Mestrado em Engenharia Informática
Ciência de Dados
Ano Letivo 2024/2025

Mini Projeto III

Trabalho Individual
Data Wrangling and Exploratory Data Analysis

© Ricardo Campos
ricardo.campos@ubi.pt

O trabalho prático é obrigatório para a obtenção de aprovação na unidade curricular. A não entrega durante o prazo previsto implica a automática reprovação dos alunos.

Objetivo: Familiarização com o processo de limpeza de dados e análise exploratória de dados. Em concreto, espera-se que o aluno desenvolva competências ao nível da:

- Importação e exploração de dados com `pandas`
- Identificação e tratamento de valores em falta
- Criação de novas variáveis
- Análise exploratória e visualização
- Formulação de hipóteses e geração de insights
- Utilização crítica e experimental de LLMs na ciência de dados

Entrega: Os trabalhos (em formato notebook – devidamente documentados) devem ser inseridos na plataforma de e-learning (moodle) até 13/05/2025, 23h59. O nome do notebook a submeter no moodle deve cumprir com o seguinte formato: *XXXX.ipynb*, onde *XXXX* é o número do aluno (e.g., *10000.ipynb*).

Realização do trabalho: Os trabalhos devem ser realizados individualmente.

Tarefa 1: Familiarização com *Data Wrangling and Exploratory Data Analysis*
Considere um ficheiro *csv* à sua escolha. Cada aluno deve escolher um ficheiro .csv com dados reais (de fontes como Kaggle, UCI, data.gov, etc.). O objetivo é explorar e transformar os dados de forma a extrair conhecimento relevante, colocando em prática as etapas iniciais de um pipeline de ciência de dados. Nota: não serão admitidos trabalhos

tendo por base o mesmo ficheiro csv. Nesse sentido, solicita-se que indique a localização/origem do seu ficheiro (*first come – first serve*) no link abaixo: https://1drv.ms/x/s!AqbUf6ry5g9tlEG8UZfAp0ENctR5?e=DkxuHQ.

Proceda à descrição do *dataset* e das suas principais colunas. Responda às seguintes questões tendo por base o ficheiro.

1. Proceda à importação do ficheiro para um Pandas *dataframe*.
2. Mostre os 5 primeiros registos.
3. Mostre o coeficiente de correlação de *pearson* entre cada par de atributos. Liste os valores de correlação de forma descendente para um atributo à sua escolha.
4. Devolva a mediana de um atributo à sua escolha (restringindo a um conjunto de dados específico. Exemplo: a mediana da idade das pessoas do sexo feminino).
5. Escreva o código que lhe permite contabilizar o número de registos *null* existente num conjunto de colunas à sua escolha.
6. Desenvolva uma função de imputação que proceda à substituição dos valores nulos de uma coluna à sua escolha com o valor da mediana desse atributo. Considere, sempre que possível, diferentes valores de mediana para cada classe (por exemplo, proceda à substituição dos valores nulos da coluna *Age* de acordo com a mediana da *Age* apurada para cada uma das três classes existentes (*Pclass* = 1, *Pclass* = 2, *Pclass* = 3)).
7. Crie novas colunas no seu *dataset*, potencialmente relacionadas com as colunas atuais (exemplo, a coluna *Title* (com os valores Mr; Miss; etc) a partir da coluna *Name* (que inclui os valores Mr. Santos; Miss Filipa).
8. Proceda a uma análise exploratória de dados que considere relevante no contexto do seu ficheiro e que revele padrões relevantes. Seja criativo.
9. Formule duas hipóteses baseadas no seu dataset e teste-as com os dados (por exemplo: "A renda média é maior entre pessoas com ensino superior?" ou "A taxa de churn é maior entre clientes mais jovens?").
10. Explore a utilização de LLMs para complementar a sua análise (responda a pelo menos duas das três questões abaixo indicadas).
    10.1. Use um LLM (ChatGPT, Gemini, Claude, etc.) para explicar uma tabela, gráfico ou insight e compare com a sua explicação.

10.2. Descreva uma tarefa em português (ex: "quero criar uma coluna com a idade em décadas") e peça ao LLM para escrever o código correspondente. Avalie o resultado obtido.

10.3. Se tiver colunas de texto (ex: comentários), peça ao LLM para sugerir rótulos (e.g., positivo/negativo).

11. Outras operações que considere relevantes para valorizar o seu trabalho.