FE05

Universidade do Minho
Escola de Engenharia

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

Ficha de Exercícios 05	
Docente:	Hugo Peixoto | José Machado
Tema:	AI Studio: Modelos de Classificação
Ano Letivo:	2026-2027 – 1º Semestre
Duração da aula:	2 horas

Introdução

Nesta ficha de exercícios irá aplicar o processo completo da metodologia CRISP-DM para resolver um problema de classificação utilizando o AI Studio. Não serão dados passos detalhados: espera-se que, com base nos conhecimentos adquiridos nas aulas anteriores, consiga tomar decisões fundamentadas em cada etapa do processo.

O dataset a utilizar é o heart_disease_uci.csv. Este dataset foi obtido a partir do repositório UCI/Kaggle e descreve fatores de risco para doenças cardíacas. O atributo num representa a classe:

classe 0: ausência de doença cardíaca
classes 1, 2, 3, 4: presença de doença cardíaca (graus crescentes)

O objetivo é construir um modelo de classificação que preveja a presença ou ausência de doença cardíaca com base nos restantes atributos.

Business Understanding

[1] Defina, numa frase, o objetivo de negócio deste problema.

[2] Defina, numa frase, o objetivo do processo de Data Mining.

[3] Que tipo de problema de Data Mining está em causa? Justifique a escolha do tipo de modelo a aplicar.

Data Understanding

Importe o dataset para o AI Studio e familiarize-se com os dados.

[4] Para cada atributo, documente: tipo, percentagem de valores ausentes, mínimo, máximo, média, moda e desvio padrão (quando aplicável). Apresente os resultados em formato de tabela.

[5] Existem atributos com registos únicos (valores que só aparecem numa instância)? Que implicações pode isso ter na modelação?

[6] Analise os histogramas dos atributos numéricos. Descreva informalmente como alguns atributos parecem influenciar o risco de doença cardíaca. Documente pelo menos 3 observações relevantes.

[7] Utilizando gráficos de dispersão (Scatter), investigue:
[a] O atributo thalach parece estar associado a doenças cardíacas? E o atributo ca?
[b] Observando o gráfico thalach (X) / oldpeak (Y), identifique possíveis regiões "densas" associadas à presença de doença cardíaca.

Data Preparation

Prepare os dados para a modelação. Todas as decisões devem ser justificadas.

[8] Tratamento de valores ausentes: escolha uma estratégia adequada e justifique. Não elimine registos sem fundamentação.

[9] Avaliação do peso dos atributos: investigue a utilização do operador "Weight by Information Gain" para selecionar um subconjunto de atributos com boa capacidade de previsão. Considere um limiar de peso adequado e justifique-o.

[10] Deteção de outliers: aplique um método de deteção de outliers. Quantos foram encontrados? Decida se devem ser removidos e justifique.

[11] Normalização: avalie se a normalização dos dados é necessária para os modelos que pretende usar. Aplique-a se justificável.

[12] Construa um processo que combine os passos anteriores (preparação completa) e guarde-o no AI Studio.

Modeling

Aplique algoritmos de classificação aos dados preparados. Para cada modelo, utilize tanto Split Data (70/30) como Cross Validation (10 folds).

[13] Comece com o classificador Decision Tree.
- [a] Compare os atributos identificados como importantes pela árvore com as conclusões da fase de Data Understanding.
- [b] Compare os resultados obtidos com Split Data e Cross Validation. Explique eventuais diferenças.
- [c] Teste o modelo com e sem pruning. Qual se comporta melhor? Justifique.

[14] Aplique pelo menos mais 2 algoritmos de classificação à sua escolha (e.g., Naive Bayes, k-NN, Random Forest). Documente os resultados de cada um.

[15] Para cada modelo, registe: accuracy, precision, recall e a matriz de confusão.

Evaluation

[16] Compare todos os modelos construídos. Resuma os resultados numa tabela com as métricas de desempenho para cada classificador e cada estratégia de validação.

[17] Calcule a baseline (modelo ingénuo que prevê sempre a classe maioritária). Algum dos seus modelos é significativamente melhor que a baseline?

[18] No contexto clínico deste problema, que métrica considera mais importante: accuracy, precision ou recall? Justifique a sua escolha considerando as consequências de falsos positivos vs. falsos negativos.

[19] Que modelo recomendaria para utilização clínica? Apresente a sua justificação de forma estruturada.

Reflexão Final

[20] Identifique pelo menos duas limitações do processo que desenvolveu e proponha melhorias concretas.

[21] Se tivesse mais dados ou atributos disponíveis, que informação adicional consideraria útil para melhorar o modelo? Justifique.