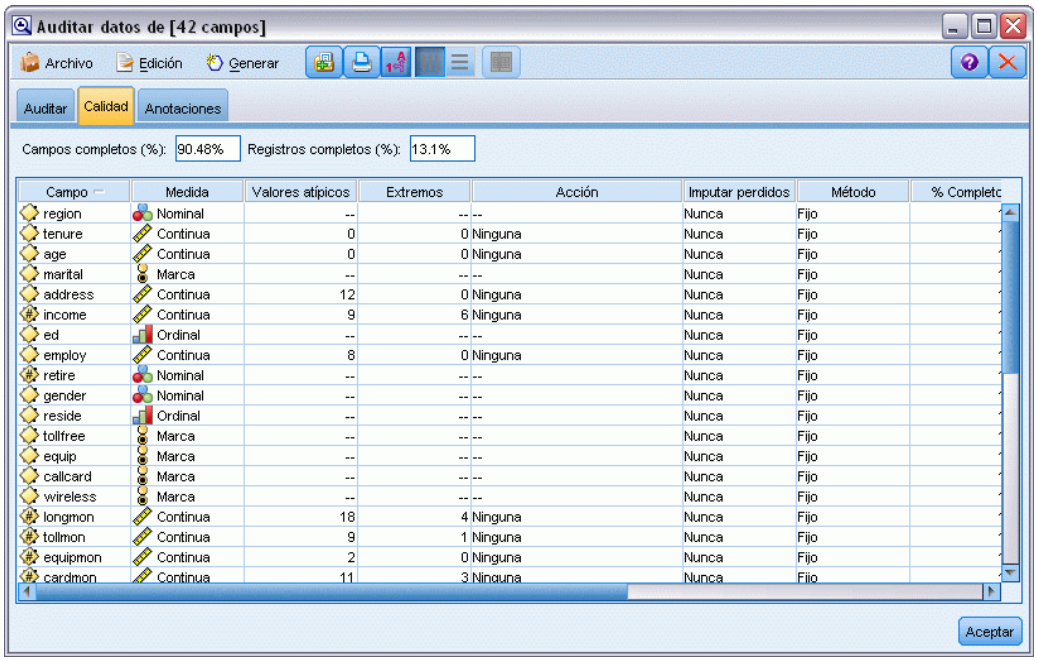
Figura 8-9
Generación de un nodo Gráfico



Gestión de valores atípicos y perdidos

La pestaña Calidad del informe de auditoría muestra información sobre valores atípicos, extremos y perdidos.

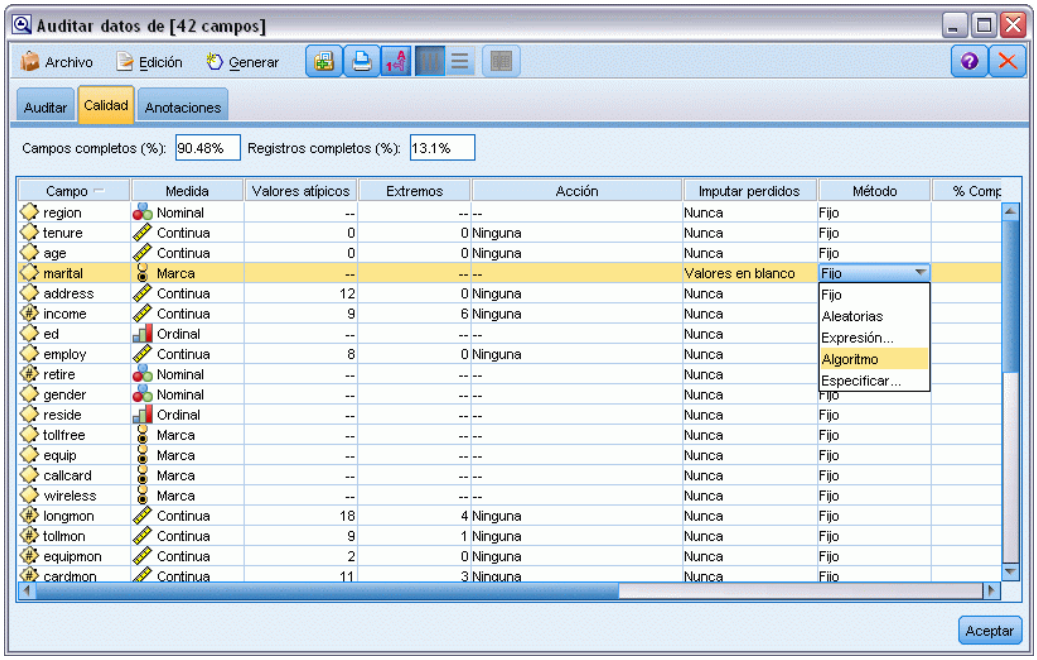
Figura 8-10
Pestaña Calidad del explorador de auditoría de datos



Campo	Medida	Valores atípicos	Extremos	Acción	Imputar perdidos	Método	% Completo
region	Nominal	--	--	--	Nunca	Fijo	
tenure	Continua	0	0	Ninguna	Nunca	Fijo	
age	Continua	0	0	Ninguna	Nunca	Fijo	
marital	Marca	--	--	--	Nunca	Fijo	
address	Continua	12	0	Ninguna	Nunca	Fijo	
income	Continua	9	6	Ninguna	Nunca	Fijo	
ed	Ordinal	--	--	--	Nunca	Fijo	
employ	Continua	8	0	Ninguna	Nunca	Fijo	
retire	Nominal	--	--	--	Nunca	Fijo	
gender	Nominal	--	--	--	Nunca	Fijo	
reside	Ordinal	--	--	--	Nunca	Fijo	
tollfree	Marca	--	--	--	Nunca	Fijo	
equip	Marca	--	--	--	Nunca	Fijo	
callcard	Marca	--	--	--	Nunca	Fijo	
wireless	Marca	--	--	--	Nunca	Fijo	
longmon	Continua	18	4	Ninguna	Nunca	Fijo	
tollmon	Continua	9	1	Ninguna	Nunca	Fijo	
equipmon	Continua	2	0	Ninguna	Nunca	Fijo	
cardmon	Continua	11	3	Ninguna	Nunca	Fijo	

También puede especificar métodos para gestionar estos valores y generar Supernodos para aplicar las transformaciones automáticamente. Por ejemplo, puede seleccionar uno o más campos e imputar o reemplazar valores perdidos para campos específicos con varios métodos, entre ellos el algoritmo C&RT.

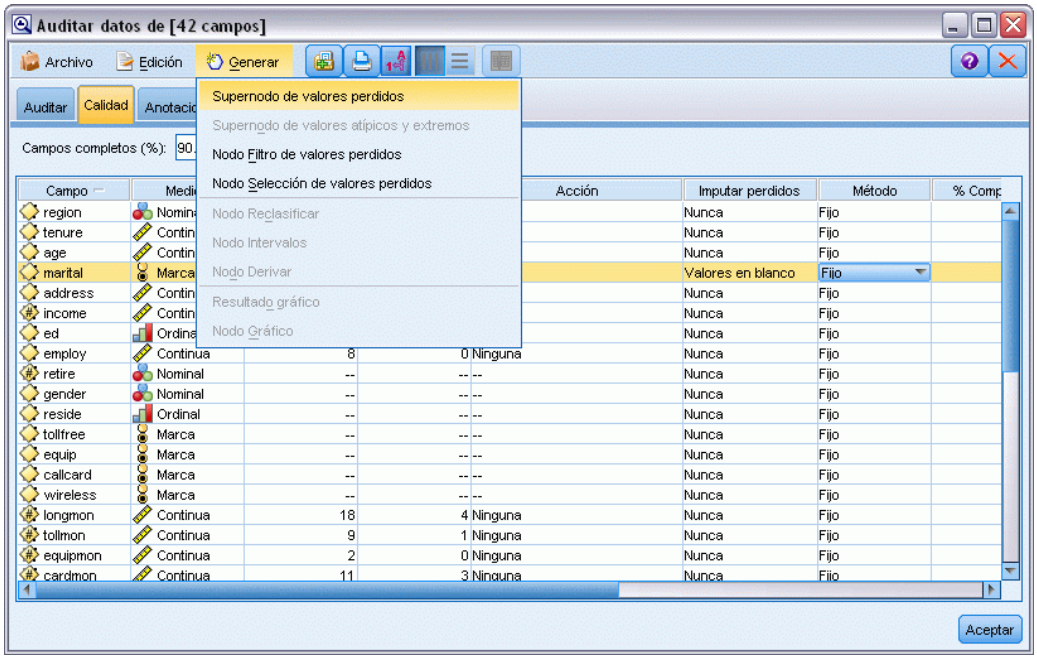
Figura 8-11
Selección de un método de imputación



Después de especificar un método de imputación para uno o más campos, para generar un Supernodo de valores perdidos, seleccione:

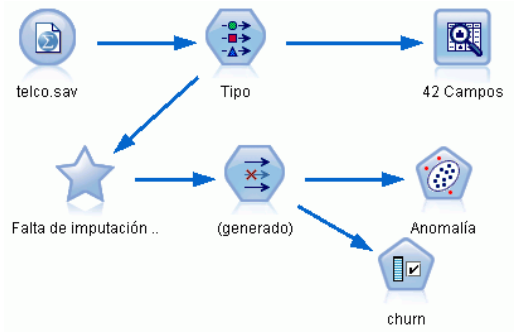
Generar > Supernodo de valores perdidos

Figura 8-12
Generación del Supernodo



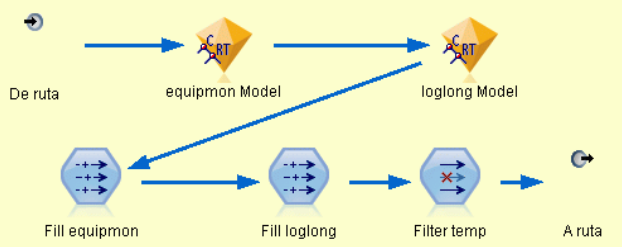
El Supernodo generado se añade al lienzo de rutas, donde lo puede conectar a la ruta para aplicar las transformaciones.

Figura 8-13
Ruta con Supernodo de valores perdidos



El Supernodo contiene una serie de nodos que realizan las transformaciones solicitadas. Para comprender cómo funciona, puede editar el Supernodo y pulsar en Acercar.

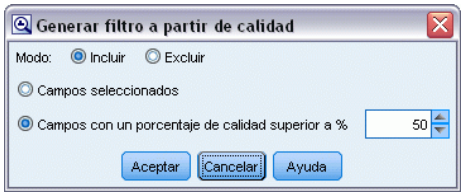
Figura 8-14
Acercamiento al Supernodo



En cada campo imputado con el método de algoritmo, por ejemplo, habrá un modelo C&RT independiente, junto con un nodo Rellenar que sustituye valores vacíos y nulos con el valor que predice el modelo. Puede añadir, editar o eliminar nodos específicos con el Supernodo para personalizar más el comportamiento.

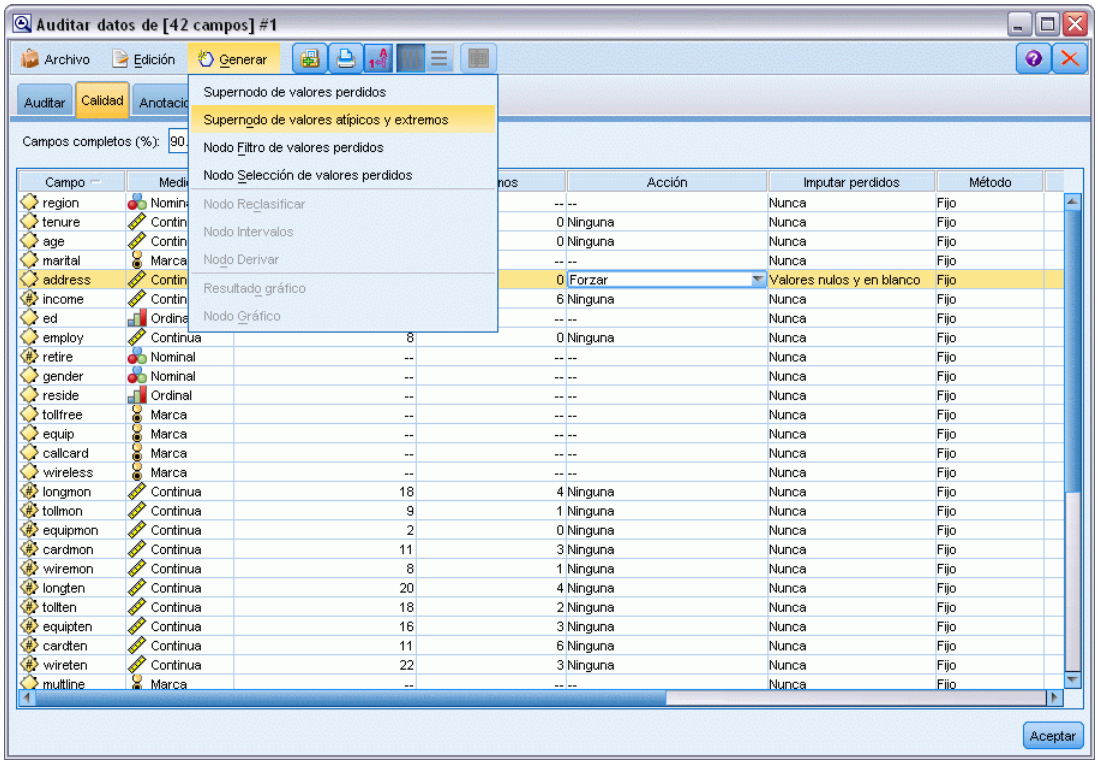
Si lo prefiere, puede generar un nodo Seleccionar o Filtro para eliminar campos o registros con valores perdidos. Por ejemplo, puede filtrar cualquier campo que tenga un porcentaje de calidad por debajo de un umbral específico.

Figura 8-15
Generación de un nodo Filtro



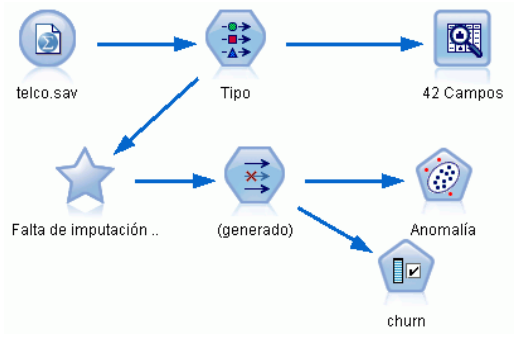
Los valores atípicos y extremos se pueden gestionar de manera similar. Especifique la acción que desea realizar en cada campo (forzar, descartar o anular) y genere un Supernodo para aplicar las transformaciones.

Figura 8-16
Generación de un nodo Filtro



Después de completar la auditoría y añadir a la ruta los nodos generados, puede continuar con el análisis. Si lo desea, puede filtrar más los datos mediante Detección de anomalías, Selección de características u otros métodos.

Figura 8-17
Ruta con Supernodo de valores perdidos



Capítulo

9

Tratamientos con medicamentos (Gráficos exploratorios/C5.0)

Para esta sección, imagine que es un investigador médico que está recopilando datos para un estudio. Ha recopilado información sobre un conjunto de pacientes, de los cuales todos sufrieron la misma enfermedad. Durante el curso del tratamiento, cada paciente respondió a un medicamento de un total de cinco. Parte de su trabajo consiste en utilizar la minería de datos para averiguar qué medicamento es el adecuado para un futuro paciente con la misma enfermedad.

Este ejemplo utiliza la ruta denominada *druglearn.str*, que hace referencia al archivo de datos denominado *DRUG1n*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *druglearn.str* se encuentra en el directorio *streams*.

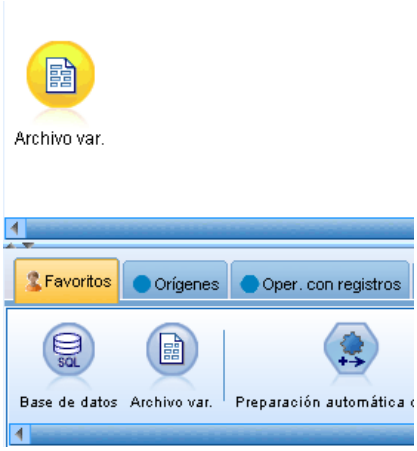
Los campos de datos que se utilizan en esta demostración son:

Campo Datos	**Descripción**
Edad	(número)
Sexo	*M* o *F*
PS	Presión sanguínea: *ALTO*, *NORMAL* o *BAJO*
Colesterol	Colesterol en sangre: *NORMAL* o *ALTO*
Na	Concentración de sodio en sangre
K	Concentración de potasio en sangre
Medicamento	Medicamento prescrito al que respondió un paciente

Lectura de datos de texto

Puede leer datos de texto delimitado utilizando un **nodo Archivo var**. Puede añadir un nodo Archivo var. desde las paletas, bien buscando el nodo en la pestaña Orígenes o bien mediante la pestaña Favoritos, que incluye este nodo por defecto. A continuación, pulse dos veces en el nuevo nodo para abrir su cuadro de diálogo.

Figura 9-1
Adición de un nodo Archivo variable



Pulse en el botón que contiene puntos suspensivos (...) y que está situado a la derecha del cuadro de texto Archivo para examinar el directorio en el que se encuentra instalado IBM® SPSS® Modeler. Abra el directorio *Demos* y seleccione el archivo *DRUG1n*.

Seleccionando la casilla Leer nombres de campo del archivo, asegúrese de que los campos y valores que se han cargado en el cuadro de diálogo.

Figura 9-2
Cuadro de diálogo Archivo var.

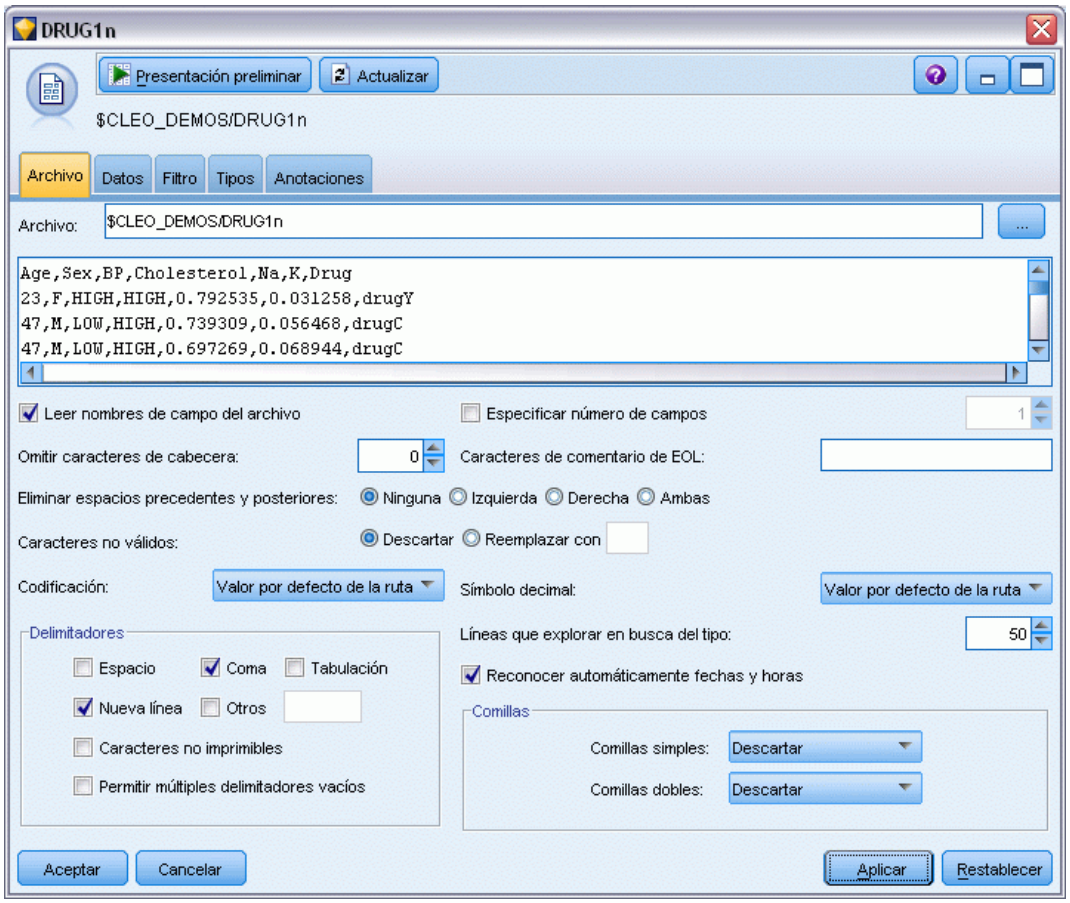


Figura 9-3
Cambio del tipo de almacenamiento para un campo

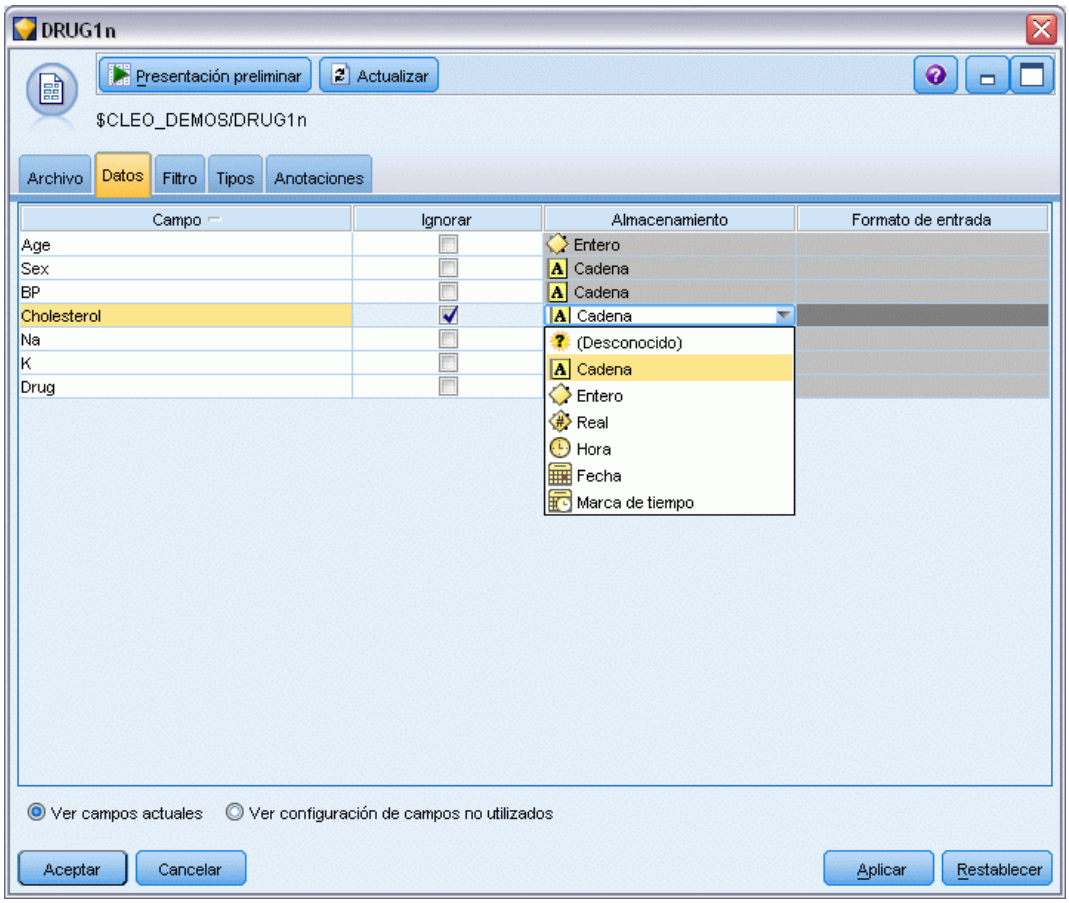
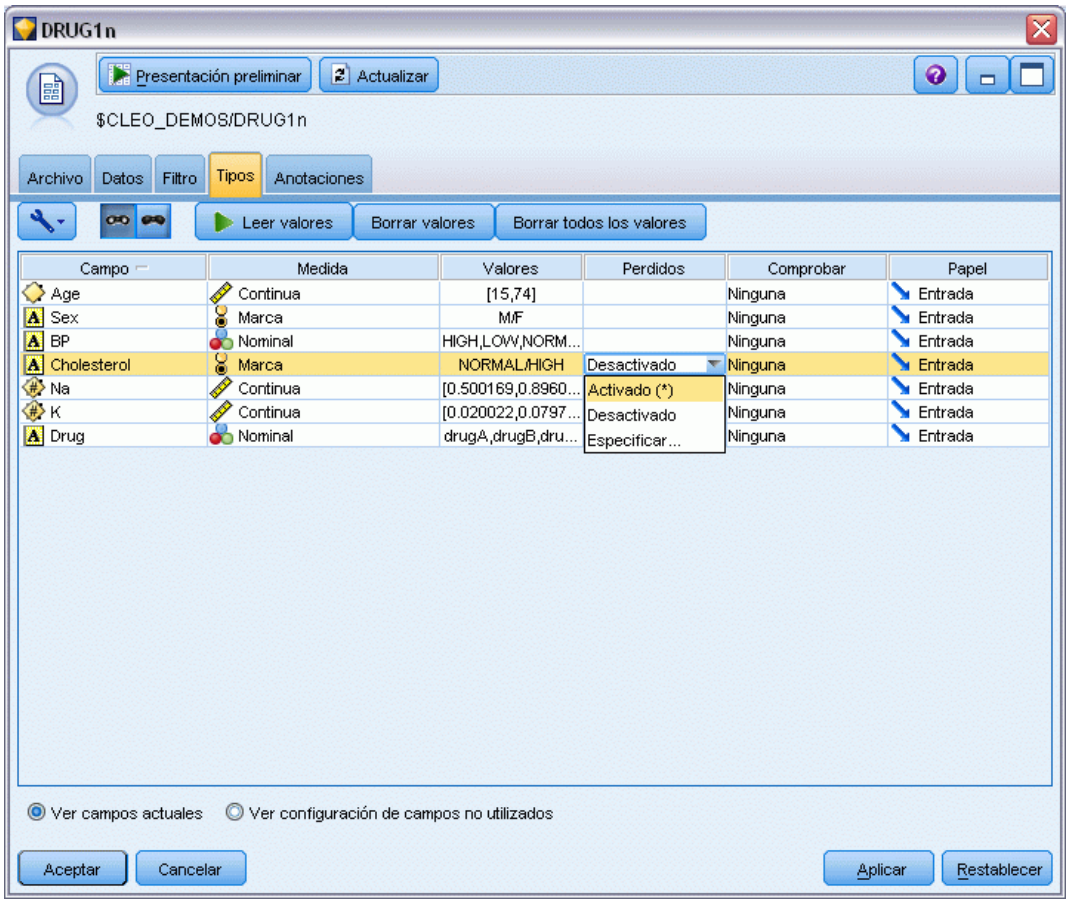


Figura 9-4
Selección de valores de la pestaña Tipos.



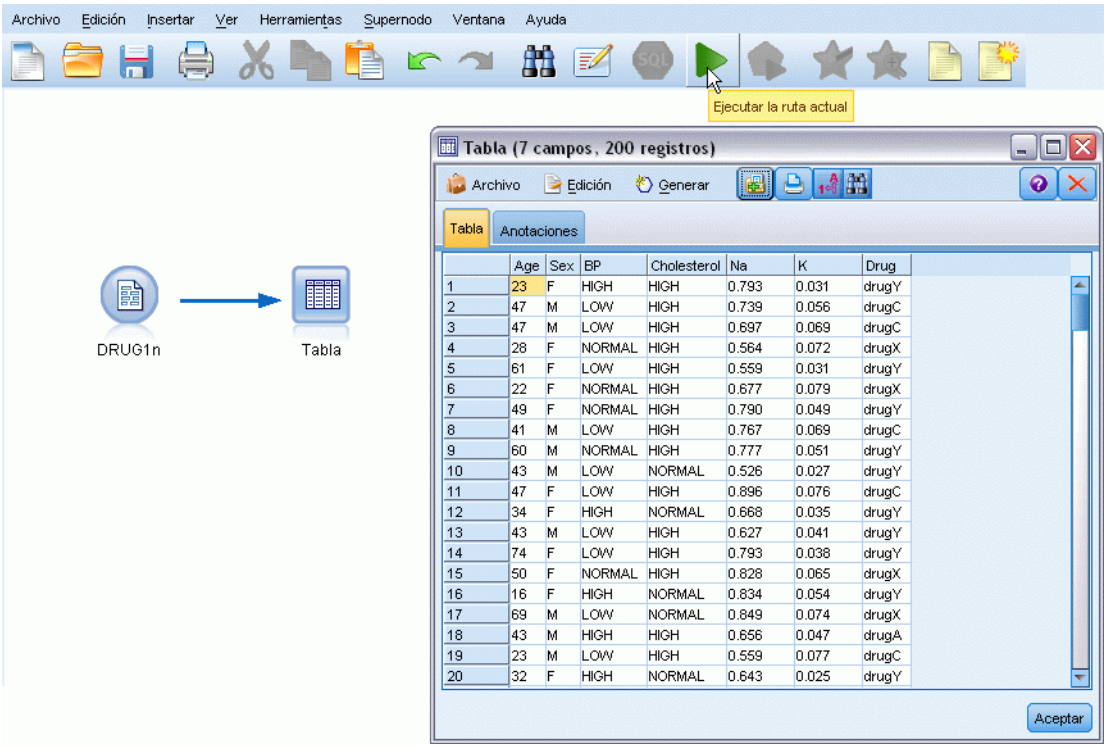
Pulse en la pestaña Datos para sustituir y cambiar los valores de **Almacenamiento** que corresponden a un campo. Tenga en cuenta que almacenamiento no es lo mismo que **Medición**, que es el nivel de medición (o tipo de uso) del campo de datos. La pestaña Tipos permite conocer mejor los tipos de campos de los datos. También puede seleccionar Leer valores para ver los valores reales de cada campo según los valores seleccionados en la columna *Valores*. Este proceso se conoce como **creación de una instancia**.

Adición de una tabla

Una vez que ha cargado el archivo de datos, puede echar un vistazo a los valores para ver el número de registros. Esto se puede hacer generando una ruta que incluya un nodo Tabla. Para colocar un nodo Tabla en una ruta, pulse dos veces en el icono de la paleta o arrastre y suelte el icono en el lienzo.

Figura 9-5
Nodo Tabla conectado al origen de datos

Figura 9-6
Ejecución de una ruta desde la barra de herramientas



	Age	Sex	BP	Cholesterol	Na	K	Drug
1	23	F	HIGH	HIGH	0.793	0.031	drugY
2	47	M	LOW	HIGH	0.739	0.056	drugC
3	47	M	LOW	HIGH	0.697	0.069	drugC
4	28	F	NORMAL	HIGH	0.564	0.072	drugX
5	61	F	LOW	HIGH	0.559	0.031	drugY
6	22	F	NORMAL	HIGH	0.677	0.079	drugX
7	49	F	NORMAL	HIGH	0.790	0.049	drugY
8	41	M	LOW	HIGH	0.767	0.069	drugC
9	60	M	NORMAL	HIGH	0.777	0.051	drugY
10	43	M	LOW	NORMAL	0.526	0.027	drugY
11	47	F	LOW	HIGH	0.896	0.076	drugC
12	34	F	HIGH	NORMAL	0.668	0.035	drugY
13	43	M	LOW	HIGH	0.627	0.041	drugY
14	74	F	LOW	HIGH	0.793	0.038	drugY
15	50	F	NORMAL	HIGH	0.828	0.065	drugX
16	16	F	HIGH	NORMAL	0.834	0.054	drugY
17	69	M	LOW	NORMAL	0.849	0.074	drugX
18	43	M	HIGH	HIGH	0.656	0.047	drugA
19	23	M	LOW	HIGH	0.559	0.077	drugC
20	32	F	HIGH	NORMAL	0.643	0.025	drugY

Al pulsar dos veces en un nodo de la paleta, se conectará automáticamente al nodo seleccionado en el lienzo de rutas. Si lo prefiere y aún no se han conectado los nodos, puede utilizar el botón central del ratón para conectar el nodo de origen al nodo Tabla. Para simular un botón central del ratón, mantenga pulsada la tecla Alt a la vez que utiliza el ratón. Para ver la tabla, pulse en el botón de flecha verde de la barra de herramientas para ejecutar la ruta o pulse con el botón derecho del ratón en el nodo Tabla y seleccione Ejecutar.

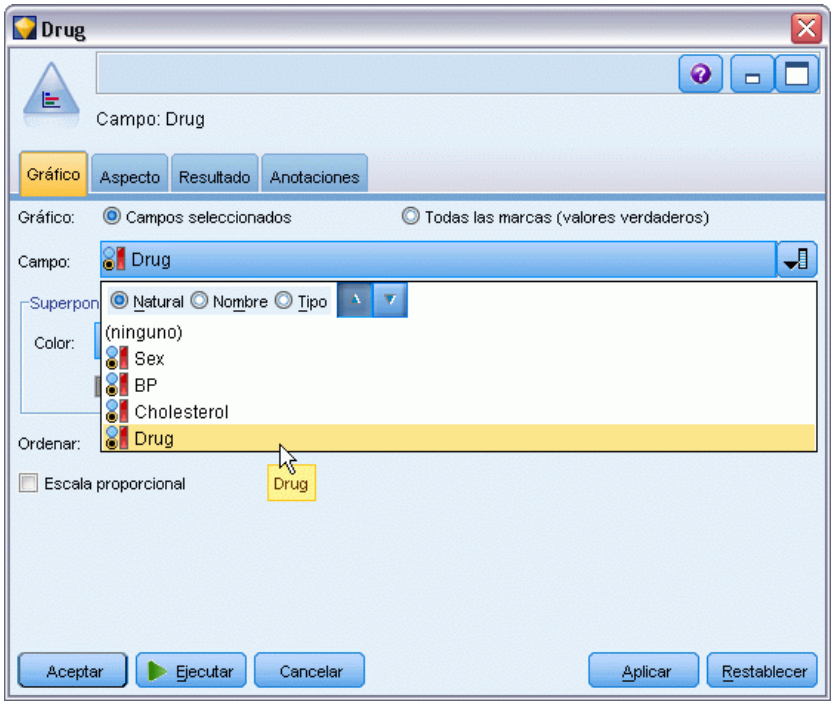
Creación de un gráfico de distribución

Durante el proceso de minería de datos, resulta útil examinar los datos mediante la creación de resúmenes visuales. IBM® SPSS® Modeler ofrece varios tipos diferentes de gráficos para seleccionar, según el tipo de datos que desea resumir. Por ejemplo, para averiguar qué proporción de pacientes respondió a cada medicamento, utilice el nodo Distribución.

Añada un nodo Distribución a la ruta y conéctelo al nodo de origen, a continuación, pulse dos veces en el nodo para editar las opciones de visualización.

Seleccione *Medicamento* como el campo objetivo cuya distribución desea mostrar. A continuación, pulse en Ejecutar en el cuadro de diálogo.

Figura 9-7
Selección de medicamento como el campo objetivo



El gráfico resultante le permite ver la "forma" de los datos. Muestra que los pacientes respondieron con más frecuencia al medicamento *Y*, y con menos frecuencia a los medicamentos *B* y *C*.

Figura 9-8
Distribución de la respuesta a un tipo de medicamento

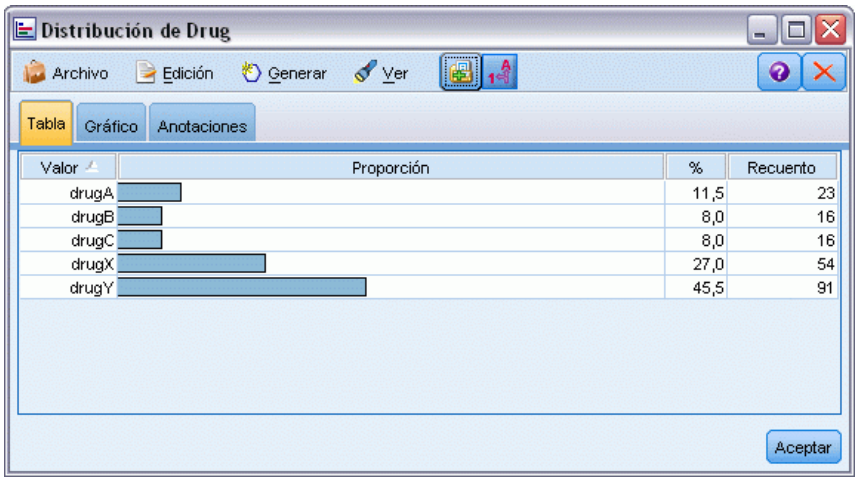
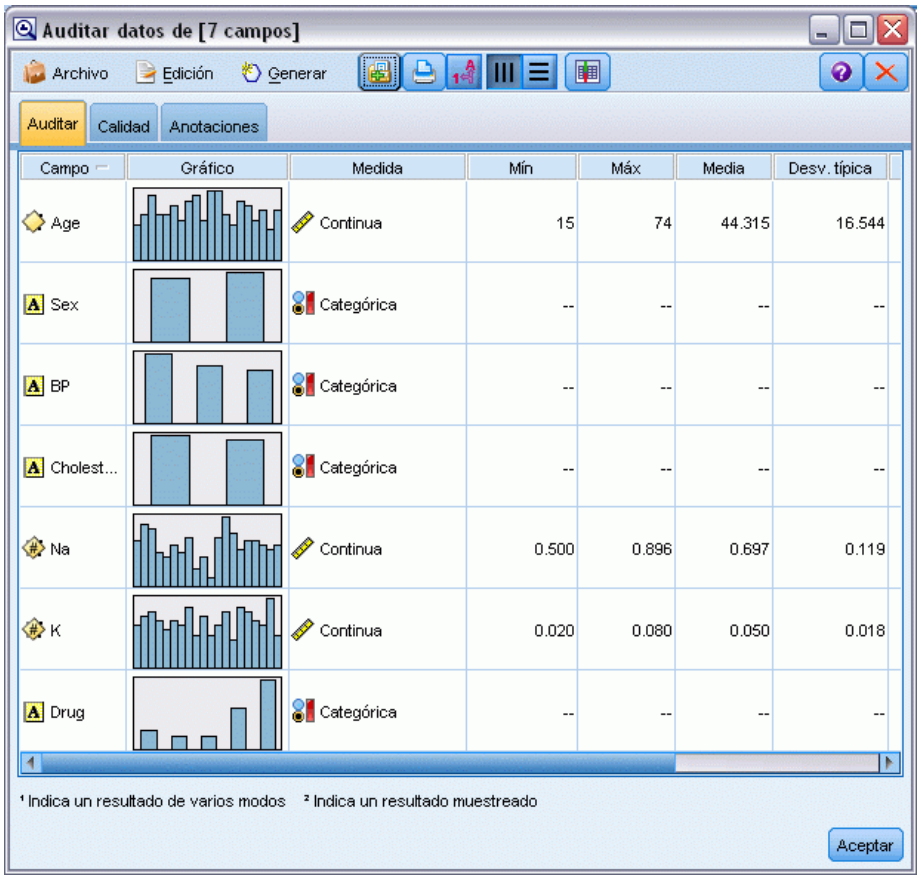


Figura 9-9
Resultados de un nodo Auditar datos



Otra posibilidad consiste en adjuntar un nodo Auditar datos para obtener una vista rápida de las distribuciones e histogramas de todos los campos a la vez. El nodo Auditar datos está disponible en la pestaña Resultados.

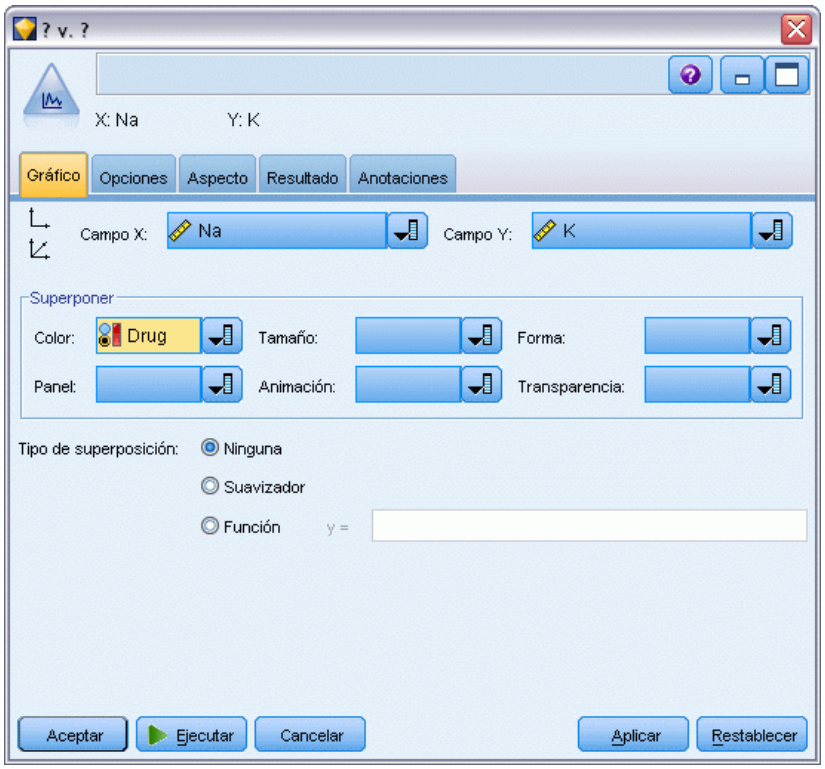
Creación de un diagrama de dispersión

Ahora, veamos los factores que pueden influir en *Medicamento*, la variable objetivo. Como investigador, sabe que las concentraciones de sodio y potasio en la sangre son factores importantes. Como se trata de valores numéricos, puede crear un diagrama de dispersión de sodio frente a potasio utilizando las categorías de medicamento como una superposición de colores.

Coloque un nodo Gráfico en el espacio de trabajo, conéctelo al nodo de origen y pulse dos veces en él para editarlo.

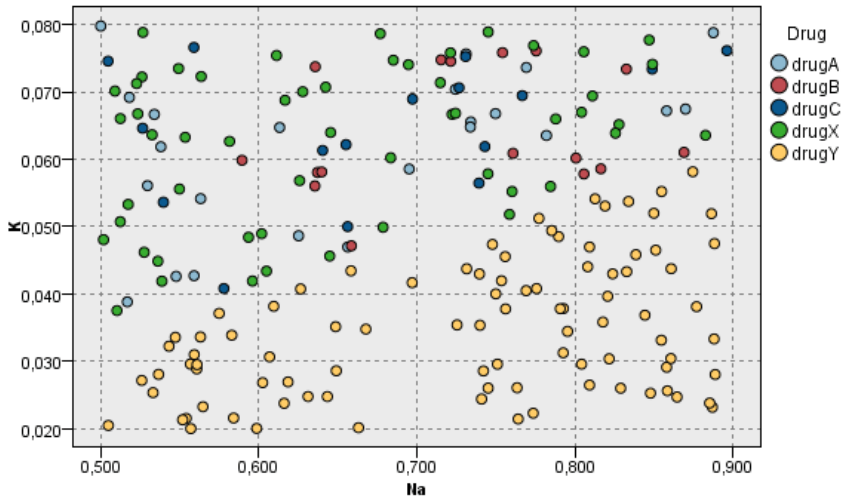
En la pestaña de gráficos, seleccione *Na* como el campo X, *K* como el campo Y, y *Medicamento* como el campo de superposición. A continuación, pulse en Ejecutar.

Figura 9-10
Creación de un diagrama de dispersión



El gráfico muestra claramente un umbral sobre el cual el medicamento correcto siempre es el medicamento *Y* y por debajo de él el medicamento correcto nunca es el medicamento *Y*. Este umbral es un cociente entre sodio (*Na*) y potasio (*K*).

Figura 9-11
Diagrama de dispersión de distribución de medicamentos

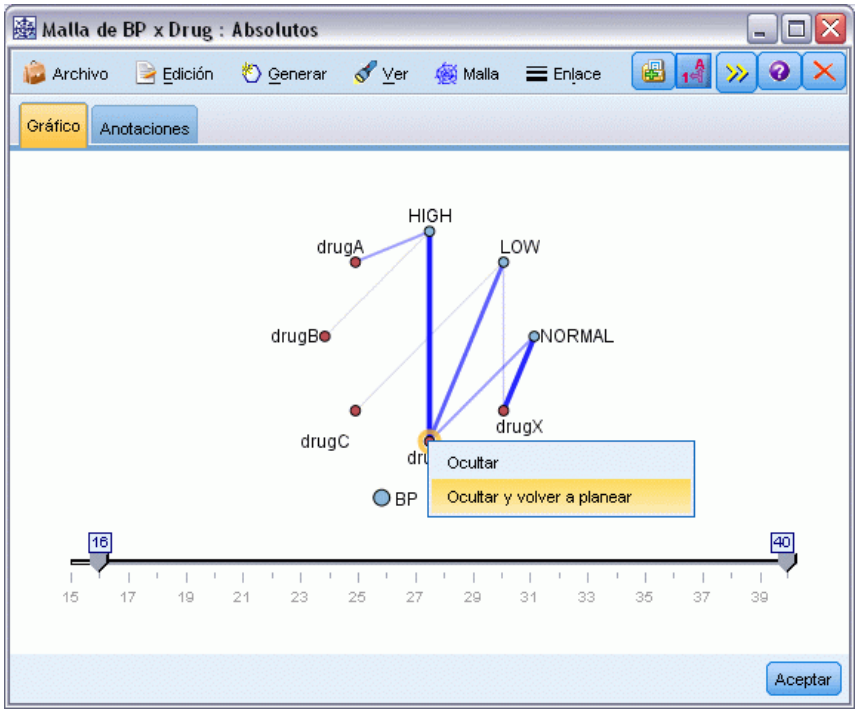


Creación de un gráfico de malla

Como algunos campos de datos son categóricos, puede intentar representar un gráfico de malla, que establece asociaciones entre distintas categorías. Empiece conectando un nodo Malla al nodo de origen en su espacio de trabajo. En el cuadro de diálogo del nodo Malla, seleccione *PS* (para presión sanguínea) y *Medicamento*. A continuación, pulse en Ejecutar.

Del gráfico, se extrae que el medicamento *Y* se asocia a los tres niveles de presión sanguínea. Esto no nos sorprende, ya que ya se ha determinado la situación en la que el medicamento *Y* es el más adecuado. Para centrarse en los demás medicamentos, puede ocultar *Y*. En el menú Ver, seleccione Modo edición, pulse con el botón derecho en el medicamento *Y* y seleccione Ocultar y volver a planear.

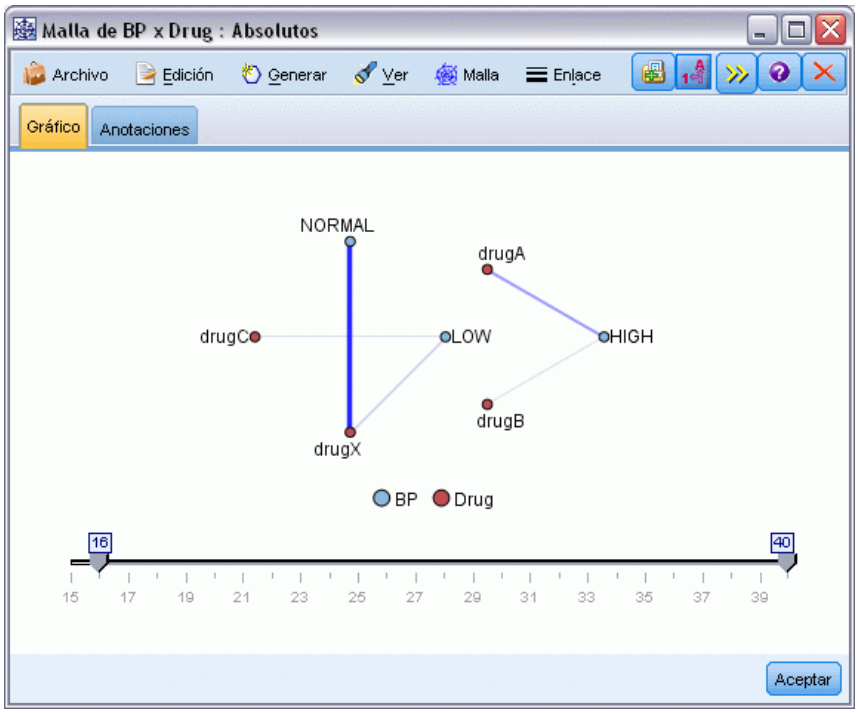
Figura 9-12
Gráfico de malla de medicamentos y presión sanguínea



En el gráfico simplificado, el medicamento *Y* y todos sus enlaces están ocultos. Ahora se puede ver claramente que sólo los medicamentos *A* y *B* están asociados a la presión sanguínea alta. Sólo los medicamentos *C* y *X* están asociados a la presión sanguínea baja. Y la presión sanguínea normal está asociada únicamente al medicamento *X*. En este punto, no obstante, aún no se sabe

cómo elegir entre los medicamentos *A* y *B*, o entre los medicamentos *C* y *X* para un paciente determinado. Es aquí donde el modelado resulta de gran utilidad.

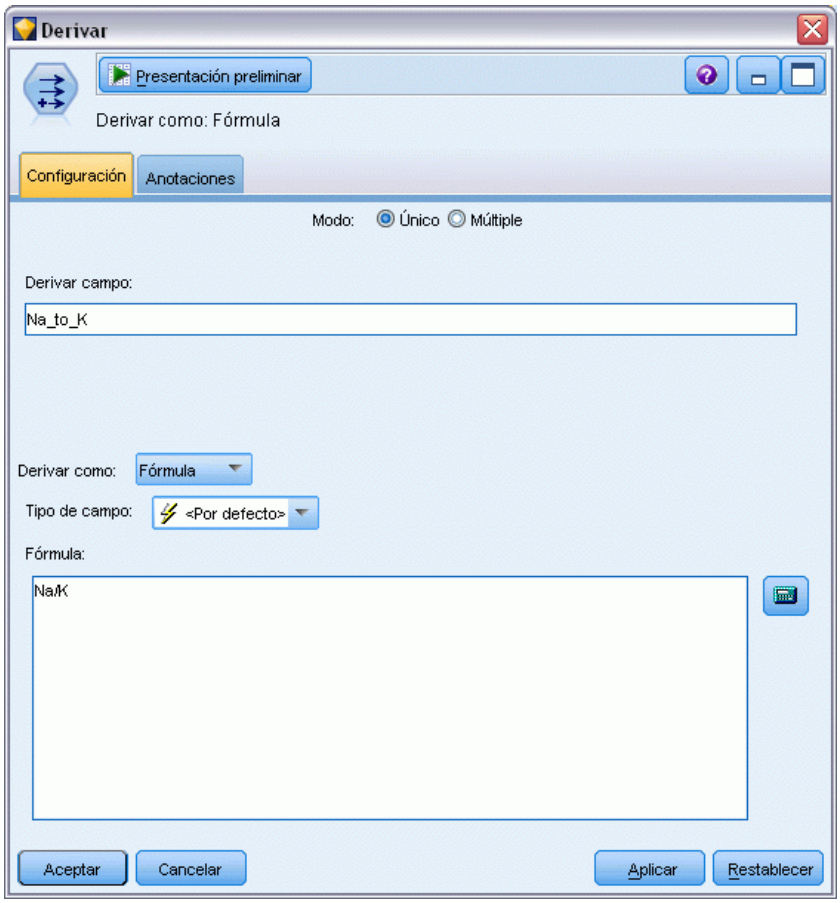
Figura 9-13
Gráfico de malla con el medicamento Y, y sus enlaces ocultos



Derivar un nuevo campo

Como el cociente de sodio-potasio parece que pronostica cuándo utilizar el medicamento *Y*, puede derivar un campo que contenga el valor de este cociente para cada registro. Este campo será de utilidad posteriormente cuando genere un modelo para pronosticar cuándo se debe utilizar cada uno de los cinco medicamentos. Para simplificar el diseño de rutas, comience eliminando todos los nodos excepto el nodo origen DRUG1n. Añada un nodo Derivar (pestaña Operaciones con campos) a DRUG1n, pulse dos veces en el nodo Derivar para editarlo.

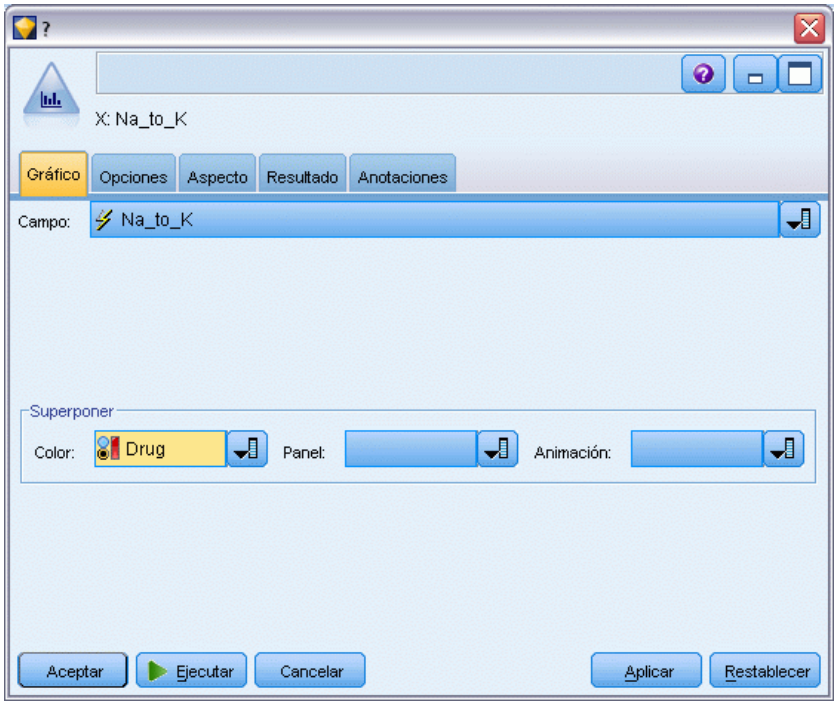
Figura 9-14
Edición del nodo Derivar



Asigne un nombre al nuevo campo *Na_to_K*. Como el nuevo campo se obtiene al dividir el valor de sodio por el valor de potasio, introduzca Na/K para la fórmula. También puede crear una fórmula pulsando en el icono situado a la derecha del campo. De esta forma se abre el Generador de expresiones, una forma de crear expresiones de forma interactiva mediante listas integradas de funciones, operandos y campos con sus valores.

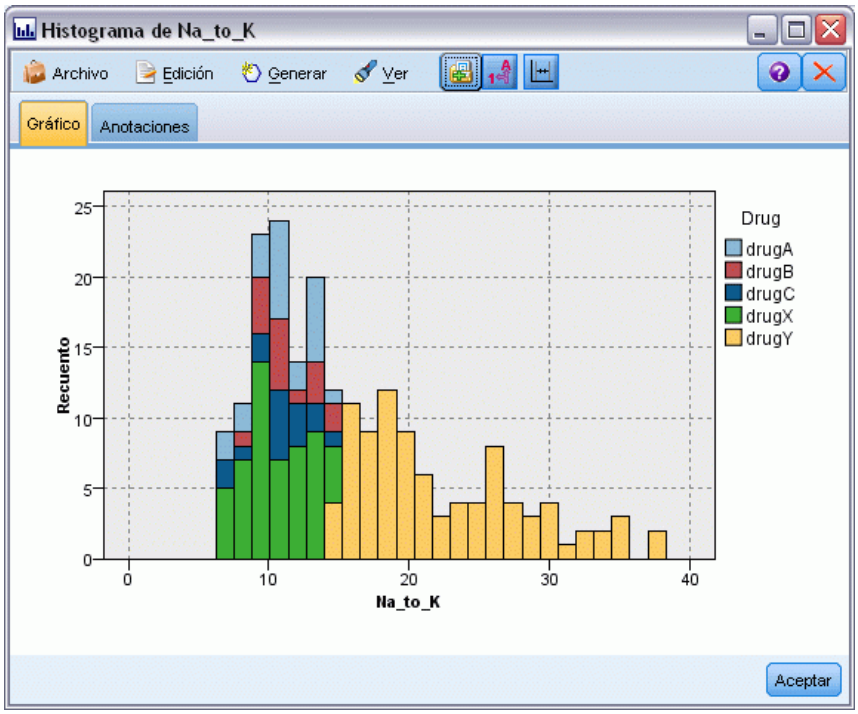
Puede comprobar la distribución del nuevo campo si añade un nodo Histograma al nodo Derivar. En el cuadro de diálogo del nodo Histograma, especifique *Na_to_K* como el campo que se va a representar y *Medicamento* como el campo de superposición.

Figura 9-15
Edición del nodo Histograma.



Cuando se ejecuta la ruta, se obtiene el siguiente gráfico. Según la presentación, se puede concluir que cuando el valor *Na_to_K* es aproximadamente 15 o mayor, el medicamento *Y* es el que se debe elegir.

Figura 9-16
Visualización del histograma

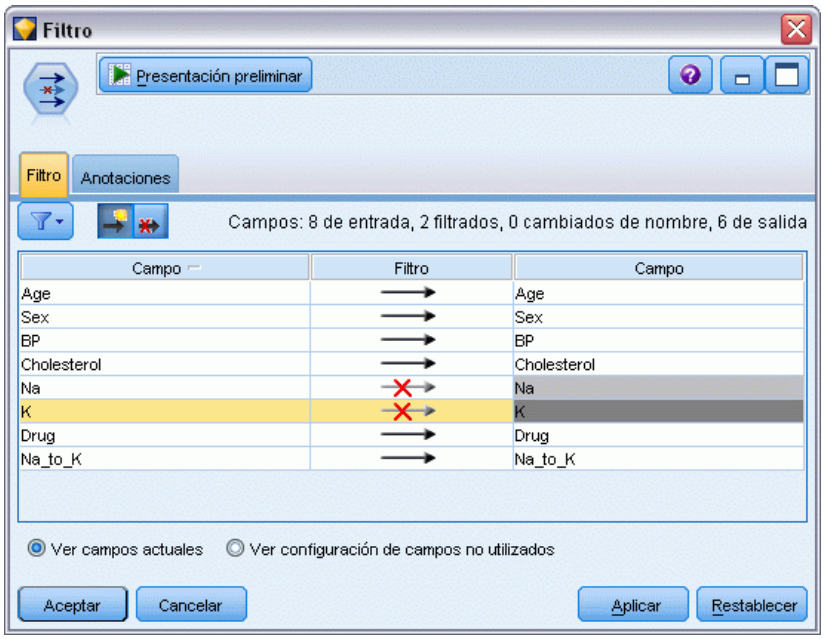


Generación de un modelo

Durante la exploración y manipulación de los datos, ha formulado algunas hipótesis. El cociente sodio-potasio en sangre parece influir en la elección del medicamento, al igual que la presión sanguínea. Sin embargo, aún no se pueden explicar todas las relaciones. Aquí es donde puede que el modelado nos dé la respuesta. En este caso, deberá intentar ajustar los datos mediante un modelo que crea reglas, el C5.0.

Como está utilizando un campo derivado, *Na_to_K*, puede filtrar para la salida los campos originales, *Na* y *K*, para que no se utilicen dos veces en el algoritmo de modelado. Puede hacerlo usando un nodo Filtro.

Figura 9-17
Edición del nodo Filtrar

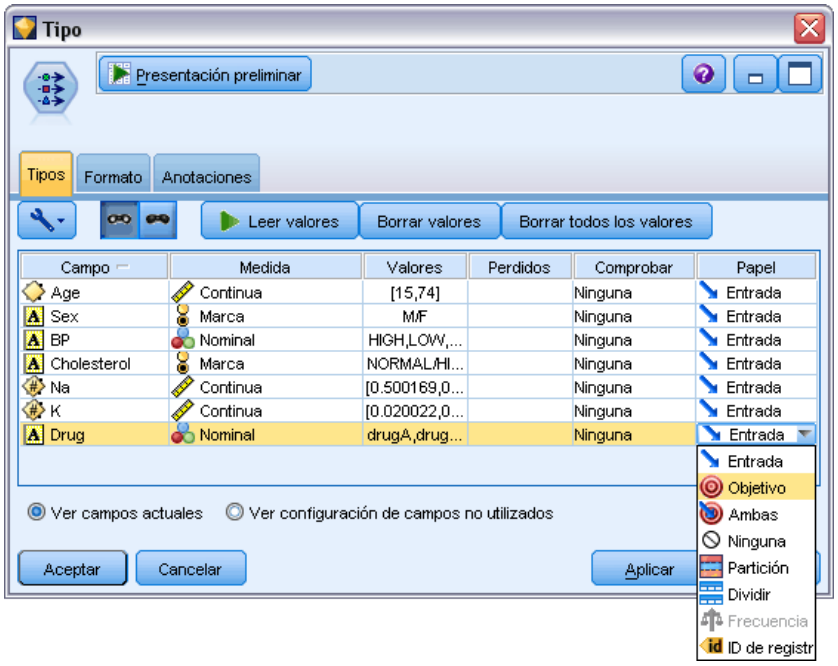


En la pestaña Filtro, pulse en las flechas situadas junto a *Na* y *K*. Aparecerá una X roja sobre cada flecha que indica que los campos están filtrados para la salida.

A continuación, conecte un nodo Tipo conectado al nodo Filtro. El nodo Tipo permite indicar los tipos de campos que está utilizando y cómo se utilizarán para pronosticar los resultados.

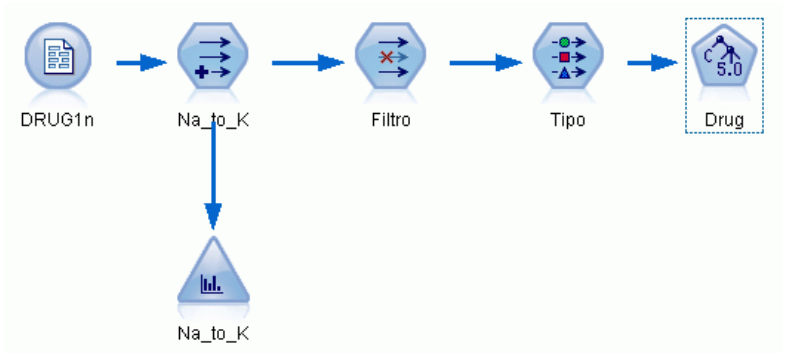
En la pestaña Tipos, defina el papel del campo *Medicamento* hacia Objetivo, lo cual indica que *Medicamento*es el campo que desea pronosticar. Deje el papel de los demás campos establecido como Entrada de forma que se utilicen como predictores.

Figura 9-18
Edición del nodo Tipo



Para estimar el modelo, coloque un nodo C5.0 en el espacio de trabajo y conéctelo al extremo de la ruta, como se muestra en la figura. A continuación, pulse el botón Ejecutar verde para ejecutar la ruta.

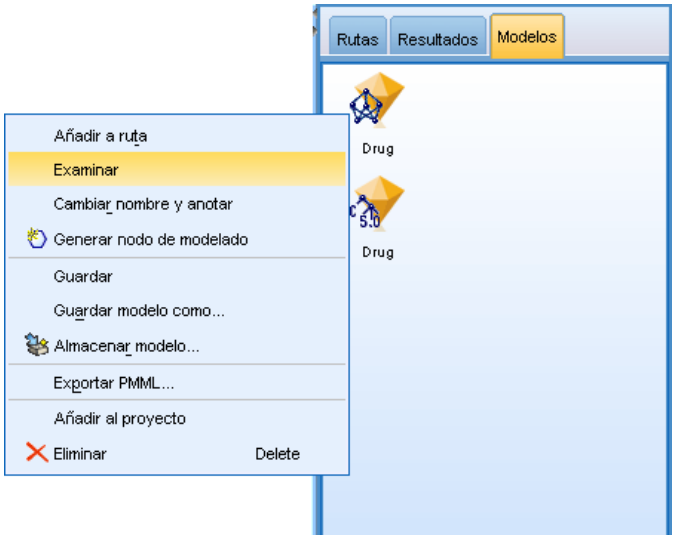
Figura 9-19
Adición de un nodo C5.0



Exploración del modelo

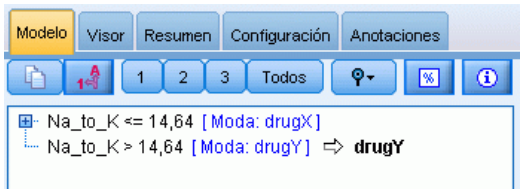
Cuando se ejecuta el nodo C5.0, el nugget del modelo se añade a la ruta y a la paleta Modelos en la esquina superior derecha de la ventana. Para examinar el modelo, pulse con el botón derecho del ratón en el icono y seleccione Editar o Examinar en el menú contextual.

Figura 9-20
Exploración del modelo



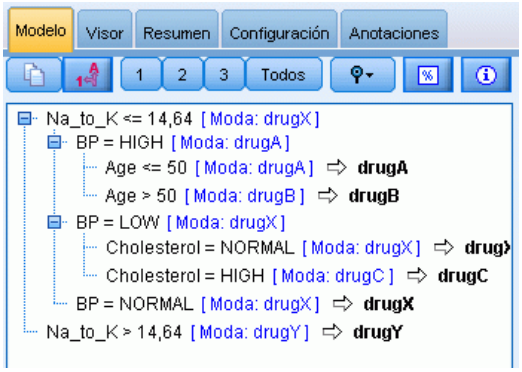
El examinador de reglas muestra el conjunto de reglas generadas por el nodo C5.0 en un formato de árbol de decisión. En un principio, el árbol está contraído. Para ampliarlo, pulse en el botón Todos para mostrar todos los niveles.

Figura 9-21
Examinador de reglas



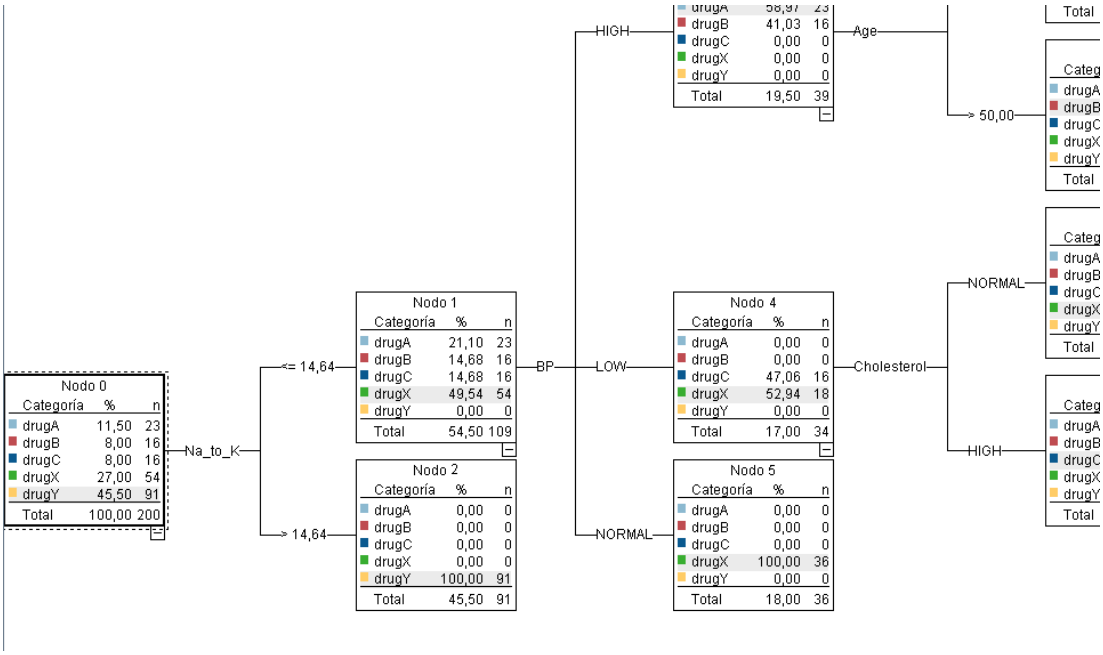
Ahora se muestran las piezas del rompecabezas que faltaban. Para aquellos sujetos con un cociente *Na* - *K* menor que 14.64 y alta presión sanguínea, la edad será la que determine la elección del medicamento. Para aquellos sujetos con una presión sanguínea baja, el colesterol parece ser el mejor predictor.

Figura 9-22
Examinador de reglas completamente expandido



El mismo árbol de decisión se puede ver en un formato gráfico más sofisticado si pulsa en la pestaña Visor. Aquí, se puede ver más fácilmente el número de casos para cada categoría de presión sanguínea, así como el porcentaje de casos.

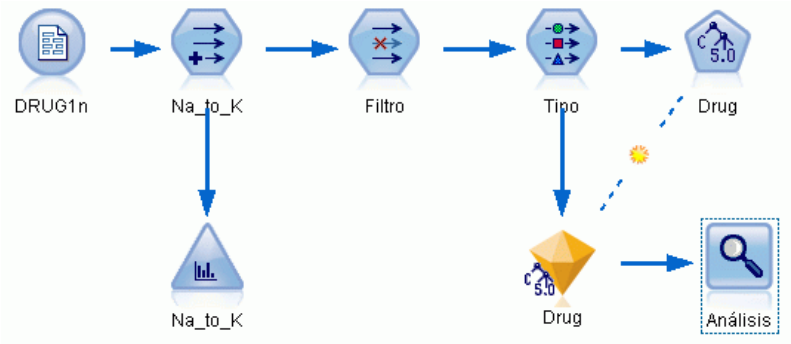
Figura 9-23
Árbol de decisión en formato gráfico



Utilización del nodo Análisis

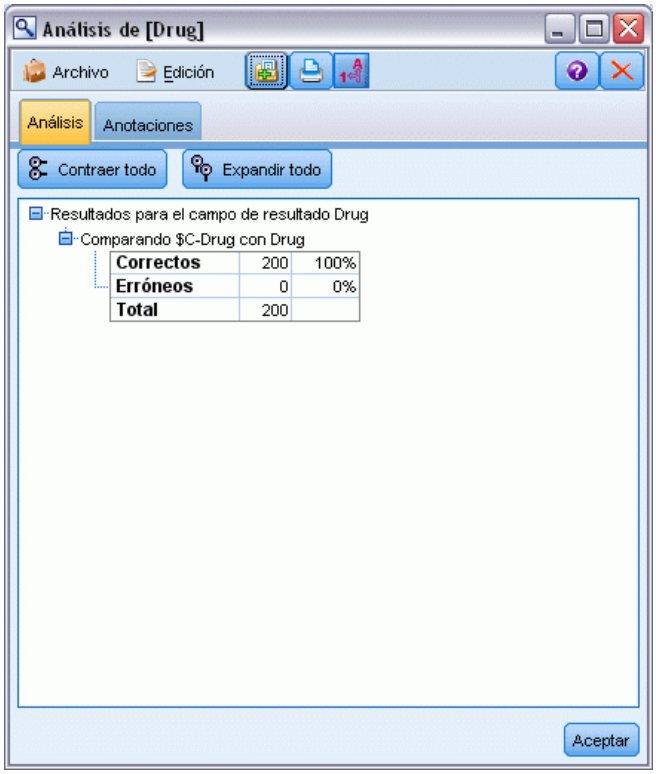
Se puede evaluar la precisión del modelo utilizando un nodo de análisis. Añada un nodo Análisis (de la paleta del nodo Resultado) al nugget de modelo, abra el nodo Análisis y pulse en Ejecutar.

Figura 9-24
Adición de un nodo Análisis



El resultado del nodo Análisis muestra que con este conjunto de datos artificial, el modelo ha pronosticado correctamente la elección del medicamento para todos los registros del conjunto de datos. Con un conjunto de datos real es poco probable ver una precisión del 100%, aunque puede utilizar el nodo Análisis para determinar si el modelo tiene una precisión aceptable para su aplicación en particular.

Figura 9-25
Resultado del nodo Análisis



Capítulo
10

Predictores de filtrado (Selección de características)

El nodo Selección de características le ayuda a identificar los campos que son más importantes para predecir determinados resultados. De un conjunto de cientos e incluso miles de predictores, el nodo Selección de características, filtra, ordena por rango y selecciona los predictores que pueden ser más importantes. En última instancia, puede lograr un modelo más eficaz y rápido, que utilice menos predictores, se ejecute de manera más rápida y sea más fácil de entender.

Los datos de este ejemplo representan los de un almacén de datos para una hipotética empresa de telefonía, y contiene información sobre las respuestas a una promoción especial de 5.000 clientes de la empresa. Los datos incluyen un gran número de campos que contienen los estadísticos del uso del teléfono, las edades de los clientes, el puesto de trabajo y los ingresos. Tres campos "objetivo" muestran si el cliente respondió a cada una de tres ofertas. La empresa desea utilizar estos datos para predecir qué clientes tienen más probabilidad de responder a ofertas similares en un futuro.

Este ejemplo utiliza la ruta denominada *featureselection.str*, que hace referencia al archivo de datos denominado *customer_dbase.sav*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *featureselection.str* se encuentra en el directorio *streams*.

Este ejemplo se centra solamente en una de las ofertas como objetivo. Utiliza el nodo de generación de árboles CHAID para desarrollar un modelo para describir qué clientes es más probable que respondan a la promoción. Contrasta dos enfoques:

- Sin selección de características. Todos los campos predictores del conjunto de datos se utilizan como entradas del árbol CHAID.
- Con selección de características. El nodo Selección de características se utiliza para seleccionar los 10 mejores predictores. Estos se introducen entonces en el árbol CHAID.

Comparando los dos modelos resultantes, podemos ver cómo la selección de características genera resultados más eficaces.

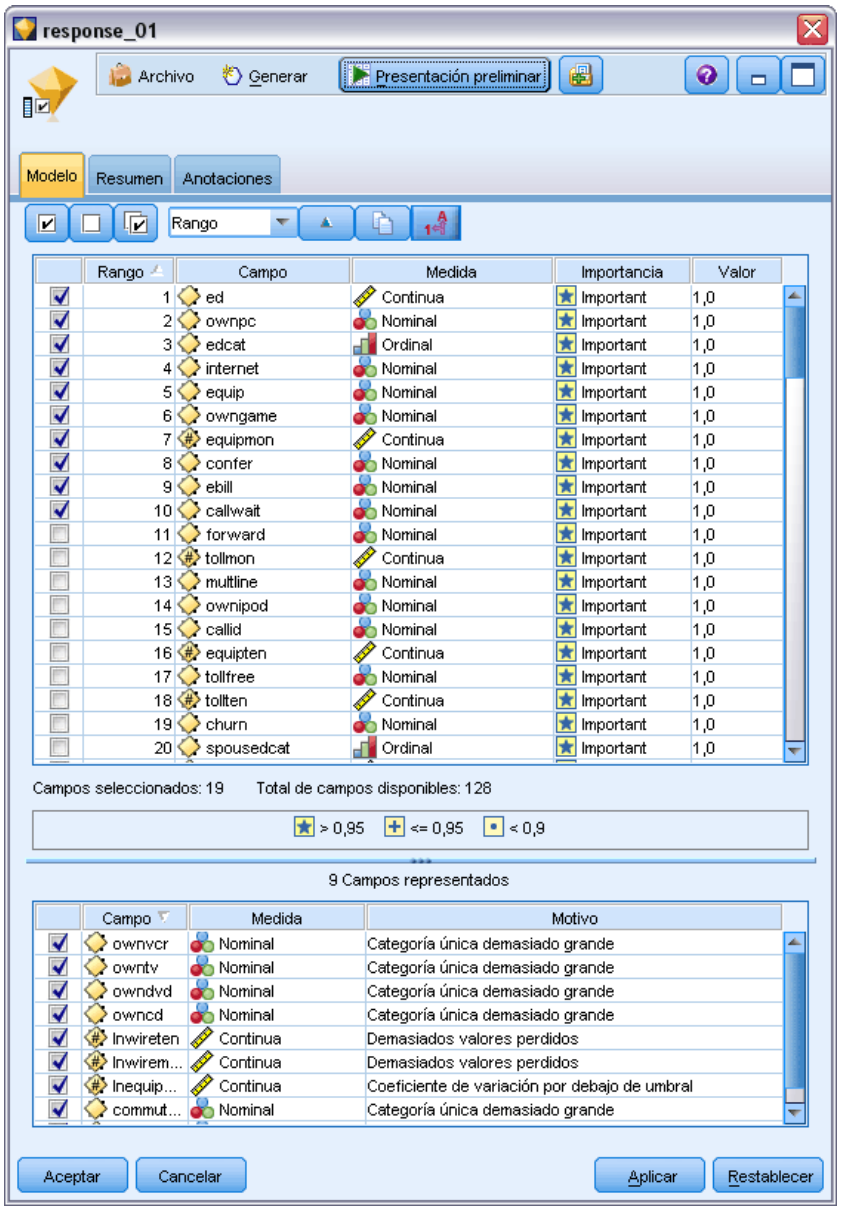
Generación de la ruta

Figura 10-1
Ruta de ejemplo de selección de características

- Añada un nodo de origen Archivo Statistics en un lienzo de rutas vacío. Apunte este nodo al archivo de datos de ejemplo *customer_dbase.sav*, que encontrará en la carpeta *Demos* dentro del directorio de instalación de IBM® SPSS® Modeler. (Si lo desea, abra el archivo de ruta de ejemplo *featureselection.str* en el directorio *streams*.)
- Adición de un nodo Tipo. En la pestaña Tipos, desplácese hasta la parte inferior y cambie el papel de *respuesta_01* a *Objetivo*. Cambie la función a *Ninguna* para el resto de campos de respuesta (*response_02* y *response_03*) y para la ID de cliente (*custid*) en la parte superior de la lista. Deje el papel definido a *Entrada* para los demás campos y pulse en el botón Leer valores; a continuación, pulse en Aceptar.
- Añada un nodo de modelado Selección de características a la ruta. En este nodo, puede especificar las reglas y criterios de los campos de filtrado o descalificación.
- Ejecute la ruta para generar el nugget de modelo de selección de características.

▶ Pulse con el botón derecho en el nugget de modelo en la ruta o en la paleta Modelos y seleccione Editar o Examinar para ver los resultados.

Figura 10-2
Pestaña Modelo en el nugget de modelo de selección de características

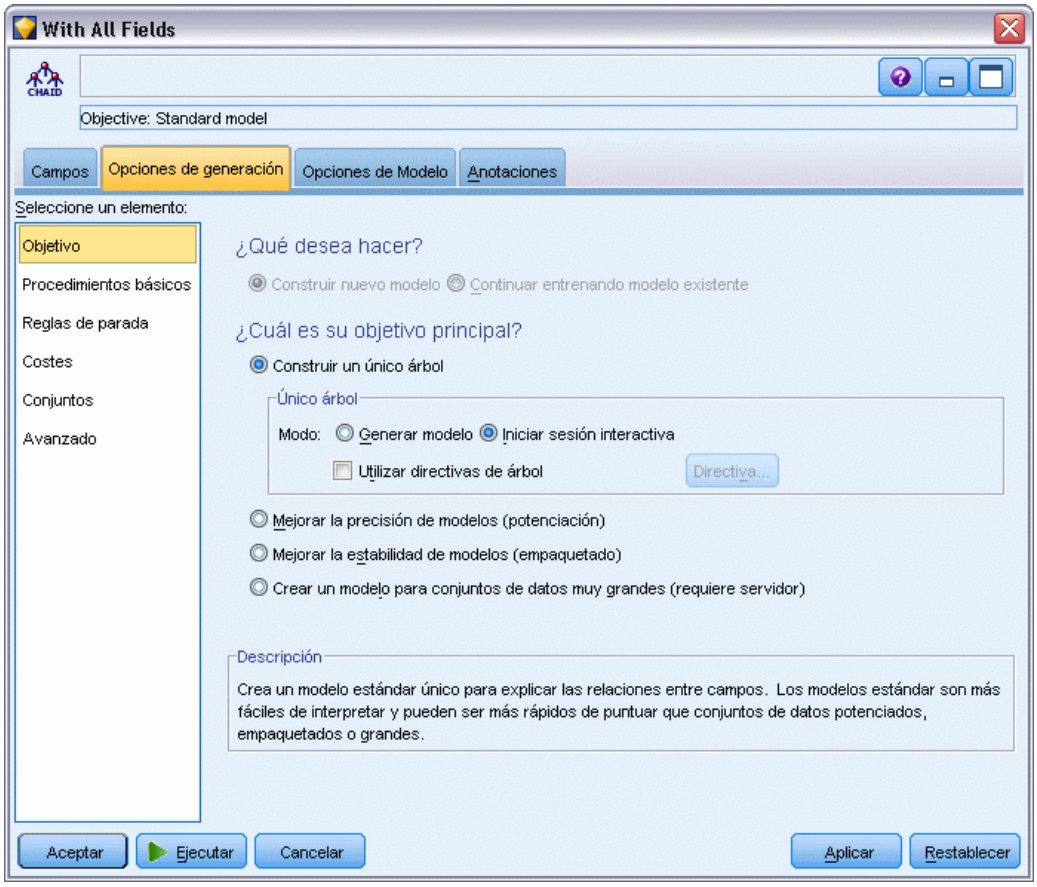


El panel superior muestra los campos que parecen ser útiles en la predicción. Se ordenan por rango según la importancia. El panel inferior muestra qué campos se han filtrado del análisis y por qué. Al examinar los campos del panel superior, es posible decidir cuáles se van a utilizar en las siguientes sesiones de modelado.

- Ahora se pueden seleccionar los campos que se utilizarán a continuación. Aunque al principio se identificaron como importantes 34 campos, queremos reducir el conjunto de predictores todavía más.
- Seleccione únicamente los 10 predictores principales con las marcas de revisión en la primera columna para cancelar la selección de los predictores que no desee. (Pulse en la marca de revisión de la fila 11, mantenga pulsada la tecla Mayús y pulse la marca de revisión de la fila 34.) Cierre el nugget de modelo.
- Para comparar los resultados sin la selección de características, debe añadir dos nodos de modelado CHAID a la ruta: uno que utilice la selección de características y otro que no la utilice.
- Añada un nodo CHAID al nodo Tipo y otro al modelo de selección de características.
- Abra cada nodo CHAID, seleccione la pestaña Opciones de generación y asegúrese de que las opciones Crear modelo nuevo, Crear un árbol único e Iniciar sesión interactiva se han seleccionado en el panel Objetivos.

  En el panel Básico, asegúrese de que Máxima profundidad de árbol se ha definido como 5.

Figura 10-3
Configuración de la pestaña Objetivos para el nodo de modelado CHAID para todos los campos de predictores

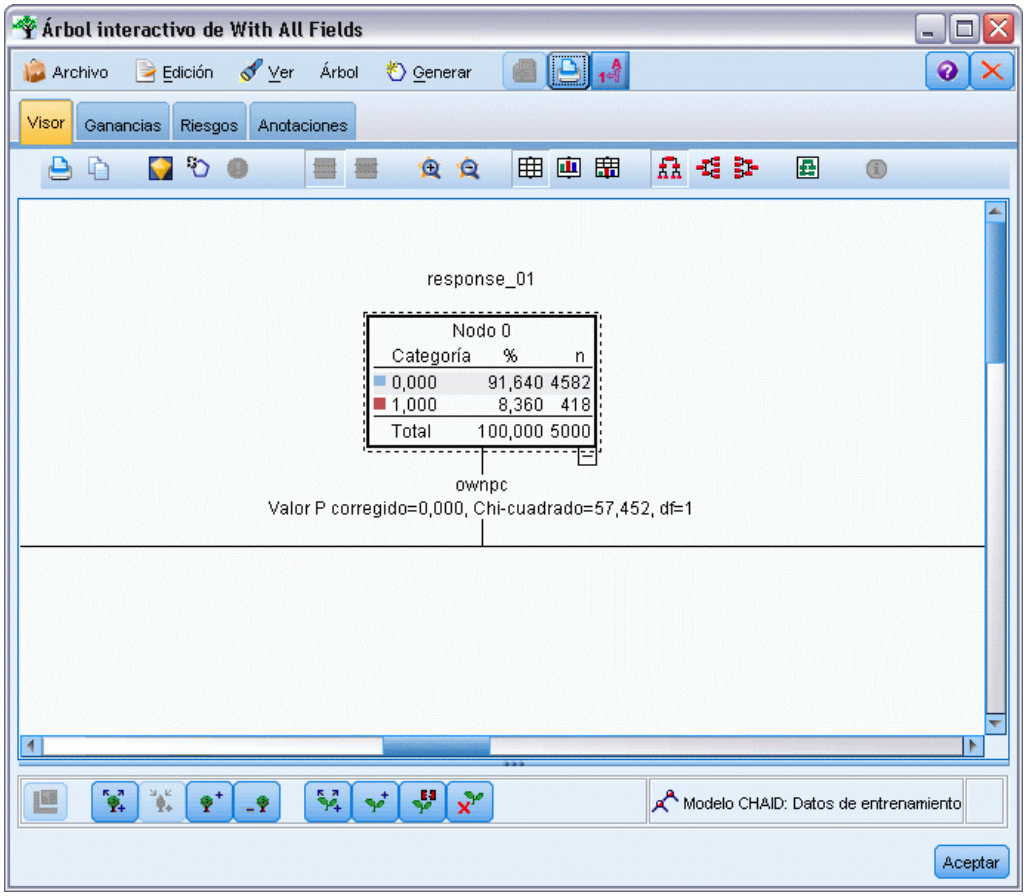


Generación de los modelos

- Ejecute el nodo CHAID que utiliza todos los predictores del conjunto de datos (el que se ha conectado al nodo Tipo). A medida que se ejecuta, observe cuánto tarda en ejecutarse. La ventana de resultados muestra una tabla.

- En los menús, seleccione Árbol > Hacer crecer árbol para ver el árbol expandido.

Figura 10-4
Crecimiento del árbol en el Generador de árboles



- Realice el mismo procedimiento para el otro nodo CHAID, que solamente utiliza 10 predictores. De nuevo, haga crecer el árbol cuando se abra el Generador de árboles.

  El segundo modelo debe haberse ejecutado más rápido que el primero. Como este conjunto de datos es relativamente pequeño, la diferencia en los tiempos de ejecución probablemente sea de unos pocos segundos; pero para conjuntos de datos reales de mayor tamaño esta diferencia puede ser considerablemente mayor, de minutos o incluso horas. Si se utiliza la selección de características, los tiempos de proceso se pueden reducir de manera significativa.

El segundo árbol también contiene menos nodos que el primero. Resulta más fácil de entender. Pero antes de decidir utilizarlo, deberá averiguar si es eficaz y cómo se compara respecto al modelo que utiliza todos los predictores.

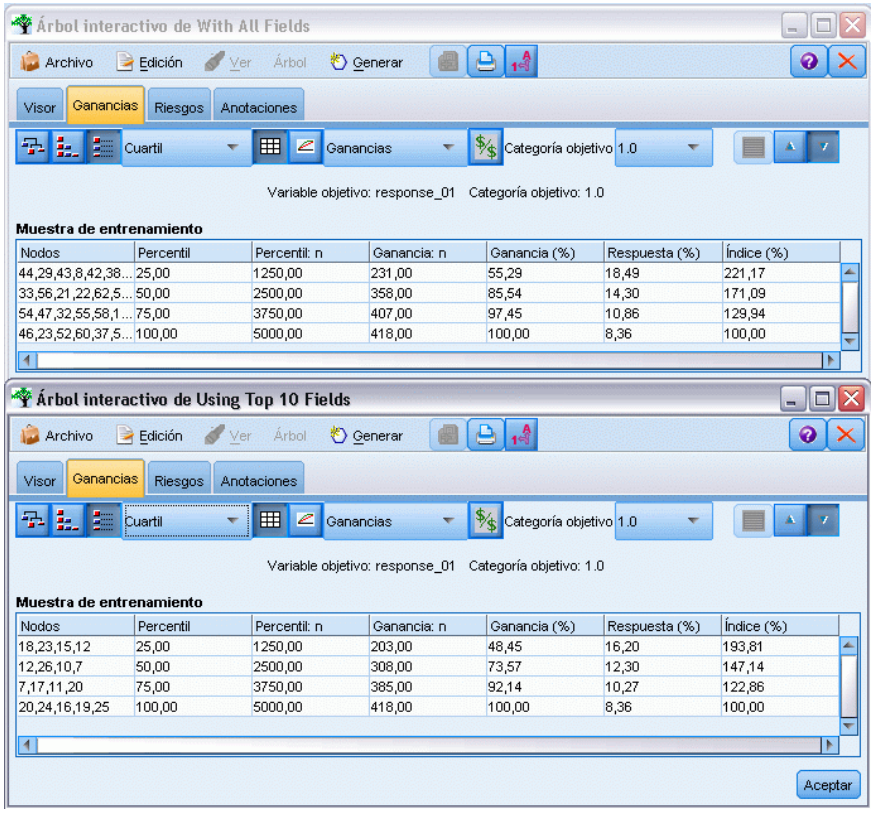
Comparación de los resultados

Para comparar los dos resultados, necesitamos una medida de la eficacia. Para ello, podemos recurrir a la pestaña Ganancias del Generador de árboles. Miraremos en **elevación**, que mide la probabilidad de que los registros de un nodo correspondan a la categoría objetivo si se comparan con todos los registros del conjunto de datos. Por ejemplo, un valor de elevación de 148% indica que la probabilidad de los registros del nodo de corresponder a la categoría objetivo es 1,48 veces mayor en relación con todos los registros del conjunto de datos. La elevación se indica en la columna *Índice* de la pestaña Ganancias.

- En el Generador de árboles para el conjunto completo de predictores, pulse en la pestaña Ganancias. Cambie la categoría objetivo a 1,0. Cambie la visualización a cuartiles pulsando en el botón Cuantiles de la barra de herramientas. A continuación seleccione Cuartil en la lista desplegable a la derecha del botón.

- Repita este procedimiento en el Generador de árboles para el conjunto de los 10 predictores de manera que pueda tener dos tablas similares Ganancias para comparar, como se muestra en las siguientes figuras.

Figura 10-5
Gráficos de ganancias para los dos modelos CHAID



Cada tabla de ganancias agrupa los nodos terminales para su árbol en cuartiles. Para comparar la eficacia de los dos modelos, mire el elevador (valor *Índice*) para el cuartil superior de cada tabla.

Cuando se incluyen todos los predictores, el modelo muestra una elevación de 221%. Esto significa que la probabilidad de los casos con las características de estos nodos de responder a la promoción objetivo es 2,2 veces mayor. Para ver cuáles son estas características, pulse para seleccionar la fila superior. Cambie a la pestaña Visor, donde los nodos correspondientes están resaltados en negro. Siga el árbol hacia abajo hasta cada nodo terminal resaltado para ver cómo se dividen los predictores. El cuartil superior solo, incluye 10 nodos. Al convertirse en modelos de puntuación reales, puede ser difícil gestionar 10 perfiles de cliente.

Con solamente los 10 mejores predictores incluidos (como se identifica en la selección de características), la elevación es de casi 194%. Aunque este modelo no es tan bueno como el que utiliza todos los predictores, resulta útil. Y aquí el cuartil superior incluye solamente 4 nodos, de manera que es más simple. Por tanto, es posible determinar que el modelo de selección de características es preferible al que tiene todos los predictores.

Resumen

Revisemos las ventajas de la selección de características. Utilizar menos predictores resulta más barato. Significa que tiene menos datos que recopilar, procesar y rellenar en los modelos. Y el tiempo de cálculo se reduce. En este ejemplo, aun con el paso adicional de selección de características, la creación de modelo fue mucho más rápida con el conjunto de predictores más pequeño. Con un conjunto de datos real de mayor tamaño, los ahorros de tiempo se incrementarán significativamente.

Al utilizar menos predictores, la puntuación es más simple. En el ejemplo puede identificar solamente 4 perfiles de clientes que probablemente respondan a la promoción. Tenga en cuenta que con números mayores de predictores, corre el riesgo de sufrir sobreajustes en su modelo. El modelo más simple puede generalizar mejor en otros conjuntos de datos (aunque necesita comprobarlo).

Podría haber utilizado un algoritmo de generación de árboles para realizar el trabajo de selección de características, permitiendo al árbol que identificara automáticamente los predictores más importantes. De hecho, el algoritmo CHAID se utiliza a menudo para este objetivo y es incluso posible hacer crecer el árbol nivel por nivel para controlar su profundidad y complejidad. Sin embargo, el nodo Selección de características es más rápido y fácil de utilizar. Ordena por rango todos los predictores en un paso rápido, para que pueda identificar rápidamente los campos más importantes. Permite modificar el número de predictores que va a incluir. Podría ejecutar fácilmente este ejemplo de nuevo utilizando los 15 ó 20 mejores predictores en lugar de 10, comparando los resultados para determinar el modelo óptimo.

Capítulo
11

Reducción de la longitud de cadena de datos de entrada (Nodo Reclasificar)

Reducción de la longitud de cadena de datos de entrada (Reclasificar)

Para los modelos de regresión logística binomial y de clasificador automático que incluyen un modelo de regresión logística binomial generado, los campos de cadena están limitados a un máximo de ocho caracteres. Si las cadenas tiene más de ocho caracteres, se pueden registrar utilizando un nodo Reclasificar.

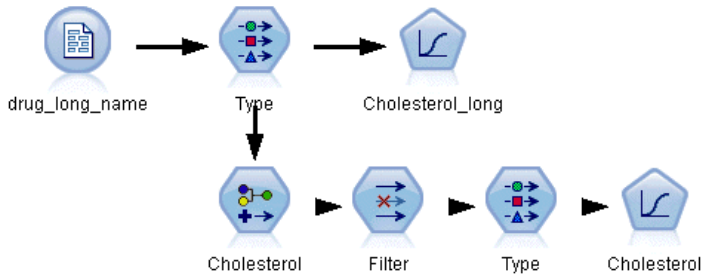
Este ejemplo utiliza la ruta denominada *reclassify_strings.str*, que hace referencia al archivo de datos denominado *drug_long_name*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *reclassify_strings.str* se encuentra en el directorio *streams*.

Este ejemplo se centra en una pequeña parte de una ruta para mostrar el orden de los errores que se pueden generar con cadenas más largas y explica cómo utilizar el nodo Reclasificar para cambiar los detalles de cadena a una longitud aceptable. Aunque el ejemplo utiliza un nodo Regresión logística binomial, es igualmente aplicable si utiliza el nodo Clasificador automático para generar un modelo de regresión logística binomial.

Reclasificación de los datos

- Si utiliza en nodo de origen Archivo variable, conéctelo a conjunto de datos *drug_long_name* en la carpeta *Demos*.

Figura 11-1
Ruta de ejemplo con reclasificación de cadena para regresión logística binomial



- Añada un nodo Tipo al nodo de origen y seleccione Colesterol_alto como objetivo.
- Añada un nodo Regresión logística al nodo Tipo.

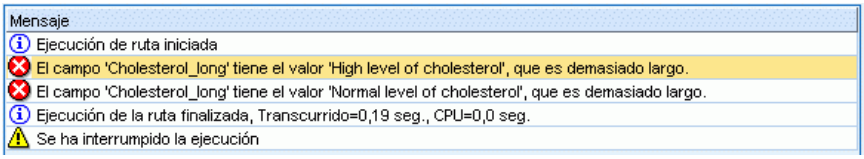
- En el nodo Regresión logística, pulse en la pestaña Modelo y seleccione el procedimiento Binomial.

Figura 11-2
Detalles de cadena larga en el campo "Colesterol_alto"

- Si ejecuta el nodo Regresión logística en *reclassify_strings.str*, aparecerá un mensaje de error advirtiéndole que los valore de la cadena Colesterol_alto son demasiado largos.

  Si encuentra este tipo de mensaje de error, realice el procedimiento que se explica a continuación para modificar los datos.

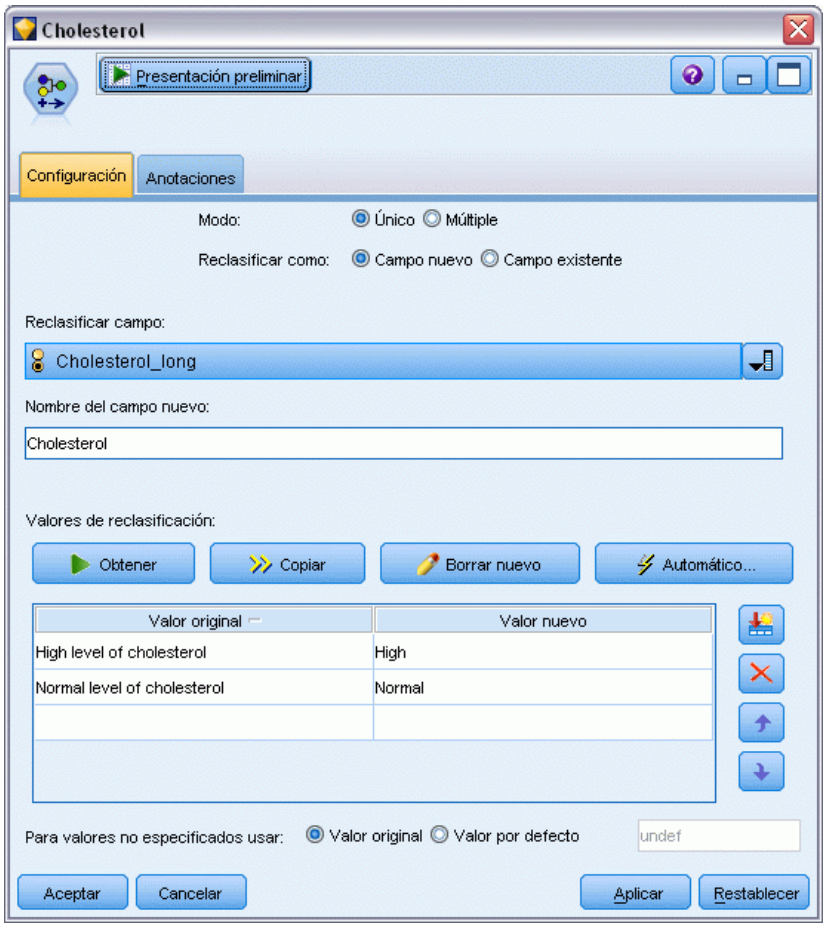
Figura 11-3
Visualización del mensaje de error cuando ejecuta el nodo de regresión logística binomial



- Añada un nodo Reclasificar al nodo Tipo.
- En el campo Reclasificar, seleccione Colesterol_alto.
- Introduzca Colesterol como el nuevo nombre del campo.
- Pulse en el botón Obtener para añadir los valores de Colesterol_alto a la columna del valor original.

- En la columna del nuevo valor, introduzca Alto junto al valor original de Alto nivel del colesterol y Normal junto al valor original de Nivel normal de colesterol.

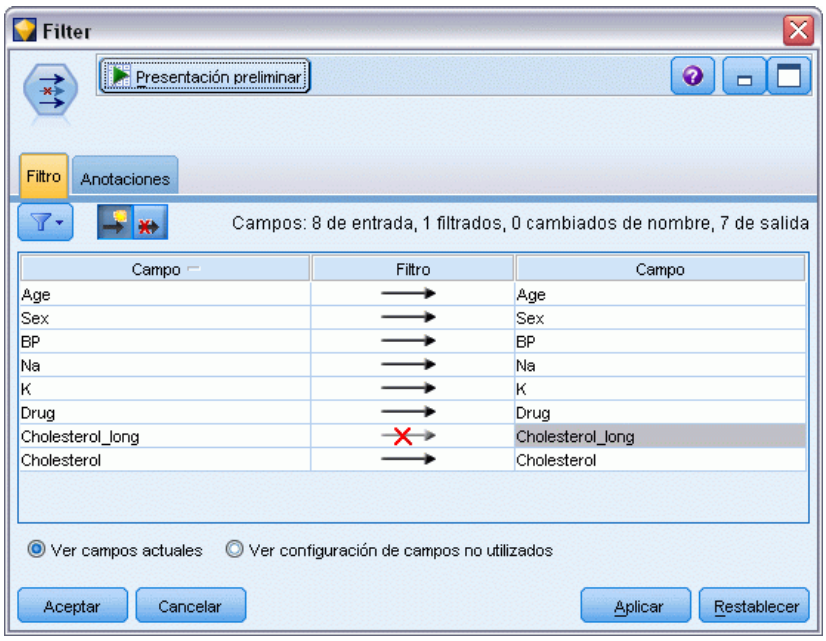
Figura 11-4
Reclasificación de cadenas largas



- Añada un nodo Filtro al nodo Reclasificar.

▶ En la columna Filtro, pulse para eliminar Colesterol_alto.

Figura 11-5
Filtrado del campo "Colesterol_alto" de los datos



▶ Añada un nodo de tipo al nodo Filtro y seleccione Colesterol como objetivo.

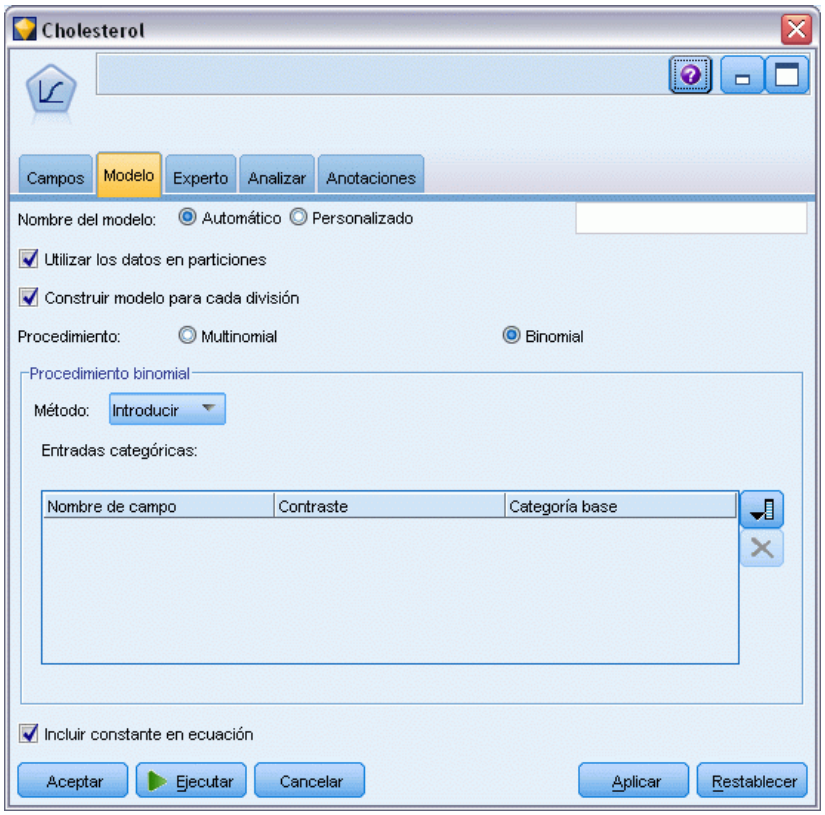
Figura 11-6
Detalles de cadena corta en el campo "Colesterol"

▶ Añada un nodo Logística al nodo Tipo.

▶ En el nodo Logística, pulse en la pestaña Modelo y seleccione el procedimiento Binomial.

▶ Ahora puede ejecutar el nodo Logística binomial y genere un modelo sin que aparezca un mensaje de error.

Figura 11-7
Selección del procedimiento binomial



Este ejemplo sólo muestra una parte de una ruta. Si necesita más información sobre los tipos de rutas en las que necesita reclasificar cadenas largas, los ejemplos siguientes están disponibles:

- Nodo Clasificador automático. Si desea obtener más información, consulte el tema Modelado de respuesta de clientes (clasificador automático) en el capítulo 5 el p. 44.
- Nodo Regresión logística binomial. Si desea obtener más información, consulte el tema Pérdida de clientes de telecomunicaciones (Regresión logística binomial) en el capítulo 14 el p. 159.

Existe más información acerca del uso de IBM® SPSS® Modeler, como una guía de usuario, referencia de nodo y guía de algoritmos, disponible en el directorio *\Documentation* del disco de instalación.

Parte III:
Ejemplos de modelado

Capítulo
12

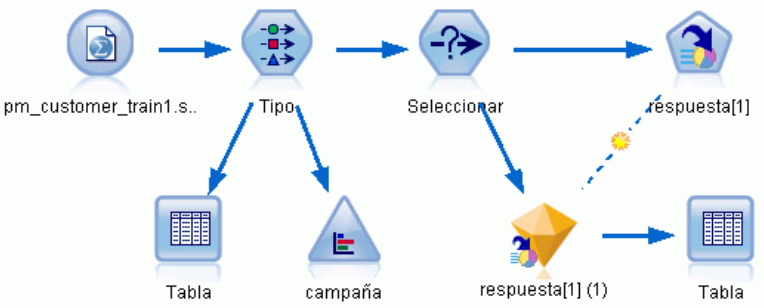
Modelado de respuesta de clientes (Lista de decisiones)

El algoritmo Lista de decisiones genera reglas que indican una mayor o menor probabilidad de obtener cierto resultado binario (sí o no). Los modelos de listas de decisiones se utilizan con frecuencia en la gestión de relaciones con los clientes, incluidos los centros de llamadas y las aplicaciones de marketing.

Este ejemplo se basa en una empresa ficticia que desea obtener resultados más rentables en las futuras campañas de marketing adaptando la oferta adecuada a cada cliente. En el ejemplo se utiliza un modelo de lista de decisiones para identificar las características de los clientes que es más probable que respondan favorablemente, teniendo en cuenta las promociones anteriores, y generar una lista de mailing a partir de estos resultados.

Los modelos de lista de decisión son especialmente adecuados para el modelo interactivo, permitiéndole ajustar los parámetros en el modelo e, inmediatamente, ver los resultados. Puede utilizar el nodo Clasificador automático como un método diferente que le permita crear automáticamente un número de modelos diferentes y ordenar los resultados.

Figura 12-1
Ejemplo de ruta de Lista de decisiones

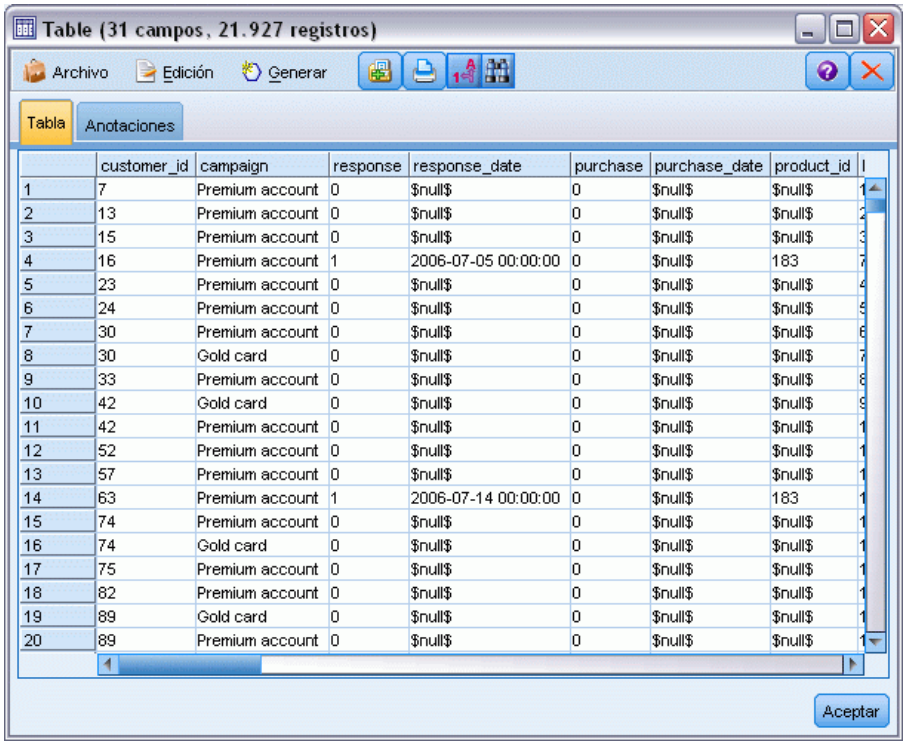


Este ejemplo utiliza la ruta denominada *pm_decisionlist.str*, que hace referencia al archivo de datos *pm_customer_train1.sav*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *pm_decisionlist.str* se encuentra en el directorio *streams*.

Datos históricos

El archivo *pm_customer_train1.sav* contiene datos históricos en los que se registran las ofertas realizadas a determinados clientes en campañas anteriores, según indica el valor del campo *campaña*. El mayor número de registros corresponden a la campaña *Cuenta principal*.

Figura 12-2
Datos sobre promociones anteriores



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	l
1	7	Premium account	0	$null$	0	$null$	$null$	1
2	13	Premium account	0	$null$	0	$null$	$null$	2
3	15	Premium account	0	$null$	0	$null$	$null$	3
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183	7
5	23	Premium account	0	$null$	0	$null$	$null$	4
6	24	Premium account	0	$null$	0	$null$	$null$	5
7	30	Premium account	0	$null$	0	$null$	$null$	6
8	30	Gold card	0	$null$	0	$null$	$null$	7
9	33	Premium account	0	$null$	0	$null$	$null$	8
10	42	Gold card	0	$null$	0	$null$	$null$	9
11	42	Premium account	0	$null$	0	$null$	$null$	1
12	52	Premium account	0	$null$	0	$null$	$null$	1
13	57	Premium account	0	$null$	0	$null$	$null$	1
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183	1
15	74	Premium account	0	$null$	0	$null$	$null$	1
16	74	Gold card	0	$null$	0	$null$	$null$	1
17	75	Premium account	0	$null$	0	$null$	$null$	1
18	82	Premium account	0	$null$	0	$null$	$null$	1
19	89	Gold card	0	$null$	0	$null$	$null$	1
20	89	Premium account	0	$null$	0	$null$	$null$	1

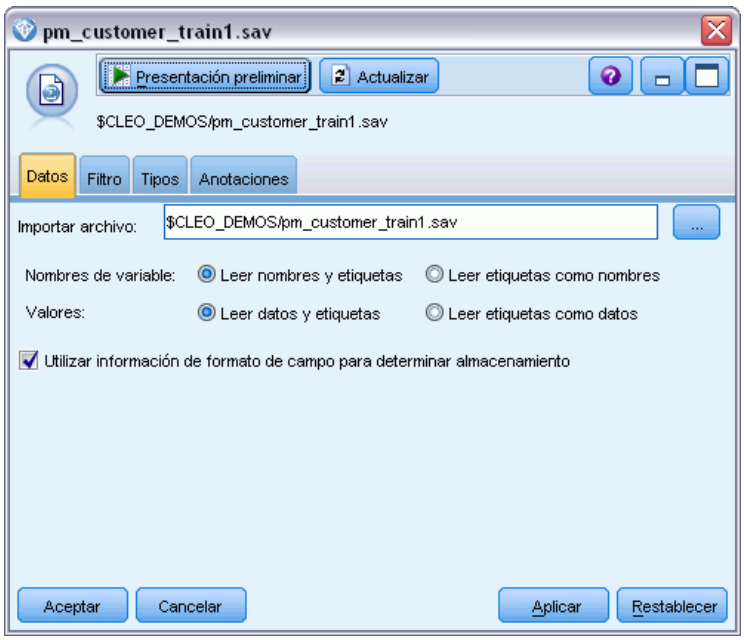
Los valores del campo *campaña* aparecen codificados como enteros en los datos, con etiquetas definidas en el nodo Tipo (por ejemplo, *2 = Cuenta principal*). Puede activar o desactivar la visualización de las etiquetas de valor en la tabla utilizando la barra de herramientas.

El archivo también incluye varios campos que contienen información demográfica y financiera acerca de cada uno de los clientes, que se puede utilizar para generar o "entrenar" un modelo que pronostique los índices de respuesta de diferentes grupos según determinadas características.

Generación de la ruta

- Añada un nodo de Archivo Statistics que apunte a *pm_customer_train1.sav*, ubicado en la carpeta *Demos* de la instalación de IBM® SPSS® Modeler. (Puede especificar `$CLEO_DEMOS/` en la ruta del archivo como acceso directo a referencia de esta carpeta.)

Figura 12-3
Lectura de datos mezclados



▶ Añada un nodo Tipo y seleccione *respuesta* como campo objetivo (Papel = Objetivo). Defina el nivel de medición de este campo como Marca.

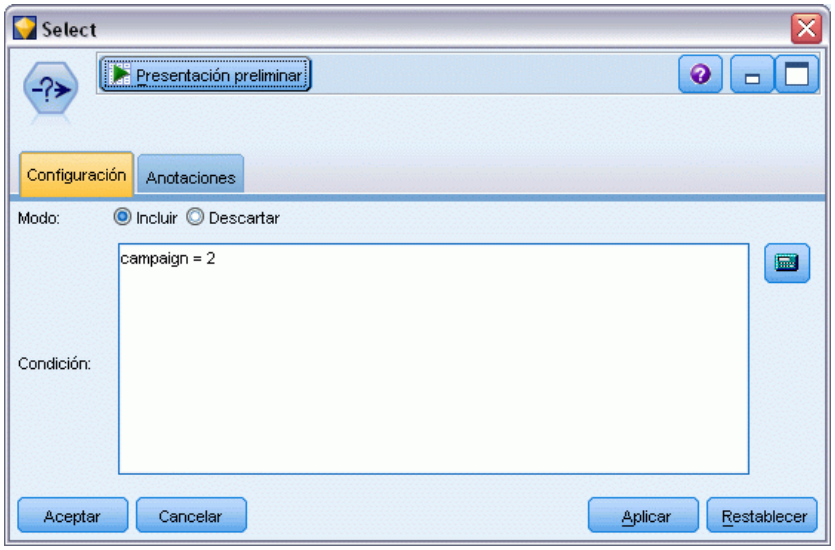
Figura 12-4
Definición del nivel de medición y el papel

▶ Establezca el papel en Ninguno para los siguientes campos: *id_cliente*, *campaña*, *fecha_respuesta*, *compra*, *fecha_compra*, *id_producto*, *Idfila* y *X_aleatorio*. Todos estos campos tienen su utilidad en los datos, pero no se utilizarán para generar el modelo real.

▶ Pulse en el botón Leer valores del nodo Tipo para asegurarse de que se crea una instancia de los valores.

Aunque los datos incluyen información acerca de cuatro campañas diferentes, el análisis lo realizaremos campaña a campaña. Como el mayor número de registros corresponden a la campaña Premium (codificada como *campaign=2* en los datos), puede utilizar un nodo Seleccionar para incluir únicamente dichos registros en la ruta.

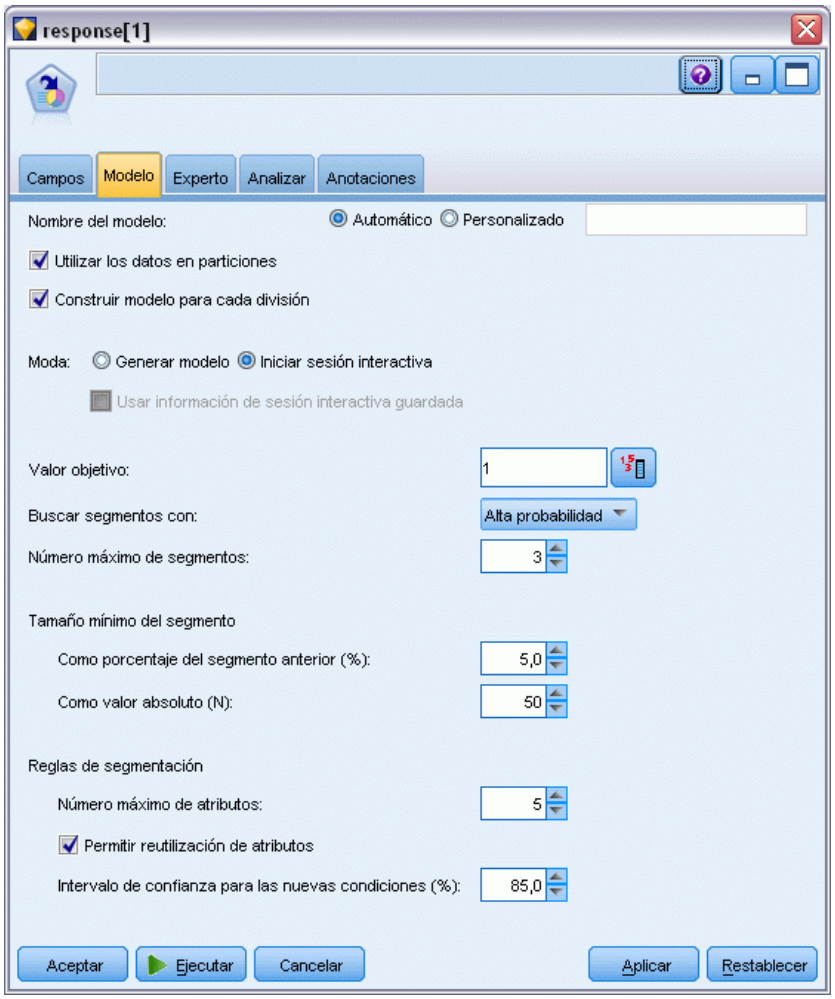
Figura 12-5
Selección de los registros correspondientes a una única campaña



Creación del modelo

- Añada un nodo Lista de decisiones a la ruta. En la pestaña Modelo, defina el valor objetivo como 1 para indicar el resultado que se desea buscar. En este caso, buscará clientes que hayan contestado *Sí* a una oferta anterior.

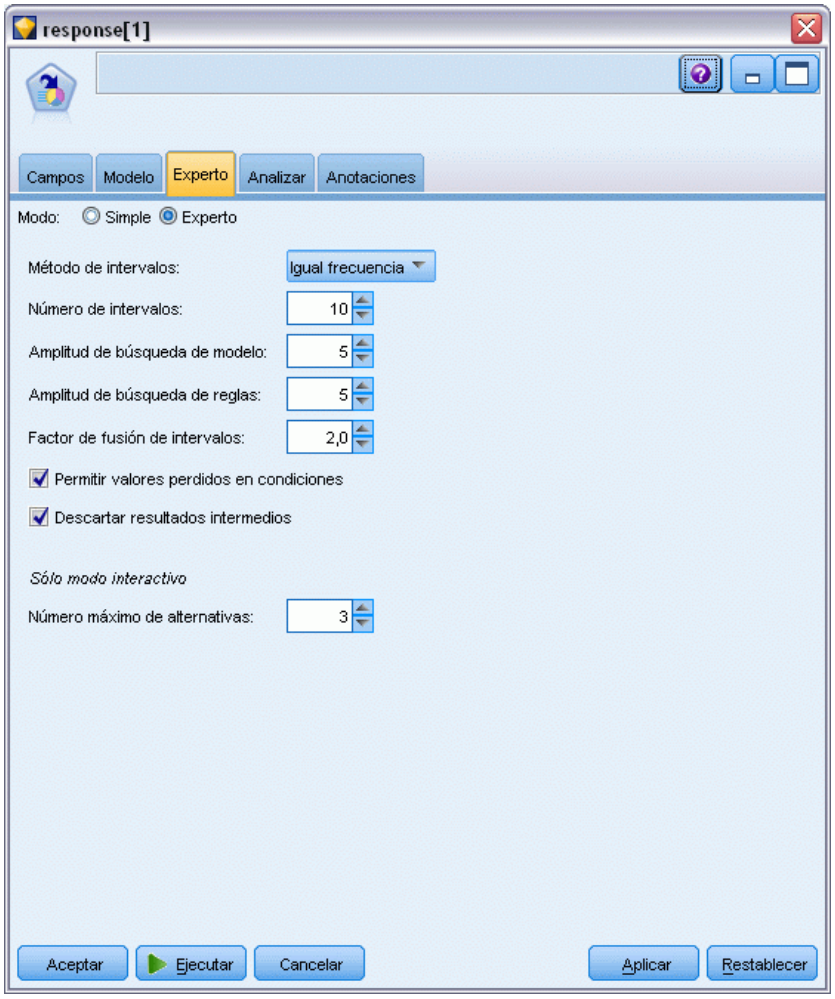
Figura 12-6
Nodo Lista de decisiones, pestaña Modelo



- Seleccione Iniciar sesión interactiva.
- Para no complicar el modelo para este ejemplo, estableceremos el número máximo de segmentos en 3.
- Cambie el intervalo de confianza de las nuevas condiciones al 85%.

▶ En la pestaña Experto, defina Modo a Experto.

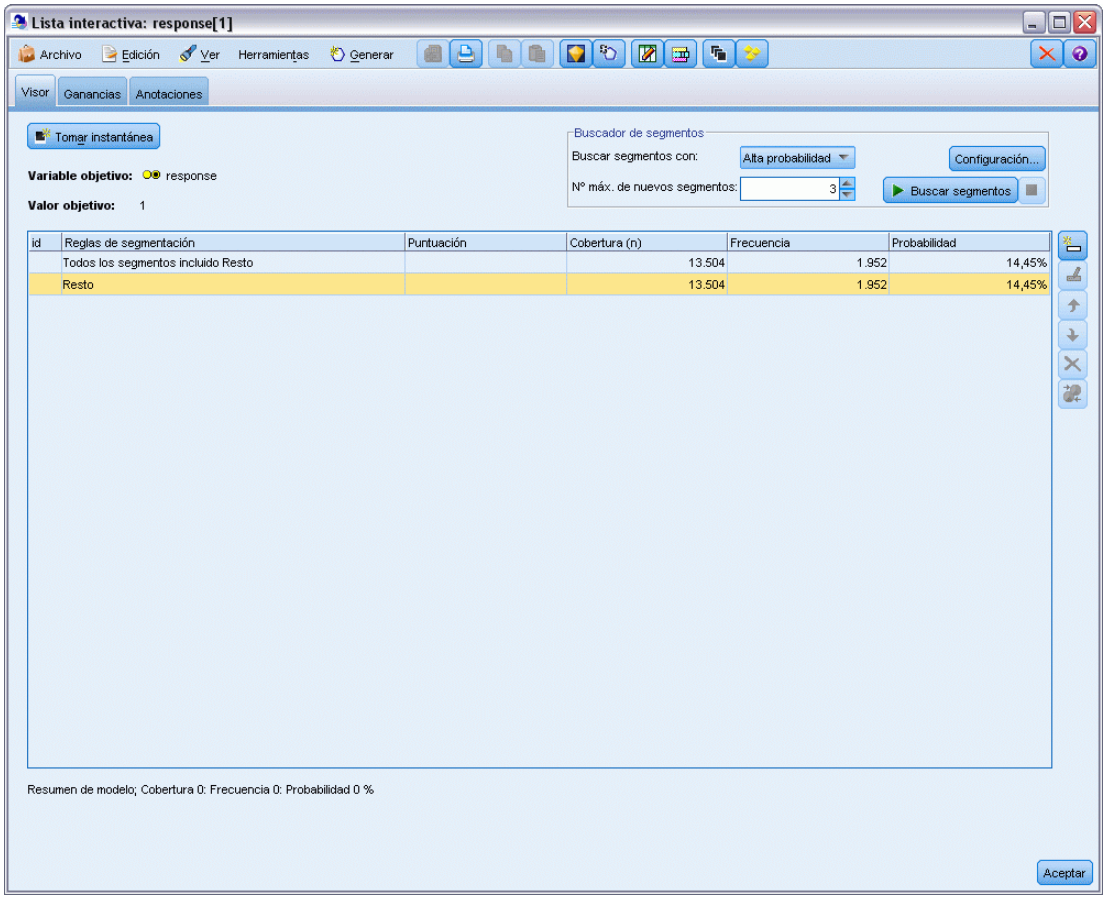
Figura 12-7
Nodo Lista de decisiones, pestaña Experto



▶ Aumente Número máximo de alternativas a 3. Esta función funciona junto con el ajuste Iniciar sesión interactiva que ha seleccionado en la pestaña Modelo.

▶ Pulse en Ejecutar para mostrar el visor de listas interactivas.

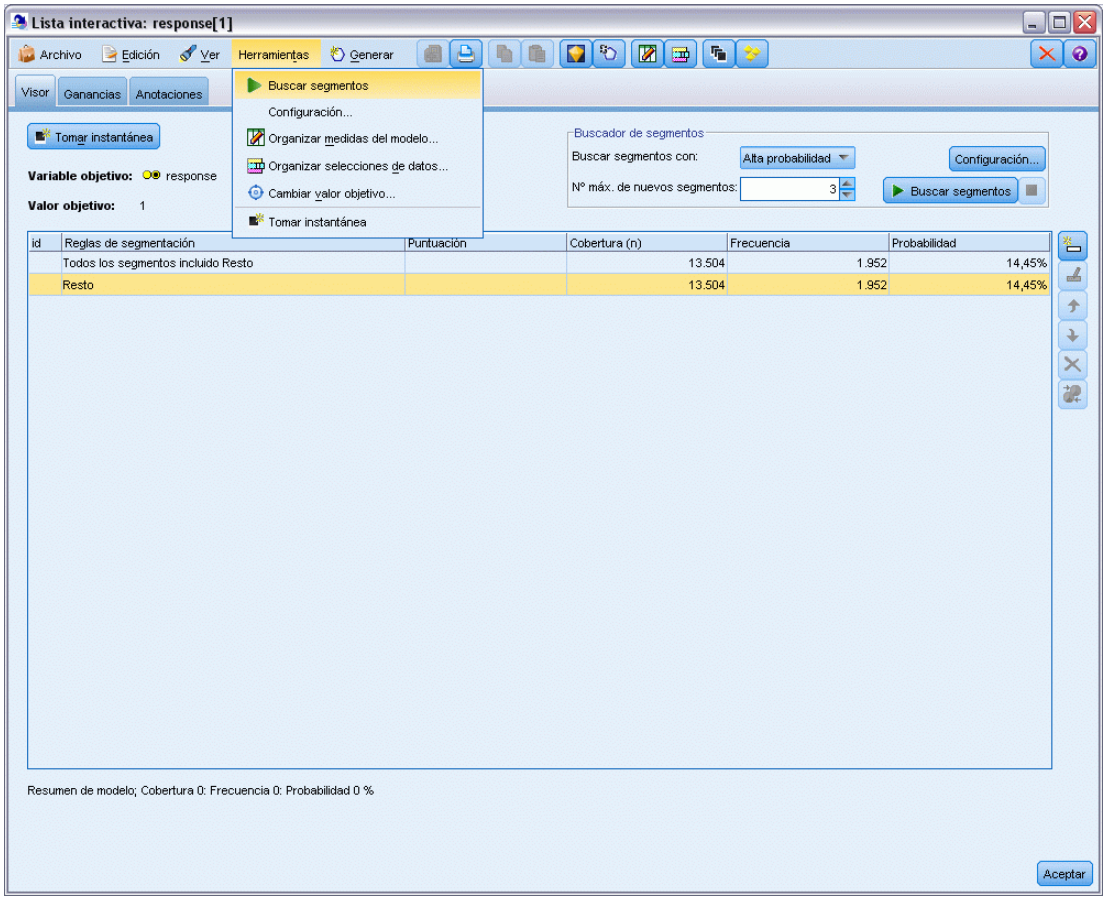
Figura 12-8
Visor de listas interactivas



Como todavía no se ha definido ningún segmento, todos los registros se encuentran en el resto. De los 13.504 registros de la muestra, 1.952 respondieron *Sí*, lo que supone una tasa de aciertos global del 14,45%. Para mejorar esta tasa, identificaremos segmentos de clientes con más (o menos) probabilidad de dar una respuesta favorable.

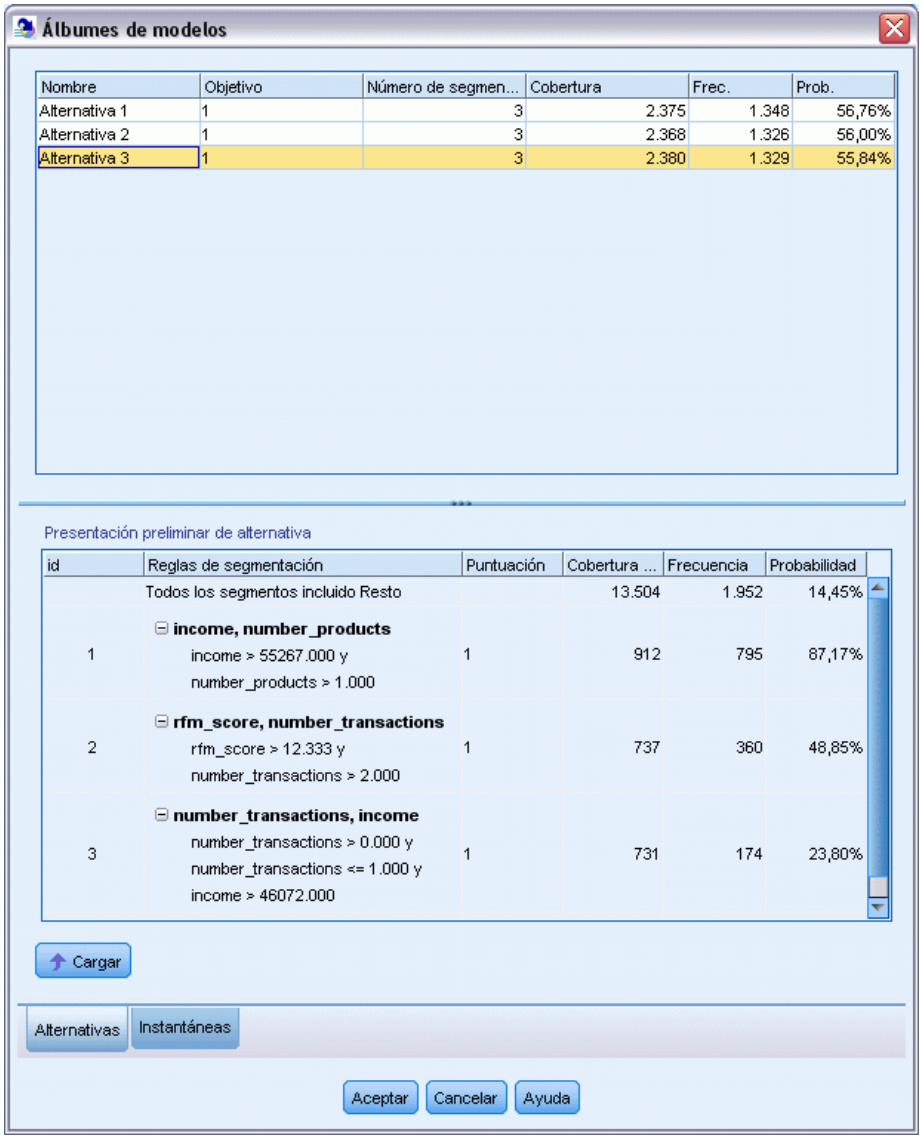
▶ En el visor de listas interactivas, seleccione:
Herramientas > Buscar segmentos

Figura 12-9
Visor de listas interactivas



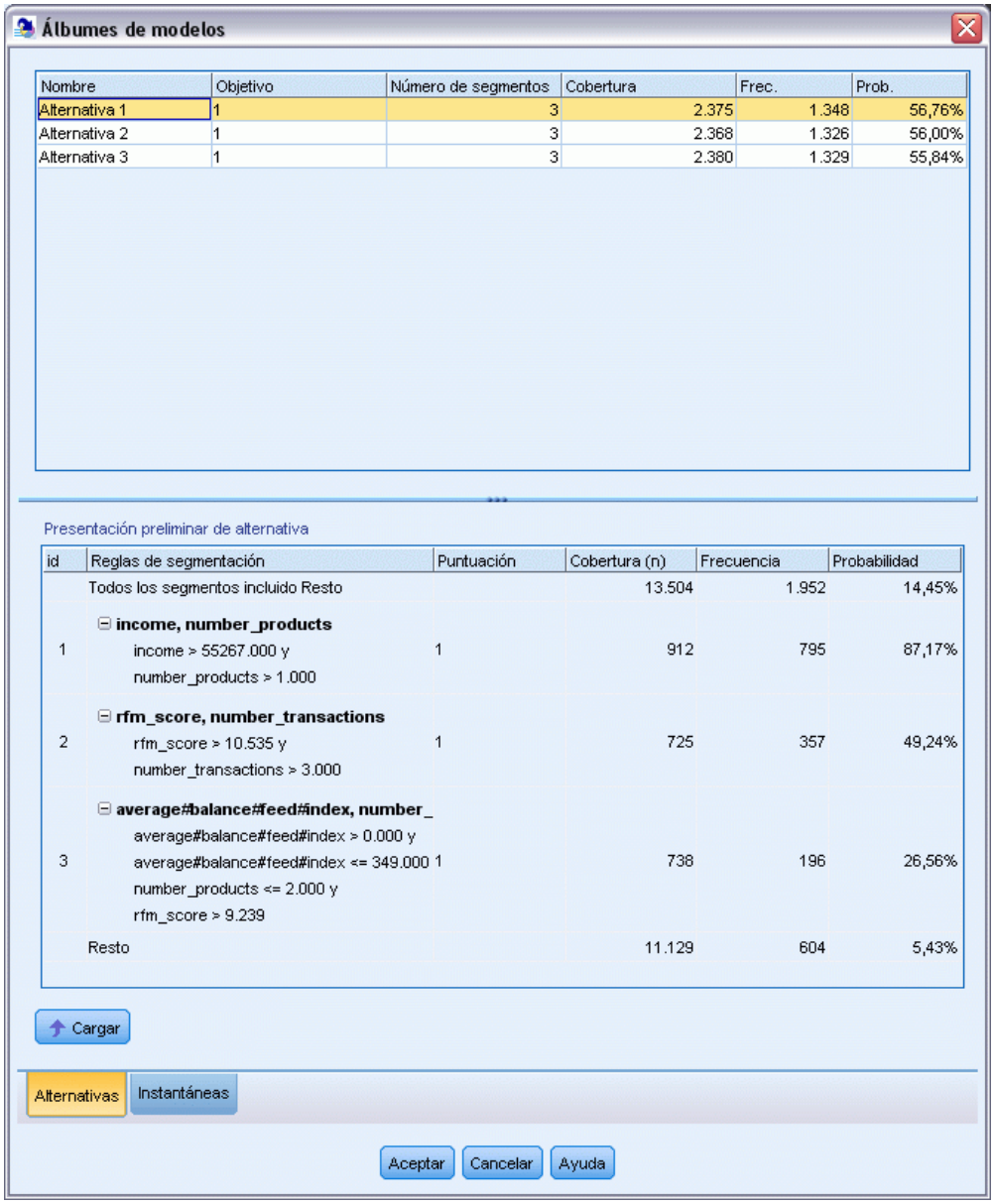
De esta manera se ejecuta la tarea de minería por defecto utilizando la configuración que especificó en el nodo Lista de decisiones. La tarea finalizada devuelve tres modelos alternativos, que se muestran en la pestaña Alternativas del cuadro de diálogo Álbumes de modelo.

Figura 12-10
Modelos alternativos disponibles



▶ Selecciones la primera alternativa de la lista; sus detalles se muestran en el panel Presentación preliminar de alternativa.

Figura 12-11
Modelo alternativo seleccionado



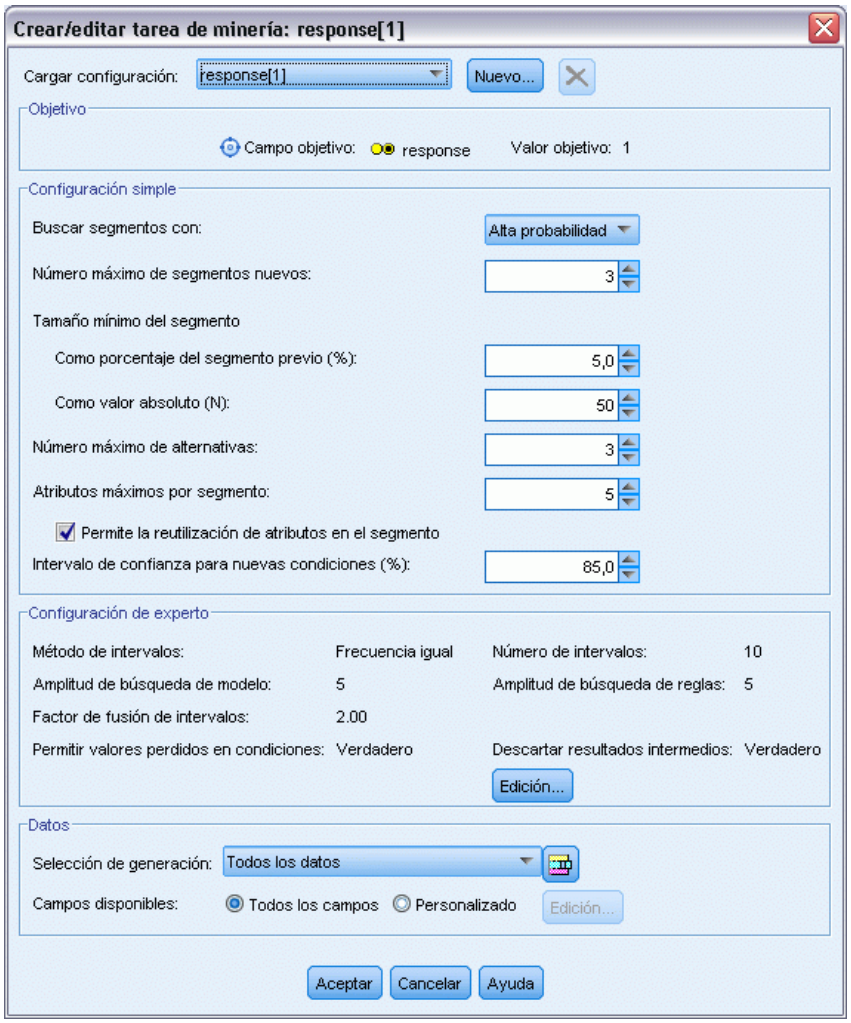
El panel Presentación preliminar de alternativa permite examinar rápidamente cualquier número de alternativas sin cambiar el modelo de trabajo, lo que facilita la experimentación con diferentes enfoques.

Nota: para lograr una mejor visión del modelo, tal vez desee maximizar el panel Presentación preliminar de alternativa dentro de la ventana, como se muestra a continuación. Esta operación se puede realizar arrastrando el borde del panel.

Mediante el uso de reglas basadas en predictores como los ingresos, el número de transacciones por mes y la puntuación RFM, el modelo identifica los segmentos con índices de respuesta mayores que los de la muestra completa. Cuando se combinan los segmentos, este modelo sugiere que es posible mejorar la tasa de acierto hasta el 56.76%. No obstante, el modelo sólo cubre una pequeña parte de la muestra y deja más de 11.000 registros (con varios cientos de aciertos entre ellos) en el resto. Lo que se necesita es un modelo que capture más aciertos de este tipo y que, al mismo tiempo, excluya los segmentos con malos resultados.

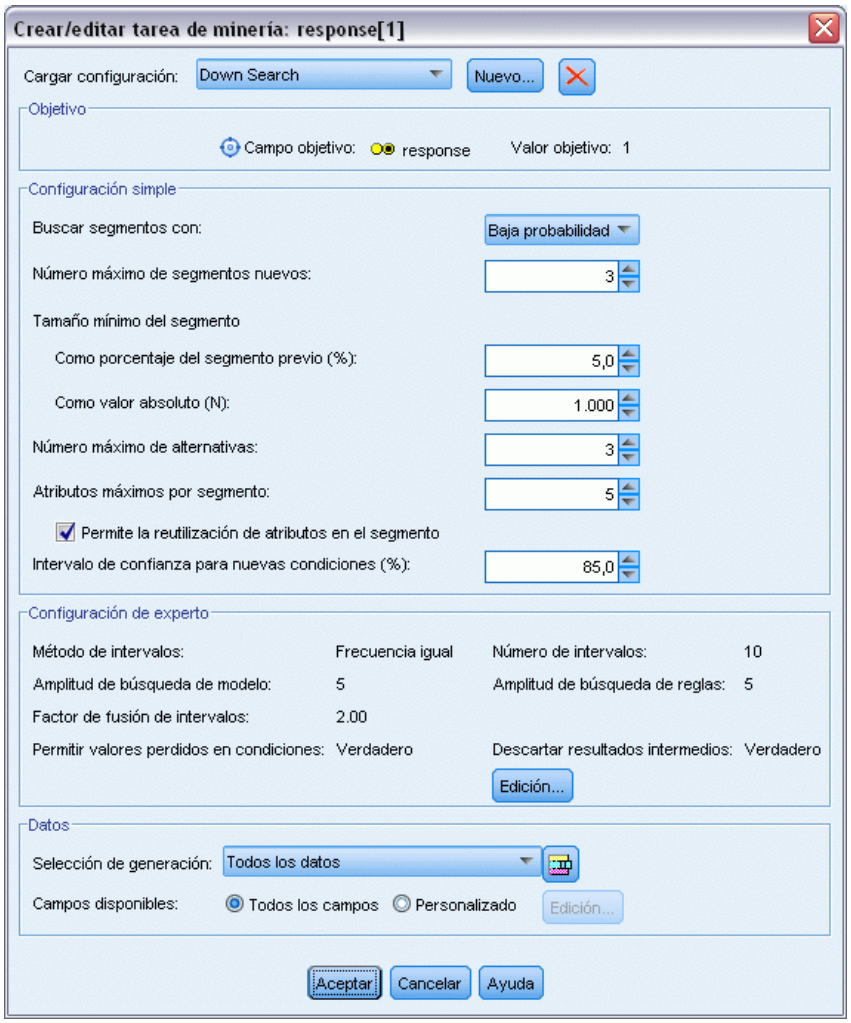
▶ Para probar otro método de modelado, seleccione en los menús:
Herramientas > Configuración

Figura 12-12
Cuadro de diálogo Crear/editar tarea de minería



▶ Pulse en el botón Nuevo (esquina superior derecha) para crea una segunda tarea de minería y especifique *Búsqueda descendente* como el nombre de la tarea en el cuadro de diálogo Nuevas configuraciones.

Figura 12-13
Cuadro de diálogo Crear/editar tarea de minería

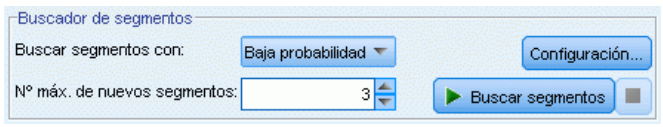


- Cambie la dirección de búsqueda a Baja probabilidad para la tarea. Al hacerlo, el algoritmo buscará los segmentos con los *menores* índices de respuesta en vez de los mayores.

- Aumente el tamaño mínimo del segmento a 1.000. Pulse en Aceptar para volver al visor de listas interactivas.

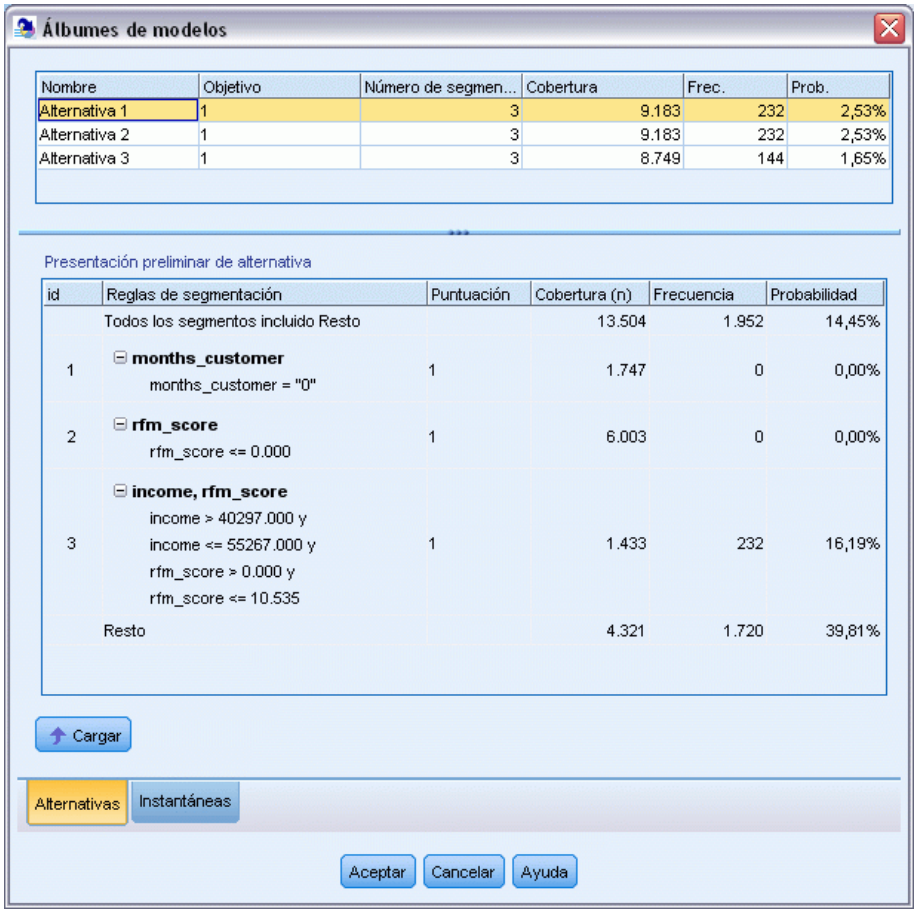
- En el visor de listas interactivas, asegúrese que el panel *Buscar segmentos* muestra los detalles de las nueva tarea y pulse en Buscar segmentos.

Figura 12-14
Buscar segmentos en nueva tarea de minería



La tarea devuelve un nuevo conjunto de alternativas, que se muestran en la pestaña Alternativas del cuadro de diálogo Álbumes de modelo y de las que se puede ver una presentación preliminar del mismo modo que los resultados anteriores.

Figura 12-15
Resultados del modelo Búsqueda descendente

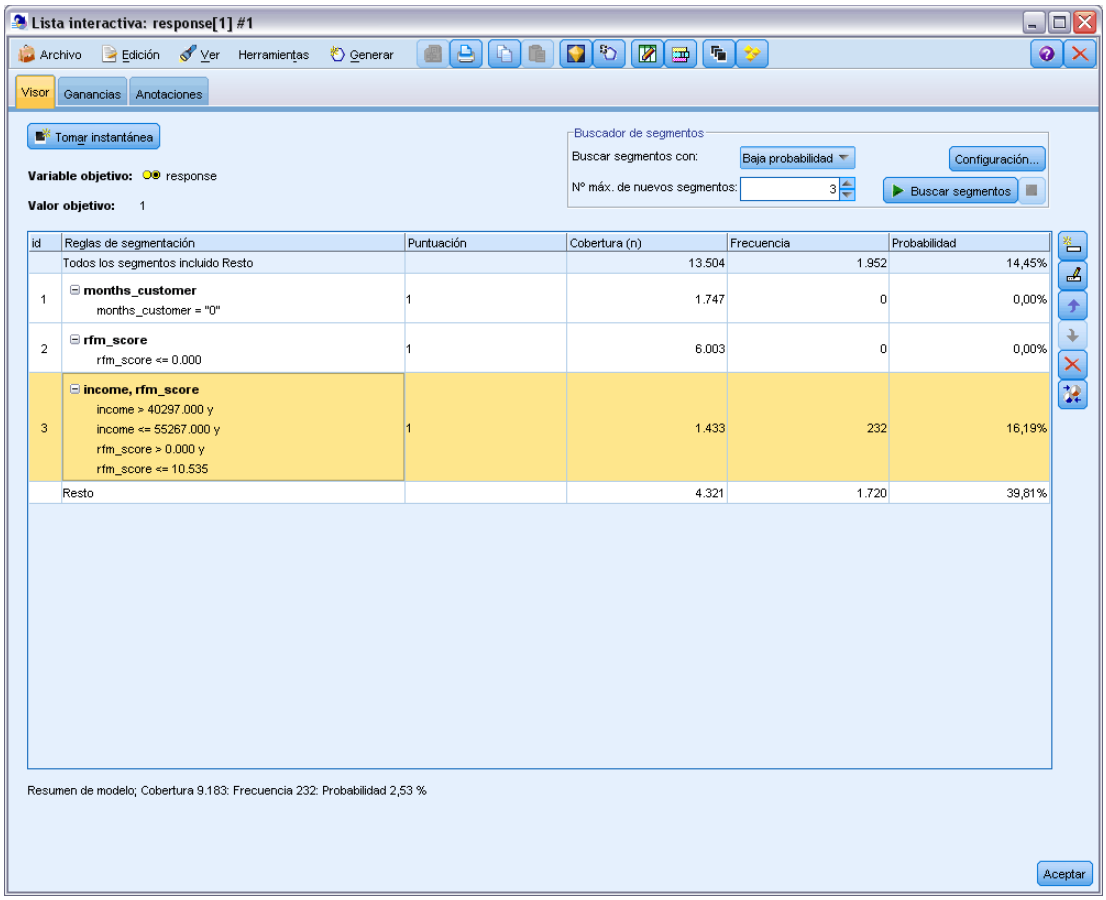


En esta ocasión, cada modelo identifica segmentos con pocas probabilidades de respuesta. Si tenemos en cuenta la primera alternativa, sólo excluir estos segmentos aumentará la tasa de aciertos del resto hasta el 39,81%. Aunque la tasa es más baja que en el modelo anterior, la cobertura es más amplia, en el sentido de que se obtiene un total de aciertos mayor.

Si se combinan los dos enfoques, utilizando una búsqueda de baja probabilidad para descartar los registros de menor interés seguida de una búsqueda de alta probabilidad, podrá mejorar este resultado.

- Pulse en Cargar para que este modelo (la primera alternativa de búsqueda descendente) sea el modelo de trabajo y pulse en Aceptar para cerrar el cuadro de diálogo Álbumes de modelo.

Figura 12-16
Exclusión de un segmento



- Pulse con el botón derecho en los dos primeros segmentos y seleccione Excluir segmento. Juntos, estos segmentos capturan casi 8.000 registros con cero aciertos en ellos, por lo que resulta lógico excluirlos de futuras ofertas. (Para indicar esto, los segmentos excluidos se puntúan con valores nulos.)

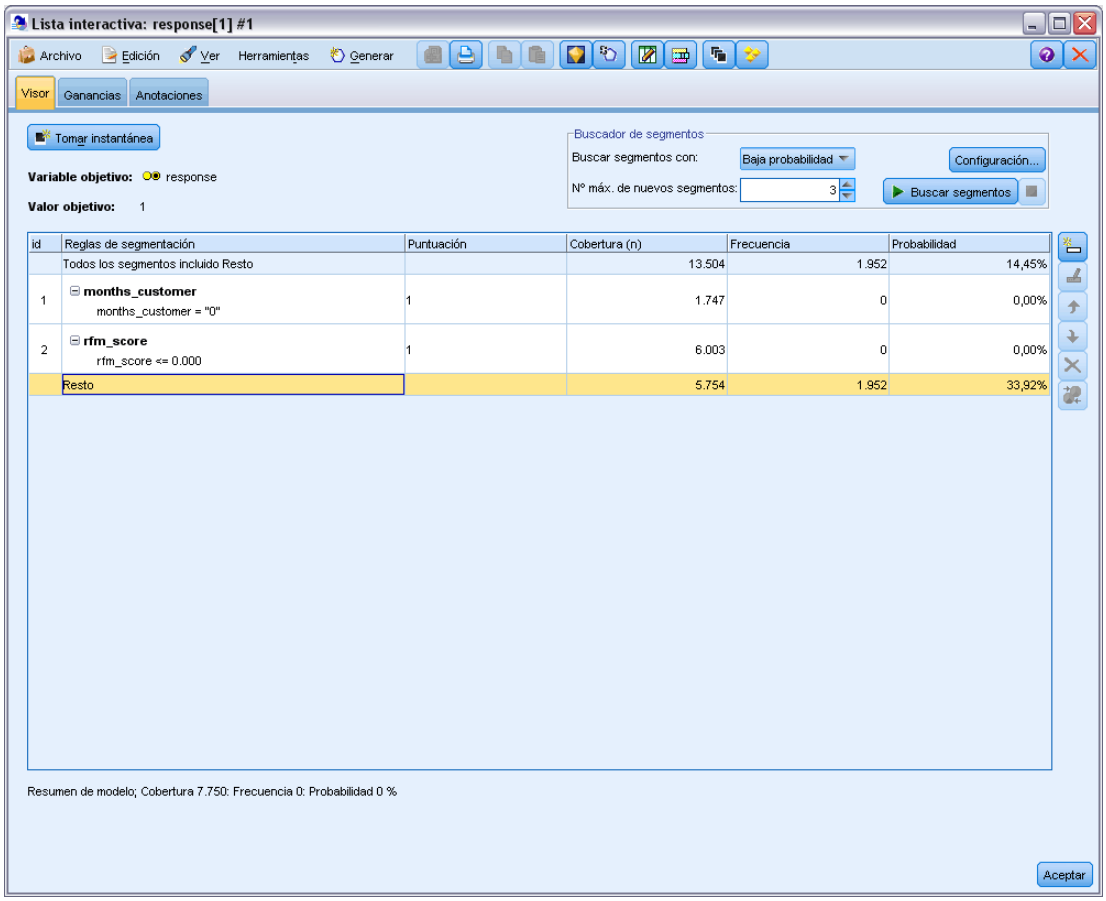
- Pulse con el botón derecho en el tercer segmento y seleccione Eliminar segmento. La tasa de acierto del 16,19% de este segmento no es muy distinta de la tasa base de 14,45%, por lo que no añade la suficiente información que justifique mantenerla.

  Nota: eliminar un segmento no es lo mismo que excluirlo. Si se excluye un segmento, cambia su puntuación, mientras que eliminarlo implica quitarlo completamente del modelo.

  Después de excluir los segmentos con peores resultados, buscaremos los segmentos con mejores resultados en el resto.

- Pulse en la fila Resto de la tabla para seleccionarla y así la próxima tarea de minería se aplicará solamente al resto.

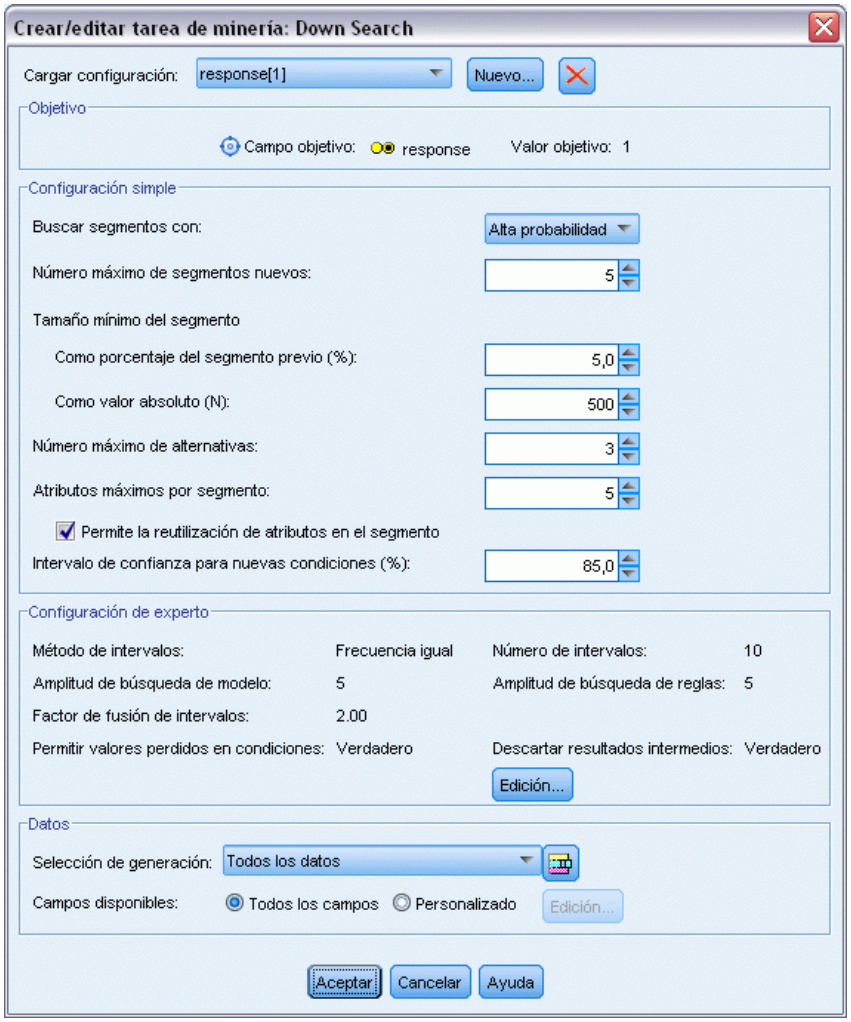
Figura 12-17
Selección de un segmento



- Con el resto seleccionado, pulse en Configuración para volver a abrir el cuadro de diálogo Crear/editar tarea de minería.
- En la parte superior de Configuración de carga, seleccione la tarea de minería por defecto: respuesta[1].
- Modifique la Configuración simple para aumentar el número de nuevos segmentos a 5 y el tamaño mínimo del segmento a 500.

▶ Pulse en Aceptar para volver al visor de listas interactivas.

Figura 12-18
Selección de la tarea de minería por defecto

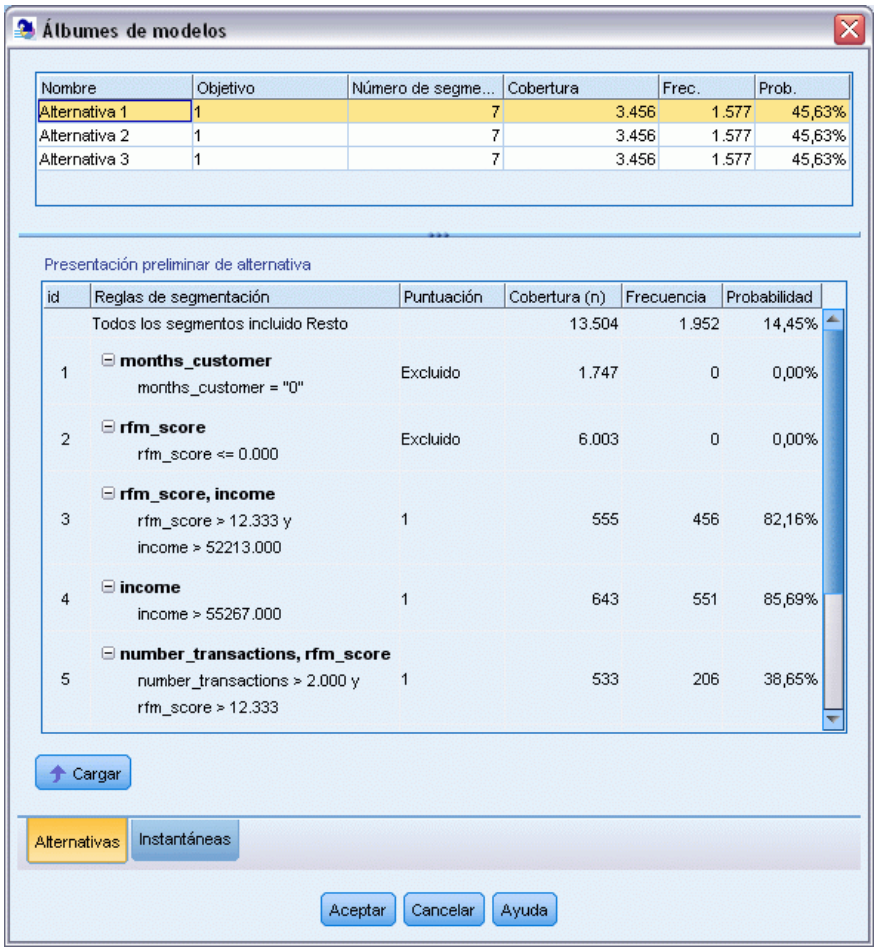


▶ Pulse en Buscar segmentos.

Se mostrará otro conjunto de modelos alternativos. Al introducir los resultados de una tarea de minería en otra, estos últimos modelos contendrán una mezcla de segmentos con buenos y malos resultados. Los segmentos con índices de respuesta bajos se excluyen, lo cual implica que se puntuarán como valores nulos. Por su parte, los segmentos incluidos se puntuarán como 1. Los estadísticos generales reflejan estas exclusiones, ya que el primer modelo alternativo muestra una

tasa de acierto del 45,63%, con una cobertura más amplia (1.577 aciertos de 3.456 registros) que cualquiera de los modelos anteriores.

Figura 12-19
Alternativas del modelo combinado

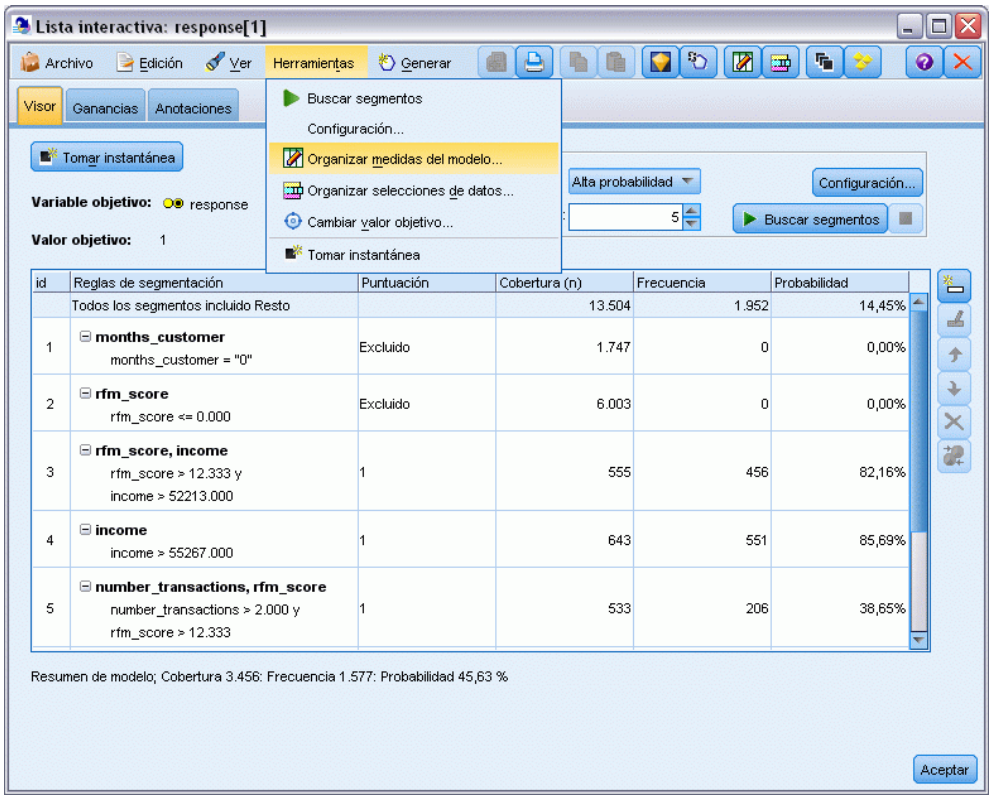


► Visualice la primera alternativa y pulse en Cargar para convertirlo en el modelo de trabajo.

Cálculo de las medidas personalizadas con Excel

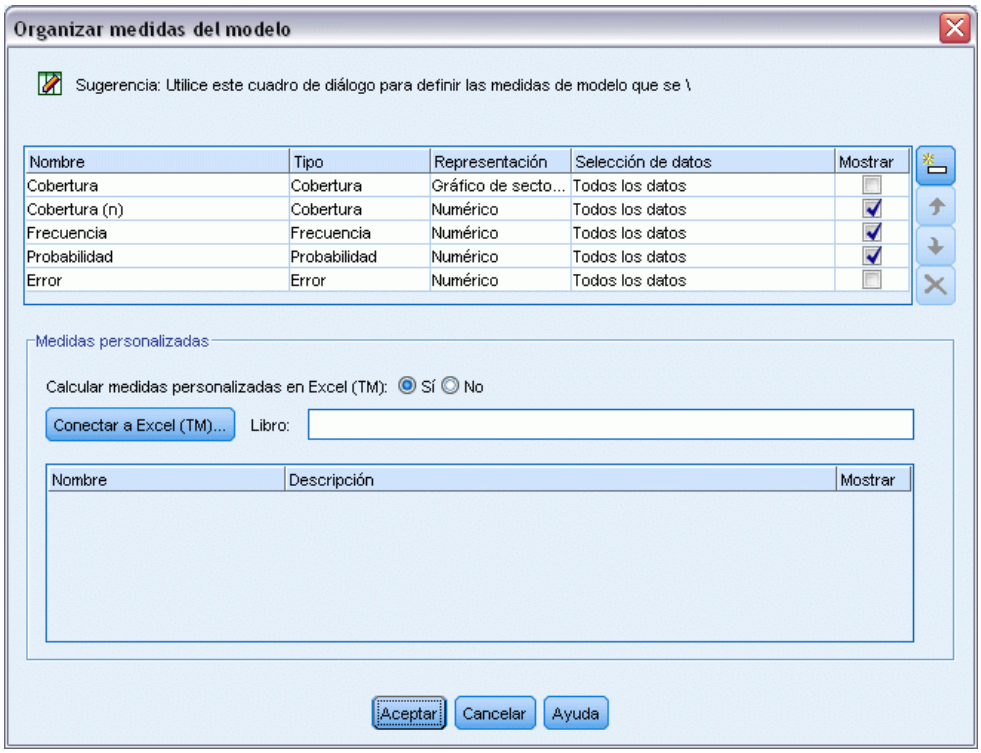
- Para obtener más información sobre el comportamiento del modelo en la práctica, elija Organizar medidas del modelo en el menú Herramientas.

Figura 12-20
Organización de las medidas del modelo



El cuadro de diálogo Organizar medidas del modelo permite elegir las medidas (o columnas) que aparecerán en el visor de listas interactivas. También es posible especificar si las medidas se calcularán utilizando todos los registros o sólo un determinado subconjunto, así como si se prefiere ver un gráfico de sectores en vez de un número en los casos pertinentes.

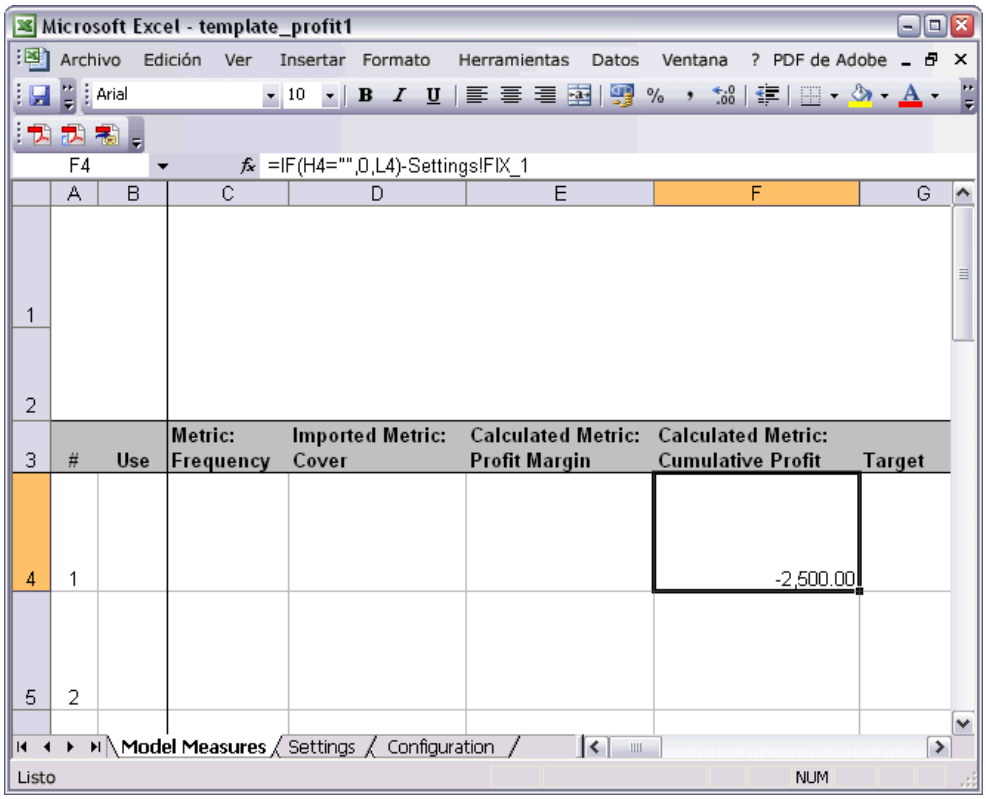
Figura 12-21
Cuadro de diálogo Organizar medidas del modelo



Además, si tiene instalado Microsoft Excel, puede enlazar con una plantilla de Excel que calcule medidas personalizadas para añadirlas a la visualización interactiva.

- En el cuadro de diálogo Organizar medidas de modelo, establezca Calcular mediciones personalizadas en Excel (TM) como Sí.
- Pulse en el botón Conectar a Excel (TM).
- Elija el libro de trabajo *template_profit.xlt*, situado en *streams* en la carpeta *Demos* de la instalación de IBM® SPSS® Modeler, y pulse en Abrir para iniciar la hoja de cálculo.

Figura 12-22
Hoja de cálculo Medidas del modelo



La plantilla de Excel contiene tres hojas de trabajo:

- Medidas de modelo muestra las medidas del modelo importadas del modelo y calcula las medidas personalizadas para exportarlas al modelo.
- Parámetros contiene parámetros que se utilizarán para calcular las medidas personalizadas.
- Configuración define las medidas que se importarán del modelo y se exportarán al modelo.

Las métricas exportadas al modelo son:

- **Margen de beneficio.** Ingresos netos del segmento
- **Beneficio acumulado.** Beneficio total de la campaña

Tal como se define mediante las siguientes fórmulas:

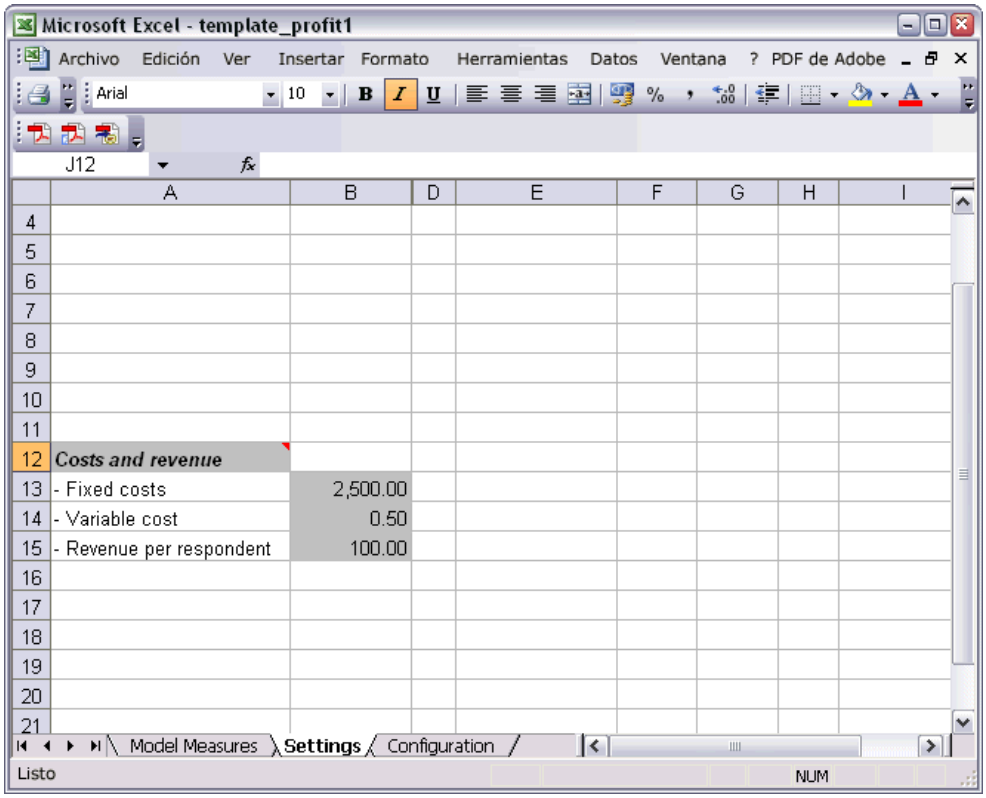
Margen de beneficio = Frecuencia * Ingreso por encuestado - Cubierto * Coste variable

Beneficio acumulado = Margen de beneficio total - Coste fijo

Observe que la frecuencia y la cobertura se importan del modelo.

El usuario debe especificar los parámetros de coste e ingresos en la hoja de cálculo Parámetros.

Figura 12-23
Hoja de cálculo de parámetros de Excel



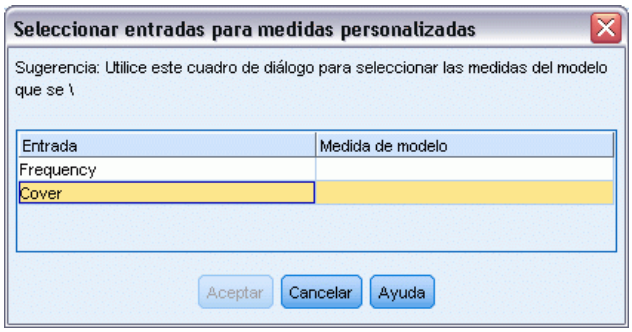
Coste fijo es el coste de preparación de la campaña; por ejemplo, el diseño y la planificación.

Coste variable es el coste de ampliar la oferta a cada cliente, por ejemplo los sobres y los sellos.

Ingreso por encuestado es el ingreso neto que se obtiene de cada cliente que responde a la oferta.

▶ Para completar el enlace con el modelo, utilice la barra de tareas de Windows (o pulse Alt+Tab) para volver a la ventana Lista interactiva.

Figura 12-24
Selección de entradas para medidas personalizadas



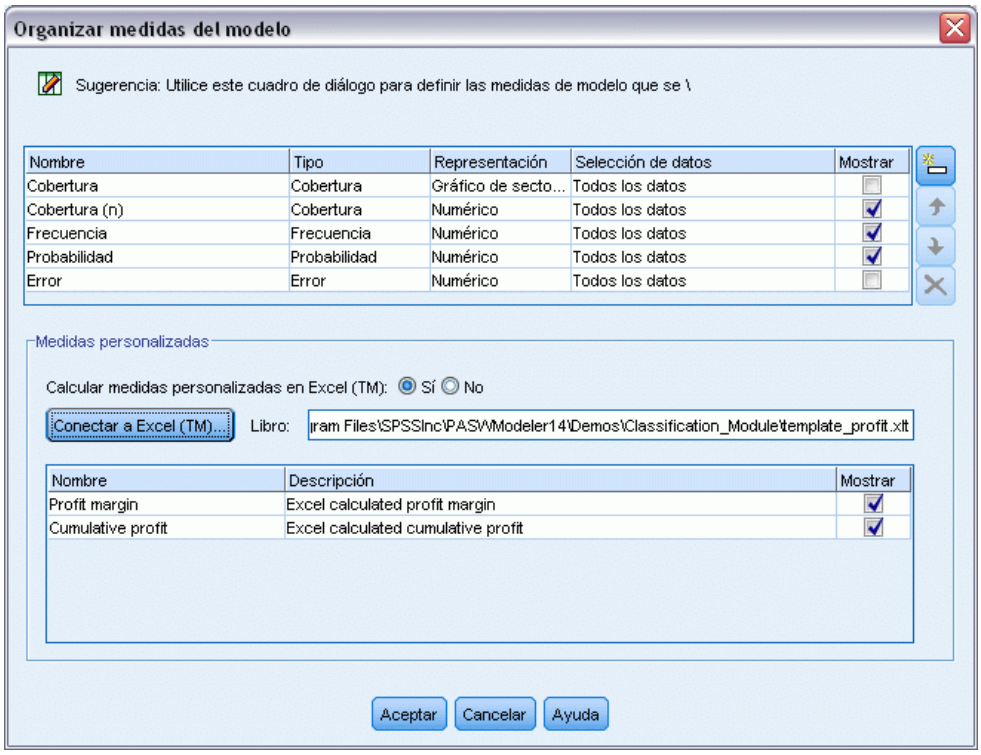
Aparecerá el cuadro de diálogo Seleccionar entradas para medidas personalizadas, que permite asignar entradas del modelo a determinados parámetros definidos en la plantilla. La columna izquierda muestra las medidas disponibles, mientras que la columna derecha asigna dichas medidas a los parámetros de la hoja de cálculo tal como se define en la hoja de cálculo Configuración.

▶ En la columna Medidas del modelo, seleccione Frecuencia y Cubierto (n) en las entradas correspondientes y pulse en Aceptar.

En este caso concreto, los nombres de los parámetros de la plantilla —Frecuencia y Cubierto— coinciden con las entradas, pero sería posible utilizar otros nombres.

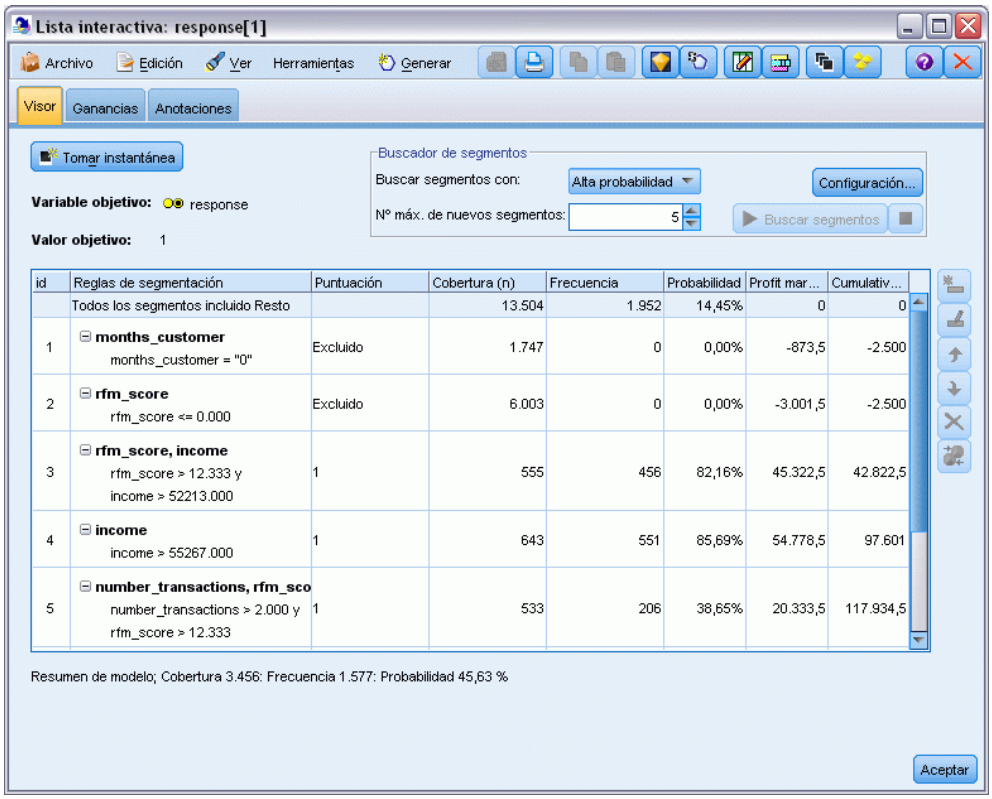
▶ Pulse en Aceptar en el cuadro de diálogo Organizar medidas del modelo para actualizar la visualización de la lista interactiva.

Figura 12-25
Cuadro de diálogo Organizar medidas del modelo con las medidas personalizadas de Excel



Las nuevas medidas ahora se añaden como nuevas columnas en la ventana y se volverán a calcular cada vez que se actualice el modelo.

Figura 12-26
Medidas personalizadas de Excel mostradas en el visor de listas interactivas



Si se edita la plantilla de Excel, es posible crear todas las medidas personalizadas que se desee.

Modificación de la plantilla de Excel

Aunque IBM® SPSS® Modeler se proporciona con una plantilla de Excel predefinida para utilizar con el visor de listas interactivas, es posible que desee modificar la configuración o agregar la suya propia. Por ejemplo, es posible que los costes de la plantilla sean incorrectos para su organización y necesite modificarlos.

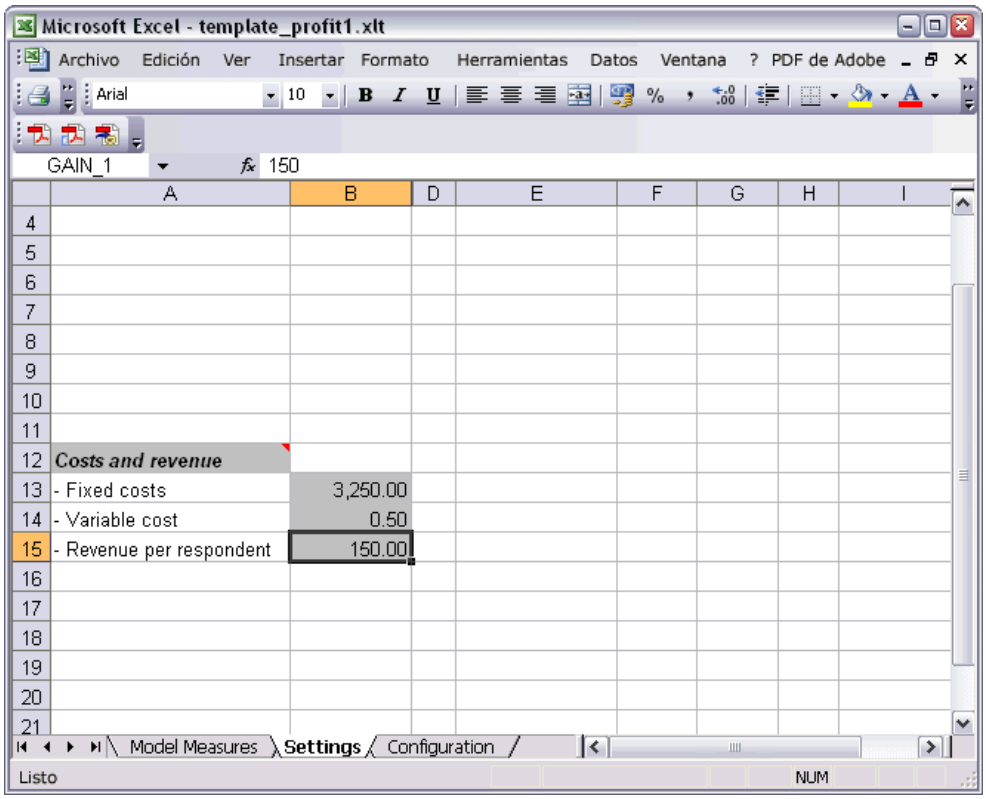
Nota: Si modifica una plantilla existente o crea una plantilla propia recuerde guardar el archivo con un sufijo *.xlt* de Excel 2003.

Para modificar la plantilla predefinida con nuevos detalles de costes y beneficios y actualizar el visor de listas interactivas con las nuevas cifras:

- En el visor de listas interactivas, seleccione Organizar medidas del modelo del menú Herramientas.
- En el cuadro de diálogo Organizar medidas del modelo, pulse en Conectar a Excel™.
- Seleccione el libro *template_profit.xlt* y pulse en Abrir para iniciar la hoja de cálculo.

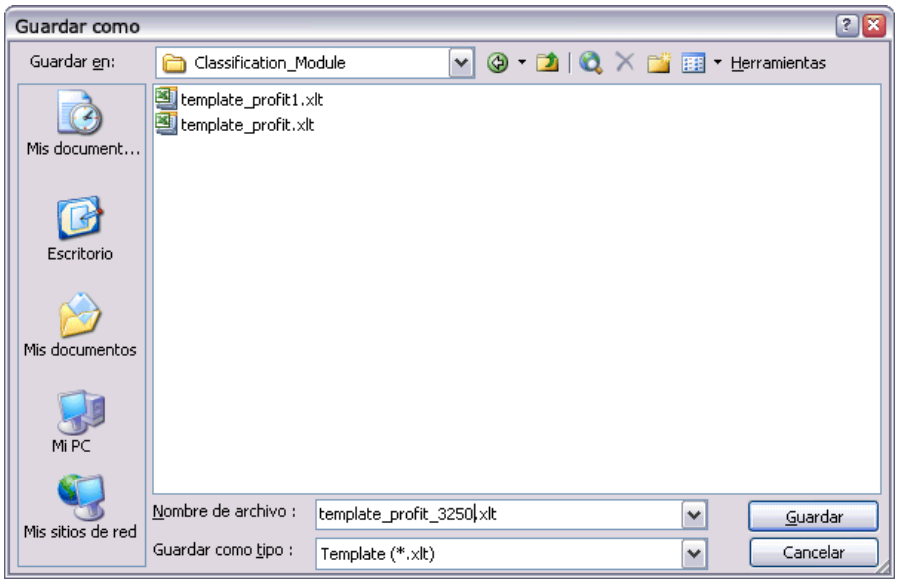
- Seleccione la hoja de cálculo Parámetros.
- Modifique Costes fijo a 3.250,00 e Ingreso por encuestado a 150,00.

Figura 12-27
Valores modificados en la hoja de cálculo Parámetros de Excel



- Guarde la plantilla modificada con un nombre único y relevante. Compruebe que tiene una extensión *.xlt* Excel 2003.

Figura 12-28
Almacenamiento de la plantilla de Excel modificada



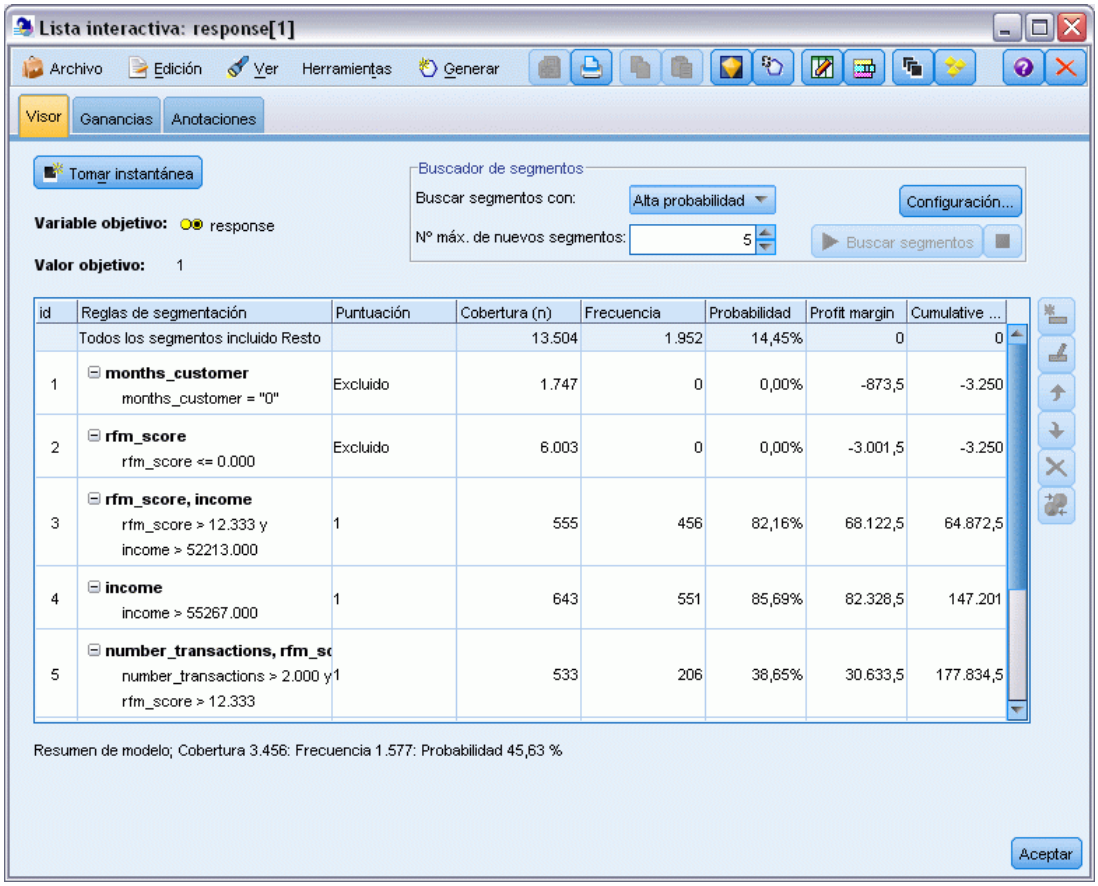
- Utilice la barra de tareas de Windows (o pulse Alt+Tab) para volver al visor de listas interactivas.

  En el cuadro de diálogo Seleccionar entradas para medidas personalizadas, seleccione las medidas que desea visualizar y pulse en Aceptar.

- Pulse en Aceptar en el cuadro de diálogo Organizar medidas del modelo para actualizar la visualización de la lista interactiva.

Obviamente, este ejemplo sólo muestra una forma simple de modificar la plantilla de Excel; puede realizar más cambios para obtener los datos y transmitir los datos a la visualización de la lista interactiva, o trabajar en Excel para producir otros resultados, como gráficos.

Figura 12-29
Medidas personalizadas modificadas de Excel mostradas en el visor de listas interactivas



Almacenamiento de resultados

Para guardar un modelo y utilizarlo más tarde durante la sesión interactiva, puede tomar una instantánea del modelo, que aparecerá en la pestaña Instantáneas. Durante la sesión interactiva se puede acceder a las instantáneas guardadas en todo momento.

Si continúa de este modo, puede experimentar con tareas de minería adicionales para buscar más segmentos. También puede editar segmentos existentes, insertar segmentos personalizados en función de sus propias reglas de negocios, crear selecciones de datos para optimizar el modelo para grupos específicos y personalizar el modelo de muchas otras maneras. Finalmente, puede incluir o excluir explícitamente cada segmento para especificar cómo se va a puntuar.

Cuando esté satisfecho con los resultados, puede utilizar el menú Generar para generar un modelo que se añada a rutas o que se distribuya para realizar la puntuación.

Si lo prefiere, para guardar su sesión interactiva y continuarla en otro momento, elija Actualizar nodo de modelado en el menú Archivo. De esta manera, el nodo de modelado de lista de decisiones se actualizará con la configuración que esté utilizando, incluidas tareas de minería, instantáneas de modelos, selecciones de datos y medidas personalizadas. La próxima vez que ejecute la ruta, asegúrese de que está seleccionada la opción Usar información de sesión guardada en el nodo de modelado Lista de decisiones para volver a iniciar la sesión en su estado actual. Si desea obtener más información, consulte el tema Lista de decisiones en el capítulo 9 en *Nodos de modelado de IBM SPSS Modeler 14.2*.

Capítulo
13

Clasificación de clientes de telecomunicaciones (Regresión logística multinomial)

La regresión logística es una técnica de estadístico para clasificar los registros en función los valores de los campos de entrada. Es análoga a la regresión lineal pero utiliza un campo objetivo categórico en lugar de uno numérico.

Por ejemplo, imagine que un proveedor de telecomunicaciones ha segmentado su base de clientes por patrones de uso de servicio, y ha categorizado a los clientes en cuatro grupos. Si los datos demográficos se pueden utilizar para predecir la pertenencia a un grupo, se pueden personalizar las ofertas para cada uno de los posibles clientes.

Este ejemplo utiliza la ruta denominada *telco_custcat.str*, que hace referencia al archivo de datos denominado *telco.sav*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *telco_custcat.str* está ubicado en el directorio *streams*.

Este ejemplo se centra en la utilización de datos demográficos para pronosticar patrones de uso. El campo objetivo *catpers* tiene cuatro posibles valores que corresponden a los cuatro grupos de clientes:

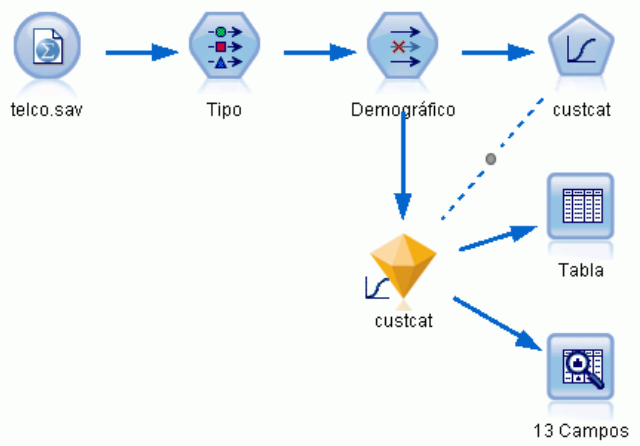
Valor	Label
1	Servicio básico
2	Servicio electrónico
3	Servicio plus
4	Servicio total

Como el objetivo tiene varias categorías, se utiliza un modelo multinomial. En el caso de un objetivo con dos categorías, como sí/no, verdadero/falso, o pérdida/retención, se puede crear un modelo binomial. Si desea obtener más información, consulte el tema Pérdida de clientes de telecomunicaciones (Regresión logística binomial) en el capítulo 14 el p. 159.

Generación de la ruta

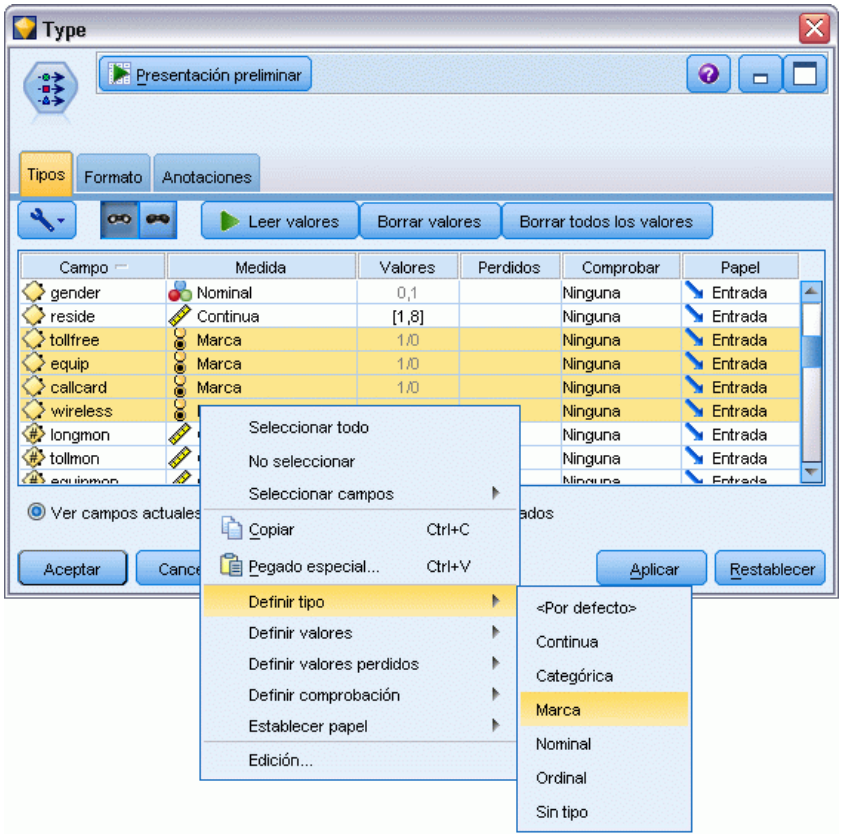
- Añada un nodo de origen Archivo Statistics apuntando a *telco.sav* en la carpeta *Demos*.

Figura 13-1
Ruta de ejemplo para clasificar a los clientes mediante regresión logística multinomial



- Añada un nodo Tipo y pulse en Leer valores, asegurándose así de que todos los niveles de medición están definidos correctamente. Por ejemplo, la mayoría de valores 0 y 1 se pueden considerar marcas.

Figura 13-2
Definición del nivel de medición para campos múltiples

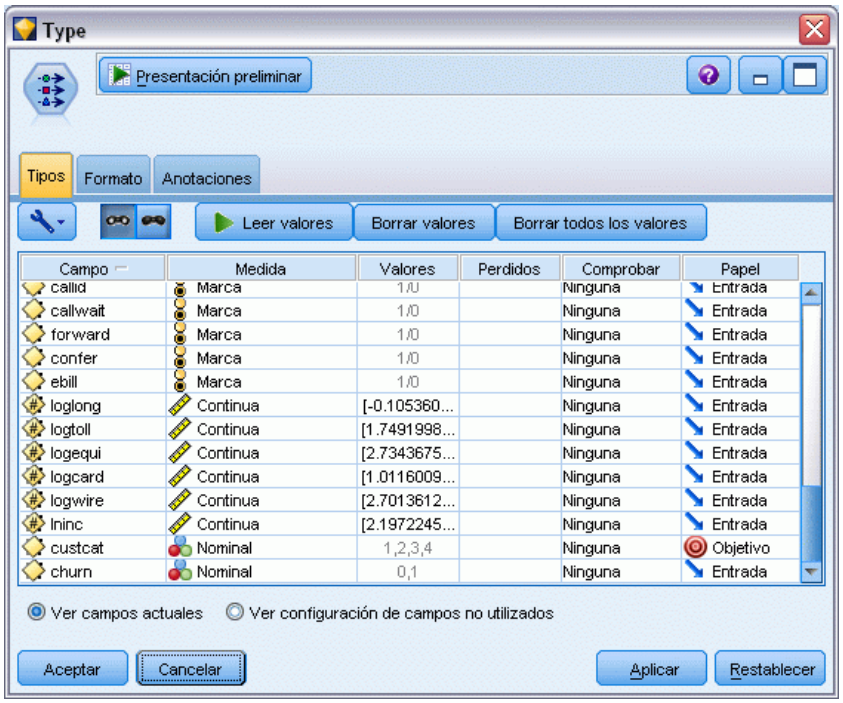


Sugerencia: para cambiar propiedades de varios campos con valores similares (como 0 y 1), pulse en la cabecera de la columna *Valores* para ordenar campos por valor y, a continuación, mantenga pulsada la tecla Mayús mientras utiliza el ratón o las teclas de flecha para seleccionar todos los campos que quiera cambiar. A continuación, puede pulsar con el botón derecho en los elementos seleccionados para cambiar el nivel de medición u otros atributos de los campos seleccionados.

Tenga en cuenta que es más correcto considerar *sexo* como campo con un conjunto de dos valores, en lugar de marca, deje su valor de medición como Nominal.

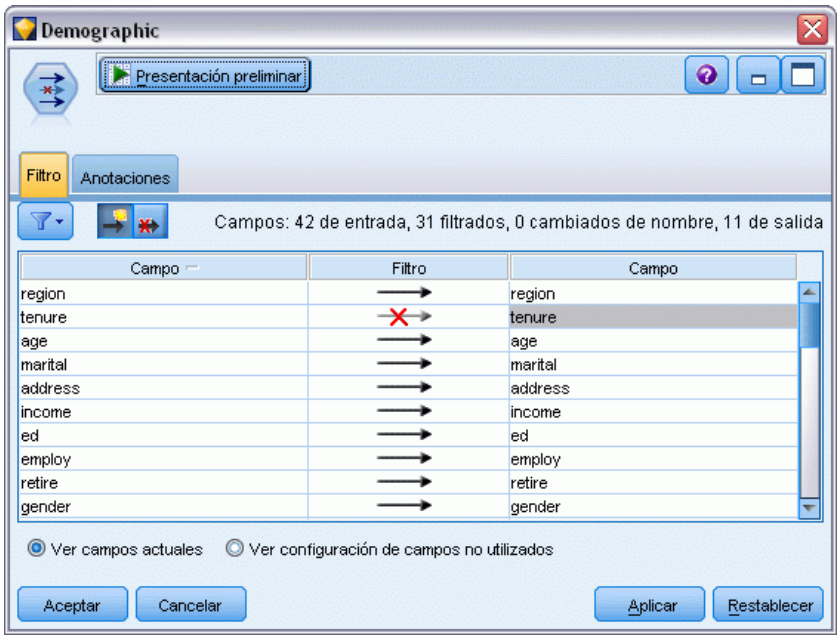
▶ Defina el papel del campo *custcat* a Objetivo. El resto de campos debe tener sus papeles definidas en Entrada.

Figura 13-3
Definición del papel de campos



Puesto que el ejemplo se centra en datos demográficos, utilice un nodo Filtrar para añadir únicamente los campos relevantes (*región*, *edad*, *estado civil*, *dirección*, *ingresos*, *educación*, *empleo*, *jubilación*, *sexo*, *residencia* y *custcat*). Los otros campos se pueden excluir para este análisis.

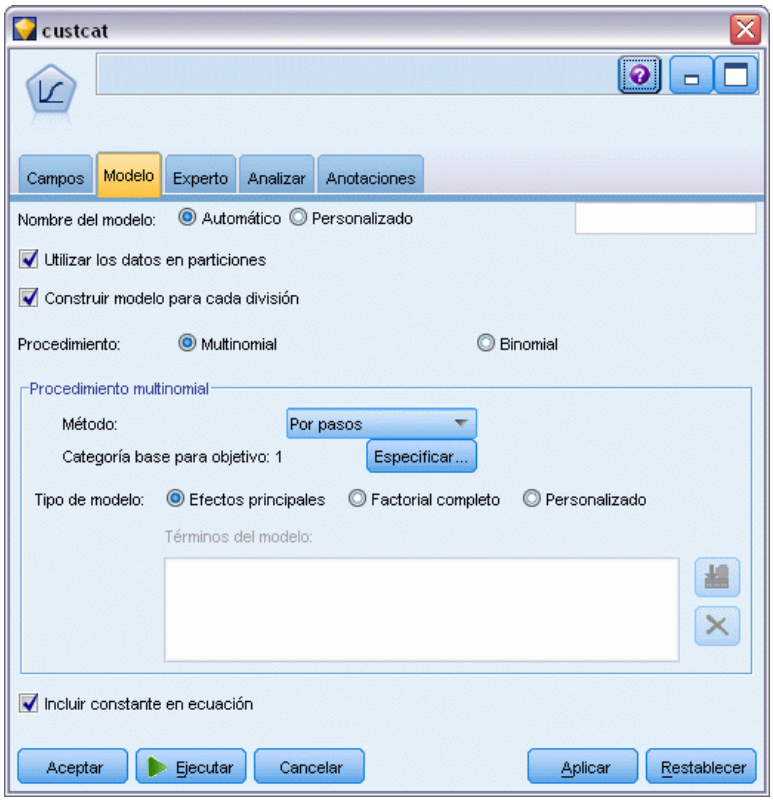
Figura 13-4
Filtrado de los campos demográficos



(Si lo prefiere, puede cambiar el papel de estos campos a Ninguno en lugar de excluirlos, o bien seleccionar los campos que desee utilizar en el nodo de modelado.)

- En el nodo Logística, pulse en la pestaña Modelo y seleccione el método Por pasos. Seleccione Multinomial, Efectos principales e Incluir constante en ecuación.

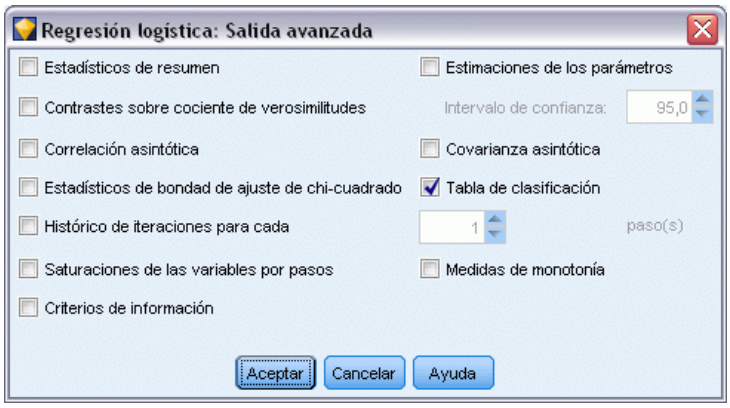
Figura 13-5
Selección de opciones del modelo



Deje la Categoría base para objetivo como 1. El modelo comparará a otros clientes con aquellos que se hayan suscrito al Servicio básico.

- En la pestaña Experto, seleccione el modo Experto, después Salida y, en el cuadro de diálogo Salida avanzada, seleccione Tabla de clasificación.

Figura 13-6
Selección de opciones de salida

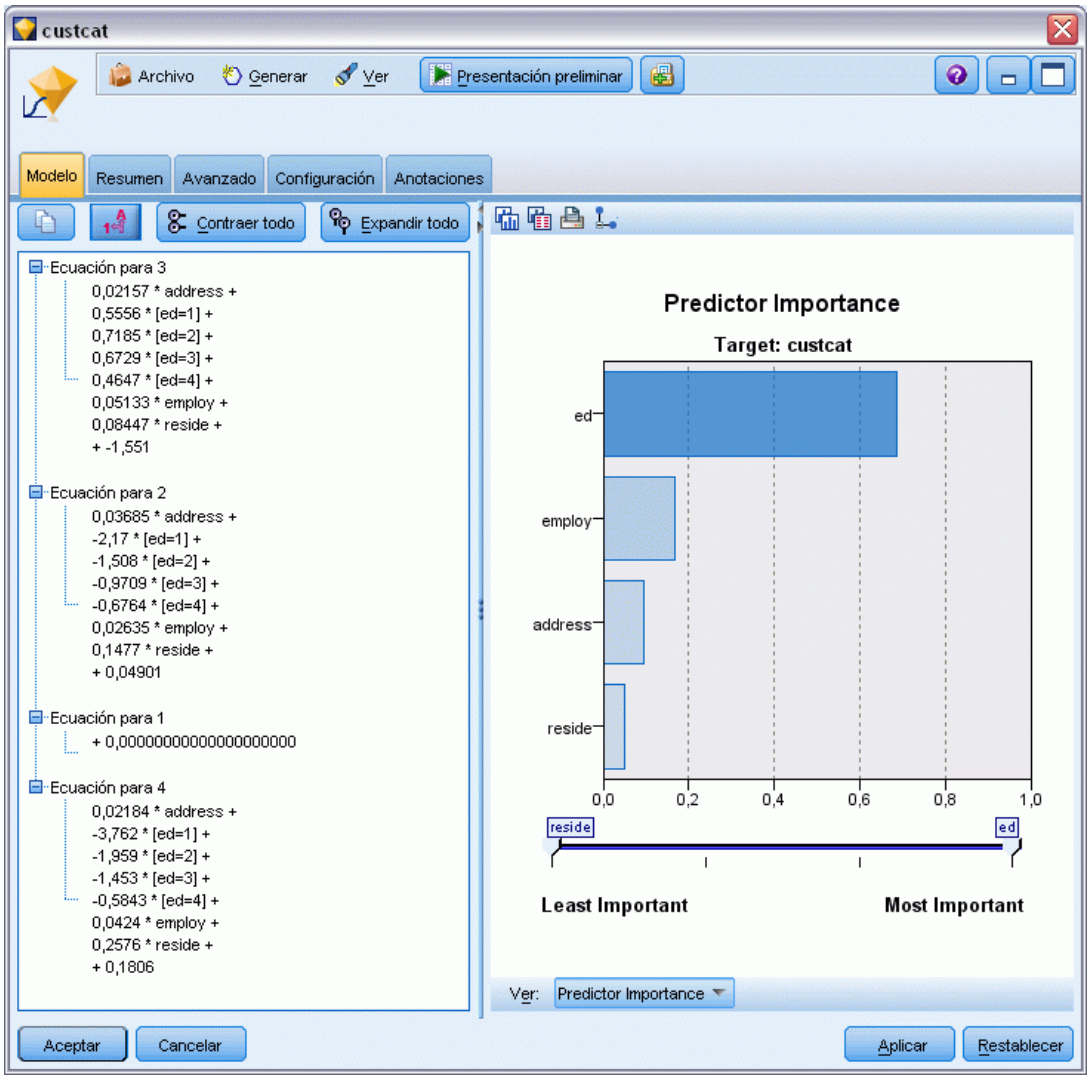


Exploración del modelo

- Ejecute el nodo para generar el modelo, que se añade a la paleta de modelos en la esquina superior derecha. Para ver sus detalles, pulse con el botón derecho en el nodo del modelo generado y seleccione Examinar.

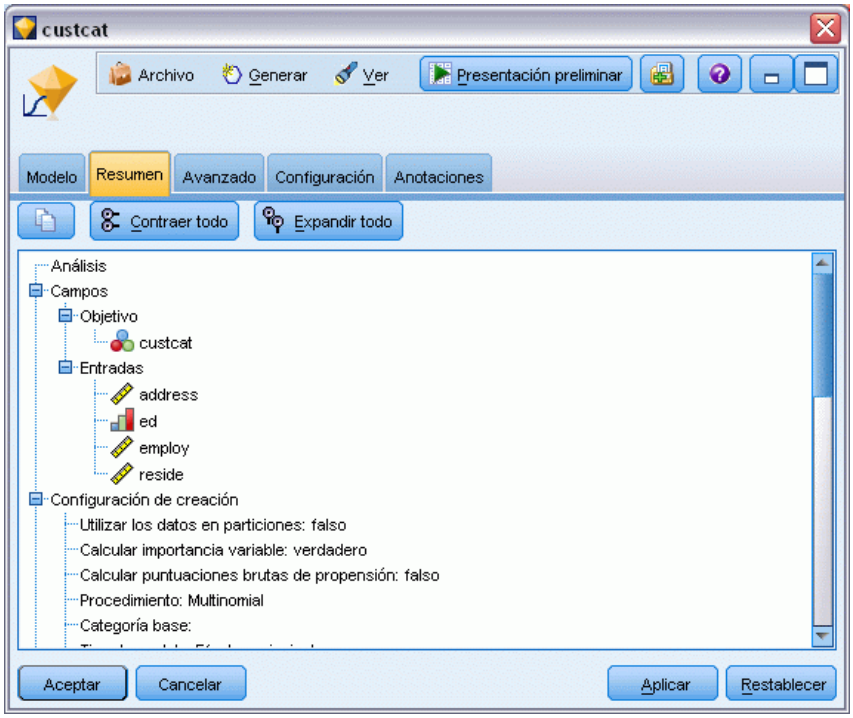
La pestaña Modelo muestra las ecuaciones utilizadas para asignar registros del campo objetivo. Hay cuatro categorías, una de las cuales es la categoría de base para la que no se muestran detalles de la ecuación. Se muestran los detallas para las otras tres ecuaciones, donde la categoría 3 representa Servicio Plus y así sucesivamente.

Figura 13-7
Exploración de los resultados del modelo



La pestaña Resumen muestra (entre otras cosas) el objetivo y las entradas (campos predictores) que utiliza el modelo. Observe que éstos son los campos que se eligieron en base al método Por pasos, no la lista completa enviada para consideración.

Figura 13-8
Resumen del modelo en el que se ven los campos Objetivo y Entrada

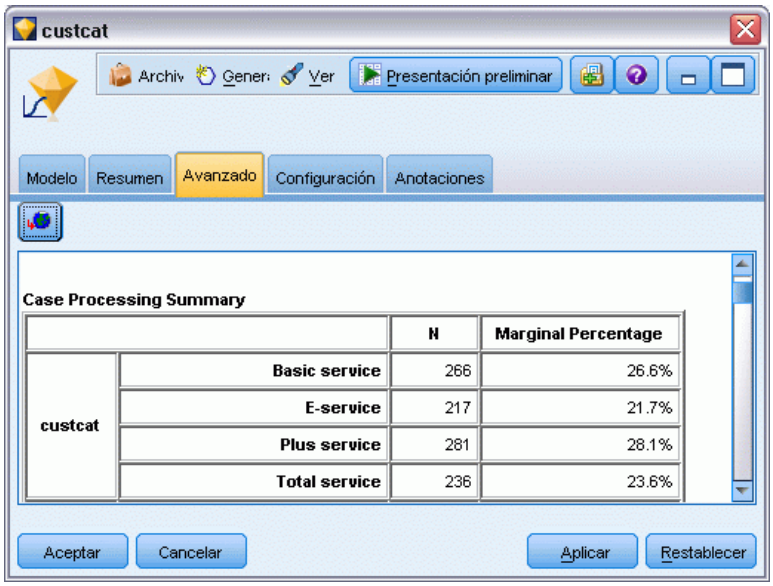


Los elementos que se muestran en la pestaña Avanzado dependen de las opciones seleccionadas en el cuadro de diálogo Salida avanzada del nodo de modelado.

Un elemento que siempre se muestra es el resumen de procesamiento de casos, que indica el porcentaje de los registros que se incluyen en cada categoría del campo objetivo. Esto le proporciona un modelo nulo que puede utilizar como base para comparar.

Sin construir un modelo que utilice predictores, su mejor opción sería asignar todos los clientes al grupo más común, que es el Servicio plus.

Figura 13-9
Resumen del procesamiento de los casos

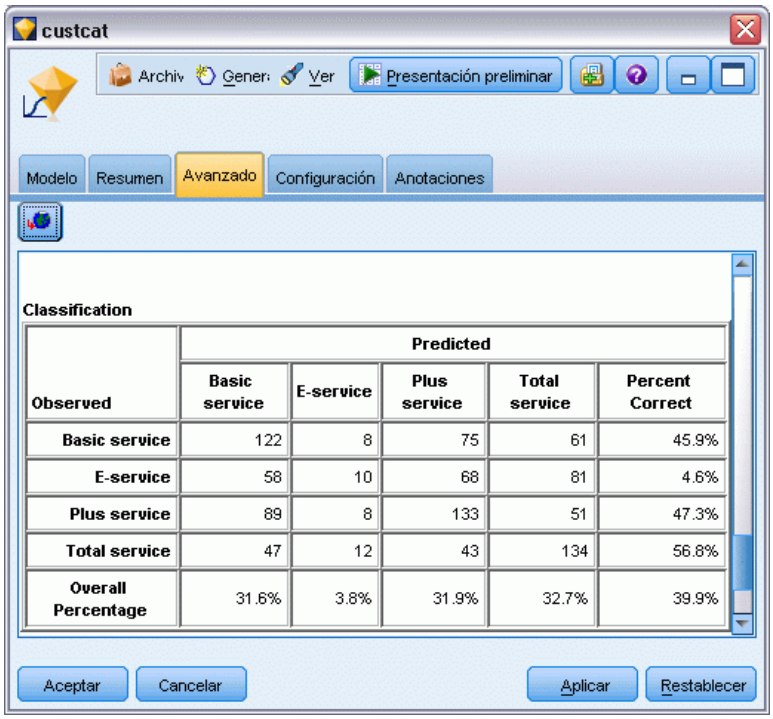


Sobre la base de los datos de entrenamiento, si asignara todos los clientes al modelo nulo acertaría 281/1000 = 28,1% de las veces. La pestaña Avanzado contiene más información que le permite examinar los pronósticos del modelo. Después, puede comparar los pronósticos con los resultados del modelo nulo para comprobar qué tal funciona el modelo con sus datos.

En la parte inferior de la pestaña Avanzado, la tabla Clasificación muestra los resultados de su modelo, que es correcto el 39,9% de las veces.

En concreto, su modelo es muy bueno en identificar clientes de Servicio total (categoría 4), pero no es fiable al identificar clientes de Servicio electrónico (categoría 2). Si desea una mayor exactitud sobre los clientes de la categoría 2, deberá encontrar otro predictor para identificarlos.

Figura 13-10
Tabla de clasificación



Observed	Predicted				
	Basic service	E-service	Plus service	Total service	Percent Correct
Basic service	122	8	75	61	45.9%
E-service	58	10	68	81	4.6%
Plus service	89	8	133	51	47.3%
Total service	47	12	43	134	56.8%
Overall Percentage	31.6%	3.8%	31.9%	32.7%	39.9%

Dependiendo de lo que quiera pronosticar, el modelo puede ser totalmente adecuado para sus necesidades. Por ejemplo, si no le interesa identificar a los clientes de la categoría 2, el modelo puede ser suficientemente exacto. Éste puede ser el caso si el Servicio electrónico se utiliza para atraer clientes pero proporciona pocos beneficios.

Si, por ejemplo, su rentabilidad más alta procede de los clientes de las categorías 3 o 4, el modelo puede darle la información que quiere.

Para evaluar cómo se ajusta el modelo a los datos, en el cuadro de diálogo Salida avanzada hay disponibles varios diagnósticos cuando se está construyendo el modelo. Si desea obtener más información, consulte el tema Resultado avanzado del nugget de modelo logístico en el capítulo 10 en *Nodos de modelado de IBM SPSS Modeler 14.2*. Puede encontrar explicaciones de los fundamentos matemáticos de los métodos de modelado utilizados en IBM® SPSS® Modeler en el *Manual de algoritmos de SPSS Modeler*, disponible en el directorio *\Documentation* del disco de instalación.

Recuerde que estos resultados están basados sólo en los datos de entrenamiento. Para evaluar qué tal se extiende el modelo a otros datos de casos reales, se utilizaría un nodo de partición para reservar un subconjunto de registros para comprobación y validación. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 14.2*.

Capítulo
14

Pérdida de clientes de telecomunicaciones (Regresión logística binomial)

La regresión logística es una técnica de estadístico para clasificar los registros en función los valores de los campos de entrada. Es análoga a la regresión lineal pero utiliza un campo objetivo categórico en lugar de uno numérico.

Este ejemplo utiliza la ruta denominada *telco_churn.str*, que hace referencia al archivo de datos denominado *telco.sav*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *telco_churn.str* está ubicado en el directorio *streams*.

Por ejemplo, suponga que un proveedor de telecomunicaciones está preocupado por el número de clientes que se pasan a la competencia. Si pudiera utilizar los datos para pronosticar qué clientes es más probable que se pasen a otro proveedor, podría personalizar las ofertas para retener a tantos clientes como sea posible.

Este ejemplo se centra en el uso de datos de uso para pronosticar la pérdida de clientes (churn). Como el objetivo tiene dos categorías distintas, se utiliza un modelo binomial. Si un objetivo tiene varias categorías, se puede crear un modelo multinomial. Si desea obtener más información, consulte el tema Clasificación de clientes de telecomunicaciones (Regresión logística multinomial) en el capítulo 13 el p. 149.

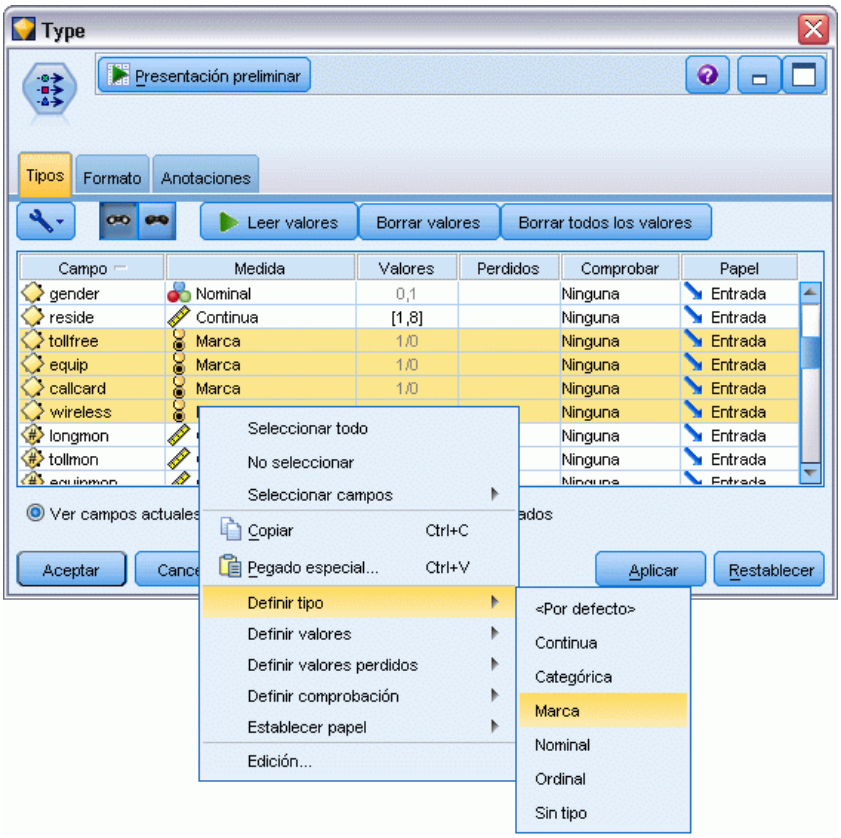
Generación de la ruta

- Añada un nodo de origen Archivo Statistics apuntando a *telco.sav* en la carpeta *Demos*.

Figura 14-1
Ruta de ejemplo para clasificar a los clientes mediante regresión logística binomial

- Añada un nodo Tipo para definir los campos, asegurándose así de que todos los niveles de medición están definidos correctamente. Por ejemplo, la mayoría de los campos con valores 0 y 1 se pueden considerar como marcas, pero algunos campos, como Sexo, se ven con más precisión como un campo nominal con dos valores.

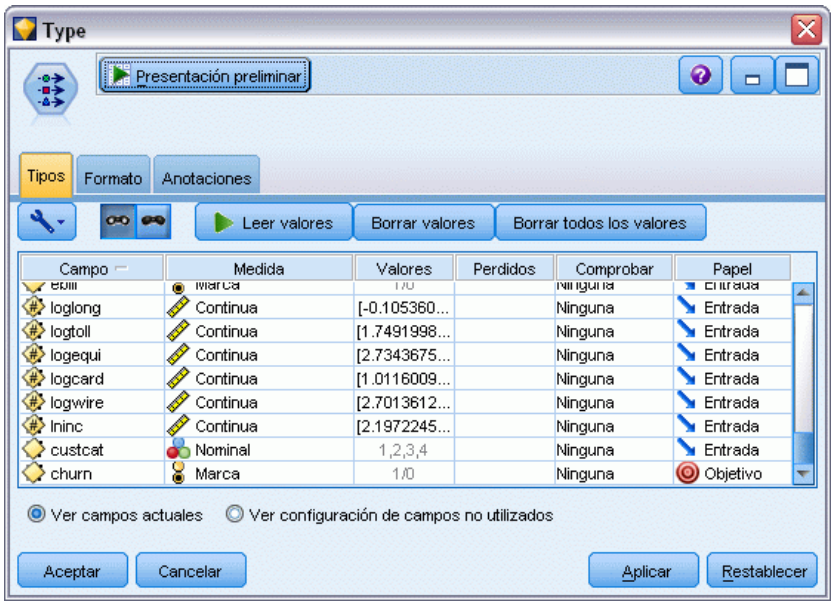
Figura 14-2
Definición del nivel de medición para campos múltiples



Sugerencia: para cambiar las propiedades de varios campos con valores similares (como 0 y 1), pulse en la cabecera de la columna *Valores* para ordenar campos por valor y, a continuación, mantenga pulsada la tecla Mayús mientras utiliza el ratón o las teclas de flecha para seleccionar todos los campos que desee cambiar. A continuación, puede pulsar con el botón derecho en los elementos seleccionados para cambiar el nivel de medición u otros atributos de los campos seleccionados.

- Defina el nivel de medición del campo *abandono* a Marca y defina el papel a Objetivo. El resto de campos debe tener sus papeles definidas en Entrada.

Figura 14-3
Definición del nivel de medición y papel para el campo abandono



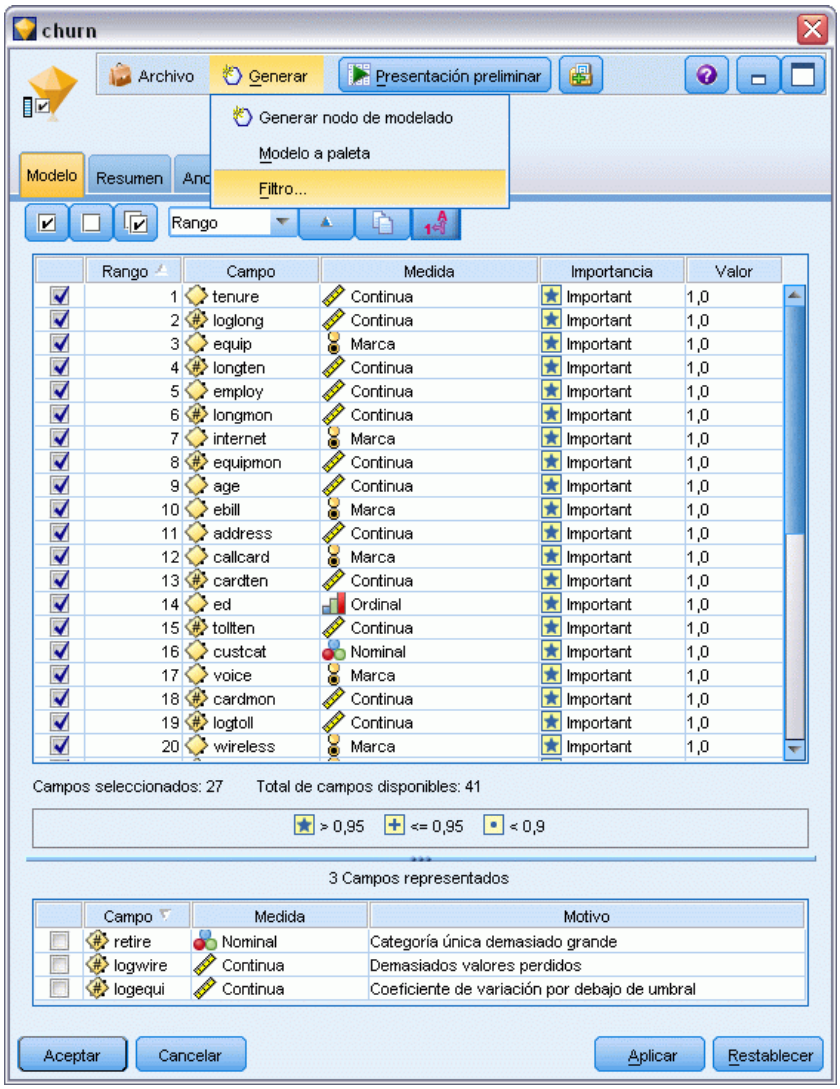
- Añada un nodo de modelado Selección de características al nodo Tipo.

  El uso de un nodo Selección de características permite eliminar predictores o datos que no aportan ninguna información útil en cuanto a la relación predictor/objetivo.

- Ejecute la ruta.

▶ Abra el nugget de modelo resultante, y desde el menú Generar, seleccione Filtrar para crear un nodo Filtrar.

Figura 14-4
Generación de un nodo Filtro desde el nodo Selección de características

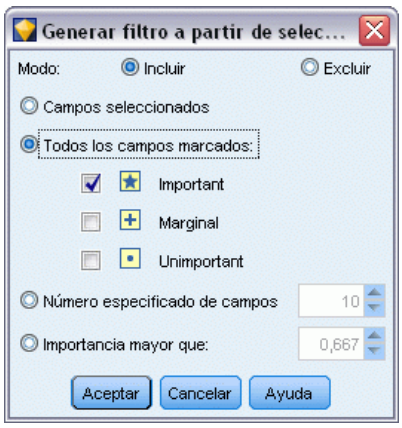


No todos los datos del archivo *telco.sav* serán útiles para pronosticar la pérdida de clientes. Puede utilizar un filtro para seleccionar sólo los datos que se consideren importantes como predictores.

▶ En el cuadro de diálogo Generar filtro, seleccione Todos los campos marcados: Importante y pulse en Aceptar.

- Conecte el nodo Filtro generado al nodo Tipo.

Figura 14-5
Selección de campos importantes



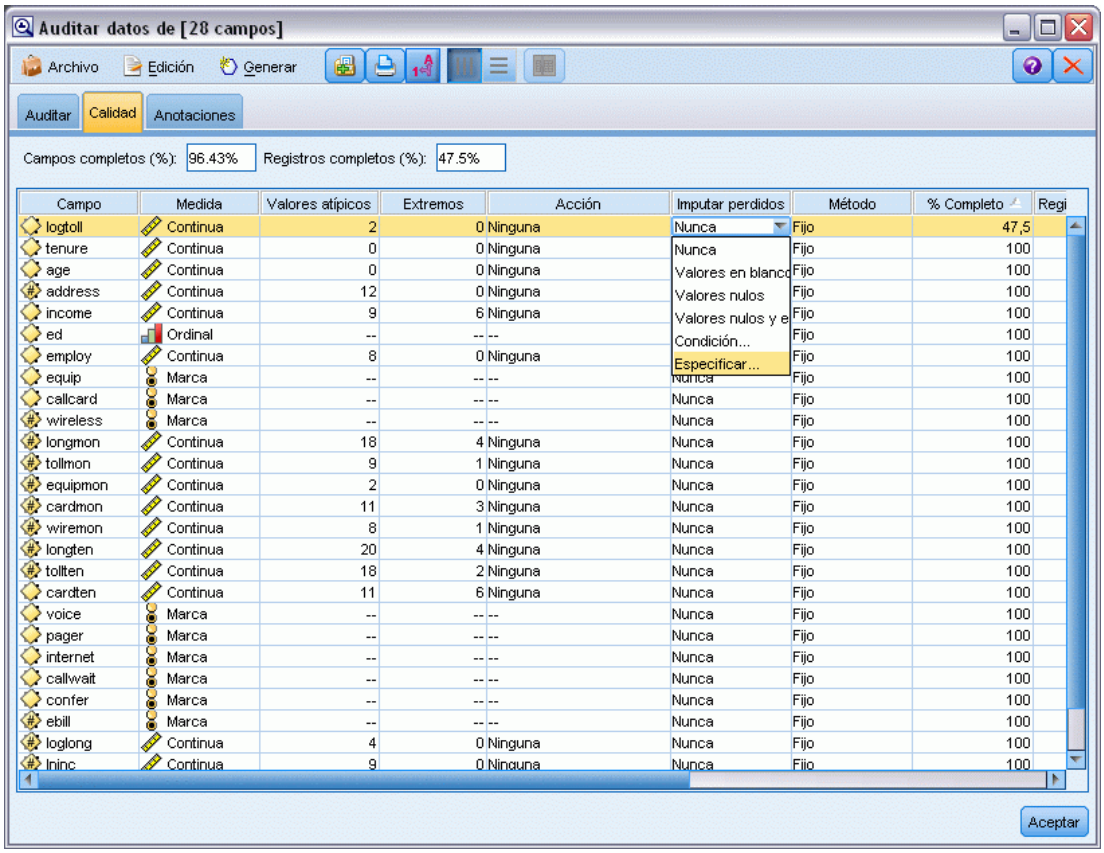
- Conecte al nodo Filtrar generado un nodo Auditar datos.

  Abra el nodo Auditar datos y pulse en Ejecutar.

- En la pestaña Calidad del explorador de auditoría de datos, pulse en la columna *% Completo* para ordenar la columna por orden numérico ascendente. Esto le permite identificar todos los campos que contienen grandes cantidades de datos perdidos. En este caso, el único campo que tiene que corregir es *logtoll*, que está completo en menos de un 50%.

▶ En la columna *Imputar perdidos* de *logtoll*, pulse en Especificar.

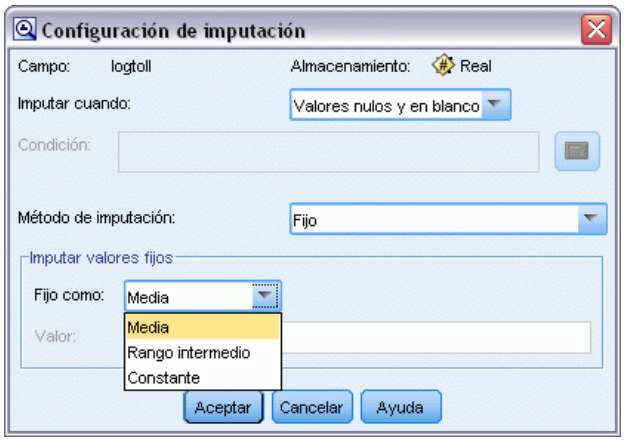
Figura 14-6
Imputación de valores perdidos de logtoll



▶ En Imputar cuando, seleccione Valores vacíos y nulos. En Fijo como, seleccione Media y pulse en Aceptar.

Si selecciona Media, se asegura que los valores imputados no afectan negativamente a la media de todos los valores del conjunto completo de datos.

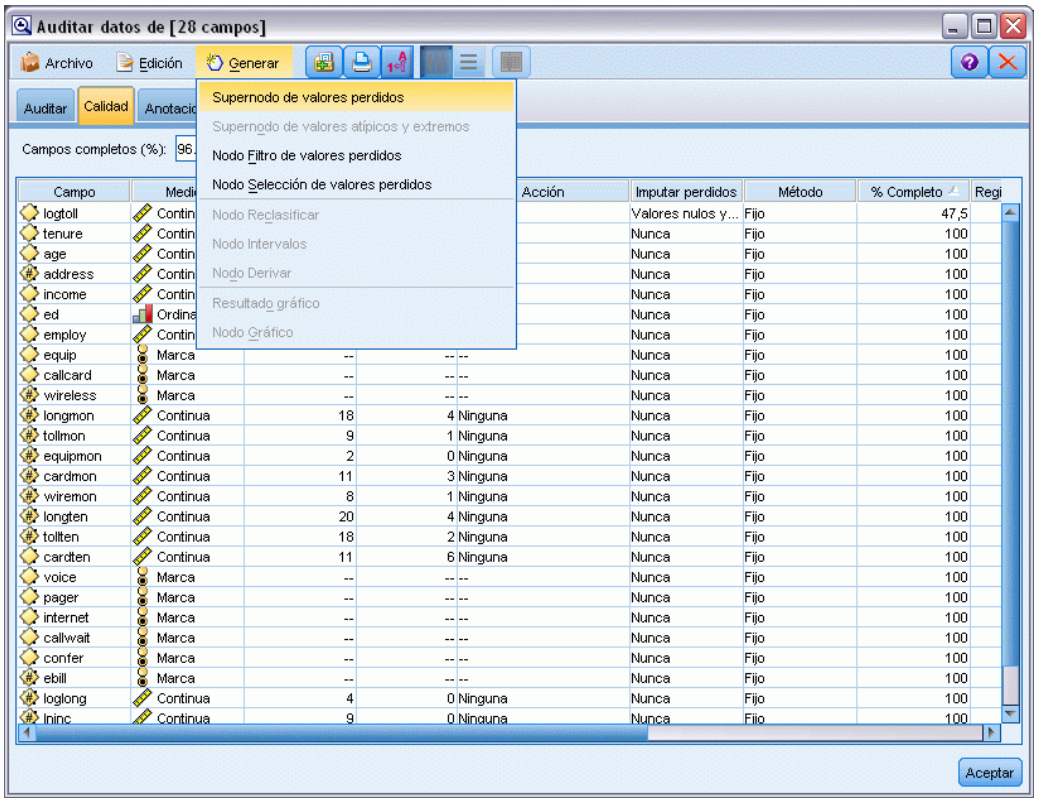
Figura 14-7
Configuración de imputación



- En la pestaña Calidad del explorador de auditoría de datos, genere el Supernodo de valores perdidos. Para ello, elija en los menús:

  Generar > Supernodo de valores perdidos

Figura 14-8
Generación de un Supernodo de valores perdidos

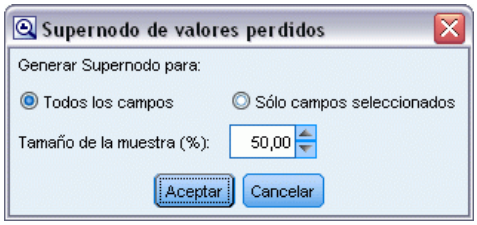


En el cuadro de diálogo Supernodo de valores perdidos, aumente el Tamaño de la muestra al 50% y pulse en Aceptar.

El Supernodo se muestra en el lienzo de rutas, con el título: *Imputación de valores perdidos*.

- Conecte el Supernodo al nodo Filtro.

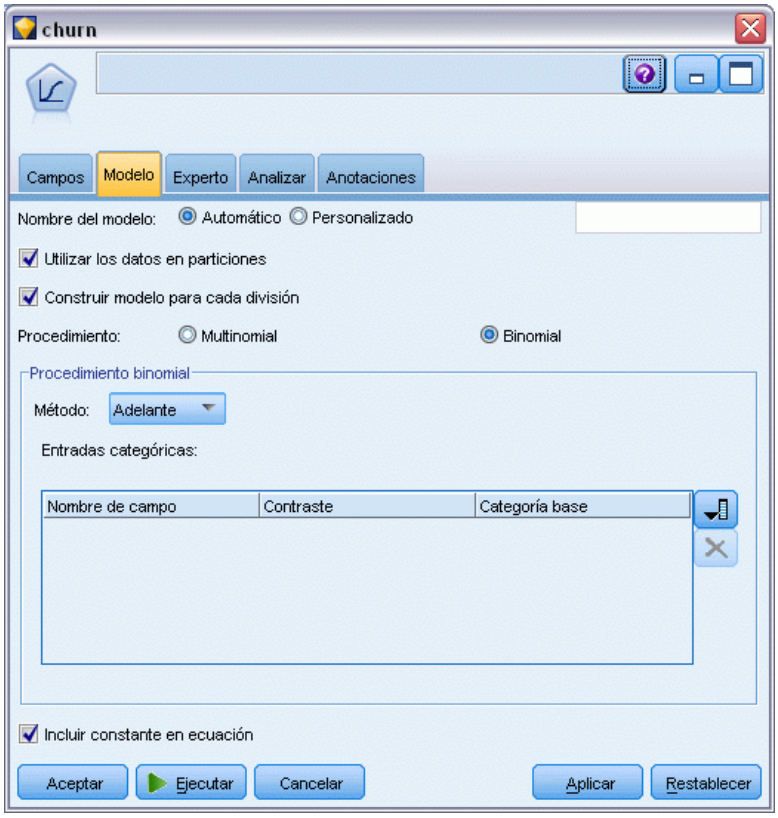
Figura 14-9
Especificación del tamaño de la muestra



- Añada un nodo Logística al Supernodo.

- En el nodo Logística, pulse en la pestaña Modelo y seleccione el procedimiento Binomial. En el área *Procedimiento binomial*, seleccione el método Adelante.

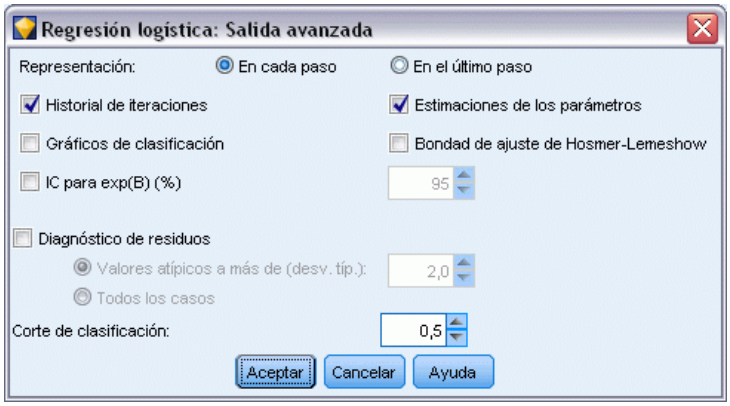
Figura 14-10
Selección de opciones del modelo



- En la pestaña Experto, seleccione el modo Experto y, a continuación, pulse en Resultado. Aparecerá el cuadro de diálogo Salida avanzada.

- En el cuadro de diálogo Salida avanzada, seleccione En cada paso como tipo de *Representación*. Seleccione Historial de iteraciones y Estimaciones de los parámetros y pulse en Aceptar.

Figura 14-11
Selección de opciones de salida



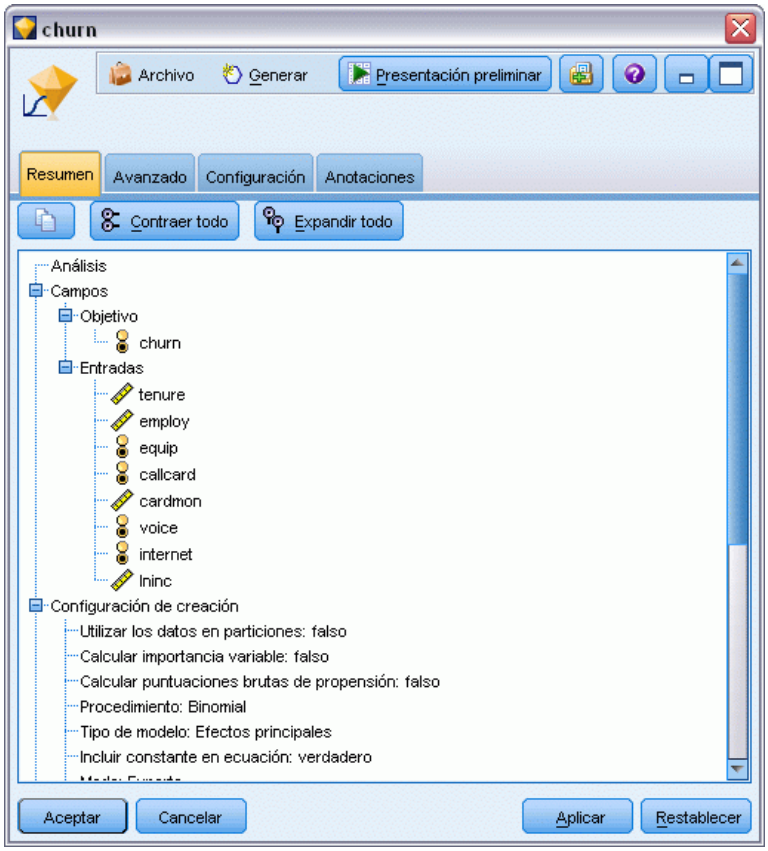
Exploración del modelo

- En el nodo Logística, pulse en Ejecutar para crear el modelo.

El nugget del modelo se añade al lienzo de rutas y a la paleta Modelos en la esquina superior derecha. Para ver los detalles, pulse con el botón derecho en el nugget de modelo y seleccione Editar o Examinar.

La pestaña Resumen muestra (entre otras cosas) el objetivo y las entradas (campos predictores) que utiliza el modelo. Observe que éstos son los campos que se eligieron según el método Adelante, no la lista completa enviada para tener en cuenta.

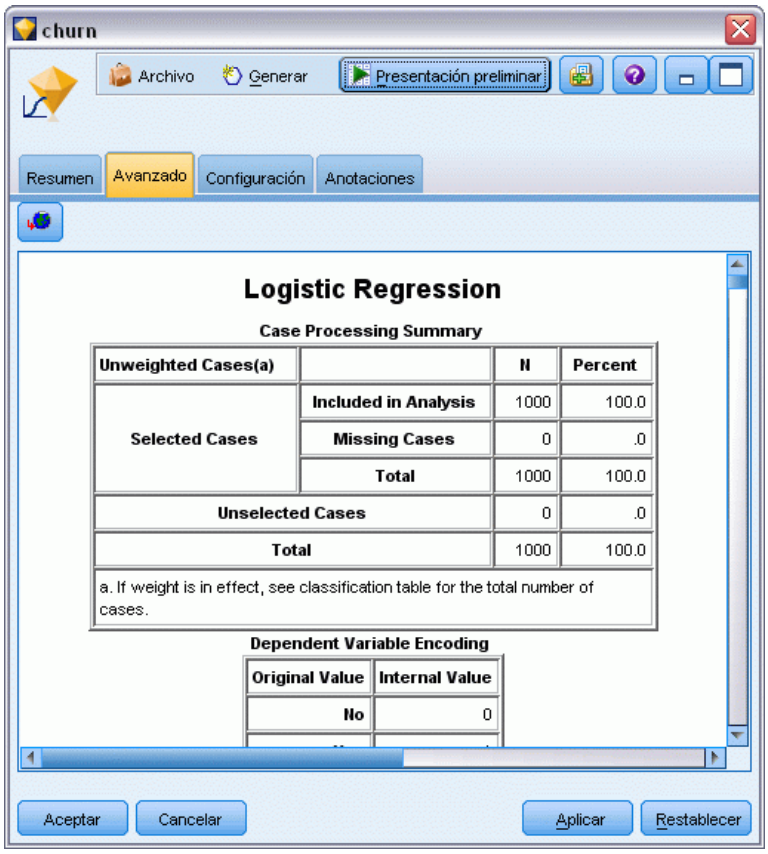
Figura 14-12
Resumen del modelo en el que se ven los campos Objetivo y Entrada



Los elementos que se muestran en la pestaña Avanzado dependen de las opciones seleccionadas en el cuadro de diálogo Salida avanzada del nodo Logística. Un elemento que siempre se muestra es el resumen de procesamiento de casos, que indica el número y el porcentaje de los registros que

se incluyen en el análisis. Además, muestra el número de casos perdidos (si los hay) en los que uno o varios campos de entrada no están disponibles y los casos que no se seleccionaron.

Figura 14-13
Resumen del procesamiento de los casos

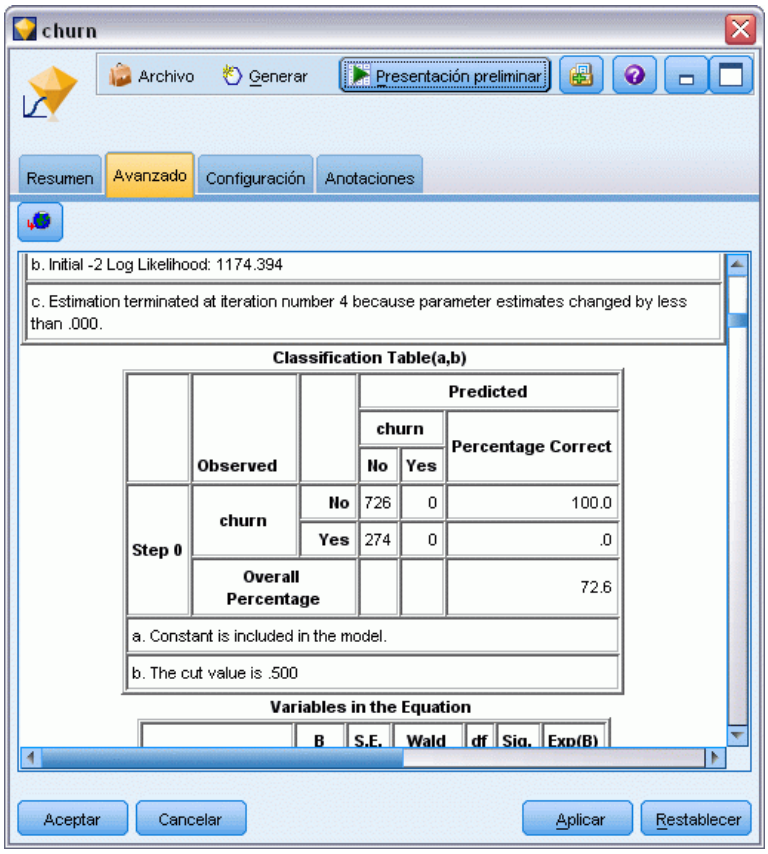


► Desplácese hacia abajo en el Resumen de procesamiento de casos para mostrar la Tabla de clasificación que se encuentra bajo Bloque 0: Bloque de comienzo.

El método Pasos sucesivos hacia adelante comienza con un modelo nulo (es decir, un modelo sin predictores) que se puede utilizar como base para comparar con el modelo final construido. Por convención, el modelo nulo lo pronostica todo como 0, por lo que el modelo nulo tiene una precisión del 72,6% sólo porque se pronostican correctamente los 726 clientes que no se pasaron a

la competencia. Sin embargo, los clientes que sí se pasaron a la competencia no se pronostican de manera correcta en absoluto.

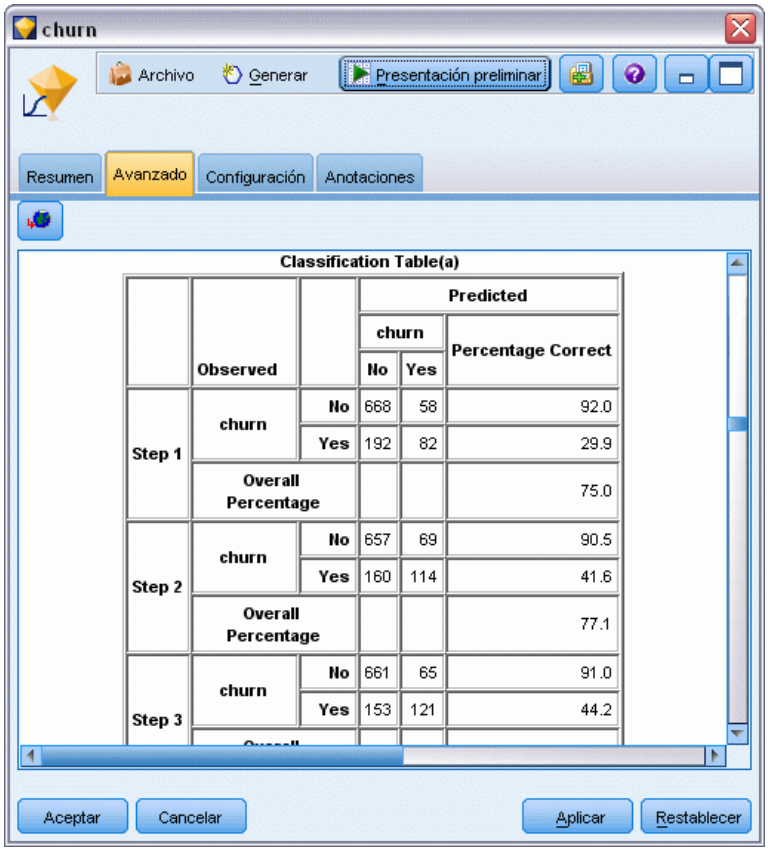
Figura 14-14
Inicio de tabla de clasificación: Bloque 0



- Desplácese hacia abajo para mostrar la Tabla de clasificación que se encuentra bajo Bloque 1: Método = Pasos sucesivos hacia adelante.

Esta tabla de clasificación muestra los resultados de su modelo a medida que se añade un predictor en cada paso. Ya en el primer paso (después de haber utilizado sólo un predictor) el modelo ha aumentado la precisión de la predicción de pérdida de clientes del 0,0% al 29,9%.

Figura 14-15
Tabla de clasificación: Bloque 1



Classification Table(a)

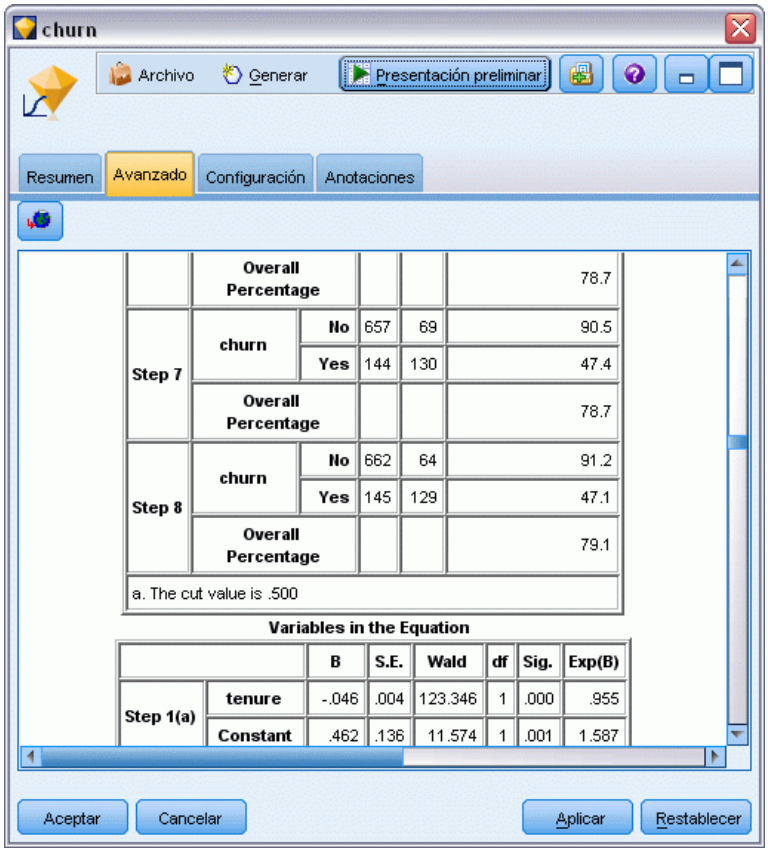
	Observed		Predicted		
			churn		Percentage Correct
			No	Yes	
Step 1	churn	No	668	58	92.0
		Yes	192	82	29.9
	Overall Percentage				75.0
Step 2	churn	No	657	69	90.5
		Yes	160	114	41.6
	Overall Percentage				77.1
Step 3	churn	No	661	65	91.0
		Yes	153	121	44.2

▶ Desplácese hasta la parte inferior de esta tabla de clasificación.

La tabla de clasificación muestra que el último paso es el número 8. En esta etapa, el algoritmo ha decidido que ya no tiene que añadir más predictores al modelo. Pese a que la precisión de los clientes que no se pasan a la competencia ha disminuido ligeramente hasta el 91.2%, la precisión

de la predicción de los que sí lo han hecho ha aumentado del 0% inicial al 47,1%. Esta es una importante mejora con respecto al modelo nulo original que no utilizaba predictores.

Figura 14-16
Tabla de clasificación: Bloque 1



	Overall Percentage				78.7
Step 7	churn	No	657	69	90.5
		Yes	144	130	47.4
	Overall Percentage				78.7
Step 8	churn	No	662	64	91.2
		Yes	145	129	47.1
	Overall Percentage				79.1

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	tenure	-.046	.004	123.346	1	.000	.955
	Constant	.462	.136	11.574	1	.001	1.587

Para un cliente que quiere disminuir la cantidad de clientes que pierde, una reducción a casi la mitad es un paso muy importante para proteger su flujo de ingresos.

Nota: este ejemplo también demuestra que utilizar el porcentaje global como guía de la precisión de un modelo puede ser equívoco en algunos casos. El modelo nulo original tenía una precisión general del 72,6%, mientras que el modelo final pronosticado tiene una precisión general del 79.1%. Sin embargo, como hemos visto, la precisión de las predicciones de categorías individuales era ampliamente diferente.

Para evaluar cómo se ajusta el modelo a los datos, en el cuadro de diálogo Salida avanzada hay disponibles varios diagnósticos cuando se está construyendo el modelo. Si desea obtener más información, consulte el tema Resultado avanzado del nugget de modelo logístico en el capítulo 10 en *Nodos de modelado de IBM SPSS Modeler 14.2*. Puede encontrar explicaciones de los fundamentos matemáticos de los métodos de modelado utilizados en IBM® SPSS® Modeler en el *Manual de algoritmos de SPSS Modeler*, disponible en el directorio *\Documentation* del disco de instalación.

Recuerde que estos resultados están basados sólo en los datos de entrenamiento. Para evaluar qué tal se extiende el modelo a otros datos de casos reales, se utilizaría un nodo de partición para reservar un subconjunto de registros para comprobación y validación. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 14.2*.

Capítulo
15

Predicción del uso de la banda ancha (serie temporal)

Predicciones con el nodo Serie temporal

Un analista que trabaja para un proveedor de banda ancha a nivel nacional debe generar predicciones de las suscripciones de usuarios para predecir la utilización de la banda ancha. Las predicciones se deben realizar para cada uno de los mercados locales que conforman la base nacional de suscriptores. Utilizaremos el modelado de series temporales para generar predicciones acerca de los tres meses siguientes para varios mercados locales. En un segundo ejemplo se muestra cómo puede convertir datos de origen si no están en el formato adecuado para introducirlos en el nodo Serie temporal.

Estos ejemplos usan la ruta llamada *broadband_create_models.str*, que hace referencia al archivo de datos *broadband_1.sav*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *broadband_create_models.str* se encuentra en la carpeta *streams*.

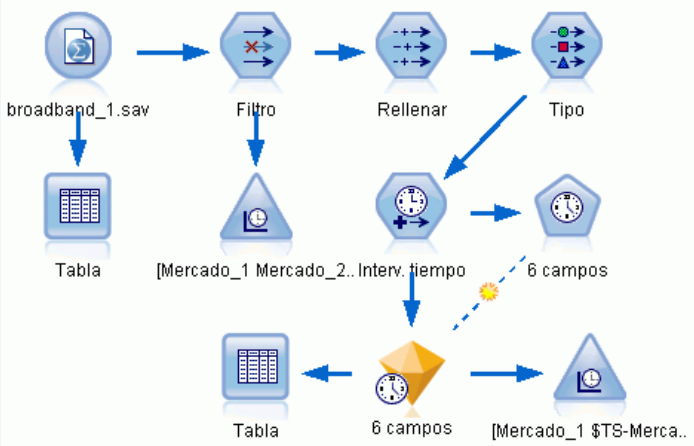
En el último ejemplo se muestra cómo aplicar los modelos guardados a un conjunto de datos actualizado para ampliar las predicciones tres meses más.

En SPSS Modeler, puede generar varios modelos de series temporales en una única operación. El archivo fuente que utilizará tiene datos de series temporales para 85 mercados distintos, aunque por motivos de simplicidad sólo vamos a modelar cinco de éstos y uno total para todos los mercados.

El archivo de datos *broadband_1.sav* tiene datos de uso mensuales para cada uno de los 85 mercados locales. Para este ejemplo, sólo se utilizarán las cinco primeras series; se creará un modelo distinto para cada una de estas series y uno total.

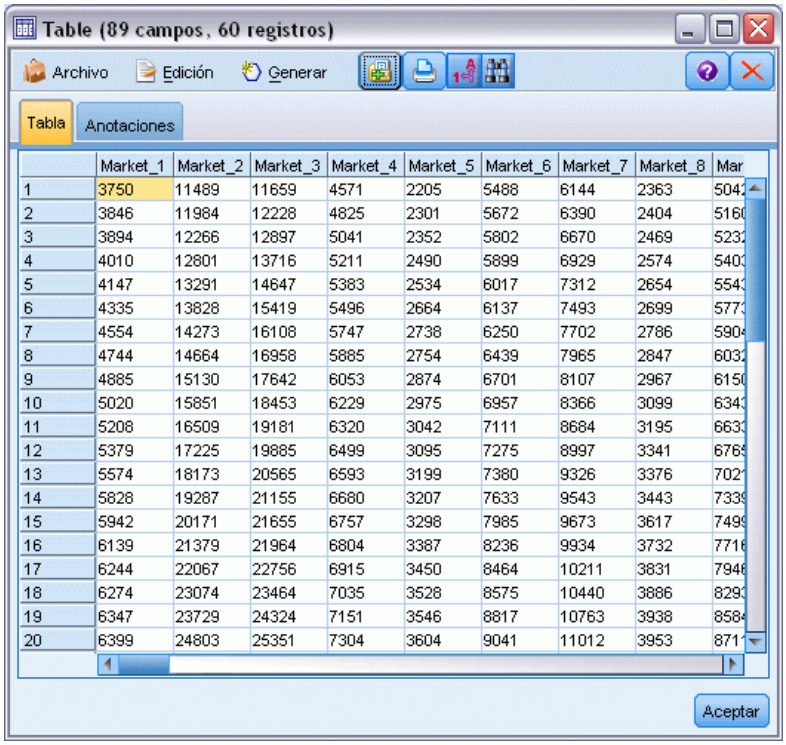
El archivo también incluye un campo de fecha que indica el mes y el año de cada registro. Se usará este campo en un nodo Intervalos de tiempo para etiquetar los registros. SPSS Modeler lee el campo de fecha como si fuera una cadena, por lo que para poder usarlo en SPSS Modeler deberá convertir el tipo de almacenamiento en un formato de fecha numérico mediante un nodo Rellenar.

Figura 15-1
Ruta de ejemplo para mostrar el modelado de series temporales



El nodo Serie temporal exige que cada serie esté en una columna diferente, con una fila para cada intervalo. SPSS Modeler proporciona métodos para transformar los datos de manera que coincidan con este formato si es necesario.

Figura 15-2
Datos de suscripción mensuales para mercados locales de banda ancha



	Market_1	Market_2	Market_3	Market_4	Market_5	Market_6	Market_7	Market_8	Mar
1	3750	11489	11659	4571	2205	5488	6144	2363	504
2	3846	11984	12228	4825	2301	5672	6390	2404	516
3	3894	12266	12897	5041	2352	5802	6670	2469	523
4	4010	12801	13716	5211	2490	5899	6929	2574	540
5	4147	13291	14647	5383	2534	6017	7312	2654	554
6	4335	13828	15419	5496	2664	6137	7493	2699	577
7	4554	14273	16108	5747	2738	6250	7702	2786	590
8	4744	14664	16958	5885	2754	6439	7965	2847	603
9	4885	15130	17642	6053	2874	6701	8107	2967	615
10	5020	15851	18453	6229	2975	6957	8366	3099	634
11	5208	16509	19181	6320	3042	7111	8684	3195	663
12	5379	17225	19885	6499	3095	7275	8997	3341	676
13	5574	18173	20565	6593	3199	7380	9326	3376	702
14	5828	19287	21155	6680	3207	7633	9543	3443	733
15	5942	20171	21655	6757	3298	7985	9673	3617	749
16	6139	21379	21964	6804	3387	8236	9934	3732	771
17	6244	22067	22756	6915	3450	8464	10211	3831	794
18	6274	23074	23464	7035	3528	8575	10440	3886	829
19	6347	23729	24324	7151	3546	8817	10763	3938	858
20	6399	24803	25351	7304	3604	9041	11012	3953	871

Creación de la ruta

- Cree una nueva ruta y añada un nodo de origen de archivo Statistics que apunte a *broadband_1.sav*.
- Use un nodo Filtro para filtrar los campos de *Mercado_6* a *Mercado_85*, así como los campos *MES_* y *AÑO_*, para simplificar el modelo.

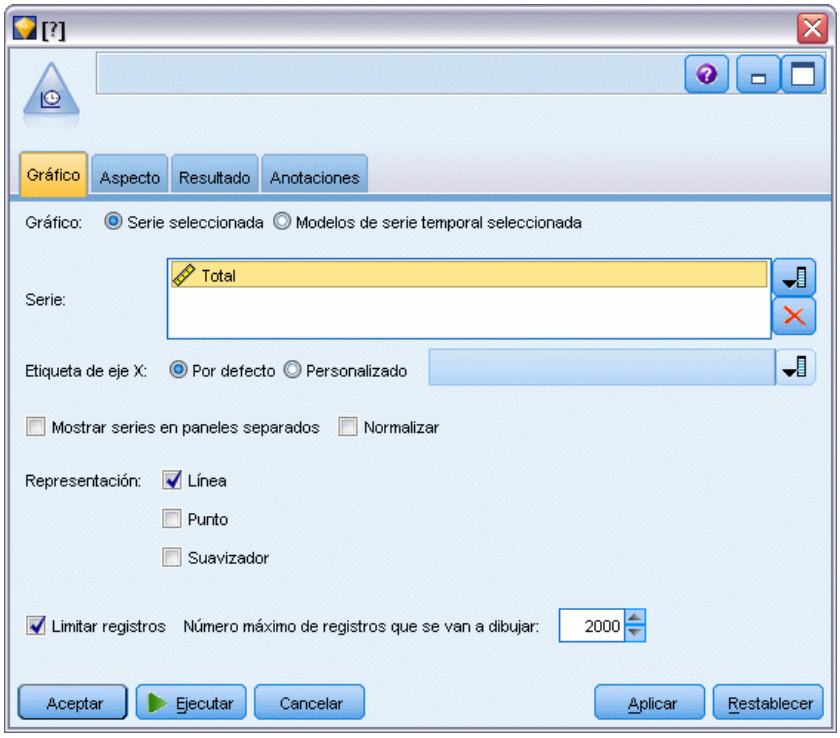
Sugerencia: para seleccionar varios campos adyacentes en una única operación, pulse en el campo *Mercado_6*, mantenga pulsado el botón izquierdo del ratón y arrástrelo hasta el campo *Mercado_85*. Los campos seleccionados se resaltarán en azul. Para añadir los otros campos, mantenga pulsada la tecla Ctrl y pulse en los campos *MES_* y *AÑO_*.

Figura 15-3
Simplificación del modelo

Examen de los datos

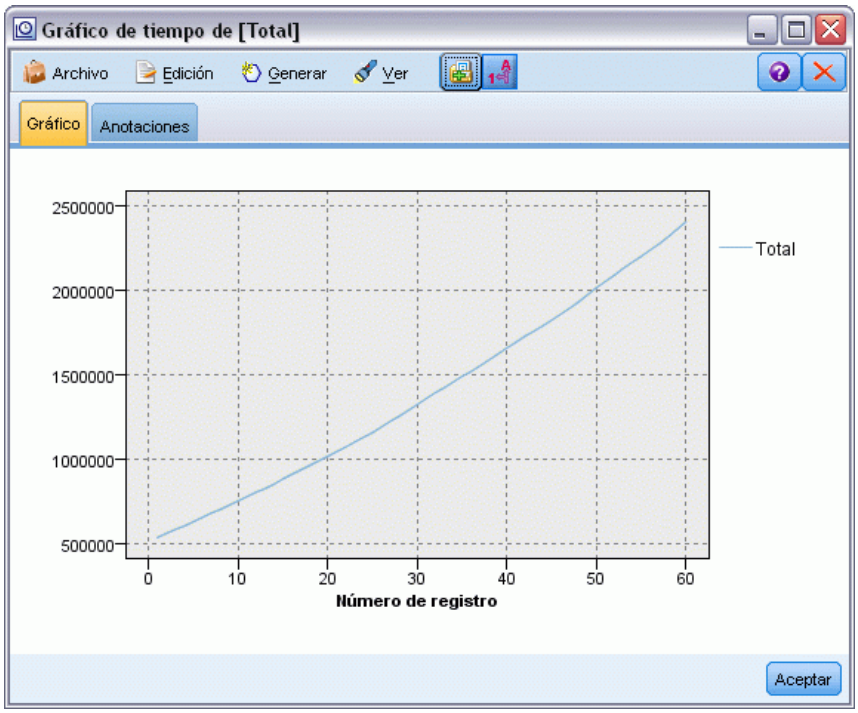
Siempre es conveniente conocer la naturaleza de los datos antes de generar un modelo. ¿Los datos muestran variaciones estacionales? Aunque el modelizador experto puede buscar automáticamente el mejor modelo estacional o no estacional para cada serie, a menudo puede obtener resultados de manera más rápida si limita la búsqueda a modelos no estacionales cuando no haya estacionalidad en los datos. Sin examinar los datos para cada uno de los 85 mercados locales, podemos obtener una imagen aproximada de la presencia o ausencia de estacionalidad al realizar un gráfico del número total de personas suscritas en los cinco mercados.

Figura 15-4
Representación del número total de suscriptores



- En la paleta Gráficos, añada un nodo Gráfico de tiempo al nodo Filtro.
- Añada el campo *Total* a la lista Series.
- Desactive las casillas de verificación Mostrar series en paneles separados y Normalizar.
- Pulse en Ejecutar.

Figura 15-5
Gráfico de tiempo del campo Total

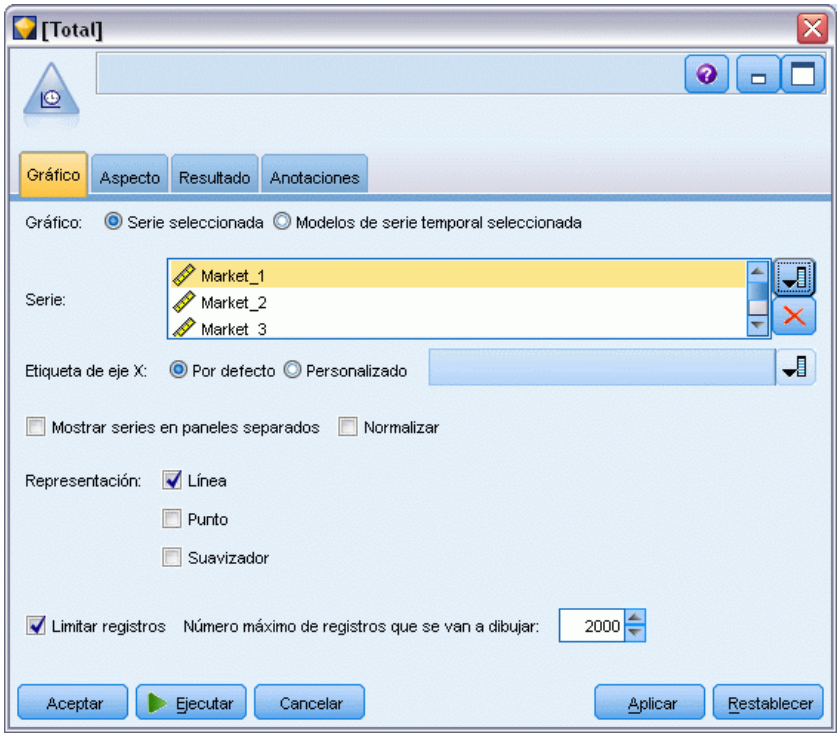


La serie muestra una tendencia ascendente muy suave sin indicios de variaciones estacionales. Puede haber series individuales con estacionalidad, aunque parece que dicha estacionalidad no es una característica prominente de los datos en general.

Por supuesto, debe inspeccionar cada una de las series antes de descartar los modelos estacionales. A continuación, puede separar las series que muestren estacionalidad y realizar sus modelos independientemente.

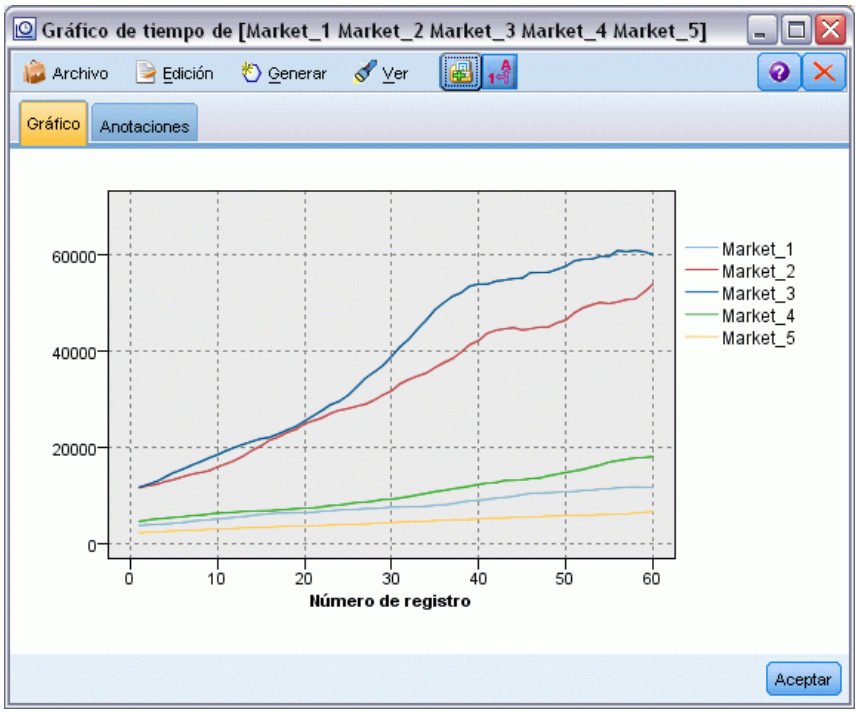
IBM® SPSS® Modeler facilita la representación de varias series a la vez.

Figura 15-6
Representación de varias series temporales



- Vuelva a abrir el nodo Gráfico de tiempo.
- Elimine el campo *Total* de la lista Series (selecciónelo y pulse en el botón X rojo).
- Añada los campos desde *Mercado_1* hasta *Mercado_5* a la lista.
- Pulse en Ejecutar.

Figura 15-7
Gráfico de tiempo de varios campos



El examen de estos mercados revela una tendencia ascendente continua en cada caso. Aunque algunos son un poco más erráticos que otros, no presentan muestras de estacionalidad.

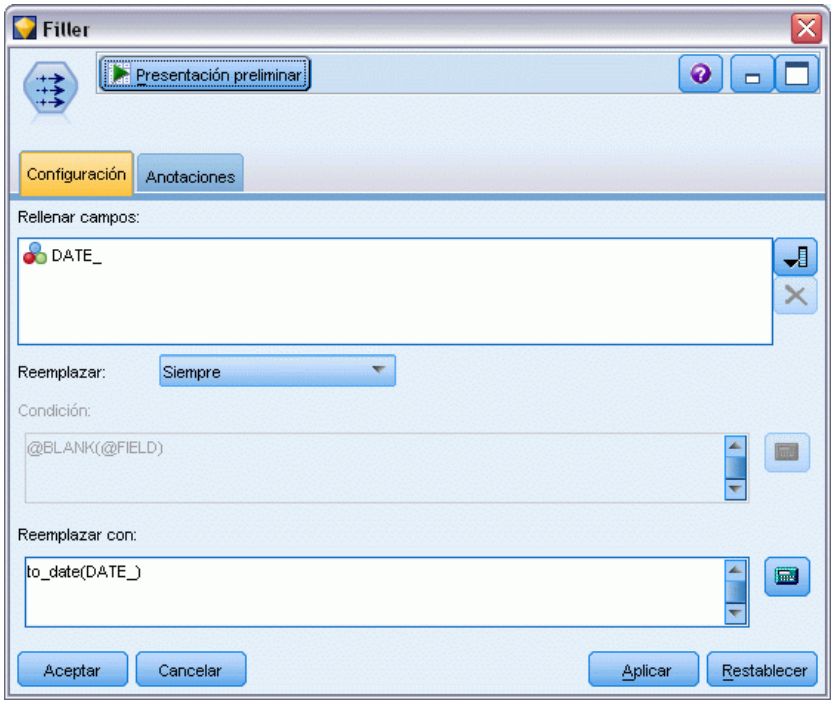
Definición de las fechas

Ahora tiene que cambiar el tipo de almacenamiento del campo *DATE_* al formato de fecha.

- Conecte un nodo Rellenar al nodo Filtro.
- Abra el nodo Rellenar y pulse en el botón selector de campos.
- Seleccione DATE_ para añadirlo a Rellenar campos.
- Defina la condición Reemplazar en Siempre.

▶ Defina el valor de Reemplazar con en to_date(FECHA_).

Figura 15-8
Configuración del tipo de almacenamiento de fecha

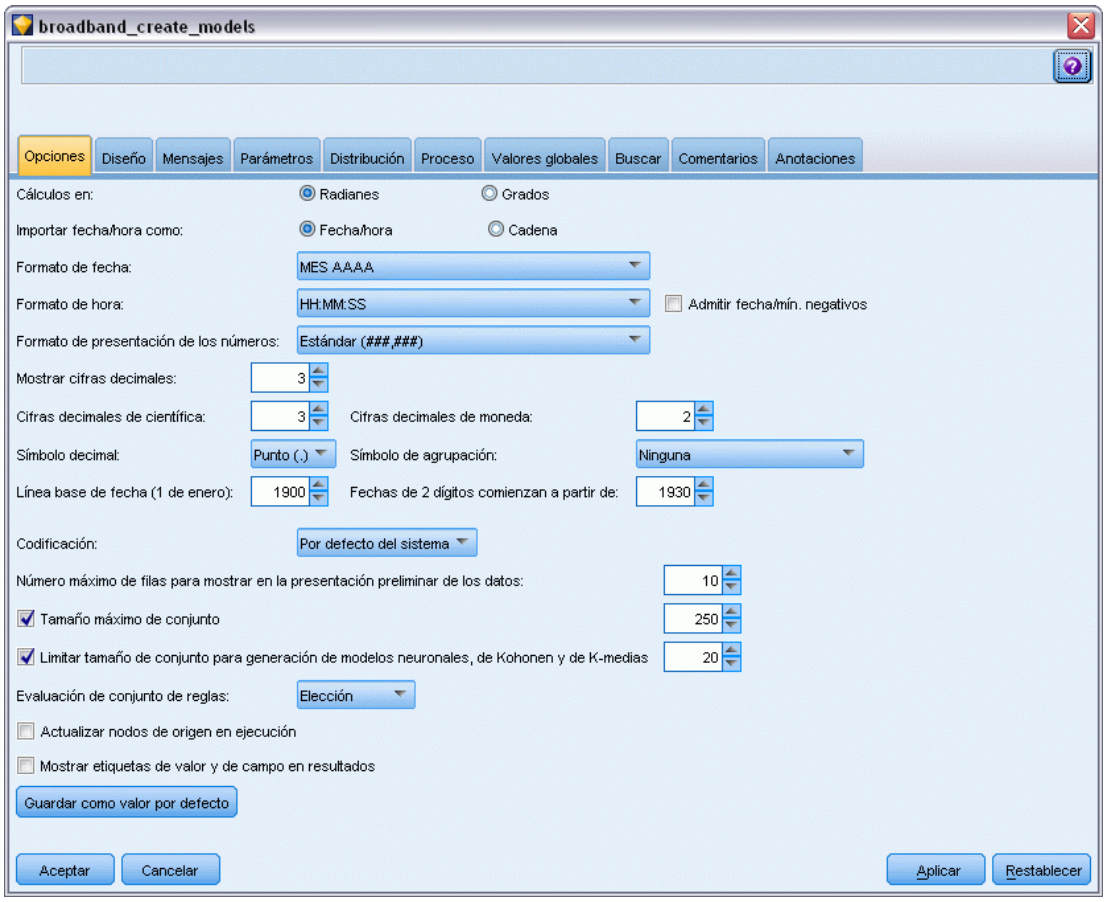


Cambie el formato de fecha por defecto para que coincida con el formato del campo Fecha. Esto es necesario para que la conversión del campo Fecha se lleve a cabo como se esperaba.

▶ En el menú, seleccione Herramientas > Propiedades de ruta > Opciones para abrir el cuadro de diálogo de opciones de rutas.

▶ Defina el formato de fecha por defecto en MES AAAA.

Figura 15-9
Configuración del formato de fecha



Definición de los objetivos

▶ Añada un nodo Tipo para definir el papel del campo DATE_ en *Ninguna*. Defina el papel a Objetivo en el resto de campos (los campos *Mercado_n* y el campo *Total*).

▶ Pulse en el botón Leer valores para rellenar la columna.

Figura 15-10
Definición del papel de varios campos

Configuración del intervalo de tiempo

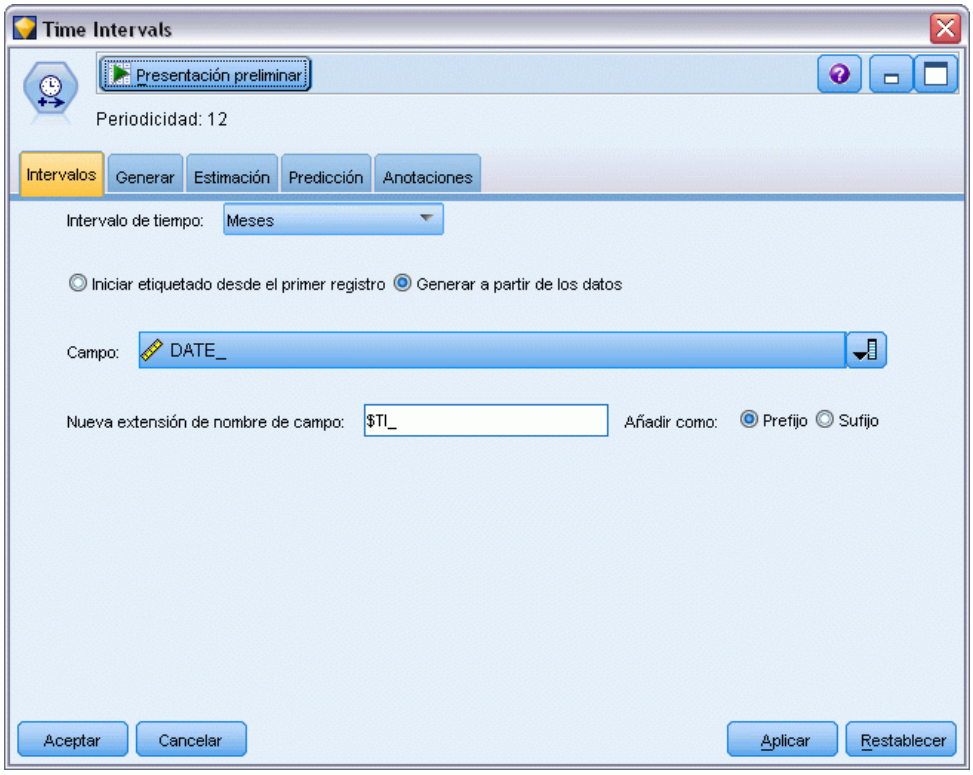
▶ Añada un nodo Intervalos de tiempo (en la paleta de operaciones con campos).

▶ En la pestaña Intervalos, seleccione Meses como intervalo de tiempo.

▶ Seleccione la opción Generar a partir de los datos.

▶ Seleccione DATE_ como campo de generación.

Figura 15-11
Configuración del intervalo de tiempo

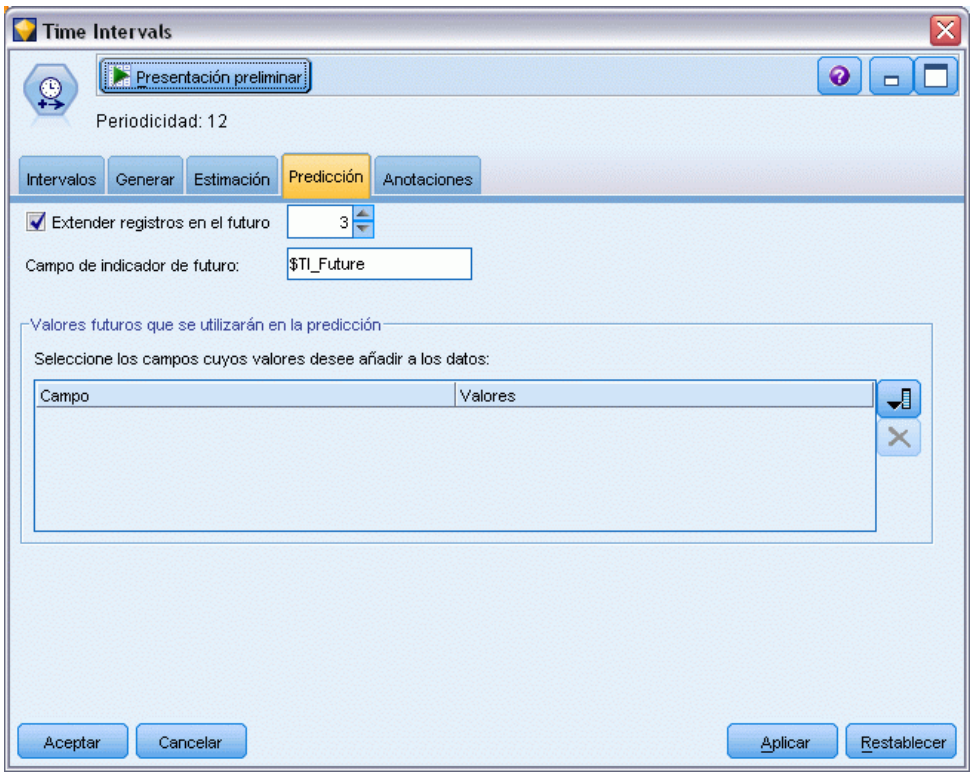


▶ En la pestaña Predicción, seleccione la casilla de verificación Extender registros en el futuro.

▶ Defina el valor en 3.

▶ Pulse en Aceptar.

Figura 15-12
Configuración del período de predicción

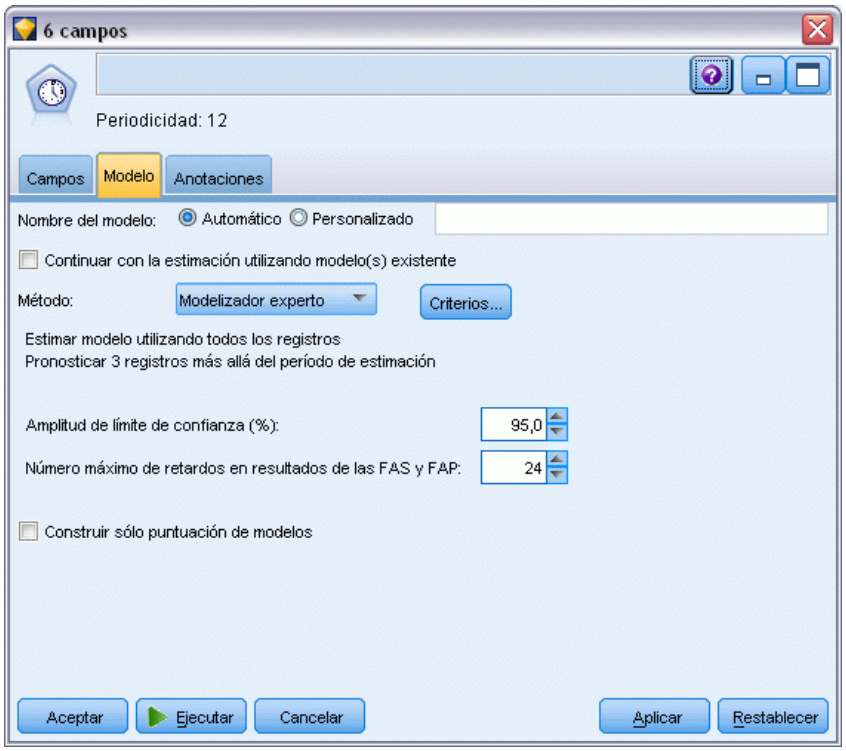


Creación del modelo

▶ En la paleta de modelado, añada un nodo Serie temporal a la ruta y conéctelo con el nodo Intervalos de tiempo.

- Pulse en Ejecutar en el nodo Intervalos de tiempo con toda la configuración por defecto. De esta forma se activa el modelizador experto para decidir cuál es el modelo más adecuado para cada serie temporal.

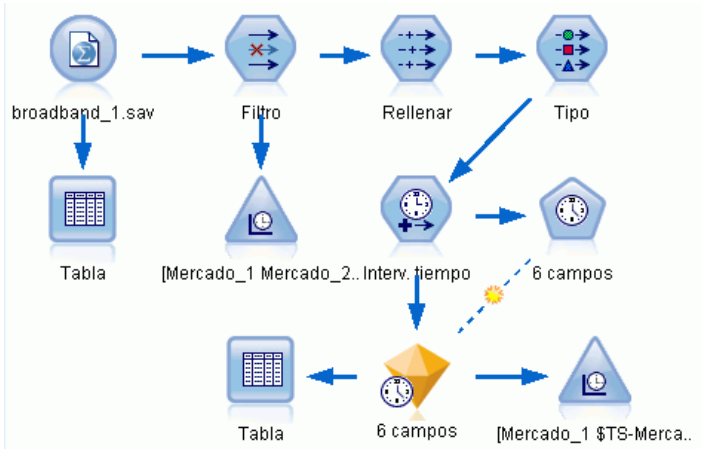
Figura 15-13
Selección del modelizador experto para series temporales



- Añada el nugget de modelo de serie temporal al nodo Intervalos de tiempo.

- Conecte un nodo Tabla al modelo de serie temporal y pulse en Ejecutar.

Figura 15-14
Ruta de ejemplo para mostrar el modelado de series temporales



Ahora hay tres nuevas filas (de la 61 a la 63) añadidas a los datos originales. Éstas son las filas para el período de predicción, en este caso de enero a marzo de 2004.

También hay varias columnas nuevas: varias columnas *$TI_*, añadidas por el nodo Intervalos de tiempo, y las columnas *$TS-*, añadidas por el nodo Serie temporal. Las columnas indican lo siguiente para cada fila (esto es, cada intervalo de los datos de las series temporales):

Columna	**Descripción**
$TI_ÍndiceTiempo	Valor de índice del intervalo de tiempo para esta fila.
$TI_EtiquetaTiempo	Etiqueta del intervalo de tiempo para esta fila.
$TI_Año	Indicadores de mes y año para los datos generados en esta fila.
$TI_Mes	
$TI_Recuento	Número de registros implicados en la determinación de nuevos datos para esta fila.
$TI_Futuro	Indica si esta fila contiene datos de predicciones.
$TS-*nombrecol*	Datos del modelo generado para cada columna de datos originales.
$TSLCI-*nombrecol*	Valor del intervalo de confianza inferior para cada columna de datos del modelo generado.
$TSUCI-*nombrecol*	Valor del intervalo de confianza superior para cada columna de datos del modelo generado.
$TS-Total	Total de los valores de $TS-*nombrecol* de esta fila.
$TSLCI-Total	Total de los valores de $TSLCI-*nombrecol* de esta fila.
$TSUCI-Total	Total de los valores de $TSUCI-*nombrecol* de esta fila.

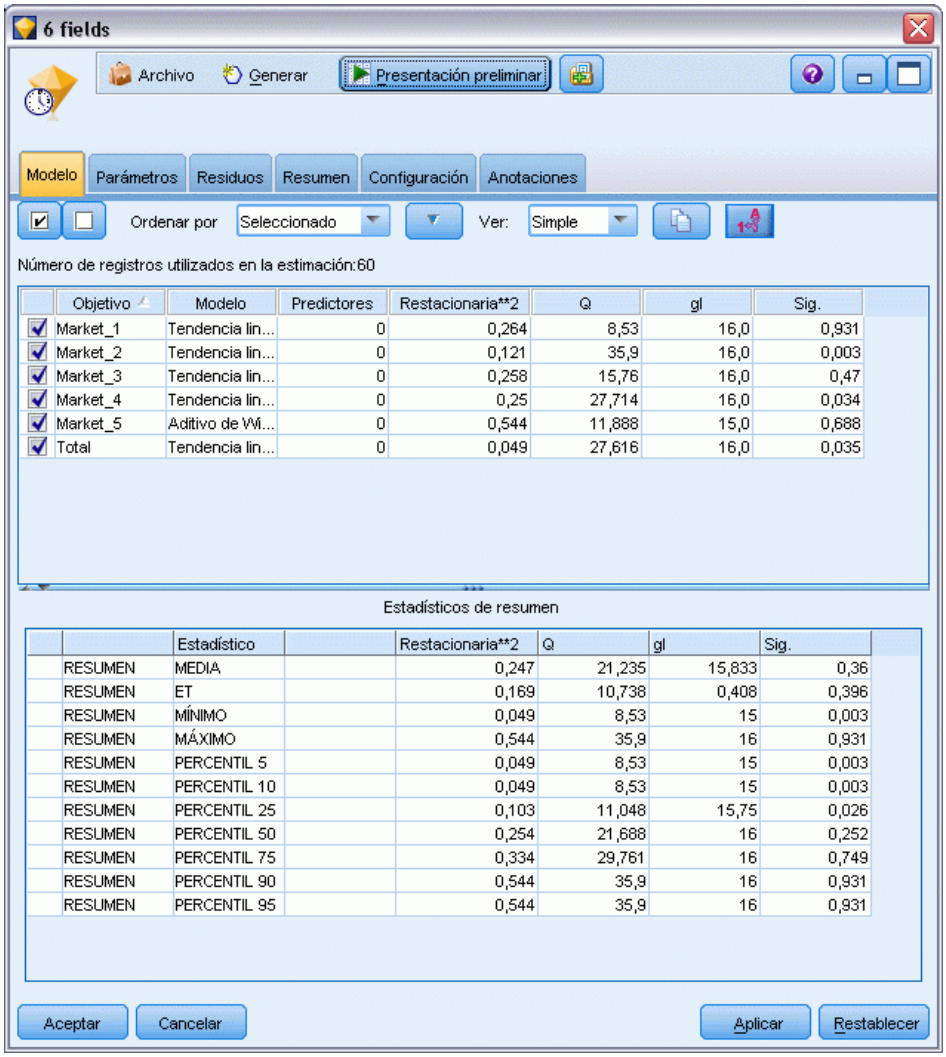
Las columnas de mayor relevancia para la operación de predicción son *$TS-Mercado_n*, *$TSLCI-Mercado_n* y *$TSUCI-Mercado_n*. En concreto, estas columnas contienen en las filas desde la 61 hasta la 63 los datos de predicciones de suscripciones de usuarios y los intervalos de confianza para cada mercado local.

Examen del modelo

- Pulse dos veces en el nugget de modelo de serie temporal para mostrar datos de los modelos generados para cada mercado.

Observe cómo el modelizador experto ha decidido crear un tipo distinto de modelo para Mercado 5 respecto al tipo que ha generado para el resto de mercados.

Figura 15-15
Modelos de series temporales generados para los mercados



La columna Predictores muestra cuántos campos se han usado como predictores para cada objetivo (en este caso, ninguno).

El resto de columnas de esta vista muestra varias medidas de bondad de ajuste para cada modelo. La columna Restacionaria**2 muestra el valor de *R* cuadrado estacionaria. Este estadístico proporciona una estimación de la proporción de la variación total de la serie que se explica con el modelo. Cuanto mayor sea el valor (hasta un máximo de 1.0), mejor se ajustará el modelo.

Las columnas Q, df y Sig. relacionan el estadístico de Box-Ljung, una prueba de la aleatoriedad de los errores residuales en el modelo. Cuanto más aleatorios sean los errores, más posibilidades hay de que sea un buen modelo. Q es el estadístico de Box-Ljung, mientras que df (grados de

libertad) muestra el número de parámetros del modelo que pueden variar libremente cuando estiman un objetivo concreto.

La columna Sig. ofrece el valor de significación del estadístico de Box-Ljung, que aporta otra indicación de si el modelo se ha especificado correctamente. Un valor de significación inferior a 0,05 indica que los errores residuales no son aleatorios, lo que implica que existe una estructura en la serie observada que el modelo no explica.

Considerando los valores *R* cuadrado estacionaria y significación, los modelos que el modelizador experto ha seleccionado para *Mercado_1*, *Mercado_3* y *Mercado_5* son muy aceptables. Los valores Sig. de *Mercado_2* y *Mercado_4* son inferiores a 0,05, lo que indica que puede ser necesario experimentar con modelos que se ajusten mejor a estos mercados.

Los valores de resumen que aparecen en la parte inferior de la representación proporcionan información sobre la distribución de los estadísticos en todos los modelos. Por ejemplo, el valor medio de *R* cuadrado estacionaria de todos los modelos es 0,247, mientras que el mínimo de ese valor es 0,049 (el del modelo *Total*) y, el máximo, 0,544 (valor de *Mercado_5*).

ET denota el error típico en todos los modelos de cada estadístico. Por ejemplo, el error típico del valor de *R* cuadrado estacionaria en todos los modelos es 0,169.

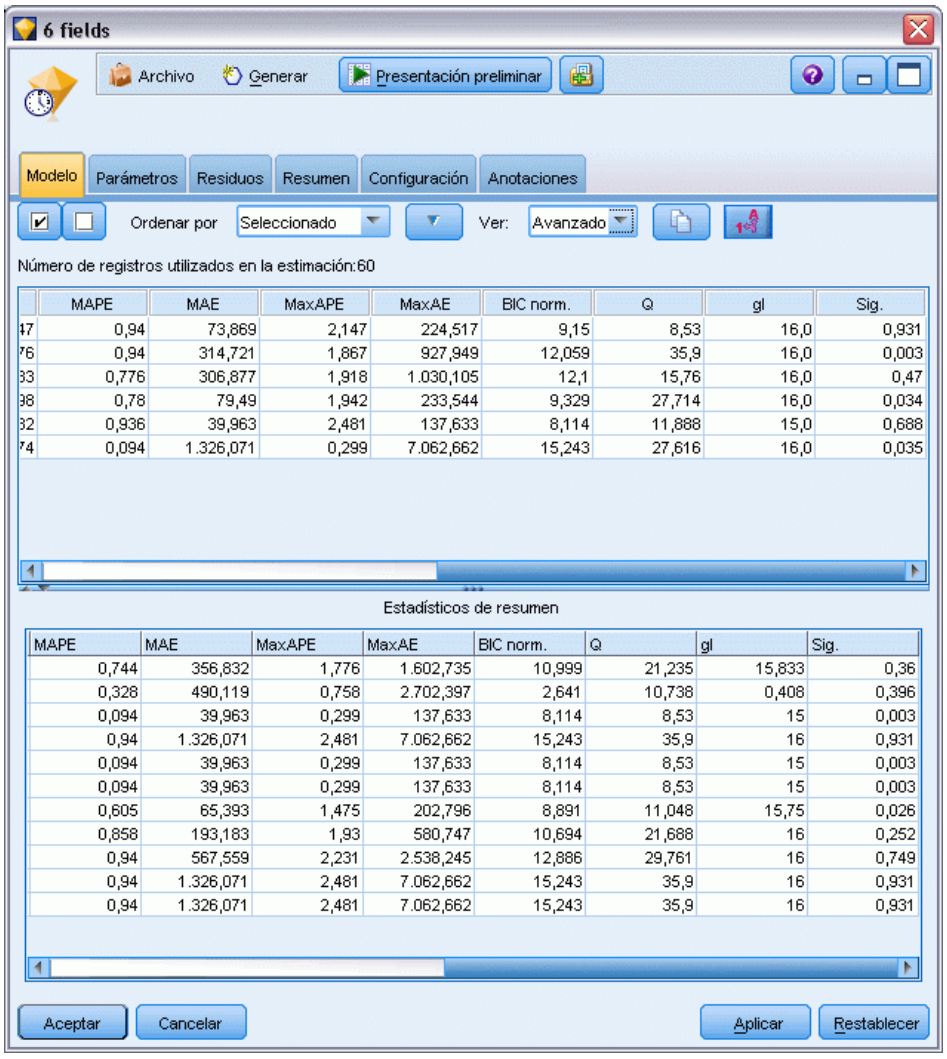
La sección de resumen también incluye valores de percentiles que ofrecen información sobre la distribución de los estadísticos en todos los modelos. Para cada percentil, ese porcentaje de modelos tiene un valor del estadístico de ajuste por debajo del valor establecido.

Así, por ejemplo, sólo el 25% de los modelos tienen un valor de *R* cuadrado estacionaria inferior a 0,121.

- Pulse en la lista desplegable Ver y seleccione Avanzado.

La representación muestra varias medidas adicionales de bondad de ajuste. R**2 es el valor *R* cuadrado, una estimación de la variación total en una serie temporal que se puede explicar mediante el modelo. Como el valor máximo de la estadística es 1,0, los modelos adecuados en este sentido.

Figura 15-16
Representación avanzada de modelos de series temporales



RMSE es el raíz del error cuadrático promedio, una medida que indica cuánto difieren los valores reales de una serie de los valores pronosticados por el modelo, y se expresa en las mismas unidades que las utilizadas para las series. Como se trata de una medición de un error, es deseable que este valor sea el menor posible. A primera vista, parece que los modelos de *Mercado_2* y *Mercado_3*, son aceptables según las estadísticas que se han obtenido hasta ahora, si bien son menos precisas que las obtenidas para los otros tres mercados.

Estas medidas de bondad de ajuste adicionales incluyen los errores absolutos porcentuales promedio (MAPE y MaxAPE). El error absoluto porcentual mide lo que varía una serie objetivo respecto al nivel pronosticado por el modelo, expresado como un valor de porcentaje. Al examinar la media y el máximo en todos los modelos, puede obtener una indicación de la incertidumbre de las predicciones.

El valor MAPE muestra que todos los modelos muestran una media de incertidumbre inferior al 1%, que es un valor muy bajo. El valor MaxAPE muestra el error absoluto máximo porcentual y resulta útil para imaginar un escenario del peor de los casos para las predicciones. Muestra que el error porcentual más grande de cada modelo pertenece al rango comprendido entre 1,8 y 2,5% aproximadamente, de nuevo unos valores muy bajos.

MAE el valor (error absoluto medio) muestra la media de los valores absolutos de los errores de predicción. Al igual que el valor RMSE, se expresa en las mismas unidades que las empleadas para las series. MaxAE muestra el mayor error pronosticado en las mismas unidades e indica el peor de los casos para las predicciones.

Aunque estos valores absolutos son interesantes, también lo son los valores de los errores de porcentaje (MAPE y MaxAPE) que son más útiles en este caso, ya que las series objetivo representan los números de suscriptores para mercados de tamaños distintos.

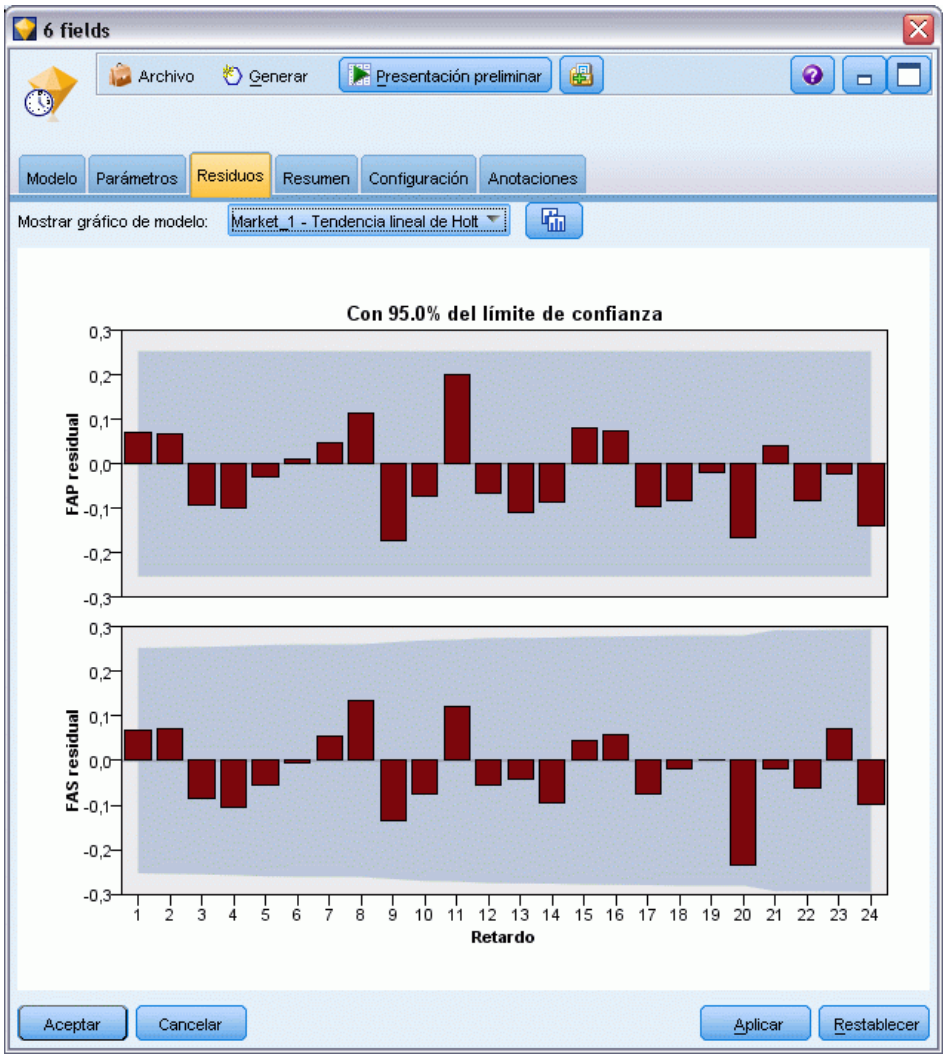
¿Los valores MAPE y MaxAPE representan una cantidad aceptable de incertidumbre con los modelos? Son verdaderamente muy bajos. En situaciones como ésta, entra en escena el sentido común empresarial, ya que el riesgo aceptable irá cambiando según el problema. Asumiremos que los estadísticos de bondad de ajuste están dentro de los límites aceptables y continuaremos observando los errores residuales.

Examinar los valores de las funciones de autocorrelación (FAS) y las autocorrelación parcial (FAP) de los residuos del modelo ayuda a comprender los modelos mejor que si sólo se consultan los estadísticos de bondad de ajuste.

Un modelo de serie temporal bien especificada capturará todas las variaciones no aleatorias, incluyendo estacionalidad, tendencia o cíclica y otros factores importantes. En este caso, un error no se debe correlacionar con sí mismo (autocorrelacionado) con el tiempo. Una estructura significativa en alguna de las funciones de correlación implicaría que el modelo subyacente está incompleto.

- Pulse en la pestaña Residuos para ver los valores de la función de autocorrelación (FAS) y la función de autocorrelación parcial (FAP) de los errores residuales del modelo del primer mercado local.

Figura 15-17
Valores de FAS y FAP de los mercados



En estos gráficos, los valores originales del error variable se han retardado en periodos de 24 horas y se comparan con el valor original para ver si existirá algún tipo de correlación con el tiempo. Para que el modelo sea aceptable, ninguna de las barras del gráfico superior (FAS) se debe extender fuera del área sombreada, en una dirección positiva (arriba) o negativa (abajo).

En este caso, debe comprobar el gráfico inferior (FAP) para ver si la estructura se confirma. El gráfico FAP controla las correlaciones después de controlar los valores de las series en los puntos temporales intercalados.

Los valores de *Mercado_1* están en el área sombreada, por lo que podemos continuar y comprobar los valores del resto de mercados.

- Pulse en la lista desplegable Mostrar gráfico de modelo para ver estos valores en el resto de mercados y los totales.

  Los valores de *Mercado_2* y *Mercado_4* son una posible causa de preocupación, confirmando nuestras sospechas de sus valores Sig.. Necesitamos experimentar con algunos modelos diferentes en esos mercados en algunos puntos para ver si podemos obtener mejores resultados, pero para el resto de este ejemplos, nos concentraremos en lo que podemos aprender del modelo *Mercado_1*.

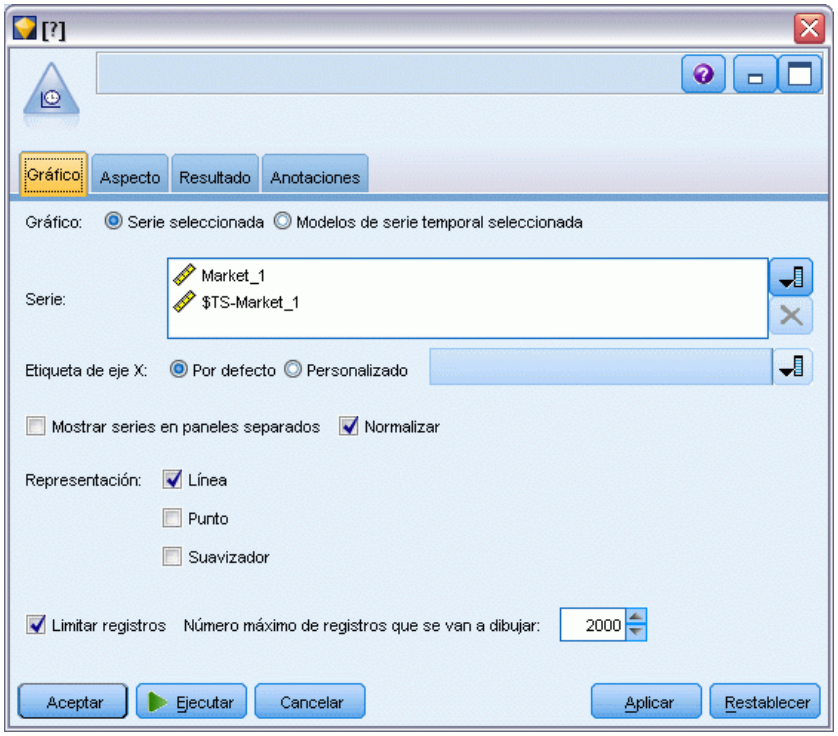
- En la paleta Gráficos, añada un nodo Gráfico de tiempo al nugget de modelo Serie temporal.

- En la pestaña Gráfico, desactive la casilla de verificación Mostrar series en paneles separados.

- En la lista Serie, pulse en el botón selector de campos, seleccione los campos *Mercado_1* y *$TS-Mercado_1*, y pulse en Aceptar para añadirlos a la lista.

- Pulse en Ejecutar para ver un gráfico de líneas de los campos reales y de predicciones del primer mercado local.

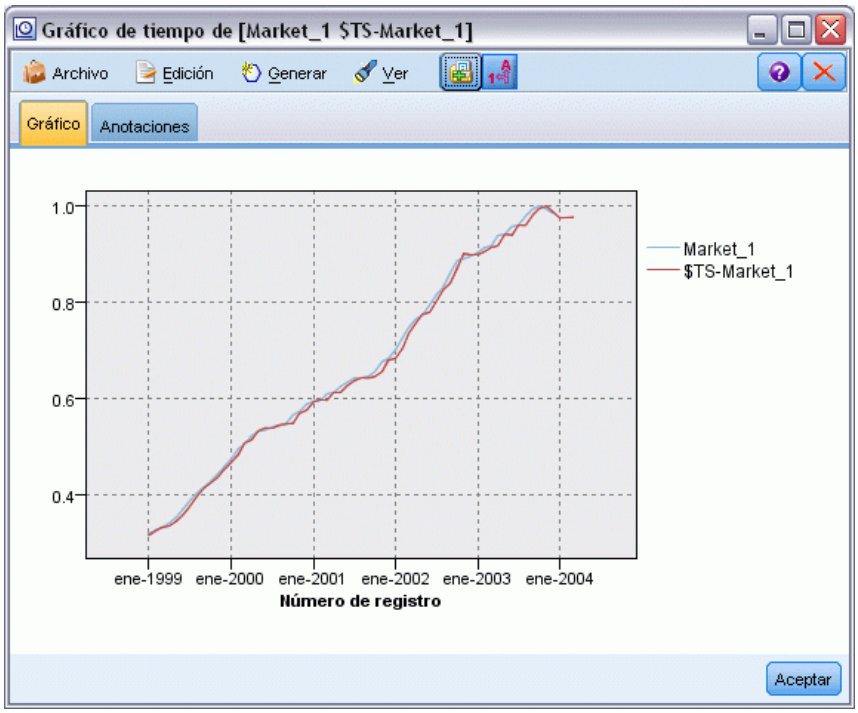
Figura 15-18
Selección de los campos que se van a representar



Observe cómo se extiende la línea de predicción (*$TS-Mercado_1*) más allá del final de los datos reales. Ahora tiene una predicción de la demanda esperada para los tres meses siguientes en este mercado.

Las líneas de los datos reales y de predicciones de toda la serie temporal están muy cerca en el gráfico, lo que indica que es un modelo fiable para esta serie temporal en particular.

Figura 15-19
Gráfico de tiempo de datos reales y de predicciones de Mercado_1



Guarde el modelo en un archivo para usarlo en un futuro ejemplo:

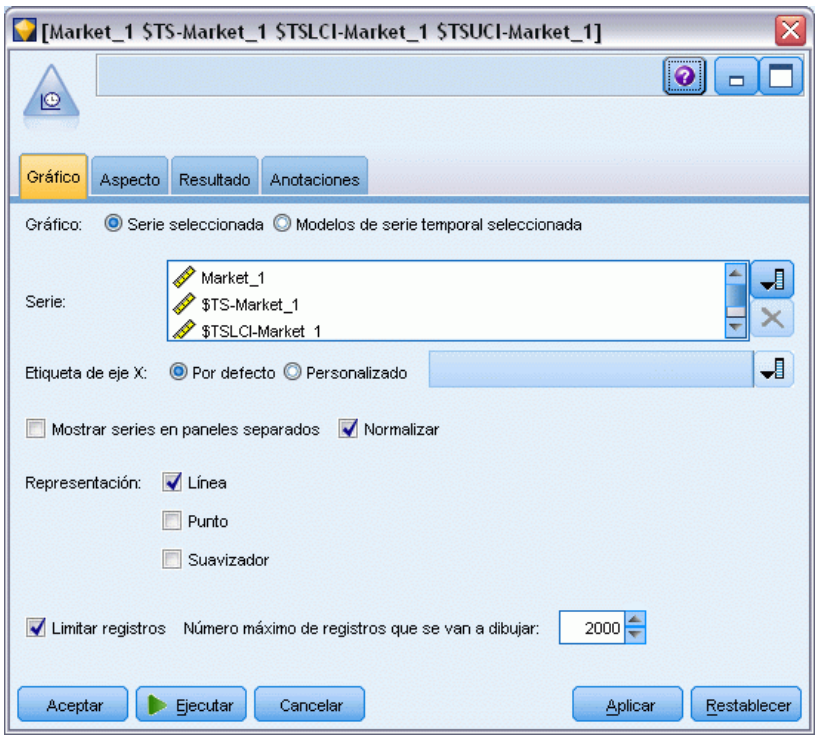
- Pulse en Aceptar para cerrar el gráfico actual.
- Abra el nugget de modelo Serie temporal.
- Seleccione Archivo > Guardar nodo y especifique la ubicación del archivo.
- Pulse en Guardar.

Tiene un modelo fiable para este mercado en particular, pero ¿qué margen de error tiene la predicción? Puede obtener una indicación de esto si examina el intervalo de confianza.

- Pulse dos veces en el último nodo Serie temporal de la ruta (con la etiqueta Mercado_1 $TS-Mercado_1) para volver a abrir este cuadro de diálogo.
- Pulse en el botón selector de campos y añada los campos *$TSLCI-Mercado_1* y *$TSUCI-Mercado_1* a la lista Series.

▶ Pulse en Ejecutar.

Figura 15-20
Adición de campos para representar

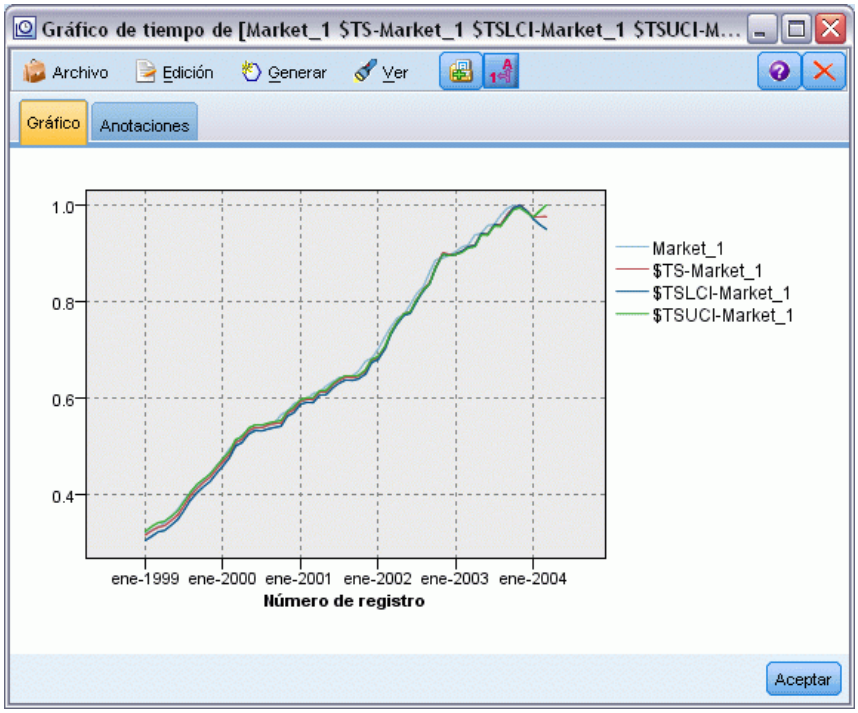


Ahora tiene el mismo gráfico de antes, pero con los límites superior (*$TSUCI*) e inferior (*$TSLCI*) del intervalo de confianza añadidos.

Observe cómo divergen los límites del intervalo de confianza a lo largo del período de predicción, lo que indica que aumenta la incertidumbre al pronosticar más lejos en el tiempo.

No obstante, a medida que transcurra cada período de tiempo, tendrá datos de uso reales correspondientes a otro mes (en este caso), en los que podrá basar la predicción. Puede leer los nuevos datos en la ruta y volver a aplicar el modelo ahora que sabe que es fiable. Si desea obtener más información, consulte el tema Nueva aplicación de modelos de series temporales el p. 197.

Figura 15-21
Gráfico de tiempo con intervalo de confianza añadido



Resumen

Ha aprendido a usar el modelizador experto para generar predicciones para varias series temporales y ha guardado los modelos resultantes en un archivo externo.

En el ejemplo siguiente, verá cómo se transforman datos de series temporales no estándar en un formato adecuado para realizar introducir datos en un nodo Serie temporal.

Nueva aplicación de modelos de series temporales

En este ejemplo se aplican los mismos modelos de series temporales del primer ejemplo de serie temporal, pero también se puede usar de manera independiente. Si desea obtener más información, consulte el tema Predicciones con el nodo Serie temporal el p. 174.

Como en el escenario original, un analista que trabaja para un proveedor de banda ancha a nivel nacional debe generar predicciones mensuales de suscripciones de usuarios para cada mercado local con el objetivo de poder pronosticar los requisitos de ancho de banda. Ya ha utilizado el modelizador experto para crear modelos y hacer una predicción de tres meses.

Se ha actualizado el almacén de datos con los datos reales del período de predicción original, por lo que desea usar esos datos para ampliar las predicciones tres meses más.

Este ejemplo utiliza la ruta denominada *broadband_apply_models.str*, que hace referencia al archivo de datos denominado *broadband_2.sav*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *broadband_apply_models.str* se encuentra en la carpeta *streams*.

Recuperación de la ruta

En este ejemplo, volverá a crear un nodo Serie temporal a partir del modelo de serie temporal guardado en el primer ejemplo. No se preocupe si no ha guardado ningún modelo: hemos incluido uno en el directorio *Demos*.

- Abra la ruta *broadband_apply_models.str* del directorio *streams* en *Demos*.

Figura 15-22
Apertura de la ruta

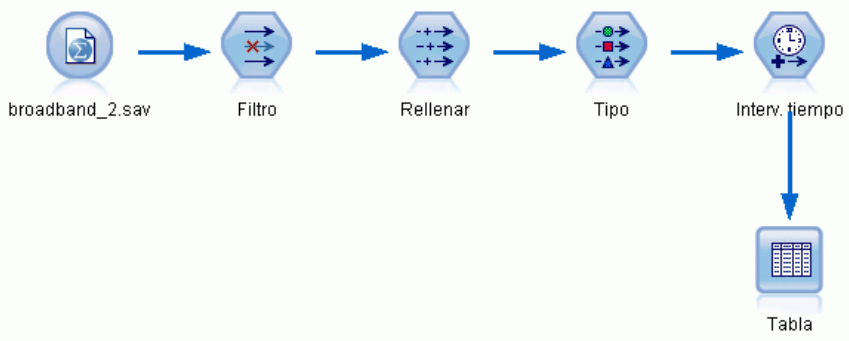
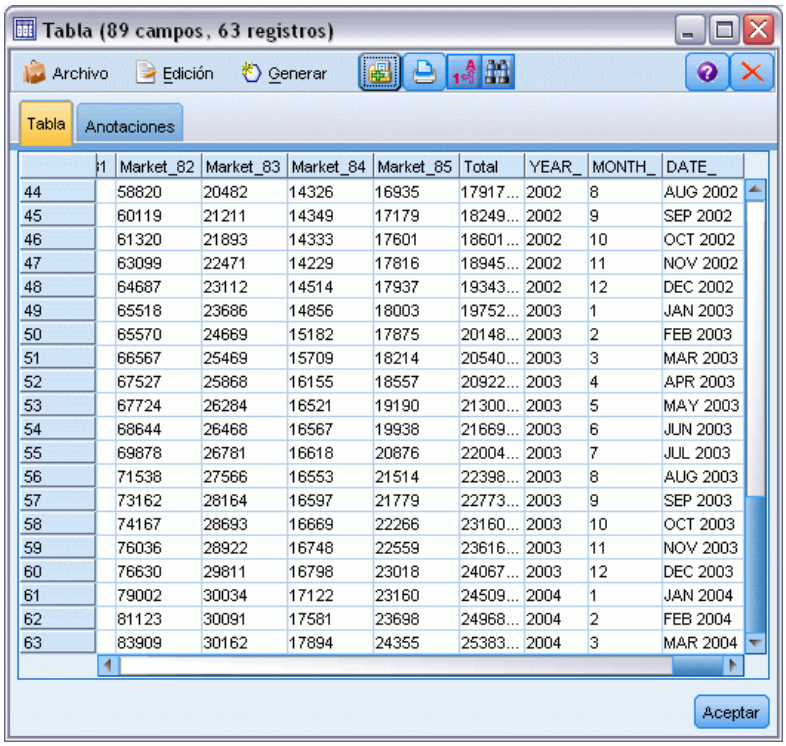


Figura 15-23
Datos de ventas actualizados



	Market_82	Market_83	Market_84	Market_85	Total	YEAR_	MONTH_	DATE_
44	58820	20482	14326	16935	17917...	2002	8	AUG 2002
45	60119	21211	14349	17179	18249...	2002	9	SEP 2002
46	61320	21893	14333	17601	18601...	2002	10	OCT 2002
47	63099	22471	14229	17816	18945...	2002	11	NOV 2002
48	64687	23112	14514	17937	19343...	2002	12	DEC 2002
49	65518	23686	14856	18003	19752...	2003	1	JAN 2003
50	65570	24669	15182	17875	20148...	2003	2	FEB 2003
51	66567	25469	15709	18214	20540...	2003	3	MAR 2003
52	67527	25868	16155	18557	20922...	2003	4	APR 2003
53	67724	26284	16521	19190	21300...	2003	5	MAY 2003
54	68644	26468	16567	19938	21669...	2003	6	JUN 2003
55	69878	26781	16618	20876	22004...	2003	7	JUL 2003
56	71538	27566	16553	21514	22398...	2003	8	AUG 2003
57	73162	28164	16597	21779	22773...	2003	9	SEP 2003
58	74167	28693	16669	22266	23160...	2003	10	OCT 2003
59	76036	28922	16748	22559	23616...	2003	11	NOV 2003
60	76630	29811	16798	23018	24067...	2003	12	DEC 2003
61	79002	30034	17122	23160	24509...	2004	1	JAN 2004
62	81123	30091	17581	23698	24968...	2004	2	FEB 2004
63	83909	30162	17894	24355	25383...	2004	3	MAR 2004

Los datos mensuales actualizados se recopilan en *broadband_2.sav*.

- Conecte un nodo Tabla al nodo Archivo IBM® SPSS® Statistics, abra el nodo Tabla y pulse en Ejecutar.

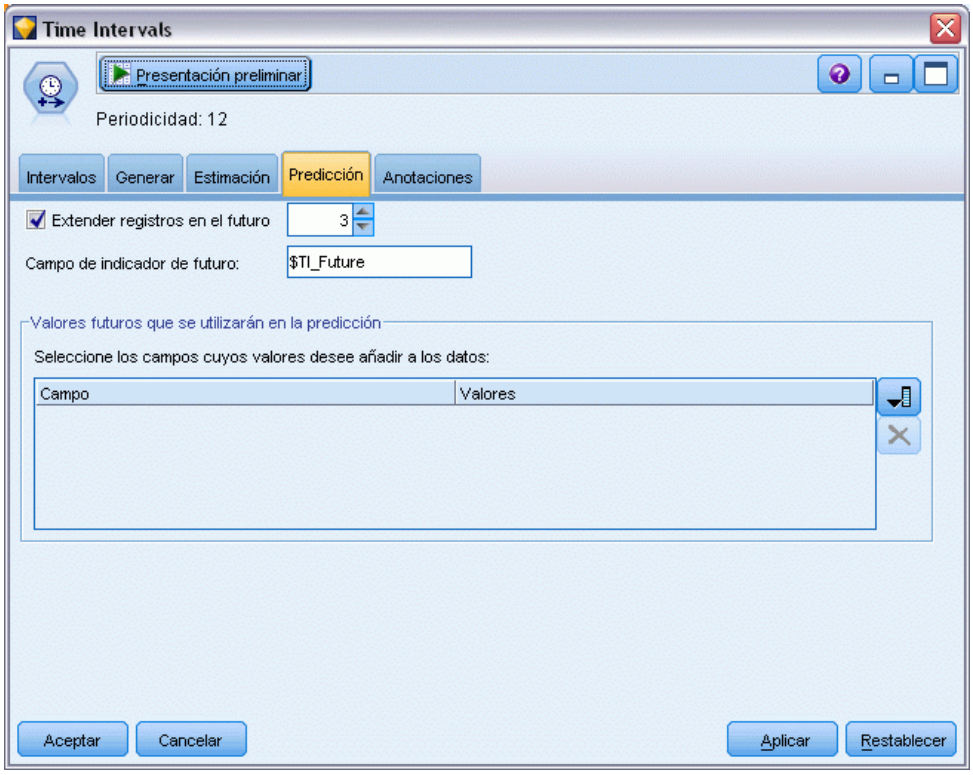
  Nota: el archivo de datos se ha actualizado con los datos reales de las ventas de enero a marzo de 2004, en las filas 61 a 63.

- Abra el nodo Intervalos de tiempo en la ruta.

- Pulse en la pestaña Predicción.

- Asegúrese de que Extender registros en el futuro está definido en 3.

Figura 15-24
Comprobación de la configuración del período de predicción

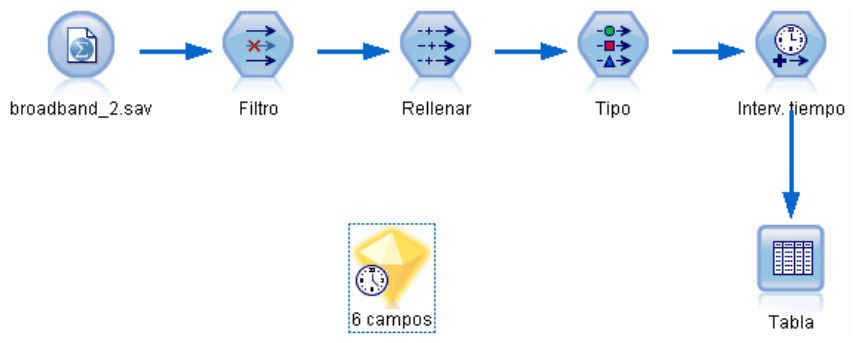


Recuperación del modelo guardado

- En el menú de IBM® SPSS® Modeler, seleccione Insertar > Nodo de archivo y seleccione el archivo *TSmodel.nod* en el directorio *Demos* (o use el modelo de serie temporal que guardó en el primer ejemplo de serie temporal).

Este archivo contiene los modelos de series temporales del ejemplo anterior. La operación de inserción coloca el correspondiente nugget de modelo de serie temporal en el lienzo.

Figura 15-25
Adición del nugget de modelo

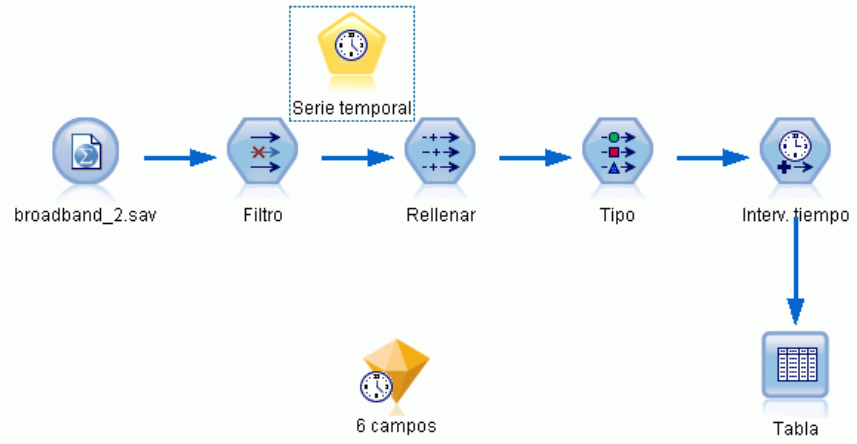


Generación de un nodo de modelado

- Abra el nugget de modelo Serie temporal y seleccione Generar > Generar nodo de modelado.

De esta forma se coloca un nodo de modelado Serie temporal en el lienzo.

Figura 15-26
Creación de nodos de modelado a partir del nugget de modelo



Generación de nuevos modelos

- Cierre el nugget de modelo Serie temporal y elimínelo del lienzo.

  El modelo antiguo se creó utilizando 60 filas de datos. Tiene que generar un nuevo modelo basado en los datos de ventas actualizados (63 filas).

- Conecte el nodo de generación Serie temporal que acaba de crear a la ruta.

Figura 15-27
Adición del nodo de modelado a la ruta

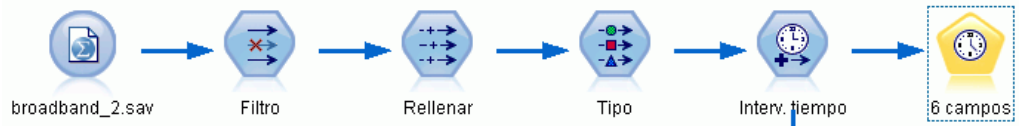
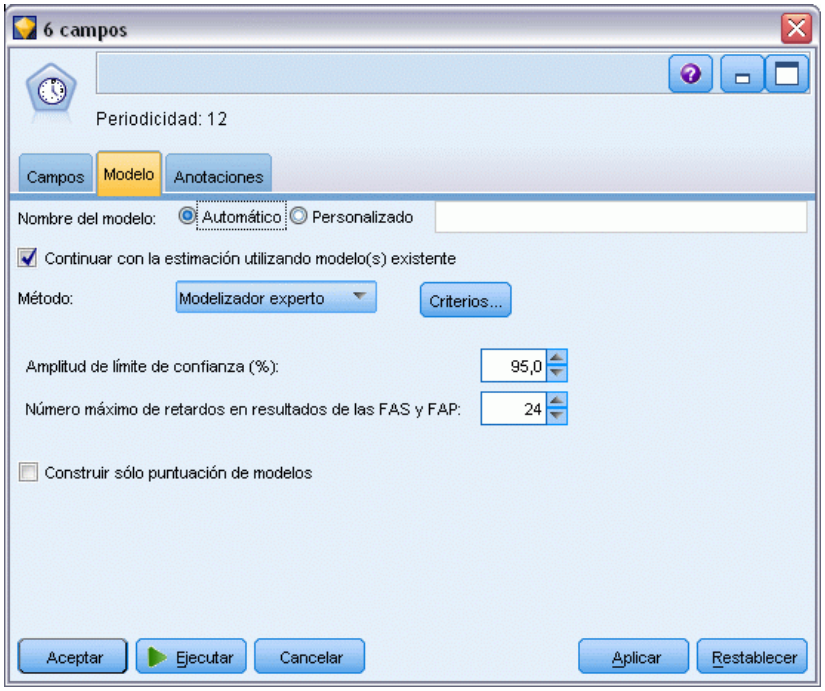


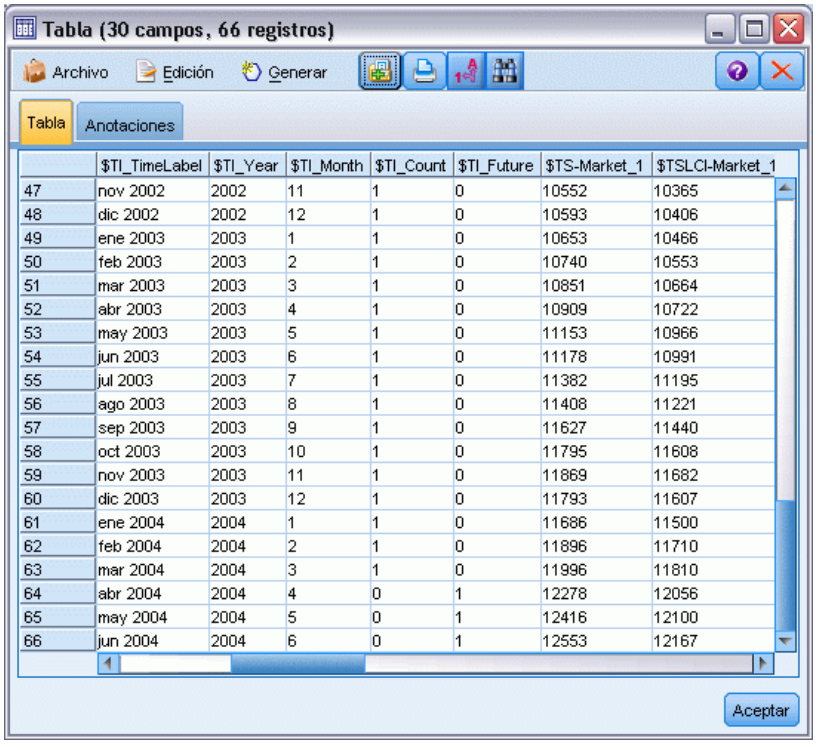
Figura 15-28
Reutilización de configuraciones almacenadas para modelos de series temporales



- Abra el nodo Serie temporal.
- En la pestaña Modelo, compruebe que ha activado Continuar con la estimación utilizando modelo(s) existente.
- Pulse en Ejecutar para colocar un nuevo nugget de modelo en el lienzo y en la paleta Modelos.

Examen del nuevo modelo

Figura 15-29
Tabla con un nuevo pronóstico

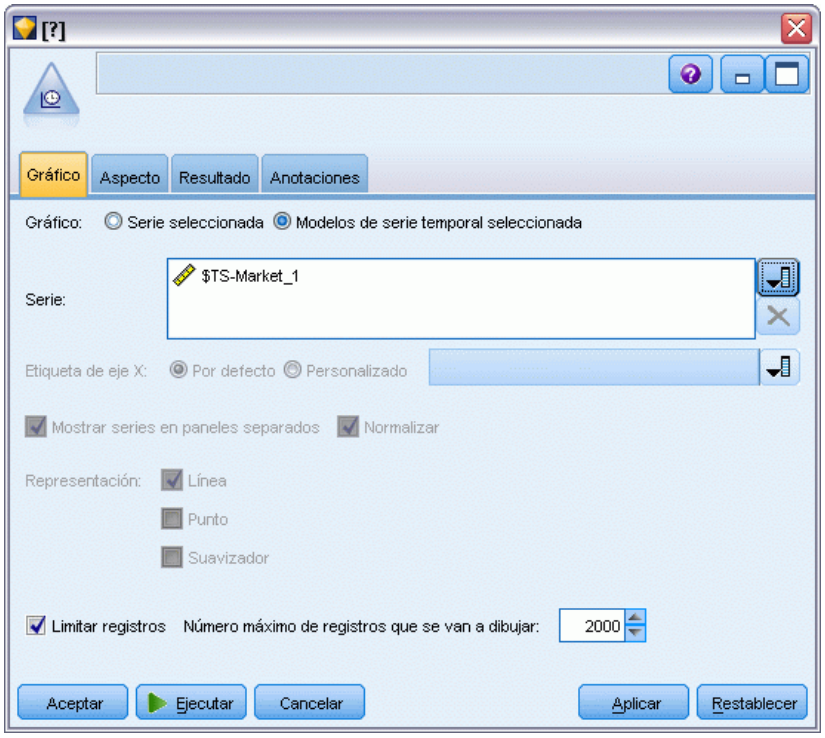


	$TI_TimeLabel	$TI_Year	$TI_Month	$TI_Count	$TI_Future	$TS-Market_1	$TSLCI-Market_1
47	nov 2002	2002	11	1	0	10552	10365
48	dic 2002	2002	12	1	0	10593	10406
49	ene 2003	2003	1	1	0	10653	10466
50	feb 2003	2003	2	1	0	10740	10553
51	mar 2003	2003	3	1	0	10851	10664
52	abr 2003	2003	4	1	0	10909	10722
53	may 2003	2003	5	1	0	11153	10966
54	jun 2003	2003	6	1	0	11178	10991
55	jul 2003	2003	7	1	0	11382	11195
56	ago 2003	2003	8	1	0	11408	11221
57	sep 2003	2003	9	1	0	11627	11440
58	oct 2003	2003	10	1	0	11795	11608
59	nov 2003	2003	11	1	0	11869	11682
60	dic 2003	2003	12	1	0	11793	11607
61	ene 2004	2004	1	1	0	11686	11500
62	feb 2004	2004	2	1	0	11896	11710
63	mar 2004	2004	3	1	0	11996	11810
64	abr 2004	2004	4	0	1	12278	12056
65	may 2004	2004	5	0	1	12416	12100
66	jun 2004	2004	6	0	1	12553	12167

- Conecte un nodo Tabla al nuevo nugget de modelo Serie temporal del lienzo.
- Abra el nodo Tabla y pulse en Ejecutar.

  El nuevo modelo sigue pronosticando con tres meses de antelación, ya que se está reutilizando la configuración almacenada. Sin embargo, en este ejemplo pronostica de abril a junio porque el período de estimación (especificado en el nodo Intervalos de tiempo) termina ahora en marzo en lugar de en enero.

Figura 15-30
Especificación de los campos que se van a representar



- Conecte un nodo de gráfico de tiempo al nugget de modelo de serie temporal generado.

  Esta vez vamos a usar la representación de un gráfico de tiempo especialmente diseñada para modelos de series temporales.

- En la pestaña Gráfico, seleccione la opción Modelos de serie temporal seleccionada.

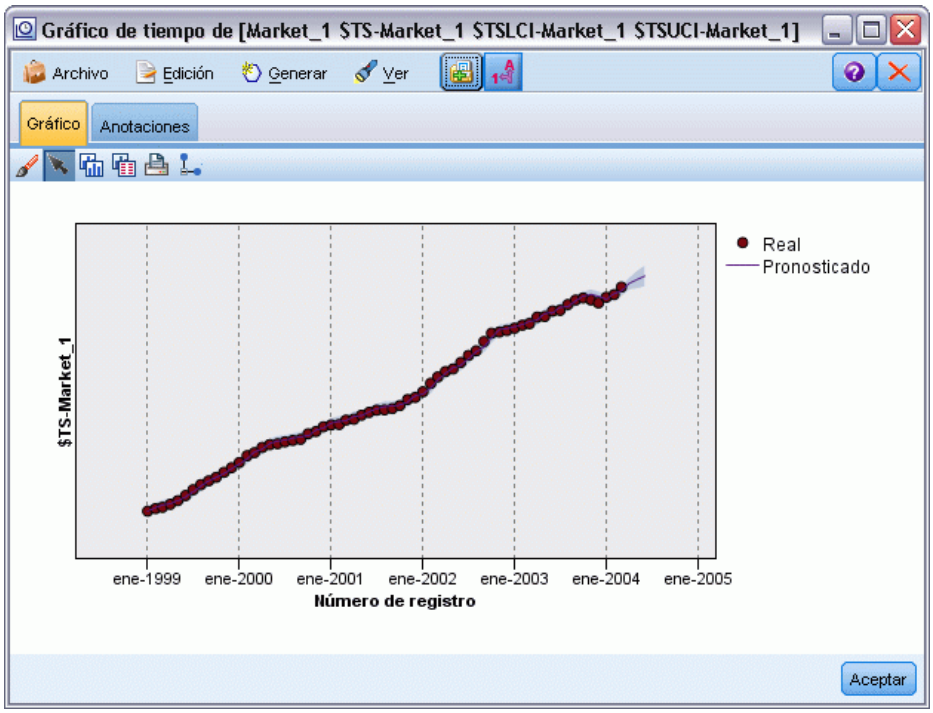
- En la lista Serie, pulse en el botón selector de campos, seleccione el campo *$TS-Mercado_1* y pulse en Aceptar para añadirlo a la lista.

- Pulse en Ejecutar.

  Ahora ya tiene un gráfico que muestra las ventas reales de *Mercado_1* hasta marzo de 2004, además de la predicción (pronóstico) de ventas y el intervalo de confianza (indicado por la zona sombreada en azul) hasta junio de 2004.

Como en el primer ejemplo, los valores de predicción siguen fielmente los datos reales a lo largo de todo el período de tiempo, lo que indica una vez más que tiene un buen modelo.

Figura 15-31
Predicción ampliada hasta junio



Resumen

Ha aprendido a aplicar modelos guardados para ampliar las predicciones anteriores cuando hay más datos actuales disponibles sin necesidad de volver a generar los modelos. Obviamente, si hay motivos para pensar que un modelo ha cambiado, deberá volver a generarlo.

Capítulo
16

Predicción de ventas por catálogo (Serie temporal)

Una compañía de venta por catálogo está interesada en pronosticar las ventas mensuales de su línea de ropa masculina en base a los datos de ventas de los últimos 10 años.

Este ejemplo utiliza la ruta denominada *catalog_forecast.str*, que hace referencia al archivo de datos denominado *catalog_seasfac.sav*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *catalog_forecast.str* se encuentra en el directorio *streams*.

En un ejemplo anterior hemos visto cómo se puede permitir que el modelizador experto decida cuál es el modelo más adecuado para la serie temporal. Ahora veremos más detenidamente los dos métodos disponibles cuando el usuario elige un modelo: suavizado exponencial y ARIMA.

Para ayudarle a elegir un modelo adecuado, es recomendable representar primero la serie temporal. La inspección visual de una serie temporal puede, por lo general, ser una buena guía para elegir. En concreto, debe preguntarse:

- ¿Dispone la serie de una tendencia global? Si es así, ¿la tendencia parece constante o, por el contrario, parece desaparecer con el tiempo?
- ¿La serie muestra estacionalidad? Si es así, ¿parece que las fluctuaciones estacionales crecen con el tiempo, o parecen ser constantes a lo largo de períodos sucesivos?

Creación de la ruta

▶ Cree una nueva ruta y añada un nodo de origen de archivo Statistics que apunte a *catalog_seasfac.sav*.

Figura 16-1
Predicción de ventas por catálogo

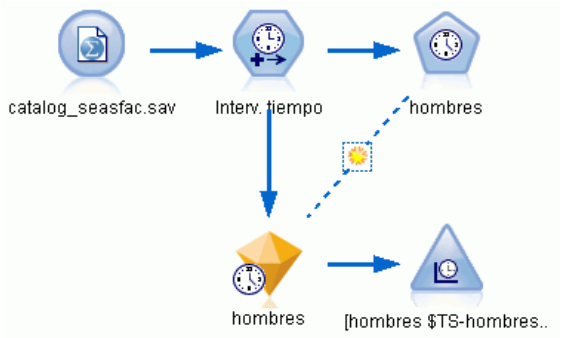
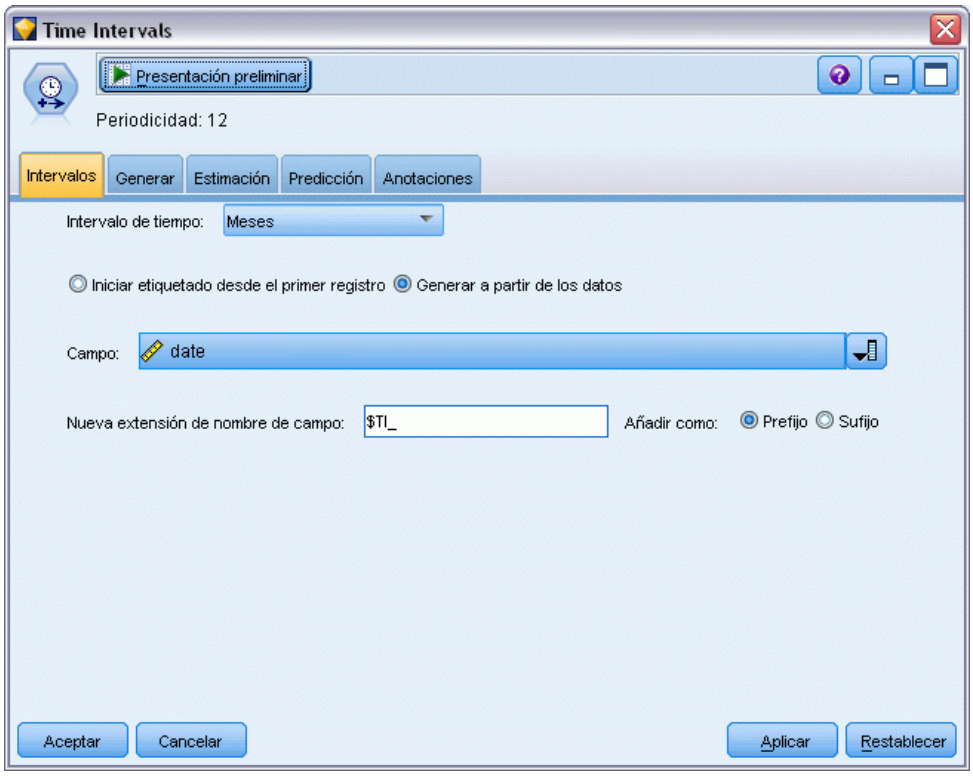


Figura 16-2
Especificación del campo objetivo

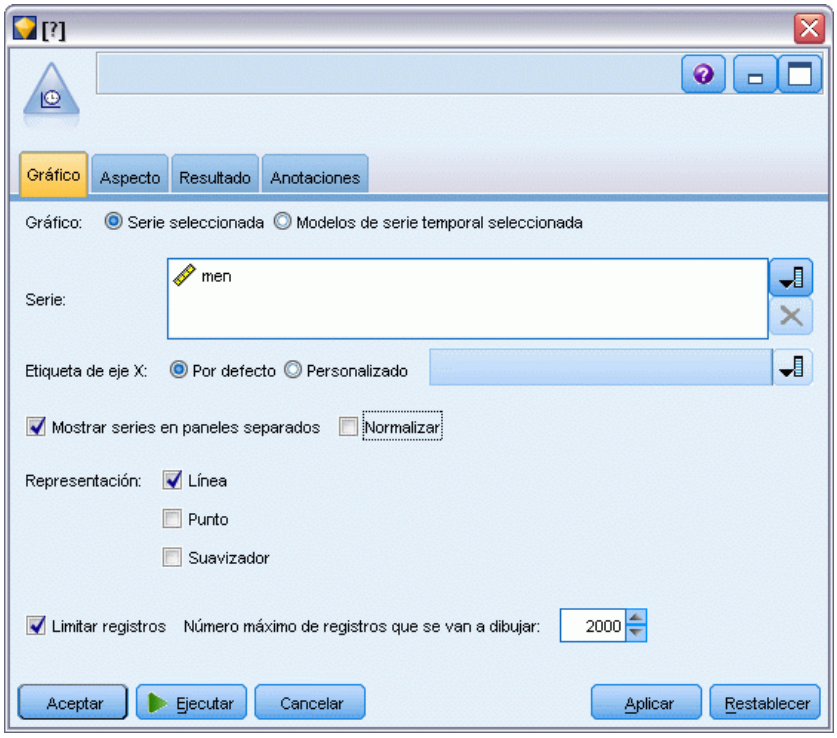
- Abra el nodo de origen de IBM® SPSS® Statistics y seleccione la pestaña Tipos.
- Pulse en Leer valores y, a continuación, en Aceptar.
- Pulse en la columna *Papel* del campo *men* y defina el papel a Objetivo.
- Defina el papel del resto de los campos como Ninguna y pulse en Aceptar.

Figura 16-3
Configuración del intervalo de tiempo



- Conecte un nodo Intervalos de tiempo al nodo de origen de SPSS Statistics.
- Abra el nodo Intervalos de tiempo y establezca Intervalo de tiempo en Meses.
- Seleccione Generar a partir de los datos.
- Establezca Campo como fecha y pulse en Aceptar.

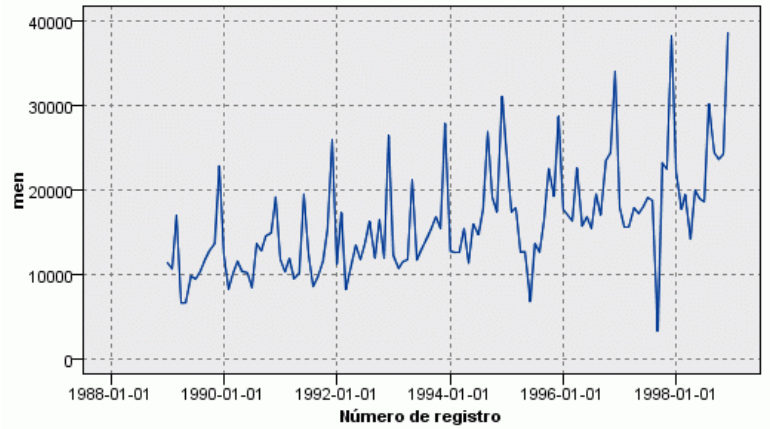
Figura 16-4
Representación de la serie temporal



- Conecte un nodo Gráfico de tiempo al nodo Intervalos de tiempo.
- En la pestaña Gráfico, añada men a la lista Series.
- Desactive la casilla de verificación Normalizar.
- Pulse en Ejecutar.

Examen de los datos

Figura 16-5
Ventas reales de ropa masculina



La serie muestra una tendencia ascendente general, es decir, los valores de la serie tienden a aumentar con el tiempo. La tendencia ascendente es aparentemente constante, lo que indica una tendencia lineal.

La serie también tiene un marcado patrón estacional con máximos anuales en diciembre, como indican las líneas verticales del gráfico. Las variaciones estacionales parecen crecer con la tendencia ascendente de la serie, que sugiere que la estacionalidad es más multiplicativa que aditiva.

▶ Pulse en Aceptar para cerrar el gráfico.

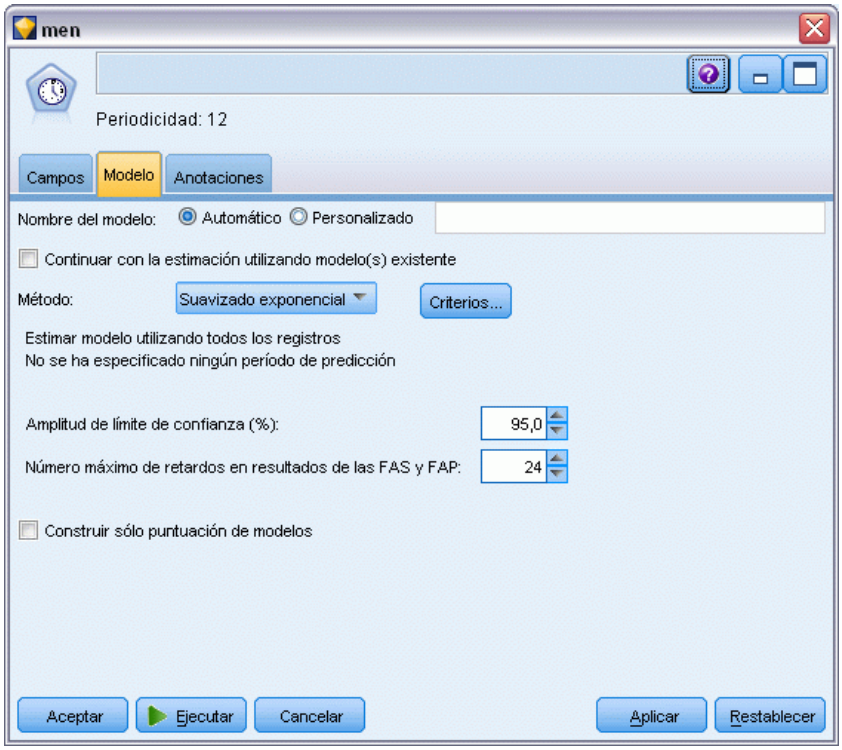
Una vez que ha identificado las características de la serie, puede intentar modelarla. El método de suavizado exponencial es útil para pronosticar las series que muestran una tendencia, estacionalidad o ambas. Como hemos visto, sus datos tienen ambas características.

Suavizado exponencial

Generar el modelo de suavizado exponencial que mejor se ajusta implica determinar el tipo de modelo (si debe incluir tendencia, estacionalidad o ambas) y, a continuación, obtener los parámetros que mejor se ajustan para el modelo elegido.

El gráfico de ventas de prendas para hombre a lo largo del tiempo sugiere un modelo con un componente de tendencia lineal y uno de estacionalidad multiplicativa. Esto implica un modelo Winters. En primer lugar, sin embargo, exploraremos un modelo simple (sin tendencia ni estacionalidad) y, a continuación, un modelo Holt (que incorpora tendencia lineal pero no estacionalidad). lo que le permitirá practicar la identificación de los casos en los que un modelo no se ajusta bien a los datos, habilidad esencial para generar un modelo correctamente.

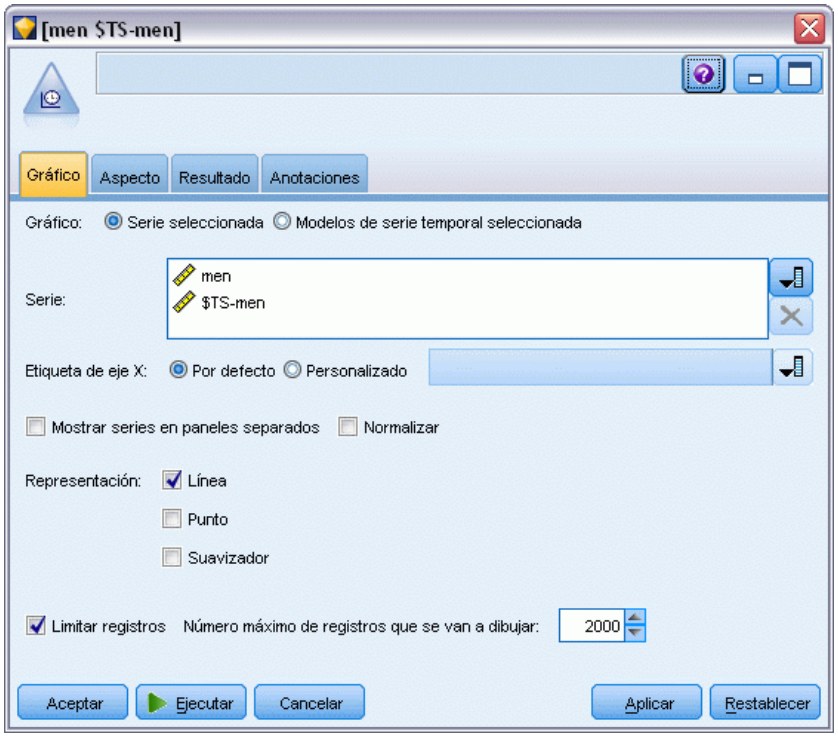
Figura 16-6
Especificación de suavizado exponencial



Comenzaremos con un modelo de suavizado exponencial simple.

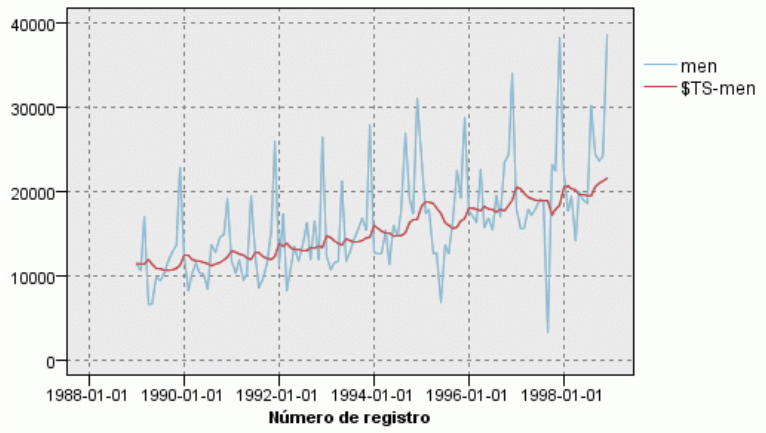
- Conecte un nodo Serie temporal al nodo Intervalos de tiempo.
- En la pestaña Modelo, en Método, seleccione Suavizado exponencial.
- Pulse en Ejecutar para generar el nugget.

Figura 16-7
Representación del modelo de serie temporal



- Conecte un nodo Gráfico de tiempo al nugget de modelo.
- En la pestaña Gráfico, añada *men* y *$TS-men* a la lista Series.
- Desactive las casillas de verificación Mostrar series en paneles separados y Normalizar.
- Pulse en Ejecutar.

Figura 16-8
Modelo de suavizado exponencial simple

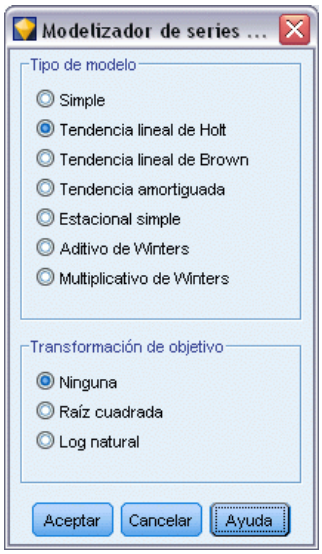


El gráfico men representa los datos reales y $TS-men denota el modelo de serie temporal.

Aunque el modelo simple muestra una tendencia ascendente gradual (y bastante marcada), no tiene en cuenta la estacionalidad. Puede rechazar este modelo sin ningún problema.

- Pulse en Aceptar para cerrar la ventana del gráfico de tiempo.

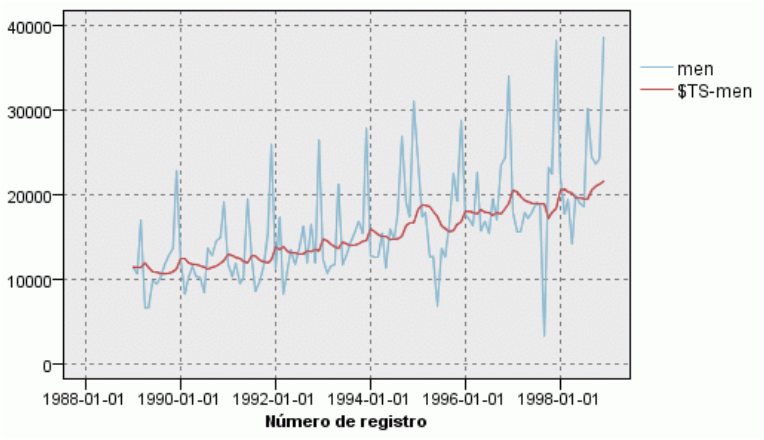
Figura 16-9
Selección de modelo de Holt



Probemos el modelo lineal de Holt. Debería crear un modelo de la tendencia mejor que el modelo simple, aunque también es improbable que capture la estacionalidad.

- Vuelva a abrir el nodo Serie temporal.
- En la pestaña Modelo, con Suavizado exponencial seleccionado como método, pulse en Criterios.
- En el cuadro de diálogo Criterios de suavizado exponencial, seleccione Tendencia lineal de Holt.
- Pulse en Aceptar para cerrar el cuadro de diálogo.
- Pulse en Ejecutar para volver a generar el nugget.
- Vuelva a abrir el nodo Gráfico de tiempo y pulse en Ejecutar.

Figura 16-10
Modelo de tendencia lineal de Holt

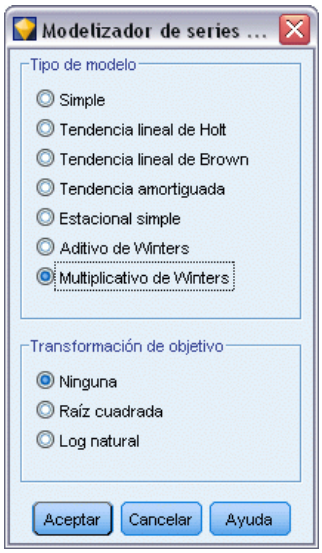


El modelo de Holt muestra una tendencia ascendente más suave que el modelo simple, pero sigue sin tener en cuenta la estacionalidad, por lo que también se puede descartar.

- Cierre la ventana del gráfico de tiempo.

Recordará que el primer gráfico de ventas de ropa masculina a lo largo del tiempo sugería un modelo que incorporase una tendencia lineal y estacionalidad multiplicativa. Por lo tanto, el modelo de Winters podría ser un candidato más adecuado.

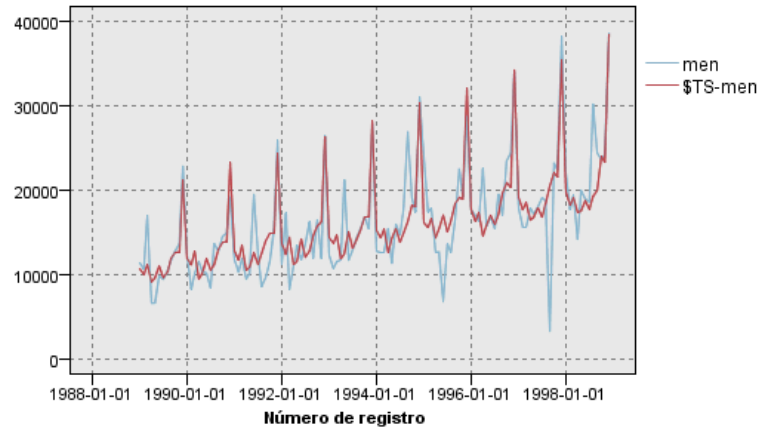
Figura 16-11
Selección del modelo de Winters



- Vuelva a abrir el nodo Serie temporal.
- En la pestaña Modelo, con Suavizado exponencial seleccionado como método, pulse en Criterios.
- En el cuadro de diálogo Criterios de suavizado exponencial, seleccione Multiplicativo de Winters.
- Pulse en Aceptar para cerrar el cuadro de diálogo.

- Pulse en Ejecutar para volver a generar el nugget.
- Abra el nodo Gráfico de tiempo y pulse en Ejecutar.

Figura 16-12
Modelo multiplicativo de Winters



Esto está mejor: el modelo refleja la tendencia y la estacionalidad de los datos.

El conjunto de datos cubre un período de 10 años e incluye 10 picos estacionales que tienen lugar en diciembre de cada año. Los 10 picos presentes en los resultados pronosticados coinciden correctamente con los 10 picos anuales de los datos reales.

Sin embargo, los resultados también subrayan las limitaciones del procedimiento Suavizado exponencial. Al observar los picos ascendentes y descendentes, nos damos cuenta de que hay una estructura significativa que no se ha tenido en cuenta.

Si está interesado principalmente en la creación de un modelo de tendencia a largo plazo con variación estacional, el suavizado exponencial puede ser una buena elección. Para crear un modelo de una estructura más compleja, como ésta, debemos considerar el uso del procedimiento ARIMA.

ARIMA

El procedimiento ARIMA permite crear un modelo de media móvil integrado autorregresivo (ARIMA) ideal para la generación de modelos correctamente ajustados de series temporales. Los modelos ARIMA proporcionan métodos más sofisticados para crear modelos de los componentes de tendencia y estacionales que los modelos de suavizado exponencial y disponen de la ventaja añadida de incluir variables predictoras en el modelo.

En el ejemplo de una compañía de venta por catálogo que quiere desarrollar un modelo de predicción, hemos visto que la empresa ha recopilado datos de las ventas mensuales de ropa masculina junto con varias series que podrían utilizarse para explicar parte de la variación en las ventas. Los posibles predictores incluyen el número de catálogos enviados por correo y el número de páginas del catálogo, el número de líneas telefónicas abiertas para realizar pedidos, el capital invertido en publicidad impresa, así como el número de representantes del servicio de atención al cliente.

¿Alguno de estos predictores es útil para la predicción? ¿Es en realidad un modelo con predictores mejor que uno sin ellos? Con el procedimiento ARIMA podemos crear modelos de predicción con predictores y observar si hay alguna diferencia significativa en su capacidad de pronóstico en comparación con el modelo de suavizado exponencial sin predictores.

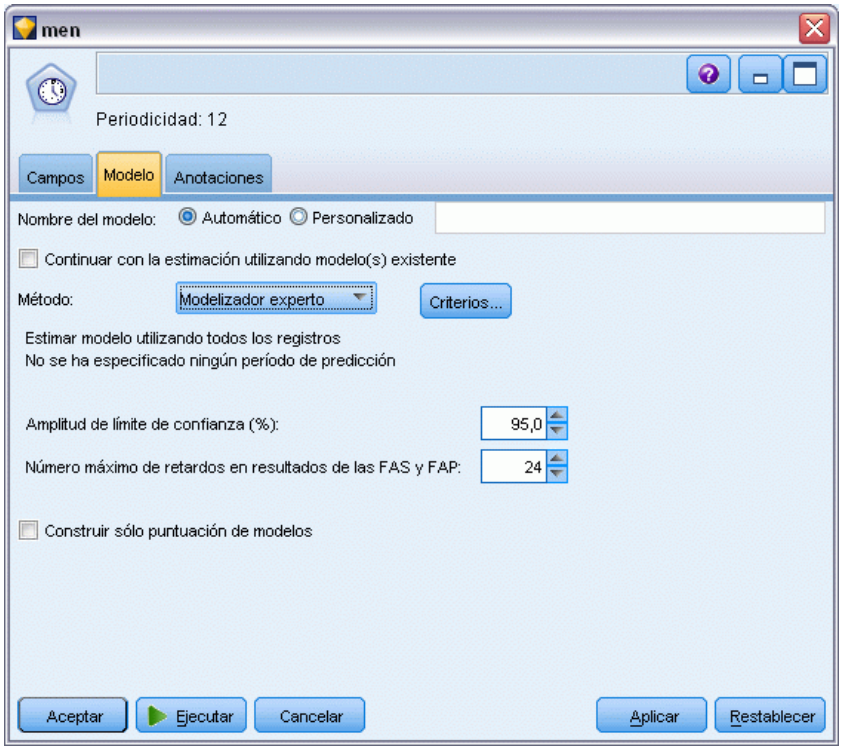
El método ARIMA permite ajustar el modelo con órdenes de autorregresión, diferenciación y media móvil, así como los valores estacionales correspondientes para estos componentes. Determinar manualmente los mejores valores para estos componentes puede llevar mucho tiempo y un gran número de ensayos y errores, así que en este ejemplo permitiremos que el modelizador experto elija un modelo ARIMA por nosotros.

Intentaremos construir un modelo mejor tratando algunas de las otras variables del conjunto de datos como variables predictoras. Las que aparentemente son más útiles para incluir como predictoras son el número de catálogos enviados (*correo*), el número de páginas del catálogo (*página*), el número de líneas telefónicas abiertas para realizar pedidos (*teléfono*), el importe invertido en publicidad impresa (*impresa*) y el número de representantes del servicio de atención al cliente (*servicio*).

Figura 16-13
Configuración de los campos predictores

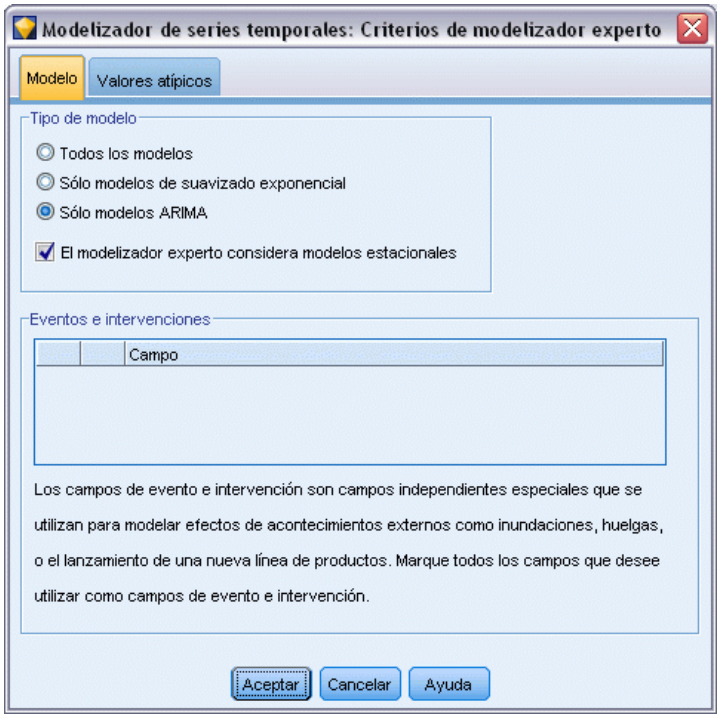
- Abra el nodo de origen del archivo de IBM® SPSS® Statistics.
- En la pestaña Tipos, defina el *Papel* de *correo*, *página*, *teléfono*, *impresa* y *servicio* como Entrada.
- Compruebe que el papel de men esté establecida como Objetivo y que el resto de los campos están establecidos como Ninguna.
- Pulse en Aceptar.

Figura 16-14
Selección del modelizador experto



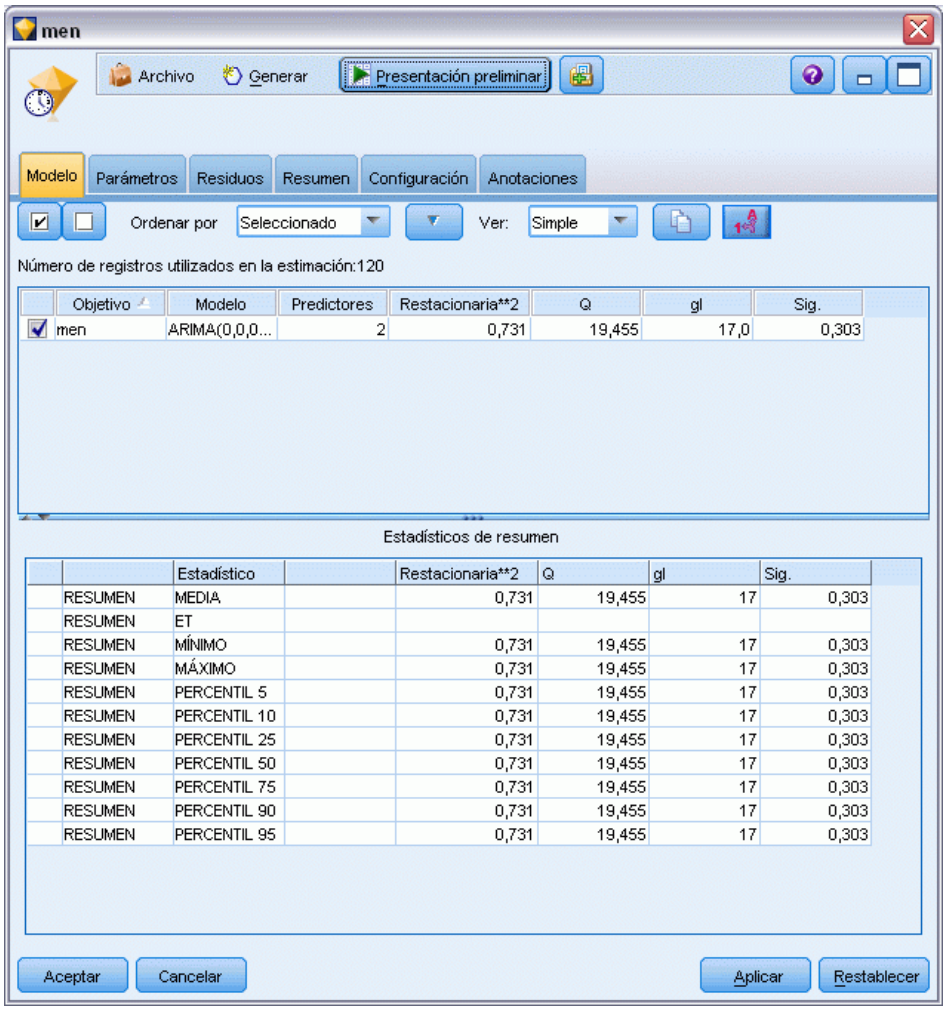
- Abra el nodo Serie temporal.
- En la pestaña Modelo, en Método, seleccione Modelizador experto y pulse en Criterios.

Figura 16-15
Selección de modelos ARIMA únicamente



- En el cuadro de diálogo Criterios de modelizador experto, seleccione la opción Sólo modelos ARIMA y compruebe que la opción El modelizador experto considera modelos estacionales está seleccionada.
- Pulse en Aceptar para cerrar el cuadro de diálogo.
- Pulse en Ejecutar en la pestaña Modelo para volver a generar el nugget de modelo.

Figura 16-16
El modelizador experto selecciona dos predictores



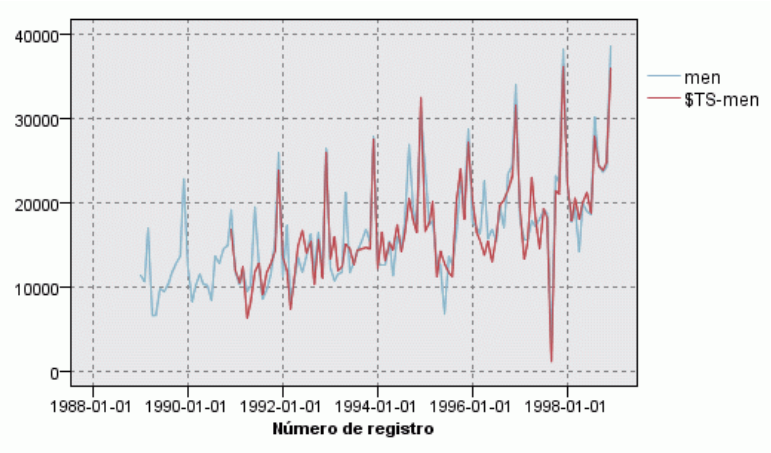
- Abra el nugget de modelo.

  Observe cómo, de los cinco predictores especificados, el modelizador experto ha seleccionado sólo dos como significativos para el modelo.

- Pulse en Aceptar para cerrar el nugget de modelo.

- Abra el nodo Gráfico de tiempo y pulse en Ejecutar.

Figura 16-17
Modelo ARIMA con predictores especificados



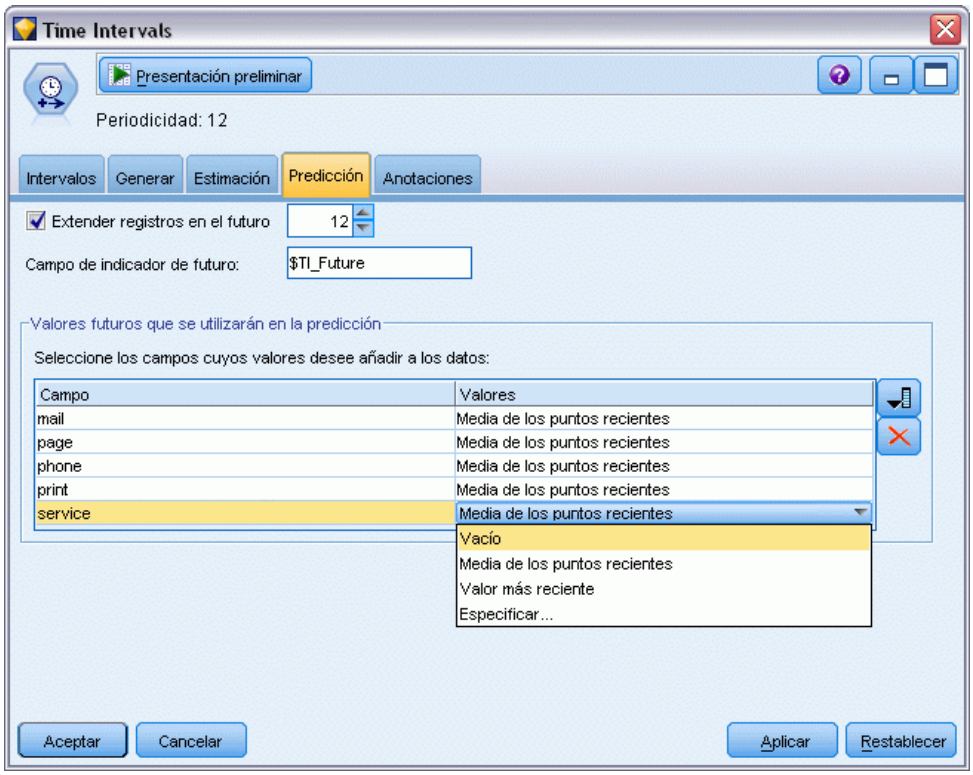
Este modelo es mejor que el anterior porque también captura el gran pico descendente, lo que lo convierte en el más adecuado hasta ahora.

Podríamos intentar refinar aún más el modelo, pero es probable que las mejoras sean mínimas a partir de ahora. Hemos comprobado que es preferible el modelo ARIMA con predictores, así que utilizaremos el modelo que acabamos de construir. En este ejemplo, pronosticaremos las ventas del próximo año.

- Pulse en Aceptar para cerrar la ventana del gráfico de tiempo.
- Abra el nodo Intervalos de tiempo y seleccione la pestaña *Predicción*.
- Active la casilla de verificación *Extender registros en el futuro* y establezca su valor en 12.

El uso de predictores para realizar predicciones requiere que el usuario especifique valores estimados en los campos del período de predicción para que el modelizador pueda predecir con más exactitud el campo objetivo.

Figura 16-18
Especificación de valores futuros para los campos predictores

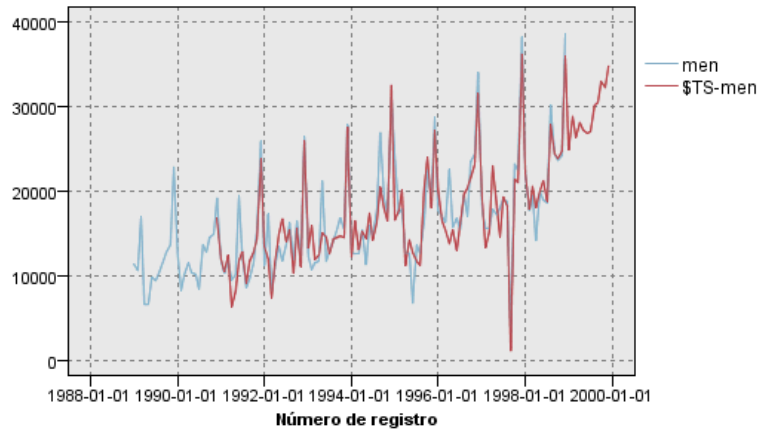


- En el grupo Valores futuros que se utilizarán en la predicción, pulse en el botón selector de campos que hay a la derecha de la columna Valores.
- En el cuadro de diálogo Seleccionar campos, seleccione desde correo hasta servicio y pulse en Aceptar.

  En un caso real, en este punto especificaría los valores futuros manualmente, ya que estos cinco predictores están relacionados con elementos que están bajo su control. En este ejemplo, utilizaremos una de las funciones predefinidas para evitar la necesidad de especificar 12 valores para cada predictor. (Cuando se familiarice con este ejemplo, podrá experimentar con diferentes valores futuros para comprobar su efecto sobre el modelo).
- En cada campo, pulse en el campo Valores para mostrar la lista de posibles valores y seleccione Media de los puntos recientes. Esta opción calcula la media de los tres últimos puntos de datos de este campo y la utiliza como el valor estimado en cada caso.
- Pulse en Aceptar.
- Abra el nodo Serie temporal y pulse en Ejecutar para volver a generar el nugget del modelo.
- Abra el nodo Gráfico de tiempo y pulse en Ejecutar.

La predicción para 1999 es buena. Como se esperaba, se vuelve a niveles normales de ventas después del pico de diciembre y hay una tendencia ascendente continua en la segunda mitad del año. Por lo general, las ventas son bastante superiores a las del año anterior.

Figura 16-19
Predicción de ventas con predictores especificados



Resumen

Ya ha creado un modelo correcto de una serie temporal compleja que incorpora no sólo una tendencia ascendente sino también variaciones estacionales y de otro tipo. También ha visto cómo, mediante ensayo y error, puede acercarse cada vez más a un modelo preciso, que es el que ha utilizado para pronosticar ventas futuras.

En la práctica, tendría que volver a aplicar el modelo a medida que los datos reales de ventas se actualicen (por ejemplo, cada mes o cada trimestre) y generar predicciones actualizadas. Si desea obtener más información, consulte el tema Nueva aplicación de modelos de series temporales en el capítulo 15 el p. 197.

Capítulo
17

Realización de ofertas a clientes (Autoaprendizaje)

El nodo de modelo de respuesta de autoaprendizaje (SLRM, del inglés Self-Learning Response Model) genera y permite actualizar un modelo con el fin de pronosticar cuáles son las ofertas más adecuadas para los clientes, y la probabilidad de que éstos acepten las ofertas. Estos tipos de modelos son muy beneficiosos en la gestión de relaciones con los clientes, incluidas las aplicaciones de marketing y los centros de llamadas.

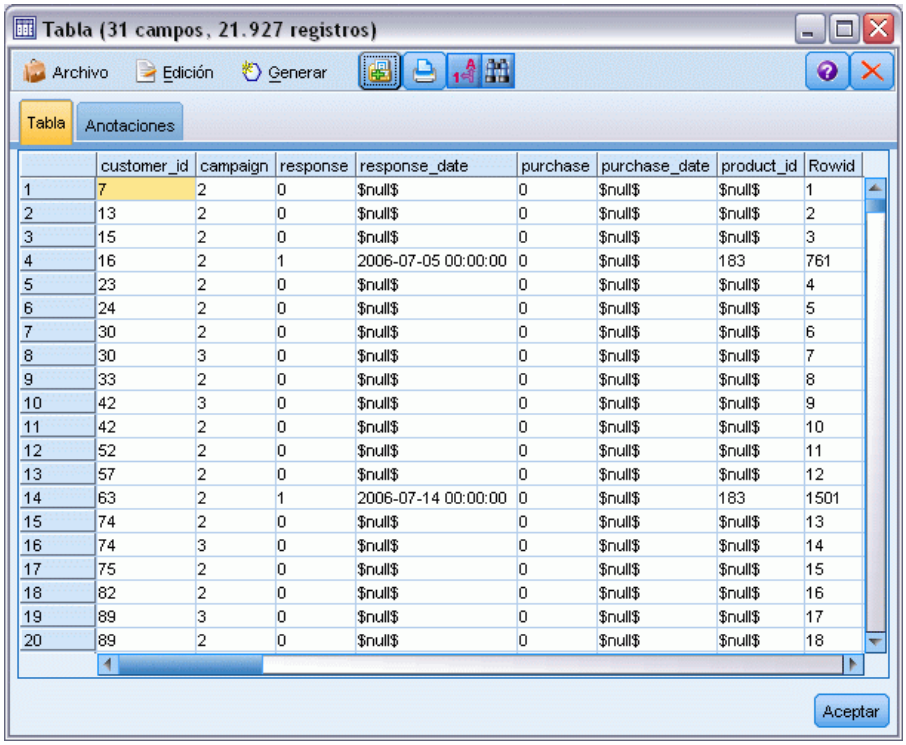
Este ejemplo se basa en una empresa bancaria ficticia. El departamento de marketing desea obtener resultados más rentables en las futuras campañas adaptando la oferta de servicios financieros a cada cliente. Concretamente, en el ejemplo se utiliza un modelo de respuesta de autoaprendizaje para identificar las características de los clientes que es más probable que respondan favorablemente, teniendo en cuenta ofertas y respuestas anteriores, y promocionar la mejor oferta existente a partir de estos resultados.

Este ejemplo utiliza la ruta denominada *pm_selflearn.str*, que hace referencia a los archivos de datos *pm_customer_train1.sav*, *pm_customer_train2.sav* y *pm_customer_train3.sav*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *pm_selflearn.str* se encuentra en la carpeta *streams*.

Datos existentes

La empresa tiene un historial de datos en el que se registran las ofertas realizadas a los clientes en campañas anteriores, así como las respuestas a dichas ofertas. Estos datos también incluyen información demográfica y financiera que se puede utilizar para pronosticar el índice de respuesta de distintos clientes.

Figura 17-1
Respuestas a ofertas anteriores

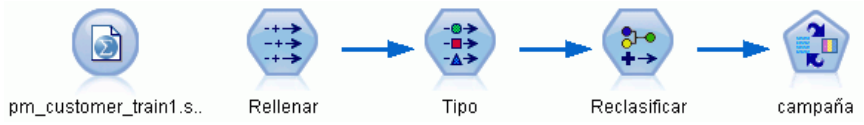


	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

Generación de la ruta

- Añada un nodo de origen de archivo Statistics que apunte a *pm_customer_train1.sav*, ubicado en la carpeta *Demos* de la instalación de IBM® SPSS® Modeler.

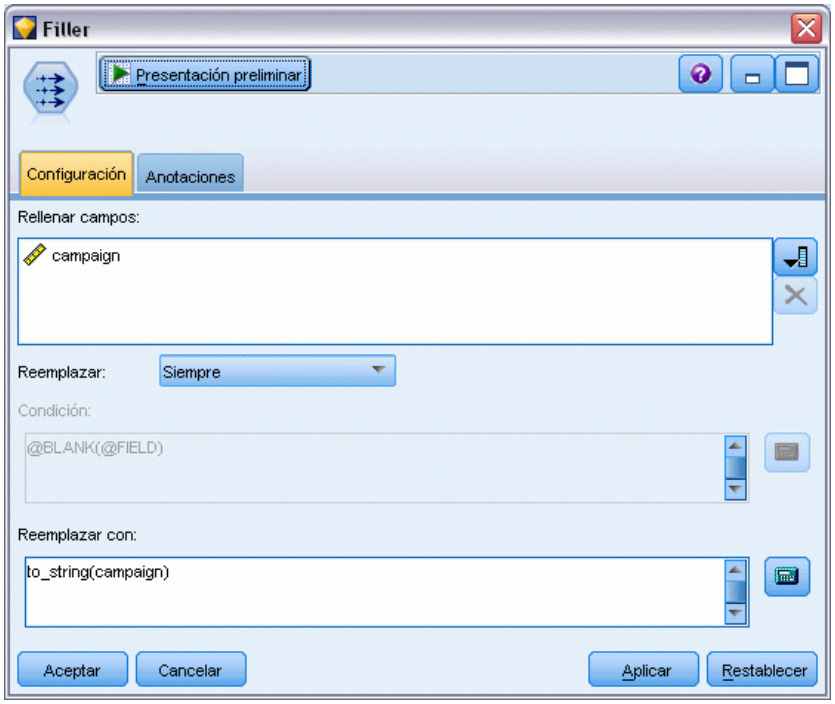
Figura 17-2
Ruta de ejemplo de SLRM



- Añada un nodo Rellenar y seleccione campaña para cumplimentar el campo.
- Seleccione un tipo de sustitución de Siempre.

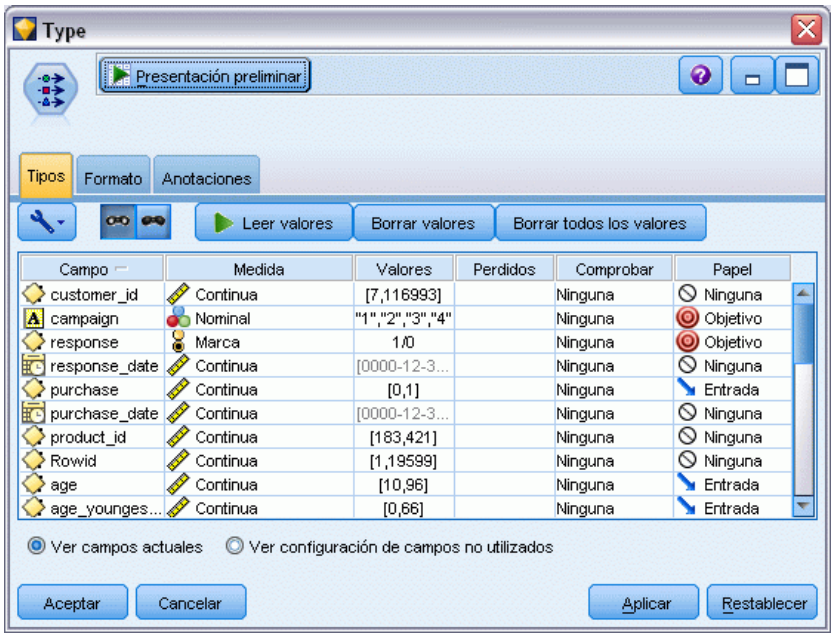
▶ En el cuadro de texto Reemplazar con, escriba to_string(campaign) y pulse en Aceptar.

Figura 17-3
Derivación del campo campaign



- Añada un nodo Tipo y defina *Papel* a Ninguno para los campos *id_cliente*, *fecha_respuesta*, *fecha_compra*, *id_producto*, *Idfila* y *X_aleatorio*.

Figura 17-4
Cambio de configuración del nodo Tipo



- Defina el *Papel* a Objetivo para los campos *campaña* y *respuesta*. Éstos son los campos en los que desea basar las predicciones.

  Defina la Medida a Marca en el campo *respuesta*.

- Pulse en Leer valores y, a continuación, en Aceptar.

  Como los datos del campo campaña aparecen como una lista de números (1, 2, 3 y 4), puede reclasificar los campos para tener unos títulos más significativos.

- Añada un nodo Reclasificar al nodo Tipo.

- En el campo Reclasificar, seleccione Campo existente.

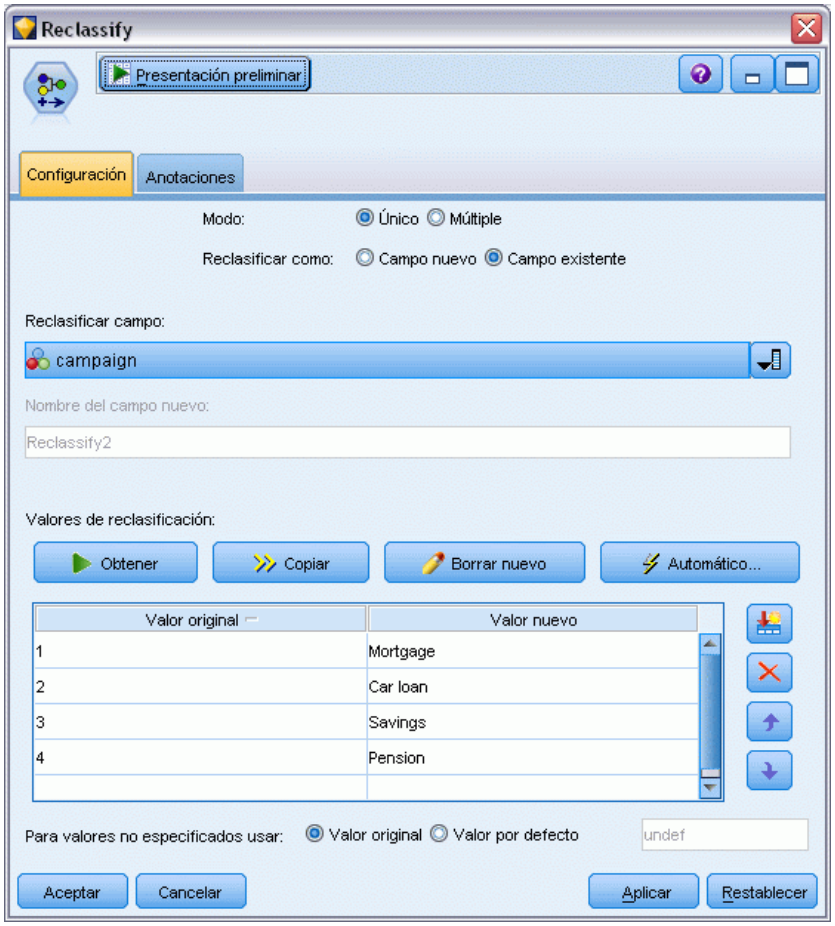
- En el campo Reclasificar, seleccione campaña.

- Pulse en el botón Obtener y los valores de campaña se añadirán a la columna *Valor original*.

- En la columna *Valor nuevo*, introduzca los siguientes nombres de campaña en las cuatro primeras filas:
  - Hipoteca
  - Préstamo coche
  - Ahorros
  - Pensión

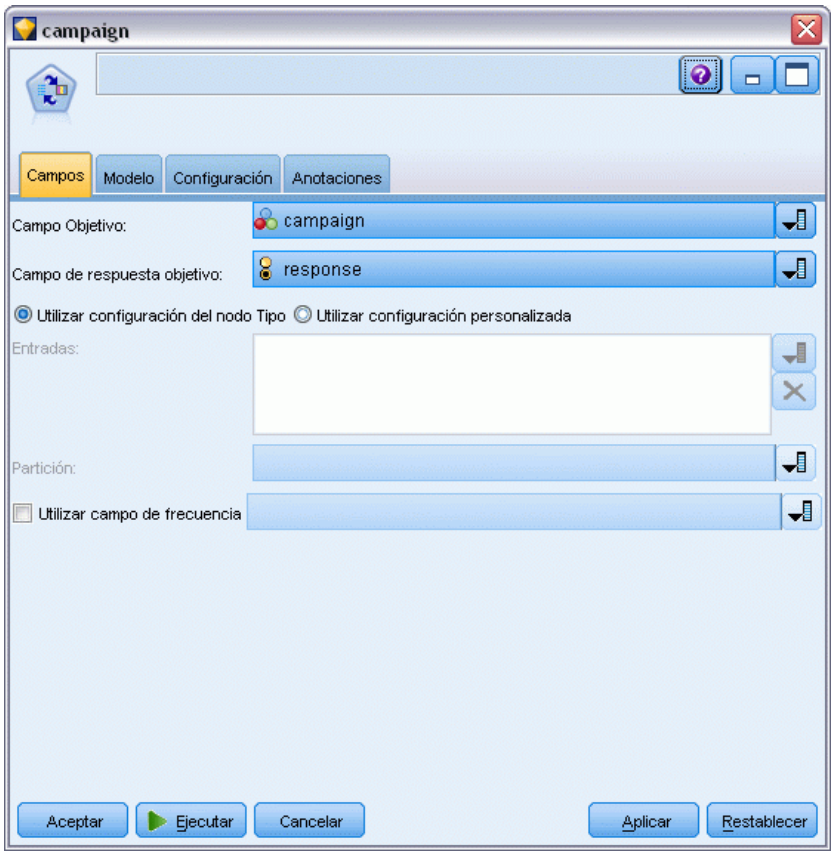
▶ Pulse en Aceptar.

Figura 17-5
Reclasificación de los nombres de campaña



- Conecte un nodo de modelado SLRM al nodo Reclasificar. En la pestaña Campos, seleccione campaña para el campo Objetivo y respuesta para el campo de respuesta objetivo.

Figura 17-6
Selección del objetivo y la respuesta objetivo

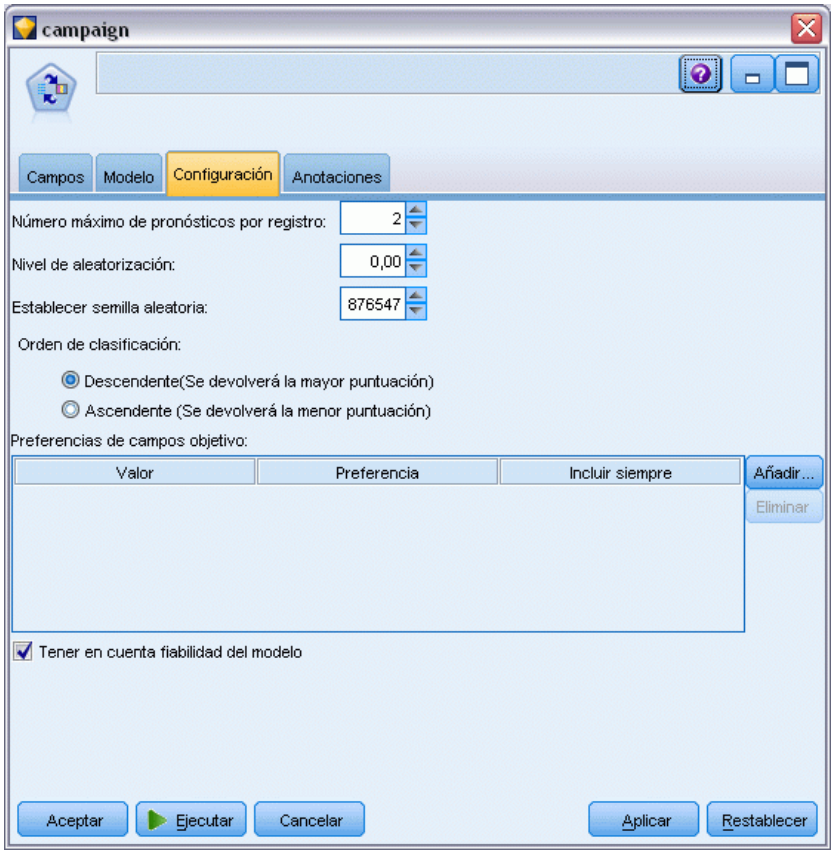


- En la pestaña Configuración, en el campo Número máximo de pronósticos por registro, reduzca el número a 2.

  Este número indica que, para cada cliente, habrá dos ofertas identificadas que tendrán la mayor probabilidad de ser aceptadas.

▶ Asegúrese de que Tener en cuenta fiabilidad del modelo se ha seleccionado y pulse en Ejecutar.

Figura 17-7
Configuración del nodo SLRM



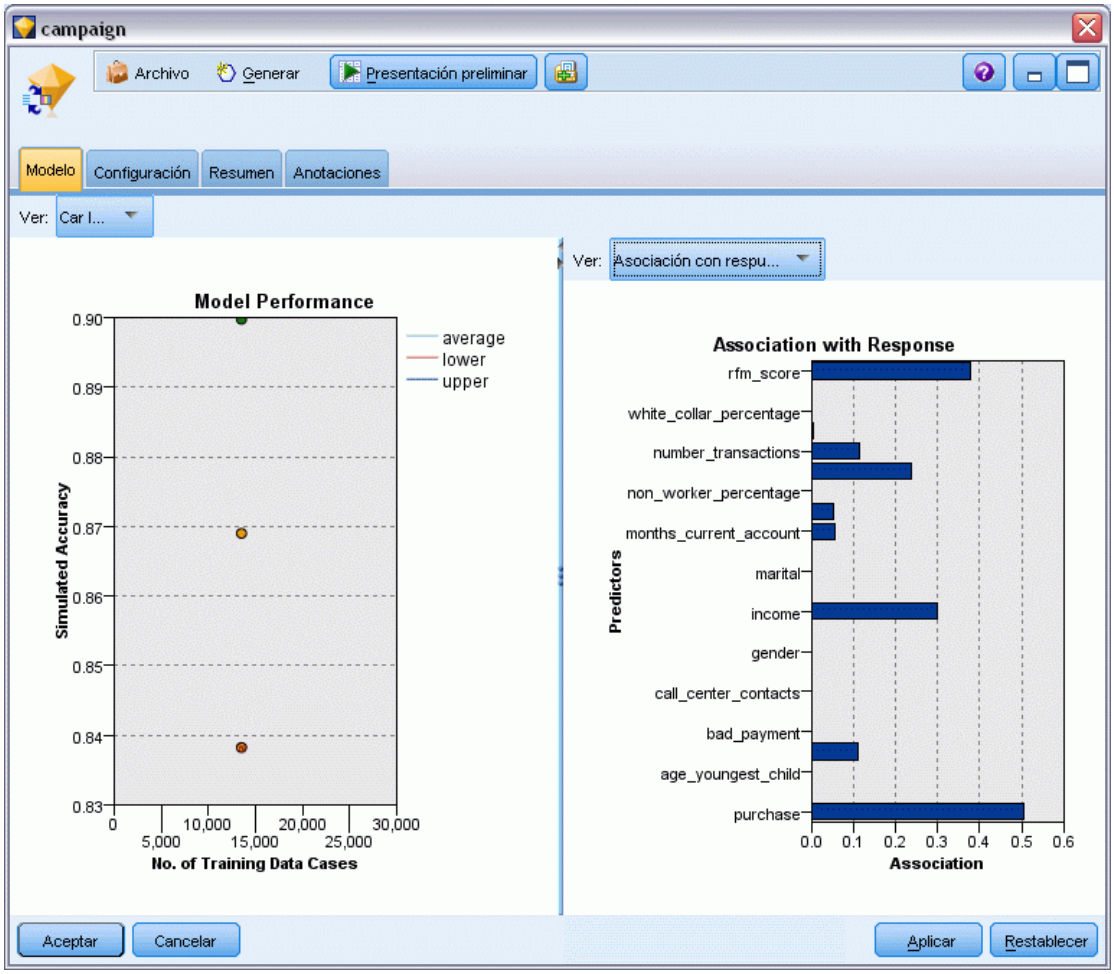
Exploración del modelo

- Abra el nugget de modelo. La pestaña Modelo muestra inicialmente la estimación de la precisión de las predicciones para cada oferta y la importancia relativa de cada predictor en la estimación del modelo.

  Para mostrar la correlación de cada predictor con la variable de objetivo, seleccione Asociación con respuesta de la lista Ver en el panel derecho.

- Para alternar entre cada una de las cuatro ofertas para las que hay pronósticos, seleccione la oferta necesaria en la lista Ver en el panel izquierdo.

Figura 17-8
Nugget de modelo SLRM

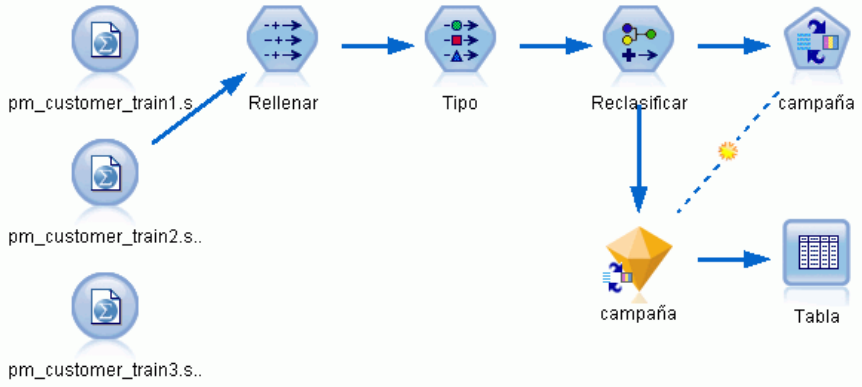


- Cierre la ventana de nugget de modelo.

- En el lienzo de rutas, desconecte el nodo de origen de IBM® SPSS® Statistics que apunta a *pm_customer_train1.sav*.

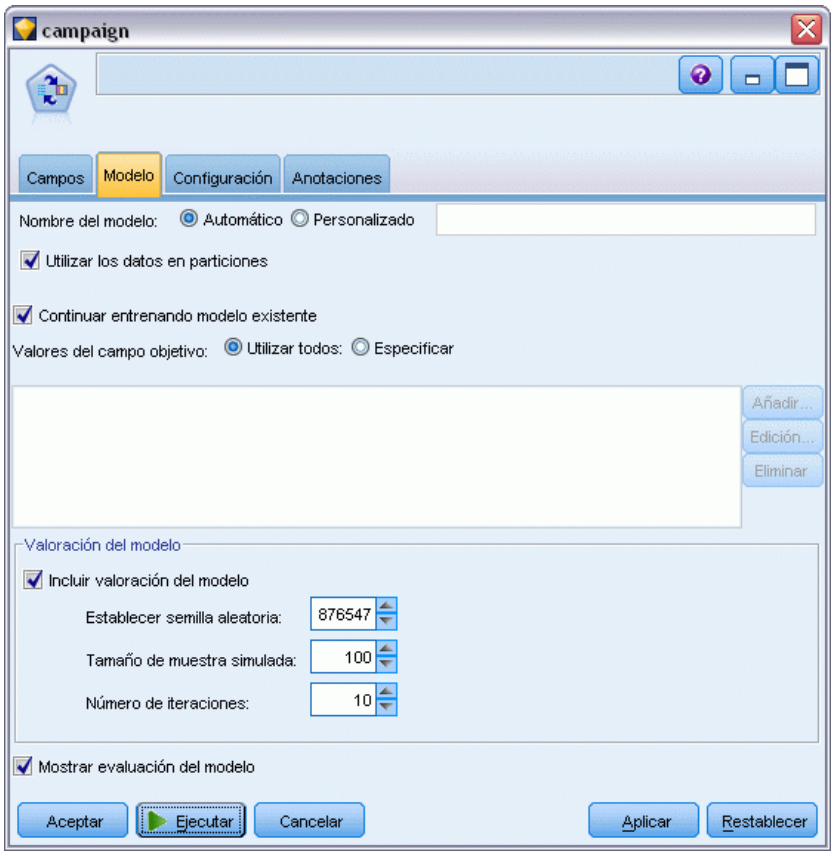
- Añada un nodo de origen de archivo Statistics que apunte a *pm_customer_train2.sav*, que se encuentra en la carpeta *Demos* de la instalación de IBM® SPSS® Modeler, y añádalo al nodo Rellenar.

Figura 17-9
Conexión del segundo origen de datos a la ruta de SLRM



- En la pestaña Modelo del nodo SLRM, seleccione Continuar entrenando modelo existente.

Figura 17-10
Continuar entrenando modelo.

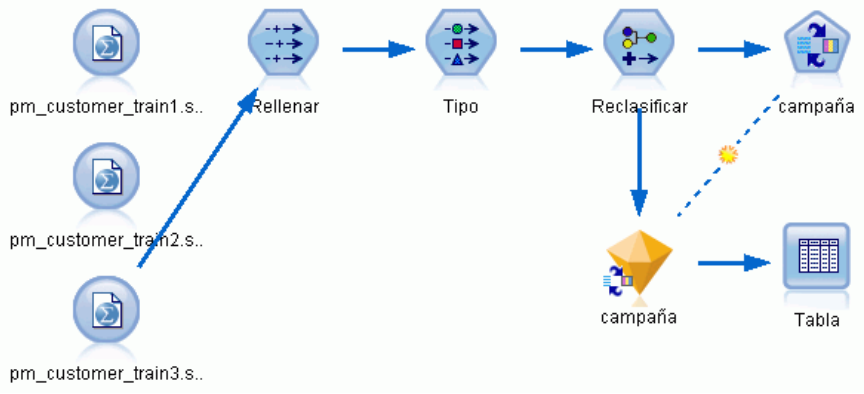


- Pulse en Ejecutar para volver a generar el nugget. Para ver los detalles, pulse con el botón derecho del ratón en el nugget del lienzo.

  La pestaña Modelo muestra ahora las estimaciones revisadas de la precisión de las predicciones para cada oferta.

- Añada un nodo de origen Archivo Statistics que apunte a *pm_customer_train3.sav*, que se encuentra en la carpeta *Demos* de la instalación de SPSS Modeler, y añádalo al nodo Rellenar.

Figura 17-11
Conexión del tercer origen de datos a la ruta de SLRM

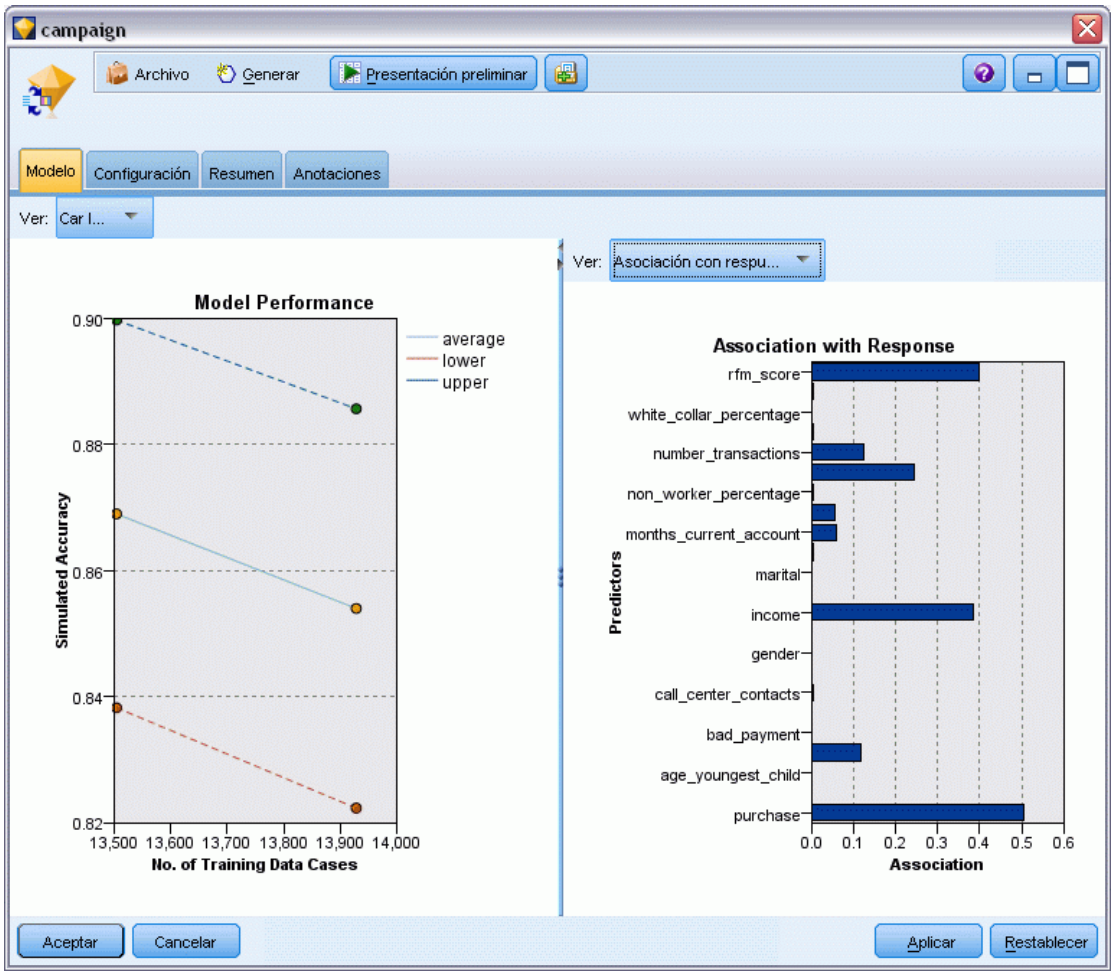


- Pulse en Ejecutar para volver a generar el nugget una vez más. Para ver los detalles, pulse con el botón derecho del ratón en el nugget del lienzo.

- La pestaña Modelo muestra ahora la precisión final estimada de las predicciones para cada oferta.

Tal como podemos ver, la precisión media desciende ligeramente (de 86,9% a 85,4%) a medida que añade los orígenes de datos adicionales; no obstante, esta fluctuación es mínima y puede atribuirse a pequeñas anomalías de los datos disponibles.

Figura 17-12
Nugget de modelo SLRM actualizado



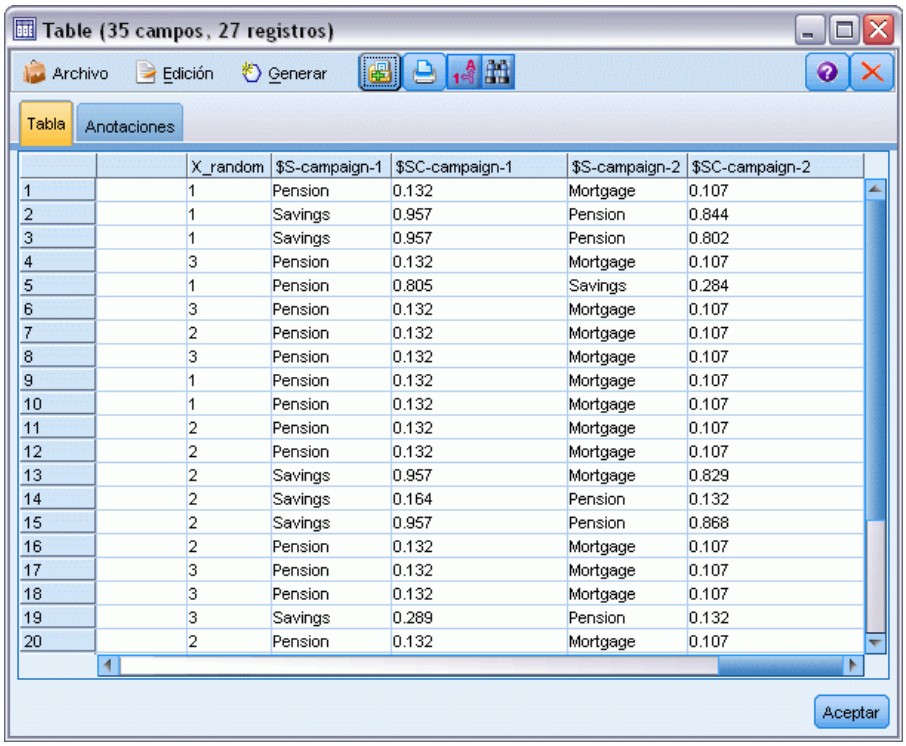
- Conecte un nodo Tabla al último modelo generado (el tercero) y ejecute el nodo Tabla.

- Desplácese hasta la parte derecha de la tabla. Las predicciones muestran las ofertas que es más probable que un cliente acepte y la confianza en que las aceptarán, según los detalles de cada cliente.

  Por ejemplo, en la primera línea de la tabla mostrada, hay un índice de confianza de tan sólo el 13,2% (se distingue por el valor 0,132 en la columna *$SC-campaign-1*) de que un cliente que previamente ha recibido un préstamo para un coche aceptará una pensión si se le ofrece. No obstante, las líneas segunda y tercera muestran dos clientes más que también recibieron un préstamo para un coche; en sus casos, hay una confianza del 95,7% de que ellos, así como otros

clientes con historiales similares, abrirán una cuenta de ahorro si se les ofrece una y más del 80% de la confianza por la que aceptarían una pensión.

Figura 17-13
Resultados del modelo: ofertas pronosticadas y confianzas



		X_random	$S-campaign-1	$SC-campaign-1	$S-campaign-2	$SC-campaign-2
1		1	Pension	0.132	Mortgage	0.107
2		1	Savings	0.957	Pension	0.844
3		1	Savings	0.957	Pension	0.802
4		3	Pension	0.132	Mortgage	0.107
5		1	Pension	0.805	Savings	0.284
6		3	Pension	0.132	Mortgage	0.107
7		2	Pension	0.132	Mortgage	0.107
8		3	Pension	0.132	Mortgage	0.107
9		1	Pension	0.132	Mortgage	0.107
10		1	Pension	0.132	Mortgage	0.107
11		2	Pension	0.132	Mortgage	0.107
12		2	Pension	0.132	Mortgage	0.107
13		2	Savings	0.957	Mortgage	0.829
14		2	Savings	0.164	Pension	0.132
15		2	Savings	0.957	Pension	0.868
16		2	Pension	0.132	Mortgage	0.107
17		3	Pension	0.132	Mortgage	0.107
18		3	Pension	0.132	Mortgage	0.107
19		3	Savings	0.289	Pension	0.132
20		2	Pension	0.132	Mortgage	0.107

Puede encontrar explicaciones de los fundamentos matemáticos de los métodos de modelado utilizados en SPSS Modeler en el *Manual de algoritmos de SPSS Modeler*, disponible en el directorio *\Documentation* del DVD del producto.

Recuerde que estos resultados están basados sólo en los datos de entrenamiento. Para evaluar qué tal se extiende el modelo a otros datos de casos reales, se utilizaría un nodo de partición para reservar un subconjunto de registros para comprobación y validación. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 14.2*. Si desea obtener más información acerca del nodo SLRM, consulte el capítulo 14 en la referencia sobre nodos.

Capítulo
18

Predicción de moras en préstamos (red bayesiana)

Las redes bayesianas le permiten crear un modelo de probabilidad combinando pruebas observadas y registradas con conocimiento del mundo real de "sentido común" para establecer la probabilidad de instancias utilizando atributos aparentemente no vinculados.

Este ejemplo utiliza la ruta denominada *bayes_bankloan.str*, que hace referencia al archivo de datos denominado *bankloan.sav*. Estos archivos están disponibles en el directorio *Demos* de cualquier instalación de IBM® SPSS® Modeler y se puede acceder desde el grupo de programas de IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *bayes_bankloan.str* se encuentra en el directorio *streams*.

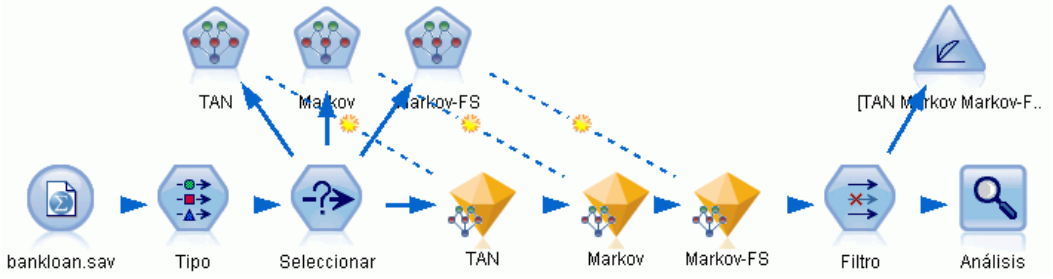
Por ejemplo, supongamos que un banco está preocupado por el posible impago de sus créditos. Si se pueden utilizar datos de créditos anteriores para pronosticar los clientes potenciales que tendrán problemas para pagar sus créditos, a estos clientes de alto riesgo se les puede negar un crédito u ofrecer otros productos.

Este ejemplo utiliza los datos de créditos existentes para pronosticar posibles morosos y observa los tres modelos diferentes de redes bayesianas para establecer cuál es el mejor modelo para pronosticar esta situación.

Generación de la ruta

- Añada un nodo de origen de archivo Statistics apuntando a *bankloan.sav* en la carpeta *Demos*.

Figura 18-1
Ruta de ejemplo de red bayesiana



- Añada un nodo Tipo al nodo de origen y defina el papel del campo predefinido a Objetivo. El resto de campos debe tener sus papeles definidas en Entrada.

▶ Pulse en el botón Leer valores para rellenar la columna *Valores*.

Figura 18-2
Selección de un campo de objetivo

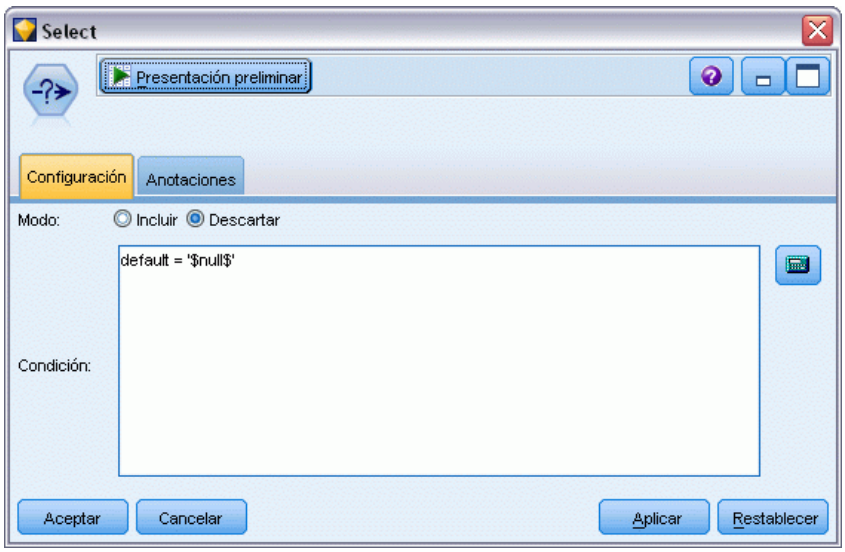
Los casos en los que el objetivo tenga un valor nulo no se utilizan cuando se genera el modelo. Puede excluir esos casos para evitar que se utilicen en una evaluación de modelo.

▶ Añada un nodo Seleccionar al nodo Tipo.

▶ En Modo, seleccione Descartar.

- En la casilla de verificación Condición, introduzca default = '$null$'.

Figura 18-3
Descarte de objetivos nulos

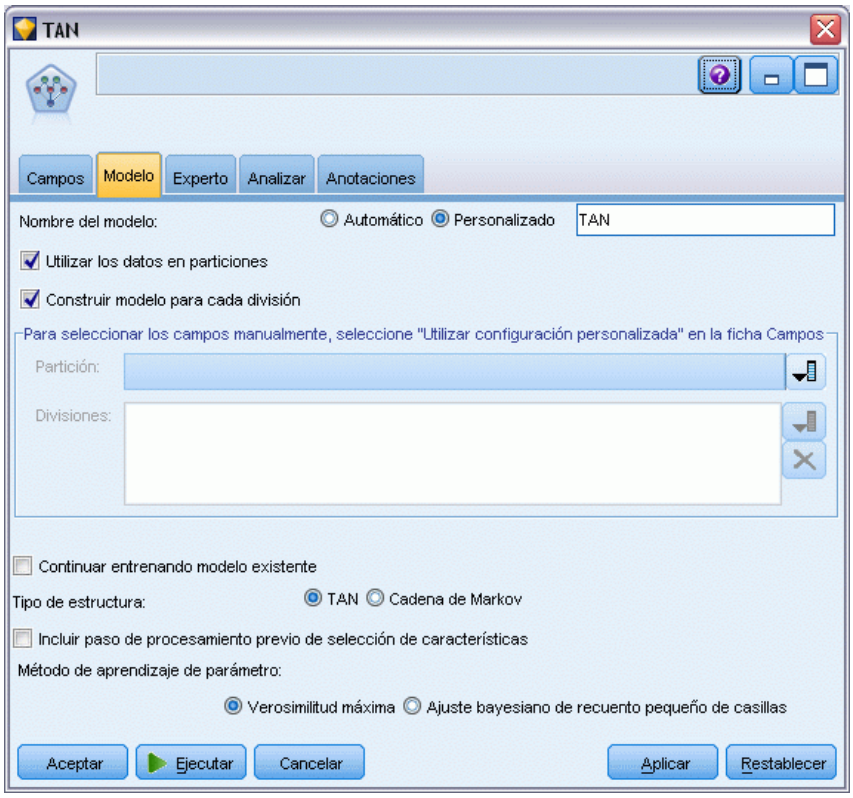


Como puede generar diferentes tipos de redes bayesianas, es recomendable comparar varios tipos para ver qué modelo proporciona los mejores pronósticos. El primero que se debe crear es un modelo redes Naïve Bayes aumentado a árbol (TAN).

- Añada un nodo Red bayesiana al nodo Seleccionar.
- En la pestaña Modelo, seleccione Personalizado para el nombre del modelo e introduzca TAN en el cuadro de texto.

▶ En el tipo de estructura, seleccione TAN y pulse en Aceptar.

Figura 18-4
Creación de un modelo redes Naïve Bayes aumentado a árbol



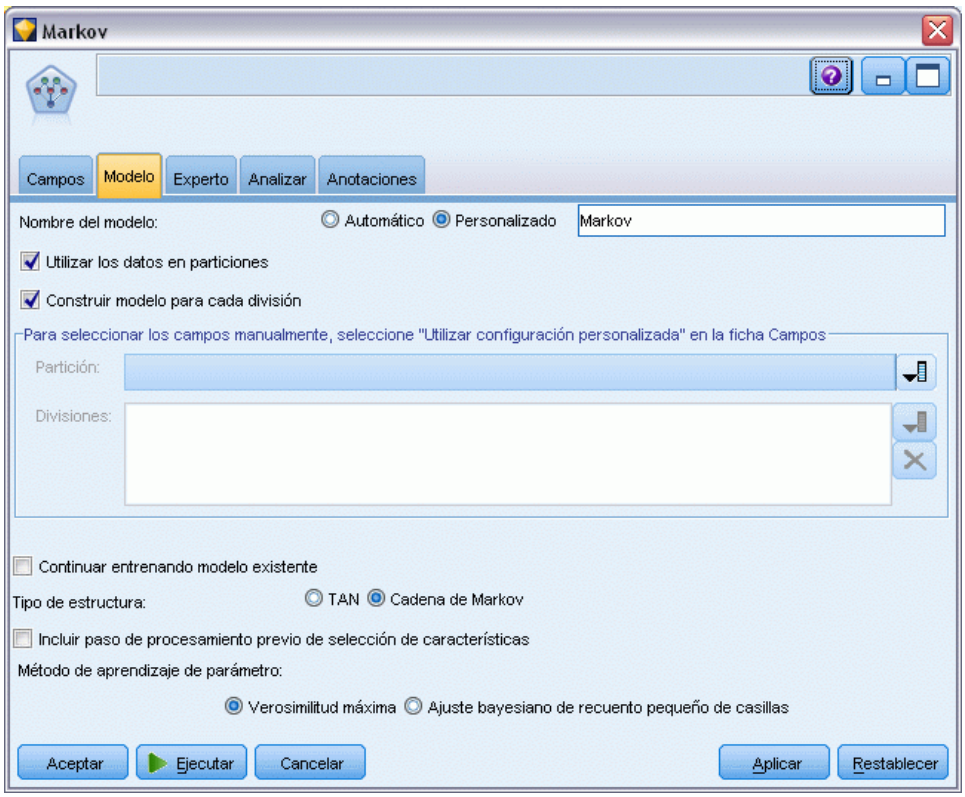
El segundo tipo de modelo tiene una estructura de cadena de Markov.

▶ Añada un segundo nodo Red bayesiana al nodo Seleccionar.

▶ En la pestaña Modelo, seleccione Personalizado para el nombre del modelo e introduzca Markov en el cuadro de texto.

- En el tipo de estructura, seleccione Cadena de Markov y pulse en Aceptar.

Figura 18-5
Creación de un modelo de cadena de Markov

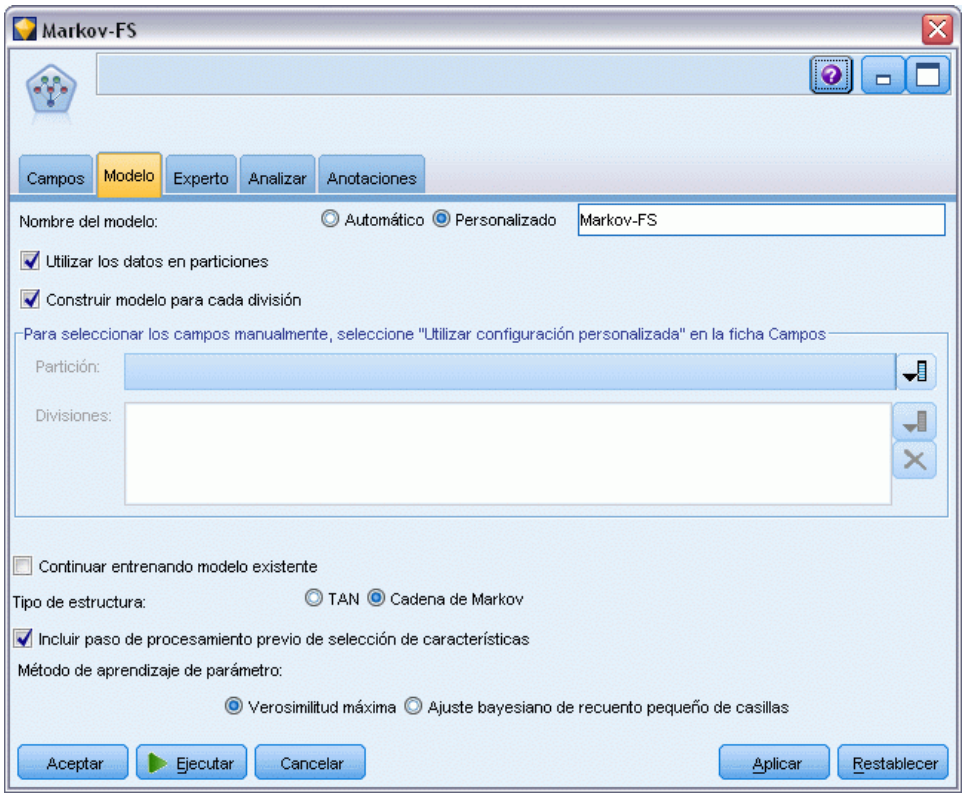


El tercer tipo de modelo tiene una estructura de cadena de Markov y utiliza el procesamiento previo de selección de características para seleccionar las entradas que están relacionadas de forma significativa a la variable de objetivo.

- Añada un tercer nodo Red bayesiana al nodo Seleccionar.
- En la pestaña Modelo, seleccione Personalizado para el nombre del modelo e introduzca Markov-FS en el cuadro de texto.
- En el tipo de estructura, seleccione Cadena de Markov.

▶ Seleccione Incluir paso de procesamiento previo de selección de características y pulse en Aceptar.

Figura 18-6
Creación de un modelo de cadena de Markov con procesamiento previo de selección de características



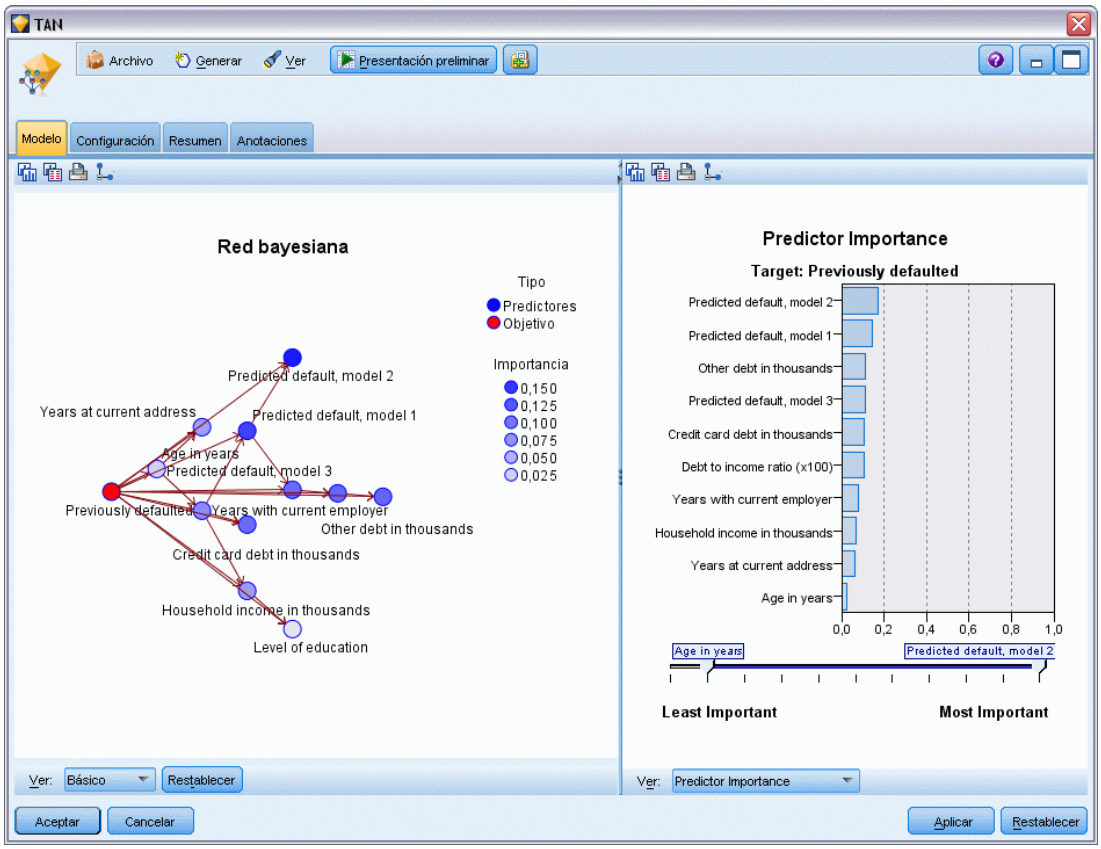
Exploración del modelo

▶ Ejecute la ruta para crear los nuggets de modelo, que se añaden a la ruta y a la paleta Modelos en la esquina superior derecha. Para ver sus detalles, pulse con el botón derecho en cualquiera de los nugget de modelo de la ruta.

La pestaña Modelo del nugget de modelo se dividirá en dos paneles. El panel izquierdo contiene una red de gráficos de nodos que muestra la relación entre el objetivo y sus predictores más importantes, así como las relaciones entre los predictores.

El panel derecho muestra *Importancia de predictores*, que indica la importancia relativa de cada predictor en la estimación del modelo, o *Probabilidades condicionales*, que contiene el valor de probabilidad condicional para cada valor del nodo y cada combinación de valores en sus nodos principales.

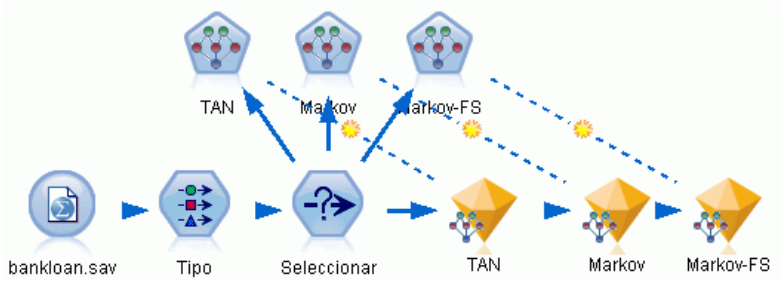
Figura 18-7
Visualización de un modelo redes Naïve Bayes aumentado a árbol



- Conecte el nugget del modelo TAN al nugget de modelo Markov (seleccione Reemplazar en el cuadro de diálogo de advertencia).
- Conecte el nugget Markov al nugget de Markov-FS (seleccione Reemplazar en el cuadro de diálogo de advertencia).

- Alinee los tres nuggets con el nodo Seleccionar para facilitar la visualización.

Figura 18-8
Alineación de los nuggets en la ruta



- Para cambiar el nombre de los resultados del modelo para mayor claridad del gráfico de evaluación que va a crear, añada un nodo Filtro al nugget de modelo de Markov-FS.
- A la derecha de la columna *Campo*, cambie el nombre de $B-default a TAN, de $B1-default a Markov y de $B2-default a Markov-FS.

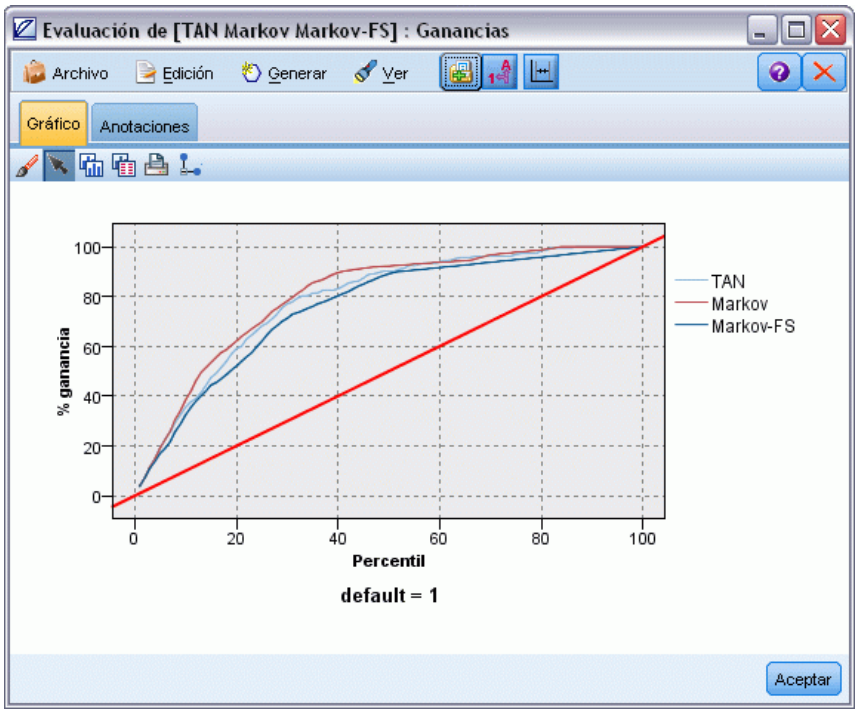
Figura 18-9
Cambio del nombre del campo de modelo

Para comparar la precisión pronosticada de los modelos, puede generar un gráfico de ganancias.

- Añada un nodo de gráfico de evaluación al nodo Filtro y ejecute el nodo de gráfico utilizando su configuración predeterminada.

El gráfico muestra que cada tipo de modelo produce resultados similares; sin embargo, el modelo de Markov es ligeramente mejor.

Figura 18-10
Evaluación de la precisión de los modelos



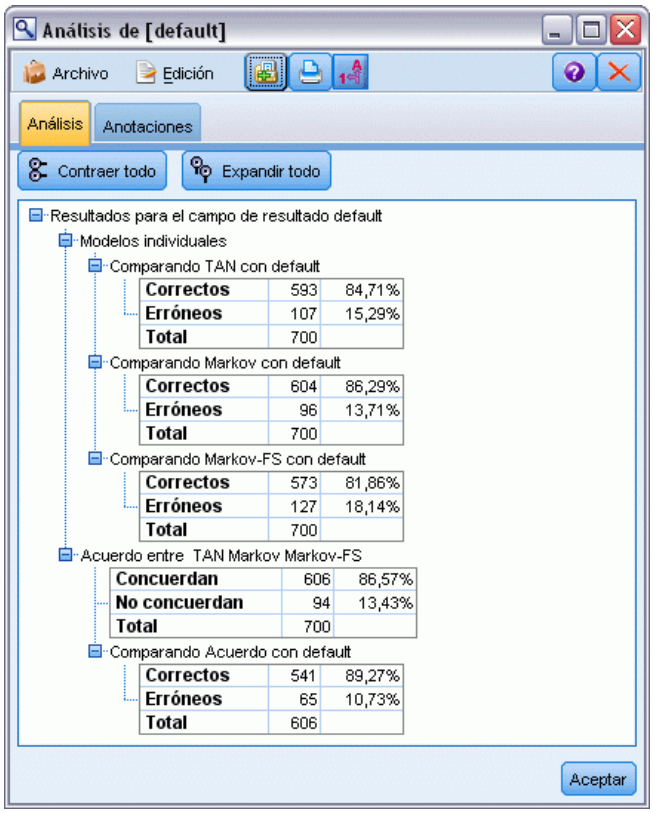
Para comprobar la precisión de los pronósticos de los modelos, puede utilizar un nodo Análisis en lugar del gráfico Evaluación. Muestra la precisión en términos del porcentaje de la precisión de los pronósticos correctos e incorrectos.

- Añada un nodo Análisis al nodo Filtro y ejecute el nodo Análisis utilizado su configuración predeterminada.

Al igual que el gráfico de evaluación, muestra que el modelo de Markov se ligeramente mejor realizando pronósticos correctos, pero el modelo Markov-FS sólo es un par de unidades inferior al del modelo de Markov. Puede significar que es mejor utilizar el modelo Markov-FS ya que

utiliza menos entradas para calcular los resultados, recopilando menos datos y el tiempo de entradas y de procesamiento.

Figura 18-11
Análisis de precisión del modelo



Puede encontrar explicaciones de los fundamentos matemáticos de los métodos de modelado utilizados en IBM® SPSS® Modeler en el *Manual de algoritmos de SPSS Modeler*, disponible en el directorio *\Documentation* del disco de instalación.

Recuerde que estos resultados están basados sólo en los datos de entrenamiento. Para evaluar qué tal se extiende el modelo a otros datos de casos reales, se utilizaría un nodo de partición para reservar un subconjunto de registros para comprobación y validación. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 14.2*.

Capítulo
19

Reentrenamiento de un modelo mensualmente (red bayesiana)

Las redes bayesianas le permiten crear un modelo de probabilidad combinando pruebas observadas y registradas con conocimiento del mundo real de "sentido común" para establecer la probabilidad de instancias utilizando atributos aparentemente no vinculados.

Este ejemplo utiliza la ruta denominada *bayes_churn_retrain.str*, que hace referencia al archivo de datos denominado *telco_Jan.sav* y *telco_Feb.sav*. Estos archivos están disponibles en el directorio *Demos* de cualquier instalación de IBM® SPSS® Modeler y se puede acceder desde el grupo de programas de IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *bayes_churn_retrain.str* se encuentra en el directorio *streams*.

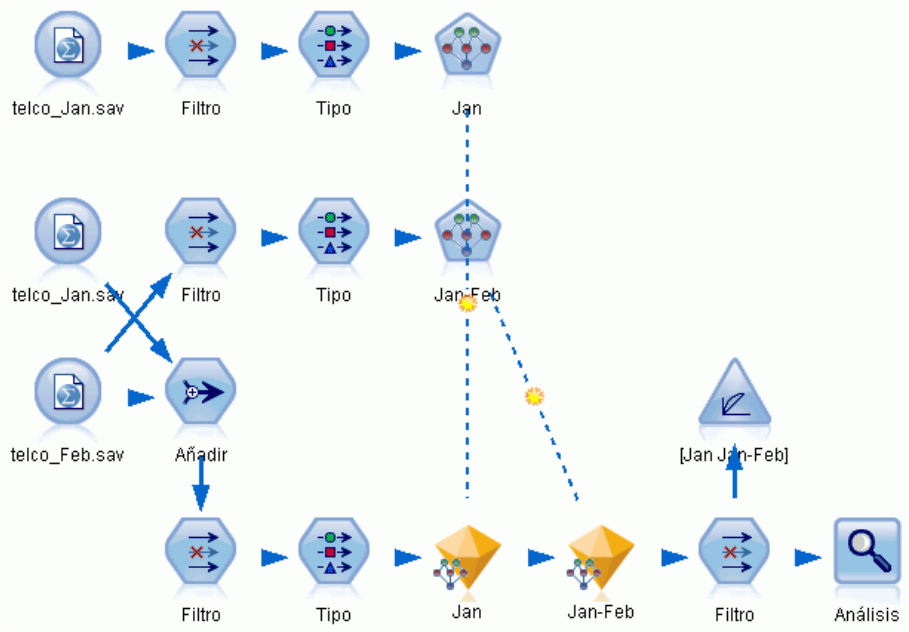
Por ejemplo, suponga que un proveedor de telecomunicaciones está preocupado por el número de clientes que se pasan a la competencia (abandono). Si se pueden utilizar datos históricos de clientes para pronosticar los clientes con más probabilidades de abandono en el futuro, se puede ofrecer a estos clientes incentivos u otras ofertas para evitar que se vayan a otro proveedor de servicios.

Este ejemplo se centra en el uso de los datos existentes de abandono de un mes para pronosticar los clientes con más probabilidades de abandono futuro y añadirlos a los datos del mes siguiente para refinar y volver a entrenar el modelo.

Generación de la ruta

- Añada un nodo de origen de archivo Statistics apuntando a *telco_Jan.sav* en la carpeta *Demos*.

Figura 19-1
Ruta de ejemplo de red bayesiana

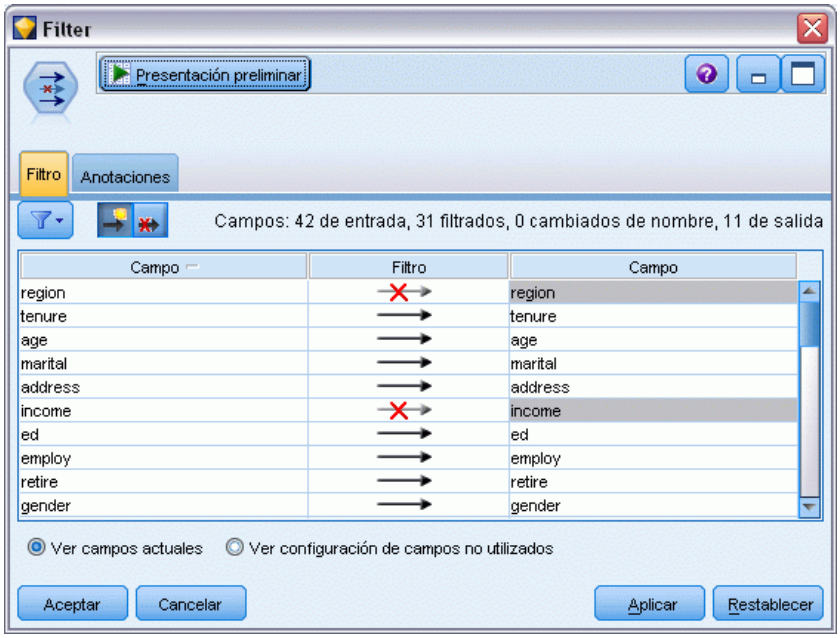


Análisis previos muestran que numerosos campos de datos tienen poca importancia a la hora de pronosticar la tasa de abandono. Estos campos se pueden filtrar por sus conjuntos de datos para aumentar la velocidad de procesamiento cuando genera y puntúa modelos.

- Añada un nodo Filtro al nodo de origen.
- Excluya todos los campos excepto *dirección*, *edad*, *abandono*, *catpers*, *educ*, *empleo*, *género*, *marital*, *residen*, *jubilación* y *periodo*.

▶ Pulse en Aceptar.

Figura 19-2
Filtrado de campos innecesarios



▶ Añada un nodo Tipo al nodo Filtro.

▶ Abra el nodo Tipo y pulse en el botón Leer valores para rellenar la columna *Valores*.

- Para que el nodo Evaluación pueda acceder al valor que es verdadero y falso, defina el nivel de medición para el campo *abandono* a Marca y defina su papel a Objetivo. Pulse en Aceptar.

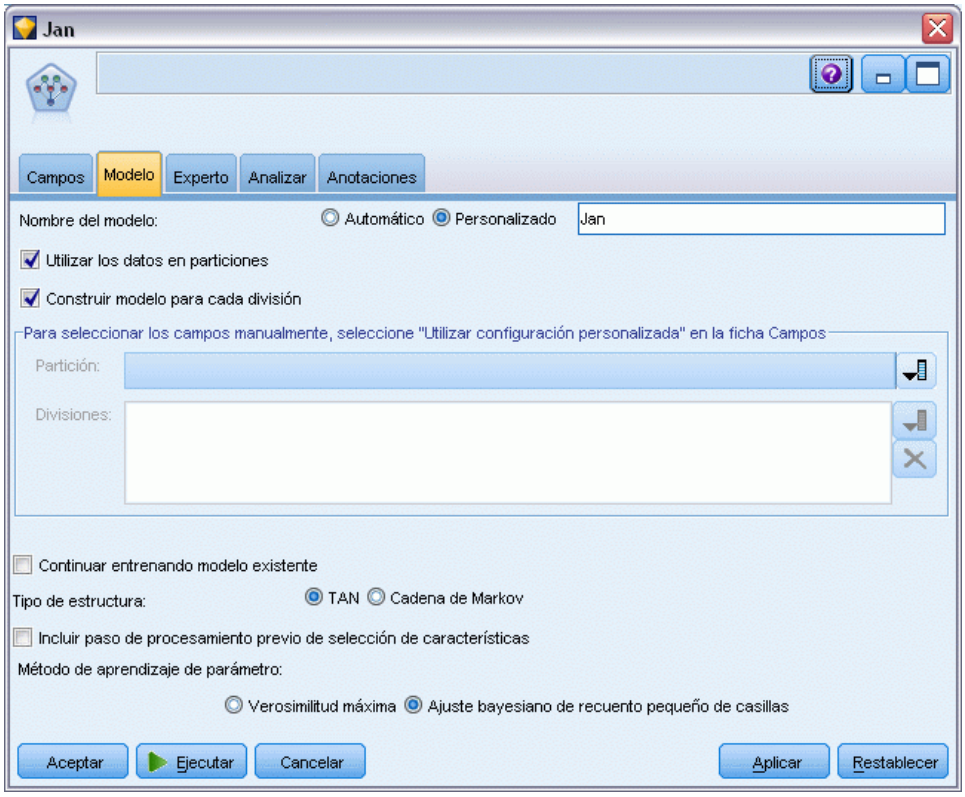
Figura 19-3
Selección de un campo de objetivo

Puede generar diferentes tipos de redes bayesianas; sin embargo, para este ejemplo va a generar un modelo Naïve Bayes aumentado a árbol (TAN). Este modelo crea una red de grandes dimensiones y garantiza que ha incluido todos los enlaces posibles entre las variables de datos, generando un modelo inicial robusto.

- Añada un nodo Red bayesiana al nodo Tipo.
- En la pestaña Modelo, seleccione Personalizado para el nombre del modelo e introduzca Ene en el cuadro de texto.
- Para el método de aprendizaje de parámetro, seleccione Ajuste bayesiano de recuentos de casillas de tamaño reducido.

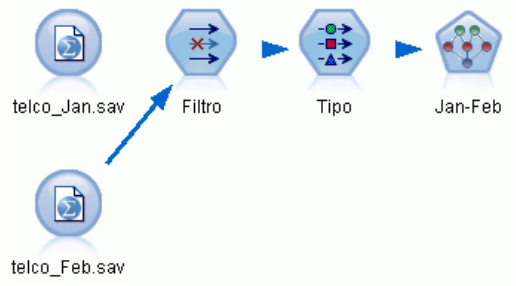
- Pulse en Ejecutar. El nugget del modelo se añade a la ruta y a la paleta Modelos en la esquina superior derecha.

Figura 19-4
Creación de un modelo redes Naïve Bayes aumentado a árbol



- Añada un nodo de origen de archivo Statistics apuntando a *telco_Feb.sav* en la carpeta *Demos*.
- Añada este nuevo nodo de origen al nodo Filtro (en el cuadro de diálogo de advertencia, seleccione Reemplazar para sustituir la conexión con el nodo origen anterior).

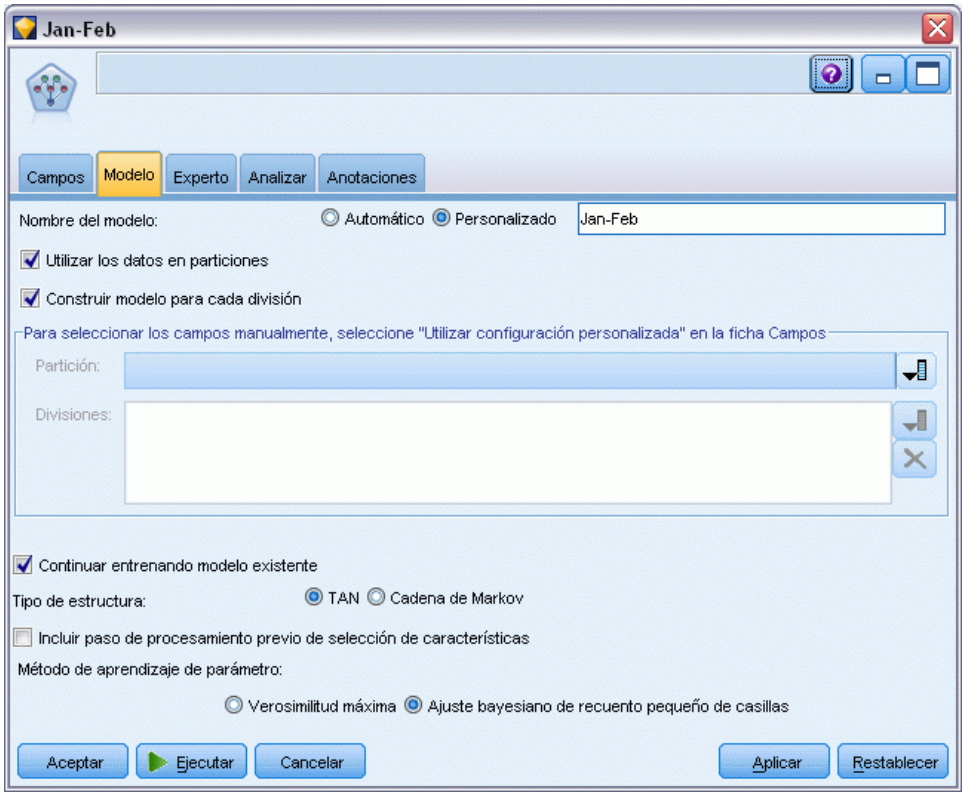
Figura 19-5
Adición de los datos del segundo mes



- En la pestaña Modelo del nodo de red bayesiana, seleccione Personalizado para el nombre del modelo e introduzca Ene-Feb en el cuadro de texto.
- Seleccione Continuar entrenando modelo existente.

▶ Pulse en Ejecutar. El nugget modelo sobrescribe el nugget existente en la ruta, pero también se añade a la paleta Modelos en la esquina superior derecha.

Figura 19-6
Reentrenamiento del modelo



Evaluación del modelo

Para comparar los modelos, debe combinar los dos conjuntos de datos.

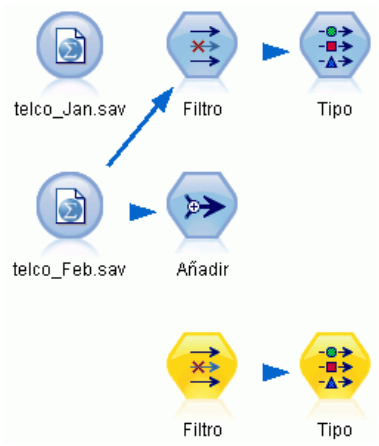
▶ Añada un nodo Añadir y añádales los nodos de origen *telco_Jan.sav* y *telco_Feb.sav*.

Figura 19-7
Añada los dos orígenes de datos

▶ Copie los nodos Filtro y Tipo anteriores de la ruta y péguelos en el lienzo de rutas.

▶ Añada el nodo Añadir al nodo Filtro que ha copiado.

Figura 19-8
Copia de los nodos en la ruta

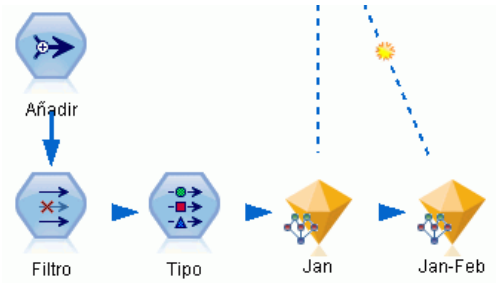


Los nuggets de los dos modelos de red bayesiana se encuentran en la paleta Modelos en la esquina superior derecha.

▶ Pulse dos veces en el nugget de modelo para llevarlo a la ruta y añadirlo al nodo Tipo recién copiado.

- Añada el nugget del modelo Ene-Feb que ya está en la ruta al nugget de modelo Ene.
- Abra el nugget de modelo Ene.

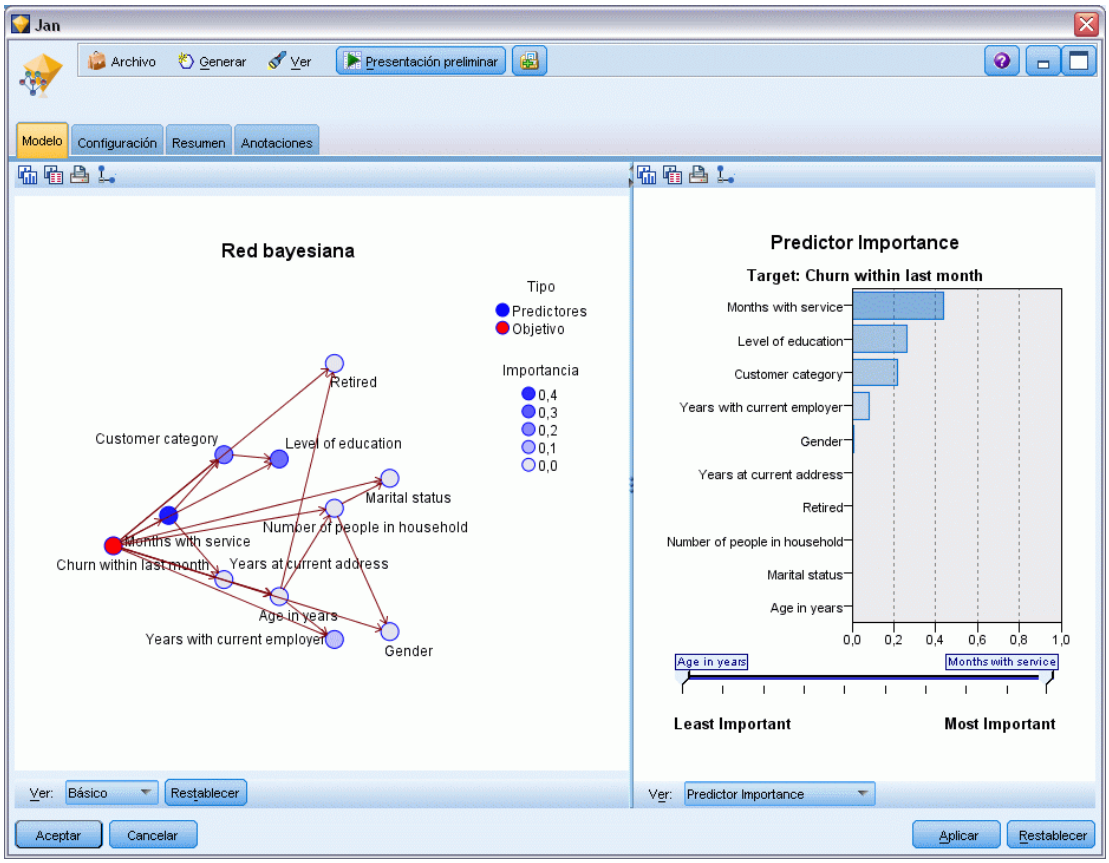
Figura 19-9
Adición de los nuggets a la ruta



La pestaña Modelo del nugget de modelo de red bayesiana se dividirá en dos columnas. La columna izquierda contiene una red de gráficos de nodos que muestra la relación entre el objetivo y sus predictores más importantes, así como las relaciones entre los predictores.

La columna derecha muestra *Importancia de predictores*, que indica la importancia relativa de cada predictor en la estimación del modelo, o *Probabilidades condicionales*, que contiene el valor de probabilidad condicional para cada valor del nodo y cada combinación de valores en sus nodos principales.

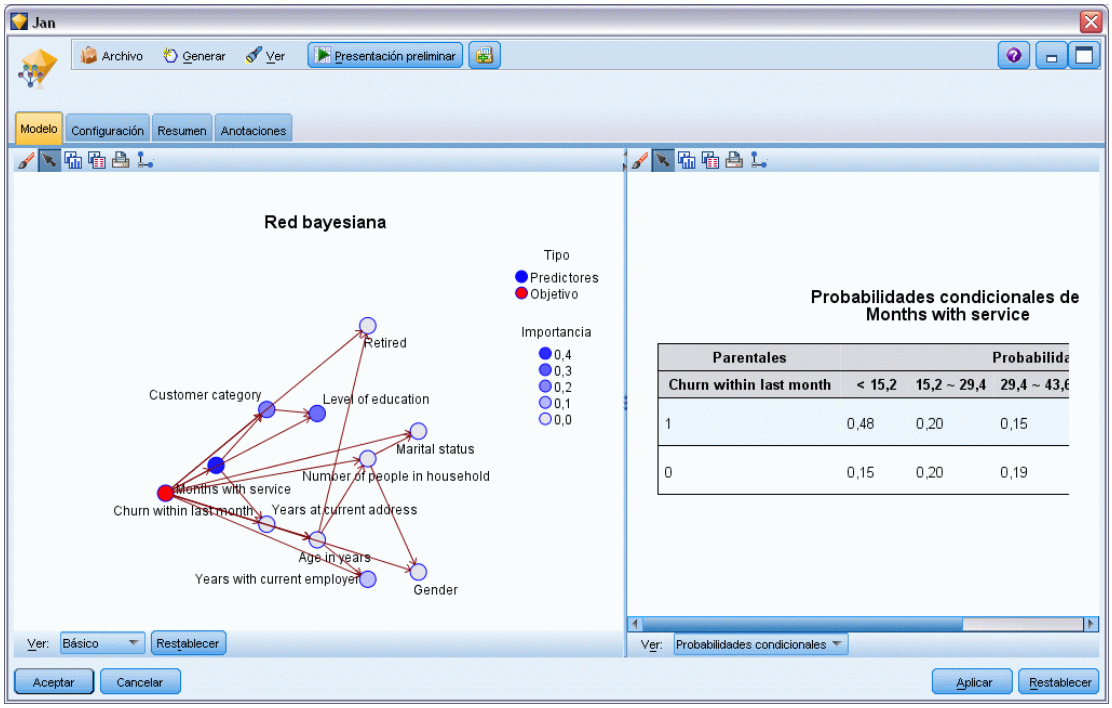
Figura 19-10
Modelo de red bayesiana mostrando la importancia de predictor



Para mostrar las probabilidades condicionales de un código, pulse en un nodo en la columna izquierda. La columna derecha se actualiza para mostrar los detalles necesarios.

Se muestran las probabilidades condicionales de cada intervalo en los que se han dividido los valores de datos en relación a los nodos hermanos y nodos parentales.

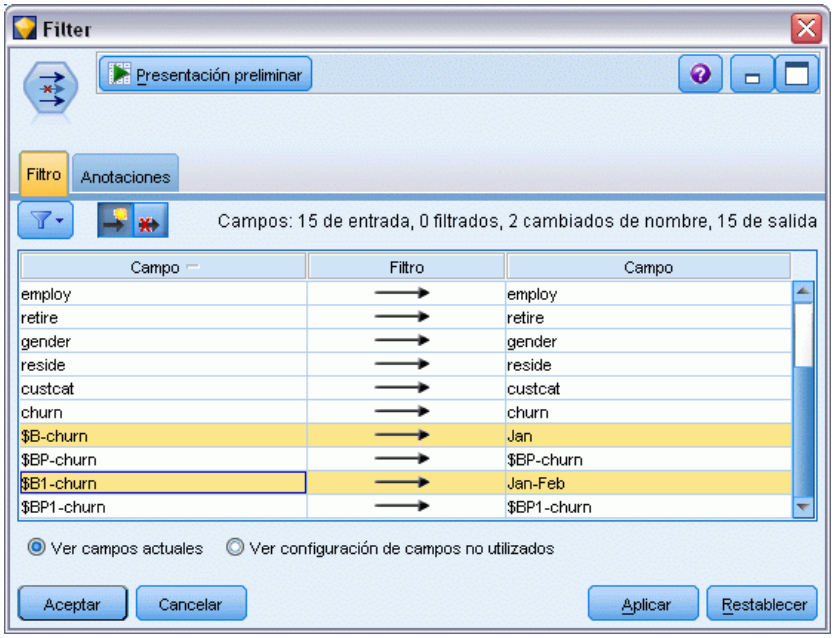
Figura 19-11
Modelo de red bayesiana con probabilidades condicionales



- Para cambiar el nombre los resultados del modelo, añada un nodo Filtro al nugget del modelo Ene-Feb.

▶ En la columna derecha *Campo*, cambie el nombre de $B-churn a Ene y $B1-churn a Ene-Feb.

Figura 19-12
Cambio del nombre del campo de modelo



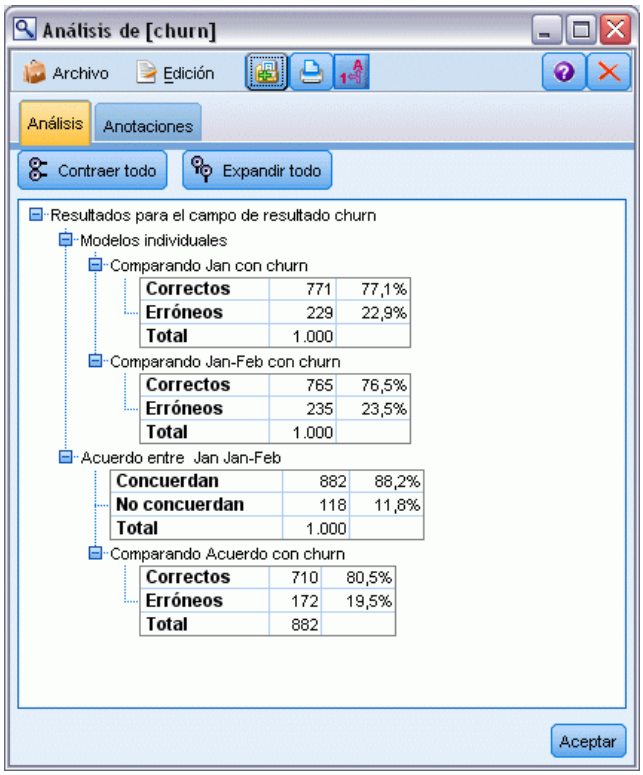
Para comprobar la calidad con la que cada modelo pronostica el abandono, utilice un nodo Análisis. Este nodo muestra el porcentaje de precisión ende los pronósticos correctos e incorrectos.

▶ Añada un nodo Análisis al nodo Filtro.

▶ Abra el nodo Análisis y pulse en Ejecutar.

Mostrará que ambos modelos tienen un grado similar de precisión cuando se pronostican abandonos.

Figura 19-13
Análisis de precisión del modelo



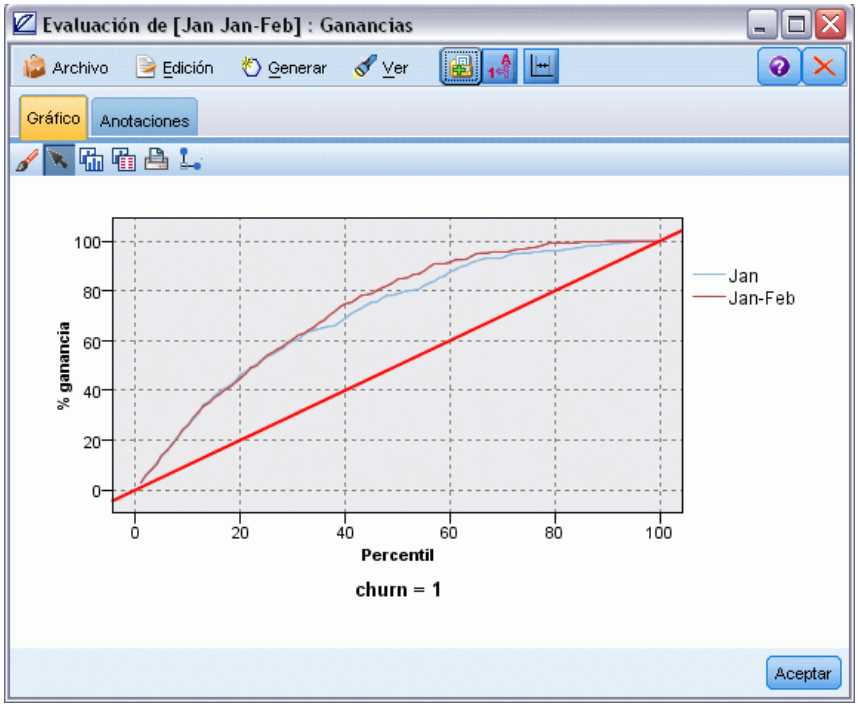
Como alternativa al nodo Análisis, puede utilizar un gráfico de evaluación para comparar la precisión de los pronósticos de los modelos, generando un gráfico de ganancias.

- Añada un nodo de gráfico de evaluación al nodo Filtro.

y ejecute el nodo de gráfico utilizando su configuración predefinida.

Al igual que el nodo Análisis, el gráfico muestra que cada tipo de modelo produce resultados similares; sin embargo, el modelo reentrenado que utiliza los datos de ambos meses es ligeramente mejor, porque tiene un mayor nivel de confianza en sus predicciones.

Figura 19-14
Evaluación de la precisión de los modelos



Puede encontrar explicaciones de los fundamentos matemáticos de los métodos de modelado utilizados en IBM® SPSS® Modeler en el *Manual de algoritmos de SPSS Modeler*, disponible en el directorio *\Documentation* del disco de instalación.

Recuerde que estos resultados están basados sólo en los datos de entrenamiento. Para evaluar qué tal se extiende el modelo a otros datos de casos reales, se utilizaría un nodo de partición para reservar un subconjunto de registros para comprobación y validación. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 14.2*.

Capítulo
20

Promoción de ventas al por menor (Red neuronal/C&RT)

Este ejemplo está relacionado con los datos que describen la gama de productos en venta y los efectos de la promoción en las ventas. (Este dato es totalmente ficticio.) Su objetivo en el ejemplo es predecir los efectos de las promociones en las ventas futuras. Similar al ejemplo del control de estado, el proceso de minería de datos consta de las fases de exploración, preparación de datos, entrenamiento y comprobación.

Este ejemplo utiliza las rutas denominadas *goodsplot.str* y *goodslearn.str*, que hacen referencia a los archivos de datos denominados *GOODS1n* y *GOODS2n*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. La ruta *goodsplot.str* está en la carpeta *streams*, mientras que el archivo *goodslearn.str* se encuentra en el directorio *streams*.

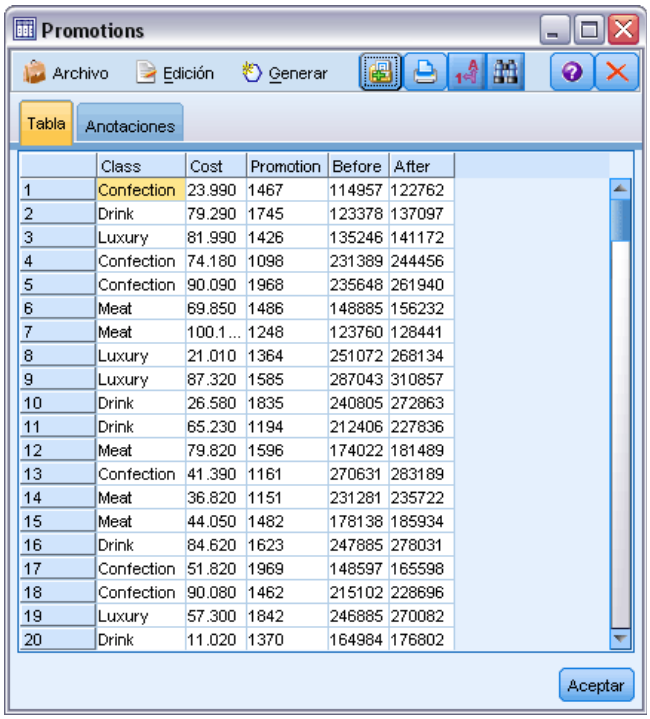
Examen de los datos

Cada registro contiene:

- *Clase*. Tipo de producto.
- *Coste*. Precio unitario.
- *Promoción*. Índice de cantidades gastadas en una promoción determinada.
- *Antes*. Ingresos antes de la promoción.
- *Después*. Ingresos después de la promoción.

La ruta *goodsplot.str* contiene una ruta simple para mostrar los datos en una tabla. Los dos campos de ingresos *Antes* y *Después*) se expresan en términos absolutos. Sin embargo, es probable que sea más útil la figura del aumento de los ingresos después de la promoción (y que es de suponer que se produce como resultado de la misma).

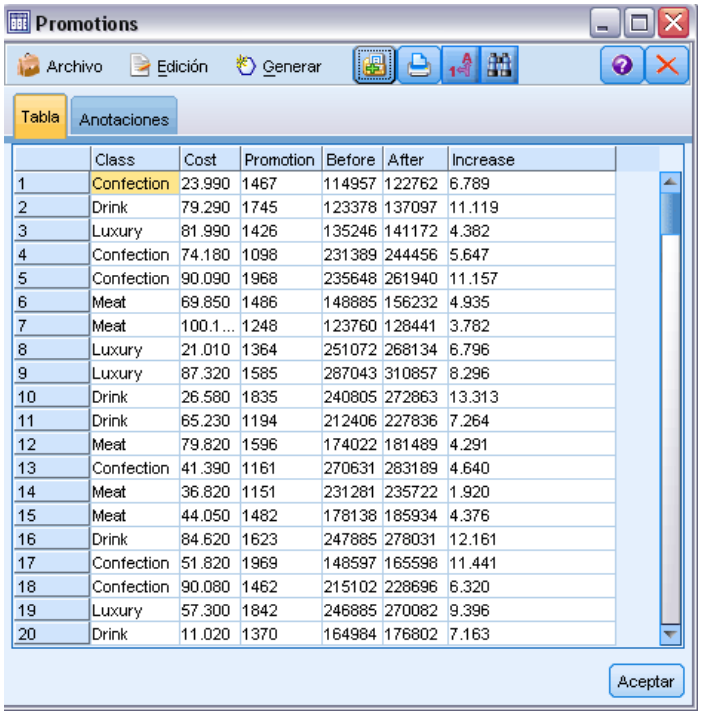
Figura 20-1
Efectos de la promoción en las ventas de productos



	Class	Cost	Promotion	Before	After
1	Confection	23.990	1467	114957	122762
2	Drink	79.290	1745	123378	137097
3	Luxury	81.990	1426	135246	141172
4	Confection	74.180	1098	231389	244456
5	Confection	90.090	1968	235648	261940
6	Meat	69.850	1486	148885	156232
7	Meat	100.1...	1248	123760	128441
8	Luxury	21.010	1364	251072	268134
9	Luxury	87.320	1585	287043	310857
10	Drink	26.580	1835	240805	272863
11	Drink	65.230	1194	212406	227836
12	Meat	79.820	1596	174022	181489
13	Confection	41.390	1161	270631	283189
14	Meat	36.820	1151	231281	235722
15	Meat	44.050	1482	178138	185934
16	Drink	84.620	1623	247885	278031
17	Confection	51.820	1969	148597	165598
18	Confection	90.080	1462	215102	228696
19	Luxury	57.300	1842	246885	270082
20	Drink	11.020	1370	164984	176802

*goodsplot.str*también contiene un nodo derivar este valor, expresado como un porcentaje de los ingresos antes de la promoción, en un campo llamado *Aumento* y muestra una tabla con dicho campo.

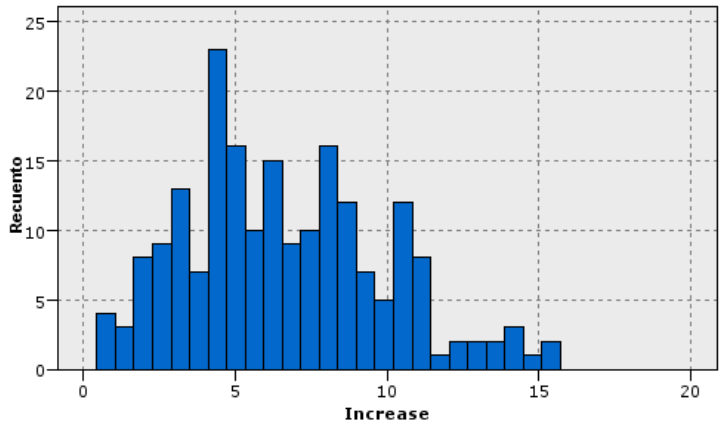
Figura 20-2
Aumento de los ingresos después de la promoción



	Class	Cost	Promotion	Before	After	Increase
1	Confection	23.990	1467	114957	122762	6.789
2	Drink	79.290	1745	123378	137097	11.119
3	Luxury	81.990	1426	135246	141172	4.382
4	Confection	74.180	1098	231389	244456	5.647
5	Confection	90.090	1968	235648	261940	11.157
6	Meat	69.850	1486	148885	156232	4.935
7	Meat	100.1...	1248	123760	128441	3.782
8	Luxury	21.010	1364	251072	268134	6.796
9	Luxury	87.320	1585	287043	310857	8.296
10	Drink	26.580	1835	240805	272863	13.313
11	Drink	65.230	1194	212406	227836	7.264
12	Meat	79.820	1596	174022	181489	4.291
13	Confection	41.390	1161	270631	283189	4.640
14	Meat	36.820	1151	231281	235722	1.920
15	Meat	44.050	1482	178138	185934	4.376
16	Drink	84.620	1623	247885	278031	12.161
17	Confection	51.820	1969	148597	165598	11.441
18	Confection	90.080	1462	215102	228696	6.320
19	Luxury	57.300	1842	246885	270082	9.396
20	Drink	11.020	1370	164984	176802	7.163

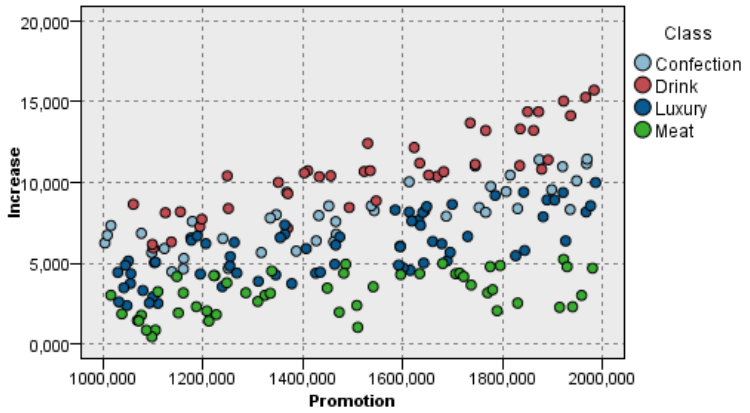
Además, la ruta muestra un histograma del aumento y un diagrama del aumento frente a los costes de promoción, superpuestos con la categoría del producto en cuestión.

Figura 20-3
Histograma del aumento de ingresos



El diagrama muestra que para cada clase de producto existe una relación casi lineal entre el aumento de los ingresos y el coste de la promoción. Por lo tanto, parece probable que un árbol de decisión o red neuronal pueda pronosticar, con una precisión razonable, el aumento de los ingresos de los otros campos disponibles.

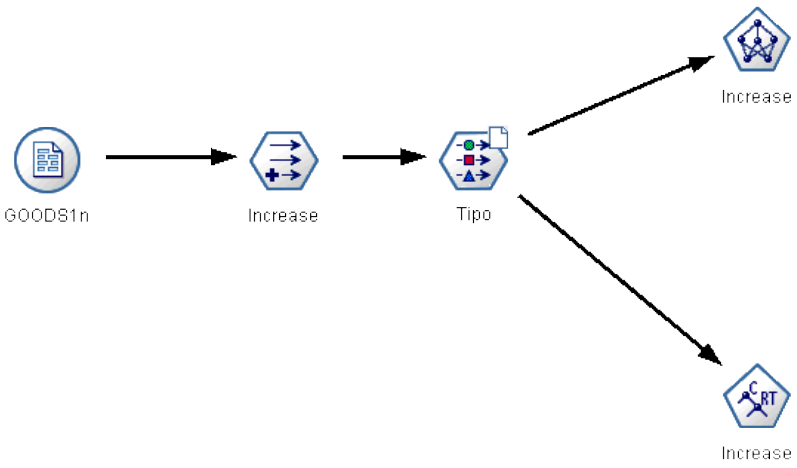
Figura 20-4
Aumento de los ingresos frente a gastos de promoción



Aprendizaje y comprobación

La ruta *goodslearn.str* entrena una red neuronal y un árbol de decisión para realizar el pronóstico de aumento de los ingresos.

Figura 20-5
Ruta de modelado goodslearn.str



Una vez que haya ejecutado los nodos de modelos y generado los modelos reales, puede comprobar los resultados del proceso de aprendizaje. Hágalo conectando el árbol de decisión y la red en serie entre el nodo Tipo y un nodo Análisis nuevo, cambiando el archivo de entrada (de datos)

GOODS2n y ejecutando el nodo Análisis. A partir de los resultados de este nodo, en concreto a partir de la correlación lineal entre el aumento pronosticado y la respuesta correcta, verá que los sistemas entrenados pronostican el aumento de los ingresos con un alto grado de corrección.

Una exploración en detalle se podría centrar en los casos en los que los sistemas entrenados cometen errores relativamente grandes. Podría identificarse representando el aumento de los ingresos pronosticado frente al aumento real. Los valores atípicos de este gráfico podrían seleccionarse utilizando los gráficos interactivos de IBM® SPSS® Modeler y, a partir de sus propiedades, se podría ajustar la descripción de los datos o el proceso de aprendizaje para mejorar la precisión.

Capítulo
21
Control de estado (Red neuronal/C5.0)

Este ejemplo se refiere a la información del estado de control de un equipo y al problema para reconocer y pronosticar estados de error. Los datos se crean a partir de una simulación ficticia y consisten en un conjunto de series concatenadas medidas durante un período. Cada registro es un informe instantáneo del equipo en cuanto a lo siguiente:

- *Hora*. Un entero.
- *Potencia*. Un entero.
- *Temperatura*. Un entero.
- *Presión*. 0 si es normal, 1 si es una advertencia de presión pasajera.
- *Tiempo funcionamiento*. Fecha desde la última revisión.
- *Estado*. Normalmente, 0; cambia a código de error cuando hay un error (101, 202 o 303).
- *Resultado*. En esta serie temporal aparece el código de error, o bien 0 si no se produce ningún error. (Estos códigos están sólo disponibles a posteriori.)

Este ejemplo utiliza las rutas denominadas *condplot.str* y *condlearn.str*, que hacen referencia a los archivos de datos denominados *COND1n* y *COND2n*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. Los archivos *condplot.str* y *condlearn.str* se encuentran en el directorio *streams*.

En cada serie temporal hay una serie de registros de un período de funcionamiento normal seguido de un período que conduce al error, como se muestra en la siguiente tabla:

Time	Potencia	Temperatura	Presión	Tiempo funcionamiento	Estado	Resultado
0	1059	259	0	404	0	0
1	1059	259	0	404	0	0
			...			
51	1059	259	0	404	0	0
52	1059	259	0	404	0	0
53	1007	259	0	404	0	303
54	998	259	0	404	0	303
			...			
89	839	259	0	404	0	303
90	834	259	0	404	303	303
0	965	251	0	209	0	0
1	965	251	0	209	0	0
			...			
51	965	251	0	209	0	0
52	965	251	0	209	0	0
53	938	251	0	209	0	101

Time	Potencia	Temperatura	Presión	Tiempo funcionamiento	Estado	Resultado
54	936	251	0	209	0	101
			...			
208	644	251	0	209	0	101
209	640	251	0	209	101	101

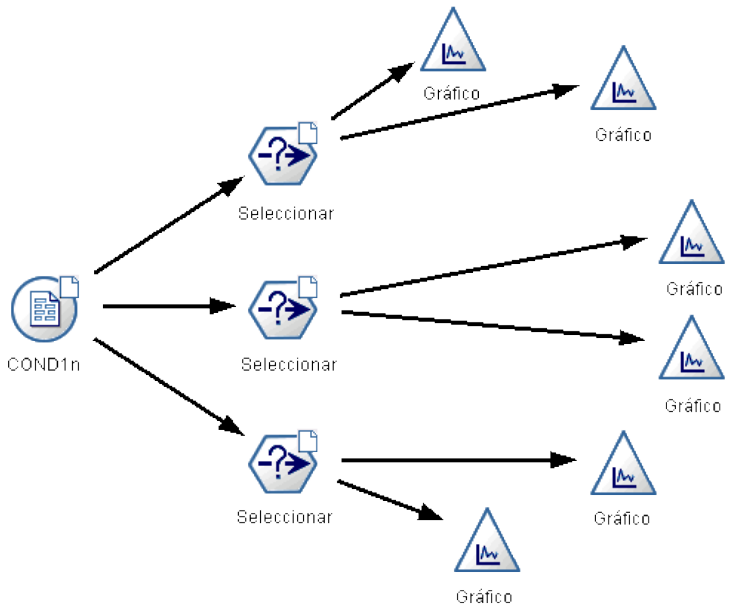
El siguiente proceso es habitual en la mayoría de los proyectos de minería de datos:

- Examine los datos para determinar qué atributos pueden ser relevantes para pronosticar o reconocer estados de interés.
- Conserve esos atributos (si todavía están presentes) o derívelos y añádalos a los datos si fuese necesario.
- Utilice los datos resultantes para entrenar reglas y redes neuronales.
- Compruebe los sistemas de entrenamiento utilizando datos de comprobación independientes.

Examen de los datos

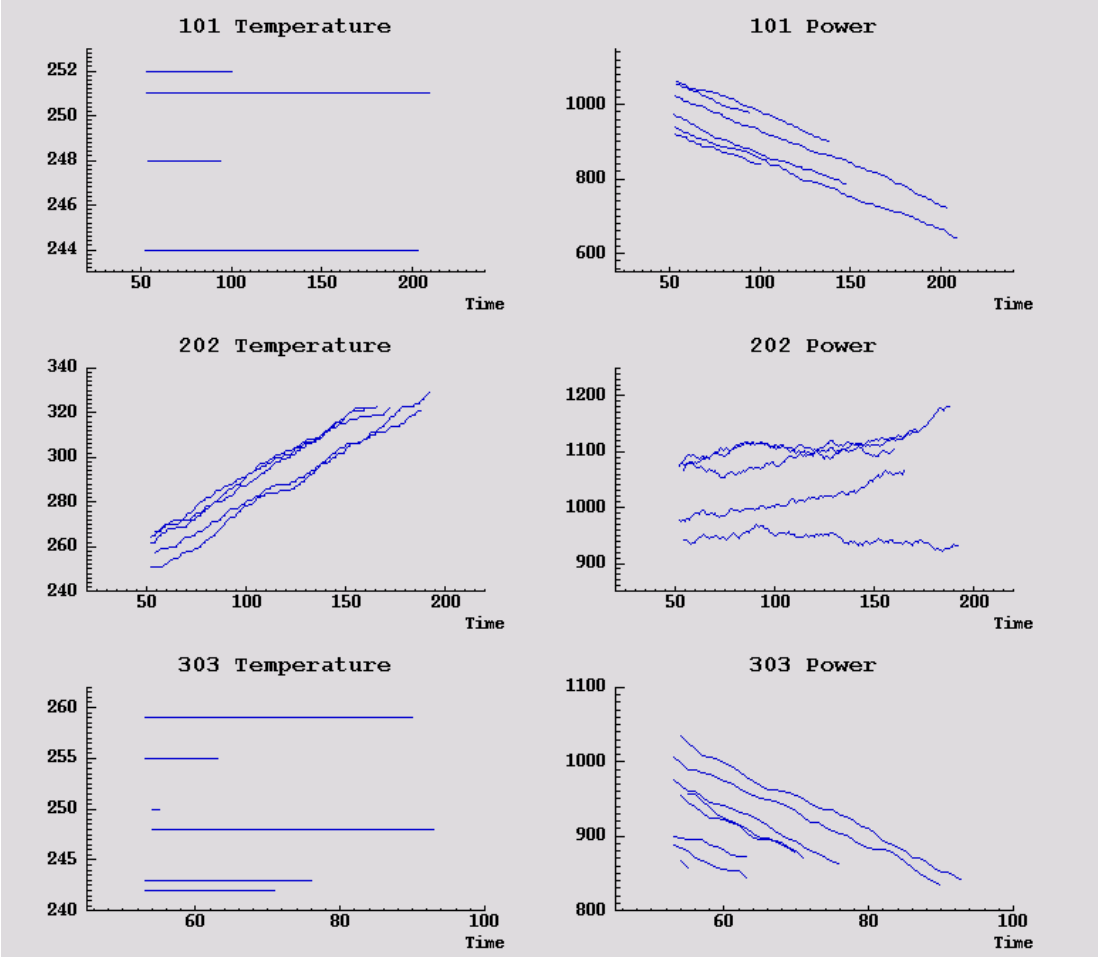
El archivo *condplot.str* muestra la primera parte del proceso. Contiene una ruta que representa un número de gráficos. Si la serie temporal de temperatura o potencia contiene patrones visibles, puede diferenciar entre condiciones de error inminentes o pronosticar quizás su ocurrencia. Tanto para la temperatura como para la potencia, la ruta que hay debajo muestra la serie temporal asociada con los tres códigos de error diferentes en gráficos separados, lo que produce seis gráficos. Los nodos de selección separan los datos asociados con los diferentes códigos de error.

Figura 21-1
Ruta condplot



Los resultados de esta ruta se muestran en la siguiente figura.

Figura 21-2
Temperatura y potencia durante un período de tiempo



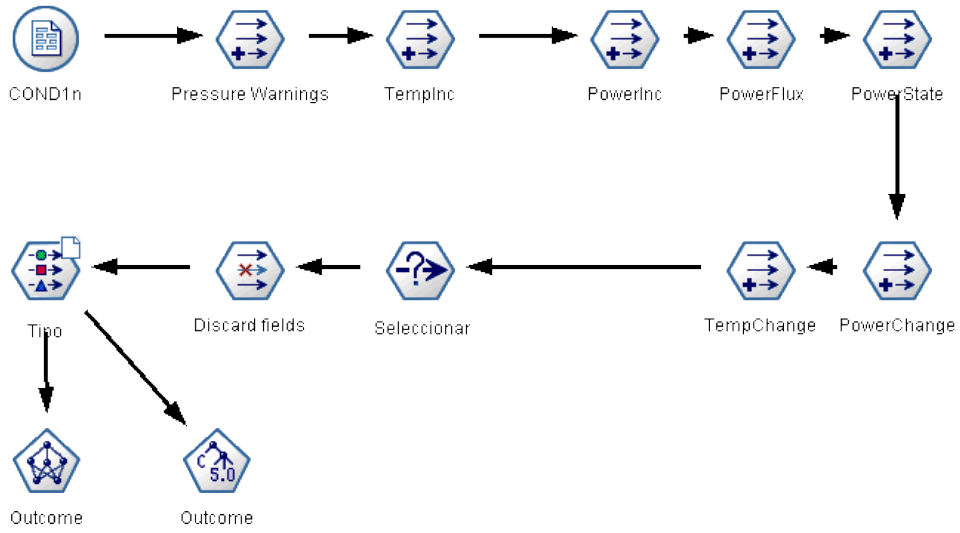
Los gráficos muestran con claridad patrones que distinguen los errores 202 de los errores 101 y 303. Los errores 202 muestran el aumento de temperatura y las fluctuaciones de potencia durante un período de tiempo; los otros errores, no. Sin embargo, los patrones que distinguen entre los errores 101 y 303 son menos claros. Ambos errores muestran una temperatura constante y una bajada de potencia, pero dicha bajada parece más pronunciada en el caso de los errores 303.

Según estos gráficos, parece que la presencia y la tasa de cambio tanto de la temperatura como de la potencia así como la presencia y el grado de fluctuación son relevantes para predecir y distinguir errores. Por lo tanto, estos atributos se deben añadir a los datos antes de aplicar los sistemas de aprendizaje.

Preparación de datos

Según los resultados de la exploración de los datos, la ruta *condlearn.str* proporciona los datos relevantes y aprende a pronosticar errores.

Figura 21-3
Ruta condlearn



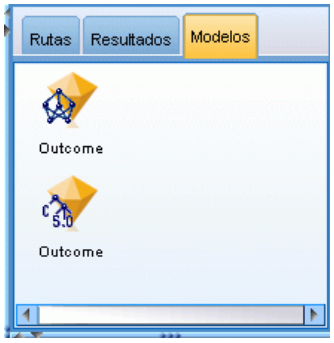
La ruta utiliza un número de nodos Derivar para preparar los datos para el modelado.

- **Nodo Archivo var**. Lee el archivo de datos *COND1n*.
- **Derivar advertencias de presión**. Cuenta el número de advertencias de presión pasajeras. Restablecer cuando el tiempo vuelve a 0.
- **Derivar Cambtemp**. Calcula la tasa pasajera de cambio de temperatura utilizando @DIFF1.
- **Derivar Cambpot**. Calcula la tasa pasajera de cambio de potencia utilizando @DIFF1.
- **Derivar Flujopot**. Una marca, que es verdadera si la potencia varió en direcciones opuestas en el último registro y en el actual (es decir, durante un pico o una bajada de potencia).
- **Derivar Estadopot**. Estado que comienza como *Estable* y cambia a *Fluctuante* cuando se detectan dos flujos de potencia sucesivos. Vuelve a cambiar a *Estable* sólo cuando ha habido un flujo de potencia durante cinco intervalos de tiempo o cuando se restablece la *Hora*.
- **Cambiopotencia**. Promedio de *Cambpot* durante los últimos cinco intervalos de tiempo.
- **Cambtemp**. Promedio de *Cambtemp* durante los últimos cinco intervalos de tiempo.
- **Desechar inicial (seleccionar)**. Descarta el primer registro de cada serie temporal para evitar saltos grandes (incorrectos) de *potencia* y *temperatura* en los límites.
- **Desechar campos**. Filtra los registros *Tiempo funcionamiento*, *Estado*, *Resultado*, *Advertencias de presión*, *Estadopot*, *Cambiopotencia* y *Cambtemp*.
- **Tipo**. Define el papel del nodo *Resultado* como Objetivo (el campo que se ha de pronosticar). Además, define el nivel de medición de *Resultado* como Nominal, *Advertencias de presión* como Continuo y *Estadopot* como Marca.

Aprendiendo

La ejecución de la ruta en *condlearn.str* entrena la regla C5.0 y la red neuronal. El entrenamiento de la red puede tomarse algún tiempo, pero el entrenamiento se puede interrumpir antes de tiempo para guardar una red que produzca resultados razonables. Una vez que se completa el aprendizaje, la pestaña Modelos en la parte superior derecha de la ventana Administradores parpadea para avisarle de que se crearon dos nuevos nuggets: uno representa la red neuronal y el otro representa la regla.

Figura 21-4
Administrador de modelos con nuggets de modelos



Los nuggets de modelos también se añaden a la ruta existente para comprobar el sistema o exportar los resultados del modelo. En este ejemplo, comprobaremos los resultados del modelo.

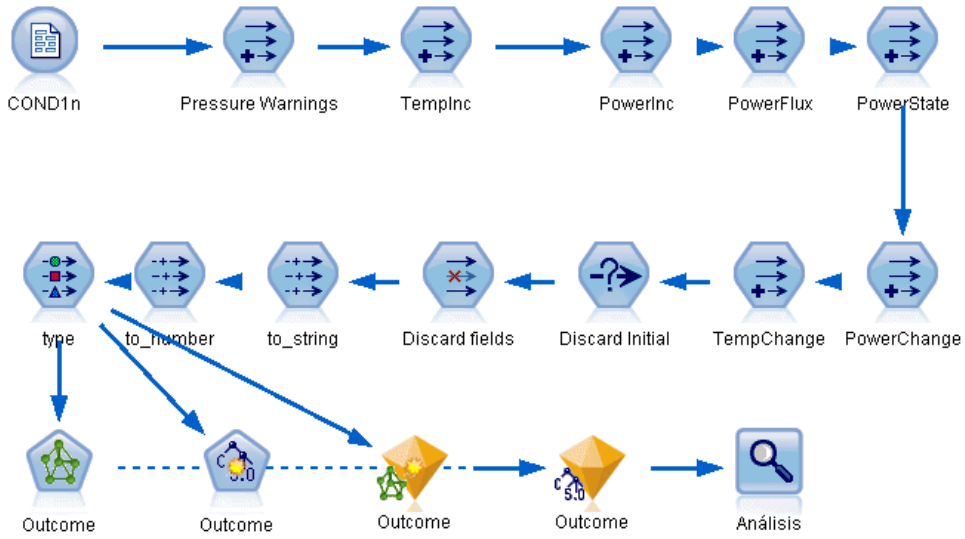
Comprobación

Los nuggets de modelos se añaden a la ruta, ambos conectados al nodo Tipo.

- Vuelva a posicionar los nuggets como se muestra, de modo que el nodo Tipo se conecte con el nugget de red neuronal, que se conecta con el nugget C5.0.
- Añada un nodo Análisis al nugget C5.0.

- Edite el nodo de origen original se edita a continuación para leer el archivo *COND2n* (en lugar de *COND1n*), ya que *COND2n* contiene datos de comprobación no mostrados.

Figura 21-5
Comprobación de la red entrenada



- Abra el nodo Análisis y pulse en Ejecutar.

  Al hacerlo se generan cifras que reflejan la precisión de la regla y la red entrenadas.

Capítulo
22

Clasificación de clientes de telecomunicaciones (Análisis discriminante)

El análisis discriminante es una técnica de estadístico para clasificar los registros en función de los valores de los campos de entrada. Es análoga a la regresión lineal pero utiliza un campo objetivo categórico en lugar de uno numérico.

Por ejemplo, imagine que un proveedor de telecomunicaciones ha segmentado su base de clientes por patrones de uso de servicio, y ha categorizado a los clientes en cuatro grupos. Si los datos demográficos se pueden utilizar para predecir la pertenencia a un grupo, se pueden personalizar las ofertas para cada uno de los posibles clientes.

Este ejemplo utiliza la ruta denominada *telco_custcat_discriminant.str*, que hace referencia al archivo de datos denominado *telco.sav*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *telco_custcat_discriminant.str* está ubicado en el directorio *streams*.

Este ejemplo se centra en la utilización de datos demográficos para pronosticar patrones de uso. El campo objetivo*catpers* tiene cuatro posibles valores que corresponden a los cuatro grupos de clientes:

Valor	Label
1	Servicio básico
2	Servicio electrónico
3	Servicio plus
4	Servicio total

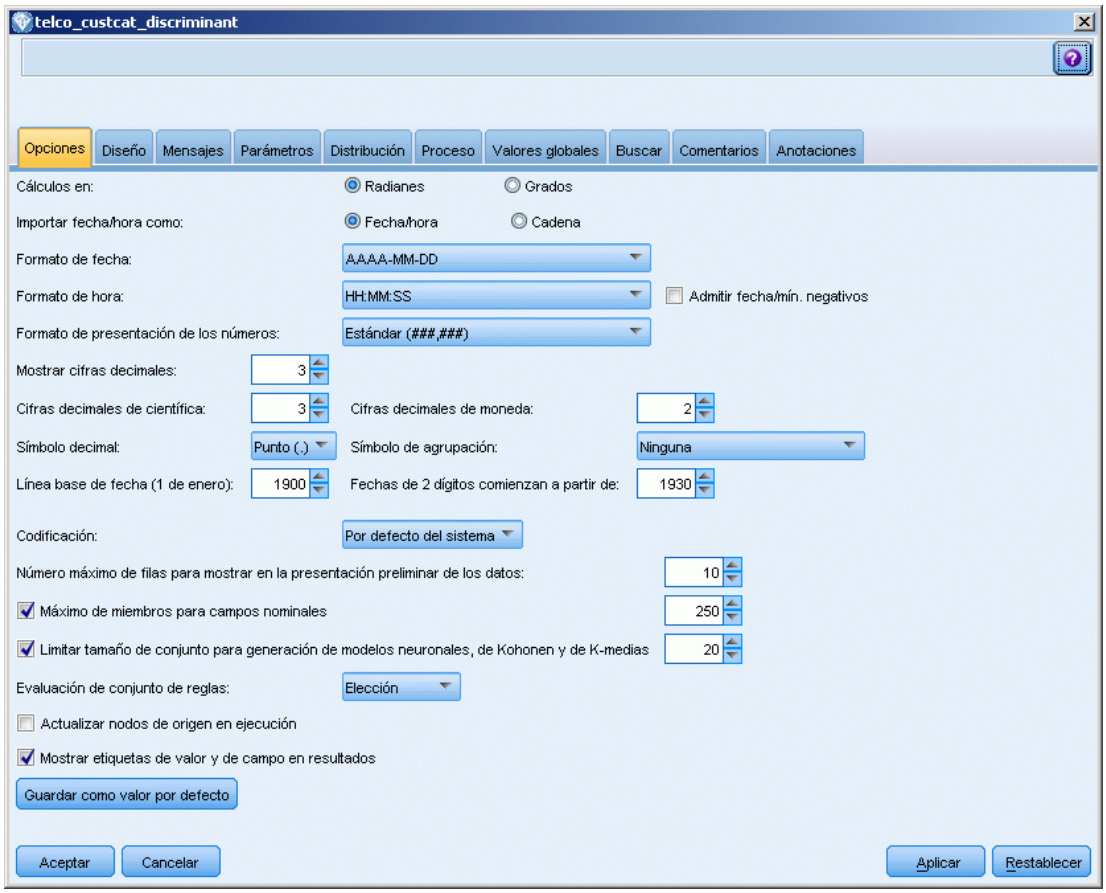
Creación de la ruta

- Primero, configure las propiedades de la ruta para mostrar las etiquetas de valor y de campo en el resultado. Elija en los menús:

  File > Propiedades de ruta...

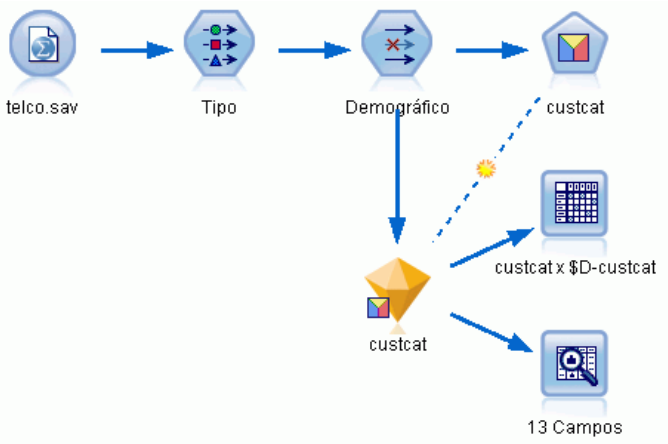
▶ Asegúrese de que se ha seleccionado Mostrar etiquetas de valor y de campo en resultados y haga clic en Aceptar.

Figura 22-1
Propiedades de ruta



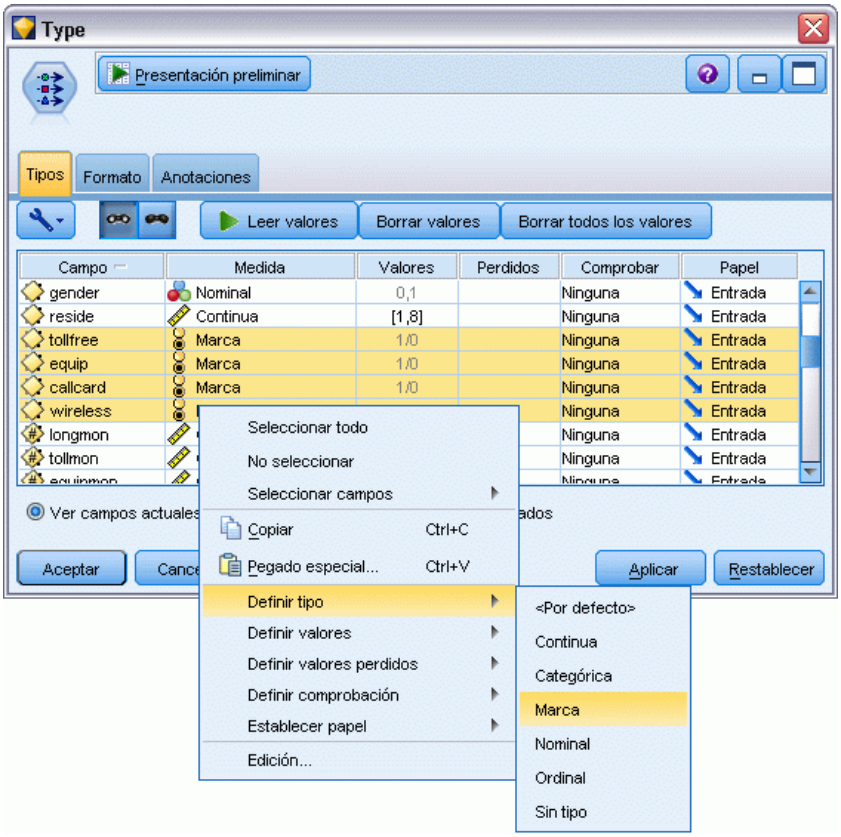
- Añada un nodo de origen Archivo Statistics apuntando a *telco.sav* en la carpeta *Demos*.

Figura 22-2
Ruta de ejemplo para clasificar a los clientes mediante análisis discriminante



- Añada un nodo Tipo y pulse en Leer valores, asegurándose así de que todos los niveles de medición están definidos correctamente. Por ejemplo, la mayoría de valores 0 y 1 se pueden considerar marcas.

Figura 22-3
Definición del nivel de medición para campos múltiples

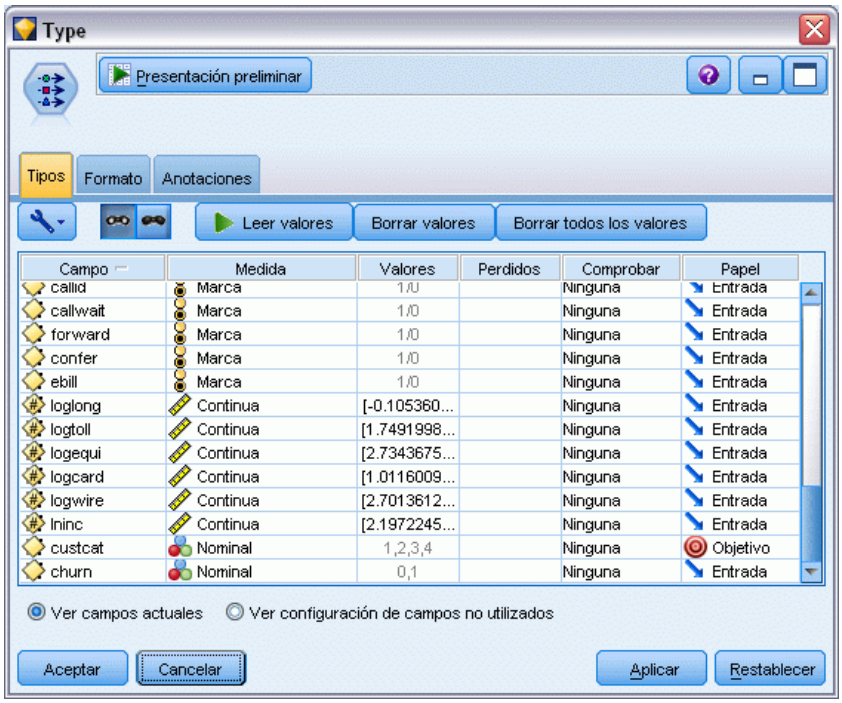


Sugerencia: para cambiar propiedades de varios campos con valores similares (como 0 y 1), pulse en la cabecera de la columna *Valores* para ordenar campos por valor y, a continuación, mantenga pulsada la tecla Mayús mientras utiliza el ratón o las teclas de flecha para seleccionar todos los campos que quiera cambiar. A continuación, puede pulsar con el botón derecho en los elementos seleccionados para cambiar el nivel de medición u otros atributos de los campos seleccionados.

Tenga en cuenta que es más correcto considerar *sexo* como campo con un conjunto de dos valores, en lugar de marca, deje su valor de medición como Nominal.

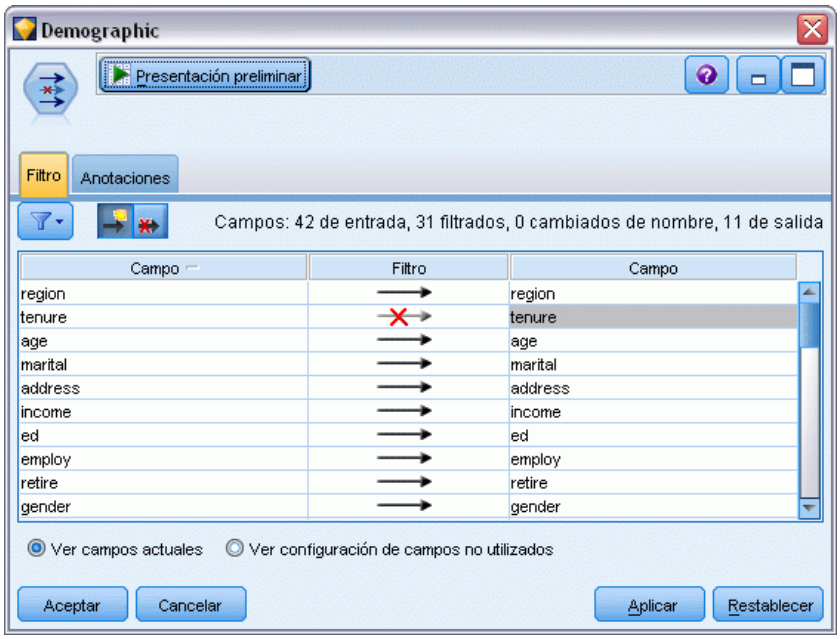
▶ Defina el papel del campo *custcat* a Objetivo. El resto de campos debe tener sus papeles definidas en Entrada.

Figura 22-4
Definición del papel de campos



Puesto que el ejemplo se centra en datos demográficos, utilice un nodo Filtrar para añadir únicamente los campos relevantes (*región*, *edad*, *estado civil*, *dirección*, *ingresos*, *educación*, *empleo*, *jubilación*, *sexo*, *residencia* y *custcat*). Los otros campos se pueden excluir para este análisis.

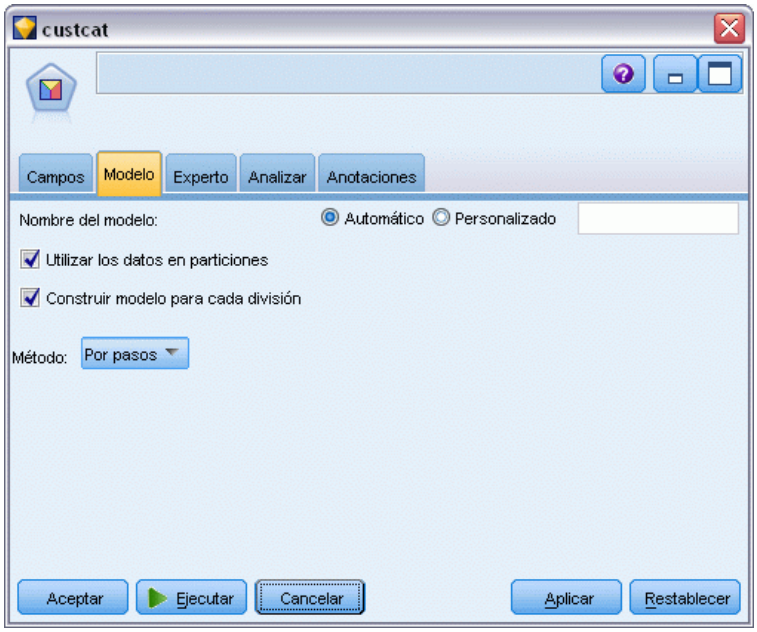
Figura 22-5
Filtrado de los campos demográficos



(Si lo prefiere, puede cambiar el papel de estos campos a Ninguno en lugar de excluirlos, o bien seleccionar los campos que desee utilizar en el nodo de modelado.)

- En el nodo Discriminante, pulse en la pestaña Modelo y seleccione el método Por pasos.

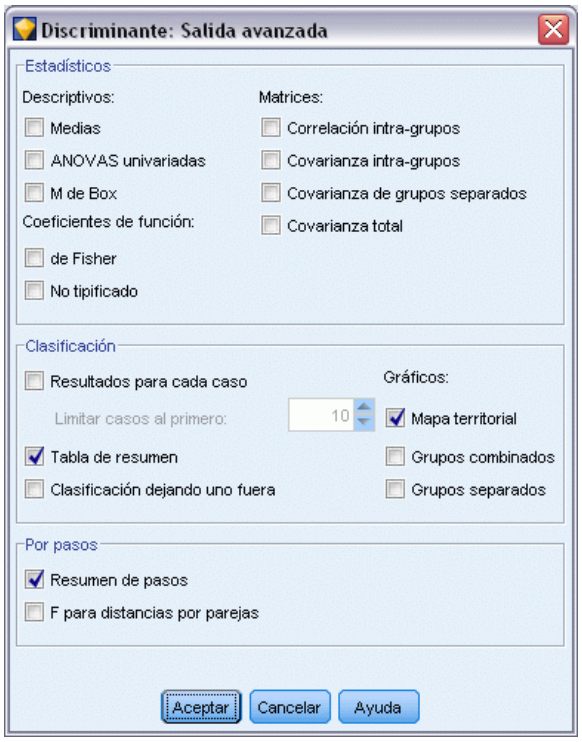
Figura 22-6
Selección de opciones del modelo



- En la pestaña Experto, seleccione el modo Experto y pulse en Resultado.

▶ En el cuadro de diálogo Salida avanzada, seleccione Tabla de resumen, Mapa territorial y Resumen de los pasos y pulse en Aceptar.

Figura 22-7
Selección de opciones de salida

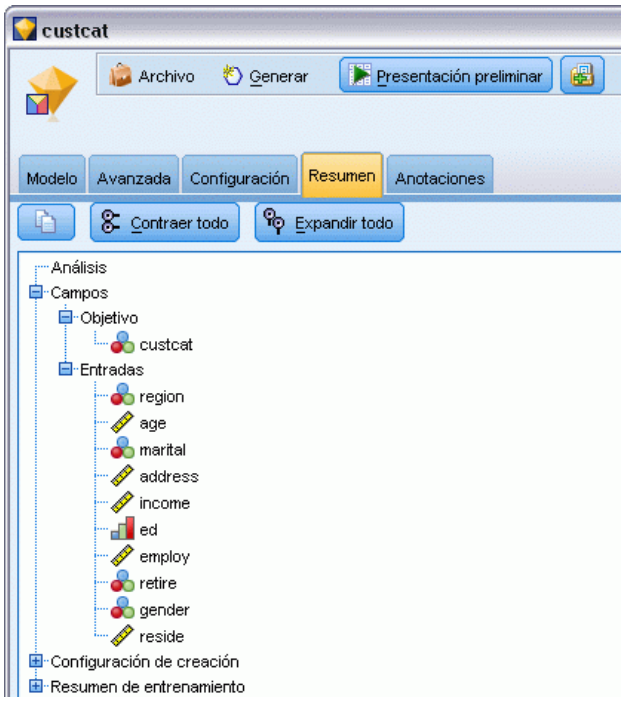


Examen del modelo

▶ Pulse en Ejecutar para crear el modelo que se añadirá a la ruta y a la paleta Modelos en la esquina superior derecha. Para ver los detalles, pulse en el nugget de modelo de la ruta.

La pestaña Resumen muestra (entre otras cosas) el objetivo y la lista completa de entradas (campos predictores) enviadas para consideración.

Figura 22-8
Resumen del modelo en el que se ven los campos Objetivo y Entrada



Para ver más detalles de los resultados del análisis discriminante:

- Pulse en la pestaña Avanzado.
- Pulse en el botón “Abrir en explorador externo” (justo debajo de la pestaña Modelo) para ver los resultados en su explorador Web.

Análisis discriminante por pasos

Figura 22-9
Variables que no aparecen en el análisis, paso 0

Paso		Tolerancia	Tolerancia mín.	F para entrar	Lambda de Wilks
0	Age in years	1,000	1,000	7,521	,978
	Marital status	1,000	1,000	3,500	,990
	Years at current address	1,000	1,000	8,433	,975
	Household income in thousands	1,000	1,000	6,689	,980
	Level of education	1,000	1,000	61,454	,844
	Years with current employer	1,000	1,000	16,976	,951
	Retired	1,000	1,000	3,005	,991
	Gender	1,000	1,000	,373	,999
	Number of people in household	1,000	1,000	3,976	,988

Cuando se tiene un gran número de predictores, el método por pasos puede ser útil al seleccionar automáticamente las "mejores" variables que se utilizarán en el modelo. El método por pasos comienza con un modelo que no incluye ninguno de los predictores. En cada paso, el predictor con el mayor valor *F para entrar* que supera los criterios de entrada (por defecto, 3,84) se añade al modelo.

Figura 22-10
Variables que no aparecen en el análisis, paso 3

Paso		Tolerancia	Tolerancia mín.	F para entrar	Lambda de Wilks
3	Age in years	,535	,535	,252	,795
	Marital status	,605	,593	1,507	,792
	Years at current address	,776	,771	3,514	,787
	Household income in thousands	,688	,657	,687	,794
	Retired	,917	,880	,353	,795
	Gender	,997	,931	,395	,795

Todas las variables que no se han incluido en el análisis tras el último paso tienen valores *F para entrar* inferiores a 3,84, por lo que no se añade ninguna más.

Figura 22-11
Variables en el análisis

Paso		Tolerancia	F para salir	Lambda de Wilks
1	Level of education	1,000	61,454	
2	Level of education	,953	59,108	,951
	Years with current employer	,953	14,933	,844
3	Level of education	,951	60,046	,940
	Years with current employer	,934	15,824	,834
	Number of people in household	,979	4,841	,807

Esta tabla muestra los estadísticos para las variables que se encuentran en el análisis en cada paso. *Tolerancia* es la proporción de su varianza no explicada por las otras variables independientes de la ecuación. Una variable con una tolerancia muy baja contribuye con poca información a un modelo y puede causar problemas de cálculo.

Los valores *F para quitar* son útiles para describir lo que ocurre si una variable se elimina del modelo actual (teniendo en cuenta que otras variables permanecen). *F para quitar* para la variable de entrada es igual que *F para entrar* en el paso anterior (mostrado en las variables no en la tabla de análisis).

Nota de advertencia sobre los métodos por pasos

Los métodos por pasos son cómodos, pero tienen sus limitaciones. No olvide que como los métodos por pasos seleccionan los modelos únicamente según su mérito estadístico, es posible que elijan predictores que no tengan **significado práctico**. Si tiene cierta experiencia con los datos y tiene ciertas expectativas acerca de los predictores que son importantes, deberá utilizar dichos conocimientos y abstenerse de utilizar métodos por pasos. Si, por el contrario, tiene un gran número de predictores y no sabe por dónde empezar, la ejecución de un análisis por pasos y el ajuste del modelo seleccionado es mejor que si no se tiene ningún modelo en absoluto.

Comprobación del ajuste del modelo

Figura 22-12
Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	,198(a)	80,2	80,2	,407
2	,048(a)	19,4	99,6	,214
3	,001(a)	,4	100,0	,031

a. Se han empleado las 3 primeras funciones discriminantes canónicas en el análisis.

Casi toda la varianza explicada por el modelo se debe a las dos primeras funciones discriminantes. Tres funciones se ajustan automáticamente, pero debido a su minúsculo autovalor, la tercera se puede prácticamente ignorar.

Figura 22-13
lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1 a la 3	,796	227,345	9	,000
2 a la 3	,953	47,486	4	,000
3	,999	,929	1	,335

La lambda de Wilks está de acuerdo en que sólo las dos primeras funciones son útiles. Para cada conjunto de funciones, esto comprueba la hipótesis de que las medias de las funciones enumeradas son iguales entre grupos. La comprobación de la función 3 tiene un valor de significación mayor de 0,10, de modo que esta función contribuye poco al modelo.

Matriz de estructura

Figura 22-14
Matriz de estructura

	Función		
	1	2	3
Level of education	,966(*)	-,090	-,244
Years with current employer	-,182	,964(*)	-,193
Age in years(a)	-,162	,598(*)	-,285
Household income in thousands(a)	,109	,514(*)	-,190
Years at current address(a)	-,151	,394(*)	-,214
Retired(a)	-,108	,230(*)	-,137
Gender(a)	,008	,054(*)	,009
Number of people in household	,232	,097	,968(*)
Marital status(a)	,132	,134	,600(*)
Correlaciones intra-grupo combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas Variables ordenadas por el tamaño de la correlación con la función.			
*. Mayor correlación absoluta entre cada variable y cualquier función discriminante.			
a. Esta variable no se emplea en el análisis.			

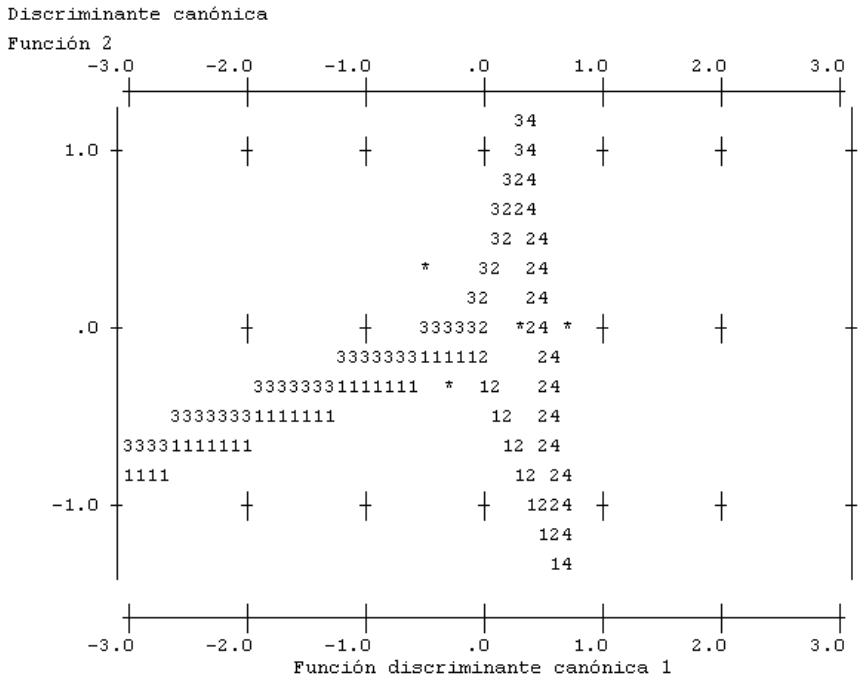
Cuando hay más de una función discriminante, un asterisco (*) marca la mayor correlación absoluta de cada variable con una de las funciones canónicas. Dentro de cada función, estas variables marcadas se ordenan por el tamaño de la correlación.

- *Nivel educativo* está más fuertemente correlacionado con la primera función y es la única variable más fuertemente correlacionada con esta función.

- *Años con empresa actual*, *Edad en años*, *Ingresos del hogar en miles*, *Años en la dirección actual*, *Retirado* y *Sexo* están más fuertemente correlacionados con las segunda función, aunque *Sexo* y *Jubilación* están más débilmente correlacionados que los otros. Las demás variables marcan esta función como función de "estabilidad".
- *Número de personas en el hogar* y *Estado civil* están más fuertemente correlacionados con la tercera función discriminante, pero esta es una función sin utilidad, así que estos predictores son prácticamente inútiles.

Mapa territorial

Figura 22-15
Mapa territorial



El mapa territorial ayuda a estudiar las relaciones entre los grupos y las funciones discriminantes. Combinado con los resultados de la matriz de estructura, ofrece una interpretación gráfica de la relación entre predictores y grupos. La primera función, mostrada en el eje horizontal, separa el grupo 4 (clientes de *servicio total*) de los demás. Ya que *Nivel educativo* está fuertemente correlacionado de forma positiva con la primera función, esto sugiere que los clientes de *Servicio total* son, en general, los más educados. La segunda función separa los grupos 1 y 3 (clientes de *Servicio básico* y de *Servicio plus*). Los clientes del *Servicio plus* tienden a haber trabajado más y a ser mayores que los clientes del *Servicio básico*. Los clientes de *Servicio electrónico* no están bien separados de los demás, aunque el mapa sugiere que tienden a estar bien educados y a tener una moderada experiencia laboral.

En general, la cercanía de los centroides del grupo, marcados con asteriscos (*), a la líneas territoriales sugiere que la separación entre todos los grupos no es muy fuerte.

Sólo las dos primeras funciones discriminantes están representadas, pero ya que la tercera función resultó ser bastante insignificante, el mapa territorial ofrece una vista amplia del modelo discriminante.

Resultados de la clasificación

Figura 22-16
Resultados de clasificación

			Grupo de pertenencia pronosticado				Total
		Customer category	Basic service	E-service	Plus service	Total service	
Original	Recuento	Basic service	125	11	61	69	266
		E-service	49	15	58	95	217
		Plus service	102	14	112	53	281
		Total service	40	16	37	143	236
	%	Basic service	47,0	4,1	22,9	25,9	100,0
		E-service	22,6	6,9	26,7	43,8	100,0
		Plus service	36,3	5,0	39,9	18,9	100,0
		Total service	16,9	6,8	15,7	60,6	100,0
a. Clasificados correctamente el 39,5% de los casos agrupados originales.							

A partir de la lambda de Wilk, se sabe que el modelo está haciendo algo más que adivinar, pero hace falta comprobar los resultados de la clasificación para determinar cuánto más está haciendo. Dados los datos observados, el modelo "nulo" (es decir, el que no incluye ningún predictor) clasificaría a todos los clientes en el grupo modal, *Servicio plus*. Por tanto, el modelo nulo sería correcto 281/1000 = 28,1% de las veces. El modelo consigue un 11,4% más o el 39,5% de los clientes. En concreto, el modelo es particularmente bueno para identificar los clientes de *Servicio total*. Sin embargo, funciona excepcionalmente mal para clasificar los clientes de *Servicio electrónico*. Tal vez necesite encontrar otro predictor para separar estos clientes.

Resumen

Ha creado un modelo que clasifica los clientes en uno de cuatro grupos de "uso de servicio" predefinidos, en función de los datos demográficos de cada cliente. Mediante la matriz de estructura y el mapa territorial, ha identificado las variables más útiles para segmentar la base de clientes. Por último, los resultados de la clasificación muestran que el modelo no clasifica correctamente los clientes de *Servicio electrónico*. Habrá que continuar con el estudio para determinar otra variable predictora que realice una mejor clasificación de estos clientes, pero dependiendo de lo que desee pronosticar, el modelo podrá adecuarse perfectamente a sus necesidades. Por ejemplo, si no está preocupado por identificar a los clientes del *Servicio electrónico* el modelo puede ser suficientemente preciso. Este puede ser el caso cuando el Servicio electrónico es un líder con pérdidas que aporta pocos beneficios. Si, por ejemplo, el mayor retorno

de la inversión proviene de clientes de *Servicio plus* o *Servicio total*, puede que el modelo le dé la información necesaria.

Recuerde que estos resultados están basados sólo en los datos de entrenamiento. Para evaluar qué tal se extiende el modelo a otros datos, se utilizaría un nodo de partición para reservar un subconjunto de registros para comprobación y validación. Si desea obtener más información, consulte el tema Nodo Partición en el capítulo 4 en *Nodos de origen, proceso y resultado de IBM SPSS Modeler 14.2*.

Las explicaciones de los fundamentos matemáticos de los métodos de modelado que se utilizan en IBM® SPSS® Modeler se enumeran en el Manual de algoritmos de SPSS Modeler. Estos archivos están disponibles en el directorio *Documentation* del disco de instalación.

Capítulo
23

Análisis de datos de supervivencia censurados por intervalos (modelos lineales generalizados)

Al analizar datos de supervivencia con censura por intervalos (esto es, cuando no se conoce la hora exacta del evento de interés, sino que sólo se sabe que se ha producido dentro de un intervalo determinado) y aplicar después el modelo de Cox a los impactos de los eventos de los intervalos, se genera un modelo de regresión log-log complementaria.

Hay información parcial de un estudio diseñado para comparar la eficacia de dos terapias de prevención de las úlceras recurrentes recopilada en *ulcer_recurrence.sav*. Este conjunto de datos se ha presentado y analizado en más sitios. Si usa modelos lineales generalizados, puede replicar los resultados de los modelos de regresión log-log complementaria.

Este ejemplo usa la ruta denominada *ulcer_genlin.str*, que hace referencia al archivo de datos *ulcer_recurrence.sav*. El archivo de datos está en la carpeta *Demos* y el archivo de ruta está en la subcarpeta *streams*. Si desea obtener más información, consulte el tema Carpeta Demos en el capítulo 1 en *Manual de usuario de IBM SPSS Modeler 14.2*.

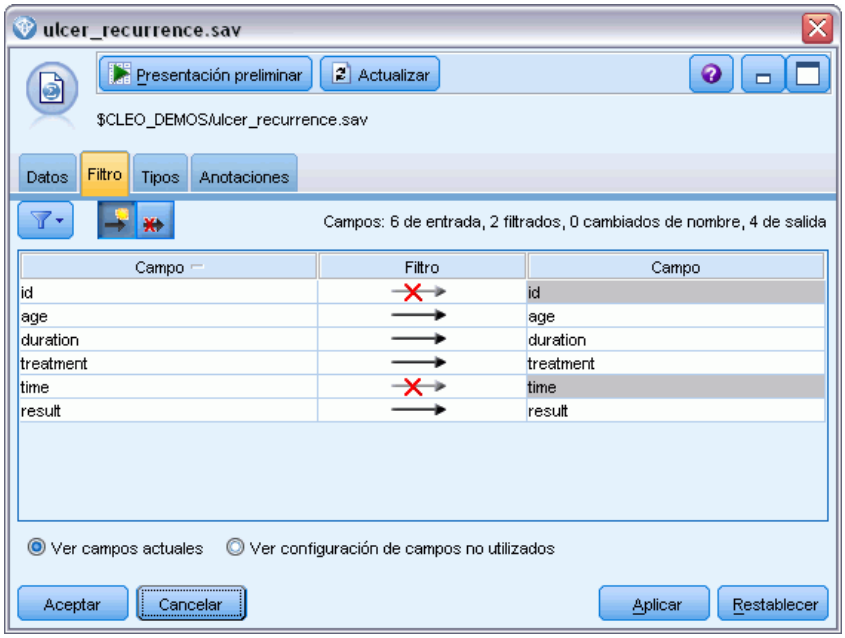
Creación de la ruta

- Añada un nodo de origen Archivo Statistics que apunte a *ulcer_recurrence.sav* en la carpeta *Demos*.

Figura 23-1
Ruta de ejemplo para predecir la recurrencia de las úlceras

▶ En la pestaña Filtro del nodo de origen, filtre *id* y *time*.

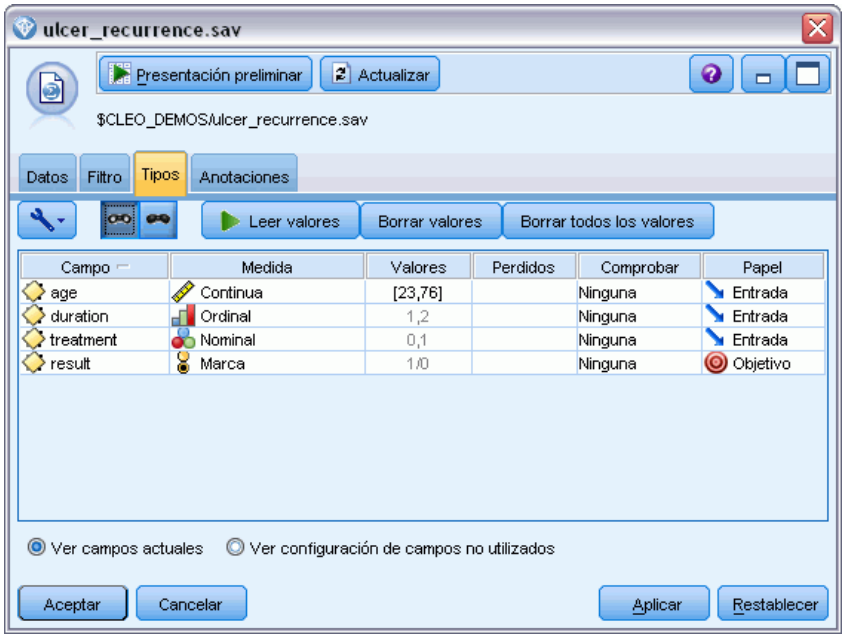
Figura 23-2
Filtrado de campos no deseados



▶ En la pestaña Tipos del nodo de origen, configure el papel del campo *resultado* como Objetivo y defina su nivel de medición como Marca. Un resultado de 1 indica que la úlcera se ha repetido. El resto de campos debe tener sus papeles definidas en Entrada.

▶ Pulse en Leer valores para instanciar los datos.

Figura 23-3
Definición del papel de campos



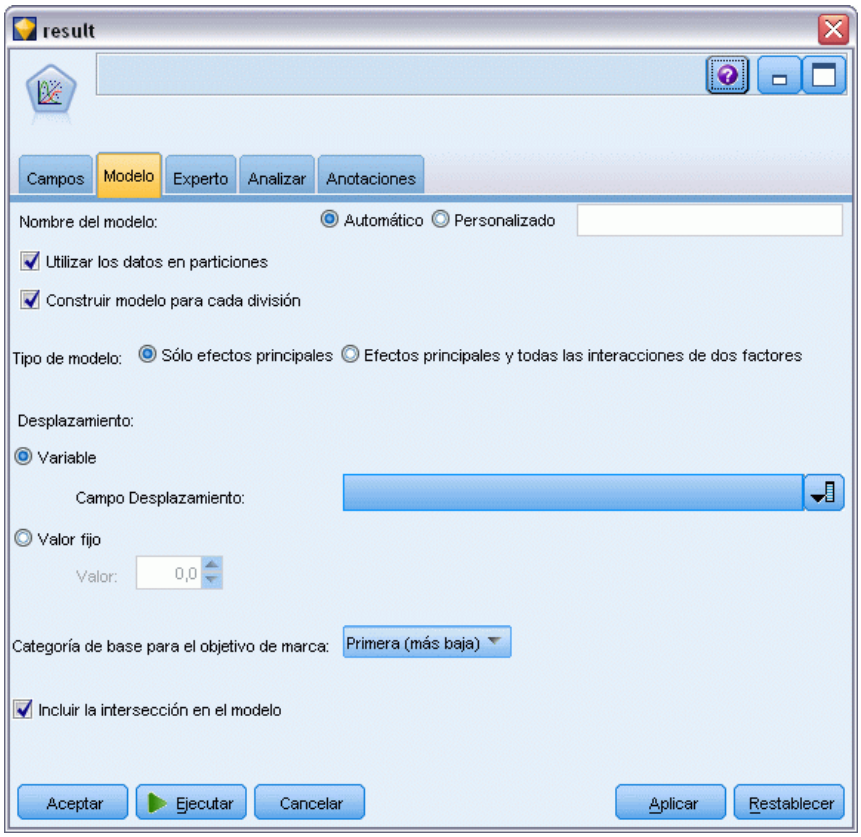
- Añada un nodo Reorg. campos y especifique *duración*, *tratamiento* y *edad* como el orden de las entradas. Esto determinará el orden en el que se introducen los campos en el modelo y le ayudará a replicar los resultados de Collett.

Figura 23-4
Ejemplo de campos reordenados de manera que se introduzcan en el modelo como desee

- Añada un nodo Genlin al nodo de origen; en el nodo Genlin, pulse en la pestaña Campos.
- Seleccione Primera (menor valor) como categoría de referencia para el objetivo. Esto indica que la segunda categoría es el evento de interés, y su efecto en el modelo está en la interpretación de estimaciones de parámetros. Un predictor continuo con coeficiente positivo indica probabilidad aumentada de la recurrencia con valores crecientes del predictor; las categorías de un predictor

nominal con coeficientes mayores indican probabilidad aumentada de la recurrencia con respecto a otras categorías del conjunto.

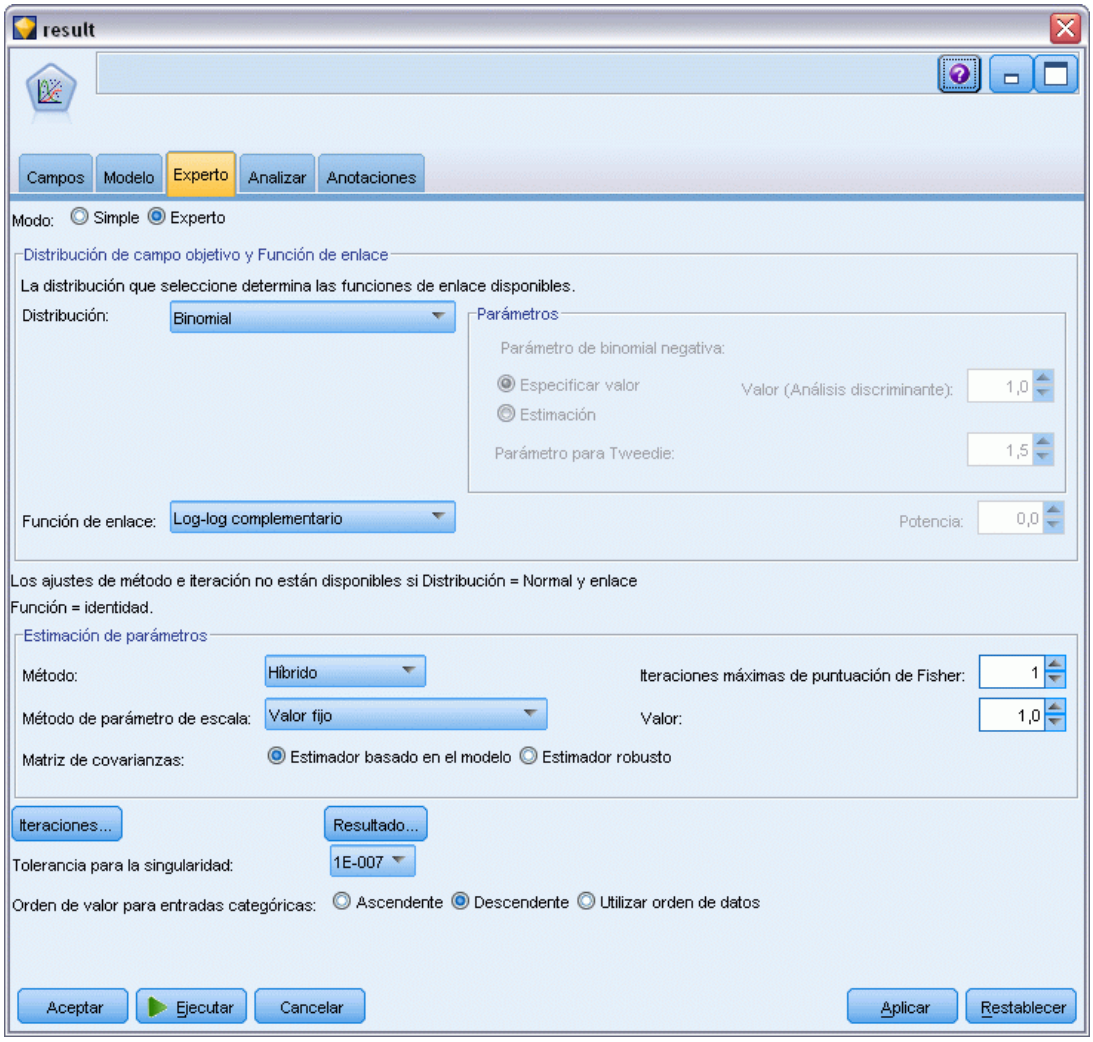
Figura 23-5
Selección de opciones del modelo



- Pulse en la pestaña Experto y seleccione Experto para activar las opciones de modelado experto.
- Seleccione Binomial como distribución y Log-log complementario como función de enlace.
- Seleccione Valor fijo como método de estimación del parámetro de escala y deje el valor por defecto de 1.0.

- Seleccione Descendente como orden de categoría para los factores. Esto indica que la primera categoría de cada factor será su categoría de referencia; el efecto de esta selección en el modelo se aprecia en la interpretación de estimaciones de los parámetros.

Figura 23-6
Selección de opciones de experto



- Ejecute la ruta para crear el nugget de modelo, que se añade al lienzo de rutas y también a la paleta Modelos en la esquina superior derecha. Para ver los detalles de modelo, pulse con el botón derecho en el nugget y seleccione Editar o Examinar.

Pruebas de efectos del modelo

Figura 23-7
Pruebas de los efectos del modelo para el modelo de efectos principales

Origen	Tipo III		
	Chi-cuadrado de Wald	gl	Sig.
(Intersección)	,536	1	,464
duration	,003	1	,958
treatment	,382	1	,537
age	,358	1	,550
Variable dependiente: ResultModelo: (Intersección), duration, treatment, age			

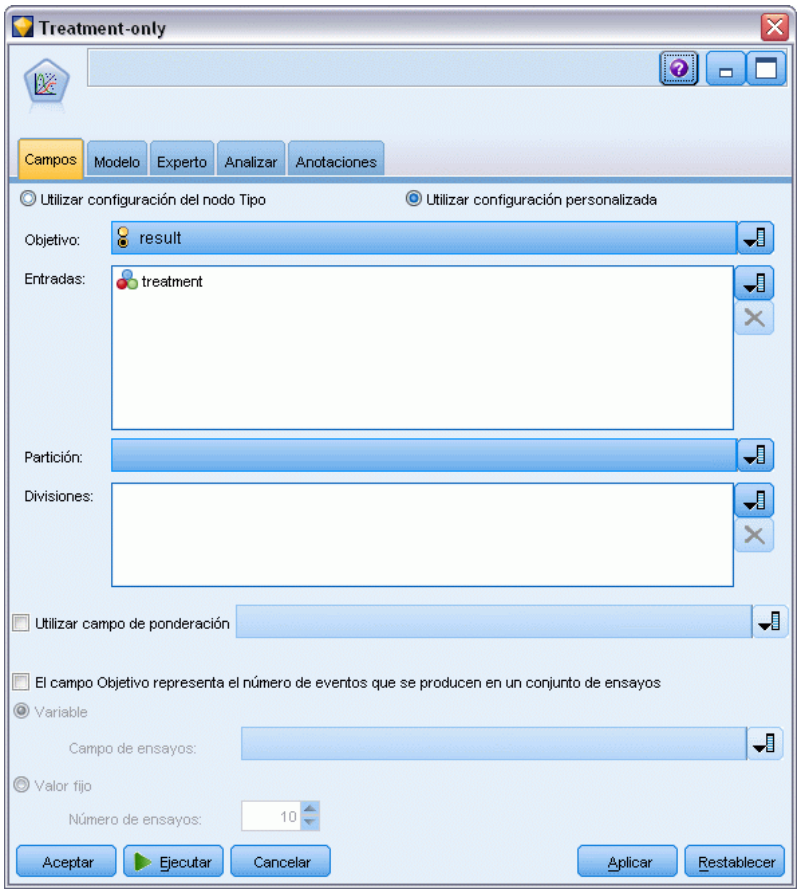
Ningún efecto del modelo es estadísticamente significativo; sin embargo, cualquier diferencia apreciable en los efectos del tratamiento son de interés clínico, por lo que ajustaremos un modelo reducido con el tratamiento exclusivamente como término del modelo.

Ajuste de los modelos exclusivos de tratamiento

- En la pestaña Campos del nodo Genlin, pulse en Utilizar configuración personalizada.
- Seleccione *resultado* como objetivo.

▶ Seleccione *tratamiento* como única entrada.

Figura 23-8
Selección de opciones de campo



▶ Ejecute la ruta y abra el nugget de modelo resultante.

En el nugget de modelo, seleccione la pestaña Avanzado y desplácese hasta la parte inferior.

Estimaciones de los parámetros

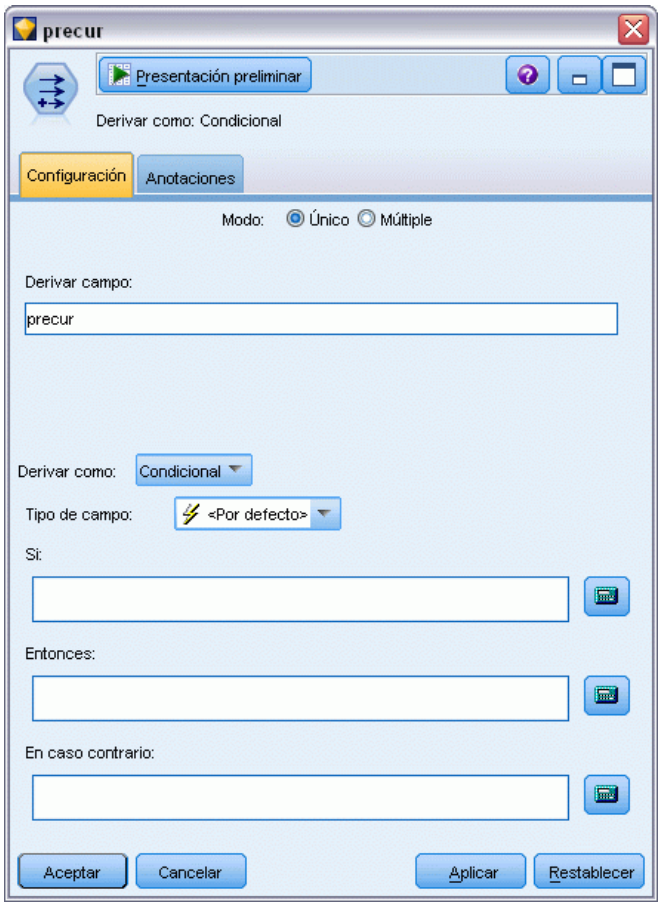
Figura 23-9
Estimaciones de parámetros para modelos exclusivos de tratamiento

Parámetro	B	Típ. Error	Intervalo de confianza de Wald 95%		Contraste de hipótesis		
			Inferior	Superior	Chi-cuadrado de Wald	gl	Sig.
(Intersección)	-1,442	,5012	-2,425	-,460	8,282	1	,004
[treatment=1]	,378	,6288	-,855	1,610	,361	1	,548
[treatment=0]	0(a)	.	.	.	.	.	.
(Escala)	1(b)						
Variable dependiente: ResultModelo: (Intersección), treatment, offset = 0							
a. Establecido en cero ya que este parámetro es redundante.							
b. Fijado en el valor mostrado.							

El efecto del tratamiento (diferencia del predictor lineal entre los dos niveles del tratamiento; esto es, el coeficiente para *[tratamiento=1]*) no es estadísticamente significativo, sino que sólo sugiere que el tratamiento *A [tratamiento=0]* puede ser mejor que el *B [tratamiento=1]* porque la estimación del parámetro para el tratamiento *B* es mayor que para la del *A* y, por tanto, está asociada a una probabilidad aumentada de la recurrencia en los 12 primeros meses. El predictor lineal, (intersección + efecto del tratamiento) es una estimación del logaritmo($-\log(1-P(recur_{12,t})$), donde $P(recur_{12,t})$ es la probabilidad de la recurrencia en los 12 meses de tratamiento t(=*A* o *B*). Se generan estas probabilidades pronosticadas para cada observación del conjunto de datos.

Probabilidades de recurrencia pronosticada y supervivencia

Figura 23-10
Opciones de configuración del nodo Derivar



- Para cada paciente, el modelo puntúa el resultado pronosticado y la probabilidad de dicho resultado. Para poder ver las probabilidades de la recurrencia pronosticada, copie el modelo generado en la paleta y añada un nodo Derivar.
- En la pestaña Configuración, introduzca precur como el campo de derivación.
- Seleccione la derivación como Condicional.
- Pulse en el botón de calculadora para abrir el generador de expresiones de la condición Si.

Figura 23-11
Nodo Derivar: Generador de expresiones de la condición Si

- Introduzca el campo *$G-result* en la expresión.
- Pulse en Aceptar.

El campo de derivación *precur* tomará el valor de la expresión Entonces si *$G-result* es igual a 1 y el valor de la expresión En caso contrario cuando sea igual a 0.

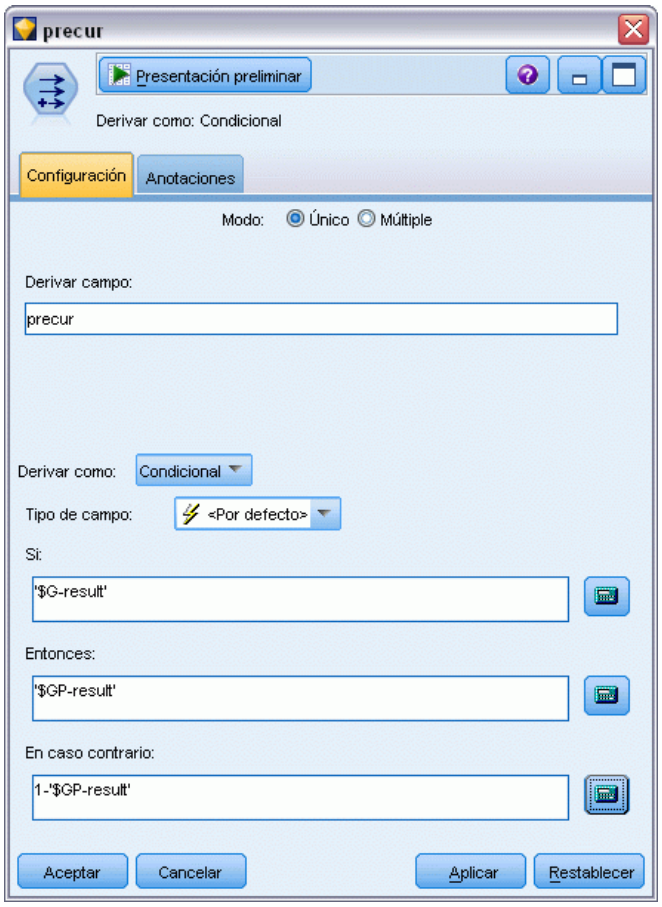
Figura 23-12
Nodo Derivar: Generador de expresiones de la expresión Entonces

- Pulse en el botón de calculadora para abrir el generador de expresiones de la expresión Entonces.
- Introduzca el campo *$GP-result* en la expresión.
- Pulse en Aceptar.

Figura 23-13
Nodo Derivar: Generador de expresiones de la expresión En caso contrario

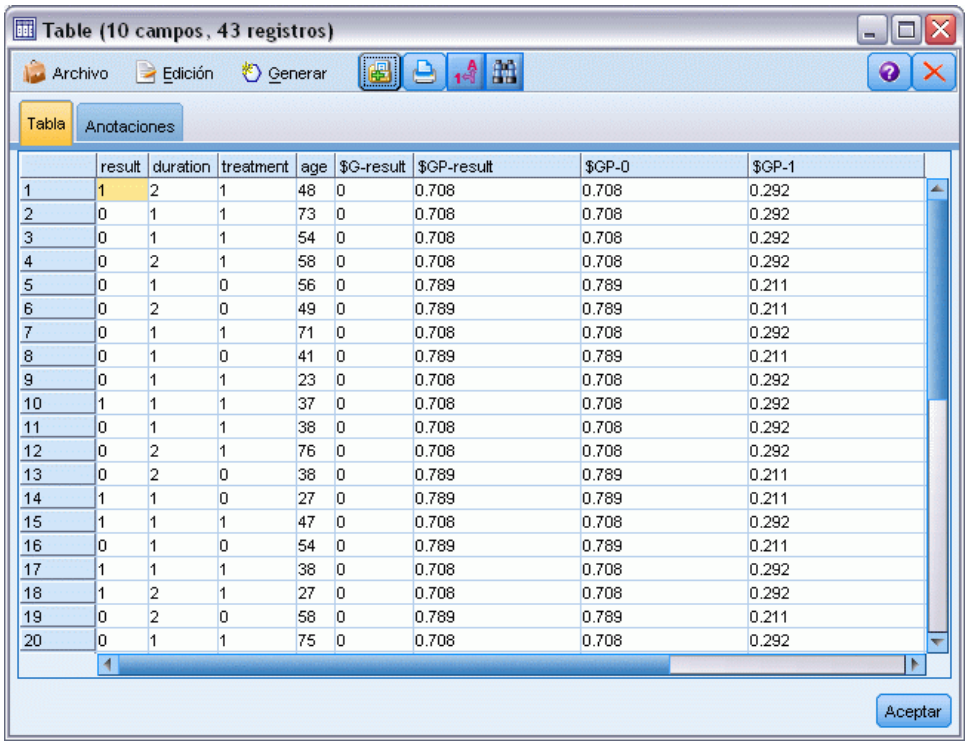
- Pulse en el botón de calculadora para abrir el generador de expresiones de la expresión En caso contrario.
- Introduzca 1- en la expresión e introduzca el campo *$GP-result* en la expresión.
- Pulse en Aceptar.

Figura 23-14
Opciones de configuración del nodo Derivar



► Añada un nodo de tabla al nodo Derivar y ejecute la ruta.

Figura 23-15
Probabilidades pronosticadas



	result	duration	treatment	age	$G-result	$GP-result	$GP-0	$GP-1
1	1	2	1	48	0	0.708	0.708	0.292
2	0	1	1	73	0	0.708	0.708	0.292
3	0	1	1	54	0	0.708	0.708	0.292
4	0	2	1	58	0	0.708	0.708	0.292
5	0	1	0	56	0	0.789	0.789	0.211
6	0	2	0	49	0	0.789	0.789	0.211
7	0	1	1	71	0	0.708	0.708	0.292
8	0	1	0	41	0	0.789	0.789	0.211
9	0	1	1	23	0	0.708	0.708	0.292
10	1	1	1	37	0	0.708	0.708	0.292
11	0	1	1	38	0	0.708	0.708	0.292
12	0	2	1	76	0	0.708	0.708	0.292
13	0	2	0	38	0	0.789	0.789	0.211
14	1	1	0	27	0	0.789	0.789	0.211
15	1	1	1	47	0	0.708	0.708	0.292
16	0	1	0	54	0	0.789	0.789	0.211
17	1	1	1	38	0	0.708	0.708	0.292
18	1	2	1	27	0	0.708	0.708	0.292
19	0	2	0	58	0	0.789	0.789	0.211
20	0	1	1	75	0	0.708	0.708	0.292

Hay una probabilidad estimada de 0,211 de que los pacientes a los que se ha asignado el tratamiento *A* experimenten una recurrencia en los 12 primeros meses; y de 0,292 para el tratamiento *B*. Tenga en cuenta que $1-P(recur_{12,\,t})$ es la probabilidad de supervivencia en los 12 meses, lo que puede resultar más interesante para los analistas de supervivencia.

Modelado de probabilidades de recurrencia por período

Un problema que presenta el modelo tal y como está es que ignora la información recopilada en el primer examen; es decir, muchos pacientes no experimentaron una recurrencia en los seis primeros meses. Un modelo "mejor" modelaría una respuesta binaria que registraría si se produjo o no el evento durante cada intervalo. El ajuste de este modelo exige una reconstrucción del conjunto de datos original, que se puede encontrar en *ulcer_recurrence_recoded.sav*. Si desea obtener más información, consulte el tema Carpeta Demos en el capítulo 1 en *Manual de usuario de IBM SPSS Modeler 14.2*. Este archivo incluye otras dos variables:

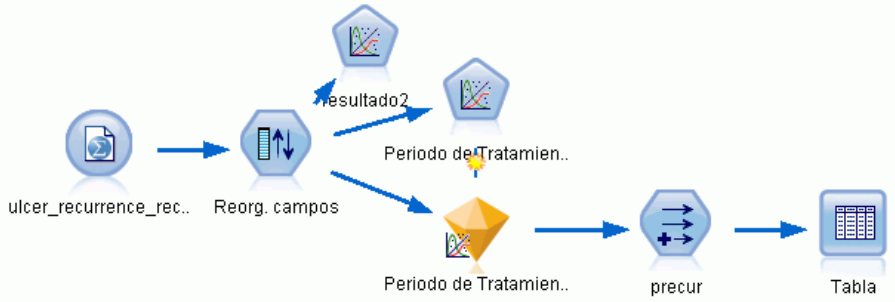
- *Periodo*, que registra si el caso se corresponde con el primer o el segundo período de examen.
- *Resultado por periodo*, que registra si se produjo una recurrencia en un paciente determinado durante un período concreto.

Cada caso original (paciente) aporta un caso por intervalo en el que permanece en el conjunto de riesgos. Así, por ejemplo, el paciente 1 aporta dos casos: uno para el primer período de examen, en el que no se produjo ninguna recurrencia, y otro para el segundo período de examen, en el que

se registró una recurrencia. Por otro lado, el paciente 10 aporta un único caso, ya que se registró una recurrencia en el primer período. Los pacientes 16, 28 y 34 se eliminaron del estudio después de seis meses y, por tanto, sólo aportan un único caso al nuevo conjunto de datos.

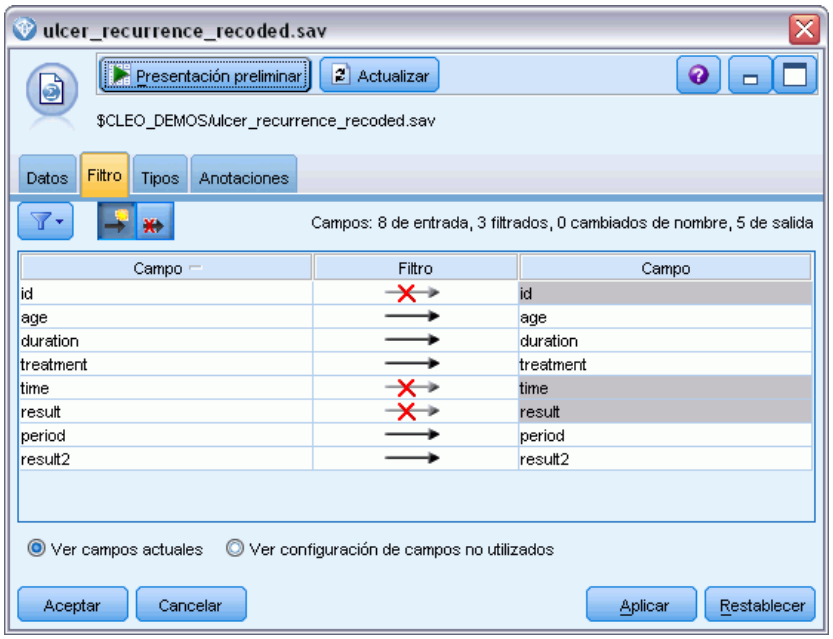
- Añada un nodo de origen Archivo Statistics que apunte a *ulcer_recurrence_recoded.sav* en la carpeta *Demos*.

Figura 23-16
Ruta de ejemplo para predecir la recurrencia de las úlceras



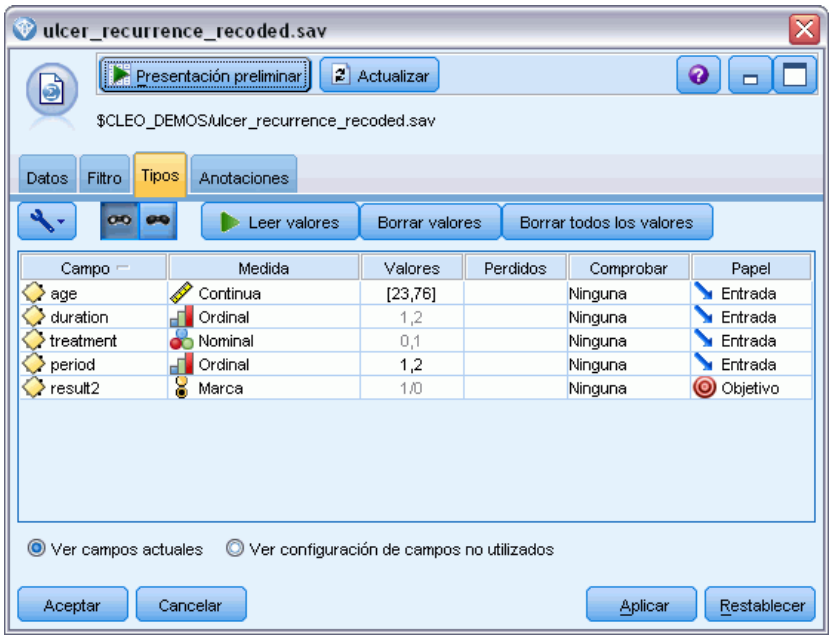
- En la pestaña Filtro del nodo de origen, filtre *id* y *hora* y *resultado*.

Figura 23-17
Filtrado de campos no deseados



- En la pestaña Tipos del nodo de origen, configure el papel del campo *result2* como Objetivo y defina su nivel de medición como Marca. El resto de campos debe tener sus papeles definidas en Entrada.

Figura 23-18
Definición del papel de campos



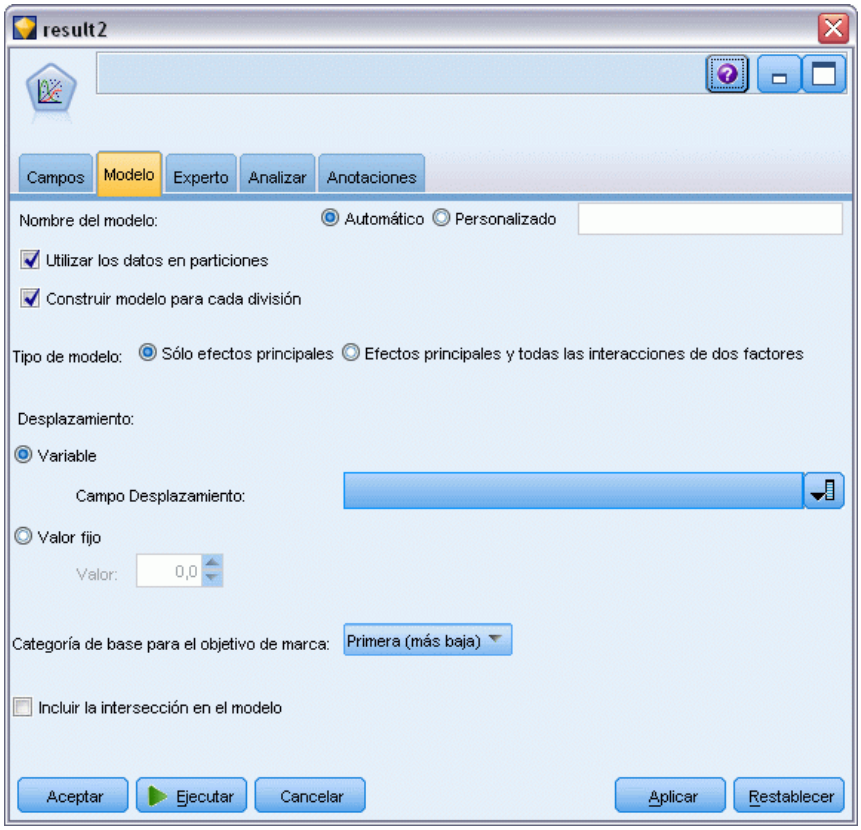
- Añada un nodo Reorg. campos y especifique *periodo*, *duración*, *tratamiento* y *edad* como el orden de las entradas. Si *periodo* se coloca como primera entrada (y no se incluye el término de

intersección en el modelo), podrá ajustar un conjunto completo de variables dummy para capturar los efectos del período.

Figura 23-19
Ejemplo de campos reordenados de manera que se introduzcan en el modelo como desee

- En el nodo GenLin, pulse en la pestaña Modelo.

Figura 23-20
Selección de opciones del modelo

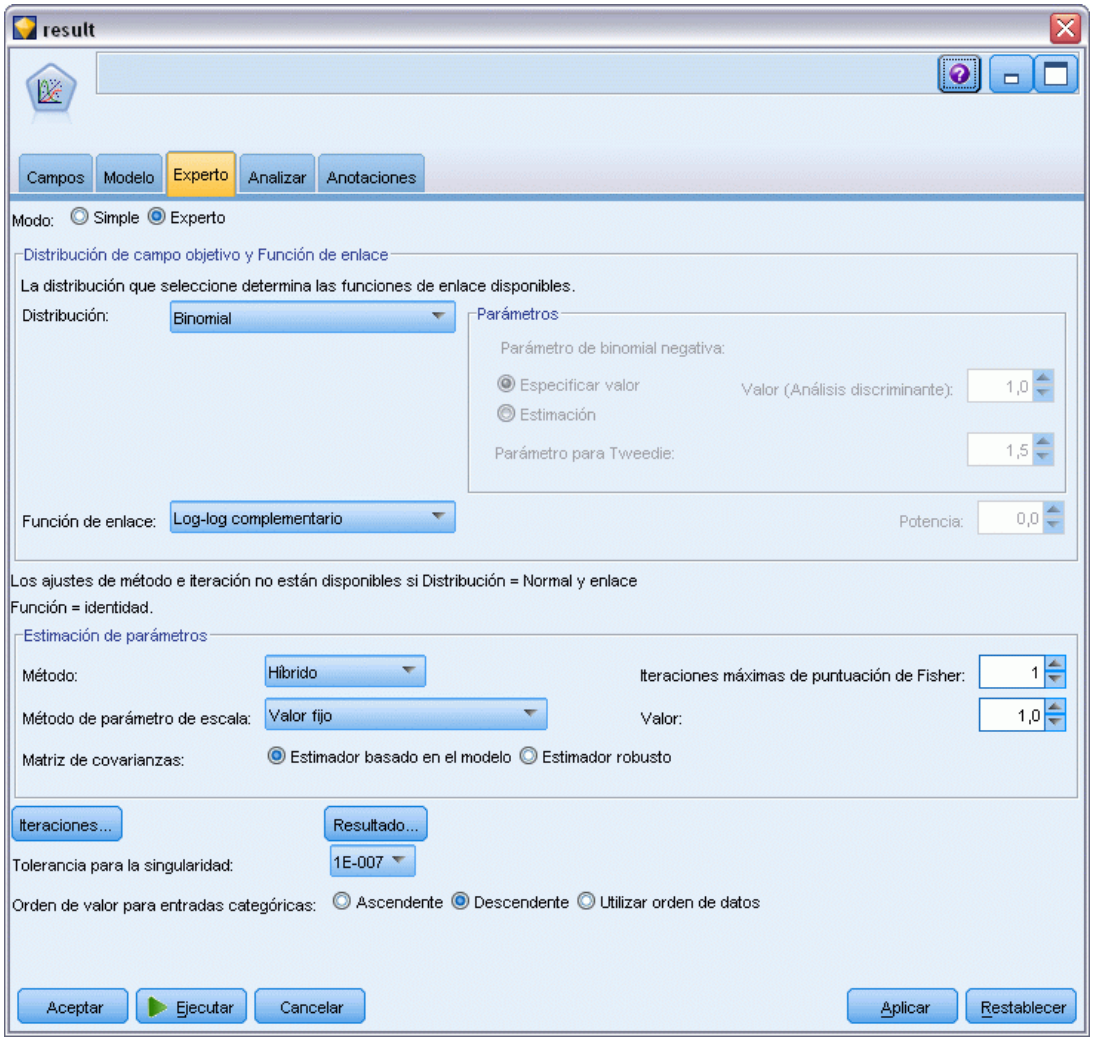


- Seleccione Primera (menor valor) como categoría de referencia para el objetivo. Esto indica que la segunda categoría es el evento de interés, y su efecto en el modelo está en la interpretación de estimaciones de parámetros.

- Desactive la casilla de verificación Incluir la intersección en el modelo.

- Pulse en la pestaña Experto y seleccione Experto para activar las opciones de modelado experto.

Figura 23-21
Selección de opciones de experto



- Seleccione Binomial como distribución y Log-log complementario como función de enlace.
- Seleccione Valor fijo como método de estimación del parámetro de escala y deje el valor por defecto de 1.0.
- Seleccione Descendente como orden de categoría para los factores. Esto indica que la primera categoría de cada factor será su categoría de referencia; el efecto de esta selección en el modelo se aprecia en la interpretación de estimaciones de los parámetros.
- Ejecute la ruta para crear el nugget de modelo, que se añade al lienzo de rutas y también a la paleta Modelos en la esquina superior derecha. Para ver los detalles de modelo, pulse con el botón derecho en el nugget y seleccione Editar o Examinar.

Pruebas de efectos del modelo

Figura 23-22
Pruebas de los efectos del modelo para el modelo de efectos principales

Origen	Tipo III		
	Chi-cuadrado de Wald	gl	Sig.
period	,464	1	,496
duration	,000	1	,988
treatment	,117	1	,732
age	,314	1	,575
Variable dependiente: Result by periodModelo: period, duration, treatment, age			

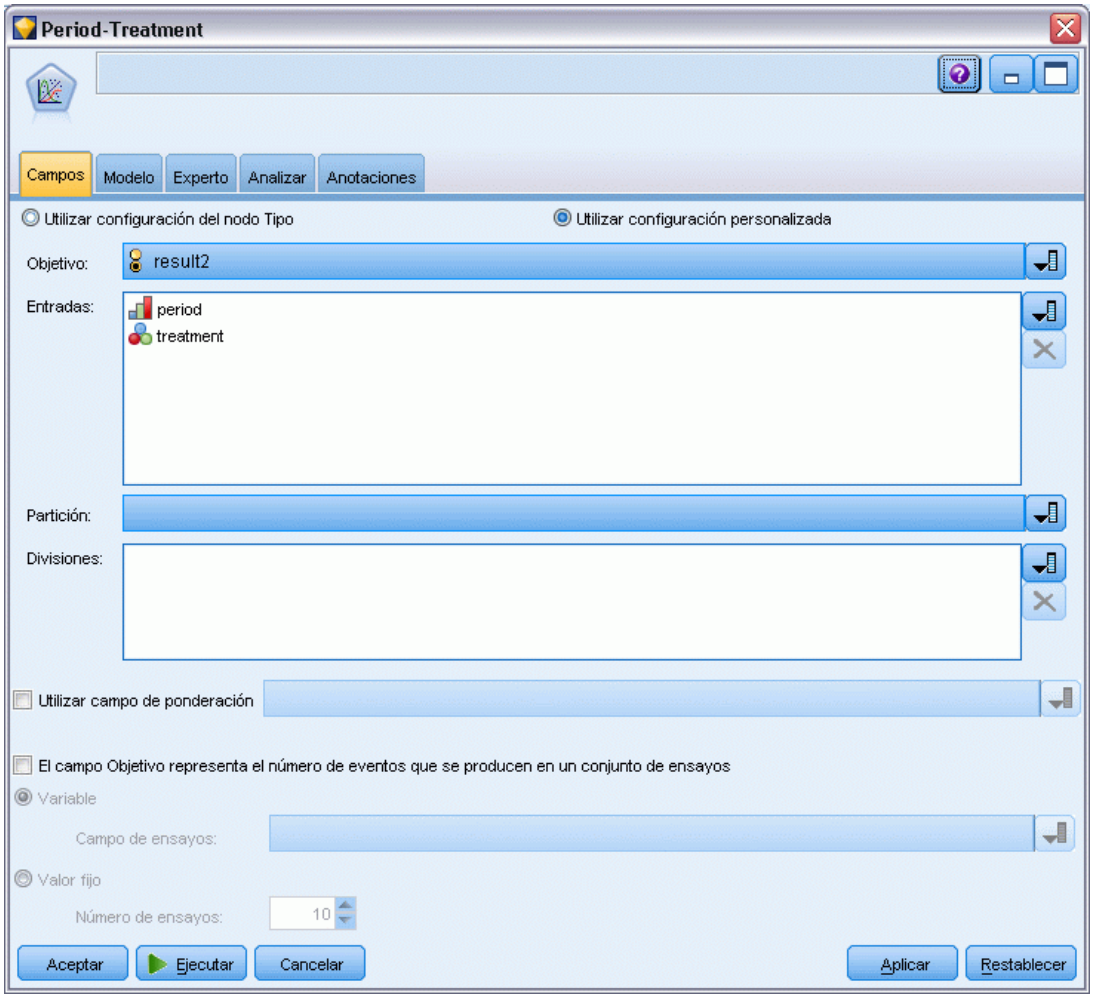
Ningún efecto del modelo es estadísticamente significativo; sin embargo, cualquier diferencia apreciable en los efectos del período y el tratamiento son de interés clínico, por lo que ajustaremos un modelo reducido sólo con esos términos del modelo.

Ajuste de modelos reducidos

- En la pestaña Campos del nodo Genlin, pulse en Utilizar configuración personalizada.
- Seleccione *result2* como objetivo.

- Seleccione *periodo* y *tratamiento* como entradas.

Figura 23-23
Selección de opciones de campo



- Ejecute el nodo, examine el modelo generado y, a continuación, copie dicho modelo en la paleta, añada un nodo de tabla y ejecútelo.

Estimaciones de los parámetros

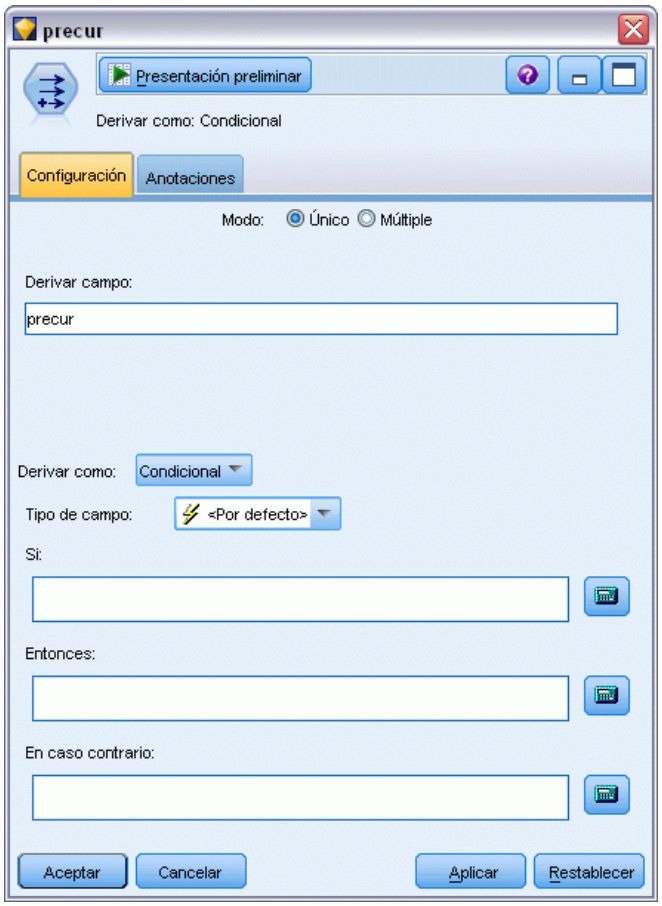
Figura 23-24
Estimaciones de parámetros para modelos exclusivos de tratamiento

Parámetro	B	Típ. Error	Intervalo de confianza de Wald 95%		Contraste de hipótesis		
			Inferior	Superior	Chi-cuadrado de Wald	gl	Sig.
[period=2]	-1,794	,5792	-2,929	-,659	9,597	1	,002
[period=1]	-2,206	,5912	-3,365	-1,047	13,926	1	,000
[treatment=1]	,195	,6279	-1,035	1,426	,097	1	,756
[treatment=0]	0(a)	.	.	.	.	.	.
(Escala)	1(b)						
Variable dependiente: Result by periodModelo: period, treatment							
a. Establecido en cero ya que este parámetro es redundante.							
b. Fijado en el valor mostrado.							

El efecto del tratamiento no es estadísticamente significativo, sino que sólo sugiere que el tratamiento *A* puede ser mejor que el *B* porque la estimación del parámetro para el tratamiento *B* está asociada a una probabilidad aumentada de la recurrencia en los 12 primeros meses. Los valores del período tienen una diferencia de 0 estadísticamente significativa, pero esto se debe a que existe un término de intersección que no se ha ajustado. El efecto del período (diferencia entre los valores del predictor lineal para *[periodo=1]* y *[periodo=2]*) no es estadísticamente significativo, como se puede comprobar en las pruebas de los efectos del modelo. El predictor lineal (efecto del período + efecto del tratamiento) es una estimación del logaritmo($-\log(1-P(recur_{p,\,t})$), donde $P(recur_{p,\,t})$ es la probabilidad de la recurrencia en el período p(=1 ó 2, que representa a 6 meses o 12 meses) dado el tratamiento t(=*A* o *B*). Se generan estas probabilidades pronosticadas para cada observación del conjunto de datos.

Probabilidades de recurrencia pronosticada y supervivencia

Figura 23-25
Opciones de configuración del nodo Derivar



- Para cada paciente, el modelo puntúa el resultado pronosticado y la probabilidad de dicho resultado. Para poder ver las probabilidades de la recurrencia pronosticada, copie el modelo generado en la paleta y añada un nodo Derivar.
- En la pestaña Configuración, introduzca precur como el campo de derivación.
- Seleccione la derivación como Condicional.
- Pulse en el botón de calculadora para abrir el generador de expresiones de la condición Si.

Figura 23-26
Nodo Derivar: Generador de expresiones de la condición Si

- Introduzca el campo *$G-result2* en la expresión.
- Pulse en Aceptar.

El campo de derivación *precur* tomará el valor de la expresión Entonces si *$G-result2* es igual a 1 y el valor de la expresión En caso contrario cuando sea igual a 0.

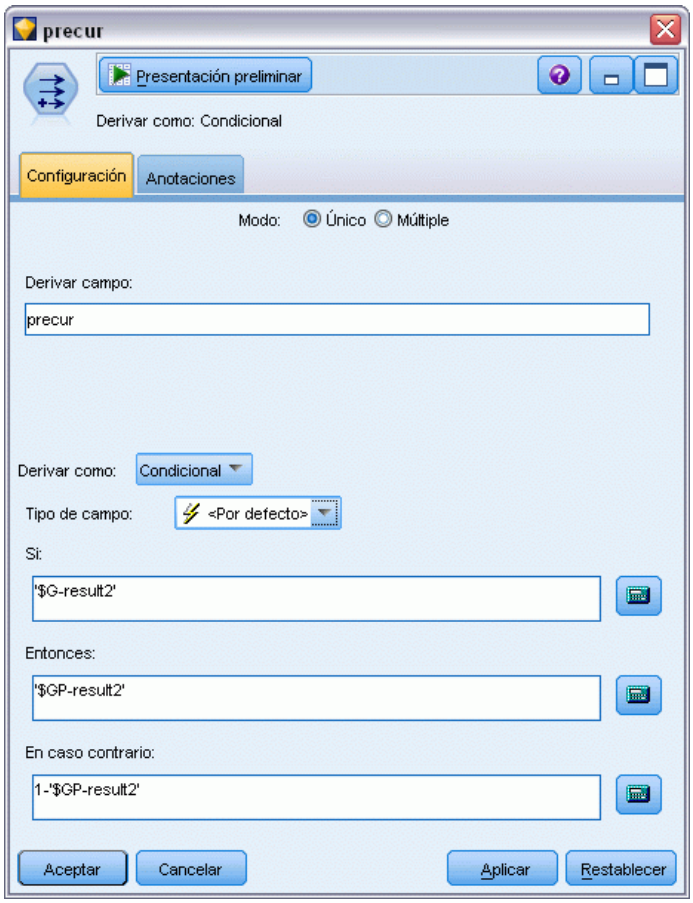
Figura 23-27
Nodo Derivar: Generador de expresiones de la expresión Entonces

- Pulse en el botón de calculadora para abrir el generador de expresiones de la expresión Entonces.
- Introduzca el campo *$GP-result2* en la expresión.
- Pulse en Aceptar.

Figura 23-28
Nodo Derivar: Generador de expresiones de la expresión En caso contrario

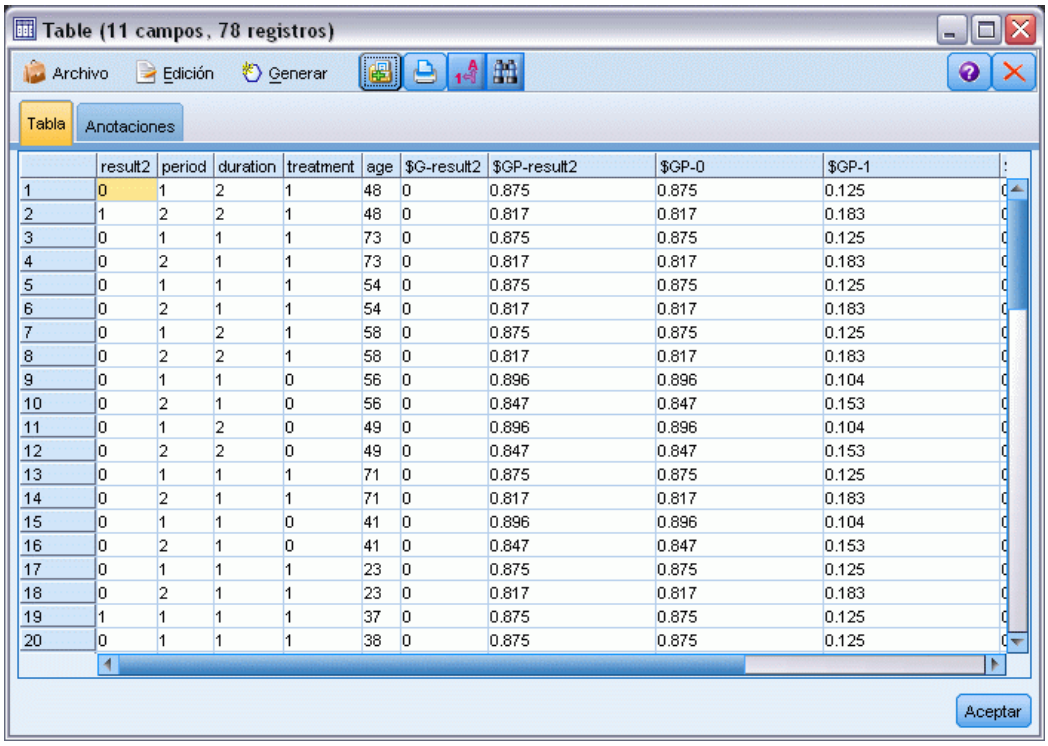
- Pulse en el botón de calculadora para abrir el generador de expresiones de la expresión En caso contrario.
- Introduzca 1- en la expresión e introduzca el campo *$GP-result2* en la expresión.
- Pulse en Aceptar.

Figura 23-29
Opciones de configuración del nodo Derivar



▶ Añada un nodo de tabla al nodo Derivar y ejecute la ruta.

Figura 23-30
Probabilidades pronosticadas



	result2	period	duration	treatment	age	$G-result2	$GP-result2	$GP-0	$GP-1
1	0	1	2	1	48	0	0.875	0.875	0.125
2	1	2	2	1	48	0	0.817	0.817	0.183
3	0	1	1	1	73	0	0.875	0.875	0.125
4	0	2	1	1	73	0	0.817	0.817	0.183
5	0	1	1	1	54	0	0.875	0.875	0.125
6	0	2	1	1	54	0	0.817	0.817	0.183
7	0	1	2	1	58	0	0.875	0.875	0.125
8	0	2	2	1	58	0	0.817	0.817	0.183
9	0	1	1	0	56	0	0.896	0.896	0.104
10	0	2	1	0	56	0	0.847	0.847	0.153
11	0	1	2	0	49	0	0.896	0.896	0.104
12	0	2	2	0	49	0	0.847	0.847	0.153
13	0	1	1	1	71	0	0.875	0.875	0.125
14	0	2	1	1	71	0	0.817	0.817	0.183
15	0	1	1	0	41	0	0.896	0.896	0.104
16	0	2	1	0	41	0	0.847	0.847	0.153
17	0	1	1	1	23	0	0.875	0.875	0.125
18	0	2	1	1	23	0	0.817	0.817	0.183
19	1	1	1	1	37	0	0.875	0.875	0.125
20	0	1	1	1	38	0	0.875	0.875	0.125

Las probabilidades de recurrencia estimada se pueden resumir de la siguiente manera:

Tratamiento	**6 meses**	**12 meses**
A	0.104	0.153
B	0.125	0.183

A partir de estos datos, la probabilidad de supervivencia a lo largo de 12 meses se puede estimar como $1-(P(recur_{1,t}) + P(recur_{2,t})\times(1-P(recur_{1,t})))$; por tanto, para cada tratamiento:

A: 1 − (0.104 + 0.153*0.896) = 0.759

B: 1 − (0.125 + 0.183*0.875) = 0.715

lo que vuelve a demostrar un apoyo sin relevancia estadística para *A* como mejor tratamiento.

Resumen

Ha ajustado una serie de modelos de regresión log-log complementaria para datos de supervivencia censurados por intervalos con modelos lineales generalizados. Aunque existen datos que avalan la elección del tratamiento *A*, puede que sea necesario emprender un estudio exhaustivo para conseguir un resultado estadísticamente significativo. Sin embargo, existen otros métodos de exploración con los datos existentes.

- Puede que valga la pena reajustar el modelo con los efectos de interacción, en especial los incluidos entre *Periodo* y *Grupo de tratamiento*.

Las explicaciones de los fundamentos matemáticos de los métodos de modelado que se utilizan en IBM® SPSS® Modeler se enumeran en el *Manual de algoritmos de SPSS Modeler*.

Capítulo
24

Uso de la regresión de Poisson para analizar las tasas de daños sufridos por barcos (modelos lineales generalizados)

Se puede usar un modelo lineal generalizado para ajustar una regresión de Poisson para el análisis de datos de frecuencias. Por ejemplo, un conjunto de datos presentados y analizados en otro sitio se refiere al daño que causan las olas a los cargueros. Se pueden modelar los recuentos de incidentes con una tasa de Poisson a partir de los valores de los predictores, y el modelo resultante puede ayudarle a determinar los tipos de barco que son más propensos a sufrir daños.

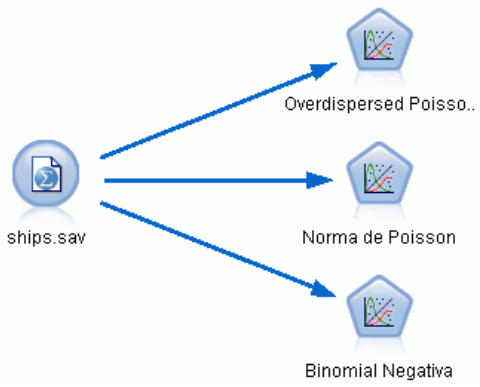
Este ejemplo usa la ruta *ships_genlin.str*, que hace referencia al archivo de datos *ships.sav*. El archivo de datos está en la carpeta *Demos* y el archivo de ruta está en la subcarpeta *streams*. Si desea obtener más información, consulte el tema Carpeta Demos en el capítulo 1 en *Manual de usuario de IBM SPSS Modeler 14.2*.

El modelado de recuentos de casillas brutos puede ser engañoso en este caso, ya que la variable *Meses de servicio agregados* varía según el tipo de barco. Las variables de este tipo, que miden la cantidad de "exposición" a riesgos, se tratan dentro del modelo lineal generalizado como variables de desplazamiento. Además, una regresión de Poisson supone que el logaritmo de la variable dependiente es lineal en los predictores. De esta forma, tendrá que usar *Logaritmo de meses de servicio agregados* para utilizar modelos lineales generalizados para ajustar una regresión de Poisson a las tasas de accidentes.

Ajuste de una regresión de Poisson "sobredispersada"

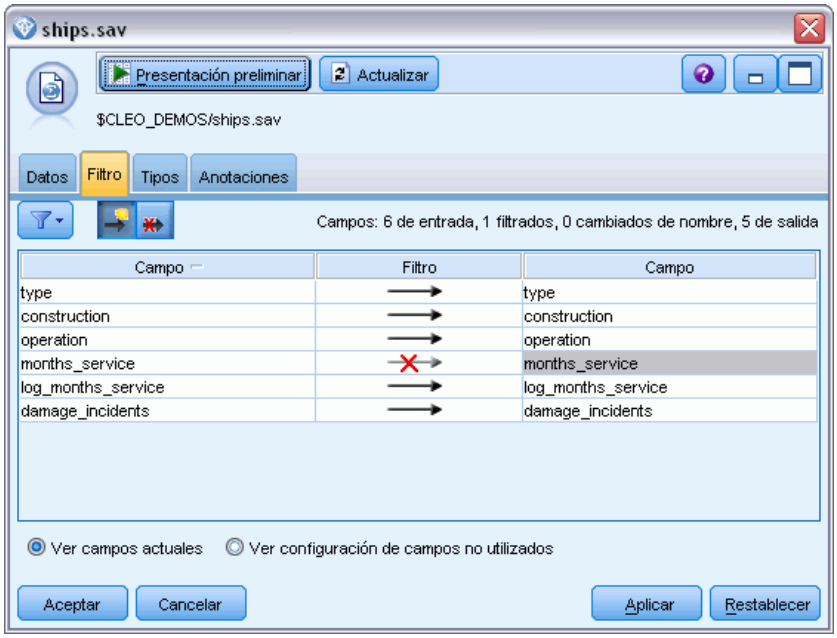
- Añada un nodo de origen Archivo Statistics que apunte a *ships.sav* en la carpeta *Demos*.

Figura 24-1
Ruta de ejemplo para analizar tasas de daños



- En la pestaña Filtro del nodo de origen, excluya el campo *meses_servicio*. Los valores transformados logarítmicamente de esta variable se incluyen en *registro_meses_servicio*, que se utilizará en el análisis.

Figura 24-2
Filtrado de un campo innecesario

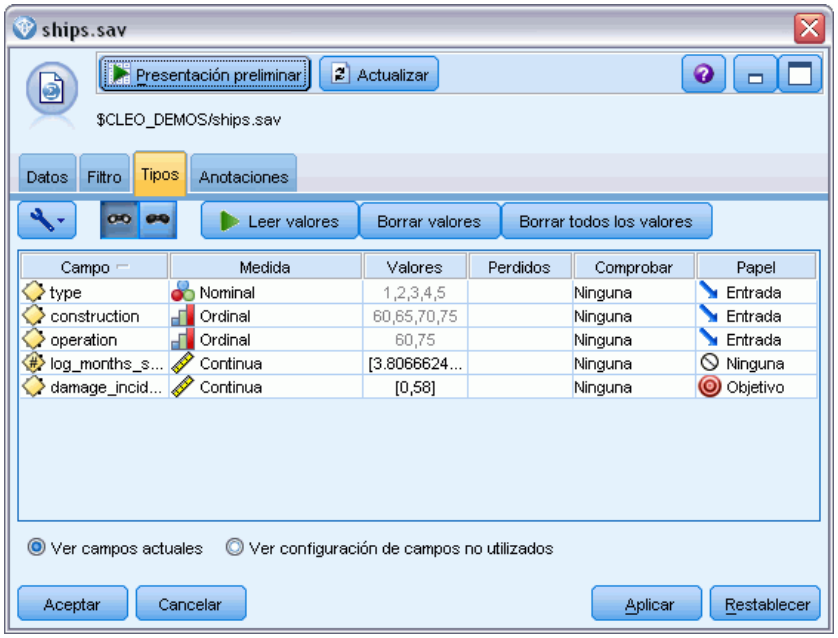


(Si lo prefiere, puede cambiar el papel de este campo a Ninguno en la pestaña Tipos en lugar de excluirla, o bien seleccionar los campos que desee utilizar en el nodo de modelado.)

- Establezca el papel del campo *incidentes_daño* como Objetivo en la pestaña Tipos del nodo de origen. El resto de campos debe tener sus papeles definidas en Entrada.

▶ Pulse en Leer valores para instanciar los datos.

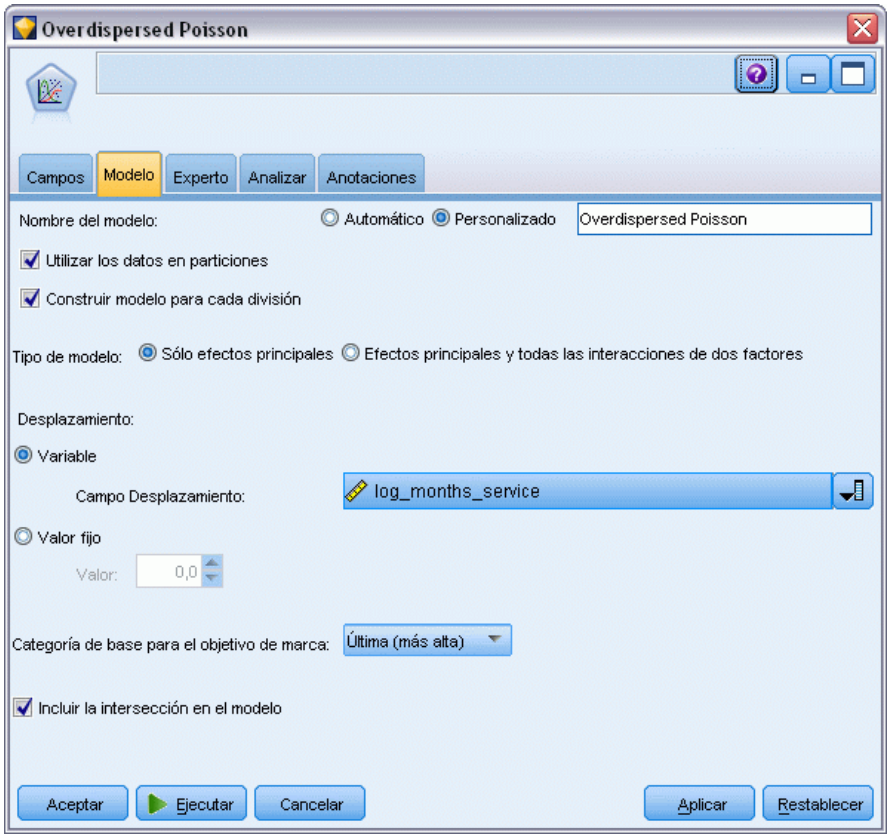
Figura 24-3
Definición del papel de campos



▶ Añada un nodo Genlin al nodo de origen; en el nodo Genlin, pulse en la pestaña Campos.

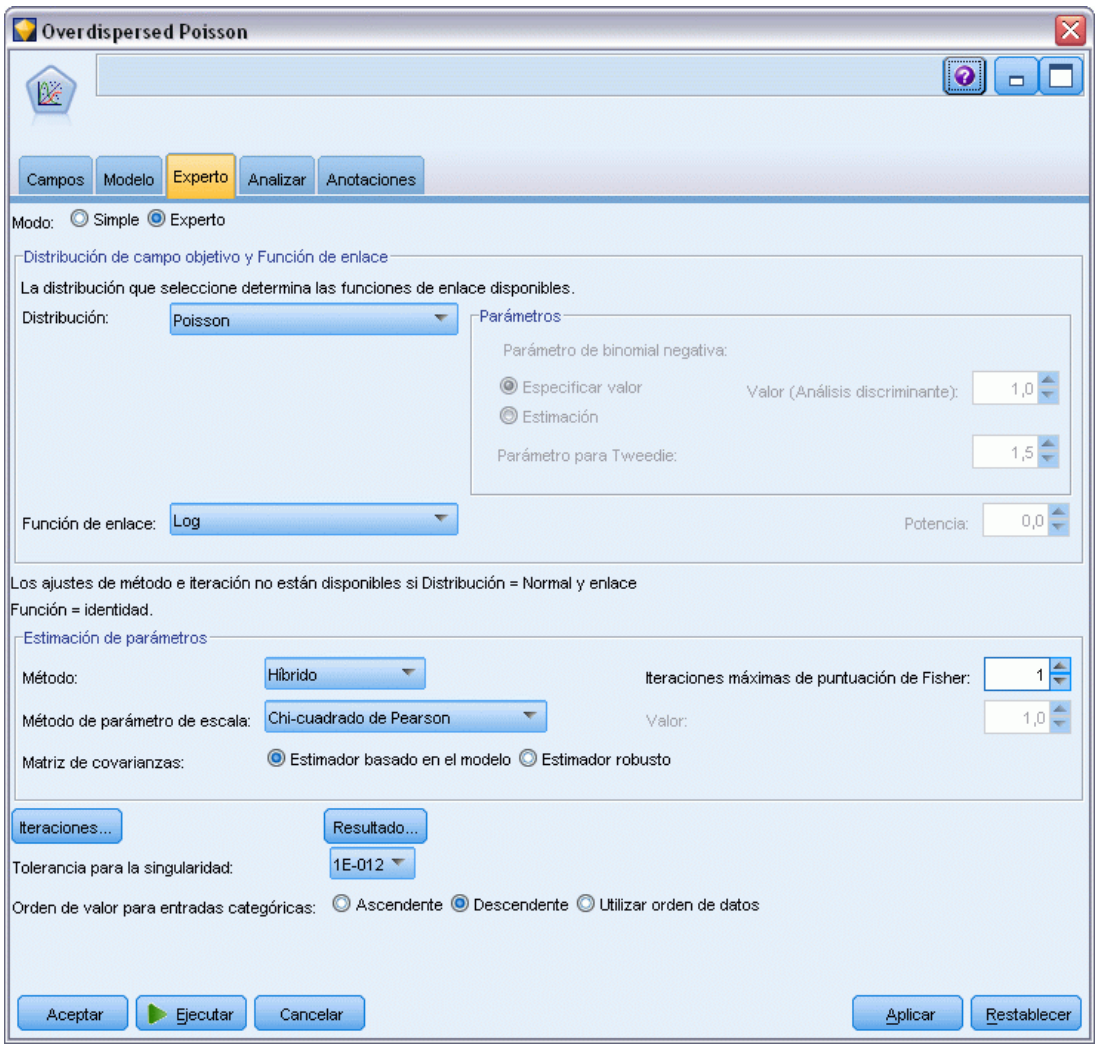
▶ Seleccione *registro_meses_servicio* como variable de desplazamiento.

Figura 24-4
Selección de opciones del modelo



- Pulse en la pestaña Experto y seleccione Experto para activar las opciones de modelado experto.

Figura 24-5
Selección de opciones de experto



- Seleccione Poisson como distribución de la respuesta y Log como función de enlace.
- Seleccione Chi-cuadrado de Pearson como método de estimación del parámetro de escala. Normalmente se supone que el parámetro de escala es 1 en una regresión de Poisson, pero McCullagh y Nelder usan la estimación de chi-cuadrado de Pearson para obtener estimaciones de la varianza y niveles de significación más conservadores.
- Seleccione Descendente como orden de categoría para los factores. Esto indica que la primera categoría de cada factor será su categoría de referencia; el efecto de esta selección en el modelo se aprecia en la interpretación de estimaciones de los parámetros.
- Pulse en Ejecutar para crear el nugget del modelo que se añadirá al lienzo de rutas y a la paleta Modelos en la esquina superior derecha. Para ver los detalles del modelo, pulse con el botón derecho en el nugget y seleccione Editar o Examinar y, a continuación, pulse en la pestaña Avanzado.

Estadísticos de bondad de ajuste

Figura 24-6
Estadísticos de bondad de ajuste

	Valor	gl	Valor/gl
Desvianza	38,695	25	1,548
Desvianza escalada	22,883	25	
Chi-cuadrado de Pearson	42,275	25	1,691
Chi-cuadrado de Pearson escalado	25,000	25	
Log verosimilitud(b,c)	-68,281		
Log-verosimilitud corregido(d)	-40,379		
Criterio de información de Akaike (AIC)	154,562		
AIC corregido para muestras finitas (AICC)	162,062		
Criterio de información bayesiano (BIC)	168,299		
AIC consistente (CAIC)	177,299		
Variable dependiente: Number of damage incidentsModelo: (Intersección), type, construction, operation, offset = log_months_service			
a. Los criterios de información están en forma "mejor cuanto más pequeño".			
b. La función de log-verosimilitud completa se muestra y se utiliza para calcular los criterios de información.			
c. El log-verosimilitud se basa en un parámetro de escala fijado en 1.			
d. El log-verosimilitud corregido se basa en un parámetro de escala estimado y se utiliza en el contraste ómnibus del ajuste del modelo.			

La tabla de estadísticos de bondad de ajuste proporciona medidas útiles para comparar diferentes modelos. Además, el *Valor/gl* de los estadísticos de desvianza y de chi-cuadrado de Pearson proporciona las estimaciones correspondientes para el parámetro de escala. Estos valores deben acercarse a 1,0 para una regresión de Poisson. Al ser mayores que 1,0, indican que puede ser conveniente ajustar el modelo sobredispersado.

Contraste Omnibus

Figura 24-7
Contraste Omnibus

Chi-cuadrado de la razón de verosimilitudes	gl	Sig.
107,633	8	,000
Variable dependiente: Number of damage incidentsModelo: (Intersección), type, construction, operation, offset = log_months_service		
a. Compara el modelo ajustado con el modelo con sólo la intersección.		

El contraste Omnibus es una prueba de chi-cuadrado de la razón de verosimilitud del modelo actual frente al modelo nulo (en este caso, de intersección). Si el valor de significación es inferior al 0,05, el modelo actual funciona mejor que el modelo nulo.

Pruebas de efectos del modelo

Figura 24-8
Contrastes de los efectos del modelo

Origen	Tipo III		
	Chi-cuadrado de Wald	gl	Sig.
(Intersección)	2138,657	1	,000
type	15,415	4	,004
construction	17,242	3	,001
operation	6,249	1	,012
Variable dependiente: Number of damage incidentsModelo: (Intersección), type, construction, operation, offset = log_months_service			

Cada término del modelo se prueba para ver si tiene algún efecto. Los términos con valores de significación inferiores a 0,05 tienen algún efecto perceptible. Todos los términos de efectos principales hacen contribuciones al modelo.

Estimaciones de los parámetros

Figura 24-9
Estimaciones de los parámetros

Parámetro	B	Típ. Error	Intervalo de confianza de Wald 95%		Contraste de hipótesis		
			Inferior	Superior	Chi-cuadrado de Wald	gl	Sig.
(Intersección)	-6,406	,2828	-6,960	-5,852	513,238	1	,000
[type=5]	,326	,3067	-,276	,927	1,127	1	,288
[type=4]	-,076	,3779	-,817	,665	,040	1	,841
[type=3]	-,687	,4279	-1,526	,151	2,581	1	,108
[type=2]	-,543	,2309	-,996	-,091	5,536	1	,019
[type=1]	0(a)	.	.	.	.	.	.
[construction=75]	,453	,3032	-,141	1,048	2,236	1	,135
[construction=70]	,818	,2208	,386	1,251	13,743	1	,000
[construction=65]	,697	,1946	,316	1,079	12,835	1	,000
[construction=60]	0(a)	.	.	.	.	.	.
[operation=75]	,384	,1538	,083	,686	6,249	1	,012
[operation=60]	0(a)	.	.	.	.	.	.
(Escala)	1,691(b)						

Variable dependiente: Number of damage incidentsModelo: (Intersección), type, construction, operation, offset = log_months_service

a. Establecido en cero ya que este parámetro es redundante.

b. Calculado basado en la chi-cuadrado de Pearson.

La tabla de estimaciones de los parámetros resume el efecto de cada predictor. Mientras que la interpretación de los coeficientes de este modelo es difícil por la naturaleza de la función de enlace, los signos de los coeficientes de las covariables y los valores relativos de los valores de los coeficientes de los niveles de factor pueden aportar información importante sobre los efectos de los predictores en el modelo.

- Para las covariables, los coeficientes positivos (negativos) indican relaciones positivas (negativas) entre predictores y resultados. El valor creciente de una covariable con un coeficiente positivo se corresponde con una tasa creciente de incidentes debidos a daños.
- En los factores, un nivel de factor con un coeficiente mayor indica una mayor incidencia de daños. El signo de un coeficiente para un nivel de factor depende del efecto del nivel de factor relativo a la categoría de referencia.

Puede realizar las siguientes interpretaciones a partir de las estimaciones de los parámetros:

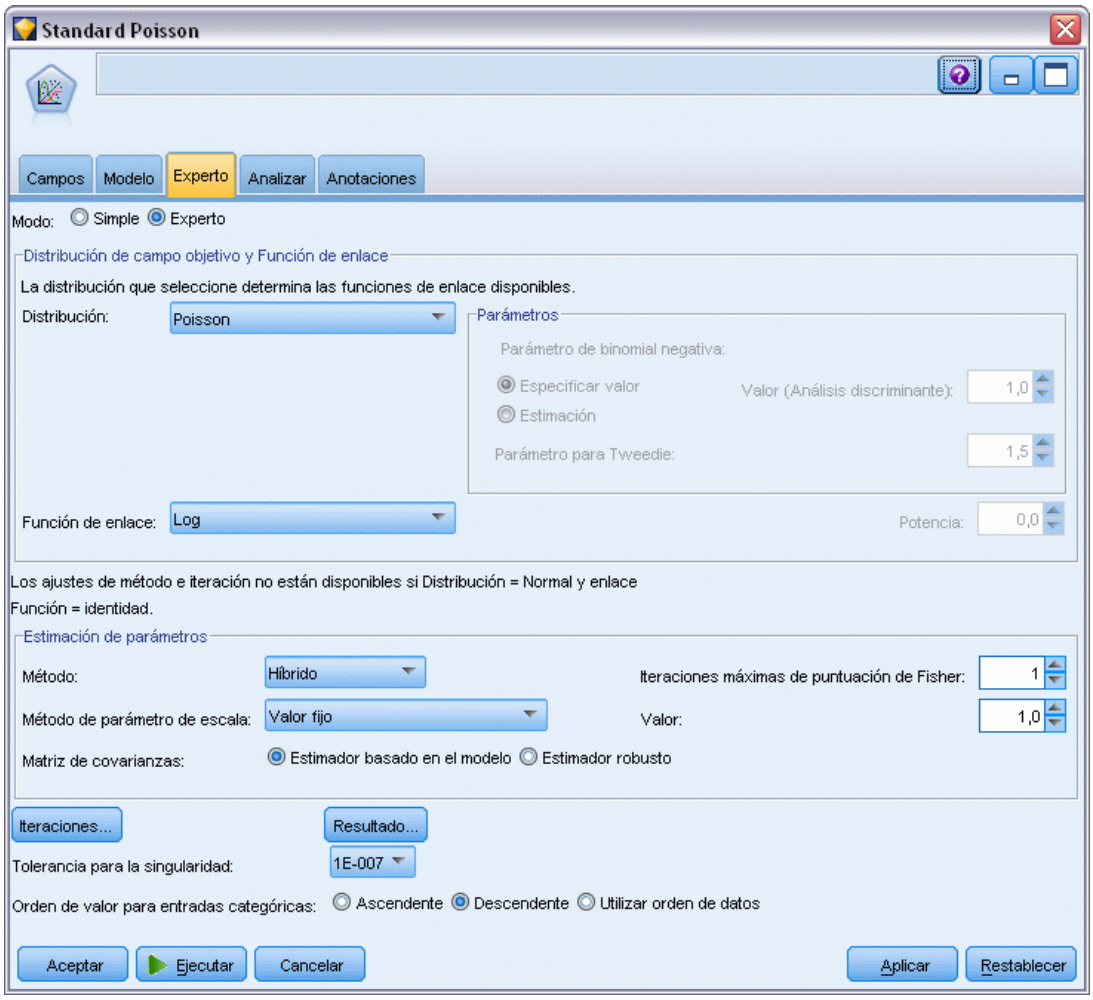
- El barco de tipo *B [tipo=2]* tiene una tasa de daños inferior (coeficiente estimado de –0,543) de manera estadísticamente significativa (valor *p* de 0,019) a la del tipo *A [tipo=1]*, la categoría de referencia. El tipo *C [tipo=3]* tiene en realidad un parámetro estimado inferior al del tipo *B*, pero la variabilidad de la estimación del *C* enmascara el efecto. Consulte las medias marginales estimadas para ver todas las relaciones entre los niveles de factor.

- Los barcos construidos entre 1965 y 1969 *[construcción=65]* y entre 1970 y 1974 *[construcción=70]* tienen tasas de daños superiores (estimaciones de coeficientes de 0,697 y 0,818, respectivamente) de manera estadísticamente significativa (valores p <0,001) a las de los construidos entre 1960 y 1964 *[construcción=60]*, la categoría de referencia. Consulte las medias marginales estimadas para ver todas las relaciones entre los niveles de factor.
- Los barcos operativos entre 1975 y 1979 *[funcionamiento=75]* tienen tasas de daños superiores (coeficiente estimado de 0,384) de manera estadísticamente significativa (valor p de 0,012) a las de los barcos operativos entre 1960 y 1974 *[funcionamiento=60]*.

Ajuste de modelos alternativos

Un problema que plantea la regresión de Poisson "sobredispersada" es que no hay una manera formal de probarla frente a la regresión de Poisson "estándar". Sin embargo, una posible prueba formal para determinar si hay sobredispersión consiste en realizar un contraste de razón de verosimilitud entre una regresión de Poisson "estándar" y una regresión binomial negativa con el resto de parámetros de configuración iguales. Si no hay sobredispersión en la regresión de Poisson, el estadístico −2×(log-verosimilitud del modelo de Poisson − log-verosimilitud del modelo binomial negativo) debe tener una distribución mixta con la mitad de su masa de probabilidad en 0 y, el resto, en una distribución chi-cuadrado con 1 grado de libertad.

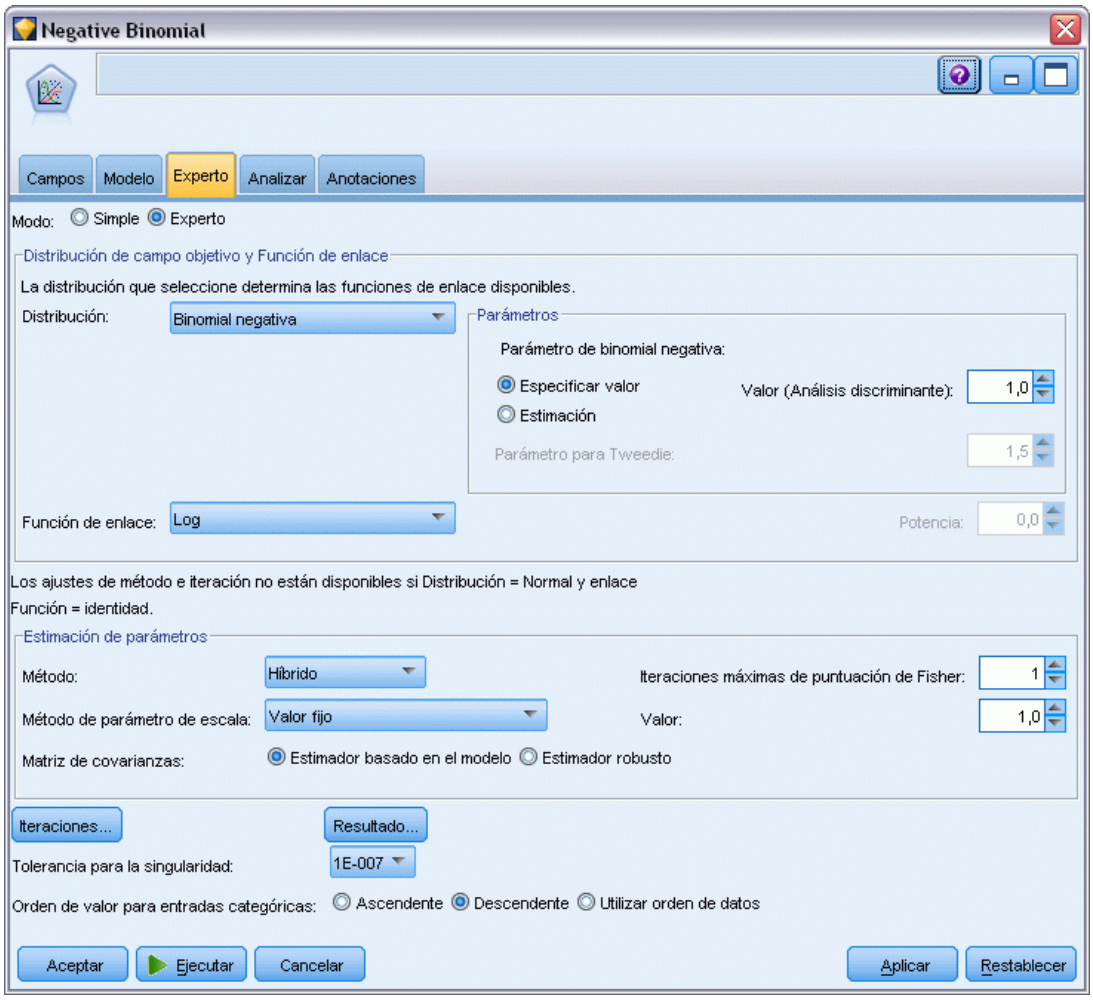
Figura 24-10
Pestaña Experto



Para ajustar la regresión de Poisson “estándar”, copie y pegue el nodo Genlin, conéctelo al nodo de origen, abra el nuevo nodo y pulse en la pestaña Experto.

- Seleccione Valor fijo como método de estimación del parámetro de escala. Este valor es 1 por defecto.

Figura 24-11
Pestaña Experto



- Para ajustar la regresión binomial negativa, copie y pegue el nodo Genlin, conéctelo al nodo de origen, abra el nuevo nodo y pulse en la pestaña Experto.
- Seleccione Binomial negativa como distribución. Deje el valor por defecto de 1 para el parámetro auxiliar.
- Ejecute la ruta y, en la pestaña Avanzado, examine los nuggets de modelo recién creados.

Estadísticos de bondad de ajuste

Figura 24-12
Estadísticos de bondad de ajuste para la regresión de Poisson estándar

	Valor	gl	Valor/gl
Desvianza	38,695	25	1,548
Desvianza escalada	38,695	25	
Chi-cuadrado de Pearson	42,275	25	1,691
Chi-cuadrado de Pearson escalado	42,275	25	
Log verosimilitud(b)	-68,281		
Criterio de información de Akaike (AIC)	154,562		
AIC corregido para muestras finitas (AICC)	162,062		
Criterio de información bayesiano (BIC)	168,299		
AIC consistente (CAIC)	177,299		
Variable dependiente: Number of damage incidentsModelo: (Intersección), type, construction, operation, offset = log_months_service			
a. Los criterios de información están en forma "mejor cuanto más pequeño".			
b. La función de log-verosimilitud completa se muestra y se utiliza para calcular los criterios de información.			

El log-verosimilitud notificado para la regresión de Poisson estándar es –68,281. Compare esto con el modelo binomial negativo.

Figura 24-13
Estadísticos de bondad de ajuste para la regresión binomial negativa

	Valor	gl	Valor/gl
Desvianza	11,145	25	,446
Desvianza escalada	11,145	25	
Chi-cuadrado de Pearson	8,815	25	,353
Chi-cuadrado de Pearson escalado	8,815	25	
Log verosimilitud(b)	-83,725		
Criterio de información de Akaike (AIC)	185,450		
AIC corregido para muestras finitas (AICC)	192,950		
Criterio de información bayesiano (BIC)	199,187		
AIC consistente (CAIC)	208,187		
Variable dependiente: Number of damage incidentsModelo: (Intersección), type, construction, operation, offset = log_months_service			
a. Los criterios de información están en forma "mejor cuanto más pequeño".			
b. La función de log-verosimilitud completa se muestra y se utiliza para calcular los criterios de información.			

El log-verosimilitud notificado para la regresión binomial negativa es –83,725. En realidad, es *más pequeño* que el log-verosimilitud para la regresión de Poisson, lo que indica (sin necesidad de realizar un contraste de razón de verosimilitud) que esta regresión binomial negativa no supone una mejora sobre la regresión de Poisson.

Sin embargo, puede que el valor seleccionado de 1 para el parámetro auxiliar de la distribución binomial negativa no sea óptimo para este conjunto de datos. Otra forma de comprobar si existe sobredispersión consiste en ajustar un modelo binomial negativo con un parámetro auxiliar igual a 0 y solicitar el contraste de multiplicadores de Lagrange en el cuadro de diálogo Resultado de la pestaña Experto. Si el contraste no arroja datos significativos, la sobredispersión no debe ser un problema para este conjunto de datos.

Resumen

Utilizando modelos lineales generalizados, ha ajustado tres modelos diferentes para los datos de frecuencias. Se ha demostrado que la regresión binomial no supone una mejora respecto a la regresión de Poisson. La regresión de Poisson sobredispersada parece ofrecer una alternativa razonable al modelo de Poisson estándar, pero no hay una prueba formal para optar por una u otra opción.

Las explicaciones de los fundamentos matemáticos de los métodos de modelado que se utilizan en IBM® SPSS® Modeler se enumeran en el *Manual de algoritmos de SPSS Modeler*.

Capítulo
25

Ajuste de una regresión gamma a reclamaciones de seguros de coches (modelos lineales generalizados)

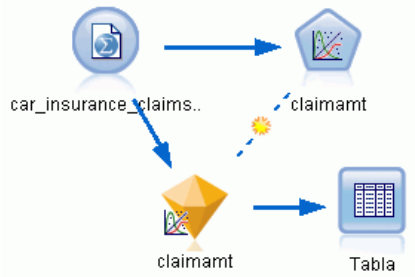
Se puede usar un modelo lineal generalizado para ajustar una regresión gamma para el análisis de datos de rango positivo. Por ejemplo, un conjunto de datos presentado y analizado en otros sitios esta relacionado con reclamaciones por daños a coches. La cantidad media de reclamaciones se puede modelar como si tuviera una distribución gamma, utilizando una función de enlace inversa para relacionar la media de la variable dependiente con una combinación lineal de los predictores. Para tener en cuenta el número variable de reclamaciones utilizado para calcular la cantidad variable de reclamaciones, especifique el *número de reclamaciones* como la ponderación de escalamiento.

Este ejemplo utiliza la ruta denominada *car-insurance_genlin.str*, que hace referencia al archivo de datos denominado *car_insurance_claims.sav*. El archivo de datos está en la carpeta *Demos* y el archivo de ruta está en la subcarpeta *streams*. Si desea obtener más información, consulte el tema Carpeta Demos en el capítulo 1 en *Manual de usuario de IBM SPSS Modeler 14.2*.

Creación de la ruta

- Añada un nodo de origen de archivo Statistics apuntando a *car_insurance_claims.sav* en la carpeta *Demos*.

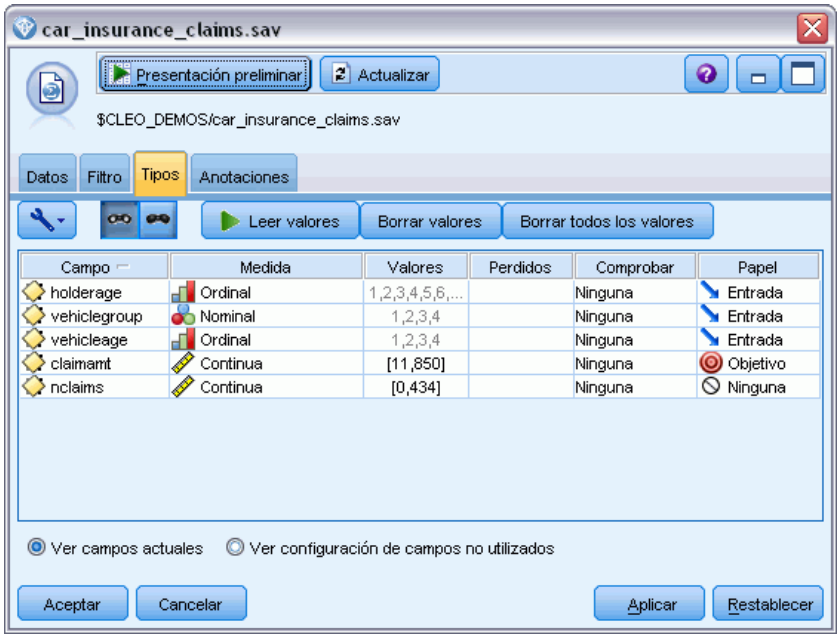
Figura 25-1
Ruta de muestra para pronosticar reclamaciones de seguros de coches



- Establezca el papel del campo *cantrecla* como Objetivo en la pestaña Tipos del nodo de origen. El resto de campos debe tener sus papeles definidas en Entrada.

▶ Pulse en Leer valores para instanciar los datos.

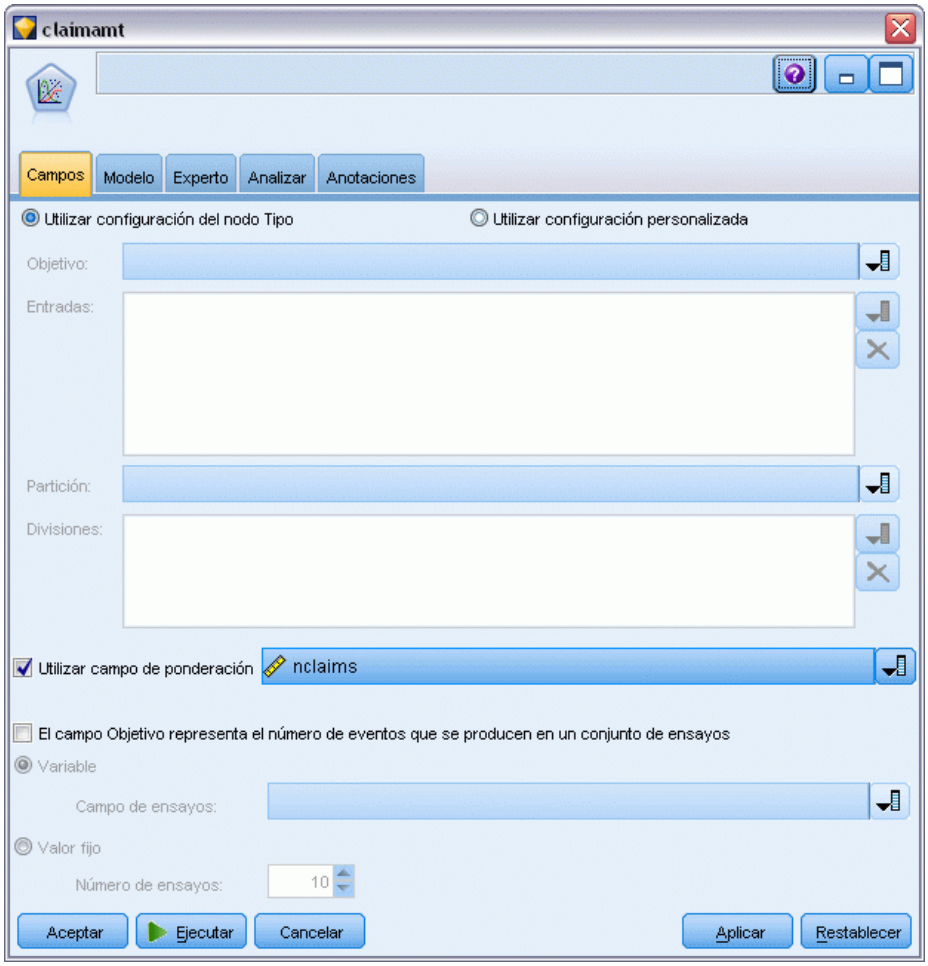
Figura 25-2
Definición del papel de campos



▶ Añada un nodo Genlin al nodo de origen; en el nodo Genlin, pulse en la pestaña Campos.

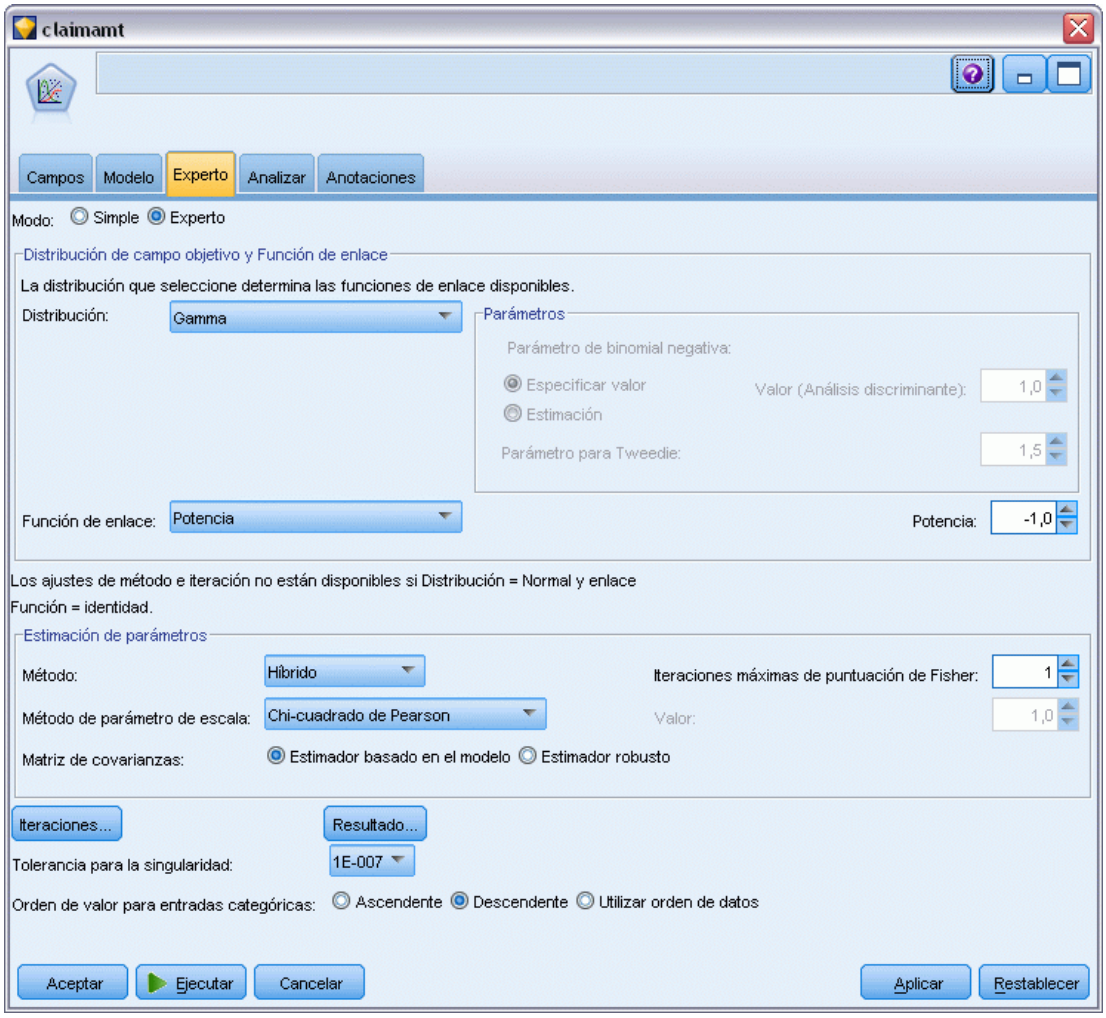
▶ Seleccione *reclamacionesn* como el campo de ponderación de escala.

Figura 25-3
Selección de opciones de campo



▶ Pulse en la pestaña Experto y seleccione Experto para activar las opciones de modelado experto.

Figura 25-4
Selección de opciones de experto



▶ Seleccione Gamma como distribución de la respuesta.

▶ Seleccione Potencia como la función de enlace y especifique -1,0 como el exponente de la función exponencial. Este es un enlace inverso.

▶ Seleccione Chi-cuadrado de Pearson como método de estimación del parámetro de escala. Este es el método utilizado por McCullagh y Nelder, aquí lo seguimos para replicar sus resultados.

▶ Seleccione Descendente como orden de categoría para los factores. Esto indica que la primera categoría de cada factor será su categoría de referencia; el efecto de esta selección en el modelo se aprecia en la interpretación de estimaciones de los parámetros.

▶ Pulse en Ejecutar para crear el nugget del modelo que se añadirá al lienzo de rutas y a la paleta Modelos en la esquina superior derecha. Para ver los detalles del modelo, pulse con el botón

derecho en el nugget de modelo y seleccione Editar o Examinar y, a continuación, seleccione la pestaña Avanzado.

Estimaciones de los parámetros

Figura 25-5
Estimaciones de los parámetros

Parámetro	B	Típ. Error	Intervalo de confianza de Wald 95%		Contraste de hipótesis		
			Inferior	Superior	Chi-cuadrado de Wald	gl	Sig.
(Intersección)	,003	,0004	,003	,004	66,593	1	,000
[holderage=8]	,001	,0004	,000	,002	4,898	1	,027
[holderage=7]	,001	,0004	,000	,002	5,046	1	,025
[holderage=6]	,001	,0004	,000	,002	5,740	1	,017
[holderage=5]	,001	,0004	,001	,002	10,682	1	,001
[holderage=4]	,000	,0004	,000	,001	1,268	1	,260
[holderage=3]	,000	,0004	,000	,001	,720	1	,396
[holderage=2]	,000	,0004	-,001	,001	,054	1	,816
[holderage=1]	0(a)	.	.	.	.	.	.
[vehiclegroup=4]	-,001	,0002	-,002	-,001	61,883	1	,000
[vehiclegroup=3]	-,001	,0002	-,001	,000	13,039	1	,000
[vehiclegroup=2]	3,77E-005	,0002	,000	,000	,050	1	,823
[vehiclegroup=1]	0(a)	.	.	.	.	.	.
[vehicleage=4]	,004	,0004	,003	,005	88,175	1	,000
[vehicleage=3]	,002	,0002	,001	,002	53,013	1	,000
[vehicleage=2]	,000	,0001	,000	,001	13,191	1	,000
[vehicleage=1]	0(a)	.	.	.	.	.	.
(Escala)	1,209(b)						

Variable dependiente: Average cost of claimsModelo: (Intersección), holderage, vehiclegroup, vehicleage

a. Establecido en cero ya que este parámetro es redundante.

b. Calculado basado en la chi-cuadrado de Pearson.

El contraste ómnibus y las pruebas de los efectos del modelo (no se muestran) indican que el modelo funciona mejor que el modelo nulo y que cada uno de los términos de efectos principales contribuyen al modelo. La tabla de estimaciones de parámetros muestra los mismos valores obtenidos por McCullagh y Nelder para los niveles de factor y el parámetro de escala.

Resumen

Al utilizar los modelos lineales generalizados, se ha ajustado una regresión gamma a los datos de reclamación. Tenga en cuenta que aunque la función de enlace canónica para la distribución gamma se utilizó en este modelo, un enlace de logaritmo también proporcionaría resultados razonables. En general, es difícil, por no decir imposible, comparar directamente modelos con diferentes funciones de enlace; no obstante, el enlace de logaritmo es un caso especial de enlace de potencia donde el exponente es 0, así se pueden comparar las desviaciones de un modelo con un enlace de logaritmo y un modelo con un enlace de potencia para determinar cuál se ajusta mejor (consulte, por ejemplo, la sección 11.3 de McCullagh y Nelder).

Las explicaciones de los fundamentos matemáticos de los métodos de modelado que se utilizan en IBM® SPSS® Modeler se enumeran en el *Manual de algoritmos de SPSS Modeler*.

Capítulo
26

Clasificación de muestras de células (SVM)

Máquina de vectores de soporte (SVM) es una clasificación y técnica de regresión especialmente adecuada para conjuntos de datos de grandes dimensiones. Un conjunto de datos de grandes dimensiones es uno con un amplio número de predictores, como el que se puede encontrar en el campo de bioinformática (la aplicación de tecnología de la información a la bioquímica y a los datos biológicos).

Un investigador médico ha obtenido un conjunto de datos con las características de un número de muestras de células humanas extraídas de pacientes con riesgo de desarrollar un cáncer. El análisis de los datos originales demostró que muchas de las características de las muestras benignas y malignas eran muy diferentes. El investigador quiere desarrollar un modelo SVM que pueda utilizar los valores de estas características de las células en las muestras de otros pacientes para indicar si las muestras pueden ser benignas o malignas.

Este ejemplo utiliza la ruta denominada *svm_cancer.str*, disponible en la carpeta *Demos* bajo la subcarpeta *streams*. El archivo de datos es *cell_samples.data*. Si desea obtener más información, consulte el tema Carpeta Demos en el capítulo 1 en *Manual de usuario de IBM SPSS Modeler 14.2*.

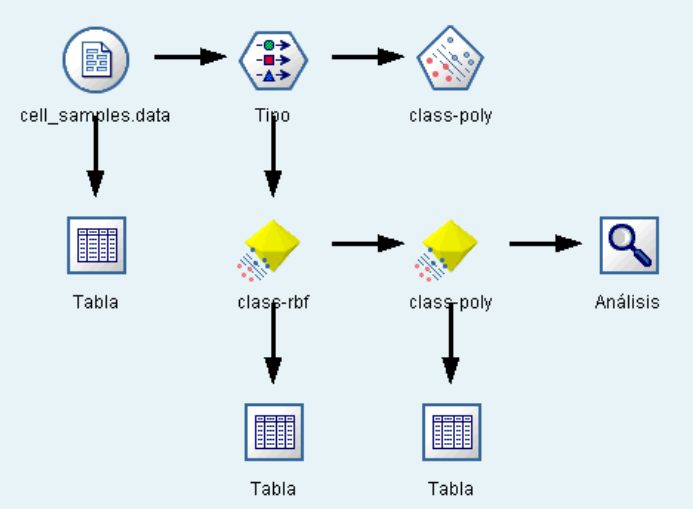
El ejemplo está basado en un conjunto de datos está disponible de forma pública en UCI Machine Learning Repository (Asunción y Newman, 2007). El conjunto de datos contiene varios cientos de muestras de células humanas y cada una contiene los valores de un conjunto de características de celdas. Los campos de cada registro son:

Nombre de campo	Descripción
ID	Identificador de paciente
Grupo	Grosor de grupo
UnifTamaño	Uniformidad del tamaño de célula
UnifForma	Uniformidad de la forma del tamaño de célula
MargAdh	Adhesión marginal
TamEpiSim	Tamaño de célula epitelial simple
NucDes	Núcleo desnudo
CromBland	Cromatina blanda
NuclNorm	Nucleolos normales
Mit	Mitosis
Class	Benigna o maligna

En este ejemplo se utiliza un conjunto de datos con un número relativamente pequeño de predictores en cada registro.

Creación de la ruta

Figura 26-1
Ruta de ejemplo para el modelado de SVM



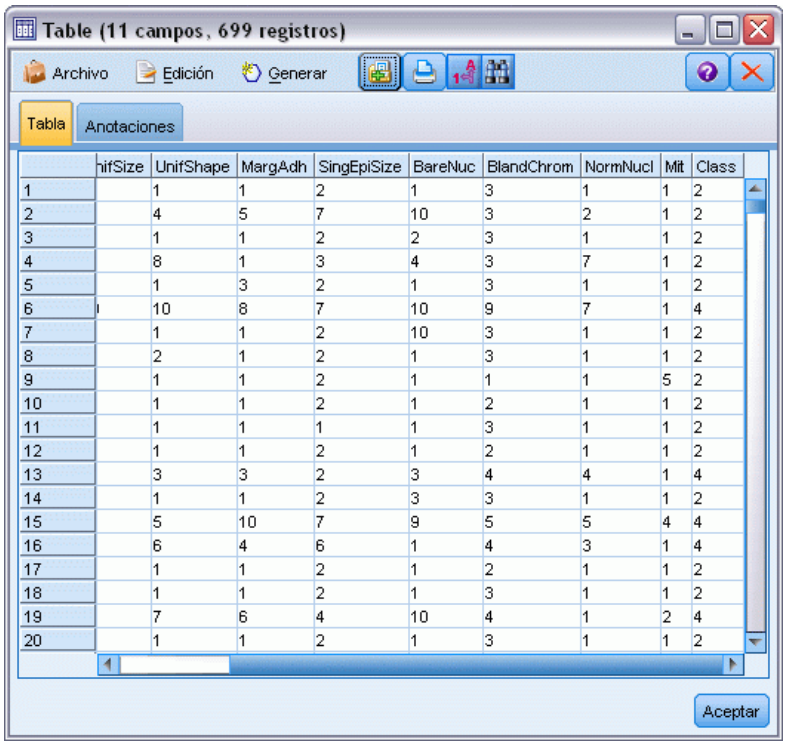
- Cree una nueva ruta y añada un nuevo núcleo de origen Archivo var. que apunte a *cell_samples.data* en la carpeta *Demos* de su instalación de IBM® SPSS® Modeler.

  Vamos a echar un vistazo a los datos del archivo de origen.

- Añada un nodo Tabla a la ruta.

- Añada un nodo Tabla al nodo Archivo var. y ejecute la ruta.

Figura 26-2
Datos de origen de SVM



	hifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1		1	1	2	1	3	1	1	2
2		4	5	7	10	3	2	1	2
3		1	1	2	2	3	1	1	2
4		8	1	3	4	3	7	1	2
5		1	3	2	1	3	1	1	2
6		10	8	7	10	9	7	1	4
7		1	1	2	10	3	1	1	2
8		2	1	2	1	3	1	1	2
9		1	1	2	1	1	1	5	2
10		1	1	2	1	2	1	1	2
11		1	1	1	1	3	1	1	2
12		1	1	2	1	2	1	1	2
13		3	3	2	3	4	4	1	4
14		1	1	2	3	3	1	1	2
15		5	10	7	9	5	5	4	4
16		6	4	6	1	4	3	1	4
17		1	1	2	1	2	1	1	2
18		1	1	2	1	3	1	1	2
19		7	6	4	10	4	1	2	4
20		1	1	2	1	3	1	1	2

El campo *ID* contiene los identificadores de pacientes. Las características de las muestras de células de cada paciente se encuentran en los campos *Grupo* a *Mit*. Los valores se clasifican del 1 al 10, siendo 1 el valor más cercano a benigno.

El campo *Clase* contiene el diagnóstico, confirmado por procedimientos médicos independientes, que definen si las muestras son benignas (valor = 2) o malignas (valor = 4).

Figura 26-3
Configuración del nodo Tipo

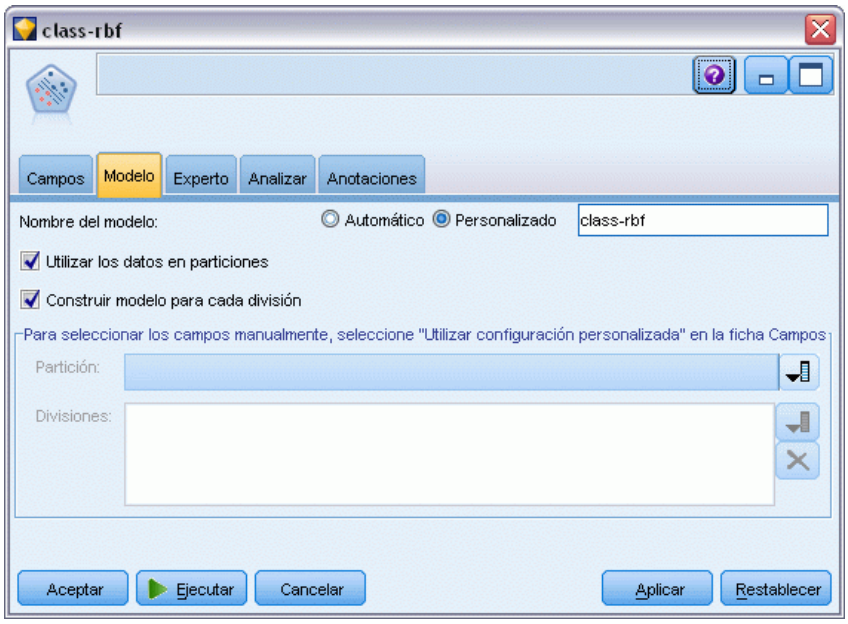
- Añada un nodo Tipo al nodo Archivo var.
- Abra el nodo Tipo.

  Queremos que el modelo pronostique el valor de *Clase* (es decir, benigno (=2) o maligno (=4)). Como este campo sólo puede tener dos valores posibles, necesitamos cambiar su nivel de medición para reflejar este hecho.

- En la columna Medición del campo *Clase* (el último de la lista), pulse en el valor Continuo y cámbielo a Marca.
- Pulse en Leer valores.
- En la columna Papel, defina el papel de *ID* (identificador de paciente) a Ninguno, ya que no se utilizará como predictor u objetivo para el modelo.
- Defina el papel del objetivo, *Clase* a Objetivo y deje el papel del resto de campos (predictores) como Entrada.
- Pulse en Aceptar.

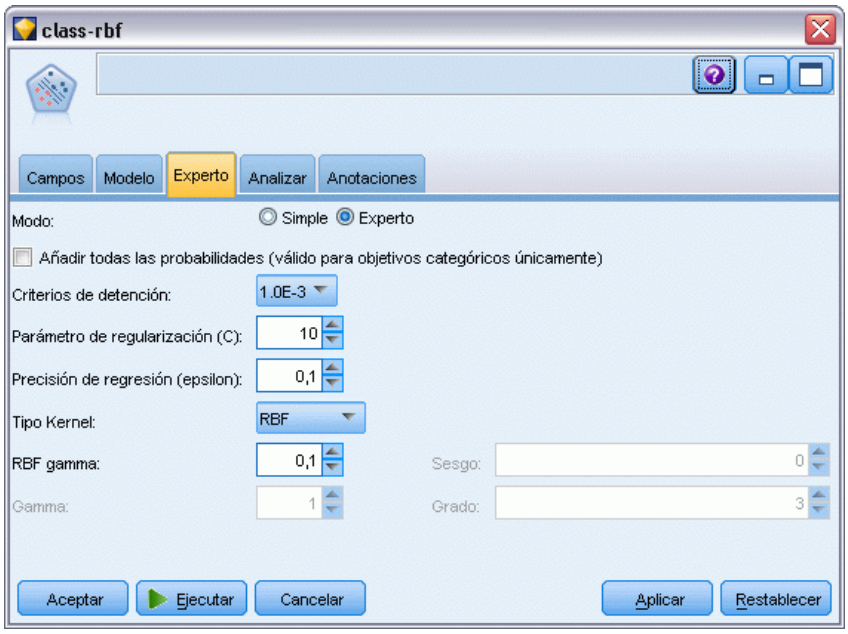
  El nodo SVM ofrece una selección de las funciones de kernel que ejecutan este procesamiento. Como no existe una forma fácil de saber la función que se comporta mejor con un conjunto de datos, vamos a seleccionar funciones diferentes y comparar sus resultados. Comencemos por la función predefinida, RBF (Función de base radial).

Figura 26-4
Configuración de la pestaña Modelo



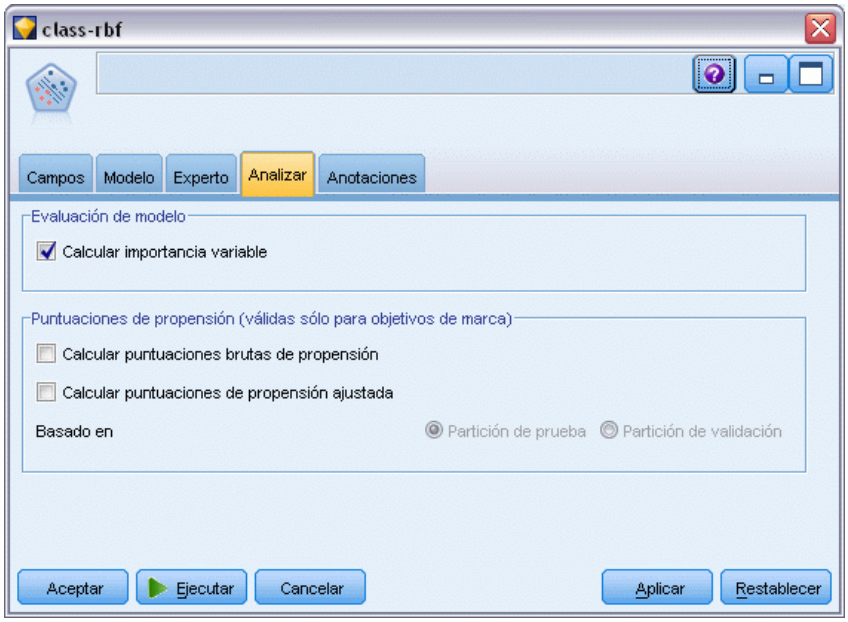
- En la paleta Modelado, añada un nodo SVM al nodo Tipo.
- Abra el nodo SVM. En la pestaña Modelo, pulse en la opción Personalizado de Nombre del modelo e introduzca *clase-rbf* en el campo de texto adyacente.

Figura 26-5
Configuración predefinida de la pestaña Experto



- En la pestaña Experto, defina el Modo a Experto para mejorar la legibilidad pero deje todas las opciones predefinidas tal cual. Tenga en cuenta que el tipo de Kernel está definido a RBF por defecto. Todas las opciones aparecen atenuadas en modo Simple.

Figura 26-6
Configuración de la pestaña Analizar

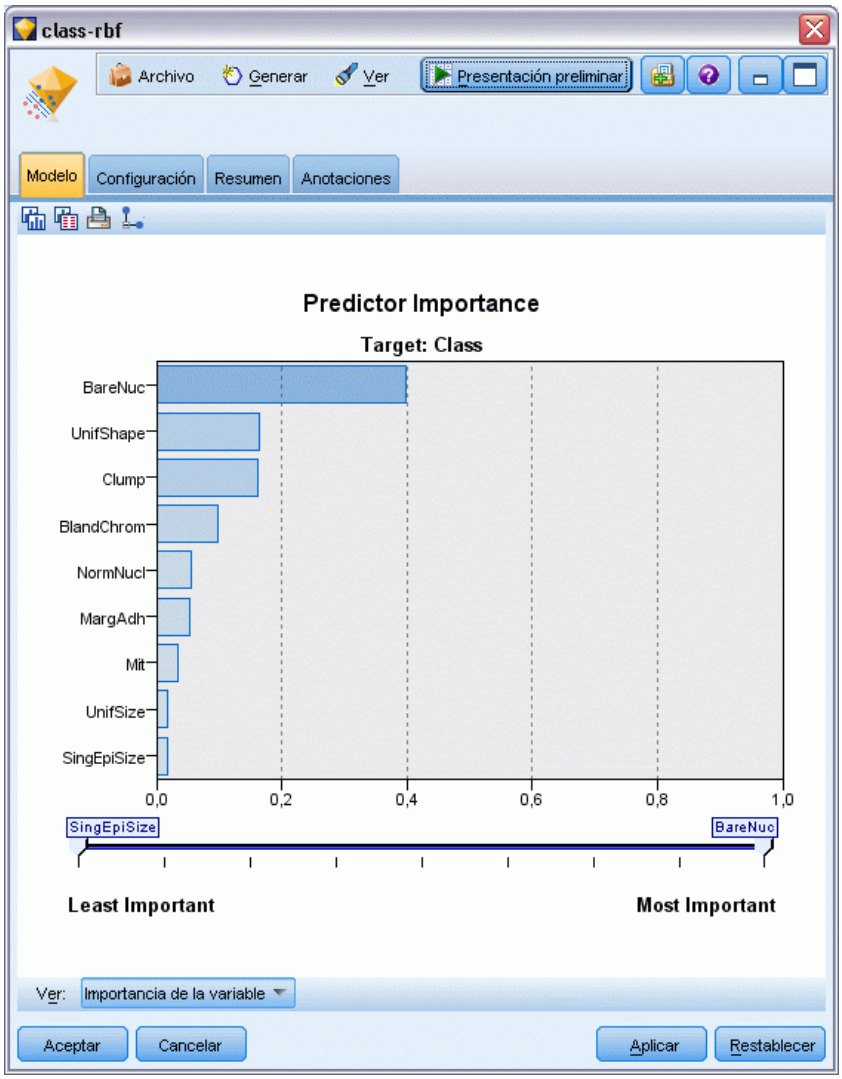


- En la pestaña Analizar, active la casilla de verificación Calcular importancia variable.

- Pulse en Ejecutar. El nugget de modelo se coloca en la ruta, y en la paleta Modelos en la parte derecha de la pantalla.
- Pulse dos veces en el nugget de modelo de la ruta.

Examen de los datos

Figura 26-7
Gráfico Importancia del predictor

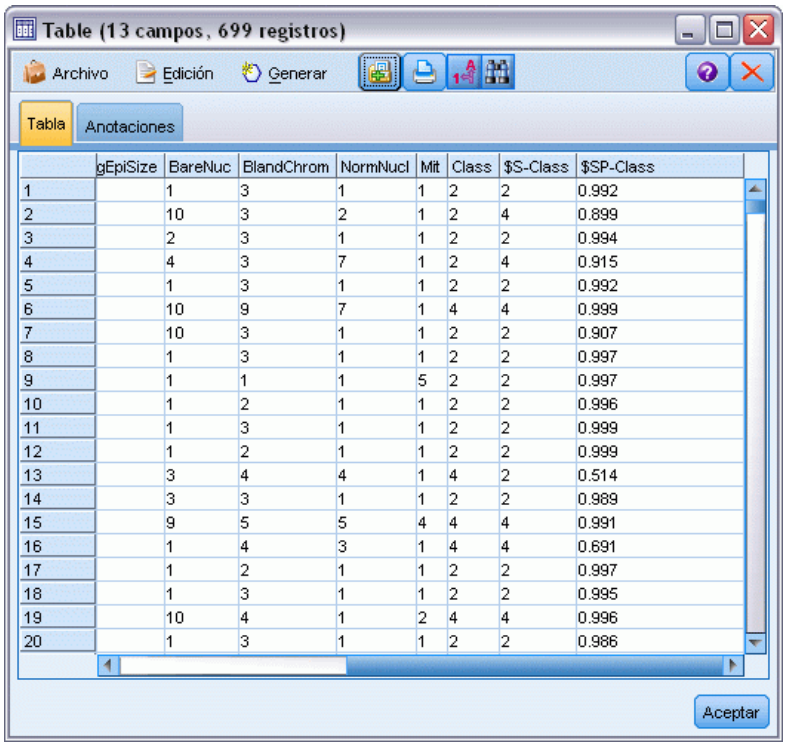


En la pestaña Modelo, el gráfico Importancia del predictor muestra el efecto relativo de los diferentes campos en la predicción. Muestra que *NucDes* es el mayor afectado, mientras que *UnifForma* y *Grupo* son también significativos.

- Pulse en Aceptar.

- Añada un nodo Tabla al nugget de modelo *clase-rbf*.
- Abra el nodo Tabla y pulse en Ejecutar.

Figura 26-8
Campos añadidos para el valor de pronóstico y confianza



	gEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class	$S-Class	$SP-Class
1		1	3	1	1	2	2	0.992
2		10	3	2	1	2	4	0.899
3		2	3	1	1	2	2	0.994
4		4	3	7	1	2	4	0.915
5		1	3	1	1	2	2	0.992
6		10	9	7	1	4	4	0.999
7		10	3	1	1	2	2	0.907
8		1	3	1	1	2	2	0.997
9		1	1	1	5	2	2	0.997
10		1	2	1	1	2	2	0.996
11		1	3	1	1	2	2	0.999
12		1	2	1	1	2	2	0.999
13		3	4	4	1	4	2	0.514
14		3	3	1	1	2	2	0.989
15		9	5	5	4	4	4	0.991
16		1	4	3	1	4	4	0.691
17		1	2	1	1	2	2	0.997
18		1	3	1	1	2	2	0.995
19		10	4	1	2	4	4	0.996
20		1	3	1	1	2	2	0.986

- El modelo ha creado dos campos extra. Desplace la tabla a la derecha para verlos:

Nombre del campo nuevo	**Descripción**
$S-Class	Los valores de *Clase* pronosticados por el modelo.
$SP-Class	Puntuación de propensión de este pronóstico (la posibilidad de que este pronóstico sea verdadero, un valor de 0,0 a 1,0).

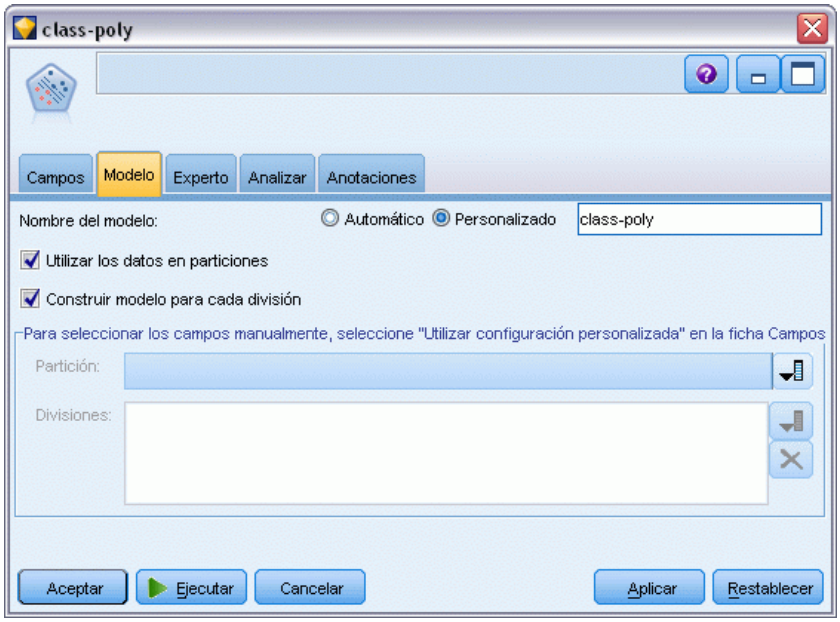
Sólo con mirar la tabla podemos ver que la puntuación de propensión (en la columna *$SP-Class*) de la mayoría de registros es razonablemente alta.

Sin embargo, hay algunas excepciones significativas; por ejemplo, el registro del paciente 1041801 en la línea 13, donde el valor de 0,514 es inaceptablemente bajo. Además, si compara *Clase* con *$S-Class*, queda claro que este modelo ha realizado numerosos pronósticos incorrectos, incluso si la puntuación de propensión era relativamente alta (por ejemplo, líneas 2 y 4).

Veamos si podemos mejorar los resultados con un tipo de función diferente.

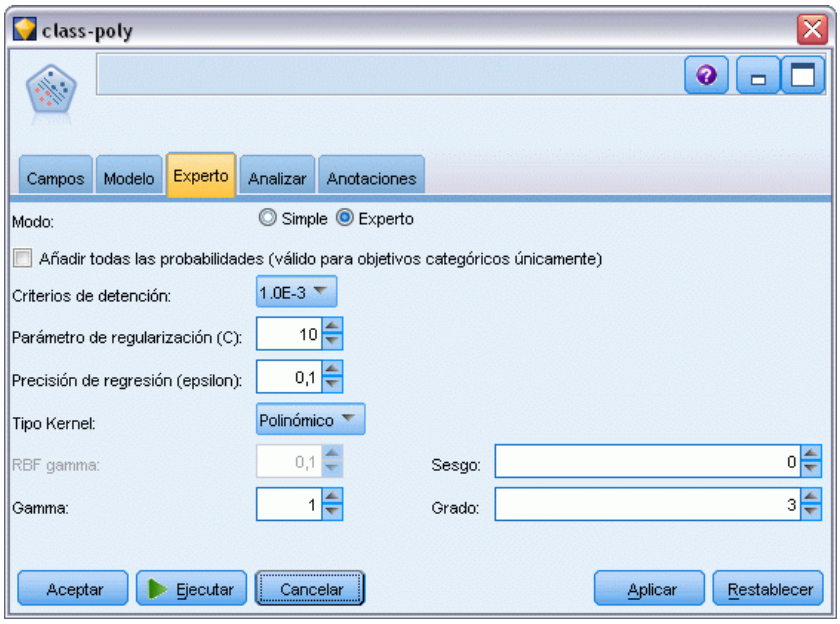
Prueba de una función diferente

Figura 26-9
Configuración de un nombre nuevo para el modelo



- Cierre la ventana de resultado de la tabla.
- Conecte un segundo de modelado SVM al nodo Tipo.
- Abra el nuevo nodo SVM.
- En la pestaña Modelo, seleccione Personalizado e introduzca *clase-poli* como el nombre del modelo.

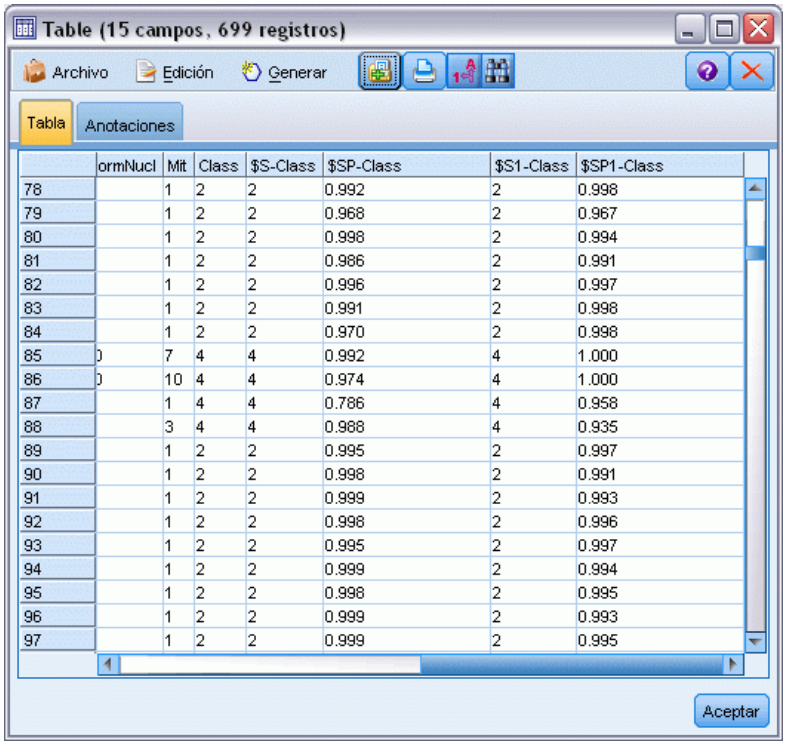
Figura 26-10
Configuración de la pestaña Experto para Polinómica



- En la pestaña Experto, defina Modo a Experto.
- Defina Tipo Kernel a Polinómica y pulse en Ejecutar. El nugget de modelo *clase-poli* se añade a la ruta y también a la paleta Modelos en la parte superior derecha de la pantalla.
- Conecte el nugget de modelo *clase-rbf* al nugget de modelo *clase-poli* (seleccione Reemplazar en el cuadro de diálogo de advertencia).
- Añada un nodo Tabla al nugget de modelo *clase-poli*.
- Abra el nodo Tabla y pulse en Ejecutar.

Comparación de los resultados

Figura 26-11
Campos añadidos para la función polinómica



	ormNucl	Mit	Class	$S-Class	$SP-Class	$S1-Class	$SP1-Class
78		1	2	2	0.992	2	0.998
79		1	2	2	0.968	2	0.967
80		1	2	2	0.998	2	0.994
81		1	2	2	0.986	2	0.991
82		1	2	2	0.996	2	0.997
83		1	2	2	0.991	2	0.998
84		1	2	2	0.970	2	0.998
85	)	7	4	4	0.992	4	1.000
86	)	10	4	4	0.974	4	1.000
87		1	4	4	0.786	4	0.958
88		3	4	4	0.988	4	0.935
89		1	2	2	0.995	2	0.997
90		1	2	2	0.998	2	0.991
91		1	2	2	0.999	2	0.993
92		1	2	2	0.998	2	0.996
93		1	2	2	0.995	2	0.997
94		1	2	2	0.999	2	0.994
95		1	2	2	0.998	2	0.995
96		1	2	2	0.999	2	0.993
97		1	2	2	0.999	2	0.995

- Desplace la tabla a la derecha para ver los nuevos campos añadidos:

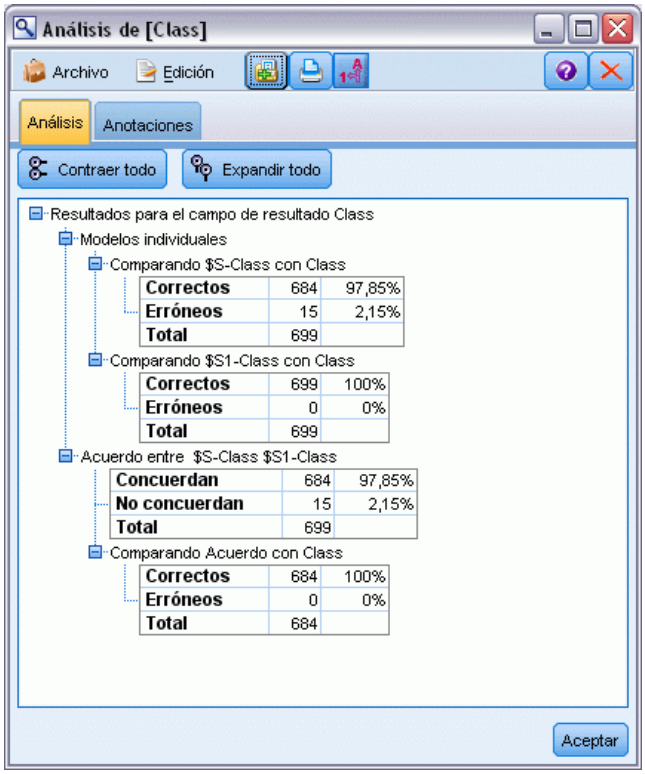
  Los campos generados para el tipo de función polinómica se denominan *$S1-Class* y *$SP1-Class*.

  Los resultados de la función polinómica parecen mucho mejores. La mayoría de puntuaciones de propensión son 0,995 o mejores, lo que es muy esperanzador.

- Para confirmar la mejora en el modelo, añada un nodo Análisis al nugget de modelo *clase-poli*.

  Abra el nodo Análisis y pulse en Ejecutar.

Figura 26-12
Nodo Análisis



Esta técnica con el nodo Análisis le permite comparar dos o más nuggets de modelos al mismo tiempo. El resultado del nodo Análisis muestra que la función RBF pronostica correctamente el 97,85% de los casos, lo que es muy positivo. Sin embargo, los resultados muestran que la función polinómica ha pronosticado correctamente el diagnóstico en cada caso concreto. En la práctica es poco probable ver una precisión del 100%, aunque puede utilizar el nodo Análisis para determinar si el modelo tiene una precisión aceptable para su aplicación en particular.

De hecho, ninguno del resto de tipos de funciones (Sigmoide y Lineal) se comporta como la función polinómica en este conjunto de datos concreto. Sin embargo, con un conjunto de datos diferente, los resultados pueden ser muy diferentes, por lo que siempre merece la pena intentar todas las opciones.

Resumen

Ha utilizado diferentes tipos de funciones de kernel SVM para pronosticar una clasificación de diferentes atributos. Ha comprobado cómo diferentes modelos de kernel ofrecen diferentes resultados para el mismo conjunto de datos y cómo puede medir la mejora del modelo con respecto a otro.

Capítulo
27

Uso de la regresión de Cox en el modelo de tiempo de abandono de cliente

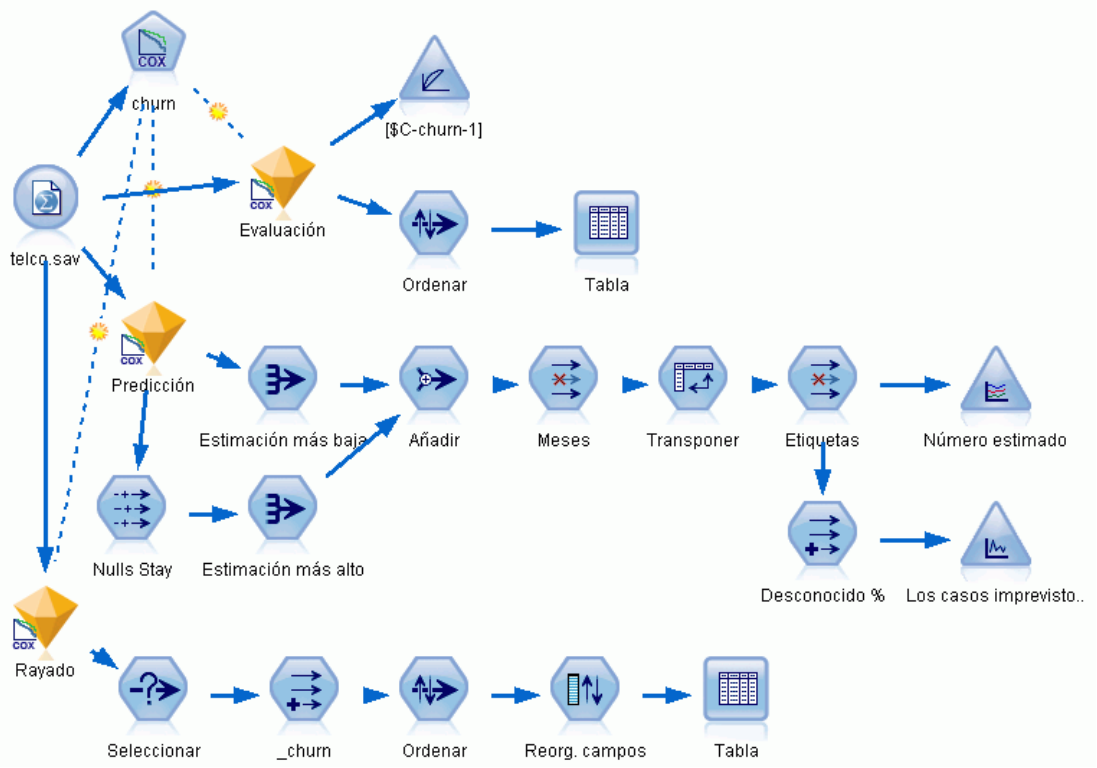
Como parte de su esfuerzo por reducir el abandono de clientes, una empresa de telecomunicaciones se ha interesado en el modelado del "tiempo de abandono" para determinar los factores que se asocian a los clientes que están a punto de cambiarse de servicio. Para este propósito, se ha seleccionado una muestra aleatoria de clientes y se ha extraído de la base de datos su duración como cliente (si aún son o no clientes activos) y distintos campos.

Este ejemplo usa la ruta *telco_coxreg.str*, que hace referencia al archivo de datos *telco.sav*. El archivo de datos está en la carpeta *Demos* y el archivo de ruta está en la subcarpeta *streams*. Si desea obtener más información, consulte el tema Carpeta Demos en el capítulo 1 en *Manual de usuario de IBM SPSS Modeler 14.2*.

Generación de un modelo adecuado

- Añada un nodo de origen Archivo Statistics apuntando a *telco.sav* en la carpeta *Demos*.

Figura 27-1
Ruta de ejemplo para analizar el tiempo de abandono



- En la pestaña Filtro del nodo de origen, excluya los campos *región*, *ingresos*, *longten* a *wireten* y *loglong* a *logwire*.

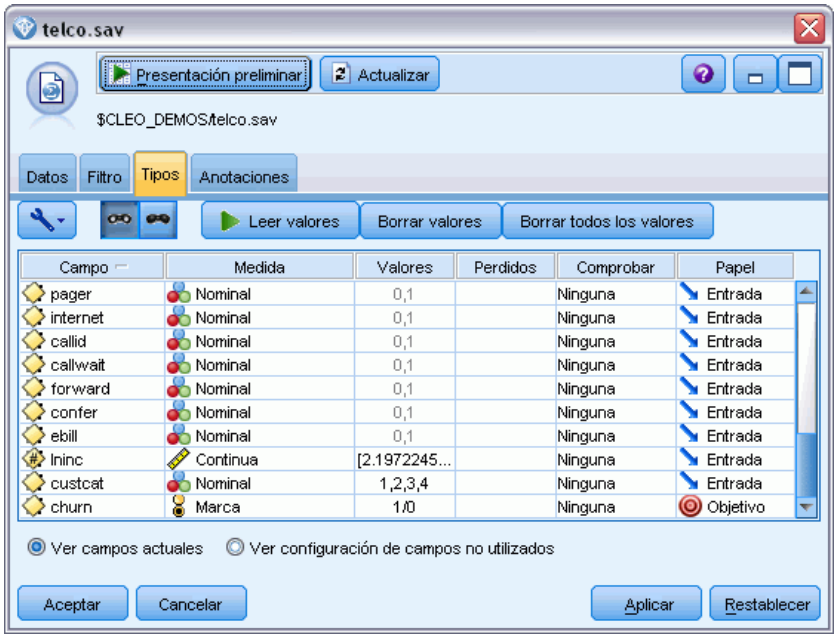
Figura 27-2
Filtrado de campos innecesarios

(Si lo prefiere, puede cambiar el papel de este campo a Ninguno en la pestaña Tipos en lugar de excluirla, o bien seleccionar los campos que desee utilizar en el nodo de modelado.)

- En la pestaña Tipos del nodo de origen, configure el papel del campo *abandono* como Objetivo y defina su nivel de medición como Marca. El resto de campos debe tener sus papeles definidas en Entrada.

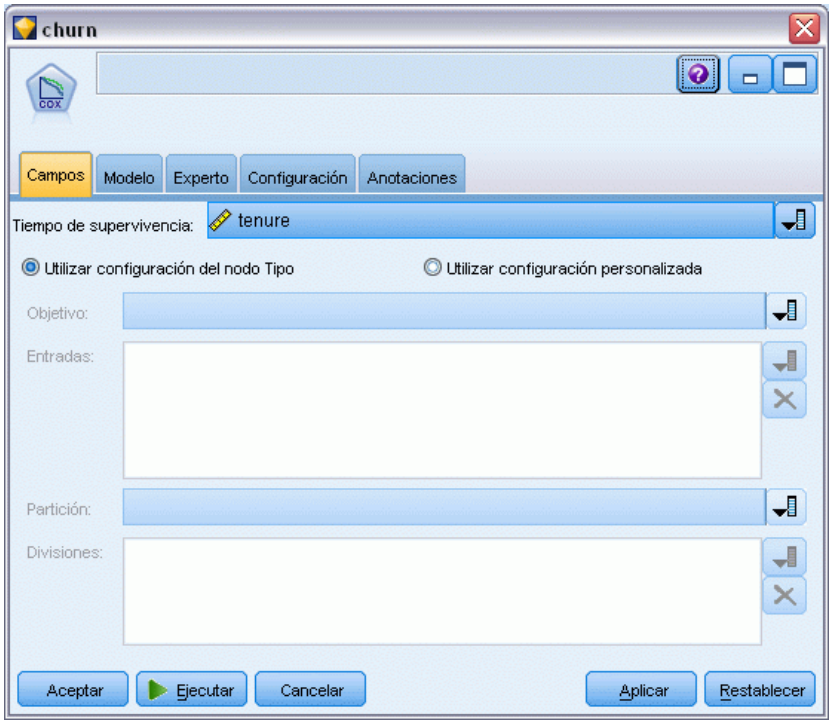
▶ Pulse en Leer valores para instanciar los datos.

Figura 27-3
Definición del papel de campos



- Añada un nodo Cox al nodo de origen; en la pestaña Campos, seleccione *periodo* como la variable temporal de supervivencia.

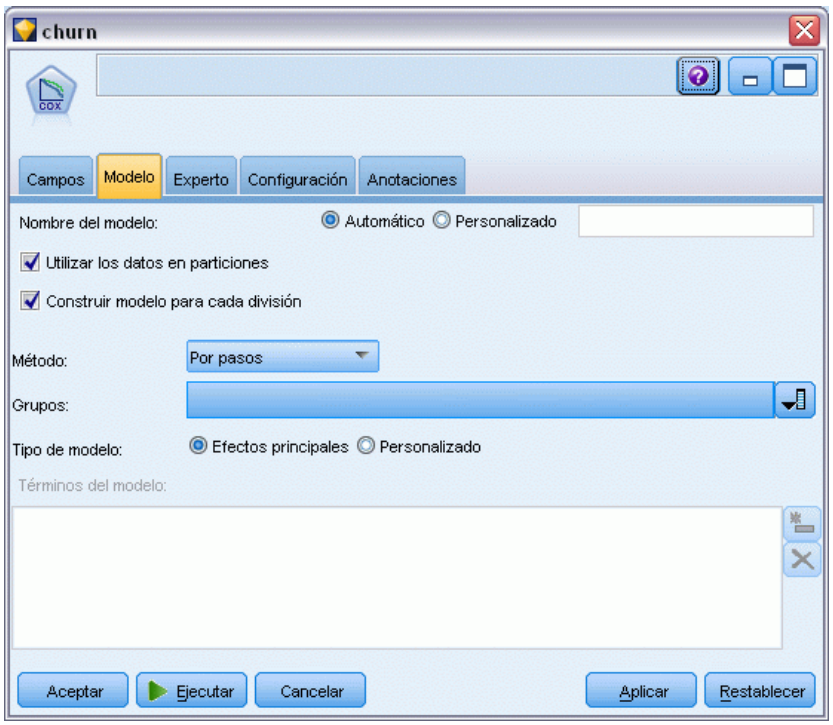
Figura 27-4
Selección de opciones de campo



- Pulse en la pestaña Modelo.

▶ Seleccione el método Por pasos como el método de selección de variables.

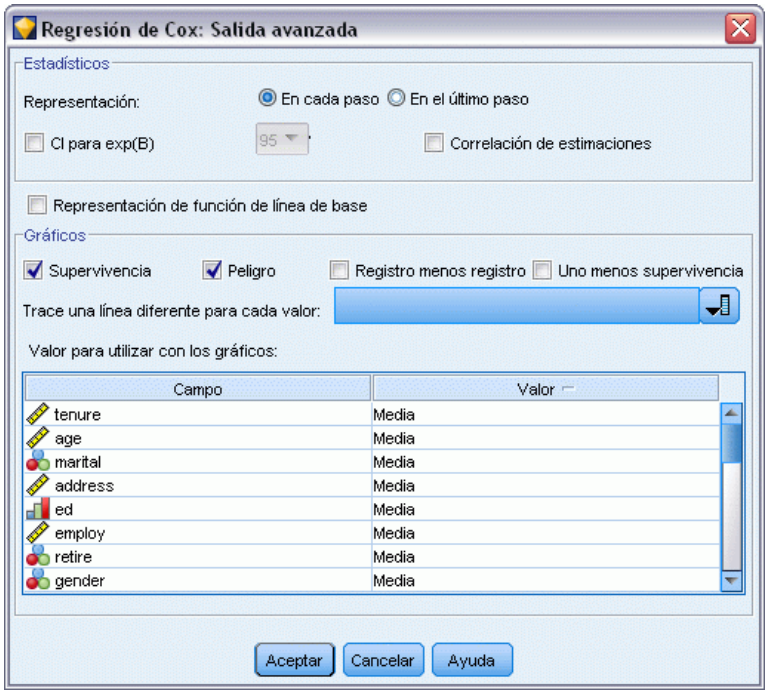
Figura 27-5
Selección de opciones del modelo



▶ Pulse en la pestaña Experto y seleccione Experto para activar las opciones de modelado experto.

▶ Pulse en Resultados.

Figura 27-6
Selección de opciones avanzadas de salida



▶ Seleccione Supervivencia y Peligro como los gráficos que se producirán y, a continuación, pulse en Aceptar.

▶ Pulse en Ejecutar para crear el nugget del modelo que se añadirá a la ruta y a la paleta Modelos en la esquina superior derecha. Para ver los detalles, pulse con el botón derecho del ratón en el nugget de la ruta. En primer lugar, observe la pestaña Resultado avanzado.

Casos censurados

Figura 27-7
Resumen del procesamiento de los casos

		N	Porcentaje
Casos disponibles en el análisis	**Evento(a)**	274	27,4%
	Censurado	726	72,6%
	Total	1000	100,0%
Casos excluidos	**Casos con valores perdidos**	0	,0%
	Casos con tiempo negativo	0	,0%
	Casos censurados antes del evento más temprano en un estrato	0	,0%
	Total	0	,0%
Total		1000	100,0%

a. Variable dependiente: Months with service

La variable de estado identifica si el evento se ha producido para un caso concreto. Si el evento no se ha producido, el caso se considera censurado. Los casos censurados no se utilizan en el cómputo de los coeficientes de regresión, pero se utilizan para calcular el peligro de línea base. El resumen de procesamiento de casos muestra que se han censurado 726 casos. Hay clientes que no han abandonado.

Iconos de variable categórica

Figura 27-8
Codificaciones de variable categórica

		Frecuencia	(1)(s)	(2)	(3)	(4)
marital(t)	0=Unmarried	505	1			
	1=Married	495	0			
ed(t)	1=Did not complete high school	204	1	0	0	0
	2=High school degree	287	0	1	0	0
	3=Some college	209	0	0	1	0
	4=College degree	234	0	0	0	1
	5=Post-undergraduate degree	66	0	0	0	0
retire(t)	0=No	953	1			
	1=Yes	47	0			
gender(t)	0=Male	483	1			
	1=Female	517	0			
tollfree(t)	0=No	526	1			
	1=Yes	474	0			
equip(t)	0=No	614	1			
	1=Yes	386	0			
callcard(t)	0=No	322	1			
	1=Yes	678	0			
wireless(t)	0=No	704	1			
	1=Yes	296	0			
multline(t)	0=No	525	1			
	1=Yes	475	0			
voice(t)	0=No	696	1			
	1=Yes	304	0			
pager(t)	0=No	739	1			
	1=Yes	261	0			
internet(t)	0=No	632	1			
	1=Yes	368	0			
callid(t)	0=No	519	1			
	1=Yes	481	0			
callwait(t)	0=No	515	1			
	1=Yes	485	0			
forward(t)	0=No	507	1			
	1=Yes	493	0			
confer(t)	0=No	498	1			
	1=Yes	502	0			
ebill(t)	0=No	629	1			
	1=Yes	371	0			
	1=Basic service	266	1	0	0	

Las codificaciones de variable categórica son una referencia de gran utilidad para interpretar los coeficientes de regresión de las covariables categóricas, especialmente las variables dicotómicas. Por defecto, la categoría de referencia es la "última" categoría. Además, por ejemplo, incluso si los clientes *Casados* tienen un valor de variable de 1 en el archivo de datos, se codifican como 0 para la regresión.

Selección de las variables

Figura 27-9
Contrastes Omnibus

Paso	-2 log de la verosimilitud	Global (puntuación)			Cambio desde el paso anterior			Cambio desde el bloque anterior		
		Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.
1(c)	3392,536	162,303	1	,000	133,828	1	,000	133,828	1	,000
2(d)	3087,314	249,392	2	,000	305,222	1	,000	439,050	2	,000
3(e)	3027,085	328,426	3	,000	60,229	1	,000	499,279	3	,000
4(f)	2990,790	347,197	4	,000	36,294	1	,000	535,574	4	,000
5(g)	2973,790	362,673	5	,000	17,000	1	,000	552,574	5	,000
6(h)	2958,796	376,140	6	,000	14,994	1	,000	567,568	6	,000
7(i)	2945,503	384,717	7	,000	13,293	1	,000	580,861	7	,000
8(j)	2936,993	417,341	8	,000	8,510	1	,004	589,371	8	,000
9(k)	2926,000	423,911	9	,000	10,994	1	,001	600,364	9	,000
10(l)	2917,551	428,078	10	,000	8,449	1	,004	608,813	10	,000
11(m)	2913,308	436,837	11	,000	4,243	1	,039	613,056	11	,000
12(n)	2908,078	440,158	12	,000	5,230	1	,022	618,286	12	,000

a. Bloque inicial número 0, función log de la verosimilitud inicial: -2 log de la verosimilitud: 3526,364

b. Bloque inicial número 1. Método = Por pasos hacia adelante (Razón de verosimilitud)

c. Variables introducidas en el paso número 1: callcard

d. Variables introducidas en el paso número 2: longmon

e. Variables introducidas en el paso número 3: equip

f. Variables introducidas en el paso número 4: employ

g. Variables introducidas en el paso número 5: multline

h. Variables introducidas en el paso número 6: voice

i. Variables introducidas en el paso número 7: address

j. Variables introducidas en el paso número 8: equipmon

k. Variables introducidas en el paso número 9: ebill

l. Variables introducidas en el paso número 10: callid

m. Variables introducidas en el paso número 11: internet

n. Variables introducidas en el paso número 12: reside

El proceso de creación de modelos utiliza un algoritmo de selección por pasos hacia adelante. Los contrastes omnibus son medidas de contrastes para comprobar la ejecución del modelo. El cambio del Chi-cuadrado del paso anterior es la diferencia entre el log-verosimilitud −2 del modelo del paso anterior y del paso actual. Si el paso consistía en agregar una variable, la inclusión tiene sentido si la significación del cambio es inferior a 0,05. Si el paso consistía en eliminar una variable, la exclusión tiene sentido si la significación del cambio es superior a 0,10. En doce pasos se agregan doce variables al modelo.

Figura 27-10
Variables en la ecuación (paso 12 únicamente)

		B	ET	Wald	gl	Sig.	Exp(B)
Paso 12	address	-,035	,009	14,543	1	,000	,966
	employ	-,051	,010	25,767	1	,000	,950
	reside	-,103	,046	5,037	1	,025	,902
	equip	-1,948	,381	26,180	1	,000	,143
	callcard	,777	,151	26,451	1	,000	2,175
	longmon	-,233	,022	115,619	1	,000	,792
	equipmon	-,042	,011	15,377	1	,000	,959
	multline	,612	,145	17,854	1	,000	1,844
	voice	-,501	,157	10,197	1	,001	,606
	internet	-,362	,160	5,114	1	,024	,697
	callid	-,464	,148	9,790	1	,002	,629
	ebill	-,399	,156	6,557	1	,010	,671

El modelo final incluye *dirección*, *empleo*, *residen*, *equipo*, *tarjetallamada*, *longmon*, *equipmon*, *multilínea*, *voz*, *internet*, *idllamada* y *efactura*. Para comprender el efecto de los predictores individuales, observe Exp(B), que se puede interpretar como el cambio pronosticado en el peligro para un aumento de unidades en el predictor.

- El valor de Exp(B) para *dirección* significa que el impacto de abandono es del 100%−(100%×0,966)=3,4% para cada año que un cliente ha vivido en la misma dirección. El impacto de abandono de un cliente que ha vivido en la misma dirección durante cinco años se reduce en un $100\%-(100\%\times 0{,}966^5)=15{,}88\%$.
- El valor de Exp(B) para *tarjetallamada* significa que el impacto de abandono de un cliente no suscrito al servicio de tarjeta de llamada es 2,175 veces más que un cliente con el servicio. Recuerde que para las codificaciones de variable categórica *No* = 1 para la regresión.
- El valor de Exp(B) para *internet* significa que el impacto de abandono de un cliente no suscrito al servicio de Internet es 0,697 veces más que un cliente con el servicio. Es un indicativo preocupante, ya que sugiere que los clientes con el servicio abandonan la compañía antes que los clientes sin el servicio.

Figura 27-11
Variables no incluidas en el modelo (paso 12 únicamente)

		Puntuación	gl	Sig.
Paso 12	age	,122	1	,726
	marital	,648	1	,421
	ed	6,328	4	,176
	ed(1)	,007	1	,934
	ed(2)	,203	1	,652
	ed(3)	,835	1	,361
	ed(4)	5,773	1	,016
	retire	,013	1	,908
	gender	,214	1	,644
	tollfree	3,243	1	,072
	wireless	,668	1	,414
	tollmon	,000	1	,987
	cardmon	3,163	1	,075
	wiremon	1,084	1	,298
	pager	1,808	1	,179
	callwait	,266	1	,606
	forward	2,201	1	,138
	confer	2,568	1	,109
	lninc	2,853	1	,091
	custcat	,864	3	,834
	custcat(1)	,466	1	,495
	custcat(2)	,450	1	,502
	custcat(3)	,019	1	,889

Todas las variables no incluidas en el modelo tienen estadísticos de puntuación con valores de significación superiores a 0,05. Sin embargo, los valores de significación de *numgratuito* y *cardmon*, son muy cercanos, mientras no sean inferiores a 0,05. Puede ser interesante su inclusión en otros estudios.

Medias de covariables

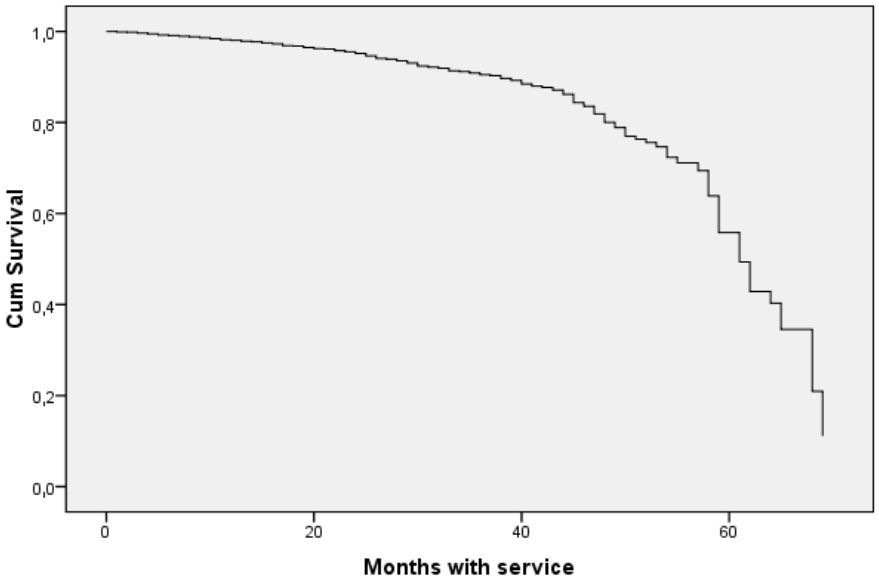
Figura 27-12
Medias de covariables

	Media
age	41,684
marital	,505
address	11,551
ed(1)	,204
ed(2)	,287
ed(3)	,209
ed(4)	,234
employ	10,987
retire	,953
gender	,483
reside	2,331
tollfree	,526
equip	,614
callcard	,322
wireless	,704
longmon	11,723
tollmon	13,274
equipmon	14,220
cardmon	13,781
wiremon	11,584
multline	,525
voice	,696
pager	,739
internet	,632
callid	,519
callwait	,515
forward	,507
confer	,498
ebill	,629
lninc	3,957
custcat(1)	,266
custcat(2)	,217
custcat(3)	,281

Esta tabla muestra el valor medio de cada variable de predictor. Esta tabla es una referencia de gran utilidad si observa gráficos de supervivencia, que se generan para los valores medios. Tenga en cuenta, sin embargo, que el cliente "promedio" no existe realmente cuando observa las medias de las variables del indicador de los predictores categóricos. Incluso con todos los predictores de escala, es poco probable que encuentre un cliente cuyos valores de covariable sean cercanos a la media. Si desea ver la curva de supervivencia de un caso concreto, puede cambiar los valores de covariable donde la curva de supervivencia se traza en el cuadro de diálogo Gráficos. Si desea ver la curva de supervivencia de un caso concreto, puede cambiar los valores de covariable donde la curva de supervivencia se traza en el grupo de gráficos del cuadro de diálogo Resultado avanzado.

Curva de supervivencia

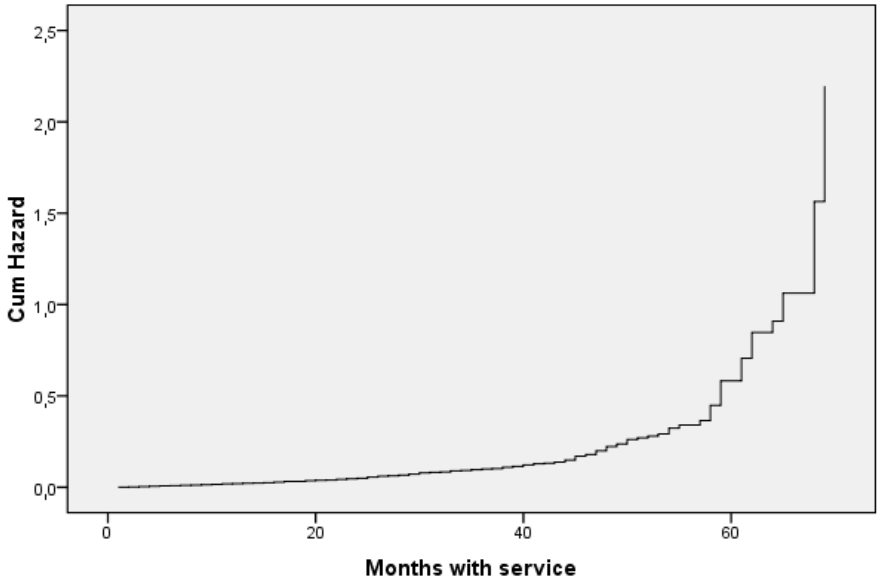
Figura 27-13
Curva de supervivencia de cliente "promedio"



La curva de supervivencia básica es una visualización del tiempo de abandono del cliente "promedio" pronosticado por el modelo. El eje horizontal muestra la hora del evento. El eje vertical muestra la probabilidad de supervivencia. Además, cualquier punto de la curva de supervivencia muestra la probabilidad de que el cliente "promedio" siga siendo un cliente después de ese tiempo. Tras 55 meses, la curva de supervivencia es menos suave. Hay menos clientes que han permanecido tanto tiempo en la compañía, por lo que hay menos información disponible y la curva tiene forma de bloque.

Curva de impacto

Figura 27-14
Curva de impacto de cliente "promedio"

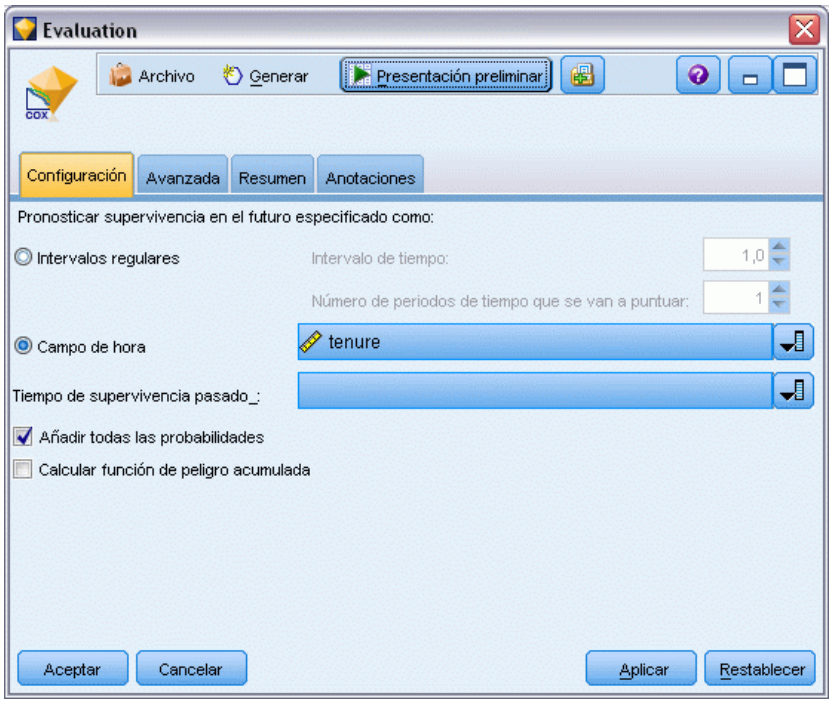


La curva de impacto básica es una visualización del potencial acumulado de abandono del cliente "promedio" pronosticado por el modelo. El eje horizontal muestra la hora del evento. El eje vertical muestra el impacto acumulado, igual al logaritmo negativo de la probabilidad de supervivencia. Transcurridos 55 meses, la curva de impacto, como la curva de supervivencia, es menos suave por la misma razón.

Evaluación

Los métodos de selección por pasos garantizan que su modelo sólo contendrá predictores "estadísticamente significativos", pero no garantizan que el modelo realice buenos pronósticos. Para ello, debe volver a analizar los registros puntuados.

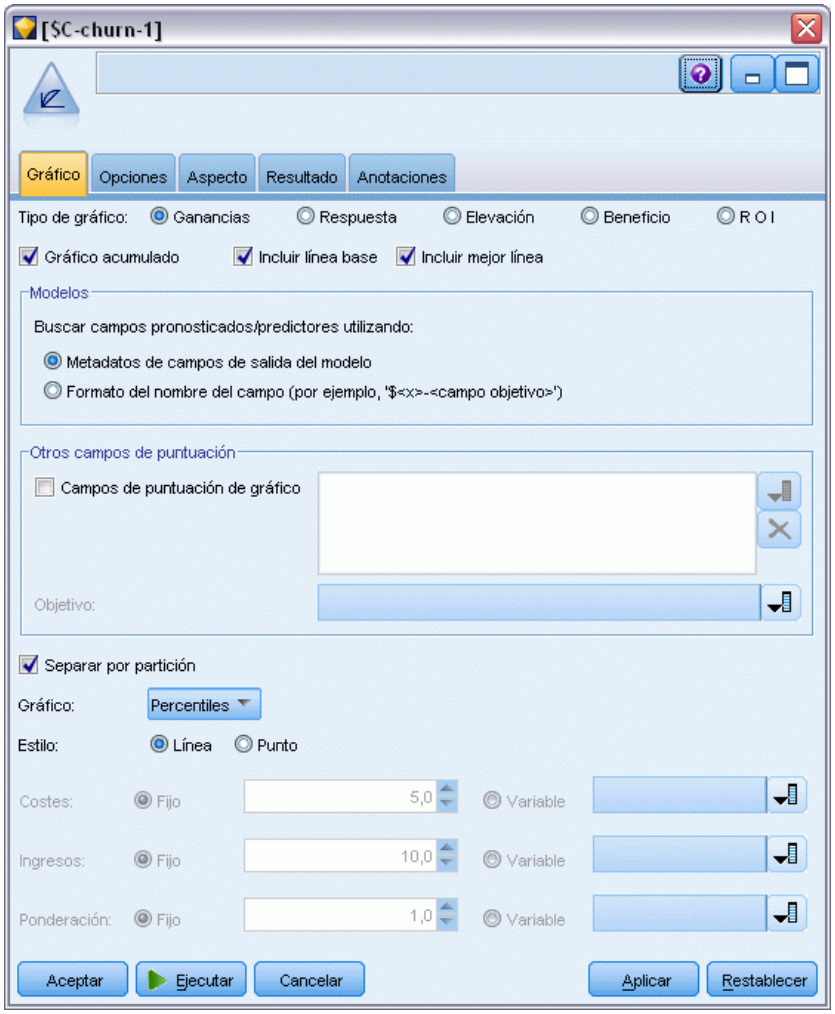
Figura 27-15
Nugget de Cox: Pestaña Configuración



- Coloque el nugget de modelo en el lienzo y adjúntelo en el nodo de origen, abra el nugget y pulse en la pestaña Configuración.
- Seleccione el campo Tiempo y especifique el *periodo*. Cada registro se puntuará en función de la longitud de su periodo.
- Seleccione Añadir todas las probabilidades.

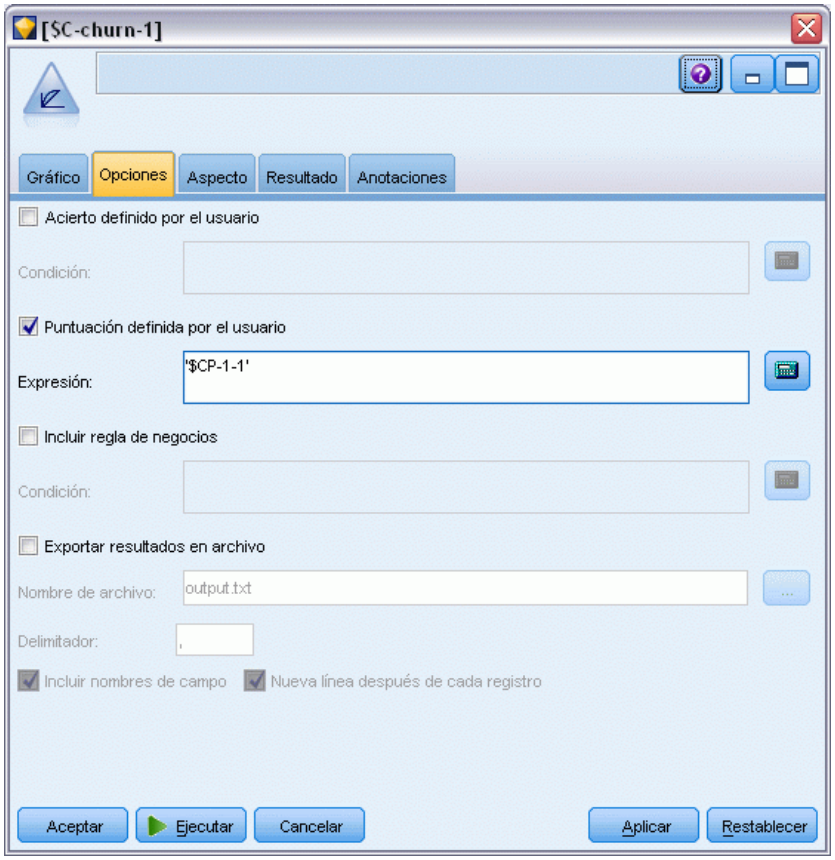
Crea puntuaciones utilizando 0,5 como el corte de abandono de cliente; si su propensión de abandono es superior a 0,5, se puntúan como abandono. No hay nada mágico en este número y se puede definir un corte diferente para obtener resultados más deseables. Para poder seleccionar un corte, utilice un nodo Evaluación.

Figura 27-16
nodo Evaluación: Pestaña Gráfico



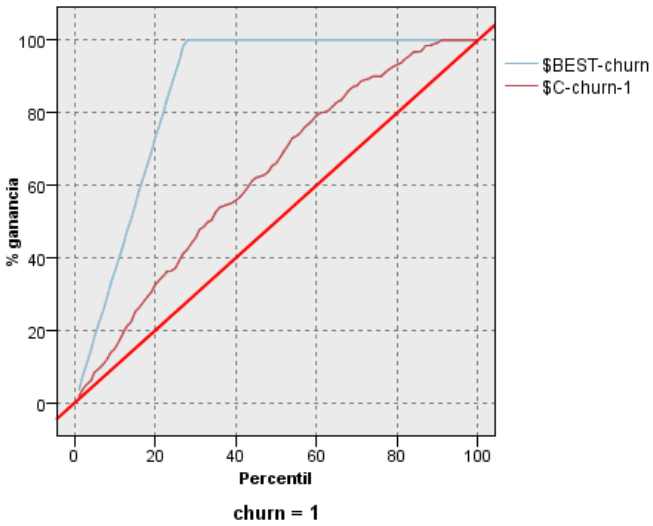
- Añada un nodo Evaluación al nugget de modelo; en la pestaña Gráfico, seleccione Incluir mejor línea.
- Pulse en la pestaña Opciones.

Figura 27-17
nodo Evaluación: Pestaña Opciones



- Seleccione Puntuación definida por el usuario e introduzca "$CP-1-1" como la expresión. Es un campo generado por el modelo que se corresponde con la propensión de abandono.
- Pulse en Ejecutar.

Figura 27-18
Gráfico de ganancias



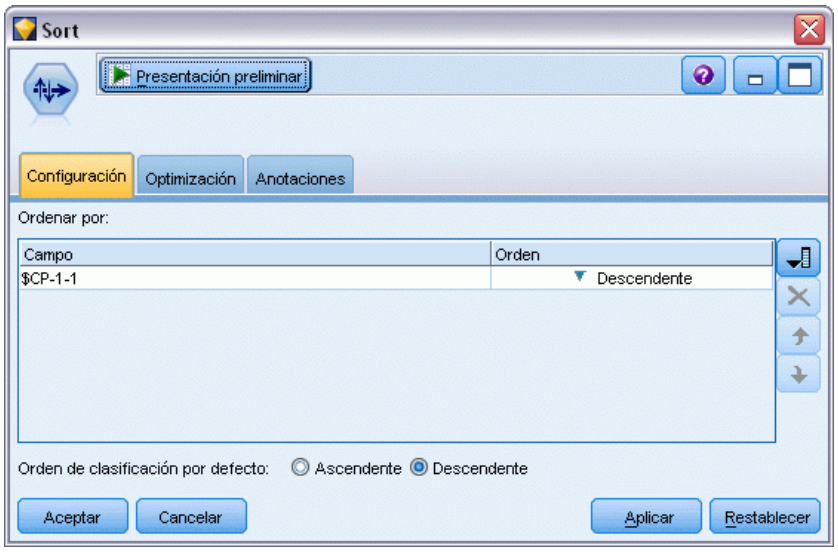
El gráfico de ganancias acumuladas muestra el porcentaje del número total de casos de una categoría dada "ganada" al dirigirse a un porcentaje del número total de casos. Por ejemplo, un punto de la curva está en (10%, 15%), lo que significa que si puntúa un conjunto de datos con el modelo y ordena todos los casos por su propensión pronosticada de abandono, debería esperar que el 10% principal contenga aproximadamente el 15% de todos los casos en la categoría *1* (usuarios que abandonan). Del mismo modo, el 60% contiene aproximadamente el 79,2% de los usuarios que abandonan. Si selecciona el 100% del conjunto de datos puntuados, obtendrá todos los usuarios que abandonan en el conjunto de datos.

La línea diagonal es la curva de "nivel básico"; si selecciona el 20% de los registros del conjunto de datos puntuados de forma aleatoria, debería esperar "ganar" aproximadamente el 20% de todos los registros de la categoría *1*. Cuanto más arriba está la curva de línea base, mayor es la ganancia. La "mejor línea" muestra la curva de un modelo "perfecto" que asigna una mayor puntuación de propensión de abandono a cada usuario que abandona que a los usuarios que no abandonan. Puede usar el gráfico de ganancias acumuladas para seleccionar un corte de clasificación al seleccionar un porcentaje que corresponde a una ganancia deseada y, a continuación, asignar ese porcentaje al valor de corte adecuado.

La definición de ganancia "deseada" depende del coste de los errores de Tipo I y Tipo II. Es decir, ¿cuál es el coste de clasificar un usuario que abandona como un usuario que no abandona (Tipo I)? ¿Cuál es el coste de clasificar un usuario que no abandona como un usuario que abandona (Tipo II)? Si la retención de clientes es la preocupación principal, es posible que desee reducir el error de tipo I; en el gráfico de ganancias acumuladas, puede corresponder con un servicio de atención al cliente mejorado en el 60% principal de propensión pronosticada de *1*, que incluye el 79,2% de los posibles usuarios que abandonan que consumen tiempo y recursos que se pueden emplear en nuevos clientes. Si la prioridad es reducir el coste de mantener su base de clientes actual, es posible que desee reducir su error de tipo II. En el gráfico, puede corresponder al aumento del servicio de atención al cliente para el 20% principal, que incluye al 32,5% de los usuarios que abandonan. Normalmente, ambas son cuestiones importantes, así

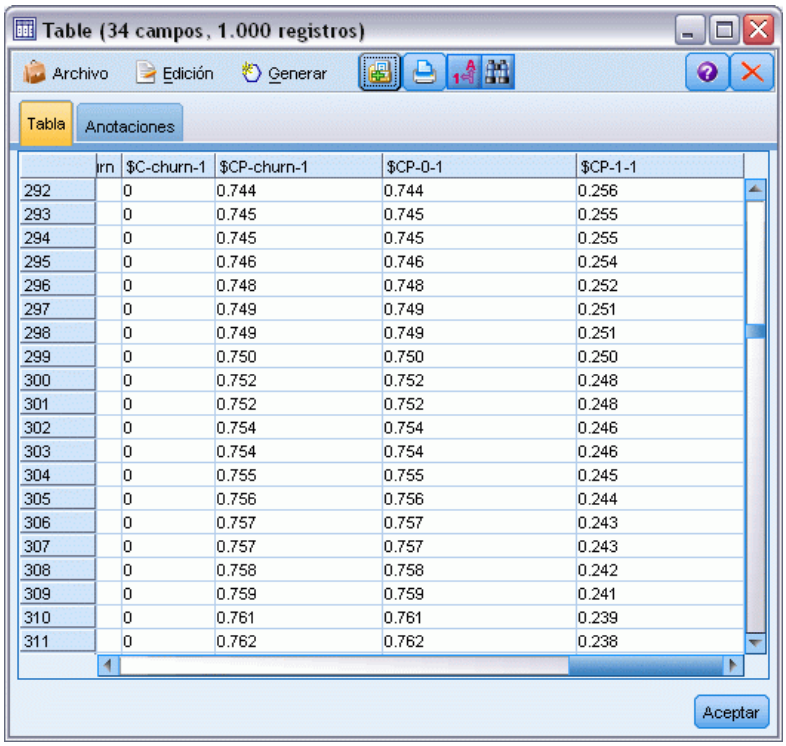
que se deberá elegir una regla de decisión para clasificar los clientes que ofrezcan la mejor combinación de susceptibilidad y especificidad.

Figura 27-19
Nodo Ordenar: Pestaña Configuración



- Por ejemplo, ha decidido que el 45,6% es una ganancia deseable, que se corresponde a tomar el 30% principal de los registros. Para buscar una clasificación adecuada, añada un nodo Ordenar al nugget de modelo.
- En la pestaña Configuración, seleccione clasificar *$CP-1-1* en orden descendente y pulse en Aceptar.

Figura 27-20
Tabla



	irn	$C-churn-1	$CP-churn-1	$CP-0-1	$CP-1-1
292		0	0.744	0.744	0.256
293		0	0.745	0.745	0.255
294		0	0.745	0.745	0.255
295		0	0.746	0.746	0.254
296		0	0.748	0.748	0.252
297		0	0.749	0.749	0.251
298		0	0.749	0.749	0.251
299		0	0.750	0.750	0.250
300		0	0.752	0.752	0.248
301		0	0.752	0.752	0.248
302		0	0.754	0.754	0.246
303		0	0.754	0.754	0.246
304		0	0.755	0.755	0.245
305		0	0.756	0.756	0.244
306		0	0.757	0.757	0.243
307		0	0.757	0.757	0.243
308		0	0.758	0.758	0.242
309		0	0.759	0.759	0.241
310		0	0.761	0.761	0.239
311		0	0.762	0.762	0.238

- Conecte un nodo Tabla al nodo Clasificar.
- Abra el nodo Tabla y pulse en Ejecutar.

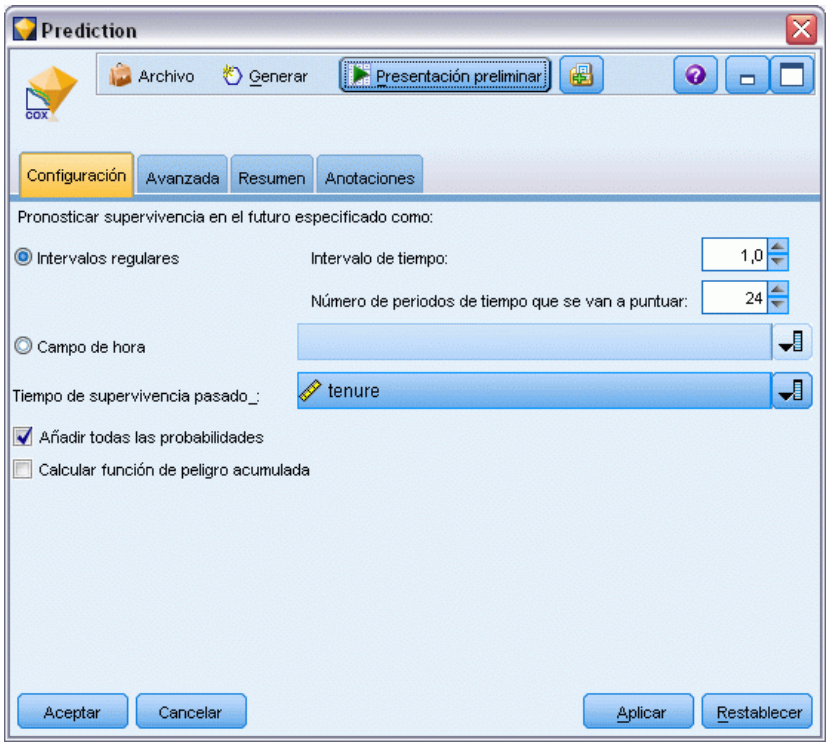
Si analiza los resultados, verá que el valor *$CP-1-1* es 0,248 en el registro número 300. Si utiliza 0,248 como corte de clasificación obtendrá como resultado que aproximadamente el 30% de los clientes se clasifican como usuarios que abandonan, incluyendo aproximadamente el 45% del total de los usuarios que abandonan.

Seguimiento del número de clientes mantenidos esperados

Cuando esté satisfecho con un modelo, es posible que desee realizar el seguimiento del número esperado de clientes en el conjunto de datos que se mantienen en los dos siguientes años. Los valores nulos, que son clientes cuyo periodo total (tiempo futuro + *periodo*) están dentro del intervalo de horas de supervivencia en el conjunto de datos utilizado para entrenar el modelo, son un dato interesante. Una forma de trabajar con ellos es crear dos conjuntos de pronósticos, uno cuyos valores nulos se consideran clientes que abandonan y otro que se consideran mantenidos.

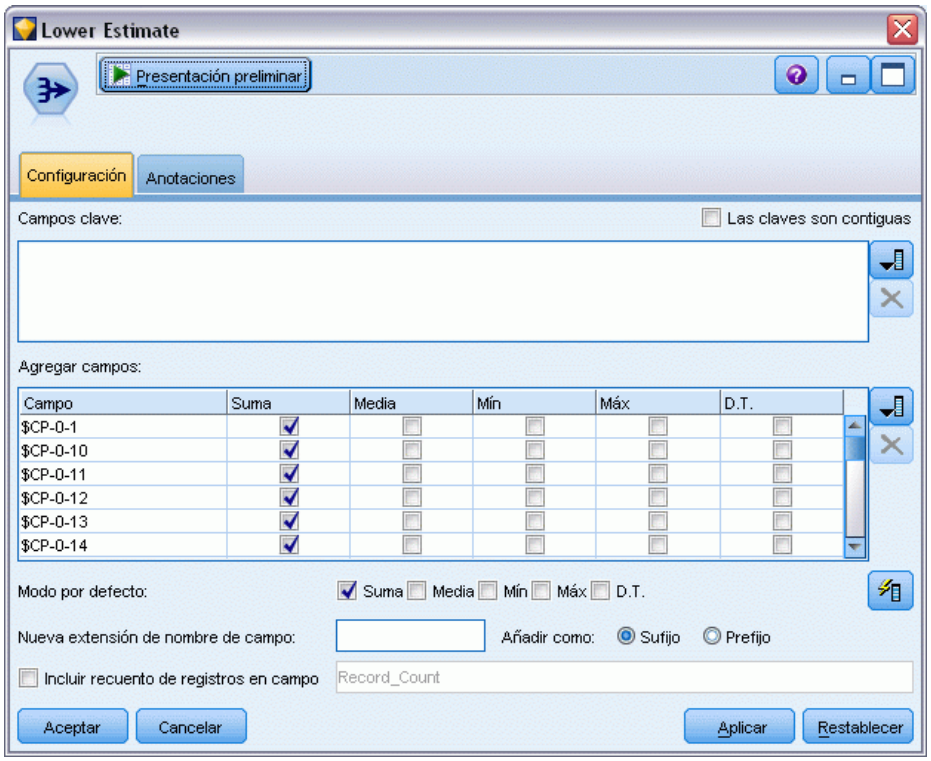
De esta forma puede establecer los límites superiores e inferiores del número de clientes mantenidos esperado.

Figura 27-21
Nugget de Cox: Pestaña Configuración



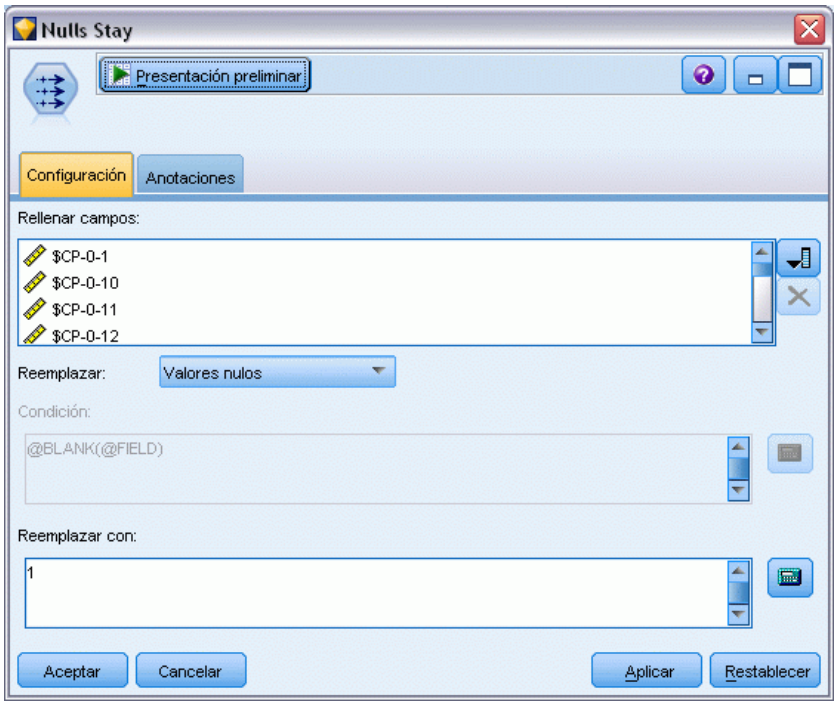
- Pulse dos veces en el nugget del modelo en la paleta Modelos (o copie y pegue el nugget en el lienzo de rutas) y conecte el nuevo nugget al nodo Origen.
- Abra el nugget en la pestaña Configuración.
- Asegúrese de que ha seleccionado Intervalos regulares e introduzca 1.0 como el intervalo de tiempo y 24 como el número de periodos que se van a puntuar. Indica que cada registro se puntuará los siguientes 24 meses.
- Seleccione *periodo* como el campo para especificar el tiempo de supervivencia anterior. El algoritmo de puntuación tendrá en cuenta la permanencia de cada usuario como cliente de la compañía.
- Seleccione Añadir todas las probabilidades.

Figura 27-22
Nodo Agregar: Pestaña Configuración



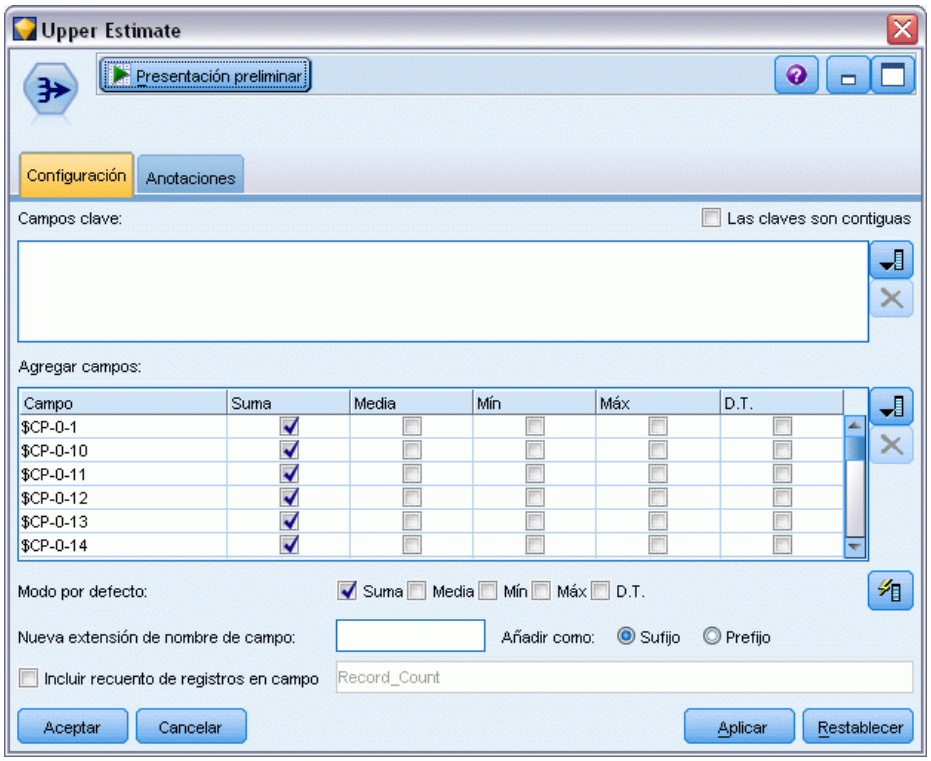
- Añada un nodo Agregar al nugget de modelo. En la pestaña Configuración cancele la selección de Media como el modo predefinido.
- Seleccione *$CP-0-1* a *$CP-0-24*, los campos de forma *$CP-0-n*, como los campos que se van a agregar. Es la forma más simple, si, en el cuadro de diálogo Seleccionar campos, ordena los campos por nombre (es decir, por orden alfabético).
- Cancele la selección de Incluir recuento de registros en campo.
- Pulse en Aceptar. Este nodo crea las predicciones “límite inferior”.

Figura 27-23
Nodo Rellenar: Pestaña Configuración



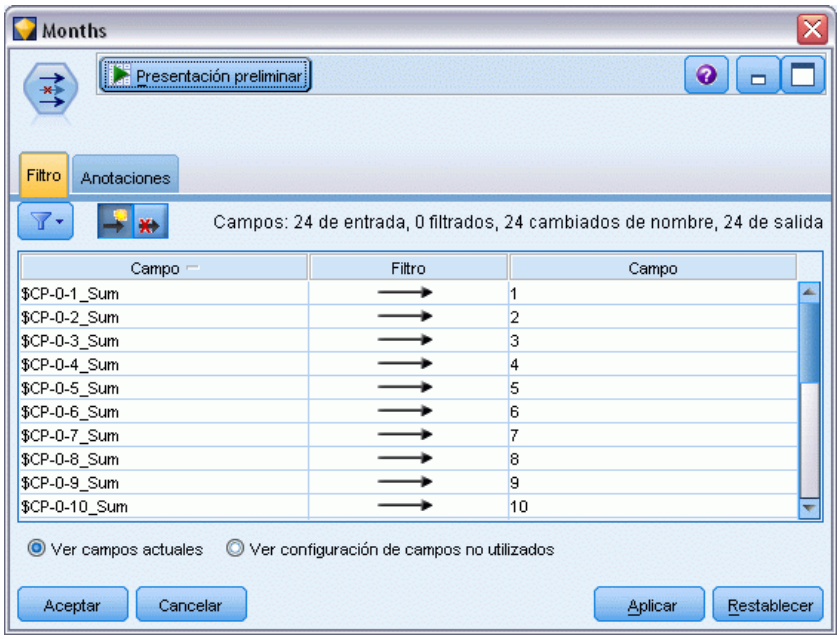
- Añada un nodo Rellenar al nugget Coxreg al que ha agregado el nodo Agregar. En la pestaña Configuración, seleccione *$CP-0-1* a *$CP-0-24*, los campos con forma *$CP-0-n*, como los campos que se van a rellenar. Es la forma más simple, si, en el cuadro de diálogo Seleccionar campos, ordena los campos por nombre (es decir, por orden alfabético).
- Sustituya Valores nulos por 1.
- Pulse en Aceptar.

Figura 27-24
Nodo Agregar: Pestaña Configuración



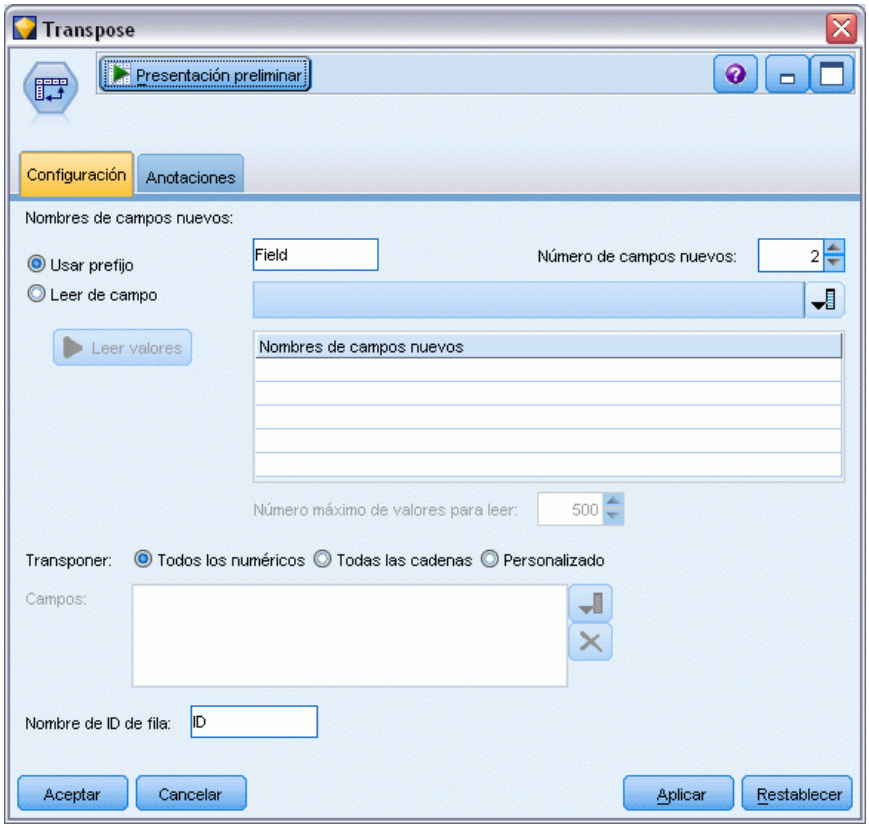
- Añada un nodo Agregar al nodo Rellenar. En la pestaña Configuración cancele la selección de Media como el modo predefinido.
- Seleccione *$CP-0-1* a *$CP-0-24*, los campos de forma *$CP-0-n*, como los campos que se van a agregar. Es la forma más simple, si, en el cuadro de diálogo Seleccionar campos, ordena los campos por nombre (es decir, por orden alfabético).
- Cancele la selección de Incluir recuento de registros en campo.
- Pulse en Aceptar. Este nodo crea las predicciones “límite superior”.

Figura 27-25
Nodo Filtro: Pestaña Configuración



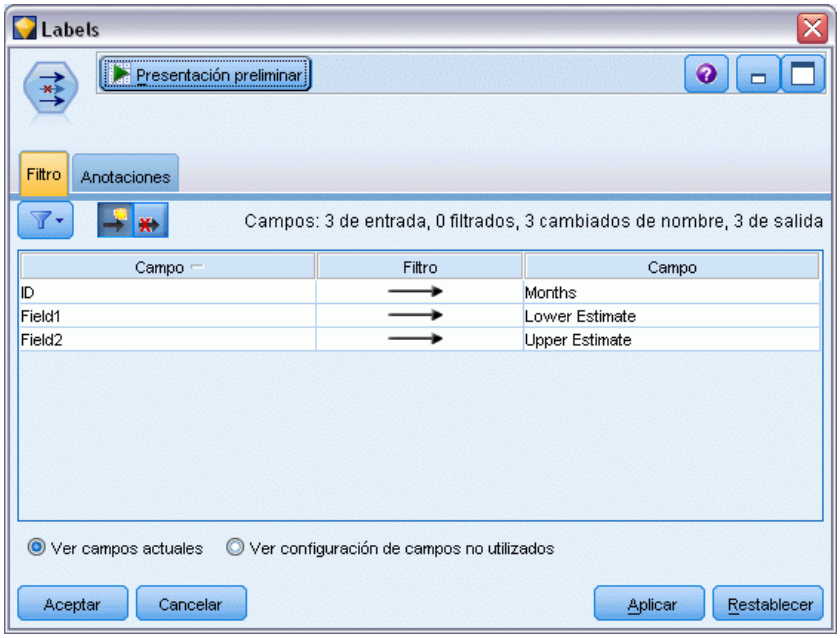
- Añada un nodo Añadir a los dos nodos Agregar y añada el nodo Filtro al nodo Añadir.
- En la pestaña Configuración del nodo Filtro, cambie el nombre de los campos *1* a *24*. Mediante un nodo Transponer, los nombres de estos campos serán los valores del eje *x* en gráficos hacia abajo.

Figura 27-26
Nodo Transponer: Pestaña Configuración



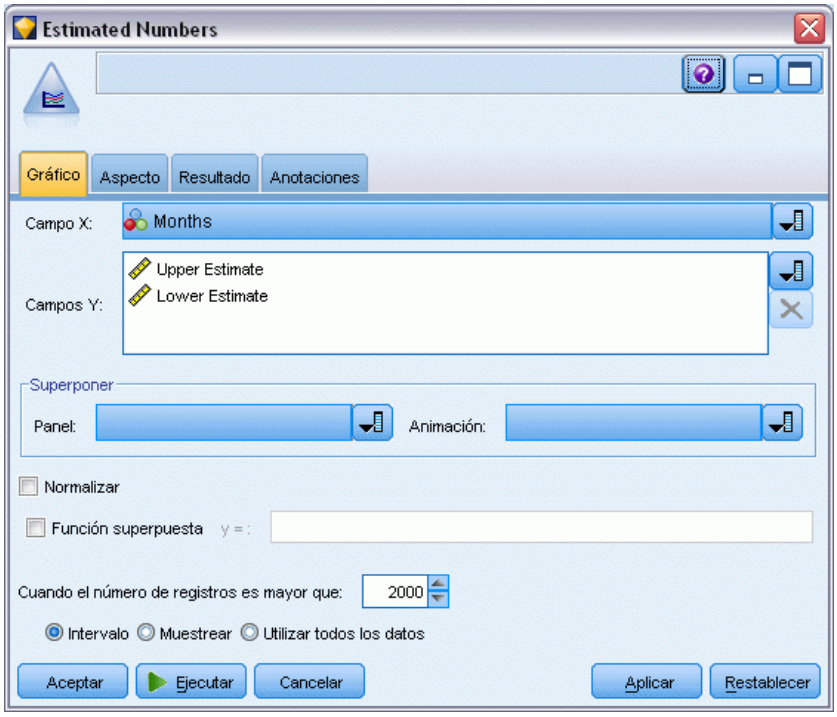
- Añada un nodo Transponer al nodo Filtro.
- Escriba 2 como el número de nuevos campos.

Figura 27-27
Nodo Filtro: Pestaña Filtro



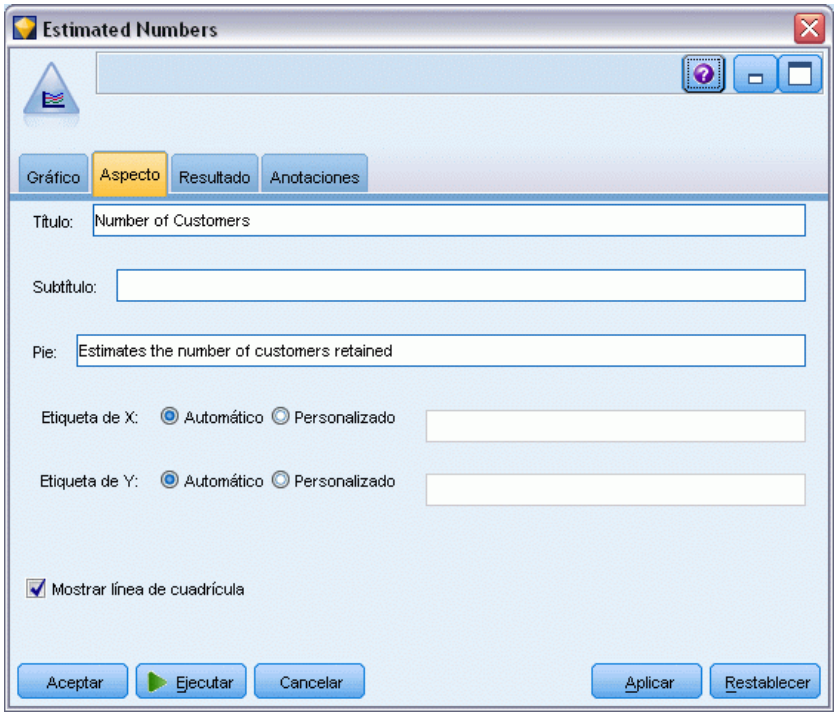
- Añada un nodo Filtro al nodo Transponer.
- En la pestaña Configuración del nodo Filtro, cambie el nombre de *ID* a *Meses*, *Campo1* a *Estimación inferior* y *Campo2* a *Estimación superior*.

Figura 27-28
Nodo G. múltiple: Pestaña Gráfico



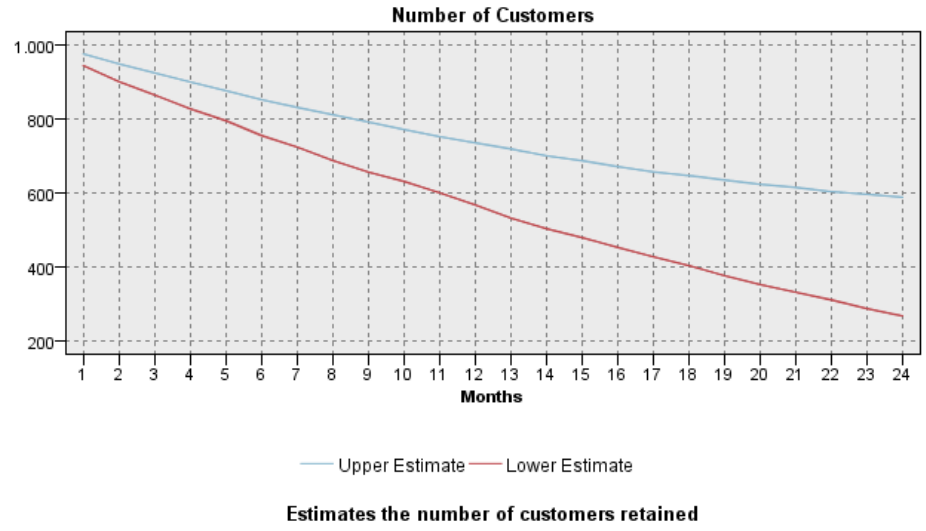
- Añada un nodo G. múltiple al nodo Filtro.
- En la pestaña Gráfico, defina *Meses* como el campo X, *Estimación inferior* y *Estimación superior* como el campo Y.

Figura 27-29
Nodo G. múltiple: Pestaña Aspecto



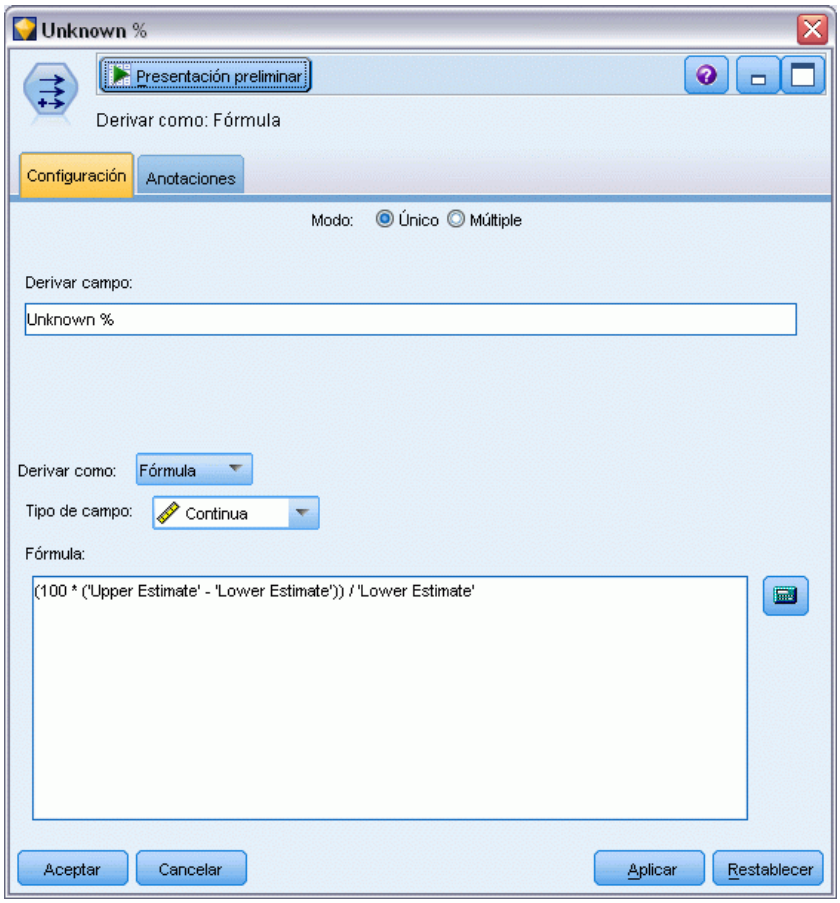
- Pulse en la pestaña Aspecto.
- Introduzca Número de clientes como el título.
- Introduzca Estimaciones del número de clientes mantenidos como captura.
- Pulse en Ejecutar.

Figura 27-30
Gráfico múltiple calculando el número de clientes mantenidos



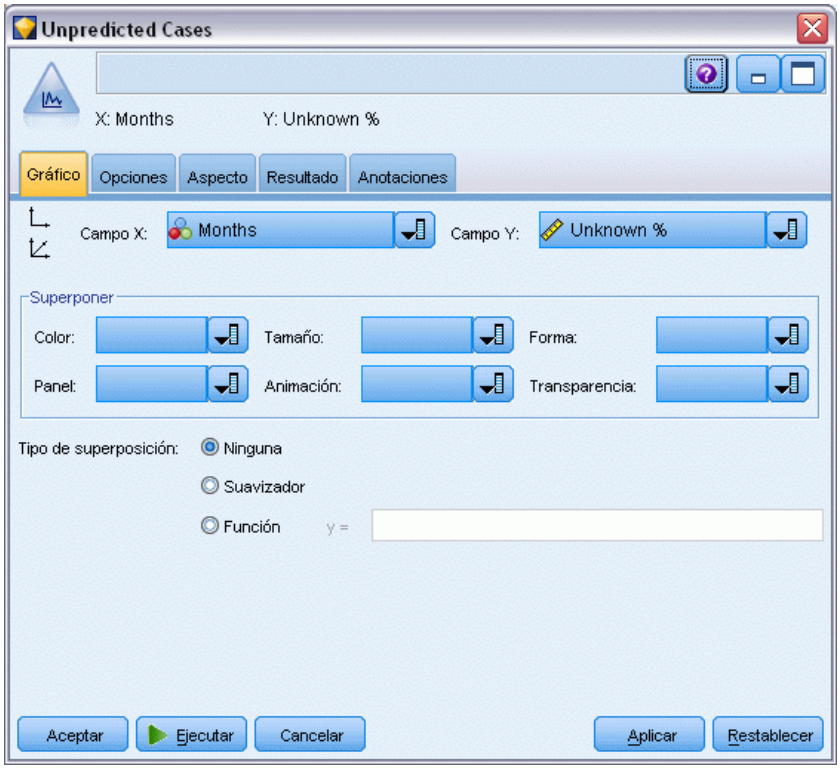
Se trazan los límites superiores e inferiores del número de clientes mantenidos estimados. La diferencia entre las dos líneas es el número de clientes puntuados como nulos, y, por lo tanto, cuyo estado es incierto. Con el tiempo se aumentará el número de estos clientes. Tras 12 meses, puede esperar retener entre 601 y 735 de los clientes originales del conjunto de datos y después de 24 meses, entre 288 y 597.

Figura 27-31
Nodo Derivar: Pestaña Configuración



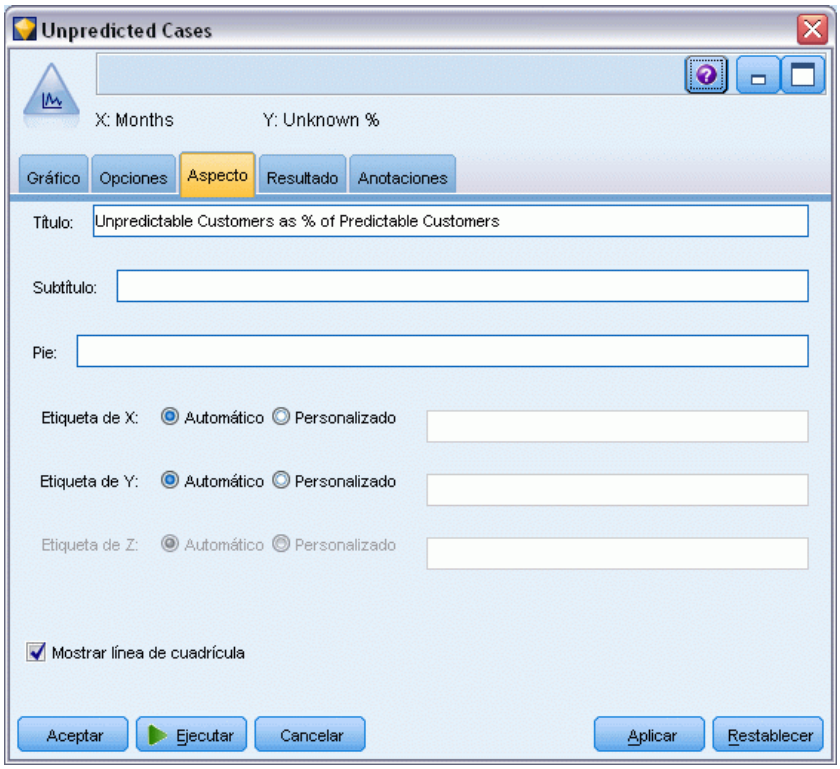
- Para ver otra forma de comprobar la inexactitud de las estimaciones del número de clientes que se retienen, añada un nodo Derivar al nodo Filtrar.
- En la pestaña Configuración del nodo Derivar, introduzca *Desconocido %* como el campo de derivación.
- Seleccione Continuo como el tipo de campo.
- Introduzca (100 * ("Estimación superior" - "Estimación inferior")) / "Estimación inferior" como fórmula. *Desconocido %* es el número de clientes “dudosos” como porcentaje de la estimación inferior.
- Pulse en Aceptar.

Figura 27-32
Nodo Gráfico: Pestaña Gráfico



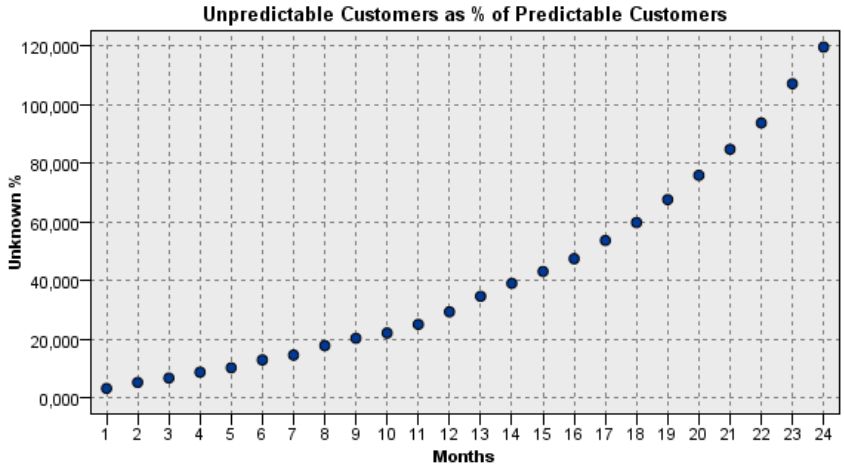
- Añada un nodo Gráfico al nodo Derivar.
- En la pestaña Gráfico del nodo Gráfico, seleccione *Meses* como el campo X y *Desconocido %* como el campo Y.
- Pulse en la pestaña Aspecto.

Figura 27-33
Nodo Gráfico: Pestaña Aspecto



- Introduzca Clientes impredecibles como % de clientes predecibles como título.
- Ejecute el nodo.

Figura 27-34
Gráfico de clientes impredecibles

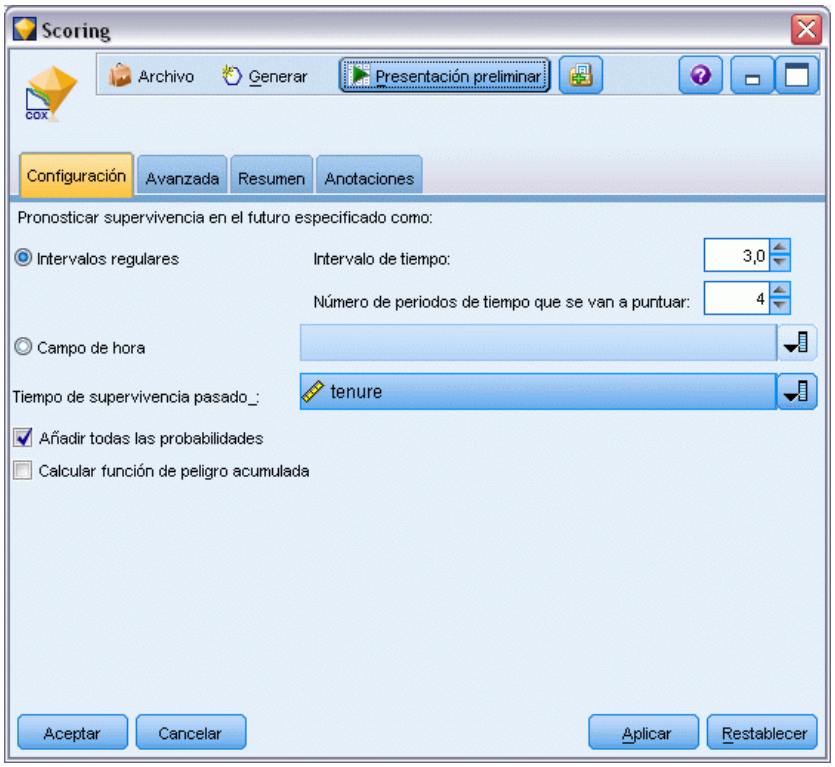


En el primer año, el porcentaje de clientes impredecibles aumenta en una proporción lineal, pero el porcentaje aumenta durante el segundo año, hasta el mes 23, en el que el número de clientes con valores nulos sobrepasa el número esperado de clientes mantenidos.

Puntuación

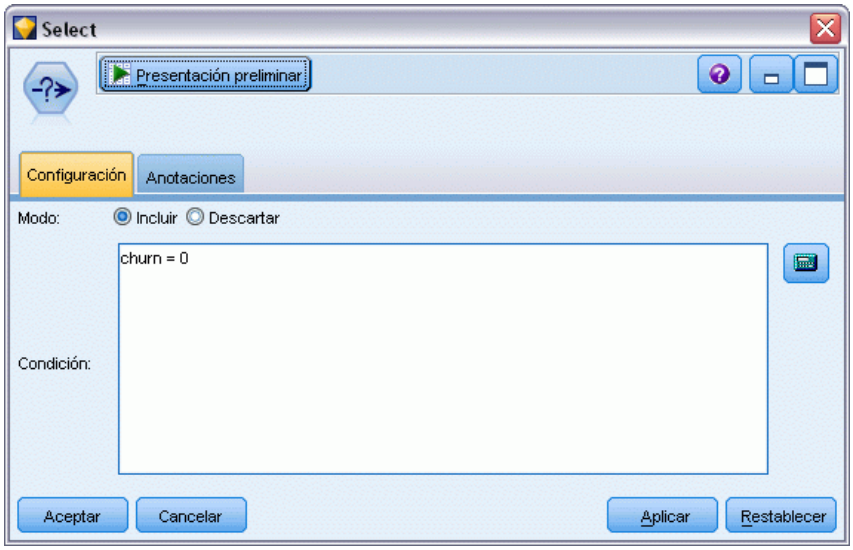
Una vez satisfecho con el modelo, es posible que desee puntuar los clientes para identificar los individuos con mayor probabilidad de abandono el año siguiente, por trimestre.

Figura 27-35
Nugget Coxreg: Pestaña Configuración



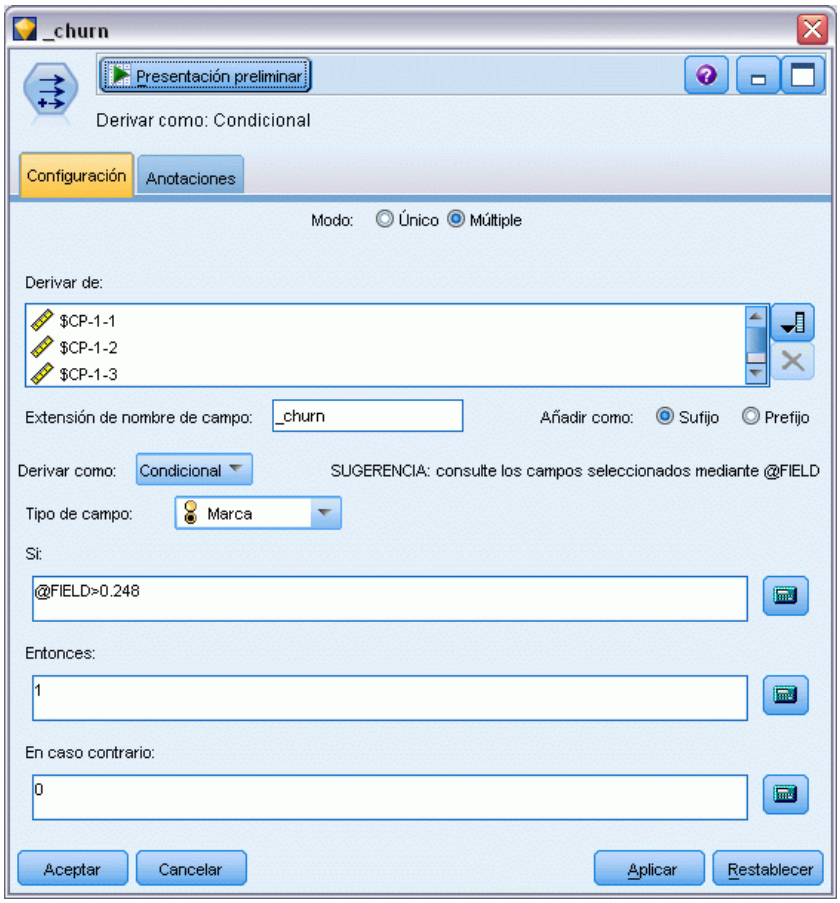
- Añada un tercer modelo al nodo Origen y abra el nugget de modelo.
- Asegúrese de que ha seleccionado Intervalos regulares e introduzca 3,0 como el intervalo de tiempo y 4 como el número de periodos que se van a puntuar. Indique que cada registro se puntuará los siguientes 4 trimestres.
- Seleccione *periodo* como el campo para especificar el tiempo de supervivencia anterior. El algoritmo de puntuación tendrá en cuenta la permanencia de cada usuario como cliente de la compañía.
- Seleccione Añadir todas las probabilidades. Estos campos extra facilitan clasificar los registros para ver una tabla.

Figura 27-36
Nodo Seleccionar: Pestaña Configuración



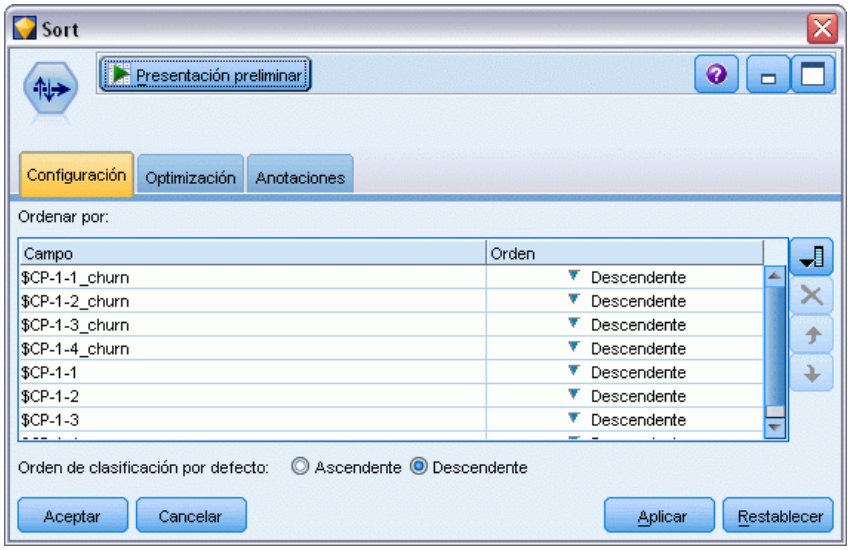
- Añada un nodo Seleccionar al nugget del modelo; en la pestaña Configuración, introduzca abandono=0 como condición. Los clientes que hayan abandonado se eliminarán de la tabla.

Figura 27-37
Nodo Derivar: Pestaña Configuración



- Añada un nodo Derivar al nodo Seleccionar; en la pestaña Configuración, seleccione Múltiple como el modo.
- Derive de *$CP-1-1* a *$CP-1-4*, los campos de forma *$CP-1-n* e introduzca _abandono como sufijo. Es la forma más simple, si, en el cuadro de diálogo Seleccionar campos, ordena los campos por nombre (es decir, por orden alfabético).
- Seleccione derivar el campo como Condicional.
- Seleccione Marca como nivel de medición.
- Introduzca @FIELD>0,248 como condición Si. Recuerde que este fue el primer corte de clasificación identificado durante la evaluación.
- Introduzca 1 como expresión Entonces.
- Introduzca 0 como expresión En caso contrario.
- Pulse en Aceptar.

Figura 27-38
Nodo Ordenar: Pestaña Configuración



- Añada un nodo Ordenar al nodo Derivar. En la pestaña Configuración, seleccione clasificar por *$CP-1-1_abandono* a *$CP-1-4-abandono* y *$CP-1-1* a *$CP-1-4*, en orden descendente. Los clientes pronosticados como abandono aparecerán al principio.

Figura 27-39
Nodo Reorg. campos: Pestaña Reordenar

- Añada un nodo Reorg. campos al nodo Ordenar. En la pestaña Reordenar, coloque *$CP-1-1_abandono* a *$CP-1-4* delante del resto de los campos. Simplemente facilita la lectura de

la tabla de resultados y es opcional. Necesitará utilizar los botones para mover los campos en la posición que aparece en la figura.

Figura 27-40
Tabla con puntuaciones de clientes

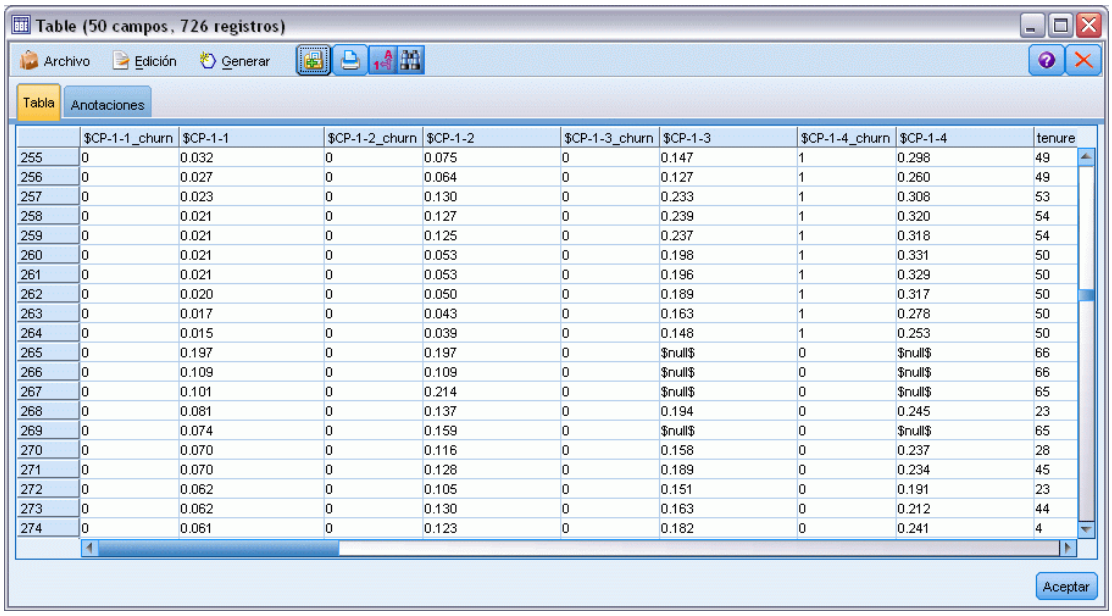


Table (50 campos, 726 registros)

Archivo Edición Generar

Tabla Anotaciones

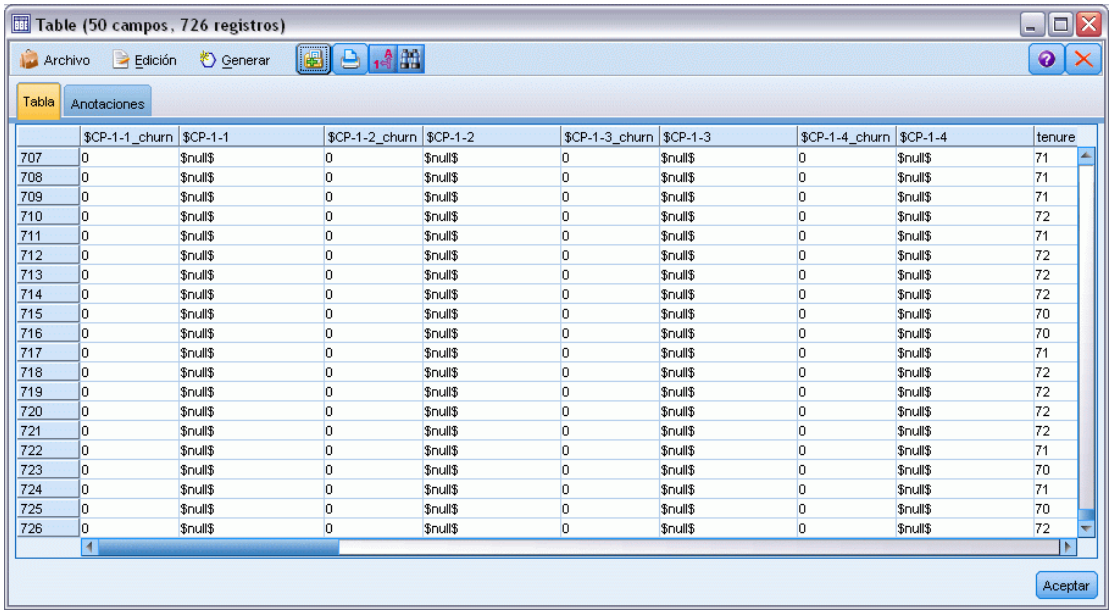
	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenure
255	0	0.032	0	0.075	0	0.147	1	0.298	49
256	0	0.027	0	0.064	0	0.127	1	0.260	49
257	0	0.023	0	0.130	0	0.233	1	0.308	53
258	0	0.021	0	0.127	0	0.239	1	0.320	54
259	0	0.021	0	0.125	0	0.237	1	0.318	54
260	0	0.021	0	0.053	0	0.198	1	0.331	50
261	0	0.021	0	0.053	0	0.196	1	0.329	50
262	0	0.020	0	0.050	0	0.189	1	0.317	50
263	0	0.017	0	0.043	0	0.163	1	0.278	50
264	0	0.015	0	0.039	0	0.148	1	0.253	50
265	0	0.197	0	0.197	0	$null$	0	$null$	66
266	0	0.109	0	0.109	0	$null$	0	$null$	66
267	0	0.101	0	0.214	0	$null$	0	$null$	65
268	0	0.081	0	0.137	0	0.194	0	0.245	23
269	0	0.074	0	0.159	0	$null$	0	$null$	65
270	0	0.070	0	0.116	0	0.158	0	0.237	28
271	0	0.070	0	0.128	0	0.189	0	0.234	45
272	0	0.062	0	0.105	0	0.151	0	0.191	23
273	0	0.062	0	0.130	0	0.163	0	0.212	44
274	0	0.061	0	0.123	0	0.182	0	0.241	4

Aceptar

- Añada un nodo Tabla al nodo Reorg. campos y ejecútelo.

Se espera que 264 abandonen al final del año, 184 al final del tercer trimestre, 103 en el segundo y 31 en el primero. Observe que dos clientes cualesquiera, uno de ellos con una alta propensión de abandono en el primer trimestre no tiene necesariamente una mayor propensión de abandono en otros trimestres; por ejemplo, consulte los registros 256 y 260. Es muy probable que se deba a la forma de la función de impacto de los meses posteriores al periodo actual; por ejemplo, los clientes que han contratado el servicio por una promoción tienen más posibilidades de abandono que los clientes que contrataron el servicio por una recomendación personal, pero si no lo hacen serán más leales durante el periodo restante. Es posible que desee volver a ordenar los clientes para tener vistas diferentes de los clientes con más probabilidades de abandono.

Figura 27-41
Tabla con clientes con valores nulos



	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenure
707	0	$null$	0	$null$	0	$null$	0	$null$	71
708	0	$null$	0	$null$	0	$null$	0	$null$	71
709	0	$null$	0	$null$	0	$null$	0	$null$	71
710	0	$null$	0	$null$	0	$null$	0	$null$	72
711	0	$null$	0	$null$	0	$null$	0	$null$	71
712	0	$null$	0	$null$	0	$null$	0	$null$	72
713	0	$null$	0	$null$	0	$null$	0	$null$	72
714	0	$null$	0	$null$	0	$null$	0	$null$	72
715	0	$null$	0	$null$	0	$null$	0	$null$	70
716	0	$null$	0	$null$	0	$null$	0	$null$	70
717	0	$null$	0	$null$	0	$null$	0	$null$	71
718	0	$null$	0	$null$	0	$null$	0	$null$	72
719	0	$null$	0	$null$	0	$null$	0	$null$	72
720	0	$null$	0	$null$	0	$null$	0	$null$	72
721	0	$null$	0	$null$	0	$null$	0	$null$	72
722	0	$null$	0	$null$	0	$null$	0	$null$	71
723	0	$null$	0	$null$	0	$null$	0	$null$	70
724	0	$null$	0	$null$	0	$null$	0	$null$	71
725	0	$null$	0	$null$	0	$null$	0	$null$	70
726	0	$null$	0	$null$	0	$null$	0	$null$	72

En la parte inferior de la tabla se encuentran los clientes con valores nulos pronosticados. Hay clientes cuyo periodo total (tiempo futuro + *periodo*) está dentro del intervalo de horas de supervivencia en el conjunto de datos utilizado para entrenar el modelo.

Resumen

Mediante la regresión de Cox, ha identificado un modelo aceptable del tiempo de abandono, ha trazado el número esperado de clientes mantenidos en los dos años siguientes e identificado los clientes con más posibilidades de abandono el año que viene. Tenga en cuenta que aunque sea un modelo aceptable, es posible que no sea el mejor modelo. Lo ideal es que compare este modelo, obtenido con el método de selección por pasos hacia adelante, con el que ha creado mediante el método de selección por pasos hacia atrás.

Las explicaciones de los fundamentos matemáticos de los métodos de modelado que se utilizan en IBM® SPSS® Modeler se enumeran en el *Manual de algoritmos de SPSS Modeler*.

Capítulo
28

Análisis de la cesta del supermercado (Reglas de inducción/C5.0)

Este ejemplo está relacionado con datos ficticios que describen el contenido de cestas de supermercado (es decir, una colección de artículos comprados a la vez) junto con los datos personales del comprador, que pueden obtenerse a través de las tarjetas de fidelidad. El objetivo es descubrir grupos de clientes que compren productos parecidos calificables desde el punto de vista demográfico, como por edad, ingresos, etc.

Este ejemplo muestra dos fases de la minería de datos:

- Modelado de reglas de asociación y una visualización de malla que muestra enlaces entre los artículos comprados
- Perfilado de reglas de inducción C5.0 de los compradores de grupos identificados de productos

Nota: Esta aplicación no utiliza directamente el modelado predictivo y, por tanto, no hay una medida de precisión para los modelos resultantes ni entrenamiento asociado/distinción de comprobaciones en el proceso de minería de datos.

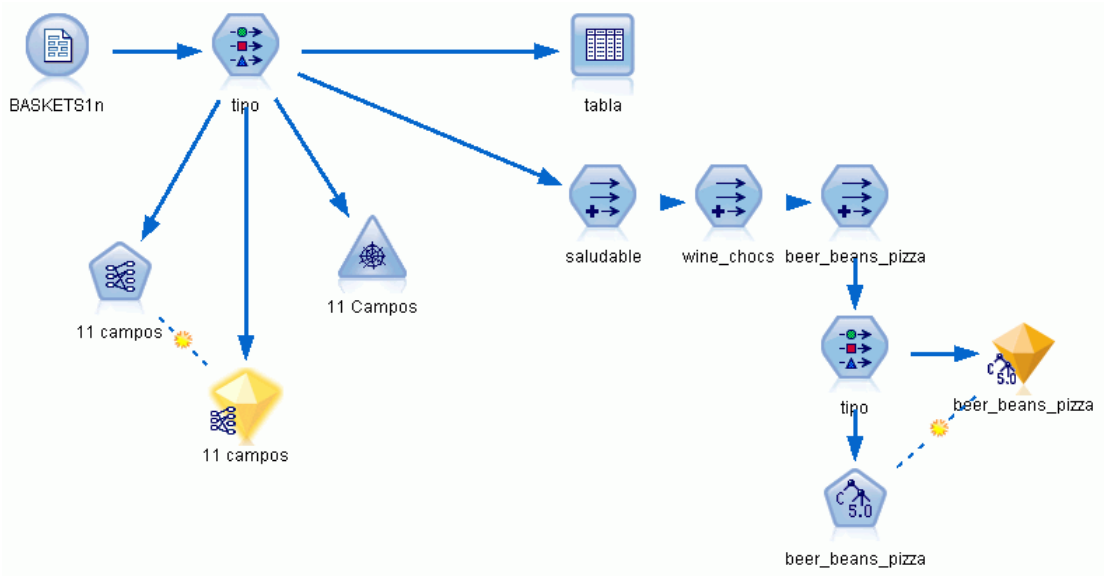
Este ejemplo utiliza la ruta denominada *baskrule*, que hace referencia al archivo de datos denominado *BASKETS1n*. Estos archivos están disponibles en el directorio *Demos* de la instalación de IBM® SPSS® Modeler. Puede acceder desde el grupo de programas IBM® SPSS® Modeler en el menú Inicio de Windows. El archivo *baskrule* se encuentra en el directorio *streams*.

Acceso a los datos

Utilizando un nodo Archivo variable, conéctese al conjunto de datos *BASKETS1n* para leer los nombres de campos del archivo. Conecte un nodo Tipo al origen de datos y, a continuación, conecte el nodo a un nodo Tabla. Defina el nivel de medición de campo *id_tarjeta* como *Sin tipo* (porque cada identificación de las tarjetas de fidelidad sólo aparece una vez en el conjunto de datos y, por lo tanto, puede no ser utilizada en el modelado). Seleccione *Nominal* como nivel

de medición para el campo *sexo* (para asegurar que el algoritmo de modelado Apriori no trate *sexo* como una marca).

Figura 28-1
ruta baskrule



Ahora, ejecute la ruta para instanciar el nodo Tipo y mostrar la tabla. El conjunto de datos contiene 18 campos y cada registro representa una cesta.

Los 18 campos están representados en los siguientes encabezados.

Resumen de los campos de cesta:

- *id_tarjeta*. Identificación de tarjetas de fidelidad para el cliente que compre esta cesta.
- *valor*. Precio de compra total de la cesta.
- *forma_pago*. Forma de pago de la cesta.

Datos personales del titular de la tarjeta:

- *sexo*
- *casa_propia*. Si el titular posee o no una casa propia.
- *ingresos*
- *edad*

Contenido de la cesta (marcas para la presencia de categorías de productos):

- *frutería*
- *carne*
- *lácteos*
- *lata_veg*
- *embutidos*

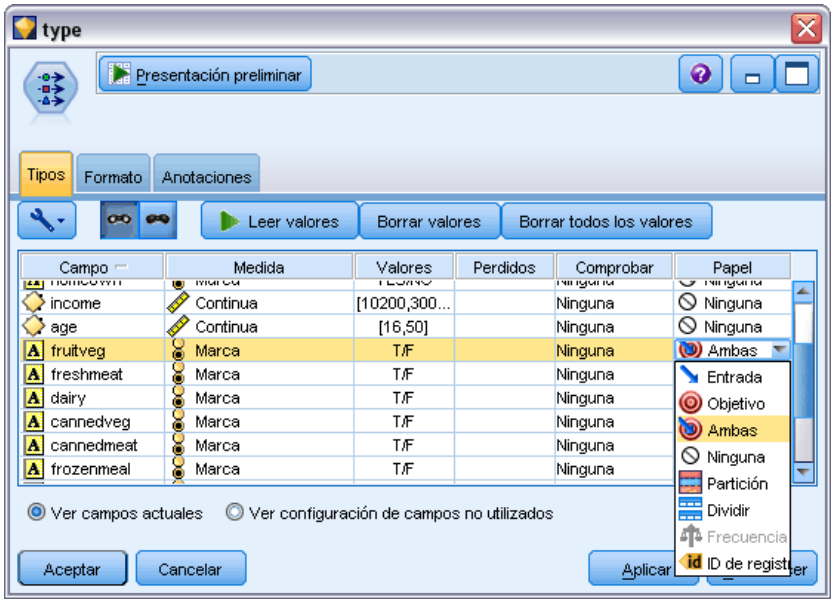
- *congelados*
- *cerveza*
- *vino*
- *refrescos*
- *pescado*
- *pastelería*

Descubrimiento de afinidades en el contenido de las cestas

Primero, debe obtener una visión general de las afinidades (asociaciones) del contenido de las cestas utilizando Apriori para crear reglas de asociación. Seleccione los campos que va a utilizar en este proceso de modelado editando el nodo Tipo y definiendo el papel de todas las categorías de productos como *Ambas* y el resto de papeles como *Ninguno*. (*Ambas* significa que el campo puede ser de entrada o de salida en el modelo resultante).

Nota: puede establecer las opciones de varios campos a la vez pulsando la tecla Mayús para seleccionarlos antes de especificar una opción de las columnas.

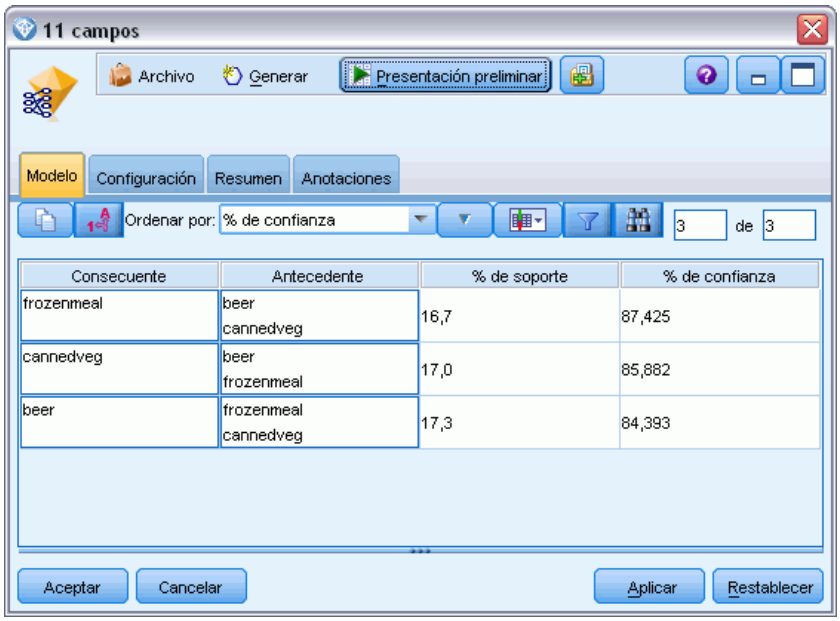
Figura 28-2
Selección de campos para el modelado



Una vez que haya especificado los campos para el modelado, conecte un nodo Apriori al nodo Tipo, edítelo, seleccione la opción Sólo valores verdaderos para las marcas y pulse en ejecutar el nodo Apriori. El resultado, un modelo de la pestaña Modelos en la parte superior derecha de

la ventana Administradores, contiene reglas de asociación que puede ver utilizando el menú contextual y seleccionando Examinar.

Figura 28-3
Reglas de asociación



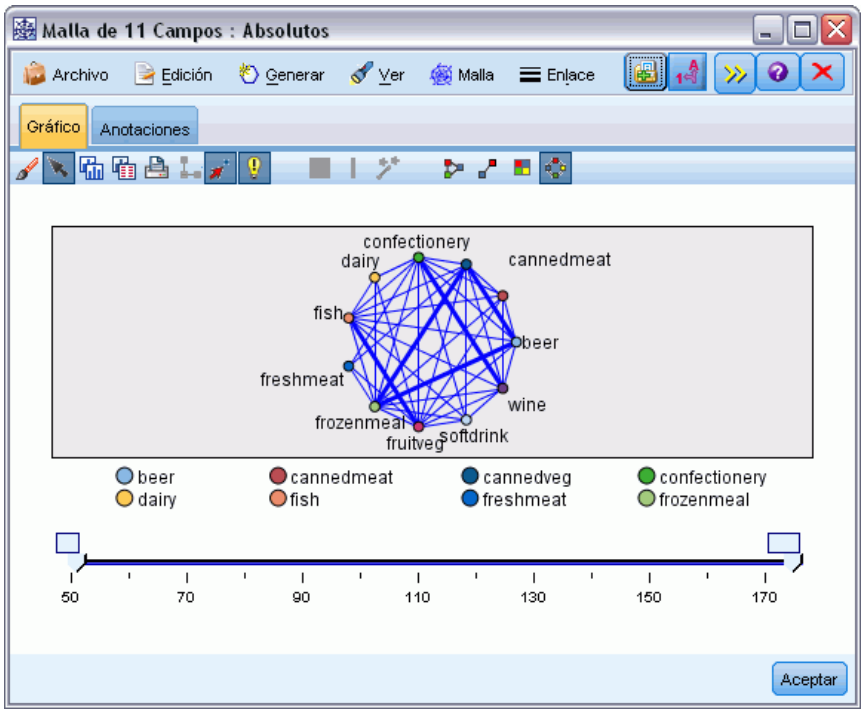
Estas reglas muestran una variedad de asociaciones entre congelados, latas de verduras y cerveza. La presencia de reglas de asociación de dos factores como:

```
congelados -> cerveza
cerveza -> congelados
```

sugiere que una visualización de malla (que muestre sólo asociaciones de dos factores) puede resaltar algunos de los patrones de estos datos.

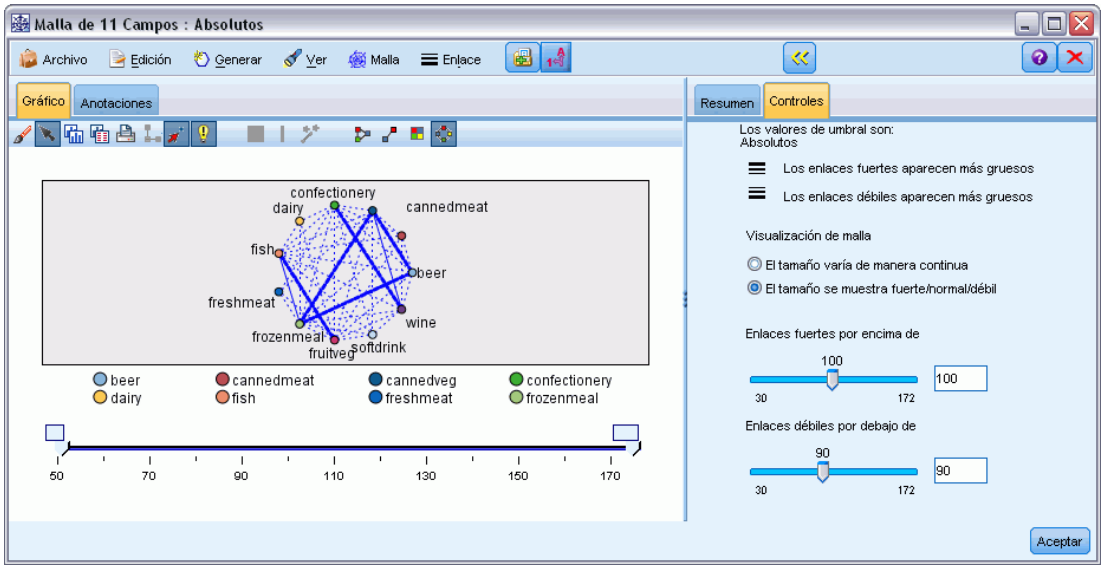
Conecte un nodo Malla al nodo Tipo, edite el nodo Malla, seleccione todo el contenido de la cesta, seleccione Mostrar sólo marcas verdaderas y pulse en ejecutar el nodo Malla.

Figura 28-4
Visualización de malla de asociaciones de productos



Puesto que la mayoría de las combinaciones de categorías de productos se producen en varias cestas, los enlaces fuertes de esta malla son demasiado numerosos para mostrar los grupos de clientes sugeridos por el modelo.

Figura 28-5
Visualización de malla restringida



- Para especificar conexiones débiles y fuertes, pulse en el botón de flecha doble amarilla de la barra de herramientas. Esto expande el cuadro de diálogo que muestra los controles y el resumen del resultado de la malla.
- Seleccione El tamaño se muestra fuerte/normal/débil.
- Establezca enlaces débiles por debajo de 90.
- Establezca enlaces fuertes por encima de 100.

En la visualización, sobresalen tres grupos de clientes:

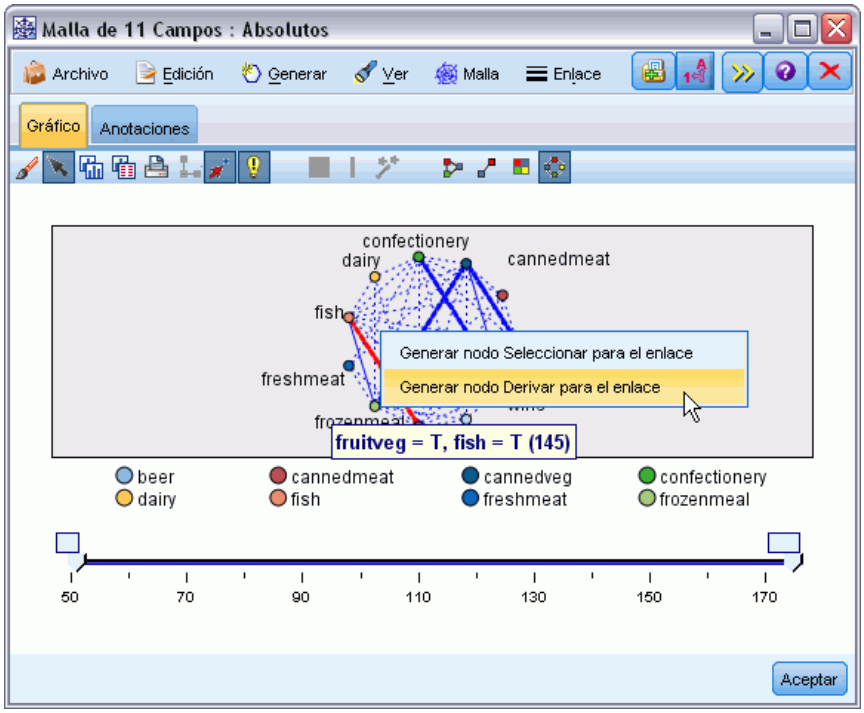
- Aquellos que compran pescado, fruta y verdura, a los que se podría denominar “consumidores sanos”.
- Aquellos que compran vino y productos de pastelería.
- Aquellos que compran cerveza, congelados y latas de verdura (“cerveza, judías y pizza”)

Perfilado de los grupos de clientes

Ahora, ha identificado tres grupos de clientes según los tipos de productos que compran, pero también quiere saber quiénes son estos clientes, es decir, su perfil demográfico. Puede lograrlo etiquetando a cada cliente con una marca de cada uno de estos grupos y utilizando una regla de inducción (C5.0) para generar reglas basadas en los perfiles de dichas marcas.

Primero debe derivar una marca para cada grupo. Esto se puede hacer de forma automática utilizando la visualización de malla que acaba de crear. Con el botón derecho del ratón, pulse en el enlace entre *frutería* y *pescado* para resaltarlo y pulse con el botón derecho y seleccione Generar nodo Derivar para el enlace.

Figura 28-6
Derivar una marca para cada grupo de clientes



Edite el nodo Derivar resultante para cambiar el nombre del campo Derivar a *sano*. Repita el ejercicio con el enlace de *vino* a *pastelería* y llame al campo Derivar resultante *vino_choco*.

Para el tercer grupo (que implica tres enlaces), asegúrese primero de que ningún enlace está seleccionado. A continuación, seleccione los tres enlaces en el triángulo *lata_veg*, *cerveza* y *congelados*. Para ello, mantenga pulsada la tecla Mayús mientras pulsa el botón izquierdo del ratón. (Asegúrese de estar en modo interactivo, y no en modo de edición). A continuación, en el menú de la visualización de malla elija:

Generar > Nodo Derivar (“Y”)

Cambie el nombre del campo Derivar resultante a *cerveza_judías_pizza*.

Para perfilar estos grupos de clientes, conecte el nodo Tipo existente a esos tres nodos Derivar en serie y, a continuación, conecte otro nodo Tipo. En el nuevo nodo Tipo, defina el papel de todos los campos como *Ninguno*, excepto para *valor*, *forma_pago*, *sexo*, *casa_propia*, *ingresos* y *edad*, que deberían establecerse como *Entrada* y el grupo de clientes relevante (por ejemplo, *cerveza_judías_pizza*), que debería establecerse como *Objetivo*. Adjunte un nodo C5.0, establezca

el tipo Salida en Conjunto de reglas y pulse en ejecutar el nodo. El modelo resultante (para *cerveza_judías_pizza*) contiene un perfil demográfico claro para este grupo de clientes:

```
Regla 1 para T:
si sexo = M
y los ingresos <= 16,900
por tanto T
```

El mismo método puede aplicarse a las marcas de los grupos de clientes seleccionándolos como salida en el segundo nodo Tipo. En este contexto, se puede generar un rango más amplio de perfiles alternativos utilizando Apriori en lugar de C5.0. Apriori también puede utilizarse para perfilar las marcas de grupos de clientes de forma simultánea porque no se restringen a un único campo de salida.

Resumen

Este ejemplo muestra cómo puede utilizarse IBM® SPSS® Modeler para descubrir afinidades, o enlaces, en una base de datos tanto por modelado (utilizando Apriori) como por visualización (utilizando una visualización de malla). Estos enlaces se corresponden con agrupaciones de casos de los datos. Dichas agrupaciones pueden investigarse detalladamente y perfilarse mediante modelado (utilizando conjuntos de reglas C5.0).

En el dominio de ventas, tales agrupaciones de clientes pueden utilizarse, por ejemplo, para identificar las ofertas especiales que mejoren el índice de respuesta a campañas de correo directas o para personalizar la gama de existencias almacenadas en un establecimiento para ajustarla a las necesidades de su base demográfica.

Capítulo
29

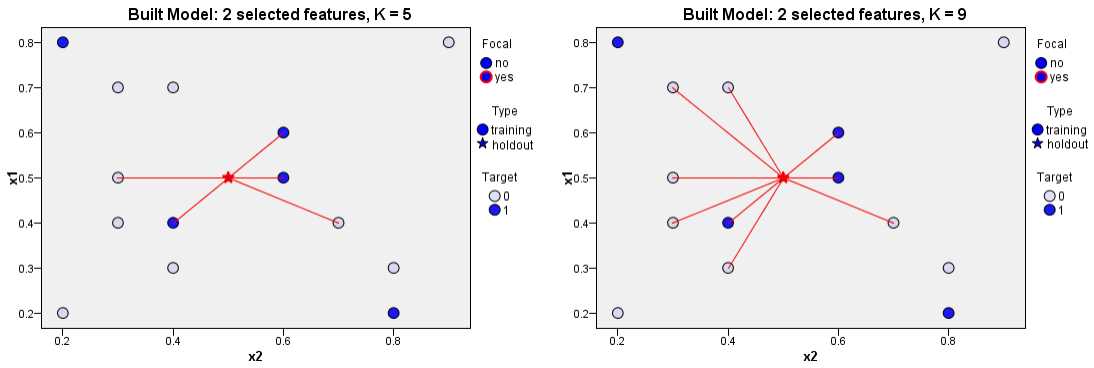
Evaluación de las nuevas ofertas de vehículos (KNN)

Análisis de vecino más próximo es un método de clasificación de casos basado en su similaridad con otros casos. En aprendizaje de máquinas, se ha desarrollado como una forma de reconocer patrones de datos sin requerir una coincidencia exacta con patrones o casos almacenados. Los casos similares están cercanos entre sí y los casos no similares están distantes entre sí. Además, la distancia entre dos casos es una medida de sus diferencias.

Los casos muy cercanos a otros se denominan "vecinos". Cuando se presenta un nuevo caso (reserva), se calcula su distancia desde cada caso del modelo. Las clasificaciones de la mayoría de casos similares (los vecinos más próximos) se anotan y el nuevo caso se coloca en la categoría que contiene el mayor número de vecinos más próximos.

Puede especificar el número de vecinos más próximos que se van a examinar; este valor se denomina *k*. Las imágenes muestran cómo se clasifica un nuevo caso utilizando dos valores diferentes de *k*. Si *k* = 5, el nuevo caso se coloca en la categoría *1* porque una mayoría de los vecinos más próximos pertenecen a esa categoría *1*. Sin embargo, si *k* = 9, el nuevo caso se coloca en la categoría *0* porque una mayoría de los vecinos más próximos pertenecen a esa categoría *0*.

Figura 29-1
Los efectos de modificar k en la clasificación



El análisis de vecino más próximo también se puede utilizar para calcular los valores de un objetivo continuo. En esta situación, la media o el valor objetivo medio de los vecinos más próximos se utiliza para obtener el valor pronosticado del nuevo caso.

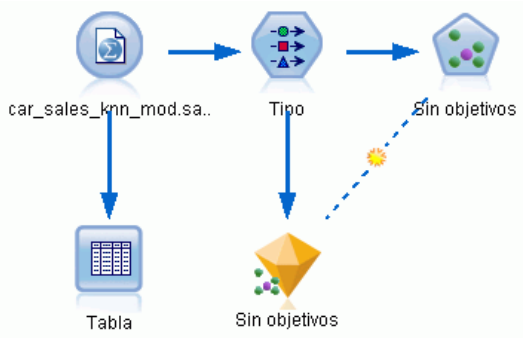
Un fabricante de automóviles ha desarrollado prototipos para dos nuevos vehículos, un coche y una furgoneta. Antes de presentar los nuevos modelos en su gama, el fabricante desea determinar qué vehículos existentes en el mercado se asemejan más a los prototipos, o sea, qué vehículos representan su "competencia directa".

El fabricante ha recopilado datos sobre modelos existentes, bajo un número de categorías, y ha añadido los detalles de sus prototipos. Las categorías bajo las que se compararán los modelos incluyen el precio en miles (*precio*), cubicaje del motor (*c_motor*), caballos (*caballos*), distancia entre ejes (*batalla*), anchura (*anchura*), longitud (*longitud*), peso en vacío (*peso_vacío*), capacidad de combustible (*cap_combustible*) y consumo de combustible (*autonomía*).

Este ejemplo utiliza la ruta denominada *car_sales_knn.str*, disponible en la carpeta *Demos* bajo la subcarpeta *streams*. El archivo de datos es *car_sales_knn_mod.sav*. Si desea obtener más información, consulte el tema Carpeta Demos en el capítulo 1 en *Manual de usuario de IBM SPSS Modeler 14.2*.

Creación de la ruta

Figura 29-2
Ruta de ejemplo para modelado KNN

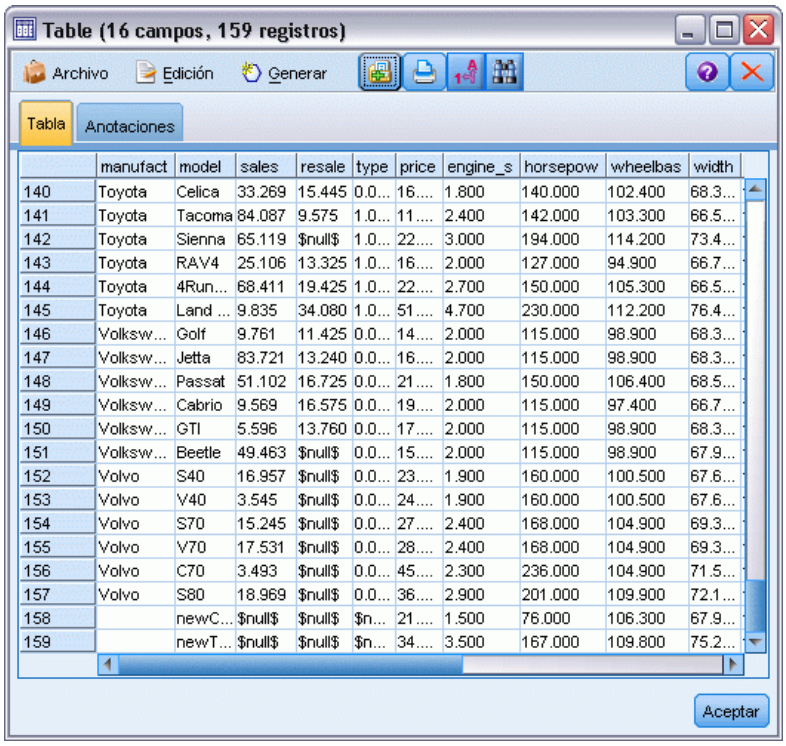


Cree una nueva ruta y añada un nuevo nodo de origen de Archivo Statistics que apunte a *car_sales_knn_mod.sav* en la carpeta *Demos* de su instalación de IBM® SPSS® Modeler.

En primer lugar, veamos qué datos ha recopilado el fabricante.

- Conecte un nodo Tabla al nodo de origen de Archivo Statistics.
- Abra el nodo Tabla y pulse en Ejecutar.

Figura 29-3
Datos de origen para coches y furgonetas



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

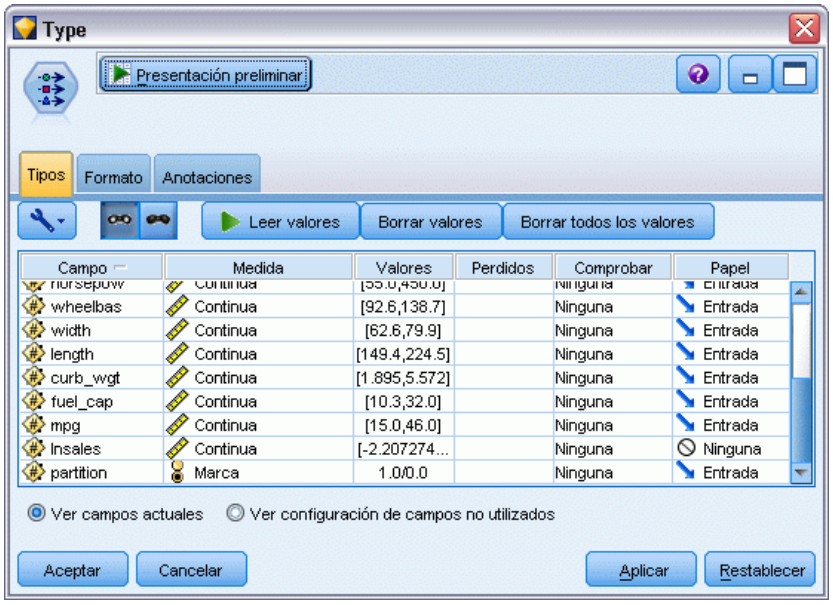
Los detalles para los dos prototipos, con los nombres *newCar* y *newTruck*, se han añadido al final del archivo.

Podemos ver en los datos de origen que el fabricante está utilizando la clasificación de “furgoneta” (valor de 1 en la columna *tipo*) de forma poco rigurosa para que implique cualquier tipo de vehículo que no sea automóvil.

La última columna, *partición*, es necesaria para que los dos prototipos puedan designarse como reservados cuando se llegue al punto de identificar su competencia directa. De esta forma, sus datos no tendrán repercusión en los cálculos, ya que es el resto del mercado lo que queremos considerar. El establecimiento del valor *partición* de los dos registros reservados a 1, mientras que el resto de los registros tienen 0 en este campo, nos permite utilizar este campo más adelante cuando tengamos que establecer los registros focales, que son los registros en los que queremos calcular la competencia directa.

Deje la ventana de resultados de la tabla abierta por el momento, ya que la necesitaremos más adelante.

Figura 29-4
Configuración del nodo Tipo

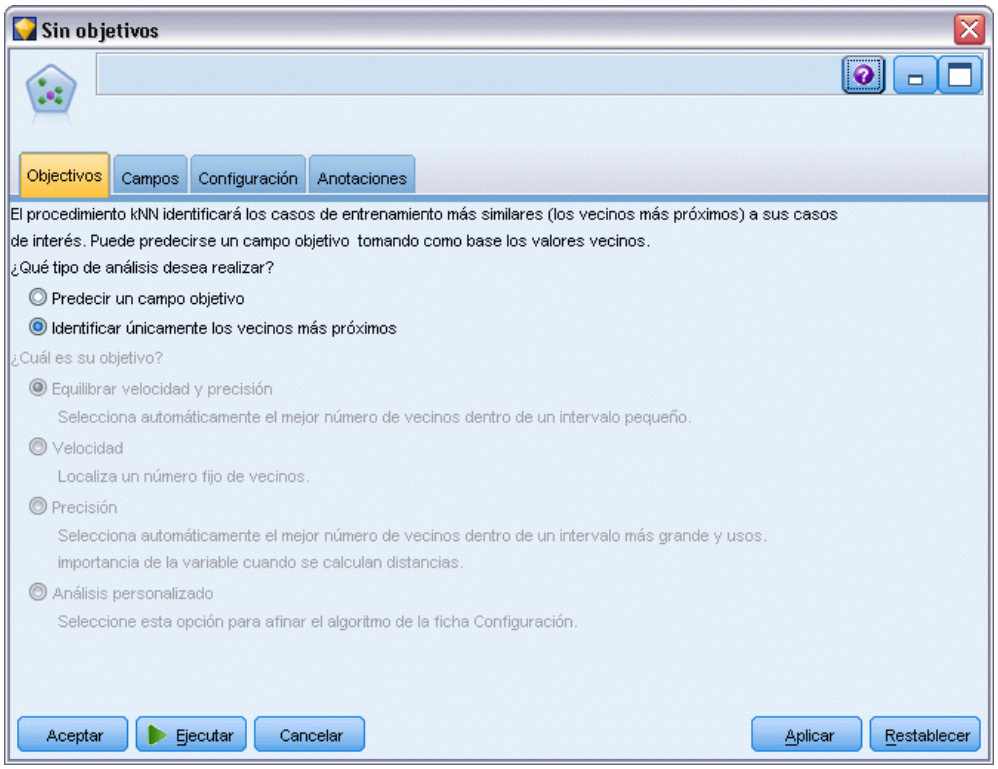


- Añada un nodo Tipo a la ruta.
- Conecte un nodo Tipo al nodo de origen de Archivo Statistics.
- Abra el nodo Tipo.

  Deseamos realizar la comparación únicamente en los campos *precio* hasta *autonomía*, de forma que dejaremos el papel para todos estos campos establecidos en Entrada.

- Establezca el papel para el resto de los campos (*fabricante* a *tipo*, junto con *Enventas*) a Ninguno.
- Establezca el nivel de medición para el último campo, *partición* a Marca. Asegúrese de que su papel se ha establecido en Entrada.
- Pulse en Leer valores para leer los valores de los datos de la ruta.
- Pulse en Aceptar.

Figura 29-5
Selección de la identificación de la competencia directa

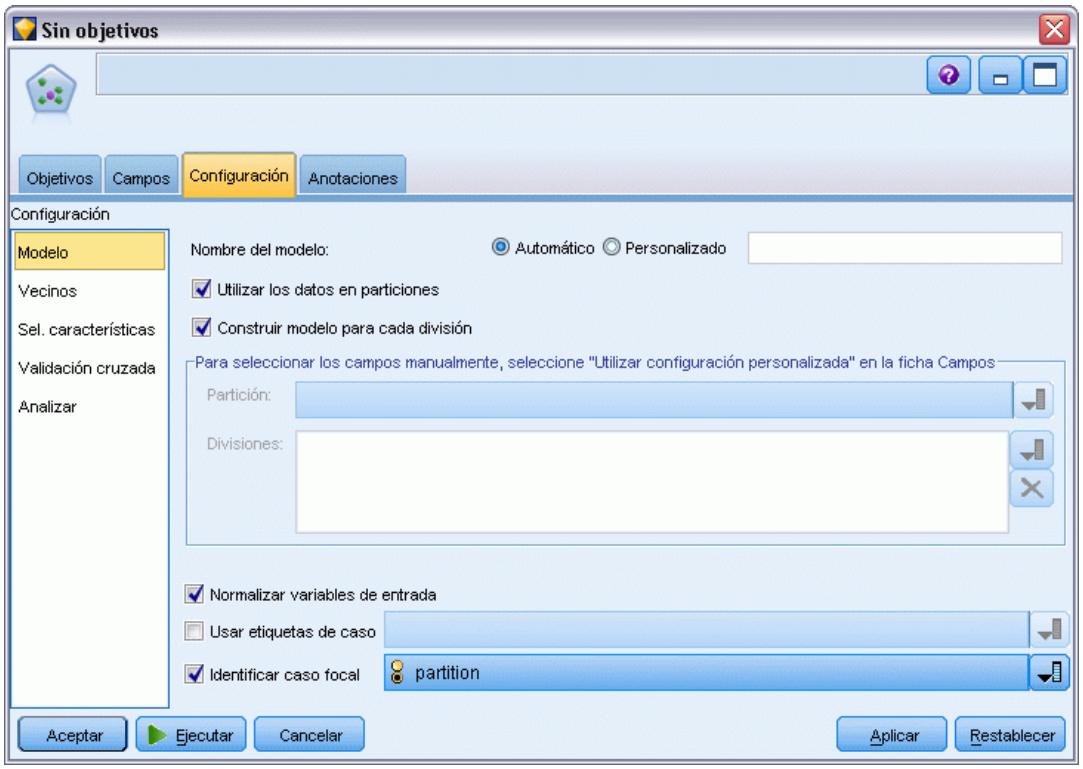


- Conecte un nodo KNN al nodo Tipo.
- Abra el nodo KNN.

  No vamos a predecir un campo objetivo en este momento, ya que sólo deseamos encontrar la competencia directa para nuestros dos prototipos.

- En la pestaña Objetivos, seleccione Identificar sólo los vecinos más próximos.
- Pulse en la pestaña Configuración.

Figura 29-6
Uso del campo partición para identificar los registros focales



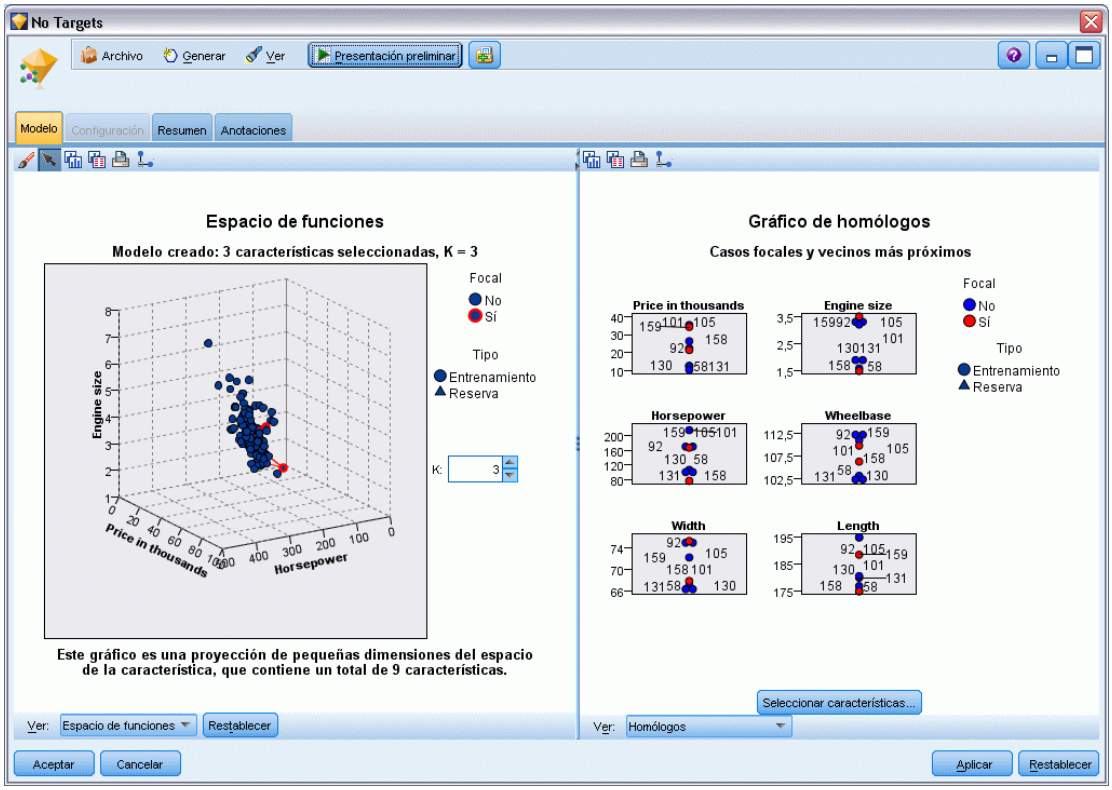
Ahora podemos utilizar el campo *partición* para identificar los registros focales, que son los registros en los que deseamos identificar la competencia directa. Utilizando un campo marca, nos aseguramos de que nos registros donde el valor de este campo está establecido como 1 se convierten en nuestros registros focales.

Como hemos visto, sólo los registros que tienen un valor de 1 en este campo son *newCar* y *newTruck*, de modo que serán nuestros registros focales.

- En el panel Modelo de la pestaña Configuración, seleccione la casilla Identificar registro focal.
- En la lista desplegable de este campo, seleccione partición.
- Pulse en el botón Ejecutar.

Examen de los resultados

Figura 29-7
La ventana Model Viewer



Se ha creado un nugget de modelo en el lienzo de rutas y en la paleta Modelos. Abra cualquiera de los nuggets para ver la visualización de Model Viewer, que tiene una ventana de dos paneles:

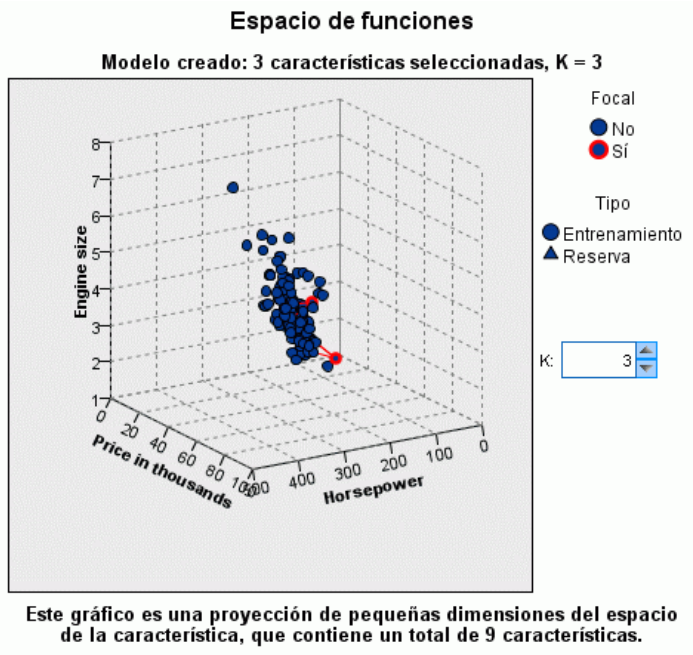
- El primer panel muestra una descripción general del modelo denominado vista principal. La vista principal del modelo Vecino más próximo se conoce como el **espacio predictor**.
- El segundo panel muestra uno de los dos tipos de vistas:

  Una vista de modelos auxiliar muestra más información sobre el modelo, pero no se centra en el propio modelo.

  Una vista enlazada es una vista que muestra detalles sobre una función del modelo cuando se desglosa parte de la vista principal.

Espacio predictor

Figura 29-8
Gráfico espacio predictor



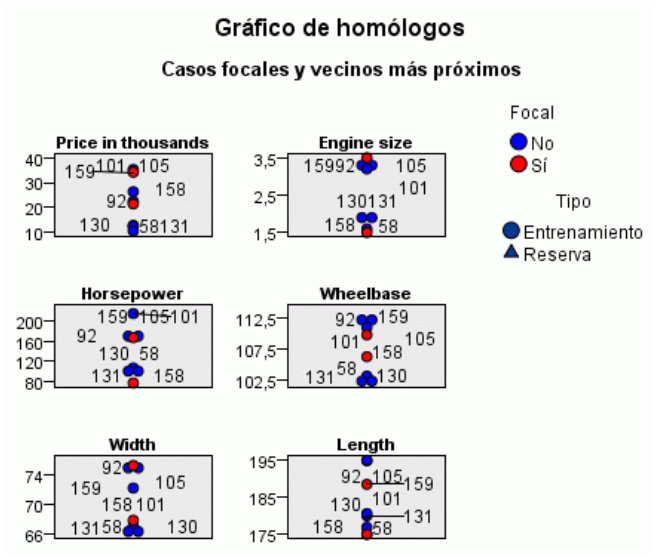
El gráfico espacio predictor es un gráfico interactivo en 3-D que representa puntos de datos para las tres funciones (los tres primeros campos de entrada de los datos de origen), representando el precio, el cubicaje y los caballos.

Nuestros dos registros focales están resaltados en rojo, con líneas que los conectan a sus vecinos *k* más próximos.

Ha pulsar y arrastrar el gráfico, podrá girarlo para obtener una mejor visión de la distribución de los puntos en el espacio predictor. Pulse en el botón Restablecer para volver a la vista por defecto.

Gráfico Homólogos

Figura 29-9
Gráfico de homólogos



La vista auxiliar por defecto es el gráfico de homólogos, que resalta los dos registros focales seleccionados en el espacio predictor y sus vecinos *k* más próximos en las seis funciones: los primeros seis campos de entrada de los datos de origen.

Los vehículos están representados por sus números de registro en los datos de origen. Aquí es donde necesitamos los resultados del nodo de Tabla para ayudarnos a su identificación.

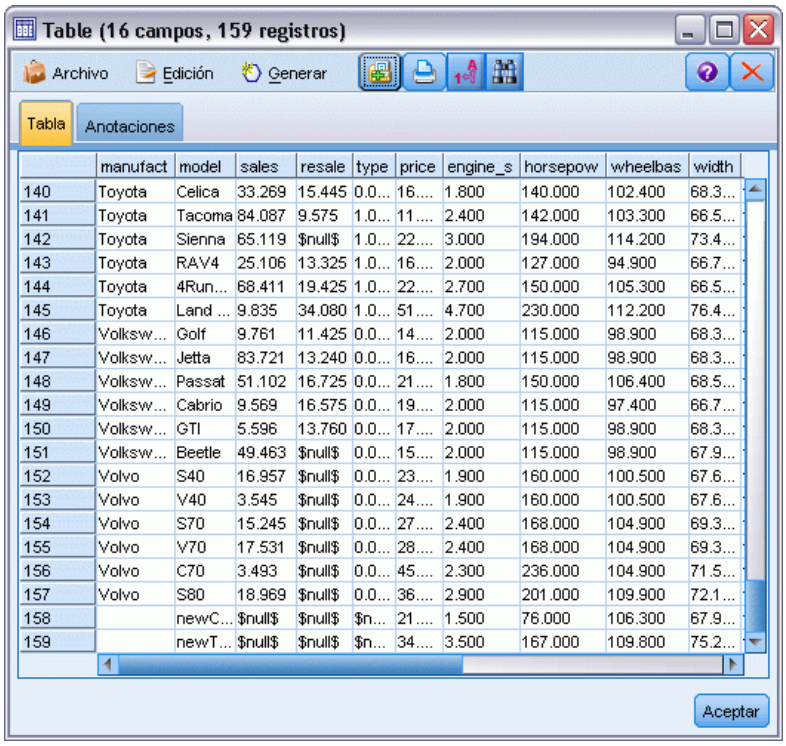
Si el resultado del nodo de Tabla está aún disponible:

- Pulse la pestaña Resultados del panel de administrador en la parte superior derecha de la ventana principal de IBM® SPSS® Modeler.
- Pulse dos veces en la entrada Tabla (16 campos, 159 registros).

Si el resultado de la tabla ya no está disponible:

- En la ventana principal de SPSS Modeler, abra el nodo Tabla.
- Pulse en Ejecutar.

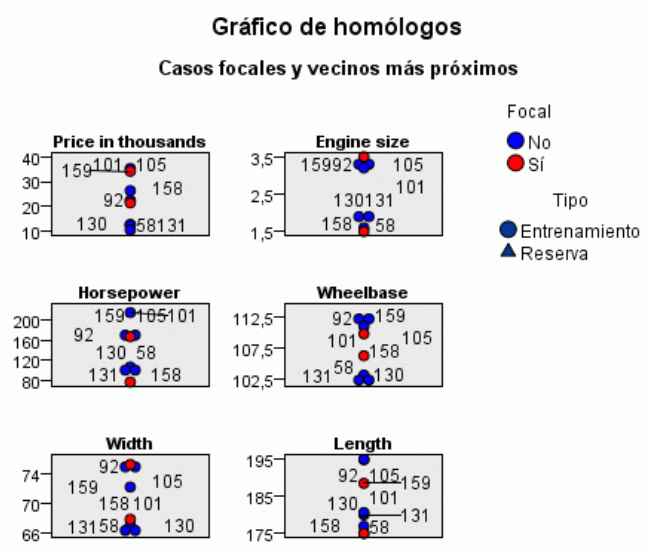
Figura 29-10
Identificación de registros por número de registro



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

Al desplazarnos hasta el final de la tabla, podemos ver que *newCar* y *newTruck* son los dos últimos registros en los datos, con los números 158 y 159 respectivamente.

Figura 29-11
Comparación de funciones en el gráfico de homólogos



Desde aquí podemos ver en el gráfico de homólogos, por ejemplo, que *newTruck* (159) tiene un cubicaje mayor que cualquiera de sus vecinos más próximos, mientras que *newCar* (158) tiene un motor más pequeño que cualquiera de *sus* vecinos más próximos.

Puede mover el ratón sobre cualquiera de los puntos individuales en las seis funciones para ver el valor real de cada función para ese caso en particular.

Pero ¿qué vehículos representan la competencia directa de *newCar* y *newTruck*?

El gráfico de homólogos tiene demasiados datos, de modo que habrá que cambiar a una vista más simple.

- Pulse la lista desplegable Ver en la parte inferior del gráfico de homólogos (la entrada que dice Homólogos).
- Seleccione Tabla de vecinos y distancias.

Tabla de vecinos y distancias

Figura 29-12
Tabla de vecinos y distancias

k distancias y vecinos más próximos

Visualizado para casos focales iniciales

Caso focal	Vecinos más próximos			Distancias más próximas		
	1	2	3	1	2	3
158	131	130	58	0,979	0,990	1,011
159	105	92	101	0,580	0,634	0,644

Ahora se ve mejor. Ahora podemos ver los tres modelos que más se acercan a nuestros dos prototipos en el mercado.

Para *newCar* (registro focal 158) son el Saturn SC (131), el Saturn SL (130) y el Honda Civic (58).

No resulta una gran sorpresa, los tres son berlinas de tamaño medio, de modo que *newCar* debería tener una buena cuota de mercado, especialmente por su excelente autonomía.

Para *newTruck* (registro focal 159), la competencia directa es el Nissan Quest (105), el Mercury Villager (92) y el Clase M de Mercedes (101).

Como hemos visto antes, no son necesariamente furgonetas en el sentido tradicional, son simplemente vehículos que están clasificados como automóviles especiales. Al mirar al resultado del nodo Tabla para su competencia directa, podemos ver que *newTruck* tiene un precio relativamente caro, así como uno de los más pesados de su segmento. Sin embargo, su autonomía es de nuevo mejor que la de sus rivales más cercanos, por lo que debe contar a su favor.

Resumen

Hemos visto cómo puede utilizar el análisis de vecinos más próximos para comparar un conjunto de funciones con un amplio abanico en casos a partir de un conjunto de datos en particular. También hemos calculado, para dos registros reservados muy diferentes, los casos que recuerdan mejor estos registros reservados.

Apéndice

A

Avisos

This information was developed for products and services offered worldwide.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

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

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

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 en el Reino Unido ni en cualquier otro país en los que dichas provisiones sean contrarias a la legislación local: SPSS INC., AN IBM COMPANY, PROPORCIONA ESTA PUBLICACIÓN "TAL CUAL" SIN GARANTÍAS DE NINGÚN TIPO, YA SEA EXPRESAS O IMPLÍCITAS, INCLUYENDO, SIN LIMITAR LA GENERALIDAD DE LAS GARANTÍAS IMPLÍCITAS DE NO INFRACCIÓN, COMERCIALIZACIÓN O IDONEIDAD PARA UN FIN DETERMINADO. Algunos estados no permiten el descargo de responsabilidad de garantías expresas o implícitas en determinadas transacciones, por lo que esta declaración no será aplicable.

Esta información puede incluir imprecisiones técnicas o errores tipográficos. La información que se contiene se puede modificar periódicamente; estos cambios se incorporarán en las nuevas ediciones de la publicación. SPSS Inc. puede realizar mejoras y/o cambios en el producto(s) y/o el programa(s) descrito en esta publicación en cualquier momento sin notificación.

Las referencias a esta información en sitios web ajenos a SPSS y a IBM se proporcionan únicamente por motivos de comodidad y no servirán de ninguna forma como aprobación de esos sitios web. Los materiales de esos sitios web no forman parte de los materiales de este producto de SPSS Inc. y el uso de esos siitios web se realiza bajo su responsabilidad.

Al enviar información a IBM o SPSS, el usuario concede a IBM y a SPSS el derecho no exclusivo de utilizar o distribuir la información de la forma que estime adecuada sin incurrir en obligaciones con el usuario.

La información relacionada con productos ajenos a productos SPSS se ha obtenido de los proveedores de esos productos, de sus anuncios publicados u otros orígenes disponibles de forma pública. SPSS no ha comprobado esos productos y no puede confirmar la precisión del rendimiento, compatibilidad o cualquier otras reclamaciones relacionadas con productos ajenos a SPSS. Las cuestiones sobre las responsabilidades de productos ajenos a SPSS se deben dirigir a los proveedores de esos productos.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Esta información contiene ejemplos de datos e informes utilizados en operaciones comerciales habituales. Para ilustrarlas de la forma más completa posible, los ejemplos incluyen los nombres de personas, empresas, marcas y productos. Todos estos nombres son inventados y cualquier similitud con los nombres y direcciones de una empresa real es una coincidencia.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Marcas comerciales

IBM, el logotipo de IBM e ibm.com son marcas comerciales de IBM Corporation, registradas en múltiples jurisdicciones en todo el mundo. Existe una lista actualizada de las marcas comerciales de IBM disponible en Internet en *http://www.ibm.com/legal/copytrade.shmtl*.

SPSS es una marca comercial de SPSS Inc., an IBM Company, registradas en múltiples jurisdicciones en todo el mundo.

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

IT Infrastructure Library es una marca comercial registrada de la Agencia central de telecomunicaciones y computación central que ahora forma parte de la Oficina de comercio gubernamental.

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

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

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

ITIL es una marca comercial registrada y una marca comercial comunitaria registrada de la Oficina de Comercio Gubernamental y está registrada en la Oficina de patentes y marcas comerciales de los Estados Unidos.

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

Cell Broadband Engine es una marca comercial de Sony Computer Entertainment, Inc. en los Estados Unidos, en otros países o ambos y se utiliza con licencia.

Java y todas las marcas comerciales y logotipos basados en Java son marcas comerciales de Sun Microsystems, Inc. en los Estados Unidos, en otros países o ambos.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

Otros nombres de productos y servicios pueden ser marcas comerciales de IBM, SPSS u otras empresas.

Bibliografía

Asunción, A., y D. Newman. 2007. "UCI Machine Learning Repository." Available at http://mlearn.ics.uci.edu/MLRepository.html.

Índice