

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

Ficha de Exercícios 02	
Docente:	Hugo Peixoto José Machado
Tema:	Preparação dos Dados
Ano Letivo:	2026-2027 – 1º Semestre
Duração da aula:	4 horas

Enunciado

Um serviço de engenharia clínica disponibiliza dados do seu parque de equipamentos hospitalares e das leituras periódicas dos respetivos sensores. Os dados vêm em duas tabelas relacionadas: equipamentos.csv (uma linha por equipamento) e leituras.csv (várias leituras por equipamento, ligadas por id_equipamento). A descrição das variáveis de cada tabela está disponível no ficheiro de apoio. O objetivo desta ficha é preparar um dataset final, limpo e integrado, pronto para modelação, replicando a fase de Data Preparation do CRISP-DM. Todo o trabalho é feito no Altair AI Studio. No fim, guarde o processo e exporte o dataset final.

Antes de começar: faça primeiro uma breve Compreensão dos Dados sobre cada ficheiro, para localizar os problemas que vai ter de resolver.

Passo 01 – Carregar os dados e remover duplicados

[1] Importe equipamentos.csv e leituras.csv para o AI Studio (Import Data).

[2] Conte o número de linhas de equipamentos e o número de id_equipamento distintos. Vai encontrar 540 linhas mas um número menor de equipamentos únicos. Qual a diferença, e o que significa?

[3] Adicione o operador Remove Duplicates sobre equipamentos. Quantas linhas foram removidas?

[4] Em leituras, há dois atributos que não acrescentam informação para a modelação ao nível do equipamento. Identifique-os e remova-os com Select Attributes.

Passo 02 – Normalizar categorias inconsistentes

[1] No separador Statistics, verifique os valores distintos de criticidade e de em_uso. Quantos valores aparecem em cada? Quantos deveriam existir?

[2] Com o operador Map (ou Replace) sobre criticidade, uniformize todas as grafias para apenas alta e baixa.

[3] Faça o mesmo sobre em_uso, de modo a ficar apenas com sim e nao.

[4] Confirme nas Statistics que cada atributo ficou só com 2 valores.

Passo 03 – Tratar valores impossíveis

[1] Usando mínimos/máximos das Statistics, identifique valores impossíveis (não apenas extremos) em: ano_instalacao, horas_uso_dia, temperatura_operacao_c e custo_aquisicao_eur.

[2] Faça o mesmo em leituras para pressao_bar e corrente_a (procure valores negativos).

[3] Substitua cada valor impossível por missing (por exemplo, com Declare Missing Value ou um filtro adequado). Justifique porque é preferível marcar como missing em vez de apagar a linha inteira.

Passo 04 – Lidar com valores em falta

- [1] Quantifique, nas Statistics, a percentagem de valores em falta por atributo (já incluindo os que marcou no passo anterior). Qual o atributo mais afetado em cada tabela?
- [2] Aplique o operador Replace Missing Values aos atributos numéricos. Escolha uma estratégia (ex.: média ou mediana) e justifique a escolha para um atributo claramente enviesado.
- [3] Confirme que deixaram de existir valores em falta nos atributos tratados.

Passo 05 – Construir atributos (combinar duas features numa)

- [1] Os atributos custo_pecas_eur e custo_maoobra_eur representam, separadamente, duas parcelas do mesmo custo. Faz sentido combiná-los num único atributo.
- [2] Com o operador Generate Attributes, crie: custo_manutencao_eur = custo_pecas_eur + custo_maoobra_eur.
- [3] Depois de criar o atributo derivado, remova as duas colunas originais (custo_pecas_eur, custo_maoobra_eur) com Select Attributes.

Passo 06 – Normalizar

- [1] Aplique o operador Normalize (método z-transformation) aos atributos numéricos de equipamentos (horas_uso_dia, temperatura_operacao_c, consumo_medio_kwh, custo_aquisicao_eur, custo_manutencao_eur, ano_instalacao).
- [2] Depois da normalização, qual passa a ser (aproximadamente) a média e o desvio padrão de cada atributo? Porque é que este passo é útil antes da modelação?

Passo 07 – Detetar e remover outliers

- [1] Adicione o operador Detect Outlier (Distances). Repare que coloca 10 outliers por defeito. Faz sentido esse número para estes dados? Porquê?
- [2] Em alternativa, use o Detect Outlier (LOF) e observe a coluna de outlier score. Ordene os valores e procure o salto (a diferença grande entre um grupo de scores altos e o resto).
- [3] Com base nesse salto, defina um critério por score (ex.: outlier se $LOF > k$) em vez de um número fixo. Quantos equipamentos são apanhados? O que têm de invulgar (cruze com horas_uso_dia, temperatura_operacao_c e consumo_medio_kwh)?
- [4] Depois de detetar e remover os outliers, remova a coluna outlier criada pelo operador (com Select Attributes).

Passo 08 – Agregar e juntar os dois datasets

- [1] equipamentos tem 1 linha por equipamento, mas leituras tem várias. Antes do join, é preciso resolver a diferença de granularidade. Como lidar com várias leituras por equipamento?
- [2] Com o operador Aggregate, calcule, por id_equipamento, a média de pressao_bar, rpm_motor, corrente_a e ruido_db. Use Rename para dar nomes claros às novas colunas (ex.: pressao_media).

[3] Verifique a integridade referencial: existe algum `id_equipamento` em leituras que não existe em equipamentos? O que fazer a essa leitura órfã?

[4] Junte as leituras agregadas a equipamentos com o operador `Join` (por `id_equipamento`). Que tipo de `join` garante que mantém todos os equipamentos? Confirme que o resultado continua com 1 linha por equipamento.

Passo 09 – Guardar o dataset final

[1] Exporte o dataset final com o operador `Write CSV`.

[2] Agrupe todos os passos anteriores num subprocesso de "Preparação dos Dados", para o processo ficar legível e reutilizável.

[3] Guarde o processo no AI Studio. O processo é tão importante como o dataset: é o que garante que a preparação é reproduzível.

Síntese

[1] Preencha, em três linhas, que problemas de qualidade encontrou em cada tabela e como os resolveu.

[2] O dataset final está pronto para a fase de Construção de Modelos (próxima aula). Formule uma hipótese (uma frase) sobre que atributos parecem mais promissores para prever avaria.