Guia de mineração em banco de dados do IBM SPSS Modeler 18.6

Nota

Antes de utilizar essas informações e o produto que elas suportam, leia as informações em “Avisos” na página 99.

Informações do produto

Esta edição se aplica à versão 18, release 4, modificação 0 de IBM® SPSS Modeler e a todos os lançamentos e modificações subsequentes até indicado de outra forma em novas edições.

© Copyright International Business Machines Corporation .

Índice

Prefácio

IBM SPSS Modeler é o ambiente de trabalho de mineração de dados de força corporativa IBM . O ModeladorSPSS ajuda as organizações a melhorarem as relações com o cliente e com o cidadão por meio de um entendimento profundo dos dados. As organizações utilizam o insight adquirido do ModeladorSPSS para reter clientes rentáveis, identificar oportunidades de venda cruzada, atrair novos clientes, detectar fraude, reduzir o risco e melhorar a entrega de serviço de governo.

A interface visual do ModeladorSPSS convida os usuários a aplicarem seus conhecimentos de negócios específicos, levando a modelos preditivos mais poderosos e reduzindo o tempo para a solução. O ModeladorSPSS oferece muitas técnicas de modelagem, como predição, classificação, segmentação e algoritmos de detecção de associação. Quando os modelos são criados, o Editor de soluçõesIBM SPSS Modeler permite entregá-los aos tomadores de decisão na empresa ou a um banco de dados.

Sobre o IBM Business Analytics

O software IBM Business Analytics fornece informações completas, consistentes e exatas nas quais os tomadores de decisão confiam para melhorar o desempenho de negócios. Um portfólio abrangente de inteligência de negócios, analise preditiva, gerenciamento de desempenho financeiro e estratégiae aplicativos analíticos fornecem insight claro, imediato e prático sobre o desempenho atual e a capacidade de prever resultados futuros. Combinado com soluções para segmentos do mercado, práticas comprovadas e serviços profissionais completos, organizações de qualquer tamanho poderão conduzir maior produtividade, automatizar as decisões de modo confiável e entregar melhores resultados.

Como parte deste dossier, o software IBM SPSS Predictive Analytics ajuda as organizações a prever futuros eventos e agir proativamente com esse insight para melhores resultados de negócios. Os clientes acadêmicos, comerciais e do governo no mundo todo se baseiam na tecnologia do IBM SPSS como uma vantagem competitiva para atrair, manter e aumentar seus clientes, enquanto reduz fraudes e minimiza riscos. Ao incorporar o software IBM SPSS em suas operações diárias, as organizações tornam-se empreendimentos preditivos-capazes de direcionar e automatizar as decisões para cumprir metas de negócios e obter vantagem competitiva mensurável. Para obter informações adicionais ou para entrar em contato com um representante, visite http://www.ibm.com/spss.

Suporte técnico

O suporte técnico está disponível para manutenção dos clientes. Os clientes podem entrar em contato com o Suporte Técnico para assistência no uso de produtos IBM ou para ajuda de instalação para um dos ambientes de hardware suportados. Para chegar ao Suporte Técnico, consulte o site da IBM em http://www.ibm.com/support. Esteja preparado para se identificar, sua organização e seu contrato de suporte ao solicitar assistência.

Capítulo 1. SobreIBM SPSS Modeler

O IBM SPSS Modeler é um conjunto de ferramentas de mineração de dados que permite desenvolver rapidamente modelos preditivos usando o conhecimento de negócios, e implementá-los em operações de negócios para melhorar a tomada de decisão. Projetado em torno do modelo CRISP-DM padrão de mercado, o IBM SPSS Modeler suporta todo o processo de mineração de dados, a partir dos dados para melhores resultados de negócios.

OIBM SPSS Modeler oferece uma variedade de métodos de modelagem tomados do aprendizado de máquina, inteligência artificial e estatística. Os métodos disponíveis na paleta Modelagem permitem derivar informações novas a partir dos dados, e desenvolver modelos preditivos. Cada método possui certas forças e é mais adequado para certos tipos de problemas.

O ModeladorSPSS pode ser comprado como um produto independente, ou usado como um cliente na combinação com o Servidor doSPSS Modeler. Várias opções adicionais também estão disponíveis, conforme resumidas nas seções a seguir. Para mais informações, consulte https://www.ibm.com/analytics/us/en/technology/spss/.

Produtos IBM SPSS Modeler

A família de produtos IBM SPSS Modeler e o software associado abrangem o seguinte.

- IBM SPSS Modeler
- ServidorIBM SPSS Modeler
- Console de administraçãoIBM SPSS Modeler (incluído com IBM SPSS Deployment Manager)
- IBM SPSS Modeler Batch
- Editor de soluçõesIBM SPSS Modeler
- ServidorIBM SPSS Modeler adaptadores para Serviços de Colaboração e ImplementaçãoIBM SPSS

IBM SPSS Modeler

ModeladorSPSS é uma versão funcionalmente completa do produto que você instala e executa em seu computador pessoal. É possível executar o ModeladorSPSS no modo local como um produto independente ou usá-lo no modo distribuído com ServidorIBM SPSS Modeler para melhorar o desempenho em conjuntos de dados grandes.

Com o ModeladorSPSS, é possível construir modelos preditivos exatos de maneira rápida e intuitiva, sem programação. Usando a interface visual exclusiva, é possível visualizar facilmente o processo de mineração de dados. Com o suporte da análise avançada integrada ao produto, é possível descobrir tendências e padrões ocultos anteriormente em seus dados. É possível modelar resultados e entender os fatores que os influenciam, permitindo que você aproveite as vantagens das oportunidades de negócios e diminua os riscos.

ModeladorSPSS está disponível em duas edições: SPSS Modeler Professional e SPSS Modeler Premium. Consulte o tópico “Edições do IBM SPSS Modeler” na página 2 para obter mais informações.

ServidorIBM SPSS Modeler

ModeladorSPSS usa uma arquitetura de cliente/servidor para distribuir solicitações para operações cheias de recursos para poderosos softwares de servidor, resultando em desempenho mais rápido em conjuntos de dados maiores.

Servidor doSPSS Modeler é um produto licenciado separadamente que é executado de forma contínua no modo de análise distribuído em um host do servidor com uma ou mais instalações do IBM SPSS Modeler. Dessa maneira, o Servidor doSPSS Modeler fornece desempenho superior em conjuntos de dados grandes, pois operações com uso intensivo de memória podem ser executadas no servidor sem

fazer download dos dados no computador cliente. ServidorIBM SPSS Modeler também fornece suporte para otimização de SQL e capacidades de modelagem dentro da base de dados, entregando mais benefícios para o desempenho e a automação.

Console de administraçãoIBM SPSS Modeler

O Console de administração do Modeler é uma interface gráfica de usuário para o gerenciamento de muitas das opções de configuração Servidor doSPSS Modeler , que também são configuráveis por meio de um arquivo de opções. O console é incluído em IBM SPSS Deployment Manager, pode ser usado para monitorar e configurar suas instalações Servidor doSPSS Modeler , e está disponível gratuitamente para os clientes atuais Servidor doSPSS Modeler . O aplicativo pode ser instalado somente em computadores Windows; no entanto, ele pode administrar um servidor instalado em qualquer plataforma suportada.

IBM SPSS Modeler Batch

Embora geralmente a mineração de dados seja um processo interativo, também é possível executar o ModeladorSPSS a partir de uma linha de comandos, sem a necessidade de uma interface gráfica com o usuário. Por exemplo, você pode ter tarefas repetidas ou de longa execução que deseja executar sem intervenção do usuário. SPSS Modeler Batch é uma versão especial do produto que fornece suporte para capacidades de análise completa do ModeladorSPSS sem acessar a interface com o usuário regular. Servidor doSPSS Modeler é necessário para usar o SPSS Modeler Batch.

Editor de soluçõesIBM SPSS Modeler

Editor de soluçõesSPSS Modeler é uma ferramenta que permite criar uma versão do pacote de um fluxo do ModeladorSPSS que pode ser executado por um mecanismo de tempo de execução externo ou integrado a um aplicativo externo. Dessa maneira, é possível publicar e implementar fluxos completos do ModeladorSPSS para uso em ambientes que não têm o ModeladorSPSS instalado. Editor de soluçõesSPSS Modeler é distribuído como parte do serviço IBM SPSS Collaboration and Deployment Services - Pontuação, para o qual uma licença separada é necessária. Com essa licença, você recebe o Tempo de execuçãoSPSS Modeler Solution Publisher, que permite executar os fluxos publicados.

Para obter mais informações sobre Editor de soluçõesSPSS Modeler, consulte a documentação do Serviços de Colaboração e ImplementaçãoIBM SPSS. A Serviços de Colaboração e ImplementaçãoIBM SPSS IBM Documentation contém seções chamadas "IBM SPSS Modeler Solution Publisher" e "IBM SPSS Analytics Toolkit."

ServidorIBM SPSS Modeler Adaptadores para Serviços de Colaboração e ImplementaçãoIBM SPSS

Inúmeros adaptadores para o Serviços de Colaboração e ImplementaçãoIBM SPSS estão disponíveis para permitir que o ModeladorSPSS e o Servidor doSPSS Modeler interajam com um repositório do Serviços de Colaboração e ImplementaçãoIBM SPSS. Dessa forma, um fluxo do ModeladorSPSS implementado no repositório pode ser compartilhado por diversos usuários ou acessado a partir do aplicativo thin client VantagensIBM SPSS Modeler. Você instala o adaptador no sistema que hospeda o repositório.

Edições do IBM SPSS Modeler

ModeladorSPSS está disponível nas seguintes edições.

SPSS Modeler Professional

SPSS Modeler Professional fornece todas as ferramentas necessárias para você trabalhar com a maioria dos tipos de dados estruturados, como comportamentos e interações controlados em sistemas CRM, demográficos, comportamento de compra e dados de vendas.

SPSS Modeler Premium

SPSS Modeler Premium é um produto licenciado separadamente que se estende SPSS Modeler Professional para trabalhar com dados especializados e com dados de texto não estruturados. SPSS Modeler Premium inclui Análise de texto doIBM SPSS Modeler:

Análise de texto doIBM SPSS Modeler usa tecnologias de linguística avançada e processamento de linguagem natural (NLP) para processar rapidamente uma grande variedade de dados de texto não estruturados, extrair e organizar conceitos chave e agrupar esses conceitos em categorias. Categorias e conceitos extraídos podem ser combinados com dados estruturados existentes, como demográficos, e aplicados à modelagem usando o conjunto completo de ferramentas de mineração de dados do IBM SPSS Modeler para gerar decisões melhores e mais focadas.

AssinaturaIBM SPSS Modeler

AssinaturaIBM SPSS Modeler fornece todas as mesmas capacidades de analítica preditiva que o cliente IBM SPSS Modeler tradicional. Com a edição de Assinaturas, é possível fazer o download de atualizações do produto regularmente.

Documentação

A documentação está disponível no menu **Ajuda** em ' ModeladorSPSS. Isso abre a IBM Documentation on-line, que está sempre disponível fora do produto.

A documentação completa de cada produto (incluindo instruções de instalação) também está disponível em formato PDF. Consulte a página de suporte a seguir: **DocumentaçãoSPSS Modeler 18.6**.

Documentação do SPSS Modeler Professional

O conjunto de documentações do SPSS Modeler Professional (excluindo instruções de instalação) é o seguinte.

- **IBM SPSS Modeler User's Guide.** Introdução geral para usar ModeladorSPSS, incluindo como construir fluxos de dados, manipular valores ausentes, construir expressões CLEM , trabalhar com projetos e relatórios, e streams de pacotes para implementação em Serviços de Colaboração e ImplementaçãoIBM SPSS ou VantagensIBM SPSS Modeler.
- **Nós de Origem, de Processo e de Saída do IBM SPSS Modeler.** Descrições de todos os nós usados para ler, processar e emitir dados em diferentes formatos. Efetivamente, isso significa todos os nós além dos de modelagem.
- **Nós de Modelagem do IBM SPSS Modeler.** Descrições de todos os nós usados para criar modelos de mineração de dados. OIBM SPSS Modeler oferece uma variedade de métodos de modelagem tomados do aprendizado de máquina, inteligência artificial e estatística.
- **Guia de Aplicativos do IBM SPSS Modeler.** Os exemplos neste guia fornecem introduções sintetizadas e direcionadas para técnicas e métodos de modelagem específicos. Uma versão online deste guia também está disponível no menu Ajuda. Veja o tópico “Exemplos de Aplicação” na página 4 para obter mais informações.
- **Script e Automação Python do IBM SPSS Modeler.** Informações sobre como automatizar o sistema por meio de script Python, incluindo as propriedades que podem ser usadas para manipular nós e fluxos.
- **Guia de Implementação do IBM SPSS Modeler.** Informações sobre a execução de fluxos IBM SPSS Modeler como etapas de processamento de tarefas sob IBM SPSS Deployment Manager.
- **Guia de Mineração Dentro do Banco de Dados do IBM SPSS Modeler.** Informações sobre como usar o poder do seu banco de dados para melhorar o desempenho e ampliar o intervalo de capacidades analíticas por meio de algoritmos de terceiros.
- **Guia de Desempenho e de Administração do ServidorIBM SPSS Modeler.** Informações sobre como configurar e administrar o ServidorIBM SPSS Modeler.

- **Guia do Usuário do IBM SPSS Deployment Manager.** Informações sobre o uso da interface de usuário do console de administração incluídas no aplicativo Gerente de implantação para monitoramento e configuração ServidorIBM SPSS Modeler.
- **IBM SPSS Modeler Guia CRISP-DM.** Guia passo a passo para o uso da metodologia CRISP-DM para mineração de dados com ModeladorSPSS.
- **IBM SPSS Modeler Batch User's Guide.** Guia completo para o uso do IBM SPSS Modeler no modo em lote, incluindo detalhes da execução do modo em lote e argumentos de linha de comandos. Este guia está disponível somente em formato PDF.

SPSS Modeler Premium documentação

O conjunto de documentações do SPSS Modeler Premium (excluindo instruções de instalação) é o seguinte.

- **Análise de texto doSPSS Modeler User's Guide.** Informações sobre o uso de analítica de texto com ModeladorSPSS, cobrindo os nós de mineração de texto, ambiente de trabalho interativo, modelos e outros recursos.

Exemplos de Aplicação

Enquanto as ferramentas de mineração de dados no ModeladorSPSS podem ajudar a resolver uma ampla variedade de negócios e problemas organizacionais, os exemplos de aplicativos fornecem introduções breves e destinadas aos métodos e técnicas de modelagem específicos. Os conjuntos de dados utilizados aqui são muito menores do que as enormes lojas de dados gerenciadas por alguns mineiros de dados, mas os conceitos e métodos que estão envolvidos são escaláveis para aplicações do mundo real.

Para acessar os exemplos, clique em **Exemplos de aplicativos** no menu Ajuda em ModeladorSPSS.

Os arquivos de dados e os fluxos de amostra são instalados na pasta `Demos` no diretório de instalação do produto. Para obter mais informações, consulte “Pasta Demos” na página 4.

Exemplos de modelagem da base de dados. Consulte os exemplos no *Guia de Mineração dentro do Banco de Dados do IBM SPSS Modeler*.

Exemplos de script. Consulte os exemplos no *Guia de Script e Automação do IBM SPSS Modeler*.

Pasta Demos

Os arquivos de dados e fluxos de amostra que são utilizados com os exemplos de aplicação são instalados na pasta `Demos` sob o diretório de instalação do produto (por exemplo: `C:\Program Files\IBM\SPSS\Modeler\<version>\Demos`). Esta pasta também pode ser acessada a partir do grupo de programas IBM ModeladorSPSS no menu Iniciar do Windows, ou clicando em `Demos` na lista de diretórios recentes na caixa de diálogo **Arquivo** > **Open Stream** .

Rastreamento de Licença

Quando você usa o ModeladorSPSS, o uso sob licença é controlado e registrado em intervalos regulares. As métricas de licença que são registradas são *AUTHORIZED_USER* e *CONCURRENT_USER* e o tipo de métrica que é registrado depende do tipo de licença que você possui para o ModeladorSPSS.

Os arquivos de log que são produzidos podem ser processados pelo IBM License Metric Tool, do qual é possível gerar relatórios de uso sob licença.

Os arquivos de log de licença são criados no mesmo diretório onde os arquivos de log do Client log do ModeladorSPSS são registrados (por padrão, `%ALLUSERSPROFILE%/IBM/SPSS/Modeler/<version>/log`).

Capítulo 2. Mineração Dentro da Base de Dados

Visão Geral da Modelagem da Base de Dados

O ServidorIBM SPSS Modeler suporta a integração com ferramentas de mineração de dados e de modelagem que estão disponíveis a partir de fornecedores de banco de dados, incluindo IBM Netezza, Oracle Data Miner e Microsoft Analysis Services. É possível construir, escorar e armazenar modelos dentro do banco de dados-tudo isso de dentro do aplicativo IBM SPSS Modeler. Isso permite combinar os recursos de análise e a facilidade de uso do IBM SPSS Modeler com o poder e o desempenho de um banco de dados, enquanto aproveita os algoritmos nativos de banco de dados enviados por esses fornecedores. Os modelos são construídos dentro do banco de dados, que podem, então, ser procurados e escorados por meio da interface do IBM SPSS Modeler de maneira normal e podem ser implementados utilizando o Editor de soluçõesIBM SPSS Modeler, se necessário. Os algoritmos suportados estão na paleta Modelagem do Banco de Dados no IBM SPSS Modeler.

Usar o IBM SPSS Modeler para acessar os algoritmos nativos de banco de dados oferece várias vantagens:

- Os algoritmos dentro do banco de dados são muitas vezes estreitamente integrados com o servidor de base de dados e podem oferecer melhor desempenho.
- Os modelos construídos armazenados "dentro do banco de dados" podem ser mais facilmente implementados e compartilhados com qualquer aplicativo que possa acessar o banco de dados.

Geração SQL. A modelagem dentro da base de dados é diferente da geração SQL, de outra forma conhecida como “SQL Pushback”. Este recurso permite gerar instruções SQL para operações nativas do IBM SPSS Modeler que podem ser "enviadas por push" (ou seja, executadas no) para o banco de dados para aprimorar o desempenho. Por exemplo, os nós Mesclagem, Agregado e Seleção podem gerar código SQL que pode ser enviado por push de volta para o banco de dados desta maneira. O uso da geração de SQL em combinação com a modelagem da base de dados pode resultar em fluxos que podem ser executados do início ao fim no banco de dados, resultando em ganhos de desempenho significativos sobre os fluxos em execução no IBM SPSS Modeler.

Nota: A modelagem da base de dados e a otimização de SQL requerem que a conectividade do ServidorIBM SPSS Modeler esteja ativada no computador do IBM SPSS Modeler. Com essa configuração ativada, é possível acessar os algoritmos de banco de dados, realizar SQL pushback diretamente do IBM SPSS Modeler e acessar o ServidorIBM SPSS Modeler. Para verificar o atual status da licença, escolha o seguinte no menu do IBM SPSS Modeler.

Ajuda > **Sobre** > **Detalhes Adicionais**

Se a conectividade estiver ativada, você verá a opção **Ativação do Servidor** na guia Status da Licença.

Para obter informações sobre algoritmos suportados, consulte as seções subsequentes sobre fornecedores específicos.

O que é necessário

Para executar modelagem da base de dados, a configuração a seguir é necessária:

- Uma conexão ODBC a um banco de dados apropriado, com componentes analíticos necessários instalados (Microsoft Analysis Services ou Oracle Data Miner).
- No IBM SPSS Modeler, a modelagem de banco de dados deve ser ativada na caixa de diálogo Aplicativos do Helper (**Ferramentas> Aplicativos Helper**).
- As configurações de **Gerar SQL** e **Otimização de SQL** devem ser ativadas na caixa de diálogo Opções de Usuário no IBM SPSS Modeler e também no ServidorIBM SPSS Modeler (se usado). Observe que a otimização de SQL não é estritamente necessária para que a modelagem da base de dados funcione, mas é altamente recomendada por motivos de desempenho.

Nota: A modelagem da base de dados e a otimização de SQL requerem que a conectividade do ServidorIBM SPSS Modeler esteja ativada no computador do IBM SPSS Modeler. Com essa configuração ativada, é possível acessar os algoritmos de banco de dados, realizar SQL pushback diretamente do IBM SPSS Modeler e acessar o ServidorIBM SPSS Modeler. Para verificar o atual status da licença, escolha o seguinte no menu do IBM SPSS Modeler.

Ajuda > **Sobre** > **Detalhes Adicionais**

Se a conectividade estiver ativada, você verá a opção **Ativação do Servidor** na guia Status da Licença.

Para obter informações detalhadas, consulte as seções subsequentes sobre fornecedores específicos.

Construção de modelo

O processo de construção e de escoragem de modelos usando algoritmos de banco de dados é semelhante a outros tipos de mineração de dados no IBM SPSS Modeler. O processo geral de trabalhar com nós e de modelar "nuggets" é semelhante a qualquer outro fluxo quando trabalhar no IBM SPSS Modeler. A única diferença é que o processamento e a construção de modelo reais são enviados por push de volta para o banco de dados.

Um fluxo de modelagem de bando de dados é conceitualmente idêntico a outros fluxos de dados no IBM SPSS Modeler; no entanto, esse fluxo executa todas as operações em um banco de dados, incluindo, por exemplo, a construção de modelo utilizando o nó da Árvore de Decisão da Microsoft. Ao executar o fluxo, o IBM SPSS Modeler instrui o banco de dados a construir e a armazenar o modelo resultante, e os detalhes são transferidos por download para o IBM SPSS Modeler. A execução dentro da base de dados é indicada pelo uso de nós púrpuros sombreados no fluxo.

Preparação de Dados

Independentemente se os algoritmos nativos do banco de dados forem utilizados ou não, as preparações de dados devem ser enviadas por push de volta para o banco de dados sempre que possível a fim de melhorar o desempenho.

- Se os dados originais estiverem armazenados no banco de dados, o objetivo é mantê-los lá ao assegurar que toda as operações de envio de dados necessárias possam ser convertidas em SQL. Isso evitará que os dados sejam transferidos por download para o IBM SPSS Modeler – evitando um gargalo que possa anular quaisquer ganhos - e permitirá que o fluxo inteiro seja executado no banco de dados.
- Se os dados originais *não* estiverem armazenados no banco de dados, modelagem do banco de dados ainda poderá ser utilizada. Nesse caso, a preparação de dados é conduzida no IBM SPSS Modeler, e o conjunto de dados preparados é transferido por upload automaticamente para o banco de dados para construção de modelo.

Pontuação do modelo

Os modelos gerados a partir do IBM SPSS Modeler usando mineração dentro da base de dados são diferentes dos modelos regulares do IBM SPSS Modeler. Embora eles apareçam no Gerenciador de Modelos como “nuggets” do modelo gerado, na realidade eles são modelos remotos mantidos no servidor de mineração de dados ou de base de dados remoto. O que você vê no IBM SPSS Modeler são simplesmente referências a esses modelos remotos. Em outras palavras, o modelo IBM SPSS Modeler que você vê é um modelo "oco" que contém informações como o hostname do servidor de banco de dados, nome do banco de dados e o nome do modelo. É importante entender essa distinção conforme você navega e escora os modelos criados utilizando algoritmos nativos do banco de dados.

Após ter criado um modelo, ele poderá ser incluído no fluxo para escoragem como qualquer outro modelo gerado no IBM SPSS Modeler. Toda a escoragem é feita dentro do banco de dados, mesmo se as operações de envio de dados não forem. (as operações de envio de dados ainda podem ser enviadas por push de volta para o banco de dados, se possível, para melhorar o desempenho, mas isso não é um requisito para que a escoragem ocorra). Também é possível procurar o modelo gerado na maioria dos casos utilizando o navegador padrão enviado pelo fornecedor de base de dados.

Para tanto a navegação como a pontuação, é necessária uma conexão ao vivo com o servidor em execução do Oracle Data Miner ou Microsoft Analysis Services.

Visualizando Resultados e Especificando Configurações

Para visualizar resultados e especificar configurações para escoragem, dê um clique duplo no modelo na tela de fluxo. Como alternativa, é possível clicar com o botão direito no modelo e escolher **Procurar** ou **Editar**. Configurações específicas dependem do tipo de modelo.

Exportando e Salvando Modelos de Banco de Dados

Os modelos e sumarizações de banco de dados podem ser exportados do navegador do modelo da mesma maneira que outros modelos criados no IBM SPSS Modeler, utilizando as opções no menu Arquivo.

1. No menu Arquivo no navegador do modelo, escolha qualquer uma das seguintes opções:

- **Exportar Texto** exporta a sumarização do modelo em um arquivo de texto.
- **Exportar HTML** exporta a sumarização do modelo em um arquivo HTML.
- **Exportar PMML** (suportado para modelos do IBM DB2 IM somente) exporta o modelo como linguagem de marcações de modelo preditivo (PMML), que pode ser utilizada com outro software compatível com PMML.

Nota: Você também pode salvar um modelo gerado escolhendo **Salvar Nó** no menu Arquivo.

Consistência de Modelo

Para cada modelo de banco de dados gerado, o IBM SPSS Modeler armazena uma descrição da estrutura do modelo, junto de uma referência ao modelo com o mesmo nome que é armazenado no banco de dados. A guia Servidor de um modelo gerado exibe uma chave exclusiva gerada para esse modelo, que corresponde o modelo real no banco de dados.

O IBM SPSS Modeler utiliza essa chave gerada aleatoriamente para verificar se o modelo ainda é consistente. Esta chave é armazenada na descrição de um modelo quando ela é construída. É recomendado verificar se as chaves correspondem antes de executar um fluxo de implementação.

1. Para verificar a consistência do modelo armazenado no banco de dados ao comparar sua descrição com a chave aleatória armazenada pelo IBM SPSS Modeler, clique no botão **Verificar**. Se o modelo de banco de dados não puder ser localizado ou se a chave não corresponder, um erro será relatado.

Visualizando e Exportando SQL Gerada

O código do SQL gerado pode ser visualizado antes da execução, o que pode ser útil para propósitos de depuração.

Capítulo 3. Modelagem da base de dados com o Microsoft Analysis Services

IBM SPSS Modeler e Microsoft Analysis Services

O IBM SPSS Modeler suporta integração com o Microsoft SQL Server Analysis Services. Essa funcionalidade é implementada como nós de modelagem no IBM SPSS Modeler e está disponível a partir da paleta Modelagem da Base de Dados. Se a paleta não estiver visível, será possível ativá-la ao ativar a integração do MS Analysis Services, disponível na guia Microsoft, na caixa de diálogo Aplicativos Auxiliares. Consulte o tópico “Ativando a Integração com o Analysis Services” na página 11 para obter mais informações.

O IBM SPSS Modeler suporta integração dos seguintes algoritmos do Analysis Services:

- Árvores de decisão
- Armazenamento em cluster
- Regras de associação
- Naive Bayes
- Regressão linear
- Rede Neural
- Regressão Logística
- Séries temporais
- Armazenamento em Cluster de Sequências

O diagrama a seguir ilustra o fluxo de dados do cliente para o servidor no qual a mineração dentro da base de dados é gerenciada pelo ServidorIBM SPSS Modeler. A construção de modelo é executada utilizando o Analysis Services. O modelo resultante é armazenado pelo Analysis Services. Uma referência a este modelo é mantida dentro dos fluxos do IBM SPSS Modeler. Em seguida, o modelo é transferido por download do Analysis Services para o Microsoft SQL Server ou para o IBM SPSS Modeler para escoragem.

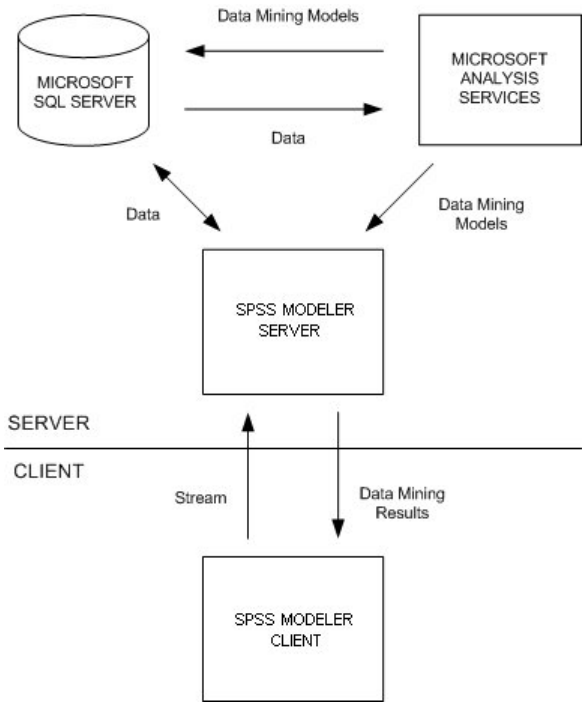

Figura 1. Fluxo de dados entre o IBM SPSS Modeler, o Microsoft SQL Server e o Microsoft Analysis Services durante a construção de modelo

Nota: o ServidorIBM SPSS Modeler não é necessário, embora ele possa ser utilizado. O cliente do IBM SPSS Modeler é capaz de processar cálculos de mineração dentro da base de dados por si só.

Requisitos para integração com o Microsoft Analysis Services

A seguir há pré-requisitos para conduzir modelagem dentro da base de dados utilizando algoritmos do Analysis Services com o IBM SPSS Modeler. Poderá ser necessário consultar seu administrador de base de dados para assegurar que essas condições sejam atendidas.

- IBM SPSS Modeler em execução com relação a uma instalação do ServidorIBM SPSS Modeler (modo distribuído) no Windows. Plataformas UNIX não são suportadas nesta integração com o Analysis Services.

  Importante: IBM SPSS Modeler usuários devem configurar uma conexão ODBC usando o driver de SQL Native Client disponível da Microsoft na URL listada abaixo em *Requisitos Adicionais ServidorIBM SPSS Modeler*. *O driver fornecido com o Pacote de acesso a dadosIBM SPSS (e geralmente recomendado para outros usos com o IBM SPSS Modeler) não é recomendado para este propósito.* O driver deve ser configurado para utilizar o SQL Server **Com Autenticação Integrada do Windows** ativados, uma vez que o IBM SPSS Modeler não suporta autenticação do SQL Server. Se você tiver questões referentes sobre como criar ou configurar permissões de origens de dados ODBC, entre em contato com seu administrador de banco de dados.
- SQL Server deve ser instalado, embora não necessariamente no mesmo host que IBM SPSS Modeler. Os usuários do IBM SPSS Modeler devem ter permissões suficientes para ler e gravar dados e eliminar e criar tabelas e visualizações.

  Nota: O SQL Server Enterprise Edition é recomendado. O Enterprise Edition oferece flexibilidade adicional ao fornecer parâmetros avançados para ajustar os resultados do algoritmo. A versão do Standard Edition fornece os mesmos parâmetros, mas não permite que os usuários editem alguns dos parâmetros avançados.
- O Microsoft SQL Server Analysis Services deve ser instalado no mesmo host que o SQL Server.

Requisitos adicionais do ServidorIBM SPSS Modeler

Para utilizar algoritmos do Analysis Services com o ServidorIBM SPSS Modeler, os seguintes componentes devem ser instalados na máquina host do ServidorIBM SPSS Modeler.

Nota: Se o SQL Server for instalado no mesmo host que o ServidorIBM SPSS Modeler, esses componentes já estarão disponíveis.

- Microsoft SQL Server Analysis Services 10.0 OLE DB Provider (certifique-se de selecionar a variante correta para seu sistema operacional)
- Microsoft SQL Server Native Client (certifique-se de selecionar a variante correta para seu sistema operacional)
- Se você estiver usando o Microsoft SQL Server 2008 ou 2012, também poderá precisar do Microsoft Core XML Services (MSXML) 6.0.

Para fazer o download desses componentes, acesse *www.microsoft.com/downloads*, procure por **.NET Framework** ou (para todos os outros componentes) **SQL Server Feature Pack** e selecione o pacote mais recente para sua versão do SQL Server.

Isso pode requerer que outros pacotes sejam instalados primeiro, que também deverão estar disponíveis no website Downloads da Microsoft.

Requisitos adicionais do IBM SPSS Modeler

Para utilizar algoritmos do Analysis Services com o IBM SPSS Modeler, os mesmos componentes devem ser instalados como acima, com a inclusão do seguinte no cliente:

- Microsoft SQL Server Datamining Viewer Controls (certifique-se de selecionar a variante correta para seu sistema operacional) - isto também requer:
- Microsoft ADOMD.NET

Para fazer o download desses componentes, acesse *www.microsoft.com/downloads*, procure por **SQL Server Feature Pack** e selecione o pacote mais recente para sua versão do SQL Server.

Nota: A modelagem da base de dados e a otimização de SQL requerem que a conectividade do ServidorIBM SPSS Modeler esteja ativada no computador do IBM SPSS Modeler. Com essa configuração ativada, é possível acessar os algoritmos de banco de dados, realizar SQL pushback diretamente do IBM SPSS Modeler e acessar o ServidorIBM SPSS Modeler. Para verificar o atual status da licença, escolha o seguinte no menu do IBM SPSS Modeler.

Ajuda > **Sobre** > **Detalhes Adicionais**

Se a conectividade estiver ativada, você verá a opção **Ativação do Servidor** na guia Status da Licença.

Ativando a Integração com o Analysis Services

Para ativar a integração do IBM SPSS Modeler com o Analysis Services, será necessário configurar o SQL Server e o Analysis Services, criar uma origem ODBC, ativar a integração na caixa de diálogo Aplicativos Auxiliares do IBM SPSS Modeler e ativar a geração e otimização SQL.

Nota: o Microsoft SQL Server e o Microsoft Analysis Services devem estar disponíveis. Veja o tópico “Requisitos para integração com o Microsoft Analysis Services” na página 10 para obter mais informações.

Configurando o SQL Server

Configure o SQL Server para permitir que a escoragem ocorra dentro do banco de dados.

1. Criar a chave de registro a seguir na máquina host do SQL Server:

```
HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\MSSQLServer\Providers\MSOLAP
```

2. Inclua o seguinte valor DWORD nessa chave:

```
AllowInProcess 1
```

3. Reinicie o SQL Server após fazer essa mudança.

Configurando o Analysis Services

Antes que o IBM SPSS Modeler possa se comunicar com o Analysis Services, deve-se primeiro definir manualmente duas configurações na caixa de diálogo Propriedades do Analysis Server:

1. Efetue login no Analysis Server por meio do MS SQL Server Management Studio.
2. Acesse a caixa de diálogo Propriedades ao clicar com o botão direito no nome do servidor e escolher **Propriedades**.
3. Marque a caixa de seleção **Mostrar Propriedades Avançadas (Todas)**.
4. Altere as seguintes propriedades:

- Altere o valor de `DataMining\AllowAdHocOpenRowsetQueries` para `True` (o valor padrão é `False`).
- Altere o valor de `DataMining\AllowProvidersInOpenRowset` para `[all]` (não há valor padrão).

Criando um DSN ODBC para o SQL Server

Para ler ou gravar em um banco de dados, você deverá ter uma origem de dados ODBC instalada e configurada para o banco de dados relevante e com permissões de leitura ou gravação, conforme necessário. O driver ODBC do Microsoft SQL Native Client é necessário e instalado automaticamente com o SQL Server. *O driver fornecido com o Pacote de acesso a dadosIBM SPSS (e geralmente recomendado para outros usos com o IBM SPSS Modeler) não é recomendado para este propósito.* Se o IBM SPSS Modeler e o SQL Server residirem em hosts diferentes, será possível fazer o download do driver ODBC Microsoft SQL Native Client. Veja o tópico "Requisitos para integração com o Microsoft Analysis Services" na página 10 para obter mais informações.

Se você tiver questões referentes sobre como criar ou configurar permissões de origens de dados ODBC, entre em contato com seu administrador de banco de dados.

1. Utilizando o driver ODBC do Microsoft SQL Native Client, crie um DSN do ODBC que aponta para o banco de dados do SQL Server utilizado no processo de mineração de dados. As configurações restantes do driver padrão devem ser utilizadas.
2. Para este DSN, assegure-se de que **Com Autenticação Integrada do Windows** esteja selecionada.

- Se o IBM SPSS Modeler e o ServidorIBM SPSS Modeler estiverem sendo executados em hosts diferentes, crie o mesmo DSN do ODBC em cada um dos hosts. Assegure-se de que o nome do mesmo DSN seja utilizado em cada host.

Habilitando a Integração de Serviços de Análise em IBM SPSS Modeler

Para permitir que o IBM SPSS Modeler use o Analysis Services, deve-se primeiro fornecer especificações do servidor na caixa de diálogo Aplicativos Auxiliares.

1. Nos menus do IBM SPSS Modeler, escolha:

   Ferramentas > **Opções** > **Aplicativos auxiliares**
2. Clique na guia **Microsoft**.

- **Ativar Integração com Microsoft Analysis Services.** Ativa a paleta Modelagem da Base de Dados (se ainda não estiver exibida) na parte inferior da janela do IBM SPSS Modeler e inclui os nós para os algoritmos do Analysis Services.
- **Host do Analysis Server.** Especifique o nome da máquina na qual o Analysis Services está em execução.
- **Banco de Dados do Analysis Server.** Selecione o banco de dados desejado clicando no botão de reticências (...) para abrir uma subcaixa de diálogo na qual é possível escolher a partir dos bancos de dados disponíveis. A lista é preenchida com os bancos de dados disponíveis para o servidor Analysis Services especificado. Como o Microsoft Analysis Services armazena modelos de mineração de dados nos bancos de dados nomeados, deve-se selecionar o banco de dados apropriado no qual os modelos da Microsoft criados pelo IBM SPSS Modeler são armazenados.
- **Conexão do SQL Server.** Especifique as informações do DSN utilizadas pelo banco de dados SQL Server para armazenar os dados que são transmitidos para o Analysis Server. Escolha a origem de dados ODBC

que será utilizada para fornecer os dados para construção dos modelos de mineração de dados do Analysis Services. Se estiver construindo modelos do Analysis Services a partir de dados fornecidos em arquivos simples ou origens de dados ODBC, os dados serão transferidos por upload automaticamente para uma tabela temporária criada no banco de dados SQL Server para o qual esta origem de dados ODBC aponta.

- **Avisar quando estiver prestes a substituir um modelo de mineração de dados.** Selecione para assegurar que os modelos armazenados no banco de dados não sejam sobrescritos pelo IBM SPSS Modeler sem aviso.

Nota: as configurações feitas na caixa de diálogo Aplicativos Auxiliares podem ser substituídas dentro dos vários nós do Analysis Services.

Ativando geração e otimização de SQL

1. Nos menus do IBM SPSS Modeler, escolha:

   Ferramentas > **Propriedades do Fluxo** > **Opções**
2. Clique na opção **Otimização** na área de janela de navegação.
3. Confirme se a opção **Gerar SQL** está ativada. Essa configuração é necessária para que a modelagem da base de dados funcione.
4. Selecione **Otimizar Geração de SQL** e **Otimizar outra execução** (não é estritamente necessário, mas é altamente recomendado para um desempenho otimizado).

Construindo Modelos com Analysis Services

A construção de modelo do Analysis Services requer que o conjunto de dados de treinamento esteja localizado em uma tabela ou visualização dentro do banco de dados SQL Server. Se os dados não estiverem localizados no SQL Server ou precisarem ser processados no IBM SPSS Modeler como parte de preparação de dados que não pode ser executada no SQL Server, os dados serão automaticamente transferidos por upload para uma tabela temporária no SQL Server antes da construção do modelo.

Gerenciando Modelos do Analysis Services

Construir um modelo do Analysis Services por meio do IBM SPSS Modeler cria um modelo no IBM SPSS Modeler e cria ou substitui um modelo no banco de dados do SQL Server. O modelo do IBM SPSS Modeler referencia o conteúdo de um modelo de banco de dados armazenado em um servidor de base de dados. O IBM SPSS Modeler pode executar verificação de consistência ao armazenar uma sequência de caracteres de chave de modelo gerada idêntica nos modelos do IBM SPSS Modeler e do SQL Server.

O nó de modelagem **Árvore de Decisão da MS** é utilizado na modelagem preditiva dos atributos categóricos e contínuos. Para atributos categóricos, o nó faz predições com base nos relacionamentos entre as colunas de entrada em um conjunto de dados. Por exemplo, em um cenário para prever quais clientes poderão comprar uma bicicleta, se nove de dez clientes mais jovens comprarem uma bicicleta, e somente dois de dez clientes mais velhos fizerem o mesmo, o nó concluirá que idade é um bom preditor de compra de bicicleta. A árvore de decisão faz predições com base nessa tendência de um resultado específico. Para atributos contínuos, o algoritmo utiliza regressão linear para determinar onde uma árvore de decisão é dividida. Se mais de uma coluna for configurada como previsível, ou se os dados de entrada contiverem uma tabela aninhada que esteja configurada para previsível, o nó construirá uma árvore de decisão separada para cada coluna previsível.

O nó de modelagem **Armazenamento em Cluster da MS** utiliza técnicas iterativas para agrupar casos em um conjunto de dados em clusters que contêm características semelhantes. Esses agrupamentos são úteis para exploração de dados, identificação de anomalias nos dados e criação de predições. Os modelos de armazenamento em cluster identificam relacionamentos em um conjunto de dados que podem não ser derivados logicamente através de uma observação casual. Por exemplo, é possível compreender logicamente que as pessoas que vão para o trabalho de bicicleta geralmente moram perto do local de trabalho. O algoritmo, no entanto, pode localizar outras características não tão óbvias de pessoas que vão para o trabalho de bicicleta. O nó de armazenamento em cluster difere de outros nós de mineração de dados em que nenhum campo de destino é especificado. O nó de armazenamento em cluster treina o modelo estritamente a partir dos relacionamentos que existem nos dados e a partir dos clusters que o nó identifica.

O nó de modelagem **Regras de Associação da MS** é útil para mecanismos de recomendação. Um mecanismo de recomendação recomenda produtos para clientes com base nos itens que eles já compraram ou nos quais demonstraram interesse. Os modelos de associação são construídos em conjuntos de dados que contêm identificadores tanto de casos individuais quanto dos itens que os casos contêm. Um grupo de itens em um caso é chamado de **itemset**. Um modelo de associação é constituído de uma série de itemsets e das regras que descrevem como esses itens são agrupados dentro dos casos. As regras que o algoritmo identifica podem ser utilizadas para prever compras futuras prováveis de um cliente, com base nos itens que já existirem no carrinho de compras desse cliente.

O nó de modelagem **Naive Bayes da MS** calcula a probabilidade condicional entre os campos de destino e de preditor e supõe que as colunas sejam independentes. O modelo é chamado de naïve por tratar todas as variáveis de predição propostas como sendo independentes umas das outras. Este método requer menos cálculo computacional do que outros algoritmos do Analysis Services e, portanto, é útil para descobrir rapidamente relacionamentos durante os estágios preliminares de modelagem. É possível utilizar esse nó para realizar explorações iniciais de dados e, em seguida, aplicar os resultados para criar modelos adicionais com outros nós que podem levar mais tempo para calcular, porém fornecem resultados mais precisos.

O Nó de modelagem **MS linear regression** é uma variação do nó de Trees de Decisão, onde o parâmetro `MINIMUM_LEAF_CASES` é configurado como maior ou igual ao número total de casos no dataset que o nó utiliza para treinar o modelo de mineração. Com o parâmetros configurado dessa maneira, o nó nunca criará uma divisão e, portanto, executará uma regressão linear.

O nó de modelagem **Rede Neural da MS** é semelhante ao nó Árvore de Decisão da MS em que o nó Rede Neural da MS calcula probabilidades para cada estado possível do atributo de entrada, quando é fornecido cada estado do atributo previsível. Posteriormente, é possível utilizar estas probabilidades para prever um resultado do atributo predito, com base nos atributos de entrada.

O nó de modelagem **MS Logistic Regression** é uma variação do nó do MS Neural Network, onde o parâmetro `HIDDEN_NODE_RATIO` é configurado como 0. Essa configuração cria um modelo de rede neural que não contém uma camada oculta e, portanto, é equivalente a regressão logística.

O nó de modelagem **Séries Temporais da MS** fornece algoritmos de regressão que são otimizados para a previsão de valores contínuos, como vendas de produtos, ao longo do tempo. Ao passo que outros algoritmos da Microsoft, como árvores de decisão, requerem colunas adicionais de novas informações como entrada para prever uma tendência, um modelo de série temporal não requer. Um modelo de série temporal pode prever tendências com base apenas no conjunto de dados original que é utilizado para criar o modelo. Também é possível incluir novos dados no modelo quando fizer uma predição e incorporar automaticamente os novos dados na análise de tendência. Veja o tópico "Nó Séries Temporais da MS" na página 17 para obter mais informações.

O nó de modelagem **Armazenamento em Cluster de Sequências da MS** identifica sequências ordenadas nos dados e combina os resultados dessa análise com técnicas de armazenamento em cluster para gerar clusters com base nas sequências e em outros atributos. Veja o tópico "Nó Armazenamento em Cluster de Sequências da MS" na página 18 para obter mais informações.

Também é possível acessar cada nó a partir da paleta Modelagem de Banco de Dados na parte inferior da janela do IBM SPSS Modeler.

Configurações comuns para todos os nós de algoritmos

As configurações a seguir são comuns para todos os algoritmos do Analysis Services.

Opções do Servidor

Na guia Servidor, é possível configurar o host do servidor de Análise, o banco de dados e a origem de dados SQL Server. As opções especificadas aqui substituirão aquelas especificadas na guia Microsoft na caixa de diálogo Aplicativos Auxiliares. Consulte o tópico "Ativando a Integração com o Analysis Services" na página 11 para obter informações adicionais.

Nota: uma variação desta guia também está disponível quando escorar os modelos do Analysis Services. Veja o tópico "Guia Servidor do Nugget do Modelo do Analysis Services" na página 20 para obter mais informações.

Opções do modelo

Para construir o modelo mais básico, é necessário especificar as opções na guia Modelo antes de continuar. O método de escoragem e outras opções avançadas estão disponíveis na guia Especialista.

As opções de modelagem básica a seguir estão disponíveis:

Nome do Modelo. Especifica o nome designado ao modelo que é criado quando o nó é executado.

- **Automático.** Gera o nome do modelo automaticamente com base nos nomes de campo de destino ou de ID ou no nome do tipo de modelo nos casos em que nenhum destino é especificado (como modelos de armazenamento em cluster).
- **Customizado.** Permite especificar um nome customizado para o modelo criado.

Utilizar dados particionados. Divide os dados em subconjuntos ou amostras separados para treinamento, teste e validação com base no campo de partição atual. Utilizar uma amostra para criar o modelo e uma amostra separada para testá-lo poderá fornecer uma indicação de quão bem o modelo será generalizado para conjuntos de dados maiores que forem semelhantes aos dados atuais. Se nenhum campo de partição for especificado no fluxo, essa opção será ignorada.

Com Drill through. Se mostrada, esta opção permite consultar o modelo para aprender detalhes sobre os casos incluídos no modelo.

Campo único. Na lista suspensa, selecione um campo que identifique exclusivamente cada caso. Normalmente, este é um campo de ID, como **CustomerID**.

Opções Avançadas da Árvore de Decisão da MS

As opções disponíveis na guia Especialista podem flutuar dependendo da estrutura do fluxo selecionado. Consulte a ajuda de nível de campo da interface com o usuário para obter detalhes completos sobre as opções avançadas para o nó de modelo do Analysis Services selecionado.

Opções Avançadas de Armazenamento em Cluster da MS

As opções disponíveis na guia Especialista podem flutuar dependendo da estrutura do fluxo selecionado. Consulte a ajuda de nível de campo da interface com o usuário para obter detalhes completos sobre as opções avançadas para o nó de modelo do Analysis Services selecionado.

Opções Avançadas do Naive Bayes da MS

As opções disponíveis na guia Especialista podem flutuar dependendo da estrutura do fluxo selecionado. Consulte a ajuda de nível de campo da interface com o usuário para obter detalhes completos sobre as opções avançadas para o nó de modelo do Analysis Services selecionado.

Opções Avançadas de Regressão Linear da MS

As opções disponíveis na guia Especialista podem flutuar dependendo da estrutura do fluxo selecionado. Consulte a ajuda de nível de campo da interface com o usuário para obter detalhes completos sobre as opções avançadas para o nó de modelo do Analysis Services selecionado.

Opções Avançadas de Rede Neural da MS

As opções disponíveis na guia Especialista podem flutuar dependendo da estrutura do fluxo selecionado. Consulte a ajuda de nível de campo da interface com o usuário para obter detalhes completos sobre as opções avançadas para o nó de modelo do Analysis Services selecionado.

Opções Avançadas de Regressão Logística da MS

As opções disponíveis na guia Especialista podem flutuar dependendo da estrutura do fluxo selecionado. Consulte a ajuda de nível de campo da interface com o usuário para obter detalhes completos sobre as opções avançadas para o nó de modelo do Analysis Services selecionado.

Nó Regras de Associação da MS

O nó de modelagem Regras de Associação da MS é útil para mecanismos de recomendação. Um mecanismo de recomendação recomenda produtos para clientes com base nos itens que eles já compraram ou nos quais demonstraram interesse. Os modelos de associação são construídos em conjuntos de dados que contêm identificadores tanto de casos individuais quanto dos itens que os casos contêm. Um grupo de itens em um caso é chamado de **itemset**.

Um modelo de associação é constituído de uma série de itemsets e das regras que descrevem como esses itens são agrupados dentro dos casos. As regras que o algoritmo identifica podem ser utilizadas para prever compras futuras prováveis de um cliente, com base nos itens que já existirem no carrinho de compras desse cliente.

Para os dados de formato tabular, o algoritmo cria escores que representam probabilidade ($MP-*field*) para cada recomendação gerada ($M-*field*). Para os dados de formato transacional, os escores são criados para suporte ($MS-*field*), probabilidade ($MP-*field*) e probabilidade ajustada ($MAP-*field*) para cada recomendação gerada ($M-*field*).

Requisitos

Os requisitos para um modelo de associação transacional são os seguintes:

- **Campo único.** Um modelo de regras de associação requer uma chave que identifique exclusivamente os registros.

- **Campo de ID.** Ao construir um modelo de Regras de Associação da MS com dados em formato transacional, um campo de ID que identifica cada transação é necessário. Os campos de ID podem ser configurados para o mesmo que o campo exclusivo.
- **Pelo menos um campo de entrada.** O algoritmo de Regras de Associação requer pelo menos um campo de entrada.
- **Campo de destino.** Ao construir um modelo de Associação da MS com dados transacionais, o campo de destino deve ser o campo de transação, por exemplo, produtos que um usuário comprou.

Opções Avançadas de Regras de Associação da MS

As opções disponíveis na guia Especialista podem flutuar dependendo da estrutura do fluxo selecionado. Consulte a ajuda de nível de campo da interface com o usuário para obter detalhes completos sobre as opções avançadas para o nó de modelo do Analysis Services selecionado.

Nó Séries Temporais da MS

O nó de modelagem Séries Temporais da MS suporta dois tipos de predições:

- futuro
- ou seja,

As **predições futuras** estimam valores de campo de destino para um número especificado de períodos de tempo além do término de seus dados históricos e são sempre executadas. Já as **predições históricas** são valores de campo de destino estimados para um número especificado de períodos de tempo para os quais você possui os valores reais em seus dados históricos. É possível utilizar as predições históricas para avaliar a qualidade do modelo ao comparar os valores históricos reais com os valores preditos. O valor do ponto de início para as predições determina se as predições históricas são executadas.

Diferentemente do nó Séries Temporais do IBM SPSS Modeler, o nó Séries Temporais da MS não precisa de um nó Intervalos de Tempo anterior. Uma outra diferença é que, por padrão, os escores são produzidos apenas para as linhas preditas, não para todas as linhas de histórico nos dados de séries temporais.

Requisitos

Os requisitos para um modelo de Séries Temporais da MS são os seguintes:

- **Campo chave de tempo único** Cada modelo deve conter um campo numérico de data que é usado como séries de caso, definindo as fatias de tempo que o modelo utilizará. O tipo de dados para o campo chave de tempo pode ser um tipo de dados de data/hora ou um tipo de dados numérico. No entanto, o campo deve conter valores contínuos e os valores devem ser exclusivos para cada série.
- **Campo de destino único.** É possível especificar apenas um campo de destino em cada modelo. O tipo de dados do campo de destino deve ter valores contínuos. Por exemplo, é possível prever como atributos numéricos, como receita, vendas ou temperatura, alteram ao longo do tempo. Entretanto, não é possível utilizar um campo que contém valores categóricos, como status de compra ou nível de educação, como o campo de destino.
- **Pelo menos um campo de entrada.** O algoritmo de Séries Temporais da MS requer pelo menos um campo de entrada. O tipo de dados do campo de entrada deve ter valores contínuos. Campos de entrada não contínuos são ignorados ao construir o modelo.
- **Conjunto de dados deve ser ordenado.** O conjunto de dados de entrada deve ser ordenado (no campo chave de tempo), caso contrário, a construção do modelo será interrompida com um erro.

Opções do Modelo de Série Temporal da MS

Nome do Modelo. Especifica o nome designado ao modelo que é criado quando o nó é executado.

- **Automático.** Gera o nome do modelo automaticamente com base nos nomes de campo de destino ou de ID ou no nome do tipo de modelo nos casos em que nenhum destino é especificado (como modelos de armazenamento em cluster).

- **Customizado.** Permite especificar um nome customizado para o modelo criado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Com Drill through. Se mostrada, esta opção permite consultar o modelo para aprender detalhes sobre os casos incluídos no modelo.

Campo único. Na lista suspensa, selecione o campo chave de tempo, que é utilizado para construir o modelo de série temporal.

Opções Avançadas de Séries Temporais da MS

As opções disponíveis na guia Especialista podem flutuar dependendo da estrutura do fluxo selecionado. Consulte a ajuda de nível de campo da interface com o usuário para obter detalhes completos sobre as opções avançadas para o nó de modelo do Analysis Services selecionado.

Se estiver fazendo predições históricas, o número de passos históricos que podem ser incluídos no resultado da escoragem é decidido pelo valor de (HISTORIC_MODEL_COUNT * HISTORIC_MODEL_GAP). Por padrão, essa limitação é 10, o que significa que apenas 10 predições históricas serão feitas. Nesse caso, por exemplo, um erro ocorrerá se você inserir um valor menor que -10 para **Predição histórica** na guia Configurações do nugget do modelo (consulte "Guia Configurações do Nugget do Modelo de Série Temporal da MS" na página 21). Se desejar ver mais predições históricas, será possível aumentar o valor de HISTORIC_MODEL_COUNT ou de HISTORIC_MODEL_GAP, mas isso aumentará o tempo de construção do modelo.

Opções de Configurações de Séries Temporais da MS

Iniciar Estimativa. Especifique o período de tempo em que deseja que as predições iniciem.

- **Iniciar A Partir De: Nova Predição.** O período de tempo no qual você deseja que futuras predições iniciem, expresso como uma compensação do último período de tempo de seus dados histórico. Por exemplo, se seus dados históricos terminaram em 12/99 e você desejou iniciar as predições em 01/00, um valor de 1 seria utilizado; entretanto, se quisesse que as predições iniciassem em 03/00, um valor de 3 seria utilizado.
- **Iniciar A Partir De: Predição Histórica.** O período de tempo no qual deseja que as predições históricas iniciem, expresso como um offset negativo do último período de tempo de seus dados históricos. Por exemplo, se seus dados históricos terminaram em 12/99 e você desejou fazer predições históricas para os últimos cinco períodos de tempo de seus dados, um valor de -5 seria utilizado.

Terminar Estimativa. Especifique o período de tempo em que deseja que as predições parem.

- **Terminar passo da predição.** O período de tempo no qual deseja que as predições parem, expresso como uma compensação do último período de tempo de seus dados histórico. Por exemplo, se seus dados históricos terminarem em 12/99 e desejar que as predições parem em 6/00, um valor de 6 seria utilizado aqui. Para predições futuras, o valor deve sempre ser maior ou igual ao valor **Iniciar De**.

Nó Armazenamento em Cluster de Sequências da MS

O nó Armazenamento em Cluster de Sequências da MS utiliza um algoritmo de análise de sequência que explora os dados contendo eventos que podem ser vinculados pelos caminhos ou *sequências* a seguir. Alguns exemplos disso podem ser caminhos de cliques criados quando os usuários navegam ou fazem procuras em um website, ou a ordem na qual um cliente inclui itens em um carrinho de compras em um varejista online. O algoritmo localiza as sequências mais comuns ao agrupar ou *armazenar em cluster* sequências que forem idênticas.

Requisitos

Os requisitos para um modelo de Armazenamento em Cluster da Microsoft são:

- **Campo de ID.** O algoritmo Armazenamento em Cluster de Sequências da Microsoft requer que as informações de sequência sejam armazenadas no formato transacional. Para isso, um campo de ID que identifica cada transação é necessário.

- **Pelo menos um campo de entrada.** O algoritmo requer pelo menos um campo de entrada.
- **Campo de sequência.** O algoritmo também requer um campo de identificador de sequência, que deve ter um nível de medição de Contínuo. Por exemplo, é possível utilizar um identificador de página da web, um número inteiro ou uma sequência de caracteres de texto, desde que o campo identifique os eventos em uma sequência. Apenas um identificador de sequência é permitido para cada sequência e apenas um tipo de sequência é permitido em cada modelo. O campo Sequência deve ser diferente dos campos ID e Único.
- **Campo de destino.** Um campo de destino é necessário quando construir um modelo de armazenamento de sequência.
- **Campo único.** Um modelo de armazenamento em cluster de sequência requer um campo-chave que identifica exclusivamente os registros. É possível configurar o campo Único para ser o mesmo que o campo ID.

Opções dos Campos de Armazenamento em Cluster de Sequências da MS

Todos os nós de modelagem possuem uma guia Campos, na qual você especifica os campos a serem utilizados na construção do modelo.

Antes de poder construir um modelo de armazenamento em cluster de sequências, é necessário especificar quais campos você deseja utilizar como destinos e como entradas. Observe que para o nó Armazenamento em Cluster de Sequências da MS, não é possível usar as informações de campo a partir de um nó Tipo de envio de dados; deve-se especificar as configurações do campo aqui.

ID. Selecione um campo de ID na lista. Campos numéricos ou simbólicos podem ser utilizados como o campo de ID. Cada valor exclusivo deste campo deve indicar uma unidade específica de análise. Por exemplo, em um aplicativo de cesta de mercado, cada ID pode representar um cliente único. Para um aplicativo de análise de log da web, cada ID pode representar um computador (pelo endereço IP) ou um usuário (pelos dados de login).

Entradas. Selecione um ou mais campos de entrada para o modelo. Estes são os campos que contêm os eventos de interesse na modelagem de sequência.

Sequência. Escolha um campo na lista a ser utilizado como o campo identificador de sequência. Por exemplo, é possível utilizar um identificador de página da web, um número inteiro ou uma sequência de caracteres de texto, desde que o campo identifique os eventos em uma sequência. Apenas um identificador de sequência é permitido para cada sequência e apenas um tipo de sequência é permitido em cada modelo. O campo Sequência deve ser diferente do campo ID (especificado nesta guia) e do campo Exclusivo (especificado na guia Modelo).

Destino. Escolha um campo a ser utilizado como o campo de destino, ou seja, o campo cujo valor você está tentando prever com base nos dados de sequência.

Opções Avançadas de Armazenamento em Cluster de Sequências da MS

As opções disponíveis na guia Especialista podem flutuar dependendo da estrutura do fluxo selecionado. Consulte a ajuda de nível de campo da interface com o usuário para obter detalhes completos sobre as opções avançadas para o nó de modelo do Analysis Services selecionado.

Escorando Modelos do Analysis Services

A escoragem de modelo ocorre dentro do SQL Server e é executada pelo Analysis Services. O conjunto de dados poderá precisar ser transferido por upload para uma tabela temporária, se os dados originarem dentro do IBM SPSS Modeler ou precisarem ser preparados dentro do IBM SPSS Modeler. Os modelos que você criar a partir do IBM SPSS Modeler utilizando mineração dentro da base de dados na realidade são um modelo remoto mantido no servidor de mineração de dados ou de banco de dados remoto. É importante entender essa distinção conforme você navega e escora os modelos criados utilizando algoritmos nativos do Microsoft Analysis Services.

No IBM SPSS Modeler, geralmente uma única predição com probabilidade ou confiança associada é entregue.

Para obter exemplos de escoragem de modelo, consulte “Exemplos de Mineração do Analysis Services” na página 22.

Configurações comuns para todos os modelos do Analysis Servicess

As configurações a seguir são comuns para todos os modelos do Analysis Services.

Guia Servidor do Nugget do Modelo do Analysis Services

A guia Servidor é utilizada para especificar conexões para mineração dentro da base de dados. A guia também fornece a chave de modelo exclusiva. A chave é gerada aleatoriamente quando o modelo é construído e armazenada no modelo no IBM SPSS Modeler e também dentro da descrição do objeto de modelo armazenada no banco de dados do Analysis Services.

Na guia Servidor, é possível configurar o host e o banco de dados do servidor de Análise e a origem de dados SQL Server para a operação de escoragem. As opções especificadas aqui substituirão aquelas especificadas nas caixas de diálogo Aplicativos Auxiliares ou Modelo de Construção no IBM SPSS Modeler. Veja o tópico “Ativando a Integração com o Analysis Services” na página 11 para obter mais informações.

GUID de Modelo. A chave de modelo é mostrada aqui. A chave é gerada aleatoriamente quando o modelo é construído e armazenada no modelo no IBM SPSS Modeler e também dentro da descrição do objeto de modelo armazenada no banco de dados do Analysis Services.

Verificar. Clique neste botão para verificar a chave do modelo com relação à chave no modelo armazenado no banco de dados do Analysis Services. Isso permite verificar se o modelo ainda existe no servidor de Análise e indica que a estrutura do modelo não foi alterada.

Nota: O botão Check está disponível apenas para modelos adicionados à tela do fluxo em preparação para pontuação. Se a verificação falhar, cheque se o modelo foi excluído ou substituído por um modelo diferente no servidor.

Visualização. Clique para abrir uma visualização gráfica do modelo de árvore de decisão. O Visualizador de Árvore de Decisão é compartilhado por outros algoritmos de árvore de decisão no IBM SPSS Modeler e a funcionalidade é idêntica.

Guia Sumarização do Nugget do Modelo do Analysis Services

A guia Sumarização de um nugget do modelo exibe informações sobre o modelo em si (*Análise*), sobre os campos usados no modelo (*Campos*), sobre as configurações utilizadas ao construir o modelo (*Configurações de Construção*) e sobre o treinamento do modelo (*Sumarização do Treinamento*).

Ao procurar o nó pela primeira vez, os resultados da guia Sumarização são reduzidos. Para ver os resultados de interesse, utilize o controle expansor à esquerda de um item para desdobrá-lo ou clique no botão **Expandir Tudo** para mostrar todos os resultados. Para ocultar os resultados após terminar de visualizá-los, use o controle expansor para reduzir os resultados específicos que deseja ocultar ou clique no botão **Reduzir Tudo** para reduzir todos os resultados.

Análise. Exibe informações sobre o modelo específico. Se tiver executado um nó Análise anexado a este nugget do modelo, as informações dessa análise também serão exibidas nesta seção.

Campos. Lista os campos utilizados como o destino e as entradas na construção do modelo.

Configurações de construção. Contém informações sobre as configurações utilizadas na construção do modelo.

Sumarização do Treinamento. Mostra o tipo de modelo, o fluxo utilizado para criá-lo, o usuário que o criou, quando ele foi construído e o tempo decorrido para construir o modelo.

Nugget do Modelo de Série Temporal da MS

O modelo de Série Temporal da MS produz escores somente para os períodos de tempo preditos, não para dados históricos.

A tabela a seguir mostra os campos que são incluídos no modelo.

Tabela 1. Campos incluídos no modelo	
Nome do Campo	**Descrição**
$M-*field*	Valor predito de *field*
$Var-*field*	Variância calculada de *field*
$Stdev-*field*	Desvio padrão de *field*

Guia Servidor do Nugget do Modelo de Série Temporal da MS

A guia Servidor é utilizada para especificar conexões para mineração dentro da base de dados. A guia também fornece a chave de modelo exclusiva. A chave é gerada aleatoriamente quando o modelo é construído e armazenada no modelo no IBM SPSS Modeler e também dentro da descrição do objeto de modelo armazenada no banco de dados do Analysis Services.

Na guia Servidor, é possível configurar o host e o banco de dados do servidor de Análise e a origem de dados SQL Server para a operação de escoragem. As opções especificadas aqui substituirão aquelas especificadas nas caixas de diálogo Aplicativos Auxiliares ou Modelo de Construção no IBM SPSS Modeler. Veja o tópico “Ativando a Integração com o Analysis Services” na página 11 para obter mais informações.

GUID de Modelo. A chave de modelo é mostrada aqui. A chave é gerada aleatoriamente quando o modelo é construído e armazenada no modelo no IBM SPSS Modeler e também dentro da descrição do objeto de modelo armazenada no banco de dados do Analysis Services.

Verificar. Clique neste botão para verificar a chave do modelo com relação à chave no modelo armazenado no banco de dados do Analysis Services. Isso permite verificar se o modelo ainda existe no servidor de Análise e indica que a estrutura do modelo não foi alterada.

Nota: O botão Check está disponível apenas para modelos adicionados à tela do fluxo em preparação para pontuação. Se a verificação falhar, cheque se o modelo foi excluído ou substituído por um modelo diferente no servidor.

Visualização. Clique para abrir uma visualização gráfica do modelo de série temporal. O Analysis Services exibe o modelo concluído como uma árvore. Também é possível visualizar um gráfico que mostra o valor histórico do campo de destino no decorrer do tempo, junto dos valores futuros preditos.

Para obter mais informações, consulte a descrição do visualizador Séries Temporais na biblioteca MSDN em http://msdn.microsoft.com/en-us/library/ms175331.aspx.

Guia Configurações do Nugget do Modelo de Série Temporal da MS

Iniciar Estimativa. Especifique o período de tempo em que deseja que as predições iniciem.

- **Iniciar A Partir De: Nova Predição.** O período de tempo no qual você deseja que futuras predições iniciem, expresso como uma compensação do último período de tempo de seus dados histórico. Por exemplo, se seus dados históricos terminaram em 12/99 e você desejou iniciar as predições em 01/00, um valor de 1 seria utilizado; entretanto, se quisesse que as predições iniciassem em 03/00, um valor de 3 seria utilizado.
- **Iniciar A Partir De: Predição Histórica.** O período de tempo no qual deseja que as predições históricas iniciem, expresso como um offset negativo do último período de tempo de seus dados históricos. Por exemplo, se seus dados históricos terminaram em 12/99 e você desejou fazer predições históricas para os últimos cinco períodos de tempo de seus dados, um valor de -5 seria utilizado.

Terminar Estimativa. Especifique o período de tempo em que deseja que as predições parem.

- **Terminar passo da predição.** O período de tempo no qual deseja que as predições parem, expresso como uma compensação do último período de tempo de seus dados histórico. Por exemplo, se seus dados históricos terminarem em 12/99 e desejar que as predições parem em 6/00, um valor de 6 seria utilizado aqui. Para predições futuras, o valor deve sempre ser maior ou igual ao valor **Iniciar De**.

Nugget do Modelo de Armazenamento em Cluster de Sequências da MS

A tabela a seguir mostra os campos que são incluídos no modelo de Armazenamento em Cluster de Sequências da MS (em que *field* é o nome do campo de destino).

Tabela 2. Campos incluídos no modelo

Nome do Campo	Descrição
$MC-*field*	Predição do cluster ao qual essa sequência pertence.
$MCP-*field*	Probabilidade que esta sequência pertença ao cluster predito.
$MS-*field*	Valor predito de *field*
$MSP-*field*	Probabilidade de o valor de $MS-*field* estar correto.

Exportando Modelos e Gerando Nós

É possível exportar uma sumarização e estrutura de modelo em arquivos de formato de texto e HTML. É possível gerar os nós Seleção e Filtro adequados quando apropriado.

Semelhantes aos outros nuggets do modelo no IBM SPSS Modeler, os nuggets do modelo do Microsoft Analysis Services suportam a geração direta dos nós de operações de registro e de campo. Usando as opções do menu Gerar do nugget do modelo, é possível gerar os nós a seguir:

- Selecione o nó (apenas se um item estiver selecionado na guia Modelo)
- nó Filtro

Exemplos de Mineração do Analysis Services

Diversos fluxos de amostra são incluídos que demonstram o uso da mineração de dados do MS Analysis Services com o IBM SPSS Modeler. Estes fluxos podem ser localizados na pasta de instalação do IBM SPSS Modeler em:

\Demos\Database_Modelling\Microsoft

Nota: a pasta Demos pode ser acessada a partir do grupo do programa do IBM SPSS Modeler no menu Iniciar do Windows.

Fluxos de Exemplo: Árvores de Decisão

Os fluxos a seguir podem ser utilizados juntos e em sequência como um exemplo do processo de mineração da base de dados utilizando o algoritmo Árvores de Decisão fornecido pelos Serviços de Análise da MS.

Tabela 3. Árvores de Decisão - fluxos de exemplos

Fluxo	Descrição
1_upload_data.str	Utilizado para limpar e fazer upload de dados de um arquivo simples para o banco de dados.
2_explore_data.str	Fornece um exemplo de exploração de dados com IBM SPSS Modeler
3_build_model.str	Constrói o modelo utilizando o algoritmo nativo do banco de dados.
4_evaluate_model.str	Usado como exemplo de avaliação de modelo com IBM SPSS Modeler

Tabela 3. Árvores de Decisão - fluxos de exemplos (continuação)	
Fluxo	**Descrição**
5_deploy_model.str	Implementa o modelo para escoragem dentro do bando de dados.

Nota: para executar o exemplo, os fluxos devem ser executados em ordem. Além disso, os nós de origem e de modelagem em cada fluxo devem ser atualizados para referenciar uma origem de dados válida para o banco de dados que deseja utilizar.

O conjunto de dados utilizado nos fluxos de exemplos refere-se a aplicativos de cartão de crédito e apresenta um problema de classificação com uma combinação de preditores categóricos e contínuos. Para obter mais informações sobre esse conjunto de dados, consulte o arquivo *crx.names* na mesma pasta que os fluxos de amostra.

Este conjunto de dados está disponível a partir do UCI Machine Learning Repository em *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/*.

Fluxo de Exemplo: Upload de Dados

O primeiro fluxo de exemplo, *1_upload_data.str*, é utilizado para limpar e fazer upload de dados de um arquivo simples para o SQL Server.

Uma vez que a mineração de dados de Analysis Services requer um campo chave, este fluxo inicial usa um nó de Derivação para adicionar um novo campo ao dataset chamado *KEY* com valores exclusivos *1,2,3* usando a função IBM SPSS Modeler `@INDEX` .

O nó Preenchimento subsequente é utilizado para manipulação de valores omissos e substitui campos vazios que são lidos a partir do arquivo de texto *crx.data* por valores *NULL*.

Fluxo de Exemplo: Explorar Dados

O segundo fluxo de exemplo, *2_explore_data.str*, é utilizado para demonstrar o uso de um nó de Auditoria de Dados para obter uma visão geral dos dados, incluindo estatísticas de sumarização e gráficos.

Dar um clique duplo em um gráfico no Relatório de Auditoria de Dados produz um gráfico mais detalhado para exploração mais profunda de um campo determinado.

Fluxo de Exemplo: Construir o Modelo

O terceiro fluxo de exemplo, *3_build_model.str*, ilustra a construção de modelo no IBM SPSS Modeler. É possível anexar o modelo de banco de dados ao fluxo e dar um clique duplo para especificar as configurações de construção.

Na guia Modelo da caixa de diálogo, é possível especificar o seguinte:

1. Selecione o campo **Chave** como o campo ID exclusivo.

Na guia Especialista, é possível fazer um ajuste preciso das configurações para construção do modelo.

Antes da execução, assegure-se de ter especificado o banco de dados correto para construção de modelo. Utilize a guia Servidor para ajustar quaisquer configurações.

Fluxo de Exemplo: Avaliar o Modelo

O quarto fluxo de exemplo, *4_evaluate_model.str*, ilustra as vantagens de utilizar o IBM SPSS Modeler para modelagem dentro da base de dados. Após ter executado o modelo, será possível incluí-lo de volta no fluxo de dados e avaliar o modelo utilizando várias ferramentas oferecidas no IBM SPSS Modeler.

Visualizando Resultados de Modelagem

É possível dar um clique duplo no nugget do modelo para explorar seus resultados. A guia Sumarização fornece uma visualização em árvore de regra de resultados. Também é possível clicar no botão

Visualizar, localizado na guia Servidor, para abrir uma visualização gráfica do modelo de Árvore de Decisão.

Avaliando Resultados de Modelo

O nó Análise no fluxo de amostra cria uma matriz de coincidência mostrando o padrão de correspondências entre cada campo predito e seu campo de destino. Execute o nó de Análise para visualizar os resultados.

O nó Avaliação no fluxo de amostra pode criar um gráfico de ganhos projetado para mostrar melhorias de precisão feitas pelo modelo. Execute o nó Avaliação para visualizar os resultados.

Fluxo de Exemplo: Implementar o Modelo

Quando estiver satisfeito com a precisão do modelo, ele poderá ser implementado para uso com aplicativos externos ou para publicar de volta no banco de dados. No fluxo de exemplo final, *5_deploy_model.str*, os dados são lidos a partir da tabela CREDIT e, em seguida, escorados e publicados na tabela CREDITSCORES utilizando um nó Exportação de Banco de Dados.

Executar o fluxo gera a SQL a seguir:

```
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. Modelagem da base de dados com a Mineração de Dados do Oracle

Sobre a Mineração de Dados do Oracle

O IBM SPSS Modeler suporta integração com a Mineração de Dados do Oracle (ODM), que fornece uma família de algoritmos de mineração de dados concisamente integrada no RDBMS Oracle. Estes recursos podem ser acessados por meio da interface gráfica com o usuário e do ambiente de desenvolvimento orientado a fluxo de trabalho do IBM SPSS Modeler, permitindo que os clientes utilizem os algoritmos de mineração de dados oferecidos pelo ODM.

O IBM SPSS Modeler suporta integração dos seguintes algoritmos de Mineração de Dados do Oracle:

- Naive Bayes
- Adaptive Bayes
- Support Vector Machine (SVM)
- Modelos Lineares Generalizados (GLM)*
- Árvore de decisão
- Cluster-O
- k-Médias
- Nonnegative Matrix Factorization (NMF)
- A priori
- Minimum Descriptor Length (MDL)
- Attribute Importance (AI)

Somente * 11*g* R1

Requisitos para Integração com o Oracle

As condições a seguir são pré-requisitos para conduzir modelagem dentro da base de dados usando a Mineração de Dados do Oracle. Poderá ser necessário consultar seu administrador de base de dados para assegurar que essas condições sejam atendidas.

- IBM SPSS Modeler em execução no modo local ou com relação a uma instalação do ServidorIBM SPSS Modeler no Windows ou UNIX.
- Oracle 10*g*R2 ou 11*g*R1 (Banco de dados 10.2 ou superior) com a opção de Mineração de Dados do Oracle.

Nota: 10*g*R2 fornece suporte para todos os algoritmos de modelagem de banco de dados, exceto Modelos Lineares Generalizados (requer 11*g*R1).

- Uma origem de dados ODBC para conexão com o Oracle, conforme descrito a seguir.

Nota: A modelagem da base de dados e a otimização de SQL requerem que a conectividade do ServidorIBM SPSS Modeler esteja ativada no computador do IBM SPSS Modeler. Com essa configuração ativada, é possível acessar os algoritmos de banco de dados, realizar SQL pushback diretamente do IBM SPSS Modeler e acessar o ServidorIBM SPSS Modeler. Para verificar o atual status da licença, escolha o seguinte no menu do IBM SPSS Modeler.

Ajuda > **Sobre** > **Detalhes Adicionais**

Se a conectividade estiver ativada, você verá a opção **Ativação do Servidor** na guia Status da Licença.

Ativando a Integração com Oracle

Para ativar a integração do IBM SPSS Modeler com a Mineração de Dados do Oracle, será necessário configurar o Oracle, criar uma origem ODBC, ativar a integração na caixa de diálogo Aplicativos Auxiliares do IBM SPSS Modeler e ativar a geração e a otimização de SQL.

Configurando o Oracle

Para instalar e configurar a Mineração de Dados do Oracle, consulte a documentação do Oracle - em particular, o *Guia do Administrador Oracle* - para mais detalhes.

Criando uma Origem ODBC para Oracle

Para ativar a conexão entre o Oracle e o IBM SPSS Modeler, é necessário criar um nome da origem de dados (DSN) do sistema ODBC.

Antes de criar um DSN, você deverá ter um entendimento básico das origens de dados e drivers ODBC, bem como do suporte ao banco de dados no IBM SPSS Modeler.

Se estiver executando no modo distribuído com relação ao ServidorIBM SPSS Modeler, crie o DSN no computador servidor. Se estiver executando no modo local (cliente), crie o DSN no computador cliente.

1. Instale os drivers ODBC. Eles estão disponíveis no disco de instalação do Pacote de acesso a dadosIBM SPSS fornecido com esta liberação. Execute o arquivo *setup.exe* para iniciar o instalador e selecione todos os drivers relevantes. Siga as instruções na tela para instalar os drivers.
   a. Criar o DSN.

      Nota: A sequência do menu depende da sua versão do Windows.

      - **Windows XP.** No menu Iniciar, escolha **Painel de Controle**. Dê um clique duplo em **Ferramentas Administrativas** e, em seguida, um clique duplo em **Origens de Dados (ODBC)**.
      - **Windows Vista.** No menu Iniciar, escolha **Painel de Controle**, em seguida, **Manutenção do Sistema**. Dê um clique duplo em **Ferramentas Administrativas**, selecione **Origens de Dados (ODBC)** e, em seguida, clique em **Abrir**.
      - **Windows 7.** No menu Iniciar, escolha **Painel de Controle**, em seguida, **Sistema & Segurança**, em seguida, **Ferramentas Administrativas**. Selecione **Origens de Dados (ODBC)** e, em seguida, clique em **Abrir**.

   b. Vá até a aba **DSN do Sistema** e, em seguida, clique em **Adicionar**.
2. Selecione o driver **SPSS OEM 6.0 Oracle Wire Protocol**.
3. Clique em **Finish**.
4. Na tela Configuração do Driver ODBC Oracle Wire Protocol, insira um nome de origem de dados de sua escolha, o nome do host do servidor Oracle, o número da porta para a conexão e o SID para a instância Oracle que você está utilizando.

   O nome do host, a porta e o SID podem ser obtidos a partir do arquivo *tnsnames.ora* na máquina servidor se tiver implementado o TNS com um arquivo *tnsnames.ora*. Entre em contato com o administrador do Oracle para obter mais informações.
5. Clique no botão **Testar** para testar a conexão.

Ativando o Oracle Data Mining Integration em IBM SPSS Modeler

1. Nos menus do IBM SPSS Modeler, escolha:

   Ferramentas > **Opções** > **Aplicativos auxiliares**
2. Clique na guia **Oracle**.

Ativar Integração da Mineração de Dados do Oracle. Ativa a paleta de Modelagem da Base de Dados (se ainda não estiver exibida) na parte inferior da janela do IBM SPSS Modeler e inclui os nós para algoritmos de Mineração de Dados do Oracle.

Conexão do Oracle. Especifique a origem de dados ODBC Oracle padrão utilizada para construir e armazenar modelos, junto de um nome de usuário e senha válidos. Esta configuração pode ser substituída nos nós de modelagem individuais e nos nuggets do modelo.

Nota: a conexão com o banco de dados utilizada para propósitos de modelagem pode ou não ser a mesma conexão utilizada para acessar dados. Por exemplo, é possível ter um fluxo que acessa dados de um banco de dados Oracle, faz o download dos dados para o IBM SPSS Modeler para limpeza ou outras manipulações e, em seguida, faz upload dos dados para um banco de dados Oracle diferente para fins de modelagem. Como alternativa, os dados originais podem residir em um arquivo simples ou em outra origem (não Oracle), caso em que eles precisariam ser transferidos por upload para o Oracle para modelagem. Em todos os casos, os dados serão transferidos por upload automaticamente para uma tabela temporária criada no banco de dados que é utilizado para modelagem.

Avisar quando estiver prestes a sobrescrever um modelo de Mineração de Dados do Oracle. Selecione esta opção para assegurar que os modelos armazenados no banco de dados não sejam sobrescritos pelo IBM SPSS Modeler sem aviso.

Listar Modelos de Mineração de Dados do Oracle. Exibe os modelos de mineração de dados disponíveis.

Ativar iniciação do Oracle Data Miner. (opcional) Quando ativado, permite que o IBM SPSS Modeler ative o aplicativo Oracle Data Miner. Consulte "Oracle Data Miner" na página 43 para obter informações adicionais.

Caminho para o executável do Oracle Data Miner. (opcional) Especifica o local físico do Oracle Data Miner para o arquivo executável do Windows (por exemplo, *C:\odm\bin\odminerw.exe*). O Oracle Data Miner não é instalado com o IBM SPSS Modeler, portanto, deve-se fazer download da versão correta a partir do website da Oracle (*http://www.oracle.com/technology/products/bi/odm/odminer.html*) e instalá-lo no cliente.

Ativando geração e otimização de SQL

1. Nos menus do IBM SPSS Modeler, escolha:

   Ferramentas > **Propriedades do Fluxo** > **Opções**
2. Clique na opção **Otimização** na área de janela de navegação.
3. Confirme se a opção **Gerar SQL** está ativada. Essa configuração é necessária para que a modelagem da base de dados funcione.
4. Selecione **Otimizar Geração de SQL** e **Otimizar outra execução** (não é estritamente necessário, mas é altamente recomendado para um desempenho otimizado).

Construindo Modelos com a Mineração de Dados do Oracle

Os nós de construção de modelo do Oracle funcionam exatamente como outros nós de modelagem no IBM SPSS Modeler, com algumas exceções. É possível acessar esses nós a partir da paleta Modelagem da Base de Dados na parte inferior da janela do IBM SPSS Modeler.

Considerações de dados

O Oracle requer que os dados categóricos sejam armazenados em um formato de sequência de caracteres (CHAR ou VARCHAR2). Como resultado, o IBM SPSS Modeler não permitirá que campos de armazenamento numérico com um nível de medição de *sinalização* ou *Nominal* (categórico) sejam especificados como entrada para modelos do ODM. Se necessário, os números podem ser convertidos em sequências de caracteres no IBM SPSS Modeler usando o nó Reclassificar.

Campo de destino. Apenas um campo pode ser selecionado como o campo de saída (destino) nos modelos de classificação do ODM.

Nome do Modelo. A partir do Oracle 11*g*R1, o nome `unique` é uma palavra-chave e não pode ser usada como um nome de modelo customizado.

Campo exclusivo. Especifica o campo utilizado para identificar exclusivamente cada caso. Por exemplo, esse pode ser um campo de ID, como *CustomerID*. O IBM SPSS Modeler impõe uma restrição de que esse campo-chave deve ser numérico.

Nota: Este campo é opcional para todos os nós Oracle , exceto Oracle Adaptive Bayes, Oracle O-Cluster e Oracle Apriori.

Comentários Gerais

- A Exportação/Importação do PMML não é fornecida a partir do IBM SPSS Modeler para modelos criados pela Mineração de Dados do Oracle.
- A escoragem de modelo sempre ocorre no ODM. O conjunto de dados poderá precisar ser transferido por upload para uma tabela temporária se os dados originarem ou precisarem ser preparados dentro do IBM SPSS Modeler.
- No IBM SPSS Modeler, geralmente uma única predição com probabilidade ou confiança associada é entregue.
- O IBM SPSS Modeler restringe o número de campos que podem ser utilizados na construção de modelo e a escoragem para 1.000.
- O IBM SPSS Modeler pode escorar os modelos do ODM dentro de fluxos publicados para execução usando o Editor de soluçõesIBM SPSS Modeler.

Opções do Servidor de Modelos do Oracle

Especifique a conexão Oracle que é utilizada para fazer upload dos dados para modelagem. Se necessário, é possível selecionar uma conexão na guia Servidor para cada nó de modelagem para substituir a conexão do Oracle padrão especificada na caixa de diálogo Aplicativos Auxiliares. Consulte o tópico “Ativando a Integração com Oracle” na página 26 para obter informações adicionais.

Comentários

- A conexão utilizada para modelagem pode ou não ser a mesma conexão utilizada no nó de origem para um fluxo. Por exemplo, é possível ter um fluxo que acessa dados de um banco de dados Oracle, faz o download dos dados para o IBM SPSS Modeler para limpeza ou outras manipulações e, em seguida, faz upload dos dados para um banco de dados Oracle diferente para fins de modelagem.
- O nome da origem de dados ODBC é efetivamente integrado em cada fluxo do IBM SPSS Modeler. Se um fluxo que é criado em um host for executado em um host diferente, o nome da origem de dados deverá ser o mesmo em cada host. Como alternativa, uma origem de dados diferente pode ser selecionada na guia Servidor em cada nó de origem ou de modelagem.

Custos de classificação errada

Em alguns contextos, determinados tipos de erros são mais caros que outros. Por exemplo, pode ser mais caro classificar um solicitante de crédito de alto risco como baixo risco (um tipo de erro) do que classificar um solicitante de baixo risco como alto risco (um tipo diferente de erro). Os custos de classificação errada permitem especificar a importância relativa de diferentes tipos de erros de predição.

Os custos de classificação errada são basicamente ponderações aplicadas a resultados específicos. Essas ponderações são fatoradas no modelo e podem, na realidade, alterar a predição (como uma forma de proteger contra erros caros).

Com exceção dos modelos do C5.0, os custos de classificação errada não serão aplicados ao escorar um modelo e não são levados em conta quando classificar ou comparar modelos usando um nó Classificador Automático, gráfico de avaliação, ou nó Análise. Um modelo que inclui custos poderá não produzir menos erros do que aquele que não inclui e poderá não ter uma classificação mais alta em termos de precisão geral, mas provavelmente executará melhor em termos práticos por possuir um viés integrado a favor de erros *menos caros*.

A matriz de custo mostra o custo para cada combinação possível de categoria predita e categoria real. Por padrão, todos os custos de classificação errada são configurados como 1,0. Para inserir valores de custo

customizado, selecione **Usar custos de classificação errada** e insira os valores customizados na matriz de custo.

Para alterar um custo de classificação errada, selecione a célula correspondente à combinação desejada de valores preditos e reais, exclua o conteúdo existente da célula e insira o custo desejado para a célula. Os custos não são simétricos automaticamente. Por exemplo, se você configurar o custo de classificação errada de *A* como *B* para 2,0, o custo da classificação errada de *B* como *A* ainda terá o valor padrão de 1,0, a menos que você também o altere explicitamente.

Nota: apenas o modelo Árvores de Decisão permite que os custos sejam especificados no momento da construção.

Oracle Naive Bayes

O Naive Bayes é um algoritmo bem conhecido para problemas de classificação. O modelo é chamado de *naïve* por tratar todas as variáveis de predição propostas como sendo independentes umas das outras. O Naive Bayes é um algoritmo rápido e escalável que calcula probabilidades condicionais para combinações de atributos e o atributo de destino. A partir dos dados de treinamento, uma probabilidade independente é estabelecida. Essa probabilidade fornece a verossimilhança de cada classe de destino, dada a ocorrência de cada categoria de valor a partir de cada variável de entrada.

- A validação cruzada é utilizada para testar a precisão do modelo nos mesmos dados que foram utilizados para construir o modelo. Isso é especialmente útil quando o número de casos disponíveis para construir um modelo é pequeno.
- A saída de modelo pode ser procurada em um formato de matriz. Os números na matriz são as probabilidades condicionais que relacionam as classes preditas (colunas) e as combinações de variável preditora-valor (linhas).

Opções do Modelo Naive Bayes

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Campo exclusivo. Especifica o campo utilizado para identificar exclusivamente cada caso. Por exemplo, esse pode ser um campo de ID, como *CustomerID*. O IBM SPSS Modeler impõe uma restrição de que esse campo-chave deve ser numérico.

Nota: Este campo é opcional para todos os nós Oracle , exceto Oracle Adaptive Bayes, Oracle O-Cluster e Oracle Apriori.

Preparação de Dados Automática. (11*g* somente) Ativa (padrão) ou desativa o modo de preparação de dados automatizada de Mineração de Dados do Oracle. Se essa caixa estiver marcada, o ODM executa automaticamente as transformações de dados necessárias pelo algoritmo. Para obter mais informações, consulte *Conceitos de Mineração de Dados do Oracle*.

Opções Avançadas do Naive Bayes

Quando o modelo é criado, os valores ou pares de valores de atributos de preditor individuais são ignorados, a menos que existam ocorrências suficientes de um determinado valor ou par nos dados de treinamento. Os limites para ignorar valores são especificados como frações com base no número de registros nos dados de treinamento. O ajuste desses limites pode reduzir o ruído e melhorar a capacidade do modelo para generalizar para outros conjuntos de dados.

- **Limite de Singleton.** Especifica o limite para um valor de atributo de preditor especificado. O número de ocorrências de um determinado valor deverá ser igual ou exceder a fração especificada ou o valor será ignorado.

- **Limite Entre Pares.** Especifica o limite para um determinado par de valores de atributo e do preditor. O número de ocorrências de um determinado par de valores deverá ser igual ou exceder a fração especificada ou o par será ignorado.

Probabilidade de predição. Permite que o modelo inclua a probabilidade de uma predição correta de um resultado possível do campo de destino. Para ativar esse recurso, escolha **Selecionar**, clique no botão **especificar** , escolha um dos resultados possíveis e, em seguida, clique em **Inserir**.

Usar Conjunto de Predição. Gera uma tabela para todos os resultados possíveis do campo de destino.

Oracle Adaptive Bayes

A Adaptive Bayes Network (ABN) constrói classificadores de rede bayesiana usando o Comprimento Mínimo da Descrição (MDL) e a seleção de variável automática. A ABN tem um bom desempenho em determinadas situações em que o Naive Bayes tem desempenho inferior e executa ao menos tão bem quanto o Naive Bayes na maioria das outras situações, embora esse desempenho possa ser menor. O algoritmo ABN fornece a capacidade de construir três tipos de modelos bayesianos avançados, incluindo árvore de decisão simplificada (variável única), Naive Bayes podado e modelos de várias variáveis impulsionadas.

Nota: O algoritmo Oracle Adaptive Bayes foi eliminado do Oracle 12C e não é suportado no IBM SPSS Modeler ao utilizar o Oracle 12c. Consulte https://docs.oracle.com/en/database/oracle/oracle-database/18/upgrd/behavior-changes-oracle-database-12c-121.html#GUID-C7DB3E44-D55A-41E7-A99B-291DBD71D87D.

Modelos Gerados

No modo de construção de variável única, a ABN produz uma árvore de decisão simplificada com base em um conjunto de regras legíveis, que permitem que o usuário ou o analista de negócios entendam a base das predições do modelo e atuem ou expliquem aos outros de forma apropriada. Isso pode ser uma vantagem significativa sobre os modelos Naive Bayes e de várias variáveis. Essas regras podem ser procuradas como um conjunto de regras padrão no IBM SPSS Modeler. Um conjunto de regras simples pode ser semelhante ao seguinte:

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

Os modelos Naive Bayes podados e de várias variáveis não podem ser procurados no IBM SPSS Modeler.

Opções do Modelo Bayes do ISW

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Campo exclusivo. Especifica o campo utilizado para identificar exclusivamente cada caso. Por exemplo, esse pode ser um campo de ID, como *CustomerID*. O IBM SPSS Modeler impõe uma restrição de que esse campo-chave deve ser numérico.

Nota: Este campo é opcional para todos os nós Oracle , exceto Oracle Adaptive Bayes, Oracle O-Cluster e Oracle Apriori.

Tipo de modelo

É possível escolher entre três modos diferentes para construir o modelo.

- **Diversas variáveis.** Constrói e compara um número de modelos, incluindo um modelo NB e modelos de probabilidade de produto de única e diversas variáveis. Esse é o modo mais exaustivo e geralmente leva mais tempo para calcular como resultado. As regras serão produzidas somente se o modelo de variável

única ficar de fora para ser o melhor. Se um modelo de diversas variáveis ou NB for escolhido, nenhuma regra será produzida.

- **Variável única.** Cria uma árvore de decisão simplificada com base em um conjunto de regras. Cada regra contém uma condição junto das probabilidades associadas a cada resultado. As regras são mutuamente exclusivas e fornecidas em um formato que possa ser lido por pessoas, o que pode ser uma enorme vantagem sobre modelos Naive Bayes e de diversas variáveis.
- **Naive Bayes.** Constrói um modelo NB único e o compara com a amostra global anterior (a distribuição de valores de resposta na amostra global). O modelo NB será produzido como saída somente se ele passar a ser um melhor preditor dos valores de resposta do que o modelo global anterior. Caso contrário, nenhum modelo será produzido como saída.

Opções Avançadas do Adaptive Bayes

Limitar tempo de execução. Selecione esta opção para especificar um tempo de construção máximo em minutos. Isso permite produzir modelos em menos tempo, embora o modelo resultante possa ser menos preciso. Em cada marco no processo de modelagem, o algoritmo verifica se ele conseguirá concluir o próximo marco dentro do período de tempo especificado antes de continuar e retornar o melhor modelo disponível quando o limite é atingido.

Máx. de Preditores. Essa opção permite limitar a complexidade do modelo e melhorar o desempenho ao limitar o número de preditores utilizados. Os preditores são classificados com base em uma medida MDL da correlação deles com o destino como uma medida da probabilidade de serem incluídos no modelo.

Máx. de Preditores Naive Bayes. Essa opção especifica o número máximo de preditores a serem utilizados no modelo Naive Bayes.

Oracle Support Vector Machine (SVM)

O Support Vector Machine (SVM) é um algoritmo de classificação e de regressão que usa a teoria de aprendizado por máquina para maximizar a precisão preditiva sem super ajustar os dados. O SVM utiliza uma transformação não linear opcional dos dados de treinamento, seguida pela procura por equações de regressão nos dados transformados para separar as classes (para variáveis resposta categóricas) ou ajustar a resposta (para variáveis resposta contínuas). A implementação do SVM da Oracle permite que os modelos sejam construídos utilizando um dos dois kernels disponíveis, linear ou Gaussiano. O kernel linear omite a transformação não linear completamente para que o modelo resultante seja essencialmente um modelo de regressão.

Para obter mais informações, consulte o *Guia do Oracle Data Mining Application Developer* e os *Conceitos de Mineração de Dados do Oracle*.

Opções do Modelo SVM do Oracle

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Campo exclusivo. Especifica o campo utilizado para identificar exclusivamente cada caso. Por exemplo, esse pode ser um campo de ID, como *CustomerID*. O IBM SPSS Modeler impõe uma restrição de que esse campo-chave deve ser numérico.

Nota: Este campo é opcional para todos os nós Oracle , exceto Oracle Adaptive Bayes, Oracle O-Cluster e Oracle Apriori.

Preparação de Dados Automática. (11*g* somente) Ativa (padrão) ou desativa o modo de preparação de dados automatizada de Mineração de Dados do Oracle. Se essa caixa estiver marcada, o ODM executa automaticamente as transformações de dados necessárias pelo algoritmo. Para obter mais informações, consulte *Conceitos de Mineração de Dados do Oracle*.

Aprendizado ativo. Fornece uma maneira de lidar com grandes conjuntos de construção. Com o aprendizado ativo, o algoritmo cria um modelo inicial com base em uma amostra pequena antes de

aplicá-lo ao conjunto de dados de treinamento completo e, em seguida, atualiza incrementalmente a amostra e o modelo com base nos resultados. O ciclo é repetido até que o modelo convirja nos dados de treinamento ou o número máximo permitido de vetores de suporte seja atingido.

Função Kernel. Selecione **Linear** ou **Gaussiana** ou deixe o padrão **Determinado pelo Sistema** para permitir que o sistema escolha o kernel mais adequado. Os kernels gaussianos são capazes de aprender relacionamentos mais complexos, mas geralmente demoram mais tempo para calcular. Talvez você queira iniciar com o kernel linear e tentar o kernel gaussiano apenas se o kernel linear não localizar um bom ajuste. Isto é mais provável de acontecer com um modelo de regressão, em que a opção de kernel interessa mais. Além disso, observe que os modelos do SVM construídos com o kernel gaussiano não podem ser procurados no IBM SPSS Modeler. Os modelos construídos com o kernel linear podem ser procurados no IBM SPSS Modeler da mesma maneira que os modelos de regressão padrão.

Método de Normalização. Especifica o método de normalização para campos de entrada e de destino contínuos. É possível escolher **Escore Z**, **Mín-Máx** ou **Nenhum**. O Oracle executará normalização automaticamente se a caixa de seleção **Preparação Automática de Dados** estiver selecionada. Desmarque essa caixa para selecionar o método de normalização manualmente.

Opções Avançadas do SVM do Oracle

Tamanho do Cache de Kernel. Especifica, em bytes, o tamanho do cache a ser utilizado para armazenar kernels calculados durante a operação de construção. Como é de se esperar, caches maiores geralmente resultam em construções mais rápidas. O padrão é 50 MB.

Tolerância de Convergência. Especifica o valor de tolerância que é permitido antes do término para a construção do modelo. O valor deve estar entre 0 e 1. O valor padrão é 0,001. Valores maiores tendem a resultar em modelos de construção mais rápida, mas menos precisos.

Especificar Desvio Padrão. Especifica o parâmetro de desvio padrão utilizado pelo kernel gaussiano. Esse parâmetro afeta o trade-off entre a complexidade do modelo e a possibilidade de generalizar para outros conjuntos de dados (causando um super ajuste ou subajuste dos dados). Valores de desvio padrão maiores favorecem o subajuste. Por padrão, esse parâmetro é estimado a partir dos dados de treinamento.

Especificar Epsilon. Para modelos de regressão somente, especifica o valor do intervalo do erro permitido na construção de modelos que não fazem distinção de epsilon. Em outras palavras, ele diferencia erros pequenos (que são ignorados) de erros grandes (que não são ignorados). O valor deve estar entre 0 e 1. Por padrão, isso é estimado a partir dos dados de treinamento.

Especificar Fator de Complexidade. Especifica o fator de complexidade, que faz um trade-off do erro do modelo (conforme medido com relação aos dados de treinamento) e da complexidade do modelo para evitar super ajuste ou subajuste dos dados. Valores mais altos impõem uma penalidade maior sobre os erros, com um risco maior de super ajuste dos dados, ao passo que valores mais baixos impõem uma penalidade menor sobre os erros, podendo levar a subajuste.

Especificar Taxa de Valores Discrepantes. Especifica a taxa desejada de valores discrepantes nos dados de treinamento. Essa opção é válida somente para modelos SVM de Primeira Classe. Ela não pode ser utilizada com a configuração **Especificar Fator de Complexidade**.

Probabilidade de predição. Permite que o modelo inclua a probabilidade de uma predição correta de um resultado possível do campo de destino. Para ativar esse recurso, escolha **Selecionar**, clique no botão **especificar** , escolha um dos resultados possíveis e, em seguida, clique em **Inserir**.

Usar Conjunto de Predição. Gera uma tabela para todos os resultados possíveis do campo de destino.

Opções de Ponderações do SVM do Oracle

Em um modelo de classificação, o uso de ponderações permite especificar a importância relativa dos diversos valores de destino possíveis. Isso pode ser útil, por exemplo, se os pontos de dados em seus dados de treinamento não estiverem realisticamente distribuídos entre as categorias. As ponderações permitem causar viés no modelo para poder compensar as categorias que forem menos

bem representadas nos dados. O aumento da ponderação de um valor de destino deve aumentar a porcentagem de predições corretas para essa categoria.

Existem três métodos de configuração de ponderações:

- **Baseado em dados de treinamento.** Este é o padrão. As ponderações baseiam-se nas frequências relativas das categorias nos dados de treinamento.
- **Igual para todas as classes.** As ponderações de todas as categorias são definidas como 1/*k*, em que *k* é o número de categorias de destino.
- **Customizado.** É possível especificar suas próprias ponderações. Os valores iniciais das ponderações são configurados como iguais para todas as classes. É possível ajustar as ponderações de categorias individuais para valores definidos pelo usuário. Para ajustar a ponderação de uma categoria específica, selecione a célula de ponderação na tabela correspondente à categoria desejada, exclua o conteúdo da célula, e digite o valor desejado.

As ponderações de todas as categorias devem somar 1,0. Se elas não somarem 1,0, um aviso será exibido, com uma opção para normalizar automaticamente os valores. Esse ajustamento automático preserva as proporções entre as categorias ao aplicar a restrição de ponderação. É possível executar este ajustamento a qualquer momento clicando no botão **Normalizar**. Para reconfigurar a tabela para valores iguais para todas as categorias, clique no botão **Igualar**.

Modelos Lineares Generalizados (GLM) do Oracle

(somente 11g) Os modelos lineares generalizados amenizam as suposições restritivas feitas pelos modelos lineares. Estas incluem, por exemplo, as suposições de que a variável de destino possui uma distribuição normal e que o efeito dos preditores na variável de destino for linear por natureza. Um modelo linear generalizado é adequado para predições em que a distribuição do destino deve ter uma distribuição não normal, como uma distribuição multinomial ou de Poisson. Da mesma forma, um modelo linear generalizado é útil nos casos em que o relacionamento ou a ligação entre os preditores e o destino deve ser não linear.

Para obter mais informações, consulte o *Guia do Oracle Data Mining Application Developer* e os *Conceitos de Mineração de Dados do Oracle*.

Opções de Modelo GLM do Oracle

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Campo exclusivo. Especifica o campo utilizado para identificar exclusivamente cada caso. Por exemplo, esse pode ser um campo de ID, como *CustomerID*. O IBM SPSS Modeler impõe uma restrição de que esse campo-chave deve ser numérico.

Nota: Este campo é opcional para todos os nós Oracle , exceto Oracle Adaptive Bayes, Oracle O-Cluster e Oracle Apriori.

Preparação de Dados Automática. (11*g* somente) Ativa (padrão) ou desativa o modo de preparação de dados automatizada de Mineração de Dados do Oracle. Se essa caixa estiver marcada, o ODM executa automaticamente as transformações de dados necessárias pelo algoritmo. Para obter mais informações, consulte *Conceitos de Mineração de Dados do Oracle*.

Método de Normalização. Especifica o método de normalização para campos de entrada e de destino contínuos. É possível escolher **Escore Z**, **Mín-Máx** ou **Nenhum**. O Oracle executará normalização automaticamente se a caixa de seleção **Preparação Automática de Dados** estiver selecionada. Desmarque essa caixa para selecionar o método de normalização manualmente.

Manipulação de Valor Omisso. Especifica como processar valores omissos nos dados de entrada:

- **Substituir pela média ou modo** substitui valores omissos de atributos numéricos pelo valor médio e substitui valores omissos de atributos categóricos pelo modo.

- **Usar somente registros completos** ignora registros com valores omissos.

Opções Avançadas de GLM do Oracle

Usar Ponderações da Linha. Marque esta caixa para ativar a lista suspensa adjacente, na qual é possível selecionar uma coluna que contém um fator de ponderação para as linhas.

Salvar Diagnósticos de Linha na Tabela. Marque esta caixa para ativar o campo de texto adjacente, no qual é possível inserir o nome de uma tabela para conter diagnósticos no nível da linha.

Nível de Confiança de Coeficiente. O grau de certeza, de 0,0 a 1,0, de que o valor predito para a resposta irá estar dentro de um intervalo de confiança calculado pelo modelo. Os limites de confiança são retornados com as estatísticas de coeficiente.

Categoria de Referência para Destino. Selecione **Customizado** para escolher um valor para o campo de destino a ser utilizado como uma categoria de referência ou deixe o valor padrão **Automático**.

Enrugar regressão. A regressão Ridge é uma técnica que compensa a situação em que houver um grau de correlação muito alto nas variáveis. É possível utilizar a opção **Automático** para permitir que o algoritmo controle o uso dessa técnica ou é possível controlar isso manualmente por meio das opções **Desativar** e **Ativar**. Se escolher ativar a regressão Ridge manualmente, será possível substituir o valor padrão do sistema para o parâmetro ridge ao inserir um valor no campo adjacente.

Produzir VIF para Regressão Ridge. Marque essa caixa se desejar produzir estatísticas de Variance Inflation Factor (VIF) quando o Ridge estiver sendo usado para regressão linear.

Probabilidade de predição. Permite que o modelo inclua a probabilidade de uma predição correta de um resultado possível do campo de destino. Para ativar esse recurso, escolha **Selecionar**, clique no botão **especificar** , escolha um dos resultados possíveis e, em seguida, clique em **Inserir**.

Usar Conjunto de Predição. Gera uma tabela para todos os resultados possíveis do campo de destino.

Opções de Ponderações de GLM do Oracle

Em um modelo de classificação, o uso de ponderações permite especificar a importância relativa dos diversos valores de destino possíveis. Isso pode ser útil, por exemplo, se os pontos de dados em seus dados de treinamento não estiverem realisticamente distribuídos entre as categorias. As ponderações permitem causar viés no modelo para poder compensar as categorias que forem menos bem representadas nos dados. O aumento da ponderação de um valor de destino deve aumentar a porcentagem de predições corretas para essa categoria.

Existem três métodos de configuração de ponderações:

- **Baseado em dados de treinamento.** Este é o padrão. As ponderações baseiam-se nas frequências relativas das categorias nos dados de treinamento.
- **Igual para todas as classes.** As ponderações de todas as categorias são definidas como 1/*k*, em que *k* é o número de categorias de destino.
- **Customizado.** É possível especificar suas próprias ponderações. Os valores iniciais das ponderações são configurados como iguais para todas as classes. É possível ajustar as ponderações de categorias individuais para valores definidos pelo usuário. Para ajustar a ponderação de uma categoria específica, selecione a célula de ponderação na tabela correspondente à categoria desejada, exclua o conteúdo da célula, e digite o valor desejado.

As ponderações de todas as categorias devem somar 1,0. Se elas não somarem 1,0, um aviso será exibido, com uma opção para normalizar automaticamente os valores. Esse ajustamento automático preserva as proporções entre as categorias ao aplicar a restrição de ponderação. É possível executar este ajustamento a qualquer momento clicando no botão **Normalizar**. Para reconfigurar a tabela para valores iguais para todas as categorias, clique no botão **Igualar**.

Árvore de decisão da Oracle

A Mineração de Dados do Oracle oferece um recurso clássico de Árvore de Decisão, com base nos algoritmos populares de Classificação e Árvore de Regressão. O modelo de Árvore de Decisão do ODM contém informações completas sobre cada nó, incluindo Confiança, Suporte e Critério de Divisão. A Regra completa para cada nó pode ser exibida e, além disso, um atributo substituto será fornecido para cada nó, a ser utilizado como substituto ao aplicar o modelo em um caso com valores omissos.

As árvores de decisão são populares porque elas são universalmente fáceis de aplicar e de entender. As árvores de decisão examinam cada atributo de entrada potencial a procura do “melhor” divisor, ou seja, o ponto de corte do atributo (`AGE > 55`, por exemplo) que divide os registros de dados de recebimento de dados em populações mais homogêneas. Após cada decisão de divisão, o ODM repete o processo de crescimento da árvore inteira e de criação de "folhas" terminais que representam populações semelhantes de registros, itens ou pessoas. Olhando para baixo a partir do nó da árvore raiz (por exemplo, a população total), as árvores de decisão fornecem regras legíveis humanas de instruções `IF A, then B`. Essas regras de árvore de decisão também fornecem o suporte e a confiança para cada nó da árvore.

Ao passo que as Adaptive Bayes Networks também podem fornecer regras simples e curtas que podem ser úteis no fornecimento de explicações para cada predição, as Árvores de Decisão fornecem regras de Mineração de Dados do Oracle integrais para cada decisão de divisão. As Árvores de Decisão também são úteis para o desenvolvimento de perfis detalhados dos melhores clientes, pacientes saudáveis, fatores associados a uma fraude, e assim por diante.

Opções de Modelo de Árvore de Decisão

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Campo exclusivo. Especifica o campo utilizado para identificar exclusivamente cada caso. Por exemplo, esse pode ser um campo de ID, como *CustomerID*. O IBM SPSS Modeler impõe uma restrição de que esse campo-chave deve ser numérico.

Nota: Este campo é opcional para todos os nós Oracle , exceto Oracle Adaptive Bayes, Oracle O-Cluster e Oracle Apriori.

Preparação de Dados Automática. (11*g* somente) Ativa (padrão) ou desativa o modo de preparação de dados automatizada de Mineração de Dados do Oracle. Se essa caixa estiver marcada, o ODM executa automaticamente as transformações de dados necessárias pelo algoritmo. Para obter mais informações, consulte *Conceitos de Mineração de Dados do Oracle*.

Métrica de impureza. Especifica qual métrica é utilizada para buscar a melhor questão de teste para dividir dados em cada nó. Os melhores divisor e valor de divisão são aqueles que resultam no maior aumento na homogeneidade do valor de destino para as entidades no nó. A homogeneidade é medida de acordo com uma métrica. As métricas suportadas são **gini** e **entropia**.

Opções Avançadas da Árvore de Decisão

Profundidade Máxima. Configura a profundidade máxima do modelo de árvore a ser construído.

Porcentagem mínima de registros em um nó. Configura a porcentagem do número mínimo de registros por nó.

Porcentagem mínima de registros para uma divisão. Configura o número mínimo de registros em um nó pai expresso como uma porcentagem do número total de registros usados para treinar o modelo. Nenhuma divisão será tentada se o número de registros estiver abaixo dessa percentagem.

Mínimo de registros em um nó. Configura o número mínimo de registros retornados.

Registros mínimos para uma divisão. Configura o número mínimo de registros em um nó pai expresso como um valor. Nenhuma divisão será tentada se o número de registros estiver abaixo desse valor.

Identificador de regra. Se selecionada, inclui no modelo uma sequência de caracteres para identificar o nó na árvore na qual uma determinada divisão é feita.

Probabilidade de predição. Permite que o modelo inclua a probabilidade de uma predição correta de um resultado possível do campo de destino. Para ativar esse recurso, escolha **Selecionar**, clique no botão **especificar** , escolha um dos resultados possíveis e, em seguida, clique em **Inserir**.

Usar Conjunto de Predição. Gera uma tabela para todos os resultados possíveis do campo de destino.

O-Cluster Oracle

O algoritmo Cluster-O do Oracle identifica agrupamentos que ocorrem naturalmente dentro de uma população de dados. Um armazenamento em cluster de particionamento Ortogonal (Cluster-O) é um algoritmo de clusterização proprietário da Oracle que cria um modelo de armazenamento em cluster hierárquico com base em grade, ou seja, cria partições de eixos paralelos (ortogonais) no espaço de atributo de entrada. O algoritmo opera recursivamente. A estrutura hierárquica resultante representa uma grade irregular que transforma em ladrilho o espaço do atributo nos clusters.

O algoritmo Cluster-O manipula atributos numéricos e categóricos, e o ODM seleciona automaticamente as melhores definições de cluster. O ODM fornece informações detalhadas do cluster, regras de cluster, valores de centroide de cluster, e pode ser utilizado para escorar uma população sobre sua associação de cluster.

Opções do Modelo de Cluster-O

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Campo exclusivo. Especifica o campo utilizado para identificar exclusivamente cada caso. Por exemplo, esse pode ser um campo de ID, como *CustomerID*. O IBM SPSS Modeler impõe uma restrição de que esse campo-chave deve ser numérico.

Nota: Este campo é opcional para todos os nós Oracle , exceto Oracle Adaptive Bayes, Oracle O-Cluster e Oracle Apriori.

Preparação de Dados Automática. (11*g* somente) Ativa (padrão) ou desativa o modo de preparação de dados automatizada de Mineração de Dados do Oracle. Se essa caixa estiver marcada, o ODM executa automaticamente as transformações de dados necessárias pelo algoritmo. Para obter mais informações, consulte *Conceitos de Mineração de Dados do Oracle*.

Número máximo de clusters. Configura o número máximo de clusters gerados.

Opções Avançadas de Cluster-O

Buffer Máximo. Configura o tamanho máximo do buffer.

Sensibilidade. Configura uma fração que especifica a densidade máxima necessária para separar um novo cluster. A fração é relacionada à densidade uniforme global.

k-Médias do Oracle

O algoritmo k-Médias do Oracle identifica agrupamentos que ocorrem naturalmente dentro de uma população de dados. O algoritmo k-Médias é um algoritmo de clusterização baseado em distância que particiona os dados em um número predeterminado de clusters (desde que haja casos distintos suficientes). Os algoritmos baseados em distância dependem de uma métrica de distância (função) para medir a semelhança entre os pontos de dados. Os pontos de dados são designados ao cluster mais próximo de acordo com a métrica de distância utilizada. O ODM fornece uma versão aprimorada do k-Médias.

O algoritmo k-Médias suporta clusters hierárquicos, manipula atributos numéricos e categóricos e corta a população no número de clusters especificado pelo usuário. O ODM fornece informações detalhadas

do cluster, regras de cluster, valores de centroide de cluster, e pode ser utilizado para escorar uma população sobre sua associação de cluster.

Opções do Modelo de K-Médias

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Campo exclusivo. Especifica o campo utilizado para identificar exclusivamente cada caso. Por exemplo, esse pode ser um campo de ID, como *CustomerID*. O IBM SPSS Modeler impõe uma restrição de que esse campo-chave deve ser numérico.

Nota: Este campo é opcional para todos os nós Oracle , exceto Oracle Adaptive Bayes, Oracle O-Cluster e Oracle Apriori.

Preparação de Dados Automática. (11*g* somente) Ativa (padrão) ou desativa o modo de preparação de dados automatizada de Mineração de Dados do Oracle. Se essa caixa estiver marcada, o ODM executa automaticamente as transformações de dados necessárias pelo algoritmo. Para obter mais informações, consulte *Conceitos de Mineração de Dados do Oracle*.

Número de clusters. Configura o número de clusters gerados

Função de Distância. Especifica qual função de distância é utilizada para Armazenamento em Cluster de k-Médias.

Critério de divisão. Especifica qual critério de divisão é utilizado para Armazenamento em Cluster de k-Médias.

Método de Normalização. Especifica o método de normalização para campos de entrada e de destino contínuos. É possível escolher **Escore Z**, **Mín-Máx** ou **Nenhum**.

Opções Avançadas de K-Médias

Iterações. Configura o número de iterações para o algoritmo k-Médias.

Tolerância de convergência. Configura a tolerância de convergência para o algoritmo k-Médias.

Número de categorias. Especifica o número de categorias no histograma de atributo produzida pelo k-Médias. Os limites de categoria para cada atributo são calculados globalmente no conjunto de dados de treinamento inteiro. O método de categorização é Equi-width. Todos os atributos possuem o mesmo número de categorias, com exceção dos atributos com um valor único que possui apenas uma categoria.

Bloquear crescimento. Configura o fator de crescimento para memória alocada para retenção dos dados do cluster.

Suporte de Atributo de Porcentagem Mínima. Configura a fração dos valores de atributo que devem ser não nulos para que o atributo seja incluído na descrição da regra para o cluster. Configurar o valor de parâmetro para muito alto em dados com valores omissos pode resultar em regras muito curtas ou até mesmo vazias.

Oracle Nonnegative Matrix Factorization (NMF)

O Nonnegative Matrix Factorization (NMF) é útil para reduzir um conjunto de dados grande em atributos representativos. Semelhante ao Principal Components Analysis (PCA) em conceito, mas capaz de manipular quantias maiores de atributos e em um modelo de representação aditivo, o NMF é um algoritmo de mineração de dados moderno e poderoso que pode ser usado para uma variedade de casos de uso.

O NMF pode ser usado para reduzir grandes quantias de dados, por exemplo, dados de texto, em representações menores e mais esparsas que reduzem a dimensionalidade dos dados (as mesmas informações podem ser preservadas usando muito menos variáveis). A saída de modelos do NMF pode ser analisada utilizando técnicas de aprendizado supervisionado, como SVMs, ou técnicas de aprendizado

não supervisionado, como técnicas de armazenamento em cluster. A Mineração de Dados do Oracle utiliza os algoritmos NMF e SVM para extrair dados de texto não estruturado.

Opções do Modelo de NMF

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Campo exclusivo. Especifica o campo utilizado para identificar exclusivamente cada caso. Por exemplo, esse pode ser um campo de ID, como *CustomerID*. O IBM SPSS Modeler impõe uma restrição de que esse campo-chave deve ser numérico.

Nota: Este campo é opcional para todos os nós Oracle , exceto Oracle Adaptive Bayes, Oracle O-Cluster e Oracle Apriori.

Preparação de Dados Automática. (11*g* somente) Ativa (padrão) ou desativa o modo de preparação de dados automatizada de Mineração de Dados do Oracle. Se essa caixa estiver marcada, o ODM executa automaticamente as transformações de dados necessárias pelo algoritmo. Para obter mais informações, consulte *Conceitos de Mineração de Dados do Oracle*.

Método de Normalização. Especifica o método de normalização para campos de entrada e de destino contínuos. É possível escolher **Escore Z**, **Mín-Máx** ou **Nenhum**. O Oracle executará normalização automaticamente se a caixa de seleção **Preparação Automática de Dados** estiver selecionada. Desmarque essa caixa para selecionar o método de normalização manualmente.

Opções Avançadas do NMF

Especificar o número de variáveis. Especifica o número de variáveis a serem extraídas.

Semente aleatória. Configura a semente aleatória para o algoritmo NMF.

Número de iterações. Configura o número de iterações para o algoritmo NMF.

Tolerância de convergência. Configura a tolerância de convergência para o algoritmo NMF.

Exibir todas as variáveis. Exibe o ID e a confiança de variável para todas as variáveis, ao invés dos valores somente da melhor variável.

A priori da Oracle

O algoritmo a priori descobre as regras de associação nos dados. Por exemplo, "se um cliente comprar uma lâmina de barbear e uma loção pós-barba, então esse cliente comprará um creme de barbear com 80% de confiança". O problema de mineração de associação pode ser decomposto em dois subproblemas:

- Localizar todas as combinações de itens, chamados de conjuntos de itens frequentes, cujo suporte é maior que o suporte mínimo.
- Utilizar os conjuntos de itens frequentes para gerar as regras desejadas. A ideia é que, por exemplo, se ABC e BC forem frequentes, então a regra "A implica BC" se aplicará se a razão de `support(ABC)` para `support(BC)` for pelo menos tão grande quanto a confiança mínima. Observe que a regra terá um suporte mínimo porque ABCD é frequente. A Associação do ODM suporta apenas as regras subsequentes únicas (ABC implica D).

O número de conjuntos de itens frequente é controlado pelos parâmetros de suporte mínimo. O número de regras geradas é controlado pelo número de conjuntos de itens frequentes e o pelo parâmetro de confiança. Se o parâmetro de confiança for configurado muito alto, poderá haver conjuntos de itens frequentes no modelo de associação, mas nenhuma regra.

O ODM utiliza uma implementação baseada em SQL do algoritmo a priori. Os passos de contagem de geração e de suporte de candidato são implementados utilizando queries SQL. Estruturas de dados na memória especializadas não são utilizadas. As queries SQL estão ajustadas de modo preciso para execução eficiente no servidor da base de dados utilizando várias sugestões.

Opções de Campos a priori

Todos os nós de modelagem possuem uma guia Campos, na qual é possível especificar os campos a serem utilizados na construção do modelo.

Antes de poder construir um modelo a priori, é necessário especificar quais campos você deseja utilizar como os itens de interesse na modelagem de associação.

Usar configurações de nó Tipo. Esta opção instrui o nó a usar as informações de campo de um nó Tipo de envio de dados. Este é o padrão.

Usar configurações customizadas. Essa opção instrui o nó a utilizar as informações de campo especificadas aqui ao invés das informações fornecidas em qualquer nó ou nós Tipo de envio de dados. Após selecionar essa opção, especifique os campos restantes no diálogo, que dependem se você está utilizando o formato transacional.

Se você *não* estiver usando o formato transacional, especifique:

- **Entradas.** Selecione um ou mais campos de entrada. Isso é semelhante a configurar o papel do campo para *Entrada* em um nó Tipo.
- **Partição.** Este campo permite especificar um campo utilizado para particionar os dados em amostras separadas para os estágios de treinamento, de teste e de validação de construção de modelo.

Se você *estiver* usando o formato transacional, especifique:

Usar formato transacional. Utilize esta opção se desejar transformar os dados de uma linha por item em uma linha por caso.

Selecionar essa opção altera os controles de campo na parte inferior da caixa de diálogo:

Para o formato transacional, especifique:

- **ID.** Selecione um campo de ID na lista. Campos numéricos ou simbólicos podem ser utilizados como o campo de ID. Cada valor exclusivo deste campo deve indicar uma unidade específica de análise. Por exemplo, em um aplicativo de cesta de mercado, cada ID pode representar um cliente único. Para um aplicativo de análise de log da web, cada ID pode representar um computador (pelo endereço IP) ou um usuário (pelos dados de login).
- **Conteúdo.** Especifique o campo conteúdo para o modelo. Esse campo contém o item de interesse na modelagem de associação.
- **Partição.** Este campo permite especificar um campo utilizado para particionar os dados em amostras separadas para os estágios de treinamento, de teste e de validação de construção de modelo. Ao utilizar uma amostra para criar o modelo e uma amostra diferente para testá-lo, é possível obter uma boa indicação do quão bem o modelo será generalizado para conjuntos de dados maiores que forem semelhantes aos dados atuais. Se diversos campos de partição tiverem sido definidos usando os nós Tipo ou Partição, um campo de partição único deverá ser selecionado na guia Campos em cada nó de modelagem que utiliza particionamento. (Se apenas uma partição estiver presente, ela será utilizada automaticamente sempre que o particionamento estiver ativado). Além disso, observe que para aplicar a partição selecionada à sua análise, o particionamento também deverá ser ativado na guia Opções de Modelo para o nó. (Desmarcar esta opção permite desativar o particionamento sem alterar as configurações do campo).

Opções do Modelo a priori

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Campo exclusivo. Especifica o campo utilizado para identificar exclusivamente cada caso. Por exemplo, esse pode ser um campo de ID, como *CustomerID*. O IBM SPSS Modeler impõe uma restrição de que esse campo-chave deve ser numérico.

Nota: Este campo é opcional para todos os nós Oracle , exceto Oracle Adaptive Bayes, Oracle O-Cluster e Oracle Apriori.

Preparação de Dados Automática. (11*g* somente) Ativa (padrão) ou desativa o modo de preparação de dados automatizada de Mineração de Dados do Oracle. Se essa caixa estiver marcada, o ODM executa automaticamente as transformações de dados necessárias pelo algoritmo. Para obter mais informações, consulte *Conceitos de Mineração de Dados do Oracle*.

Comprimento máximo da regra. Configura o número máximo de condições prévias para qualquer regra, um número inteiro de 2 a 20. Esta é uma maneira de limitar a complexidade das regras. Se as regras forem muito complexas ou muito específicas, ou se o seu conjunto de regras está levando muito tempo para treinar, tente diminuir esta configuração.

Confiança mínima. Estabelece o nível de confiança mínimo, um valor entre 0 e 1. As regras com confiança menor que o critério especificado são descartadas.

Suporte mínimo. Configura o limite mínimo de suporte, um valor entre 0 e 1. Apriori descobre padrões com frequência acima do limite mínimo de suporte.

Oracle Minimum Description Length (MDL)

O algoritmo Oracle Minimum Description Length (MDL) ajuda a identificar os atributos que tiverem a maior influência sobre um atributo de destino. Muitas vezes, saber quais atributos são mais influentes ajuda a entender e a gerenciar melhor seus negócios e pode ajudar a simplificar as atividades de modelagem. Além disso, esses atributos podem indicar os tipos de dados que talvez você queira incluir para aumentar seus modelos. O MDL pode ser utilizado, por exemplo, para localizar os atributos de processo mais relevantes para prever a qualidade de uma peça fabricada, os fatores associados à perda de clientes ou os genes que mais podem estar envolvidos no tratamento de uma doença específica.

O Oracle MDL descarta campos de entrada que considerar como não importantes na predição do destino. Com os campos de entrada restantes, ele, então, constrói um nugget do modelo não refinado que está associado a um modelo do Oracle, visível no Oracle Data Miner. Navegar no modelo do Oracle Data Miner exibe um gráfico mostrando os campos de entrada restantes, classificados em ordem de sua significância na predição do destino.

Um ranqueamento negativo indica ruído. Os campos de entrada classificados em zero ou menos não contribuem com a predição devem provavelmente ser removidos dos dados.

Para exibir o gráfico

1. Clique com o botão direito no nugget do modelo não refinado na paleta Modelos e escolha **Procurar**.
2. Na janela modelo, clique no botão para ativar o Oracle Data Miner.
3. Conecte-se ao Oracle Data Miner. Consulte o tópico “Oracle Data Miner” na página 43 para obter informações adicionais.
4. No painel do navegador do Oracle Data Miner, expanda **Modelos** e, em seguida, **Importância de Atributo**.
5. Selecione o modelo Oracle relevante (ele terá o mesmo nome que o campo de destino especificado no IBM SPSS Modeler). Se você não tiver certeza de qual é o correto, selecione a pasta Importância de Atributo e procure um modelo por data de criação.

Opções do Modelo de MDL

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Campo exclusivo. Especifica o campo utilizado para identificar exclusivamente cada caso. Por exemplo, esse pode ser um campo de ID, como *CustomerID*. O IBM SPSS Modeler impõe uma restrição de que esse campo-chave deve ser numérico.

Nota: Este campo é opcional para todos os nós Oracle , exceto Oracle Adaptive Bayes, Oracle O-Cluster e Oracle Apriori.

Preparação de Dados Automática. (11*g* somente) Ativa (padrão) ou desativa o modo de preparação de dados automatizada de Mineração de Dados do Oracle. Se essa caixa estiver marcada, o ODM executa automaticamente as transformações de dados necessárias pelo algoritmo. Para obter mais informações, consulte *Conceitos de Mineração de Dados do Oracle*.

Oracle Attribute Importance (AI)

O objetivo da importância de atributo é descobrir quais atributos no conjunto de dados estão relacionados ao resultado e o grau para o qual eles influenciam o resultado final. O nó Oracle Attribute Importance analisa dados, localiza padrões e prevê resultados com um nível de confiança associado.

Opções do Modelo AI

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Preparação de Dados Automática. (11*g* somente) Ativa (padrão) ou desativa o modo de preparação de dados automatizada de Mineração de Dados do Oracle. Se essa caixa estiver marcada, o ODM executa automaticamente as transformações de dados necessárias pelo algoritmo. Para obter mais informações, consulte *Conceitos de Mineração de Dados do Oracle*.

Opções de Seleção AI

A guia Opções permite especificar as configurações padrão para selecionar ou excluir campos de entrada no nugget do modelo. Em seguida, é possível incluir o modelo em um fluxo para selecionar um subconjunto de campos para uso em esforços de construção de modelo subsequentes. Como alternativa, é possível substituir essas configurações ao selecionar ou cancelar seleção de campos adicionais no navegador do modelo após gerar o modelo. No entanto, as configurações padrão permitem aplicar o nugget do modelo sem mudanças adicionais, o que pode ser particularmente útil para propósitos de script.

As opções a seguir estão disponíveis:

Todos os campos classificados. Seleciona campos com base em seu ranqueamento, como *importante*, *marginal* ou *não importante*. É possível editar o rótulo de cada ranqueamento e também os valores de corte utilizados para designar registros para um ranqueamento ou outro.

Número máximo de campos. Seleciona os *n* principais campos com base na importância.

Importância maior que. Seleciona todos os campos com uma importância maior que o valor especificado.

O campo de destino é sempre preservado, independentemente da seleção.

Guia Modelo do Nugget do Modelo da AI

A guia Modelo de um nugget do modelo Oracle AI exibe o ranqueamento e a importância de todas as entradas e permite selecionar os campos para filtragem utilizando as caixas de seleção na coluna à esquerda. Ao executar o fluxo, apenas os campos verificados são preservados, junto da predição de destino. Os outros campos de entrada são descartados. As seleções padrão baseiam-se nas opções especificadas no nó de modelagem, e também é possível selecionar ou cancelar a seleção de campos adicionais conforme necessário.

- Para ordenar a lista por classificação, nome do campo, importância ou qualquer uma das outras colunas exibidas, clique no cabeçalho da coluna. Como alternativa, selecione o item desejado na lista ao lado do botão Ordenar Por e utilize as setas para cima e para baixo para alterar a direção da ordenação.

- É possível utilizar a barra de ferramentas para marcar ou desmarcar todos os campos e acessar a caixa de diálogo Verificar Campos, que permite selecionar os campos por ranqueamento ou importância. Também é possível pressionar as teclas Shift ou Ctrl enquanto clica nos campos para estender a seleção.
- Os valores limite para classificar entradas como importantes, marginais ou não importantes são exibidos na legenda abaixo da tabela. Esses valores são especificados no nó de modelagem.

Gerenciando Modelos do Oracle

Os modelos do Oracle são incluídos na paleta Modelos exatamente como outros modelos do IBM SPSS Modeler e podem ser utilizados de forma muito semelhante. No entanto, há algumas diferenças importantes, dado que cada modelo do Oracle criado no IBM SPSS Modeler, na verdade, referencia um modelo armazenado em um servidor de base de dados.

Guia Servidor do Nugget do Modelo do Oracle

Construir um modelo do ODM por meio do IBM SPSS Modeler cria um modelo no IBM SPSS Modeler e cria ou substitui um modelo no banco de dados do Oracle. Um modelo do IBM SPSS Modeler desse tipo referencia o conteúdo de um modelo de banco de dados armazenado em um servidor de banco de dados. O IBM SPSS Modeler pode executar verificação de consistência ao armazenar um **modelo de chave** gerado idêntico no modelo do IBM SPSS Modeler e no modelo do Oracle.

A sequência de caracteres de chave para cada modelo do Oracle é exibida na coluna *Informações do Modelo* na caixa de diálogo Modelos de Lista. A sequência de caracteres de chave para um modelo do IBM SPSS Modeler é exibida como a **Chave de Modelo** na guia Servidor de um modelo do IBM SPSS Modeler (quando colocada em um fluxo).

O botão Verificar na guia Servidor de um nugget do modelo pode ser utilizado para verificar se as chaves de modelo no modelo do IBM SPSS Modeler e no modelo do Oracle correspondem. Se nenhum modelo com o mesmo nome puder ser localizado no Oracle ou se as chaves de modelo não corresponderem, o modelo Oracle foi excluído ou reconstruído desde que o modelo do IBM SPSS Modeler foi construído.

Guia Sumarização do Nugget do Modelo do Oracle

A guia Sumarização de um nugget do modelo exibe informações sobre o modelo em si (*Análise*), sobre os campos usados no modelo (*Campos*), sobre as configurações utilizadas ao construir o modelo (*Configurações de Construção*) e sobre o treinamento do modelo (*Sumarização do Treinamento*).

Ao procurar o nó pela primeira vez, os resultados da guia Sumarização são reduzidos. Para ver os resultados de interesse, utilize o controle expansor à esquerda de um item para desdobrá-lo ou clique no botão **Expandir Tudo** para mostrar todos os resultados. Para ocultar os resultados após terminar de visualizá-los, use o controle expansor para reduzir os resultados específicos que deseja ocultar ou clique no botão **Reduzir Tudo** para reduzir todos os resultados.

Análise. Exibe informações sobre o modelo específico. Se tiver executado um nó Análise anexado a este nugget do modelo, as informações dessa análise também serão exibidas nesta seção.

Campos. Lista os campos utilizados como o destino e as entradas na construção do modelo.

Configurações de construção. Contém informações sobre as configurações utilizadas na construção do modelo.

Sumarização do Treinamento. Mostra o tipo de modelo, o fluxo utilizado para criá-lo, o usuário que o criou, quando ele foi construído e o tempo decorrido para construir o modelo.

Guia Configurações do Nugget do Modelo do Oracle

A guia Configurações no nugget do modelo permite substituir a configuração de determinadas opções no nó de modelagem para propósitos de escoragem.

Árvore de decisão da Oracle

Usar custos de classificação errada. Determina se os custos de classificação errada devem ser usados no modelo de Árvore de Decisão do Oracle. Veja o tópico "Custos de classificação errada" na página 28 para obter mais informações.

Identificador de regra. Se selecionada (marcada), inclui uma coluna do identificador da regra no modelo de Árvore de Decisão do Oracle. O identificador da regra identifica o nó na árvore em que uma determinada divisão é feita.

NMF da Oracle

Exibir todas as variáveis. Se selecionada (marcada), exibe o ID e a confiança de variável para todas as variáveis, ao invés dos valores somente da melhor variável, no modelo do Oracle NMF.

Listando Modelos do Oracle

O botão Listar Modelos de Mineração de Dados do Oracle ativa uma caixa de diálogo que lista os modelos de banco de dados existentes e permite que os modelos sejam removidos. Essa caixa de diálogo pode ser ativada a partir da caixa de diálogo Aplicativos Auxiliares e a partir das caixas de diálogo de construção, navegação e aplicação de nós relacionados ao ODM.

As informações a seguir são exibidas para cada modelo:

- **Nome do Modelo.** O nome do modelo, que é utilizado para ordenar a lista.
- **Informações de modelo.** Informações chave de modelo compostas de data/hora da construção e do nome da coluna de destino
- **Tipo de modelo.** Nome do algoritmo que construiu este modelo

Oracle Data Miner

O Oracle Data Miner é a interface com o usuário para Mineração de Dados do Oracle (ODM) e substitui a interface com o usuário anterior do IBM SPSS Modeler para ODM. O Oracle Data Miner é projetado para aumentar a taxa de sucesso do analista no uso correto de algoritmos ODM. Esses objetivos são tratados de várias maneiras:

- Os usuários precisam de assistência maior na aplicação de uma metodologia que aborda a preparação de dados e a seleção de algoritmo. O Oracle Data Miner supre essa necessidade ao fornecer Atividades de Mineração de Dados para guiar os usuários através da metodologia adequada.
- O Oracle Data Miner inclui heurística melhorada e expandida nos assistentes de construção e de transformação de modelo para reduzir a chance de erro na especificação de configurações de modelo e de transformação.

Definindo uma Conexão com o Oracle Data Miner

1. O Oracle Data Miner pode ser ativado a partir de todos os nós de aplicação de construção Oracle e de caixas de diálogo de saída por meio do botão **Iniciar o Oracle Data Miner**.

   

   Figura 2. Botão Iniciar o Oracle Data Miner
2. A caixa de diálogo **Editar Conexão** do Oracle Data Miner é apresentada ao usuário antes de o aplicativo externo Oracle Data Miner ser ativado (desde que a opção Aplicativo Auxiliar esteja definida corretamente).

   Nota: esta caixa de diálogo é exibida apenas na ausência de um nome de conexão definido.

   - Forneça um nome de conexão do Data Miner e insira as informações apropriadas do servidor Oracle 10gR2 ou 10gR1. O servidor Oracle deve ser o mesmo servidor especificado no IBM SPSS Modeler.
3. A caixa de diálogo **Escolher Conexão** do Oracle Data Miner fornece opções para especificar qual nome de conexão definido no passo acima é utilizado.

Consulte o Oracle Data Miner no website da Oracle para obter mais informações sobre os requisitos, instalação e uso do Oracle Data Miner.

Preparando os Dados

Dois tipos de preparação de dados podem ser úteis quando estiver utilizando o Naive Bayes, o Adaptive Bayes e o Support Vector Machine fornecidos com os algoritmos de Mineração de Dados do Oracle na modelagem:

- **Categorização**, ou conversão de campos de amplitude numérica contínua para categorias de algoritmos que não podem aceitar dados contínuos.
- **Normalização**, ou transformações aplicadas a intervalos numéricos para que eles tenham médias e desvios padrão similares.

Categorização

O nó Categorização do IBM SPSS Modeler oferece um número de técnicas para executar operações de categorização. Uma operação de categorização é definida e pode ser aplicada a um ou diversos campos. Executar a operação de categorização em um conjunto de dados cria os limites e permite que um nó Derivar do IBM SPSS Modeler seja criado. A operação de derivação pode ser convertida em SQL e aplicada antes da construção e da escoragem do modelo. Esta abordagem cria uma dependência entre o modelo e o nó Derivar que executa a categorização, e também permite que as especificações de categorização sejam reutilizadas por diversas tarefas de modelagem.

Normalização

Campos contínuos (intervalo numérico) que são utilizados como entradas para os modelos do Support Vector Machine devem ser normalizados antes da construção de modelo. No caso de modelos de regressão, a normalização também deve ser revertida para reconstruir o escore na saída do modelo. As configurações de modelo SVM permitem escolher **Escore Z**, **Mín-Máx** ou **Nenhum**. Os coeficientes de normalização são construídos pelo Oracle como um passo no processo de construção de modelo, e os coeficientes são transferidos por upload para o IBM SPSS Modeler e armazenados com o modelo. No tempo de aplicação, os coeficientes são convertidos em expressões de derivação do IBM SPSS Modeler e utilizados para preparar os dados para escoragem antes de transmitir esses dados para o modelo. Neste caso, a normalização está estritamente associada à tarefa de modelagem.

Exemplos de Mineração de Dados do Oracle

Diversos fluxos de amostra são incluídos que demonstram o uso de ODM com o IBM SPSS Modeler. Esses fluxos podem ser localizados na pasta de instalação do IBM SPSS Modeler em *\Demos\Database_Modelling\Oracle Data Mining*.

Nota: a pasta Demos pode ser acessada a partir do grupo do programa do IBM SPSS Modeler no menu Iniciar do Windows.

Os fluxos na tabela a seguir podem ser utilizados juntos e em sequência como um exemplo do processo de mineração da base de dados, utilizando o algoritmo Support Vector Machine (SVM) que é fornecido com a Mineração de Dados do Oracle:

Tabela 4. Mineração da base de dados - fluxos de exemplo

Fluxo	**Descrição**
1_upload_data.str	Utilizado para limpar e fazer upload de dados de um arquivo simples para o banco de dados.
2_explore_data.str	Fornece um exemplo de exploração de dados com IBM SPSS Modeler
3_build_model.str	Constrói o modelo utilizando o algoritmo nativo do banco de dados.
4_evaluate_model.str	Usado como exemplo de avaliação de modelo com IBM SPSS Modeler

Tabela 4. Mineração da base de dados - fluxos de exemplo (continuação)	
Fluxo	**Descrição**
5_deploy_model.str	Implementa o modelo para escoragem dentro do bando de dados.

Nota: para executar o exemplo, os fluxos devem ser executados em ordem. Além disso, os nós de origem e de modelagem em cada fluxo devem ser atualizados para referenciar uma origem de dados válida para o banco de dados que deseja utilizar.

O conjunto de dados utilizado nos fluxos de exemplos refere-se a aplicativos de cartão de crédito e apresenta um problema de classificação com uma combinação de preditores categóricos e contínuos. Para obter mais informações sobre esse conjunto de dados, consulte o arquivo *crx.names* na mesma pasta que os fluxos de amostra.

Este conjunto de dados está disponível a partir do UCI Machine Learning Repository em *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/*.

Fluxo de Exemplo: Upload de Dados

O primeiro fluxo de exemplo, *1_upload_data.str*, é utilizado para limpar e fazer upload de dados de um arquivo simples para o Oracle.

Já que o Oracle Data Mining requer um campo ID exclusivo, este fluxo inicial usa um nó de Derivação para adicionar um novo campo ao dataset chamado *ID*, com valores exclusivos 1,2,3, usando a função IBM SPSS Modeler `@INDEX` .

O nó Preenchimento é utilizado para manipulação de valores omissos e substitui campos vazios que são lidos a partir do arquivo de texto *crx.data* por valores *NULL*.

Fluxo de Exemplo: Explorar Dados

O segundo fluxo de exemplo, *2_explore_data.str*, é utilizado para demonstrar o uso de um nó de Auditoria de Dados para obter uma visão geral dos dados, incluindo estatísticas de sumarização e gráficos.

Dar um clique duplo em um gráfico no Relatório de Auditoria de Dados produz um gráfico mais detalhado para exploração mais profunda de um campo determinado.

Fluxo de Exemplo: Construir o Modelo

O terceiro fluxo de exemplo, *3_build_model.str*, ilustra a construção de modelo no IBM SPSS Modeler. Dê um clique duplo no nó de origem do banco de dados (denominado CREDIT) para especificar a origem de dados. Para especificar as configurações de construção, dê um clique duplo no nó de construção (inicialmente rotulado como CLASS, que é alterado para FIELD16 quando a origem de dados for especificada).

Na guia Modelo da caixa de diálogo:

1. Assegure-se de que **ID** esteja selecionado como o campo Exclusivo.
2. Assegure-se de que **Linear** esteja selecionado como a função de kernel e que **z** seja o método de normalização.

Fluxo de Exemplo: Avaliar o Modelo

O quarto fluxo de exemplo, *4_evaluate_model.str*, ilustra as vantagens de utilizar o IBM SPSS Modeler para modelagem dentro da base de dados. Após ter executado o modelo, será possível incluí-lo de volta no fluxo de dados e avaliar o modelo ao utilizar várias ferramentas oferecidas no IBM SPSS Modeler.

Visualizando Resultados de Modelagem

Anexe um nó Tabela ao nugget do modelo para explorar os resultados. O campo **$O-field16** mostra o valor predito para *field16* em cada caso, e o campo **$OC-field16** mostra o valor de confiança para essa predição.

Avaliando Resultados de Modelo

É possível utilizar o nó Análise para criar uma matriz de coincidência mostrando o padrão de correspondências entre cada campo predito e seu campo de destino. Execute o nó Análise para ver os resultados.

É possível utilizar o nó Avaliação para criar um gráfico de ganhos designado a mostrar as melhorias de precisão feitas pelo modelo. Execute o nó Avaliação para ver os resultados.

Fluxo de Exemplo: Implementar o Modelo

Quando estiver satisfeito com a precisão do modelo, ele poderá ser implementado para uso com aplicativos externos ou para publicar de volta no banco de dados. No fluxo de exemplo final, *5_deploy_model.str*, os dados são lidos a partir da tabela CREDITDATA e, em seguida, escorados e publicados na tabela CREDITSCORES usando o nó Publicação chamado *implementar solução*.

Capítulo 5. Modelagem de Banco de Dados com IBM Data Warehouse e AnáliseIBM Netezza

ModeladorSPSS com IBM Data Warehouse e AnáliseIBM Netezza

IBM SPSS Modeler suporta integração com o IBM Data Warehouse e AnáliseIBM Netezza, que fornece a capacidade de executar analytics avançados nesses servidores IBM . Esses recursos podem ser acessados por meio da interface gráfica com o usuário do IBM SPSS Modeler e do ambiente de desenvolvimento orientado a fluxo de trabalho, permitindo que você execute os algoritmos de mineração de dados diretamente no ambiente IBM Netezza ou IBM Data Warehouse.

ModeladorSPSS suporta a integração dos seguintes algoritmos a partir de **AnáliseIBM Netezza**:

- Árvores de decisão
- K-Médias
- TwoStep
- Rede Bayes
- Naive Bayes
- KNN
- Armazenamento em Cluster Decisivo
- PCA
- Árvore de Regressão
- Regressão linear
- Séries temporais
- Linear generalizado

Para obter mais informações sobre esses algoritmos, consulte o *AnáliseIBM Netezza Developer's Guide* e o *AnáliseIBM Netezza Guia de Referência*.

ModeladorSPSS suporta a integração dos seguintes algoritmos a partir do **IBM Data Warehouse** (Bayes Net, Divisive Clustering e Time Series não são suportados):

- Árvores de decisão
- K-Médias
- TwoStep
- Naive Bayes
- KNN
- PCA
- Árvore de Regressão
- Regressão linear
- Linear generalizado

Nota: O AIX não é suportado.

Requisitos de integração

As condições a seguir são pré-requisitos para a realização de modelagem em banco de dados usando AnáliseIBM Netezza ou IBM Data Warehouse. Poderá ser necessário consultar seu administrador de base de dados para assegurar que essas condições sejam atendidas.

- O IBM SPSS Modeler em execução com relação a uma instalação do ServidorIBM SPSS Modeler no Windows ou UNIX (exceto zLinux, para o qual os drivers ODBC do IBM Netezza não estão disponíveis).
- IBM Netezza Performance Server, executando o pacote AnáliseIBM Netezza.

  Nota: a versão mínima do Netezza Performance Server (NPS) que é necessária depende da versão do INZA necessária e é a seguinte:

  – Qualquer versão maior que o NPS 6.0.0 P8 suportará versões do INZA anteriores a 2.0.

  – Usar o INZA 2.0 ou superior requer o NPS P5 6.0.5 ou superior.

  Os nós Linear Generalizado do Netezza e Série Temporal do Netezza requerem o INZA 2.0 e superior para funcionar. Todos os outros nós dentro da base de dados do Netezza precisam do INZA 1.1 ou posterior.
- Uma origem de dados ODBC para conexão com um banco de dados do IBM Netezza. Consulte o tópico "Ativação da integração" na página 48 para obter informações adicionais.
- Uma fonte de dados ODBC para conexão com um banco de dados do IBM Data Warehouse.
- Geração e otimização de SQL ativadas no IBM SPSS Modeler. Veja o tópico "Ativação da integração" na página 48 para obter mais informações.

Nota: A modelagem da base de dados e a otimização de SQL requerem que a conectividade do ServidorIBM SPSS Modeler esteja ativada no computador do IBM SPSS Modeler. Com essa configuração ativada, é possível acessar os algoritmos de banco de dados, realizar SQL pushback diretamente do IBM SPSS Modeler e acessar o ServidorIBM SPSS Modeler. Para verificar o atual status da licença, escolha o seguinte no menu do IBM SPSS Modeler.

Ajuda > **Sobre** > **Detalhes Adicionais**

Se a conectividade estiver ativada, você verá a opção **Ativação do Servidor** na guia Status da Licença.

Ativação da integração

A ativação da integração com o AnáliseIBM Netezza ou o IBM Data Warehouse consiste nas seguintes etapas.

- Configurando AnáliseIBM Netezza ou IBM Data Warehouse
- Criar uma origem ODBC
- Habilitando a integração em IBM SPSS Modeler
- Habilitando a geração SQL e a otimização em IBM SPSS Modeler

Esses passos são descritos nas seções a seguir.

Configurando AnáliseIBM Netezza ou IBM Data Warehouse

Para instalar e configurar AnáliseIBM Netezza ou IBM Data Warehouse, consulte a documentação apropriada do IBM . Por exemplo, para AnáliseIBM Netezza, consulte o *AnáliseIBM Netezza Guia de Instalação* fornecido com aquele produto. A seção *Configurando Permissões do Banco de Dados* no guia contém detalhes de scripts que precisam ser executados para permitir que os fluxos do IBM SPSS Modeler sejam gravados no banco de dados.

Nota: Se você estará usando nós que contam com cálculo matricial, o Matrix Engine deve ser inicializado por rodar `CALL NZM..INITIALIZE();` caso contrário, execução de procedimentos armazenados falhará. Inicialização é um passo de configuração único para cada banco de dados.

Criando uma Fonte ODBC para AnáliseIBM Netezza

Para ativar a conexão entre o banco de dados IBM Netezza e o IBM SPSS Modeler, é necessário criar um nome da origem de dados (DSN) do sistema ODBC.

Antes de criar um DSN, você deverá ter um entendimento básico das origens de dados e drivers ODBC, bem como do suporte ao banco de dados no IBM SPSS Modeler.

Se estiver executando no modo distribuído com relação ao ServidorIBM SPSS Modeler, crie o DSN no computador servidor. Se estiver executando no modo local (cliente), crie o DSN no computador cliente.

Clientes Windows

1. No CD *Netezza Client*, execute o arquivo *nzodbcsetup.exe* para iniciar o instalador. Siga as instruções na tela para instalar o driver. Para obter instruções completas, consulte o Guia de instalação e configuração do IBM Netezza ODBC, JDBC e OLE DB.
   a. Criar o DSN.

      Nota: A sequência do menu depende da sua versão do Windows.

      - **Windows XP.** No menu Iniciar, escolha **Painel de Controle**. Dê um clique duplo em **Ferramentas Administrativas** e, em seguida, um clique duplo em **Origens de Dados (ODBC)**.
      - **Windows Vista.** No menu Iniciar, escolha **Painel de Controle**, em seguida, **Manutenção do Sistema**. Dê um clique duplo em **Ferramentas Administrativas**, selecione **Origens de Dados (ODBC)** e, em seguida, clique em **Abrir**.
      - **Windows 7.** No menu Iniciar, escolha **Painel de Controle**, em seguida, **Sistema & Segurança**, em seguida, **Ferramentas Administrativas**. Selecione **Origens de Dados (ODBC)** e, em seguida, clique em **Abrir**.

   b. Vá até a aba **DSN do Sistema** e, em seguida, clique em **Adicionar**.
2. Selecione **NetezzaSQL** na lista e clique em **Concluir**.
3. Na guia **Opções do DSN** da tela Configuração do Driver ODBC Netezza, digite um nome de origem de dados de sua escolha, o nome do host ou o endereço IP do servidor IBM Netezza, o número da porta para a conexão, o banco de dados da instância do IBM Netezza que estiver usando e seus detalhes de nome do usuário e de senha para a conexão com o banco de dados. Clique no botão **Ajuda** para obter uma explicação dos campos.
4. Clique no botão **Testar Conexão** e verifique se é possível conectar-se ao banco de dados.
5. Quando tiver uma conexão bem-sucedida, clique em **OK** repetidamente para sair da tela Administrador da Origem de Dados ODBC.

Servidores Windows

O procedimento para o Servidor Windows é o mesmo que o procedimento do cliente para Windows XP.

Servidores UNIX ou Linux

O procedimento a seguir se aplica aos servidores UNIX ou Linux (exceto zLinux, para o qual os drivers ODBC do IBM Netezza não estão disponíveis).

1. A partir do seu Netezza Client CD/DVD, copie o arquivo `<platform>cli.package.tar.gz` relevante para um local temporário no servidor.
2. Extraia o conteúdo do archive por meio dos comandos **`gunzip`** e **`untar`**
3. Inclua permissões de execução no script *unpack* que é extraído.
4. Execute o script, respondendo os prompts na tela.
5. Edite o arquivo `modelersrv.sh` para incluir as seguintes linhas.

```
. <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 exemplo:

```
. /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. Localize o arquivo `/usr/local/nz/lib64/odbc.ini` e copie seu conteúdo para o arquivo `odbc.ini` que é instalado com o SDAP (aquele definido pela variável de ambiente $ODBCINI).

   Nota: para sistemas Linux de 64 bits, o parâmetro **`Driver`** referencia incorretamente o driver de 32 bits. Ao copiar o conteúdo do `odbc.ini` no passo anterior, edite o caminho dentro deste parâmetro de maneira adequada, por exemplo:

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

7. Edite os parâmetros na definição DSN do Netezza para refletir o banco de dados a ser utilizado.
8. Reinicie o ServidorIBM SPSS Modeler e teste os nós de mineração dentro da base de dados do Netezza no cliente.

Ativando a integração em ModeladorSPSS

1. No menu principal do IBM SPSS Modeler, escolha

   Ferramentas > **Opções** > **Aplicações de Helper.**

2. Clique na guia **IBM Data Warehouse** .

Habilitar o IBM Data Warehouse Analytics Integration. Possibilita a paleta de Modelagem de Banco de Dados (se não já foi exibida) na parte inferior da janela IBM SPSS Modeler e adiciona os nós para algoritmos de Mineração de Dados IBM e Netezza .

IBM Data Warehouse Connection. Clique no botão **Editar** e escolha a string de conexão do IBM Data Warehouse que você configurou ao criar a fonte ODBC . Para obter mais informações, consulte o console admin do IBM Data Warehouse.

Ativando geração e otimização de SQL

Devido à possibilidade de trabalhar com conjuntos de dados muito grandes, por motivos de desempenho, deve-se ativar as opções de geração e de otimização de SQL no IBM SPSS Modeler.

1. Nos menus do IBM SPSS Modeler, escolha:

   Ferramentas > **Propriedades do Fluxo** > **Opções**

2. Clique na opção **Otimização** na área de janela de navegação.
3. Confirme se a opção **Gerar SQL** está ativada. Essa configuração é necessária para que a modelagem da base de dados funcione.
4. Selecione **Otimizar Geração de SQL** e **Otimizar outra execução** (não é estritamente necessário, mas é altamente recomendado para um desempenho otimizado).

Construindo modelos com AnáliseIBM Netezza e IBM Data Warehouse

Cada um dos algoritmos suportados possui um nó de modelagem correspondente. É possível acessar os nós de modelagem IBM Data Warehouse e IBM Netezza na guia **Modelagem do Banco de Dados** na paleta de nós.

Considerações de dados

Os campos na origem de dados podem conter variáveis de diferentes tipos de dados, dependendo do nó de modelagem. No IBM SPSS Modeler, os tipos de dados são conhecidos como *níveis de medição*. A guia Campos do nó de modelagem utiliza ícones para indicar os tipos de nível de medição permitidos para seus campos de entrada e de destino.

Campo de destino O campo de destino é o campo cujo valor você está tentando prever. Quando um destino pode ser especificado, apenas um dos campos de dados de origem pode ser selecionado como o campo de destino.

Campo ID do registro Especifica o campo usado para identificar com exclusividade cada caso. Por exemplo, esse pode ser um campo de ID, como *CustomerID*. Se os dados de origem não incluírem um campo de ID, será possível criar este campo por meio de um nó Derivar, como mostra o procedimento a seguir.

1. Selecione o nó de origem.
2. Na guia Operações de Campo na paleta de nós, dê um clique duplo no nó Derivar.
3. Abra o nó Derivar ao dar um clique duplo em seu ícone na tela.
4. No campo **Derivar campo**, digite (por exemplo) `ID`.
5. No campo **Fórmula**, digite `@INDEX` e clique em **OK**.
6. Conecte o nó Derivar ao restante do fluxo.

Nota: Se você recuperar dados numéricos longos a partir de um banco de dados Netezza utilizando o tipo de dados `NUMERIC(18,0)`, o ModeladorSPSS poderá, às vezes, arredondar os dados durante a importação. Para evitar esse problema, armazene seus dados utilizando o tipo de dados `BIGINT` ou `NUMERIC(36,0)`.

Nota: Devido às limitações nos tipos de campos que podem ser usados, um campo com um nível de Medição de tipelidade e um papel de Registro ID não aparece em um nó de modelagem Netezza In-Database (por exemplo, K-Means).

Manipulando valores nulos

Se os dados de entrada contiverem valores nulos, o uso de alguns dos nós do Netezza poderá resultar em mensagens de erro ou fluxos de longa execução, portanto, é recomendado remover registros que contiverem valores nulos. Use o método a seguir.

1. Anexe um nó Seleção ao nó de origem.
2. Configure a opção **Modo** do nó Seleção para **Descartar**.
3. Insira o seguinte no campo **Condição**:

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

   Assegure-se de incluir cada campo de entrada.
4. Conecte o nó Seleção ao restante do fluxo.

Saída de modelo

É possível que um fluxo contendo um nó de modelagem de Data Warehouse ou Netezza produza resultados um pouco diferentes cada vez que for executado. Isso ocorre porque a ordem na qual o nó lê os dados de origem nem sempre é a mesma, já que os dados são lidos em tabelas temporárias antes da construção de modelo. No entanto, as diferenças produzidas por este efeito são insignificantes.

Comentários gerais

- No Serviços de Colaboração e ImplementaçãoIBM SPSS, não é possível criar configurações de pontuação usando fluxos que contêm IBM Data Warehouse ou IBM Netezza nós de modelagem de banco de dados
- Exportação ou importação PMML não é possível para modelos criados pelos nós do Data Warehouse ou Netezza .

Opções de Campo

Na guia Campos, você escolhe se deseja utilizar as configurações de papel do campo já definidas em nós de envio de dados ou fazer as designações de campo manualmente.

Use funções predefinidas. Essa opção utiliza as configurações de papel (destinos, preditores, e assim por diante) a partir de um nó Tipo de envio de dados (ou na guia Tipos de um nó de origem de envio de dados).

Utilize designações de campo customizadas. Escolha esta opção se desejar designar destinos, preditores e outros papéis manualmente nessa tela.

Campos. Utilize os botões de seta para designar itens manualmente a partir desta lista para os vários campos de papel à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo de papel.

Clique no botão **Tudo** para selecionar todos os campos na lista ou clique em um botão de nível de medição individual para selecionar todos os campos com esse nível de medição.

Destino. Escolha um campo como o destino para a predição. Para modelos Lineares Generalizados, consulte também o campo **Avaliações** nessa tela.

ID do registro. O campo a ser utilizado como o identificador de registro exclusivo.

Preditores (Inputs). Escolha um ou mais campos como entradas para a predição.

Opções do Servidor

Na guia do Servidor, você especifica o banco de dados do IBM Data Warehouse onde o modelo deve ser construído.

IBM Data Warehouse Server Details. Aqui você especifica os detalhes da conexão para o banco de dados que você deseja utilizar para o modelo.

- **Utilizar conexão de envio de dados**. (padrão) Usa os detalhes de conexão especificados em um nó de envio de dados, por exemplo, o nó de origem do Banco de Dados. Esta opção funciona apenas se todos os nós de upstream forem capazes de usar o pushback SQL. Neste caso, não há necessidade de mover os dados para fora do banco de dados, já que a SQL implementa completamente todos os nós de envio de dados.
- **Mover dados para conexão**. Move os dados para o banco de dados que você especificar aqui. Fazer isso permite que a modelagem funcione se os dados estão em outro banco de dados do IBM Data Warehouse, ou um banco de dados de outro fornecedor, ou mesmo se os dados estão em um arquivo flat. Além disso, os dados serão movidos de volta para o banco de dados especificado aqui se os dados tiverem sido extraídos porque um nó não executou SQL pushback. Clique no botão **Editar** para procurar e selecionar uma conexão.

Cuidado: AnáliseIBM Netezza e o IBM Data Warehouse é geralmente usado com conjuntos de dados muito grandes. A transferência de grandes quantias de dados entre os bancos de dados, fora do banco de dados ou de volta a ele pode ser muito demorada e deve ser evitada sempre que possível.

Nota: O nome da origem de dados ODBC é efetivamente integrado em cada fluxo do IBM SPSS Modeler. Se um fluxo que é criado em um host for executado em um host diferente, o nome da origem de dados deverá ser o mesmo em cada host. Como alternativa, uma origem de dados diferente pode ser selecionada na guia Servidor em cada nó de origem ou de modelagem.

Opções de modelo

Na guia Opções do modelo, é possível escolher se deseja especificar um nome para o modelo ou gerar um nome automaticamente. Também é possível configurar valores padrão para opções de escoragem.

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Substituir existente se o nome tiver sido usado. Se marcar essa caixa de seleção, qualquer modelo existente com o mesmo nome será sobrescrito.

Disponibilizar para Escoragem. É possível configurar os valores padrão aqui para as opções de escoragem que aparecem no diálogo para o nugget do modelo. Para obter detalhes das opções, consulte o tópico de ajuda para a guia Configurações desse nugget específico.

Gerenciando modelos

Construir um modelo do IBM Netezza ou IBM Data Warehouse via ModeladorSPSS cria um modelo no ModeladorSPSS e cria ou substitui um modelo no banco de dados do IBM Data Warehouse. O modelo do ModeladorSPSS desse tipo referencia o conteúdo de um modelo de banco de dados armazenado em um servidor de base de dados. O ModeladorSPSS pode executar a verificação de consistência armazenando uma sequência de chaves de modelo gerada idêntica no modelo ModeladorSPSS e no modelo Netezza ou Data Warehouse.

O nome do modelo para cada modelo Netezza ou Data Warehouse é exibido sob a coluna *Modelo de Informações* na caixa de diálogo Listar Modelos de Banco de Dados. O nome do modelo para um modelo do ModeladorSPSS é exibido como uma Chave de Modelo na guia Servidor de um modelo do ModeladorSPSS (quando colocado em um fluxo).

O botão Check pode ser usado para verificar se as chaves do modelo no modelo ModeladorSPSS e no modelo Netezza ou Data Warehouse combinam. Se nenhum modelo de mesmo nome puder ser encontrado em Netezza ou Data Warehouse, ou se as teclas do modelo não combinam, o modelo Netezza ou Data Warehouse foi excluído ou reconstruído desde que o modelo ModeladorSPSS foi construído.

Listando Modelos de Banco de Dados

ModeladorSPSS fornece uma caixa de diálogo para listar os modelos que são armazenados no IBM Data Warehouse e possibilita que os modelos sejam excluídos. Essa caixa de diálogo é acessível a partir da caixa de diálogo IBM Helper Applications e a partir das caixas de diálogo de construção, navegação e aplicação para IBM Data Warehouse e IBM Netezza nós relacionados à mineração de dados. As informações a seguir são exibidas para cada modelo:

- O nome do modelo (nome do modelo, que é utilizado para ordenar a lista).
- Nome do proprietário.
- O algoritmo utilizado no modelo.
- O estado atual do modelo, por exemplo, Concluído.
- A data na qual o modelo foi criado.

Árvore de Regressão do IBM Data WH

Uma árvore de regressão é um algoritmo baseado em árvore que divide uma amostra de casos repetidamente para derivar subconjuntos do mesmo tipo, com base nos valores de um campo de destino numérico. Assim como as árvores de decisão, as árvores de regressão decompõem os dados em subconjuntos nos quais as folhas da árvore correspondem a subconjuntos suficientemente pequenos ou suficientemente uniformes. As divisões são selecionadas para reduzir a dispersão dos valores de atributo de destino, de modo que eles possam ser razoavelmente bem preditos pelos seus valores médios nas folhas.

IBM Data WH Regression Tree Build Options-Tree Growth

É possível configurar as opções de construção para o crescimento e poda da árvore.

As opções de construção a seguir estão disponíveis para crescimento da árvore:

Profundidade máxima da árvore. O número máximo de níveis até o qual a árvore pode crescer abaixo do nó raiz, ou seja, o número de vezes em que a amostra é dividida recursivamente. O padrão é 62, que é a profundidade máxima da árvore para propósitos de modelagem.

Nota: Se o visualizador no nugget do modelo mostrar a representação textual do modelo, um máximo de 12 níveis da árvore será exibido.

Critérios de Divisão. Estas opções controlam quando parar a divisão da árvore. Se não desejar utilizar os valores padrão, clique em **Customizar** e altere os valores.

- **Medida de avaliação de divisão.** Essa medida de avaliação de classe avalia o melhor local para dividir a árvore.

  Nota: Atualmente, a variância é a única opção possível.
- **Melhoria mínima para divisões.** O valor mínimo pelo qual a impureza deve ser reduzida antes de uma nova divisão ser criada na árvore. O objetivo da construção de árvore é criar subgrupos com valores de saída semelhantes para minimizar a impureza dentro de cada nó. Se a melhor divisão de uma ramificação reduzir a impureza em um nível menor que a quantia especificada pelo critério de divisão, a ramificação não será dividida.
- **Número mínimo de instâncias para uma divisão.** O número mínimo de registros que podem ser divididos. Quando uma quantia menor que esse número de registros não divididos permanece, nenhuma divisão adicional é feita. É possível utilizar esse campo para evitar a criação de subgrupos pequenos na árvore.

Estatísticas. Esse parâmetro define quantas estatísticas são incluídas no modelo. Selecione uma das opções a seguir:

- **Todos.** Todas as estatísticas relacionadas a coluna e todas as estatísticas relacionadas a valor são incluídas.

  Nota: Esse parâmetro inclui o número máximo de estatísticas e pode, portanto, afetar o desempenho do seu sistema. Se não desejar visualizar o modelo no formato gráfico, especifique **Nenhum**.
- **Colunas.** Estatísticas relacionadas a coluna são incluídas.
- **Nenhum.** Apenas estatísticas que são necessárias para escorar o modelo são incluídas.

IBM Data WH Tree Build Options-Tree Pruning

É possível usar as opções de poda para especificar os critérios de poda para a árvore de regressão. A intenção da poda é reduzir o risco de super ajuste ao remover subgrupos crescidos demasiadamente que não melhoram a precisão esperada nos novos dados.

Medida de poda. A medida de poda assegura que a precisão estimada do modelo permaneça dentro dos limites aceitáveis após remover uma folha da árvore. É possível selecionar uma das medidas a seguir.

- **mse.** Erro quadrático médio (padrão) - mede o quão próxima uma linha ajustada está dos pontos de dados.
- **r2.** R-quadrado - mede a proporção de variação na variável dependente explicada pelo modelo de regressão.
- **Pearson.** coeficiente de correlação de Pearson - mede a intensidade do relacionamento entre as variáveis linearmente dependentes que são normalmente distribuídas.
- **Spearman.** coeficiente de correlação de Spearman - detecta relacionamentos não lineares que parecem fracas de acordo com a correlação Pearson, mas que podem realmente ser fortes.

Dados para poda. É possível usar alguns ou todos os dados de treinamento para estimar a precisão esperada nos novos dados. Como alternativa, é possível usar um conjunto de dados de poda separado de uma tabela especificada para esse propósito.

- **Usar todos os dados de treinamento**. Essa opção (a padrão) usa todos os dados de treinamento para estimar a precisão do modelo.
- **Usar % de dados de treinamento para poda**. Use esta opção para dividir os dados em dois conjuntos, um para treinamento e outro para poda, utilizando a porcentagem especificada aqui para a poda de dados.

  Selecione **Replicar resultados** se quiser especificar uma semente aleatória para assegurar que os dados sejam particionados da mesma maneira toda vez que executar o fluxo. É possível especificar um número inteiro no campo **Valor semente usado para poda** ou clicar em **Gerar**, que criará um pseudonúmero inteiro aleatório.

- **Usar dados de uma tabela existente.** Especifique o nome da tabela de um conjunto de dados de poda separado para estimar a precisão do modelo. Fazer isso é considerado mais confiável do que utilizar dados de treinamento. No entanto, essa opção poderá resultar na remoção de um subconjunto de dados grande do conjunto de treinamento, reduzindo, assim, a qualidade da árvore de decisão.

Cluster de divisão Netezza

O armazenamento em cluster de divisão é um método de análise de cluster em que o algoritmo é executado repetidamente para dividir os clusters em subclusters até que um ponto de parada especificado seja atingido.

A formação do cluster começa com um único cluster contendo todas as instâncias de treinamento (registros). A primeira iteração do algoritmo divide o conjunto de dados em dois subclusters, com iterações subsequentes dividindo-os em mais subclusters. Os critérios de parada são especificados como o número máximo de iterações, como um número máximo de níveis para os quais o conjunto de dados é dividido e como um número mínimo necessário de instâncias para particionamento adicional.

A árvore de armazenamento em cluster hierárquico resultante pode ser utilizada para classificar instâncias ao propagá-las para baixo do cluster raiz, como no exemplo a seguir.

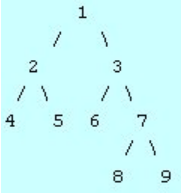

Figura 3. Exemplo de uma árvore de armazenamento em cluster de divisão

Em cada nível, o melhor subcluster correspondente é escolhido com relação à distância da instância dos centros de subcluster.

Quando as instâncias são escoradas com um nível de hierarquia aplicado de -1 (o padrão), a escoragem retorna apenas um cluster de folha, já que as folhas são designadas por um número negativo. No exemplo, este seria um dos clusters 4, 5, 6, 8 ou 9. No entanto, se o nível de hierarquia for configurado para 2, por exemplo, a pontuação retornaria um dos clusters no segundo nível abaixo do cluster raiz, a saber: 4, 5, 6 ou 7.

Opções do Campo de Armazenamento em Cluster de Divisão Netezza

Na guia Campos, você escolhe se deseja utilizar as configurações de papel do campo já definidas em nós de envio de dados ou fazer as designações de campo manualmente.

Usar funções predefinidas. Essa opção utiliza as configurações de papel (destinos, preditores, e assim por diante) a partir de um nó Tipo de envio de dados (ou na guia Tipos de um nó de origem de envio de dados).

Usar designações de campo customizado. Escolha esta opção se desejar designar destinos, preditores e outros papéis manualmente nessa tela.

Campos. Utilize os botões de seta para designar itens manualmente a partir desta lista para os vários campos de papel à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo de papel.

Clique no botão **Tudo** para selecionar todos os campos na lista ou clique em um botão de nível de medição individual para selecionar todos os campos com esse nível de medição.

ID do registro. O campo a ser utilizado como o identificador de registro exclusivo.

Preditores (Entradas). Escolha um ou mais campos como entradas para a predição.

Opções de Construção de Armazenamento em Cluster de Divisão Netezza

A guia Opções de Criação é onde você configura todas as opções para construir o modelo. Obviamente, é possível clicar somente no botão **Executar** para construir um modelo com todas as opções padrão, no entanto, você normalmente deseja customizar a construção para seus próprios propósitos.

Medida de distância. O método a ser utilizado para medir a distância entre pontos de dados, em que distâncias maiores indicam dissimilaridades maiores. As opções são:

- **Euclidean**. (padrão) A distância entre dois pontos é calculada pela união deles com uma linha reta.
- **Manhattan**. A distância entre dois itens é calculada como a soma das diferenças absolutas entre suas coordenadas.
- **Canberra**. Semelhante à distância de Manhattan, porém mais sensível a pontos de dados mais próximos da origem.
- **Máximo**. A distância entre dois pontos é calculada como a maior de suas diferenças ao longo de qualquer dimensão de coordenada.

Número máximo de iterações. O algoritmo opera ao executar várias iterações do mesmo processo. Esta opção permite parar o treinamento do modelo após o número de iterações especificado.

Profundidade máxima das árvores de cluster. O número máximo de níveis para os quais o conjunto de dados pode ser subdividido.

Replicar resultados. Marque esta caixa se desejar configurar uma semente aleatória, que permitirá replicar as análises. É possível especificar um número inteiro ou clicar em **Gerar**, que cria um pseudonúmero inteiro aleatório.

Número mínimo de instâncias para uma divisão. O número mínimo de registros que podem ser divididos. Quando uma quantia menor que esse número de registros não divididos permanece, nenhuma divisão adicional é feita. É possível utilizar esse campo para evitar a criação de subgrupos muito pequenos na árvore de cluster.

IBM Data WH Generalized Linear

A regressão linear é uma técnica estatística consagrada para classificar registros com base nos valores de campos de entrada numéricos. A regressão linear se ajusta a uma linha reta ou superfície que minimiza as discrepâncias entre os valores de saída preditos e reais. Os modelos lineares são úteis para modelar uma ampla variedade de fenômenos do mundo real devido a sua simplicidade no treinamento e na aplicação do modelo. No entanto, os modelos lineares supõem uma distribuição normal na variável dependente (resposta) e um impacto linear das variáveis independentes (preditoras) na variável dependente.

Há muitas situações em que uma regressão linear é útil, mas as suposições acima não se aplicam. Por exemplo, ao modelar a escolha do consumidor entre um número discreto de produtos, a variável dependente provavelmente terá uma distribuição multinomial. Da mesma forma, ao modelar a renda com relação à idade, a renda geralmente aumenta à medida que a idade aumenta, mas é improvável que a ligação entre os dois seja tão simples quanto uma linha reta.

Para essas situações, um modelo linear generalizado pode ser utilizado. Os modelos lineares generalizados expandem o modelo de regressão linear para que a variável dependente esteja relacionada às variáveis preditoras por meio de uma função de ligação especificada, para a qual há uma opção de funções adequadas. Além disso, o modelo permite que a variável dependente tenha um distribuição não normal, como Poisson.

O algoritmo busca iterativamente o modelo de melhor ajuste, até um número especificado de iterações. Para o cálculo do melhor ajuste, o erro é representado pela soma dos quadrados das diferenças entre o valor predito e real da variável dependente.

IBM Data WH Generalized Linear Model Field Options

Na guia Campos, você escolhe se deseja utilizar as configurações de papel do campo que já estiverem definidas em nós de envio de dados ou fazer as designações de campo manualmente.

Usar funções predefinidas. Esta opção utiliza as configurações de papel, como destinos ou preditores, a partir de um nó Tipo de envio de dados ou da guia Tipos de um nó de origem de envio de dados.

Usar designações de campo customizado. Escolha esta opção se desejar designar destinos, preditores e outros papéis manualmente nessa tela.

Campos. Utilize os botões de seta para designar itens manualmente a partir desta lista para os vários campos de papel à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo de papel.

Clique no botão **Tudo** para selecionar todos os campos na lista ou clique em um botão de nível de medição individual para selecionar todos os campos com esse nível de medição.

Destino. Escolha um campo como o destino para a predição.

ID do registro. O campo a ser utilizado como o identificador de registro exclusivo. Os valores deste campo devem ser exclusivos para cada registro, por exemplo, números de ID do cliente.

Ponderação de Instância. Especifique um campo para utilizar ponderações de instância. Uma ponderação de instância é uma ponderação por linha de dados de entrada. Por padrão, supõe-se que todos os registros de entrada tenham uma importância relativa igual. É possível alterar a importância ao designar ponderações individuais para os registros de entrada. O campo que você especificar deve conter uma ponderação numérica para cada linha de dados de entrada.

Preditores (Entradas). Selecione o campo ou os campos de entrada. Essa ação é semelhante a configurar o papel do campo para *Entrada* em um nó Tipo.

IBM Data WH Generalized Linear Model Options-Geral

Na guia Opções do modelo, é possível escolher se deseja especificar um nome para o modelo ou gerar um nome automaticamente. Também é possível fazer várias configurações referentes ao modelo, à função de ligação, às interações do campo de entrada (se houver alguma), bem como configurar valores padrão para as opções de escoragem.

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Opções de campo. É possível especificar as funções dos campos de entrada para construir o modelo.

Configurações Gerais. Essas configurações estão relacionadas aos critérios de parada para o algoritmo.

- **Número máximo de iterações.** Número máximo de iterações que o algoritmo executará; o mínimo é 1 e o padrão é 20.
- **Erro máximo (1e).** O valor máximo de erro (em notação científica) no qual o algoritmo deve parar de localizar o melhor modelo de ajuste. O mínimo é 0 e o padrão é -3, significando 1E-3 ou 0,001.
- **Limite de valores de erro insignificantes (1e).** O valor (em notação científica) abaixo do qual os erros são tratados como tendo um valor de zero. O mínimo é -1 e o padrão é -7, significando que os valores de erro abaixo de 1E-7 (ou 0,0000001) são contados como insignificantes.

Configurações de Distribuição. Essas configurações estão relacionadas à distribuição da variável dependente (resposta).

- **Distribuição de variável de resposta.** O tipo de distribuição, que é um de **Bernoulli** (padrão), **Gaussiana**, **Poisson**, **Binomial**, **Binomial Negativa**, **Wald** (Gaussiana Inversa) e **Gama**.
- **Parâmetros.** (Poisson ou distribuição binomial somente) Deve-se especificar uma das seguintes opções no campo **Especificar parâmetro**:
  - Para que o parâmetro seja estimado automaticamente a partir dos dados, selecione **Padrão**.
  - Para permitir otimização da distribuição quasi-verossimilhança, selecione **Quasi**.
  - Para especificar explicitamente o valor do parâmetro, selecione **Explícito**.

(Distribuição Binomial somente) Deve-se especificar a coluna da tabela de entrada que deve ser utilizada como o campo de avaliações, conforme necessário pela distribuição binomial. Esta coluna contém o número de avaliações para a distribuição binomial.

(Distribuição binomial negativa somente) É possível utilizar o padrão de -1 ou especificar um valor de parâmetro diferente.

Configurações da Função de Ligação. Configurações referentes à função de ligação, que relaciona a variável dependente às variáveis preditoras.

- **Função de ligação.** A função a ser usada, que é uma de **Identity**, **Inverse**, **Invnegative**, **Invsquare**, **Sqrt**, **Power**, **Oddspower**, **Log**, **Clog**, **Loglog**, **Cloglog**, **Logit** (padrão), **Probit**, **Gaussit**, **Cauchit**, **Canbinom**, **Cangeom**, **Cannegbinom**.
- **Parâmetros.** (Funções de ligação Power ou Oddspower somente) Um valor de parâmetro poderá ser especificado se a função de ligação for **Power** ou **Oddspower**. Escolha para especificar um valor ou utilize o padrão de 1.

IBM Data WH Generalized Model Options-Interação

O painel Interação contém as opções para especificar interações (ou seja, efeitos multiplicadores entre os campos de entrada).

Interação da Coluna. Marque essa caixa de seleção para especificar interações entre os campos de entrada. Deixe a caixa desmarcada se não houver interações.

Insira as interações no modelo ao selecionar um ou mais campos na lista de origem e arrastá-los para a lista de interações. O tipo de interação criado depende do ponto de acesso no qual você soltar a seleção.

- **Principal.** Os campos eliminados aparecem como interações principais separadas na parte inferior da lista de efeitos.
- **bidirecional.** Todos os pares possíveis dos campos eliminados aparecem como interações de duas vias na parte inferior da lista de interações.
- **de três direções.** Todos os trios possíveis dos campos eliminados aparecem como interações de três vias na parte inferior da lista de interações.
- ***.** A combinação de todos os campos eliminados aparece como uma única interação na parte inferior da lista de interações.

Incluir Intercepto. O intercepto é geralmente incluído no modelo. Se você conseguir presumir as passagens de dados por meio da origem, será possível excluir o intercepto.

Botões da caixa de diálogo

Os botões à direita da exibição permitem fazer mudanças nos termos usados no modelo.

Figura 4. botão Excluir

Excluir termos do modelo ao selecionar os termos que você deseja excluir e clicando no botão Excluir.

Figura 5. Botões Reordenar

Reordene os termos dentro do modelo ao selecionar os termos que você deseja reordenar e clicando na seta para cima ou para baixo.

Figura 6. Botão de interação customizado

Incluir um termo customizado

É possível especificar interações customizadas no formato *n1*x1*x1*x1.*. Selecione um campo na lista **Campos**, clique no botão de seta para a direita para incluir o campo no **Termo Customizado**, clique em **Por***, selecione o próximo campo, clique no botão de seta para a direita, e assim por diante. Quando tiver construído a interação customizada, clique em **Incluir termo** para retornã-lo ao painel de Interação.

IBM Data WH Generalized Linear Model Opções-Opções de Scoring

Disponibilizar para Escoragem. É possível configurar os valores padrão aqui para as opções de escoragem que aparecem no diálogo para o nugget do modelo. Veja o tópico "IBM Data WH Generalized Linear Modelo Nugget-Guia de Configurações" na página 83 para obter mais informações.

- **Incluir campos de entrada.** Marque esta caixa de seleção se desejar exibir os campos de entrada na saída do modelo, bem como as predições.

IBM

Uma árvore de decisão é uma estrutura hierárquica que representa um modelo de classificação. Com um modelo de árvore de decisão, é possível desenvolver um sistema de classificação para prever ou classificar futuras observações a partir de um conjunto de dados de treinamento. A classificação toma a forma de uma estrutura em árvore na qual as ramificações representam pontos de divisão na classificação. As divisões dividem os dados em subgrupos recursivamente até que um ponto de interrupção seja atingido. Os nós da árvore nos pontos de parada são conhecidos como **folhas**. Cada folha designa um rótulo, conhecido como um **rótulo de classe**, aos membros de seu subgrupo ou classe.

Ponderações de Instância e Ponderações de Classe

Por padrão, supõe-se que todos os registros de entrada e as classes tenham uma importância relativa igual. É possível alterar isso ao designar ponderações individuais para os membros de um ou para os dois itens. Isso pode ser útil, por exemplo, se os pontos de dados em seus dados de treinamento não estiverem realisticamente distribuídos entre as categorias. As ponderações permitem causar viés no modelo para poder compensar as categorias que forem menos bem representadas nos dados. O aumento da ponderação de um valor de destino deve aumentar a porcentagem de predições corretas para essa categoria.

No nó de modelagem de Árvore de Decisão, é possível especificar dois tipos de ponderações. **Ponderações de Instância** designa uma ponderação para cada linha de dados de entrada. As ponderações são geralmente especificadas como 1,0 para a maioria dos casos, com valores mais altos ou mais baixos fornecido apenas para os casos que forem mais ou menos importantes do que a maioria, conforme mostrado na tabela a seguir.

Tabela 5. Exemplo de ponderação de instância

ID de Registro	Destino	Ponderação da instância
1	drugA	1.1
2	drugB	1.0
3	drugA	1.0
4	drugB	0.3

As **Ponderações de classe** designam uma ponderação para cada categoria do campo de destino, conforme mostrado na tabela a seguir.

Tabela 6. Exemplo de ponderação de classe

Classe	**Ponderação de Classe**
drugA	1.0
drugB	1.5

Ambos os tipos de ponderações podem ser utilizados ao mesmo tempo, caso em que eles são multiplicados e utilizados como ponderações de instância. Assim, se os dois exemplos anteriores forem utilizados juntos, o algoritmo poderá utilizar as ponderações de instância conforme mostrado na tabela a seguir.

Tabela 7. Exemplo do cálculo de ponderação de instância

ID de Registro	**Cálculo**	**Ponderação da instância**
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

Opções do Campo de Árvore de Decisão Netezza

Na guia Campos, você escolhe se deseja utilizar as configurações de papel do campo já definidas em nós de envio de dados ou fazer as designações de campo manualmente.

Usar papéis predefinidos Esta opção utiliza as configurações de papel (destinos, preditores e assim por diante) a partir de um nó Tipo de envio de dados (ou na guia Tipos de um nó de origem de envio de dados).

Usar designações de campo customizado. Para designar manualmente destinos, preditores e outros papéis, selecione esta opção.

Campos. Utilize os botões de seta para designar itens manualmente a partir desta lista para os vários campos de papel à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo de papel.

Para selecionar todos os campos na lista, clique no botão **Tudo**, ou clique em um botão de nível de medição individual para selecionar todos os campos com esse nível de medição.

Destino. Selecione um campo como o destino para a predição.

ID do registro. O campo a ser utilizado como o identificador de registro exclusivo. Os valores deste campo devem ser exclusivos para cada registro (por exemplo, números de ID do cliente).

Ponderação de Instância. Especificar um campo aqui permite usar ponderações de instância (uma ponderação por linha de dados de entrada) ao invés ou além das ponderações de classe padrão (uma ponderação por categoria para o campo de destino). O campo que você especificar aqui deve ser um que contenha uma ponderação numérica para cada linha dos dados de entrada. Veja o tópico “Ponderações de Instância e Ponderações de Classe” na página 59 para obter mais informações.

Preditores (Entradas). Selecione o campo ou os campos de entrada. Isso é semelhante a configurar o papel do campo como *Entrada* em um nó Tipo.

Opções de Construção da árvore de decisão do IBM Data WH

As opções de construção a seguir estão disponíveis para crescimento da árvore:

Medida de Crescimento. Essas opções controlam a maneira como o crescimento da árvore é medido.

- **Medida de Impureza.** Esta medida avalia o melhor local para dividir a árvore. É uma medição da variabilidade em um subgrupo ou segmento de dados. Uma medição de impureza baixa indica um grupo no qual a maioria dos membros possui valores semelhantes para o campo de critério ou de destino.

  As medições suportadas são **Entropia** e **Gini**. Essas medições baseiam-se em probabilidades da associação de categoria para a ramificação.
- **Profundidade máxima da árvore.** O número máximo de níveis até o qual a árvore pode crescer abaixo do nó raiz, ou seja, o número de vezes em que a amostra é dividida recursivamente. O valor-padrão desta propriedade é 10, e o valor máximo que pode ser configurado para essa propriedade é 62.

  Nota: Se o visualizador no nugget do modelo mostrar a representação textual do modelo, um máximo de 12 níveis da árvore será exibido.

Critérios de Divisão. Estas opções controlam quando parar a divisão da árvore.

- **Melhoria mínima para divisões.** O valor mínimo pelo qual a impureza deve ser reduzida antes de uma nova divisão ser criada na árvore. O objetivo da construção de árvore é criar subgrupos com valores de saída semelhantes para minimizar a impureza dentro de cada nó. Se a melhor divisão de uma ramificação reduzir a impureza em um nível menor que a quantia especificada pelo critério de divisão, a ramificação não será dividida.
- **Número mínimo de instâncias para uma divisão.** O número mínimo de registros que podem ser divididos. Quando uma quantia menor que esse número de registros não divididos permanece, nenhuma divisão adicional é feita. É possível utilizar esse campo para evitar a criação de subgrupos pequenos na árvore.

Estatísticas. Esse parâmetro define quantas estatísticas são incluídas no modelo. Selecione uma das opções a seguir:

- **Todos.** Todas as estatísticas relacionadas a coluna e todas as estatísticas relacionadas a valor são incluídas.

  Nota: Esse parâmetro inclui o número máximo de estatísticas e pode, portanto, afetar o desempenho do seu sistema. Se não desejar visualizar o modelo no formato gráfico, especifique **Nenhum**.
- **Colunas.** Estatísticas relacionadas a coluna são incluídas.
- **Nenhum.** Apenas estatísticas que são necessárias para escorar o modelo são incluídas.

IBM Data WH Decision Tree Node-Classe Weights

Aqui é possível designar ponderações para classes individuais. O padrão é designar um valor de 1 para todas as classes, tornando-as igualmente ponderadas. Ao especificar diferentes ponderações numéricas para rótulos de classe diferentes, você instrui o algoritmo a ponderar os conjuntos de treinamento de classes específicas de modo apropriado.

Para alterar uma ponderação, dê um clique duplo nela na coluna **Ponderação** e faça as mudanças desejadas.

Valor. O conjunto de rótulos de classe derivados dos valores possíveis do campo de destino.

Peso. A ponderação a ser designada a uma classe específica. Designar uma ponderação maior para uma classe torna o modelo mais sensível a essa classe com relação a outras classes.

As ponderações de classe podem ser usadas em combinação com as ponderações de instância. Consulte o tópico “Ponderações de Instância e Ponderações de Classe” na página 59 para obter informações adicionais.

IBM Data WH Decision Tree Node-Tree Pruning

É possível usar as opções de poda para especificar os critérios de poda para a árvore de decisão. A intenção da poda é reduzir o risco de super ajuste ao remover subgrupos crescidos demasiadamente que não melhoram a precisão esperada nos novos dados.

Medida de poda. A medida de poda padrão, **Precisão**, assegura que a precisão estimada do modelo permaneça dentro dos limites aceitáveis após remover uma folha da árvore. Utilize a alternativa, **Precisão Ponderada**, se desejar levar em conta as ponderações de classe ao aplicar a poda.

Dados para poda. É possível usar alguns ou todos os dados de treinamento para estimar a precisão esperada nos novos dados. Como alternativa, é possível usar um conjunto de dados de poda separado de uma tabela especificada para esse propósito.

- **Usar todos os dados de treinamento**. Essa opção (a padrão) usa todos os dados de treinamento para estimar a precisão do modelo.
- **Usar % de dados de treinamento para poda**. Use esta opção para dividir os dados em dois conjuntos, um para treinamento e outro para poda, utilizando a porcentagem especificada aqui para a poda de dados.

  Selecione **Replicar resultados** se quiser especificar uma semente aleatória para assegurar que os dados sejam particionados da mesma maneira toda vez que executar o fluxo. É possível especificar um número inteiro no campo **Valor semente usado para poda** ou clicar em **Gerar**, que criará um pseudonúmero inteiro aleatório.
- **Usar dados de uma tabela existente.** Especifique o nome da tabela de um conjunto de dados de poda separado para estimar a precisão do modelo. Fazer isso é considerado mais confiável do que utilizar dados de treinamento. No entanto, essa opção poderá resultar na remoção de um subconjunto de dados grande do conjunto de treinamento, reduzindo, assim, a qualidade da árvore de decisão.

IBM Data WH Regressão Linear

Os modelos lineares preveem uma variável resposta contínua com base em relacionamentos lineares entre o destino e um ou mais preditores. Embora limitados a modelar diretamente apenas relacionamentos lineares, os modelos de regressão linear são relativamente simples e fornecem uma fórmula matemática fácil de interpretar para escoragem. Os modelos lineares são rápidos, eficientes e fáceis de usar, embora sua aplicabilidade seja limitada em comparação com aqueles produzidos por algoritmos de regressão mais refinados.

IBM Data WH Linear Regression Build Options

A guia Opções de Criação é onde você configura todas as opções para construir o modelo. Obviamente, é possível clicar somente no botão **Executar** para construir um modelo com todas as opções padrão, no entanto, você normalmente deseja customizar a construção para seus próprios propósitos.

Usar Decomposição em Valores Singulares para resolver equações. A vantagem de usar a matriz de Decomposição em Valores Singulares ao invés da matriz original é que ela é mais robusta com relação a erros numéricos e também pode acelerar o cálculo.

Incluir intercepto no modelo. Incluir o intercepto aumenta a precisão geral da solução.

Calcular diagnósticos de modelo. Essa opção faz com que um número de diagnósticos seja calculado no modelo. Os resultados são armazenados em matrizes ou tabelas para revisão posterior. Os diagnósticos incluem r-quadrado, soma dos quadrados residual, estimativa da variância, desvio padrão, valor *p* e valor *t*.

Esses diagnósticos referem-se à validade e à utilidade do modelo. Deve-se executar diagnósticos separados nos dados subjacentes para assegurar que eles atendam às suposições de linearidade.

IBM Data WH KNN

Análise do Vizinho mais Próximo é um método de classificação de casos com base na sua similaridade com outros casos. Em aprendizado por máquina, ela foi desenvolvida como uma maneira de reconhecer padrões de dados sem requerer uma correspondência exata com nenhum dos padrões ou casos armazenados. Casos semelhantes ficam próximos uns dos outros e os casos diferentes ficam distantes uns dos outros. Portanto, a distância entre dois casos é uma medida de sua dissimilaridade.

Casos próximos são chamados de "vizinhos". Quando um novo caso (validação) é apresentado, sua distância de cada um dos casos no modelo é calculada. As classificações dos casos mais similares – os vizinhos mais próximos – são verificadas e o novo caso é colocado na categoria que contiver o maior número de vizinhos mais próximos.

É possível especificar o número de vizinhos mais próximos a serem examinados; este valor é denominado *k*. As fotos mostram como um novo caso seria classificado usando dois valores diferentes de *k*. Quando *k* = 5, o novo caso é colocado na categoria *1* porque a maioria dos vizinhos mais próximos pertence à categoria *1*. No entanto, quando *k* = 9, o novo caso é colocado na categoria *0* porque uma maioria dos vizinhos mais próximos pertence à categoria *0*.

A análise do vizinho mais próximo também pode ser utilizada para calcular valores para uma variável resposta contínua. Nesta situação, a média ou mediana do valor dos vizinhos mais próximos é utilizada para obter o valor predito para o novo caso.

Opções de Modelo KNN IBM Data WH KNN

Na guia Opções do Modelo - Geral, é possível escolher se deseja especificar um nome para o modelo ou gerar um nome automaticamente. Também é possível configurar opções que controlam como o número de vizinhos mais próximos é calculado, além de configurar opções para melhor desempenho e precisão do modelo.

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Vizinhos

Medida de distância. O método a ser utilizado para medir a distância entre pontos de dados, em que distâncias maiores indicam dissimilaridades maiores. As opções são:

- **Euclidean**. (padrão) A distância entre dois pontos é calculada pela união deles com uma linha reta.
- **Manhattan**. A distância entre dois itens é calculada como a soma das diferenças absolutas entre suas coordenadas.
- **Canberra**. Semelhante à distância de Manhattan, porém mais sensível a pontos de dados mais próximos da origem.
- **Máximo**. A distância entre dois pontos é calculada como a maior de suas diferenças ao longo de qualquer dimensão de coordenada.

Número de Vizinhos Mais Próximos (k). O número de vizinhos mais próximos para um caso específico. Observe que utilizar um número maior de vizinhos não resultará necessariamente em um modelo mais preciso.

A opção de *k* controla o balanceamento entre a prevenção de super ajuste (isso pode ser importante, principalmente para dados "ruidosos") e a resolução (produzindo predições diferentes para instâncias semelhantes). Geralmente é necessário ajustar o valor de *k* para cada conjunto de dados, com valores típicos variando de 1 a várias dezenas.

Aprimorar desempenho e exatidão

Padronizar medições antes de calcular a distância. Se selecionada, essa opção padroniza as medições para campos de entrada contínuos antes de calcular os valores de distância.

Use conjuntos principais para aumentar o desempenho de grandes conjuntos de dados. Se selecionada, essa opção utiliza amostragem de conjunto principal para acelerar o cálculo quando grandes conjuntos de dados estão envolvidos.

IBM Data WH KNN Opções de Modelo-Opções de Scoring

Na guia Opções de Modelo - Opções de Escoragem, é possível configurar o valor padrão para a opção de escoragem e designar ponderações relativas para classes individuais.

Disponibilizar para Escoragem

Incluir campos de entrada. Especifica se os campos de entrada são incluídos na escoragem por padrão.

Ponderações de classe

Utilize essa opção se desejar alterar a importância relativa de classes individuais na construção do modelo.

Nota: esta opção será ativada somente se você estiver utilizando KNN para classificação. Se estiver executando regressão (ou seja, se o tipo de campo de destino for Contínuo), a opção será desativada.

O padrão é designar um valor de 1 para todas as classes, tornando-as igualmente ponderadas. Ao especificar diferentes ponderações numéricas para rótulos de classe diferentes, você instrui o algoritmo a ponderar os conjuntos de treinamento de classes específicas de modo apropriado.

Para alterar uma ponderação, dê um clique duplo nela na coluna **Ponderação** e faça as mudanças desejadas.

Valor. O conjunto de rótulos de classe derivados dos valores possíveis do campo de destino.

Peso. A ponderação a ser designada a uma classe específica. Designar uma ponderação maior para uma classe torna o modelo mais sensível a essa classe com relação a outras classes.

IBM Data WH K-Means

O nó K-Médias implementa o algoritmo *k*-médias, que fornece um método de análise de cluster. É possível utilizar esse nó para armazenar em cluster um conjunto de dados em grupos distintos.

O algoritmo é um algoritmo de clusterização baseado em distância que depende de uma métrica de distância (função) para medir a similaridade entre os pontos de dados. Os pontos de dados são designados ao cluster mais próximo de acordo com a métrica de distância utilizada.

O algoritmo opera ao executar várias iterações do mesmo processo básico, em que cada instância de treinamento é designada ao cluster mais próximo (com relação à função de distância especificada, aplicada à instância e ao centro do cluster). Todos os centros do cluster são, então, recalculados como os vetores do valor do atributo médio das instâncias designadas para clusters específicos.

Opções de Campo K-Means IBM Data WH

Na guia Campos, você escolhe se deseja utilizar as configurações de papel do campo já definidas em nós de envio de dados ou fazer as designações de campo manualmente.

Usar funções predefinidas. Essa opção utiliza as configurações de papel (destinos, preditores, e assim por diante) a partir de um nó Tipo de envio de dados (ou na guia Tipos de um nó de origem de envio de dados).

Usar designações de campo customizado. Escolha esta opção se desejar designar destinos, preditores e outros papéis manualmente nessa tela.

Campos. Utilize os botões de seta para designar itens manualmente a partir desta lista para os vários campos de papel à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo de papel.

Clique no botão **Tudo** para selecionar todos os campos na lista ou clique em um botão de nível de medição individual para selecionar todos os campos com esse nível de medição.

ID do registro. O campo a ser utilizado como o identificador de registro exclusivo.

Preditores (Entradas). Escolha um ou mais campos como entradas para a predição.

IBM Data WH K-Means Build Options Tab

Ao configurar as opções de construção, é possível customizar a construção do modelo para seus próprios propósitos.

Se desejar construir um modelo com as opções padrão, clique em **Executar**.

Medida de distância. Esse parâmetro define o método de medida para a distância entre pontos de dados. Distâncias maiores indicam dissimilaridades maiores. Selecione uma das opções a seguir:

- **Euclidiana.** A medida euclidiana é a distância em linha reta entre dois pontos de dados.
- **Euclidiana Normalizada.** A medida Euclidiana Normalizada é semelhante à medida Euclidiana, mas é normalizada pelo desvio padrão quadrado. Ao contrário da medida Euclidiana, a medida Euclidiana Normalizada também possui escala invariável.
- **Mahalanobis.** A medida de Mahalanobis é uma medida Euclidiana generalizada que leva em conta correlações de dados de entrada. Assim como a medida Euclidiana Normalizada, a medida de Mahalanobis possui escala invariável.
- **Manhattan.** A medida de Manhattan é a distância entre dois pontos de dados que é calculada como a soma das diferenças absolutas entre suas coordenadas.
- **Canberra.** A medida de Canberra é semelhante à medida de Manhattan, mas é mais sensível aos pontos de dados que estiverem mais próximos da origem.
- **Máxima.** A medida Máxima é a distância entre dois pontos de dados que é calculada como a maior de suas diferenças ao longo de qualquer dimensão de coordenada.

Número de clusters. Esse parâmetro define o número de clusters a serem criados.

Número máximo de iterações. O algoritmo faz várias iterações do mesmo processo. Esse parâmetro define o número de iterações após o qual cada treinamento de modelo para.

Estatísticas. Esse parâmetro define quantas estatísticas são incluídas no modelo. Selecione uma das opções a seguir:

- **Todos.** Todas as estatísticas relacionadas a coluna e todas as estatísticas relacionadas a valor são incluídas.

  Nota: Esse parâmetro inclui o número máximo de estatísticas e pode, portanto, afetar o desempenho do seu sistema. Se não desejar visualizar o modelo no formato gráfico, especifique **Nenhum**.
- **Colunas.** Estatísticas relacionadas a coluna são incluídas.
- **Nenhum.** Apenas estatísticas que são necessárias para escorar o modelo são incluídas.

Replicar resultados. Marque essa caixa de seleção se desejar configurar uma semente aleatória para replicar análises. É possível especificar um número inteiro ou criar um pseudonúmero inteiro aleatório clicando em **Gerar**.

IBM Data WH Naive Bayes

O Naive Bayes é um algoritmo bem conhecido para problemas de classificação. O modelo é chamado de *naïve* por tratar todas as variáveis de predição propostas como sendo independentes umas das outras. O Naive Bayes é um algoritmo rápido e escalável que calcula probabilidades condicionais para combinações de atributos e o atributo de destino. A partir dos dados de treinamento, uma probabilidade independente é estabelecida. Essa probabilidade fornece a verossimilhança de cada classe de destino, dada a ocorrência de cada categoria de valor a partir de cada variável de entrada.

Rede bayesiana Netezza

Uma rede bayesiana é um modelo que exibe variáveis em um conjunto de dados e as independências probabilísticas ou condicionais entre elas. Usando o nó Netezza Bayes Net, você pode construir um modelo de probabilidade combinando evidências observadas e gravadas com conhecimentos do mundo

real de bom senso para estabelecer a probabilidade de ocorrências usando atributos aparentemente desvinculados.

Opções do Campo de Rede Bayes Netezza

Na guia Campos, você escolhe se deseja utilizar as configurações de papel do campo já definidas em nós de envio de dados ou fazer as designações de campo manualmente.

Para este nó, o campo de destino é necessário apenas para escoragem, portanto, ele não é exibido nesta guia. É possível configurar ou alterar o destino em um nó Tipo, na guia Opções do Modelo desse nó ou na guia Configurações do nugget do modelo. Consulte o tópico "Nugget de Rede Netezza Bayes - Guia Configurações" na página 77 para obter informações adicionais.

Usar funções predefinidas. Essa opção utiliza as configurações de papel (destinos, preditores, e assim por diante) a partir de um nó Tipo de envio de dados (ou na guia Tipos de um nó de origem de envio de dados).

Usar designações de campo customizado. Escolha esta opção se desejar designar destinos, preditores e outros papéis manualmente nessa tela.

Campos. Utilize os botões de seta para designar itens manualmente a partir desta lista para os vários campos de papel à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo de papel.

Clique no botão **Tudo** para selecionar todos os campos na lista ou clique em um botão de nível de medição individual para selecionar todos os campos com esse nível de medição.

Preditores (Entradas). Escolha um ou mais campos como entradas para a predição.

Opções de Construção de Rede do Netezza Bayes

A guia Opções de Criação é onde você configura todas as opções para construir o modelo. Obviamente, é possível clicar somente no botão **Executar** para construir um modelo com todas as opções padrão, no entanto, você normalmente deseja customizar a construção para seus próprios propósitos.

Índice de base. O identificador numérico a ser designado ao primeiro atributo (campo de entrada) para facilitar o gerenciamento interno.

Tamanho da amostra. O tamanho da amostra a ser obtida se o número de atributos for grande a ponto de causar um tempo de processamento inaceitavelmente longo.

Exibir informações adicionais durante a execução. Se essa caixa for selecionada (padrão), informações de progresso adicionais serão exibidas em uma caixa de diálogo de mensagens.

Séries temporais Netezza

Uma **série temporal** é uma sequência de valores de dados numéricos, medidos em momentos sucessivos (embora não necessariamente regulares)--por exemplo, os preços de ações diários ou dados de vendas semanais. Analisar esses dados pode ser útil, por exemplo, para destacar o comportamento como tendências e sazonalidade (um padrão repetitivo) e prever o comportamento futuro de eventos passados.

As Séries Temporais Netezza suportam os algoritmos de série temporal a seguir.

- análise espectral
- suavização exponencial
- Média Móvel Integrada Autorregressiva (ARIMA)
- decomposição de tendência sazonal

Esses algoritmos dividem uma série temporal em uma tendência e em um componente sazonal. Esses componentes são, então, analisados a fim de construir um modelo que possa ser utilizado para predição.

A **análise espectral** é utilizada para identificar o comportamento periódico na série temporal. Para séries temporais compostas por diversas periodicidades subjacentes ou quando uma quantia considerável

de ruído aleatório está presente nos dados, a análise espectral fornece os meios mais claros de identificar componentes periódicos. Este método detecta as frequências de comportamento periódico ao transformar a série do domínio de tempo em uma série de domínio de frequência.

A **suavização exponencial** é um método de previsão que usa valores ponderados de observações de séries anteriores para prever valores futuros. Com a suavização exponencial, a influência das observações diminui ao longo do tempo de maneira exponencial. Este método prevê um ponto por vez, ajustando suas previsões conforme novos dados chegam e levando em conta inclusão, tendência e sazonalidade.

Os modelos **ARIMA** fornecem métodos mais sofisticados para modelagem de tendência e de componentes sazonais do que os modelos de suavização exponencial. Esse método envolve especificar explicitamente as ordens autorregressivas e de média móvel, bem como o grau de diferenciação.

Nota: na prática, os modelos ARIMA serão mais úteis se desejar incluir preditores que possam ajudar a explicar o comportamento das séries que estiverem sendo previstas, como o número de catálogos enviados por email ou o número de ocorrências em uma página da web da empresa. Os modelos de suavização exponencial descrevem o comportamento da série temporal sem tentar explicar por que ela se comporta dessa maneira.

A **decomposição de tendência sazonal** remove o comportamento periódico da série temporal para executar uma análise de tendência e, em seguida, seleciona uma forma de base para a tendência, como uma função quadrática. Essas formas básicas têm um número de parâmetros cujos valores são determinados de forma a minimizar o erro quadrático médio dos resíduos (ou seja, as diferenças entre os valores ajustados e observados das séries temporais).

Interpolação de Valores nas Séries Temporais Netezza

A **Interpolação** é o processo de estimativa e de inserção de valores omissos em dados de séries temporais.

Se os intervalos das séries temporais forem regulares, mas alguns valores simplesmente não estiverem presentes, os valores omissos poderão ser estimados utilizando a interpolação linear. Considere a seguinte série de chegadas de passageiros mensais em um terminal de aeroporto.

Tabela 8. Chegadas mensais em um terminal de passageiros

Mês	**Passageiros**
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
22	6.000.000

Nesse caso, a interpolação linear pode estimar o valor omisso para o mês 5 como 3.650.000 (o ponto médio entre os meses 4 e 6).

Intervalos irregulares são manipulados de maneira diferente. Considere a seguinte série de leituras de temperatura.

Tabela 9. Leituras de temperatura

Data	**Hora**	**Temperatura**
24/07/2011	7h	57

Tabela 9. Leituras de temperatura (continuação)		
Data	**Hora**	**Temperatura**
24/07/2011	14h	75
24/07/2011	21h	77
25/07/2011	7h15	59
25/07/2011	14h	77
25/07/2011	20h55	74
27/07/2011	7h	60
27/07/2011	14h	78
27/07/2011	22h	74

Aqui temos as leituras obtidas em três pontos durante três dias, mas em vários momentos, em que somente alguns deles são comuns entre os dias. Além disso, apenas dois dos dias são consecutivos.

Essa situação pode ser manipulada de uma de duas maneiras: cálculo de agregados ou determinação de um tamanho do passo.

Os agregados podem ser agregados calculados diariamente de acordo com uma fórmula baseada no conhecimento semântico dos dados. Isso pode resultar no seguinte conjunto de dados.

Tabela 10. Leituras de temperatura (agregadas)		
Data	**Hora**	**Temperatura**
24/07/2011	24h	69
25/07/2011	24h	82
26/07/2011	24h	nulo
27/07/2011	24h	77

Como alternativa, o algoritmo pode tratar a série como uma série distinta e determinar um tamanho de passo apropriado. Nesse caso, o tamanho do passo determinado pelo algoritmo pode ser de 8 horas, resultando no seguinte.

Tabela 11. Leituras de temperatura com o tamanho do passo calculado		
Data	**Hora**	**Temperatura**
24/07/2011	6h	
24/07/2011	14h	75
24/07/2011	22h	
25/07/2011	6h	
25/07/2011	14h	77
25/07/2011	22h	
26/07/2011	6h	
26/07/2011	14h	
26/07/2011	22h	
27/07/2011	6h	

Tabela 11. Leituras de temperatura com o tamanho do passo calculado (continuação)		
Data	**Hora**	**Temperatura**
27/07/2011	14h	78
27/07/2011	22h	74

Aqui, apenas quatro leituras correspondem às medições originais, mas, com a ajuda dos outros valores conhecidos na série original, os valores omissos podem novamente ser calculados por interpolação.

Opções do Campo de Séries Temporais Netezza

Na guia Campos, você especifica papéis para os campos de entrada nos dados de origem.

Campos. Utilize os botões de seta para designar itens manualmente a partir desta lista para os vários campos de papel à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo de papel.

Destino. Escolha um campo como o destino para a predição. Este deve ser um campo com um nível de medição de Contínuo.

(Preditor) Pontos de Tempo. (obrigatório) O campo de entrada que contém os valores de data ou hora para as séries temporais. Isso deve ser um campo com um nível de medição Contínuo ou Categórico e um tipo de armazenamento de dados de Data, Hora, Registro de Data e Hora ou Numérico. O tipo de armazenamento de dados do campo que você especificar aqui também define o tipo de entrada para alguns campos em outras guias deste nó de modelagem.

(Preditor) IDs de Séries Temporais (Por). Um campo contendo IDs de séries temporais; use se a entrada contiver mais de uma série temporal.

Opções de Construção de Séries Temporais Netezza

Há dois níveis de opções de construção:

- Básico – as configurações para a escolha do algoritmo, interpolação e o intervalo de tempo a serem utilizados.
- Avançado - configurações para previsão

Esta seção descreve as opções básicas.

A guia Opções de Criação é onde você configura todas as opções para construir o modelo. Obviamente, é possível clicar somente no botão **Executar** para construir um modelo com todas as opções padrão, no entanto, você normalmente deseja customizar a construção para seus próprios propósitos.

Algoritmo

Essas são as configurações relacionadas ao algoritmo de série temporal para ser utilizado.

Nome do Algoritmo. Escolha o algoritmo de série temporal que deseja utilizar. Os algoritmos disponíveis são **Análise Espectral**, **Suavização exponencial** (padrão), **ARIMA** ou **Decomposição de Tendência Sazonal**. Consulte o tópico “Séries temporais Netezza” na página 66 para obter mais informações.

Tendência. (Suavização exponencial somente) A suavização exponencial simples não executará bem se a série temporal exibir uma tendência. Utilize esse campo para especificar a tendência, se houver, para que o algoritmo possa considerá-lo.

- **Determinado pelo Sistema.** (padrão) O sistema tenta encontrar o valor ideal para esse parâmetro.
- **Nenhum(N).** A série temporal não apresenta uma tendência.
- **Aditiva(A).** Uma tendência que aumenta constantemente ao longo do tempo.
- **Aditiva amortecida (DA).** Uma tendência aditiva que eventualmente desaparece.
- **Multiplicativa(M).** Uma tendência que aumenta ao longo do tempo, normalmente mais rapidamente do que uma tendência aditiva constante.

- **Multiplicativa amortecida (DM).** Uma tendência multiplicativa que eventualmente desaparece.

Sazonalidade. (Suavização Exponencial somente) Utilize este campo para especificar se a série temporal exibe quaisquer padrões sazonais nos dados.

- **Determinado pelo Sistema.** (padrão) O sistema tenta encontrar o valor ideal para esse parâmetro.
- **Nenhum(N).** A série temporal não apresenta padrões sazonais.
- **Aditiva(A).** O padrão de flutuações sazonais exibe uma tendência constante ascendente ao longo do tempo.
- **Multiplicativa(M).** Mesmo que sazonalidade aditiva, mas, além da amplitude (a distância entre os pontos altos e baixos) das flutuações sazonais, aumenta com relação à tendência ascendente geral das flutuações.

Usar configurações determinadas pelo sistema para ARIMA. (ARIMA somente) Escolha esta opção se desejar que o sistema determine as configurações para o algoritmo ARIMA.

Especificar. (ARIMA somente) Escolha esta opção e clique no botão para especificar as configurações do ARIMA manualmente.

Interpolação

Se os dados de origem de séries temporais tiverem valores omissos, escolha um método para inserir valores estimados para preencher as diferenças nos dados. Consulte o tópico "Interpolação de Valores nas Séries Temporais Netezza" na página 67 para obter informações adicionais.

- **Linear.** Escolha este método se os intervalos das séries temporais forem regulares, mas alguns valores simplesmente não estiverem presentes.
- **Splines Exponenciais.** Ajusta uma curva suave na qual os valores de ponto de dados conhecidos aumentam ou diminuem a uma taxa alta.
- **Splines Cúbicos.** Ajusta uma curva suave nos pontos de dados conhecidos para estimar os valores omissos.

Intervalo de tempo

Aqui é possível escolher se deseja utilizar todo o intervalo de dados na série temporal ou um subconjunto contínuo desses dados para criar o modelo. Uma entrada válida para esses campos é definida pelo tipo de armazenamento de dados do campo especificado para Pontos de Tempo na guia Campos. Consulte o tópico "Opções do Campo de Séries Temporais Netezza" na página 69 para obter informações adicionais.

- **Usar tempos mais antigos e mais recentes disponíveis nos dados.** Escolha esta opção se desejar utilizar o intervalo completo dos dados de séries temporais.
- **Especificar espaço de tempo.** Escolha esta opção se desejar utilizar apenas uma parte da série temporal. Utilize os campos **Tempo mais antigo (de)** e **Tempo mais recente (até)** para especificar os limites.

Estrutura ARIMA

Especifique os valores dos diversos componentes não sazonais e sazonais do modelo ARIMA. Em cada caso, configure o operador para **=** (igual a) ou **< =** (menor ou igual a), em seguida, especifique o valor no campo adjacente. Os valores devem ser números inteiros não negativos especificando os graus.

Não sazonal. Os valores para os vários componentes não sazonais do modelo.

- **Graus of autocorrelação (p).** O número de ordens autorregressivas no modelo. As ordens autorregressivas especificam quais valores anteriores da série são utilizados para prever valores atuais. Por exemplo, uma ordem autorregressiva de 2 especifica que o valor de dois períodos de tempo da série no passado será utilizado para prever o valor atual.
- **Derivação (d).** Especifica a ordem de diferenciação aplicada à série antes de estimar os modelos. A diferenciação é necessária quando tendências estiverem presentes (as séries com tendências normalmente são não estacionárias e a modelagem ARIMA assume estacionariedade) e é utilizada para remover seus efeitos. A ordem da diferenciação corresponde ao grau de tendência das séries --

a diferenciação de primeira ordem considera tendências lineares, a diferenciação de segunda ordem considera tendências quadráticas, e assim por diante.

- **Média móvel (q).** O número de ordens de média móvel no modelo. As ordens de média móvel especificam como os desvios da média de série para valores anteriores são utilizados para prever valores atuais. Por exemplo, as ordens de média móvel de 1 e 2 especificam que os desvios do valor médio das séries de cada um dos dois últimos períodos de tempo são considerados ao prever valores atuais da série.

Sazonal. Os componentes de autocorrelação sazonal (SP), de derivação (SD) e de média móvel (SQ) desempenham o mesmo papel que seus correspondentes não sazonais. Para ordens sazonais, no entanto, os valores atuais da série são afetados pelos valores anteriores da série separados por um ou mais períodos sazonais. Por exemplo, para dados mensais (período de sazonal de 12), uma ordem sazonal de 1 significa que o valor de série atual é afetado pelo valor de série 12 períodos anteriores ao período atual. Em seguida, uma ordem sazonal de 1, para os dados mensais, será o mesmo que especificar uma ordem não sazonal de 12.

As configurações sazonais serão consideradas apenas se a sazonalidade for detectada nos dados, ou se as configurações de Período forem especificadas na guia Avançado.

Opções de Construção de Séries Temporais Netezza - Avançado

É possível utilizar as configurações avançadas para especificar opções para previsão.

Usar configurações determinadas pelo sistema para opções de construção de modelo. Escolha esta opção se desejar que o sistema determine as configurações avançadas.

Especificar. Escolha essa opção se desejar especificar as opções avançadas manualmente. (A opção não estará disponível se o algoritmo for Análise Espectral).

- **Período/Unidades para período.** O período de tempo após o qual algum comportamento característico da série temporal se repete. Por exemplo, para uma série temporal de números de vendas semanais, você especificaria 1 para o período e `Weeks` para as unidades O **Período** deve ser um número inteiro não negativo e as **Unidades para período** podem ser **Milissegundos**, **Segundos**, **Minutos**, **Horas**, **Dias**, **Semanas**, **Trimestres** ou **Anos**. Não configure **Unidades para período** se **Período** não estiver configurado ou se o tipo de tempo não for numérico. No entanto, se você especificar **Período**, deve-se também especificar **Unidades para período**.

Configurações para previsão. É possível escolher fazer previsões até um determinado ponto no tempo ou em momentos específicos. Uma entrada válida para esses campos é definida pelo tipo de armazenamento de dados do campo especificado para Pontos de Tempo na guia Campos. Veja o tópico “Opções do Campo de Séries Temporais Netezza” na página 69 para obter mais informações.

- **Horizonte de previsão.** Escolha esta opção se desejar especificar somente um terminal para previsão. As previsões serão feitas até este momento.
- **Tempos de previsão.** Selecione esta opção para especificar um ou mais momentos nos quais fazer previsões. Clique em **Incluir** para incluir uma nova linha na tabela de pontos de tempo. Para excluir uma linha, selecione a linha e clique em **Excluir**.

Opções do Modelo de Série Temporal Netezza

Na guia Opções do modelo, é possível escolher se deseja especificar um nome para o modelo ou gerar um nome automaticamente. Também é possível configurar valores padrão para as opções de saída do modelo.

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Disponibilizar para Escoragem. É possível configurar os valores padrão aqui para as opções de escoragem que aparecem no diálogo para o nugget do modelo.

- **Incluir valores históricos no resultado.** Por padrão, a saída do modelo não inclui os valores de dados históricos (aqueles utilizados para fazer a predição). Marque essa caixa de seleção para incluir estes valores.
- **Incluir valores interpolados no resultado.** Se escolher incluir valores históricos na saída, selecione esta caixa se você também desejar incluir valores interpolados, se houver. Observe que a interpolação funciona apenas nos dados históricos, portanto, esta caixa não estará disponível se **Incluir valores históricos no resultado** estiver desmarcada. Consulte o tópico "Interpolação de Valores nas Séries Temporais Netezza" na página 67 para obter mais informações.

IBM Data WH TwoStep

O nó TwoStep implementa o algoritmo TwoStep que fornece um método para armazenar dados em cluster em grandes conjuntos de dados.

É possível utilizar esse nó para armazenar em cluster dados enquanto os recursos disponíveis, por exemplo, restrições de memória e de tempo, são considerados.

O algoritmo TwoStep é um algoritmo de mineração da base de dados que armazena em cluster dados da seguinte maneira:

1. Uma árvore de variável de armazenamento em cluster (CF) é criada. Essa árvore altamente balanceada armazena variáveis de armazenamento em cluster para armazenamento em cluster hierárquico em que os registros de entrada semelhantes se tornam parte dos mesmos nós de árvore.
2. As folhas da árvore CF são agrupadas hierarquicamente na memória para gerar o resultado de armazenamento em cluster final. O melhor número de clusters é determinado automaticamente. Se você especificar um número máximo de clusters, o melhor número de clusters dentro do limite especificado será determinado.
3. O resultado de armazenamento em cluster é refinado em um segundo passo, em que um algoritmo semelhante ao algoritmo K-Médias é aplicado aos dados.

Opções do Campo IBM Data WH TwoStep

Ao configurar as opções de campo, é possível especificar a utilização das configurações de papel do campo que são definidas em nós de envio de dados. Também é possível fazer as designações de campo manualmente.

Selecionar um item. Escolha essa opção para utilizar as configurações de papel de um nó Tipo de envio de dados ou a partir da guia Tipos de um nó de origem de envio de dados. As configurações de papel são, por exemplo, destinos e preditores.

Usar designações de campo customizado. Escolha esta opção se desejar designar destinos, preditores e outros papéis manualmente.

Campos. Utilize as setas para designar itens manualmente a partir desta lista para os campos de papel à direita. Os ícones indicam os níveis de medição válidos para cada campo de papel.

ID do registro. O campo a ser utilizado como o identificador de registro exclusivo.

Preditores (Entradas). Escolha um ou mais campos como entradas para a predição.

Opções de Criação do IBM Data WH TwoStep

Ao configurar as opções de construção, é possível customizar a construção do modelo para seus próprios propósitos.

Se desejar construir um modelo com as opções padrão, clique em **Executar**.

Medida de distância. Esse parâmetro define o método de medida para a distância entre pontos de dados. Distâncias maiores indicam dissimilaridades maiores. As opções são:

- **Log da verossimilhança.** A medida de probabilidade coloca uma distribuição de probabilidade nas variáveis. Variáveis contínuas são consideradas como sendo distribuídas normalmente, ao passo que as

variáveis categóricas são consideradas como sendo multinomiais. Todas as variáveis são consideradas independentes.

- **Euclidiana.** A medida euclidiana é a distância em linha reta entre dois pontos de dados.
- **Euclidiana Normalizada.** A medida Euclidiana Normalizada é semelhante à medida Euclidiana, mas é normalizada pelo desvio padrão quadrado. Ao contrário da medida Euclidiana, a medida Euclidiana Normalizada também possui escala invariável.

Número do cluster. Esse parâmetro define o número de clusters a serem criados. As opções são:

- **Calcular automaticamente o número de clusters.** O número de clusters é calculado automaticamente. É possível especificar o número máximo de clusters no campo **Número**.
- **Especificar o número de clusters.** Especifique quantos clusters devem ser criados.

Estatísticas. Esse parâmetro define quantas estatísticas são incluídas no modelo. As opções são:

- **Todos.** Todas as estatísticas relacionadas a coluna e todas as estatísticas relacionadas a valor são incluídas.

  Nota: Esse parâmetro inclui o número máximo de estatísticas e pode, portanto, afetar o desempenho do seu sistema. Se não desejar visualizar o modelo no formato gráfico, especifique **Nenhum**.
- **Colunas.** Estatísticas relacionadas a coluna são incluídas.
- **Nenhum.** Apenas estatísticas que são necessárias para escorar o modelo são incluídas.

Replicar resultados. Marque essa caixa de seleção se desejar configurar uma semente aleatória para replicar análises. É possível especificar um número inteiro ou criar um pseudonúmero inteiro aleatório clicando em **Gerar**.

IBM Data WH PCA

A análise de componente principal (PCA) é uma poderosa técnica de redução de dados projetada para reduzir a complexidade dos dados. A PCA localiza combinações lineares dos campos de entrada que executam a melhor tarefa de capturar a variância no conjunto inteiro de campos, em que os componentes são ortogonais (não correlacionados) entre si. O objetivo é localizar um número pequeno de campos derivados (os componentes principais) que sumarizem efetivamente as informações no conjunto original de campos de entrada.

Nota: Pode ocorrer um erro ao marcar o modelo se nomes de campo minúsizar forem usados. Trata-se de um defeito conhecido do Db2 Data Warehouse, com a workaround sendo para renomear todos os campos para maiúsculas antes da pontuação.

Opções de Campo PCA IBM Data WH

Na guia Campos, você escolhe se deseja utilizar as configurações de papel do campo já definidas em nós de envio de dados ou fazer as designações de campo manualmente.

Usar funções predefinidas. Essa opção utiliza as configurações de papel (destinos, preditores, e assim por diante) a partir de um nó Tipo de envio de dados (ou na guia Tipos de um nó de origem de envio de dados).

Usar designações de campo customizado. Escolha esta opção se desejar designar destinos, preditores e outros papéis manualmente nessa tela.

Campos. Utilize os botões de seta para designar itens manualmente a partir desta lista para os vários campos de papel à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo de papel.

Clique no botão **Tudo** para selecionar todos os campos na lista ou clique em um botão de nível de medição individual para selecionar todos os campos com esse nível de medição.

ID do registro. O campo a ser utilizado como o identificador de registro exclusivo.

Preditores (Entradas). Escolha um ou mais campos como entradas para a predição.

Opções de Construção PCA IBM Data WH

A guia Opções de Criação é onde você configura todas as opções para construir o modelo. Obviamente, é possível clicar somente no botão **Executar** para construir um modelo com todas as opções padrão, no entanto, você normalmente deseja customizar a construção para seus próprios propósitos.

Centralizar dados antes de calcular o PCA. Se selecionada (padrão), essa opção executa a centralização de dados (também conhecida como "subtração da média") antes da análise. A centralização de dados é necessária para assegurar que o primeiro componente principal descreva a direção da variância máxima, caso contrário, o componente poderá corresponder mais estreitamente à média dos dados. Normalmente você desmarca esta opção para melhoria de desempenho somente se os dados já tiverem sido preparados desta maneira.

Executar ajuste de escala de dados antes de calcular o PCA. Esta opção executa ajuste de escala de dados antes da análise. Fazer isso pode tornar a análise menos arbitrária quando variáveis diferentes forem medidas em unidades diferentes. Em sua forma mais simples, o ajuste de escala de dados pode ser obtido ao dividir cada variável pela sua variação padrão.

Usar o método mais rápido, mas menos preciso para calcular o PCA. Esta opção faz com que o algoritmo utilize um método mais rápido, mas menos preciso (forceEigensolve) de localizar os componentes principais.

Gerenciando os Modelos IBM Data WH e Netezza

Os modelos IBM Data Warehouse e AnáliseIBM Netezza são adicionados à tela e à paleta de Models da mesma forma que outros modelos IBM SPSS Modeler , e podem ser usados de muita da mesma forma. No entanto, há algumas diferenças importantes, dado que cada um IBM Data Warehouse ou modelo AnáliseIBM Netezza criado em IBM SPSS Modeler realmente referencia um modelo armazenado em um servidor de banco de dados. Assim, para que um fluxo funcione corretamente, ele deverá conectar-se ao banco de dados no qual o modelo foi criado, e a tabela de modelo não deverá ter sido alterada por um processo externo.

Marcando os modelos IBM Data Warehouse e AnáliseIBM Netezza

Os modelos são representados na tela por um ícone de nugget do modelo (pepita de ouro). O principal propósito de um nugget é escorar dados para gerar predições ou permitir análise adicional das propriedades do modelo. Os escores são incluídos na forma de um ou mais campos de dados extras que podem ser tornados visíveis ao anexar um nó Tabela ao nugget e executar essa ramificação do fluxo, conforme descrito posteriormente nesta seção. Algumas caixas de diálogo de nugget, como aquelas da Árvore de Decisão ou da Árvore de Regressão, possuem adicionalmente uma guia Modelo que fornece uma representação visual do modelo.

Os campos extras são distinguidos pelo prefixo `$<id>-` adicionado ao nome do campo de destino, em que `<id>` depende do modelo, e identifica o tipo de informação que está sendo adicionada. Os identificadores diferentes são descritos nos tópicos para cada nugget do modelo.

Para visualizar os escores, conclua os seguintes passos:

1. Anexe um nó Tabela ao nugget do modelo.
2. Abra o nó Tabela.
3. Clique em **Executar** .
4. Role para a direita da janela de saída de tabela para visualizar os campos extras e seus escores.

Guia do Servidor de Nugget do IBM Data WH e Netezza

Na guia Configurações, é possível configurar opções do servidor para escorar o modelo. É possível continuar utilizando uma conexão do servidor que foi especificada anteriormente ou mover os dados para outro banco de dados que você especificar aqui.

IBM Data Warehouse Server Details. Aqui você especifica os detalhes da conexão para o banco de dados que você deseja utilizar para o modelo.

- **Utilizar conexão de envio de dados**. (padrão) Usa os detalhes de conexão especificados em um nó de envio de dados, por exemplo, o nó de origem do Banco de Dados. Esta opção funciona apenas se todos os nós de upstream forem capazes de usar o pushback SQL. Neste caso, não há necessidade de mover os dados para fora do banco de dados, já que a SQL implementa completamente todos os nós de envio de dados.
- **Mover dados para conexão**. Move os dados para o banco de dados que você especificar aqui. Fazer isso permite que a modelagem funcione se os dados estão em outro banco de dados do IBM Data Warehouse, ou um banco de dados de outro fornecedor, ou mesmo se os dados estão em um arquivo flat. Além disso, os dados serão movidos de volta para o banco de dados especificado aqui se os dados tiverem sido extraídos porque um nó não executou SQL pushback. Clique no botão **Editar** para procurar e selecionar uma conexão.

Cuidado: AnáliseIBM Netezza e o IBM Data Warehouse é geralmente usado com conjuntos de dados muito grandes. A transferência de grandes quantias de dados entre os bancos de dados, fora do banco de dados ou de volta a ele pode ser muito demorada e deve ser evitada sempre que possível.

Nome do Modelo. O nome do modelo. O nome é mostrado apenas para sua informação; ele não pode ser alterado aqui.

IBM Data WH Decision Tree Model Nuggets

O nugget do modelo de Árvore de Decisão exibe a saída a partir da operação de modelagem, e também permite configurar algumas opções para escoragem do modelo.

Ao executar um fluxo contendo um nugget do modelo de Árvore de Decisão, o nó inclui, por padrão, um novo campo, cujo nome é derivado do nome de destino.

Tabela 12. Campo de escoragem de modelo para Árvore de Decisão

Nome do campo incluído	**Significado**
$I-*target_name*	Valor predito para o registro atual.

Se selecionar a opção **Calcular probabilidades das classes designadas para registros de escoragem** no nó de modelagem ou no nugget do modelo e executar o fluxo, um campo adicional é incluído.

Tabela 13. Campo de escoragem de modelo para Árvore de Decisão - adicional

Nome do campo incluído	**Significado**
$IP-*target_name*	Valor de confiança (de 0,0 a 1,0) para a predição.

IBM Data WH Decision Tree Nugget-Modelo Tab

A guia **Modelo** mostra o Importância do Preditor do modelo de árvore de decisão em formato gráfico. O comprimento da barra representa a importância do preditor.

Nota: Quando estiver trabalhando com o IBM Netezza Analytics Versão 2.x ou anterior, o conteúdo do modelo de árvore de decisão é mostrado somente em formato textual.

Para estas versões, as seguintes informações são mostradas:

- Cada linha de texto corresponde a um nó ou a uma folha.
- A indentação reflete o nível da árvore.
- Para um nó, a condição de divisão é exibida.
- Para uma folha, o rótulo de classe designado é mostrado.

IBM Data WH Decision Tree Nugget-Configurações Tab

A guia Configurações permite configurar algumas opções para escoragem do modelo.

Incluir campos de entrada. Se selecionada, esta opção transmite todos os campos de entrada originais posteriormente, anexando um ou mais campos de modelagem extras a cada linha de dados. Se limpar essa caixa de seleção, somente o campo ID de registro e os campos de modelagem extras serão transmitidos, fazendo o fluxo executar mais rapidamente.

Calcular probabilidades das classes designadas para escoragem de registros. (Árvore de Decisão e Naive Bayes somente) Se selecionada, esta opção significa que os campos de modelagem extras incluem um campo de confiança (ou seja, uma probabilidade) e também o campo de predição. Se limpar esta caixa de seleção, somente o campo de predição será produzido.

Usar dados de entrada determinísticos. Se selecionada, essa opção assegura que qualquer algoritmo Netezza que executa diversas passagens da mesma visualização utilize o mesmo conjunto de dados para cada passagem. Se desmarcar essa caixa de seleção para mostrar que dados não determinísticos estão sendo utilizados, uma tabela temporária será criada para reter a saída de dados para processamento, como aquela produzida por um nó de partição, e será excluída após o modelo ser criado.

IBM Data WH Decision Tree Nugget-Viewer Tab

A guia **Visualizador** mostra uma apresentação em árvore do modelo de árvore da mesma forma que o SPSS Modeler faz para seu modelo de árvore de decisão.

Nota: Se o modelo for construído com o IBM Netezza Analytics Versão 2.x ou anterior, a guia **Visualizador** estará vazia.

IBM Data WH K-Means Modelo Nugget

Os nuggets do modelo K-Médias contêm todas as informações capturadas pelo modelo de armazenamento em cluster, bem como informações sobre os dados de treinamento e o processo de estimação.

Ao executar um fluxo contendo um nugget do modelo K-Médias, o nó inclui dois novos campos contendo a associação de cluster e distância do centro do cluster designado para esse registro. O novo campo com o nome $KM-K-Means é para a associação de cluster e o novo campo com o nome $KMD-K-Means é para a distância do centro do cluster.

IBM Data WH K-Means Nugget-Modelo Tab

A guia **Modelo** contém várias visualizações gráficas que mostram estatísticas de sumarização e distribuições para campos de clusters. É possível exportar os dados do modelo ou exportar a visualização como um gráfico.

Quando estiver trabalhando com o IBM Netezza Analytics Versão 2.x ou anterior, ou quando construir o modelo com o Mahalanobis como uma medida de distância, o conteúdo do modelo K-Médias é mostrado somente em formato textual.

Para estas versões, as seguintes informações são mostradas:

- **Estatísticas de Sumarização.** Para os clusters menor e maior, as estatísticas de sumarização mostram o número de registros. As estatísticas de sumarização também mostram a porcentagem do conjunto de dados que é usado por esses clusters. A lista também mostra a razão de tamanho do cluster maior com o menor.
- **Sumarização de Armazenamento em Cluster.** A sumarização de armazenamento em cluster lista os clusters que são criados pelo algoritmo. Para cada cluster, a tabela mostra o número de registros nesse cluster, junto da distância média do centro do cluster para esses registros.

IBM Data WH K-Means Nugget-Configurações Tab

A guia Configurações permite configurar algumas opções para escoragem do modelo.

Incluir campos de entrada. Se selecionada, esta opção transmite todos os campos de entrada originais posteriormente, anexando um ou mais campos de modelagem extras a cada linha de dados. Se limpar essa caixa de seleção, somente o campo ID de registro e os campos de modelagem extras serão transmitidos, fazendo o fluxo executar mais rapidamente.

Medida de distância. O método a ser utilizado para medir a distância entre pontos de dados, em que distâncias maiores indicam dissimilaridades maiores. As opções são:

- **Euclidean**. (padrão) A distância entre dois pontos é calculada pela união deles com uma linha reta.
- **Manhattan**. A distância entre dois itens é calculada como a soma das diferenças absolutas entre suas coordenadas.
- **Canberra**. Semelhante à distância de Manhattan, porém mais sensível a pontos de dados mais próximos da origem.
- **Máximo**. A distância entre dois pontos é calculada como a maior de suas diferenças ao longo de qualquer dimensão de coordenada.

Nuggets do Modelo de Rede Bayes Netezza

O nugget do modelo Rede Bayes fornece uma maneira de configurar opções para escoragem do modelo.

Ao executar um fluxo contendo um nugget do modelo de Rede Bayes, o nó inclui um novo campo, cujo nome é derivado do nome de destino.

Tabela 14. Campo de escoragem de modelo para Rede Bayes

Nome do campo incluído	**Significado**
$BN-*target_name*	Valor predito para o registro atual.

É possível visualizar o campo extra ao anexar um nó Tabela ao nugget do modelo e executar o nó Tabela.

Nugget de Rede Netezza Bayes - Guia Configurações

Na guia Configurações, é possível configurar opções para escorar o modelo.

Destino. Se desejar escorar um campo de destino que é diferente do destino atual, escolha o novo destino aqui.

ID do registro. Se nenhum campo de ID de Registro for especificado, escolha o campo a ser utilizado aqui.

Tipo de predição. A variação do algoritmo de predição que deseja utilizar:

- **Melhor (vizinho mais correlacionado).** (padrão) Utiliza o nó vizinho mais correlacionado.
- **Vizinhos (predição ponderada dos vizinhos).** Utiliza uma predição ponderada de todos os nós vizinhos.
- **Vizinhos NN (vizinhos não nulos).** Igual à opção anterior, exceto que ela ignora nós com valores nulos (ou seja, nós que correspondem aos atributos que possuem valores omissos para a instância para a qual a predição é calculada).

Incluir campos de entrada. Se selecionada, esta opção transmite todos os campos de entrada originais posteriormente, anexando um ou mais campos de modelagem extras a cada linha de dados. Se limpar essa caixa de seleção, somente o campo ID de registro e os campos de modelagem extras serão transmitidos, fazendo o fluxo executar mais rapidamente.

IBM Data WH Naive Bayes Modelo Nuggets

O nugget do modelo Naive Bayes fornece uma maneira de configurar opções para escoragem do modelo.

Ao executar um fluxo contendo um nugget do modelo Naive Bayes, o nó inclui, por padrão, um novo campo, cujo nome é derivado do nome de destino.

Tabela 15. Campo de escoragem de modelo para Naive Bayes - padrão	
Nome do campo incluído	**Significado**
$I-*target_name*	Valor predito para o registro atual.

Se selecionar a opção **Calcular probabilidades das classes designadas para registros de escoragem** no nó de modelagem ou no nugget do modelo e executar o fluxo, dois campos adicionais são incluídos.

Tabela 16. Campos de escoragem de modelo para Naive Bayes - adicionais	
Nome do campo incluído	**Significado**
$IP-*target_name*	O numerador bayesiano da classe para a instância (ou seja, o produto da probabilidade da classe anterior pelas probabilidades condicionais do valor do atributo de instância).
$ILP-*target_name*	O logaritmo natural do último.

É possível visualizar os campos extras ao anexar um nó Tabela ao nugget do modelo e executar o nó Tabela.

IBM Data WH Ingênua Bayes Nugget-Guia de Configurações

Na guia Configurações, é possível configurar opções para escorar o modelo.

Incluir campos de entrada. Se selecionada, esta opção transmite todos os campos de entrada originais posteriormente, anexando um ou mais campos de modelagem extras a cada linha de dados. Se limpar essa caixa de seleção, somente o campo ID de registro e os campos de modelagem extras serão transmitidos, fazendo o fluxo executar mais rapidamente.

Calcular probabilidades das classes designadas para escoragem de registros. (Árvore de Decisão e Naive Bayes somente) Se selecionada, esta opção significa que os campos de modelagem extras incluem um campo de confiança (ou seja, uma probabilidade) e também o campo de predição. Se limpar esta caixa de seleção, somente o campo de predição será produzido.

Melhorar a precisão da probabilidade para conjuntos de dados pequenos ou altamente não balanceados. Ao calcular as probabilidades, esta opção chama a técnica de estimação *m* para evitar probabilidades zero durante a estimação. Esse tipo de estimação de probabilidades pode ser mais lento, mas pode dar melhores resultados para conjuntos de dados pequenos ou altamente não balanceados.

IBM Data WH KNN Modelo Nuggets

O nugget do modelo KNN fornece uma maneira de configurar opções para escoragem do modelo.

Ao executar um fluxo contendo um nugget do modelo KNN, o nó inclui um novo campo, cujo nome é derivado do nome de destino.

Tabela 17. Campo de escoragem de modelo para KNN	
Nome do campo incluído	**Significado**
$KNN-*target_name*	Valor predito para o registro atual.

É possível visualizar o campo extra ao anexar um nó Tabela ao nugget do modelo e executar o nó Tabela.

IBM Data WH KNN Nugget-Configurações Tab

Na guia Configurações, é possível configurar opções para escorar o modelo.

Medida de distância. O método a ser utilizado para medir a distância entre pontos de dados, em que distâncias maiores indicam dissimilaridades maiores. As opções são:

- **Euclidean**. (padrão) A distância entre dois pontos é calculada pela união deles com uma linha reta.
- **Manhattan**. A distância entre dois itens é calculada como a soma das diferenças absolutas entre suas coordenadas.
- **Canberra**. Semelhante à distância de Manhattan, porém mais sensível a pontos de dados mais próximos da origem.
- **Máximo**. A distância entre dois pontos é calculada como a maior de suas diferenças ao longo de qualquer dimensão de coordenada.

Número de Vizinhos Mais Próximos (k). O número de vizinhos mais próximos para um caso específico. Observe que utilizar um número maior de vizinhos não resultará necessariamente em um modelo mais preciso.

A opção de *k* controla o balanceamento entre a prevenção de super ajuste (isso pode ser importante, principalmente para dados "ruidosos") e a resolução (produzindo predições diferentes para instâncias semelhantes). Geralmente é necessário ajustar o valor de *k* para cada conjunto de dados, com valores típicos variando de 1 a várias dezenas.

Incluir campos de entrada. Se selecionada, esta opção transmite todos os campos de entrada originais posteriormente, anexando um ou mais campos de modelagem extras a cada linha de dados. Se limpar essa caixa de seleção, somente o campo ID de registro e os campos de modelagem extras serão transmitidos, fazendo o fluxo executar mais rapidamente.

Padronizar medições antes de calcular a distância. Se selecionada, essa opção padroniza as medições para campos de entrada contínuos antes de calcular os valores de distância.

Use conjuntos principais para aumentar o desempenho de grandes conjuntos de dados. Se selecionada, essa opção utiliza amostragem de conjunto principal para acelerar o cálculo quando grandes conjuntos de dados estão envolvidos.

Nuggets do Modelo de Armazenamento em Cluster de Divisão Netezza

O nugget do modelo Armazenamento em Cluster de Divisão fornece uma maneira de configurar opções para escoragem do modelo.

Ao executar um fluxo contendo um nugget do modelo Armazenamento em Cluster de Divisão, o nó inclui um novo campo, cujo nome é derivado do nome de destino.

Tabela 18. Campos de escoragem de modelo para Armazenamento em Cluster de Divisão

Nome do campo incluído	**Significado**
$DC-*target_name*	Identificador do subcluster ao qual o registro atual é designado.
$DCD-*target_name*	Distância do centro do subcluster ao registro atual.

É possível visualizar os campos extras ao anexar um nó Tabela ao nugget do modelo e executar o nó Tabela.

Nugget de Armazenamento em Cluster de Divisão Netezza - Guia Configurações

Na guia Configurações, é possível configurar opções para escorar o modelo.

Incluir campos de entrada. Se selecionada, esta opção transmite todos os campos de entrada originais posteriormente, anexando um ou mais campos de modelagem extras a cada linha de dados. Se limpar essa caixa de seleção, somente o campo ID de registro e os campos de modelagem extras serão transmitidos, fazendo o fluxo executar mais rapidamente.

Medida de distância. O método a ser utilizado para medir a distância entre pontos de dados, em que distâncias maiores indicam dissimilaridades maiores. As opções são:

- **Euclidean**. (padrão) A distância entre dois pontos é calculada pela união deles com uma linha reta.
- **Manhattan**. A distância entre dois itens é calculada como a soma das diferenças absolutas entre suas coordenadas.
- **Canberra**. Semelhante à distância de Manhattan, porém mais sensível a pontos de dados mais próximos da origem.
- **Máximo**. A distância entre dois pontos é calculada como a maior de suas diferenças ao longo de qualquer dimensão de coordenada.

Nível de hierarquia aplicado. O nível de hierarquia que deve ser aplicado aos dados.

IBM Data WH PCA Modelo Nuggets

O nugget do modelo PCA fornece uma maneira de configurar opções para escoragem do modelo.

Ao executar um fluxo contendo um nugget do modelo PCA, o nó inclui, por padrão, um novo campo, cujo nome é derivado do nome de destino.

Tabela 19. Campo de escoragem de modelo para PCA

Nome do campo incluído	Significado
$F-*target_name*	Valor predito para o registro atual.

Se você especificar um valor maior que 1 no **Número de componentes principais ...** campo no nó de modelagem ou no nugget do modelo e executar o stream, o nó adiciona um novo campo para cada componente. Neste caso, os nomes de campo são sufixados por *–n*, em que *n* é o número do componente. Por exemplo, se seu modelo for denominado *pca* e contiver três componentes, os novos campos serão denominados *$F-pca-1*, *$F-pca-2* e *$F-pca-3*.

É possível visualizar os campos extras ao anexar um nó Tabela ao nugget do modelo e executar o nó Tabela.

Nota: Pode ocorrer um erro ao marcar o modelo se nomes de campo minúsizar forem usados. Trata-se de um defeito conhecido do Db2 Data Warehouse, com a workaround sendo para renomear todos os campos para maiúsculas antes da pontuação.

IBM Data WH PCA Nugget-Configurações Tab

Na guia Configurações, é possível configurar opções para escorar o modelo.

Número de componentes principais a serem utilizados na projeção. O número de componentes principais para os quais você deseja reduzir o conjunto de dados. Este valor não deve exceder o número de atributos (campos de entrada).

Incluir campos de entrada. Se selecionada, esta opção transmite todos os campos de entrada originais posteriormente, anexando um ou mais campos de modelagem extras a cada linha de dados. Se limpar essa caixa de seleção, somente o campo ID de registro e os campos de modelagem extras serão transmitidos, fazendo o fluxo executar mais rapidamente.

Nuggets do Modelo de Árvore de Regressão Netezza

O nugget do modelo de Árvore de Regressão fornece uma maneira de configurar opções para escoragem do modelo.

Ao executar um fluxo contendo um nugget do modelo de Árvore de Regressão, o nó inclui, por padrão, um novo campo, cujo nome é derivado do nome de destino.

Tabela 20. Campo de escoragem de modelo para Árvore de Regressão

Nome do campo incluído	Significado
$I-*target_name*	Valor predito para o registro atual.

Se selecionar a opção **Calcular variância estimada** no nó de modelagem ou no nugget do modelo e executar o fluxo, um campo adicional é incluído.

Tabela 21. Campo de escoragem de modelo para Árvore de Regressão - adicional

Nome do campo incluído	Significado
$IV-*target_name*	Variâncias estimadas do valor predito.

É possível visualizar os campos extras ao anexar um nó Tabela ao nugget do modelo e executar o nó Tabela.

Nugget de Árvore de Regressão Netezza - Guia Modelo

A guia **Modelo** mostra o Importância do Preditor do modelo de árvore de regressão em formato gráfico. O comprimento da barra representa a importância do preditor.

Nota: Quando estiver trabalhando com o IBM Netezza Analytics Versão 2.x ou anterior, o conteúdo do modelo de árvore de regressão é mostrado somente em formato textual.

Para estas versões, as seguintes informações são mostradas:

- Cada linha de texto corresponde a um nó ou a uma folha.
- A indentação reflete o nível da árvore.
- Para um nó, a condição de divisão é exibida.
- Para uma folha, o rótulo de classe designado é mostrado.

Nugget de Árvore de Regressão Netezza - Guia Configurações

Na guia Configurações, é possível configurar opções para escorar o modelo.

Incluir campos de entrada. Se selecionada, esta opção transmite todos os campos de entrada originais posteriormente, anexando um ou mais campos de modelagem extras a cada linha de dados. Se limpar essa caixa de seleção, somente o campo ID de registro e os campos de modelagem extras serão transmitidos, fazendo o fluxo executar mais rapidamente.

Calcular variância estimada. Indica se as variâncias das classes designadas devem ser incluídas na saída.

Nugget de Árvore de Regressão Netezza - Guia Visualizador

A guia **Visualizador** mostra uma apresentação em árvore do modelo de árvore da mesma forma que o SPSS Modeler faz para seu modelo de árvore de regressão.

Nota: Se o modelo for construído com o IBM Netezza Analytics Versão 2.x ou anterior, a guia **Visualizador** estará vazia.

IBM Data WH Linear Regression Model Nuggets

O nugget do modelo de Regressão Linear fornece uma maneira de configurar opções para escoragem do modelo.

Ao executar um fluxo contendo um nugget do modelo de Regressão Linear, o nó inclui um novo campo, cujo nome é derivado do nome de destino.

Tabela 22. Campo de escoragem de modelo para Regressão Linear

Nome do campo incluído	Significado
$LR-*target_name*	Valor predito para o registro atual.

IBM Data WH Linear Regression Nugget-Guia de Configurações

Na guia Configurações, é possível configurar opções para escorar o modelo.

Incluir campos de entrada. Se selecionada, esta opção transmite todos os campos de entrada originais posteriormente, anexando um ou mais campos de modelagem extras a cada linha de dados. Se limpar essa caixa de seleção, somente o campo ID de registro e os campos de modelagem extras serão transmitidos, fazendo o fluxo executar mais rapidamente.

Nugget do Modelo de Série Temporal do Netezza

O nugget do modelo fornece acesso à saída da operação de modelagem de série temporal. A saída consiste nos campos a seguir.

Tabela 23. Campos de saída do modelo de Série Temporal

Campo	Descrição
TSID	O identificador da série temporal; o conteúdo do campo especificado para IDs de Séries Temporais na guia Campos do nó de modelagem. Consulte o tópico "Opções do Campo de Séries Temporais Netezza" na página 69 para obter informações adicionais.
Horário	O período de tempo dentro da série temporal atual.
HISTÓRICO	Os valores de dados históricos (aqueles utilizados para fazer a previsão). Esse campo será incluído somente se a opção **Incluir valores históricos no resultado** estiver selecionada na guia Configurações do nugget do modelo.
$TS-INTERPOLATED	Os valores interpolados, onde utilizados. Esse campo será incluído somente se a opção **Incluir valores interpolados no resultado** estiver selecionada na guia Configurações do nugget do modelo. Interpolação é uma opção na guia Opções de Construção do nó de modelagem.
$TS-FORECAST	Os valores de previsão para a série temporal.

Para visualizar a saída do modelo, anexe um nó Tabela (na guia Saída da paleta do nó) ao nugget do modelo e execute o nó Tabela.

Nugget de Séries Temporais do Netezza - Guia Configurações

Na guia Configurações, é possível especificar opções para customizar a saída de modelo.

Nome do Modelo. O nome do modelo, conforme especificado na guia Opções de Modelo do nó de modelagem.

As outras opções são as mesmas que aquelas na guia Opções de Modelagem do nó de modelagem.

IBM Data WH Generalized Linear Model Nugget

O nugget do modelo fornece acesso à saída da operação de modelagem.

Ao executar um fluxo contendo um nugget do modelo Linear Generalizado, o nó inclui um novo campo, cujo nome é derivado do nome de destino.

Tabela 24. Campo de escoragem de modelo para Linear Gereneralizado

Nome do campo incluído	Significado
$GLM-*target_name*	Valor predito para o registro atual.

A guia Modelo exibe várias estatísticas relacionadas ao modelo.

A saída consiste nos campos a seguir.

Tabela 25. Campos de saída a partir do modelo Linear Generalizado	
Campo de saída	**Descrição**
Parâmetro	Os parâmetros (ou seja, as variáveis preditoras) usados pelo modelo. Estas são as colunas numéricas e nominais, bem como o intercepto (o termo constante no modelo de regressão).
Beta	O coeficiente de correlação (ou seja, o componente linear do modelo).
Erro padrão	O desvio padrão para o beta.
Testar	As estatísticas de teste utilizadas para avaliar a validade do parâmetro.
valor p	A probabilidade de um erro ao assumir que o parâmetro é significativo.
Sumarização de Residuais	
Tipo de Residual	O tipo de residual da predição para o qual os valores de sumarização são mostrados.
RSS	O valor do residual.
df	Os graus de liberdade para os residuais.
valor p	A probabilidade de um erro. Um valor alto indica um modelo insuficientemente ajustado, e um valor baixo indica um bom ajuste.

IBM Data WH Generalized Linear Modelo Nugget-Guia de Configurações

Na guia Configurações, é possível customizar a saída do modelo.

A opção é a mesma que aquela mostrada para Opções de Escoragem no nó de modelagem. Consulte o tópico "IBM Data WH Generalized Linear Model Opções-Opções de Scoring" na página 59 para obter informações adicionais.

IBM Data WH TwoStep Modelo Nugget

Ao executar um fluxo que contém um nugget do modelo TwoStep, o nó inclui dois novos campos contendo a associação e a distância do cluster do centro do cluster designado para esse registro. O novo campo com o nome $TS-Twostep é para a associação de cluster e o novo campo com o nome $TSP-Twostep é para a distância do centro do cluster.

IBM Data WH TwoStep Guia Nugget-Modelo

A guia **Modelo** contém várias visualizações gráficas que mostram estatísticas de sumarização e distribuições para campos de clusters. É possível exportar os dados do modelo ou exportar a visualização como um gráfico.

Capítulo 6. Modelagem da base de dados com o IBM DB2 for z/OS

IBM SPSS Modeler e o IBM DB2 for z/OS

O SPSS Modeler suporta integração com o DB2 for z/OS, que fornece a capacidade de executar analítica avançada em servidores DB2 for z/OS. É possível acessar esses recursos através da interface gráfica com o usuário e do ambiente de desenvolvimento orientado a fluxo de trabalho do SPSS Modeler. Desta forma, é possível executar os algoritmos de mineração de dados diretamente no ambiente do DB2 for z/OS ao alavancar o IBM DB2 Analítica Accelerator.

O SPSS Modeler suporta integração dos seguintes algoritmos do DB2 for z/OS.

- Árvores de decisão
- K-Médias
- Naive Bayes
- Árvore de Regressão
- TwoStep

Requisitos para Integração com o IBM DB2 for z/OS

As condições a seguir são pré-requisitos para conduzir a modelagem dentro da base de dados ao utilizar o DB2 for z/OS e o IBM DB2 Analytics Accelerator for z/OS. Para assegurar que essas condições sejam atendidas, poderá ser necessário consultar seu administrador de base de dados. Para requisitos detalhados, incluindo versões suportadas, consulte os Relatórios de compatibilidade do Produto de Software.

- IBM SPSS Modeler em execução no modo local ou com relação a uma instalação do SPSS Modeler Server no Windows ou UNIX
- Db2 para z/OS junto com o Db2 Analytics Accelerator para z/OS
- IBM SPSS Data Access Pack
- No servidor que estiver executando o SPSS Modeler Server, um dos seguintes sistemas:
  - IBM Db2 Data Server Driver for ODBC and CLI
  - Qualquer versão do DB2 para Linux®, UNIX e Windows com uma origem de dados ODBC que é configurada para o DB2 for z/OS
- Licença para o DB2 Connect para System z
- Geração e otimização de SQL ativadas no SPSS Modeler
- Db2 z/OS em-banco de dados em banco de dados requer apenas tabelas de acelerador-apenas tabelas (AOT) ou tabelas aceleradas, e suporte INZA. O IDAA INZA foi introduzido no IDAA 5.1.This significa que os nós de mineração Db2 z/OS em banco de dados não funcionarão com versões anteriores do IDAA.

  Se você usar um DSN habilitado pelo IDAA no Modeler, as únicas tabelas que serão exibidas na lista de tabelas retornadas no nó de origem do Banco de Dados usando esse DSN serão AOT ou tabelas aceleradas.

Ativando a Integração com o IBM DB2 Analytics Accelerator for z/OS

A ativação da integração com o DB2 Analytics Accelerator for z/OS consiste nos seguintes passos:

- Configurar o DB2 for z/OS e o DB2 Analytics Accelerator for z/OS
- Criar uma origem ODBC
- Ativar a integração do IBM DB2 for z/OS no IBM SPSS Modeler
- Ativar a geração e a otimização de SQL no SPSS Modeler
- Ativando o IBM SPSS Modeler Server Scoring Adapter para Db2 para z/OS
- Configurando o DSN usando IBM Db2 Cliente em IBM SPSS Modelador

Configurando o IBM DB2 for z/OS e o IBM Analytics Accelerator for z/OS

Instruções sobre como configurar o DB2 for z/OS e o Analytics Accelerator for z/OS são descritas no website a seguir:

Db2 Acelerador de Analytics para z/OS.

Criando uma Origem ODBC para o IBM DB2 for z/OS e para o IBM DB2 Accelerator Analytics

Para obter informações sobre como ativar uma conexão entre o DB2 for z/OS e o IBM DB2 Analytics Accelerator, consulte os seguintes websites:

- Para versão 4: DB2 Analytics Accelerator for z/OS 4.1.0
- Para versão 3: DB2 Analytics Accelerator for z/OS 3.1.0
- Ativando a aceleração de query com o IBM DB2 Analytics Accelerator para aplicativos ODBC e JDBC sem modificar os aplicativos
- Erro SQL a partir do driver ODBC ao executar uma query no DB2 Analytics Accelerator for z/OS

Ativar a integração do IBM DB2 for z/OS no IBM SPSS Modeler

Para ativar a integração do Db2 para o z/OS no SPSS Modeler, executar as seguintes etapas:

1. No diretório SPSS Modeler `config` , abra o arquivo `odbc-db2-accelerator-names.cfg` .

   Se o arquivo não existir, ele deverá ser criado.
2. Inclua os nomes de todas as fontes de dados e os nomes de todos os aceleradores. Por exemplo:

   ```
   dsn1, acceleratorname1
   dsn2, acceleratorname2
   ```
3. O CCSID padrão para acelerador apenas tabelas (AOT) é Unicode; para substituir isso, modifique as entradas adicionando strings de codificação aos nomes do acelerador. Por exemplo:

   ```
   dsn1, acceleratorname1, EBCDIC
   dsn2, acceleratorname2, UNICODE
   ```
4. Salve e feche o arquivo `odbc-db2-accelerator-names.cfg` e, em seguida, abra o arquivo `odbc-db2-custom-properties.cfg` do mesmo diretório.
5. SPSS Modelador usa SQL para configurar os registros IDAA. Se necessário, você pode substituir essas entradas, alterando o SQL para os valores necessários. Por exemplo:

   ```
   current_query_sql_acc, "SET CURRENT QUERY ACCELERATION = ELIGIBLE"
   current_get_archive_acc, "SET CURRENT GET_ACCEL_ARCHIVE = NO"
   ```
6. Por padrão, o Modelador SPSS usa SQL para criar tabelas temporárias para um cache de banco de dados. Se necessário, você pode substituir isso especificando o nome do banco de dados esperado. Por exemplo:

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

7. Por padrão, o SPSS Modeler considera que consultas SQL escritas em um nó de origem ODBC são não rejogáveis, significando que a consulta é considerada como retornar resultados diferentes ao ser executada várias vezes. No entanto, em alguns cenários, isso pode evitar que o Modeler gerem SQL para nós de downstream e possa ser substituído alterando o valor relevante para Y. Por exemplo:

   ```
   assume_custom_sql_replayable, Y
   ```

8. A partir do menu principal SPSS Modeler, clique em **Ferramentas** > **Opções** > **Aplicações de Ajuda**.
9. Clique na guia **IBM DB2 for z/OS**.
10. Selecione **Ativar IBM Db2 for z/OS Data Mining Integration** e, em seguida, clique em **OK**.

Nota: Não é possível visualizar tabelas IDAA e não IDAA ao mesmo tempo em Modeler.

Ativando geração e otimização de SQL

Devido à possibilidade de trabalhar com conjuntos de dados muito grandes, por motivos de desempenho, deve-se ativar as opções de geração e de otimização de SQL no IBM SPSS Modeler.

Para configurar o SPSS Modeler, execute os passos a seguir:

1. A partir dos IBM SPSS Os menus do Modelador escolhem **Ferramentas** > **Propriedades do Fluxo** > **Opções**
2. Clique na opção **Otimização** na área de janela de navegação.
3. Confirme se a opção **Gerar SQL** está ativada. Essa configuração é necessária para que a modelagem da base de dados funcione.
4. Selecione **Otimizar Geração de SQL e Otimizar outra execução** (não é estritamente necessário, mas é altamente recomendado para um desempenho otimizado).

Configurando o DSN usando IBM Db2 Cliente em IBM SPSS Modelador

Se necessário, para configurar um nome de origem de dados (DSN) usando o Db2 Client para Db2 em Modelador SPSS , preencha as seguintes etapas:

1. Se não já estiver instalado, instale o Db2 Client no sistema operacional onde o Modeler Server está instalado.
2. Usando o comando **`db2 catalog`** , catalogue o banco de dados e inclua uma nova origem de dados no arquivo `db2cli.ini` no Db2 Client. Certifique-se de apontar para o alias de banco de dados definido.
3. Configurar acesso a dados; etapas detalhadas estão disponíveis na documentação do Modelador.

   Para obter mais informações, consulte o tópico ' **Recomendações de arquitetura e hardware** > **Acesso aos dados** no ' *Guia de Administração e Desempenho do Modeler Server* (ModelerServerAdminPerformance.pdf).
4. Crie uma nova fonte de dados ODBC em `odbc.ini` referenciando o alias de banco de dados definido na etapa 2.
5. Para usuários do Linux ou UNIX:
   a. Certifica-se de que a biblioteca do driver `libdb2o.so` é usada (em vez de `libdb2.so`), e certifica-se de que `'DriverUnicodeType=1'` esteja definido para a nova fonte de dados.
   b. Na instalação do IBM SPSS Data Access Pack , assegure-se de que o caminho da biblioteca do Db2 Client seja incluído no `odbc.sh`.
   c. Certifica-se de que o Modeler Server usa uma biblioteca de wrapper ODBC Driver com codificação UTF-16 (isto é chamado `'libspssodbc_datadirect_utf16.so'`).
6. Certise-se de que o usuário que se conecta ao Db2 tem os privilégios necessários para executar a seguinte consulta:

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

Modelos de Construção com o IBM DB2 for z/OS

Cada um dos algoritmos suportados possui um nó de modelagem correspondente. É possível acessar os nós de modelagem do DB2 for z/OS a partir da guia Modelagem de Banco de Dados na paleta de nós.

Considerações de dados

Os campos na origem de dados podem conter variáveis de diferentes tipos de dados, dependendo do nó de modelagem. No SPSS Modeler, os tipos de dados são conhecidos como *Níveis de medição*. A guia Campos do nó de modelagem utiliza ícones para indicar os tipos de nível de medição permitidos para seus campos de entrada e de destino.

Campo de destino. O campo de destino é o campo cujo valor você está tentando prever. Quando um destino pode ser especificado, apenas um dos campos de dados de origem pode ser selecionado como o campo de destino.

Campo de ID de registro. Especifica o campo utilizado para identificar exclusivamente cada caso. Por exemplo, esse pode ser um campo de ID, como *CustomerID*. Se os dados de origem não incluírem um campo de ID, será possível criar este campo por meio de um nó Derivar, como mostra o procedimento a seguir.

1. Selecione o nó de origem.
2. Na guia Operações de Campo na paleta de nós, dê um clique duplo no nó Derivar.
3. Abra o nó Derivar ao dar um clique duplo em seu ícone na tela.
4. No campo **Derivar campo**, digite (por exemplo) `ID`.
5. No campo **Fórmula**, digite `@INDEX` e clique em **OK**.
6. Conecte o nó Derivar ao restante do fluxo.

Manipulando valores nulos

Se os dados de entrada contiverem valores nulos, o uso de alguns dos nós do DB2 for z/OS poderá resultar em mensagens de erro ou fluxos de longa execução, portanto, é recomendado remover registros que contiverem valores nulos. Use o método a seguir.

1. Anexe um nó Seleção ao nó de origem.
2. Configure a opção **Modo** do nó Seleção para **Descartar**.
3. Insira o seguinte no campo **Condição**:

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

   Assegure-se de incluir cada campo de entrada.
4. Conecte o nó Seleção ao restante do fluxo.

Saída de modelo

Um fluxo contendo um nó de modelagem do DB2 para z/OS pode produzir resultados um pouco diferentes toda vez que for executado. Isso ocorre porque a ordem na qual o nó lê os dados de origem nem sempre é a mesma, já que os dados são lidos em tabelas temporárias antes da construção de modelo. No entanto, as diferenças produzidas por este efeito são insignificantes.

Comentários gerais

- No SPSS Collaboration and Deployment Services, não é possível criar as configurações de escoragem usando fluxos contendo nós de modelagem do DB2 for z/OS.
- A exportação ou importação do PMML não é possível para modelos criados pelos nós do DB2 for z/OS.

Modelos do IBM DB2 for z/OS - Opções de Campo

Na guia Campos, você escolhe se deseja utilizar as configurações de papel do campo já definidas em nós de envio de dados ou fazer as designações de campo manualmente.

Usar funções predefinidas. Essa opção utiliza as configurações de papel (destinos, preditores, e assim por diante) a partir de um nó Tipo de envio de dados (ou na guia Tipos de um nó de origem de envio de dados).

Usar designações de campo customizado. Escolha esta opção se desejar designar destinos, preditores e outros papéis manualmente nessa tela.

Campos. Utilize os botões de seta para designar itens manualmente a partir desta lista para os vários campos de papel à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo de papel.

Clique no botão **Tudo** para selecionar todos os campos na lista ou clique em um botão de nível de medição individual para selecionar todos os campos com esse nível de medição.

Destino. Escolha um campo como o destino para a predição. Para modelos Lineares Generalizados, consulte também o campo **Avaliações** nessa tela.

ID do registro. O campo a ser utilizado como o identificador de registro exclusivo.

Preditores (Entradas). Escolha um ou mais campos como entradas para a predição.

Modelos do IBM DB2 for z/OS - opções do servidor

Na guia Servidor, especifique o sistema do DB2 for z/OS no qual o modelo deve ser construído.

- **Usar conexão de envio de dados.** (padrão) Usa os detalhes de conexão especificados em um nó de envio de dados, por exemplo, o nó de origem do Banco de Dados. *Nota:* essa opção funcionará somente se todos os nós de envio de dados forem capazes de utilizar o SQL pushback. Neste caso, não há necessidade de mover os dados para fora do banco de dados, já que a SQL implementa completamente todos os nós de envio de dados.
- **Mover dados para conexão**. Move os dados para o banco de dados que você especificar aqui. Fazer isso permite que a modelagem funcione se os dados estiverem em outro banco de dados IBM ou em um banco de dados de outro fornecedor, ou mesmo se os dados estiverem em um arquivo simples. Além disso, os dados serão movidos de volta para o banco de dados especificado aqui se os dados tiverem sido extraídos porque um nó não executou SQL pushback. Clique no botão **Editar** para procurar e selecionar uma conexão.

Nota: O nome da origem de dados ODBC é integrado efetivamente em cada fluxo do SPSS Modeler. Se um fluxo que é criado em um host for executado em um host diferente, o nome da origem de dados deverá ser o mesmo em cada host. Como alternativa, uma origem de dados diferente pode ser selecionada na guia Servidor em cada nó de origem ou de modelagem.

Modelos do IBM DB2 for z/OS - opções do modelo

Na guia Opções do modelo, é possível escolher se deseja especificar um nome para o modelo ou gerar um nome automaticamente.

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Substituir existente se o nome tiver sido usado. Se marcar essa caixa de seleção, qualquer modelo existente com o mesmo nome será sobrescrito.

Modelos do IBM DB2 for z/OS - K-Médias

O nó K-Médias implementa o algoritmo *k*-médias, que fornece um método de análise de cluster. É possível utilizar esse nó para armazenar em cluster um conjunto de dados em grupos distintos.

O algoritmo é um algoritmo de clusterização baseado em distância que depende de uma métrica de distância (função) para medir a similaridade entre os pontos de dados. Os pontos de dados são designados ao cluster mais próximo de acordo com a métrica de distância utilizada.

O algoritmo opera ao executar várias iterações do mesmo processo básico, em que cada instância de treinamento é designada ao cluster mais próximo (com relação à função de distância especificada, aplicada à instância e ao centro do cluster). Todos os centros do cluster são, então, recalculados como os vetores do valor do atributo médio das instâncias designadas para clusters específicos.

Modelos do IBM DB2 for z/OS - opções de campo K-Médias

Na guia Campos, você escolhe se deseja utilizar as configurações de papel do campo já definidas em nós de envio de dados ou fazer as designações de campo manualmente.

Usar funções predefinidas. Essa opção utiliza as configurações de papel (destinos, preditores, e assim por diante) a partir de um nó Tipo de envio de dados (ou na guia Tipos de um nó de origem de envio de dados).

Usar designações de campo customizado. Escolha esta opção se desejar designar destinos, preditores e outros papéis manualmente nessa tela.

Campos. Utilize os botões de seta para designar itens manualmente a partir desta lista para os vários campos de papel à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo de papel.

Clique no botão **Tudo** para selecionar todos os campos na lista ou clique em um botão de nível de medição individual para selecionar todos os campos com esse nível de medição.

ID do registro. O campo a ser utilizado como o identificador de registro exclusivo.

Preditores (Entradas). Escolha um ou mais campos como entradas para a predição.

Modelos do IBM DB2 for z/OS - Opções de Construção de K-Médias

Ao configurar as opções de construção, é possível customizar a construção do modelo para seus próprios propósitos.

Se desejar construir um modelo com as opções padrão, clique em **Executar**.

Medida de Distância. Esse parâmetro define o método de medida para a distância entre pontos de dados. Distâncias maiores indicam dissimilaridades maiores. Selecione uma das opções a seguir:

- **Euclidiana.** A medida euclidiana é a distância em linha reta entre dois pontos de dados.
- **Euclidiana Normalizada.** A medida Euclidiana Normalizada é semelhante à medida Euclidiana, mas é normalizada pelo desvio padrão quadrado. Ao contrário da medida Euclidiana, a medida Euclidiana Normalizada também possui escala invariável.

Número de clusters. Esse parâmetro define o número de clusters a serem criados.

Número máximo de iterações. O algoritmo faz várias iterações do mesmo processo. Esse parâmetro define o número de iterações após o qual cada treinamento de modelo para.

Estatísticas. Esse parâmetro define quantas estatísticas são incluídas no modelo. Selecione uma das opções a seguir:

- **Todos.** Todas as estatísticas relacionadas a coluna e todas as estatísticas relacionadas a valor são incluídas.

  Nota: Esse parâmetro inclui o número máximo de estatísticas e pode, portanto, afetar o desempenho do seu sistema. Se não desejar visualizar o modelo no formato gráfico, especifique **Nenhum**.
- **Colunas.** Estatísticas relacionadas a coluna são incluídas.
- **Nenhum.** Apenas estatísticas que são necessárias para escorar o modelo são incluídas.

Replicar resultados. Marque essa caixa de seleção se desejar configurar uma semente aleatória para replicar análises. É possível especificar um número inteiro ou criar um pseudonúmero inteiro aleatório clicando em **Gerar**.

Modelos do IBM DB2 for z/OS - Naive Bayes

O Naive Bayes é um algoritmo bem conhecido para problemas de classificação. O modelo é chamado de naïve por tratar todas as variáveis de predição propostas como sendo independentes umas das outras. O Naive Bayes é um algoritmo rápido e escalável que calcula probabilidades condicionais para combinações de atributos e o atributo de destino. A partir dos dados de treinamento, uma probabilidade independente é estabelecida. Essa probabilidade fornece a verossimilhança de cada classe de destino, dada a ocorrência de cada categoria de valor a partir de cada variável de entrada.

Modelos do IBM DB2 for z/OS - Árvores de Decisão

Uma árvore de decisão é uma estrutura hierárquica que representa um modelo de classificação. Com um modelo de árvore de decisão, é possível desenvolver um sistema de classificação para prever ou classificar futuras observações a partir de um conjunto de dados de treinamento. A classificação toma a forma de uma estrutura em árvore na qual as ramificações representam pontos de divisão na classificação. As divisões dividem os dados em subgrupos recursivamente até que um ponto de interrupção seja atingido. Os nós da árvore nos pontos de parada são conhecidos como *folhas*. Cada folha designa um rótulo, conhecido como um *rótulo de classe*, aos membros de seu subgrupo ou classe.

Modelos do IBM DB2 for z/OS - Opções do Campo de Árvore de Decisão

Na guia Campos, você escolhe se deseja utilizar as configurações de papel do campo já definidas em nós de envio de dados ou fazer as designações de campo manualmente.

Usar funções predefinidas. Essa opção utiliza as configurações de papel (destinos, preditores, e assim por diante) a partir de um nó Tipo de envio de dados (ou na guia Tipos de um nó de origem de envio de dados).

Usar designações de campo customizado. Escolha esta opção se desejar designar destinos, preditores e outros papéis manualmente nessa tela.

Campos. Utilize os botões de seta para designar itens manualmente a partir desta lista para os vários campos de papel à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo de papel.

Clique no botão **Tudo** para selecionar todos os campos na lista ou clique em um botão de nível de medição individual para selecionar todos os campos com esse nível de medição.

Destino. Escolha um campo como o destino para a predição.

ID do registro. O campo a ser utilizado como o identificador de registro exclusivo. Os valores deste campo devem ser exclusivos para cada registro (por exemplo, números de ID do cliente).

Peso da Instância. Especificar um campo aqui permite usar ponderações de instância (uma ponderação por linha de dados de entrada) ao invés ou além das ponderações de classe padrão (uma ponderação por categoria para o campo de destino). O campo que você especificar aqui deve ser um que contenha uma ponderação numérica para cada linha dos dados de entrada.

Preditores (Entradas). Selecione o campo ou os campos de entrada. Isso é semelhante a configurar o papel do campo como *Entrada* em um nó Tipo.

Modelos do IBM DB2 for z/OS - Opções de Construção de Árvore de Decisão

As opções de construção a seguir estão disponíveis para crescimento da árvore:

Medida de Crescimento. Essas opções controlam a maneira como o crescimento da árvore é medido.

- **Medida de Impureza.** Esta medida avalia o melhor local para dividir a árvore. É uma medição da variabilidade em um subgrupo ou segmento de dados. Uma medição de impureza baixa indica um grupo no qual a maioria dos membros possui valores semelhantes para o campo de critério ou de destino.

  As medições suportadas são **Entropia** e **Gini**. Essas medições baseiam-se em probabilidades da associação de categoria para a ramificação.
- **Profundidade máxima da árvore.** O número máximo de níveis até o qual a árvore pode crescer abaixo do nó raiz, ou seja, o número de vezes em que a amostra é dividida recursivamente. O valor-padrão desta propriedade é 10, e o valor máximo que pode ser configurado para essa propriedade é 62.

  Nota: Se o visualizador no nugget do modelo mostrar a representação textual do modelo, um máximo de 12 níveis da árvore será exibido.

Critérios de Divisão. Estas opções controlam quando parar a divisão da árvore.

- **Melhoria mínima para divisões.** O valor mínimo pelo qual a impureza deve ser reduzida antes de uma nova divisão ser criada na árvore. O objetivo da construção de árvore é criar subgrupos com valores de saída semelhantes para minimizar a impureza dentro de cada nó. Se a melhor divisão de uma ramificação reduzir a impureza em um nível menor que a quantia especificada pelo critério de divisão, a ramificação não será dividida.
- **Número mínimo de instâncias para uma divisão.** O número mínimo de registros que podem ser divididos. Quando uma quantia menor que esse número de registros não divididos permanece, nenhuma divisão adicional é feita. É possível utilizar esse campo para evitar a criação de subgrupos pequenos na árvore.

Estatísticas. Esse parâmetro define quantas estatísticas são incluídas no modelo. Selecione uma das opções a seguir:

- **Todos.** Todas as estatísticas relacionadas a coluna e todas as estatísticas relacionadas a valor são incluídas.

  Nota: Esse parâmetro inclui o número máximo de estatísticas e pode, portanto, afetar o desempenho do seu sistema. Se não desejar visualizar o modelo no formato gráfico, especifique **Nenhum**.
- **Colunas.** Estatísticas relacionadas a coluna são incluídas.
- **Nenhum.** Apenas estatísticas que são necessárias para escorar o modelo são incluídas.

Modelos do IBM DB2 for z/OS - Nó Árvore de Decisão - Ponderações de Classe

Aqui é possível designar ponderações para classes individuais. O padrão é designar um valor de 1 para todas as classes, tornando-as igualmente ponderadas. Ao especificar diferentes ponderações numéricas para rótulos de classe diferentes, você instrui o algoritmo a ponderar os conjuntos de treinamento de classes específicas de modo apropriado.

Para alterar uma ponderação, dê um clique duplo nela na coluna **Ponderação** e faça as mudanças desejadas.

Valor. O conjunto de rótulos de classe derivados dos valores possíveis do campo de destino.

Peso. A ponderação a ser designada a uma classe específica. Designar uma ponderação maior para uma classe torna o modelo mais sensível a essa classe com relação a outras classes.

As ponderações de classe podem ser usadas em combinação com as ponderações de instância.

Modelos do IBM DB2 for z/OS - Nó Árvore de Decisão - Poda da Árvore

É possível usar as opções de poda para especificar os critérios de poda para a árvore de decisão. A intenção da poda é reduzir o risco de super ajuste ao remover subgrupos crescidos demasiadamente que não melhoram a precisão esperada nos novos dados.

Medida de poda. A medida de poda padrão, **Precisão**, assegura que a precisão estimada do modelo permaneça dentro dos limites aceitáveis após remover uma folha da árvore. Utilize a alternativa, **Precisão Ponderada**, se desejar levar em conta as ponderações de classe ao aplicar a poda.

Dados para poda. É possível usar alguns ou todos os dados de treinamento para estimar a precisão esperada nos novos dados. Como alternativa, é possível usar um conjunto de dados de poda separado de uma tabela especificada para esse propósito.

- **Usar todos os dados de treinamento.** Essa opção (a padrão) usa todos os dados de treinamento para estimar a precisão do modelo.
- **Usar % de dados de treinamento para poda.** Use esta opção para dividir os dados em dois conjuntos, um para treinamento e outro para poda, utilizando a porcentagem especificada aqui para a poda de dados.
- Selecione **Replicar resultados** se quiser especificar uma semente aleatória para assegurar que os dados sejam particionados da mesma maneira toda vez que executar o fluxo. É possível especificar um número inteiro no campo **Valor semente usado para poda** ou clicar em **Gerar**, que criará um pseudonúmero inteiro aleatório.
- **Usar dados de uma tabela existente.** Especifique o nome da tabela de um conjunto de dados de poda separado para estimar a precisão do modelo. Fazer isso é considerado mais confiável do que utilizar dados de treinamento.

Modelos do IBM DB2 for z/OS - Árvore de Regressão

Uma árvore de regressão é um algoritmo baseado em árvore que divide uma amostra de casos repetidamente para derivar subconjuntos do mesmo tipo, com base nos valores de um campo de destino numérico. Assim como as árvores de decisão, as árvores de regressão decompõem os dados em subconjuntos nos quais as folhas da árvore correspondem a subconjuntos suficientemente pequenos ou suficientemente uniformes. As divisões são selecionadas para reduzir a dispersão dos valores de atributo de destino, de modo que eles possam ser razoavelmente bem preditos pelos seus valores médios nas folhas.

Modelos do IBM DB2 for z/OS - opções de construção da Árvore de Regressão - crescimento da árvore

É possível configurar as opções de construção para o crescimento e poda da árvore.

As opções de construção a seguir estão disponíveis para crescimento da árvore:

Profundidade máxima da árvore. O número máximo de níveis até o qual a árvore pode crescer abaixo do nó raiz, ou seja, o número de vezes em que a amostra é dividida recursivamente. O padrão é 62, que é a profundidade máxima da árvore para propósitos de modelagem.

Nota: Se o visualizador no nugget do modelo mostrar a representação textual do modelo, um máximo de 12 níveis da árvore será exibido.

Critérios de Divisão. Estas opções controlam quando parar a divisão da árvore.

- **Medida de avaliação de divisão.** Essa medida de avaliação de classe avalia o melhor local para dividir a árvore.

  Nota: Atualmente, a variância é a única opção possível.
- **Melhoria mínima para divisões.** O valor mínimo pelo qual a impureza deve ser reduzida antes de uma nova divisão ser criada na árvore. O objetivo da construção de árvore é criar subgrupos com valores de saída semelhantes para minimizar a impureza dentro de cada nó. Se a melhor divisão de uma ramificação reduzir a impureza em um nível menor que a quantia especificada pelo critério de divisão, a ramificação não será dividida.
- **Número mínimo de instâncias para uma divisão.** O número mínimo de registros que podem ser divididos. Quando uma quantia menor que esse número de registros não divididos permanece,

nenhuma divisão adicional é feita. É possível utilizar esse campo para evitar a criação de subgrupos pequenos na árvore.

Estatísticas. Esse parâmetro define quantas estatísticas são incluídas no modelo. Selecione uma das opções a seguir:

- **Todos.** Todas as estatísticas relacionadas a coluna e todas as estatísticas relacionadas a valor são incluídas.

  Nota: Esse parâmetro inclui o número máximo de estatísticas e pode, portanto, afetar o desempenho do seu sistema. Se não desejar visualizar o modelo no formato gráfico, especifique **Nenhum**.
- **Colunas.** Estatísticas relacionadas a coluna são incluídas.
- **Nenhum.** Apenas estatísticas que são necessárias para escorar o modelo são incluídas.

Modelos do IBM DB2 for z/OS - opções de construção da Árvore de Regressão - poda da árvore

É possível usar as opções de poda para especificar os critérios de poda para a árvore de regressão. A intenção da poda é reduzir o risco de super ajuste ao remover subgrupos crescidos demasiadamente que não melhoram a precisão esperada nos novos dados.

Medida de poda. A medida de poda assegura que a precisão estimada do modelo permaneça dentro dos limites aceitáveis após remover uma folha da árvore. É possível selecionar uma das medidas a seguir.

- **mse.** Erro quadrático médio (padrão) - mede o quão próxima uma linha ajustada está dos pontos de dados.
- **r2.** R-quadrado - mede a proporção de variação na variável dependente explicada pelo modelo de regressão.
- **Pearson.** coeficiente de correlação de Pearson - mede a intensidade do relacionamento entre as variáveis linearmente dependentes que são normalmente distribuídas.
- **Spearman.** coeficiente de correlação de Spearman - detecta relacionamentos não lineares que parecem fracas de acordo com a correlação Pearson, mas que podem realmente ser fortes.

Dados para poda. É possível usar alguns ou todos os dados de treinamento para estimar a precisão esperada nos novos dados. Como alternativa, é possível usar um conjunto de dados de poda separado de uma tabela especificada para esse propósito.

- **Usar todos os dados de treinamento.** Essa opção (a padrão) usa todos os dados de treinamento para estimar a precisão do modelo.
- **Usar % de dados de treinamento para poda.** Use esta opção para dividir os dados em dois conjuntos, um para treinamento e outro para poda, utilizando a porcentagem especificada aqui para a poda de dados.

  Selecione **Replicar resultados** se quiser especificar uma semente aleatória para assegurar que os dados sejam particionados da mesma maneira toda vez que executar o fluxo. É possível especificar um número inteiro no campo **Valor semente usado para poda** ou clicar em **Gerar**, que criará um pseudonúmero inteiro aleatório.
- **Usar dados de uma tabela existente.** Especifique o nome da tabela de um conjunto de dados de poda separado para estimar a precisão do modelo. Fazer isso é considerado mais confiável do que utilizar dados de treinamento.

Modelos do IBM DB2 for z/OS - TwoStep

O nó TwoStep implementa o algoritmo TwoStep que fornece um método para armazenar dados em cluster em grandes conjuntos de dados.

É possível utilizar esse nó para armazenar em cluster dados enquanto os recursos disponíveis, por exemplo, restrições de memória e de tempo, são considerados.

O algoritmo TwoStep é um algoritmo de mineração da base de dados que armazena em cluster dados da seguinte maneira:

1. Uma árvore de variável de armazenamento em cluster (CF) é criada. Essa árvore altamente balanceada armazena variáveis de armazenamento em cluster para armazenamento em cluster hierárquico em que os registros de entrada semelhantes se tornam parte dos mesmos nós de árvore.
2. As folhas da árvore CF são agrupadas hierarquicamente na memória para gerar o resultado de armazenamento em cluster final. O melhor número de clusters é determinado automaticamente. Se você especificar um número máximo de clusters, o melhor número de clusters dentro do limite especificado será determinado.
3. O resultado de armazenamento em cluster é refinado em um segundo passo, em que um algoritmo semelhante ao algoritmo K-Médias é aplicado aos dados.

Modelos do IBM DB2 for z/OS - opções de campo do TwoStep

Ao configurar as opções de campo, é possível especificar a utilização das configurações de papel do campo que são definidas em nós de envio de dados. Também é possível fazer as designações de campo manualmente.

Selecionar um item. Escolha essa opção para utilizar as configurações de papel de um nó Tipo de envio de dados ou a partir da guia Tipos de um nó de origem de envio de dados. As configurações de papel são, por exemplo, destinos e preditores.

Usar designações de campo customizado. Escolha esta opção se desejar designar destinos, preditores e outros papéis manualmente.

Campos. Utilize as setas para designar itens manualmente a partir desta lista para os campos de papel à direita. Os ícones indicam os níveis de medição válidos para cada campo de papel.

ID do Registro. O campo a ser utilizado como o identificador de registro exclusivo.

Preditores (Entradas). Escolha um ou mais campos como entradas para a predição.

Modelos do IBM DB2 for z/OS - opções de construção do TwoStep

Ao configurar as opções de construção, é possível customizar a construção do modelo para seus próprios propósitos.

Se desejar construir um modelo com as opções padrão, clique em **Executar**.

Medida de Distância. Esse parâmetro define o método de medida para a distância entre pontos de dados. Distâncias maiores indicam dissimilaridades maiores. A opção é:

- **Log da verossimilhança.** A medida de probabilidade coloca uma distribuição de probabilidade nas variáveis. Variáveis contínuas são consideradas como sendo distribuídas normalmente, ao passo que as variáveis categóricas são consideradas como sendo multinomiais. Todas as variáveis são consideradas independentes.

Número do Cluster. Esse parâmetro define o número de clusters a serem criados. As opções são:

- **Calcular automaticamente o número de clusters.** O número de clusters é calculado automaticamente. É possível especificar o número máximo de clusters no campo **Número**.
- **Especificar o número de clusters.** Especifique quantos clusters devem ser criados.

Estatísticas. Esse parâmetro define quantas estatísticas são incluídas no modelo. As opções são:

- **Todos.** Todas as estatísticas relacionadas a coluna e todas as estatísticas relacionadas a valor são incluídas.

  Nota: Esse parâmetro inclui o número máximo de estatísticas e pode, portanto, afetar o desempenho do seu sistema. Se não desejar visualizar o modelo no formato gráfico, especifique **Nenhum**.
- **Colunas.** Estatísticas relacionadas a coluna são incluídas.
- **Nenhum.** Apenas estatísticas que são necessárias para escorar o modelo são incluídas.

Replicar resultados. Marque essa caixa de seleção se desejar configurar uma semente aleatória para replicar análises. É possível especificar um número inteiro ou criar um pseudonúmero inteiro aleatório clicando em **Gerar**.

Modelos do IBM DB2 for z/OS - Nugget TwoStep - guia Modelo

A guia **Modelo** contém várias visualizações gráficas que mostram estatísticas de sumarização e distribuições para campos de clusters. É possível exportar os dados do modelo ou exportar a visualização como um gráfico.

Gerenciando Modelos do IBM DB2 for z/OS

Os modelos do DB2 for z/OS são incluídos na tela e na paleta de Modelos da mesma forma que outros modelos do IBM SPSS Modeler e podem ser utilizados em grande parte da mesma forma.

Para escorar os dados diretamente no DB2 for z/OS, execute os seguintes passos:

1. Instale o SPSS Scoring Adapter no banco de dados do DB2 for z/OS no qual os dados estão localizados.
2. Assegure-se de que o fluxo se conecte ao banco de dados do DB2 for z/OS no qual os dados estão localizados.

Escorando Modelos do IBM DB2 for z/OS

Os modelos são representados na tela por um ícone de nugget do modelo (pepita de ouro). O principal propósito de um nugget é escorar dados para gerar predições ou permitir análise adicional das propriedades do modelo. Os escores são incluídos na forma de um ou mais campos de dados extras que podem ser tornados visíveis ao anexar um nó Tabela ao nugget e executar essa ramificação do fluxo, conforme descrito posteriormente nesta seção. Algumas caixas de diálogo de nugget, como aquelas da Árvore de Decisão ou da Árvore de Regressão, possuem adicionalmente uma guia Modelo que fornece uma representação visual do modelo.

Os campos extras são distinguidos pelo prefixo `$<id>-` adicionado ao nome do campo de destino, em que `<id>` depende do modelo, e identifica o tipo de informação que está sendo adicionada. Os identificadores diferentes são descritos nos tópicos para cada nugget do modelo.

Para visualizar os escores, conclua os seguintes passos:

1. Anexe um nó Tabela ao nugget do modelo.
2. Abra o nó Tabela.
3. Clique em **Executar** .
4. Role para a direita da janela de saída de tabela para visualizar os campos extras e seus escores.

Nota: O processo de escoragem não é executado no acelerador, mas no DB2, requerendo, consequentemente, que a tabela de entrada para o escore esteja localizada fisicamente no DB2. Portanto, como entrada de escoragem, apenas uma tabela baseada em DB2 ou uma tabela acelerada pode ser utilizada. Se o fluxo usar uma tabela somente do acelerador, ocorrerá o seguinte erro: `"THE STATEMENT CANNOT BE EXECUTED BY DB2 OR IN THE ACCELERATOR."`

Nuggets do Modelo de Árvore de Decisão do IBM DB2 for z/OS

O nugget do modelo de Árvore de Decisão exibe a saída a partir da operação de modelagem e também permite configurar algumas opções para escoragem do modelo.

Ao executar um fluxo que contém um nugget do modelo de Árvore de Decisão, o nó inclui dois novos campos, cujos nomes são derivados do destino.

Tabela 26. Campo de escoragem de modelo para Árvore de Decisão	
Nome do campo incluído	**Significado**
$I-*target_name*	Valor predito para o registro atual.
$IP-*target_name*	Valor de confiança (de 0,0 a 1,0) para a predição.

Nota: Devido às limitações no DB2 for z/OS, os nomes de colunas podem ser truncados.

Nugget de Árvore de Decisão do IBM DB2 for z/OS - guia Modelo

A guia **Modelo** mostra o Importância do Preditor do modelo de árvore de decisão em formato gráfico. O comprimento da barra representa a importância do preditor.

Nugget de Árvore de Decisão do IBM DB2 for z/OS - Guia Visualizador

A guia **Visualizador** mostra uma apresentação em árvore do modelo de árvore da mesma forma que o SPSS Modeler faz para seu modelo de árvore de decisão.

Nugget do Modelo K-Médias do IBM DB2 for z/OS

Os nuggets do modelo K-Médias contêm todas as informações capturadas pelo modelo de armazenamento em cluster, bem como informações sobre os dados de treinamento e o processo de estimação.

Ao executar um fluxo que contém um nugget do modelo K-Médias, o nó inclui dois novos campos contendo a associação e a distância do cluster do centro do cluster designado para esse registro. Os novos nomes de campo são derivados do nome do modelo, prefixados com `$KM-` para a associação de cluster e com `$KMD-` para a distância do centro do cluster. Por exemplo, se o seu modelo for denominado `Kmeans`, os novos campos serão denominados `$KM-Kmeans` e `$KMD-Kmeans`.

Nota: Devido às limitações no DB2 for z/OS, os nomes de colunas podem ser truncados.

Nugget de K-Médias do IBM DB2 for z/OS - guia Modelo

A guia **Modelo** contém várias visualizações gráficas que mostram estatísticas de sumarização e distribuições para campos de clusters. É possível exportar os dados do modelo ou exportar a visualização como um gráfico.

Nuggets do Modelo Naive Bayes do IBM DB2 for z/OS

Ao executar um fluxo que contém um nugget do modelo Naive Bayes, o nó inclui dois novos campos, cujos nomes são derivados do nome de destino.

Tabela 27. Campo de escoragem de modelo para Naive Bayes	
Nome do campo incluído	**Significado**
$I-*target_name*	Valor predito para o registro atual.
$IP-*target_name*	Valor de confiança (de 0,0 a 1,0) para a predição.

Nota: Devido às limitações no DB2 for z/OS, os nomes de colunas podem ser truncados.

É possível visualizar os campos extras ao anexar um nó Tabela ao nugget do modelo e executar o nó Tabela.

Nuggets do Modelo de Árvore de Regressão do IBM DB2 for z/OS

Ao executar um fluxo que contém um nugget do modelo de Árvore de Regressão, o nó inclui dois novos campos, cujos nomes são derivados do nome de destino.

Tabela 28. Campo de escoragem de modelo para Árvore de Regressão	
Nome do campo incluído	**Significado**
$I-*target_name*	Valor predito para o registro atual.
$IS-*target_name*	Desvio padrão estimado do valor predito.

Nota: Devido às limitações no DB2 for z/OS, os nomes de colunas podem ser truncados.

É possível visualizar os campos extras ao anexar um nó Tabela ao nugget do modelo e executar o nó Tabela.

Nugget de Árvore de Regressão do IBM DB2 for z/OS - guia Modelo

A guia **Modelo** mostra o Importância do Preditor do modelo de árvore de regressão em formato gráfico. O comprimento da barra representa a importância do preditor.

Nugget de Árvore de Regressão do IBM DB2 for z/OS - guia Visualizador

A guia **Visualizador** mostra uma apresentação em árvore do modelo de árvore da mesma forma que o SPSS Modeler faz para seu modelo de árvore de regressão.

Nugget do Modelo TwoStep do IBM DB2 for z/OS

Ao executar um fluxo que contém um nugget do modelo TwoStep, o nó inclui dois novos campos contendo a associação e a distância do cluster do centro do cluster designado para esse registro. Os novos nomes de campo são derivados do nome do modelo, prefixados com `$TS-` para a associação de cluster e com `$TSD-` para a distância do centro do cluster. Por exemplo, se o seu modelo for denominado `MDL`, os novos campos serão denominados `$TS-MDL` e `$TSD-MDL`.

Avisos

Estas informações foram desenvolvidas para os produtos e serviços oferecidos nos EUA. Este material pode estar disponível na IBM em outros idiomas. No entanto, pode ser necessário possuir uma cópia do produto ou da versão do produto no mesmo idioma para acessá-lo.

É possível que a IBM não ofereça os produtos, serviços ou recursos discutidos nesta publicação em outros países. Consulte um representante IBM local para obter informações sobre produtos e serviços disponíveis atualmente em sua área. Qualquer referência a produtos, programas ou serviços IBM não significa que apenas produtos, programas ou serviços IBM possam ser utilizados. Qualquer outro produto, programa ou serviço, funcionalmente equivalente, poderá ser utilizado em substituição daqueles, desde que não infrinja nenhum direito de propriedade intelectual da IBM. Entretanto, a avaliação e verificação da operação de qualquer produto, programa ou serviço não IBM são de responsabilidade do Cliente.

A IBM pode ter patentes ou solicitações de patentes pendentes relativas a assuntos tratados nesta publicação. O fornecimento desta publicação não lhe garante direito algum sobre tais patentes. Pedidos de licença devem ser enviados, por escrito, para:

Gerência de Relações Comerciais e Industriais da IBM Brasil
IBM Corporation
Botafogo
Rio de Janeiro, RJ
EUA

Para pedidos de licença relacionados a informações de Conjunto de Caracteres de Byte Duplo (DBCS), entre em contato com o Departamento de Propriedade Intelectual da IBM em seu país ou envie pedidos de licença, por escrito, para:

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

A INTERNATIONAL BUSINESS MACHINES CORPORATION FORNECE ESTA PUBLICAÇÃO "NO ESTADO EM QUE SE ENCONTRA", SEM GARANTIA DE NENHUM TIPO, SEJA EXPRESSA OU IMPLÍCITA, INCLUINDO, MAS NÃO SE LIMITANDO ÀS GARANTIAS IMPLÍCITAS DE MERCADO OU DE ADEQUAÇÃO A UM DETERMINADO PROPÓSITO. Alguns países não permitem a exclusão de garantias expressas ou implícitas em certas transações; portanto, essa disposição pode não se aplicar ao Cliente.

Essas informações podem conter imprecisões técnicas ou erros tipográficos. São feitas alterações periódicas nas informações aqui contidas; tais alterações serão incorporadas em futuras edições desta publicação. A IBM pode, a qualquer momento, aperfeiçoar e/ou alterar os produtos e/ou programas descritos nesta publicação, sem aviso prévio.

Referências nestas informações a Web sites não IBM são fornecidas apenas por conveniência e não representam de forma alguma um endosso a esses websites. Os materiais contidos nestes documentos ou Web sites não fazem parte dos materiais deste produto IBM e a utilização destes documentos ou Web sites é de inteira responsabilidade do Cliente.

A IBM pode usar ou distribuir quaisquer informações fornecidas da forma que julgar apropriada sem incorrer em qualquer obrigação para com o Cliente.

Licenciados deste programa que desejam obter informações sobre este assunto com objetivo de permitir: (i) a troca de informações entre programas criados independentemente e outros programas (incluindo este) e (ii) a utilização mútua das informações trocadas, devem entrar em contato com:

Gerência de Relações Comerciais e Industriais da IBM Brasil
IBM Corporation

Botafogo
Rio de Janeiro, RJ
EUA

Tais informações podem estar disponíveis, sujeitas a termos e condições apropriadas, incluindo em alguns casos o pagamento de uma taxa.

O programa licenciado descrito nesta publicação e todo o material licenciado disponível são fornecidos pela IBM sob os termos do Contrato com o Cliente IBM, do Contrato Internacional de Licença do Programa IBM ou de qualquer outro contrato equivalente.

Os exemplos de clientes e dados de desempenho citados são apresentados com propósitos meramente ilustrativos. Os resultados reais de desempenho podem variar, dependendo das configurações e condições operacionais específicas.

As informações relativas a produtos não IBM foram obtidas junto aos fornecedores dos respectivos produtos, de seus anúncios publicados ou de outras fontes disponíveis publicamente. A IBM não testou estes produtos e não pode confirmar a precisão de seu desempenho, compatibilidade nem qualquer outra reivindicação relacionada a produtos não IBM. Dúvidas sobre os recursos de produtos não IBM devem ser encaminhadas diretamente a seus fornecedores.

As declarações relacionadas aos objetivos e intenções futuras da IBM estão sujeitas a alterações ou cancelamento sem aviso prévio e representam apenas metas e objetivos.

Estas informações contêm exemplos de dados e relatórios utilizados nas operações diárias de negócios. Para ilustrá-los da forma mais completa possível, os exemplos podem incluir nomes de indivíduos, empresas, marcas e produtos. Todos esses nomes são fictícios e qualquer semelhança com pessoas ou empresas reais é mera coincidência.

Marcas comerciais

IBM, o logotipo IBM e ibm.com são marcas comerciais ou marcas registradas da International Business Machines Corp., registradas em várias jurisdições no mundo todo. Outros nomes de produto e de serviço podem ser marcas registradas da IBM ou de outras empresas. Uma lista atual de marcas registradas IBM está disponível na web em"Informações de copyright e marca registrada" em www.ibm.com/legal/copytrade.shtml.

Adobe, o logotipo Adobe, PostScript e o logotipo PostScript são marcas registradas ou marcas comerciais da Adobe Systems Incorporated nos Estados Unidos e/ou em outros países.

Intel, o logotipo Intel, Intel Inside, o logotipo Intel Inside, Intel Centrino, o logotipo do Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium são marcas comerciais ou marcas registradas da Intel Corporation ou suas subsidiárias nos Estados Unidos e em outros países.

Linux é uma marca registrada da Linus Torvalds nos Estados Unidos e/ou em outros países.

Microsoft, Windows, Windows NT e o logotipo Windows são marcas comerciais da Microsoft Corporation nos Estados Unidos e/ou em outros países.

UNIX é uma marca registrada do The Open Group nos Estados Unidos e/ou em outros países.

Java e todas as marcas comerciais e logotipos baseados em Java são marcas comerciais ou marcas registradas da Oracle e/ou de suas afiliadas.

Termos e condições para documentação do produto

As permissões para uso destas publicações são concedidas sujeitas aos seguintes termos e condições.

Aplicabilidade

Estes termos e condições estão em adição aos termos de uso para o website da IBM.

Uso pessoal

É possível reproduzir estas publicações para seu uso pessoal não comercial, desde que todos os avisos do proprietário sejam preservados. Você pode não distribuir, exibir ou fazer trabalhos derivados dessas publicações, ou de qualquer parte delas, sem o consentimento expresso da IBM.

Uso Comercial

O Cliente pode reproduzir, distribuir e exibir estas publicações unicamente dentro da empresa, desde que todos os avisos do proprietário sejam preservados. Você pode não fazer trabalhos derivados dessas publicações, ou reproduzir, distribuir ou exibir estas publicações ou qualquer parte delas fora de sua empresa, sem o consentimento expresso da IBM.

Direitos

Exceto quando expressamente concedido nesta permissão, nenhuma outra permissão, licença ou direito é concedido, seja de maneira expressa ou implícita, para as publicações ou quaisquer informações, dados, software ou outras propriedades intelectuais aqui contidas.

A IBM reserva-se o direito de retirar as permissões concedidas aqui sempre que, a seu critério, o uso das publicações seja prejudicial a seus interesses ou, conforme determinado pela IBM, as instruções acima não estejam sendo seguidas corretamente.

O Cliente não pode fazer download, exportar ou re-exportar estas informações, exceto se estiver em conformidade total com todas as leis e regulamentos aplicáveis, incluindo todas as leis e regulamentos de exportação dos Estados Unidos.

A IBM NÃO FAZ QUALQUER TIPO DE GARANTIA QUANTO AO CONTEÚDO DESTAS PUBLICAÇÕES. AS PUBLICAÇÕES SÃO FORNECIDAS "NO ESTADO EM QUE SE ENCONTRAM" E SEM GARANTIA DE NENHUM TIPO, SEJA EXPRESSA OU IMPLÍCITA, INCLUINDO, MAS A ELAS NÃO SE LIMITANDO, AS GARANTIAS IMPLÍCITAS DE COMERCIALIZAÇÃO, NÃO INFRAÇÃO E ADEQUAÇÃO A UM DETERMINADO PROPÓSITO.

Índice remissivo

Caracteres Especiais

A

B

C

D

E

F

G

I

K

L

M

N

O

P

R

S

T

V