# Trabalho de grupo IMD

Licenciatura em Ciência de Dados (2º ano)

22 de Fevereiro de 2025

Conceição Figueiredo (conceicao.figueiredo@iscte-iul.pt)
João Araújo (Joao.Miguel.Araujo@iscte-iul.pt )
Rodrigo Alvarez (Rodrigo_Alvarez@iscte-iul.pt )

**Deadline entrega:** 21 de março de 2025

**Tema:Previsão da percentagem da despesa total alocada para alimentação num agregado familiar de Espanha (por mês)** - desenvolver um modelo de regressão múltipla que permita estimar a respetiva percentagem com base nas variáveis contidas na base de dados (BudgetFood.csv ).

Base de dados: BudgetFood.csv
Variáveis:

- wfood: percentagem da despesa total que uma família gasta para alimentação /por mês
- totexp: despesas totais do agregado familiar (moeda em peseta Espanhola, transformar em euro, sabendo que 1000 pesetas = 6 euros)
- age: idade da pessoa de referência do agregado familiar

- size: número de elementos do agregado familiar
- town: tamanho da cidade onde a família tem o domicilio (5 categórias: 1 – cidade pequena,...,5 - cidade grande)
- sex: sexo da pessoa de referência do agregado familiar (man, woman)

A base de dados BudgetFood.csv tem 23972 linhas (famílias).

Cada grupo de trabalho vai considerar subamostras com **2500 observações**, na ordem indicada no ficheiro Grupos_IMD_2024_25.

*(Portanto, (não esquecer!) devem extrair a vossa subamostra da base de dados total e transformar a moeda - de pesetas para euro, considerando 1000 pesetas = 6 euros)*

Tirar tempo para procurar o máximo de informação de contexto que permita a **compreensão** dos **dados** e do **domínio** de onde provêm.

Neste trabalho recomenda-se usar o software R (script em RStudio ou Notebook em Jupyter Notebook, IRkernel).

## Pontos principais:

O desenvolvimento do projeto e do relatório final deve seguir a metodologia CRISP-DM.

1. Com base nos dados e a sua pesquisa sobre os dados (*Business Understanding*), definir o(s) *problema(s)*/*questões* que gostariam de *resolver*/*responder* e justificar a escolha das variáveis.
2. Fazer a limpeza dos dados.
3. Analisar gráficos e estatísticas descritivas das variáveis.
4. Analisar correlação e causalidade entre as variáveis.

5. Pré-processamento dos dados e manipulação de *features* (agrupar, juntar, eliminar, transformar as variáveis).
6. Usar algoritmos de aprendizagem supervisionada (regressão linear, regressão polinomial, interação de variáveis, regressão não-linear) sobre o seu conjunto de dados.
7. Dividir o dataset em conjuntos de treino e de teste.
8. Validar o modelo escolhido e fazer a previsão da variável dependente/alvo (sobre o conjunto de teste).
9. Avaliar a performance da previsão feita (sobre o conjunto de teste).
10. Interpretação/explicação dos resultados obtidos a partir dos dados (prós e contras).
11. **Relatório**: podem usar *dashboards*, editores de texto, ou outras ferramentas.
12. **O relatório final deve ser enviado por e-mail em formato pdf. Devem ainda submeter um ficheiro zippado com o dataset final e o ficheiro com o código utilizado.**
13. O **código** tem de poder ser executado AS-IS (diretamente) e deverá estar devidamente comentado.
14. Deverá incluir todas as experiências feitas, incluindo a análise e preparação de dados, modelação, e avaliação do modelo.
15. A interpretação dos resultados deve ser efetuada de forma crítica e não apenas factual.
16. O relatório tem um **limite de 25 páginas**.