Trabalho Prático

Análise de Dados em Informática

Engenharia Informática - 3º ano 2º semestre
Ano Letivo 2024/2025

---

1. Objetivos

Objetivo Geral:

- Análise Exploratórias de Dados
- Análise Inferencial
- Correlação e Regressão

Objetivos específicos:

- Definir a metodologia de trabalho
- Análise e discussão dos resultados com recurso ao Phyton
- Escrita de Artigo Técnico com a Análise de Dados

2. Calendarização

Lançamento das propostas de trabalhos: até 1 de março de 2025
Entrega do trabalho: até **29 de março de 2025** (23:55)
Defesa e discussão: em data a marcar pelo professor de TP

3. Normas

- O grupo (**máx 3 elementos**) deve ser o mesmo nas 2 iterações do Trabalho Prático.
- Deverá ser usado o Phyton como ferramenta de suporte ao tratamento de dados.
- A **data final de ENTREGA** do 1º Trabalho Prático é **29 de março de 2025**, no moodle. Independentemente deste prazo, os grupos deverão ser capazes de, quando o professor o solicitar, reportar o estado de desenvolvimento do trabalho.
- A entrega do trabalho consta de um Artigo Científico. Deverá submeter todos os documentos num ficheiro compactado. O zip file deve conter:
    - artigo científico em pdf
    - dados utilizados em formato csv
    - script completo (e comentado) do código criado em Phyton para resolver o problema proposto
- O nome do ficheiro deverá seguir a seguinte notação:

  ANADI_YYY_XXX_Nºaluno1_Nºaluno2_Nºaluno3.zip, onde **YYY** representa a sigla do docente das TP, e **XXX** representa a turma TP.

  Exemplo: **ANADI_AIM_3DA_7777777_8888888_9999999.zip.**
- Trabalhos cuja designação não respeite a notação indicada, **serão penalizados em 10%.**
- **A entrega do trabalho deverá ser submetida no moodle até à data de entrega definida. Não serão aceites trabalhos fora do prazo.**
- A defesa e discussão dos trabalhos decorrerá em dia e hora a marcar por cada professor das teórico-práticas. No dia da apresentação, **TODOS** os elementos do grupo deverão estar presentes. Os elementos ausentes não terão classificação. A defesa e discussão serão realizadas em grupo com questões direcionadas a cada elemento individualmente.
- Cada grupo é responsável por gerir o seu processo de desenvolvimento. Dificuldades e problemas deverão ser comunicados atempadamente ao professor das aulas teórico-práticas.
- Código de conduta: (cf. Regulamento Disciplinar dos Estudantes do IPP)
    - Nenhum estudante ou grupo pode assumir pertença de trabalho realizado por outrem ou desenvolvido em conluio.
    - É expressamente proibido o uso de materiais, artefactos ou código de outrem sem a devida, e explícita indicação de origem.
    - Código de outras fontes deve ser claramente identificado no próprio código, indicando a fonte.
    - Casos de apropriação ilícita de materiais, artefactos e ou código sujeito a avaliação serão reportados à Presidência do ISEP.
    - A utilização de ferramentas com IA de assistência à codificação/desenho (e.g. chatGPT) deve ser mencionada.
- É obrigatório o uso da ferramenta de controle de versões Bitbucket.

3.1. Artigo Científico

No artigo científico deverão ser documentadas todas as fases da metodologia de trabalho seguida, preparação e exploração dos dados, análise e discussão dos resultados e conclusões (**máximo de 8 páginas** com o *template* do IEEE disponibilizado no moodle). Considere os seguintes aspetos:

- O artigo deverá ter uma secção inicial com o resumo de todas as técnicas estatísticas usadas na

realização da análise dos dados.

- O artigo deverá ter uma secção final com um resumo das principais conclusões retiradas da resolução das diferentes questões colocadas no enunciado.
- Cada problema resolvido deverá ter uma breve explicação da técnica estatística e uma conclusão baseada na interpretação e análise dos resultados obtidos. Caso considere relevante deverá incluir uma síntese destas conclusões na secção final de conclusões.

3.2. Avaliação

Na avaliação do trabalho serão considerados os seguintes aspetos e as ponderações especificadas na tabela 1:

a) Contextualização e objetivos (Sumário e Introdução)
b) Qualidade do código Phyton e respetiva documentação
c) A qualidade do processo de análise de dados seguido, a organização do código, a avaliação dos modelos criados e as conclusões alcançadas
d) Organização, qualidade da escrita, apresentação e clareza do relatório
e) A defesa e discussão
f) Participação individual de cada um dos elementos

Tabela 1 – Grelha de avaliação do Trabalho Prático 1

Sumário	15%
Questão 1	30%
Questão 2	20%
Questão 3	20%
Conclusão e referências	15%

Nota: A nota de cada um dos elementos do grupo será definida de acordo com a sua participação (em %). A equipa de avaliação de trabalhos práticos irá validar, no momento da defesa do trabalho (que poderá ser por videoconferência), a participação de cada um dos elementos do grupo na concretização dos objetivos do trabalho e do grupo. **Os elementos ausentes não terão classificação.**

4. Descrição do Trabalho

Na realização do Trabalho Prático 1 pretende-se que os alunos desenvolvam o processo de Análise Exploratória de Dados, Análise Inferencial, Correlação e Regressão [1-3].

A poluição ambiental é um dos maiores problemas deste século, estando na origem de enormes danos para o ecossistema biológico, condições meteorológicas, saúde humana e organismos vivos. De acordo com o relatório da Organização Mundial de Saúde, nove em cada dez pessoas respiram ar poluído, causando aproximadamente sete milhões de mortes em todo o mundo todos os anos (segundo a OMS). Na Europa estão identificados alguns dos principais poluentes atmosféricos: amoníaco (**NH_3**), dióxido de azoto (**NO_2**), partículas em suspensão com uma componente aerodinâmica de diâmetro inferior a 2.5µm (**PM2.5**) e ozono troposférico (**O_3**). Nas cidades, onde vive 74% da população da União Europeia, os níveis de **PM2.5** e O_3 têm efeitos nocivos na saúde humana associados a doenças respiratórias e doenças cardiovasculares. A poluição atmosférica também danifica ecossistemas, sendo **O_3** considerado o poluente atmosférico mais prejudicial á vegetação e biodiversidade.
Este trabalho tem como objetivo analisar os níveis de poluentes em diferentes zonas geográficas de países europeus no ano 2022. Para a realização desta análise deve utilizar os dados disponíveis no ficheiro ***AIRPOL_data.csv*** contém dados de níveis dos poluentes atmosféricos (**PM2.5**, **NO_2**, **O_3**) de várias regiões geográficas europeias, relativos ao ano de 2022**.** Informação sobre os rótulos das colunas encontram-se no

ficheiro ***AIRPOL_metadata.xlsx***. Use um grau de significância de 5% para todos os testes de hipótese efetuados.

4.1. Análise e exploração de dados

1. Construa um gráfico que permita visualizar os níveis médios do poluente **O_3** nas diversas regiões de Portugal (**NUTS Code**). Indique a região com nível médio de **O_3** mais elevado.

2. Usando *boxplots*, construa um gráfico que permita comparar as distribuições dos níveis médios do poluente **PM2.5** em Portugal, Espanha, Franca e Itália.

3. Usando uma representação gráfica adequada, compare as distribuições do número de mortes prematuras (**Value**) em Portugal, Espanha, França e Itália.

4. Construa uma tabela que indique, os valores da média, quartis, desvio padrão, assimetria e curtose, do número de mortes prematuras (**Value**) associado a **STROKE** para cada um dos países: Espanha, França, Itália e Grécia (apresente os resultados com 4 casas decimais)

4.2. Inferência Estatística

1. Considere na base de dados os valores relativos a Portugal e selecione aleatoriamente uma amostra aleatória de 50 registos dos níveis médios de poluição atmosférica (**AirpolPT**).

2. Use esta amostra para testar se o número médio dos níveis de poluição atmosférica em Portugal é inferior ao número homólogo na Albânia.

3. Selecione aleatoriamente duas amostras de 20 registos dos níveis médios de poluição atmosférica em Espanha e Franca. Use estas amostras para testar se os níveis de poluição atmosférica apresentam diferenças significativas entre os dois países.

4. De modo semelhante considere quatro amostras aleatórias de 20 registos dos níveis médios de poluição atmosférica nos seguintes países: Portugal, Albânia, Espanha e França. Use estas amostras para testar a existência de diferenças significativas nos níveis médios de poluição atmosférica dos 4 países. **Nota:** Caso necessário efetue uma análise **post-hoc** adequada.

4.3. Correlação e Regressão

1. Considere os dados relativos aos níveis médios do poluente **PM2.5** em Portugal, Espanha, França e Itália e construa uma tabela de correlação entre estes países.

2. Selecione os dados relativos á Alemanha e o poluente **PM2.5**. Considere as seguintes variáveis explicativas:

- X_1 - Nível medio de poluição
- X_2 - Área da região afetada

e a variável dependente

- Y - Número de mortes prematuras

a) Determine o modelo de regressão linear.

b) Verifique as condições sobre os resíduos.

c) Verifique se existe colinearidade (VIF).

d) Comente o modelo obtido tendo em conta todas as características relevantes para a qualidade do modelo.

e) Estime o número de mortes para a regiões com **NUTS Code**: DE131,DE132,DE133, DE134 ,DE135, DE136, DE137,DE138 e DE139 e compare com os valores reais

4.4. Análise e Discussão de Resultados

Efetue uma síntese dos resultados e das conclusões, obtidos neste trabalho, que considera mais importantes, justificando sempre que necessário (conclusão).

5. Referências Bibliográficas

[1]. C. HEUMANN and SHALABH M. SCHOMAKER, Introduction to statistics and data analysis, Springer International Publishing, 2016.

[2]. DOUGLAS C. MONTGOMERY, Design and Analysis of Experiments, 8th edition. John Wiley & Sons, New York, 2013

[3]. WES MCKINNEY, PYTHON FOR DATA ANALYSIS: DATA WRANGLING WITH PANDAS, NUMPY, AND JUPYTER, 3RD EDITION, HTTPS://WESMCKINNEY.COM/BOOK/,O'REILLY MEDIA, 2022.