IBM SPSS Data Preparation 24

Comunicado

Antes de usar estas informações e o produto suportado por elas, leia as informações nos "Avisos" na página 33.

Informações sobre o produto

Esta edição aplica-se à versão 24, liberação 0, modificação 0 do IBM® SPSS Statistics e a todas as liberações e modificações subsequentes até que seja indicado de outra forma em novas edições.

Índice

Capítulo 1. 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 validar dados.

O módulo complementar Preparação de Dados permite identificar casos incomuns e casos, variáveis e valores de dados inválidos em seu conjunto de dados ativo, e preparar dados para modelagem.

Uso de procedimentos de Preparação de Dados

O 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 precisar 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 possíveis valores discrepantes estatísticos que podem causar problemas para muitos 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 a partir de outros módulos complementares.

Capítulo 2. 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 a partir do dicionário de dados. Quando não houver regras, uma nova regra de item temporário que pode ser modificada de acordo com seus propósitos é 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.
- **Exclusão.** 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 de 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 CrossVarRule 1," e assim por diante.
- **Exclusão.** 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 variável selecionada atualmente é exibida em uma área reservada na caixa de diálogo.
- As constantes de sequência devem ser colocadas entre aspas ou apóstrofos.
- Se os valores contiverem decimais, (.) deve ser usado como o indicador decimal.

Capítulo 3. 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 satisfação do cliente mensal 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 de 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.

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

Como opção, você pode:

- 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 de caso. 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 seguintes verificações de sua validade. 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 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 algumas 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 algumas 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 algumas 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 algumas 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 identificadoras de casos na guia Variáveis, será possível selecionar qualquer uma das seguintes verificações de sua validade.

- **Sinalizar IDs incompletos.** Esta opção relata casos com identificadores de casos incompletos. Para um caso específico, um identificador é 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 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 3 para obter mais informações

Variáveis de análise. A lista mostra variáveis de análise, sumariza suas distribuições e mostra o número de regras aplicadas a cada variável. Observe que os valores omissos de 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 as 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 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 4 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 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 exibir todos eles.

- **Sumarizar violações por variável de análise.** Para cada variável análise, esta 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. Esta opção permite solicitar estatísticas descritivas para variáveis de análise. Uma tabela de frequências é gerada para cada variável categórica. Uma tabela de estatísticas básicas, 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 leitura fácil.

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.** Os casos vazios são designados ao valor 1. Todos os outros casos são codificados com 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) são designados ao mesmo número de grupo. Casos com identificadores exclusivos ou incompletos são codificados com 0.
- **Indicador de ID incompleto.** Casos com identificadores de caso vazios ou incompletos são designados ao valor 1. Todos os outros casos são codificados com 0.
- **Violações de regra de validação.** Essa é a contagem total entre casos 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 um aplicativo de uma regra de validação e tem um valor 1 se o caso violar a regra, e um valor 0, se ele não violar.

Capítulo 4. Preparação de dados automatizados

Preparar dados para análise é um dos passos mais importantes em qualquer projeto - e, tradicionalmente, um dos mais demorados. A Preparação de Dados Automatizados (ADP) manipula a tarefa para você, analisando os seus dados e identificando correções, fazendo a triagem de campos que são problemáticos ou provavelmente inúteis, derivando novos atributos quando apropriado e melhorando o desempenho através de técnicas inteligentes de triagem. É possível usar o algoritmo de modo totalmente **automático**, permitindo que ele escolha e aplique correções ou é possível usá-lo de modo **interativo**, visualizando as mudanças antes de elas serem feitas e aceitar e rejeitá-las conforme você desejar.

Usar ADP possibilita que você torne os seus dados prontos para construção de modelo fácil e rapidamente, sem precisar de conhecimento anterior dos conceitos estatísticos envolvidos. Os modelos tenderão a ser construídos e pontuados com mais agilidade; além disso, usar ADP melhora a robustez de 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 antigo campo não é usado em análise posterior; o seu papel está configurado como Nenhum. Além disso, observe que qualquer informação de valor omisso de usuário não é transferida para esses campos recém-criados e qualquer valor omisso no novo campo é omisso para o sistema.

Exemplo. Uma empresa de seguros com recursos limitados para investigar indenizações de seguro dos proprietários de imóveis quer construir um modelo para sinalizar indenizações suspeitas, potencialmente fraudulentas. Antes de construir o modelo, eles irão deixar os dados prontos para modelagem usando preparação de dados automatizados. Como eles querem poder revisar as transformações propostas antes que elas sejam aplicadas, eles usarão preparação de dados automatizados no modo interativo.

Um grupo do segmento de mercado automotivo mantém o controle das vendas para uma variedade de veículos motorizados pessoais. Em um esforço para serem capazes de identificar modelos de desempenho superior e inferior, eles querem estabelecer um relacionamento entre vendas de veículo e características de veículo. Eles usarão preparação de dados automatizados para preparar os dados para análise e construir modelos usando os dados de "antes" e "depois" da preparação para verem como os resultados diferem.

Qual é o seu objetivo? A preparação de dados automatizados recomenda passos de preparação de dados que afetarão a velocidade com a qual outros algoritmos podem construir modelos e melhorar o poder preditivo desses modelos. Isso pode incluir a transformação, construção e seleção de recursos. 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 fornecer prioridade igual para ambas: à velocidade com as quais dados são processados por algoritmos de construção de modelo e à precisão das predições.
- **Otimizar para velocidade.** Essa opção prepara os dados para fornecer prioridade para a velocidade com a qual os dados são processados por algoritmos de construção de modelo. Quando você estiver trabalhando com conjuntos de dados muitos grandes ou estiver procurando uma resposta rápida, selecione essa opção.
- **Otimizar para precisão.** Essa opção prepara os dados para fornecer prioridade para a precisão de predições produzidas por algoritmos de construção de modelo.
- **Análise customizada.** Quando você desejar mudar manualmente o algoritmo na guia Configurações, selecione essa opção. Observe que essa configuração será selecionada automaticamente se você subsequentemente fizer mudanças em opções na guia Configurações que forem incompatíveis com um dos outros objetivos.

Para Obter Preparação de Dado Automático

Nos menus, escolha:

1. Nos menus, escolha:

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

Como alternativa, você pode:

- Especificar um objetivo na guia Objetivo.
- Especificar designações de campo na guia Campos.
- Especificar configurações de especialista 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 sugerida.
4. Se estiver satisfeito, clique em **Executar**. Caso contrário, clique em **Limpar Análise**, mude qualquer configuração que desejar e clique em **Analisar**.

Como opção, você pode:

- Especificar um objetivo na guia Objetivo.
- Especificar designações de campo na guia Campos.
- Especificar configurações de especialista 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 posterior.

Usar funções predefinidas. Essa opção usa informação de campo. Se houver um campo único com um papel como uma Resposta, ele será usado como a resposta; caso contrário, não haverá nenhuma resposta. Todos os campos com uma função predefinida como uma Entrada serão usados como entradas. Pelo menos um campo de entrada é necessário.

Usar designações de campo customizado. Quando você substitui papéis do campo movendo campos a partir de suas listas padrão, o diálogo automaticamente é alternado para essa opção. Ao fazer designações de campo customizado, especifique os campos a seguir:

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

Guia Configurações

A guia Configurações abrange vários grupos diferentes de configurações que é possível modificar 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 podem manipular diretamente detalhes de data e hora; essas configurações permitem que você derive novos dados de duração que podem ser usados como entradas de modelo a partir de datas e horas em seus dados existentes. Os campos que contêm datas e horas devem ser predefinidos com tipos de armazenamento de data e hora. Os campos originais de data e hora não serão recomendados como entrada de modelo seguindo a preparação de dados automatizados.

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

Calcular 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 considerando as informações sobre a data nos dados de entrada. Selecionar **Data de hoje** significa que a data do sistema atual é sempre usada quando a ADP é executada. Para usar uma data específica, selecione **Data fixa** e insira a data requerida.
- **Unidades para Duração de Data.** Especifique se a ADP deve decidir automaticamente sobre 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 horário de referência para cada variável que contém horários.

- **Horário de Referência.** Especifique o horário a partir do qual a duração será calculada considerando as informações sobre o horário 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 a ADP deve decidir automaticamente sobre a unidade de duração de tempo ou selecione a partir de **Unidades fixas** de Horas, Minutos ou Segundos.

Extrair Elementos Cíclicos de Tempo. Use 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 a partir das datas.** Para qualquer entrada de data, especifique se você deseja extrair anos, meses, dias ou qualquer combinação.
- **Extrair a partir de horários.** Para qualquer entrada de horário, especifique se você deseja extrair horas, minutos, segundos ou qualquer combinação.

Excluir Campos

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

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

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

Excluir campos nominais com demasiadas categorias exclusivas. Campos nominais com mais do que o número especificado de categorias são removidos da análise posterior. Especifique um número inteiro

positivo. O padrão é 100. Isso é útil para remover automaticamente da modelagem campos que contêm informações exclusivas de registro, como ID, endereço ou nome.

Excluir campos categóricos com demasiados valores em uma única categoria. Campos ordinais e nominais com uma categoria que contém mais do que a porcentagem especificada dos registros são removidos da análise posterior. 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.

Ajustar Medição

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

Tratamento do Valor Discrepante. Especifique se deve-se substituir valores discrepantes para as entradas e a resposta; se for assim, especifique um critério de corte do valor discrepante, medido em desvios padrão e um método para substituir valores discrepantes. Valores discrepantes podem ser substituídos por corte (configuração para o valor de corte) ou configurando-os como valores omissos. Qualquer valor discrepante configurado como valores omissos segue as configurações de tratamento de valor omisso selecionadas abaixo.

Substituir Valores Omissos. Especifique se deve-se substituir valores omissos de campos contínuos, nominais ou ordinais.

Reordenar Campos Nominais. Selecione isso para recodificar os valores de campos nominais (configurar) a partir da categoria menor (que ocorre menos frequentemente) até a maior (que ocorre mais frequentemente). 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", então, a preparação de dados automatizados iria 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 se mantém as seleções.

Ponderação de Análise. Essa variável contém ponderações de análise (regressão ou amostragem). Ponderações de análise são usadas para explicar as diferenças em variância em 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. Usando as estimativas de desvio médio e padrão observadas como parâmetro de preenchimento, os campos são padronizados e, em seguida, os escores 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.

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 de volta para a escala original. Consulte o tópico para obter mais informações Consulte o tópico “Pontuações de transformação retrospectiva” na página 24 para obter mais informações

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 se 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 tornar um modelo mais econômico reduzindo o número de campos a serem processados em associação com a resposta. Categorias semelhantes são identificadas com base no relacionamento entre a entrada e a resposta. 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, as versões originais e derivadas do campo serão excluídas da análise posterior, porque elas não têm nenhum valor como um preditor.
- **Quando não existe nenhuma resposta, mescle as categorias de dispersão com base em contagens.** Se o conjunto de dados não tiver nenhuma resposta, é possível escolher 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. Mesclando paradas 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 de resposta categórica, é 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 categorização única para um determinado campo, as versões originais e categorizadas do campo serão excluídas porque elas não têm 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 ordenar 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 ordenar 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 na resposta.

Selecionar e Construir

Para melhorar o poder preditivo de seus dados, é possível construir novos campos 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 a partir de uma combinação de várias variáveis existentes. As variáveis antigas não são usadas em análise posterior. Essa opção aplica-se apenas a variáveis de entrada contínuas nas quais a resposta é contínua ou na qual não há nenhuma resposta.

Nomes do Campo

Para identificar facilmente variáveis novas e transformadas, a ADP cria e aplica novos nomes básicos, prefixos ou sufixos. É possível corrigir esses nomes para serem mais relevantes para as suas necessidades e dados.

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

Além disso, especifique o nome do prefixo a ser aplicado em qualquer variável que for construída através 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 como muitas variáveis novas são derivadas, por exemplo:

- 1 a 9 variáveis construídas serão denominadas: variável1 a variável9.
- 10 a 99 variáveis construídas serão denominadas: variável10 a variável99.
- 100 a 999 variáveis construídas serão denominadas: variável100 a variável999, e assim por diante.

Isso assegura que as variáveis construídas serão ordenadas em uma ordem sensata, independentemente de quantas houver.

Durações Calculadas a partir de Datas e Horas. Especifique as extensões de nome a serem aplicadas em durações calculadas a partir de ambas: datas e horas.

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

Aplicando e Salvando Transformações

Dependendo de você estar 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.** Campos recomendados pela preparação de dados automatizados 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 campos que contêm informações não usadas em 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 Dado Automático, 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 usada no diálogo Preparação de Dados Interativos para permitir que você revise as transformações recomendadas. O diálogo Preparação de Dado Automático não inclui esse passo.

1. Quando você estiver satisfeito com as configurações de ADP, incluindo qualquer mudança feita 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 sumariza o processamento de seus dados e exibe recomendações como para de que maneira os dados podem ser modificados ou melhorados para escoragem. É possível, então, revisar e aceitar ou rejeitar essas recomendações.

A guia Análise é composta de dois painéis, a visualização principal à esquerda e a visualização vinculada ou auxiliar à direita. Existem três visualizações principais:

- Sumarização de Processamento de Campo (o padrão). Consulte o tópico “Sumarização de Processamento de Campo” na página 18 para obter mais informações
- Campos. Consulte o tópico “Campos” na página 18 para obter mais informações
- Sumarização de Ação. Consulte o tópico “Sumarização de Ação” na página 19 para obter mais informações

Existem quatro visualizações vinculadas/auxiliares:

- Poder Preditivo (o padrão). Consulte o tópico “Poder Preditivo” na página 20 para obter mais informações
- Tabela de Campos. Consulte o tópico “Tabela de Campos” na página 20 para obter mais informações
- Detalhes de Campo. Consulte o tópico “Detalhes do Campo” na página 20 para obter mais informações

- Detalhes de Ação. Consulte o tópico “Detalhes de Ação” na página 22 para obter mais informações

Links entre visualizações

Dentro da visualização principal, o texto sublinhado nas tabelas controla a exibição na visualização vinculada. Clicar no texto permite que você obtenha detalhes sobre um determinado campo, conjunto de campos ou passo de processamento. O link que você selecionou por último é mostrado em cor escura; isso o ajuda 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 qualquer mudança que você fez nas visualizações de Análise, clique em **Reconfigurar** na parte inferior do painel de visualização principal.

Sumarização de Processamento de Campo

A Tabela de Sumarização de Processamento de Campo fornece uma captura instantânea do impacto geral planejado de processamento, incluindo mudanças no estado das variáveis e o número de variáveis construídas.

Observe que nenhum modelo está realmente construído, então não existe uma medida ou gráfico da mudança no poder preditivo geral antes e depois da preparação de dados; em vez disso, é possível exibir gráficos do poder preditivo de preditores recomendados individuais.

A tabela exibe as seguintes informações:

- O número de campo de destino.
- O número de preditores originais (de entrada).
- Os preditores recomendados para uso em análise e 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. Consulte o tópico “Tabela de Campos” na página 20 para obter mais informações. **Variáveis recomendadas para uso em análise** são exibidas na visualização vinculada de Poder Preditivo. Consulte o tópico “Poder Preditivo” na página 20 para obter mais informações

Campos

A visualização principal de 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 deve-se aceitar a transformação sugerida ou usar a versão original.

A visualização de Campos consiste em duas tabelas, uma para a resposta e uma 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 do campo de destino; 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 a resposta.

Tabela de preditores

A tabela de **Preditores** é sempre mostrada. Cada linha da tabela representa um campo. Por padrão, as linhas são ordenadas 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 versões transformadas de campos mostradas na tabela sempre representam as versões finais.

Por padrão, somente campos recomendados são mostrados na tabela de Preditores. Para exibir os campos restantes, selecione a caixa **Incluir campos não recomendados na tabela** acima da tabela; esses campos serão, então, 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 mais tarde e se deve-se usar as transformações sugeridas. Por padrão, a lista suspensa reflete as recomendações.

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

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

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

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

  Nota: Para preditores com ambas as versões: original e transformada, a mudança entre versões **Original** e **Transformada** 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 20 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.** Poder preditivo é exibido somente para campos que a ADP recomenda. Essa coluna não será exibida se não houver nenhuma resposta definida. Poder preditivo varia de 0 a 1, com maiores valores indicando "melhores" preditores. Em geral, o poder preditivo é útil para comparar preditores dentro de uma análise de ADP, mas valores de poder preditivo não devem ser comparados entre análises.

Sumarização de Ação

Para cada ação tomada pela preparação de dados automatizados, preditores de entrada são transformados e/ou removidos; campos que sobrevivem a uma ação são usados na próxima. Os campos

que sobrevivem até o último passo são, então, recomendados para uso em modelagem, enquanto as entradas para preditores transformados e construídos são removidas.

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 de Ação" na página 22 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 você seleciona **Preditores recomendados para uso em análise** na visualização principal de Sumarização de Processamento de Campo, o gráfico exibe o poder preditivo de preditores recomendados. Campos são ordenados por 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 a sua opção de sufixo no painel de Nomes do Campo da guia Configurações; por exemplo: *_transformed*.

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

O poder preditivo de cada poder preditivo recomendado é calculado a partir de uma regressão linear ou de um modelo de Bayes simples, dependendo de a resposta ser contínua ou categórica.

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 respostas, o nome ou rótulo original do campo é usado, mesmo se a resposta tiver sido transformada.

  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 usado; por exemplo: *Preditor1*.
- **Nível de Medição.** Isso exibe o ícone que representa o tipo de dados.

  Para a Resposta, o **Nível de Medição** sempre reflete a versão transformada (se a resposta tiver sido transformada); por exemplo, mudada de ordinal (conjunto ordenado) para contínua (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 e hora derivados e preditores construídos, os gráficos são mostrados apenas para o novo preditor.

Nota: Se um campo for excluído devido a ter demasiadas categorias somente o histórico de processamento será mostrado.

Gráfico de Distribuição

Distribuição de campo contínua é mostrada como um histograma, com uma sobreposição de curva normal e uma linha de referência vertical para o valor médio; campos categóricos são exibidos como um gráfico de barras.

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 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 sobre a contagem e percentagem do número total de registros para categorias em gráficos de barras.

Gráfico de Valor Omisso

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

Se a ADP executou manipulação de valor omisso, o gráfico de pizza de pós-transformação também incluirá o valor de substituição como um rótulo - ou seja, o valor usado no lugar de 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, gráficos de barras exibem o poder preditivo antes e depois da transformação. Se a resposta foi transformada, o poder preditivo calculado será em relação à resposta transformada.

Nota: Gráficos de poder preditivo não serão mostrados se nenhuma resposta estiver definida ou se a resposta for clicada 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 pela ordem em que foram executadas; no entanto, para certos passos podem ter sido executadas várias ações para um campo específico.

Nota: Essa 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 de Ação" na página 22 para obter mais informações
- **Detalhes.** A lista de processamento executada. Por exemplo, Transformar em unidades padrão.
- **Função.** Mostrada somente para preditores construídos, isso mostra a combinação linear dos campos de entrada, por exemplo, 0,06*idade + 1,21*altura.

Detalhes de 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 de ação são exibidos abaixo do título e podem incluir detalhes do número de preditores derivados, reformulação de campos, transformações de resposta, 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, conforme os preditores são excluídos ou mesclados.

Nota: Se uma ação foi desligada ou nenhuma resposta foi especificada, uma mensagem de erro é exibida no lugar dos detalhes de ação quando a ação é clicada na visualização principal de Sumarização de Ação.

Há nove ações possíveis; no entanto, nem todas são 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 qualquer duração de data foi calculada.

Tabela de Rastreamento de Preditor

A tabela exibe o número dos seguintes preditores excluídos do processamento:
- Constantes.
- Preditores com demasiados valores omissos.
- Preditores com demasiados casos em uma única categoria.
- Campos nominais (conjuntos) com demasiadas categorias.
- Preditores colocados fora da tela, no total.

Verificar Tabela de Nível de Medição

A tabela exibe os números de reformulações de campos, divididos como a seguir:
- Reformulação de campos ordinais (conjuntos ordenados) como campos contínuos.
- Reformulação de campos contínuos como campos ordinais.
- Número total de reformulações.

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

Tabela de Valores Discrepantes

A tabela exibe contagens de como qualquer valor discrepante foi tratado.

- O número de campos contínuos para os quais valores discrepantes foram encontrados e aparados ou o número de campos contínuos para os quais valores discrepantes foram encontrados e configurados como omissos, dependendo de suas configurações no painel Preparar Entradas e Resposta na guia Configurações.
- O número de campos contínuos foi excluído porque eles eram constantes, após o tratamento do valor discrepante.

Uma nota de rodapé mostra o valor de corte do valor discrepante; enquanto uma outra nota de rodapé é mostrada se nenhum campo de entrada (de resposta ou preditor) era contínuo.

Tabela de Valores Omissos

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

- Resposta. Essa linha não é mostrada se nenhuma resposta for especificada.
- Preditores. Isso é mais dividido em número de nominal (conjunto), ordinal (conjunto ordenado) e contínuo.
- O número total de valores omissos substituídos.

Tabela de Destino

A tabela exibe se a resposta foi transformada, mostrada como:

- Transformação Box-Cox para normalidade. Isso é mais dividido 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 a partir do menor para o maior a fim de melhorar a estabilidade.
- Cujas categorias foram mescladas para maximizar a associação com a resposta.
- Cujas categorias foram mescladas para manipular categorias de dispersão.
- Excluídas em razão da baixa a associação com a resposta.
- Excluídas porque elas eram constantes após a mesclagem.

Uma nota de rodapé é mostrada se não havia 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 uma amplitude comum. Além disso, isso mostra o número de preditores transformados usando uma transformação mín-máx, assim como os valores mínimos e máximos especificados.
- Valores do preditor categorizados e o número de preditores categorizados.

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

- Construídos.

- Excluídos em razão da baixa a associação com a resposta.
- Excluídos porque eles eram constantes após a categorização.
- Excluídos porque eles eram constantes após a construção.

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

Pontuações de transformação retrospectiva

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 de volta para a escala original.

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

   Transformar > **Preparar Dados para Modelagem** > **Transformar Escores de Volta...**
2. Selecione um campo para transformar de volta. Esse campo deve conter valores preditos pelo modelo da resposta transformada.
3. Especifique um sufixo para o novo campo. Esse novo campo conterá valores preditos pelo modelo na escala original da resposta não transformada.
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 Dado Automático. Consulte o tópico “Aplicando e Salvando Transformações” na página 16 para obter mais informações

Capítulo 5. 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 no passo de análise de dados exploratória, antes de qualquer análise de dados inferencial. Esse algoritmo foi projetado para detecção de anomalias genéricas; ou seja, a definição de um caso anômalo não é específica de nenhum aplicativo específico, como a detecção de padrões de pagamento incomuns no segmento de mercado de assistência médica ou detecção de lavagem de dinheiro no segmento de mercado de finanças, no qual a definição de uma anomalia pode ser bem definida.

Exemplo. Um analista de dados contratado para construir modelos preditivos para resultados de tratamento de AVC está preocupado com a qualidade de dados, porque esses modelos podem ser sensíveis a observações incomuns. Algumas dessas observações distantes 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 marcar a saída, mas 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 aleatoriamente os casos. 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 interno empírico 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 índice 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.

Sumarizações. 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), frequência e 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 casos 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 de usuário e omissos do sistema.

- **Excluir valores omissos da análise.** 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 for 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 anomalia consistirá, no máximo, de 50 casos, cada um com um valor de índice de anomalia maior ou igual a 2.

Número de grupos de peers. O procedimento procurará 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, o 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 da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 6. Discretização 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 em vez 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 ativo 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 a categorização ideal

1. Nos 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 omissos, 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 tratados usando a exclusão de listwise ou dos pares. Os valores ausentes do usuário são sempre tratados como inválidos. Ao recodificar os valores de variáveis originais em uma nova variável, os valores omissos de 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 guia e de categorização.
- **Listwise** Esta opção opera em todas as variáveis especificadas na guia Variáveis. Se alguma variável estiver omissa 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 $i = 2, ..., n_{final}$ (a segunda categoria de valor mais baixo até a segunda categoria de valor mais alto), calcule

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

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 da sintaxe de comando* para obter informações de sintaxe completa.

Avisos

Essas informações foram desenvolvidas para produtos e serviços oferecidos nos Estados Unidos. Esse material pode estar disponível a partir da IBM em outros idiomas. No entanto, pode ser necessário possuir uma cópia do produto ou da versão do produto nesse idioma para acessá-lo.

É possível que a IBM não ofereça produtos, serviços ou recursos discutidos neste documento 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 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. Pedidos de licença podem ser enviados, por escrito, para:

Gerência de Relações Comerciais e Industriais da IBM Brasil
Av. Pasteur, 138-146
CEP 22290-240
Rio de Janeiro, RJ
Brasil

Para pedidos de 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 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 NÃO-VIOLAÇÃO, COMERCIALIZAÇÃO OU ADEQUAÇÃO A UM DETERMINADO PROPÓSITO. Alguns países não permitem a exclusão de garantias explícitas ou implícitas em certas transações; portanto, esta instrução 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 o(s) produto(s) e/ou programa(s) 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 para esse produto IBM e o uso 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 o mesmo com o objetivo de permitir: (i) a troca de informações entre programas criados independentemente e outros programas (incluindo este) e (ii) o uso mútuo de informações trocadas, devem entrar em contato com:

Gerência de Relações Comerciais e Industriais da IBM Brasil
Av. Pasteur, 138-146
CEP 22290-240
Rio de Janeiro, RJ
Brasil

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. Resultados de desempenho reais podem variar dependendo das configurações específicas e das 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.

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 assuntos, 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 amostra 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 qualquer 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:

© nome de sua empresa) (ano). Partes deste código são derivadas dos Programas de Amostra da IBM Corp.

Marcas comerciais

IBM, o logotipo IBM e ibm.com são marcas comerciais ou marcas registradas da International Business Machines Corp., registradas em muitos países no mundo todo. Outros nomes de produtos e serviços podem ser marcas comerciais 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 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 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 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 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 suas afiliadas.

Índice Remissivo

Impresso no Brasil