Guía de minería interna de bases de datos de IBM SPSS Modeler 16

Nota

Antes de utilizar esta información y el producto al que da soporte, lea la información de la sección “Noticias” en la página 115.

Información del producto

Esta edición se aplica a la versión 16, release 0, modificación 0 de IBM(r) SPSS(r) Modeler y a todos los releases y las modificaciones posteriores, hasta que se indique lo contrario en nuevas ediciones.

Contenido

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 predicción, clasificación, 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 proporciona una información completa, coherente y precisa en la que confían los responsables de tomar decisiones para mejorar el rendimiento de la empresa. Un conjunto integral de inteligencia empresarial, análisis predictivo, rendimiento financiero y gestión de estrategias y aplicaciones de análisis que ofrece una perspectiva clara, inmediata e interactiva del rendimiento actual y la capacidad de 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, automatizar las decisiones de forma fiable y alcanzar mejores resultados.

Como parte de esta cartera, el software IBM SPSS Predictive Analytics ayuda a las organizaciones a predecir sucesos futuros y a actuar de forma proactiva sobre esa visión para alcanzar mejores resultados empresariales. 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 a los clientes, reduciendo al mismo tiempo el fraude y el riesgo. Al incorporar el software 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 ponerse en contacto con un representante, visite http://www.ibm.com/spss.

Soporte técnico

El soporte técnico está disponible para clientes que tienen servicio de mantenimiento. Los clientes pueden ponerse en contacto con el Soporte técnico si desean recibir ayuda al utilizar productos de IBM Corp. o para la instalación en uno de los entornos de hardware soportados. Para ponerse en contacto con el Soporte técnico, consulte el sitio web de IBM Corp. en http://www.ibm.com/support. Tenga a mano su acuerdo de asistencia y esté preparado para identificarse a sí mismo y a su organización al solicitar ayuda.

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 desplegarlos en operaciones empresariales para mejorar la toma de decisiones. Con un diseño que sigue el modelo CRISP-DM, estándar del sector, IBM SPSS Modeler admite el proceso completo de minería de datos, desde los propios datos hasta obtener los mejores resultados empresariales.

IBM 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 como cliente junto con SPSS Modeler Server. También hay disponible cierto número de opciones adicionales que se resumen en las siguientes secciones. Para obtener más información, consulte http://www.ibm.com/software/analytics/spss/products/modeler/.

Productos IBM SPSS Modeler

La familia de productos IBM SPSS Modeler y su software asociado se componen de lo siguiente:

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Serveradaptadores paraIBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler es una versión con todas las funcionalidades del producto que puede instalar y ejecutar en su ordenador personal. Puede ejecutar SPSS Modeler en modo local como un producto independiente o utilizarla en modo distribuido junto con IBM SPSS Modeler Server para mejorar el rendimiento a la hora de trabajar con grandes conjuntos de datos.

Con SPSS Modeler, puede crear modelos predictivos precisos de forma rápida e intuitiva sin necesidad de programación. Mediante su exclusiva interfaz visual, podrá visualizar fácilmente el proceso de minería de datos. Con ayuda del análisis avanzado incrustado en el producto podrá detectar patrones y tendencias en sus datos que anteriormente estaban ocultos. Podrá modelar los resultados y comprender los factores que influyen en ellos, lo que le permitirá aprovechar oportunidades comerciales y mitigar los riesgos.

SPSS Modeler está disponible en dos ediciones: SPSS Modeler Professional y SPSS Modeler Premium. Para obtener más información, consulte el tema “Ediciones de IBM SPSS Modeler” en la página 2.

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.

SPSS Modeler Server es un producto con licencia independiente que se ejecuta de manera continua en modo de análisis distribuido en un host de servidor junto con una o más instalaciones de IBM SPSS

Modeler. De este modo, SPSS Modeler Server ofrece un mejor rendimiento cuando se trabaja con grandes conjuntos de datos, ya que las operaciones que requieren un uso intensivo de memoria se pueden realizar en el servidor sin tener que descargar datos al equipo cliente. IBM SPSS Modeler Server también ofrece asistencia para las capacidades de optimización de SQL y modelado interno de bases de datos, lo que proporciona mayores ventajas en cuanto al rendimiento y la automatización.

IBM SPSS Modeler Administration Console

Modeler Administration Console es una aplicación gráfica para administrar muchas de las opciones de configuración de SPSS Modeler Server, las cuales también pueden configurarse a través de un archivo de opciones. La aplicación proporciona una interfaz de usuario de la consola para supervisar y configurar las instalaciones de SPSS Modeler Server y está disponible de forma completamente gratuita para los clientes actuales de SPSS Modeler Server. La aplicación solamente se puede instalar en los ordenadores con Windows; sin embargo, puede administrar un servidor que esté instalado en cualquier plataforma compatible.

IBM SPSS Modeler Batch

Aunque la minería de datos suele ser un proceso interactivo, también es posible ejecutar SPSS Modeler desde una línea de comandos, sin necesidad de la interfaz gráfica del usuario. Por ejemplo, puede que tenga tareas repetitivas o cuya ejecución sea de larga duración que quiera realizar sin intervención del usuario. SPSS Modeler Batch es una versión especial del producto que proporciona soporte para todas las capacidades analíticas de SPSS Modeler sin el acceso a la interfaz de usuario habitual. Es necesario disponer de una licencia de SPSS Modeler Server para utilizar SPSS Modeler Batch.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher es una herramienta que le permite crear una versión empaquetada de una ruta de SPSS Modeler que se puede ejecutar en un motor de tiempo de ejecución externo o incrustado en una aplicación externa. De este modo, puede publicar y desplegar rutas completas de SPSS Modeler par su uso en entornos que no tienen instalado SPSS Modeler. SPSS Modeler Solution Publisher se distribuye como parte del servicio de IBM SPSS Collaboration and Deployment Services - Puntuación, para el cual se requiere una licencia separada. Con esta licencia, recibirá SPSS Modeler Solution Publisher Runtime, que le permite ejecutar las rutas publicadas.

IBM SPSS Modeler ServerAdaptadores paraIBM SPSS Collaboration and Deployment Services

Tiene a su disposición un determinado número de adaptadores para IBM SPSS Collaboration and Deployment Services que permiten que SPSS Modeler y SPSS Modeler Server interactúen con un repositorio de IBM SPSS Collaboration and Deployment Services. De este modo, varios usuarios podrán compartir una ruta de SPSS Modeler desplegada en el repositorio, o bien se podrá acceder a ella desde la aplicación cliente de baja intensidad IBM SPSS Modeler Advantage. Debe instalar el adaptador en el sistema donde se aloje el repositorio.

Ediciones de IBM SPSS Modeler

SPSS Modeler está disponible en las siguientes ediciones.

SPSS Modeler Professional

SPSS Modeler Professional proporciona todas las herramientas que necesita para trabajar con la mayoría de los tipos de datos estructurados, como los comportamientos e interacciones registrados en los sistemas de CRM, datos demográficos, comportamientos de compra y datos de ventas.

SPSS Modeler Premium

SPSS Modeler Premium es un producto con licencia independiente que amplía SPSS Modeler Professional para poder trabajar con datos especializados, como los utilizados para el análisis de entidades o las redes sociales, así como con datos de texto no estructurados. SPSS Modeler Premium está formado por los siguientes componentes:

IBM SPSS Modeler Entity Analytics incorpora una dimensión adicional al análisis predictivo de IBM SPSS Modeler. Mientras que el análisis predictivo trata de predecir comportamientos futuros a partir de datos del pasado, el análisis de entidades se centra en mejorar la coherencia de los datos actuales mediante la resolución de conflictos de identidades dentro de los propios registros. La identidad de un individuo, una organización, un objeto o cualquier otra entidad puede estar expuesta a ambigüedades. La resolución de identidades puede ser vital en diversos campos, entre los que se incluyen la gestión de la relación con el cliente, la detección de fraudes, la lucha contra el blanqueo de dinero y la seguridad nacional e internacional.

IBM SPSS Modeler Social Network Analysis transforma la información sobre relaciones en campos que caracterizan el comportamiento social de individuos y grupos. Mediante el uso de datos que describen las relaciones subyacentes de las redes sociales, IBM SPSS Modeler Social Network Analysis identifica a los líderes sociales que influyen en el comportamiento de otros en la red. Además, puede determinar qué personas se ven más afectadas por otros participantes de la red. Al combinar estos resultados con otras medidas, puede crear perfiles completos de individuos en los que basar sus modelos predictivos. Los modelos que incluyan esta información social tendrán un mejor rendimiento que los modelos que no la incluyan.

IBM SPSS Modeler Text Analytics utiliza tecnologías de lingüística avanzada y Procesamiento del lenguaje natural (PLN) 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.

Documentación de IBM SPSS Modeler

Tiene a su disposición 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 (incluidas las instrucciones de instalación) en formato PDF está disponible en la carpeta *Documentation* en cada DVD del producto. También es posible descargar los documentos de instalación en Internet en http://www-01.ibm.com/support/docview.wss?uid=swg27038316.

La documentación en ambos formatos también está disponible desde el centro de información de SPSS Modeler en http://publib.boulder.ibm.com/infocenter/spssmodl/v16r0m0/.

Documentación de SPSS Modeler Professional

El conjunto de documentación de SPSS Modeler Professional (excluidas las instrucciones de instalación) es el siguiente.

- **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 despliegue en IBM SPSS Collaboration and Deployment Services, Predictive Applications o IBM SPSS Modeler Advantage.
- **Nodos de 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.** Las descripciones de todos los nodos utilizados para crear modelos de minería de datos. IBM SPSS Modeler ofrece una gran variedad de métodos de modelado procedentes del aprendizaje automático, la inteligencia artificial y el estadístico.
- **IBM SPSS Modeler Algorithms Guide.** Descripciones de los fundamentos matemáticos de los métodos de modelado que se utilizan en IBM SPSS Modeler. Esta guía está disponible únicamente en formato PDF.
- **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. Para obtener más información, consulte el tema "Ejemplos de aplicaciones" en la página 5.
- **Scripts y automatización de IBM SPSS Modeler.** Información sobre la automatización del sistema a través de scripts, incluidas las propiedades que se pueden utilizar para manipular nodos y rutas.
- **IBM SPSS ModelerManual de despliegue.** Información sobre la ejecución de rutas y escenarios de IBM 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 proporciona la capacidad de integrar programas de otros fabricante como rutinas de procesamiento de datos o algoritmos de modelado como nodos en IBM 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 administración y rendimiento de IBM SPSS Modeler Server.** 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 IBM SPSS Modeler Server. La consola se implementa como complemento de la aplicación Gestor de despliegue.
- **IBM SPSS Modeler CRISP-DM Guide.** Manual que explica paso a paso cómo utilizar la metodología de CRISP-DM en la minería de datos con SPSS Modeler.
- **Manual del usuario de IBM SPSS Modeler Batch.** Guía completa de cómo utilizar IBM SPSS Modeler en modo por lotes, incluida información detallada sobre la ejecución del modo por lotes y argumentos de línea de comandos. Esta guía está disponible únicamente en formato PDF.

Documentación de SPSS Modeler Premium

El conjunto de documentación de SPSS Modeler Premium (excluidas las instrucciones de instalación) es el siguiente.

- **IBM SPSS Modeler Entity Analytics - Manual del usuario.** Información sobre cómo utilizar el análisis de entidades con SPSS Modeler, que cubre la instalación y configuración de repositorios, nodos de análisis de entidades y tareas administrativas.
- **IBM SPSS Modeler Social Network Analysis - Manual del usuario.** Una guía para realizar análisis de redes sociales con SPSS Modeler, incluido el análisis de grupos y el análisis de difusión.
- **SPSS Modeler Text Analytics - Manual del usuario.** Información sobre cómo utilizar el análisis de texto con SPSS Modeler, que cubre los nodos de minería de texto, programa interactivo, plantillas y otros recursos.
- **Manual del usuario de IBM SPSS Modeler Text Analytics Administration Console.** Información sobre cómo instalar y utilizar la interfaz de usuario de la consola para supervisar y configurar IBM SPSS Modeler Server para su uso con SPSS Modeler Text Analytics . La consola se implementa como complemento de la aplicación Gestor de despliegue.

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 ejemplo se instalan en la carpeta *Demos* en el directorio de instalación del producto. Para obtener más información, consulte el tema "Carpeta Demos".

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

Ejemplos de scripts. Consulte los ejemplos en la *Guía de scripts y automatización de IBM SPSS Modeler*.

Carpeta Demos

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

Capítulo 2. Minería interna de bases de datos

Conceptos básicos del modelado de bases de datos

IBM SPSS Modeler Server admite la integración con herramientas de modelado y minería de datos que están disponibles en proveedores de bases de datos como IBM Netezza, IBM DB2 InfoSphere Warehouse, Oracle Data Miner y Microsoft Analysis Services. Podrá crear, puntuar y almacenar modelos dentro de la base de datos, todo desde la aplicación IBM SPSS Modeler. Esto permite combinar las capacidades analíticas y la facilidad de uso de IBM SPSS Modeler con la potencia y el rendimiento de una base de datos, al mismo tiempo que se saca partido de los algoritmos nativos de bases de datos proporcionados por estos proveedores. Los modelos se generan en la base de datos, que podrá explorar y puntuar a través de la interfaz de IBM SPSS Modeler de la forma habitual y desplegar utilizando IBM SPSS Modeler Solution Publisher si fuera necesario. Los algoritmos admitidos se encuentran en la paleta de modelado de bases de datos en IBM SPSS Modeler.

Utilizar IBM SPSS Modeler para acceder a los algoritmos nativos de la base de datos ofrece varias ventajas:

- Los algoritmos internos de la base de datos suelen estar muy integrados con el servidor de la base de datos y pueden ofrecer un mejor rendimiento.
- Los modelos creados y almacenados "dentro de la base de datos" se pueden desplegar en, y compartir con, cualquier aplicación que pueda acceder a la base de datos, de una manera más fácil.

Generación de SQL. El modelado interno de bases de datos es distinto de la generación de SQL, que también se denomina "retrotracción de SQL". Esta característica permite generar sentencias SQL para operaciones nativas de IBM SPSS Modeler que se pueden "retrotraer" (es decir, ejecutar) en la base de datos con el fin de mejorar el rendimiento. Por ejemplo, los nodos Fundir, Agregar y Seleccionar generan código SQL que se puede devolver a la base de datos de esta forma. Al utilizar la generación de SQL combinada con el modelado de bases datos se pueden obtener rutas que se pueden ejecutar desde el comienzo hasta el final en la base de datos, de forma que se mejora considerablemente el rendimiento con respecto a las rutas ejecutadas enIBM SPSS Modeler.

Nota: optimización de SQL y modelado de bases de datos requieren que la conectividad de IBM SPSS Modeler Server esté activada en el equipo con IBM SPSS Modeler. Con esta configuración activada, puede acceder a los algoritmos de bases de datos, devolver SQL directamente desde IBM SPSS Modeler y acceder a IBM SPSS Modeler Server. Para verificar el estado de la licencia actual, seleccione las siguientes opciones en el menú de IBM SPSS Modeler.

Ayuda > **Acerca de** > **Detalles adicionales**

Si la conectividad está activada, verá la opción **Activación de servidor** en la pestaña Estado de licencia.

Si desea obtener información sobre los algoritmos admitidos, consulte las siguientes secciones acerca de los proveedores específicos.

Requisitos

Para realizar el modelado de bases de datos, necesita la siguiente configuración:

- Una conexión ODBC a una base de datos adecuada, con los componentes analíticos necesarios instalados (Analysis Services de Microsoft, Oracle Data Miner o IBM DB2 InfoSphere Warehouse).
- En IBM SPSS Modeler, el modelado de bases de datos debe estar activado en el cuadro de diálogo Aplicaciones de ayuda (**Herramientas > Aplicaciones de ayuda**).

- Los ajustes **Generar SQL** y **Optimizar generación de SQL** deben estar activados en el cuadro de diálogo Opciones de usuario de IBM SPSS Modeler, así como de IBM SPSS Modeler Server (si se utiliza). Tenga en cuenta que la optimización de SQL no es estrictamente obligatoria para que el modelado de bases de datos funcione, pero se recomienda por motivos de rendimiento.

Nota: optimización de SQL y modelado de bases de datos requieren que la conectividad de IBM SPSS Modeler Server esté activada en el equipo con IBM SPSS Modeler. Con esta configuración activada, puede acceder a los algoritmos de bases de datos, devolver SQL directamente desde IBM SPSS Modeler y acceder a IBM SPSS Modeler Server. Para verificar el estado de la licencia actual, seleccione las siguientes opciones en el menú de IBM SPSS Modeler.

Ayuda > **Acerca de** > **Detalles adicionales**

Si la conectividad está activada, verá la opción **Activación de servidor** en la pestaña Estado de licencia.

Para obtener información detallada, consulte las siguientes secciones acerca de los proveedores específicos.

Generación de modelos

El proceso de creación y puntuación de modelos utilizando algoritmos de la base de datos es similar a otros tipos de minería de datos en IBM SPSS Modeler. El proceso general de trabajo con nodos y modelado de "nuggets" es similar a cualquier otra ruta al trabajar en IBM SPSS Modeler. La única diferencia es que el proceso real y creación de modelos se retrotraen a la base de datos.

Una ruta de modelado de bases de datos es conceptualmente idéntica a otras rutas de datos de IBM SPSS Modeler; sin embargo, esta ruta realiza todas las operaciones en una base de datos, incluida, por ejemplo, la creación de modelos utilizando el nodo Árbol de decisión de Microsoft. Cuando se ejecuta la ruta, IBM SPSS Modeler indica a la base de datos que cree y almacene el modelo resultante, y los detalles se descargan en IBM SPSS Modeler. La ejecución interna de la base de datos se indica mediante el uso de nodos sombreados en púrpura en la ruta.

Preparación de datos

Se utilicen o no algoritmos nativos de base de datos, las preparaciones de datos deberían retrotraerse a la base de datos cuando fuera posible para mejorar el rendimiento.

- Si los datos originales se almacenan en la base de datos, el objetivo es mantenerlos allí asegurándose de que todas las operaciones anteriores de la ruta necesarias se pueden convertir a SQL. Esto evitará que los datos se descarguen en IBM SPSS Modeler, evitando un cuello de botella que podría anular cualquier ganancia, y permitirá que toda la ruta se ejecute en la base de datos.
- El modelado de bases de datos se podrá utilizar aunque los datos originales *no* estén almacenados en la base de datos. En este caso, la preparación de datos se lleva a cabo en IBM SPSS Modeler y el conjunto de datos preparados se carga automáticamente en la base de datos para la creación del modelo.

Puntuación de modelos

Los modelos generados desde IBM SPSS Modeler utilizando la minería interna de bases de datos son distintos de los modelos de IBM SPSS Modeler habituales. Aunque aparecen en el gestor de modelos como “nuggets” del modelo generado, en realidad son modelos remotos que se guardan en el servidor remoto de la base de datos o la minería de datos. Lo que el usuario ve en IBM SPSS Modeler son simplemente referencias a estos modelos remotos. Es decir, el modelo de IBM SPSS Modeler que ve es un modelo “falso” que contiene información como el nombre de host del servidor de base de datos, el nombre de la base de datos y el nombre del modelo. Es importante comprender esta distinción cuando se examinan y puntúan modelos creados con algoritmos nativos de base de datos.

Una vez creado un modelo, puede añadirlo a la ruta para su puntuación como cualquier otro modelo generado en IBM SPSS Modeler. Todas las puntuaciones se llevan a cabo dentro de la base de datos, aunque esto no suceda con las operaciones anteriores de la ruta. (Siempre que sea posible, las operaciones anteriores de la ruta se pueden retrotraer a la base de datos para mejorar el rendimiento, pero no es un requisito para la puntuación.) También, en la mayoría de los casos, es posible examinar el modelo generado mediante el explorador estándar suministrado por el proveedor de la base de datos.

Tanto para examinar como para puntuar, se requiere una conexión activa con el servidor que ejecuta Oracle Data Miner, IBM DB2 InfoSphere Warehouse o Analysis Services de Microsoft.

Visualización de los resultados y especificación de la configuración

Para ver los resultados y especificar la configuración de la puntuación, pulse dos veces en el modelo en el lienzo de rutas. También puede pulsar con el botón derecho en el modelo y seleccionar **Examinar** o **Edición**. La configuración específica depende del tipo de modelo.

Exportación y almacenamiento de modelos de base de datos

Los modelos y resúmenes de bases de datos se pueden exportar desde el explorador de modelos del mismo modo que los demás modelos creados en IBM SPSS Modeler: utilizando las opciones del menú Archivo.

1. En el menú Archivo del explorador de modelos, seleccione alguna de las siguientes opciones:

- **Exportar texto** exporta el resumen de modelo a un archivo de texto
- **Exportar HTML** exporta el resumen de modelo a un archivo HTML
- **Exportar PMML** (compatible con modelos IBM DB2 IM únicamente) exporta el modelo como lenguaje de códigos para modelos predictivos (PMML), que se puede utilizar con otro software compatible con PMML.

Nota: también puede guardar un modelo generado seleccionando **Guardar nodo** en el menú Archivo.

Coherencia de modelos

Para cada modelo de base de datos generado, IBM SPSS Modeler almacena una descripción de la estructura del modelo junto con una referencia al modelo con el mismo nombre almacenado dentro de la base de datos. La pestaña Servidor de un modelo generado muestra una clave exclusiva generada para este modelo, que coincide con el modelo real de la base de datos.

IBM SPSS Modeler utiliza esta clave generada al azar para comprobar que el modelo sigue siendo consistente. Dicha clave se almacena en la descripción de un modelo cuando éste se crea. Es recomendable comprobar que las claves coinciden antes de ejecutar una ruta de despliegue.

1. Para comprobar la coherencia del modelo almacenado en la base de datos comparando su descripción con la clave aleatoria almacenada por IBM SPSS Modeler, pulse el botón **Comprobar**. Si no se puede encontrar el modelo de la base de datos o la clave no coincide, se notifica un error.

Presentación y exportación del SQL generado

Se puede realizar una presentación preliminar del código SQL generado antes de la ejecución, lo cual puede resultar útil para la depuración.

Capítulo 3. Modelado de bases de datos con Microsoft Analysis Services

IBM SPSS Modeler y Microsoft Analysis Services

IBM SPSS Modeler admite la integración con Microsoft SQL Server Analysis Services. Esta funcionalidad se ha implementado como nodos de modelado de IBM SPSS Modeler y está disponible en la paleta de modelado de bases de datos. Si esta paleta no está visible, puede activarla habilitando la integración de MS Analysis Services, disponible en la pestaña Microsoft del cuadro de diálogo Aplicaciones de ayuda. Para obtener más información, consulte el tema "Activación de la integración con Analysis Services" en la página 13.

IBM SPSS Modeler admite la integración de los siguientes algoritmos de Analysis Services:

- Árboles de decisión
- Clústeres
- Reglas de asociación
- bayesiano ingenuo
- Regresión lineal
- Red neuronal
- Regresión Logística
- Serie temporal
- Clúster de secuencia

El siguiente diagrama ilustra el flujo de datos desde el cliente al servidor, donde la minería interna de bases de datos es gestionada por IBM SPSS Modeler Server. La generación de modelos se realiza con Analysis Services. El modelo resultante se almacena en Analysis Services, y en las rutas de IBM SPSS Modeler se conserva una referencia a este modelo, el cual se descarga después desde Analysis Services a Microsoft SQL Server o IBM SPSS Modeler para su puntuación.

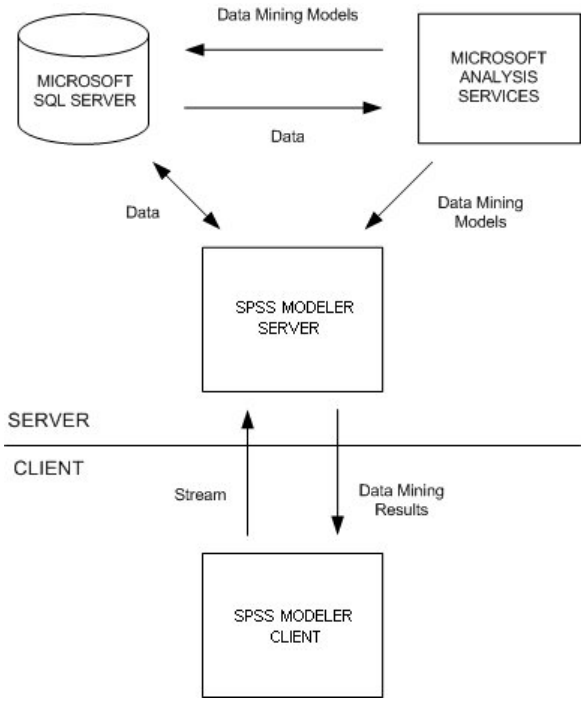

Figura 1. Flujo de datos entre IBM SPSS Modeler, Microsoft SQL Server y Microsoft Analysis Services durante la generación de modelos

Nota: el IBM SPSS Modeler Server no es obligatorio, aunque se puede utilizar. El cliente de IBM SPSS Modeler es capaz de realizar los cálculos de minería interna de bases de datos él mismo.

Requisitos para la integración con Microsoft Analysis Services

A continuación se detallan los requisitos previos para el modelado interno de bases de datos mediante algoritmos de Analysis Services con IBM SPSS Modeler. Es posible que necesite consultar con el administrador de la base de datos para asegurarse de que se reúnen las condiciones.

- Ejecución de IBM SPSS Modeler con respecto a una instalación de IBM SPSS Modeler Server (modo distribuido) en Windows. Las plataformas UNIX no se admiten en esta integración con Analysis Services.

  Importante: los usuarios de IBM SPSS Modeler deben configurar una conexión ODBC mediante el controlador SQL Native Client disponible de Microsoft en la dirección URL siguiente en *Requisitos adicionales de IBM SPSS Modeler Server*. *No se recomienda utilizar en este caso el controlador proporcionado con IBM SPSS Data Access Pack (y que normalmente se recomienda para otros usos con IBM SPSS Modeler).* El controlador debe configurarse para usar SQL Server con la opción **Con autenticación integrada de Windows** activada, dado que IBM SPSS Modeler no es compatible con la autenticación de SQL Server. Si tiene alguna pregunta acerca de la creación o configuración de permisos de los orígenes de datos ODBC, póngase en contacto con el administrador de la base de datos.
- Se debe instalar SQL Server 2005 ó 2008, aunque no necesariamente en el mismo host que IBM SPSS Modeler. Los usuarios de IBM SPSS Modeler deben tener permisos suficientes para leer y escribir datos y para eliminar y crear tablas y vistas.

  Nota: se recomienda SQL Server Enterprise Edition. Enterprise Edition ofrece flexibilidad adicional, ya que proporciona parámetros avanzados para ajustar los resultados de los algoritmos. La versión Standard Edition ofrece los mismos parámetros, pero no permite que los usuarios editen algunos parámetros avanzados.
- Microsoft SQL Server Analysis Services debe estar instalado en el mismo host que SQL Server.

Requisitos adicionales de IBM SPSS Modeler Server

Para poder utilizar los algoritmos de Analysis Services con IBM SPSS Modeler Server, deberá tener los siguientes componentes instalados en el equipo host de IBM SPSS Modeler Server.

Nota: si SQL Server está instalado en el host de IBM SPSS Modeler Server, ya estarán disponibles estos componentes.

- Microsoft .NET Framework Version 2.0 Redistributable Package (x86)
- Microsoft Core XML Services (MSXML) 6.0
- Microsoft SQL Server 2008 Analysis Services 10.0 OLE DB Provider (asegúrese de seleccionar la variación correcta de su sistema operativo)
- Microsoft SQL Server 2008 Native Client (asegúrese de seleccionar la variante correcta de su sistema operativo)

Para descargar estos componentes, vaya a *www.microsoft.com/downloads*, busque **.NET Framework** o (para el resto de componentes) **SQL Server Feature Pack** y seleccione el último paquete de su versión de SQL Server.

Es posible que esto requiera la instalación de otros paquetes antes, que también estarán disponibles a través del sitio Web de descargas de Microsoft.

Requisitos adicionales de IBM SPSS Modeler

Para poder utilizar los algoritmos de Analysis Services con IBM SPSS Modeler, debe tener instalados en el cliente los mismos componentes anteriores, además de los siguientes:

- Microsoft SQL Server 2008 Datamining Viewer Controls (asegúrese de seleccionar la variante correcta de su sistema operativo); también requiere:
- Microsoft ADOMD.NET

Para descargar estos componentes, vaya a *www.microsoft.com/downloads*, busque **SQL Server Feature Pack** y seleccione el último paquete de su versión de SQL Server.

Nota: optimización de SQL y modelado de bases de datos requieren que la conectividad de IBM SPSS Modeler Server esté activada en el equipo con IBM SPSS Modeler. Con esta configuración activada, puede acceder a los algoritmos de bases de datos, devolver SQL directamente desde IBM SPSS Modeler y acceder a IBM SPSS Modeler Server. Para verificar el estado de la licencia actual, seleccione las siguientes opciones en el menú de IBM SPSS Modeler.

Ayuda > **Acerca de** > **Detalles adicionales**

Si la conectividad está activada, verá la opción **Activación de servidor** en la pestaña Estado de licencia.

Activación de la integración con Analysis Services

Para activar la integración de IBM SPSS Modeler con Analysis Services, necesitará configurar SQL Server y Analysis Services, crear un origen ODBC, activar la integración en el cuadro de diálogo Aplicaciones de ayuda de IBM SPSS Modeler y activar la generación y optimización de SQL.

Nota: Microsoft SQL Server y Microsoft Analysis Services deben estar disponibles. Para obtener más información, consulte el tema “Requisitos para la integración con Microsoft Analysis Services” en la página 12.

Configuración de SQL Server

Configure SQL Server para que se pueda realizar la puntuación dentro de la base de datos.

1. Cree la siguiente clave de registro en el equipo host de SQL Server:
   `HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\MSSQLServer\Providers\MSOLAP`
2. Añada el siguiente valor `DWORD` a esta clave:
   `AllowInProcess 1`
3. Reinicie SQL Server después de realizar este cambio.

Configuración de Analysis Services

Para que IBM SPSS Modeler pueda comunicarse con Analysis Services, deberá configurar manualmente dos valores en el cuadro de diálogo Propiedades de Analysis Server:

1. Inicie sesión en Analysis Server mediante MS SQL Server Management Studio.
2. Acceda al cuadro de diálogo Propiedades pulsando con el botón derecho del ratón en el nombre del servidor y seleccionando **Propiedades**.
3. Active la casilla de verificación **Mostrar propiedades avanzadas (todas)**.
4. Cambie las propiedades siguientes:

- Cambie el valor de `DataMining\AllowAdHocOpenRowsetQueries` a `Verdadero` (el valor predeterminado es `Falso`).
- Cambie el valor de `DataMining\AllowProvidersInOpenRowset` a `[all]` (no hay valor predeterminado).

Creación de un DSN ODBC para SQL Server

Para leer una base de datos o escribir en ella, es necesario tener un origen de datos ODBC instalado y configurado para la base de datos en cuestión, así como contar con los permisos de lectura o escritura necesarios para ello. El controlador ODBC de Microsoft SQL Native Client es necesario y se instala automáticamente con SQL Server. *No se recomienda utilizar en este caso el controlador proporcionado con IBM SPSS Data Access Pack (y que normalmente se recomienda para otros usos con IBM SPSS Modeler).* Si IBM SPSS Modeler y SQL Server se encuentran en hosts diferentes, puede descargar el controlador ODBC de Microsoft SQL Native Client. Para obtener más información, consulte el tema "Requisitos para la integración con Microsoft Analysis Services" en la página 12.

Si tiene alguna pregunta acerca de la creación o configuración de permisos de los orígenes de datos ODBC, póngase en contacto con el administrador de la base de datos.

1. Con dicho controlador ODBC de Microsoft SQL Native Client, cree un DNS ODBC que señale la base de datos de SQL Server utilizada en el proceso de minería de datos. Se deben usar el resto de valores de configuración predeterminada del controlador.
2. Para este DSN, asegúrese de que **Con autenticación integrada de Windows** está seleccionada.

- Si IBM SPSS Modeler e IBM SPSS Modeler Server se están ejecutando en hosts distintos, cree el mismo DSN ODBC en cada uno de ellos. Asegúrese de que se usa el mismo nombre DSN en cada host.

Activación de la integración con Analysis Services en IBM SPSS Modeler

Para activar IBM SPSS Modeler para que utilice Analysis Services, necesitará proporcionar antes algunas especificaciones del servidor en el cuadro de diálogo Aplicaciones de ayuda.

1. Seleccione en los menús de IBM SPSS Modeler:
   Herramientas > **Opciones** > **Aplicaciones de ayuda**
2. Pulse en la pestaña **Microsoft**.

- **Activar integración de Microsoft Analysis Services.** Habilita la paleta de modelado de bases de datos (si todavía no se muestra) en la parte inferior de la ventana de IBM SPSS Modeler y añade los nodos para algoritmos de Analysis Services.
- **Host del servidor de análisis.** Especifique el nombre del equipo en el que se está ejecutando Analysis Services.

- **Base de datos del servidor de análisis.** Seleccione la base de dtos que desee pulsando en el botón de puntos suspensivos (...) para abrir un cuadro de subdiálogo en el que puede elegir una de las bases de datos disponibles. La lista contiene bases de datos disponibles para el servidor de Analysis Server especificado. Como Microsoft Analysis Services almacena los modelos de minería de datos en bases de datos con nombre, debe seleccionar la base de datos adecuada en la que se almacenan los modelos de Microsoft creados por IBM SPSS Modeler.
- **Conexión de SQL Server.** Especifique la información de DSN utilizada por la base de datos de SQL Server para almacenar los datos que pasan al servidor de Analysis Server. Seleccione el origen de datos ODBC que se va a utilizar para proporcionar los datos de la generación de los modelos de minería de datos de Analysis Services. Si está generando modelos de Analysis Services a partir de datos proporcionados en archivos planos u orígenes de datos ODBC, los datos se cargarán automáticamente en una tabla temporal creada en la base de datos de SQL Server a la que apunta este origen de datos ODBC.
- **Avisar cuando haya de sobrescribirse un modelo de minería de datos.** Seleccione esta función para asegurarse de que IBM SPSS Modeler no sobrescribe los modelos almacenados en la base de datos sin avisar con anterioridad.

Nota: la configuración establecida en el cuadro de diálogo Aplicaciones de ayuda se puede sobrescribir en los diversos nodos de Analysis Services.

Activación de optimización y generación de SQL

1. Seleccione en los menús de IBM SPSS Modeler:

   Herramientas > **Propiedades de ruta** > **Opciones**
2. Pulse en la opción **Optimización** en el panel de navegación.
3. Confirme que la opción **Generar SQL** está activada. Esta configuración es necesaria para que el modelado de bases de datos funcione.
4. Seleccione las opciones **Optimizar generación de SQL** y **Optimizar otra ejecución** (no obligatorias, pero recomendadas para un rendimiento optimizado).

Generación de modelos con Analysis Services

La generación de modelos con Analysis Services requiere que el conjunto de datos de entrenamiento se encuentre en una tabla o vista dentro de la base de datos de SQL Server. Si los datos no se encuentran en SQL Server o se tienen que procesar en IBM SPSS Modeler como parte de la preparación de datos que no se puede realizar en SQL Server, los datos se cargarán automáticamente en una tabla temporal de SQL Server antes de la generación de modelos.

Gestión de modelos de Analysis Services

La generación de un modelo de Analysis Services mediante IBM SPSS Modeler implica la generación de un modelo en IBM SPSS Modeler y la generación o sustitución de un modelo en la base de datos de SQL Server. El modelo de IBM SPSS Modeler hace referencia al contenido de un modelo de base de datos almacenado en un servidor de base de datos. IBM SPSS Modeler puede realizar la comprobación de la coherencia almacenando una cadena clave del modelo generada idéntica, tanto en el modelo de IBM SPSS Modeler como en el de SQL Server.

El nodo de modelado **Árbol de decisión de MS** se utiliza en modelado predictivo tanto en atributos continuos como categóricos. Para los atributos categóricos, el nodo hace predicciones basadas en las relaciones entre columnas de entradas de un conjunto de datos. Por ejemplo, en un escenario en el que se va a predecir qué clientes tienen mayor probabilidad de comprar una bicicleta, si nueve de cada diez clientes jóvenes compran una bicicleta pero solamente dos de cada diez clientes mayores lo hacen, el nodo infiere que la edad es un buen predictor de la compra de bicicletas. El árbol de decisión realiza predicciones según esta tendencia hacia un resultado en particular. Para atributos continuos, el algoritmo utiliza la regresión lineal para determinar dónde se divide un árbol de decisión. Si se establece más de una columna como predecible o si los datos de entrada contienen una tabla anidada que se establece como predecible, el nodo crea un árbol de decisión diferente para cada columna predecible.

El nodo de modelado **Agrupación en clústeres de MS** utiliza técnicas iterativas para agrupar casos de conjuntos de datos en clústeres que contienen características similares. Estos grupos resultan útiles para explorar datos, identificar anomalías en los datos y crear predicciones. Los modelos de clúster identifican relaciones en una base de datos que es posible que no pueda derivar de forma lógica mediante la observación informal. Por ejemplo, puede deducir de manera lógica que la gente que va al trabajo en bicicleta no suele vivir lejos de su lugar de trabajo. Sin embargo, el algoritmo puede detectar otras características sobre las personas que van en bicicleta que no son tan obvias. El nodo de agrupación en clústeres se distingue de otros modelos de minería de datos en que no se especifica ningún campo objetivo. El nodo de agrupación en clústeres entrena el modelo de estrictamente a partir de las relaciones que existen en los datos y de los clústeres que identifica el nodo.

El nodo de modelado **Reglas de asociación de MS** es útil para motores de recomendación. El modelo de recomendación recomienda productos a los clientes basándose en los elementos que ya se han adquirido o en los que se ha indicado un interés. Los modelos de asociación se crean en conjuntos de datos que contienen identificadores, tanto para casos individuales como para los elementos que contienen los casos. El grupo de elementos de un caso se denomina **conjunto de elementos**. Los modelos de asociación se componen de una serie de conjuntos de elementos y las reglas que describen la forma de agruparse enq los casos. Las reglas que identifica el algoritmo se pueden utilizar para predecir las posibles adquisiciones futuras del cliente en función de los elementos que ya existen en el carro de la compra del mismo.

El nodo de modelado **bayesiano ingenuo de MS** calcula la probabilidad condicional entre los campos objetivo y predictor y asume que las columnas son independientes. El modelo se denomina ingenuo porque trata todas las variables de predicción propuestas como independientes unas de otras. Este método es computacionalmente menos intenso que otros algoritmos de Analysis Services y, por lo tanto, resulta útil para descubrir con rapidez relaciones durante las etapas preliminares del modelado. Puede utilizar este nodo para realizar exploraciones iniciales de los datos y, a continuación, aplicar los resultados para crear modelos adicionales con otros nodos que puedan tardar más en calcularse pero ofrecen unos resultados más precisos.

El nodo de modelado **Regresión lineal de MS** es una variación del nodo Árboles de decisión, donde el parámetro `MINIMUM_LEAF_CASES` está establecido para que sea mayor o igual que el número total de casos del conjunto de datos que utiliza el nodo para entrenar el modelo de minería. Con el parámetro establecido de esta forma, el nodo nunca creará una división y realizará, por tanto, una regresión lineal.

El nodo de modelado **Red neuronal de MS** se parece al nodo Árbol de decisión de MS en que el nodo Red neuronal de MS calcula probabilidades para cada estado posible del atributo de entrada cuando se proporcionan todos los estados del atributo predecible. Más adelante se pueden utilizar estas probabilidades para predecir un resultado del atributo predicho basado en los atributos de entrada.

El nodo de modelado **Regresión logística de MS** es una variación del nodo Red neuronal de MS, donde el parámetro HIDDEN_NODE_RATIO está establecido como 0. Este ajuste crea un modelo de red neuronal que no contiene una capa oculta y, por tanto, equivale a la regresión logística.

El nodo de modelado **Serie temporal de MS** proporciona algoritmos de regresión optimizados para la previsión de valores continuos, como ventas de productos en el tiempo. Mientras que los algoritmos de Microsoft, como los árboles de decisión requieren más columnas de nueva información como entrada para predecir una tendencia, un modelo de serie temporal no. Un modelo de serie temporal puede predecir tendencias basándose únicamente en el conjunto de datos original que se utiliza para crear el modelo. También puede agregar nuevos datos al modelo cuando realice una predicción e incorporar automáticamente los nuevos datos en el análisis de tendencias. Para obtener más información, consulte el tema "Nodo Serie temporal de MS" en la página 19.

El nodo de modelado **Agrupación en clústeres de secuencias de MS** de identifica las secuencias ordenadas en datos y combina los resultados de este análisis con técnicas de agrupación en clústeres para generar clústeres basados en las secuencias y otros atributos. Para obtener más información, consulte el tema "Nodo Agrupación en clústeres de secuencias de MS" en la página 21.

Puede acceder a cada nodo desde la paleta Modelado de bases de datos de la parte inferior de la ventana de IBM SPSS Modeler.

Configuración común a todos los nodos de algoritmo

La configuración de las siguientes opciones es común a todos los algoritmos de Analysis Services:

Opciones de Servidor

En la pestaña Servidor se pueden configurar la base de datos y el host de Analysis Server y el origen de datos de SQL Server. Las opciones especificadas aquí sobrescriben las especificadas en la pestaña Microsoft del cuadro de diálogo Aplicaciones de ayuda. Para obtener más información, consulte el tema "Activación de la integración con Analysis Services" en la página 13.

Nota: también existe una variación de esta pestaña disponible al puntuar modelos de Analysis Services. Para obtener más información, consulte el tema "Pestaña Servidor de modelo generado de Analysis Services" en la página 22.

Opciones de modelo

Para generar el modelo más básico, debe especificar las opciones de la pestaña Modelo antes de continuar. El método de puntuación y otras opciones avanzadas están disponibles en la pestaña Experto.

Están disponibles las siguientes opciones de modelado básico:

Nombre del modelo. Especifica el nombre asignado al modelo creado al ejecutar el nodo.

- **Automático.** Genera el nombre del modelo de forma automática basándose en los nombres de los campos objetivo o de ID, o en el nombre del tipo de modelo en los casos en los que no se especifique ningún campo objetivo (como en los modelos de agrupación en clústeres).
- **Personalizado.** Permite especificar un nombre personalizado para el modelo creado.

Utilizar los datos en particiones. Divide los datos en muestras o subconjuntos independientes para entrenamiento, comprobación y validación, en función del campo de partición actual. Si se utiliza una muestra para crear el modelo y otra muestra independiente para comprobarlo, se puede obtener una

indicación de la bondad del modelo a la hora de generalizar conjuntos de datos de mayor tamaño similares a los datos actuales. Si no se especifica ningún campo de partición en la ruta, se ignorará esta opción.

Con exploración. Si se muestra, esta opción permite consultar el modelo para obtener información acerca de los casos que se incluyen.

Campo exclusivo. En la lista desplegable, seleccione un campo que identifique de manera exclusiva cada caso. Normalmente es un campo de ID, como por ejemplo **CustomerID**.

Opciones de Experto para los árboles de decisión de MS

Las opciones disponibles en la pestaña Experto pueden fluctuar dependiendo de la estructura de la ruta seleccionada. Consulte la ayuda de nivel de campo de la interfaz de usuario para obtener información detallada sobre las opciones de experto para el nodo de modelo de Analysis Services seleccionado.

Opciones de Experto para los clústeres de MS

Las opciones disponibles en la pestaña Experto pueden fluctuar dependiendo de la estructura de la ruta seleccionada. Consulte la ayuda de nivel de campo de la interfaz de usuario para obtener información detallada sobre las opciones de experto para el nodo de modelo de Analysis Services seleccionado.

Opciones de experto para el bayesiano ingenuo de MS

Las opciones disponibles en la pestaña Experto pueden fluctuar dependiendo de la estructura de la ruta seleccionada. Consulte la ayuda de nivel de campo de la interfaz de usuario para obtener información detallada sobre las opciones de experto para el nodo de modelo de Analysis Services seleccionado.

Opciones de experto de regresión lineal de MS

Las opciones disponibles en la pestaña Experto pueden fluctuar dependiendo de la estructura de la ruta seleccionada. Consulte la ayuda de nivel de campo de la interfaz de usuario para obtener información detallada sobre las opciones de experto para el nodo de modelo de Analysis Services seleccionado.

Opciones de experto de red neuronal de MS

Las opciones disponibles en la pestaña Experto pueden fluctuar dependiendo de la estructura de la ruta seleccionada. Consulte la ayuda de nivel de campo de la interfaz de usuario para obtener información detallada sobre las opciones de experto para el nodo de modelo de Analysis Services seleccionado.

Opciones de experto de regresión logística de MS

Las opciones disponibles en la pestaña Experto pueden fluctuar dependiendo de la estructura de la ruta seleccionada. Consulte la ayuda de nivel de campo de la interfaz de usuario para obtener información detallada sobre las opciones de experto para el nodo de modelo de Analysis Services seleccionado.

Nodo Reglas de asociación de MS

El nodo de modelado Reglas de asociación de MS es útil para motores de recomendación. El modelo de recomendación recomienda productos a los clientes basándose en los elementos que ya se han adquirido o en los que se ha indicado un interés. Los modelos de asociación se crean en conjuntos de datos que contienen identificadores, tanto para casos individuales como para los elementos que contienen los casos. El grupo de elementos de un caso se denomina **conjunto de elementos**.

Los modelos de asociación se componen de una serie de conjuntos de elementos y las reglas que describen la forma de agruparse enq los casos. Las reglas que identifica el algoritmo se pueden utilizar para predecir las posibles adquisiciones futuras del cliente en función de los elementos que ya existen en el carro de la compra del mismo.

En el caso de datos con formato tabular, el algoritmo crea puntuaciones que representan la probabilidad (*campo* $MP) de cada recomendación generada (*campo* $M). En el caso de datos con formato transaccional, las puntuaciones se crean para compatibilidad (*campo* $MS), probabilidad (*campo* $MP) y probabilidad ajustada (*campo* $MAP) para cada recomendación generada (*campo* $M).

Requisitos

Los requisitos de un modelo de asociación transaccional son las siguientes:

- **Campo exclusivo.** Un modelo de reglas de asociación requiere una clave que identifique los registros de forma exclusiva.
- **Campo de ID.** Cuando se genera un modelo de reglas de asociación de MS con datos de formatos transaccionales, se requiere un campo de ID que identifique cada transacción. Los campos de ID se pueden definir igual que el campo exclusivo.
- **Al menos un campo de entrada.** El algoritmo de reglas de asociación requiere al menos un campo de entrada.
- **Campo objetivo.** Cuando crea un modelo de asociación de MS con datos transaccionales, el campo de destino debe ser el campo de transacción, por ejemplo, productos que ha comprado un usuario.

Opciones de Experto para las reglas de asociación de MS

Las opciones disponibles en la pestaña Experto pueden fluctuar dependiendo de la estructura de la ruta seleccionada. Consulte la ayuda de nivel de campo de la interfaz de usuario para obtener información detallada sobre las opciones de experto para el nodo de modelo de Analysis Services seleccionado.

Nodo Serie temporal de MS

El nodo de modelado de serie temporal de MS admite dos tipos de predicciones:

- futuras
- históricas

Predicciones futuras valores de campo objetivo estimados para un número específico de períodos temporales más allá del final de sus datos históricos, y se realizan siempre. **Predicciones históricas** son valores de campo objetivo estimados durante un número especificado de períodos temporales para los que tiene los valores reales en sus datos históricos. Puede utilizar predicciones históricas para evaluar la calidad del modelo, comparando los valores históricos reales con los valores predichos. El valor del punto inicial de las predicciones determina si se realizarán predicciones históricas.

Al contrario que el nodo de series temporales de IBM SPSS Modeler, el nodo Series Temporales de MS no necesita un nodo Intervalos temporales anterior. Otra diferencia es que, de forma predeterminada, las puntuaciones solamente se producen para las filas predichas, no solamente para todas las filas históricas de los datos de la serie temporal.

Requisitos

Los requisitos para un modelo de serie temporal de MS son los siguientes:

- **Campo temporal clave único.** Cada modelo debe contener un campo de fecha o numérico que se utiliza como la serie de casos, definiendo las porciones de tiempo que utilizará el modelo. El tipo de datos del campo temporal puede ser un tipo de dates de fecha/tiempo o un tipo de datos numérico. Sin embargo, el campo debe contener valores continuos y los valores deben ser exclusivos para cada serie.
- **Campo objetivo único.** Solamente puede especificar un campo objetivo en cada modelo. El tipo de datos del campo objetivo debe tener valores continuos. Por ejemplo, puede predecir cómo cambiarán con el tiempo los atributos numéricos, como ingresos, ventas o temperatura. Sin embargo, no puede utilizar un campo que contenga valores categóricos, como el estado de compra o nivel de educación, como el campo de destino.

- **Al menos un campo de entrada.** El algoritmo de serie temporal de MS requiere al menos un campo de entrada. El tipo de datos del campo de entrada debe tener valores continuos. Los campos de entrada no continuos se ignoran cuando genera el modelo.
- **El conjunto de datos se debe ordenar.** El conjunto de datos de entrada se debe clasificar (en el campo temporal clave), de lo contrario, el modelo se interrumpirá con un error.

Opciones del modelo de serie temporal de MS

Nombre del modelo. Especifica el nombre asignado al modelo creado al ejecutar el nodo.

- **Automático.** Genera el nombre del modelo de forma automática basándose en los nombres de los campos objetivo o de ID, o en el nombre del tipo de modelo en los casos en los que no se especifique ningún campo objetivo (como en los modelos de agrupación en clústeres).
- **Personalizado.** Permite especificar un nombre personalizado para el modelo creado.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo.

Con exploración. Si se muestra, esta opción permite consultar el modelo para obtener información acerca de los casos que se incluyen.

Campo exclusivo. En la lista desplegable, seleccione el campo de tiempo clave, que se utilizará para construir el modelo de serie temporal.

Opciones del experto de serie temporal de MS

Las opciones disponibles en la pestaña Experto pueden fluctuar dependiendo de la estructura de la ruta seleccionada. Consulte la ayuda de nivel de campo de la interfaz de usuario para obtener información detallada sobre las opciones de experto para el nodo de modelo de Analysis Services seleccionado.

Si está realizando predicciones históricas, el número de pasos históricos que pueden incluirse en el resultado de puntuación viene decidido por el valor de (HISTORIC_MODEL_COUNT * HISTORIC_MODEL_GAP). De forma predeterminada, este límite es 10, lo que significa que solamente se realizarán 10 predicciones históricas. En este caso, por ejemplo, se produce un error si introduce un valor inferior a -10 para **Predicción histórica** en la pestaña Configuraciones del nugget de modelo (consulte "Pestaña Configuración de nugget de modelo Serie temporal de MS" en la página 24). Si desea ver más predicciones históricas, puede aumentar el valor de HISTORIC_MODEL_COUNT o HISTORIC_MODEL_GAP, pero se aumentará el tiempo de creación del modelo.

Opciones del configuración de serie temporal de MS

Comenzar estimación. Especifique el periodo de tiempo en el que desea que empiecen las predicciones.

- **Iniciar desde: Nueva predicción.** El periodo de tiempo en el que desea que se inicien los futuras predicciones, expresado como un desplazamiento desde el último periodo de sus datos históricos. Por ejemplo, si sus datos históricos finalizaron en el 12/99 y desea que sus predicciones comiencen en 01/00, debería utilizar un valor de 1; sin embargo, si desea que las predicciones comiencen el 03/00, debería utilizar un valor de 3.
- **Iniciar desde: Predicción histórica.** El periodo de tiempo en el que desea que se inicien las predicciones históricas, expresado como un desplazamiento negativo desde el último periodo de sus datos históricos. Por ejemplo, si sus datos históricos finalizaron el 12/99 y desea realizar predicciones históricas de los cinco últimos periodos de tiempo de sus datos, debería utilizar un valor de -5.

Terminar estimación. Especifique el periodo de tiempo en el que desea que terminen las predicciones.

- **Paso final de la predicción.** El periodo de tiempo en el que desea que se terminen los futuras predicciones, expresado como un desplazamiento desde el último periodo de sus datos históricos. Por ejemplo, si sus datos históricos terminan el 12/99 y desea que sus predicciones terminen el 6/00, debería utilizar un valor de 6. Para predicciones futuras, el valor debe ser siempre mayor o igual que el valor **Iniciar desde**.

Nodo Agrupación en clústeres de secuencias de MS

El nodo Clúster de secuencia de MS utiliza un algoritmo de análisis de secuencia que explora los datos que contienen eventos que pueden enlazarse por las siguientes rutas, o *secuencias*. Algunos de los ejemplos pueden ser las rutas creadas cuando los usuarios navegan o exploran un sitio Web o el orden en que un cliente añade elementos a un carrito de la compra en una tienda en línea. El algoritmo busca las secuencias más comunes agrupando o *agrupando en clústeres*, secuencias idénticas.

Requisitos

Los requisitos de un modelo de agrupación en clústeres de secuencias de Microsoft son los siguientes:

- **Campo de ID.** El algoritmo de Agrupación en clústeres de secuencias de Microsoft requiere que la información de la secuencia se guarde en formato transaccional. Para ello, se requiere un campo de ID que identifique cada transacción.
- **Al menos un campo de entrada.** El algoritmo requiere al menos un campo de entrada.
- **Campo Secuencia.** El algoritmo también requiere un campo identificador de secuencias, que debe tener un nivel de medición de Continuo. Por ejemplo, puede utilizar un identificador de páginas Web, un entero o una cadena de texto, siempre que el campo identifique los eventos en una secuencia. Solamente se permite un identificador de secuencia para cada secuencia y solamente se permite un tipo de secuencia en cada modelo. El campo Secuencia debe ser diferente de los campos ID y Exclusivo.
- **Campo objetivo.** Se requiere un campo objetivo cuando genere un modelo de agrupación en clústeres de secuencias.
- **Campo exclusivo.** Un modelo de agrupación en clústeres de secuencias requiere una clave que identifique los registros de forma exclusiva. Puede definir el campo Exclusivo con el mismo valor que el campo ID.

Opciones de campos de agrupación en clústeres de secuencias de MS

Todos los nodos de modelado tienen una pestaña Campos en la que se pueden especificar los campos que se usarán para generar el modelo.

Para generar un modelo de agrupación en clústeres de secuencias, se deben especificar los campos que se desea usar como objetivos y como entradas. Tenga en cuenta que el nodo Clúster de secuencia de MS, no puede utilizar información de campo de un nodo Tipo anterior; debe especificar los ajustes del campo aquí.

ID. Seleccione un campo de ID de la lista. Los campos numéricos o simbólicos se pueden utilizar como campo de ID. Cada valor exclusivo de este campo debe indicar una unidad de análisis específica. Por ejemplo, en una aplicación de la cesta de la compra, cada ID puede representar a un sólo cliente. Para una aplicación de análisis del registro Web, cada ID puede representar un equipo (con la dirección IP) o un usuario (con los datos de inicio de sesión).

Entradas. Seleccione el campo(s) de entrada para el modelo. Estos son los campos que contienen los eventos de interés del modelado de secuencia.

Secuencia. Seleccione un campo de la lista, que se utilizará como el campo identificador de la secuencia. Por ejemplo, puede utilizar un identificador de páginas Web, un entero o una cadena de texto, siempre que el campo identifique los eventos en una secuencia. Solamente se permite un identificador de secuencia para cada secuencia y solamente se permite un tipo de secuencia en cada modelo. El campo Secuencia debe ser diferente del campo ID (especificado en esta pestaña) y el campo Exclusivo (especificado en la pestaña Modelo).

Objetivo. Seleccione un campo que se utilizará como campo objetivo, o sea, el campo cuyo valor está intentando predecir está basado en datos de secuencias.

Opciones de Experto para los clústeres de secuencias de MS

Las opciones disponibles en la pestaña Experto pueden fluctuar dependiendo de la estructura de la ruta seleccionada. Consulte la ayuda de nivel de campo de la interfaz de usuario para obtener información detallada sobre las opciones de experto para el nodo de modelo de Analysis Services seleccionado.

Puntuación de modelos de Analysis Services

La puntuación de modelos tiene lugar en SQL Server y se lleva a cabo mediante Analysis Services. Es posible que sea necesario cargar el conjunto de datos en una tabla temporal si los datos se originan dentro de IBM SPSS Modeler o es necesario prepararlos dentro de IBM SPSS Modeler. Los modelos que se crean desde IBM SPSS Modeler utilizando la minería interna de bases de datos son realmente un modelo remoto que se guarda en el servidor de la base de datos o la minería de datos remota. Es importante comprender esta distinción cuando se examinan y puntúan modelos creados con algoritmos de Microsoft Analysis Services.

En IBM SPSS Modeler, generalmente, solamente se proporciona una única predicción y una confianza o probabilidad asociada.

Para obtener ejemplos de puntuación de modelos, consulte "Ejemplos de minería de datos con Analysis Services" en la página 25.

Configuración común a todos los modelos de Analysis Services

La configuración de las siguientes opciones es común a todos los modelos de Analysis Services:

Pestaña Servidor de modelo generado de Analysis Services

La pestaña Servidor se utiliza para especificar las conexiones para la minería interna de bases de datos. La pestaña también proporciona la clave de modelo exclusiva. Esta clave se genera aleatoriamente cuando el modelo se crea y se almacena tanto en el modelo de IBM SPSS Modeler como en la descripción del objeto de modelo almacenado en la base de datos de Analysis Services.

En la pestaña Servidor se pueden configurar la base de datos y el host del servidor de análisis, así como el origen de datos de SQL Server para la operación de puntuación. Las opciones especificadas aquí sobrescriben a las especificadas en los cuadros de diálogo Aplicaciones de ayuda o Generar modelo de IBM SPSS Modeler. Para obtener más información, consulte el tema "Activación de la integración con Analysis Services" en la página 13.

GUID de modelo. Aquí se muestra la clave del modelo. Esta clave se genera aleatoriamente cuando el modelo se crea y se almacena tanto en el modelo de IBM SPSS Modeler como en la descripción del objeto de modelo almacenado en la base de datos de Analysis Services.

Comprobar. Pulse en este botón para verificar la clave del modelo con respecto a la clave del modelo almacenado en la base de datos de Analysis Services. De este modo podrá verificar que el modelo aún existe en el servidor de análisis e indica que su estructura no ha cambiado.

Nota: el botón Comprobar está disponible sólo para los modelos añadidos al lienzo de rutas en la preparación para la puntuación. Si se produce un error al realizar la comprobación, investigue si el modelo ha sido eliminado o sustituido por otro modelo en el servidor.

Ver. Pulse para obtener una vista gráfica del modelo de árbol de decisión. El visor de árboles de decisión lo comparten todos los algoritmos de árboles de decisión de IBM SPSS Modeler, y su funcionalidad es idéntica.

Pestaña Resumen de modelo generado de Analysis Services

La pestaña Resumen de un nugget de modelo muestra información sobre el propio modelo (*Análisis*), los campos utilizados en el modelo (*Campos*), la configuración utilizada al generar el modelo (*Configuración de creación*) y el entrenamiento del modelo (*Resumen de entrenamiento*).

Cuando se examina el nodo por primera vez, los resultados de la pestaña Resumen aparecen contraídos. Para ver los resultados de interés, utilice el control de expansión situado a la izquierda de un elemento con objeto de desplegarlo, o bien pulse en el botón **Expandir todo** para mostrar todos los resultados. Para ocultar los resultados cuando haya terminado de consultarlos, utilice el control de expansión con objeto de contraer los resultados específicos que desee ocultar o pulse en el botón **Contraer todo** para contraer todos los resultados.

Análisis. Muestra información sobre el modelo específico. Si ha ejecutado un nodo Análisis conectado a este nugget de modelo, la información de dicho análisis también se mostrará en esta sección.

Campos. Enumera los campos utilizados como objetivo y entradas en la generación del modelo.

Configuración de creación. Contiene información sobre la configuración que se utiliza en la generación del modelo.

Resumen de entrenamiento. Muestra el tipo del modelo, la ruta utilizada para crearlo, el usuario que lo creó, cuándo se generó y el tiempo que se tardó en generar el modelo.

Nugget de modelo Serie temporal de MS

El modelo de Serie temporal de MS genera puntuaciones solamente para los períodos de tiempo predichos, no para los datos históricos.

La tabla siguiente muestra los campos que se añaden al modelo.

Tabla 1. Campos añadidos al modelo

Nombre del campo	Descripción
$M-*campo*	Valor predicho del *campo*
$Var-*campo*	Varianza calculada de *campo*
$Stdev-*campo*	Desviación estándar de *campo*

Pestaña Servidor de nugget de modelo Serie temporal de MS

La pestaña Servidor se utiliza para especificar las conexiones para la minería interna de bases de datos. La pestaña también proporciona la clave de modelo exclusiva. Esta clave se genera aleatoriamente cuando el modelo se crea y se almacena tanto en el modelo de IBM SPSS Modeler como en la descripción del objeto de modelo almacenado en la base de datos de Analysis Services.

En la pestaña Servidor se pueden configurar la base de datos y el host del servidor de análisis, así como el origen de datos de SQL Server para la operación de puntuación. Las opciones especificadas aquí sobrescriben a las especificadas en los cuadros de diálogo Aplicaciones de ayuda o Generar modelo de IBM SPSS Modeler. Para obtener más información, consulte el tema “Activación de la integración con Analysis Services” en la página 13.

GUID de modelo. Aquí se muestra la clave del modelo. Esta clave se genera aleatoriamente cuando el modelo se crea y se almacena tanto en el modelo de IBM SPSS Modeler como en la descripción del objeto de modelo almacenado en la base de datos de Analysis Services.

Comprobar. Pulse en este botón para verificar la clave del modelo con respecto a la clave del modelo almacenado en la base de datos de Analysis Services. De este modo podrá verificar que el modelo aún existe en el servidor de análisis e indica que su estructura no ha cambiado.

Nota: el botón Comprobar está disponible sólo para los modelos añadidos al lienzo de rutas en la preparación para la puntuación. Si se produce un error al realizar la comprobación, investigue si el modelo ha sido eliminado o sustituido por otro modelo en el servidor.

Ver. Pulse para obtener una vista gráfica del modelo de Serie temporal. Analysis Services muestra el modelo completo como un árbol. También puede ver un gráfico que muestre el valor histórico del campo objetivo con el paso del tiempo, junto con los valores futuros predichos.

Para obtener más información, consulte la descripción del visor de Serie temporal en la biblioteca MSDN en http://msdn.microsoft.com/en-us/library/ms175331.aspx.

Pestaña Configuración de nugget de modelo Serie temporal de MS

Comenzar estimación. Especifique el periodo de tiempo en el que desea que empiecen las predicciones.

- **Iniciar desde: Nueva predicción.** El periodo de tiempo en el que desea que se inicien los futuras predicciones, expresado como un desplazamiento desde el último periodo de sus datos históricos. Por ejemplo, si sus datos históricos finalizaron en el 12/99 y desea que sus predicciones comiencen en 01/00, debería utilizar un valor de 1; sin embargo, si desea que las predicciones comiencen el 03/00, debería utilizar un valor de 3.
- **Iniciar desde: Predicción histórica.** El periodo de tiempo en el que desea que se inicien las predicciones históricas, expresado como un desplazamiento negativo desde el último periodo de sus datos históricos. Por ejemplo, si sus datos históricos finalizaron el 12/99 y desea realizar predicciones históricas de los cinco últimos periodos de tiempo de sus datos, debería utilizar un valor de -5.

Terminar estimación. Especifique el periodo de tiempo en el que desea que terminen las predicciones.

- **Paso final de la predicción.** El periodo de tiempo en el que desea que se terminen los futuras predicciones, expresado como un desplazamiento desde el último periodo de sus datos históricos. Por ejemplo, si sus datos históricos terminan el 12/99 y desea que sus predicciones terminen el 6/00, debería utilizar un valor de 6. Para predicciones futuras, el valor debe ser siempre mayor o igual que el valor **Iniciar desde**.

Nugget de modelo de clúster de secuencias de MS

La tabla siguiente muestra los campos que se añaden al modelo de agrupación en clústeres de secuencias de MS (donde *campo* es el nombre del campo objetivo).

Tabla 2. Campos añadidos al modelo

Nombre del campo	Descripción
$MC-*campo*	Predicción del clúster al que pertenece la secuencia.
$MCP-*campo*	Probabilidad de que esta secuencia pertenezca al clúster predicho.
$MS-*campo*	Valor predicho del *campo*
$MSP-*campo*	Probabilidad de que el valor de $MS-*campo* sea correcto.

Exportación de modelos y generación de nodos

Puede exportar una estructura y un resumen del modelo a archivos con formato de texto y HTML. Puede generar los nodos Seleccionar y Filtrar adecuados donde proceda.

Al igual que otros nuggets de modelo de IBM SPSS Modeler, los nuggets de modelos de Microsoft Analysis Services admiten la generación directa de nodos de operaciones con campos y registros. Utilizando las opciones del menú Generar del nugget de modelo, puede generar los siguientes nodos:

- Nodo Seleccionar (solamente si se selecciona un elemento en la pestaña Modelo)
- nodo Filtrar

Ejemplos de minería de datos con Analysis Services

Se incluyen varias rutas de ejemplo que ejemplifican el uso de la minería de datos de MS Analysis Services con IBM SPSS Modeler. Estas rutas se encuentran en la carpeta de instalación de IBM SPSS Modeler en:

\Demos\Database_Modelling\Microsoft

Nota: se puede acceder a la carpeta Demos desde el grupo de programas IBM SPSS Modeler en el menú Inicio de Windows.

Rutas de ejemplo: Árboles de decisión

Las siguientes rutas pueden utilizarse juntas en una secuencia como un ejemplo del proceso de minería de bases de datos mediante el algoritmo de árboles de decisión proporcionado por MS Analysis Services.

Tabla 3. Árboles de decisión - rutas de ejemplo

Ruta	Descripción
1_upload_data.str	Se utiliza para depurar y cargar los datos desde un archivo sin formato a la base de datos.
2_explore_data.str	Ofrece un ejemplo de exploración de los datos con IBM SPSS Modeler.
3_build_model.str	Genera el modelo mediante el algoritmo nativo de base de datos.
4_evaluate_model.str	Se utiliza como ejemplo de evaluación del modelo con IBM SPSS Modeler.
5_deploy_model.str	Se utiliza para desplegar el modelo para la puntuación interna de la base de datos.

Nota: para poder ejecutar el ejemplo, las rutas se deben procesar en orden. Además, los nodos de origen y modelado de cada ruta se deben actualizar para que hagan referencia a un origen de datos válido para la base de datos que desee utilizar.

El conjunto de datos utilizado en las rutas de ejemplo está relacionado con aplicaciones de tarjetas de crédito y presenta un problema de clasificación con una mezcla de predictores continuos y categóricos. Si desea obtener más información acerca de este conjunto de datos, consulte el archivo *crx.names* ubicado en la misma carpeta que las rutas de ejemplo.

Este conjunto de datos se puede descargar desde UCI Machine Learning Repository, en *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/*.

Ruta de ejemplo: cargar datos

La primera ruta de ejemplo, *1_upload_data.str*, se utiliza para depurar y cargar los datos desde un archivo sin formato a SQL Server.

Puesto que la minería de datos de Analysis Services necesita un campo de clave, esta ruta inicial utiliza un nodo Derivar para añadir un nuevo campo al conjunto de datos llamado *KEY* con los valores exclusivos *1,2,3* utilizando la función `@INDEX` de IBM SPSS Modeler.

El siguiente nodo Rellenar posterior se utiliza para gestionar los valores perdidos y sustituye los campos vacíos de lectura del archivo de texto *crx.data* por valores *NULOS*.

Ruta de ejemplo: explorar datos

La segunda ruta de ejemplo, *2_explore_data.str*, se utiliza para demostrar el uso del nodo Auditoría de datos para conocer los conceptos básicos de los datos, incluyendo estadísticos de resumen y gráficos.

Si pulsa dos veces en un gráfico en el informe de auditoría de datos, aparecerá un gráfico más detallado donde podrá explorar un campo específico con más profundidad.

Ruta de ejemplo: generar modelo

La tercera ruta de ejemplo, *3_build_model.str*, ilustra la generación del modelo en IBM SPSS Modeler. Puede adjuntar el modelo de la base de datos a la ruta y pulsar dos veces en él para especificar la configuración de generación.

En la pestaña Modelo del cuadro de diálogo puede especificar lo siguiente:

1. Seleccione **Clave** como campo de ID exclusivo.

En la pestaña Experto puede ajustar con precisión la configuración de generación del modelo.

Antes de la ejecución, asegúrese de que ha especificado la base de datos correcta para la generación del modelo. Utilice la pestaña Servidor para ajustar cualquier configuración.

Ruta de ejemplo: evaluar modelo

La cuarta ruta de ejemplo, *4_evaluate_model.str*, ilustra las ventajas de utilizar IBM SPSS Modeler para el modelado interno de bases de datos. Una vez ejecutado el modelo, puede volver a añadirlo a la ruta de datos y evaluarlo con varias herramientas que se ofrecen en IBM SPSS Modeler.

Consulta de los resultados del modelado

Puede pulsar dos veces en el nugget de modelo para examinar los resultados. La pestaña Resumen ofrece una vista de árbol-regla de los mismos. También puede pulsar en el botón **Ver** (que se encuentra en la pestaña Servidor) para obtener una vista gráfica del modelo de árbol de decisiones.

Evaluación de los resultados del modelo

El nodo Análisis de la ruta de ejemplo crea una matriz de coincidencias que muestre el patrón de coincidencias entre cada campo predicho y su campo objetivo. Ejecute el nodo Análisis para ver los resultados.

El nodo Evaluación de la ruta de ejemplo puede crear un gráfico de elevación diseñado para mostrar las mejoras de precisión realizadas por el modelo. Ejecute el nodo Evaluación para ver los resultados.

Ruta de ejemplo: desplegar modelo

Una vez satisfecho con la precisión del modelo, puede desplegarlo para su uso con aplicaciones externas o para volver a publicarlo en la base de datos. En la ruta de ejemplo final, *5_deploy_model.str*, se leen los datos desde la tabla `CREDIT` y, a continuación, se puntúan y se publican en la tabla `CREDITSCORES` mediante el nodo Base de datos.

La ejecución de la ruta genera el siguiente código SQL:

```
DROP TABLE CREDITSCORES

CREATE TABLE CREDITSCORES ( "field1" varchar(1),"field2" varchar(255),"field3" f
loat,"field4" varchar(1),"field5" varchar(2),"field6" varchar(2),"field7" varcha
r(2),"field8" float,"field9" varchar(1),"field10" varchar(1),"field11" int,"fiel
d12" varchar(1),"field13" varchar(1),"field14" int,"field15" int,"field16" varch
ar(1),"KEY" int,"$M-field16" varchar(9),"$MC-field16" float )

INSERT INTO CREDITSCORES ("field1","field2","field3","field4","field5","field6","field7","field8",
"field9","field10","field11","field12","field13","field14","field15","field16",
"KEY","$M-field16","$MC-field16")
SELECT T0.C0 AS C0,T0.C1 AS C1,T0.C2 AS C2,T0.C3 AS C3,T0.C4 AS C4,T0.C5 AS C5,
 T0.C6 AS C6,T0.C7 AS C7,T0.C8 AS C8,T0.C9 AS C9,T0.C10 AS C10,
```

```
 T0.C11 AS C11,T0.C12 AS C12,T0.C13 AS C13,T0.C14 AS C14,
 T0.C15 AS C15,T0.C16 AS C16,T0.C17 AS C17,T0.C18 AS C18
FROM (
 SELECT CONVERT(NVARCHAR,[TA].[field1])  AS C0, CONVERT(NVARCHAR,[TA].[field2])  AS C1,
  [TA].[field3] AS C2, CONVERT(NVARCHAR,[TA].[field4])  AS C3,
  CONVERT(NVARCHAR,[TA].[field5])  AS C4, CONVERT(NVARCHAR,[TA].[field6]) AS C5,
  CONVERT(NVARCHAR,[TA].[field7])  AS C6, [TA].[field8] AS C7,
  CONVERT(NVARCHAR,[TA].[field9])  AS C8, CONVERT(NVARCHAR,[TA].[field10])  AS C9,
  [TA].[field11] AS C10, CONVERT(NVARCHAR,[TA].[field12])  AS C11,
  CONVERT(NVARCHAR,[TA].[field13])  AS C12, [TA].[field14] AS C13,
  [TA].[field15] AS C14, CONVERT(NVARCHAR,[TA].[field16])  AS C15,
  [TA].[KEY] AS C16, CONVERT(NVARCHAR,[TA].[$M-field16])  AS C17,
  [TA].[$MC-field16] AS C18
 FROM openrowset('MSOLAP',
  'Datasource=localhost;Initial catalog=FoodMart 2000',
  'SELECT [T].[C0] AS [field1],[T].[C1] AS [field2],[T].[C2] AS [field3],
   [T].[C3] AS [field4],[T].[C4] AS [field5],[T].[C5] AS [field6],
   [T].[C6] AS [field7],[T].[C7] AS [field8],[T].[C8] AS [field9],
   [T].[C9] AS [field10],[T].[C10] AS [field11],[T].[C11] AS [field12],
   [T].[C12] AS [field13],[T].[C13] AS [field14],[T].[C14] AS [field15],
   [T].[C15] AS [field16],[T].[C16] AS [KEY],[CREDIT1].[field16] AS [$M-field16],
   PredictProbability([CREDIT1].[field16]) AS [$MC-field16]
  FROM [CREDIT1] PREDICTION JOIN
   openrowset(''MSDASQL'',
   ''Dsn=LocalServer;Uid=;pwd='',''SELECT T0."field1" AS C0,T0."field2" AS C1,
   T0."field3" AS C2,T0."field4" AS C3,T0."field5" AS C4,T0."field6" AS C5,
   T0."field7" AS C6,T0."field8" AS C7,T0."field9" AS C8,T0."field10" AS C9,
   T0."field11" AS C10,T0."field12" AS C11,T0."field13" AS C12,
   T0."field14" AS C13,T0."field15" AS C14,T0."field16" AS C15,
   T0."KEY" AS C16 FROM "dbo".CREDITDATA T0'') AS [T]
  ON [T].[C2] = [CREDIT1].[field3] and [T].[C7] = [CREDIT1].[field8]
   and [T].[C8] = [CREDIT1].[field9] and [T].[C9] = [CREDIT1].[field10]
   and [T].[C10] = [CREDIT1].[field11] and [T].[C11] = [CREDIT1].[field12]
   and [T].[C14] = [CREDIT1].[field15]') AS [TA]
)  T0
```

Capítulo 4. Modelado de bases de datos con Oracle Data Mining

Acerca de Oracle Data Mining

IBM SPSS Modeler admite la integración con Oracle Data Mining (ODM) de Oracle, lo cual proporciona una familia de algoritmos de minería de datos incrustados con gran cohesión en Oracle RDBMS. Es posible acceder a estas características mediante la interfaz gráfica del usuario de IBM SPSS Modeler y el entorno de desarrollo orientados al flujo de trabajo, lo que permite a los clientes aumentar los algoritmos de minería de datos ofrecidos por ODM.

IBM SPSS Modeler admite la integración de los siguientes algoritmos desde Oracle Data Mining:

- bayesiano ingenuo
- Bayesiano adaptativo
- Máquina de vectores de soporte (SVM)
- Modelos lineales generalizados (GLM)*
- Árbol de decisiones
- O-clúster
- k-medias
- Factorización de matrices no negativas (NMF)
- Apriori
- Longitud mínima de la descripción (LMD)
- Importancia del atributo (AI)

* 11*g* R1 únicamente

Requisitos para la integración con Oracle

Las siguientes condiciones constituyen los requisitos previos para realizar el modelado interno de bases de datos mediante Oracle Data Mining. Es posible que necesite consultar con el administrador de la base de datos para asegurarse de que se reúnen las condiciones.

- IBM SPSS Modeler ejecutándose en modo local o en una instalación de IBM SPSS Modeler Server en Windows o UNIX.
- Oracle 10*g*R2 ó 11*g*R1 (10.2 Database o posterior) con la opción de Oracle Data Mining.

Nota: 10*g*R2 proporciona soporte para todos los algoritmos de modelado de bases de datos, a excepción de los modelos lineales generalizados (requiere 11*g*R1).

- Un origen de datos ODBC para conectarse a Oracle, tal y como se describe a continuación.

Nota: optimización de SQL y modelado de bases de datos requieren que la conectividad de IBM SPSS Modeler Server esté activada en el equipo con IBM SPSS Modeler. Con esta configuración activada, puede acceder a los algoritmos de bases de datos, devolver SQL directamente desde IBM SPSS Modeler y acceder a IBM SPSS Modeler Server. Para verificar el estado de la licencia actual, seleccione las siguientes opciones en el menú de IBM SPSS Modeler.

Ayuda > **Acerca de** > **Detalles adicionales**

Si la conectividad está activada, verá la opción **Activación de servidor** en la pestaña Estado de licencia.

Activación de la integración con Oracle

Para activar la integración de IBM SPSS Modeler con Oracle Data Mining, necesitará configurar Oracle, crear un origen ODBC, permitir la integración en el cuadro de diálogo Aplicaciones de ayuda de IBM SPSS Modeler y activar la generación y optimización de SQL.

Configuración de Oracle

Para instalar y configurar Oracle Data Mining, consulte la documentación de Oracle (en concreto, la *Guía del administrador de Oracle*) para obtener más detalles.

Creación de un origen ODBC para Oracle

Para activar la conexión entre Oracle y IBM SPSS Modeler, debe crear un nombre de origen de datos del sistema ODBC (DSN).

Antes de crear un DSN, debe tener un conocimiento básico de las unidades y los orígenes de datos ODBC y el soporte de la base de datos en IBM SPSS Modeler.

Si se está ejecutando en modo distribuido con respecto al servidor IBM SPSS Modeler Server, cree el DSN en el equipo servidor. Si se está ejecutando en modo local (cliente), cree el DSN en el equipo cliente.

1. Instale los controladores ODBC. Se encuentran en el disco de instalación de IBM SPSS Data Access Pack enviado con esta versión. Ejecute el archivo *setup.exe* para iniciar el instalador y seleccione todos los controladores relevantes. Siga las instrucciones que aparecen en pantalla para instalar los controladores.
   a. Cree el DSN.

      Nota: la secuencia de menú depende de la versión de Windows.

      - **Windows XP.** En el menú Inicio, seleccione **Panel de control**. Pulse dos veces en **Herramientas administrativas** y, a continuación, pulse dos veces en **Orígenes de datos (ODBC)**.
      - **Windows Vista.** En el menú Inicio, seleccione **Panel de control** y, a continuación, **Sistema y mantenimiento**. Pulse dos veces en **Herramientas administrativas**, seleccione **Orígenes de datos (ODBC)** y, a continuación, pulse **Abrir**.
      - **Windows 7.** En el menú Inicio, seleccione **Panel de control**, luego **Sistema y seguridad** y, a continuación, **Herramientas administrativas**. Seleccione **Orígenes de datos (ODBC)** y, a continuación, pulse en **Abrir**.

   b. Pulse en la pestaña **DSN de sistema** y, a continuación, **Añadir**.
2. Seleccione el controlador **SPSS OEM 6.0 Oracle Wire Protocol**.
3. Pulse **Finalizar**.
4. En la pantalla de configuración del controlador ODBC Oracle Wire Protocol, introduzca el nombre de origen de datos que desee, el nombre de host del servidor de Oracle, el número de puerto para la conexión y el identificador de seguridad para la instancia de Oracle que se está utilizando.

   Puede obtener el nombre de host, el puerto y el Id. de seguridad del archivo *tnsnames.ora* en la máquina del servidor si se ha implementado TNS con un archivo *tnsnames.ora*. Póngase en contacto con el administrador de SPSS si desea obtener más información.
5. Pulse en el botón **de comprobación** para comprobar la conexión.

Activación de la integración de Oracle Data Mining en IBM SPSS Modeler

1. Seleccione en los menús de IBM SPSS Modeler:

   Herramientas > **Opciones** > **Aplicaciones de ayuda**
2. Pulse en la pestaña **Oracle**.

Habilitar integración de Oracle Data Mining. Habilita la paleta de modelado de bases de datos (si todavía no se muestra) en la parte inferior de la ventana de IBM SPSS Modeler y añade los nodos para algoritmos de Oracle Data Mining.

Conexión de Oracle. Especifique el origen de datos ODBC de Oracle que se utiliza de forma predeterminada para generar y almacenar modelos, un nombre de usuario válido y una contraseña. Esta configuración se puede omitir en los nodos de modelado individual y nuggets de modelo.

Nota: la conexión de base de datos que se utiliza para el modelado puede ser la misma que se utiliza para acceder a los datos, pero también puede ser otra diferente. Por ejemplo, puede tener una ruta que accede a los datos desde una base de datos de Oracle, descarga los datos a IBM SPSS Modeler para realizar limpiezas y otras manipulaciones y, a continuación, carga los datos en una base de datos de Oracle diferente para realizar modelados. Si lo prefiere, los datos originales pueden encontrarse en un archivo sin formato u otro origen (no perteneciente a Oracle), en cuyo caso sería necesario cargarlo a Oracle para realizar el modelado. En todos los casos, los datos se cargarán de manera automática en una tabla temporal creada en una base de datos que se utiliza para modelado.

Avisar cuando haya de sobrescribirse un modelo de Oracle Data Mining. Seleccione esta opción para asegurarse de que IBM SPSS Modeler no sobrescribe los modelos almacenados en la base de datos sin avisar con anterioridad.

Crear lista de modelos de Oracle Data Mining Muestra los modelos de minería de datos disponibles.

Activar inicio de Oracle Data Miner. (opcional) Cuando está activada, permite a IBM SPSS Modeler iniciar la aplicación Oracle Data Miner. Consulte “Oracle Data Miner” en la página 48 para obtener más información.

Ruta del ejecutable de Oracle Data Miner. (opcional) Especifica la ubicación física del archivo ejecutable de Oracle Data Miner de Windows (normalmente en *C:\odm\bin\odminerw.exe*). Oracle Data Miner no se instala con IBM SPSS Modeler; debe descargar la versión correcta del sitio Web de Oracle (*http://www.oracle.com/technology/products/bi/odm/odminer.html*) e instalarla en el cliente.

Activación de optimización y generación de SQL

1. Seleccione en los menús de IBM SPSS Modeler:

   Herramientas > **Propiedades de ruta** > **Opciones**
2. Pulse en la opción **Optimización** en el panel de navegación.
3. Confirme que la opción **Generar SQL** está activada. Esta configuración es necesaria para que el modelado de bases de datos funcione.
4. Seleccione las opciones **Optimizar generación de SQL** y **Optimizar otra ejecución** (no obligatorias, pero recomendadas para un rendimiento optimizado).

Generación de modelos con Oracle Data Mining

Los nodos de generación de modelos de Oracle funcionan exactamente igual que otros nodos de modelado de IBM SPSS Modeler, con contadas excepciones. Es posible acceder a estos nodos desde la paleta Modelado de bases de datos de la parte inferior de la ventana de IBM SPSS Modeler.

Consideraciones sobre los datos

Oracle necesita que los datos categóricos se almacenen en un formato de cadena (CHAR o VARCHAR2). Como resultado, IBM SPSS Modeler no permitirá que se especifique como entrada de modelos de ODM el almacenaje de campos numéricos con un nivel de medición de *Marca* o *Nominal* (categórico). Si fuera necesario, los números pueden convertirse en cadenas en IBM SPSS Modeler mediante el nodo Reclasificar.

Campo objetivo. Solamente se puede seleccionar un campo como resultado (objetivo) en los modelos de clasificación ODM.

Nombre del modelo. Desde Oracle 11*g*R1 en adelante, el nombre `unique` es una palabra clave y no se puede utilizar como un nombre de modelo personalizado.

Campo exclusivo. Especifica el campo utilizado para identificar de manera exclusiva cada caso. Por ejemplo, puede tratarse de un campo de ID, como *CustomerID*. IBM SPSS Modeler impone la restricción de que este campo debe ser numérico.

Nota: este campo es opcional para todos los nodos Oracle, excepto Bayesiano adaptativo de Oracle, O-clúster de Oracle y Apriori de Oracle.

Comentarios generales

- IBM SPSS Modeler no proporciona funciones de exportación/importación PMML para modelos creados por Oracle Data Mining.
- La puntuación de modelos siempre ocurre en ODM. Es posible que sea necesario cargar el conjunto de datos en una tabla temporal si los datos se originan o se tienen que preparar dentro de IBM SPSS Modeler.
- En IBM SPSS Modeler, generalmente, solamente se proporciona una única predicción y una confianza o probabilidad asociada.
- IBM SPSS Modeler restringe tanto el número de campos que pueden utilizarse en la generación de modelos como la puntuación a 1,000.
- IBM SPSS Modeler puede puntuar los modelos de ODM desde dentro de las rutas publicadas para ejecutarlos mediante IBM SPSS Modeler Solution Publisher.

Opciones del servidor de modelos de Oracle

Especifique la conexión de Oracle que se utiliza para cargar los datos para el modelado. Si fuera necesario, puede seleccionar una conexión para cada nodo de modelado en la pestaña del servidor con objeto de omitir la conexión predeterminada especificada de Oracle en el cuadro de diálogo Aplicaciones de ayuda. Para obtener más información, consulte el tema “Activación de la integración con Oracle” en la página 30.

Comentarios

- La conexión que se utiliza para el modelado puede ser la misma que se utiliza en el nodo de origen de una ruta, o puede ser otra diferente. Por ejemplo, puede tener una ruta que accede a los datos desde una base de datos de Oracle, descarga los datos a IBM SPSS Modeler para realizar limpiezas y otras manipulaciones y, a continuación, carga los datos en una base de datos de Oracle diferente para realizar modelados.
- El nombre de origen de datos ODBC se incrusta de manera efectiva en cada ruta de IBM SPSS Modeler. Si una ruta creada en un host se ejecuta en un host diferente, el nombre del origen de datos debe ser el mismo en cada host. Si lo prefiere, se puede seleccionar un origen de datos diferente en la pestaña Servidor de cada nodo de origen o de modelado.

Costes de clasificación errónea

En algunos contextos, ciertos tipos de errores son más costosos que otros. Por ejemplo, puede resultar más costoso clasificar a un solicitante de crédito de alto riesgo como de bajo riesgo (un tipo de error) que clasificar a un solicitante de crédito de bajo riesgo como de alto riesgo (otro tipo de error). Los costes de clasificación errónea permiten especificar la importancia relativa de los diversos tipos de errores de predicción.

Los costes de clasificación errónea son básicamente ponderaciones aplicadas a resultados específicos. Estas ponderaciones se extraen en el modelo y pueden realmente cambiar la predicción (como forma de protección frente a errores costosos).

Salvo los modelos C5.0, los costes de clasificación errónea no se aplican cuando se puntúa un modelo y no se tienen en cuenta cuando se clasifican o comparan modelos utilizando un nodo Clasificador automático, un diagrama de evaluación o un nodo Análisis. Es posible que un modelo que incluya costes no produzca menos errores que uno que no lo haga, y es posible que no ordene ningún valor mayor en términos de precisión global, pero es probable que funcione mejor en términos prácticos porque contiene un sesgo integrado en favor de errores *más baratos*.

La matriz de costes muestra el coste para cada combinación posible de categoría predicha y categoría real. De forma predeterminada, todos los costes de clasificación errónea se establecen en 1,0. Para introducir valores de coste personalizados, seleccione **Utilizar costes de clasificación errónea** e introduzca valores personalizados en la matriz de costes.

Para cambiar un coste de clasificación errónea, seleccione la casilla correspondiente a la combinación deseada de valores predichos y reales, elimine el contenido existente de la casilla e introduzca en ella el coste deseado. Los costes no son simétricos automáticamente. Por ejemplo, si establece el coste de clasificación errónea *A* como *B* para que sea 2,0, el coste de clasificación errónea de *B* como *A* aún tendrá el valor predeterminado 1,0 hasta que también se modifique explícitamente.

Nota: solamente los modelos de árboles de decisión permiten especificar los costes durante la generación.

Bayesiano ingenuo de Oracle

Bayesiano ingenuo es un algoritmo muy utilizado para resolver problemas de clasificación. El modelo se denomina *ingenuo* porque trata todas las variables de predicción propuestas como independientes unas de otras. El bayesiano ingenuo es un algoritmo rápido y escalable que calcula las probabilidades condicionales para las combinaciones de atributos y el atributo de objetivo. A partir de los datos de entrenamiento se establece una probabilidad independiente. Esta probabilidad proporciona la verosimilitud de cada clase objetivo, una vez dada la instancia de cada categoría de valor a partir de cada variable de entrada.

- La validación cruzada se utiliza para comprobar la precisión del modelo en los mismos datos que se utilizaron para generar el modelo. Resulta especialmente útil cuando el número de casos disponible para generar un modelo es pequeño.
- El resultado del modelo se puede examinar en un formato de matriz. Los números de la matriz son probabilidades condicionales relacionadas con las clases (columnas) predichas y las combinaciones (filas) del valor de la variable predictora.

Opciones del modelo bayesiano ingenuo

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo.

Campo exclusivo. Especifica el campo utilizado para identificar de manera exclusiva cada caso. Por ejemplo, puede tratarse de un campo de ID, como *CustomerID*. IBM SPSS Modeler impone la restricción de que este campo debe ser numérico.

Nota: este campo es opcional para todos los nodos Oracle, excepto Bayesiano adaptativo de Oracle, O-clúster de Oracle y Apriori de Oracle.

Preparación automática de datos. (11*g* únicamente) Activa (valor predeterminado) o desactiva el modo de preparación de datos automatizada para Oracle Data Mining. Si esta casilla está marcada, ODM realiza automáticamente las transformaciones de datos requeridas por el algoritmo. Para obtener más información, consulte *Conceptos de Oracle Data Mining*.

Opciones de experto para el bayesiano ingenuo

Una vez creado el modelo, los valores predictores individuales del atributo o los pares de valores se omiten a menos que existan instancias suficientes de un valor o una pareja en particular en los datos de entrenamiento. Los umbrales para ignorar los valores se especifican como fracciones basadas en el número de registros de los datos de entrenamiento. El ajuste de estos umbrales puede reducir el ruido y aumentar la capacidad del modelo para extenderse a otros conjuntos de datos.

- **Umbral de singleton.** Especifica el umbral de un valor de atributo de un predictor dado. El número de instancias de un valor dado debe igualar o sobrepasar la fracción especificada o el valor se ignorará.
- **Umbral por parejas.** Especifica el umbral de una pareja de valores determinada de atributo y predictor. El número de instancias de una pareja de valores en particular debe igualar o sobrepasar la fracción especificada o la pareja se ignorará.

Probabilidad de predicción. Permite que el modelo incluya la probabilidad de una predicción correcta para un posible resultado del campo objetivo. Para activar esta característica, seleccione **Seleccionar**, pulse en el botón **Especificar**, seleccione uno de los resultados posibles y pulse en **Insertar**.

Utilizar conjunto de predicciones. Genera una tabla de todos los resultados posibles del campo destino.

Bayesiano adaptativo de Oracle

La red de bayesiano adaptativo (RBA) construye clasificadores de redes bayesianas mediante la longitud mínima de la descripción (LMD) y la selección de características automática. RBA funciona bien en ciertas ocasiones en las que el bayesiano ingenuo no funciona con precisión y funciona, como mínimo, igual de bien en el resto de situaciones, aunque el rendimiento puede ser más lento. El algoritmo RBA proporciona la capacidad de generar tres tipos de modelos avanzados basados en los bayesianos, incluido el árbol de decisión simplificado (mono-característica), el bayesiano ingenuo podado y los modelos multi-característica aumentados.

Modelos generados

En el modo de generación mono-característica, RBA crea un árbol de decisión simplificado, basado en un conjunto de reglas legibles para los humanos, que permiten al usuario empresarial o al analista comprender la base de las predicciones del modelo y actuar en consecuencia o explicarlas a otros. Esto puede suponer una ventaja significativa sobre los modelos bayesiano ingenuo y multi-característica. Estas reglas pueden examinarse como un conjunto de reglas estándar en IBM SPSS Modeler. Un conjunto de reglas simple puede tener un aspecto parecido a éste:

```
IF MARITAL_STATUS = "Casado"
AND EDUCATION_NUM = "13-16"
THEN CHURN= "TRUE"
Confidence = .78, Support = 570 cases
```

Los modelos multi-característica y bayesiano ingenuo podado no pueden examinarse en IBM SPSS Modeler.

Opciones del modelo bayesiano adaptativo

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo.

Campo exclusivo. Especifica el campo utilizado para identificar de manera exclusiva cada caso. Por ejemplo, puede tratarse de un campo de ID, como *CustomerID*. IBM SPSS Modeler impone la restricción de que este campo debe ser numérico.

Nota: este campo es opcional para todos los nodos Oracle, excepto Bayesiano adaptativo de Oracle, O-clúster de Oracle y Apriori de Oracle.

Tipo de modelo

Puede seleccionar entre los tres modos diferentes para la generación del modelo.

- **Multi-característica.** Genera y compara un número de modelos, incluido un modelo bayesiano ingenuo, así como los modelos únicos y multi-característica de probabilidad de productos. Éste es el modo más exhaustivo y, generalmente, el que tarda más en calcularse como un resultado. Sólo se producen reglas si el modelo mono-característica se presenta como el mejor. Si se elige un modelo multi-característica o NB, no se producirán reglas.
- **Mono-característica.** Crea un árbol de decisión simplificado basado en un conjunto de reglas. Cada regla contiene una condición junto con las probabilidades asociadas con cada resultado. Estas reglas son mutuamente excluyentes y se proporcionan en un formato legible para los humanos, lo que puede ser una importante ventaja sobre los modelos bayesiano ingenuo y multi-característica.
- **Bayesiano ingenuo.** Genera un único modelo bayesiano ingenuo y lo compara con la previa de la muestra global (la distribución de los valores de objetivo en la muestra global). El modelo bayesiano ingenuo se genera como salida solamente si se presenta como mejor predictor de los valores objetivo que la previa global. Si no, no se genera ningún modelo como resultado.

Opciones de experto para el bayesiano adaptativo

Limitar tiempo de procesamiento. Seleccione esta opción para especificar un tiempo de generación máximo en minutos. Esto posibilita la generación de modelos en menor tiempo, aunque el modelo resultante puede ser menos exacto. En cada hito del proceso de modelado, el algoritmo comprueba si será capaz de completar el siguiente hito en la cantidad de tiempo especificada antes de continuar y devuelve el mejor modelo disponible al alcanzar el límite.

Máx. de predictores. Esta opción permite limitar la complejidad del modelo y aumentar el rendimiento limitando el número de predictores utilizados. Los predictores se ordenan en función de una medida LMD de su correlación con el objetivo como medida de verosimilitud incluida en el modelo.

Máx. de predictores de bayesiano ingenuo. Esta opción especifica el número máximo de predictores que se utilizan en el modelo bayesiano ingenuo.

Máquina de vectores de soporte de Oracle (SVM)

La máquina de vectores de soporte (SVM) es un algoritmo de clasificación y regresión que utiliza la teoría de aprendizaje de las máquinas para maximizar la precisión de las predicciones sin ajustar excesivamente los datos. SVM utiliza una transformación no lineal opcional de los datos de entrenamiento, seguida de la búsqueda de ecuaciones de regresión en los datos transformados para separar las clases (para objetivos categóricos) o ajustar el objetivo (para los objetivos continuos). La implementación de SVM de Oracle permite que se generen modelos mediante el uso de los dos kernels disponibles: lineal o gausiano. El kernel lineal omite la transformación no lineal de una vez, de tal forma que el modelo resultante sea, en esencia, un modelo de regresión.

Para obtener más información, consulte el manual *Oracle Data Mining Application Developer's Guide* y *Oracle Data Mining Concepts*.

Opciones del modelo SVM

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Campo exclusivo. Especifica el campo utilizado para identificar de manera exclusiva cada caso. Por ejemplo, puede tratarse de un campo de ID, como *CustomerID*. IBM SPSS Modeler impone la restricción de que este campo debe ser numérico.

Nota: este campo es opcional para todos los nodos Oracle, excepto Bayesiano adaptativo de Oracle, O-clúster de Oracle y Apriori de Oracle.

Preparación automática de datos. (11*g* únicamente) Activa (valor predeterminado) o desactiva el modo de preparación de datos automatizada para Oracle Data Mining. Si esta casilla está marcada, ODM realiza automáticamente las transformaciones de datos requeridas por el algoritmo. Para obtener más información, consulte *Conceptos de Oracle Data Mining*.

Aprendizaje activo. Proporciona un método para gestionar los conjuntos generados de gran tamaño. Con el aprendizaje activo, el algoritmo crea un modelo inicial en base a una pequeña muestra antes de aplicarlo a todo el conjunto de datos de entrenamiento y, a continuación, actualiza la muestra y el modelo de forma gradual en función de los resultados. Este ciclo se repite hasta que el modelo converge en los datos de entrenamiento o hasta que se alcanza el número máximo de vectores de soporte permitidos.

Función de kernel. Seleccione **Lineal** o **Gaussiano**, o bien deja la opción predeterminada **Determinado por el sistema** para que el sistema seleccione el kernel más adecuado. Los kernels gaussianos pueden aprender relaciones más complejas, aunque normalmente tardan más en realizar los cálculos. Es posible que desee comenzar con el kernel lineal e intentarlo con el kernel gaussiano sólo si el kernel lineal no es capaz de encontrar un ajuste adecuado. Esto es más probable que suceda con un modelo de regresión, donde la elección del kernel es más importante. Asimismo, observe que los modelos SVM creados con el kernel gaussiano no pueden examinarse en IBM SPSS Modeler. Los modelos construidos con el kernel lineal pueden examinarse en IBM SPSS Modeler del mismo modo que los modelos de regresión estándar.

Método de normalización. Especifica el método de normalización utilizado para la entrada continua y los campos objetivo. Es posible seleccionar **Puntuaciones Z**, **Mín.-Máx.** o **Ninguna**. Oracle realiza la normalización automáticamente si la casilla de verificación **Preparación de datos automática** está seleccionada. No marque este casilla para seleccionar manualmente el método de normalización.

Opciones de experto de SVM

Tamaño de caché de Kernel. Especifica el tamaño de caché, en bytes, que se utiliza para almacenar los kernels calculados durante la operación de generación. Como es de esperar, las cachés de mayor tamaño generalmente originan construcciones más rápidas. El valor predeterminado es 50 MB.

Tolerancia de convergencia. Especifica el valor de tolerancia permitido para la generación del modelo antes de terminar. El valor debe estar comprendido entre 0 y 1. El valor predeterminado es 0,001. Los valores mayores tienden a originar generaciones más rápidas, aunque modelos menos exactos.

Especificar desviación estándar. Permite especificar el parámetro de desviación estándar que el kernel gaussiano utiliza. Este parámetro afecta al equilibrio entre la complejidad del modelo y la capacidad para generalizar a otros conjuntos de datos (sobreajustando y subajustando los datos). Los valores de desviación estándar mayores favorecen el subajuste. De forma predeterminada, se estima este parámetro a partir de los datos de entrenamiento.

Especificar épsilon. Solamente para los modelos de regresión, especifica el valor del intervalo del error permitido en la generación de modelos no sensibles a épsilon. En otras palabras, distingue pequeños

errores (que se ignoran) de los grandes (que no se ignoran). El valor debe estar comprendido entre 0 y 1. De forma predeterminada, se calcula a partir de los datos de entrenamiento.

Especificar factor de complejidad. Permite determinar el factor de complejidad, que equilibra el error del modelo (como se mide con respecto a los datos de entrenamiento) y la complejidad del modelo a fin de evitar el sobreajuste o el subajuste de los datos. Los valores mayores proporcionan una penalización mayor a los errores, lo que supone un mayor riesgo de sobreajuste de los datos; los valores menores proporcionan una penalización menor en los errores y pueden originar subajustes.

Especificar tasa de valores atípicos. Especifica la tasa de valores atípicos deseada en los datos de entrenamiento. Sólo es válida en los modelos de una clase de SVM. No se puede utilizar con la configuración **Especificar factor de complejidad**.

Probabilidad de predicción. Permite que el modelo incluya la probabilidad de una predicción correcta para un posible resultado del campo objetivo. Para activar esta característica, seleccione **Seleccionar**, pulse en el botón **Especificar**, seleccione uno de los resultados posibles y pulse en **Insertar**.

Utilizar conjunto de predicciones. Genera una tabla de todos los resultados posibles del campo destino.

Opciones de ponderaciones de SVM Oracle

En un modelo de clasificación, el uso de ponderaciones le permite especificar la importancia relativa de varios valores objetivo posibles. El hacerlo puede ser de utilidad, por ejemplo, si los puntos de datos en sus datos de entrenamiento no están distribuidos de forma realista entre las categorías. Las ponderaciones le permiten sesgar el modelo, de forma que puede compensar esas categorías que están peor representadas en los datos. Al aumentar la ponderación de un valor destino debería aumentar el porcentaje de las predicciones correctas para esa categoría.

Hay tres métodos de configuración de ponderaciones:

- **Basadas en datos de entrenamiento.** Es el valor por omisión. Las ponderaciones se basan en las frecuencias relativas de las categorías en los datos de entrenamiento.
- **Igual para todas las clases.** Las ponderaciones de todas las categorías se definen como $1/k$, donde k es el número de categorías objetivo.
- **Personalizado.** Puede especificar sus propias ponderaciones. Los valores iniciales de las ponderaciones se configuran como iguales para todas las clases. Puede ajustar las ponderaciones de cada categoría individualmente con valores personalizados. Para ajustar la ponderación de una categoría específica, seleccione la casilla Ponderar de la tabla correspondiente a la probabilidad que desee, elimine el contenido de la casilla e introduzca el valor que desee.

Las ponderaciones de todas las categorías deberían sumar 1.0. En caso contrario, aparecerá una advertencia con una opción para normalizar los valores automáticamente. Este ajuste automático conserva las proporciones en todas las categorías a la vez que fuerza la restricción de ponderación. Puede llevar a cabo este ajuste en cualquier momento pulsando en el botón **Normalizar**. Para restablecer la tabla de modo que todas las categorías tengan el mismo valor, pulse en el botón **Igualar**.

Modelos lineales generalizados de Oracle (GLM)

(11g únicamente) Los modelos lineales generalizados relajan los supuestos restrictivos de los modelos lineales. Entre ellos se incluyen, por ejemplo, los supuestos de que la variable objetivo tiene una distribución normal y que el efecto de los predictores de la variable objetivo es lineal por naturaleza. Un modelo lineal generalizado es el adecuado para suposiciones en las que el objetivo es posible que tenga una distribución no normal, como una distribución multinomial o de Poisson. De forma similar, un modelo lineal generalizado es de gran utilidad en los casos en los que es probable que la relación o enlace entre los predictores y objetivo sea no lineal.

Para obtener más información, consulte el manual *Oracle Data Mining Application Developer's Guide* y *Oracle Data Mining Concepts*.

Opciones del modelo GLM

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Campo exclusivo. Especifica el campo utilizado para identificar de manera exclusiva cada caso. Por ejemplo, puede tratarse de un campo de ID, como *CustomerID*. IBM SPSS Modeler impone la restricción de que este campo debe ser numérico.

Nota: este campo es opcional para todos los nodos Oracle, excepto Bayesiano adaptativo de Oracle, O-clúster de Oracle y Apriori de Oracle.

Preparación automática de datos. (11*g* únicamente) Activa (valor predeterminado) o desactiva el modo de preparación de datos automatizada para Oracle Data Mining. Si esta casilla está marcada, ODM realiza automáticamente las transformaciones de datos requeridas por el algoritmo. Para obtener más información, consulte *Conceptos de Oracle Data Mining*.

Método de normalización. Especifica el método de normalización utilizado para la entrada continua y los campos objetivo. Es posible seleccionar **Puntuaciones Z**, **Mín.-Máx.** o **Ninguna**. Oracle realiza la normalización automáticamente si la casilla de verificación **Preparación de datos automática** está seleccionada. No marque este casilla para seleccionar manualmente el método de normalización.

Gestión de valores perdidos. Especifica cómo se procesarán los valores perdidos en los datos de entrada:

- **Sustituir con media o modo** sustituye los valores perdidos de los atributos numéricos con el valor de la media y sustituye los valores perdidos de los atributos categóricos con el modo.
- **Solamente utilizar registros completos** ignora los registros con valores perdidos.

Opciones de experto de GLM

Utilizar ponderaciones de fila. Active esta casilla de verificación para activar la lista desplegable adyacente, desde donde podrá seleccionar una columna que contiene un factor de ponderación para las filas.

Guardar diagnósticos de fila en la tabla. Active esta casilla de verificación para activar el campo de texto adyacente, donde podrá introducir el nombre de una tabla que contiene diagnósticos de nivel de fila.

Nivel de confianza de coeficiente. El grado de certidumbre, entre 0,0 y 1,0, del valor predicho para el objetivo en un intervalo de confianza calculado para el modelo. Los límites de confianza se devuelven con los estadísticos de coeficientes.

Categoría de referencia para el objetivo. Seleccione **Personalizada** para seleccionar un valor del campo objetivo y utilizarlo como categoría de referencia, o deje el valor predeterminado **Auto**.

Regresión contraída. La regresión contraída es una técnica que compensa la situación en la que existe un grado de correlación demasiado alto en las variables. Puede utilizar la opción **Auto** para permitir que el algoritmo controle el uso de esta técnica, o bien, puede controlarlo manualmente mediante las opciones **Desactivar** y **Activar**. Si selecciona activar la regresión contraída manualmente, puede sustituir el valor predeterminado del sistema por el parámetro contraído introduciendo un valor en el campo adyacente.

Producir VIF para una regresión contraída. Active esta casilla de verificación si desea producir estadísticos de Factor de inflación de la varianza (FIV) si la contracción se utiliza para la regresión lineal.

Probabilidad de predicción. Permite que el modelo incluya la probabilidad de una predicción correcta para un posible resultado del campo objetivo. Para activar esta característica, seleccione **Seleccionar**, pulse en el botón **Especificar**, seleccione uno de los resultados posibles y pulse en **Insertar**.

Utilizar conjunto de predicciones. Genera una tabla de todos los resultados posibles del campo destino.

Opciones de ponderaciones de GLM Oracle

En un modelo de clasificación, el uso de ponderaciones le permite especificar la importancia relativa de varios valores objetivo posibles. El hacerlo puede ser de utilidad, por ejemplo, si los puntos de datos en sus datos de entrenamiento no están distribuidos de forma realista entre las categorías. Las ponderaciones le permiten sesgar el modelo, de forma que puede compensar esas categorías que están peor representadas en los datos. Al aumentar la ponderación de un valor destino debería aumentar el porcentaje de las predicciones correctas para esa categoría.

Hay tres métodos de configuración de ponderaciones:

- **Basadas en datos de entrenamiento.** Es el valor por omisión. Las ponderaciones se basan en las frecuencias relativas de las categorías en los datos de entrenamiento.
- **Igual para todas las clases.** Las ponderaciones de todas las categorías se definen como $1/k$, donde k es el número de categorías objetivo.
- **Personalizado.** Puede especificar sus propias ponderaciones. Los valores iniciales de las ponderaciones se configuran como iguales para todas las clases. Puede ajustar las ponderaciones de cada categoría individualmente con valores personalizados. Para ajustar la ponderación de una categoría específica, seleccione la casilla Ponderar de la tabla correspondiente a la probabilidad que desee, elimine el contenido de la casilla e introduzca el valor que desee.

Las ponderaciones de todas las categorías deberían sumar 1.0. En caso contrario, aparecerá una advertencia con una opción para normalizar los valores automáticamente. Este ajuste automático conserva las proporciones en todas las categorías a la vez que fuerza la restricción de ponderación. Puede llevar a cabo este ajuste en cualquier momento pulsando en el botón **Normalizar**. Para restablecer la tabla de modo que todas las categorías tengan el mismo valor, pulse en el botón **Igualar**.

Árbol de decisión de Oracle

Oracle Data Mining ofrece una característica de árbol de decisión clásica, basada en el popular algoritmo Árbol de clasificación y regresión. El modelo de árbol de decisión de ODM contiene información detallada sobre cada nodo, incluyendo la confianza, el soporte y los criterios de división. Se puede mostrar la regla para cada nodo y, además, se proporciona un atributo de sustitución para los nodos, que puede utilizarse como sustituto cuando al aplicar el modelo a un caso con valores perdidos.

Los árboles de decisión son populares porque su aplicación es universal y sencilla, y son fáciles de comprender. Los árboles de decisión criban a través de cada atributo de entrada potencial en busca del mejor "divisor," es decir, el punto de corte del atributo (por ejemplo, `AGE > 55`) que divide los registros de datos posteriores de la ruta en varias poblaciones homogéneas. Tras cada decisión de división, ODM repite el proceso desarrollando el árbol entero y creando "hojas" terminales que representan poblaciones similares de registros, elementos o personas. Al descender desde el nodo raíz del árbol (por ejemplo, la población total), los árboles de decisión proporcionan reglas legibles para los humanos de las instrucciones `Si A, entonces B`. Estas reglas del árbol de decisión también proporcionan el soporte y la confianza para cada nodo del árbol.

Mientras que las redes de bayesiano adaptativo pueden proporcionar también reglas cortas y sencillas que pueden resultar útiles para proporcionar explicaciones para cada predicción, los árboles de decisión proporcionan reglas de Oracle Data Mining para cada decisión de división. Los árboles de decisión también resultan útiles para desarrollar perfiles detallados de mejores clientes, pacientes saludables, factores asociados al fraude, etc.

Opciones de modelo para los árboles de decisión

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Campo exclusivo. Especifica el campo utilizado para identificar de manera exclusiva cada caso. Por ejemplo, puede tratarse de un campo de ID, como *CustomerID*. IBM SPSS Modeler impone la restricción de que este campo debe ser numérico.

Nota: este campo es opcional para todos los nodos Oracle, excepto Bayesiano adaptativo de Oracle, O-clúster de Oracle y Apriori de Oracle.

Preparación automática de datos. (11*g* únicamente) Activa (valor predeterminado) o desactiva el modo de preparación de datos automatizada para Oracle Data Mining. Si esta casilla está marcada, ODM realiza automáticamente las transformaciones de datos requeridas por el algoritmo. Para obtener más información, consulte *Conceptos de Oracle Data Mining*.

Métrica de la impureza. Especifica la métrica que se utiliza para buscar la mejor pregunta de comprobación para dividir los datos en cada nodo. Los mejores divisores y valores de división son aquellos que dan como resultado el mayor aumento en la homogeneidad del valor objetivo para las entidades del nodo. La homogeneidad se calcula según una métrica. Las medidas admitidas son **gini** y **entropía**.

Opciones de Experto para los árboles de decisión

Profundidad máxima. Establece la profundidad máxima del modelo de árbol que se va a generar.

Porcentaje mínimo de registros en un nodo. Establece el porcentaje de número mínimo de registros por nodo.

Porcentaje mínimo de registros para una división. Establece el número mínimo de registros en un nodo padre expresado como porcentaje del número total de registros empleados para entrenar el modelo. No se intenta dividir cuando el número de registros es inferior a este porcentaje.

Mínimo de registros en un nodo. Establece el número mínimo de registros devueltos.

Mínimo de registros para una división. Establece el número mínimo de registros en un nodo padre expresado como un valor. No se intenta dividir cuando el número de registros es inferior a este valor.

Identificador de regla. Si está marcado, incluye en el modelo una cadena para identificar el nodo en un árbol en el que se ha realizado una división en particular.

Probabilidad de predicción. Permite que el modelo incluya la probabilidad de una predicción correcta para un posible resultado del campo objetivo. Para activar esta característica, seleccione **Seleccionar**, pulse en el botón **Especificar**, seleccione uno de los resultados posibles y pulse en **Insertar**.

Utilizar conjunto de predicciones. Genera una tabla de todos los resultados posibles del campo destino.

O-clúster de Oracle

El algoritmo O-clúster de Oracle identifica las agrupaciones que se producen de forma natural en una población de datos. El clúster de partición ortogonal (O-clúster) es un algoritmo de clúster propiedad de Oracle que crea un modelo de clúster jerárquico basado en la cuadrícula, es decir, crea particiones de eje paralelo (ortogonal) en el espacio del atributo de entrada. El algoritmo funciona de forma recursiva. La estructura jerárquica resultante representa una cuadrícula irregular que forma un mosaico de clústeres en el espacio del atributo.

El algoritmo O-clúster gestiona atributos numéricos y categóricos, y ODM selecciona de forma automática las mejores definiciones de clúster. ODM proporciona información detallada, reglas y valores centroides del clúster, y se puede utilizar para puntuar una población en relación con su pertenencia a un clúster.

Opciones de Modelo para O-clúster

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Campo exclusivo. Especifica el campo utilizado para identificar de manera exclusiva cada caso. Por ejemplo, puede tratarse de un campo de ID, como *CustomerID*. IBM SPSS Modeler impone la restricción de que este campo debe ser numérico.

Nota: este campo es opcional para todos los nodos Oracle, excepto Bayesiano adaptativo de Oracle, O-clúster de Oracle y Apriori de Oracle.

Preparación automática de datos. (11*g* únicamente) Activa (valor predeterminado) o desactiva el modo de preparación de datos automatizada para Oracle Data Mining. Si esta casilla está marcada, ODM realiza automáticamente las transformaciones de datos requeridas por el algoritmo. Para obtener más información, consulte *Conceptos de Oracle Data Mining*.

Número máximo de clústeres. Establece el número máximo de clústeres generados.

Opciones de Experto para O-clúster

Búfer máximo. Establece el tamaño máximo del búfer.

Sensibilidad. Establece una fracción que especifica la densidad máxima necesaria para separar un nuevo clúster. La fracción está relacionada con la densidad uniforme global.

K-medias de Oracle

El algoritmo K-medias de Oracle identifica los clústeres que se producen de forma natural en una población de datos. El algoritmo K-medias es un algoritmo de clúster basado en la distancia que divide los datos en un número de clústeres predeterminado (siempre que haya suficientes casos distintos). Los algoritmos basados en la distancia confían en una métrica de distancia (función) para calcular la similitud entre los puntos de datos. Los puntos de datos se asignan al clúster más próximo en función de la métrica de distancia empleada. ODM proporciona una versión mejorada de K-medias.

El algoritmo K-medias admite clústeres jerárquicos, trata atributos numéricos y categóricos, y divide la población en el número de clústeres especificado por el usuario. ODM proporciona información detallada, reglas y valores centroides del clúster, y se puede utilizar para puntuar una población en relación con su pertenencia a un clúster.

Opciones de Modelo para K-medias

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Campo exclusivo. Especifica el campo utilizado para identificar de manera exclusiva cada caso. Por ejemplo, puede tratarse de un campo de ID, como *CustomerID*. IBM SPSS Modeler impone la restricción de que este campo debe ser numérico.

Nota: este campo es opcional para todos los nodos Oracle, excepto Bayesiano adaptativo de Oracle, O-clúster de Oracle y Apriori de Oracle.

Preparación automática de datos. (11*g* únicamente) Activa (valor predeterminado) o desactiva el modo de preparación de datos automatizada para Oracle Data Mining. Si esta casilla está marcada, ODM realiza automáticamente las transformaciones de datos requeridas por el algoritmo. Para obtener más información, consulte *Conceptos de Oracle Data Mining*.

Número de clústeres. Establece el número máximo de clústeres generados.

Función de distancia. Especifica qué función de distancia se va a utilizar para los clústeres de K-medias.

Criterio de división. Especifica qué criterio de división se va a utilizar para los clústeres de K-medias.

Método de normalización. Especifica el método de normalización utilizado para la entrada continua y los campos objetivo. Es posible seleccionar **Puntuaciones Z**, **Mín.-Máx.** o **Ninguna**.

Opciones de Experto para K-medias

Iteraciones. Establece el número de iteraciones para el algoritmo K-medias.

Tolerancia de convergencia. Establece la tolerancia de convergencia para el algoritmo K-medias.

Número de intervalos. Especifica el número de intervalos en el histograma del atributo producidos por K-medias. Los límites de intervalo de cada atributo se calculan globalmente en el conjunto de datos de entrenamiento completo. El método de intervalos es equitativo. Todos los atributos tienen el mismo número de intervalos, excepto los atributos con un único valor, que tienen sólo un intervalo.

Crecimiento en bloques. Establece el factor de crecimiento para la memoria asignada a los datos de clústeres.

Soporte de atributo de porcentaje mínimo. Establece la fracción de los valores de atributo que no deben ser nulos para que el atributo se incluya en la descripción de reglas del clúster. Si se establece un valor de parámetro demasiado alto en los datos con valores perdidos, se obtendrán muy pocas reglas o incluso reglas vacías.

Factorización de matrices no negativas (NMF) de Oracle

La factorización de matrices no negativas (NMF) permite reducir los grandes conjuntos de datos en atributos representativos. Conceptualmente, es similar al análisis de componentes principales (PCA) pero puede gestionar un mayor número de atributos en modelos de representación aditivos; NMF es un algoritmo de minería de datos potente y actual que se puede usar para una amplia variedad de casos.

NMF permite reducir grandes cantidades de datos, por ejemplo datos de texto, en representaciones más pequeñas y dispersas que reducen la dimensionalidad de los datos (se puede conservar la misma información con muchas menos variables). El resultado de los modelos de NMF se puede analizar mediante técnicas de aprendizaje supervisado, como las de SVM, o técnicas de aprendizaje no supervisado, como las técnicas de clúster. Oracle Data Mining utiliza algoritmos NMF y SVM para analizar datos de texto no estructurados.

Opciones de Modelo para NMF

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Campo exclusivo. Especifica el campo utilizado para identificar de manera exclusiva cada caso. Por ejemplo, puede tratarse de un campo de ID, como *CustomerID*. IBM SPSS Modeler impone la restricción de que este campo debe ser numérico.

Nota: este campo es opcional para todos los nodos Oracle, excepto Bayesiano adaptativo de Oracle, O-clúster de Oracle y Apriori de Oracle.

Preparación automática de datos. (11*g* únicamente) Activa (valor predeterminado) o desactiva el modo de preparación de datos automatizada para Oracle Data Mining. Si esta casilla está marcada, ODM realiza automáticamente las transformaciones de datos requeridas por el algoritmo. Para obtener más información, consulte *Conceptos de Oracle Data Mining*.

Método de normalización. Especifica el método de normalización utilizado para la entrada continua y los campos objetivo. Es posible seleccionar **Puntuaciones Z**, **Mín.-Máx.** o **Ninguna**. Oracle realiza la normalización automáticamente si la casilla de verificación **Preparación de datos automática** está seleccionada. No marque este casilla para seleccionar manualmente el método de normalización.

Opciones de Experto para NMF

Especificar número de características. Especifica el número de características que se desea extraer.

Semilla aleatoria. Establece la semilla aleatoria para el algoritmo NMF.

Número de iteraciones. Establece el número de iteraciones para el algoritmo NMF.

Tolerancia de convergencia. Establece la tolerancia de convergencia para el algoritmo NMF.

Mostrar todas las características. Muestra la identificación de las características y la confianza de todas las características, en vez de aquellos valores solamente para la mejor característica.

Apriori de Oracle

El algoritmo Apriori encuentra reglas de asociación en los datos. Por ejemplo, "si un cliente compra una cuchilla y una loción para después del afeitado, hay un 80% de posibilidades de que el cliente compre también crema de afeitado". El problema de análisis de asociaciones se divide en dos problemas secundarios:

- Encontrar todas las combinaciones de elementos, denominadas conjuntos de elementos frecuentes, cuyo soporte es superior al soporte mínimo.
- Utilizar los conjuntos de elementos frecuentes para generar las reglas deseadas. La idea es que si, por ejemplo, ABC y BC son frecuentes, entonces la regla "A implica BC", siempre que el cociente de `soporte(ABC)` y `soporte(BC)` sea como mínimo igual de grande que la confianza mínima. Observe que la regla tendrá un soporte mínimo debido a que ABCD es frecuente. La asociación de ODM solamente admite reglas únicas consecuentes (ABC implica D).

El número de conjuntos de elementos frecuentes se rige por los parámetros de soporte mínimo. El número de reglas generadas se rige por el número de conjuntos de elementos frecuentes y el parámetro de confianza. Si el parámetro de confianza se define con un valor demasiado alto, es posible que haya conjuntos de elementos frecuentes en el modelo de asociación, pero no reglas.

ODM emplea una implementación del algoritmo Apriori basada en SQL. La generación del candidato y los pasos de recuento de soporte se han implementado mediante consultas de SQL. No se utilizan estructuras de datos en memoria especializadas. Las consultas de SQL se han ajustado con precisión para ejecutarse de forma eficaz en el servidor de la base de datos mediante diversas sugerencias.

Opciones de los campos Apriori

Todos los nodos de modelado tienen una pestaña Campos en la que se pueden especificar los campos que se usarán para generar el modelo.

Antes de poder crear un modelo Apriori, es necesario que especifique qué campos desea utilizar como los elementos de interés en el modelado de asociación.

Utilizar configuración del nodo Tipo. Esta opción permite indicar al nodo que use la información de campo de un nodo Tipo situado en un punto anterior de la ruta. Éste es el valor por omisión.

Utilizar configuración personalizada. Esta opción permite indicar al nodo que use la información de campo especificada aquí en lugar de la proporcionada en nodos Tipo situados en cualquier punto anterior de la ruta. Después de seleccionar esta opción, especifique los campos restantes en el cuadro de diálogo, que dependen de si está utilizando formato transaccional.

Si *no* está utilizando formato transaccional, especifique:

- **Entradas.** Seleccione los campos de entrada. Se trata de una acción similar a establecer un rol de un campo a *Entrada* en un nodo Tipo.
- **Partición.** Este campo permite especificar un campo usado para dividir los datos en muestras independientes para las fases de entrenamiento, prueba y validación en la generación del modelo.

Si *está* utilizando formato transaccional, especifique:

Utilizar formato transaccional. Utilice esta opción si desea transformar datos de una fila por elemento a una fila por caso.

Al seleccionar esta opción se cambia los controles del campo en la parte inferior de este cuadro de diálogo:

Para formato transaccional, especifique:

- **ID.** Seleccione un campo de ID de la lista. Los campos numéricos o simbólicos se pueden utilizar como campo de ID. Cada valor exclusivo de este campo debe indicar una unidad de análisis específica. Por ejemplo, en una aplicación de la cesta de la compra, cada ID puede representar a un sólo cliente. Para una aplicación de análisis del registro Web, cada ID puede representar un equipo (con la dirección IP) o un usuario (con los datos de inicio de sesión).
- **Contenido.** Especifique el campo de contenido del modelo. Este campo contiene el elemento de interés del modelo de asociación.
- **Partición.** Este campo permite especificar un campo usado para dividir los datos en muestras independientes para las fases de entrenamiento, prueba y validación en la generación del modelo. Si usa una muestra para crear el modelo y otra muestra distinta para comprobarlo, podrá obtener una buena indicación de la bondad del modelo a la hora de generar conjuntos de datos de mayor tamaño similares a los datos actuales. Si se han definido varios campos de partición mediante nodos Tipo o Partición, se deberá seleccionar un campo de partición simple en la pestaña Campos en todos los nodos de modelado que usen la partición. (Si solamente hay una partición, se usará automáticamente siempre que se active la partición.) Debe tener en cuenta que al aplicar la partición seleccionada en su análisis, también debe activar la partición en la pestaña Opciones del modelo para el nodo. (Si se elimina la selección de esta opción, se posibilita la desactivación de la partición sin cambiar la configuración del campo.)

Opciones de Modelo para Apriori

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Campo exclusivo. Especifica el campo utilizado para identificar de manera exclusiva cada caso. Por ejemplo, puede tratarse de un campo de ID, como *CustomerID*. IBM SPSS Modeler impone la restricción de que este campo debe ser numérico.

Nota: este campo es opcional para todos los nodos Oracle, excepto Bayesiano adaptativo de Oracle, O-clúster de Oracle y Apriori de Oracle.

Preparación automática de datos. (11*g* únicamente) Activa (valor predeterminado) o desactiva el modo de preparación de datos automatizada para Oracle Data Mining. Si esta casilla está marcada, ODM realiza automáticamente las transformaciones de datos requeridas por el algoritmo. Para obtener más información, consulte *Conceptos de Oracle Data Mining*.

Longitud máxima de la regla. Establece el número máximo de precondiciones de una regla, un entero entre 2 y 20. Se trata de una forma de limitar la complejidad de las reglas. Si las reglas son demasiado complejas o demasiado específicas, o si el conjunto de reglas se tarda demasiado en entrenar, pruebe a reducir este ajuste.

Confianza mínima. Define el valor de confianza mínimo; un valor entre 0 y 1. Las reglas con un nivel de confianza inferior a los criterios especificados se descartan.

Soporte mínimo. Define el umbral mínimo compatible, un valor entre 0 y 1. Apriori descubre los patrones con una frecuencia superior al umbral mínimo compatible.

Longitud mínima de la descripción de Oracle (LMD)

El algoritmo Longitud mínima de la descripción (LMD) de Oracle ayuda a identificar los atributos con mayor influencia sobre el atributo objetivo. A menudo, conocer los atributos con mayor influencia ayuda a comprender y gestionar mejor el negocio y a simplificar las actividades de modelado. Además, estos atributos pueden indicar los tipos de datos que se desean añadir para argumentar los modelos. LMD se puede utilizar, por ejemplo, para encontrar los atributos del proceso más relevantes para predecir la calidad de una pieza fabricada, los factores asociados con el abandono de clientes o los genes de mayor implicación en el tratamiento de una enfermedad determinada.

LMD de Oracle descarta los campos de entrada que se considera irrelevante en la predicción del objetivo. Con los campos restantes genera un nugget de modelo sin refinar que se asocia con un modelo Oracle, visible en Oracle Data Miner. Al explorar el modelo en Oracle Data Miner se muestra un gráfico que muestra el resto de campos de entrada, ordenados por el orden de significación en el objetivo predicho.

Una clasificación negativa indica ruido. Los campos de entrada clasificados como cero o menos no contribuyen a la predicción y se deben eliminar de los datos.

Para mostrar el gráfico

1. Pulse con el botón derecho en el nugget de modelo sin refinar en la paleta Modelos y seleccione **Buscar**.
2. En la ventana del modelo, pulse en el botón para iniciar Oracle Data Miner.
3. Conéctese a Oracle Data Miner. Para obtener más información, consulte el tema "Oracle Data Miner" en la página 48.
4. En el panel de navegación de Oracle Data Miner, expanda **Modelos** y seleccione **Importancia del atributo**.
5. Seleccione el modelo de Oracle relevante (tendrá el mismo nombre que el campo de objetivo que especifique en IBM SPSS Modeler). Si no está seguro de cual es el correcto, seleccione la carpeta Importancia del atributo y busque un modelo por la fecha de creación.

Opciones de Modelo para LMD

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Campo exclusivo. Especifica el campo utilizado para identificar de manera exclusiva cada caso. Por ejemplo, puede tratarse de un campo de ID, como *CustomerID*. IBM SPSS Modeler impone la restricción de que este campo debe ser numérico.

Nota: este campo es opcional para todos los nodos Oracle, excepto Bayesiano adaptativo de Oracle, O-clúster de Oracle y Apriori de Oracle.

Preparación automática de datos. (11*g* únicamente) Activa (valor predeterminado) o desactiva el modo de preparación de datos automatizada para Oracle Data Mining. Si esta casilla está marcada, ODM realiza automáticamente las transformaciones de datos requeridas por el algoritmo. Para obtener más información, consulte *Conceptos de Oracle Data Mining*.

Importancia del atributo de Oracle (AI)

El objetivo de la importancia del atributo es descubrir los atributos del conjunto de datos que están relacionados con el resultado y el grado en el que influyen en el resultado final. El nodo Importancia del atributo de Oracle analiza los datos, busca patrones y predice resultados con un nivel de confianza asociado.

Opciones de modelo de AI

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo.

Preparación automática de datos. (11*g* únicamente) Activa (valor predeterminado) o desactiva el modo de preparación de datos automatizada para Oracle Data Mining. Si esta casilla está marcada, ODM realiza automáticamente las transformaciones de datos requeridas por el algoritmo. Para obtener más información, consulte *Conceptos de Oracle Data Mining*.

Opciones de selección de AI

La pestaña Opciones permite especificar la configuración predeterminada para seleccionar o excluir campos de entrada en el nugget de modelo. Tras ello, se puede añadir el modelo a una ruta para seleccionar un subconjunto de campos para usarlo en generaciones de modelos posteriores. Opcionalmente, se puede sobrescribir esta configuración seleccionando o anulando la selección de campos adicionales en el explorador de modelos cuando haya generado el modelo. Sin embargo, la configuración predeterminada permite aplicar el nugget de modelo sin más cambios, lo que puede ser especialmente útil para scripts.

Se encuentran disponibles las siguientes opciones:

Todos los campos clasificados. Selecciona los campos según la clasificación como *important*, *marginal* o *unimportant*. Se puede editar la etiqueta de cada clasificación, así como los valores de corte que se utilizan para asignar los registros a un rango u otro.

Número especificado de campos. Selecciona los *n* campos principales en función de su importancia.

Importancia mayor que. Selecciona todos los campos con una importancia superior al valor especificado.

El campo objetivo siempre se conserva, independientemente de la selección.

Pestaña Modelo de nugget de modelo de AI

La pestaña Modelo de un nugget de modelo AI de Oracle muestra el rango y la importancia de todas las entradas en el panel superior y, asimismo, permite seleccionar los campos que se van a filtrar utilizando las casillas de verificación de la columna de la izquierda. Cuando se ejecuta la ruta, solamente se conservan los campos marcados, junto con la predicción del objetivo. El resto de campos de entradas se descartan. Las selecciones predeterminadas se basan en las opciones especificadas en el nodo de modelado, pero se puede seleccionar o anular la selección de campos adicionales según sea necesario.

- Para ordenar la lista por rango, nombre del campo, importancia o cualquiera de las columnas que aparecen, pulse en la cabecera de la columna. También puede seleccionar el elemento que desee de la lista Ordenar por y usar las flechas hacia arriba y hacia abajo para cambiar la dirección de la ordenación.
- Puede utilizar la barra de herramientas para seleccionar o anular la selección de cualquier campo y para acceder al cuadro de diálogo Seleccionar campos, que le permite seleccionar campos por rango o importancia. También puede pulsar las teclas Mayús o Ctrl mientras pulsa los campos para ampliar la selección.
- Los valores de umbral para clasificar las entradas como importantes, marginales o sin importancia se muestran en la leyenda bajo la tabla. Estos valores se especifican en el nodo de modelado.

Gestión de modelos de Oracle

Los modelos de Oracle se añaden a la paleta de modelos al igual que cualquier otro modelo de IBM SPSS Modeler y puede utilizarse de un modo muy parecido. Sin embargo, existen un par de diferencias significativas, ya que cada modelo de Oracle generado en IBM SPSS Modeler se refiere en realidad a un modelo almacenado en el servidor de una base de datos.

Pestaña Servidor del nugget de modelo de Oracle

La generación de un modelo ODM mediante IBM SPSS Modeler implica la generación de un modelo en IBM SPSS Modeler y la generación o sustitución de un modelo en la base de Oracle. Un modelo de IBM SPSS Modeler de este tipo hace referencia al contenido de un modelo de base de datos almacenado en un servidor de base de datos. IBM SPSS Modeler puede realizar la comprobación de la coherencia almacenando una cadena **clave del modelo** generada idéntica, tanto en el modelo de IBM SPSS Modeler como en el de Oracle.

La cadena clave para cada modelo de Oracle se muestra debajo de la columna *Información del modelo*, en el cuadro de diálogo Crear lista de modelos. La cadena clave para un modelo de IBM SPSS Modeler se muestra como **Clave de modelo** en la pestaña Servidor de un modelo de IBM SPSS Modeler (cuando se sitúa en una ruta).

El botón Comprobar de la pestaña Servidor de un nugget de modelo puede utilizarse para comprobar que las claves del modelo de IBM SPSS Modeler y el de Oracle coinciden. Si no es posible encontrar un modelo con el mismo nombre en Oracle o si las claves de modelos no coinciden, significa que el modelo de Oracle se ha eliminado o se ha generado de nuevo desde que se generó el modelo de IBM SPSS Modeler.

Pestaña Resumen del nugget de modelo de Oracle

La pestaña Resumen de un nugget de modelo muestra información sobre el propio modelo (*Análisis*), los campos utilizados en el modelo (*Campos*), la configuración utilizada al generar el modelo (*Configuración de creación*) y el entrenamiento del modelo (*Resumen de entrenamiento*).

Cuando se examina el nodo por primera vez, los resultados de la pestaña Resumen aparecen contraídos. Para ver los resultados de interés, utilice el control de expansión situado a la izquierda de un elemento con objeto de desplegarlo, o bien pulse en el botón **Expandir todo** para mostrar todos los resultados. Para

ocultar los resultados cuando haya terminado de consultarlos, utilice el control de expansión con objeto de contraer los resultados específicos que desee ocultar o pulse en el botón **Contraer todo** para contraer todos los resultados.

Análisis. Muestra información sobre el modelo específico. Si ha ejecutado un nodo Análisis conectado a este nugget de modelo, la información de dicho análisis también se mostrará en esta sección.

Campos. Enumera los campos utilizados como objetivo y entradas en la generación del modelo.

Configuración de creación. Contiene información sobre la configuración que se utiliza en la generación del modelo.

Resumen de entrenamiento. Muestra el tipo del modelo, la ruta utilizada para crearlo, el usuario que lo creó, cuándo se generó y el tiempo que se tardó en generar el modelo.

Pestaña Configuración del nugget de modelo de Oracle

La pestaña Configuración del nugget de modelo permite sustituir el ajuste de algunas opciones en el nodo de modelado para la puntuación.

Árbol de decisión de Oracle

Utilizar costes de clasificación errónea. Determina si se utilizarán costes de clasificación errónea en el modelo de árbol de decisión de Oracle. Para obtener más información, consulte el tema "Costes de clasificación errónea" en la página 32.

Identificador de regla. Si está seleccionada (activada), añade una columna de identificador de regla al modelo de árbol de decisión de Oracle. El identificador de regla identifica el nodo del árbol en el que se realiza una división en particular.

NMF de Oracle

Mostrar todas las características. Si está seleccionada (activada), muestra la identificación de las características y la confianza de todas las características, en vez de aquellos valores solamente para la mejor característica, en el modelo NMF de Oracle.

Enumeración de modelos de Oracle

El botón Crear lista de modelos de Oracle Data Mining abre un cuadro de diálogo que enumera los modelos de base de datos existentes y permite eliminar los modelos. Este cuadro de diálogo se abre desde el cuadro de diálogo Aplicaciones de ayuda o desde los cuadros de diálogo de generación, búsqueda y aplicación para los nodos relacionados con ODM.

La siguiente información se muestra para cada modelo:

- **Nombre del modelo.** Nombre del modelo utilizado para ordenar la lista
- **Información del modelo.** Información clave del modelo compuesta por la fecha y hora de la generación y el nombre de la columna de objetivo
- **Tipo de modelo.** Nombre del algoritmo que creó este modelo

Oracle Data Miner

Oracle Data Miner es la interfaz de usuario para Oracle Data Mining (ODM) y sustituye a la interfaz de usuario anterior de IBM SPSS Modeler para ODM. Oracle Data Miner se ha diseñado para incrementar la tasa de éxito del analista mediante la utilización adecuada de algoritmos de ODM. Estos objetivos se cubren de varias formas:

- Los usuarios necesitan ayuda adicional para aplicar una metodología dirigida tanto a la preparación de datos como a la selección de algoritmos. Oracle Data Miner satisface esta necesidad mediante actividades de minería de datos que guían a los usuarios paso a paso a través de la metodología adecuada.
- Oracle Data Miner incluye heurísticas ampliadas y mejoradas para la generación de modelos y asistentes de transformación, con el fin de reducir la probabilidad de error al especificar la configuración del modelo y la transformación.

Definición de una conexión con Oracle Data Miner

1. Oracle Data Miner se puede ejecutar desde cualquier cuadro de diálogo de generación, aplicación de nodos o salida de Oracle, a través del botón **Iniciar Oracle Data Miner**.

Figura 2. Botón Iniciar Oracle Data Miner

2. El cuadro de diálogo **Editar conexión** de Oracle Data Miner se muestra al usuario antes de que se ejecute la aplicación externa Oracle Data Miner (siempre que la opción Aplicaciones de ayuda se haya definido correctamente).

   Nota: este cuadro de diálogo solamente se muestra cuando no se ha definido un nombre de conexión.

   - Especifique un nombre para la conexión Data Miner e introduzca la información adecuada del servidor de Oracle 10gR1 o 10gR2. El servidor de Oracle debería ser el mismo especificado en IBM SPSS Modeler.
3. El cuadro de diálogo **Seleccionar conexión** de Oracle Data Miner incluye opciones para especificar el nombre de conexión, definido en el paso anterior, que se desea utilizar.

Consulte Oracle Data Miner en el sitio Web de Oracle para obtener información adicional sobre los requisitos, la instalación y el uso de Oracle Data Miner.

Preparación de los datos

Pueden ser útiles dos tipos de preparaciones de datos al utilizar el bayesiano ingenuo, el bayesiano adaptativo y la máquina de vectores de soporte que se proporciona con los algoritmos de Oracle Data Mining en el modelado de:

- **Intervalos** o conversión de campos numéricos continuos de rango a categorías para los algoritmos que no pueden aceptar los datos continuos.
- **Normalización** o las transformaciones que se aplican a los rangos numéricos a fin de que dispongan de medias similares y desviaciones estándar.

Intervalos

El nodo de intervalos de IBM SPSS Modeler ofrece una serie de técnicas para realizar operaciones de intervalos. Una operación de intervalo se define de tal manera que puede aplicarse a uno o varios campos. La ejecución de la operación de intervalos sobre un conjunto de datos crea umbrales y permite la creación de un nodo Derivar de IBM SPSS Modeler. La operación de derivación puede convertirse a SQL y aplicarse antes de la generación y puntuación de modelos. Este método crea una dependencia entre el modelo y el nodo Derivar que realiza el intervalo y que permite reutilizar las especificaciones del intervalo mediante varias tareas de modelado.

Normalización

Los campos continuos (rango numérico) utilizados como entradas en los modelos de la máquina de vectores de soporte deben normalizarse antes de la generación del modelo. En el caso de los modelos de

regresión, la normalización debe estar revertida para generar de nuevo la puntuación del resultado del modelo. La configuración del modelo SVM permite seleccionar **Puntuaciones Z**, **Mín.-Máx.** o **Ninguna**. Los coeficientes de normalización son generados por Oracle como un paso en el proceso de creación del modelo y, a continuación, se cargan a IBM SPSS Modeler y se almacenan con el modelo. En el momento en que se aplican, los coeficientes se convierten en expresiones de derivación de IBM SPSS Modeler y se utilizan para preparar los datos para la puntuación antes de pasar los datos al modelo. En este caso, la normalización está estrechamente relacionada con la tarea de modelado.

Ejemplos de Oracle Data Mining

Existen varias rutas de ejemplo que ejemplifican el uso de ODM con IBM SPSS Modeler. Estas rutas se pueden encontrar en la carpeta de instalación de IBM SPSS Modeler en *\Demos\Database_Modelling\Oracle Data Mining*.

Nota: se puede acceder a la carpeta Demos desde el grupo de programas IBM SPSS Modeler en el menú Inicio de Windows.

Las rutas de la tabla siguiente se pueden utilizar juntas en una secuencia como un ejemplo del proceso de minería de bases de datos, utilizando el algoritmo de la máquina de vectores de soporte (SVM) que se proporciona con Oracle Data Mining:

Tabla 4. Minería de bases de datos - rutas de ejemplo

Ruta	Descripción
1_upload_data.str	Se utiliza para depurar y cargar los datos desde un archivo sin formato a la base de datos.
2_explore_data.str	Ofrece un ejemplo de exploración de los datos con IBM SPSS Modeler.
3_build_model.str	Genera el modelo mediante el algoritmo nativo de base de datos.
4_evaluate_model.str	Se utiliza como ejemplo de evaluación del modelo con IBM SPSS Modeler.
5_deploy_model.str	Se utiliza para desplegar el modelo para la puntuación interna de la base de datos.

Nota: para poder ejecutar el ejemplo, las rutas se deben procesar en orden. Además, los nodos de origen y modelado de cada ruta se deben actualizar para que hagan referencia a un origen de datos válido para la base de datos que desee utilizar.

El conjunto de datos utilizado en las rutas de ejemplo está relacionado con aplicaciones de tarjetas de crédito y presenta un problema de clasificación con una mezcla de predictores continuos y categóricos. Si desea obtener más información acerca de este conjunto de datos, consulte el archivo *crx.names* ubicado en la misma carpeta que las rutas de ejemplo.

Este conjunto de datos se puede descargar desde UCI Machine Learning Repository, en *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/*.

Ruta de ejemplo: cargar datos

La primera ruta de ejemplo, *1_upload_data.str*, se utiliza para depurar y cargar los datos desde un archivo sin formato a Oracle.

Puesto que Oracle Data Mining necesita un campo de ID exclusivo, esta ruta inicial utiliza un nodo Derivar para añadir un nuevo campo al conjunto de datos llamado *ID*, con los valores exclusivos 1,2,3, utilizando la función `@INDEX` de IBM SPSS Modeler.

El nodo Rellenar se utiliza para gestionar los valores perdidos y sustituye los campos vacíos que se leen en el archivo de texto *crx.data* por valores *NULOS*.

Ruta de ejemplo: explorar datos

La segunda ruta de ejemplo, *2_explore_data.str*, se utiliza para demostrar el uso del nodo Auditoría de datos para conocer los conceptos básicos de los datos, incluyendo estadísticos de resumen y gráficos.

Si pulsa dos veces en un gráfico en el informe de auditoría de datos, aparecerá un gráfico más detallado donde podrá explorar un campo específico con más profundidad.

Ruta de ejemplo: generar modelo

La tercera ruta de ejemplo, *3_build_model.str*, ilustra la generación del modelo en IBM SPSS Modeler. Pulse dos veces en el nodo de origen de base de datos (etiquetado como CREDIT) para especificar el origen de datos. Para especificar la configuración de creación, haga doble clic en el nodo de creación (etiquetado inicialmente como CLASS, que cambia a FIELD16 cuando se especifica el origen de datos).

En la pestaña Modelo del cuadro de diálogo:

1. Asegúrese de que **ID** se selecciona como campo Exclusivo.
2. Asegúrese de que **lineal** esté seleccionado como la función kernel y **Puntuaciones z** como el método de normalización.

Ruta de ejemplo: evaluar modelo

La cuarta ruta de ejemplo, *4_evaluate_model.str*, ilustra las ventajas de utilizar IBM SPSS Modeler para el modelado interno de bases de datos. Una vez ejecutado el modelo, puede volver a añadirlo a la ruta de datos y evaluarlo con varias herramientas que se ofrecen en IBM SPSS Modeler.

Consulta de los resultados del modelado

Conecte un nodo de tabla al nugget de modelo para explorar sus resultados. El campo **$O-field16** muestra el valor predicho para *field16* en cada caso, y el campo **$OC-field16** muestra el valor de confianza para esta predicción.

Evaluación de los resultados del modelo

Puede usar el nodo Análisis para crear una matriz de coincidencias que muestre el patrón de coincidencias entre cada campo predicho y su campo objetivo. Ejecute el nodo Análisis para ver los resultados.

Asimismo, también puede utilizar el nodo Evaluación para crear un gráfico de elevación diseñado para mostrar las mejoras de precisión realizadas por el modelo. Ejecute el nodo Evaluación para ver los resultados.

Ruta de ejemplo: desplegar modelo

Una vez satisfecho con la precisión del modelo, puede desplegarlo para su uso con aplicaciones externas o para volver a publicarlo en la base de datos. En la ruta de ejemplo final, *5_deploy_model.str*, se leen los datos desde la tabla `CREDITDATA` y, a continuación, se puntúan y se publican en la tabla `CREDITSCORES` mediante el nodo Editor llamado *desplegar solución*.

Capítulo 5. Modelado de bases de datos con IBM InfoSphere Warehouse

IBM InfoSphere Warehouse y IBM SPSS Modeler

IBM InfoSphere Warehouse (ISW) proporciona un grupo de algoritmos de minería de datos incrustado en IBM DB2 RDBMS. IBM SPSS Modeler proporciona nodos que admiten la integración de los siguientes algoritmos de IBM:

- Árboles de decisión
- Reglas de asociación
- Clústeres demográficos
- Agrupación en clústeres de Kohonen
- Reglas de secuencias
- Regresión de transformación
- Regresión lineal
- Regresión polinómica
- bayesiano ingenuo
- Regresión Logística
- Serie temporal

Si desea obtener más información acerca de estos algoritmos, consulte la documentación de su instalación de IBM InfoSphere Warehouse.

Requisitos para la integración con IBM InfoSphere Warehouse

Las siguientes condiciones constituyen los requisitos previos para realizar el modelado interno de bases de datos mediante la minería de datos de InfoSphere Warehouse. Es posible que necesite consultar con el administrador de la base de datos para asegurarse de que se reúnen las condiciones.

- Ejecución de IBM SPSS Modeler con respecto a una instalación de IBM SPSS Modeler Server en Windows o UNIX.
- IBM DB2 Data Warehouse Edition versión 9.1

  o
- IBM InfoSphere Warehouse versión 9.5 Enterprise Edition
- Un origen de datos ODBC para conectarse a DB2, tal y como se describe a continuación.

Nota: optimización de SQL y modelado de bases de datos requieren que la conectividad de IBM SPSS Modeler Server esté activada en el equipo con IBM SPSS Modeler. Con esta configuración activada, puede acceder a los algoritmos de bases de datos, devolver SQL directamente desde IBM SPSS Modeler y acceder a IBM SPSS Modeler Server. Para verificar el estado de la licencia actual, seleccione las siguientes opciones en el menú de IBM SPSS Modeler.

Ayuda > **Acerca de** > **Detalles adicionales**

Si la conectividad está activada, verá la opción **Activación de servidor** en la pestaña Estado de licencia.

Activación de la integración con IBM InfoSphere Warehouse

Para activar la integración de IBM SPSS Modeler con IBM InfoSphere Warehouse Data Mining, necesitará configurar ISW y crear un origen ODBC, permitir la integración en el cuadro de diálogo Aplicaciones de ayuda de IBM SPSS Modeler y activar la generación y optimización de SQL.

Configuración de ISW

Para instalar y configurar ISW, siga las instrucciones de la *guía de instalación de InfoSphere Warehouse*.

Creación de un origen ODBC para ISW

Para activar la conexión entre ISW y IBM SPSS Modeler, debe crear un nombre de origen de datos del sistema ODBC (DSN).

Antes de crear un DSN, debe tener un conocimiento básico de las unidades y los orígenes de datos ODBC y el soporte de la base de datos en IBM SPSS Modeler.

Si IBM SPSS Modeler Server e IBM InfoSphere Warehouse Data Mining se están ejecutando en hosts distintos, cree el mismo DSN ODBC en cada uno de ellos. Asegúrese de utilizar el mismo nombre para este DSN en los dos hosts.

1. Instale los controladores ODBC. Se encuentran en el disco de instalación de IBM SPSS Data Access Pack enviado con esta versión. Ejecute el archivo *setup.exe* para iniciar el instalador y seleccione todos los controladores relevantes. Siga las instrucciones que aparecen en pantalla para instalar los controladores.
   a. Cree el DSN.

      Nota: la secuencia de menú depende de la versión de Windows.

      - **Windows XP.** En el menú Inicio, seleccione **Panel de control**. Pulse dos veces en **Herramientas administrativas** y, a continuación, pulse dos veces en **Orígenes de datos (ODBC)**.
      - **Windows Vista.** En el menú Inicio, seleccione **Panel de control** y, a continuación, **Sistema y mantenimiento**. Pulse dos veces en **Herramientas administrativas**, seleccione **Orígenes de datos (ODBC)** y, a continuación, pulse **Abrir**.
      - **Windows 7.** En el menú Inicio, seleccione **Panel de control**, luego **Sistema y seguridad** y, a continuación, **Herramientas administrativas**. Seleccione **Orígenes de datos (ODBC)** y, a continuación, pulse en **Abrir**.

   b. Pulse en la pestaña **DSN de sistema** y, a continuación, **Añadir**.
2. Seleccione el controlador **SPSS OEM 6.0 DB2 Wire Protocol**.
3. Pulse **Finalizar**.
4. En el cuadro de diálogo Configuración del controlador ODBC DB2 Wire Protocol:
   - Especifique un nombre de origen de datos.
   - Para la dirección IP, especifique el nombre de host del servidor en el que se encuentra DB2 RDBMS.
   - Acepte el valor predeterminado del puerto TCP (50000).
   - Especifique el nombre de la base de datos con la que va a conectar.
5. Pulse **Probar conexión**.
6. En el cuadro de diálogo de conexión con DB2 Wire Protocol, introduzca el nombre de usuario y la contraseña que le proporcionó el administrador de la base de datos y pulse en **Aceptar**.

   El mensaje **Conexión establecida** .

   IBM DB2 ODBC DRIVER. Si el controlador ODBC es IBM DB2 ODBC DRIVER, siga estos pasos para crear un DSN ODBC:
7. En el administrador de origen de datos ODBC, pulse en la pestaña **DSN de sistema** y, a continuación, **Añadir**.
8. Seleccione el controlador **IBM DB2 ODBC DRIVER** y pulse en **Finalizar**.
9. En la ventana Añadir de IBM DB2 ODBC DRIVER, especifique un nombre de origen de datos y, a continuación, para el alias de la base de datos, pulse **Añadir**.

10. En la ventana <Nombre de origen de datos> de configuración de CLI/ODBC, en la pestaña Origen de datos, especifique el ID de usuario y la contraseña que le proporcionó el administrador de la base de datos y, a continuación, pulse la pestaña **TCP/IP**.
11. En la pestaña TCP/IP, introduzca:
    - El nombre de la base de datos con la que desea conectar.
    - Un nombre de alias de la base de datos (no más de ocho caracteres).
    - El nombre de host del servidor de la base de datos con el que desea conectar.
    - El número de puerto para la conexión.
12. Pulse en la pestaña de **opciones de seguridad**, seleccione la opción de **especificación de las opciones de seguridad (opcional)** y acepte la opción predeterminada: **Utilizar valor de autenticación en la configuración DBM del servidor**.
13. Pulse en la pestaña **Origen de los datos** y, a continuación, en **Conectar**.

Aparecerá un mensaje indicando que **la conexión se ha probado correctamente**.

Configuración de ODBC para comentarios (Opcional)

Para recibir comentarios de IBM InfoSphere Warehouse Data Mining durante la generación de modelos y permitir a IBM SPSS Modeler cancelar la generación del modelo, lleve a cabo los siguientes pasos para configurar el origen de datos ODBC que se creó en la sección anterior. Tenga en cuenta que este paso de configuración permite a IBM SPSS Modeler leer datos de DB2 que pueden no confirmarse en la base de datos ejecutando las transacciones simultáneamente. Si tiene alguna duda sobre las implicaciones de este cambio, consulte con el administrador de la base de datos.

Controlador SPSS OEM 6.0 DB2 Wire Protocol. Para el controlador de Connect ODBC, siga estos pasos:

1. Inicie el administrador de origen de datos ODBC, seleccione el origen de datos creado en la sección anterior y pulse en el botón **Configurar**.
2. En el cuadro de diálogo de configuración del controlador ODBC DB2 Wire Protocol, pulse en la pestaña de **opciones avanzadas**.
3. Establezca el nivel de aislamiento predeterminado en **0-LECTURA NO CONFIRMADA** y, a continuación, pulse en **Aceptar**.

Controlador IBM DB2 ODBC. Para el controlador de IBM DB2, siga estos pasos:

4. Inicie el administrador de origen de datos ODBC, seleccione el origen de datos creado en la sección anterior y, a continuación, pulse en el botón **Configurar**.
5. En el cuadro de diálogo Configuración de CLI/ODBC, pulse en la pestaña **Configuración avanzada** y, a continuación, en el botón **Añadir**.
6. En el cuadro de diálogo de adición de parámetros de CLI/ODBC, seleccione el parámetro **AISLAMIENTO TXN** y pulse en **Aceptar**.
7. En el cuadro de diálogo del nivel de aislamiento, seleccione la opción de **lectura no confirmada** y pulse en **Aceptar**.
8. En el cuadro de diálogo Configuración de CLI/ODBC, pulse en **Aceptar** para completar la configuración.

Observe que los comentarios de IBM InfoSphere Warehouse Data Mining aparecen en el siguiente formato:

```
<NºITERACIONES> / <PROGRESO> / <FASEKERNEL>
```

donde:

- `<NºITERACIONES>` indica el número de pasadas actuales por los datos, comenzando por 1.
- `<PROGRESO>` indica el progreso de la iteración actual con un número de 0,0 a 1,0.
- `<FASEKERNEL>` describe la fase actual del algoritmo de minería de datos.

Habilitar la integración de IBM InfoSphere Warehouse Data Mining en IBM SPSS Modeler

Para permitir que IBM SPSS Modeler utilice DB2 con IBM InfoSphere Warehouse Data Mining, primero debe proporcionar algunas especificaciones en el cuadro de diálogo Aplicaciones de ayuda.

1. Seleccione en los menús de IBM SPSS Modeler:
   Herramientas > **Opciones** > **Aplicaciones de ayuda**
2. Pulse en la pestaña **IBM InfoSphere Warehouse**.

Habilite integración de InfoSphere Warehouse Data Mining. Habilita la paleta de modelado de bases de datos (si todavía no se muestra) en la parte inferior de la ventana de IBM SPSS Modeler y añade los nodos para algoritmos de ISW Data Mining.

Conexión de DB2. Especifica el origen de datos ODBC de DB2 que se utiliza de forma predeterminada para la generación y el almacenamiento de los modelos. Esta configuración se puede omitir en la generación individual de modelos y nodos de modelos generados. Pulse el btón de puntos suspensivos (**...**) para seleccionar el origen de datos.

La conexión de base de datos que se utiliza para el modelado puede ser la misma que se utiliza para acceder a los datos, pero también puede ser otra diferente. Por ejemplo, puede tener una ruta que accede a los datos desde una base de datos de DB2, descarga los datos a IBM SPSS Modeler para realizar depuraciones y otras manipulaciones y, a continuación, carga los datos en una base de datos de DB2 diferente para realizar el modelado. Si lo prefiere, los datos originales pueden encontrarse en un archivo sin formato u otro origen (no perteneciente a DB2), en cuyo caso sería necesario cargar los datos a DB2 para realizar el modelado. En todos los casos, los datos se cargarán de manera automática en una tabla temporal creada en la base de datos que se utiliza para el modelado, si es necesario.

Avisar cuando haya de sobrescribirse un modelo de integración de InfoSphere Warehouse Data Mining. Seleccione esta opción para asegurarse de que IBM SPSS Modeler no sobrescribe los modelos almacenados en la base de datos sin avisar con anterioridad.

Crear lista de modelos de InfoSphere Warehouse Data Mining. Esta opción permite enumerar y eliminar los modelos almacenados en DB2. Para obtener más información, consulte el tema “Creación de lista de modelos de la base de datos” en la página 59.

Habilitar el inicio de la visualización de InfoSphere Warehouse Data Mining. Si ha instalado el módulo de visualización, debe activarlo aquí para utilizarlo en IBM SPSS Modeler.

Ruta del ejecutable de visualización. La ubicación del ejecutable del módulo de visualización (si está instalado); por ejemplo, *C:\Archivos de programa\IBM\ISWarehouse\Im\IMVisualization\bin\imvisualizer.exe*.

Directorio del complemento de visualización de series temporales. La ubicación del complemento de visualización de series temporales (si está instalado); por ejemplo, *C:\Archivos de programa\IBM\ISWShared\plugins\com.ibm.datatools.datamining.imvisualization.flash_2.2.1.v20091111_0915*.

Habilitar opciones de consumo de InfoSphere Warehouse Data Mining. Puede establecer un límite de consumo de memoria en un algoritmo de minería interna de bases de datos y especificar otras opciones arbitrarias en forma de línea de comandos para modelos específicos. El límite de memoria permite controlar el consumo de memoria y especificar un valor para la opción de potencia `-buf`. Aquí se pueden especificar otras opciones de potencia en forma de línea de comandos y también pasarlas a IBM InfoSphere Warehouse Data Mining. Para obtener más información, consulte el tema “Opciones de potencia” en la página 60.

Comprobar versión de InfoSphere Warehouse. Comprueba la versión de IBM InfoSphere Warehouse que está utilizando e indica un error si intenta utilizar una característica de minería de datos que no sea compatible con su versión.

Activación de optimización y generación de SQL

1. Seleccione en los menús de IBM SPSS Modeler:
   Herramientas > **Propiedades de ruta** > **Opciones**
2. Pulse en la opción **Optimización** en el panel de navegación.
3. Confirme que la opción **Generar SQL** está activada. Esta configuración es necesaria para que el modelado de bases de datos funcione.
4. Seleccione las opciones **Optimizar generación de SQL** y **Optimizar otra ejecución** (no obligatorias, pero recomendadas para un rendimiento optimizado).

Generación de modelos con IBM InfoSphere Warehouse Data Mining

La generación de modelos con IBM InfoSphere Warehouse Data Mining requiere que el conjunto de datos de entrenamiento se encuentre en una tabla o vista dentro de la base de datos de DB2. Si los datos no se encuentran en DB2 o se tienen que procesar en IBM SPSS Modeler como parte de la preparación de datos que no se puede realizar en DB2, los datos se cargarán automáticamente en una tabla temporal de DB2 antes de la generación de modelos.

Despliegue y puntuación de modelos

La puntuación de modelos siempre ocurre dentro de DB2 y se realiza siempre por medio de IBM InfoSphere Warehouse Data Mining. Es posible que sea necesario cargar el conjunto de datos en una tabla temporal si los datos se originan o se tienen que preparar dentro de IBM SPSS Modeler. En el caso de los modelos Árbol de decisión, Regresión y Agrupación en clústeres de IBM SPSS Modeler, generalmente, sólo se proporciona una única predicción y una confianza o probabilidad asociada. Además, una opción de usuario para mostrar las probabilidades o confianzas de cada resultado (similar a la de la regresión logística) es una opción de puntuación de tiempo disponible en la pestaña Configuración del nugget de modelo (la casilla de verificación **Incluir confianzas para todas las clases**). En el caso de modelos Asociación y Secuencia de IBM SPSS Modeler, se proporcionan varios valores. IBM SPSS Modeler puede puntuar modelos de IBM InfoSphere Warehouse Data Mining desde dentro de rutas publicadas para su ejecución utilizando IBM SPSS Modeler Solution Publisher.

La tabla siguiente explica los campos generados por los modelos de puntuación.

Tabla 5. Campos de puntuación de modelos

Tipo de modelo	Columnas de puntuación	Descripción
Árboles de decisión	$I-*campo*	La mejor predicción para *campo*.
	$IC-*campo*	Confianza de la mejor predicción para *campo*.
	valor 1 de $IC, ..., *valor N* de $IC	Confianza de cada valor posible de *N* de *campo* (opcional).
Regresión	$I-*campo*	La mejor predicción para *campo*.
	$IC-*campo*	Confianza de la mejor predicción para *campo*.
Clústeres	$I-*nombre_modelo*	Mejor asignación de clúster para el registro de entrada.
	$IC-*nombre_modelo*	Confianza de la mejor asignación de clúster para el registro de entrada.
Association	$I-*nombre_modelo*	Identificador de regla coincidente.

Tabla 5. Campos de puntuación de modelos (continuación)

Tipo de modelo	Columnas de puntuación	Descripción
	$IH-*nombre_modelo*	Elemento principal.
	$IHN-*nombre_modelo*	Nombre de elemento principal.
	$IS-*nombre_modelo*	Valor de soporte de regla coincidente.
	$IC-*nombre_modelo*	Valor de confianza de regla coincidente.
	$IL-*nombre_modelo*	Valor de elevación de regla coincidente.
	$IMB-*nombre_modelo*	Número de elementos del cuerpo o conjuntos de elementos del cuerpo coincidentes (como todos los elementos del cuerpo o conjuntos de elementos del cuerpo deben coincidir con este número, éste es igual al número de elementos del cuerpo o conjuntos de elementos del cuerpo).
Secuencia	$I-*nombre_modelo*	Identificador de regla coincidente
	$IH-*nombre_modelo*	Conjunto de elementos principales de regla coincidente
	$IHN-*nombre_modelo*	Nombres de los elementos del conjunto de elementos principales de regla coincidente
	$IS-*nombre_modelo*	Valor de soporte de regla coincidente
	$IC-*nombre_modelo*	Valor de confianza de regla coincidente
	$IL-*nombre_modelo*	Valor de elevación de regla coincidente
	$IMB-*nombre_modelo*	Número de elementos del cuerpo o conjuntos de elementos del cuerpo coincidentes (como todos los elementos del cuerpo o conjuntos de elementos del cuerpo deben coincidir con este número, éste es igual al número de elementos del cuerpo o conjuntos de elementos del cuerpo)
bayesiano ingenuo	$I-*campo*	La mejor predicción para *campo*.
	$IC-*campo*	Confianza de la mejor predicción para *campo*.
Regresión Logística	$I-*campo*	La mejor predicción para *campo*.
	$IC-*campo*	Confianza de la mejor predicción para *campo*.

Gestión de modelos DB2

La generación de un modelo de IBM InfoSphere Warehouse Data Mining mediante IBM SPSS Modeler implica la generación de un modelo en IBM SPSS Modeler y la generación o sustitución de un modelo en la base de datos de DB2. El modelo de IBM SPSS Modeler de este tipo hace referencia al contenido de un modelo de base de datos almacenado en el servidor de una base de datos. IBM SPSS Modeler puede

realizar la comprobación de coherencia almacenando una cadena clave del modelo generada idéntica, tanto en el modelo de IBM SPSS Modeler como en el de DB2.

La cadena clave para cada modelo de DB2 se muestra debajo de la columna de *información del modelo* en el cuadro de diálogo de creación de lista de modelos de la base de datos. La cadena clave para un modelo de IBM SPSS Modeler se muestra como Clave de modelo en la pestaña Servidor de un modelo de IBM SPSS Modeler (cuando se sitúa en una ruta).

El botón Comprobar se puede utilizar para comprobar que las claves de los modelos de IBM SPSS Modeler y DB2 coinciden. Si no es posible encontrar un modelo con el mismo nombre en DB2 o si las claves de modelos no coinciden, significa que el modelo de DB2 se ha eliminado o se ha generado de nuevo desde que se generó el modelo de IBM SPSS Modeler. Para obtener más información, consulte el tema “Pestaña Servidor del nugget de modelo de ISW” en la página 76.

Creación de lista de modelos de la base de datos

IBM SPSS Modelerofrece un cuadro de diálogo que permite enumerar los modelos almacenados en IBM InfoSphere Warehouse Data Mining y también eliminarlos. Se puede acceder a este cuadro de diálogo desde el cuadro de diálogo Aplicaciones de ayuda de IBM y desde los cuadros de diálogo de generación, exploración y aplicación de los nodos relacionados con IBM InfoSphere Warehouse Data Mining. La siguiente información se muestra para cada modelo:

- Nombre del modelo (nombre del modelo, utilizado para ordenar la lista).
- Información del modelo (información de la clave del modelo a partir de una clave aleatoria generada cuando IBM SPSS Modeler genera el modelo).
- Tipo de modelo (la tabla de DB2 en la que IBM InfoSphere Warehouse Data Mining ha almacenado el modelo).

Exploración de modelos

La herramienta Visualizer es el único método para explorar los modelos de InfoSphere Warehouse Data Mining. La herramienta puede instalarse opcionalmente con InfoSphere Warehouse Data Mining. Para obtener más información, consulte el tema “Activación de la integración con IBM InfoSphere Warehouse” en la página 53.

- Pulse en **Ver** para ejecutar la herramienta de visualización. Lo que muestra la herramienta depende del tipo de nodo generado. Por ejemplo, la herramienta de visualización mostrará una vista de clases predichas cuando se inicie desde un nugget de modelo Árbol de decisión de ISW.
- Pulse en **Resultados de prueba** (sólo árboles de decisión y secuencia) para ejecutar la herramienta de visualización y ver la calidad general del modelo generado.

Exportación de modelos y generación de nodos

Puede realizar acciones de importación y exportación PMML en los modelos de IBM InfoSphere Warehouse Data Mining. El PMML que se exporta es el PMML original generado por IBM InfoSphere Warehouse Data Mining. La función de exportación devuelve el modelo en formato PMML.

Puede exportar una estructura y un resumen del modelo a archivos con formato de texto y HTML. Puede generar los nodos Filtrar, Seleccionar y Derivar adecuados donde proceda. Para obtener más información, consulte "Exportación de modelos" en el *Manual del usuario de IBM SPSS Modeler*.

Configuración de nodos común a todos los algoritmos

La siguiente configuración es común en muchos de los algoritmos de IBM InfoSphere Warehouse Data Mining:

Objetivo y predictores. Puede especificar un objetivo y predictores mediante el nodo Tipo o utilizando manualmente la pestaña Campos del nodo de generación de modelos, como es estándar en IBM SPSS Modeler.

Origen de datos ODBC Esta configuración permite al usuario reemplazar los datos ODBC predeterminados para el modelo actual. (El valor predeterminado se especifica en el cuadro de diálogo Aplicaciones de ayuda. Para obtener más información, consulte el tema "Activación de la integración con IBM InfoSphere Warehouse" en la página 53.)

Opciones de la pestaña Servidor de ISW

Puede especificar la conexión de DB2 utilizada para cargar los datos para el modelado. Si fuera necesario, puede seleccionar una conexión en la pestaña Servidor para cada nodo de modelado a fin de omitir la conexión de DB2 predeterminada especificada en el cuadro de diálogo Aplicaciones de ayuda. Para obtener más información, consulte el tema "Activación de la integración con IBM InfoSphere Warehouse" en la página 53.

La conexión que se utiliza para el modelado puede ser la misma que se utiliza en el nodo de origen de una ruta, o puede ser otra diferente. Por ejemplo, puede tener una ruta que accede a los datos desde una base de datos de DB2, descarga los datos a IBM SPSS Modeler para realizar depuraciones y otras manipulaciones y, a continuación, carga los datos en una base de datos de DB2 diferente para realizar el modelado.

El nombre de origen de datos ODBC se incrusta de manera efectiva en cada ruta de IBM SPSS Modeler. Si una ruta creada en un host se ejecuta en un host diferente, el nombre del origen de datos debe ser el mismo en cada host. Si lo prefiere, se puede seleccionar un origen de datos diferente en la pestaña Servidor de cada nodo de origen o de modelado.

Puede obtener comentarios a medida que genera un modelo, utilizando las siguientes opciones:

- **Activar comentarios.** Seleccione esta opción para obtener comentarios durante la generación de un modelo (el valor predeterminado es desactivada).
- **Intervalo de comentarios (en segundos).** Especifique la frecuencia con la que IBM SPSS Modeler recupera comentarios en el progreso de generación de un modelo.

Habilitar opciones de consumo de InfoSphere Warehouse Data Mining. Seleccione esta opción para activar el botón **Opciones de potencia**, que permite especificar un conjunto de opciones avanzadas, como un límite de memoria y SQL personalizado. Para obtener más información, consulte el tema "Opciones de potencia".

La pestaña Servidor de un nodo generado incluye una opción para realizar una comprobación de coherencia almacenando una cadena clave del modelo generado idéntica tanto en el modelo de IBM SPSS Modeler como en el de DB2. Para obtener más información, consulte el tema "Pestaña Servidor del nugget de modelo de ISW" en la página 76.

Opciones de potencia

La pestaña Servidor de los algoritmos existentes incluye una casilla de verificación para activar las opciones de potencia de modelado de ISW. Cuando se pulsa en el botón **Opciones de potencia**, aparece el cuadro de diálogo de opciones de potencia de ISW, con funciones para:

- Límite de memoria.
- Otras opciones de potencia.
- SQL personalizado de datos de análisis.
- SQL personalizado de datos lógicos.
- SQL personalizado de configuración de análisis.

Límite de memoria. Limita el consumo de memoria de un algoritmo de generación de modelos. Tenga en cuenta que la opción de potencia estándar establece un límite en el número de valores discretos de los datos categóricos.

Otras opciones de potencia. Permite especificar opciones de potencia arbitrarias en forma de línea de comandos para modelos o soluciones en concreto. Los valores específicos pueden variar, dependiendo de la implementación o solución. Puede ampliar manualmente el SQL generado por IBM SPSS Modeler para definir una tarea de generación de modelos.

SQL personalizado de datos de análisis. Puede añadir llamadas a métodos para modificar el objeto `DM_MiningData`. Por ejemplo, si introduce el siguiente SQL, se añade un filtro basado en un campo de *partición* a los datos utilizados en la generación de modelos:

```
..DM_setWhereClause('"partición" = 1')
```

SQL personalizado de datos lógicos. Puede añadir llamadas a métodos para modificar el objeto `DM_LogicalDataSpec`. Por ejemplo, el siguiente SQL elimina un campo del conjunto de campos utilizado para la generación de modelos:

```
..DM_remDataSpecFld('campo6')
```

SQL personalizado de configuración de análisis. Puede añadir llamadas a métodos para modificar el objeto `DM_ClasSettings/DM_RuleSettings/DM_ClusSettings/DM_RegSettings`. Por ejemplo, si introduce el siguiente SQL, se indica a IBM InfoSphere Warehouse Data Mining que establezca el campo de *partición* en activo (lo que significa que siempre se incluirá en el modelo resultante):

```
..DM_setFldUsageType('partición',1)
```

Opciones de costes de ISW

En la pestaña Costes, puede ajustar los costes de clasificación errónea, lo que le permite especificar la importancia relativa de los distintos tipos de errores de predicción.

En algunos contextos, ciertos tipos de errores son más costosos que otros. Por ejemplo, puede resultar más costoso clasificar a un solicitante de crédito de alto riesgo como de bajo riesgo (un tipo de error) que clasificar a un solicitante de crédito de bajo riesgo como de alto riesgo (otro tipo de error). Los costes de clasificación errónea permiten especificar la importancia relativa de los diversos tipos de errores de predicción.

Los costes de clasificación errónea son básicamente ponderaciones aplicadas a resultados específicos. Estas ponderaciones se extraen en el modelo y pueden realmente cambiar la predicción (como forma de protección frente a errores costosos).

Salvo los modelos C5.0, los costes de clasificación errónea no se aplican cuando se puntúa un modelo y no se tienen en cuenta cuando se clasifican o comparan modelos utilizando un nodo Clasificador automático, un diagrama de evaluación o un nodo Análisis. Es posible que un modelo que incluya costes no produzca menos errores que uno que no lo haga, y es posible que no ordene ningún valor mayor en términos de precisión global, pero es probable que funcione mejor en términos prácticos porque contiene un sesgo integrado en favor de errores *más baratos*.

La matriz de costes muestra el coste para cada combinación posible de categoría predicha y categoría real. De forma predeterminada, todos los costes de clasificación errónea se establecen en 1,0. Para introducir valores de coste personalizados, seleccione **Utilizar costes de clasificación errónea** e introduzca valores personalizados en la matriz de costes.

Para cambiar un coste de clasificación errónea, seleccione la casilla correspondiente a la combinación deseada de valores predichos y reales, elimine el contenido existente de la casilla e introduzca en ella el coste deseado. Los costes no son simétricos automáticamente. Por ejemplo, si establece el coste de clasificación errónea *A* como *B* para que sea 2,0, el coste de clasificación errónea de *B* como *A* aún tendrá el valor predeterminado 1,0 hasta que también se modifique explícitamente.

Árbol de decisión de ISW

Los modelos de árboles de decisión permiten desarrollar sistemas de clasificación que predicen o clasifican observaciones según un conjunto de reglas de decisión. Si dispone de datos divididos en clases que le interesan (por ejemplo, préstamos de alto riesgo frente a préstamos de bajo riesgo, suscriptores frente a no suscriptores, votantes frente a no votantes o tipos de bacterias), puede usar los datos para generar reglas que pueda usar para clasificar casos antiguos o recientes con la máxima precisión. Por ejemplo, podría generar un árbol que clasificara el riesgo de crédito o la intención de compra basándose en la edad y otros factores.

El algoritmo de árboles de decisión de ISW genera árboles de clasificación en datos de entrada categóricos. El árbol de decisión resultante es binario. Se pueden aplicar varias configuraciones para generar el modelo, incluyendo costes de clasificación errónea.

La herramienta ISW Visualizer es el único método para explorar los modelos de IBM InfoSphere Warehouse Data Mining.

Opciones de modelo para los árboles de decisión de ISW

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilizar los datos en particiones. Si define un campo de partición, seleccione **Utilizar los datos en particiones**.

Realizar ejecución de comprobación. Puede elegir realizar una ejecución de comprobación. Se realizará una ejecución de comprobación de IBM InfoSphere Warehouse Data Mining una vez generado el modelo en la partición de entrenamiento. De este modo se realizará una pasada por la partición de comprobación para establecer información de calidad del modelo, gráficos de elevación, etc.

Máxima profundidad de árbol. Se puede especificar la profundidad máxima del árbol. De este modo se limita la profundidad del árbol al número especificado de niveles. Si no selecciona esta opción, no se aplicará ningún límite. Para evitar demasiados modelos complejos, no se suele recomendar un valor superior a 5.

Opciones de Experto para los árboles de decisión de ISW

Pureza máxima. Esta opción establece la pureza máxima de los nodos internos. Si la división de un nodo hace que uno de las hijos supere la medida de pureza especificada (por ejemplo, más del 90% de los casos corresponde a una categoría especificada), el nodo no se dividirá.

Número mínimo de casos por nodo interno. Si la división de un nodo resulta en un nodo con menos casos que el mínimo especificado, el nodo no se dividirá.

Asociación de ISW

Puede utilizar el nodo Asociación de ISW para buscar reglas de asociación entre elementos que están presentes en un conjunto de grupos. Las reglas de asociación asocian una conclusión concreta (por ejemplo, la compra de un producto en especial) con un conjunto de condiciones (la compra de varios productos).

Puede incluir o excluir reglas de asociación del modelo especificando **restricciones**. Si selecciona incluir un campo de entrada particular, las reglas de asociación que contienen al menos uno de los elementos específicos se incluyen en el modelo. Si excluye un campo de entrada, las reglas de asociación que contienen algunos de los elementos especificados se descartan de los resultados.

Los algoritmos de ISW y Sequence pueden utilizar las **taxonomías**. Las taxonomías establecen correlaciones entre valores individuales y conceptos de un nivel superior. Por ejemplo, los bolígrafos y los lápices se pueden correlacionar a una categoría de papelería.

Las reglas de asociación tienen una única consecuente (la conclusión) y antecedentes múltiples (el conjunto de condiciones). A continuación se ofrece un ejemplo:

```
[Pan, Mermelada]  ⌂  [Mantequilla]
[Pan, Mermelada]
 ⌂  [Margarina]
```

En este caso, `Pan` y `Mermelada` son los antecedentes (también conocidos como **cuerpo de la regla**) y `Mantequilla` o `Margarina` son ejemplos de una consecuente (también conocida como **cabecera de regla**). La primera regla indica que una persona que ha comprado pan y mermelada también ha comprado mantequilla al mismo tiempo. La segunda regla identifica un cliente que, cuando compra la misma combinación (pan y mermelada) también ha comprado margarina en la misma visita a la tienda.

La herramienta Visualizer es el único método para explorar los modelos de IBM InfoSphere Warehouse Data Mining.

Opciones de campo de Asociación de ISW

En la pestaña Campos, puede especificar los campos que se van a utilizar para generar el modelo.

Para generar un modelo, primero se deben especificar los campos que se desea usar como objetivos y como entradas. Salvo algunas excepciones, todos los nodos de modelado usarán la información de los campos procedente de un nodo Tipo anterior en la ruta. Con el ajuste predeterminado mediante el cual se utiliza el nodo Tipo para seleccionar los campos de entrada y objetivo, el único ajuste que puede cambiar en esta pestaña es el diseño de tabla para datos no transaccionales.

Utilizar configuración del nodo Tipo. Esta opción especifica el uso de la información de campo de un nodo Tipo situado en un punto anterior de la ruta. Éste es el valor por omisión.

Utilizar configuración personalizada. Esta opción especifica el uso de la información de campo introducida aquí en lugar de la proporcionada en nodos Tipo situados en cualquier punto anterior de la ruta. Después de seleccionar esta opción, especifique los campos siguientes si es necesario.

Uso del formato transaccional. Seleccione la casilla de verificación si los datos de origen están en el **formato transaccional**. Los registros de este formato tienen dos campos, uno para una ID y otro para el contenido. Cada registro representa un único elemento o transacción y los elementos asociados se enlazan usando el mismo ID. Cancele esta selección si los datos están en **formato tabular**, en los que los elementos se representan por marcas separadas, donde cada campo de marca representa la presencia o ausencia de un elemento específico y cada registro representa un conjunto completo de elementos asociados.

- **ID.** Para los datos transaccionales, seleccione el campo de ID de la lista. Los campos numéricos o simbólicos se pueden utilizar como campo de ID. Cada valor exclusivo de este campo debe indicar una unidad de análisis específica. Por ejemplo, en una aplicación de la cesta de la compra, cada ID puede representar a un sólo cliente. Para una aplicación de análisis del registro Web, cada ID puede representar un equipo (con la dirección IP) o un usuario (con los datos de inicio de sesión).
- **Contenido.** Especifique el campo o campos de contenido del modelo. Estos campos contienen los elementos de interés del modelo de asociación. Puede especificar un único campo nominal cuando los datos estén en formato transaccional.

Uso del formato tabular. Anule la selección de la casilla de verificación **Utilizar formato transaccional** si los datos de origen están en el formato tabular.

- **Entradas.** Seleccione el campo(s) de entrada. Se trata de una acción similar a establecer el rol del campo en *Entrada* en un nodo Tipo.

- **Partición.** Este campo permite especificar un campo usado para dividir los datos en muestras independientes para las fases de entrenamiento, prueba y validación en la generación del modelo. Si usa una muestra para generar el modelo y otra muestra distinta para comprobarlo, podrá obtener una buena indicación de la bondad del modelo a la hora de generar conjuntos de datos de mayor tamaño similares a los datos actuales. Si se han definido varios campos de partición mediante nodos Tipo o Partición, se deberá seleccionar un campo de partición simple en la pestaña Campos en todos los nodos de modelado que usen la partición. (Si solamente hay una partición, se usará automáticamente siempre que se active la partición.) Debe tener en cuenta que al aplicar la partición seleccionada en su análisis, también debe activar la partición en la pestaña Opciones del modelo para el nodo. (Si se elimina la selección de esta opción, se posibilita la desactivación de la partición sin cambiar la configuración del campo.)

Diseño de tabla para datos no transaccionales. Para datos tabulares, puede seleccionar un diseño de tabla estándar (valor predeterminado) o un diseño de longitud de elemento limitada.

En el diseño predeterminado, el número de columnas está determinado por el número total de elementos asociados.

Tabla 6. Diseño de tabla predeterminado.

Group ID	Cuenta corriente	Cuenta de ahorro	Tarjeta de crédito	Loan	Cuenta de custodia
Herrero	Y	Y	Y	-	-
Jackson	Y	-	Y	Y	Y
Douglas	Y	-	-	-	Y

En el diseño de longitud de elemento limitada, el número de columnas está determinado por el número mayor de elementos asociados en cualquiera de las filas.

Tabla 7. Diseño de tabla de longitud de elemento limitada.

Group ID	Elemento1	Elemento2	Elemento3	Elemento4
Herrero	cuenta corriente	cuenta de ahorro	Tarjeta de crédito	-
Jackson	cuenta corriente	Tarjeta de crédito	préstamo	cuenta de custodia
Douglas	cuenta corriente	cuenta de custodia	-	-

Opciones de modelos de asociación de ISW

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo.

Soporte mínimo de las reglas (%). Nivel de contabilidad mínima de las reglas de asociación o de secuencia. El modelo incluye las reglas que alcanzan como mínimo este nivel de soporte únicamente. El valor se calcula como A/B*100, donde A es el número de grupos que contienen todos los elementos que aparecen en la regla, y B es el número total de los grupos que se toman en consideración. Si desea centrarse en las asociaciones o secuencias más comunes, aumente el valor de este parámetro.

Confianza mínima de las reglas (%). Nivel de confianza mínima de las reglas de asociación o de secuencia. El modelo incluye las reglas que alcanzan como mínimo este nivel de confianza únicamente. El valor se calcula como m/n*100, donde m es el número de grupos que contienen la cabecera de regla unido (consecuente) y el cuerpo de la regla (antecedente), y n es el número de grupos que contienen el

cuerpo de la regla. Si se obtienen demasiadas asociaciones o secuencias sin interés, pruebe a aumentar el valor de este parámetro. Si se obtienen muy pocas asociaciones o secuencias, pruebe a disminuir el valor de este parámetro.

Tamaño máximo de regla. Número máximo de elementos permitidos en una regla, incluyendo el elemento consecuente. Si las asociaciones o secuencias de interés resultantes son pocas, se puede disminuir el valor del parámetro para que el conjunto se genere más rápido.

Nota: sólo los nodos con formato de entrada transaccional se puntúan; los formatos de tabla de Verdad (datos tabulares) permanecen sin refinar.

Opciones de Experto de Asociación de ISW

En la pestaña Experto del nodo de asociación, puede especificar las reglas de asociación que se incluirán en los resultados o se excluyen de los resultados. Si decide incluir elementos especificados, las reglas que contienen al menos uno de los elementos especificados se incluyen en el modelo. Si decide excluir elementos especificados, las reglas que contienen cualquiera de los elementos especificados se descartan de los resultados.

Si se selecciona la opción **Utilizar restricciones de elemento**, cualquier elemento que se haya añadido a la lista de restricciones se incluirá o excluirá de los resultados, en función de la configuración de **Restringir tipo**.

Restringir tipo. Seleccione si desea incluir o excluir de los resultados las reglas de asociación que contienen los elementos específicos.

Editar restricciones. Para añadir un elemento a la lista de elementos restringidos, selecciónelo de la lista de elementos y pulse en el botón de flecha derecha.

Opciones de taxonomía de ISW

Los algoritmos de ISW y Sequence pueden utilizar las **taxonomías**. Las taxonomías establecen correlaciones entre valores individuales y conceptos de un nivel superior. Por ejemplo, los bolígrafos y los lápices se pueden correlacionar a una categoría de papelería.

En la pestaña Taxonomía, puede definir correlaciones de categorías para expresar taxonomías en los datos. Por ejemplo, una taxonomía puede crear dos categorías (`normal` y `lujo`) y, a continuación asigna elementos básicos a cada una de estas categorías. Por ejemplo, `vino` se asigna a `lujo` y `pan` se asigna a `normal` la taxonomía tiene una estructura padre-hijo, como se muestra en la tabla siguiente.

Tabla 8. Ejemplo de estructura de taxonomía

Hijo	Nivel superior
vino	Lujo
pan	Grapa

Con esta taxonomía establecida, puede generar un modelo de asociación o secuencia que incluya reglas que impliquen a las categorías y a los elementos básicos.

Nota: para activar las opciones de esta pestaña, los datos de origen deben estar en formato transaccional y debe seleccionar **Utilizar formato transaccional** en la pestaña **Campos** y seleccionar **Utilizar taxonomía** en esta pestaña.

Nombre de tabla. Esta opción especifica el nombre de la tabla de DB2 para almacenar los detalles de la taxonomía.

Columna hija. Esta opción especifica el nombre de la columna hija en la tabla de la taxonomía. La columna hija contiene los nombres de los elementos o de las categorías.

Columna padre. Esta opción especifica el nombre de la columna padre en la tabla de la taxonomía. La columna padre contiene los nombres de las categorías.

Cargar detalles en tabla. Seleccione esta opción si la información de taxonomía almacenada en IBM SPSS Modeler se debe cargar a la tabla de taxonomía en el momento de la generación de modelos. Tenga en cuenta que si la tabla de taxonomía ya existe, se eliminará. La información sobre taxonomía se almacena con el nodo generador de modelos y se edita mediante los botones Editar categorías y Editar taxonomía.

Editor de categorías

El cuadro de diálogo Editar categorías permite añadir y eliminar categorías en una lista ordenada.

Para añadir una categoría, introduzca su nombre en el campo **Nueva categoría** y pulse en el botón de flecha para moverlo a la lista **Categorías**.

Para eliminar una categoría, selecciónela en la lista **Categorías** y pulse en el botón Eliminar.

Editor de taxonomías

El cuadro de diálogo Editar taxonomía permite combinar el conjunto de elementos básicos definido en los datos y el conjunto de categorías para generar una taxonomía. Para añadir entradas a la taxonomía, seleccione uno o más elementos o categorías de la lista de la izquierda y una o más categorías de la lista de la derecha y, a continuación, pulse en el botón de flecha. Tenga en cuenta que si alguna de las entradas añadidas a la taxonomía produce algún conflicto (por ejemplo, si se especifica `cat1 -> cat2` y lo opuesto, `cat2 -> cat1`), éstas no se añadirán.

Secuencia de ISW

El nodo Secuencia descubre patrones, en datos secuenciales u ordenados en el tiempo, con el formato `pan > queso`. Los elementos de una secuencia son **conjuntos de elementos** que constituyen una única transacción. Por ejemplo, si una persona va a la tienda y compra pan y leche y, varios días después, vuelve a la tienda para comprar un poco de queso, la actividad de compras de esa persona se puede representar como dos conjuntos de elementos. El primer conjunto de elementos contiene pan y leche y el segundo contiene queso. Una **secuencia** es una lista de conjuntos de elementos que tiende a producirse en un orden previsible. El nodo Secuencia detecta secuencias frecuentes y crea un nodo de modelo generado que se puede utilizar para realizar predicciones.

Puede utilizar la función de minería Normas de secuencia en varias áreas comerciales. Por ejemplo, en la industria al por menor, puede encontrar series típicas de compras. Estas series le muestran distintas combinaciones de clientes, productos y horario de la compra. Con esta información, es posible identificar los clientes potenciales para un producto específico. Además, puede ofrecer los productos a los clientes potenciales en la fecha prevista.

Una secuencia es un conjunto ordenado de conjuntos de elementos. Las secuencias contienen los siguientes niveles de agrupación.

- Eventos que suceden simultáneamente y constituyen una transacción única o un conjunto de elementos.
- Cada elemento o cada conjunto de elementos pertenece a un grupo de transacciones. Por ejemplo, un artículo adquirido pertenece a un cliente, una pulsación en una página específica pertenece a un usuario de Internet, o un componente pertenece a un vehículo fabricado. Varios conjuntos de elementos que suceden en distinto momento y pertenecen al mismo grupo de transacción forman una secuencia.

Opciones de modelos de secuencias de ISW

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo.

Soporte mínimo de las reglas (%). Nivel de contabilidad mínima de las reglas de asociación o de secuencia. El modelo incluye las reglas que alcanzan como mínimo este nivel de soporte únicamente. El valor se calcula como A/B*100, donde A es el número de grupos que contienen todos los elementos que aparecen en la regla, y B es el número total de los grupos que se toman en consideración. Si desea centrarse en las asociaciones o secuencias más comunes, aumente el valor de este parámetro.

Confianza mínima de las reglas (%). Nivel de confianza mínima de las reglas de asociación o de secuencia. El modelo incluye las reglas que alcanzan como mínimo este nivel de confianza únicamente. El valor se calcula como *m*/*n**100, donde *m* es el número de grupos que contienen la cabecera de regla unido (consecuente) y el cuerpo de la regla (antecedente), y *n* es el número de grupos que contienen el cuerpo de la regla. Si se obtienen demasiadas asociaciones o secuencias sin interés, pruebe a aumentar el valor de este parámetro. Si se obtienen muy pocas asociaciones o secuencias, pruebe a disminuir el valor de este parámetro.

Tamaño máximo de regla. Número máximo de elementos permitidos en una regla, incluyendo el elemento consecuente. Si las asociaciones o secuencias de interés resultantes son pocas, se puede disminuir el valor del parámetro para que el conjunto se genere más rápido.

Nota: sólo los nodos con formato de entrada transaccional se puntúan; los formatos de tabla de Verdad (datos tabulares) permanecen sin refinar.

Opciones de Experto de Secuencia de ISW

Puede especificar las reglas de secuencia que se incluirán o excluirán de los resultados. Si decide incluir elementos especificados, las reglas que contienen al menos uno de los elementos especificados se incluyen en el modelo. Si decide excluir elementos especificados, las reglas que contienen cualquiera de los elementos especificados se descartan de los resultados.

Si se selecciona la opción **Utilizar restricciones de elemento**, cualquier elemento que se haya añadido a la lista de restricciones se incluirá o excluirá de los resultados, en función de la configuración de **Restringir tipo**.

Restringir tipo. Seleccione si desea incluir o excluir de los resultados las reglas de asociación que contienen los elementos específicos.

Editar restricciones. Para añadir un elemento a la lista de elementos restringidos, selecciónelo de la lista de elementos y pulse en el botón de flecha derecha.

Regresión de ISW

El nodo Regresión de ISW admite los siguientes algoritmos de regresión:

- Transformación (valor predeterminado)
- Aritmético
- Polinómico
- RBF

Regresión de transformación

El algoritmo de regresión de transformación de ISW genera modelos que forman árboles de decisión con ecuaciones de regresión en sus hojas. Tenga en cuenta que el Visualizador de IBM no mostrará la estructura de estos modelos.

El explorador de IBM SPSS Modeler muestra los parámetros y las anotaciones. Sin embargo, la estructura de modelos no se puede explorar. Hay relativamente pocos parámetros de generación que puede configurar el usuario.

Regresión lineal

El algoritmo de regresión lineal de ISW asume una relación linean entre los campos explicatorios y el campo objetivo. Genera modelos que representan ecuaciones. Se espera que el valor predicho difiera del valor observado, ya que la ecuación de regresión es una aproximación del campo objetivo. La diferencia se denomina residuo.

El modelado de IBM InfoSphere Warehouse Data Mining reconoce los campos que no tienen un valor explicativo. Para determinar si un campo tiene valor explicativo, el algoritmo Regresión lineal realiza comprobaciones de estadísticos además de la selección de variables autonómica. Si conoce los campos que no tienen valor explicativo, puede seleccionar de forma automática un subconjunto de campos explicativos para tiempos de ejecución más breves.

El algoritmo Regresión lineal proporciona los siguientes métodos de selección automática de subconjuntos de campos explicativos:

Regresión por pasos. Para la regresión paso a paso, debe especificar un nivel de significancia mínimo. El algoritmo Regresión lineal utiliza solamente los campos con un nivel de significación superior al valor especificado.

Regresión R cuadrado. El método de regresión R cuadrado identifica un modelo óptimo mediante la optimización de una medida de calidad del modelo. Se emplea una de las siguientes medidas de calidad:

- Coeficiente de correlación de Pearson cuadrado.
- El coeficiente de correlación de Pearson cuadrado ajustado.

De forma predeterminada, el algoritmo Regresión lineal selecciona de forma automática subconjuntos de campos explicativos mediante el coeficiente de correlación de Pearson cuadrado ajustado para optimizar la calidad del modelo.

Regresión polinómica

El algoritmo de regresión polinómica de ISW asume una relación polinómica. Un modelo de regresión polinómica es una ecuación formada por las siguientes partes:

- El grado máximo de regresión polinómica
- Una aproximación del campo objetivo
- Los campos explicativos.

Regresión de RBF

El algoritmo de regresión de RBF de ISW asume una relación entre los campos explicativos y el campo objetivo. Esta relación se puede expresar como una combinación lineal de funciones de Gaussian. Las funciones de Gauss son funciones de base radial específicas.

Opciones de modelos de regresión de ISW

En la pestaña Modelo del nodo Regresión de ISW, puede especificar el tipo de algoritmo de regresión que desea utilizar, así como:

- Si utilizar o no datos divididos
- Si realizar o no una ejecución de comprobación
- Un límite para el valor de R^2
- Un límite para el tiempo de procesamiento

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo.

Método de regresión. Seleccione el tipo de regresión que desee ejecutar. Para obtener más información, consulte el tema "Regresión de ISW" en la página 67.

Realizar ejecución de comprobación. Puede elegir realizar una ejecución de comprobación. Se realizará una ejecución de comprobación de InfoSphere Warehouse Data Mining una vez generado el modelo en la partición de entrenamiento. De este modo se realizará una pasada por la partición de comprobación para establecer información de calidad del modelo, gráficos de elevación, etc.

Limitar R cuadrado. Esta opción especifica el error sistemático tolerado máximo (el coeficiente de correlación de Pearson cuadrado, R^2). Este coeficiente mide la correlación entre el error de predicción de los datos de verificación y los valores reales del objetivo. Tiene un valor entre 0 (sin correlación) y 1 (positiva perfecta o correlación negativa). El valor que defina define el límite superior del error sistemático aceptable del modelo.

Limitar tiempo de procesamiento. Especifique el tiempo de procesamiento máximo que desee en minutos.

Opciones de Experto de Regresión de ISW

En la pestaña Experto del nodo Regresión de ISW, puede especificar un número de opciones avanzadas de una regresión lineal, polinomial o de RBF.

Opciones de Experto para regresión lineal o polinomial

Limitar grado de polinomio. Establece el grado máximo de regresión polinómica. Si se establece el grado máximo de regresión polinómica en **1**, el algoritmo Regresión polinómica es idéntico al algoritmo Regresión lineal. Si se especifica un valor alto para el grado máximo de regresión polinómica, el algoritmo Regresión polinómica tiende a sobreajustarse. Esto significa que el modelo resultante se aproxima de manera precisa a los datos de entrenamiento; no obstante, falla cuando se aplica a datos no utilizados para el entrenamiento.

Usar interceptación. Cuando está activado, fuerza a la curva de regresión a pasar por el origen. Esto significa que el modelo no contendrá un término constante.

Utilización de selección de características automática. Cuando se activa, el algoritmo trata de determinar un subconjunto óptimo de predictores posibles di no especifica un nivel de significación mínimo.

Usar nivel de significación mínima. Cuando se especifica un nivel de significación mínimo, la regresión por pasos se utiliza para determinar un subconjunto de posible predictores. Sólo los campos independientes cuya significación es superior al valor especificado contribuyen con el cálculo del modelo de regresión.

Configuración de campos. Para especificar las opciones de los campos de entrada, pulse en la fila correspondiente de la columna Configuración de la tabla Configuración de campos y seleccione **<Especificar configuración>**. Para obtener más información, consulte el tema "Especificación de configuración de campos de regresión" en la página 70.

Opciones de Experto para Regresión de RBF

Utilizar tamaño de muestra de entrada. Define una muestra 1-de cada-N para la verificación y comprobación de modelo.

Utilizar tamaño de muestra de salida. Define una muestra 1-de cada-N para el entrenamiento.

Utilizar número máximo de centros. El número máximo de centros que se generan en cada paso. Como el número de centros puede aumentar hasta el doble del número inicial durante un paso, el número real de centros puede ser superior al número que especifique.

Utilizar tamaño mínimo de región. El número mínimo de registros que se asigna a una región.

Utilizar máximo de pasos de datos. El número máximo de pasos a través de los datos de entrada realizado por el algoritmo. Si se especifica, este valor debe ser mayor o igual que el número mínimo de pasos.

Utilizar mínimo de pasos de datos. El número mínimo de pasos a través de los datos de entrada realizado por el algoritmo. Especifique un valor elevado solamente si tiene suficientes datos de entrenamiento y si está seguro de que existe un buen modelo.

Especificación de configuración de campos de regresión

En el cuadro de diálogo Editar valores de regresión puede especificar el rango de valores de un campo de entrada individual para la regresión lineal o polinómica.

Valor MIN. El valor válido mínimo de este campo de entrada.

Valor MAX. El valor válido máximo de este campo de entrada.

Agrupación en clústeres de ISW

La función de minería de Agrupación busca en los datos de entrada las características que aparecen juntas con más frecuencia. Agrupa los datos de entrada en agrupaciones. Los miembros de cada agrupación tienen propiedades similares. No existe ninguna noción preconcebida respecto a la existencia de patrones dentro de los datos. El clúster es un proceso de descubrimiento.

El nodo Clúster de ISW permite elegir los siguientes métodos de agrupación en clústeres:

- Demográfico
- Kohonen
- BIRCH (reducción y clúster iterativo equilibrado mediante jerarquías) mejorado

La técnica de algoritmo de **agrupación en clústeres demográfica** se basa en la distribución. La agrupación en clústeres basada en distribución proporcionan un método rápido y natural para agrupar en clústeres bases de datos de gran tamaño. El número de clústeres se selecciona automáticamente (puede especificar el número máximo de clústeres). Hay un gran número de parámetros que puede configurar el usuario.

La técnica de algoritmo de **agrupación en clústeres de Kohonen** se basa en el centro. La correlación de funciones Kohonen intenta colocar los centros de la agrupación en lugares que minimicen la distancia global entre los registros y los centros de agrupación. La separación entre agrupaciones no se tiene en cuenta. Los vectores de centro se organizan en una correlación con un determinado número de columnas y de filas. Estos vectores están interconectados de forma que no sólo se ajusta el vector ganador más cercano al registro de entrenamiento, sino también los vectores que están en la proximidad. Sin embargo, cuanto más lejos estén los otros centros, menos se ajustan.

La técnica de algoritmo de **agrupación en clústeres de BIRCH** mejorada se basa en la distribución y trata de minimizar la distancia total entre los registros y sus clústeres. La distancia del logaritmo de la

verosimilitud se utiliza de forma predeterminada para determinar la distancia entre un registro y un clúster; asimismo, puede seleccionar Distancia euclídea si todos los campos activos son numéricos. El algoritmo de BIRCH realiza dos pasos independientes: primero, ordena los registros de entrada en un árbol de característica de clúster de modo que los registros similares pasan a formar parte de los nodos del mismo árbol; a continuación, agrupa las hojas de este árbol en la memoria para generar el resultado de agrupación en clústeres definitivo.

Opciones de modelos de clúster de ISW

En la pestaña Modelo del nodo Clúster, puede especificar el método que utilizará para crear los clústeres, junto con algunas opciones relacionadas.

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo.

Método de clúster. Elija el método que desea utilizar para crear los clústeres: **Demográfico**, **Kohonen**, o **BIRCH mejorado**. Para obtener más información, consulte el tema "Agrupación en clústeres de ISW" en la página 70.

Limitar número de clústeres. Limitando el número de clústeres se ahorra tiempo de procesamiento al evitar la producción de numerosos clústeres pequeños.

Número de filas/Número de columnas. (Método Kohonen únicamente) Especifica el número de filas y las columnas del mapa de características de Kohonen. (Solamente está disponible si se selecciona **Limitar número de pasadas de kohonen** y si **Limitar número de clústeres** no está seleccionado.)

Limitar número de pasadas de Kohonen. (Método de Kohonen únicamente) Especifica el número de pasadas sobre los datos del algoritmo de agrupación en clústeres durante las ejecuciones del entrenamiento. Con cada pase, los vectores del centro se ajustan a fin de minimizar la distancia total entre centros de agrupación y registros. Además, se reduce la cantidad en la que se ajustan los vectores. En el primer pase, los ajustes se realizan grosso modo. En el pase final, la cantidad en la que se ajustan los centros es bastante pequeña. Solamente se realizan ajustes pequeños.

Medida de distancia. (Método de BIRCH mejorado únicamente) Seleccione la medida de distancia desde el registro hasta el clúster utilizada por el algoritmo de BIRCH. Puede seleccionar la distancia del logaritmo de la verosimilitud, que es el valor predeterminado, o la distancia euclídea. *Nota*: solamente puede seleccionar la distancia euclídea cuando todos los campos activos sean numéricos.

Número máximo de nodos hoja. (Método de BIRCH mejorado únicamente) Número máximo de nodos hoja que quiere que tenga el árbol de características de clústeres. El árbol de características de clústeres es el resultado del primer paso en el algoritmo BIRCH mejorado, donde los registros de datos se organizan en un árbol de modo que los registros similares pertenecen al mismo nodo hoja. El tiempo de procesamiento del algoritmo aumenta con el número de nodos hoja. El valor predeterminado es 1000.

Pasadas de BIRCH. (Método de BIRCH mejorado únicamente) Número de pasadas del algoritmo sobre los datos para ajustar el resultado de la agrupación en clústeres. El número de pasadas afecta al tiempo de procesamiento de ejecuciones de entrenamiento (ya que cada pasada requiere una exploración completa de los datos) y a la calidad del modelo. Los valores mínimos dan como resultado un tiempo de procesamiento corto; sin embargo, pueden dar como resultado modelos de menor calidad. Los valores altos aumentan el tiempo de proceso y suelen generar modelos mejores. Por lo general, 3 o más pases dan buenos resultados. El valor predeterminado es 3.

Opciones de Experto para los clústeres de ISW

En la pestaña Experto del nodo Agrupación en clústeres, puede especificar opciones avanzadas, como umbrales de similitudes, límites del tiempo de ejecución y ponderaciones de campos.

Limitar tiempo de procesamiento. Seleccione esta casilla para activar las opciones que permiten controlar el tiempo necesario para crear el modelo. Puede especificar un tiempo en minutos, un porcentaje mínimo de datos de entrenamiento que deben procesarse o ambos. Además, para el método de BIRCH, puede especificar el número máximo de nodos hoja que se crearán en el árbol CF.

Especificar umbral de similitud. (Agrupación en clústeres demográfica únicamente) El límite inferior de similitud de los dos registros de datos que pertenecen al mismo clúster. Por ejemplo, un valor de 0,25 significa que los registros con valores con una similitud del 25% se pueden asignar al mismo clúster. Un valor de 1.0 significa que los registros deben ser idénticos para que aparezcan en el mismo clúster.

Configuración de campos. Para especificar las opciones de los campos de entrada, pulse en la fila correspondiente de la columna Configuración de la tabla Configuración de campos y seleccione **<Especificar configuración>**.

Especificación de configuración de campos de clúster

En el cuadro de diálogo Editar configuración de clúster puede especificar opciones para campos de entrada individuales.

Ponderación de campo. Asigne un valor mayor o menor de ponderación al campo durante el proceso de generación del modelo. Por ejemplo, si cree que este campo es relativamente menos importante para el modelo que otros campos, reduzca su ponderación de campo relativa al resto de campos.

Ponderación de valor. Asigna un valor mayor o menor de ponderación a valores concretos de este campo. Algunos valores de campos pueden ser más comunes que otros valores. La coincidencia de valores inusuales en un campo puede ser más significativa para un clúster que la coincidencia de valores frecuentes. Puede seleccionar uno de los métodos siguientes para ponderar los valores de este campo (en cada caso, los valores inusuales tienen un valor grande de ponderación, mientras que los valores comunes tienen un valor de ponderación menor):

- **Logarítmico.** Asigna una ponderación a cada valor en función del logaritmo de su probabilidad en los datos de entrada.
- **Probabilístico.** Asigna una ponderación a cada valor en función de su probabilidad en los datos de entrada.

En cada método también puede seleccionar una opción **con compensación** para compensar el valor de ponderación aplicado a cada campo. Si realiza la compensación de la ponderación de valores, la importancia general del campo ponderado es igual a la de un campo no ponderado. Esto se cumple cualquiera que sea el número de valores posibles. La ponderación compensada afecta únicamente a la importancia relativa de las coincidencias en el conjunto de valores posibles.

Utilizar escala de similitudes. Seleccione esta opción si desea utilizar una escala de similitudes para controlar el cálculo de la medición de similitud de un campo. Especifique la escala de similitud como número absoluto. La especificación sólo es efectiva para los campos numéricos activos. Si no especifica una escala de similitudes, se utiliza el valor predeterminado (la mitad de la desviación estándar). Para obtener un mayor número de clústeres, reduzca la similitud media entre pares de clústeres por escalas de similitudes menores de campos numéricos.

Tratamiento de valores atípicos. Los valores atípicos son valores de campo que caen fuera del rango de valores especificados para el campo, tal y como definen los valores **MIN** y **MAX**. Puede seleccionar cómo desea que los valores atípicos se gestionen en este campo.

- El valor predeterminado, **ninguno**, significa que los valores atípicos no realizan ninguna acción.

- Si selecciona **reemplazar con MIN o MAX**, un valor de campo menor que el **valor MIN** o mayor que el **valor MAX** se sustituye por los valores de MIN o MAX que procedan. Puede definir los valores de MIN y MAX en este caso.
- Si selecciona **tratar como perdido**, los valores atípicos se consideran como valores perdidos y se descartan. Puede definir los valores de MIN y MAX en este caso.

Bayesiano ingenuo ISW

Bayesiano ingenuo es un algoritmo muy utilizado para resolver problemas de clasificación. El modelo se denomina *ingenuo* porque trata todas las variables de predicción propuestas como independientes unas de otras. El bayesiano ingenuo es un algoritmo rápido y escalable que calcula las probabilidades condicionales para las combinaciones de atributos y el atributo de objetivo. A partir de los datos de entrenamiento se establece una probabilidad independiente. Esta probabilidad proporciona la verosimilitud de cada clase objetivo, una vez dada la instancia de cada categoría de valor a partir de cada variable de entrada.

El algoritmo de clasificación bayesiano ingenuo de ISW es un clasificador probabilístico. Se basa en modelos de probabilidad que incorporan fuertes supuestos de independencia.

Opciones del modelo bayesiano ingenuo de ISW

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo.

Realizar ejecución de comprobación. Puede elegir realizar una ejecución de comprobación. Se realizará una ejecución de comprobación de IBM InfoSphere Warehouse Data Mining una vez generado el modelo en la partición de entrenamiento. De este modo se realizará una pasada por la partición de comprobación para establecer información de calidad del modelo, gráficos de elevación, etc.

Umbral de probabilidad. El umbral de probabilidad define una probabilidad para cualquier combinación de predictor y valores de objetivo que no se ven en los datos de entrenamiento. Esta probabilidad debería estar entre 0 y 1. El valor predeterminado es 0.001.

Regresión logística ISW

La regresión logística, también denominada regresión nominal, es una técnica estadística para clasificar los registros a partir de los valores de los campos de entrada. Es equivalente a la regresión lineal, pero el algoritmo de regresión logística de ISW toma un campo objetivo de marcas (binario) en lugar de uno numérico.

Opciones del modelo de regresión logística de ISW

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilizarán los datos de la partición de entrenamiento para la generación del modelo.

Realizar ejecución de comprobación. Puede elegir realizar una ejecución de comprobación. Se realizará una ejecución de comprobación de IBM InfoSphere Warehouse Data Mining una vez generado el modelo en la partición de entrenamiento. De este modo se realizará una pasada por la partición de comprobación para establecer información de calidad del modelo, gráficos de elevación, etc.

Serie temporal ISW

El algoritmo de series temporales de ISW le permite predecir eventos futuros basándose en eventos conocidos del pasado.

De forma parecida a los métodos de regresión comunes, los algoritmos de serie pronostican un valor numérico. A diferencia de los métodos de regresión comunes, las predicciones de serie temporal se centran en los valores de una serie ordenada en el futuro. Estas predicciones se suelen denominar previsiones.

Los algoritmos de serie temporal son algoritmos univariables. Esto significa que la variable independiente es una columna de tiempo o una columna de orden. Las previsiones se basan en los valores del pasado. No se basan en otras columnas independientes.

Los algoritmos de series temporales son diferentes de los algoritmos de regresión comunes porque no solamente predicen valores futuros sino que también incorporan ciclos estacionales en la previsión.

La función de minería Serie temporal proporciona los siguientes algoritmos para predecir tendencias futuras:

- Modelo Autorregresivo Integrado de Media Móvil (ARIMA)
- Ajuste exponencial
- Descomposición de tendencia estacional

El algoritmo que crea la mejor previsión de sus datos depende de diferentes supuestos de modelo. Puede calcular todas las previsiones al mismo tiempo. Los algoritmos calculan una previsión detallada, incluido el comportamiento estacional de la serie temporal original. Si tiene el cliente de IBM InfoSphere Warehouse instalado, puede utilizar el visualizador de series temporales para evaluar y comparar las curvas resultantes.

Opciones de los campos de Serie temporal de ISW

Hora. Seleccione el campo de entrada que contiene la serie temporal. Debe ser un campo con un tipo de almacenamiento de fecha, hora, marca de tiempo, real o entero.

Utilizar configuración del nodo Tipo. Esta opción permite indicar al nodo que use la información de campo de un nodo Tipo situado en un punto anterior de la ruta. Éste es el valor por omisión.

Utilizar configuración personalizada. Esta opción permite indicar al nodo que use la información de campo especificada aquí en lugar de la proporcionada en nodos Tipo situados en cualquier punto anterior de la ruta. Después de seleccionar esta opción, especifique los campos siguientes si es necesario.

Objetivos. Seleccione uno o varios campos objetivo. Se trata de una acción similar a establecer el rol del campo en *Objetivo* en un nodo Tipo.

Opciones del modelo de series temporales de ISW

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Algoritmos de previsión. Seleccione los algoritmos que se utilizarán para el modelado. Puede elegir uno de los siguientes o una mezcla de ellos:

- ARIMA
- Suavizado exponencial
- Descomposición de tendencias estacional.

Tiempo de finalización de previsión. Especifique si el tiempo de finalización de previsión se calculará automáticamente o se especificará manualmente.

Valor de campo de tiempo. Cuando **Tiempo de finalización de previsión** se establezca como manual, introduzca el tiempo de finalización de previsión. El valor que puede introducir depende del tipo de campo de tiempo; por ejemplo, si el tipo es un entero que representa las horas, puede introducir 48 para detener la previsión después de procesar los datos de 48 horas. Además, este campo puede pedirle que introduzca una fecha u hora como valor de finalización.

Opciones de Experto de Serie temporal de ISW

Utilizar todos los registros para generar el modelo. Este es el ajuste predeterminado; todos los registros se analizan cuando se genera el modelo.

Utilizar un subconjunto de los registros para generar el modelo. Si solamente quiere crear el modelo a partir de una parte de los datos disponibles, seleccione esta opción. Por ejemplo, esto puede ser necesario cuando tenga una cantidad excesiva de datos repetitivos.

Introduzca **Valor de hora de inicio** y **Valor de hora de finalización** para identificar los datos que deben utilizarse. Tenga en cuenta que los valores que puede introducir en estos campos dependen del tipo de campo de tiempo; por ejemplo, puede ser un número de horas o días, o bien fechas u horas específicas.

Método de interpolación para valores objetivo perdidos. Si está procesando datos con uno o más valores perdidos, seleccione el método que debe utilizarse para calcularlos. Puede elegir uno de los siguientes:

- Aritmético
- LíneasSp exponenciales
- LíneaSp cúbicas

Visualización de modelos de series temporales de ISW

Los modelos de series temporales de ISW se obtienen bajo la forma de un modelo sin refinar, que contiene información extraída de los datos pero que no está diseñado para generar predicciones directamente.

Figura 3. Icono de modelo sin refinar

Si tiene el cliente de IBM InfoSphere Warehouse instalado, puede utilizar la herramienta de visualización de series temporales para obtener una representación gráfica de sus datos de series temporales.

Para utilizar la herramienta de visualización de series temporales:

1. Asegúrese de que ha completado las tareas de integración de IBM SPSS Modeler con IBM InfoSphere Warehouse. Para obtener más información, consulte el tema "Activación de la integración con IBM InfoSphere Warehouse" en la página 53.
2. Pulse dos veces en el icono del modelo sin refinar en la paleta de modelos.
3. En la pestaña Servidor del cuadro de diálogo, pulse en el botón Ver para mostrar el visualizador en su explorador Web predeterminado.

Nuggets de modelos de ISW Data Mining

Puede crear modelos desde los nodos de ISW Árbol de decisión, Asociación, Secuencia, Regresión y Agrupación en clústeres, que se incluyen con IBM SPSS Modeler.

Pestaña Servidor del nugget de modelo de ISW

La pestaña Servidor proporciona opciones para realizar comprobaciones de coherencia y abrir la herramienta de visualización de IBM.

IBM SPSS Modeler puede realizar la comprobación de coherencia almacenando una cadena clave del modelo generada idéntica, tanto en el modelo de IBM SPSS Modeler como en el de ISW. La comprobación de coherencia se realiza pulsando en el botón **Comprobar** de la pestaña Servidor. Para obtener más información, consulte el tema "Gestión de modelos DB2" en la página 58.

La herramienta Visualizer es el único método para explorar los modelos de InfoSphere Warehouse Data Mining. La herramienta puede instalarse opcionalmente con InfoSphere Warehouse Data Mining. Para obtener más información, consulte el tema "Activación de la integración con IBM InfoSphere Warehouse" en la página 53.

- Pulse en **Ver** para ejecutar la herramienta de visualización. Lo que muestra la herramienta depende del tipo de nodo generado. Por ejemplo, la herramienta de visualización mostrará una vista de clases predichas cuando se inicie desde un nugget de modelo Árbol de decisión de ISW.
- Pulse en **Resultados de prueba** (sólo árboles de decisión y secuencia) para ejecutar la herramienta de visualización y ver la calidad general del modelo generado.

Pestaña Configuración del nugget de modelo de ISW

En IBM SPSS Modeler, generalmente, sólo se proporciona una única predicción y una confianza o probabilidad asociada. Además, una opción de usuario para mostrar las probabilidades de cada resultado (similar a la de la regresión logística) es una opción de puntuación de tiempo disponible en el nugget de modelo de la pestaña Configuración.

Incluir confianzas para todas las clases. Para cada uno de los posibles resultados del campo objetivo, añade una columna ofreciendo el nivel de confianza.

Pestaña Resumen del nugget de modelo de ISW

La pestaña Resumen de un nugget de modelo muestra información sobre el propio modelo (*Análisis*), los campos utilizados en el modelo (*Campos*), la configuración utilizada al generar el modelo (*Configuración de creación*) y el entrenamiento del modelo (*Resumen de entrenamiento*).

Cuando se examina el nodo por primera vez, los resultados de la pestaña Resumen aparecen contraídos. Para ver los resultados de interés, utilice el control de expansión situado a la izquierda de un elemento con objeto de desplegarlo, o bien pulse en el botón **Expandir todo** para mostrar todos los resultados. Para ocultar los resultados cuando haya terminado de consultarlos, utilice el control de expansión con objeto de contraer los resultados específicos que desee ocultar o pulse en el botón **Contraer todo** para contraer todos los resultados.

Análisis. Muestra información sobre el modelo específico. Si ha ejecutado un nodo Análisis conectado a este nugget de modelo, la información de dicho análisis también se mostrará en esta sección.

Campos. Enumera los campos utilizados como objetivo y entradas en la generación del modelo.

Configuración de creación. Contiene información sobre la configuración que se utiliza en la generación del modelo.

Resumen de entrenamiento. Muestra el tipo del modelo, la ruta utilizada para crearlo, el usuario que lo creó, cuándo se generó y el tiempo que se tardó en generar el modelo.

Ejemplos de ISW Data Mining

IBM SPSS Modeler para Windows incluye varias rutas de demostración que ilustran el proceso de minería de bases de datos. Estas rutas se encuentran en la carpeta de instalación de IBM SPSS Modeler en:

\Demos\Database_Modeling\IBM DB2 ISW

Nota: se puede acceder a la carpeta Demos desde el grupo de programas IBM SPSS Modeler en el menú Inicio de Windows.

Las siguientes rutas se pueden utilizar juntas en secuencia como ejemplo del proceso de minería de bases de datos:

- *1_upload_data.str*: se utiliza para depurar y cargar datos desde un archivo sin formato a DB2.
- *2_explore_data.str*: se utiliza como un ejemplo de exploración de los datos con IBM SPSS Modeler.
- *3_build_model.str*: se utiliza para generar un modelo de árbol de decisión de ISW.
- *4_evaluate_model.str*: se utiliza como un ejemplo de evaluación del modelo con IBM SPSS Modeler.
- *5_deploy_model.str*: se utiliza para desplegar el modelo para la puntuación interna de la base de datos.

El conjunto de datos utilizado en las rutas de ejemplo está relacionado con aplicaciones de tarjetas de crédito y presenta un problema de clasificación con una mezcla de predictores continuos y categóricos. Si desea obtener más información acerca de este conjunto de datos, consulte el siguiente archivo de la carpeta de instalación de IBM SPSS Modeler en:

\Demos\Database_Modeling\IBM DB2 ISW\crx.names

Este conjunto de datos está disponible desde UCI Machine Learning Repository en http://archive.ics.uci.edu/ml/.

Ruta de ejemplo: cargar datos

La primera ruta de ejemplo, *1_upload_data.str*, se utiliza para depurar y cargar los datos desde un archivo sin formato a DB2.

El nodo Rellenar se utiliza para gestionar los valores perdidos y sustituye los campos vacíos de lectura del archivo de texto *crx.data* por valores *NULOS*.

Ruta de ejemplo: explorar datos

La segunda ruta de ejemplo, *2_explore_data.str*, se utiliza para demostrar la exploración de datos en IBM SPSS Modeler.

Un paso habitual en la exploración de datos es adjuntar un nodo Auditoría de datos a los datos. El nodo Auditoría de datos está disponible desde la paleta de nodos de resultado.

Puede utilizar los resultados del nodo Auditoría de datos para conocer los conceptos básicos de la distribución de los datos y los campos. Si pulsa dos veces en un gráfico en la ventana Auditoría de datos, aparecerá un gráfico más detallado donde podrá explorar un campo específico con más profundidad.

Ruta de ejemplo: generar modelo

La tercera ruta de ejemplo, *3_build_model.str*, ilustra la generación del modelo en IBM SPSS Modeler. Puede adjuntar el nodo de modelado de bases de datos a la ruta y pulsar dos veces en él para especificar la configuración de generación.

Mediante las pestañas Modelo y Experto del nodo de modelado, puede ajustar la profundidad máxima del árbol y detener la posterior división de un nodo desde que se genera el árbol de decisión inicial configurando la pureza máxima y el número mínimo de casos por nodo interno. Para obtener más información, consulte el tema "Árbol de decisión de ISW" en la página 62.

Ruta de ejemplo: evaluar modelo

La cuarta ruta de ejemplo, *4_evaluate_model.str*, ilustra las ventajas de utilizar IBM SPSS Modeler para el modelado interno de bases de datos. Una vez ejecutado el modelo, puede volver a añadirlo a la ruta de datos y evaluarlo con varias herramientas que se ofrecen en IBM SPSS Modeler.

Cuando abra la ruta por primera vez, el nugget de modelo (*field16*) no se incluye en la ruta. Abra el nodo de origen CREDIT y compruebe que ha especificado un origen de datos. A continuación, si ha ejecutado la ruta *3_build_model.str* para crear un nugget *field16* en la paleta Modelos, puede ejecutar los nodos desconectados pulsando el botón **Run** de la barra de herramientas (el botón con un triángulo verde). Se ejecuta un script que copia el nugget *field16* en la ruta, lo conecta a los nodos existentes y ejecuta los nodos del terminal en la ruta.

Se puede adjuntar un nodo Análisis (disponible en la paleta Resultado) para crear una matriz de coincidencias que muestre el patrón de coincidencias entre cada campo generado (predicho) y su campo objetivo. Ejecute el nodo Análisis para ver los resultados.

Asimismo, puede crear un gráfico de ganancias para mostrar las mejoras de precisión realizadas por el modelo. Adjunte un nodo Evaluación al modelo generado y, a continuación, ejecute la ruta para ver los resultados.

Ruta de ejemplo: desplegar modelo

Una vez satisfecho con la precisión del modelo, puede desplegarlo para su uso con aplicaciones externas o para volver a escribir las puntuaciones en la base de datos. En la ruta de ejemplo *5_deploy_model.str*, los datos se leen desde la tabla `CREDIT`. Si se ejecuta el nodo de exportación de bases de datos *desplegar solución*, los datos no se puntúan realmente. En su lugar, la ruta crea el archivo de imagen publicado *credit_scorer.pim* y el archivo de parámetros publicado *credit_scorer.par*.

Como en el ejemplo anterior, se ejecuta un script que copia el nugget *field16* en la ruta de la paleta Modelos, lo conecta a los nodos existentes y ejecuta los nodos del terminal en la ruta. En este caso debe especificar primero un origen de datos tanto en el origen de la base de datos como en los nodos de exportación.

Capítulo 6. Modelado de bases de datos con IBM Netezza Analytics

IBM SPSS Modeler y IBM Netezza Analytics

IBM SPSS Modeler admite la integración con IBM Netezza Analytics, que proporciona la capacidad de ejecutar análisis avanzados en servidores IBM Netezza. Es posible acceder a estas características mediante la interfaz gráfica de usuario de IBM SPSS Modeler y el entorno de desarrollo orientado al flujo de trabajo, lo que le permite ejecutar los algoritmos de minería de datos directamente en el entorno de IBM Netezza.

IBM SPSS Modeler admite la integración de los siguientes algoritmos de IBM Netezza Analytics.

- Árboles de decisión
- K-medias
- Red bayesiana
- bayesiano ingenuo
- KNN
- Clúster divisivo
- PCA
- Árbol de regresión
- Regresión lineal
- Serie temporal
- Lineal generalizado

Para obtener más información sobre los algoritmos, consulte el manual *IBM Netezza Analytics Developer's Guide* y el manual *IBM Netezza Analytics Reference Guide*.

Requisitos para la integración con IBM Netezza Analytics

Las siguientes condiciones constituyen los requisitos previos para realizar el modelado interno de bases de datos mediante IBM Netezza Analytics. Es posible que necesite consultar con el administrador de la base de datos para asegurarse de que se reúnen las condiciones.

- IBM SPSS Modeler ejecutado en modo local o en una instalación de IBM SPSS Modeler Server Windows o UNIX (excepto zLinux, para el que los controladores ODBC de IBM Netezza no están disponibles).
- IBM Netezza Performance Server 6.0 o posterior, ejecutando el paquete de IBM SPSS In-Database Analytics.
- Un origen de datos ODBC para conectar a una base de datos IBM Netezza. Para obtener más información, consulte el tema "Activación de la integración con IBM Netezza Analytics" en la página 80.
- Optimización y generación de SQL activada en IBM SPSS Modeler. Para obtener más información, consulte el tema "Activación de la integración con IBM Netezza Analytics" en la página 80.

Nota: optimización de SQL y modelado de bases de datos requieren que la conectividad de IBM SPSS Modeler Server esté activada en el equipo con IBM SPSS Modeler. Con esta configuración activada, puede acceder a los algoritmos de bases de datos, devolver SQL directamente desde IBM SPSS Modeler y acceder a IBM SPSS Modeler Server. Para verificar el estado de la licencia actual, seleccione las siguientes opciones en el menú de IBM SPSS Modeler.

Ayuda > **Acerca de** > **Detalles adicionales**

Si la conectividad está activada, verá la opción **Activación de servidor** en la pestaña Estado de licencia.

Activación de la integración con IBM Netezza Analytics

La activación de la integración con IBM Netezza Analytics se compone de los pasos siguientes.

- Configuración de IBM Netezza Analytics
- Creación de un origen ODBC
- Activación de la integración en IBM SPSS Modeler
- Activación de optimización y generación de SQL en IBM SPSS Modeler

Se describen en las secciones siguientes.

Configuración de IBM Netezza Analytics

Para instalar y configurar IBM Netezza Analytics, consulte la documentación de IBM Netezza Analytics, en particular la *Guía de instalación de IBM Netezza Analytics*, para obtener más detalles. La sección *Configuración de los permisos de la base de datos* en esa guía contiene detalles de scripts que se deben ejecutar para permitir a las IBM SPSS Modeler rutas escribir en la base de datos.

Nota: si va a utilizar nodos basados en cálculos de matriz (PCA de Netezza y regresión lineal de Netezza), el motor de matriz de Netezza debe inicializarse ejecutando `CALL NZM..INITIALIZE();` de lo contrario, la ejecución de los procedimientos almacenados fallará. La inicialización es un paso único de configuración en cada base de datos.

Creación de un origen ODBC para IBM Netezza Analytics

Para activar la conexión entre la base de datos de IBM Netezza y IBM SPSS Modeler, debe crear un nombre de origen de datos del sistema ODBC (DSN).

Antes de crear un DSN, debe tener un conocimiento básico de las unidades y los orígenes de datos ODBC y el soporte de la base de datos en IBM SPSS Modeler.

Si se está ejecutando en modo distribuido con respecto al servidor IBM SPSS Modeler Server, cree el DSN en el equipo servidor. Si se está ejecutando en modo local (cliente), cree el DSN en el equipo cliente.

Clientes de Windows

1. Desde su CD *Netezza Client*, ejecute el archivo *nzodbcsetup.exe* para iniciar el instalador. Siga las instrucciones que aparecen en pantalla para instalar el controlador. Para obtener todas las instrucciones, consulte la *Guía de IBM Netezza de configuración e instalación de ODBC, JDBC, y OLE DB*.
   a. Cree el DSN.

      Nota: la secuencia de menú depende de la versión de Windows.

      - **Windows XP.** En el menú Inicio, seleccione **Panel de control**. Pulse dos veces en **Herramientas administrativas** y, a continuación, pulse dos veces en **Orígenes de datos (ODBC)**.
      - **Windows Vista.** En el menú Inicio, seleccione **Panel de control** y, a continuación, **Sistema y mantenimiento**. Pulse dos veces en **Herramientas administrativas**, seleccione **Orígenes de datos (ODBC)** y, a continuación, pulse **Abrir**.
      - **Windows 7.** En el menú Inicio, seleccione **Panel de control**, luego **Sistema y seguridad** y, a continuación, **Herramientas administrativas**. Seleccione **Orígenes de datos (ODBC)** y, a continuación, pulse en **Abrir**.

   b. Pulse en la pestaña **DSN de sistema** y, a continuación, **Añadir**.
2. Seleccione **NetezzaSQL** en la lista y pulse en **Finalizar**.

3. En la pestaña **Opciones de DSN** en la pantalla de configuración del controlador ODBC Netezza, escriba un nombre de origen de datos de su elección, el nombre del host o dirección IP del servidor de IBM Netezza, el número de puerto para la conexión, la base de datos de la instancia de IBM Netezza que está utilizando y su nombre de usuario y contraseña para la conexión a la base de datos. Pulse en el botón **Ayuda** para obtener una explicación de los campos.
4. Pulse en el botón **Probar conexión** y asegúrese de que puede conectarse a la base de datos.
5. Si obtiene una conexión, pulse en **ACEPTAR** repetidamente para salir de la pantalla del administrador de origen de datos ODBC.

Servidores de Windows

El procedimiento para los servidores de Windows es el mismo que el procedimiento de clientes de Windows XP.

Servidores UNIX o Linux

El siguiente procedimiento es aplicable a servidores UNIX o Linux (excepto zLinux, para el que los controladores de IBM Netezza ODBC no están disponibles).

1. En su CD de *Netezza Client*, copie el archivo *<plataforma>cli.package.tar.gz* correspondiente en una ubicación temporal del servidor.
2. Extraiga el contenido del archivo mediante los comandos `gunzip` y `untar`.
3. Añada los permisos de ejecución al script *unpack* que se ha extraído.
4. Ejecute el script, respondiendo a las preguntas que aparecen en pantalla.
5. Edite el archivo *modelersrv.sh* e incluya las siguientes líneas.

```
. /usr/IBM/SPSS/SDAP61_notfinal/odbc.sh
LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
NZ_ODBC_INI_PATH=/usr/IBM/SPSS/SDAP61_notfinal; export NZ_ODBC_INI_PATH
```

6. Busque el archivo */usr/local/nz/lib64/odbc.ini* y copie su contenido en el archivo *odbc.ini* que se ha instalado con SDAP 6.1 (el definido por la variable de entorno $ODBCINI).

   Nota: en sistemas Linux de 64 bits, el parámetro *Driver* hace referencia incorrectamente al controlador de 32 bits. Si copia el contenido de *odbc.ini* en el paso anterior, edite la ruta de este parámetro, por ejemplo:

   ```
   /usr/local/nz/lib64/libnzodbc.so
   ```

7. Edite los parámetros en la definición Netezza DSN para que refleje la base de datos que se utilizará.
8. Reinicie IBM SPSS Modeler Server y compruebe el uso de los nodos de minería interna de bases de datos de Netezza en el cliente.

Activación de la integración de IBM Netezza Analytics en IBM SPSS Modeler

1. En el menú principal de IBM SPSS Modeler, elija:

   Herramientas > **Opciones** > **Aplicaciones de ayuda.**
2. Pulse en la pestaña **IBM Netezza**.

Habilitar integración de Netezza Data Mining. Habilita la paleta de modelado de bases de datos (si todavía no se muestra) en la parte inferior de la ventana de IBM SPSS Modeler y añade los nodos para algoritmos de Netezza Data Mining.

Conexión de Netezza. Pulse en el botón **Editar** y seleccione la cadena de conexión de Netezza que configuró anteriormente al crear el origen ODBC. Para obtener más información, consulte el tema "Creación de un origen ODBC para IBM Netezza Analytics" en la página 80.

Activación de optimización y generación de SQL

Debido a la verosimilitud del trabajo con grandes conjuntos de datos, por razones de rendimiento debe activar las opciones de optimización y generación de SQL en IBM SPSS Modeler.

1. Seleccione en los menús de IBM SPSS Modeler:

   Herramientas > **Propiedades de ruta** > **Opciones**
2. Pulse en la opción **Optimización** en el panel de navegación.
3. Confirme que la opción **Generar SQL** está activada. Esta configuración es necesaria para que el modelado de bases de datos funcione.
4. Seleccione las opciones **Optimizar generación de SQL** y **Optimizar otra ejecución** (no obligatorias, pero recomendadas para un rendimiento optimizado).

Generación de modelos con IBM Netezza Analytics

Cada uno de los algoritmos compatibles tiene un nodo de modelado correspondiente. Puede acceder a los nodos de modelado de IBM Netezza desde la pestaña Modelado de bases de datos en la paleta de nodos.

Consideraciones sobre los datos

Los campos en el origen de datos pueden contener variables de varios tipos de datos, dependiendo del nodo de modelado. En IBM SPSS Modeler, los tipos de datos se denominan **niveles de medición**. La pestaña Campos del nodo de modelado utiliza iconos para indicar los tipos de nivel de medición permitidos en sus campos de entrada y de objetivo.

Campo objetivo. El campo objetivo es el campo cuyo valor está intentando predecir. Si se puede especificar un objetivo, solamente se puede seleccionar uno de los campos de datos de origen como el campo objetivo.

Campo de ID de registro. Especifica el campo utilizado para identificar de manera exclusiva cada caso. Por ejemplo, puede tratarse de un campo de ID, como *CustomerID*. Si los datos de origen no incluyen un campo ID, puede crearlo mediante un nodo Derivar, tal y como indica el siguiente procedimiento.

1. Seleccione el nodo de origen.
2. En la pestaña Operaciones con campos de la paleta de nodos, pulse dos veces en el nodo Derivar.
3. Abra el nodo Derivar pulsando dos veces en su icono en el lienzo.
4. En el campo **Nodo Derivar**, introduzca (por ejemplo) `ID`.
5. En el campo **Fórmula**, introduzca `@INDEX` y pulse en **Aceptar**.
6. Conecte el nodo Derivar al resto de la ruta.

Gestión de valores nulos

Si los datos de entrada contienen valores nulos, si utiliza algunos de los nodos de Netezza se pueden producir mensajes de error o rutas repetitivas, por lo que recomendamos eliminar los registros con valores nulos. Utilice el siguiente método.

1. Conecte un nodo Seleccionar al nodo de origen.
2. Defina la opción **Modo** del nodo Seleccionar como **Descartar**.
3. Introduzca lo siguiente en el campo **Condición**:

   ```
   @NULL(campo1)
   [o @NULL(campo2)[... o @NULL(campoN]])
   ```

   Asegúrese de incluir todos los campos de entrada.
4. Conecte el nodo Seleccionar al resto de la ruta.

Resultado de modelo

Es posible que una ruta que contenga un nodo de modelado de Netezza produzca resultados ligeramente diferentes cada vez que se ejecute. Esto se debe a que el orden en el que el nodo lee los datos de origen no es siempre el mismo, ya que los datos se leen en tablas temporales antes de la generación de modelos. Sin embargo, las diferencias producidas por este efecto carecen de significado.

Comentarios generales

- En IBM SPSS Collaboration and Deployment Services, no es posible crear configuraciones de puntuación usando rutas que contengan nodos de modelado de bases de datos de IBM Netezza.
- La exportación o importación de PMML no es posible en modelos creados por los nodos de Netezza.

Modelos de Netezza: Opciones de campo

En la pestaña Campos, puede seleccionar si desea utilizar la configuración de rol de campo ya definida en nodos anteriores o realizar las asignaciones de campos manualmente.

Utilizar roles predefinidos. Esta opción utiliza las definiciones de roles (objetivos, predictores, etcétera) desde un nodo Tipo anterior (o la pestaña Tipo de un nodo de origen anterior).

Utilizar asignaciones de campos personalizadas. Seleccione esta opción si desea asignar objetivos, predictores y otros roles manualmente en esta pantalla.

Campos. Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los diferentes campos de roles en la parte derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol.

Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse en un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

Objetivo. Seleccione un campo como el destino de la predicción. Para modelos lineales generalizados, consulte también el campo **Ensayos** de esta pantalla.

ID de registro. El campo que se utilizará como el identificador de registros exclusivo.

Predictores (Entradas). Seleccione uno o más campos como entradas de la predicción.

Modelos de Netezza: Opciones del servidor

En la pestaña Servidor, especifique la base de datos de IBM Netezza donde se almacenará el modelo.

Detalles del servidor Netezza DB. Aquí puede especificar los detalles de la conexión a la base de datos que desea utilizar para el modelo.

- **Utilice la conexión anterior**. (valor predeterminado) Utiliza los detalles de conexión especificada en un nodo anterior, por ejemplo, el nodo de origen Base de datos. *Nota*: esta opción sólo funciona si todos los nodos anteriores pueden retrotraer operaciones SQL. En este caso no es necesario extraer los datos de la base de datos, ya que SQL lo implementa todo a partir de los nodos anteriores.
- **Desplace los datos a la conexión**. Desplaza los datos a la base de datos especificada aquí. De este modo permite trabajar al modelado si los datos se encuentran en otra base de datos de IBM Netezza o en una base de datos de otro proveedor, o incluso si los datos se encuentran en un archivo sin formato. Además, los datos vuelven a la base de datos especificada aquí si se han extraído los datos porque un nodo no ha llevado a cabo la retroacción SQL. Haga clic en el botón **Edición** para buscar y seleccionar una conexión. *Precaución*: IBM Netezza Analytics suele utilizarse con conjuntos de datos muy grandes. Transferir grandes cantidades de datos entre bases de datos, o extraerlos y devolverlos a la base de datos, lleva mucho tiempo y se debería evitar en la medida de lo posible.

Nombre de tabla. El nombre de la tabla base de datos donde se almacenará el modelo. *Nota*: debe ser una nueva tabla; no puede utilizar una tabla existente para esta operación.

Comentarios

- La conexión que se utiliza para el modelado no tiene la misma que se utiliza en el nodo de origen de una ruta, o puede ser otra diferente. Por ejemplo, puede tener una ruta que acceda a los datos desde una base de datos de IBM Netezza, descargue los datos a IBM SPSS Modeler para realizar limpiezas y otras manipulaciones y, a continuación, cargue los datos en una base de datos de IBM Netezza diferente para realizar modelados. No obstante, tenga en cuenta que ese tipo de configuración puede afectar al rendimiento de forma negativa.
- El nombre de origen de datos ODBC se incrusta de manera efectiva en cada ruta de IBM SPSS Modeler. Si una ruta creada en un host se ejecuta en un host diferente, el nombre del origen de datos debe ser el mismo en cada host. Si lo prefiere, se puede seleccionar un origen de datos diferente en la pestaña Servidor de cada nodo de origen o de modelado.

Modelos de Netezza: Opciones del modelo

En la pestaña Opciones del modelo, puede seleccionar si desea especificar un nombre para el modelo o generar un nombre automáticamente. También puede definir valores predeterminados para opciones de puntuación.

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Dejar disponible para puntuar. Puede establecer aquí los valores predeterminados para las opciones de puntuación que aparecen en el cuadro de diálogo del nugget de modelo. Para ver los detalles de las opciones, consulte el tema de ayuda de la pestaña Configuración de ese nugget concreto.

Gestión de modelos Netezza

La generación de un modelo de IBM Netezza a través de IBM SPSS Modeler crea un modelo en IBM SPSS Modeler y crea o sustituye un modelo en la base de datos de Netezza. El modelo de IBM SPSS Modeler de este tipo hace referencia al contenido de un modelo de base de datos almacenado en el servidor de una base de datos. IBM SPSS Modeler puede realizar la comprobación de coherencia almacenando una serie de clave idéntica del modelo generado en el modelo de IBM SPSS Modeler y el modelo de Netezza.

El nombre del modelo para cada modelo de Netezza se visualiza debajo de la columna *Información de modelo* en el cuadro de diálogo Listado de modelos de base de datos. El nombre de modelo para un modelo de IBM SPSS Modeler se visualiza como la Clave de modelo en la pestaña Servidor de un modelo de IBM SPSS Modeler (cuando se coloca en una ruta).

El botón Comprobar se puede utilizar para comprobar que las claves del modelo de IBM SPSS Modeler y del modelo de Netezza coinciden. Si no se puede encontrar ningún modelo con el mismo nombre en Netezza, o si las claves del modelo no coinciden, el modelo de Netezza se ha suprimido o vuelto a generar desde que se creó el modelo de IBM SPSS Modeler.

Creación de lista de modelos de la base de datos

IBM SPSS Modeler proporciona un cuadro de diálogo para listar los modelos que están almacenados en IBM Netezza y permite suprimir modelos. Se puede acceder a este cuadro de diálogo desde el cuadro de diálogo de las aplicaciones de ayuda de IBM y desde los cuadros de diálogo de creación, exploración y aplicación para los nodos relacionados con la minería de datos de IBM Netezza. La siguiente información se muestra para cada modelo:

- Nombre del modelo (nombre del modelo, utilizado para ordenar la lista).
- Nombre de propietario.
- El algoritmo utilizado en el modelo.
- El estado actual del modelo; por ejemplo, Completo.

- La fecha cuando se creó el modelo.

Árboles de decisión de Netezza

Un árbol de decisión es una estructura jerárquica que representa un modelo de clasificación. Con un modelo de árbol de decisión, puede desarrollar un sistema de clasificación para predecir o clasificar observaciones futuras desde un conjunto de datos de entrenamiento. La clasificación toma la forma de una estructura de árbol donde las ramas representan puntos de división en la clasificación. Los puntos de división dividen los datos en subgrupos de forma recursiva hasta que se alcanza un punto de parada. Los nodos de árbol en los puntos de parada se conocen como **hojas**. Cada hoja asigna una etiqueta, conocida como **etiqueta clase**, para los miembros de su subgrupo o clase.

El resultado del modelo toma la forma de una representación textual del árbol. Cada línea del texto corresponde a un nodo o una hoja y el espacio refleja el nivel del árbol. Para un nodo, aparece la condición de división; para una hoja, se muestra la etiqueta de clase asignada.

Ponderaciones de instancias y ponderaciones de clases

De forma predeterminada, se supone que todos los registros de entrada y todas las clases tienen la misma importancia relativa. Puede cambiar esto asignando ponderaciones individuales a los miembros de uno o ambos elementos. El hacerlo puede ser de utilidad, por ejemplo, si los puntos de datos en sus datos de entrenamiento no están distribuidos de forma realista entre las categorías. Las ponderaciones le permiten sesgar el modelo, de forma que puede compensar esas categorías que están peor representadas en los datos. Al aumentar la ponderación de un valor destino debería aumentar el porcentaje de las predicciones correctas para esa categoría.

En el nodo de modelado del Árbol de decisión, puede especificar dos tipos de ponderación. **Ponderaciones de instancias** asignan una ponderación a cada fila de datos de entrada. Las ponderaciones se suelen especificar como 1.0 para la mayoría de los casos, con mayor o menor valor solamente a aquellos casos que son más o menos importantes que la mayoría, como se muestra en la tabla siguiente.

Tabla 9. Ejemplo de ponderación de instancia

ID de registro	destino	Ponderación de instancia
1	medicamentoA	1,1
2	medicamentoB	1,0
3	medicamentoA	1,0
4	medicamentoB	0,3

Las **ponderaciones de clases** asignan una ponderación a cada categoría del campo de destino, tal como se muestra en la tabla siguiente.

Tabla 10. Ejemplo de ponderación de clase

Clase	Ponderación de clase
medicamentoA	1,0
medicamentoB	1,5

Se pueden utilizar ambos tipos de ponderación al mismo tiempo, en caso de que se hayan multiplicado entre sí y se hayan utilizado como ponderaciones de instancias. De este modo, si los dos ejemplos anteriores se utilizaran juntos, el algoritmo utilizaría las ponderaciones de instancias como se muestra en la tabla siguiente.

Tabla 11. Ejemplo de cálculo de ponderación de instancia

ID de registro	Cálculo	Ponderación de instancia
1	1,1*1,0	1,1
2	1,0*1,5	1,5
3	1,0*1,0	1,0
4	0,3*1,5	0,45

Opciones de campo para los árboles de decisión de Netezza

En la pestaña Campos, puede seleccionar si desea utilizar la configuración de rol de campo ya definida en nodos anteriores o realizar las asignaciones de campos manualmente.

Utilizar roles predefinidos. Esta opción utiliza las definiciones de roles (objetivos, predictores, etcétera) desde un nodo Tipo anterior (o la pestaña Tipo de un nodo de origen anterior).

Utilizar asignaciones de campos personalizadas. Seleccione esta opción si desea asignar objetivos, predictores y otros roles manualmente en esta pantalla.

Campos. Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los diferentes campos de roles en la parte derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol.

Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse en un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

Objetivo. Seleccione un campo como el destino de la predicción.

ID de registro. El campo que se utilizará como el identificador de registros exclusivo. Los valores de esta campo deben ser exclusivos para cada registro (por ejemplo, número ID del cliente).

Ponderación de instancia. La especificación de un campo aquí le permite utilizar ponderaciones de instancia (una ponderación por fila de datos de entrada) en lugar de, o además de las ponderaciones de clase predeterminadas (una ponderación por categoría para el campo destino). El campo que especifica aquí debe ser uno de los que contienen una ponderación numérica para cada fila de datos de entrada. Para obtener más información, consulte el tema "Ponderaciones de instancias y ponderaciones de clases" en la página 85.

Predictores (Entradas). Seleccione el campo(s) de entrada. Se trata de una acción similar a establecer el rol del campo en *Entrada* en un nodo Tipo.

Opciones de generación de árbol de decisión de Netezza

Las opciones de generación siguientes están disponibles para el crecimiento del árbol:

Medida de crecimiento. Estas opciones controlan el modo en el que se mide el crecimiento de árbol. Si no desea utilizar los valores predeterminados, pulse **Personalizar** y cambie los valores.

- **Medida de impureza.** Esta medida evalúa el mejor lugar para dividir el árbol. Es una medición de la variabilidad en un subgrupo de segmentos de datos. Una medición de impureza baja indica un grupo donde la mayoría de los miembros tienen valores similares para el campo de criterio o de objetivo.

  Las mediciones soportadas son **Entropía** y **Gini**. Estas mediciones se basan en probabilidades de la pertenencia de categoría para la rama.
- **Máxima profundidad de árbol.** El número máximo de niveles que puede crecer el árbol por debajo del nodo raíz, es decir, el número de veces que se divide la muestra repetidamente. El valor predeterminado es 62, que es la profundidad de árbol máxima para generar modelos.

Nota: Si el visor del nugget de modelo muestra la representación textual del modelo, se visualiza un máximo de 12 niveles del árbol.

Criterios de división. Estas opciones controlan cuando hay que detener la división del árbol. Si no desea utilizar los valores predeterminados, pulse **Personalizar** y cambie los valores.

- **Mejoras mínimas para divisiones.** La cantidad mínima por la que se debe reducir la impureza antes de crear una nueva división en el árbol. El objetivo de la creación del árbol es crear subgrupos con valores de salida similares para minimizar la impureza en cada nodo. Si la mejor división para una rama reduce la impureza una cantidad menor a la especificada por los criterios de división, la rama no se divide.
- **Número mínimo de instancias para una división.** El número mínimo de registros que se pueden dividir. Cuando quedan menos registros sin dividir que este número, no se realizan más divisiones. Puede utilizar este campo para evitar crear pequeños subgrupos en el árbol.

Estadísticos. Este parámetro define cuántas estadísticas se incluyen en el modelo. Seleccione una de las opciones siguientes:

- **Todas.** Todas las estadísticas relacionadas con la columna y con el valor se incluyen.

  Nota: Este parámetro incluye el número máximo de estadísticas y, por lo tanto, podría afectar el rendimiento del sistema. Si no desea ver el modelo en formato gráfico, especifique **Ninguno**.
- **Columnas.** Se incluyen las estadísticas relacionadas con la columna.
- **Ninguno.** Solo se incluyen las estadísticas necesaria para puntuar el modelo.

Nodo de árbol de decisión de Netezza: Ponderaciones de clases

Aquí podrá asignar ponderaciones a clases individuales. El valor predeterminado es asignar un valor de 1 a todas las clases, ponderándolas de forma equitativa. Especificando diferentes ponderaciones numéricas para diferentes etiquetas de clases, indicará el algoritmo para ponderar en consecuencia los conjuntos de entrenamiento de clases particulares.

Para cambiar una ponderación, pulse dos veces en ella en la columna **Ponderación** y realice los cambios que desee.

Valor. El conjunto de etiquetas de clases, se deriva de los posibles valores del campo de destino.

Ponderación. La ponderación que se asignará a una clase particular. La asignación de una ponderación superior a una clase hace que el modelo sea más sensitivo que la clase relativa a las otras clases.

Puede utilizar la ponderación de clases junto con las ponderaciones de instancias. Para obtener más información, consulte el tema “Ponderaciones de instancias y ponderaciones de clases” en la página 85.

Nodo de árbol de decisión de Netezza: Poda de árbol

Puede utilizar las opciones de poda para especificar el criterio de poda para el árbol de decisión. La intención de la poda es reducir el riesgo de sobreajuste eliminando los subgrupos con sobrecrecimiento que no han mejorado la precisión esperada en nuevos datos.

Medida de poda. La medida de poda predeterminada, **Precisión**, garantiza que la precisión estimada del modelo permanezca en límites aceptables después de eliminar una hoja del árbol. Si desea ubicar las ponderaciones de clases en cuentas durante la aplicación de la poda, utilice la **Precisión ponderada** alternativa.

Datos para la poda. Puede utilizar algunos o todos los datos de entrenamiento para estimar la precisión esperada en nuevos datos. Si lo prefiere, puede utilizar un conjunto de datos separado para la poda de una pestaña específica para este fin.

- **Utilizar todos los datos de entrenamiento**. Esta opción (la predeterminada) utiliza todos los datos de entrenamiento para estimar la precisión del modelo.

- **Utiliza % de los datos de entrenamiento para la poda**. Use esta opción para dividir los datos en dos grupos, uno para el entrenamiento y otro para la poda, usando el porcentaje especificado aquí para los datos de la poda.

  Seleccione **Replicar resultados** si desea especificar una semilla aleatoria para asegurarse de que se han dividido los datos de la misma forma cada vez que ejecuta ruta. Puede especificar un valor entero en el campo **Semilla utilizada en el campo de poda** o pulsar en **Generar**, que creará un valor entero seudoaleatorio.
- **Utilice datos desde una pestaña existente**. Especifique el nombre de la pestaña de un conjunto de datos de poda separado para estimar la precisión del modelo. Realizar esta acción está más fiable que utilizar datos de entrenamiento. Sin embargo, esta opción puede causar la eliminación de un gran subconjunto de datos del conjunto de entrenamiento además de la reducción de la calidad del árbol de decisión.

K-medias de Netezza

El nodo K-Medias implementa el algoritmo *k*-medias, que proporciona un método de análisis de clúster. Puede utilizar este nodo para agrupar un conjunto de datos en grupos distintos.

El algoritmo es un algoritmo de agrupación en clústeres basado en la distancia que se basa en una métrica de distancia (función) para medir la similitud entre puntos de datos. Los puntos de datos se asignan al clúster más próximo en función de la métrica de distancia empleada.

El algoritmo funciona realizando varias iteraciones del mismo proceso básico en el cual se asigna cada instancia de prueba al clúster más cercano (respecto a la función de distancia especificada, aplicada al centro de instancia y clúster). Todos los centros de clústeres se vuelven a calcular como los vectores de valor de los atributos de medias de las instancias asignadas a clústeres particulares.

Opciones de campo para K-medias de Netezza

En la pestaña Campos, puede seleccionar si desea utilizar la configuración de rol de campo ya definida en nodos anteriores o realizar las asignaciones de campos manualmente.

Utilizar roles predefinidos. Esta opción utiliza las definiciones de roles (objetivos, predictores, etcétera) desde un nodo Tipo anterior (o la pestaña Tipo de un nodo de origen anterior).

Utilizar asignaciones de campos personalizadas. Seleccione esta opción si desea asignar objetivos, predictores y otros roles manualmente en esta pantalla.

Campos. Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los diferentes campos de roles en la parte derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol.

Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse en un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

ID de registro. El campo que se utilizará como el identificador de registros exclusivo.

Predictores (Entradas). Seleccione uno o más campos como entradas de la predicción.

Pestaña Opciones de generación de K-medias de Netezza

Al definir las opciones de generación, puede personaliza la generación del modelo para sus propios propósitos.

Si desea generar un modelo con las opciones predeterminadas, pulse **Ejecutar**.

Medida de distancia. Este parámetro define el método de medida para la distancia entre puntos de entradas. Las distancias mayores indican disimilaridades más grandes. Seleccione una de las opciones siguientes:

- **Euclídea.** La medida euclídea es la distancia en línea recta entre dos puntos.
- **Euclídea normalizada.** La medida euclídea normalizada es similar a la medida euclídea, pero está normalizada mediante la desviación estándar al cuadrado. A diferencia de la medida euclídea, la medida euclídea normalizada también es invariante de escala.
- **Mahalanobis.** La medida de Mahalanobis es una medida euclídea generalizada que tiene en cuenta las correlaciones de datos de entrada. Al igual que la medida euclídea normalizada, la medida de Mahalanobis es invariante de escala.
- **Manhattan.** La medida de Manhattan es la distancia entre dos puntos de datos que se calcula como la suma de las diferencias absolutas entre sus coordenadas.
- **Canberra.** La medida de Canberra es similar a la medida de Manhattan pero es más sensible a los puntos de datos que están más cerca del origen.
- **Máxima.** La medida Máxima es la distancia entre dos puntos de datos que se calcula como la mayor de sus diferencias entre cualquier dimensión de las coordenadas.

Número de clústeres. Este parámetro define el número de clústeres que se va a crear.

Número máximo de iteraciones. El algoritmo realiza varias iteraciones del mismo proceso. Este parámetro define el número de iteraciones después de las cuales la formación del modelo se detiene.

Estadísticos. Este parámetro define cuántas estadísticas se incluyen en el modelo. Seleccione una de las opciones siguientes:

- **Todas.** Todas las estadísticas relacionadas con la columna y con el valor se incluyen.

  Nota: Este parámetro incluye el número máximo de estadísticas y, por lo tanto, podría afectar el rendimiento del sistema. Si no desea ver el modelo en formato gráfico, especifique **Ninguno**.
- **Columnas.** Se incluyen las estadísticas relacionadas con la columna.
- **Ninguno.** Solo se incluyen las estadísticas necesaria para puntuar el modelo.

Replicar resultados. Seleccione esta casilla de verificación si desea definir una semilla aleatoria para replicar análisis. Puede especificar un entero, o puede crear un entero pseudo aleatorio pulsando **Generar**.

Red bayesiana de Netezza

Una red bayesiana es un modelo que muestra variables en un conjunto de datos y las independencias probabilísticas o condicionales entre ellas. Si utiliza el nodo Red bayesiana de Netezza, podrá 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.

Opciones de campo de red bayesiana de Netezza

En la pestaña Campos, puede seleccionar si desea utilizar la configuración de rol de campo ya definida en nodos anteriores o realizar las asignaciones de campos manualmente.

En este nodo, el campo de destino solo es necesario para la puntuación, por lo que no se muestra en esta pestaña. Puede definir o cambiar el destino de un nodo Tipo, en la pestaña Opciones del modelo de este nodo o en la pestaña Configuración del nugget de modelo. Para obtener más información, consulte el tema “Nugget de red bayesiana de Netezza: Pestaña Configuración” en la página 108.

Utilizar roles predefinidos. Esta opción utiliza las definiciones de roles (objetivos, predictores, etcétera) desde un nodo Tipo anterior (o la pestaña Tipo de un nodo de origen anterior).

Utilizar asignaciones de campos personalizadas. Seleccione esta opción si desea asignar objetivos, predictores y otros roles manualmente en esta pantalla.

Campos. Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los diferentes campos de roles en la parte derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol.

Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse en un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

Predictores (Entradas). Seleccione uno o más campos como entradas de la predicción.

Opciones de generación de red bayesiana de Netezza

La pestaña Opciones de generación es la ubicación en la que se definen todas las opciones para crear el modelo. Puede pulsar en el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero normalmente querrá personalizar la generación de sus tareas.

Índice básico. El identificador numérico que se asignará al primer atributo (campo de entrada) para una gestión interna más fácil.

Tamaño muestral. El tamaño de la muestra que se tomará si el número de los atributos es tan grande que causará un tiempo de procesamiento inaceptablemente largo.

Mostrar información adicional durante la ejecución. Si esta casilla está marcada (valor predeterminado), se muestra información de progreso adicional en un cuadro de diálogo de mensaje.

Bayesiano ingenuo de Netezza

Bayesiano ingenuo es un algoritmo muy utilizado para resolver problemas de clasificación. El modelo se denomina *ingenuo* porque trata todas las variables de predicción propuestas como independientes unas de otras. El bayesiano ingenuo es un algoritmo rápido y escalable que calcula las probabilidades condicionales para las combinaciones de atributos y el atributo de objetivo. A partir de los datos de entrenamiento se establece una probabilidad independiente. Esta probabilidad proporciona la verosimilitud de cada clase objetivo, una vez dada la instancia de cada categoría de valor a partir de cada variable de entrada.

KNN de Netezza

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 parecidos están próximos y los que no lo son están alejados 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 los casos más similares (los vecinos más próximos) se anotan y el nuevo caso se sitúa en la categoría que contiene el mayor número de vecinos próximos.

Puede especificar el número de vecinos más próximos para examinar; este valor se denomina *k*. Las imágenes muestran cómo un podría clasificarse un nuevo caso utilizando dos valores diferentes de *k*. Si *k*

= 5, el nuevo caso se sitúa en la categoría *1* porque una mayoría de vecinos más próximos pertenecen a la categoría *1*. Sin embargo, si *k* = 9, el nuevo caso se sitúa en la categoría *0* porque una mayoría de vecinos próximos pertenecen a la categoría *0*.

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 predicho del nuevo caso.

Opciones de modelo de KNN de Netezza: Opciones generales

En la pestaña Opciones: General del modelo, puede seleccionar si desea especificar un nombre para el modelo o generar un nombre automáticamente. También puede definir opciones que controlen cómo se calcula el número de vecinos más próximos y definir opciones para mejorar el rendimiento y precisión del modelo.

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Vecinos

Medida de distancia. El método que se debe utilizar para medir la distancia entre puntos de datos; una mayor distancia indica mayores diferencias. Las opciones son:

- **Euclídea**. (predeterminada) La distancia entre dos puntos se calcula dibujando una línea recta que los una.
- **Manhattan**. La distancia entre dos puntos se calcula con la suma de las diferencias absolutas entre sus coordenadas.
- **Canberra**. Similar a la distancia Manhattan, pero más receptiva a los puntos de datos más cercanos al origen.
- **Máximo**. La distancia entre dos puntos se calcula como la diferencia mayor en cualquiera de las dimensiones de coordenadas.

Número de vecinos más próximos (k) El número de vecinos más próximos de un caso concreto. Tenga en cuenta que el uso de un número mayor de vecinos no implica que el modelo resultante sea más preciso.

La elección de *k* controla el equilibrio entre la prevención del sobreajuste (puede ser importante, especialmente para datos "con ruido") y resolución (produciendo predicciones diferentes para instancias similares). Normalmente tendrá que ajustar el valor de *k* en cada conjunto de datos, con los valores típicos comprendidos entre 1 y varias docenas.

Mejorar rendimiento y precisión

Estandarizar mediciones antes de calcular la distancia. Si está seleccionada, esta opción estandariza las mediciones de campos de entrada continuos antes de calcular los valores de distancia.

Utilizar conjuntos principales para incrementar el rendimiento de los conjuntos de datos grandes. Si está seleccionada, esta opción utiliza el muestreo del conjunto principal para acelerar el cálculo de conjuntos de datos grandes.

Opciones de modelo de KNN de Netezza: Opciones de puntuación

En la pestaña Opciones de modelo: Opciones de puntuación, puede definir el valor predeterminado de una opción de puntuación y asignar ponderaciones relativas a clases individuales.

Dejar disponible para puntuar

Incluir campos de entrada. Especifica si los campos de entrada se incluyen en la puntuación de forma predeterminada.

Ponderaciones de clases

Utilice esta opción si desea cambiar la importancia relativa de clases individuales en la creación del modelo.

Nota: esta opción solo se activa si utiliza KNN para clasificación. Si ejecuta una regresión (es decir, si el tipo de campo de destino es Continuo), esta opción está desactivada.

El valor predeterminado es asignar un valor de 1 a todas las clases, ponderándolas de forma equitativa. Especificando diferentes ponderaciones numéricas para diferentes etiquetas de clases, indicará el algoritmo para ponderar en consecuencia los conjuntos de entrenamiento de clases particulares.

Para cambiar una ponderación, pulse dos veces en ella en la columna **Ponderación** y realice los cambios que desee.

Valor. El conjunto de etiquetas de clases, se deriva de los posibles valores del campo de destino.

Ponderación. La ponderación que se asignará a una clase particular. La asignación de una ponderación superior a una clase hace que el modelo sea más sensitivo que la clase relativa a las otras clases.

Clúster divisivo de Netezza

La agrupación en clústeres divisiva es un método de análisis de clúster donde el algoritmo se ejecuta repetidamente para dividir clústeres en subclústeres hasta que se alcanza un punto de detención especificado.

La formación del clúster comienza con un clúster único con todas las instancias de entrenamiento (registros). La primera iteración del algoritmo divide el conjunto de datos en dos subclústeres, con iteraciones posteriores dividiéndolas en más subclústeres. Los criterios de parada se especifican como un número máximo de iteraciones, un número máximo de niveles en los que se divide el conjunto de datos y un número mínimo necesario de instancias para más particiones.

El árbol de agrupación en clústeres jerárquico resultante se puede utilizar para clasificar instancias propagándolas hacia abajo desde el clúster raíz, como en el siguiente ejemplo.

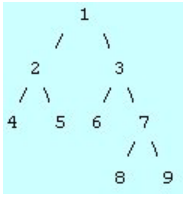

Figura 4. Ejemplo de un árbol de agrupación en clústeres divisiva

En cada nivel, se selecciona el subclúster que mejor se ajusta con respecto a la distancia de la instancia desde los centros de subclústeres.

Si las instancias se puntúan con un nivel de jerarquía aplicado de -1 (el valor predeterminado), la puntuación solo devuelve un clúster de hoja, ya que un número negativo designa las hojas. En el ejemplo, sería uno de los clústeres 4, 5, 6, 8 o 9. Sin embargo, si el nivel de jerarquía está definido como 2, por ejemplo, la puntuación devolvería uno de los clústeres del segundo nivel bajo el clúster raíz, es decir, 4, 5, 6 o 7.

Opciones de campo de clúster divisivo de Netezza

En la pestaña Campos, puede seleccionar si desea utilizar la configuración de rol de campo ya definida en nodos anteriores o realizar las asignaciones de campos manualmente.

Utilizar roles predefinidos. Esta opción utiliza las definiciones de roles (objetivos, predictores, etcétera) desde un nodo Tipo anterior (o la pestaña Tipo de un nodo de origen anterior).

Utilizar asignaciones de campos personalizadas. Seleccione esta opción si desea asignar objetivos, predictores y otros roles manualmente en esta pantalla.

Campos. Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los diferentes campos de roles en la parte derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol.

Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse en un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

ID de registro. El campo que se utilizará como el identificador de registros exclusivo.

Predictores (Entradas). Seleccione uno o más campos como entradas de la predicción.

Opciones de generación de clúster divisivo de Netezza

La pestaña Opciones de generación es la ubicación en la que se definen todas las opciones para crear el modelo. Puede pulsar en el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero normalmente querrá personalizar la generación de sus tareas.

Medida de distancia. El método que se debe utilizar para medir la distancia entre puntos de datos; una mayor distancia indica mayores diferencias. Las opciones son:

- **Euclídea**. (predeterminada) La distancia entre dos puntos se calcula dibujando una línea recta que los una.
- **Manhattan**. La distancia entre dos puntos se calcula con la suma de las diferencias absolutas entre sus coordenadas.
- **Canberra**. Similar a la distancia Manhattan, pero más receptiva a los puntos de datos más cercanos al origen.
- **Máximo**. La distancia entre dos puntos se calcula como la diferencia mayor en cualquiera de las dimensiones de coordenadas.

Número máximo de iteraciones. El algoritmo funciona realizando varias iteraciones del mismo proceso. Esta opción permite detener el entrenamiento del modelo después del número de iteraciones especificado.

Profundidad máxima de los árboles de clúster. El número máximo de niveles en que se puede subdividir el conjunto de datos.

Replicar resultados. Seleccione esta casilla si desea establecer una semilla aleatoria que le permitirá replicar análisis. Puede especificar un entero o pulsar en **Generar**, que crea un entero seudoaleatorio.

Número mínimo de instancias para una división. El número mínimo de registros que se pueden dividir. Cuando queden menos registros por dividir que este número, no se realizarán más divisiones. Puede emplear este campo para evitar la creación de subgrupos de tamaño muy reducido en el árbol de clústeres.

PCA de Netezza

El análisis de componentes principal (PCA) es una poderosa técnica de reducción de datos, diseñada para reducir la complejidad de los datos. PCA busca combinaciones lineales de los campos de entrada que realizan el mejor trabajo a la hora de capturar la varianza en todo el conjunto de campos, en el que los componentes son ortogonales (no correlacionados) entre ellos. El objetivo es encontrar un número pequeño de campos derivados (los componentes principales) que resuman de forma eficaz la información del conjunto original de campos de entrada.

Opciones de campo de PCA de Netezza

En la pestaña Campos, puede seleccionar si desea utilizar la configuración de rol de campo ya definida en nodos anteriores o realizar las asignaciones de campos manualmente.

Utilizar roles predefinidos. Esta opción utiliza las definiciones de roles (objetivos, predictores, etcétera) desde un nodo Tipo anterior (o la pestaña Tipo de un nodo de origen anterior).

Utilizar asignaciones de campos personalizadas. Seleccione esta opción si desea asignar objetivos, predictores y otros roles manualmente en esta pantalla.

Campos. Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los diferentes campos de roles en la parte derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol.

Pulse en el botón **Todos** para seleccionar todos los campos de la lista o pulse en un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

ID de registro. El campo que se utilizará como el identificador de registros exclusivo.

Predictores (Entradas). Seleccione uno o más campos como entradas de la predicción.

Opciones de generación de PCA de Netezza

La pestaña Opciones de generación es la ubicación en la que se definen todas las opciones para crear el modelo. Puede pulsar en el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero normalmente querrá personalizar la generación de sus tareas.

Centrar datos antes de calcular el PCA. Si está activada (opción predeterminada), esta opción ejecuta el centrado de datos (también conocido como "sustracción de media") antes del análisis. El centrado de datos es necesario para garantizar que el primer componente principal describe la dirección de la varianza máxima; de lo contrario es posible que el componente se corresponda más con la media de los datos. Normalmente desactivaría esta opción solo para mejorar el rendimiento si los datos ya se han preparado de esta forma.

Realizar adaptación de datos antes de calcular el PCA. Esta opción ejecuta la adaptación de los datos antes del análisis. De esta forma el análisis será menos arbitrario si las diferentes variables se miden en unidades diferentes. En su forma más simple, la adaptación de datos se puede realizar dividiendo cada variable por su variación estándar.

Utilizar un método menos preciso pero más rápido para calcular el PCA. Esta opción provoca que el algoritmo utilice un método menos preciso pero más rápido (forceEigensolve) para buscar los componentes principales.

Árbol de regresión de Netezza

Un árbol de regresión es un algoritmo basado en árbol que divide una muestra de casos de forma repetida para derivar subconjuntos del mismo tipo, en función de los valores de un campo de salida numérico. Al igual que con los árboles de decisión, los árboles de regresión descomponen los datos en subconjuntos en los que las hojas del árbol se corresponden con subconjuntos suficientemente pequeños o suficientemente uniformes. Las divisiones se seleccionan para reducir la dispersión de valores de atributos de objetivo, por lo que se pueden predecir razonablemente bien por los valores medios de sus hojas.

El resultado del modelo toma la forma de una representación textual del árbol. Cada línea del texto corresponde a un nodo o una hoja y el espacio refleja el nivel del árbol. Para un nodo, aparece la condición de división; para una hoja, se muestra la etiqueta de clase asignada.

Opciones de generación de árbol de regresión de Netezza Regression - Crecimiento del árbol

Puede definir opciones de generación para el crecimiento del árbol y la poda del árbol.

Si no desea utilizar los valores predeterminados, pulse **Personalizar** y cambie los valores.

Las opciones de generación siguientes están disponibles para el crecimiento del árbol:

Máxima profundidad de árbol. El número máximo de niveles que puede crecer el árbol por debajo del nodo raíz, es decir, el número de veces que se divide la muestra repetidamente. El valor predeterminado es 62, que es la profundidad de árbol máxima para generar modelos.

Nota: Si el visor del nugget de modelo muestra la representación textual del modelo, se visualiza un máximo de 12 niveles del árbol.

Criterios de división. Estas opciones controlan cuando hay que detener la división del árbol. Si no desea utilizar los valores predeterminados, pulse **Personalizar** y cambie los valores.

- **Medida de evaluación de división.** Esta medida de evaluación de clase evalúa el mejor lugar para dividir el árbol.

  Nota: Actualmente, la varianza es la única opción posible.
- **Mejoras mínimas para divisiones.** La cantidad mínima por la que se debe reducir la impureza antes de crear una nueva división en el árbol. El objetivo de la creación del árbol es crear subgrupos con valores de salida similares para minimizar la impureza en cada nodo. Si la mejor división para una rama reduce la impureza una cantidad menor a la especificada por los criterios de división, la rama no se divide.
- **Número mínimo de instancias para una división.** El número mínimo de registros que se pueden dividir. Cuando quedan menos registros sin dividir que este número, no se realizan más divisiones. Puede utilizar este campo para evitar crear pequeños subgrupos en el árbol.

Estadísticos. Este parámetro define cuántas estadísticas se incluyen en el modelo. Seleccione una de las opciones siguientes:

- **Todas.** Todas las estadísticas relacionadas con la columna y con el valor se incluyen.

  Nota: Este parámetro incluye el número máximo de estadísticas y, por lo tanto, podría afectar el rendimiento del sistema. Si no desea ver el modelo en formato gráfico, especifique **Ninguno**.
- **Columnas.** Se incluyen las estadísticas relacionadas con la columna.
- **Ninguno.** Solo se incluyen las estadísticas necesaria para puntuar el modelo.

Opciones de generación de árbol de regresión de Netezza: Poda de árbol

Puede utilizar las opciones de poda para especificar el criterio de poda para el árbol de regresión. La intención de la poda es reducir el riesgo de sobreajuste eliminando los subgrupos con sobrecrecimiento que no han mejorado la precisión esperada en nuevos datos.

Medida de poda. La medida de poda garantiza que la precisión estimada del modelo permanezca en límites aceptables después de eliminar una hoja del árbol. Puede elegir una de las siguientes mediciones.

- **mse.** Error cuadrático promedio: (valor predeterminado) mide la cercanía de una línea ajustada a sus puntos de datos.
- **r2.** R cuadrado: mide la proporción de variación de la variable dependiente explicada por el modelo de regresión.
- **Pearson.** Coeficiente de correlación de Pearson: mide la fuerza de la relación entre variables dependientes lineales que se distribuyen normalmente.
- **Spearman.** Coeficiente de correlación de Spearman: detecta relaciones no lineales que parecen débiles en función de la correlación de Pearson, pero que pueden ser fuertes.

Datos para la poda. Puede utilizar algunos o todos los datos de entrenamiento para estimar la precisión esperada en nuevos datos. Si lo prefiere, puede utilizar un conjunto de datos separado para la poda de una pestaña específica para este fin.

- **Utilizar todos los datos de entrenamiento**. Esta opción (la predeterminada) utiliza todos los datos de entrenamiento para estimar la precisión del modelo.
- **Utiliza % de los datos de entrenamiento para la poda**. Use esta opción para dividir los datos en dos grupos, uno para el entrenamiento y otro para la poda, usando el porcentaje especificado aquí para los datos de la poda.

  Seleccione **Replicar resultados** si desea especificar una semilla aleatoria para asegurarse de que se han dividido los datos de la misma forma cada vez que ejecuta ruta. Puede especificar un valor entero en el campo **Semilla utilizada en el campo de poda** o pulsar en **Generar**, que creará un valor entero seudoaleatorio.
- **Utilice datos desde una pestaña existente**. Especifique el nombre de la pestaña de un conjunto de datos de poda separado para estimar la precisión del modelo. Realizar esta acción está más fiable que utilizar datos de entrenamiento. Sin embargo, esta opción puede causar la eliminación de un gran subconjunto de datos del conjunto de entrenamiento además de la reducción de la calidad del árbol de decisión.

Regresión lineal de Netezza

Los modelos lineales predicen un objetivo continuo tomando como base las relaciones lineales entre el destino y uno o más predictores. Mientras están limitadas únicamente a relaciones lineales de modelado directamente, los modelos de regresión lineales son relativamente simples y proporcionan una fórmula matemática de fácil interpretación para su puntuación. Los modelos lineales son rápidos, eficientes y fáciles de usar, aunque su aplicabilidad está limitada en comparación con los producidos por algoritmos de regresión más ajustados.

Opciones de generación de regresión lineal de Netezza

La pestaña Opciones de generación es la ubicación en la que se definen todas las opciones para crear el modelo. Puede pulsar en el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero normalmente querrá personalizar la generación de sus tareas.

Utilizar descomposición de valores singulares para solucionar las ecuaciones. El uso de la matriz de descomposición de valores singulares en lugar de la matriz original tiene la ventaja de ser más robusta con errores numéricos y puede también acelerar los cálculos.

Incluir la interceptación en el modelo. Incluir la interceptación aumenta la precisión global de la solución.

Calcular diagnósticos del modelo. Esta opción produce un número de diagnósticos que se calcularán en el modelo. Los resultados se guardan en matrices o tablas para su posterior revisión. Entre los diagnósticos se incluyen R cuadrado, sumas residuales de cuadrados, estimación de varianza, desviación estándar, valor *p* y valor *t*.

Estos diagnósticos están relacionados con al validez y utilidad del modelo. Debe ejecutar diagnósticos diferentes en los datos subyacentes para garantizar que cumple con los supuestos de linealidad.

Series temporales de Netezza

Una **serie temporal** es una secuencia de valores de datos numéricos, medidos en puntos temporales sucesivos (aunque no necesariamente regulares); por ejemplo, precios de acciones diarios o datos de ventas semanales. El análisis de dichos datos puede ser de utilidad, por ejemplo, para resaltar comportamientos como tendencias y estacionalidad (un patrón que se repite) y para predecir comportamientos futuros a partir de eventos pasados.

El nodo Series temporales de Netezza admite los siguientes algoritmos de series temporales.

- análisis espectral
- suavizado exponencial
- Modelo Autorregresivo Integrado de Media Móvil (ARIMA)
- descomposición de tendencias estacional

Estos algoritmos descomponen una serie temporal en un componente de tendencias y un componente estacional. A continuación, estos componentes se analizan para generar un modelo que pueda utilizarse para la predicción.

Análisis espectral se utiliza para identificar comportamientos periódicos en series temporales. En el caso de series temporales compuestas por múltiples periodicidades subyacentes o cuando haya una cantidad considerable de ruido aleatorio presente en los datos, el análisis espectral ofrece el medio más claro para identificar componentes periódicos. Este método detecta las frecuencias del comportamiento periódico transformando la serie del dominio temporal a una serie del dominio de frecuencia.

El **suavizado exponencial** es un método de predicción que utiliza los valores ponderados de las observaciones anteriores de la serie para predecir los valores futuros. Con el suavizado exponencial, la influencia de las observaciones disminuye con el tiempo de manera exponencial. Este método prevé un punto cada vez, ajustando sus previsiones a medida que entran nuevos datos, teniendo en cuenta la adición de cuentas, las tendencias y la estacionalidad.

Los modelos **ARIMA** ofrecen métodos más sofisticados para modelar componentes de tendencias y estacionales que los modelos de suavizado exponencia. Este método implica la especificación explícita de órdenes autorregresivos y de media móvil además del grado de diferenciación.

Nota: en términos prácticos, los modelos ARIMA son especialmente útiles si desea incluir predictores que puedan ayudar a explicar el comportamiento de la serie que se está previendo, como el número de catálogos enviados por correo o el número de visitas de la página Web de una empresa. Los modelos de suavizado exponencial describen el comportamiento de la serie temporal sin tratar de explicar el motivo de su comportamiento.

La **descomposición de tendencias estacional** elimina el comportamiento periódico de las series temporales para realizar un análisis de tendencias y, a continuación, selecciona una forma básica para la tendencia, como una función cuadrática. Estas formas básicas tienen un determinado número de

parámetros cuyos valores están determinados para minimizar el error cuadrático promedio de los residuos (es decir, las diferencias entre los valores ajustados y observados de las series temporales).

Interpolación de valores en Series temporales de Netezza

La **interpolación** es el proceso de estimación e inserción de valores perdidos en datos de series temporales.

Si los intervalos de las series temporales son regulares pero algunos valores sencillamente no están presentes, los valores perdidos pueden estimarse mediante la interpolación lineal. Considere las siguientes series de llegadas de pasajeros mensuales a la terminal de un aeropuerto.

Tabla 12. Llegadas mensuales a una terminal de pasajeros

Mes	Pasajeros
3	3.500.000
4	3.900.000
5	-
6	3.400.000
7	4.500.000
8	3.900.000
9	5.800.000
10	6.000.000

En este caso, la interpolación lineal estimaría el valor perdido del mes 5 como 3.650.000 (el punto medio entre los meses 4 y 6).

Los intervalos irregulares se tratan de manera diferente. Considere las siguientes series de lecturas de temperaturas.

Tabla 13. Lecturas de temperaturas

Fecha	Hora	Temperatura:
24-07-2011	7:00	57
14-07-2011	14:00	75
24-07-2011	21:00	72
25-07-2011	7:15	59
25-07-2011	14:00	77
25-07-2011	20:55	74
27-07-2011	7:00	60
27-07-2011	14:00	78
27-07-2011	22:00	74

Aquí tenemos lecturas tomadas en tres puntos durante tres días, pero a horas diversas y únicamente algunas de ellas son comunes entre los distintos días. Además, únicamente dos de los días son consecutivos.

Esta situación se puede gestionar de una de las las dos maneras siguientes: calculando agregados o determinando un tamaño de paso.

Los agregados pueden ser agregados diarios calculados según una fórmula basada en el conocimiento semántico de los datos. De realizar esta acción, se obtendría el siguiente conjunto de datos.

Tabla 14. Lecturas de temperaturas (agregadas)

Fecha	Hora	Temperatura:
24-07-2011	24:00	69
25-07-2011	24:00	71
26-07-2011	24:00	nulo
27-07-2011	24:00	72

Por otro lado, el algoritmo puede tratar la serie como una serie distinta y determinar un tamaño de paso adecuado. En este caso, el tamaño de paso determinado por el algoritmo podrían ser 8 horas, con lo que se obtendría lo siguiente.

Tabla 15. Lecturas de temperaturas con el tamaño de paso calculado

Fecha	Hora	Temperatura:
24-07-2011	6:00	
24-07-2011	14:00	75
24-07-2011	22:00	
25-07-2011	6:00	
25-07-2011	14:00	77
25-07-2011	22:00	
26-07-2011	6:00	
26-07-2011	14:00	
26-07-2011	22:00	
27-07-2011	6:00	
27-07-2011	14:00	78
27-07-2011	22:00	74

Aquí solamente cuatro lecturas se corresponden con las mediciones originales, pero con la ayuda de los otros valores conocidos de la serie original, los valores perdidos pueden volver a calcularse mediante interpolación.

Opciones de campos de Series temporales de Netezza

En la pestaña Campos, puede especificar roles para los campos de entrada en los datos de origen.

Campos. Utilice los botones de flecha para asignar los elementos manualmente desde esta lista a los diferentes campos de roles en la parte derecha de la pantalla. Los iconos indican los niveles de medición válidos para cada campo de rol.

Objetivo. Seleccione un campo como el destino de la predicción. Éste debe ser un campo con un nivel de medición Continuo.

(Predictor) Puntos temporales. (obligatorio) El campo de entrada que contiene los valores de fecha u hora de la serie temporal. Éste debe ser un campo con un nivel de medición Continuo o Categórico y un tipo de almacenamiento de dato de Fecha, Hora, Marca de tiempo o Numérico. El tipo de almacenamiento de datos del campo que especifique aquí también define el tipo de entrada de algunos campos de otras pestañas de este nodo de modelado.

(Predictor) Id. De series temporales (Por). Un campo que contiene diversos ID de series temporales; utilice esta opción si la entrada contiene más de una serie temporal.

Opciones de generación de Series temporales de Netezza

Hay dos niveles de opciones de generación:

- Básico: ajustes para la selección de algoritmo, interpolación e intervalo de tiempo que se utilizarán.
- Avanzado: ajustes para la previsión.

Esta sección describe las opciones básicas.

La pestaña Opciones de generación es la ubicación en la que se definen todas las opciones para crear el modelo. Puede pulsar en el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero normalmente querrá personalizar la generación de sus tareas.

Algoritmo

Estos son los ajustes relacionados con el algoritmo de series temporales que debe utilizarse.

Nombre de algoritmo. Seleccione el algoritmo de series temporales que quiera utilizar. Los algoritmos disponibles son **Análisis espectral**, **Suavizado exponencial** (valor predeterminado), **ARIMA** o **Descomposición de tendencias estacional**. Para obtener más información, consulte el tema "Series temporales de Netezza" en la página 97.

Tendencia. (únicamente Suavizado exponencial) El suavizado exponencial simple no funciona correctamente si la serie temporal muestra una tendencia. Utilice este campo para especificar la tendencia, si la hay, para que el algoritmo pueda tenerla en cuenta.

- **Determinado por el sistema.** (valor predeterminado) El sistema intenta encontrar el valor óptimo para este parámetro.
- **Ninguno(N).** La serie temporal no muestra ninguna tendencia.
- **Aditivo(A).** Una tendencia que aumenta de manera constante a lo largo del tiempo.
- **Aditivos amortiguados(DA).** Una tendencia de aditivos que finalmente desaparece.
- **Multiplicativo(M).** Una tendencia que aumenta con el paso del tiempo, normalmente más rápido que una tendencia de aditivos constante.
- **Multiplicativo amortiguado(DM).** Una tendencia de multiplicativos que finalmente desaparece.

Estacionalidad. (únicamente Suavizado exponencial) Utilice este campo para especificar si la serie temporal muestra algún patrón estacional en los datos.

- **Determinado por el sistema.** (valor predeterminado) El sistema intenta encontrar el valor óptimo para este parámetro.
- **Ninguno(N).** La serie temporal no muestra patrones estacionales.
- **Aditivo(A).** El patrón de fluctuaciones estacionales muestra una tendencia ascendente constante a lo largo del tiempo.
- **Multiplicativo(M).** Igual que la estacionalidad aditiva, pero además de la amplitud (la distancia entre los puntos altos y bajos) de las fluctuaciones estacionales aumenta con respecto a la tendencia ascendente total de las fluctuaciones.

Utilice la configuración determinada del sistema para ARIMA. (únicamente ARIMA) Seleccione esta opción si quiere que el sistema determine los ajustes del algoritmo ARIMA.

Especificar. (únicamente ARIMA) Seleccione esta opción y pulse en el botón para especificar los ajustes de ARIMA manualmente.

Interpolación

Si los datos de origen de la serie temporal tienen valores perdidos, seleccione un método para insertar valores estimados y así rellenar los huecos en los datos. Para obtener más información, consulte el tema "Interpolación de valores en Series temporales de Netezza" en la página 98.

- **Lineal.** Seleccione este método si los intervalos de la serie temporal son regulares pero algunos valores sencillamente no están presentes.
- **LíneasSp exponenciales.** Ajusta una curva suavizada donde los valores de puntos de datos conocidos aumentan o disminuyen a una tasa elevada.
- **LíneasSp cúbicas.** Ajusta una curva suavizada a los puntos de datos conocidos para estimar los valores perdidos.

Rango de horas

Aquí puede elegir si desea utilizar el rango completo de datos en la serie temporal o un subconjunto contiguo de dichos datos para crear el modelo. La entrada válida de estos campos se define por el tipo de almacenamiento de datos del campo especificado para puntos temporales en la pestaña Campos. Para obtener más información, consulte el tema "Opciones de campos de Series temporales de Netezza" en la página 99.

- **Utilice las primeras y las últimas horas disponibles en los datos.** Seleccione esta opción si desea utilizar el rango completo de los datos de la serie temporal.
- **Especifique la ventana de hora.** Seleccione esta opción si desea utilizar únicamente una parte de la serie temporal. Utilice los campos **Hora más temprana (desde)** y **Hora más tardía (hasta)** para especificar los límites.

Estructura de ARIMA

Especifique los valores de los diversos componentes estacionales y no estacionales del modelo ARIMA. En cada caso, establezca el operador como = (igual a) o <= (menor o igual que) y, a continuación, especifique el valor en el campo adyacente. Los valores deben ser enteros no negativos que especifiquen los grados.

No estacional. Los valores de los diversos componentes no estacionales del modelo.

- **Grados de correlación automática (p).** Es el número de órdenes autorregresivos del modelo. Los órdenes autorregresivos especifican los valores previos de la serie utilizados para predecir los valores actuales. Por ejemplo, un orden autorregresivo igual a 2 especifica que se van a utilizar los valores de la serie correspondientes a dos períodos de tiempo del pasado para predecir el valor actual.
- **Derivación (d).** Especifica el orden de diferenciación aplicado a la serie antes de estimar los modelos. La diferenciación es necesaria si hay tendencias (las series con tendencias suelen ser no estacionarias y el modelado de ARIMA asume la estacionariedad) y se utiliza para eliminar su efecto. El orden de diferenciación corresponde al grado de la tendencia de la serie (la diferenciación de primer orden representa las tendencias lineales, la diferenciación de segundo orden representa las tendencias cuadráticas, etc).
- **Media móvil (q).** Es el número de órdenes de media móvil presentes en el modelo. Los órdenes de media móvil especifican el modo en que se utilizan las desviaciones respecto a la media de la serie para los valores previos con el fin de predecir los valores actuales. Por ejemplo, los órdenes de media móvil de 1 y 2 especifican que las desviaciones respecto al valor medio de la serie de cada uno de los dos últimos períodos de tiempo se tienen en cuenta al predecir los valores actuales de la serie.

Estacional. Los componentes de autocorrelación estacional (SP), derivación (SD) y media móvil (SQ) tienen los mismoss roles que los correspondientes componentes no estacionales. No obstante, en el caso de los órdenes estacionales, los valores de la serie actual se ven afectados por los valores de la serie anterior separados por uno o más períodos estacionales. Por ejemplo, para los datos mensuales (período

estacional de 12), un orden estacional de 1 significa que el valor de la serie actual se ve afectado por el valor de la serie 12 períodos antes del actual. Un orden estacional de 1 para los datos mensuales equivale a la especificación de un orden no estacional de 12.

Los ajustes estacionales únicamente se consideran si se detecta estacionalidad en los datos o si especifica ajustes periódicos en la pestaña Avanzado.

Opciones de generación de Series temporales de Netezza: Avanzado

Puede utilizar los ajustes avanzados para especificar opciones para la previsión.

Utilice los ajustes determinados por el sistema para las opciones de generación de modelos. Seleccione esta opción si desea que el sistema determine los ajustes avanzados.

Especificar. Seleccione esta opción si desea especificar las opciones avanzadas manualmente. (Esta opción no está disponible si el algoritmo es Análisis espectral.)

- **Período/Unidades de período.** El período de tiempo tras el que un determinado comportamiento característico de la serie temporal se repite. Por ejemplo, en una serie temporal de cifras de ventas semanales especificaría 1 para el período y `Semanas` para las unidades. **Período** debe ser un número entero no negativo; **Unidades de período** puede ser **Milisegundos**, **Segundos**, **Minutos**, **Horas**, **Días**, **Semanas**, **Trimestres** o **Años**. No defina **Unidades de período** si no se ha definido **Período**, o si el tipo de tiempo no es un valor numérico. No obstante, si especifica **Período**, también debe especificar **Unidades de período**.

Ajustes para la previsión. Puede decidir realizar previsiones hasta un punto temporal específico o hasta puntos temporales específicos. La entrada válida de estos campos se define por el tipo de almacenamiento de datos del campo especificado para puntos temporales en la pestaña Campos. Para obtener más información, consulte el tema "Opciones de campos de Series temporales de Netezza" en la página 99.

- **Horizonte de previsión.** Seleccione esta opción si desea especificar únicamente un punto final para la previsión. Las previsiones se realizarán hasta este punto temporal.
- **Horas de previsión.** Seleccione esta opción para especificar uno o más puntos temporales en los que realizar previsiones. Pulse en **Añadir** para añadir una nueva fila a la tabla de puntos temporales. Para eliminar una fila, seleccione la fila y pulse en **Eliminar**.

Opciones del modelo Series temporales de Netezza

En la pestaña Opciones del modelo, puede seleccionar si desea especificar un nombre para el modelo o generar un nombre automáticamente. También puede definir valores predeterminados para las opciones de resultados de modelo.

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Dejar disponible para puntuar. Puede establecer aquí los valores predeterminados para las opciones de puntuación que aparecen en el cuadro de diálogo del nugget de modelo.

- **Incluya valores históricos en el resultado.** De forma predeterminada, el resultado de modelo no incluye los valores de datos históricos (los utilizados para realizar la predicción). Seleccione esta casilla de verificación para incluir estos valores.
- **Incluya valores interpolados en el resultado.** Si decide incluir valores históricos en el resultado, seleccione esta casilla de verificación si también desea incluir los valores interpolados, si los hay. Tenga en cuenta que la interpolación solo funciona en datos históricos, de modo que esta casilla no estará disponible si no se ha seleccionado **Incluya valores históricos en el resultado**. Para obtener más información, consulte el tema "Interpolación de valores en Series temporales de Netezza" en la página 98.

Lineales generalizados de Netezza

La regresión lineal es una técnica estadística establecida hace tiempo para clasificar los registros en función los valores de los campos de entrada numérica. La regresión lineal se ajusta a una línea recta o una superficie que minimiza las discrepancias entre los valores de resultados predichos y reales. Los modelos lineales son de utilidad para modelar una amplia gama de fenómenos del mundo real debido a su simplicidad tanto en la formación como en la aplicación de modelos. Sin embargo, los modelos lineales asumen una distribución normal en la variable (de objetivo) dependiente y un impacto lineal de las variables independientes (predictivas) en la variable dependiente.

Hay muchas situaciones en las que una regresión lineal es de utilidad pero los supuestos anteriores no son aplicables. Por ejemplo, al modelar la elección del consumidor entre un número de productos discreto, es probable que la variable dependiente tenga una distribución multinomial. Del mismo modo, al modelar los ingresos frente a la edad, los ingresos normalmente aumentan a medida que aumenta la edad, pero no es probable que el enlace entre los dos sea tan sencillo como una línea recta.

Para estas situaciones se puede utilizar un modelo lineal generalizado. Los modelos lineales generalizados amplían el modelo de regresión lineal de modo que la variable dependiente está relacionada con las variables predictivas mediante una determinada función de enlace, para la que puede elegir entre distintas funciones adecuadas. Además, el modelo permite que la variable dependiente tenga una distribución que no sea normal, como Poisson.

El algoritmo busca de manera iterativa el modelo que mejor se ajuste, hasta un número de iteraciones especificado. Al calcular el mejor ajuste, el error se representa mediante la suma de cuadrados de las diferencias entre el valor predicho y real de la variable dependiente.

Opciones de modelo lineal generalizado de Netezza: Opciones generales

En la pestaña Opciones del modelo, puede seleccionar si desea especificar un nombre para el modelo o generar un nombre automáticamente. También puede realizar diversos ajustes relacionados con el modelo, la función de enlace y las interacciones del campo de entrada (si las hay) y establecer valores predeterminados para opciones de puntuación.

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Opciones de campo. Puede especificar los roles de los campos de entrada para generar el modelo.

Configuración general. Estos ajustes están relacionados con los criterios de parada del algoritmo.

- **Número máximo de iteraciones.** Número máximo de iteraciones que ejecutará el algoritmo; el valor mínimo es 1 y el valor predeterminado es 20.
- **Error máximo (1e).** El valor del error máximo (en notación científica) en el que el algoritmo debería dejar de buscar el modelo de mejor ajuste. El valor mínimo es 0, el valor predeterminado es -3, lo que significa 1E-3 o 0,001.
- **Umbral de valores de error insignificantes (1e).** El valor (en notación científica) por debajo del que los errores se tratan como si su valor fuera cero. El valor mínimo es -1, el valor predeterminado es -7, lo que significa que los valores de error por debajo de 1E-7 (o 0,0000001) se cuentan como insignificantes.

Configuración de la distribución. Estos ajustes están relacionados con la distribución de la variable (de objetivo) dependiente.

- **Distribución de la variable de respuesta.** El tipo de distribución; puede ser **Bernoulli** (valor predeterminado), **Gaussian**, **Poisson**, **Binomial negativa**, **Wald** (de Gauss inversa) y **Gamma**.

- **Parámetros.** (únicamente distribución binomial negativa) Puede especificar un valor de parámetro si la distribución es binomial negativa. Seleccione especificar un valor o utilizar el valor predeterminado -1.

Configuración de la función de enlace. Estos ajustes están relacionados con la función de enlace, que relaciona la variable dependiente con las variables predictivas.

- **Función de enlace.** La función que debe utilizarse; puede ser **Identidad**, **Inversa**, **Invnegative**, **Invsquare**, **Sqrt**, **Potencia**, **Oddspower**, **Log**, **Clog**, **Loglog**, **Cloglog**, **Logit** (valor predeterminado), **Probit**, **Gaussit**, **Cauchit**, **Canbinom**, **Cangeom**, **Cannegbinom**.
- **Parámetros.** (Únicamente funciones de enlace Potencia u Oddspower) Puede especificar un valor de parámetro si la función de enlace es **Potencia** u **Oddspower**. Seleccione especificar un valor o utilizar el valor predeterminado 1.

Opciones de modelo lineal generalizado de Netezza: Interacción

El panel Interacción contiene las opciones para especificar interacciones (es decir, efectos multiplicativos entre campos de entrada).

Interacción de columna. Seleccione esta casilla de verificación para especificar interacciones entre campos de entrada. Deje la casilla sin seleccionar si no hay interacciones.

Introduzca interacciones en el modelo seleccionando uno o más campos en la lista de origenes y arrastrándolos a la lista de interacciones. El tipo de interacción creada depende de la zona activa en la que suelte la selección.

- **Principal.** Los campos que suelte aparecen como interacciones principales independientes en la parte inferior de la lista de interacciones.
- **2 factores.** Todos los pares posibles de los campos que suelte aparecen como interacciones de 2 factores en la parte inferior de la lista de interacciones.
- **3 factores.** Todos los triples posibles de los campos que suelte aparecen como interacciones de 3 factores en la parte inferior de la lista de interacciones.
- ***.** La combinación de todos los campos que suelte aparece como una única interacción en la parte inferior de la lista de interacciones.

Incluir interceptación. La interceptación se incluye normalmente en el modelo. Si se puede dar por supuesto la lectura de datos en el origen, se puede excluir la interceptación.

Botones del cuadro de diálogo

Los botones a la derecha de la pantalla le permiten realizar cambios en los términos que se utilizan en el modelo.

Figura 5. Botón Suprimir

Eliminar términos del modelo seleccionando los términos que quiera eliminar y pulsando en el botón Eliminar.

Figura 6. Botones Reordenar

Reordenar los términos dentro del modelo seleccionando los términos que desea reordenar y pulsando la flecha hacia arriba o hacia abajo.

Figura 7. Botón de interacción personalizada

Añadir un término personalizado

Puede especificar interacciones personalizadas con la forma *n1*x1*x1*x1...* Seleccione un campo de la lista de **Campos**, pulse el botón de flecha derecha para añadir el campo a **Término personalizado**, pulse **Por***, seleccione el campo siguiente, pulse el botón de flecha derecha, etc. Cuando haya creado la interacción personalizada, pulse en **Añadir término** para devolverlo al panel Interacción.

Opciones de modelo lineal generalizado de Netezza: Opciones de puntuación

Dejar disponible para puntuar. Puede establecer aquí los valores predeterminados para las opciones de puntuación que aparecen en el cuadro de diálogo del nugget de modelo. Para obtener más información, consulte el tema "Nugget de modelo lineal generalizado de Netezza: Pestaña Configuración" en la página 114.

- **Incluir campos de entrada.** Seleccione esta casilla de verificación si desea mostrar los campos de entrada en el resultado de modelo así como las predicciones.

Gestión de modelos de IBM Netezza Analytics

Se añaden los modelos de IBM Netezza Analytics a los lienzos así como a la paleta Modelos del mismo modo que en otros modelos de IBM SPSS Modeler y se pueden utilizar prácticamente del mismo modo. Sin embargo, existen algunas diferencias significativas, ya que cada modelo de IBM Netezza Analytics generado en IBM SPSS Modeler se refiere en realidad a un modelo almacenado en el servidor de una base de datos. Por tanto, para que una ruta funcione correctamente, debe conectarse a la base de datos donde se creó el modelo, y la tabla del modelo no debe haber cambiado por un proceso externo.

Puntuación de modelos de IBM Netezza Analytics

Los modelos se representan en el lienzo por un icono de nugget de modelo dorado. La finalidad principal de un nugget es puntuar datos para generar predicciones o permitir nuevos análisis de propiedades de modelos. Las puntuaciones se añaden en forma de uno o más campos de datos extra que se pueden hacer visibles adjuntando un nodo Tabla al nugget y ejecutando esa rama de la ruta, tal y como se describe posteriormente en esta sección. Algunos cuadros de diálogo de nugget, como los del árbol de decisión o de regresión, tienen además una pestaña Modelo que proporciona una representación visual del modelo.

Los campos extra se distinguen mediante el prefijo `$<id>-` añadido al nombre del campo objetivo, donde `<id>` depende del modelo e identifica el tipo de información que se añade. Los diferentes identificadores se describen en los temas de cada nugget de modelo.

Para ver los resultados, complete los pasos siguientes:

1. Conecte un nodo Tabla a este nugget de modelo.

2. Abra el nodo Tabla.
3. Haga clic en **Ejecutar**.
4. Desplácese a la derecha de la ventana de resultado de la tabla para ver los campos extra y sus resultados.

Pestaña Servidor del nugget de modelo de Netezza

En la pestaña Servidor puede definir las opciones del servidor para puntuar el modelo. Puede continuar utilizando una conexión al servidor especificada anteriormente, o bien, puede mover los datos a otra base de datos que especifique aquí.

Detalles del servidor Netezza DB. Aquí puede especificar los detalles de la conexión a la base de datos que desea utilizar para el modelo.

- **Utilice la conexión anterior**. (valor predeterminado) Utiliza los detalles de conexión especificada en un nodo anterior, por ejemplo, el nodo de origen Base de datos. *Nota*: esta opción sólo funciona si todos los nodos anteriores pueden retrotraer operaciones SQL. En este caso no es necesario extraer los datos de la base de datos, ya que SQL lo implementa todo a partir de los nodos anteriores.
- **Desplace los datos a la conexión**. Desplaza los datos a la base de datos especificada aquí. De este modo permite trabajar al modelado si los datos se encuentran en otra base de datos de IBM Netezza o en una base de datos de otro proveedor, o incluso si los datos se encuentran en un archivo sin formato. Además, los datos vuelven a la base de datos especificada aquí si se han extraído los datos porque un nodo no ha llevado a cabo la retroacción SQL. Haga clic en el botón **Edición** para buscar y seleccionar una conexión. *Precaución*: IBM Netezza Analytics suele utilizarse con conjuntos de datos muy grandes. Transferir grandes cantidades de datos entre bases de datos, o extraerlos y devolverlos a la base de datos, lleva mucho tiempo y se debería evitar en la medida de lo posible.

Nombre del modelo. El nombre del modelo. El nombre se muestra únicamente a título informativo; no se puede modificar aquí.

Nuggets de modelo de árbol de decisión de Netezza

El nugget de modelo de árbol de decisión muestra los resultados de la operación de modelado y le permite establecer algunas opciones de puntuación del modelo.

Si ejecuta una ruta con un nodo de modelado Árbol de decisión, de forma predeterminada, el nodo añade un campo nuevo, cuyo nombre se deriva del nombre del modelo.

Tabla 16. Campo de puntuación de modelo del árbol de decisión

Nombre del campo añadido	Descripción
$I-*nombre_modelo*	Valor predicho del registro actual.

Si selecciona la opción **Calcular las probabilidades de las clases asignadas para puntuar registros** en el nodo de modelado o en el nugget de modelo y ejecuta la ruta, se añade un campo adicional.

Tabla 17. Campo de puntuación de modelo del árbol de decisión: Adicional

Nombre del campo añadido	Descripción
$IP-*nombre_modelo*	Valor de confianza (entre 0,0 y 1,0) de la predicción.

Nugget del árbol de decisión de Netezza: Pestaña Modelo

La pestaña **Modelo** muestra la importancia del predictor del modelo de árbol de decisión en formato gráfico. La longitud de la barra representa la importancia del predictor.

Nota: Si está trabajando con IBM Netezza Analytics Versión 2.x o anterior, el contenido del modelo de árbol de decisiones se muestra solo en formato de texto.

Para estas versiones, se muestra la información siguiente:

- Cada línea de texto corresponde a un nodo o una hoja.
- La sangría refleja el nivel del árbol.
- Para un nodo, se visualiza la condición de división.
- Para una hoja, se muestra la etiqueta de clase asignada.

Nugget del árbol de decisión de Netezza: Pestaña Configuración

La pestaña configuración le permite establecer algunas opciones para la puntuación para el modelo.

Incluir campos de entrada. Si está seleccionada, esta opción pasa todos los campos de entrada original siguientes, añadiendo los campos o el campo de modelado adicional a cada fila de datos. Si cancela la selección de esta casilla de verificación, sólo el campo ID de registro y los campos de modelados adicionales pasan y así la ruta se ejecuta con más rapidez.

Calcular las probabilidades de las clases asignadas para puntuar registros. (Sólo árboles de decisión y de bayesiano ingenuo) Si está seleccionada, esta opción significa que los campos de modelado adicional incluyen un campo de confianza (es decir, una probabilidad) así como el campo de predicción. Si anula la selección de esta casilla de verificación, sólo se produce el campo de predicción.

Nugget de modelo de K-medias de Netezza

Los nugget de modelo de K-medias contienen toda la información capturada por el modelo de agrupación en clústeres, así como información acerca de los datos de entrenamiento y el proceso de estimación.

Cuando ejecuta una ruta que contiene un nugget de modelo de K-medias, el nodo añade dos nuevos campos que contienen la pertenencia del clúster y la distancia a partir del centro del clúster asignado para ese registro. Del nombre del modelo se derivan los nuevos nombres de campos con el prefijo *$KM-* para la pertenencia del clúster y *$KMD-* para la distancia desde el centro del clúster. Por ejemplo, si el modelo se llama *Kmeans*, los nuevos campos se llamarían *$KM-Kmeans* y *$KMD-Kmeans*.

Nugget de K-medias de Netezza - pestaña Modelo

La pestaña **Modelo** contiene distintas vistas de gráfica que muestran estadísticas de resumen y distribuciones para campos de clústeres. Puede exportar los datos del modelo, o puede exportar la vista como un gráfico.

Si está trabajando con IBM Netezza Analytics Version 2.x o anterior, o si genera el modelo con Mahalanobis como medida de distancia, el contenido de los modelos de K-medias se muestra solo en formato de texto.

Para estas versiones, se muestra la información siguiente:

- **Estadísticos de resumen.** Para el clúster más pequeño y también para el más grande, las estadísticas de resumen muestran el número de registros. Las estadísticas de resumen también muestran el porcentaje del conjunto de datos que es recogido por estos clústeres. La lista muestra también la relación de tamaño del mayor clúster al menor clúster.
- **Resumen de agrupación en clúster.** El resumen de agrupación en clúster lista los clústeres creados por el algoritmo. Para cada clúster, la tabla muestra el número de registros en ese clúster, junto con la distancia media desde el centro del clúster hasta esos registros.

Nugget de K-medias de Netezza: Pestaña Configuración

La pestaña configuración le permite establecer algunas opciones para la puntuación para el modelo.

Incluir campos de entrada. Si está seleccionada, esta opción pasa todos los campos de entrada original siguientes, añadiendo los campos o el campo de modelado adicional a cada fila de datos. Si cancela la selección de esta casilla de verificación, sólo el campo ID de registro y los campos de modelados adicionales pasan y así la ruta se ejecuta con más rapidez.

Medida de distancia. El método que se debe utilizar para medir la distancia entre puntos de datos; una mayor distancia indica mayores diferencias. Las opciones son:

- **Euclídea**. (predeterminada) La distancia entre dos puntos se calcula dibujando una línea recta que los una.
- **Manhattan**. La distancia entre dos puntos se calcula con la suma de las diferencias absolutas entre sus coordenadas.
- **Canberra**. Similar a la distancia Manhattan, pero más receptiva a los puntos de datos más cercanos al origen.
- **Máximo**. La distancia entre dos puntos se calcula como la diferencia mayor en cualquiera de las dimensiones de coordenadas.

Nuggets de modelo de red bayesiana de Netezza

El nugget de modelo de red bayesiana proporciona un método para definir las opciones de puntuación del modelo.

Si ejecuta una ruta con un nodo de modelado Red bayesiana, el nodo añade un campo nuevo, cuyo nombre se deriva del nombre del modelo.

Tabla 18. Campo de puntuación de modelo de red bayesiana

Nombre del campo añadido	Descripción
$BN-*nombre_modelo*	Valor predicho del registro actual.

Puede ver el campo extra si adjunta un nodo Tabla al nugget de modelo y ejecuta el nodo Tabla. Para obtener más información, consulte el tema "Puntuación de modelos de IBM Netezza Analytics" en la página 105.

Nugget de red bayesiana de Netezza: Pestaña Configuración

En la pestaña Configuración puede definir las opciones para puntuar el modelo.

Objetivo. Si desea puntuar un campo de objetivo diferente del objetivo actual, seleccione el nuevo objetivo aquí.

ID de registro. Si no ha especificado ninguna ID de registro, seleccione el campo que utilizará aquí.

Tipo de predicción. La variación del algoritmo de predicción que desea utilizar:

- **La mejor (elemento afín más correlacionado)** (valor predeterminado) Utiliza el nodo vecino más correlacionado.
- **Vecinos (predicción ponderada de vecinos).** Utiliza una predicción ponderada de todos los nodos vecinos.
- **Vecinos NN (vecinos no nulos).** Similar a la opción anterior, salvo que ignora los nodos con valores nulos (es decir, nodos que se corresponden con atributos a los que les faltan valores para la instancia cuya predicción se calcula).

Incluir campos de entrada. Si está seleccionada, esta opción pasa todos los campos de entrada original siguientes, añadiendo los campos o el campo de modelado adicional a cada fila de datos. Si cancela la selección de esta casilla de verificación, sólo el campo ID de registro y los campos de modelados adicionales pasan y así la ruta se ejecuta con más rapidez.

Nuggets de modelo bayesiano ingenuo de Netezza

El nugget de modelo bayesiano ingenuo proporciona un método para definir las opciones de puntuación del modelo.

Si ejecuta una ruta con un nodo de modelado bayesiano ingenuo, de forma predeterminada, el nodo añade un campo nuevo, cuyo nombre se deriva del nombre del modelo.

Tabla 19. Campo de puntuación de modelo bayesiano ingenuo: Opción predeterminada

Nombre del campo añadido	Descripción
$I-*nombre_modelo*	Valor predicho del registro actual.

Si selecciona la opción **Calcular las probabilidades de las clases asignadas para puntuar registros** en el nodo de modelado o en el nugget de modelo y ejecuta la ruta, se añaden dos campos adicionales.

Tabla 20. Campos de puntuación de modelo bayesiano ingenuo: Adicional

Nombre del campo añadido	Descripción
$IP-*nombre_modelo*	El numerador bayesiano de la clase de la instancia (es decir, el producto de las probabilidades previas de clase y las probabilidades de valor de atributo de instancia condicional).
$ILP-*nombre_modelo*	El logaritmo natural del segundo.

Puede ver los campos extra si adjunta un nodo Tabla al nugget de modelo y ejecuta el nodo Tabla. Para obtener más información, consulte el tema "Puntuación de modelos de IBM Netezza Analytics" en la página 105.

Nugget de bayesiano ingenuo de Netezza: Pestaña Configuración

En la pestaña Configuración puede definir las opciones para puntuar el modelo.

Incluir campos de entrada. Si está seleccionada, esta opción pasa todos los campos de entrada original siguientes, añadiendo los campos o el campo de modelado adicional a cada fila de datos. Si cancela la selección de esta casilla de verificación, sólo el campo ID de registro y los campos de modelados adicionales pasan y así la ruta se ejecuta con más rapidez.

Calcular las probabilidades de las clases asignadas para puntuar registros. (Sólo árboles de decisión y de bayesiano ingenuo) Si está seleccionada, esta opción significa que los campos de modelado adicional incluyen un campo de confianza (es decir, una probabilidad) así como el campo de predicción. Si anula la selección de esta casilla de verificación, sólo se produce el campo de predicción.

Mejorar la precisión de probabilidad de los conjuntos de datos pequeños o no equilibrados. Cuando se calculan probabilidades, esta opción activa la técnica de estimación *m-* para evitar probabilidades de cero durante el cálculo. Este tipo de estimación de probabilidades puede ser más lenta, pero proporciona mejores resultados para conjuntos de datos pequeños o no equilibrados.

Nuggets de modelo KNN de Netezza

El nugget de modelo KNN proporciona un método para definir las opciones de puntuación del modelo.

Si ejecuta una ruta con un nodo de modelado KNN, el nodo añade un campo nuevo, cuyo nombre se deriva del nombre del modelo.

Tabla 21. Campo de puntuación de modelo de KNN

Nombre del campo añadido	Descripción
$KNN-*nombre_modelo*	Valor predicho del registro actual.

Puede ver el campo extra si adjunta un nodo Tabla al nugget de modelo y ejecuta el nodo Tabla. Para obtener más información, consulte el tema "Puntuación de modelos de IBM Netezza Analytics" en la página 105.

Nugget KNN de Netezza: Pestaña Configuración

En la pestaña Configuración puede definir las opciones para puntuar el modelo.

Medida de distancia. El método que se debe utilizar para medir la distancia entre puntos de datos; una mayor distancia indica mayores diferencias. Las opciones son:

- **Euclídea**. (predeterminada) La distancia entre dos puntos se calcula dibujando una línea recta que los una.
- **Manhattan**. La distancia entre dos puntos se calcula con la suma de las diferencias absolutas entre sus coordenadas.
- **Canberra**. Similar a la distancia Manhattan, pero más receptiva a los puntos de datos más cercanos al origen.
- **Máximo**. La distancia entre dos puntos se calcula como la diferencia mayor en cualquiera de las dimensiones de coordenadas.

Número de vecinos más próximos (k) El número de vecinos más próximos de un caso concreto. Tenga en cuenta que el uso de un número mayor de vecinos no implica que el modelo resultante sea más preciso.

La elección de *k* controla el equilibrio entre la prevención del sobreajuste (puede ser importante, especialmente para datos "con ruido") y resolución (produciendo predicciones diferentes para instancias similares). Normalmente tendrá que ajustar el valor de *k* en cada conjunto de datos, con los valores típicos comprendidos entre 1 y varias docenas.

Incluir campos de entrada. Si está seleccionada, esta opción pasa todos los campos de entrada original siguientes, añadiendo los campos o el campo de modelado adicional a cada fila de datos. Si cancela la selección de esta casilla de verificación, sólo el campo ID de registro y los campos de modelados adicionales pasan y así la ruta se ejecuta con más rapidez.

Estandarizar mediciones antes de calcular la distancia. Si está seleccionada, esta opción estandariza las mediciones de campos de entrada continuos antes de calcular los valores de distancia.

Utilizar conjuntos principales para incrementar el rendimiento de los conjuntos de datos grandes. Si está seleccionada, esta opción utiliza el muestreo del conjunto principal para acelerar el cálculo de conjuntos de datos grandes.

Nuggets de modelo de clúster divisivo de Netezza

El nugget de modelo de agrupación en clústeres divisivo proporciona un método para definir las opciones de puntuación del modelo.

Si ejecuta una ruta con un nodo de modelado de agrupación en clústeres divisivo, el nodo añade dos campos nuevos, cuyos nombres se derivan del nombre del modelo.

Tabla 22. Campos de puntuación de modelo de clúster divisivo

Nombre del campo añadido	Descripción
$DC-*nombre_modelo*	Identificador del subclúster al que se asigna el registro actual.
$DCD-*nombre_modelo*	Distancia del centro del subclúster del registro actual.

Puede ver los campos extra si adjunta un nodo Tabla al nugget de modelo y ejecuta el nodo Tabla. Para obtener más información, consulte el tema "Puntuación de modelos de IBM Netezza Analytics" en la página 105.

Nugget de clúster divisivo de Netezza: Pestaña Configuración

En la pestaña Configuración puede definir las opciones para puntuar el modelo.

Incluir campos de entrada. Si está seleccionada, esta opción pasa todos los campos de entrada original siguientes, añadiendo los campos o el campo de modelado adicional a cada fila de datos. Si cancela la selección de esta casilla de verificación, sólo el campo ID de registro y los campos de modelados adicionales pasan y así la ruta se ejecuta con más rapidez.

Medida de distancia. El método que se debe utilizar para medir la distancia entre puntos de datos; una mayor distancia indica mayores diferencias. Las opciones son:

- **Euclídea**. (predeterminada) La distancia entre dos puntos se calcula dibujando una línea recta que los una.
- **Manhattan**. La distancia entre dos puntos se calcula con la suma de las diferencias absolutas entre sus coordenadas.
- **Canberra**. Similar a la distancia Manhattan, pero más receptiva a los puntos de datos más cercanos al origen.
- **Máximo**. La distancia entre dos puntos se calcula como la diferencia mayor en cualquiera de las dimensiones de coordenadas.

Nivel de jerarquía aplicado. El nivel de jerarquía que se debe aplicar a los datos.

Nuggets de modelo PCA de Netezza

El nugget de modelo PCA proporciona un método para definir las opciones de puntuación del modelo.

Si ejecuta una ruta con un nodo de modelado PCA, de forma predeterminada el nodo añade un campo nuevo, cuyo nombre se deriva del nombre del modelo.

Tabla 23. Campo de puntuación de modelo de PCA

Nombre del campo añadido	Descripción
$F-*nombre_modelo*	Valor predicho del registro actual.

Si especifica un valor mayor que 1 en el campo **Número de componentes principales...** en el nodo de modelado o en el nugget de modelo y ejecuta la ruta, el nodo añade un campo nuevo para cada componente. En este caso, los nombres de los campos están precedidos por el sufijo *-n*, donde *n* es el número del componente. Por ejemplo, si el modelo se denomina *pca* y contiene tres componentes, los nuevos campos se llamarían *$F-pca-1*, *$F-pca-2* y *$F-pca-3*.

Puede ver los campos extra si adjunta un nodo Tabla al nugget de modelo y ejecuta el nodo Tabla. Para obtener más información, consulte el tema "Puntuación de modelos de IBM Netezza Analytics" en la página 105.

Nugget PCA de Netezza: Pestaña Configuración

En la pestaña Configuración puede definir las opciones para puntuar el modelo.

Número de componentes principales que se utilizarán en la proyección. El número de componentes principales a los que desea reducir el conjunto de datos. Este valor no debe superar el número de atributos (campos de entrada).

Incluir campos de entrada. Si está seleccionada, esta opción pasa todos los campos de entrada original siguientes, añadiendo los campos o el campo de modelado adicional a cada fila de datos. Si cancela la selección de esta casilla de verificación, sólo el campo ID de registro y los campos de modelados adicionales pasan y así la ruta se ejecuta con más rapidez.

Nuggets de modelo de árboles de regresión de Netezza

El nugget de modelo de árbol de regresión proporciona un método para definir las opciones de puntuación del modelo.

Si ejecuta una ruta con un nodo de modelado Árbol de regresión, de forma predeterminada el nodo añade un campo nuevo, cuyo nombre se deriva del nombre del modelo.

Tabla 24. Campo de puntuación de modelo del árbol de regresión

Nombre del campo añadido	Descripción
$I-*nombre_modelo*	Valor predicho del registro actual.

Si selecciona la opción **Calcular varianza estimada** en el nodo de modelado o en el nugget de modelo y ejecuta la ruta, se añade un campo adicional.

Tabla 25. Campo de puntuación de modelo del árbol de regresión: Adicional

Nombre del campo añadido	Descripción
$IV-*nombre_modelo*	Varianzas estimadas de clases asignadas.

Puede ver los campos extra si adjunta un nodo Tabla al nugget de modelo y ejecuta el nodo Tabla. Para obtener más información, consulte el tema “Puntuación de modelos de IBM Netezza Analytics” en la página 105.

Nugget de árbol de regresión de Netezza - pestaña Modelo

La pestaña **Modelo** muestra la importancia del predictor del modelo de árbol de regresión en formato gráfico. La longitud de la barra representa la importancia del predictor.

Nota: Si está trabajando con IBM Netezza Analytics Versión 2.x o anterior, el contenido del modelo de árbol de regresión se muestra solo en formato de texto.

Para estas versiones, se muestra la información siguiente:

- Cada línea de texto corresponde a un nodo o una hoja.
- La sangría refleja el nivel del árbol.
- Para un nodo, se visualiza la condición de división.
- Para una hoja, se muestra la etiqueta de clase asignada.

Nugget del árbol de regresión de Netezza: Pestaña Configuración

En la pestaña Configuración puede definir las opciones para puntuar el modelo.

Incluir campos de entrada. Si está seleccionada, esta opción pasa todos los campos de entrada original siguientes, añadiendo los campos o el campo de modelado adicional a cada fila de datos. Si cancela la

selección de esta casilla de verificación, sólo el campo ID de registro y los campos de modelados adicionales pasan y así la ruta se ejecuta con más rapidez.

Calcular varianza estimada. Indica si las varianzas de las clases asignadas se deben incluir en los resultados.

Nuggets de modelo de regresión lineal de Netezza

El nugget de modelo de regresión lineal proporciona un método para definir las opciones de puntuación del modelo.

Si ejecuta una ruta con un nodo de modelado con regresión lineal, el nodo añade un campo nuevo, cuyo nombre se deriva del nombre del modelo.

Tabla 26. Campo de puntuación de modelo de regresión lineal

Nombre del campo añadido	Descripción
$LR-*nombre_modelo*	Valor predicho del registro actual.

Nugget de regresión lineal de Netezza: Pestaña Configuración

En la pestaña Configuración puede definir las opciones para puntuar el modelo.

Incluir campos de entrada. Si está seleccionada, esta opción pasa todos los campos de entrada original siguientes, añadiendo los campos o el campo de modelado adicional a cada fila de datos. Si cancela la selección de esta casilla de verificación, sólo el campo ID de registro y los campos de modelados adicionales pasan y así la ruta se ejecuta con más rapidez.

Nugget de modelo Series temporales de Netezza

El nugget de modelo permite acceder al resultado de la operación de modelado de series temporales. El resultado consta de los siguientes campos.

Tabla 27. Campos de salida del modelo Series temporales

Campo	Descripción
TSID	El identificador de la serie temporal; el contenido del campo especificado para los ID de series temporales en la pestaña Campos del nodo de modelado. Para obtener más información, consulte el tema "Opciones de campos de Series temporales de Netezza" en la página 99.
HORA	El período de tiempo dentro de la serie temporal actual.
HISTÓRICO	Los valores de datos históricos (los utilizados para realizar la predicción). Este campo se incluye únicamente si la opción **Incluya valores históricos en el resultado** está seleccionada en la pestaña Configuraciones del nugget de modelo.
$TS-INTERPOLADO	Los valores interpolados, cuando se utilicen. Este campo se incluye únicamente si la opción **Incluya valores interpolados en el resultado** está seleccionada en la pestaña Configuraciones del nugget de modelo. La interpolación es una opción de la pestaña Opciones de generación del nodo de modelado.
$TS-FORECAST	Los valores de previsión de la serie temporal.

Para ver el resultado de modelo, adjunte un nodo Tabla (de la pestaña Resultados de la paleta de nodos) al nugget de modelo y ejecute el nodo Tabla.

Nugget Series temporales de Netezza: Pestaña Configuración

En la pestaña Configuración puede especificar opciones para personalizar el resultado de modelo.

Nombre del modelo. El nombre del modelo, como está especificado en la pestaña Opciones del modelo del nodo de modelado.

Las otras opciones son las mismas que las de la pestaña Opciones de modelo del nodo de modelado.

Nugget de modelo lineal generalizado de Netezza

El nugget de modelo permite acceder al resultado de la operación de modelado.

Si ejecuta una ruta con un nodo de modelado lineal generalizado, el nodo añade un campo nuevo, cuyo nombre se deriva del nombre del modelo.

Tabla 28. Campo de puntuación de modelo de lineal generalizado

Nombre del campo añadido	Descripción
$GLM-*nombre_modelo*	Valor predicho del registro actual.

La pestaña Modelo muestra varios estadísticos relacionados con el modelo.

El resultado consta de los siguientes campos.

Tabla 29. Campos de resultados del modelo lineal generalizado

Campo de resultado	Descripción
Parameter	Los parámetros (esto es, las variables predictivas) utilizadas por el modelo. Son columnas numéricas y nominales, así como la interceptación (el término constante en el modelo de regresión).
Beta	El coeficiente de correlación (esto es, el componente lineal del modelo).
Error Std	La desviación estándar de beta.
Test	Los estadísticos de prueba utilizados para evaluar la validez del parámetro.
valor p	La probabilidad de error si se asume que el parámetro es significativo.
Resumen de residuos	
Tipo de residuo	El tipo de residuo de la predicción para el que se muestran los valores de resumen.
RSS	El valor del residuo.
df	Los grados de libertad del residuo.
valor p	La probabilidad de error. Un valor alto indica un modelo con pocos ajustes; un valor bajo indica un buen ajuste.

Nugget de modelo lineal generalizado de Netezza: Pestaña Configuración

En la pestaña Configuración puede personalizar el resultado de modelo.

La opción es la misma que la mostrada para Opciones de puntuación en el nodo de modelado. Para obtener más información, consulte el tema “Opciones de modelo lineal generalizado de Netezza: Opciones de puntuación” en la página 105.

Noticias

Esta información se ha desarrollado para los productos y servicios ofrecidos en todo el mundo.

Es posible que IBM no ofrezca los productos, servicios o características que se tratan en este documento en otros países. Consulte al representante de IBM de su zona para obtener información acerca de los productos y servicios que están actualmente disponibles en su zona. Las referencias hechas a un producto, programa o servicio IBM no pretenden afirmar ni dar a entender que sólo se puede utilizar dicho producto, programa o servicio IBM. Se puede utilizar en su lugar cualquier producto, programa o servicio funcionalmente equivalente que no vulnere ningún derecho de propiedad intelectual de IBM. No obstante, es responsabilidad del usuario evaluar y verificar el funcionamiento de cualquier producto, programa o servicio que no sea de IBM.

IBM puede tener patentes o solicitudes de patentes pendientes que cubran el tema principal descrito en este documento. El abastecimiento de este documento no le otorga ninguna licencia sobre dichas patentes. Puede enviar consultas sobre licencias, por escrito, a:

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

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

Intellectual Property Licensing
Ley de propiedad intelectual y jurídica
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502, Japón

El párrafo siguiente no se aplica al Reino Unido ni a ningún otro país donde estas disposiciones sean incompatibles con la legislación vigente: INTERNATIONAL BUSINESS MACHINES CORPORATION PROPORCIONA ESTA PUBLICACIÓN "TAL CUAL" SIN GARANTÍAS DE NINGÚN TIPO, NI EXPLÍCITAS NI IMPLÍCITAS, INCLUYENDO PERO NO LIMITÁNDOSE A ELLAS, LAS GARANTÍAS IMPLÍCITAS DE NO INFRACCIÓN DE DERECHOS DE TERCEROS, COMERCIALIZACIÓN O IDONEIDAD PARA UN FIN CONCRETO. Algunas legislaciones no contemplan la declaración de limitación de responsabilidad, ni implícita ni explícita, en determinadas transacciones, por lo que cabe la posibilidad de que esta declaración no sea aplicable en su caso.

Esta información puede contener imprecisiones técnicas o errores tipográficos. La información aquí contenida está sometida a cambios periódicos; tales cambios se irán incorporando en nuevas ediciones de la publicación. IBM puede realizar en cualquier momento mejoras o cambios en los productos o programas descritos en esta publicación sin previo aviso.

Las referencias en este documento a sitios web que no sean de IBM se proporcionan únicamente como ayuda y no se consideran en modo alguno como una recomendación por parte de IBM de dichos sitios web. El contenido de esos sitios Web no forma parte del contenido del presente producto de IBM y la utilización de esos sitios Web corre a cuenta y riesgo del usuario.

IBM puede utilizar o distribuir cualquier información que proporcione en la forma que considere adecuada sin incurrir en ninguna obligación con el usuario.

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

Tel. 900 100 400
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
Estados Unidos

Dicha información puede estar disponible, sujeta a los términos y condiciones correspondientes, incluyendo, en algunos casos, el pago de una tarifa.

IBM proporciona el programa bajo licencia descrito en esta información y todo el material bajo licencia disponible para él de acuerdo con los términos del Contrato del cliente de IBM, el Acuerdo internacional de licencia de programas de IBM o cualquier contrato equivalente existente entre las partes.

Los datos de rendimiento incluidos en este documento se han obtenido en un entorno controlado. Por tanto, los resultados obtenidos en otros entornos operativos pueden variar significativamente. Es posible que algunas medidas se hayan desarrollado con sistemas en nivel de desarrollo y no hay garantía de que estas medidas sean las mismas en sistemas con disponibilidad general. Es más, es posible que la estimación de algunas medidas se haya realizado mediante extrapolación. Puede que los resultados reales varíen. Los usuarios de esta documentación deberían verificar los datos aplicables a su entorno específico.

La información relacionada con los productos que no son de IBM se ha obtenido de los proveedores de dichos productos, sus anuncios publicados u otras fuentes de disponibilidad pública. IBM no ha probado esos productos y no puede confirmar la exactitud del rendimiento, de la compatibilidad ni de ninguna otra declaración relacionada con productos que no sean de IBM. Las preguntas sobre las funciones de los productos que no son de IBM se deben dirigir a los proveedores de estos productos.

Todas las declaraciones sobre futuras tendencias o intenciones de IBM están sujetas a modificación o retirada sin previo aviso y representan únicamente metas y objetivos.

Esta información contiene ejemplos de datos e informes utilizados en operaciones empresariales cotidianas. Para ilustrarlos de la forma más completa posible, incluyen los nombres de personas, empresas, marcas y productos. Todos estos nombres son ficticios y cualquier similitud con nombres y direcciones utilizados por una empresa real son mera coincidencia.

Si está viendo esta información en copia software, es posible que las fotografías e ilustraciones a color no aparezcan.

Marcas registradas

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

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

Linux es una marca registrada de Linus Torvalds en EE.UU. y/o en otros países.

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

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

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

Otros nombres de productos y servicios pueden ser marcas registradas de IBM o de otras empresas.

Índice

K

L

M

N

O

P

R

S

T

U

V

Impreso en España