Guía de IBM SPSS Modeler 18.4 In-Database Mining

Nota

Antes de utilizar esta información y el producto al que da soporte, lea la información del apartado "Avisos" en la página 99.

Información del producto

Esta edición se aplica a la versión 18, release 4, modificación 0 de IBM® SPSS Modeler y a todos los releases y modificaciones posteriores hasta que se indique lo contrario en nuevas ediciones.

© Copyright International Business Machines Corporation .

Contenido

Prefacio ..... **vii**

Capítulo 1. Acerca deIBM SPSS Modeler **1**
Productos IBM SPSS Modeler 1
IBM SPSS Modeler 1
IBM SPSS Modeler Server 1
IBM SPSS Modeler Administration Console 2
IBM SPSS Modeler Batch 2
IBM SPSS Modeler Solution Publisher 2
Adaptadores de IBM SPSS Modeler Server para IBM SPSS Collaboration and Deployment Services 2
Ediciones de IBM SPSS Modeler 2
Documentación 3
Documentación de SPSS Modeler Professional 3
Documentación de SPSS Modeler Premium 4
Ejemplos de aplicación 4
Carpeta Demos 4
Rastreo de licencias 4

Capítulo 2. Minería interna de bases de datos **5**
Conceptos básicos del modelado de bases de datos 5
Qué necesita 5
Generación de modelos 6
Preparación de los datos 6
Puntuación de modelos 6
Exportar y guardar modelos de base de datos 7
Coherencia de modelos 7
Presentación y exportación del SQL generado 7

Capítulo 3. Modelado de bases de datos con Microsoft Analysis Services **9**
IBM SPSS Modeler y Microsoft Analysis Services 9
Requisitos para la integración con Microsoft Analysis Services 10
Activación de la integración con Analysis Services 11
Generación de modelos con Analysis Services 13
Gestión de modelos de Analysis Services 13
Configuración común a todos los nodos de algoritmo 15
Opciones de Experto para los árboles de decisión de MS 16
Opciones de Experto para los clústeres de MS 16
Opciones de experto para el bayesiano ingenuo de MS 16
Opciones de experto de regresión lineal de MS 16
Opciones de experto de red neuronal de MS 16
Opciones de experto de regresión logística de MS 16
Nodo Reglas de asociación de MS 16
Nodo Serie temporal de MS 17
Nodo Agrupación en clústeres de secuencias de MS 18
Puntuación de modelos de Analysis Services 19
Configuración común a todos los modelos de Analysis Services 20
Nugget de modelo Serie temporal de MS 21
Nugget de modelo de clúster de secuencias de MS 22
Exportación de modelos y generación de nodos 22

Ejemplos de minería de datos con Analysis Services ... 22
Rutas de ejemplo: Árboles de decisión ... 22

Capítulo 4. Modelado de bases de datos con Oracle Data Mining ... 25
Acerca de Oracle Data Mining ... 25
Requisitos para la integración con Oracle ... 25
Activación de la integración con Oracle ... 26
Generación de modelos con Oracle Data Mining ... 27
Opciones del servidor de modelos de Oracle ... 28
Costes de clasificación errónea ... 28
Bayesiano ingenuo de Oracle ... 29
Opciones del modelo bayesiano ingenuo ... 29
Opciones de experto para el bayesiano ingenuo ... 29
Bayesiano adaptativo de Oracle ... 30
Opciones del modelo bayesiano adaptativo ... 30
Opciones de experto para el bayesiano adaptativo ... 31
Máquina de vectores de soporte de Oracle (SVM) ... 31
Opciones del modelo SVM ... 31
Opciones de experto de SVM ... 32
Opciones de ponderaciones de SVM Oracle ... 32
Modelos lineales generalizados de Oracle (GLM) ... 33
Opciones del modelo GLM ... 33
Opciones de experto de GLM ... 34
Opciones de ponderaciones de GLM Oracle ... 34
Árbol de decisión de Oracle ... 34
Opciones de modelo para los árboles de decisión ... 35
Opciones de Experto para los árboles de decisión ... 35
O-clúster de Oracle ... 36
Opciones de Modelo para O-clúster ... 36
Opciones de Experto para O-clúster ... 36
K-medias de Oracle ... 36
Opciones de Modelo para K-medias ... 37
Opciones de Experto para K-medias ... 37
Factorización de matrices no negativas (NMF) de Oracle ... 37
Opciones de Modelo para NMF ... 38
Opciones de Experto para NMF ... 38
Apriori de Oracle ... 38
Opciones de los campos Apriori ... 39
Opciones de Modelo para Apriori ... 39
Longitud mínima de la descripción de Oracle (LMD) ... 40
Opciones de Modelo para LMD ... 40
Importancia del atributo de Oracle (AI) ... 41
Opciones de modelo de AI ... 41
Opciones de selección de AI ... 41
Pestaña Modelo de nugget de modelo de AI ... 41
Gestión de modelos de Oracle ... 42
Pestaña Servidor del nugget de modelo de Oracle ... 42
Pestaña Resumen del nugget de modelo de Oracle ... 42
Pestaña Configuración del nugget de modelo de Oracle ... 43
Enumeración de modelos de Oracle ... 43
Oracle Data Miner ... 43
Preparación de los datos ... 44
Ejemplos de Oracle Data Mining ... 44
Ruta de ejemplo: cargar datos ... 45
Ruta de ejemplo: Explorar datos ... 45
Ruta de ejemplo: Generar modelo ... 45
Ruta de ejemplo: Evaluar modelo ... 46

Ruta de ejemplo: Desplegar modelo..... 46

Capítulo 5. Modelado de bases de datos con IBM Data Warehouse y IBM Netezza Analytics..... 47

SPSS Modeler con IBM Data Warehouse y IBM Netezza Analytics..... 47
Requisitos de integración..... 47
Habilitación de la integración..... 48
Configuración de IBM Netezza Analytics o IBM Data Warehouse..... 48
Creación de un origen ODBC para IBM Netezza Analytics..... 48
Habilitación de la integración en SPSS Modeler..... 50
Activación de optimización y generación de SQL..... 50
Creación de modelos con IBM Netezza Analytics e IBM Data Warehouse..... 50
Opciones de campos..... 51
Opciones de servidor..... 52
Opciones de modelo..... 52
Gestión de modelos..... 53
Listado de modelos de base de datos..... 53
Árbol de regresión de IBM Data WH..... 53
Opciones de generación de árbol de regresión de IBM Data WH: Crecimiento del árbol..... 53
Opciones de generación de árbol de IBM Data WH: Poda de árbol..... 54
Clúster divisivo de Netezza..... 55
Opciones de campo de clúster divisivo de Netezza..... 55
Opciones de generación de clúster divisivo de Netezza..... 55
Lineal generalizado de IBM Data WH..... 56
Opciones de campo de modelo lineal generalizado de IBM Data WH..... 56
Opciones de modelo lineal generalizado de IBM Data WH: Generales..... 57
Opciones de modelo lineal generalizado de IBM Data WH: Interacciones..... 58
Opciones de modelo lineal generalizado de IBM Data WH: Opciones de puntuación..... 59
Árboles de decisión de IBM Data WH..... 59
Ponderaciones de instancia y ponderaciones de clase..... 59
Opciones de campo para los árboles de decisión de Netezza..... 60
Opciones de generación de árbol de decisión de IBM Data WH..... 61
Regresión lineal de IBM Data WH..... 62
Opciones de generación de regresión lineal de IBM Data WH..... 62
KNN de IBM Data WH..... 63
Opciones de modelo KNN de IBM Data WH: Generales..... 63
Opciones de modelo KNN de IBM Data WH: Opciones de puntuación..... 64
K-medias de IBM Data WH..... 64
Opciones de campo de k-medias de IBM Data WH..... 64
Pestaña Opciones de generación de k-medias de IBM Data WH..... 65
Bayesiano ingenuo de IBM Data WH..... 65
Red bayesiana de Netezza..... 66
Opciones de campo de red bayesiana de Netezza..... 66
Opciones de generación de red bayesiana de Netezza..... 66
Series temporales de Netezza..... 66
Interpolación de valores en Series temporales de Netezza..... 67
Opciones de campos de Series temporales de Netezza..... 69
Opciones de generación de Series temporales de Netezza..... 69
Opciones del modelo Series temporales de Netezza..... 72
Biétapico de IBM Data WH..... 72
Opciones de campo biétapico de IBM Data WH..... 72
Opciones de generación bietápica de IBM Data WH..... 73
PCA de IBM Data WH..... 73
Opciones de campo de PCA de IBM Data WH..... 73
Opciones de generación de PCA de IBM Data WH..... 74
Gestión de modelos de IBM Data WH y de Netezza..... 74
Puntuación de modelos IBM Data Warehouse y IBM Netezza Analytics..... 74

Pestaña Servidor del nugget de modelo de IBM Data WH y Netezza....75
Nuggets de modelo de árbol de decisión de IBM Data WH....75
Nugget de modelo de k-medias de IBM Data WH....76
Nuggets de modelo de red bayesiana de Netezza....77
Nuggets de modelo bayesiano ingenuo de IBM Data WH....78
Nuggets de modelo KNN de IBM Data WH....78
Nuggets de modelo de clúster divisivo de Netezza....79
Nuggets de modelo PCA de IBM Data WH....80
Nuggets de modelo de árboles de regresión de Netezza....81
Nuggets de modelo de regresión lineal de IBM Data WH....82
Nugget de modelo Series temporales de Netezza....82
Nugget de modelo lineal generalizado de IBM Data WH....83
Nugget de modelo bietápico de IBM Data WH....83

Capítulo 6. Modelador de bases de datos con IBM Db2 para z/OS....85
IBM SPSS Modeler e IBM Db2 para z/OS....85
Requisitos para la integración con IBM Db2 para z/OS....85
Habilitar la integración con IBM Db2 Analytics Accelerator for z/OS....85
Configurar IBM Db2 para z/OS e IBM Analytics Accelerator for z/OS....86
Crear un origen ODBC para IBM Db2 para z/OS e IBM Db2 Analytics Accelerator....86
Habilitación de la integración de IBM Db2 para z/OS en IBM SPSS Modeler....86
Activación de optimización y generación de SQL....87
Configuración de DSN utilizando el cliente de IBM Db2 en IBM SPSS Modeler....87
Crear modelos con IBM Db2 para z/OS....88
Modelos de IBM Db2 para z/OS - Opciones de campo....89
Modelos de IBM Db2 para z/OS - Opciones de servidor....89
Modelos de IBM Db2 para z/OS - Opciones Modelo....89
Modelos de IBM Db2 para z/OS - k-medias....89
Modelos de IBM Db2 para z/OS - Opciones de campo de k-medias....90
Modelos de IBM Db2 para z/OS - Opciones de generación de k-medias....90
Modelos de IBM Db2 para z/OS - Bayesiano ingenuo....91
Modelos de IBM Db2 para z/OS - Árboles de decisiones....91
Modelos de IBM Db2 para z/OS - Opciones de campo de árbol de decisiones....91
Modelos de IBM Db2 para z/OS - Opciones de generación de árbol de decisiones....91
Modelos de IBM Db2 para z/OS - Nodo de árbol de decisiones - Ponderaciones de clase....92
Modelos de IBM Db2 para z/OS - Nodo de árbol de decisiones - Poda de árbol....92
Modelos de IBM Db2 para z/OS - Árbol de regresión....93
Modelos de IBM Db2 para z/OS - Opciones de generación de árbol de regresión - Crecimiento de árbol....93
Modelos de IBM Db2 para z/OS - Opciones de generación de árbol de regresión - Poda de árbol....94
Modelos de IBM Db2 para z/OS - Dos fases....94
Modelos de IBM Db2 para z/OS - Opciones de campo bietápico....95
Modelos de IBM Db2 para z/OS - Opciones de generación de dos fases....95
Nugget de modelos bietápicos de IBM Db2 para z/OS - Pestaña Modelo....95
Gestionar modelos de IBM Db2 para z/OS....95
Puntuar modelos de IBM Db2 para z/OS....96
Nuggets de modelo de árbol de decisiones de IBM Db2 para z/OS....96
Nugget de modelo de k-medias de IBM Db2 para z/OS....97
Nuggets de modelo de bayesiano ingenuo de IBM Db2 para z/OS....97
Nuggets de modelo de árbol de regresión de IBM Db2 para z/OS....97
Nugget de modelo bietápico de IBM Db2 para z/OS....98

Avisos....99
Marcas comerciales....100
Términos y condiciones de la documentación de producto....100

Índice....103

Prefacio

IBM SPSS Modeler es el entorno de trabajo de minería de datos de nivel de empresa de IBM. 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 los usuarios a aplicar su experiencia empresarial específica, 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 predicciones, clasificaciones, segmentación y algoritmos de detección de asociaciones. Una vez que se crean los modelos, IBM SPSS Modeler Solution Publisher permite su distribución en toda la empresa a los encargados de tomar las decisiones o a una base de datos.

Acerca de IBM Business Analytics

IBM Business Analytics proporciona información completa, coherente y precisa en la que confían para mejorar el rendimiento de su negocio quienes toman las decisiones. 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 estos documentos, IBM SPSS Predictive Analytics ayuda a las organizaciones a predecir situaciones futuras y a actuar de forma proactiva con esa información para mejorar sus resultados. Los clientes comerciales, gubernamentales y académicos de todo el mundo confían en la tecnología 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 de IBM SPSS en sus operaciones diarias, las organizaciones se convierten en empresas predictivas, capaces de dirigir y automatizar decisiones para alcanzar los objetivos comerciales y lograr una ventaja considerable sobre la competencia. Para obtener más información o ponerse en contacto con un representante, visite http://www.ibm.com/spss.

Soporte técnico

Hay asistencia técnica disponible para los clientes de mantenimiento. Los clientes pueden ponerse en contacto con el Soporte técnico si desean recibir ayuda al utilizar productos de IBM o para la instalación en uno de los entornos de hardware soportados. Para ponerse en contacto con el servicio de asistencia técnica, consulte el sitio web de IBM 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 deIBM 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 https://www.ibm.com/analytics/us/en/technology/spss/.

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 (incluido con el IBM SPSS Deployment Manager)
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- Adaptadores de IBM SPSS Modeler Server para IBM 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 esta forma, SPSS Modeler Server proporciona un rendimiento superior en grandes conjuntos de datos, ya que las operaciones que requieren un uso intensivo de la memoria pueden realizarse en el servidor sin tener que descargar los datos en el equipo del cliente. IBM SPSS Modeler Server también

ofrece soporte para las prestaciones de optimización de SQL y modelado en la base de datos, lo que ofrece ventajas adicionales en el rendimiento y la automatización.

IBM SPSS Modeler Administration Console

El Modeler Administration Console región propietaria del archivos una interfaz gráfica de usuario para gestionar muchas de las opciones de configuración de SPSS Modeler Server, que también se pueden configurar a través de un archivo de opciones. La consola se incluye en el IBM SPSS Deployment Manager, se puede utilizar para supervisar y configurar las instalaciones de SPSS Modeler Server y está disponible de forma 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 prestaciones de análisis de SPSS Modeler sin acceso a la interfaz de usuario habitual. Es necesario disponer 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 - Scoring, para el cual se requiere una licencia separada. Con esta licencia, recibirá SPSS Modeler Solution Publisher Runtime, que le permite ejecutar las rutas publicadas.

Para obtener más información sobre SPSS Modeler Solution Publisher, consulte la documentación de IBM SPSS Collaboration and Deployment Services. La documentación de IBM de IBM SPSS Collaboration and Deployment Services contiene secciones denominadas "IBM SPSS Modeler Solution Publisher" e "IBM SPSS Analytics Toolkit."

Adaptadores de IBM SPSS Modeler Server para IBM 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 trabajar con datos especializados y con datos de texto no estructurado. SPSS Modeler Premium incluye IBM SPSS Modeler Text Analytics:

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.

IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription proporciona todas las mismas prestaciones de análisis predictivo que el cliente tradicional de IBM SPSS Modeler. Con la edición Subscription, puede descargar actualizaciones de producto con regularidad.

Documentación

La documentación está disponible en el menú **Ayuda** de SPSS Modeler. De esta forma se abre la documentación de IBM en línea, que está siempre disponible fuera del producto.

La documentación completa de cada producto (incluidas las instrucciones de instalación) también está disponible en formato PDF en https://www.ibm.com/support/pages/spss-modeler-184-documentation.

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 para utilizar SPSS Modeler, incluyendo cómo crear rutas de datos, manejar valores perdidos, crear expresiones de CLEM, trabajar con proyectos e informes y empaquetar rutas para el despliegue en IBM SPSS Collaboration and Deployment Services o IBM SPSS Modeler Advantage.
- **Nodos Origen, Proceso y Resultados de IBM SPSS Modeler.** Descripciones de todos los nodos utilizados para leer, procesar y dar salida a datos en diferentes formatos. En la práctica, esto implica todos los nodos que no sean nodos de modelado.
- **Nodos de modelado de IBM SPSS Modeler.** Descripciones de todos los nodos utilizados para crear modelos de minería de datos. IBM SPSS Modeler ofrece una gran variedad de métodos de modelado procedentes del aprendizaje automático, la inteligencia artificial y el estadístico.
- **Guía de aplicaciones de IBM SPSS Modeler.** Los ejemplos de esta guía ofrecen introducciones breves y concisas a métodos y técnicas de modelado específicos. También está disponible una versión en línea de esta guía desde el menú Ayuda. Para obtener más información, consulte el tema “Ejemplos de aplicación” en la página 4.
- **Automatización y scripts Python IBM SPSS Modeler.** Información sobre la automatización del sistema mediante scripts de Python, incluidas las propiedades que se pueden utilizar para manipular nodos y rutas.
- **Guía de despliegue de IBM SPSS Modeler.** La información sobre cómo ejecutar rutas de IBM SPSS Modeler como pasos en el proceso de trabajos en IBM SPSS Deployment Manager.
- **Guía de minería interna de base 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.

- **Guía del usuario del IBM SPSS Deployment Manager.** Información sobre cómo utilizar la interfaz de usuario de la consola de administración incluida en la aplicación Gestor de despliegue para supervisar y configurar IBM SPSS Modeler Server.
- **Guía CRISP-DM de IBM SPSS Modeler.** Manual que explica paso a paso cómo utilizar la metodología de CRISP-DM en la minería de datos con SPSS Modeler.
- **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.

- **Manual del usuario de SPSS Modeler Text Analytics .** 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.

Ejemplos de aplicación

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 que se utilizan aquí son mucho más pequeños que los enormes almacenes de datos gestionados por algunos analizadores de datos, pero los conceptos y métodos implicados son escalables a las aplicaciones reales.

Para acceder a los ejemplos, pulse **Ejemplos de aplicación** en el menú Ayuda en SPSS Modeler.

Los archivos de datos y vía de accesos de ejemplo se instalan en la carpeta `Demos` en el directorio de instalación del producto. Para obtener más información, consulte “Carpeta Demos” en la página 4.

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

Ejemplos de scripts. Consulte los ejemplos en la *Guía de minería interna de bases de datos de IBM SPSS Modeler*.

Carpeta Demos

Los archivos de datos y las corrientes de ejemplo que se utilizan con los ejemplos de aplicación se instalan en la carpeta `Demos` bajo el directorio de instalación del producto (por ejemplo: `C:\Program Files\IBM\SPSS\Modeler\<version>\Demos`). 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 del cuadro de diálogo **Archivo** > **Corriente abierta** .

Rastreo de licencias

Cuando se utiliza SPSS Modeler, el uso de las licencias se rastrea y registra a intervalos regulares. Las métricas de licencia que se registran son *AUTHORIZED_USER* y *CONCURRENT_USER*, y el tipo de métrica que se registra depende del tipo de licencia que tiene para SPSS Modeler.

IBM License Metric Tool puede procesar los archivos de registro que se generan, a partir de los cuales puede crear informes de uso de licencia.

Los archivos de registro de licencia se crean en el mismo directorio donde se registran los archivos de registro de SPSS Modeler Client (de forma predeterminada, `%ALLUSERSPROFILE%/IBM/SPSS/Modeler/<version>/log`).

Capítulo 2. Minería interna de bases de datos

Conceptos básicos del modelado de bases de datos

IBM SPSS Modeler Server supports integration with data mining and modeling tools that are available from database vendors, including IBM Netezza, Oracle Data Miner, and 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 en IBM SPSS Modeler.

Nota: El modelado de bases de datos y la optimización SQL requieren que la conectividad IBM SPSS Modeler Server esté habilitada en el sistema 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.

Qué necesita

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 (Microsoft Analysis Services o Oracle Data Miner).
- En IBM SPSS Modeler, el modelado de base de datos debe estar habilitado en el recuadro 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: El modelado de bases de datos y la optimización SQL requieren que la conectividad IBM SPSS Modeler Server esté habilitada en el sistema 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 base de datos es conceptualmente idéntica a otras rutas de datos en IBM SPSS Modeler; sin embargo, esta ruta lleva a cabo 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 en base de datos es indicada por el uso de nodos sombreados en púrpura en la ruta.

Preparación de los 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 a 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.

Para la navegación y la puntuación, es necesaria una conexión en directo con el servidor que ejecuta Oracle Data Miner o Microsoft Analysis Services.

Visualización de resultados y especificación de valores

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.

Exportar y guardar 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** (soportado solo para modelos de IM de IBM Db2) exporta el modelo como lenguaje de marcación de modelo predictivo (PMML), que se puede utilizar con otro software compatible con PMML.

Nota: También puede guardar un modelo generado eligiendo **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 consistencia del modelo almacenado en la base de datos comparando su descripción con la clave al azar 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 11.

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

- Árboles de decisión
- Agrupación en 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.

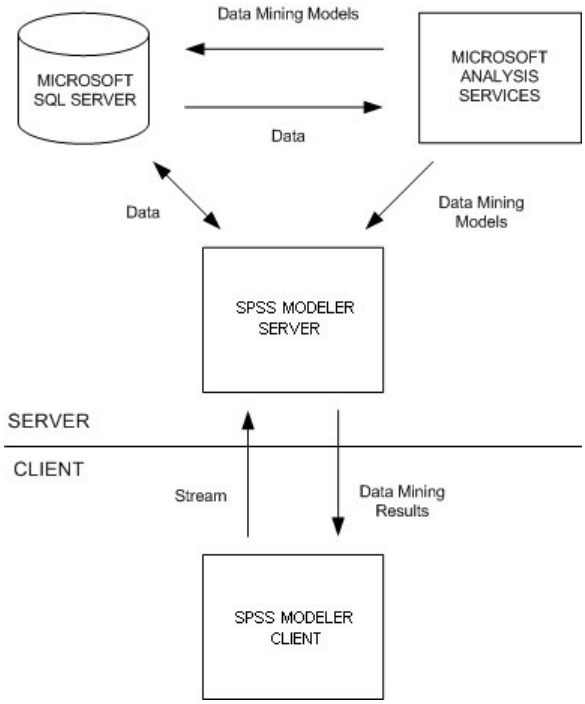

Figura 1. Flujo de datos entre IBM SPSS Modeler, Microsoft SQL Server y Microsoft Analysis Services durante la generación de modelos

Nota:: IBM SPSS Modeler Server no es necesario, aunque se puede utilizar. El cliente de IBM SPSS Modeler es capaz de procesar los cálculos de minería de base de datos en sí.

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 Modelerdeben configurar una conexión ODBC utilizando el controlador del cliente nativo SQL disponible en Microsoft en el URL listado debajo de *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.
- SQL Server debe estar instalado, 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 mismo host que IBM SPSS Modeler Server, ya estarán disponibles estos componentes.

- Microsoft SQL Server Analysis Services 10.0 OLE DB Provider (asegúrese de seleccionar la variación correcta de su sistema operativo)
- Microsoft SQL Server Native Client (asegúrese de seleccionar la variante correcta de su sistema operativo)
- Si utiliza Microsoft SQL Server 2008 o 2012, es posible que también requiera Microsoft Core XML Services (MSXML) 6.0.

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 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: El modelado de bases de datos y la optimización SQL requieren que la conectividad IBM SPSS Modeler Server esté habilitada en el sistema 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 10.

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 la 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 `True` (el valor predeterminado es `False`).
- Cambie el valor de `DataMining\AllowProvidersInOpenRowset` a `[all]` (no hay valor predeterminado).

Creación de un DSN ODBC para SQL Server

Para leer o escribir en una base de datos, debe tener un origen de datos ODBC instalado y configurado para la base de datos pertinente, con los permisos de lectura o escritura, según sea necesario. 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 10.

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. En los menús de IBM SPSS Modeler, elija:

   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 datos que desee pulsando en el botón de puntos suspensivos (...) para abrir un subcuadro de diálogo en el que puede elegir una de las 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. En los menús de IBM SPSS Modeler , elija:

   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 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 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 en 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 de **Regresión lineal de MS** es una variación del nodo Árboles de decisión, donde el parámetro `MINIMUM_LEAF_CASES` se establece para ser mayor o igual que el número total de casos del conjunto de datos que el nodo utiliza 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 de **Regresión logística de MS** es una variación del nodo de red neuronal de MS, donde el parámetro `HIDDEN_NODE_RATIO` se establece en 0. Este valor crea un modelo de red neuronal que no contiene una capa oculta y, por lo tanto, es equivalente 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 17.

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 18.

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 11.

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 20.

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.

- **Auto.** 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.

Utilice 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 en 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.

- **Auto.** 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.

Utilice datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen 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 21). 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 ruta 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 ver ejemplos de puntuación de modelo, consulte "Ejemplos de minería de datos con Analysis Services" en la página 22.

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 11.

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 solo está disponible para los modelos añadidos al lienzo de la ruta 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.

vista. Pulse para obtener una vista gráfica del modelo de árbol de decisión. El visor de árboles de decisión lo comparten otros 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 de 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 11.

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 solo está disponible para los modelos añadidos al lienzo de la ruta 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.

vista. 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 de 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 ruta 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	
Flujo	**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.

Ya 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 *CLAVE* con los valores exclusivos *1,2,3* utilizando la función IBM SPSS Modeler `@INDEX`.

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)

Sólo * 11*g* R1

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, excepto 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: El modelado de bases de datos y la optimización SQL requieren que la conectividad IBM SPSS Modeler Server esté habilitada en el sistema 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 del 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**, luego **Herramientas administrativas**. Seleccione **Orígenes de datos (ODBC)** y, a continuación, pulse en **Abrir**.

   b. Vaya a la pestaña **DSN del sistema** y, después, pulse **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. En los menús de IBM SPSS Modeler , elija:

   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 43 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. En los menús de IBM SPSS Modeler , elija:

   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 11gR1 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, Clúster-O 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 26.

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 *naïve* 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.

Utilice datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen 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, Clúster-O 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 habilitar esta característica, elija **Seleccionar**, pulse el botón **Especificar**, elija uno de los resultados posibles y, después, pulse **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.

Nota: El algoritmo bayesiano adaptativo de Oracle se ha descartado en Oracle 12C y no está soportado en IBM SPSS Modeler al utilizar Oracle 12C. Consulte http://docs.oracle.com/database/121/DMPRG/release_changes.htm#DMPRG726.

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 = "Married"
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.

Utilice datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen 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, Clúster-O 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. Solamente 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 gaussiano. 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 *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, Clúster-O 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 solamente 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, esto 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 habilitar esta característica, elija **Seleccionar**, pulse el botón **Especificar**, elija uno de los resultados posibles y, después, pulse **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.** Este es el método predeterminado. 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 *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, Clúster-O 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 habilitar esta característica, elija **Seleccionar**, pulse el botón **Especificar**, elija uno de los resultados posibles y, después, pulse **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.** Este es el método predeterminado. 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 tamiza a través de cada atributo de entrada potencial buscando el mejor "splitter", es decir, el punto de corte de atributo (`AGE > 55`, por ejemplo) que divide los registros de datos en sentido descendente en poblaciones más 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. Mirando hacia abajo desde el nodo del árbol raíz (por ejemplo, el total de la población), los árboles de decisión proporcionan reglas legibles humanas de las sentencias `IF A, then 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, Clúster-O 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 habilitar esta característica, elija **Seleccionar**, pulse el botón **Especificar**, elija uno de los resultados posibles y, después, pulse **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. La agrupación en clústeres de partición ortogonal (O-clúster) es un algoritmo de agrupación en clústeres propiedad de Oracle que crea un modelo de agrupación en clústeres jerárquica 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, Clúster-O 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 agrupación en clústeres 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, Clúster-O 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 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 solamente 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 agrupación en clústeres. 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, Clúster-O 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" se mantiene si la proporción de `support(ABC)` a `support(BC)` es al menos tan grande como la mínima confianza. 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. Este es el método predeterminado.

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).
- **principal.** 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, Clúster-O 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. Define 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. Establece el nivel de confianza mínimo, un valor entre 0 y 1. Las reglas con una confianza inferior a la especificada en el criterio de precisión se descartan.

Soporte mínimo. Establece el umbral de soporte mínimo, un valor entre 0 y 1. Apriori descubre patrones con frecuencia por encima del umbral de soporte mínimo.

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 43.
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, Clúster-O 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.

Utilice datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que sólo se utilicen 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.

Están disponibles las opciones siguientes:

Todos los campos clasificados. Selecciona los campos según la clasificación como *important*, *marginal* o *unimportant*. Se puede editar la etiqueta de 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. El modelo de IBM SPSS Modeler de este tipo hace referencia al contenido del modelo de base de datos almacenado en el servidor de base de datos. IBM SPSS Modeler puede realizar la comprobación de coherencia almacenando una cadena **clave del modelo** generada idéntica en el modelo de IBM SPSS Modeler y en el modelo 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 28.

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.

En intervalo

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 ruta 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	
Flujo	**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 Derivación para añadir un nuevo campo al conjunto de datos llamado *ID*, con los valores exclusivos 1,2,3, utilizando la función IBM SPSS Modeler `@INDEX`.

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, pulse dos veces en el nodo de generación (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 Data Warehouse y IBM Netezza Analytics

SPSS Modeler con IBM Data Warehouse y IBM Netezza Analytics

IBM SPSS Modeler admite la integración con IBM Data Warehouse y IBM Netezza Analytics, que proporciona la capacidad de ejecutar análisis avanzados en estos servidores de IBM. Se puede acceder a estas características a través de 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 o IBM Data Warehouse.

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

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

Si desea más información sobre estos algoritmos, consulte la *Guía del desarrollador de IBM Netezza Analytics* y la *Guía de referencia de IBM Netezza Analytics*.

SPSS Modeler admite la integración de los algoritmos siguientes de **IBM Data Warehouse** (la red bayesiana, la agrupación en clúster divisiva y la serie temporal no están soportados):

- Árboles de decisión
- K-medias
- TwoStep
- Bayesiano ingenuo
- KNN
- PCA
- Árbol de regresión
- Regresión lineal
- Lineal generalizado

Nota: AIX no está soportado.

Requisitos de integración

Las condiciones siguientes son requisitos previos para realizar el modelado interno de base de datos mediante IBM Netezza Analytics o IBM Data Warehouse. 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 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, que ejecuta el paquete de IBM Netezza Analytics.

  Nota: la versión mínima de Netezza Performance Server (NPS) necesaria depende de la versión de INZA necesaria, y es la siguiente:

  – Cualquiera mayor que NPS 6.0.0 P8 tendrá soporte para versiones de INZA anteriores a la 2.0.

  – Para utilizar INZA 2.0 o superior se necesita NPS 6.0.5 P5 o superior.

  Netezza Generalized Linear y Netezza Time Series requieren INZA 2.0 y superiores para ser funcionales. Todos los demás nodos de Netezza In-Database necesitan INZA 1.1 o posterior.
- Un origen de datos ODBC para conectar a una base de datos IBM Netezza. Para obtener más información, consulte el tema "Habilitación de la integración" en la página 48.
- Un origen de datos ODBC para conectarse a una base de datos IBM Data Warehouse.
- Optimización y generación de SQL activada en IBM SPSS Modeler. Para obtener más información, consulte el tema "Habilitación de la integración" en la página 48.

Nota: El modelado de bases de datos y la optimización SQL requieren que la conectividad IBM SPSS Modeler Server esté habilitada en el sistema 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.

Habilitación de la integración

La habilitación de la integración con IBM Netezza Analytics o IBM Data Warehouse consta de los pasos siguientes.

- Configuración de IBM Netezza Analytics o IBM Data Warehouse
- 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 o IBM Data Warehouse

Para instalar y configurar IBM Netezza Analytics o IBM Data Warehouse, consulte la documentación apropiada de IBM. Por ejemplo, para IBM Netezza Analytics, consulte la *Guía de instalación de IBM Netezza Analytics* que se proporciona con ese producto. 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 que se basan en cálculos de matriz, el motor de matrices se debe inicializar ejecutando `CALL NZM..INITIALIZE();` de lo contrario, fallará la ejecución de procedimientos almacenados. 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. Si desea instrucciones completas, consulte la Guía de IBM Netezza de configuración e instalación de ODBC, JDBC, y OLE DB.
   a. Cree el DSN.

      Nota: La secuencia del 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**, luego **Herramientas administrativas**. Seleccione **Orígenes de datos (ODBC)** y, a continuación, pulse en **Abrir**.

   b. Vaya a la pestaña **DSN del sistema** y, después, pulse **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. Desde el CD/DVD del cliente de Netezza, copie el archivo `<platform>cli.package.tar.gz` relevante 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.

```
. <SDAP Install Path>/odbc.sh
LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
NZ_ODBC_INI_PATH=<SDAP Install Path>; export NZ_ODBC_INI_PATH
```

Por ejemplo:

```
. /usr/IBM/SPSS/SDAP/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/SDAP; export NZ_ODBC_INI_PATH
```

6. Localice el archivo `/usr/local/nz/lib64/odbc.ini` y copie su contenido en el archivo `odbc.ini` que se ha instalado con SDAP (el definido mediante 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 vía de acceso 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.

Habilitación de la integración en SPSS Modeler

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

   Herramientas > **Opciones** > **Aplicaciones de ayuda.**

2. Pulse la pestaña **IBM Data Warehouse**.

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

Conexión de IBM Data Warehouse. Pulse el botón **Editar** y elija la serie de conexión de IBM Data Warehouse que configuró al crear el origen ODBC. Si desea más información, consulte la consola de administración de IBM Data Warehouse.

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. En los menús de IBM SPSS Modeler , elija:

   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).

Creación de modelos con IBM Netezza Analytics e IBM Data Warehouse

Cada uno de los algoritmos compatibles tiene un nodo de modelado correspondiente. Puede acceder a los nodos de modelos de IBM Data Warehouse e 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 ID de registro Especifica el campo utilizado para identificar de forma 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.

Nota: Si recupera datos numéricos largos de una base de datos Netezza utilizando el tipo de datos `NUMERIC(18,0)`, a veces, SPSS Modeler puede redondear los datos durante la importación. Para evitar este problema, guarde los datos utilizando el tipo de datos `BIGINT` o `NUMERIC(36,0)`.

Nota: Debido a las limitaciones en los tipos de campos que se pueden utilizar, un campo con un nivel de medición sin tipo y un rol de ID de registro no aparecen en un nodo de modelado interno de base de datos Netezza (por ejemplo, k-medias).

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(field1) [or @NULL(field2)[... or @NULL(fieldN]])
   ```

   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 Data Warehouse o Netezza genere resultados ligeramente diferentes cada vez que se ejecuta. 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 utilizando rutas que contengan nodos de modelado de IBM Data Warehouse o IBM Netezza.
- La exportación o importación de PMML no es posible para los modelos creados por los nodos de Data Warehouse o Netezza.

Opciones de campos

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 la configuración de rol (objetivos, predictores, etc) de un nodo de Tipo anterior (o la pestaña Tipos 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.

Opciones de servidor

En la pestaña Servidor, especifique la base de datos de IBM Data Warehouse donde se va a crear el modelo.

Detalles de IBM Data Warehouse Server. 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. Esta opción solo funciona si todos los nodos anteriores pueden utilizar 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 Data Warehouse 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 e IBM Data Warehouse normalmente se utiliza 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.

Nota: 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.

Opciones de 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.

Sustituir el existente si el nombre se ha utilizado. Si selecciona esta casilla de verificación, se sobrescribirá cualquier modelo existente con el mismo nombre.

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

La creación de un modelo de IBM Netezza o de IBM Data Warehouse a través de SPSS Modeler crea un modelo en SPSS Modeler y crea o sustituye un modelo en la base de datos IBM Data Warehouse. El modelo de 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. SPSS Modeler puede realizar una comprobación de congruencia almacenando una serie de clave idéntica del modelo generado en ambos modelos, el de SPSS Modeler y el de Netezza o Data Warehouse.

El nombre de modelo para cada modelo de Netezza o de Data Warehouse se muestra 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 SPSS Modeler se visualiza como la Clave de modelo en la pestaña Servidor de un modelo de SPSS Modeler (cuando se coloca en una ruta).

El botón Comprobar se puede utilizar para comprobar que las claves del modelo en el modelo de SPSS Modeler y el modelo de Netezza o de Data Warehouse coinciden. Si no se puede encontrar ningún modelo con el mismo nombre en Netezza o Data Warehouse, o si las claves del modelo no coinciden, el modelo de Netezza o de Data Warehouse se ha suprimido o se ha generado de nuevo desde que se generó el modelo de SPSS Modeler.

Listado de modelos de base de datos

SPSS Modeler proporciona un cuadro de diálogo para listar los modelos que están almacenados en IBM Data Warehouse y permite suprimir modelos. 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 creación, exploración y aplicación para nodos relacionados con la minería de datos de IBM Data Warehouse e 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.

Árbol de regresión de IBM Data WH

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.

Opciones de generación de árbol de regresión de IBM Data WH: Crecimiento del árbol

Puede definir opciones de generación para el crecimiento del árbol y la poda del árbol.

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.

Estadísticas. Este parámetro define cuántas estadísticas se incluyen en el modelo. Seleccione una de las opciones siguientes:

- **Todo.** 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 IBM Data WH: 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.

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.

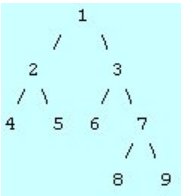

Figura 3. Ejemplo de un árbol de agrupación en clústeres divisiva

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 se establece en 2, por ejemplo, la puntuación devolverá uno de los clústeres en el segundo nivel por debajo del clúster raíz, a saber, 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 la configuración de rol (objetivos, predictores, etc) de un nodo de Tipo anterior (o la pestaña Tipos 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. Por supuesto, puede pulsar el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero, normalmente, deseará personalizar la generación para sus propios fines.

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.

Lineal generalizado de IBM Data WH

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, binomial, etc.

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 campo de modelo lineal generalizado de IBM Data WH

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 papeles predefinidos. Esta opción utiliza los valores de rol, como destinos o predictores desde un nodo Tipo anterior, o la pestaña Tipos de un nodo de origen anterior.

Utilizar asignaciones de campos personalizadas. Seleccione esta opción si desea asignar destinos, 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, los números de ID de cliente.

Ponderación de instancia. Especifique un campo para utilizar ponderaciones de instancia. Una ponderación de instancia es una ponderación por fila de datos de entrada. De forma predeterminada, se supone que todos los registros de entrada tienen la misma importancia relativa. Puede cambiar la importancia asignando ponderaciones individuales a los registros de entrada. El campo que especifique debe contener una ponderación numérica para cada fila de los datos de entrada.

Predictores (Entradas). Seleccione el campo(s) de entrada. Esta acción es similar a establecer el rol del campo en *Entrada* en un nodo Tipo.

Opciones de modelo lineal generalizado de IBM Data WH: 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; uno de **Bernoulli** (predeterminado), **Gauss**, **Poisson**, **Binomial**, **Binomial negativo**, **Wald** (Gauss inverso) y **Gamma**.
- **Parámetros.** (Solo distribución Poisson o binomial) Debe especificar una de las opciones siguientes en el campo **Especificar parámetro**:
  - Para que se estime el parámetro automáticamente a partir de los datos, seleccione **Predeterminado**.
  - Para permitir la optimización de la distribución de cuasi-verosimilitud, seleccione **Cuasi**.
  - Para especificar de forma explícita el valor de parámetro, seleccione **Explícito**.

(Solo distribución binomial) Debe especificar la columna de la tabla de entrada que se va a utilizar como el campo de ensayos como precisa la distribución binomial. Esta columna contiene el número de ensayos para la distribución binomial.

(Solo la distribución binomial negativa) Puede utilizar el valor predeterminado -1 o especificar un valor de parámetro diferente.

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 IBM Data WH: Interacciones

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 orígenes 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 descartados 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 4. Botón Suprimir

Eliminar términos del modelo seleccionando los términos que quiera eliminar y pulsando en el botón Eliminar.

Figura 5. Botones Reordenar

Reordenar los términos dentro del modelo seleccionando los términos que quiera reordenar y pulsando la flecha arriba o abajo.

Figura 6. Botón de interacción personalizada

Añadir un término personalizado

Puede especificar interacciones personalizadas en la forma*n1*x1*x1*x1...* Seleccione un campo de la lista **Campos**, pulse en el botón de flecha derecha para añadir el campo a **Término personalizado**, pulse en **Por***, seleccione el siguiente campo, pulse en el botón de flecha derecha y así sucesivamente. Cuando haya creado la interacción personalizada, pulse en **Añadir término** para devolverlo al panel Interacción.

Opciones de modelo lineal generalizado de IBM Data WH: 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 IBM Data WH: Pestaña Configuración" en la página 83.

- **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.

Árboles de decisión de IBM Data WH

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.

Ponderaciones de instancia y ponderaciones de clase

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, por ejemplo:

Tabla 5. Ejemplo de ponderación de instancia

ID de registro	Destino	Ponderación de instancia
1	medicamentoA	1,1
2	medicamentoB	1,0
3	medicamentoA	1,0

Tabla 5. Ejemplo de ponderación de instancia (continuación)		
ID de registro	**Destino**	**Ponderación de instancia**
4	medicamentoB	0,3

Ponderaciones de clase asignan una ponderación a cada categoría del campo de destino, tal como se muestra en la tabla siguiente.

Tabla 6. 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 se han utilizado dos ejemplos previos de forma conjunta, el algoritmo utilizaría ponderaciones de instancia tal como se muestra en la tabla siguiente.

Tabla 7. 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 los valores del rol (objetivos, predictores, etc.) desde un nodo Tipo en sentido ascendente (o la pestaña Tipos de un nodo de origen en sentido ascendente).

Utilizar asignaciones de campos personalizadas. Para asignar manualmente los objetivos, predictores y otros roles, seleccione esta opción.

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 un botón de nivel de medición individual para seleccionar todos los campos con ese nivel de medición.

Objetivo. Seleccione un campo como el objetivo 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 instancia y ponderaciones de clase” en la página 59.

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 IBM Data WH

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.

- **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 de esta propiedad es 10 y el valor máximo que puede establecer para esta propiedad es 62.

  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.

- **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.

Estadísticas. Este parámetro define cuántas estadísticas se incluyen en el modelo. Seleccione una de las opciones siguientes:

- **Todo.** 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 IBM Data WH: Ponderaciones de clase

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.

Peso. 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 instancia y ponderaciones de clase” en la página 59.

Nodo de árbol de decisión de IBM Data WH: 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.

Regresión lineal de IBM Data WH

Los modelos lineales predicen un objetivo continuo basándose en relaciones lineales entre el objetivo 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 IBM Data WH

La pestaña Opciones de generación es la ubicación en la que se definen todas las opciones para crear el modelo. Por supuesto, puede pulsar el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero, normalmente, deseará personalizar la generación para sus propios fines.

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 almacenan 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.

KNN de IBM Data WH

Análisis de vecino más próximo es un método de clasificación de casos basado en su similaridad con otros casos. En el aprendizaje de máquinas, se ha desarrollado como una forma para reconocer patrones de datos sin requerir una coincidencia exacta con patrones almacenados, o casos. 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 próximos entre sí 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 parecidos, los vecinos más próximos, se cuadran y el nuevo caso se incluye en la categoría que contiene el mayor número de vecinos más próximos.

Puede especificar el número de vecinos más cercanos a examinar; este valor se denomina *k*. Las imágenes muestran cómo se clasificaría un nuevo caso utilizando dos valores diferentes de *k*. Cuando *k* = 5, el nuevo caso se coloca en la categoría *1* porque la mayoría de los vecinos más cercanos pertenecen a la categoría *1*. Sin embargo, cuando *k* = 9, el nuevo caso se coloca en la categoría *0* porque la mayoría de los vecinos más cercanos 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 KNN de IBM Data WH: 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 KNN de IBM Data WH: 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 puntuación

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.

Peso. 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.

K-medias de IBM Data WH

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 de k-medias de IBM Data WH

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 la configuración de rol (objetivos, predictores, etc) de un nodo de Tipo anterior (o la pestaña Tipos 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 IBM Data WH

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.
- **Máximo.** 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.

Estadísticas. Este parámetro define cuántas estadísticas se incluyen en el modelo. Seleccione una de las opciones siguientes:

- **Todo.** 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**.

Bayesiano ingenuo de IBM Data WH

Bayesiano ingenuo es un algoritmo muy utilizado para resolver problemas de clasificación. El modelo se denomina *naïve* 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.

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. Utilizando el nodo de red bayesiana Netezza, puede crear un modelo de probabilidad combinando pruebas observadas y registradas con el conocimiento del mundo real y el sentido común para establecer la probabilidad de apariciones mediante 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 77.

Utilizar papeles predefinidos. Esta opción utiliza la configuración de rol (objetivos, predictores, etc) de un nodo de Tipo anterior (o la pestaña Tipos 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. Por supuesto, puede pulsar el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero, normalmente, deseará personalizar la generación para sus propios fines.

Índice básico. El identificador numérico que se asignará al primer atributo (campo de entrada) para una gestión interna más fácil.

Tamaño de muestra. 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 este recuadro se selecciona (por defecto), se muestra información de progreso adicional en un cuadro de diálogo de mensaje.

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 previsión 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 exponencial 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 8. 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 9. Lecturas de temperaturas

Fecha	**Tiempo**	**Temperatura:**
2011-07-24	7:00	57
2011-07-24	14:00	75
2011-07-24	21:00	72
2011-07-25	7:15	59
2011-07-25	14:00	77
2011-07-25	20:55	74
2011-07-27	7:00	60
2011-07-27	14:00	78
2011-07-27	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 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 10. Lecturas de temperaturas (agregadas)

Fecha	**Tiempo**	**Temperatura:**
2011-07-24	24:00	69
2011-07-25	24:00	71
2011-07-26	24:00	NULL
2011-07-27	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 11. Lecturas de temperaturas con el tamaño de paso calculado

Fecha	**Tiempo**	**Temperatura:**
2011-07-24	6:00	
2011-07-24	14:00	75
2011-07-24	22:00	
2011-07-25	6:00	
2011-07-25	14:00	77
2011-07-25	22:00	
2011-07-26	6:00	

Tabla 11. Lecturas de temperaturas con el tamaño de paso calculado (continuación)		
Fecha	**Tiempo**	**Temperatura:**
2011-07-26	14:00	
2011-07-26	22:00	
2011-07-27	6:00	
2011-07-27	14:00	78
2011-07-27	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. Por supuesto, puede pulsar el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero, normalmente, deseará personalizar la generación para sus propios fines.

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 66.

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 67.

- **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 69.

- **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 ARIMA

Especifique los valores de los diversos componentes estacionales y no estacionales del modelo ARIMA. En cada caso, establezca el operador en **=** (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 se corresponde con el 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 de 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 69.

- **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 **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 67.

Biétapico de IBM Data WH

El nodo TwoStep implementa el algoritmo TwoStep que proporciona un método para agrupar datos a lo largo de conjuntos de datos grandes.

Puede utilizar este nodo para agrupar datos cuando los recursos disponibles, tales como restricciones de memoria y de tiempo, son un factor importante.

El algoritmo TwoStep es un algoritmo de minería de datos que agrupa datos de la manera siguiente:

1. Se crea un árbol de dispositivo de clúster (CF). Este árbol altamente equilibrado almacena funciones de agrupación jerárquica en la que registros de entrada similares pasan a formar parte de los mismos nodos del árbol.
2. Las hojas del árbol CF están agrupadas jerárquicamente en memoria para generar el resultado final de la agrupación en clúster. El mejor número de clústeres se determina automáticamente. Si especifica un número máximo de clústeres, se determina el mejor número de clústeres dentro del límite especificado.
3. El resultado de la agrupación se refina en un segundo paso donde un algoritmo que es similar al algoritmo K-Means se aplica a los datos.

Opciones de campo biétapico de IBM Data WH

Al establecer las opciones de campo, puede especificar que se utilicen los valores de rol de campo definidos en nodos anteriores. También puede realizar asignaciones de campo manualmente.

Seleccionar un elemento. Elija esta opción para utilizar los valores de rol de un nodo de Tipo anterior o desde la pestaña Tipos de un nodo de origen anterior. Los valores de rol son, por ejemplo, destinos y predictores.

Utilizar asignaciones de campos personalizadas. Seleccione esta opción si desea asignar objetivos, predictores y otros roles manualmente.

Campos. Utilice las flechas para asignar elementos manualmente desde esta lista a los campos de rol de la derecha. Los iconos indican los niveles de medición válidos para cada campo de rol.

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 bietápica de IBM Data WH

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. Las opciones son:

- **Log-verosimilitud.** La medida de la verosimilitud realiza una distribución de probabilidad entre las variables. Las variables continuas se supone que tienen una distribución normal, mientras que las variables categóricas se supone que son multinomiales. Se supone que todas las variables son independientes.
- **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.

Número de clúster. Este parámetro define el número de clústeres que se va a crear. Las opciones son:

- **Calcular automáticamente número de clústeres.** El número de clústeres se calcula automáticamente. Puede especificar el número máximo de clústeres en el campo **Máximo**.
- **Especificar número de clústeres.** Especifique cuántos clústeres se deben crear.

Estadísticas. Este parámetro define cuántas estadísticas se incluyen en el modelo. Las opciones son:

- **Todo.** 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**.

PCA de IBM Data WH

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.

Nota: Puede que se produzca un error al puntuar el modelo si se utilizan nombres de campo en minúsculas. Este es un problema conocido de Db2 Data Warehouse que puede solucionarse cambiando el nombre de todos los campos a mayúsculas antes de puntuar.

Opciones de campo de PCA de IBM Data WH

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 la configuración de rol (objetivos, predictores, etc) de un nodo de Tipo anterior (o la pestaña Tipos 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 IBM Data WH

La pestaña Opciones de generación es la ubicación en la que se definen todas las opciones para crear el modelo. Por supuesto, puede pulsar el botón **Ejecutar** para generar un modelo con todas las opciones predeterminadas, pero, normalmente, deseará personalizar la generación para sus propios fines.

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.

Gestión de modelos de IBM Data WH y de Netezza

Los modelos de IBM Data Warehouse y de IBM Netezza Analytics se añaden al lienzo y a la paleta Modelos de la misma forma que otros modelos de IBM SPSS Modeler, y se pueden utilizar prácticamente del mismo modo. Sin embargo, existen unas pocas diferencias significativas, puesto que cada modelo de IBM Data Warehouse o de IBM Netezza Analytics creado en IBM SPSS Modeler se refiere en realidad a un modelo almacenado en un servidor de 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 IBM Data Warehouse y 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 por el prefijo `$<id>-` que se añade al nombre del campo de destino, 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. Pulse 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 IBM Data WH y 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 de IBM Data Warehouse Server. 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. Esta opción solo funciona si todos los nodos anteriores pueden utilizar 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 Data Warehouse 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 e IBM Data Warehouse normalmente se utiliza 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 IBM Data WH

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 nugget de modelo Árbol de decisión, de forma predeterminada el nodo añade un campo nuevo, cuyo nombre se deriva del nombre del objetivo.

Tabla 12. Campo de puntuación de modelo del árbol de decisión

Nombre del campo añadido	**Descripción**
$I-*nombre_objetivo*	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 13. Campo de puntuación de modelo del árbol de decisión: Adicional

Nombre del campo añadido	**Descripción**
$IP-*nombre_objetivo*	Valor de confianza (entre 0,0 y 1,0) de la predicción.

Nugget de árbol de decisión de IBM Data WH: 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 de árbol de decisión de IBM Data WH: 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.

Utilizar datos de entrada deterministas. Si se selecciona, esta opción garantiza que todo algoritmo de Netezza que ejecute varias pasadas de la misma vista utilice el mismo conjunto de datos en cada pasada. Si desmarca este recuadro de selección para indicar que se utilicen datos no deterministas, se creará una tabla temporal para conservar los resultados de datos para su proceso, como los producidos por un nodo de partición; esta tabla se suprime una vez que se crea el modelo.

Nugget de árbol de decisión de IBM Data WH - Pestaña Visor

La pestaña **Visor** muestra una presentación de árbol del modelo de árbol del mismo modo en que lo hace SPSS Modeler para su modelo de árbol de decisión.

Nota: Si el modelo está construido con IBM Netezza Analytics Versión 2.x o anterior, la pestaña **Visor** estará vacía.

Nugget de modelo de k-medias de IBM Data WH

Los nugget de modelo de K-medias contienen toda la información capturada por el modelo de agrupación en cluster, 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. El nuevo campo con nombre $KM-K-Means es para la pertenencia del clúster y el nuevo campo con nombre $KMD-K-Means es para la distancia desde el centro del clúster.

Nugget de k-medias de IBM Data WH: 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 IBM Data WH: 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 nugget de modelo de Red bayesiana, el nodo añade un campo nuevo, cuyo nombre se deriva del nombre objetivo.

Tabla 14. Campo de puntuación de modelo de red bayesiana

Nombre del campo añadido	Descripción
$BN-*nombre_objetivo*	Valor predicho del registro actual.

Puede ver el campo extra si adjunta un nodo Tabla al nugget de modelo y ejecuta el nodo Tabla.

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 IBM Data WH

El nugget de modelo bayesiano ingenuo proporciona un método para definir las opciones de puntuación del modelo.

Si ejecuta una ruta con un nugget de modelo bayesiano ingenuo, de forma predeterminada el nodo añade un campo nuevo, cuyo nombre se deriva del nombre del objetivo.

Tabla 15. Campo de puntuación de modelo bayesiano ingenuo: Opción predeterminada

Nombre del campo añadido	**Descripción**
$I-*nombre_objetivo*	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 16. Campos de puntuación de modelo bayesiano ingenuo: Adicional

Nombre del campo añadido	**Descripción**
$IP-*nombre_objetivo*	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_objetivo*	El logaritmo natural del segundo.

Puede ver los campos extra si adjunta un nodo Tabla al nugget de modelo y ejecuta el nodo Tabla.

Nugget de bayesiano ingenuo de IBM Data WH: 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 IBM Data WH

El nugget de modelo KNN proporciona un método para definir las opciones de puntuación del modelo.

Si ejecuta una ruta con un nugget de modelo KNN, el nodo añade un campo nuevo, cuyo nombre se deriva del nombre objetivo.

Tabla 17. Campo de puntuación de modelo de KNN	
Nombre del campo añadido	**Descripción**
$KNN-*nombre_objetivo*	Valor predicho del registro actual.

Puede ver el campo extra si adjunta un nodo Tabla al nugget de modelo y ejecuta el nodo Tabla.

Nugget de KNN de IBM Data WH: 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 nugget de modelo de agrupación en clústeres divisivo, el nodo añade dos campos nuevos, cuyos nombres se derivan del nombre del objetivo.

Tabla 18. Campos de puntuación de modelo de clúster divisivo	
Nombre del campo añadido	**Descripción**
$DC-*nombre_objetivo*	Identificador del subclúster al que se asigna el registro actual.
$DCD-*nombre_objetivo*	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.

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 IBM Data WH

El nugget de modelo PCA proporciona un método para definir las opciones de puntuación del modelo.

Si ejecuta una ruta con un nugget de modelo PCA, de forma predeterminada el nodo añade un campo nuevo, cuyo nombre se deriva del nombre del objetivo.

Tabla 19. Campo de puntuación de modelo de PCA

Nombre del campo añadido	**Descripción**
$F-*nombre_objetivo*	Valor predicho del registro actual.

Si especifica un valor mayor que 1 en la **Número de componentes principales ...** en el nodo de modelado o en el nugget de modelo y ejecute la ruta, el nodo añade un nuevo campo 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.

Nota: Puede que se produzca un error al puntuar el modelo si se utilizan nombres de campo en minúsculas. Este es un problema conocido de Db2 Data Warehouse que puede solucionarse cambiando el nombre de todos los campos a mayúsculas antes de puntuar.

Nugget de PCA de IBM Data WH: 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 nugget de modelo Árbol de regresión, de forma predeterminada el nodo añade un campo nuevo, cuyo nombre se deriva del nombre del objetivo.

Tabla 20. Campo de puntuación de modelo del árbol de regresión

Nombre del campo añadido	**Descripción**
$I-*nombre_objetivo*	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 21. Campo de puntuación de modelo del árbol de regresión: Adicional

Nombre del campo añadido	**Descripción**
$IV-*nombre_objetivo*	Varianzas estimadas del valor predicho.

Puede ver los campos extra si adjunta un nodo Tabla al nugget de modelo y ejecuta el nodo Tabla.

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.

Nugget de árbol de regresión de Netezza - Pestaña Visor

La pestaña **Visor** muestra una presentación de árbol del modelo de árbol del mismo modo en que lo hace SPSS Modeler para su modelo de árbol de regresión.

Nota: Si el modelo está construido con IBM Netezza Analytics Versión 2.x o anterior, la pestaña **Visor** estará vacía.

Nuggets de modelo de regresión lineal de IBM Data WH

El nugget de modelo de regresión lineal proporciona un método para definir las opciones de puntuación del modelo.

Si ejecuta una ruta con un nugget de modelo Regresión lineal, el nodo añade un campo nuevo, cuyo nombre se deriva del nombre objetivo.

Tabla 22. Campo de puntuación de modelo de regresión lineal

Nombre del campo añadido	Descripción
$LR-*nombre_destino*	Valor predicho del registro actual.

Nugget de regresión lineal de IBM Data WH - 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 23. 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 69.
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 IBM Data WH

El nugget de modelo permite acceder al resultado de la operación de modelado.

Si ejecuta una ruta con un nugget de modelo lineal generalizado, el nodo añade un campo nuevo, cuyo nombre se deriva del nombre objetivo.

Tabla 24. Campo de puntuación de modelo de lineal generalizado

Nombre del campo añadido	**Descripción**
$GLM-*nombre_objetivo*	Valor predicho del registro actual.

La pestaña Modelo muestra varios estadísticos relacionados con el modelo.

El resultado consta de los siguientes campos.

Tabla 25. Campos de resultados del modelo lineal generalizado

Campo de salida	**Descripción**
Parámetro	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.
Prueba	Los estadísticos de prueba utilizados para evaluar la validez del parámetro.
Valor de probabilidad	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 de probabilidad	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 IBM Data WH: 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 IBM Data WH: Opciones de puntuación" en la página 59.

Nugget de modelo bietápico de IBM Data WH

Cuando ejecuta una ruta que contiene un nugget de modelo TwoStep, el nodo añade dos campos nuevos que contienen la información de pertenencia al clúster y la distancia respecto al centro del clúster asignado correspondiente al registro. El nuevo campo con nombre $TS-Twostep es para la pertenencia del clúster, y el nuevo campo con nombre $TSP-Twostep es para la distancia desde el centro del clúster.

Nugget biétapico de IBM Data WH - 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.

Capítulo 6. Modelador de bases de datos con IBM Db2 para z/OS

IBM SPSS Modeler e IBM Db2 para z/OS

SPSS Modeler soporta la integración con Db2 para z/OS, que proporciona la capacidad de ejecutar análisis avanzados en servidores Db2 para z/OS. Puede acceder a estas características a través de la interfaz de usuario gráfica de SPSS Modeler y el entorno de desarrollo orientado a flujo de trabajo. De esta forma, puede ejecutar los algoritmos de minería de datos directamente en el entorno de Db2 para z/OS sacando partido de IBM Db2 Analytics Accelerator.

SPSS Modeler soporta la integración de los algoritmos siguientes de Db2 para z/OS.

- Árboles de decisión
- K-medias
- Bayesiano ingenuo
- Árbol de regresión
- TwoStep

Requisitos para la integración con IBM Db2 para z/OS

Las condiciones siguientes son requisitos previos para realizar modelado interno de base de datos utilizando Db2 para z/OS e IBM Db2 Analytics Accelerator for z/OS. Para asegurarse de que se cumplen estas condiciones, puede ser necesario que consulte al administrador de bases de datos. Para obtener requisitos detallados, incluyendo las versiones soportadas, consulte Informes de compatibilidad de productos de software.

- IBM SPSS Modeler se debe estar ejecutando en la modalidad local o para una instalación de SPSS Modeler Server en Windows o UNIX
- Db2 para z/OS junto con Db2 Analytics Accelerator para z/OS
- IBM SPSS Data Access Pack
- En el servidor donde se ejecuta SPSS Modeler Server, debe existir uno de los sistemas siguientes:
  - IBM Db2 Data Server Driver for ODBC and CLI
  - Cualquier versión de Db2 para Linux®, UNIX y Windows con un origen de datos ODBC que está configurado para Db2 para z/OS
- Licencia para Db2 Connect para System z
- Generación y optimización de SQL habilitada en SPSS Modeler
- La minería interna de base de datos Db2 z/OS requiere tablas de solo acelerador (AOT) o tablas aceleradas y soporte de INZA. IDAA INZA se presentó en IDAA 5.1. Esto significa que los nodos de minería interna de base de datos Db2 z/OS no funcionarán con versiones anteriores de IDAA.

  Si utiliza un DSN habilitado para IDAA en Modeler, las únicas tablas que se mostrarán en la lista de tablas devueltas en el nodo de origen Base de datos que utilizan dicho DSN no será AOT ni tablas aceleradas.

Habilitar la integración con IBM Db2 Analytics Accelerator for z/OS

La habilitación de la integración con Db2 Analytics Accelerator for z/OS consta de los pasos siguientes:

- Configuración de Db2 para z/OS y Db2 Analytics Accelerator for z/OS
- Creación de un origen ODBC

- Habilitación de la integración de IBM Db2 para z/OS en IBM SPSS Modeler
- Habilitar la generación y optimización de SQL en SPSS Modeler
- Habilitación del adaptador de puntuación de IBM SPSS Modeler Server para Db2 para z/OS
- Configuración de DSN utilizando el cliente de IBM Db2 en IBM SPSS Modeler

Configurar IBM Db2 para z/OS e IBM Analytics Accelerator for z/OS

Cómo configurar Db2 para z/OS y Analytics Accelerator for z/OS se describe en el sitio web siguiente:

Db2 Analytics Accelerator for z/OS.

Crear un origen ODBC para IBM Db2 para z/OS e IBM Db2 Analytics Accelerator

Si desea más información sobre cómo habilitar una conexión entre Db2 para z/OS e IBM Db2 Analytics Accelerator, consulte los sitios web siguientes:

- Para la versión 4: Db2 Analytics Accelerator for z/OS 4.1.0
- Para la versión 3: Db2 Analytics Accelerator for z/OS 3.1.0
- Habilitar la aceleración de consulta con IBM Db2 Analytics Accelerator para aplicaciones ODBC y JDBC sin modificar las aplicaciones
- Error SQL del controlador ODBC al ejecutar una consulta en Db2 Analytics Accelerator for z/OS

Habilitación de la integración de IBM Db2 para z/OS en IBM SPSS Modeler

Para habilitar la integración de Db2 para z/OS en SPSS Modeler, realice los pasos siguientes:

1. Desde el directorio `config` de SPSS Modeler, abra el archivo `odbc-db2-accelerator-names.cfg`.

   Si el archivo no existe, debe crearlo.
2. Añada los nombres de todos los orígenes de datos y los nombres de todos los aceleradores. Por ejemplo:

   ```
   dsn1, acceleratorname1
   dsn2, acceleratorname2
   ```
3. El CCSID predeterminado para tablas de solo acelerador (AOT) es Unicode; para alterar temporalmente esto, modifique las entradas añadiendo cadenas de codificación a los nombres de acelerador. Por ejemplo:

   ```
   dsn1, acceleratorname1, EBCDIC
   dsn2, acceleratorname2, UNICODE
   ```
4. Guarde y cierre el archivo `odbc-db2-accelerator-names.cfg` y, después, abra el archivo `odbc-db2-custom-properties.cfg` desde el mismo directorio.
5. SPSS Modeler utiliza SQL para establecer los registros IDAA. Si es necesario, puede alterar temporalmente estas entradas cambiando el SQL por los valores necesarios. Por ejemplo:

   ```
   current_query_sql_acc, "SET CURRENT QUERY ACCELERATION = ELIGIBLE"
   current_get_archive_acc, "SET CURRENT GET_ACCEL_ARCHIVE = NO"
   ```
6. De forma predeterminada, SPSS Modeler utiliza SQL para crear tablas temporales para una memoria caché de base de datos. Si es necesario, puede alterar temporalmente este hecho especificando el nombre previsto de la base de datos. Por ejemplo:

   ```
   [OSZ]
   table_create_temp_sql_acc, 'CREATE TABLE <table-name> <(table-columns)> IN DATABASE
   NAME_OF_DATABASE_FOR_AOT'
   ```

7. De forma predeterminada, SPSS Modeler considera que las consultas SQL escritas en un nodo de origen ODBC no se pueden reproducir, lo que significa que la consulta se considera que devolverá distintos resultados cuando se ejecute varias veces. Sin embargo, en algunos casos, esto puede impedir que Modeler genere SQL para nodos en sentido descendente y se puede alterar temporalmente cambiando el valor relevante a Y. Por ejemplo:

   ```
   assume_custom_sql_replayable, Y
   ```

8. En el menú principal de SPSS Modeler, pulse **Herramientas** > **Opciones** > **Aplicaciones de ayuda**.
9. Pulse la pestaña **IBM Db2 para z/OS**.
10. Seleccione **Habilitar la integración de minería de datos de IBM Db2 para z/OS** y, después, pulse **Aceptar**.

Nota: No puede ver tablas IDAA y tablas no IDAA a la vez en Modeler.

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.

Para configurar SPSS Modeler, realice los pasos siguientes:

1. Desde los menús de IBM SPSS Modeler elija **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).

Configuración de DSN utilizando el cliente de IBM Db2 en IBM SPSS Modeler

Si es necesario, para configurar un nombre de origen de datos (DSN) utilizando el cliente de Db2 para Db2 en SPSS Modeler, complete los pasos siguientes:

1. Si todavía no está instalado, instale el cliente de Db2 en el sistema operativo donde está instalado el Modeler Server.
2. Utilizando el comando **db2 `catalog`**, catalogue la base de datos y añada un nuevo origen de datos al archivo `db2cli.ini` en el cliente de Db2. Asegúrese de que señale al alias de base de datos definido.
3. Configure el acceso de datos; los pasos detallados están disponibles en la documentación de Modeler.

   Si desea más información, consulte el tema **Recomendaciones de arquitectura y hardware** > **Data Access** en la guía *Modeler Server Administration and Performance Guide* (ModelerServerAdminPerformance.pdf).
4. Cree un origen de datos ODBC nuevo en `odbc.ini` haciendo referencia al alias de base de datos definido en el paso 2.
5. Para usuarios Linux o UNIX:
   a. Asegúrese de que se utiliza la biblioteca del controlador `libdb2o.so` (en lugar de `libdb2.so`), y asegúrese de que `'DriverUnicodeType=1'` está definido para el origen de datos nuevo.
   b. En la instalación de IBM SPSS Data Access Pack, asegúrese de que la vía de acceso a biblioteca del cliente de Db2 se ha añadido a `odbc.sh`.
   c. Asegúrese de que Modeler Server utiliza una biblioteca de derivador de controlador ODBC con codificación UTF-16 (este se denomina `'libspssodbc_datadirect_utf16.so'`).
6. Asegúrese de que el usuario que se conecta a Db2 tiene los privilegios necesarios para ejecutar la consulta siguiente:

   ```
   SELECT ACCELERATORNAME FROM SYSACCEL.SYSACCELERATORS
   ```

Crear modelos con IBM Db2 para z/OS

Cada uno de los algoritmos compatibles tiene un nodo de modelado correspondiente. Puede acceder a los nodos de modelado de Db2 para z/OS desde la pestaña Modelador 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 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, el uso de algunos nodos de Db2 para z/OS puede generar mensajes de error o rutas de larga ejecución, así que se recomienda 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(field1) [or @NULL(field2)[... or @NULL(fieldN]])
   ```

   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 contenga un nodo de modelado de Db2 para z/OS para generar resultados ligeramente diferentes cada vez que se ejecuta. 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 SPSS Collaboration and Deployment Services, no es posible crear configuraciones de puntuación utilizando rutas que contengan nodos de modelado de Db2 para z/OS.
- La exportación o importación de PMML no es posible para los modelos creados por los nodos de Db2 para z/OS.

Modelos de IBM Db2 para z/OS - 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 la configuración de rol (objetivos, predictores, etc) de un nodo de Tipo anterior (o la pestaña Tipos 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 IBM Db2 para z/OS - Opciones de servidor

En la pestaña Servidor, especifique el sistema Db2 para z/OS donde se creará el modelo.

- **Utilizar conexión anterior en la ruta.** (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 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.

Nota: El nombre de origen de datos ODBC se incrusta de manera efectiva en cada ruta de 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 IBM Db2 para z/OS - Opciones Modelo

En la pestaña Opciones del modelo, puede seleccionar si desea especificar un nombre para el modelo o generar un nombre automáticamente.

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.

Sustituir existente si el nombre se ha utilizado. Si selecciona esta casilla de verificación, se sobrescribirá cualquier modelo existente con el mismo nombre.

Modelos de IBM Db2 para z/OS - k-medias

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.

Modelos de IBM Db2 para z/OS - Opciones de campo de k-medias

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 la configuración de rol (objetivos, predictores, etc) de un nodo de Tipo anterior (o la pestaña Tipos 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.

Modelos de IBM Db2 para z/OS - Opciones de generación de k-medias

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.

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.

Estadísticas. Este parámetro define cuántas estadísticas se incluyen en el modelo. Seleccione una de las opciones siguientes:

- **Todo.** 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**.

Modelos de IBM Db2 para z/OS - Bayesiano ingenuo

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.

Modelos de IBM Db2 para z/OS - Árboles de decisiones

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.

Modelos de IBM Db2 para z/OS - Opciones de campo de árbol de decisiones

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 la configuración de rol (objetivos, predictores, etc) de un nodo de Tipo anterior (o la pestaña Tipos 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.

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.

Modelos de IBM Db2 para z/OS - Opciones de generación de árbol de decisiones

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.

- **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 de esta propiedad es 10 y el valor máximo que puede establecer para esta propiedad es 62.

  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.

- **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.

Estadísticas. Este parámetro define cuántas estadísticas se incluyen en el modelo. Seleccione una de las opciones siguientes:

- **Todo.** 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.

Modelos de IBM Db2 para z/OS - Nodo de árbol de decisiones - Ponderaciones de clase

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.

Modelos de IBM Db2 para z/OS - Nodo de árbol de decisiones - 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.
- **Utilizar un % 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.
- **Utilizar datos desde una tabla 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.

Modelos de IBM Db2 para z/OS - Árbol de regresión

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.

Modelos de IBM Db2 para z/OS - Opciones de generación de árbol de regresión - Crecimiento de árbol

Puede definir opciones de generación para el crecimiento del árbol y la poda del árbol.

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.

- **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.

Estadísticas. Este parámetro define cuántas estadísticas se incluyen en el modelo. Seleccione una de las opciones siguientes:

- **Todo.** 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.

Modelos de IBM Db2 para z/OS - Opciones de generación de árbol de regresión - 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.
- **Utilizar un % 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.
- **Utilizar datos desde una tabla 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.

Modelos de IBM Db2 para z/OS - Dos fases

El nodo TwoStep implementa el algoritmo TwoStep que proporciona un método para agrupar datos a lo largo de conjuntos de datos grandes.

Puede utilizar este nodo para agrupar datos cuando los recursos disponibles, tales como restricciones de memoria y de tiempo, son un factor importante.

El algoritmo TwoStep es un algoritmo de minería de datos que agrupa datos de la manera siguiente:

1. Se crea un árbol de dispositivo de clúster (CF). Este árbol altamente equilibrado almacena funciones de agrupación jerárquica en la que registros de entrada similares pasan a formar parte de los mismos nodos del árbol.
2. Las hojas del árbol CF están agrupadas jerárquicamente en memoria para generar el resultado final de la agrupación en clúster. El mejor número de clústeres se determina automáticamente. Si especifica un número máximo de clústeres, se determina el mejor número de clústeres dentro del límite especificado.
3. El resultado de la agrupación se refina en un segundo paso donde un algoritmo que es similar al algoritmo K-Means se aplica a los datos.

Modelos de IBM Db2 para z/OS - Opciones de campo bietápico

Al establecer las opciones de campo, puede especificar que se utilicen los valores de rol de campo definidos en nodos anteriores. También puede realizar asignaciones de campo manualmente.

Seleccionar un elemento. Elija esta opción para utilizar los valores de rol de un nodo de Tipo anterior o desde la pestaña Tipos de un nodo de origen anterior. Los valores de rol son, por ejemplo, destinos y predictores.

Utilizar asignaciones de campos personalizadas. Seleccione esta opción si desea asignar objetivos, predictores y otros roles manualmente.

Campos. Utilice las flechas para asignar elementos manualmente desde esta lista a los campos de rol de la derecha. Los iconos indican los niveles de medición válidos para cada campo de rol.

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 IBM Db2 para z/OS - Opciones de generación de dos fases

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. La opción es:

- **Log-verosimilitud.** La medida de la verosimilitud realiza una distribución de probabilidad entre las variables. Las variables continuas se supone que tienen una distribución normal, mientras que las variables categóricas se supone que son multinomiales. Se supone que todas las variables son independientes.

Número de clúster. Este parámetro define el número de clústeres que se va a crear. Las opciones son:

- **Calcular automáticamente número de clústeres.** El número de clústeres se calcula automáticamente. Puede especificar el número máximo de clústeres en el campo **Máximo**.
- **Especificar número de clústeres.** Especifique cuántos clústeres se deben crear.

Estadísticas. Este parámetro define cuántas estadísticas se incluyen en el modelo. Las opciones son:

- **Todo.** 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**.

Nugget de modelos bietápicos de IBM Db2 para z/OS - 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.

Gestionar modelos de IBM Db2 para z/OS

Los modelos de Db2 para z/OS se añaden al lienzo y la paleta Modelos de la misma forma que otros modelos de IBM SPSS Modeler y se pueden utilizar de forma muy parecida.

Para puntuar los datos directamente en Db2 para z/OS, realice los pasos siguientes:

1. Instale el adaptador de puntuación SPSS en la base de datos de Db2 para z/OS donde se encuentran los datos.
2. Asegúrese de que la ruta se conecta a la base de datos Db2 para z/OS donde se encuentran los datos.

Puntuar modelos de IBM Db2 para z/OS

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 por el prefijo `$<id>-` que se añade al nombre del campo de destino, 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. Pulse en **Ejecutar**.
4. Desplácese a la derecha de la ventana de resultado de la tabla para ver los campos extra y sus resultados.

Nota: El proceso de puntuación no se ejecuta en el acelerador, pero sí en Db2 y, por consiguiente, requiere que la tabla de entrada para la puntuación esté situada físicamente en Db2. Por lo tanto, como entrada de puntuación, solo se puede utilizar una tabla basada en Db2 o una tabla acelerada. Si la ruta utiliza una tabla de solo acelerador, se produce el error siguiente: `"LA DECLARACIÓN NO PUEDE SER EJECUTADA POR DB2 O EN EL ACELERADOR."`

Nuggets de modelo de árbol de decisiones de IBM Db2 para z/OS

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.

Al ejecutar una ruta que contiene un nugget de modelo de árbol de decisión, el nodo añade dos campos nuevos, cuyos nombres se derivan del destino.

Tabla 26. Campo de puntuación de modelo del árbol de decisión

Nombre del campo añadido	**Descripción**
$I-*nombre_objetivo*	Valor predicho del registro actual.
$IP-*nombre_objetivo*	Valor de confianza (entre 0,0 y 1,0) de la predicción.

Nota: Debido a limitaciones en Db2 para z/OS, los nombres de columna podrían estar truncados.

Nugget de árbol de decisiones de IBM Db2 for z/OS - 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.

Nugget de árbol de decisiones de IBM Db2 para z/OS - Pestaña Visor

La pestaña **Visor** muestra una presentación de árbol del modelo de árbol del mismo modo en que lo hace SPSS Modeler para su modelo de árbol de decisión.

Nugget de modelo de k-medias de IBM Db2 para z/OS

Los nugget de modelo de K-medias contienen toda la información capturada por el modelo de agrupación en cluster, 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`.

Nota: Debido a limitaciones en Db2 para z/OS, los nombres de columna podrían estar truncados.

Nugget de k-medias de IBM Db2 para z/OS - 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.

Nuggets de modelo de bayesiano ingenuo de IBM Db2 para z/OS

Si ejecuta una ruta con un nugget de modelo bayesiano ingenuo, el nodo añade dos campos nuevos, cuyos nombres se derivan del nombre del objetivo.

Tabla 27. Campo de puntuación de modelo bayesiano ingenuo

Nombre del campo añadido	**Descripción**
$I-*nombre_objetivo*	Valor predicho del registro actual.
$IP-*nombre_objetivo*	Valor de confianza (entre 0,0 y 1,0) de la predicción.

Nota: Debido a limitaciones en Db2 para z/OS, los nombres de columna podrían estar truncados.

Puede ver los campos extra si adjunta un nodo Tabla al nugget de modelo y ejecuta el nodo Tabla.

Nuggets de modelo de árbol de regresión de IBM Db2 para z/OS

Al ejecutar una ruta que contiene un nugget de modelo de árbol de regresión, el nodo añade dos campos nuevos, cuyos nombres se derivan del nombre de destino.

Tabla 28. Campo de puntuación de modelo del árbol de regresión

Nombre del campo añadido	**Descripción**
$I-*nombre_objetivo*	Valor predicho del registro actual.
$IS-*nombre_destino*	Desviación estándar estimada del valor predicho.

Nota: Debido a limitaciones en Db2 para z/OS, los nombres de columna podrían estar truncados.

Puede ver los campos extra si adjunta un nodo Tabla al nugget de modelo y ejecuta el nodo Tabla.

Nugget de árbol de regresión de IBM Db2 para z/OS - 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.

Nugget de árbol de regresión de IBM Db2 para z/OS - Pestaña Visor

La pestaña **Visor** muestra una presentación de árbol del modelo de árbol del mismo modo en que lo hace SPSS Modeler para su modelo de árbol de regresión.

Nugget de modelo bietápico de IBM Db2 para z/OS

Cuando ejecuta una ruta que contiene un nugget de modelo TwoStep, el nodo añade dos campos nuevos que contienen la información de pertenencia al clúster y la distancia respecto al centro del clúster asignado correspondiente al registro. Los nuevos nombres de archivo se obtienen a partir del nombre del modelo, con el prefijo $TS- para la información de pertenencia al clúster y $TSD- para la distancia respecto del centro del clúster. Por ejemplo, si el modelo se llama MDL, los nuevos campos se llamarían $TS-MDL y $TSD-MDL.

Avisos

Esta información se ha desarrollado para productos y servicios ofrecidos en Estados Unidos. Es posible que este material esté disponible en IBM en otros idiomas. Sin embargo, deberá poseer una copia del producto o de la versión del producto en ese idioma para poder acceder a él.

Es posible que IBM no ofrezca los productos, servicios o características que se tratan en este documento en otros países. El representante local de IBM le puede informar sobre los productos y servicios que están actualmente disponibles en su localidad. Las referencias a productos, programas o servicios de IBM no pretenden establecer ni implicar que sólo puedan utilizarse dichos productos, programas o servicios de IBM. En su lugar, se puede utilizar cualquier producto, programa o servicio equivalente que no infrinja ninguno de los derechos 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.

Es posible que IBM tenga patentes o solicitudes de patente pendientes que traten el tema descrito en este documento. La entrega de este documento no le otorga ninguna licencia sobre dichas patentes. Puede enviar preguntas acerca de las licencias, por escrito, a:

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

Para consultas de licencia relativas a la información de doble byte (DBCS), póngase en contacto con el departamento de propiedad intelectual de IBM de su país, o envíe las consultas, por escrito, a:

Intellectual Property Licensing
Derecho de propiedad intelectual y legal
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

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

Esta información podría contener imprecisiones técnicas o errores tipográficos. Periódicamente se efectúan cambios en la información aquí contenida; estos cambios se incorporarán en nuevas ediciones de la publicación. IBM se reserva el derecho a realizar en cualquier momento y sin notificación previa, mejoras o modificaciones en los productos y programas que se describen en el presente manual.

Las referencias en esta información a sitios web no IBM se proporcionan solo por comodidad y en modo alguno constituyen un aval de dichos sitios web. El material de esos sitios web no forma parte del material de este producto de IBM y el uso de dichos sitios web es responsabilidad del usuario.

IBM puede utilizar o distribuir cualquier información que proporcione en la forma que crea adecuada sin incurrir por ello en ninguna obligación con el cliente.

Los titulares 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 intercambiada, deben ponerse en contacto con:

IBM Director of Licensing
IBM Corporation

North Castle Drive, MD-NC119
Armonk, NY 10504-1785
EE.UU.

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 y los ejemplos de cliente se presentan únicamente con fines ilustrativos. Los resultados reales del rendimiento pueden variar en función de determinadas configuraciones y condiciones de funcionamiento.

La información acerca de productos que no son de IBM se ha obtenido de los proveedores de dichos productos, sus publicaciones anunciadas u otras fuentes disponibles públicamente. IBM no ha comprobado estos productos y no puede confirmar la precisión de su rendimiento, compatibilidad ni contemplar ninguna otra reclamación relacionada con los productos que no son de IBM. Las preguntas acerca de las funciones de productos que no sean de IBM deben dirigirse a los proveedores de dichos productos.

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

Esta información contiene ejemplos de datos e informes utilizados en operaciones comerciales diarias. Para ilustrarlos de la forma más completa posible, los ejemplos incluyen nombres de personas, empresas, marcas y productos. Todos estos nombres son ficticios y cualquier similitud con nombres de personas o empresas es mera coincidencia.

Marcas comerciales

IBM, el logotipo de IBM e ibm.com son marcas comerciales o marcas registradas de International Business Machines Corp., registradas en muchas jurisdicciones en todo el mundo. Otros nombres de productos y servicios pueden ser marcas registradas de IBM o de otras empresas. Existe una lista actual de las marcas registradas de IBM en la web bajo "Copyright and trademark information" en www.ibm.com/legal/copytrade.shtml.

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

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

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

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

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

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

Términos y condiciones de la documentación de producto

Los permisos para utilizar estas publicaciones se otorgan de acuerdo con los términos y condiciones siguientes.

Ámbito de aplicación

Estos términos y condiciones son adicionales a los términos de uso del sitio web de IBM.

Uso personal

Puede reproducir estas publicaciones para el uso personal, no comercial puesto que se conservan los avisos de propiedad. No puede distribuir, visualizar o realizar trabajos derivados de estas publicaciones, o de partes de las mismas, sin el consentimiento expreso de IBM.

Uso comercial

Puede reproducir, distribuir y mostrar estas publicaciones exclusivamente en su empresa siempre que se conserven todos los avisos de propiedad. No podrá crear trabajo derivado de estas publicaciones, ni reproducir, distribuir ni visualizar estas publicaciones o cualquier parte de éstas sin el consentimiento expreso de IBM.

Derechos

Salvo que se indique expresamente lo contrario en este permiso, no se otorga ningún otro permiso, licencia o derecho, ya sea de manera implícita o explícita, a las publicaciones ni a cualquier información, datos, software u otros elementos sujetos a la propiedad intelectual contenidos en ella.

IBM se reserva el derecho a retirar los permisos otorgados en ellas siempre que, a su discreción, el uso de las publicaciones vaya en detrimento de sus intereses o no se sigan adecuadamente las instrucciones anteriores, tal como determina IBM.

No puede descargar, exportar ni volver a exportar esta información excepto con plena conformidad de todas las leyes y regulaciones aplicables, incluyendo la legislación y las regulaciones de exportación de EE.UU.

IBM NO GARANTIZA EL CONTENIDO DE ESTAS PUBLICACIONES. LAS PUBLICACIONES SE PROPORCIONAN "TAL CUAL" Y SIN GARANTÍA DE NINGUNA CLASE, NI EXPLÍCITA NI IMPLÍCITA, INCLUYENDO PERO SIN LIMITARSE A LAS GARANTÍAS IMPLÍCITAS DE COMERCIALIZACIÓN, NO VULNERACIÓN E IDONEIDAD PARA UN FIN DETERMINADO.

Índice

Caracteres Especiales

Árbol de decisiones
    IBM Db2 for z/OS 91, 92, 96–98
    IBM Netezza Analytics 59–62, 75, 76, 81
    Oracle Data Mining 34, 35
árboles de decisión
    Microsoft Analysis Services 9, 11, 19
    opciones de experto 16
    opciones de modelo 15
    opciones de servidor 15
    puntuación - opciones de resumen 20
    puntuación - opciones de servidor 20
árboles de regresión
    IBM Db2 for z/OS 93, 94, 97
    IBM Netezza Analytics 53, 54, 81

A

agrupación en clústeres
    IBM Netezza Analytics 79, 80
    opciones de experto 16
    opciones de modelo 15
    opciones de servidor 15
    puntuación - opciones de resumen 20
    puntuación - opciones de servidor 20
agrupación en clústeres de secuencias
    opciones de modelo 15
agrupación en clústeres de secuencias (Microsoft)
    opciones de campos 19
    opciones de experto 19
agrupación en clústeres divisiva
    IBM Netezza Analytics 55
análisis espectral, IBM Netezza Analytics 66
Analysis Services
    Árboles de decisión 22
    ejemplos 22
    gestión de modelos 13
Apriori
    Microsoft 16
    Oracle Data Mining 38, 39
archivo tnsnames.ora 26

B

bayesiano ingenuo
    opciones de experto 16
    opciones de modelo 15
    opciones de servidor 15
    puntuación - opciones de resumen 20
    puntuación - opciones de servidor 20
Bayesiano ingenuo
    IBM Db2 for z/OS 91, 97
    IBM Netezza Analytics 65, 78
    Oracle Data Mining 29

C

campo exclusivo
    Apriori de Oracle 35, 39
    Bayesiano ingenuo de Oracle 29
    K-medias de Oracle 37
    LMD de Oracle 40
    máquina de vectores de soporte de Oracle 31
    NMF de Oracle 38
    O-clúster de Oracle 36
    Oracle Data Mining 27
    red de bayesiano adaptativo de Oracle 30
campos de partición
    selección 39
Clúster divisivo
    IBM Netezza Analytics 79, 80
configurar
    IBM Db2 para z/OS e IBM Analytics Accelerator for z/OS 86
costes
    Oracle 28
costes de clasificación errónea
    Oracle 28
criterio de división
    K-medias de Oracle 37

D

database
    modelado interno de bases de datos 6, 9, 11, 13, 19
datos de intervalo
    modelos de Oracle 44
datos de normalización
    modelos de Oracle 44
descomposición de tendencias estacional, IBM Netezza Analytics 66
despliegue 24, 46
desviación estándar
    máquina de vectores de soporte de Oracle 32
documentación 3
DSN
    configurar 11

E

ejemplos
    Guía de aplicaciones 3
    minería de bases de datos 22–24, 45
    visión general 4
ejemplos de aplicaciones 3
epsilon
    máquina de vectores de soporte de Oracle 32
etiqueta de clase, en modelos de árbol de Netezza 59, 91
evaluación 24, 46
exploración 23, 45
exportación
    modelos de Analysis Services 22

F

factor de complejidad
  máquina de vectores de soporte de Oracle 32
función de distancia
  K-medias de Oracle 37

G

generación de nodos 22
generación de SQL 6

H

hoja, en modelos de árbol de Netezza 59, 91

I

IBM
  gestión de modelos 53
IBM Db2 for z/OS
  Árbol de regresión 93
  Árboles de decisión 91
  Bayesiano ingenuo 91
  configurar con IBM SPSS Modeler 88, 89
  configurar IBM Db2 para z/OS e IBM Analytics Accelerator for z/OS 86
  gestionar modelos Db2 para z/OS 95, 96
  integración con IBM Db2 Analytics Accelerator for z/OS 85
  K-medias 89
  Nugget de modelo bayesiano ingenuo 97
  nugget de modelo Bietápico 95, 98
  Nugget de modelo de árbol de regresión 97
  Nugget de modelo de K-medias 97
  Nugget de modelo del árbol de decisión 96–98
  Opciones de campo para K-medias 90
  Opciones de campo para los árboles de decisión 91
  opciones de campo TwoStep 95
  opciones de campos 89
  Opciones de generación de árbol de decisión 91, 92
  Opciones de generación de árbol de regresión 93, 94
  opciones de generación del modelo bietápico 95
  Opciones de generación para K-medias 90
  opciones de modelo 89
  requisitos para la integración con IBM Db2 para z/OS 85
  TwoStep 94
IBM Netezza Analytics
  Árbol de regresión 53
  Árboles de decisión 59
  Bayesiano ingenuo 65
  Clúster divisivo 55
  configurar con IBM SPSS Modeler 47, 48, 50, 52
  gestión de modelos 74, 75
  K-medias 64
  Lineal generalizado 56
  Modelo de modelo de regresión lineal 82
  Nugget de modelo bayesiano ingenuo 78
  nugget de modelo Bietápico 83, 84
  Nugget de modelo de agrupación en clústeres divisivo 79, 80
  Nugget de modelo de árbol de regresión 81
  Nugget de modelo de K-medias 76, 77
IBM Netezza Analytics *(continuación)*
  Nugget de modelo de red bayesiana 77
  Nugget de modelo del árbol de decisión 75, 76, 81
  Nugget de modelo KNN 78, 79
  Nugget de modelo lineal generalizado 56, 83
  Nugget de modelo PCA 80
  nugget de modelo Serie temporal 82
  Opciones de campo de clúster divisivo 55
  Opciones de campo de PCA 73
  Opciones de campo de red bayesiana 66
  Opciones de campo para K-medias 64
  Opciones de campo para los árboles de decisión 60
  opciones de campo TwoStep 72
  opciones de campos 51
  opciones de campos de Series temporales 69
  Opciones de generación de árbol de decisión 61, 62
  Opciones de generación de árbol de regresión 53, 54
  Opciones de generación de clúster divisivo 55
  Opciones de generación de PCA 74
  Opciones de generación de red bayesiana 66
  Opciones de generación de regresión lineal 62
  opciones de generación de Series temporales 69, 71
  opciones de generación del modelo bietápico 73
  Opciones de generación para K-medias 65
  opciones de modelo 52
  Opciones de modelo KNN 63, 64
  Opciones de modelo lineal generalizado 57, 58
  Opciones del modelo de serie temporal 72
  PCA 73
  Red bayesiana 66
  Regresión lineal 62
  Serie temporal 66
  TwoStep 72
  Vecinos más cercanos (KNN) 63
IBM SPSS Modeler
  documentación 3
  minería de bases de datos 5
IBM SPSS Modeler Server 1
IBM SPSS Modeler Solution Publisher
  modelos de Oracle Data Mining 27
id de seguridad
  conexión de Oracle 26
Importancia del atributo (AI)
  Oracle Data Mining 41
interpolación de valores, series temporales de IBM Netezza Analytics 67

K

K-medias
  IBM Db2 for z/OS 89, 90, 97
  IBM Netezza Analytics 64, 65, 76, 77
  Oracle Data Mining 36, 37
kernel gaussiano
  máquina de vectores de soporte de Oracle 31
kernel lineal
  máquina de vectores de soporte de Oracle 31

L

Llave
  claves de modelos 7
LMD 30

longitud mínima de la descripción 30
Longitud mínima de la descripción (LMD)
  Oracle Data Mining 40

M

máquina de vectores de soporte
  Oracle Data Mining 31, 32
medida de impureza de entropía 61
medida de impureza Gini 61
medidas de impureza
  Árbol de decisión de Netezza 61
  Árbol de decisiones 91
método de normalización
  K-medias de Oracle 37
  máquina de vectores de soporte de Oracle 31
  NMF de Oracle 38
métrica de la impureza
  Apriori de Oracle 35
Microsoft
  Analysis Services 9, 11, 19
  Clúster de secuencia 9
  gestión de modelos 13
  modelado bayesiano ingenuo 9, 11, 19
  modelado de Agrupación en clústeres 9, 11, 19
  modelado de árboles de decisión 9, 11, 19
  modelado de red neuronal 11, 19
  modelado de reglas de asociación 9, 11, 19
  modelado de regresión lineal 11, 19
  Modelado de regresión logística 11, 19
  Red neuronal 9
  Regresión lineal 9
  Regresión Logística 9
Microsoft Analysis Services 21, 22
mín.-máx.
  datos de normalización 31, 44
minería de bases de datos
  configuración 11
  creación de modelos 6
  ejemplo 22
  opciones de optimización 6
  preparación de datos 6
  utilizar IBM SPSS Modeler 5
modelado de bases de datos
  IBM Netezza Analytics 47, 48, 50, 52
  Oracle 25–28
modelado de Db2 para z/OS
  IBM Db2 for z/OS 85, 88, 89
modelado interno de bases de datos 20
modelos
  administración de Analysis Services 13
  almacenamiento 7
  aspectos de coherencia 7
  creación de modelo internos de la base de datos 6
  evaluación 24, 46
  exploración de Oracle 30
  exportación 7
  gestionar Netezza 53
  listado de Netezza 53
  puntuación interna modelos de la base de datos 6
modelos ARIMA
  IBM Netezza Analytics 66, 70
modelos bayesiano ingenuo podados
  red de bayesiano adaptativo de Oracle 30
modelos bayesiano ingenuo podados *(continuación)*
  red de bayesiano adaptativo de Oracle 30
modelos de bayesiano ingenuo
  IBM Netezza Analytics 78
  red de bayesiano adaptativo de Oracle 30
Modelos de redes bayesianas
  IBM Netezza Analytics 66, 77
modelos de reglas de asociación
  Microsoft 16
modelos del vecino más próximo
  IBM Netezza Analytics 63, 64, 78, 79
modelos KNN
  IBM Netezza Analytics 78, 79
modelos lineales generalizados
  IBM Netezza Analytics 56–59, 83
Modelos lineales generalizados (GLM)
  Oracle Data Mining 33, 34
modelos mono-característica
  red de bayesiano adaptativo de Oracle 30
modelos multi-característica
  red de bayesiano adaptativo de Oracle 30
modelos PCA
  IBM Netezza Analytics 73, 74, 80

N

Netezza
  gestión de modelos 53
NMF
  Oracle Data Mining 37, 38
nodo Auditoría de datos 23, 45
nodo Editor
  modelos de Oracle Data Mining 27
nodos
  generación 22
nodos de modelado
  Agrupación en clústeres de Microsoft 13
  Agrupación en clústeres de secuencias de Microsoft 13
  Árboles de decisión de Microsoft 13
  bayesiano ingenuo de Microsoft 13
  modelado interno de bases de datos 6, 9, 11, 13, 19
  red neuronal de Microsoft 13
  reglas de asociación de Microsoft 13
  regresión lineal de Microsoft 13
  regresión logística de Microsoft 13
  Series temporales de Microsoft 13
NombreHost
  conexión de Oracle 26
nuggets de modelo
  IBM Db2 for z/OS 95–98
  IBM Netezza Analytics 56, 75–84
número de clústeres
  K-medias de Oracle 37
  O-clúster de Oracle 36

O

O-clúster
  Oracle Data Mining 36
ODBC
  configuración de SQL Server 11
  configuración para Oracle 25–28
  configurar 11

ODBC *(continuación)*
configurar para IBM Db2 para z/OS 89
configurar para IBM Netezza Analytics 47, 48, 50, 52
ODM. Consulte Oracle Data Mining 25
opciones de campos
IBM Db2 for z/OS 89–91, 95
IBM Netezza Analytics 51, 55, 60, 64, 66, 69, 72–74
opciones de generación
IBM Db2 for z/OS 90–95
IBM Netezza Analytics 53–55, 61, 62, 65, 66, 69, 71, 73
opciones de modelo
IBM Db2 for z/OS 89
IBM Netezza Analytics 52, 57, 58, 63, 64, 72
Oracle Data Miner 43
Oracle Data Mining
Apriori 38, 39
Árbol de decisiones 34, 35
Bayesiano ingenuo 29
comprobación de la coherencia 42
configurar con IBM SPSS Modeler 25–28
costes de clasificación errónea 43
ejemplos 44–46
gestión de modelos 42, 43
Importancia del atributo (AI) 41
K-medias 36, 37
Longitud mínima de la descripción (LMD) 40
máquina de vectores de soporte 31, 32
Modelos lineales generalizados (GLM) 33, 34
NMF 37, 38
O-clúster 36
preparación de datos 44
red de bayesiano adaptativo 30, 31

P

partición de datos 39
penalización de complejidad 16–18
ponderación de clase, en modelos de árbol de Netezza 59
ponderación de instancia, en modelos de árbol de Netezza 59
probabilidades previas
Oracle Data Mining 32
puerto
conexión de Oracle 26
puntuación 6, 74, 96
puntuaciones z
datos de normalización 31, 44

R

red de bayesiano adaptativo
Oracle Data Mining 30, 31
red neuronal
opciones de experto 16
opciones de modelo 15
opciones de servidor 15
puntuación - opciones de resumen 20
puntuación - opciones de servidor 20
reglas de asociación
opciones de experto 17
opciones de modelo 15
opciones de servidor 15
puntuación - opciones de resumen 20
reglas de asociación *(continuación)*
puntuación - opciones de servidor 20
regresión lineal
IBM Db2 for z/OS 93
IBM Netezza Analytics 53, 62, 82
opciones de experto 16
opciones de modelo 15
opciones de servidor 15
puntuación - opciones de resumen 20
puntuación - opciones de servidor 20
Regresión logística
opciones de experto 16
opciones de modelo 15
opciones de servidor 15
puntuación - opciones de resumen 20
puntuación - opciones de servidor 20
requisitos
IBM Db2 for z/OS
85

S

Serie temporal
IBM Netezza Analytics 69, 71, 72
Serie temporal (IBM Netezza Analytics) 66
servidor
ejecución de Analysis Services 15, 20
Solution Publisher
modelos de Oracle Data Mining 27
SQL Server
Conexión ODBC 11
configurar 11
suavizado exponencial
IBM Netezza Analytics 66
SVM. Consulte Máquina de vectores de soporte 31

T

time series (IBM Netezza Analytics) 82
time series (Microsoft)
opciones de configuración 18
opciones de experto 18
opciones de modelo 17
tolerancia de convergencia
máquina de vectores de soporte de Oracle 32
twostep
IBM Db2 for z/OS 94, 95
IBM Netezza Analytics 72
TwoStep
IBM Db2 for z/OS 98
IBM Netezza Analytics 72, 73, 83, 84

U

umbral de singleton
Bayesiano ingenuo de Oracle 29
umbral por parejas
Bayesiano ingenuo de Oracle 29

V

validación cruzada
Bayesiano ingenuo de Oracle 29