Artigo Científico ANADI

1st Gonçalo Teixeira
Departamento de Engenharia Informática
ISEP-Instituto Superior de Engenharia Informática
1200882

2st Sérgio Lopes
Departamento de Engenharia Informática
ISEP-Instituto Superior de Engenharia Informática
1200625

3st Tiago Freitas
Departamento de Engenharia Informática
ISEP-Instituto Superior de Engenharia Informática
1200628

I. Introdução

O presente documento foi desenvolvido no âmbito da unidade curricular de Análise de Dados em Informática (ANADI), lecionada durante o segundo semestre do terceiro ano da Licenciatura em Engenharia Informática (LEI) do Instituto Superior de Engenharia do Porto (ISEP).

Os dados analisados, ('DADOS1.csv','DADOS2.csv' e 'DADOS3.csv'), foram fornecidos pelos docentes da unidade curricular e tratados com recurso ao software RStudio com a linguagem R.

No primeiro exercício são estudados os valores de uma bomba elétrica, com o objetivo de melhorar o seu rendimento de produção. Esta bomba foi monitorizada, tendo em vista a sua otimização.

No segundo exercício é feita a comparação sobre a precisão de algoritmos de machine learning, incidindo em 10 conjuntos de dados diferentes. Este estudo tem como objetivo a melhor escolha de algoritmo para cada tipo de problemática.

No terceiro exercício temos como objetivo explorar a relação entre a aceleração de carros e outras variáveis, tais como o número de cilindros, o peso do veículo e a potência do motor. Para isso, será realizada uma análise exploratória dos dados, seguida da aplicação de testes de normalidade e da construção de um modelo de regressão linear múltipla. Por fim, será feita a previsão da aceleração para o veiculo especificado.

II. Preparação dos dados em estudo

De modo a fazer uma análise mais intuitiva dos dados, foram alterados os nomes de algumas colunas e, consoante o ficheiro em estudo, adicionadas novas colunas com o objetivo, de facilitar o tratamento dos dados.

III. Exercícios

Discussão de resultados e procedimentos realizados.

A. Exercício 1

1) a):

- Foi adicionada uma coluna 'Anos', importando do documento 'DADOS1.csv' a coluna 's' e foi usado o sistema POSIXct para transformar a coluna que encontrava-se previamente em segundos para o formato "yy/mm/dd HH:MM:SS GMT", com ínicio no dia 1/06/2013 e término no dia 12/06/2014.
- Foram também alteradas as colunas 'bbl.d.2' e 'bbl.d.5' para Bomba 1 e Bomba 2, respetivamente.

Fig. 1. Gráfico de temperaturas do motor das bombas.

1) b):

Neste gráfico foi possível concluir:

- A Bomba K.1 é constante, mas ao longo do tempo, demonstra algumas irregularidades e decandência na temperatura do motor.
- A Bomba K.3 é a mais irregular de todas, apresenta muitos valores dispersos de temperatura entre os 367ºC e os 377ºC e é a que tem a média de temperatura mais baixa.
- A Bomba K.5 é a que tem a média de temperatura mais alta e com valores pouco dispersos, mas, em contraste com a Bomba K.1, não apresentra decadência de valores e é constante.

1) c):

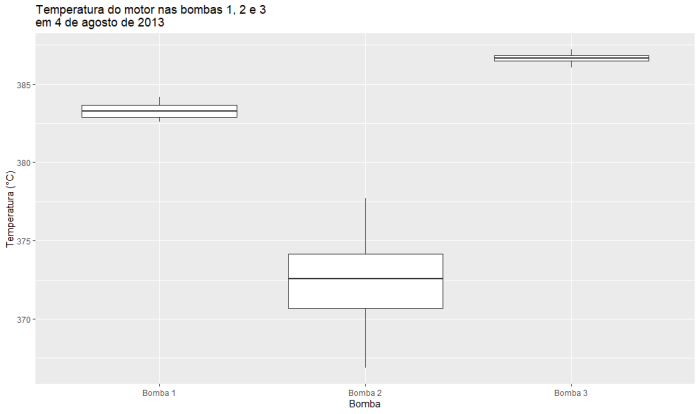

Fig. 2. BoxPlot sobre temperaturas de motor.

Neste BoxPlot foi possível concluir:

- A Bomba K.1/1 tem alguma dispersão de dados, tendo uma mediana de 383.27, valor máximo de 384.16 e mínimo 382.59. Devido há pouca dispersão de dados, os valores de primeiro e 3 quartil assemelham-se muito à mediana, o 2º quartil.
- A Bomba K.3/2 tem uma grande discrepância entre quartis, tendo uma mediana de 372.57, valor máximo de 377.69 e mínimo 366.90. Como existe uma grande dispersão o valor de diferença dos 1º e 3º quartis, em relação, à mediana, é quase de 2ºC.
- A Bomba K.5/3 tem uma dispersão de dados quase nula, tendo uma mediana de 386.67, valor máximo de 387.22 e mínimo 386.04. Devido à quase nula dispersão existe uma grande proximidade entre quartis. Por exemplo: a diferença entre o 1ºquartil e o 3ºquartil é de 0.23 ºC

	variable	value.min	value.25%	value.50%	value.75%	value.max
1	K.1	382.5949	382.8813	383.2723	383.6533	384.1590
2	K.3	366.8697	370.6827	372.5668	374.1734	377.6947
3	K.5	386.0382	386.5069	386.6658	386.8359	387.2185

Fig. 3. Informação sobre os quartis de temperaturas de motor.

1) d):

1) : De modo, a fazer a apresentação do seguinte gráfico de barras, restringiu-se os dados para a data desejada, posteriormente, converteu-se os valores para inteiros e por último fez-se a média diária de cada bomba.

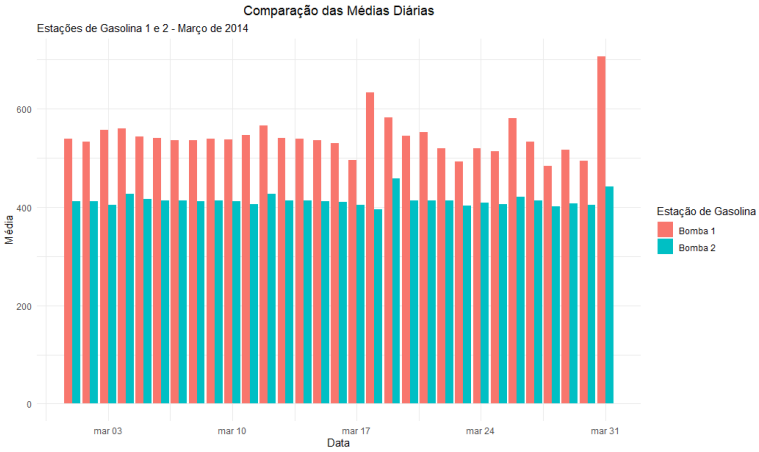

Fig. 4. Gráfico de barras de taxa de óleo mensal.

Neste gráfico de barras foi possível concluir:

- No tempo delimitado, a bomba 1 foi sempre superior a bomba 2, em termos, de média de produção de barris de óleo diária.
- Existe também uma maior descrepância de dados na bomba 1, relativamente à consistência da bomba 2, em termos, de média de produção de barris de óleo diária.

2) :

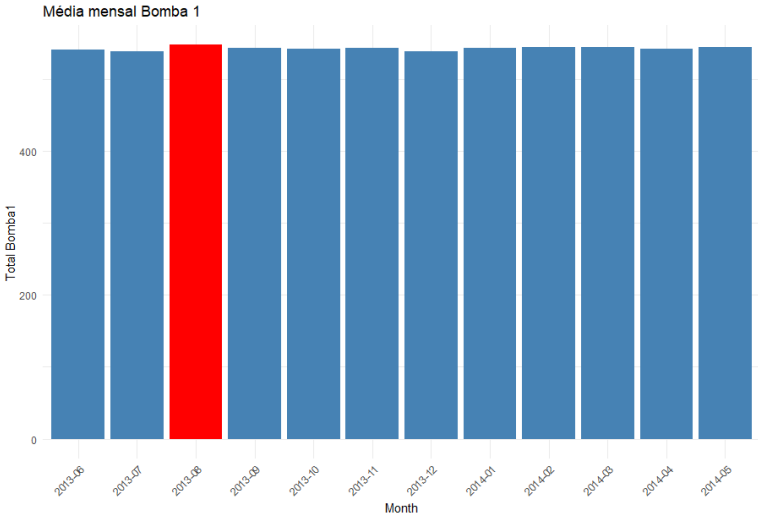

Fig. 5. Gráfico de barras do mês com mais produção.

- Foi elaborado um gráfico de barras, tendo em conta o tempo limitado, com a média de produção mensal de barris da bomba 1. Deste gráfico é possível concluir que o mês com mais produção é o de agosto.

3) :

No tempo delimitado, foi atribuído um index a cada dia para cada valor da bomba 1 e 2. Como index da tabela começa com 0, reduziu-se 1 ao número de amostra. Assim podemos concluir:

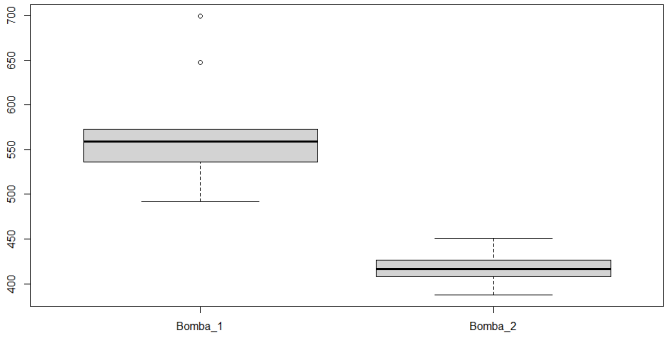

Fig. 6. BoxPlot da produção diária

- A bomba 1 contêm 2 outliers (valores extremos) e o primeiro quartil e terceiro quartil têm valores de produção diária superiores aos da bomba 2. Com estes valores, é possível depreender que apesar de ter valores mais extremos, tem uma média e mediana superior à da bomba 2.

4) :

Foi realizado o teste de levene para verificar a homogenidade das variâncias, posteriormente realizei o teste de Shapiro para verificar a normalidade das distribuições e por último fiz o teste t Student para comparar as médias das amostras.

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value  Pr(>F)
group  1   5.032 0.03939 *
      16
---
signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 7. Teste de Levene

```
        Shapiro-Wilk normality test

data:  dados_ex4$Bomba_1
W = 0.89971, p-value = 0.2503
```

Fig. 8. Teste de Shapiro 1

```
        Shapiro-Wilk normality test

data:  dados_ex4$Bomba_2
W = 0.98709, p-value = 0.9907
```

Fig. 9. Teste de Shapiro 2

- Como os valores do teste de Shapiro são ambos superiores ao nível de significância de 0.05 de variâcia, não temos evidências para rejeitar a hipótese de normalidade para ambas.
- Com o resultado do teste de levene, como este é inferior a 0.05, então, existem evidências que nos permitem rejeitar a homogenidade, sendo assim as amostras heterogéneas

Teste t de Student

- H0: A média de produção diária da Bomba 1 não é superior à média da Bomba 2, no período delimitado.
- H1: A média de produção diária da Bomba 1 é superior à média da Bomba 2, no período delimitado.

```
        Welch Two Sample t-test

data:  dados_ex4$Bomba_1 and dados_ex4$Bomba_2
t = 6.7472, df = 9.3031, p-value = 3.584e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 111.2259      Inf
sample estimates:
mean of x mean of y
 570.1566  417.6500
```

Fig. 10. BoxPlot da produção diária

Podemos assim concluir:

- como o p-value resultante do teste t de Student é inferior a 0.05, temos evidências suficientes para rejeitar H0. Assim a média de produção diária da Bomba 1 é superior à média da Bomba 2, no período delimitado.

5) : Foi efetuada a média de todos os valores presentes no documento em estudo, sendo a média da bomba 1 e 2, respetivamente, 570.16 e 417.65. Assim podemos comprovar que os dados adquiridos na alínea anterior estão de acordo com a realidade, como um todo.

B. Exercício 2

2) a): Foi feito teste shapiro para todos os algoritos para saber se os dados eram normalmente distribuídos ou não, tendo somente o teste ao algoritmo ML, rejeitado h0, a hipótese de distruibuição normal e por isso foi usado o Pearson.

```
> shapiro.test(data$SVM) # o p-value=0.2687 como é > alfa, logo não se rejeita H0

        Shapiro-Wilk normality test

data:  data$SVM
W = 0.90818, p-value = 0.2687

> shapiro.test(data$DT) # o p-value=0.0677 como é > alfa, logo não se rejeita H0

        Shapiro-Wilk normality test

data:  data$DT
W = 0.85561, p-value = 0.06772

> shapiro.test(data$KN) # o p-value=0.6926 como é > alfa, logo não se rejeita H0

        Shapiro-Wilk normality test

data:  data$KN
W = 0.95203, p-value = 0.6926
```

Fig. 11. Teste de Shapiro 1

```
> shapiro.test(data$RF) # o p-value=0.3138 como é > alfa, logo não se rejeita H0

        Shapiro-Wilk normality test

data:  data$RF
W = 0.91457, p-value = 0.3138

> shapiro.test(data$ML) # o p-value=0.0214 como é < alfa, logo rejeita-se H0 e pode-se concluir que não segue

        Shapiro-Wilk normality test

data:  data$ML
W = 0.8139, p-value = 0.02138

>                       # distribuição normal
> shapiro.test(data$GB)

        Shapiro-Wilk normality test

data:  data$GB
W = 0.93629, p-value = 0.5125
```

Fig. 12. Teste de Shapiro 1

Posteriormente, foi efetuada uma matriz de correlação com os dados em questão, e um gráfico baseado na matriz.

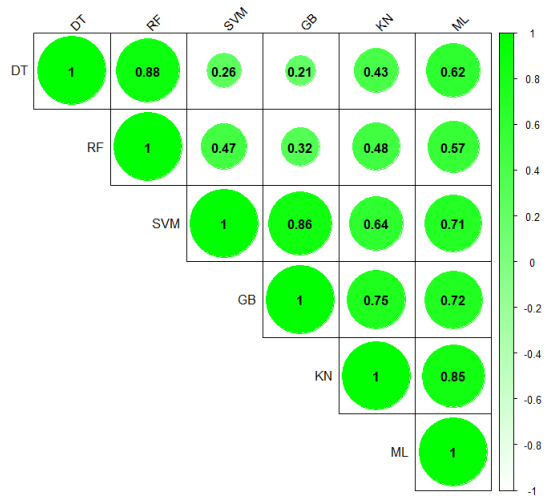

Fig. 13. Gráfico baseado na matriz de correlação.

Este gráfico é analisado através dos coeficientes de correlação de Pearson, visto serem valores compreendidos entre -1 e 1. Além disso, o coeficiente de correlação de Pearson é uma medida estatística que indica o grau de correlação linear entre duas variáveis quantitativas. Podendo ser qualificado como:

- O coeficiente de correlação de Pearson será próximo de -1, se dois algoritmos produzem resultados opostos
- O coeficiente de correlação de Pearson será próximo de 0, se dois algoritmos produzem resultados independentes
- O coeficiente de correlação de Pearson será próximo de 1, se dois algoritmos produzem resultados semelhantes

Com base na matriz de correlação foram observados fortes relacionamentos entre ML/KN, GB/SVM e RF/DT, respectivamente, enquanto os menos fortes foram GB/DT e DT/SVM, respectivamente.

2 b) e c) :

De modo a realizar este exercício, decorreu o seguinte procedimento:

- Organização dos dados em duas colunas distintas
- Boxplot para se poder poder visualizar a distribuição dos dados dos algoritmos
- Teste de normalidade de Shapiro-Wilk (realizado na alínea a))
- Realizado teste de levene para verificar a homogenidade

```
> # H0: as variâncias são iguais
> # H1: existe pelo menos um par de grupos cujas variâncias são diferentes
>
> leveneTest(Precisão ~ Algoritmo, data_melted,center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value    Pr(>F)
group  5  6.2254 0.0001259 ***
      54
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig. 14. Teste de Levene.

- Como p=0.0001259, não é maior que 0.05, então rejeita-se H0, logo as variâncias dos grupos não são todas iguais
- Podemos concluir que os testes paramétricos não são aplicáveis, e aplicamos o teste Friedman

```
> # H0: Não existem diferenças significativas entre a precisão dos algoritmos
> # H1: Existe pelo menos um algoritmo cuja precisão é diferente dos restantes
>
> # Teste Friedman
> friedman_test <- friedman.test(Precisão ~ Algoritmo | Dataset, data = data_organized); friedman_test

        Friedman rank sum test

data:  Precisão and Algoritmo and Dataset
Friedman chi-squared = 8.7097, df = 5, p-value = 0.1212
```

Fig. 15. Teste de Friedman.

Assim concluímos que, como p-value=0.1212 é maior que o alfa=0.05, não se rejeita H0, logo não se pode afirmar que existem diferenças significativas entre as precisões dos diferentes tipos de algoritmos.

C. Exercício 3

3) a):

- A resolução apresentada começa com a leitura do arquivo CSV "DADOS3.csv" através da função read.csv(). Em seguida, os dados são divididos em três grupos de viaturas as de 4 cilindros, 6 cilindros e 8 cilindros, recorrendo à função subset().
- Inicialmente verificamos se as amostras seguem distribuições normais e para tal, foi utilizado o teste de Lilliefors para cada grupo de cilindros, uma vez que se trata do teste mais aconselhado para amostras com n maior ou igual 30. Os resultados dos testes foram impressos usando a função print(), conseguindo concluir que não existem evidências suficientes para rejeitar a hipótese da normalidade da distribuição de acelerações para os carros com 4, 6 e 8 cilindros. O valor de p para os testes foi maior que 0,05 em todos os casos, indicando que não há diferenças significativas entre as distribuições observadas e a distribuição normal.
- Para verificar se há diferenças significativas na aceleração entre os grupos de cilindros, foi utilizado o teste de Kruskal-Wallis, uma alternativa não paramétrica ao teste da ANOVA. O resultado do teste indica que há diferenças significativas na aceleração entre os grupos de cilindros com um p-value de 2.795e-11 podemos concluir que há diferenças significativas na aceleração entre os grupos de 4, 6 e 8 cilindros.

3) b) i):

- O modelo de regressão linear múltipla foi criado com o objetivo de analisar a relação entre a aceleração das viaturas tendo em conta o número de cilindros, peso e potência do motor. A variável dummy foi utilizada para representar o número de cilindros do carro, com três níveis: 4, 6 e 8 cilindros.
- Os resultados indicam que o número de cilindros tem uma relação significativa com a aceleração do carro. No estudo carros com 4 cilindros e 6 cilindros têm uma aceleração média de 1.52 sendo menor por outro lado os carros com 8 cilindros têm uma aceleração média de 4.87, assim sendo as viaturas com mais cilindros tendem a ter uma aceleração mais lenta.

  se o peso do carro e a potência do motor forem considerados, é possível afirmar a sua relação significativa com a aceleração. A cada aumento de uma unidade no peso do carro, a aceleração diminui em média 0.003 unidades. Da mesma forma que aumentando os HP, a aceleração diminui em média 0.057 unidades.
- O modelo tem um R quadrado ajustado de 0.5931, significando que cerca de 59.31 por cento da variação da aceleração pode ser explicada pelas variáveis independentes. Já o F estatístico é significativo com um valor de p menor que 2.2e-16, indicando que pelo menos uma das variáveis independentes tem um efeito significativo na aceleração.

3) b) ii):

- Com um peso de 2950kg, potência de 100 HP e 4 cilindros a variável "nova obs" de acordo com o modelo de regressão linear múltipla criado, o carro previsto terá uma aceleração de 17.30784 unidades de medida.

IV. Conclusão

A. Exercício 1

Em conclusão, através do estudo feito, podemos concluir que a bomba 1, tem um maior rendimento, em relação à bomba 2, no que toca a produção de barris de petróleo. No entanto a bomba 1 funciona numa temperatura mais elevada, o que pode provocar desgaste, entre outros problemas associados, em contraste, com a bomba 2, que opera em temperaturas mais baixas.

B. Exercício 2

Em conclusão, foi realizado um estudo, com o intuito de verificar a precisão de 6 algoritmos em 10 conjuntos de dados diferentes, em que foi gerada uma matriz de correlação que apontava como mais forte os relacionamentos ML/KN, GB/SVM e RF/DT e os mais fracos GB/DT e DT/SVM. Estas informações podem ser relevantes, no momento de escolha de algoritmos em futuros projetos de Machine learning

C. Exercício 3

Em conclusão, após o estudo efetuado, com o intuito de verificar as diferenças significativas na aceleração nos três grupos identificados, em que foram feitos testes Lillie e posteriormente o teste de Kruskal-Wallis onde foi possível identificar uma diferença significativa entre os grupos através do p-value 2.795e-11, de seguida foi encontrada a regressão linear múltipla para investigar a relação entre a aceleração do carro e suas características tais como: peso, potência e cilindragem e por mim foi determinada uma previsão de aceleração de 17.30784 unidades de medida do veiculo com as características mencionadas no enunciado.

V. Conhecimento

Este documento foi desenvolvido tendo por base o conhecimento adquirido nas aulas da unidade curricular em questão e também por consulta na internet.