

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

Ficha de Exercícios Exemplo	
Docente:	Hugo Peixoto José Machado
Tema:	Aprendizagem e Extração do Conhecimento
Ano Letivo:	2025-2026 – 1º Semestre
Duração da aula:	2 horas

1. Introdução

Uma unidade hospitalar pretende identificar doentes com maior risco de **readmissão a 30 dias**, com o objetivo de planear intervenções preventivas simples (ex.: follow-up precoce, educação na alta).

Foi disponibilizado um conjunto de dados sintético, distribuído por vários ficheiros, representando diferentes fontes de informação hospitalar. Nomeadamente:

- a) 01_identificacao_demografia
- b) 02_informacao_clinica
- c) 03_internamento
- d) 04_processo_assistencial
- e) 05_variaveis_problematicas
- f) 06_target_readmissao

O objetivo não é criar um modelo “perfeito”, mas extrair conhecimento útil, explicável e eticamente defensável.

2. Business Understanding

[1] Explique, em 3–5 linhas, qual é o objetivo real do modelo neste contexto hospitalar.

[2] Indique duas consequências negativas de um modelo que tenha uma accuracy muito elevada, mas seja pouco explicável.

3. Data Understanding

[1] Depois de integrar os vários ficheiros usando uma pipeline do AI Studio:

- a) qual é a variável alvo?
- b) o problema é de classificação, regressão ou outro tipo?

[2] A variável `readmissao_30d` apresenta desbalanceamento. Devido a esse facto explique:

- a) Porque isto é esperado em contexto clínico.
- b) Porque a accuracy (acuidade) não é uma métrica adequada neste caso.

4. Data Preparation

[1] Explique porque razão as seguintes variáveis não devem entrar no modelo, mesmo existindo nos dados:

- `id_episodio`
- `codigo_servico_antigo`

[2] Deverá remover a variável `score_risco_previo` antes da modelação. Explique o conceito de data leakage e porque esta variável o poderá provocar.

[3] Utilizando o operador “Generate Attributes” deverá aplicar limites aos valores extremos das variáveis *dias_internamento*, e *tempo_espera_urgencia*.

- Explique porque não foi correto apagar os registos extremos mas sim preferível limitar os valores.

[4] Deverá aplicar o operador “Nominal to Numerical” às variáveis:

- servico
- tipo_admissao
- destino_alta
- sexo
- diagnostico_principal

Porquê?

5. Modeling

[1] Deverá usar Cross Validation (10 folds) e testar com dois modelos:

- a) Regressão Logística
- b) Decision Tree

[2] Deverá comparar os resultados de accuracy e recall (da classe positiva).

- a) Explique qual o resultado mais realista e aceitável num contexto clínico.
- b) Explique duas razões pelas quais os resultados devem ser visto com desconfiança.

6. Evaluation

[1] Explique a diferença entre falso positivo e falso negativo. Qual deles é mais crítico neste problema? Justifique do ponto de vista clínico.

[2] Indique três variáveis que considere acionáveis, ou seja, que permitam intervenção prática para reduzir o risco de readmissão.

[3] Explique porque este modelo pode ser usado como apoio à decisão e não deve ser usado para decidir altas automaticamente.