

| 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.

O objetivo do modelo é identificar doentes com maior risco de readmissão a 30 dias, de forma a apoiar o planeamento de intervenções preventivas simples (ex.: follow-up precoce, reforço de educação na alta). O modelo não substitui a decisão clínica, funcionando apenas como apoio à priorização de doentes com maior risco.

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

**Perda de confiança clínica:** modelos não explicáveis são difíceis de aceitar por profissionais de saúde, limitando a sua utilização prática.  
**Risco ético e organizacional:** decisões baseadas em modelos opacos podem reforçar vieses ou erros sem que seja possível identificar a sua origem.

## 3. Data Understanding

[1] Depois de integrar os vários ficheiros usando uma pipeline do AI Studio:  
a) qual é a variável alvo?

A variável alvo é readmissao\_30d.

b) o problema é de classificação, regressão ou outro tipo?

Trata-se de um problema de classificação binária, uma vez que o objetivo é prever se ocorre ou não readmissão a 30 dias.

[2] A variável `readmissao_30d` apresenta desbalanceamento, comprove isso com um histograma. Devido a esse facto explique:

a) Porque isto é esperado em contexto clínico.

A maioria dos doentes não é readmitida, sendo a readmissão um evento relativamente raro, o que é consistente com a prática clínica real.

b) Porque a accuracy (acuidade) não é uma métrica adequada neste caso.

Num cenário desbalanceado, um modelo pode obter elevada accuracy simplesmente prevendo sempre a classe maioritária, sem identificar corretamente os doentes de risco. Assim, métricas como recall da classe positiva são mais informativas.

## 4. Data Preparation

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

- `id_episodio`
- `codigo_servico_antigo`

**id\_episodio:** É um identificador técnico, sem significado clínico, que pode levar o modelo a memorizar casos específicos, comprometendo a generalização.

**codigo\_servico\_antigo:** Representa uma codificação administrativa obsoleta, sem relevância clínica atual, introduzindo ruído e reduzindo a interpretabilidade do modelo.

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

Data leakage ocorre quando o modelo utiliza informação que não estaria disponível no momento real da decisão.

A variável `score_risco_previo` resulta de uma avaliação de risco anterior e está fortemente correlacionada com o outcome, permitindo ao modelo “antecipar” a resposta, inflacionando artificialmente o desempenho.

[3] Utilizando o operador “Generate Attributes” deverá aplicar limites aos valores extremos das variáveis `dias_internamento`, e `tempo_espera_urgencia`. Para observar esses valores deverá utilizar o histograma.

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

Não foi correto apagar os registos extremos porque estes representam casos clinicamente plausíveis, frequentemente associados a maior gravidade.

A limitação (Winsorization manual) permite reduzir o impacto dos extremos no modelo, preservando a informação e evitando a introdução de viés por exclusão de doentes complexos.

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

- `servico`
- `tipo_admissao`
- `destino_alta`

- sexo
- diagnostico\_principal

Porquê?

Estas variáveis são nominais, não possuem ordem natural e não podem ser usadas diretamente pelos algoritmos de aprendizagem.

A conversão para formato numérico (ex.: codificação dummy) permite ao modelo utilizá-las corretamente, evitando a criação de ordens artificiais entre categorias.

## 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.

A Regressão Logística, apesar de apresentar menor accuracy, mostra valores de recall mais plausíveis e comportamento mais estável, sendo mais adequada para contexto clínico.

- b) Explique duas razões pelas quais os resultados devem ser visto com desconfiança.

**Sobreajustamento:** árvores de decisão podem memorizar padrões específicos do dataset.

**Reconstrução da regra geradora do target:** em dados sintéticos, a árvore pode aprender regras quase determinísticas, não generalizáveis a novos dados.

## 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.

**Falso positivo:** o modelo identifica risco quando não há readmissão.

**Falso negativo:** o modelo não identifica um doente que será readmitido.

Neste problema, o falso negativo é mais crítico, pois corresponde a um doente de risco que não recebe intervenção preventiva.

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

Uma variável acionável é uma variável que:

- pode ser modificada por uma decisão humana ou organizacional
- antes do outcome acontecer
- e cuja alteração pode reduzir o risco

Exemplos de variáveis não acionáveis

- idade
- sexo
- diagnóstico principal

- índice de comorbilidade

são importantes para estratificar risco, mas não podem ser mudadas.

Neste caso exemplos de variáveis acionáveis:

- **followup\_7d**: Indica se o doente teve consulta ou contacto de seguimento até 7 dias após a alta.

Se o modelo mostra que ausência de follow-up ↑ risco de readmissão, então a intervenção é clara: “Garantir follow-up precoce para doentes de risco.”

- **educacao\_alta**: Se o modelo associa ausência de educação → maior readmissão, então: “Reforçar educação na alta em doentes sinalizados.”

- **alta\_planeada**: Indica se a alta foi planeada ou ocorreu de forma reativa (ex.: pressão de camas)

Estas permitem intervenções organizacionais ou clínicas diretas.

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

O modelo pode ser usado como apoio à decisão, ajudando a priorizar doentes para intervenção.

Não deve ser usado para decidir altas automaticamente, pois:

- não incorpora todo o contexto clínico,
- pode cometer erros com impacto direto no doente,
- deve complementar, e não substituir, o julgamento clínico.