FE08

Universidade do Minho
Escola de Engenharia

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

Ficha de Exercícios 08	
Docente:	Hugo Peixoto | José Machado
Tema:	Extração de conhecimento com Python
Ano Letivo:	2024-2025 – 1º Semestre
Duração da aula:	2 horas

O Data Set usado neste exercício é sobre doenças cardíacas disponível no ficheiro heart-c.csv, obtido através do repositório da Kaggle:

https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data?select=heart_disease_uci.csv

Este Data Set descreve fatores de risco para doenças cardíacas. O atributo ***num*** representa o atributo da classe (binominal):

class 0- nenhuma doença
class 1,2,3,4 - aumento do nível de doença cardíaca.

O principal objetivo deste exercício é prever doenças cardíacas a partir de outros atributos no Data Set. Obviamente, trata-se de um problema de classificação. A linguagem de programação a ser usada será o Python, através do GoogleColab (https://colab.google/). A descrição deste exercício é gradual. Portanto, espera-se que possa entender melhor os vários aspetos e questões envolvidos no processo de Extração de Conhecimento.

1. Compreensão dos Dados

[1] Carregar os dados e imprimir as primeiras linhas. Examinar a estrutura básica e identificar o tipo de dados de cada atributo.

[2] Verificar a existência de valores em falta e identificar colunas que necessitem de limpeza ou transformação adicional.

[3] Calcular estatísticas básicas (média, mediana, etc.) e analisar a distribuição de cada atributo.

2. Exploração dos Dados

[1] Explore a distribuição da classe. Qual a variação entre casos positivos (doença) vs casos negativos (sem doença). Converta a classe em 0 – sem doença; 1- com doença.

[2] Compare diferentes atributos (*age*, *ejection fraction*, *sérum sodium*, *oldpeak*) com a classe. Recorra por exemplo a gráficos de caixa (*boxplot*).
a) Use histogramas (*histplot*) para comparar *thalach* e *ca* com a possibilidade de doença.

[3] Identifique padrões ou correlações entre os atributos, utilizando matrizes de correlação. Não utilize os atributos '*id*' e '*num*'.

3. Processamento de Dados

[1] Tratar os valores em falta identificados na fase de compreensão dos dados (Remoção vs substituição de nulos). Caso os dados em falta sejam superiores a 30%, deverá remover o atributo.

[2] Escalar ou normalizar características numéricas para garantir consistência nos dados para a fase de modelação.

[3] Codificar variáveis categóricas, se existirem, em valores numéricos.

[4] Remover atributos desnecessários.

4. Modelação

[1] Dividir o conjunto de dados em conjuntos de treino e teste (80/20).

[2] Construir e treinar um modelo de classificação simples (ex.: LogisticRegression) para prever a variável alvo.

[3] Construir e treinar 2 modelos de classificação mais complexos (ex.: Random Forest, SVM).

5. Avaliação

[1] Avaliar o desempenho de cada modelo utilizando métricas como Precisão, Recall e F1-score.

[2] Avaliar a acuidade global de cada modelo. Recorrer a um gráfico de barras para apresentar a comparação de resultados.

[3] Gerar uma matriz de confusão para cada modelo e interpretar os resultados. Tentar aprimorar a impressão da matriz de confusão.

[4] Voltar o ponto 4 – Modelação, e calcular a acuidade dos 3 modelos usando com *Cross-Validation*. Aumentou ou diminuiu?