FE04

Universidade do Minho
Escola de Engenharia

Curso: Mestrado em Engenharia Biomédica
U.C.: Aprendizagem e Extração do Conhecimento

Ficha de Exercícios 04	
Docente:	Hugo Peixoto | José Machado
Tema:	Data Preparation
Ano Letivo:	2024-2025 – 1° Semestre
Duração da aula:	2 horas

1. Juntar e Agrupar Dados

[1] Colocar os dados no workflow.
[a] Expandir o repositório "Samples" no painel de Repositório. A seguir, expandir a pasta de dados dentro do repositório de exemplo para usar os dados "*Products*" e "*Transactions*".
[b] Arrastar os dados "*Products*" e "*Transactions*" da pasta "Samples - Data" para o painel de Processos.

[2] Juntar os dados.
[a] Procurar o operador "Join" na caixa de pesquisa no topo do painel de Operadores.
Arrastar o operador "Join" para o painel de Processos.
[b] Ligar a porta de saída de "*Retrieve Products*" a uma das portas de entrada do "*Join*" (não importa qual).
[c] Ligar "*Retrieve Transactions*" à outra porta de entrada do "*Join*".
[d] Clicar em "*Join*" para o selecionar. No painel de Parâmetros, encontrar o campo de atributos-chave.
[e] Clicar em "*Edit List*". Selecionar "*Product ID*" para os atributos-chave à esquerda e à direita.

[3] Agrupar os dados para contar compras de produtos.
[a] Arrastar o operador "*Aggregate*" para o processo. Conecta-o à saída de "Join".
[b] Clicar em "*Aggregate*" para o selecionar. Fazer as seguintes alterações no painel de Parâmetros:
[b1] Clicar em "aggregation attributes".
[d] Selecionar "*Customer ID*" na caixa à esquerda e definir a função como "count" na caixa à direita.
[e] Permanecer nesta janela e adicionar outra entrada: "Product Name", com a função definida para "mode".
[f] Clicar em "group by attributes". Depois, selecionar "Product ID" e transportar para a direita.

2. Criar e Remover Atributos

[1] Elimine o operador Aggregate do processo anterior.

[2] Definir um novo atributo:
[a] Adicionar o operador "Generate Attributes".
[b] Ligar ao operador "Join".
[c] Colocar em "Edit List" nos Parâmetros para "Generate Attributes" para definir o novo atributo (coluna). Aparecerá uma janela.
[d] Na coluna da esquerda, insirir "Total" como o nome do atributo.
[e] Na coluna da direita, clicar na calculadora e multiplique "Amount*Price" para a expressão da função.

[3] Remover atributos desnecessários:
[a] Adicionar o operador "Select Attributes" ao processo.Faça as seguintes alterações nos Parâmetros:
[a.1] Definir "attribute filter type" como "subset". Isto significa que o operador será aplicado apenas aos atributos (colunas) que forem especificadas. Isto permite escolher um subconjunto de colunas para manter nos dados - todas as outras serão removidas.
[a.2] Clicar em "Select Attributes".
[a.3] Na janela resultante, selecionar os atributos "Customer ID", "Product Name" e "Total". Se a lista estiver vazia, validar a ligação do operador...
[e] Executar o processo.

Quizz:

- Na vista de Resultados, será possível descobrir o ID do cliente que pagou mais por um único produto? Quanto pagou? Podem ser ordenados os dados clicando no cabeçalho de uma coluna.
- É possível responder à mesma questão usando operadores?
- Como se descreve a forma do gráfico de distribuição para o novo atributo "Total"? É possivel encontrar isso na aba de Estatísticas ou tentar criar um gráfico para esse atributo.

3. Lidar com valores nulos

[1] Preparar os dados.

[a] Arrastar os dados "Titanic" para o processo.

[b] Colocar o cursor sobre a porta de saída e aguardar até que a tooltip mostre os metadados.

[c] Pressionar F3 enquanto a tooltip está visível. Agora, ela transforma-se numa janela, e pode ser deslocada para ver a informação de todas as colunas.

[d] Verificar as colunas com valores em falta.

[2] Remover atributos com muitos valores em falta.

[a] Adicionar um novo operador "Select Attributes".

[b] Liguar o novo operador ao operador "Retrieve" e a saída à prota de resultado "res" à direita.

[c] Nos Parâmetros, alterar o "attribute filter type" para "Subset" e selecionar todos os atributos exceto "Cabin" e "Life Boat". Isso significa que esses dois serão removidos pelo operador.

[d] Executar o processo.

[e] Clicar na aba de Estatísticas e verificar quais atributos com valores em falta ainda restam.

[3] Substituir valores em falta.

[a] Procurar o operador "Replace Missing Values" e adicionar ao processo. Largar na ligação entre "Select Attributes" e a porta de resultado. Desta forma, não é necessário de ligar manualmente os operadores.

[b] Nos Parâmetros deste operador, usar "single" para o tipo de filtro de atributos e selecione "Age" para o atributo.

[c] Executar o processo novamente.

[4] Remover exemplos com valores em falta.

[a] Procurar "Filter Examples" e largar na linha de conexão até ao ponto de resultado. Caso seja perdida a ligaçãop, pode ser ligada manualmente.

[b] Nota o link na parte inferior do painel de Parâmetros, que mostra/esconde parâmetros avançados. Clicar em "Show advanced parameters" para exibir todos os parâmetros do operador.

[c] Novos parâmetros devem aparecer. Definir "condition class" para "no_missing_attributes".

[d] Executar o processo novamente.

Quizz:

- Verificar a aba de Estatísticas - ainda mostra valores em falta para alguma das colunas?
- Por que não é uma boa ideia filtrar logo os exemplos em vez de remover atributos e substituir os valores em falta na "Age"?
- Clicar com o botão direito em "Select Attributes" e desativar a opção "Enable Operator". Fazer o mesmo com "Replace Missing Values". O que é esperado se executar o processo agora? Tentar!
- Quantos exemplos restam no conjunto de dados agora?

4. Normalização e Deteção de Outliers

[1] Preparar os dados.

[a] Arrastar os dados "Titanic" para o processo.

[b] Adicionar o operador "Select Attributes".

[c] Alterar os Parâmetros para remover "Cabin", "Life Boat", "Name" e "Ticket Number".

[2] Normalizar os intervalos de valores dos atributos.
[a] Adicionar o operador "Normalize".

[3] Detetar outliers:
[a] Procurar o operador "Detect Outlier (Distances)", adicionar e ligar ao "Normalize". Usar as configurações padrão.

[4] Remover outliers do conjunto de exemplos:
[a] Adicionar "Filter Examples" ao processo e conecte-o ao operador anterior e também à porta de resultado à direita.
[b] Nos seus Parâmetros, adicionar um novo filtro com "Outlier equals false".
[c] Executar o processo.

Quizz:

- Como seria possível mudar o processo para encontrar 20 outliers em vez de 10?
- Como seria possível alterar o processo para mostrar apenas outliers em vez de os remover?
- Substitir o operador de deteção de outliers por "Detect Outlier (LOF)" e adicionar um ponto de interrupção após este operador antes de executar. Qual a diferença em relação a antes?