IBM SPSS Data Preparation 29

Nota

Antes de usar estas informações e o produto suportado por elas, leia as informações nos “Avisos” na página 25.

Informações sobre o produto

Essa edição se aplica à versão 29, liberação 0, modificação 1 de IBM® SPSS Statistics e a todas as liberações e modificações subsequentes até que seja indicado de outra forma em novas edições.

Índice

Capítulo 1. Preparação de dados

Os recursos de preparação de dados a seguir são incluídos na Edição Base.

Introdução à Preparação de Dados

À medida que aumenta a força dos sistemas de computação, os interesses por informações crescem proporcionalmente, conduzindo a mais e mais coletas de dados, mais casos, mais variáveis e mais erros de entrada de dados. Esses erros são o flagelo das previsões de modelo preditivo que são o objetivo final de data warehousing, portanto, é necessário manter os dados "limpos". No entanto, a quantidade de dados armazenados cresceu além da capacidade de verificar os casos manualmente, o que é vital para implementar processos automatizados para validação de dados.

A preparação de dados permite identificar casos incomuns e casos inválidos, variáveis e valores de dados em seu conjunto de dados ativo, além de preparar dados para a modelagem.

Uso dos procedimentos de preparação de dados

Seu uso de procedimentos de preparação de dados depende de suas necessidades específicas. Uma rota típica, após carregar seus dados, é:

- **Preparação de metadados.** Revise as variáveis em seu arquivo de dados e determine seus valores, rótulos e níveis de medição válidos. Identifique combinações de valores da variável que são impossíveis, mas comumente codificados incorretamente. Defina regras de validação com base nestas informações. Essa pode ser uma tarefa demorada, mas vale o esforço se você precisa validar arquivos de dados com atributos semelhantes regularmente.
- **Validação de dados.** Execute verificações básicas e verificações em regras de validação definidas para identificar casos, variáveis e valores de dados inválidos. Quando forem encontrados dados inválidos, investigue e corrija a causa. Isso pode requerer outro passo por meio da preparação de metadados.
- **Preparação de modelo.** Use a preparação de dados automatizada para obter transformações dos campos originais que irão melhorar a construção de modelo. Identifique potenciais valores discrepantes estatísticos que possam causar problemas para vários modelos preditivos. Alguns valores discrepantes são o resultado de valores da variável inválidos que não foram identificados. Isso pode requerer outro passo por meio da preparação de metadados.

Quando seu arquivo de dados estiver "limpo", você estará pronto para construir modelos por meio de outros módulos complementares.

Regras de validação

Uma regra é usada para determinar se um caso é válido. Há dois tipos de regras de validação:

- **Regras de variável única.** As regras de variável única consistem em um conjunto fixo de verificações que se aplicam a uma única variável, como verificações de valores fora do intervalo. Para regras de variável única, os valores válidos podem ser expressos como um intervalo de valores ou uma lista de valores aceitáveis.
- **Regras de variável cruzada.** Regras de variável cruzada são regras definidas pelo usuário que podem ser aplicadas a uma única variável ou a uma combinação de variáveis. As regras de variável cruzada são definidas por uma expressão lógica que sinaliza valores inválidos.

As regras de validação são salvas no dicionário de dados de seu arquivo de dados. Isso permite especificar uma regra uma vez e, em seguida, reutilizá-la.

Carregar regras de validação predefinidas

É possível obter rapidamente um conjunto de regras de validação prontas para uso, carregando regras predefinidas de um arquivo de dados externo incluído na instalação.

Para carregar regras de validação predefinidas

1. Nos menus, escolha:

   Dados > **Validação** > **Carregar regras predefinidas...**

Como alternativa, é possível usar o assistente Copiar propriedades de dados para carregar regras de qualquer arquivo de dados.

Definir regras de validação

A caixa de diálogo Definir regras de validação permite criar e visualizar regras de validação de variável única e de variável cruzada.

Para criar e visualizar regras de validação

1. Nos menus, escolha:

   Dados > **Validação** > **Definir regras...**

   A caixa de diálogo é preenchida com regras de validação de variável única e de variável cruzada lidas do dicionário de dados. Quando não houver regras, uma nova regra de item temporário que pode ser modificada de acordo com seus propósitos será criada automaticamente.
2. Selecione regras individuais nas guias Regras de variável única e Regras de variável cruzada para visualizar e modificar suas propriedades.

Definir regras de variável única

A guia Regras de variável única permite criar, visualizar e modificar regras de validação de variável única.

Regras. A lista mostra regras de validação de variável única por nome e o tipo de variável à qual a regra pode ser aplicada. Quando a caixa de diálogo for aberta, ela mostrará regras definidas no dicionário de dados ou, se nenhuma regra estiver definida atualmente, uma regra de item temporário chamada "Regra de variável única 1". Os seguintes botões aparecem abaixo da lista Regras:

- **Novo.** Inclui uma nova entrada na parte inferior da lista Regras. A regra é selecionada e designada ao nome "SingleVarRule *n*," em que *n* é um número inteiro, de forma que o nome da nova regra seja exclusivo entre regras de variável única e de variável cruzada.
- **Duplicar.** Inclui uma cópia da regra selecionada na parte inferior da lista Regras. O nome da regra é ajustado para que seja exclusivo entre regras de variável única e de variável cruzada. Por exemplo, se você duplicar "SingleVarRule 1", o nome da primeira regra duplicada será "Cópia de SingleVarRule 1", o da segunda será "Cópia (2) de SingleVarRule 1" e assim por diante.
- **Excluir.** Exclui a regra selecionada.

Definição de regra. Esses controles permitem visualizar e configurar propriedades para uma regra selecionada.

- **Nome.** O nome da regra deve ser exclusivo entre regras de variável única e de variável cruzada.
- **Tipo.** Este é o tipo de variável à qual a regra pode ser aplicada. Selecione entre **Numérico**, **Sequência de caracteres** e **Data**.
- **Formato.** Isso permite selecionar o formato de data para regras que podem ser aplicadas a variáveis de data.
- **Valores válidos.** É possível especificar os valores válidos como um intervalo ou uma lista de valores.

  Definição de Intervalo

  Os controles de definição de intervalo permitem especificar um intervalo válido. Os valores fora do intervalo são sinalizados como inválidos.

Para especificar um intervalo, insira os valores mínimo ou máximo, ou ambos. Os controles da caixa de seleção permitem sinalizar valores não rotulados e de número não inteiro no intervalo.

Definição de lista

Os controles da definição de lista permitem definir uma lista de valores válidos. Os valores não incluídos na lista são sinalizados como inválidos.

Insira valores de lista na grade. A caixa de seleção determina se o campo é importante quando valores de dados de sequência de caracteres são verificados na lista de valores aceitáveis.

- **Permitir valores omissos de usuário.** Controla se os valores omissos do usuário são sinalizados como inválidos.
- **Permitir valores omissos do sistema.** Controla se os valores omissos do sistema são sinalizados como inválidos. Isso não se aplica aos tipos de regra de sequência de caracteres.
- **Permitir valores em branco.** Controla se os valores da sequência de caracteres em branco (ou seja, completamente vazios) são sinalizados como inválidos. Isso não se aplica a tipos de regras sem sequência de caracteres.

Definir regras de variável cruzada

A guia Regras de variável cruzada permite criar, visualizar e modificar regras de validação de variável cruzada.

Regras. A lista mostra regras de validação de variável cruzada por nome. Quando a caixa de diálogo é aberta, ela mostra uma regra de item temporário chamada"CrossVarRule 1". Os seguintes botões aparecem abaixo da lista Regras:

- **Novo.** Inclui uma nova entrada na parte inferior da lista Regras. A regra é selecionada e designada ao nome "CrossVarRule *n*," em que *n* é um número inteiro, de forma que o nome da nova regra seja exclusivo entre regras de variável única e de variável cruzada.
- **Duplicar.** Inclui uma cópia da regra selecionada na parte inferior da lista Regras. O nome da regra é ajustado para que seja exclusivo entre regras de variável única e de variável cruzada. Por exemplo, se você duplicar "CrossVarRule 1", o nome da primeira regra duplicada será "Cópia de CrossVarRule 1", o da segunda será "Cópia (2) de CrossVarRule1" e assim por diante.
- **Excluir.** Exclui a regra selecionada.

Definição de regra. Esses controles permitem visualizar e configurar propriedades para uma regra selecionada.

- **Nome.** O nome da regra deve ser exclusivo entre regras de variável única e de variável cruzada.
- **Expressão lógica.** Basicamente, essa é a definição de regra. Deve-se codificar a expressão para que os casos inválidos sejam avaliados como 1.

Construindo expressões

1. Para construir uma expressão, cole componentes no campo Expressão ou digite diretamente no campo Expressão.

- É possível colar funções ou variáveis do sistema comumente usadas selecionando um grupo da lista de grupos Função e dando um clique duplo na função ou variável na lista Funções e variáveis especiais (ou selecione a função ou variável e clique em **Inserir**). Insira valores para quaisquer parâmetros indicados por pontos de interrogação (aplica-se somente a funções). O grupo de funções chamado **Todos** fornece uma lista de todas as funções e variáveis do sistema disponíveis. Uma breve descrição da função ou da variável selecionada atualmente é exibida em uma área reservada na caixa de diálogo.
- Constantes da sequência de caracteres devem ser colocadas entre aspas ou apóstrofos.
- Se valores contiverem decimais, um ponto (.) deverá ser usado como o indicador decimal.

Validar dados

A caixa de diálogo Validar dados permite identificar casos, variáveis e valores de dados suspeitos e inválidos no conjunto de dados ativo.

Exemplo. Uma analista de dados deve fornecer um relatório mensal de satisfação do cliente para seu cliente. Os dados que ela recebe todos os meses precisam passar por uma verificação de qualidade em busca de IDs do cliente incompletos, valores de variáveis que estão fora do intervalo e combinações de valores de variáveis que são comumente inseridos com erro. A caixa de diálogo Validar dados permite que a analista especifique as variáveis que identificam exclusivamente os clientes, defina regras de variável única para os intervalos de variáveis válidos e defina regras de variável cruzada para capturar combinações impossíveis. O procedimento retorna um relatório dos casos e variáveis com problemas. Além disso, os dados possuem os mesmos elementos de dados todos os meses, portanto, a analista pode aplicar as regras ao novo arquivo de dados no mês seguinte.

Estatísticas. O procedimento produz listas de variáveis, casos e valores de dados que falham em várias verificações, contagens de violações de regras de variável única e de variável cruzada e sumarizações descritivas simples de variáveis de análise.

Ponderações. O procedimento ignora a especificação da variável de ponderação e, em vez disso, trata-a como qualquer outra variável de análise.

Para validar dados

1. Nos menus, escolha:

   Dados > **Validação** > **Validar dados...**
2. Selecione uma ou mais variáveis de análise para validação por verificações básicas de variáveis ou por regras de validação de variável única.

   Como alternativa, é possível:
3. Clique na guia **Regras de variável cruzada** e aplique uma ou mais regras de variável cruzada.

Opcionalmente, é possível:

- Selecione uma ou mais variáveis de identificação de caso para verificar se há IDs duplicados ou incompletos. As variáveis de ID do caso também são usadas para rotular a saída casewise. Se duas ou mais variáveis de ID do caso forem especificadas, a combinação de seus valores será tratada como um identificador de caso.

Campos com nível de medição desconhecido

O alerta de Nível de Medição é exibido quando o nível de medição para uma ou mais variáveis (campos) no conjunto de dados é desconhecido. Como o nível de medição afeta o cálculo de resultados para este procedimento, todas as variáveis devem ter um nível de medição definido.

Dados de varredura. Lê os dados no conjunto de dados ativo e designa o nível de medição padrão para quaisquer campos com um nível de medição desconhecido atualmente. Se o conjunto de dados for grande, isso poderá demorar algum tempo.

Designar Manualmente. Abre um diálogo que lista todos os campos com um nível de medição desconhecido. É possível utilizar este diálogo para designar o nível de medição para esses campos. Também é possível designar o nível de medição na Visualização de Variável do Editor de Dados.

Como o nível de medição é importante para este procedimento, não é possível acessar o diálogo para executar este procedimento até que todos os campos possuam um nível de medição definido.

Validar verificações básicas de dados

A guia Verificações básicas permite selecionar verificações básicas para variáveis de análise, identificadores de casos e casos inteiros.

Variáveis de análise. Se você selecionou quaisquer variáveis de análise na guia Variáveis, será possível selecionar qualquer uma das verificações de validade a seguir. A caixa de seleção permite ativar ou desativar as verificações.

- **Porcentagem máxima de valores omissos.** Relata variáveis de análise com uma porcentagem de valores omissos maior do que o valor especificado. O valor especificado deve ser um número positivo menor ou igual a 100.
- **Porcentagem máxima de casos em uma única categoria.** Se quaisquer variáveis de análise forem categóricas, essa opção relatará variáveis de análise categóricas com uma porcentagem de casos representando uma única categoria não omissa maior que o valor especificado. O valor especificado deve ser um número positivo menor ou igual a 100. A porcentagem é baseada em casos com valores não omissos da variável.
- **Porcentagem máxima de categorias com contagem de 1.** Se quaisquer variáveis de análise forem categóricas, essa opção relatará variáveis de análise categóricas nas quais a porcentagem de categorias da variável contendo apenas um caso é maior que o valor especificado. O valor especificado deve ser um número positivo menor ou igual a 100.
- **Coeficiente mínimo de variação.** Se quaisquer variáveis de análise forem de escala, essa opção relatará variáveis de análise de escala nas quais o valor absoluto do coeficiente de variação é menor que o valor especificado. Essa opção se aplica apenas a variáveis nas quais a média é diferente de zero. O valor especificado deve ser um número não negativo. Especificar 0 desativa a verificação do coeficiente de variação.
- **Desvio padrão mínimo.** Se quaisquer variáveis de análise forem de escala, essa opção relatará variáveis de análise de escala cujo desvio padrão é menor que o valor especificado. O valor especificado deve ser um número não negativo. Especificar 0 desativa a verificação de desvio padrão.

Identificadores de casos. Se você selecionou quaisquer variáveis de identificador de caso na guia Variáveis, será possível selecionar qualquer uma das verificações de validade a seguir.

- **Sinalizar IDs incompletos.** Esta opção relata casos com identificadores de casos incompletos. Para um caso específico, um identificador será considerado incompleto se o valor de qualquer variável de ID estiver em branco ou omisso.
- **Sinalizar IDs duplicados.** Esta opção relata casos com identificadores de casos duplicados. Os identificadores incompletos são excluídos do conjunto de possíveis duplicatas.

Sinalizar casos vazios. Esta opção relata casos em que todas as variáveis estão vazias ou em branco. Para o propósito de identificar casos vazios, é possível optar por usar todas as variáveis no arquivo (exceto quaisquer variáveis de ID) ou apenas variáveis de análise definidas na guia Variáveis.

Validar Dados - Regras de Variável Única

A guia Regras de variável única exibe regras de validação de variável única disponíveis e permite aplicá-las a variáveis de análise. Para definir regras de variável única, clique em **Definir regras**. Consulte o tópico "Definir regras de variável única" na página 2 para obter mais informações

Variáveis de análise. A lista mostra variáveis de análise, resume suas distribuições e mostra o número de regras aplicadas a cada variável. Observe que os valores omissos do usuário e do sistema não são incluídos nas sumarizações. A lista suspensa Exibição controla quais variáveis são mostradas; é possível escolher entre **Todas as variáveis**, **Variáveis numéricas**, **Variáveis de sequência de caracteres** e **Variáveis de data**.

Regras. Para aplicar regras às variáveis de análise, selecione uma ou mais variáveis e verifique todas as regras que você deseja aplicar na lista Regras. A lista Regras mostra somente regras que são apropriadas para as variáveis de análise selecionadas. Por exemplo, se variáveis de análise numéricas forem selecionadas, somente as regras numéricas serão mostradas, se uma variável de sequência de caracteres for selecionada, somente as regras de sequência de caracteres serão mostradas. Se nenhuma variável de análise for selecionada ou se elas tiverem tipos de dados mistos, nenhuma regra será mostrada.

Distribuições de variáveis. As sumarizações de distribuição mostradas na lista Variáveis de análise podem ser baseadas em todos os casos ou em uma varredura dos primeiros *n* casos, conforme especificado na caixa de texto Casos. Clicar em **Varrer novamente** atualiza as sumarizações de distribuição.

Validar dados - Regras de variável cruzada

A guia Regras de variável cruzada exibe as regras de variável cruzada disponíveis e permite aplicá-las a seus dados. Para definir regras de variável cruzada adicionais, clique em **Definir regras**. Consulte o tópico "Definir regras de variável cruzada" na página 3 para obter mais informações

Validar saída de dados

Relatório casewise. Se você aplicou quaisquer regras de validação de variável única ou de variável cruzada, será possível solicitar um relatório que lista as violações da regra de validação para casos individuais.

- **Número mínimo de violações.** Essa opção especifica o número mínimo de violações de regras necessário para que um caso seja incluído no relatório. Especifique um número inteiro positivo.
- **Número máximo de casos.** Essa opção especifica o número máximo de casos incluídos no relatório de caso. Especifique um número inteiro positivo menor ou igual a 1000.

Regras de validação de variável única. Se você aplicou quaisquer regras de validação de variável única, será possível escolher como exibir os resultados ou se todos eles devem ser exibidos.

- **Sumarizar violações por variável de análise.** Para cada variável de análise, essa opção mostra todas as regras de validação de variável única que foram violadas e o número de valores que violaram cada regra. Ela também relata o número total de violações de regra de variável única para cada variável.
- **Sumarizar violações por regra.** Para cada regra de validação de variável única, essa opção relata variáveis que violaram a regra e o número de valores inválidos por variável. Ela também relata o número total de valores que violaram cada regra nas variáveis.

Exibir estatísticas descritivas para variáveis de análise. Essa opção permite solicitar estatísticas descritivas para variáveis de análise. Uma tabela de frequências é gerada para cada variável categórica. Uma tabela de estatísticas de sumarização, incluindo a média, o desvio padrão, o mínimo e o máximo é gerada para as variáveis de escala.

Mover casos com violações de regras de validação para a parte superior do conjunto de dados ativo. Essa opção move casos com violações de regras de variável única ou de variável cruzada para a parte superior do conjunto de dados ativo para fácil leitura.

Validar dados - Salvar

A guia Salvar permite salvar variáveis que registram violações de regras no conjunto de dados ativo.

Variáveis de sumarização. Estas são as variáveis individuais que podem ser salvas. Marque uma caixa para salvar a variável. São fornecidos nomes padrão para as variáveis, é possível editá-los.

- **Indicador de caso vazio.** Aos casos vazios, é atribuído o valor 1. Todos os outros casos são codificados como 0. Os valores da variável refletem o escopo especificado na guia Verificações básicas.
- **Grupo de ID duplicado.** Casos que possuem o mesmo identificador de caso (diferente de casos com identificadores incompletos) têm o mesmo número de grupo designado. Casos com identificadores exclusivos ou incompletos são codificados como 0.
- **Indicador de ID incompleto.** Os casos com identificadores de caso vazios ou incompletos têm o valor 1 designado. Todos os outros casos são codificados como 0.
- **Violações de regra de validação.** Essa é a contagem casewise total de violações de regra de validação de variável única e de variável cruzada.

Substituir variáveis de sumarização existentes. As variáveis salvas no arquivo de dados devem ter nomes exclusivos ou variáveis de substituição com o mesmo nome.

Salvar variáveis indicadoras. Esta opção permite salvar um registro completo de violações de regra de validação. Cada variável corresponde a uma aplicação de uma regra de validação e tem um valor 1, se o caso viola a regra, e um valor 0, se ele não viola.

Preparação de dados automatizada

A preparação de dados para análise é uma das etapas mais importantes em qualquer projeto e, tradicionalmente, uma das mais demoradas. A Automated Data Preparation (ADP) manipula a tarefa para você, analisa seus dados e identifica correções, realiza uma triagem dos campos que são problemáticos ou que provavelmente não são úteis, deriva novos atributos quando apropriado e melhora o desempenho por meio de técnicas de triagem inteligentes. É possível utilizar o algoritmo de forma totalmente **automática**, permitindo que ele escolha e aplique as correções, ou é possível utilizá-lo de forma **interativa**, visualizando as mudanças antes que elas sejam feitas e aceitando-as ou rejeitando-as conforme desejado.

Usando a ADP, é possível preparar seus dados para desenvolvimento de modelo de forma rápida e fácil, sem a necessidade de conhecimento prévio dos conceitos estatísticos envolvidos. Os modelos tendem a criar e escorar mais rapidamente e, além disso, o uso da ADP melhora a robustez dos processos de modelagem automatizados.

Nota: quando a ADP prepara um campo para análise, ela cria um novo campo contendo os ajustes ou as transformações, ao invés de substituir os valores e propriedades existentes do antigo campo. O campo antigo não é usado em análise adicional, sua função é configurada como Nenhum. Observe também que qualquer informação de valor omisso do usuário não é transferida para esses campos recém-criados e quaisquer valores omissos no novo campo são omissos do sistema.

Exemplo. Uma empresa de seguros com recursos limitados para investigar solicitações de seguro de um proprietário de residência deseja construir um modelo para sinalizar solicitações suspeitas potencialmente fraudulentas. Antes de construir o modelo, eles irão deixar os dados prontos para modelagem usando a preparação de dados automatizada. Como eles desejam ser capazes de revisar as transformações propostas antes que as transformações sejam aplicadas, eles usarão a preparação de dados automatizada no modo interativo.

Um grupo da indústria automotiva mantém o controle das vendas de uma variedade de veículos automotivos pessoais. Em um esforço para que sejam capazes de identificar modelos de desempenho superior e inferior, eles querem estabelecer um relacionamento entre vendas de veículo e características do veículo. Eles usarão a preparação de dados automatizada para preparar os dados para análise e construir modelos usando os dados de "antes" e "depois" da preparação para ver como os resultados diferem.

Qual é o seu objetivo? A preparação de dados automatizada recomenda passos de preparação de dados que afetarão a velocidade com que outros algoritmos podem construir modelos e melhorar o poder preditivo desses modelos. Isso pode incluir a transformação, construção e seleção de variáveis. A resposta também pode ser transformada. É possível especificar as prioridades de construção de modelo nas quais o processo de preparação de dados deve se concentrar.

- **Balancear velocidade e precisão.** Essa opção prepara os dados para dar igual prioridade à velocidade com a qual os dados são processados por algoritmos de construção de modelo e à precisão das predições.
- **Otimizar para velocidade.** Essa opção prepara os dados para dar prioridade à velocidade com a qual os dados são processados por algoritmos de construção de modelo. Selecione essa opção quando estiver trabalhando com conjuntos de dados muito grandes ou procurando por uma resposta rápida.
- **Otimizar para precisão.** Essa opção prepara os dados para dar prioridade à precisão de predições produzidas por algoritmos de construção de modelo.
- **Análise customizada.** Selecione essa opção quando desejar mudar manualmente o algoritmo na guia Configurações. Observe que essa configuração será selecionada automaticamente se mudanças subsequentes incompatíveis com um dos outros objetivos forem feitas nas opções da guia Configurações.

Para Obter Preparação de Dado Automático

Nos menus, escolha:

1. Nos menus, escolha:

   Transformar > **Preparar Dados para Modelagem** > **Automática...**
2. Clique em **Executar** .

Opcionalmente, é possível:

- Especificar um objetivo na guia Objetivo.
- Especificar designações de campo na guia Campos.
- Especificar configurações especialistas na guia Configurações.

Para Obter Preparação de Dados Interativa

1. Nos menus, escolha:

   Transformar > **Preparar Dados para Modelagem** > **Interativa...**
2. Clique em **Analisar** na barra de ferramentas na parte superior do diálogo.
3. Clique na guia Análise e revise as etapas de preparação de dados sugeridas.
4. Se estiver satisfeito, clique em **Executar**. Caso contrário, clique em **Limpar Análise**, mude qualquer configuração que desejar e clique em **Analisar**.

Opcionalmente, é possível:

- Especificar um objetivo na guia Objetivo.
- Especificar designações de campo na guia Campos.
- Especificar configurações especialistas na guia Configurações.
- Salvar as etapas de preparação de dados sugerida em um arquivo XML, clicando em **Salvar XML**.

Guia Campos

A guia Campos especifica quais campos devem ser preparados para análise adicional.

Usar papéis predefinidos. Esta opção usa informações de campo existentes. Se houver um campo único com uma função como um Destino, ele será usado como o destino, caso contrário, não haverá nenhum destino. Todos os campos com uma função predefinida como Entrada serão usados como entradas. Pelo menos um campo de entrada é necessário. para obter mais informações.

Usar designações de campo customizado. Ao substituir funções de campo, movendo os campos de suas listas padrão, o diálogo alterna automaticamente para esta opção. Ao fazer designações de campo customizado, especifique os campos a seguir:

- **Destino (opcional).** Se você planeja construir modelos que requerem um destino, selecione o campo de destino. Isso é semelhante a configurar a função do campo como Destino.
- **Entradas.** Selecione um ou mais campos de entrada. Isso é semelhante a configurar a função do campo como Entrada.

Guia Configurações

A guia Configurações abrange vários grupos diferentes de configurações que podem ser modificadas para ajustar com precisão como o algoritmo processa os seus dados. Se você fizer qualquer mudança nas configurações padrão que for incompatível com os outros objetivos, a guia Objetivo será atualizada automaticamente para selecionar a opção **Customizar análise**.

Preparar datas e horas

Muitos algoritmos de modelagem não conseguem manipular diretamente detalhes de data e hora, essas configurações permitem derivar novos dados de duração que podem ser usados como entradas de modelo de datas e horas em seus dados existentes. Os campos que contêm datas e horas devem ser predefinidos com tipos de armazenamento de data ou hora. Os campos originais de data e hora não serão recomendados como entrada de modelo seguindo a preparação de dados automatizada.

Preparar datas e horas para modelagem. Cancelar a seleção desta opção desativa todos os outros controles de Preparar Datas e Horas enquanto mantém as seleções.

Calcular o tempo decorrido até a data de referência. Isso produz o número de anos/meses/dias desde uma data de referência para cada variável que contém datas.

- **Data de Referência.** Especifique a data a partir da qual a duração será calculada com relação às informações de data nos dados de entrada. A seleção de **Data de hoje** significa que a data atual do sistema é sempre utilizada quando o ADP é executado. Para utilizar uma data específica, selecione **Data fixa** e insira a data necessária.
- **Unidades para Duração de Data.** Especifique se o ADP deve decidir automaticamente a unidade de duração de data ou selecione a partir de **Unidades fixas** de Anos, Meses ou Dias.

Calcular tempo decorrido até o horário de referência. Isso produz o número de horas/minutos/segundos desde um tempo de referência para cada variável que contém os tempos.

- **Tempo de Referência.** Especifique o tempo a partir do qual a duração será calculada com relação às informações de tempo nos dados de entrada. Selecionar **Horário Atual** significa que o horário do sistema atual é sempre usado quando a ADP é executada. Para usar um horário específico, selecione **Horário fixo** e insira os detalhes necessários.
- **Unidades para Duração de Tempo.** Especifique se o ADP deve decidir automaticamente a unidade de duração de tempo ou selecione a partir de **Unidades fixas** de Horas, Minutos ou Segundos.

Extrair Elementos de Tempo Cíclicos. Utilize essas configurações para dividir um campo de data ou hora único em um ou mais campos. Por exemplo, se você selecionar todas as três caixas de seleção, o campo de data de entrada "23-05-1954" será dividido em três campos: 23, 05 e 1954, cada um usando o sufixo definido no painel de **Nomes de Campo** e o campo de data original será ignorado.

- **Extrair de datas.** Para quaisquer entradas de data, especifique se você deseja extrair anos, meses, dias ou qualquer combinação.
- **Extrair de tempos.** Para quaisquer entradas de horário, especifique se você deseja extrair horas, minutos, segundos ou qualquer combinação.

Excluir Campos

Dados de qualidade ruim podem afetar a precisão de suas predições, portanto, é possível especificar o nível de qualidade aceitável para variáveis de entrada. Todos os campos que são constantes ou que possuem 100% de valores omissos são automaticamente excluídos.

Excluir campos de entrada de qualidade baixa. Cancelar a seleção desta opção desativa todos os outros controles de Excluir Campos enquanto mantém as seleções.

Excluir campos com demasiados valores omissos. Os campos com mais que a porcentagem especificada de valores omissos são removidos da análise adicional. Especifique um valor maior ou igual a 0, que é equivalente a cancelar a seleção dessa opção, e menor ou igual a 100, embora campos com todos valores omissos sejam excluídos automaticamente. O padrão é 50.

Excluir campos nominais com categorias exclusivas em excesso. Os campos nominais que tiverem mais que o número especificado de categorias são removidos da análise adicional. Especifique um número inteiro positivo. O padrão é 100. Isso é útil para remover automaticamente da modelagem os campos que contêm informações exclusivas de registro, como o ID, o endereço ou o nome.

Excluir campos categóricos com valores em excesso em uma única categoria. Campos ordinais e nominais com uma categoria que contém mais que a porcentagem especificada de registros são

removidos da análise adicional. Especifique um valor maior ou igual a 0, equivalente a cancelar a seleção dessa opção, e menor ou igual a 100, embora campos constantes sejam automaticamente excluídos. O padrão é 95.

Medição de ajuste

Ajustar nível de medição. Cancelar a seleção dessa opção desativa todos os outros controles de Ajuste de Medição enquanto mantém as seleções.

Nível de Medição. Especifique se o nível de medição de campos contínuos com "muito poucos" valores pode ser ajustado para ordinal e os campos ordinais com "demasiados" valores podem ser ajustados para contínuos.

- **Número máximo de valores para campos ordinais.** Campos ordinais com mais do que o número especificado de categorias são reformulados como campos contínuos. Especifique um número inteiro positivo. O padrão é 10. Esse valor deve ser maior ou igual ao número mínimo de valores para campos contínuos.
- **Número mínimo de valores para campos contínuos.** Campos contínuos com menos do que o número especificado de valores exclusivos são reformulados como campos ordinais. Especifique um número inteiro positivo. O padrão é 5. Esse valor deve ser menor ou igual ao número máximo de valores para campos ordinais.

Melhorar a qualidade dos dados

Preparar campos para melhorar qualidade de dados. Cancelar a seleção dessa opção desativa todos os outros controles de Melhoria de Qualidade de Dados enquanto mantém as seleções.

Tratamento do Valor Discrepante. Especifique se valores discrepantes devem ser substituídos para as entradas e o destino, em caso afirmativo, especifique um critério de corte do valor discrepante, medido em desvios padrão e um método para substituir valores discrepantes. Os valores discrepantes podem ser substituídos cortando-os (configurado com o valor de corte) ou configurando-os como valores omissos. Quaisquer valores discrepantes configurados como valores omissos seguem as configurações de tratamento de valor omisso selecionadas abaixo.

Substituir Valores Omissos. Especifique se os valores omissos de campos contínuos, nominais ou ordinais devem ser substituídos.

Reordenar Campos Nominais. Selecione isso para recodificar os valores de campos nominais (configurar) da menor categoria (que ocorre com menos frequência) para a maior (que ocorre com mais frequência). Os novos valores de campo iniciam com 0 como a categoria menos frequente. Observe que o novo campo será numérico mesmo se o campo original for uma sequência de caracteres. Por exemplo, se os valores de dados do campo nominal forem "A", "A", "A", "B", "C", "C", a preparação de dados automatizada irá recodificar "B" como 0, "C" como 1 e "A" como 2.

Escalar novamente Campos

Escalar novamente campos. Cancelar a seleção dessa opção desativa todos os outros controles para Escalar novamente campos enquanto mantém as seleções.

Ponderação de Análise. Essa variável contém ponderações de análise (regressão ou amostragem). Ponderações de análise são usadas para explicar as diferenças na variância entre níveis do campo de destino. Selecione um campo contínuo.

Campos de Entrada Contínuos. Isso normalizará campos de entrada contínuos usando uma **transformação de escore z** ou **transformação mín-máx**. Escalar entradas novamente é especialmente útil quando você seleciona **Executar construção de variável** nas configurações Selecionar e Construir.

- **Transformação de escore z.** Usando as estimativas de desvio médio e padrão observadas como parâmetro de preenchimento, os campos são padronizados e, em seguida, os escores *z* são mapeados para os valores correspondentes de uma distribuição normal com o **Desvio médio final** e o **Desvio padrão final** especificados. Especifique um número para **Desvio médio final** e um número positivo

para **Desvio padrão final**. Os padrões são 0 e 1, respectivamente, correspondendo à nova escala padronizada.

- **Transformação mín-máx.** Usando o mínimo e máximo observado como parâmetro de preenchimento, os campos são mapeados para os valores correspondentes de uma distribuição uniforme com o **Mínimo** e **Máximo** especificados. Especifique números com **Máximo** maior do que **Mínimo**.

Variável de Resposta Contínua. This transforms a continuous target using the Box-Cox transformation into a field that has an approximately normal distribution with the specified **Final mean** and **Final standard deviation**. Especifique um número para **Desvio médio final** e um número positivo para **Desvio padrão final**. Os padrões são 0 e 1, respectivamente.

Nota: Se uma resposta foi transformada pela ADP, modelos subsequentes construídos usando a resposta transformada pontuam as unidades transformadas. Para interpretar e usar os resultados, deve-se converter o valor predito novamente para a escala original. Consulte o tópico para obter mais informações. Consulte o tópico para obter mais informações. “Escores de transformação retroativa” na página 19

Transformar Campos

Para melhorar o poder preditivo de seus dados, é possível transformar os campos de entrada.

Transformar campo para modelagem. Cancelar a seleção dessa opção desativa todos os outros controles para Transformar Campos enquanto mantém as seleções.

Campos de Entrada Categóricos As opções a seguir estão disponíveis:

- **Mesclar categorias de dispersão para maximizar a associação com a resposta.** Selecione isso para criar um modelo mais econômico, reduzindo o número de camposa serem processados em associação com o destino. Categorias similares são identificadas com base no relacionamento entre a entrada e o destino. Categorias que não são significativamente diferentes (ou seja, que têm um valor *p* maior do que o valor especificado) são mescladas. Especifique um valor maior que 0 e menor ou igual a 1. Se todas as categorias forem mescladas em uma só, as versões originais e derivadas do campo serão excluídas de análise posterior, porque não têm valor como um preditor.
- **Quando não houver nenhuma resposta, mesclar as categorias de dispersão com base nas contagens.** Se o conjunto de dados não tiver nenhum destino, é possível optar por mesclar categorias de dispersão de campos ordinais e nominais. O método de frequência igual é usado para mesclar categorias com menos do que a porcentagem mínima especificada do número total de registros. Especifique um valor maior ou igual a 0 e menor ou igual a 100. O padrão é 10. A mesclagem para quando não há categorias com menos do que o percentual mínimo especificado de casos ou quando existem apenas duas categorias restantes.

Campos de Entrada Contínuos. Se o conjunto de dados incluir uma variável resposta categórica, será possível categorizar entradas contínuas com associações fortes para melhorar o desempenho de processamento. Categorias são criadas com base nas propriedades de "subconjuntos homogêneos", que são identificadas pelo método de Scheffe usando o valor *p* especificado como o alpha para o valor crítico para determinar subconjuntos homogêneos . Especifique um valor maior que 0 e menor ou igual a 1. O padrão é 0,05. Se a operação de categorização resultar em uma categoria única para um campo específico, as versões originais e categorizadas do campo serão excluídas, porque elas não possuirão nenhum valor como um preditor.

Nota: A categorização em ADP difere da categorização ideal. A categorização ideal usa informações de entropia para converter um campo contínuo em um campo categórico, isso precisa classificar dados e armazenar tudo na memória. A ADP usa subconjuntos homogêneos para categorizar um campo contínuo, o que significa que a categorização de ADP não precisa classificar dados e não armazena todos os dados na memória. O uso do método de subconjunto homogêneo para categorizar um campo contínuo significa que o número de categorias após a categorização é sempre menor ou igual ao número de categorias no destino.

Selecione e construa

Para melhorar o poder preditivo de seus dados, é possível construir novos campos baseados nos campos existentes.

Executar seleção de variável. Uma entrada contínua é removida da análise se o valor *p* para a sua correlação com a resposta for maior que o valor *p* especificado.

Executar construção de variável. Selecione essa opção para derivar novas variáveis de uma combinação de várias variáveis existentes. As variáveis antigas não são usadas em análise posterior. Esta opção se aplica somente às variáveis de entrada contínua em que o destino seja contínuo, ou onde não houver nenhum destino.

Nomes do Campo

Para identificar facilmente recursos novos e transformados, a ADP cria e aplica os novos nomes, prefixos ou sufixos básicos. É possível corrigir esses nomes para serem mais relevantes para suas próprias necessidades e dados.

Campos Transformados e Construídos. Especifique as extensões de nome a serem aplicadas aos campos de destino e de entrada transformados.

Além disso, especifique o nome do prefixo a ser aplicado em quaisquer variáveis construídas por meio das configurações Selecionar e Construir. O novo nome é criado, anexando um sufixo numérico nesse nome raiz de prefixo. O formato do número depende de quantas novas variáveis são derivadas, por exemplo:

- 1 a 9 variáveis construídas serão denominadas: variável1 a variável9.
- 10 a 99 variáveis construídas serão denominadas: variável01 a variável99.
- 100 a 999 variáveis construídas serão denominadas: variável001 a variável999 e assim por diante.

Isso assegura que as variáveis construídas sejam classificadas em uma ordem sensível, não importa quantas houver.

Durações Calculadas a partir de Datas e Horas. Especifique as extensões de nome a serem aplicadas às durações calculadas a partir das datas e horas.

Elementos Cíclicos Extraídos a partir de Datas e Horas. Especifique as extensões de nome a serem aplicadas aos elementos cíclicos extraídos das datas e horas.

Aplicando e Salvando Transformações

Dependendo se você está usando os diálogos de preparação de dados interativos ou automáticos, as configurações para aplicar e salvar transformações são ligeiramente diferentes.

Preparação de Dados Interativos Aplicar Configurações de Transformações

Dados Transformados. Essas configurações especificam onde salvar os dados transformados.

- **Inclua novos campos no conjunto de dados ativo.** Qualquer campo criado pela preparação de dados automatizados é incluído como novo campo no conjunto de dados ativo. **Atualizar papéis para campos analisados** irá configurar o papel para Nenhum para qualquer campo que for excluído de análise posterior pela preparação de dados automatizados.
- **Crie um novo conjunto de dados ou um arquivo contendo os dados transformados.** Os campos recomendados pela preparação de dados automatizada são incluídos em um novo conjunto de dados ou arquivo. **Incluir campos não analisados** inclui campos no conjunto de dados original que não foram especificados na guia Campos para o novo conjunto de dados. Isso é útil para transferir os campos que contêm informações não usadas na modelagem, como ID, endereço ou nome, para o novo conjunto de dados.

Preparação de Dado Automático Aplicar e Salvar Configurações

O grupo de Dados Transformados é o mesmo que na Preparação de Dados Interativos. Na preparação de dados automáticos, as opções adicionais a seguir estão disponíveis:

Aplicar transformações. Nos diálogos de Preparação de Dado Automático, cancelar a seleção dessa opção desativa todos os outros controles de Aplicação e Salvamento enquanto se mantém as seleções.

Salvar transformações como sintaxe. Isso salva as transformações recomendadas como sintaxe de comando para um arquivo externo. O diálogo de Preparação de Dados Interativos não tem esse controle porque ele colará as transformações como sintaxe de comando na janela de sintaxe se você clicar em **Colar**.

Salvar transformações como XML. Isso salva as transformações recomendadas como XML em um arquivo externo, que pode ser mesclado com modelo PMML usando TMS MERGE ou aplicado a outro conjunto de dados usando TMS IMPORT. O diálogo de Preparação de Dados Interativos não tem esse controle porque ele salvará as transformações como XML se você clicar em **Salvar XML** na barra de ferramentas na parte superior do diálogo.

Guia Análise

Nota: a guia Análise é utilizada no diálogo Preparação de Dados Interativa para permitir revisar as transformações recomendadas. O diálogo Preparação de dados automática não inclui essa etapa.

1. Quando estiver satisfeito com as configurações do ADP, incluindo quaisquer mudanças feitas nas guias Objetivo, Campos e Configurações, clique em **Analisar Dados**; o algoritmo aplica as configurações nas entradas de dados e exibe os resultados na guia Análise.

A guia Análise contém resultado tabular e gráfico que resume o processamento de seus dados e exibe recomendações quanto à maneira como os dados podem ser modificados ou melhorados para escoragem. Em seguida, é possível revisar e aceitar ou rejeitar as recomendações.

A guia Análise é formada por dois painéis, a visualização principal à esquerda e a visualização vinculada ou auxiliar à direita. Há três visualizações principais:

- Sumarização de Processamento de Campo (o padrão). Consulte o tópico “Sumarização de Processamento de Campo ” na página 13 para obter mais informações.
- Campos. Consulte o tópico “Campos ” na página 14 para obter mais informações.
- Sumarização da ação. Consulte o tópico “Sumarização de Ação ” na página 15 para obter mais informações.

Há quatro visualizações vinculadas/auxiliares:

- Poder Preditivo (o padrão). Consulte o tópico “Poder Preditivo ” na página 15 para obter mais informações.
- Tabela de Campos. Consulte o tópico “Tabela de campos ” na página 16 para obter mais informações.
- Detalhes do campo. Consulte o tópico “Detalhes do Campo ” na página 16 para obter mais informações.
- Detalhes da ação. Consulte o tópico “Detalhes da Ação ” na página 17 para obter mais informações.

Vínculos entre visualizações

Dentro da visualização principal, o texto sublinhado nas tabelas controla a exibição na visualização vinculada. Clicar no texto permite obter detalhes sobre um determinado campo, conjunto de campos ou etapa de processamento. O link que você selecionou por último é mostrado em uma cor mais escura, isso ajuda você a identificar a conexão entre os conteúdos dos dois painéis de visualização.

Reconfigurando as visualizações

Para exibir novamente as recomendações de Análise originais e abandonar quaisquer mudanças feitas nas visualizações Análise, clique em **Reconfigurar** na parte inferior do painel de visualização principal.

Sumarização de Processamento de Campo

A tabela Sumarização do processamento de campo fornece uma captura instantânea do impacto geral projetado de processamento, incluindo mudanças no estado das variáveis e no número de variáveis construídas.

Observe que nenhum modelo é realmente construído, portanto, não há uma medida ou um gráfico da mudança no poder preditivo geral antes e após a preparação de dados, como alternativa, é possível exibir gráficos do poder preditivo de preditores individuais recomendados.

A tabela exibe as informações a seguir:

- O número de campos de destino.
- O número de preditores originais (de entrada).
- Os preditores recomendados para uso na análise e na modelagem. Isso inclui o número total de campos recomendados, o número de campos originais não transformados recomendados, o número de campos transformados recomendados (excluindo versões intermediárias de qualquer campo, campos derivados de preditores de data/hora e preditores construídos), o número de campos recomendados que são derivados de campos de data/hora e o número de preditores construídos recomendados.
- O número de preditores de entrada não recomendados para uso em qualquer forma, seja em sua forma original, como um campo derivado, ou como entrada para um preditor construído.

Onde qualquer informação dos **Campos** estiver sublinhada, clique para exibir mais detalhes em uma visualização vinculada. Detalhes da **Resposta**, das **Variáveis de Entrada** e das **Variáveis de entrada não usadas** são mostrados na visualização vinculada de Tabela de Campos. Veja o tópico “Tabela de campos ” na página 16 para obter mais informações. **Recursos recomendados para uso na análise** são exibidos na visualização vinculada Poder preditivo. Consulte o tópico “Poder Preditivo ” na página 15 para obter mais informações

Campos

A visualização principal Campos exibe os campos processados e se a ADP recomenda usá-los em modelos de recebimento de dados. É possível substituir a recomendação para qualquer campo, por exemplo, para excluir variáveis construídas ou incluir variáveis que a ADP recomenda excluir. Se um campo tiver sido transformado, é possível decidir se deseja aceitar a transformação sugerida ou usar a versão original.

A visualização Campos consiste em duas tabelas, uma para o destino e outra para preditores que foram processados ou criados.

Tabela de destino

A tabela de **Destino** é mostrada somente se uma resposta estiver definida nos dados.

A tabela contém duas colunas:

- **Nome.** Esse é o nome ou o rótulo do campo de destino e o nome original é sempre usado, mesmo se o campo tiver sido transformado.
- **Nível de Medição.** Isso exibe o ícone que representa o nível de medição. Passe o mouse sobre o ícone para exibir um rótulo (contínuo, ordinal, nominal e assim por diante) que descreve os dados.

  Se a resposta tiver sido transformada, a coluna **Nível de Medição** refletirá a versão final transformada. *Nota*: não é possível desligar transformações para o destino.

Tabela Preditores

A tabela **Preditores** é sempre mostrada. Cada linha da tabela representa um campo. Por padrão, as linhas são classificadas em ordem decrescente de poder preditivo.

Para variáveis ordinárias, o nome original é sempre usado como o nome da linha. Ambas as versões, original e derivada, de campos de data/hora aparecem na tabela (em linhas separadas), a tabela também inclui preditores construídos.

Observe que as versões transformadas de campos mostrados na tabela sempre representam as versões finais.

Por padrão, somente os campos recomendados são mostrados na tabela Preditores. Para exibir os campos restantes, selecione a caixa **Incluir campos não recomendados na tabela** acima da tabela e, em seguida, esses campos são exibidos na parte inferior da tabela.

A tabela contém as colunas a seguir:

- **Versão para Usar.** Isso exibe uma lista suspensa que controla se um campo será usado no recebimento de dados e se as transformações sugeridas devem ser usadas. Por padrão, a lista suspensa reflete as recomendações.

  Para preditores ordinários que foram transformados, a lista suspensa possui três opções: **Transformado**, **Original** e **Não utilizar**.

  Para preditores ordinários não transformados, as opções são: **Original** e **Não utilizar**.

  Para campos de data/hora derivados e preditores construídos, as opções são: **Transformado** e **Não utilizar**.

  Para campos de data originais, a lista suspensa é desativada e configurada para **Não utilizar**.

  Nota: para preditores com as versões originais e transformadas, a mudança entre as versões **Original** e **Transformado** atualiza automaticamente as configurações de **Nível de Medição** e de **Poder Preditivo** para essas variáveis.
- **Nome.** Cada nome do campo é um link. Clique em um nome para exibir mais informações sobre o campo na visualização vinculada. Consulte o tópico “Detalhes do Campo ” na página 16 para obter mais informações
- **Nível de Medição.** Isso exibe o ícone que representa o tipo de dados, passe o mouse sobre o ícone para exibir um rótulo (contínuo, ordinal, nominal e assim por diante) que descreve os dados.
- **Poder Preditivo.** O poder preditivo é exibido apenas para campos recomendados pela ADP. Essa coluna não será exibida se não houver nenhum destino definido. O poder preditivo varia de 0 a 1, com valores maiores indicando "melhores" preditores. Em geral, o poder preditivo é útil para comparar preditores dentro de uma análise da ADP, mas os valores de poder preditivo não devem ser comparados nas análises.

Sumarização de Ação

Para cada ação tomada pela preparação de dados automatizada, preditores de entrada são transformados e/ou filtrados, os campos que sobrevivem a uma ação são usados na próxima. Os campos que sobrevivem até a última etapa são, então, recomendados para uso na modelagem, enquanto as entradas para preditores transformados e construídos são filtradas.

A Sumarização de Ação é uma tabela simples que lista as ações de processamento tomadas pela ADP. Onde qualquer **Ação** estiver sublinhada, clique para exibir mais detalhes em uma visualização vinculada sobre as ações tomadas. Consulte o tópico “Detalhes da Ação ” na página 17 para obter mais informações

Nota: Somente as versões transformadas originais e finais de cada campo são mostradas, não qualquer versão intermediária que foi usada durante a análise.

Poder Preditivo

Exibido por padrão quando a análise é executada pela primeira vez, ou quando selecionar **Preditores recomendados para uso em análise** na visualização principal de Sumarização de Processamento de Campo, o gráfico exibe o poder preditivo dos preditores recomendados. Os campos são classificados pelo poder preditivo, com o campo com o valor mais alto aparecendo na parte superior.

Para versões transformadas de preditores ordinários, o nome do campo reflete sua opção de sufixo no painel Nomes de Campo da guia Configurações, por exemplo: *_transformed*.

Ícones de nível de medição são exibidos após os nomes de campos individuais.

O poder preditivo de cada preditor recomendado é calculado por meio de uma regressão linear ou modelo naïve Bayes, dependendo se o destino é contínuo ou categórico.

Tabela de campos

Exibida quando você clica em **Resposta**, **Preditores** ou **Preditores não usados** na visualização principal de Sumarização de Processamento, a visualização de Tabela de Campos exibe uma tabela simples que lista as variáveis relevantes.

A tabela contém duas colunas:

- **Nome.** O nome do preditor.

  Para destinos, o nome ou o rótulo original do campo é usado, mesmo que o destino tenha sido transformado.

  Para versões transformadas de preditores ordinários, o nome reflete a sua opção de sufixo no painel de Nomes do Campo da guia Configurações; por exemplo: *_transformed*.

  Para campos derivados de datas e horas, o nome da versão transformada final é usado; por exemplo: *bdate_years*.

  Para preditores construídos, o nome do preditor construído é utilizado, por exemplo: *Predictor1*.
- **Nível de Medição.** Isso exibe o ícone que representa o tipo de dados.

  Para o Destino, o **Nível de medição** sempre reflete a versão transformada (se o destino foi transformado), por exemplo, alterado de ordinal (conjunto ordenado) para contínuo (intervalo, escala), ou vice-versa.

Detalhes do Campo

Exibida quando você clica em qualquer **Nome** na visualização principal de Campos, a visualização de Detalhes de Campo contém valores de distribuição, omissos e gráficos de poder preditivo (se aplicável) para o campo selecionado. Além disso, o histórico de processamento para o campo e o nome do campo transformado também são mostrados (se aplicável).

Para cada conjunto de gráficos, duas versões são mostradas lado a lado para comparar o campo com e sem transformações aplicadas. Se uma versão transformada do campo não existir, um gráfico será mostrado apenas para a versão original. Para campos de data ou hora derivados e preditores construídos, os gráficos são mostrados somente para o novo preditor.

Nota: se um campo for excluído devido a categorias em excesso, somente o histórico de processamento será mostrado.

Gráfico de distribuição

A distribuição de campo contínuo é mostrada como um histograma, com uma curva normal sobreposta e uma linha de referência vertical para o valor médio, e os campos categóricos são exibidos como um gráfico de barras.

Os histogramas são rotulados para mostrar desvio padrão e assimetria, no entanto, a assimetria não será exibida se o número de valores for 2 ou menos ou se a variância do campo original for menor do que de 10 a 20.

Passe o mouse sobre o gráfico para exibir a média para histogramas ou a contagem e a porcentagem do número total de registros para categorias nos gráficos de barras.

Gráfico de valor omisso

Os gráficos de pizza comparam a porcentagem dos valores omissos com e sem as transformações aplicadas e os rótulos do gráfico mostram a porcentagem.

Se a ADP executou a manipulação de valor omisso, o gráfico de pizza pós-transformação também incluirá o valor de substituição como um rótulo, ou seja, o valor usado no lugar dos valores omissos.

Passe o mouse sobre o gráfico para exibir a contagem de valor omisso e a porcentagem do número total de registros.

Gráfico de Poder Preditivo

Para campos recomendados, os gráficos de barras exibem o poder preditivo antes e após a transformação. Se o destino tiver sido transformado, o poder preditivo calculado será referente ao destino transformado.

Nota: os gráficos de poder preditivo não serão mostrados se nenhum destino estiver definido ou se o destino for clicado no painel de visualização principal.

Passe o mouse sobre o gráfico para exibir o valor do poder preditivo.

Tabela de históricos de processamento

A tabela mostra como a versão transformada de um campo foi derivada. As ações tomadas pela ADP são listadas na ordem em que foram executadas, no entanto, para certas etapas, podem ter sido executadas várias ações para um campo específico.

Nota: esta tabela não é mostrada para os campos que não foram transformados.

As informações na tabela são divididas em duas ou três colunas:

- **Ação.** O nome da ação. Por exemplo, Preditores contínuos. Consulte o tópico "Detalhes da Ação " na página 17 para obter mais informações.
- **Detalhes.** A lista de processamentos executados. Por exemplo, Transformar em unidades padrão.
- **Função.** Mostrada somente para preditores construídos, isso exibe a combinação linear dos campos de entrada, por exemplo, 0,06*idade + 1,21*altura.

Detalhes da Ação

Exibidos quando você seleciona qualquer **Ação** sublinhada na visualização principal de Sumarização de Ação; a visualização vinculada de Detalhes de Ação exibe informações específicas de ação e comuns para cada etapa de processamento que foi executada; os detalhes específicos de ação são exibidos primeiro.

Para cada ação, a descrição é usada como o título na parte superior da visualização vinculada.
Os detalhes específicos da ação são exibidos abaixo do título e podem incluir detalhes do número de preditores derivados, campos reformulados, transformações de destino, categorias mescladas ou reordenadas e preditores construídos ou excluídos.

Conforme cada ação é processada, o número de preditores usados no processamento pode mudar, por exemplo, à medida que preditores são excluídos ou mesclados.

Nota: se uma ação foi desativada ou nenhum destino foi especificado, uma mensagem de erro será exibida no lugar dos detalhes da ação quando a ação é clicada na visualização principal Sumarização da Ação.

Há nove ações possíveis, no entanto, nem todas elas estão necessariamente ativas para cada análise.

Tabela de campos de texto

A tabela exibe o número de:

- Preditores excluídos da análise.

Tabela de Preditores de Data e Hora

A tabela exibe o número de:

- Durações derivadas de preditores de data e hora.
- Elementos de data e hora.
- Preditores de data e hora derivados, no total.

A data ou hora de referência é exibida como uma nota de rodapé se quaisquer durações de data foram calculadas.

Tabela de triagem do preditor

A tabela exibe o número dos preditores a seguir excluídos do processamento:

- Constantes.

- Preditores com valores omissos em excesso.
- Preditores com casos em excesso em uma única categoria.
- Campos nominais (conjuntos) com categorias em excesso.
- Preditores triados, no total.

Verificar Tabela de Nível de Medição

A tabela exibe os números de campos reformulados, divididos como a seguir:

- Campos ordinais (conjuntos ordenados) reformulados como campos contínuos.
- Campos contínuos reformulados como campos ordinais.
- Número total de reformulações.

Se nenhum campo de entrada (destino ou preditor) era contínuo ou ordinal, isso será mostrado como uma nota de rodapé.

Tabela de valores discrepantes

A tabela exibe contagens de quantos valores discrepantes foram manipulados.

- O número de campos contínuos para os quais valores discrepantes foram localizados e aparados ou o número de campos contínuos para os quais valores discrepantes foram localizados e configurados como omissos, dependendo de suas configurações no painel Preparar entradas e destino na guia Configurações.
- O número de campos contínuos excluídos porque eram constantes, após a manipulação de valor discrepante.

Uma nota de rodapé mostra o valor de corte do valor discrepante, enquanto outra nota de rodapé será mostrada se nenhum campo de entrada (destino ou preditor) era contínuo.

Tabela de Valores Omissos

A tabela exibe os números de campos que tiveram valores omissos substituídos, divididos em:

- Resposta. Esta linha não será mostrada se nenhum destino for especificado.
- Preditores. Isso é dividido ainda em número de nominal (conjunto), ordinal (conjunto ordenado) e contínuo.
- O número total de valores omissos substituídos.

Tabela de destino

A tabela exibe se o destino foi transformado, mostrado como:

- Transformação Box-Cox para normalidade. Isso é dividido ainda mais em colunas que mostram os critérios especificados (desvio médio e padrão) e Lambda.
- Categorias de destino reordenadas para melhorar a estabilidade.

Tabela de preditores categóricos

A tabela exibe o número de preditores categóricos:

- Cujas categorias foram reordenadas da menor para a maior para melhorar a estabilidade.
- Cujas categorias foram mescladas para maximizar a associação com o destino.
- Cujas categorias foram mescladas para manipular categorias de dispersão.
- Excluídos devido à baixa associação com o destino.
- Excluídos porque eram constantes após a mesclagem.

Uma nota de rodapé será mostrada se não houver nenhum preditor categórico.

Tabela de Preditores Contínuos

Existem duas tabelas. A primeira exibe um dos números de transformações a seguir:

- Valores do preditor transformados em unidades padrão. Além disso, isso mostra o número de preditores transformados, a média especificada e o desvio padrão.
- Valores do preditor mapeados para um intervalo comum. Além disso, isso mostra o número de preditores transformados usando uma transformação mín-máx, bem como os valores mínimo e máximo especificados.
- Valores do preditor categorizados e o número de preditores categorizados.

A segunda tabela exibe os detalhes da construção de espaço do preditor, mostrados como o número de preditores:

- Construídos.
- Excluídos devido a uma baixa associação com o destino.
- Excluídos porque eram constantes após a categorização.
- Excluídos porque eram constantes após a construção.

Uma nota de rodapé é mostrada se nenhum preditor contínuo foi inserido.

Escores de transformação retroativa

Se um destino tiver sido transformado pela ADP, os modelos subsequentes construídos usando o destino transformado escorarão as unidades transformadas. Para interpretar e usar os resultados, deve-se converter o valor predito novamente para a escala original.

1. Para transformar escores de volta, a partir dos menus escolha:

   Transformar > **Preparar dados para modelagem** > **Escores de transformação retroativa...**
2. Selecione um campo para transformar de volta. Este campo deve conter valores preditos pelo modelo do destino transformado.
3. Especifique um sufixo para o novo campo. Este novo campo conterá valores preditos pelo modelo na escala original do destino não transformado.
4. Especifique a localização do arquivo XML que contém as transformações de ADP. Isso deve ser um arquivo salvo a partir de diálogos de preparação de dados interativa ou automática. Consulte o tópico “Aplicando e Salvando Transformações ” na página 12 para obter mais informações

Identificar casos incomuns

O procedimento Detecção de anomalias procura casos incomuns com base em desvios das normas de seus grupos de clusters. O procedimento foi projetado para detectar rapidamente casos incomuns para propósitos de auditoria de dados na etapa de análise exploratória de dados, antes de qualquer análise inferencial de dados. Esse algoritmo foi projetado para detecção de anomalias genéricas, ou seja, a definição de um caso anômalo não é específica para nenhum aplicativo em particular, como a detecção de padrões de pagamento incomuns no segmento de mercado de assistência médica ou a detecção de lavagem de dinheiro no segmento de mercado de finanças, no qual a definição de uma anomalia pode ser bem delimitada.

Exemplo. Um analista de dados contratado para construir modelos preditivos para resultados de tratamento de AVC está preocupado com a qualidade dos dados, porque tais modelos podem ser sensíveis a observações incomuns. Algumas dessas observações discrepantes representam casos realmente exclusivos e, portanto, não são apropriadas para predição, enquanto outras observações são causadas por erros de entrada de dados nos quais os valores estão tecnicamente "corretos" e, portanto, não podem ser capturados por procedimentos de validação de dados. O procedimento Identificar casos incomuns localiza e relata esses valores discrepantes para que o analista possa decidir como tratá-los.

Estatísticas. O procedimento produz grupos de peers, normas do grupo de peers para variáveis contínuas e categóricas, índices de anomalia com base em desvios de normas do grupo de peers e valores de impacto de variável para variáveis que mais contribuem com um caso que está sendo considerado incomum.

Considerações de dados

Dados. Este procedimento funciona com variáveis contínuas e categóricas. Cada linha representa uma observação distinta e cada coluna representa uma variável distinta na qual os grupos de peers são baseados. Uma variável de identificação de caso pode estar disponível no arquivo de dados para marcação de saída, mas ela não será usada na análise. Os valores omissos são permitidos. A variável de ponderação, se especificada, é ignorada.

O modelo de detecção pode ser aplicado a um novo arquivo de dados de teste. Os elementos dos dados de teste devem ser iguais aos elementos dos dados de treinamento. E, dependendo das configurações do algoritmo, o tratamento de valor omisso que é usado para criar o modelo pode ser aplicado ao arquivo de dados de teste antes da escoragem.

Ordem de casos. Observe que a solução pode depender da ordem dos casos. Para minimizar os efeitos da ordem, ordene os casos de forma aleatória. Para verificar a estabilidade de uma determinada solução, talvez você queira obter várias soluções diferentes com casos ordenados em diferentes ordens aleatórias. Em situações com tamanhos de arquivos extremamente grandes, podem ser feitas várias execuções com uma amostra de casos ordenados em diferentes ordens aleatórias.

Suposições. O algoritmo considera que todas as variáveis são inconstantes e independentes e que nenhum caso possui valores omissos para qualquer uma das variáveis de entrada. Cada variável contínua é considerada como tendo uma distribuição normal (Gaussiana) e cada variável categórica é considerada como tendo uma distribuição multinomial. O teste empírico interno indica que o procedimento é bastante robusto a violações da suposição de independência e das suposições distributivas, mas esteja ciente de como essas suposições são atendidas.

Para identificar casos incomuns

1. Nos menus, escolha:

   Dados > **Identificar casos incomuns...**
2. Selecione pelo menos uma variável de análise.
3. Opcionalmente, escolha uma variável identificadora de caso para ser usada na identificação de saída.

Campos com nível de medição desconhecido

O alerta de Nível de Medição é exibido quando o nível de medição para uma ou mais variáveis (campos) no conjunto de dados é desconhecido. Como o nível de medição afeta o cálculo de resultados para este procedimento, todas as variáveis devem ter um nível de medição definido.

Dados de varredura. Lê os dados no conjunto de dados ativo e designa o nível de medição padrão para quaisquer campos com um nível de medição desconhecido atualmente. Se o conjunto de dados for grande, isso poderá demorar algum tempo.

Designar Manualmente. Abre um diálogo que lista todos os campos com um nível de medição desconhecido. É possível utilizar este diálogo para designar o nível de medição para esses campos. Também é possível designar o nível de medição na Visualização de Variável do Editor de Dados.

Como o nível de medição é importante para este procedimento, não é possível acessar o diálogo para executar este procedimento até que todos os campos possuam um nível de medição definido.

Identificar saída de casos incomuns

Lista de casos incomuns e as razões pelas quais eles são considerados incomuns. Essa opção produz três tabelas:

- A lista de índices de casos de anomalia exibe casos que são identificados como incomuns e exibe seus valores de índice de anomalia correspondentes.
- A lista de IDs de peers de casos de anomalia exibe casos incomuns e informações referentes a seus grupos de peers correspondentes.
- A lista de razões de anomalia exibe o número do caso, a variável de razão, o valor de impacto da variável, o valor da variável e a norma da variável para cada razão.

Todas as tabelas são ordenadas por índice de anomalia em ordem decrescente. Além disso, os IDs dos casos serão exibidos se a variável identificadora de caso for especificada na guia Variáveis.

Resumos. Os controles nesse grupo produzem sumarizações de distribuição.

- **Normas do grupo de peers.** Essa opção exibe a tabela de normas de variáveis contínuas (se alguma variável contínua for usada na análise) e a tabela de normas de variáveis categóricas (se alguma variável categórica for usada na análise). A tabela de normas de variáveis contínuas exibe a média e o desvio padrão de cada variável contínua para cada grupo de peers. A tabela de normas de variável categórica exibe o modo (categoria mais popular), a frequência e a porcentagem de frequência de cada variável categórica para cada grupo de peers. A média de uma variável contínua e o modo de uma variável categórica são usados como os valores de norma na análise.
- **Índices de anomalia.** A sumarização de índice de anomalia exibe estatísticas descritivas para o índice de anomalia dos casos que são identificados como os mais incomuns.
- **Ocorrência de razão por variável de análise.** Para cada razão, a tabela exibe a frequência e a porcentagem de frequência de ocorrência de cada variável como uma razão. A tabela também relata as estatísticas descritivas do impacto de cada variável. Se o número máximo de razões estiver configurado como 0 na guia Opções, essa opção não estará disponível.
- **Casos processados.** A sumarização de processamento de caso exibe as contagens e porcentagens de contagens para todos os casos no conjunto de dados ativo, os casos incluídos e excluídos na análise e os casos em cada grupo de peers.

Identificar casos incomuns - Salvar

Salvar variáveis. Os controles nesse grupo permitem salvar variáveis de modelo no conjunto de dados ativo. Também é possível optar por substituir variáveis existentes cujos nomes entram em conflito com as variáveis a serem salvas.

- **Índice de anomalia.** Salva o valor do índice de anomalia para cada caso para uma variável com o nome especificado.
- **Grupos de peers.** Salva o ID do grupo de peers, a contagem de caso e o tamanho como uma porcentagem para cada caso para variáveis com o nome raiz especificado. Por exemplo, se o nome raiz *Peer* for especificado, as variáveis *Peerid*, *PeerSize* e *PeerPctSize* serão geradas. *Peerid* é o ID do grupo de peers do caso, *PeerSize* é o tamanho do grupo e *PeerPctSize* é o tamanho do grupo como uma porcentagem.
- **Razões.** Salva conjuntos de variáveis de razão com o nome raiz especificado. Um conjunto de variáveis de razão consiste no nome da variável como a razão, sua medida de impacto da variável, seu próprio valor e o valor da norma. O número de conjuntos depende do número de razões solicitadas na guia Opções. Por exemplo, se o nome raiz *Reason* for especificado, as variáveis *ReasonVar_k*, *ReasonMeasure_k*, *ReasonValue_k* e *ReasonNorm_k* serão geradas, em que *k* é a razão *k*. Essa opção não estará disponível se o número de razões estiver configurado como 0.

Exportar arquivo de modelo. Permite salvar o modelo em formato XML.

Identificar casos incomuns - Valores omissos

A guia Valores omissos é usada para controlar o tratamento de valores omissos do usuário e omissos do sistema.

- **Excluir valores omissos da análise.** Os casos com valores omissos são excluídos da análise.
- **Incluir valores omissos na análise.** Os valores omissos de variáveis contínuas são substituídos por suas médias globais correspondentes e as categorias omissas de variáveis categóricas são agrupadas e tratadas como uma categoria válida. As variáveis processadas são então usadas na análise. Opcionalmente, é possível solicitar a criação de uma variável adicional que representa a proporção de variáveis omissas em cada caso e usar essa variável na análise.

Identificar opções de casos incomuns

Critérios para identificar casos incomuns. Essas seleções determinam quantos casos são incluídos na lista de anomalias.

- **Porcentagem de casos com valores mais altos de índice de anomalia.** Especifique um número positivo que seja menor ou igual a 100.
- **Número fixo de casos com valores mais altos de índice de anomalia.** Especifique um número inteiro positivo que seja menor ou igual ao número total de casos no conjunto de dados ativo que são usados na análise.
- **Identificar somente casos cujo valor de índice de anomalia atende ou excede um valor mínimo.** Especifique um número não negativo. Um caso é considerado anômalo se seu valor de índice de anomalia é maior ou igual ao ponto de corte especificado. Essa opção é usada junto com as opções **Porcentagem de casos** e **Número fixo de casos**. Por exemplo, se você especificar um número fixo de 50 casos e um valor de corte de 2, a lista de anomalias consistirá, no máximo, em 50 casos, cada um com um valor de índice de anomalia maior ou igual a 2.

Número de grupos de peers. O procedimento irá procurar o melhor número de grupos de peers entre os valores mínimo e máximo especificados. Os valores devem ser números inteiros positivos e o mínimo não deve exceder o máximo. Quando os valores especificados forem iguais, o procedimento considerará um número fixo de grupos de peers.

Nota: Dependendo da quantidade de variação em seus dados, pode haver situações nas quais o número de grupos de peers que os dados podem suportar é menor que o número especificado como o mínimo. Nessa situação, o procedimento pode produzir um número menor de grupos de peers.

Número máximo de razões. Uma razão consiste na medida de impacto da variável, no nome da variável para essa razão, no valor da variável e no valor do grupo de peers correspondente. Especifique um número inteiro não negativo, se esse valor for igual ou exceder o número de variáveis processadas que são usadas na análise, todas as variáveis serão mostradas.

Recursos adicionais do comando DETECTANOMALY

O idioma da sintaxe de comando também permite:

- Omita algumas variáveis no conjunto de dados ativo da análise sem especificar explicitamente todas as variáveis de análise (usando o subcomando EXCEPT).
- Especifique um ajustamento para balancear a influência de variáveis contínuas e categórica (usando a palavra-chave `MLWEIGHT` no subcomando `CRITERIA`).

Consulte a *Referência de Sintaxe de Comando* para obter informações de sintaxe completa.

Categorização ideal

O procedimento Categorização ideal distingue uma ou mais variáveis de escala (referidas de agora em diante como **variáveis de entrada de categorização**), distribuindo os valores de cada variável em categorias. A formação da categoria é ideal com relação a uma variável guia categórica que "supervisiona" o processo de categorização. As categorias podem, então, ser usadas no lugar dos valores de dados originais para análise adicional.

Exemplos. A redução do número de valores distintos usados por uma variável tem vários usos, incluindo:

- Requisitos de dados de outros procedimentos. Variáveis distintas podem ser tratadas como categóricas para uso em procedimentos que requerem variáveis categóricas. Por exemplo, o procedimento Tabulações cruzadas requer que todas as variáveis sejam categóricas.
- Privacidade de dados. Relatar valores categorizados em vez de valores reais pode ajudar a proteger a privacidade de suas origens de dados. O procedimento Categorização ideal pode orientar a escolha de categorias.
- Desempenho da velocidade. Alguns procedimentos são mais eficientes ao trabalhar com um número reduzido de valores distintos. Por exemplo, a velocidade da Regressão logística multinomial pode ser melhorada usando variáveis distintas.
- Descobrindo a separação de dados completa ou quase completa.

Categorização ideal versus visual. As caixas de diálogo Categorização visual oferecem vários métodos automáticos para criar categorias sem o uso de uma variável guia. Essas regras "não supervisionadas" são úteis para produzir estatísticas descritivas, como tabelas de frequências, mas a Categorização ideal é superior quando seu objetivo final é produzir um modelo preditivo.

Saída. O procedimento produz tabelas de pontos de corte para as categorias e estatísticas descritivas para cada variável de entrada de categorização. Além disso, é possível salvar novas variáveis no conjunto de dados ativos contendo os valores categorizados das variáveis de entrada de categorização e salvar as regras de categorização como sintaxe de comando para uso na distinção de novos dados.

Considerações de dados de categorização ideal

Dados. Esse procedimento espera que as variáveis de entrada de categorização sejam variáveis de escala numéricas. A variável guia deve ser categórica e pode ser uma sequência de caracteres ou numérica.

Para obter um agrupamento aprimorado

1. A partir dos menus, escolha:

   Transformar > **Categorização ideal...**
2. Selecione uma ou mais variáveis de entrada de categorização.
3. Selecione uma variável guia.

Variáveis contendo os valores de dados categorizados não são geradas por padrão. Use a guia Salvar para salvar essas variáveis.

Saída de categorização ideal

A guia Saída controla a exibição dos resultados.

- **Terminais para categorias.** Exibe o conjunto de terminais para cada variável de entrada de categorização.
- **Estatísticas descritivas para variáveis que são categorizadas.** Para cada variável de entrada de categorização, essa opção exibe o número de casos com valores válidos, o número de casos com valores ausentes, o número de valores válidos distintos e os valores mínimo e máximo. Para a variável guia, esta opção exibe a distribuição de classe para cada variável de entrada de categorização relacionada.
- **Entropia de modelo para variáveis que são categorizadas.** Para cada variável de entrada de categorização, essa opção exibe uma medida da precisão preditiva da variável com relação à variável guia.

Categorização ideal - Salvar

Salvar variáveis no conjunto de dados ativo. As variáveis que contêm valores de dados categorizados podem ser usadas no lugar das variáveis originais em análise adicional.

Salvar regras de categorização como sintaxe. Gera a sintaxe de comando que pode ser usada para categorizar outros conjuntos de dados. As regras de recodificação são baseadas nos pontos de corte determinados pelo algoritmo de categorização.

Valores omissos de categorização ideal

A guia Valores omissos especifica se os valores omissos são manipulados usando a exclusão listwise ou pairwise. Os valores omissos de usuário são sempre tratados como inválidos. Ao registrar os valores de variáveis originais em uma nova variável, os valores omissos do usuário são convertidos em omissos do sistema.

- **Pairwise.** Essa opção opera em cada par de variável de entrada guia e de categorização. O procedimento usará todos os casos com valores não omissos na variável de entrada de guia e de categorização.

- **Listwise** Esta opção opera em todas as variáveis especificadas na guia Variáveis. Se alguma variável estiver ausente para um caso, o caso inteiro será excluído.

Opções de categorização ideal

Pré-processamento. A "pré-categorização" de variáveis de entrada de categorização com muitos valores distintos pode melhorar o tempo de processamento sem um grande sacrifício na qualidade das categorias finais. O número máximo de categorias fornece um limite superior no número de categorias criadas. Portanto, se você especificar 1000 como o máximo, mas uma variável de entrada de categorização tiver menos de 1000 valores distintos, o número de categorias pré-processadas criadas para a variável de entrada de categorização será igual ao número de valores distintos na variável de entrada de categorização.

Categorias esparsamente preenchidas. Ocasionalmente, o procedimento pode produzir categorias com pouquíssimos casos. A seguinte estratégia exclui esses pseudopontos de corte:

Para uma determinada variável, suponha que o algoritmo localizou n_{final} pontos de corte e, portanto, categorias $n_{final}+1$. Para categorias $i = 2, \ldots, n_{final}$ (a segunda categoria com menor valor até a segunda categoria com maior valor), compute

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}), sizeof(b_{i+1}))}$$

em que *sizeof(b)* é o número de casos na categoria.

Quando esse valor for menor que o limite de mesclagem especificado, b_i será considerado esparsamente preenchido e será mesclado com b_{i-1} ou b_{i+1}, o que tiver a menor entropia de informações de classe.

O procedimento faz uma única passagem através das categorias.

Terminais de categoria. Esta opção especifica como o limite inferior de um intervalo é definido. Como o procedimento determina automaticamente os valores dos pontos de corte, isso é principalmente uma questão de preferência.

Primeira (Mais baixa) / Última (Mais alta) categoria. Essas opções especificam como os pontos de corte mínimo e máximo para cada variável de entrada de categorização são definidos. Geralmente, o procedimento considera que as variáveis de entrada de categorização podem usar qualquer valor na linha de número real, mas se você tiver alguma razão teórica ou prática para limitar o intervalo, será possível limitá-lo pelos valores mais baixo/mais alto.

Recursos adicionais do comando OPTIMAL BINNING

O idioma da sintaxe de comando também permite:

- Executar categorização não supervisionada por meio do método de frequências iguais (usando o subcomando `CRITERIA`).

Consulte a *Referência de Sintaxe de Comando* para obter informações de sintaxe completa.

Avisos

Estas informações foram desenvolvidas para produtos e serviços oferecidos nos EUA. Esse material pode estar disponível a partir da IBM em outros idiomas. Entretanto, pode ser necessário que possua uma cópia do produto ou versão de produto nesse idioma a fim de 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 um produto, programa ou serviço IBM não está destinado a declarar ou implicar que apenas esse produto, programa ou serviço IBM possa ser usado. Qualquer produto, programa ou serviço funcionalmente equivalente, que não infrinja nenhum direito de propriedade intelectual da IBM poderá ser utilizado em substituição a este produto, programa ou serviço. 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. É possível enviar consultas sobre licenças, por escrito, para:

Gerência de Relações Comerciais e Industriais da IBM Brasil

Av. Pasteur, 138-146, Botafogo
Botafogo
Rio de Janeiro, RJCEP 22290-240

Para consultas sobre licença relacionados a informações de DBCS (Conjunto de Caracteres de Byte Duplo), entre em contato com o Departamento de Propriedade Intelectual da IBM em seu país ou envie consultas sobre licença, por escrito, para:

Intellectual Property Licensing

Legal and Intellectual Property Law
IBM Japan Ltd.

19-21, Nihonbashi-Hakozakicho, Chuo-kuTokyo 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 A ELAS NÃO SE LIMITANDO, AS GARANTIAS IMPLÍCITAS DE NÃO INFRAÇÃO, COMERCIALIZAÇÃO OU 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. Periodicamente, são feitas mudanças nas informações aqui contidas; tais mudanças serão incorporadas em novas edições da publicação. A IBM pode, a qualquer momento, aperfeiçoar e/ou alterar os produtos e/ou programas descritos nesta publicação, sem aviso prévio.

Qualquer referência nestas informações a websites não IBM são fornecidas apenas por conveniência e não representam de forma alguma um endosso a esses websites. Os materiais contidos nesses websites não fazem parte dos materiais desse produto IBM e a utilização desses websites é de inteira responsabilidade do Cliente.

A IBM por usar ou distribuir as 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

Av. Pasteur, 138-146, Botafogo
Botafogo
Rio de Janeiro, RJCEP 22290-240

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 dados de desempenho e do Cliente citados são apresentados apenas para propósitos ilustrativos. Os resultados de desempenho reais podem variar dependendo das configurações específicas e condições operacionais.

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 esses produtos e não pode confirmar a precisão de desempenho, compatibilidade nem qualquer outra reivindicação relacionada a produtos não IBM. Perguntas sobre os recursos de produtos não IBM devem ser endereçadas aos fornecedores desses produtos.

Instruções relativas à direção futura ou intento da IBM estão sujeitas a mudança ou retirada sem aviso e representam metas e objetivos apenas.

Essas informações contêm exemplos de dados e relatórios utilizados em operações diárias de negócios. Para ilustrá-los da forma mais completa possível, os exemplos incluem nomes de indivíduos, empresas, marcas e produtos. Todos esses nomes são fictícios e qualquer semelhança com pessoas ou empresas reais é mera coincidência.

LICENÇA DE COPYRIGHT:

Estas informações contêm programas de aplicativos de amostra na linguagem fonte, ilustrando as técnicas de programação em diversas plataformas operacionais. O Cliente pode copiar, modificar e distribuir estes programas de exemplo sem a necessidade de pagar à IBM, com objetivos de desenvolvimento, utilização, marketing ou distribuição de programas aplicativos em conformidade com a interface de programação de aplicativo para a plataforma operacional para a qual os programas de amostra são criados. Esses exemplos não foram testados completamente em todas as condições. Portanto, a IBM não pode garantir ou implicar a confiabilidade, manutenção ou função destes programas. Os programas de amostra são fornecidos "no estado em que se encontram" sem garantia de nenhum tipo. A IBM não será responsabilizada por quaisquer danos decorrentes do uso dos programas de amostra.

Cada cópia ou parte destes programas de amostra ou qualquer trabalho derivado deve incluir um aviso de copyright com os dizeres:

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 inteiro. Outros nomes de produtos e serviços podem ser marcas registradas da IBM ou de outras empresas. ma lista atual de marcas comerciais da IBM está disponível na web em "Copyright and trademark information" em www.ibm.com/legal/copytrade.shtml.

Adobe, o logotipo Adobe, PostScript e o logotipo PostScript são marcas ou marcas registradas 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 Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium são marcas comerciais ou marcas registradas da Intel Corporation ou de suas subsidiárias nos Estados Unidos e em outros países.

Linux é 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 da The Open Group nos Estados Unidos e 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.

Índice remissivo

Caracteres Especiais

C

D

E

G

I

M

N

P

R

S

T

V