POLITÉCNICO DE LEIRIA | ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO

Departamento de Matemática

Análise Exploratória de Dados

Mestrado em Ciência de Dados

1.º Ano – 1.º Semestre 2025/2026

Enunciado do trabalho escrito

- O enunciado do trabalho é constituído por duas páginas.
- Cada grupo é constituído por 3 elementos[1] e deverá escolher uma base de dados para trabalhar, a qual deverá:
  - ser distinta da utilizada pelos restantes grupos,
  - ter pelo menos uma variável nominal,
  - ter pelo menos uma variável ordinal,
  - ter pelo menos 8 variáveis quantitativas e
  - ser constituída por pelo menos 200 indivíduos.
- Os grupos devem ser formalizados utilizando o formulário disponível em INSCRIÇÃO (*link* igualmente disponível no Moodle) até ao dia **28 de novembro de 2025**.
- A base de dados escolhida deve ser remetida por email (rui.santos@ipleiria.pt) até ao dia **28 de novembro de 2025** para validação.
- Os resultados do trabalho deverão ser obtidos através da utilização da linguagem R em ambiente RStudio.
- Deve ser elaborado um ficheiro *script* com todas as operações realizadas na elaboração do trabalho, o qual deverá ser submetido juntamente com o relatório do trabalho.
- O relatório, redigido e estruturado de forma adequada a ser entregue a um cliente que encomendou o presente trabalho, deve conter:
  - a identificação (número e nome) de todos os estudantes que realizaram o trabalho;
  - os resultados utilizados na fundamentação das respostas, devidamente enquadrados no texto (não devem ser colocados no relatório resultados que não sejam utilizados na análise apresentada);
  - a interpretação e a fundamentação de todas as conclusões enunciadas;
  - gráficos adequados, que ajudem na interpretação e/ou na fundamentação das ideias apresentadas;
  - a validação de todos os pressupostos associados às metodologias utilizadas;
  - uma análise crítica à qualidade dos modelos utilizados;
  - no máximo 15 páginas (não são contabilizadas as páginas da capa, eventuais índices, referências bibliográficas ou anexos).
- A apresentação, o rigor e a clareza da exposição são elementos importantes na apreciação do trabalho.
- A submissão do trabalho consiste no envio, via plataforma Moodle, da base de dados utilizada em formato `RData`, do relatório em formato PDF e do ficheiro *script* do RStudio.
- Prazo limite para a submissão: **05 de janeiro de 2026**.
- Prova oral associada ao trabalho: **07 de janeiro de 2026**.

[1] A constituição de grupos com um número de elementos distinto terá de ser previamente autorizado, sendo solicitado por email para rui.santos@ipleiria.pt.

O trabalho deve incluir as quatro seguintes análises estatísticas.

1. **[6 val.]** Análise exploratória geral da base de dados, com a descrição pormenorizada de duas variáveis qualitativas (uma nominal e uma ordinal) e de duas variáveis quantitativas. Estas quatro variáveis são escolhidas pelo grupo, entre as variáveis da base de dados. Em particular, esta análise deve incluir:
   - a manipulação dos dados (preparação da base de dados para o tratamento estatístico);
   - a análise à existência de observações omissas;
   - a apresentação de tabelas, gráficos e medidas adequadas ao resumo da informação das 4 variáveis escolhidas para análise (uma nominal, uma ordinal e duas quantitativas);
   - a apresentação de intervalos com 95% de confiança para a percentagem de indivíduos em cada categoria da variável nominal;
   - a análise se alguma das duas variáveis quantitativas escolhidas pode ser caracterizada por uma distribuição normal;
   - a comparação dos valores observados nas duas variáveis quantitativas nas diferentes categorias da variável ordinal, incluindo uma comparação da sua variabilidade em cada categoria, bem como do seu valor médio;
   - a análise da independência/associação entre as duas variáveis qualitativas (as variáveis nominal e ordinal escolhidas);
   - a análise da correlação entre as duas variáveis quantitativas escolhidas.

2. **[4 val.]** Análise fatorial em componentes principais a todas as variáveis quantitativas da base de dados, de forma a investigar uma possível redução da dimensão dos dados, nomeadamente:
   - a análise da correlação entre as variáveis e do índice KMO (Kaiser-Meyer-Olkin);
   - a fundamentação do número de componentes utilizadas, com referência à variância explicada;
   - a apresentação de gráficos que ilustrem as conclusões, avaliando se as componentes principais permitem distinguir os indivíduos das diferentes categorias de cada uma das duas variáveis qualitativas utilizadas na primeira questão.

3. **[6 val.]** Análise de *clusters* a todas as variáveis quantitativas da base de dados, em particular:
   - a aplicação de um método hierárquico (apresentando o respetivo dendrograma);
   - a aplicação de um método não hierárquico (*k-means*);
   - a justificação do número de *clusters* escolhido (em cada método);
   - a comparação dos grupos obtidos com os grupos definidos por cada uma das duas variáveis qualitativas utilizadas na primeira questão.

4. **[4 val.]** Aplicação do algoritmo de classificação *naïve* Bayes, em particular:
   - a aplicação do algoritmo para classificar (isoladamente) cada uma das variáveis qualitativas utilizadas na primeira questão;
   - a análise da fiabilidade das classificações obtidas, utilizando a matriz de confusão e medidas associadas.

Qualquer dúvida sobre o trabalho deve ser remetida por email para rui.santos@ipleiria.pt.