Trabalho Prático

Análise de Dados em Informática

Engenharia Informática - 3º ano 2º semestre
Ano Letivo 2022/2023

---

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 R
- Escrita de Artigo Técnico com a Análise de Dados

2. Calendarização

Lançamento das propostas de trabalhos: até 8 de março de 2023
Entrega do trabalho: até **16 de abril de 2023** (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 R como ferramenta de suporte ao tratamento de dados.
- A **data final de ENTREGA** do 1º Trabalho Prático é **16 de abril de 2023,** 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 R para resolver o problema
- 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).

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 R
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].

4.1. Enunciado

1. Uma bomba elétrica para poços profundos (ESP - Electric Submersible Pump) é um dispositivo que transporta fluídos de grandes profundidades até à superfície. Tendo em vista otimizar o rendimento da produção de uma empresa de extração de petróleo foram monitorizados dados de funcionamento, em tempo real, de 3 ESP (ESP01, ESP02 e ESP03).

Tabela 1 - Descrição das variáveis

Labels			Variable
Pump 1	**Pump 2**	**Pump 3**	
ESP01	ESP02	ESP03	Discharge-Pressure (psia)
ESP01.1	ESP02.1	ESP03.1	Intake-Pressure (psia)
ESP01.2	ESP02.2	ESP03.2	Intake-Temperature (K)
ESP01.3	ESP02.3	ESP03.3	Motor-Temperature (K)
ESP01.4	ESP02.4	ESP03.4	VSDFREQOUT (Hz)
ESP01.5	ESP02.5	ESP03.5	V SDMOTAMPS (A)
ICO1	ICO2		Choke position (closed =0%)
ICO1.1	ICO2.1		Pressure 1 (psia)
ICO1.2	ICO2.2		Pressure 2 (psia)
ICO1.3	ICO2.3		Temperature 1 (deg F)
ICO1.4	ICO2.4		Temperature 2 (deg F)
ICO1.5	ICO2.5		Water cut (%) water produced/volume of total liquids
ICO1.6	ICO2.6		Liquid rate (bbl/d) total flow produced
ICO1.7	ICO2.7		Water rate (bbl/d) Liquid rate*Water cut
ICO1.8	ICO2.8		Oil rate (bbl/d) Liquid rate - Water cut

O ficheiro **DADOS1.csv** contém dados medidos, a cada 5 minutos no espaço temporal de 1 de junho de 2013 às 00:00 até 12 de junho de 2014 às 14:50. A tabela 2 descreve os parâmetros em questão:

a) Acrescente aos dados importados uma coluna com o tempo no sistema ***POSIXct*** no formato "yy/mm/dd HH:MM:SS GMT". Deve usar as opções origin = "1970-01-01" e tz = "GMT" para transformar o tempo (em segundos) para o sistema pedido.

b) Efetue um gráfico que permita comparar a temperatura do motor nas bombas 1,2 e 3, no dia 4 de agosto de 2013.

c) Efetue um boxplot com os dados da alínea anterior. Comente os resultados obtidos.

d) Uma forma de avaliar a quantidade de barris produzida num dia é calcular a média das medições do "oil rate" efetuadas no dia em questão:

   i. Efetue um gráfico de barras que compare os barris de petróleo produzidos diariamente pelas bombas 1 e 2 no mês de março de 2014.

   ii. Em que mês a bomba 1 extraiu mais barris de petróleo?
   N.B. Considere os meses compreendidos entre os dias 1-6-2013 e 31-5-2014.

   iii. Extraiu-se uma amostra aleatória de dias entre os dias 1-6-2013 e 31-5-2014 usando as seguintes instruções:

```
set.seed(300)
sample(1:365,10)
##  [1]  78 362  66 156 277  41 339 169 272 268
Sendo que os dias foram numerados por ordem crescente, ou seja, 1
representa o dia 1 de junho de 2013, 2 representa o dia 2 de junho
de 2013,… etc.
```

   Calcule a produção diária, nos dias da amostra aleatória, para as bombas 1 e 2 e efetue um *boxplot* dos dados obtidos para a bomba 1 e para a bomba 2.

   iv. Utilize as amostras aleatórias da alínea anterior para efetuar um teste de hipóteses que permita verificar se a média da produção diária de petróleo da bomba 1 foi superior à da bomba 2 no período de 1-6-2013 e 31-5-2014.

   v. Confirme se a decisão obtida no teste da alínea anterior corresponde à "realidade".

2. Pretende-se comparar a precisão de 6 algoritmos de Machine Learning: SVM, DT, KN, RF, ML e GB. Para o efeito calculou-se a precisão de cada algoritmo sobre 10 conjuntos de dados: D1, D2, …, D10. Os dados encontram-se guardados no ficheiro **DADOS2.csv**:

   a) Averigue se existe correlação entre a precisão de cada par de algoritmos (Apresente a matriz de correlações e comente os resultados).

   b) Efetue um teste de hipótese para averiguar se existem diferenças significativas entre a precisão dos diferentes algoritmos.

   c) Justifique todas as opções e decisões tomadas na alínea anterior. No caso de haver diferenças significativas faça um estudo *post-hoc* do teste que efetuou.

3. O ficheiro **DADOS3.csv** contém dados de 4 variáveis (aceleração, número de cilindros, peso e potência) de 99 viaturas escolhidas aleatoriamente:

   a) Divida as 99 viaturas em três grupos: viaturas com 4 cilindros, com 6 e com 8. Existirá diferenças significativas na aceleração entre os três grupos?

   b) Supondo que a aceleração é a variável dependente e as restantes variáveis são independentes:

i. Encontre o modelo de regressão linear. N.B. considere a variável número de cilindros uma variável "*Dummy*".

ii. Use o modelo encontrado na alínea anterior para estimar a aceleração de uma viatura com: um peso de 2950 kg, potência de 100 Hp e 4 cilindros.

4. 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]. JOAQUIM P. MARQUES DE SÁ, Applied Statistics Using SPSS, STATISTICA, MATLAB and R, 2nd Edition, John Wiley & Sons, 2007.