Análise Exploratória de Dados, Análise Inferencial e Correlação e Regressão

Marta Portugal
Licenciatura em Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1201845@isep.ipp.pt

Rui Dias
Licenciatura em Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1200963@isep.ipp.pt

André Ferreira
Licenciatura em Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1200605@isep.ipp.pt

O presente documento descreve o processo de análise exploratória de dados, análise inferencial, correlação e regressão aplicados a dados de cujos dados nos foram fornecidos.
Tomando como exemplo, a primeira situação, objetivo com estas análises é otimizar o rendimento da produção de uma empresa de extração de petróleo. A análise exploratória de dados é realizada para identificar padrões, tendências e anomalias nos dados coletados. De seguida, a análise inferencial para obter insights estatísticos e identificar possíveis relações entre as variáveis. Finalmente, uma análise de correlação para avaliar a força e direção das relações entre as variáveis, e, será aplicada a regressão para modelar e prever o rendimento das bombas ESP.

Palavras-chave: Análise Exploratória de Dados, Análise Inferencial, Correlação, Regressão, Bomba Elétrica, Poço Profundo, Otimização, Rendimento.

I. INTRODUÇÃO

A análise de dados é uma ferramenta fundamental na otimização de processos de produção, incluindo a indústria de extração de petróleo.

Este artigo tem como foco a análise exploratória de dados, análise inferencial, correlação e regressão aplicadas aos dados de funcionamento de três bombas elétricas para poços profundos (*ESP - Electric Submersible Pump*), denominadas *ESP01, ESP02 e ESP03*. O objetivo é desenvolver um processo abrangente de análise desses dados em tempo real, visando a otimização do rendimento da produção de uma empresa de extração de petróleo.

As técnicas estatísticas utilizadas neste estudo têm como objetivo obter insights valiosos a partir dos dados fornecidos, com potencial para influenciar tomadas de decisão estratégicas na indústria petrolífera. A análise exploratória, a análise inferencial, a correlação e a regressão serão aplicadas de forma integrada e contínua, proporcionando uma melhor e mais aprofundada compreensão, no que toca o desempenho e possíveis melhorias das bombas elétricas em estudo.

II. REVISÃO DE LITERATURA

A. Análise Exploratória de Dados

A Análise Exploratória de Dados (AED) é uma abordagem de investigação estatística que visa entender melhor um conjunto de dados e identificar padrões, tendências, anomalias e relacionamentos entre as variáveis presentes. Através da Análise exploratória de dados os analistas podem usar várias técnicas estatísticas para explorar os dados e identificar padrões interessantes. Isso inclui visualizações gráficas, estatísticas descritivas e testes de hipóteses simples. O objetivo é descobrir tendências, padrões e anomalias que possam estar presentes nos dados.

B. Análise Inferencial

A Análise Inferencial é uma abordagem estatística que visa fazer generalizações sobre uma população a partir de um conjunto de dados amostral. O objetivo da análise inferencial é obter conclusões que possam ser aplicadas a uma população maior com um certo grau de certeza.

Durante a análise inferencial, os analistas usam técnicas estatísticas para estimar parâmetros desconhecidos da população. Eles fazem isso usando informações contidas nos dados amostrais, como a média da amostra ou a variância. A partir dessas estimativas, é possível fazer inferências sobre a população, como a média ou a proporção de uma característica em toda a população.

Para fazer inferências precisas sobre a população, os analistas usam técnicas estatísticas como testes de hipóteses e intervalos de confiança. Essas técnicas ajudam a avaliar a probabilidade de que as conclusões tiradas a partir dos dados amostrais sejam válidas para toda a população.

A análise inferencial só funciona quando a amostra é selecionada de maneira aleatória e representativa da população de interesse. A análise inferencial é usada em muitas áreas, como pesquisa de mercado, ciência, engenharia e medicina.

C. Regressão Linear

A Regressão Linear é um dos algoritmos estatísticos mais utilizados na análise de dados e *machine learning*. É aplicado para fazer previsões com base em outras variáveis do modelo. Ele usa o coeficiente de correlação para medir a relação entre as variáveis, expressando-a em valores no intervalo [-1;1]. Um valor próximo de 0 indica pouca relação, enquanto valores próximos de 1 ou -1 indicam uma

forte relação. O coeficiente de correlação é calculado pelo teste de Pearson se as variáveis forem contínuas, ou pelo teste de Spearman se não forem. O coeficiente de determinação é calculado como o quadrado do coeficiente de correlação, representando a porcentagem da variável dependente explicada pelo modelo de regressão linear.As regressões lineares podem ser simples ou múltiplas. No caso de serem simples, a previsão da variável dependente é feita a partir da análise dos valores de somente uma única variável independente. Da comparação destas duas variáveis, resulta uma linha reta, representando a relação direta entre elas. No entanto, existem casos em que a análise de uma única variável independente não é suficiente para prever o comportamento de uma variável dependente. Assim, é necessário recorrer a uma regressão linear múltipla, que conta com a utilização de mais do que um variável independente.

D. Correlação Linear

A correlação linear é uma medida estatística que descreve a relação linear entre duas variáveis quantitativas. É calculada através do coeficiente de correlação, que varia de -1 a 1 [1]. Um valor próximo de 1 indica uma correlação positiva forte, ou seja, à medida que uma variável aumenta, a outra também aumenta. Um valor próximo de -1 indica uma correlação negativa forte, ou seja, à medida que uma variável aumenta, a outra diminui. Um valor próximo de 0 indica que não há uma relação linear aparente entre as variáveis.

A correlação linear é amplamente utilizada em análises de dados para entender a relação entre duas variáveis e pode ser calculada usando diferentes métodos, como o coeficiente de correlação de Pearson, que é aplicado a variáveis quantitativas com distribuição normal [2], ou o coeficiente de correlação de Spearman, que é aplicado a variáveis quantitativas com distribuição não normal ou a variáveis ordinais [1].

E. Testes de Hipótese

O Teste de Hipótese é uma metodologia de estatística que nos auxilia a tomar decisões sobre uma ou mais populações baseadas na informação obtida da amostra. Permite verificar se os dados amostrais trazem evidência que apoiem ou não uma hipótese estatística formulada.

Em muitas situações práticas o interesse do é verificar a veracidade sobre um ou mais parâmetros populacionais (μ,σ2,p) ou sobre a distribuição de uma variável aleatória.

Um dos primeiros trabalhos sobre testes foi publicado em 1710 (John Arbuthnot);

[https://www.inf.ufsc.br/~andre.zibetti/probabilidade/teste-de-hipoteses.html].

Num teste de hipóteses há que definir duas hipóteses designadas por hipótese nula H0 e hipótese alternativa H1. A hipótese alternativa está associada à "conjectura" que pretendemos verificar se é válida, no contexto do problema em análise. A hipótese nula é a hipótese complementar de H1.

A estratégia básica seguida no método do teste de hipóteses consiste em tentar suportar a validade de H1, uma vez provada a inverosimilhança de H0, isto é, conseguindo-se mostrar que com elevada probabilidade a hipótese nula é falsa, fica corroborada a validade da hipótese alternativa. Se não for possível rejeitar H0, a hipótese H1 não será reforçada pelo teste.

[file:///C:/Users/andre/Downloads/NotasApoioPE_233_254%20(2).pdf]

Diferença entre Testes Paramétricos e Não Paramétricos:

Os testes de hipótese podem ser classificados como paramétricos ou não paramétricos, dependendo das suposições feitas sobre a distribuição subjacente dos dados. Os testes paramétricos assumem que os dados possuem uma distribuição específica, geralmente a distribuição normal. Este tipo de testes usam os parâmetros dessa distribuição, como a média e o desvio padrão, para fazer inferências estatísticas sobre a população a partir da qual os dados foram amostrados. Exemplos de testes paramétricos incluem o teste t de Student, a análise de variância (ANOVA) e o teste de regressão linear. Normalmente os Testes Paramétricos são mais robustos.

Por outro lado temos os Testes não Paramétricos, estes não assumem nenhuma distribuição de probabilidade específica para os dados e não dependem de parâmetros fixos. Esses testes são geralmente usados quando os dados são categóricos ou não seguem uma distribuição normal.

Exemplos de Testes não Paramétricos, são o Teste de Friedman, Kruskal-Wallis, entre outros...

III. PREPARAÇÃO E SELEÇÃO DOS DADOS

Para conseguir prever ou classificar os dados desejados, é preciso começar com a preparação e seleção dos atributos do ***dataset***.

O modelo de dados deve estar armazenado em um arquivo com a extensão "***.csv***", para ser importado pelo R Studio e processado pelos diferentes algoritmos.

Após a importação do arquivo, foi feita uma cópia da tabela em uma nova instância, e em seguida as duas primeiras linhas foram removidas e demos outos nomes às colunas de forma a que ficasse mais legível utilizando o comando subset.

Em seguida, os dados foram traduzidos para valores numéricos, para serem normalizados.

As colunas que continham apenas números foram facilmente traduzidas com a função "as.numeric" do R.

Para as colunas com mais de duas opções diferentes, como no exercício 3, transformamos as variáveis em fator usando a funcao "as.factor".

Uma variável dummy, também conhecida como variável indicadora ou variável binária, é uma variável categórica em que cada valor possível é representado por um valor binário, geralmente 0 ou 1. Uma variável dummy é criada através da função "factor" ou "as.factor", que converte uma variável categórica em uma série de variáveis binárias. Cada valor possível da variável categórica é representado por uma variável binária distinta, que assume o valor 1 se o valor da variável categórica correspondente for igual ao valor representado pela variável binária, e 0 caso contrário. A utilização de variáveis dummy é comum em análises estatísticas, especialmente em regressão, quando se deseja incluir variáveis categóricas em um modelo de análise. Para gerar as regressões lineares, usou-se a função "lm".

IV. Solução e Análise Dos Dados

Para realizar a análise da situação na empresa de extração de petróleo, anteriormente mencionada, foram propostas três circunstâncias a analisar e superar.
Testaram-se as metodologias nas propostas e para tal, foi fornecido um conjunto de dados provenientes das três bombas ESP e três ficheiros *.csv* com os dados medidos.

A. ***DADOS1.csv*** *- 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*

Todo este processo foi inicializado com a importação do ficheiro pela opção incluída no RStudio na aba

Environment, clicar em *File — Import Dataset — From Text (readr)*

Procedemos ao seu processamento, no caso, o problema requeria a **adição aos dados importados duma coluna com o tempo em segundos no sistema *POSIXct* no formato *"yy/mm/dd HH:MM:SS GMT", com origin = "1970-01-01" e tz = "GMT".***

Para uma melhor organização e proteção de dados, copiamos a tabela para uma nova instância, eliminamos as duas primeiras linhas e renomeamos as colunas,

```
dados <- DADOS1
dados <- subset(dados, !(row.names(dados) %in% c("1", "2")))

colnames(dados) <- c("DischargePressureESP01",⋯)
```

Posto isto passamos à criação da nova coluna formatada chamada *TimeFormatted* com o comando *as.POSIXct* e formatamos os dados para segundos e finalmente exportamos a tabela criada para um arquivo CSV para poder ser guardada, tal como demonstra figura seguinte:

```
dados$TimeFormatted <- as.POSIXct(dados$Time,
origin = "1970-01-01", tz = "GMT", format = "%s")

write.csv(dados, file = "dadosTime.csv", row.names = FALSE)
```

No âmbito **comparar a temperatura do motor nas bombas 1,2 e 3, no dia 4 de agosto de 2013**, começamos por filtrar os dados desse dia, passá-los para numéricos e criar o gráfico,

```
dados_subset <- subset(dados, format(TimeFormatted, "%Y-%m-%d") == "2013-08-04")
dados_subset$MotorTemperatureESP01.3<-
as.numeric(dados_subset$MotorTemperatureESP01.3)
dados_subset$MotorTemperatureESP02.3<-
as.numeric(dados_subset$MotorTemperatureESP02.3)
dados_subset$MotorTemperatureESP03.3<-
as.numeric(dados_subset$MotorTemperatureESP03.3)
                library(ggplot2)
grafico <- ggplot(dados_subset, aes(x = TimeFormatted)) +geom_line(aes(y = MotorTemperatureESP01.3, color = "Bomba 1", group=1), linetype = "solid") + geom_line(aes(y = MotorTemperatureESP02.3, color = "Bomba 2", group=2), linetype = "solid") + geom_line(aes(y = MotorTemperatureESP03.3, color = "Bomba 3", group=3), linetype = "solid") +labs(x = "Tempo", y = "Temperatura da bomba", title = "Comparação da Temperatura do motor das bombas 1, 2 e 3 no dia 4 de Agosto de 2013") + scale_color_manual(values = c("Bomba 1" = "red", "Bomba 2" = "green", "Bomba 3" = "blue")) +theme_minimal()
```

Observando o gráfico seguinte, é possível verificar diferenças significativas no que toca as temperaturas do motor nas bombas. A bomba 1 (vermelha) verifica uma temperatura média entre os 382.5- 385, consideravelmente constante ao longo do tempo. O mesmo se verifica na bomba 3(azul), mas no intervalo de temperatura de 385-387.5. Por outro lado, a bomba 2 (verde) verifica múltiplas oscilações e inconstâncias, no que consta um intervalo de temperaturas muito mais baixas (367,5-377,5).

Figura 1 -gráfico de comparação das temperaturas

Consequentemente, o seu ***boxplot*** será:

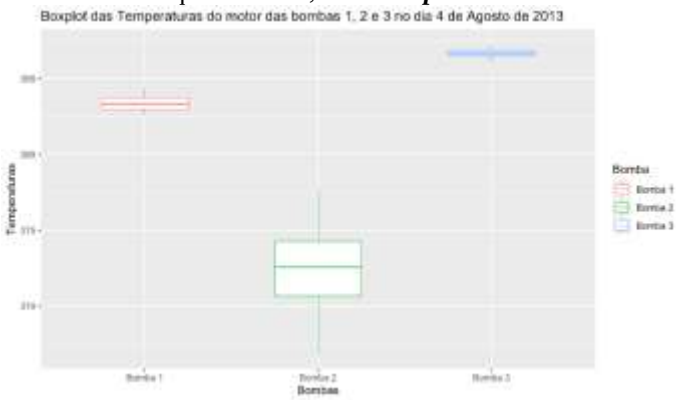

Figura 2 -boxplot de comparação das temperaturas

A **quantidade de barris produzida** num dia pode ser calculada pela média das medições do "oil rate" efetuadas no dia em questão:

Nesse sentido foi elaborado um **gráfico de barras que compare os barris de petróleo produzidos diariamente pelas bombas 1 e 2 no mês de março de 2014**.

Primeiramente, convertemos a coluna "TimeFormatted" para formato de data, filtramos os dados para o mês de março e convertemos para numérico.

```
dados$TimeFormatted <- as.Date(dados$TimeFormatted)
dados_subset <- subset(dados, format(TimeFormatted, "%m") == "03")
dados_subset$OilRateICO1.8<- as.numeric(dados_subset$OilRateICO1.8)
dados_subset$OilRateICO2.8<- as.numeric(dados_subset$OilRateICO2.8)
```

De seguida, calculamos a média da taxa de petróleo para cada dia das bombas para podermos criar e formatar o gráfico:

```
df_avg <- dados_subset %>%
 group_by(TimeFormatted) %>%
 summarise(avg_oil_rate1 = mean(OilRateICO1.8),
        avg_oil_rate2 = mean(OilRateICO2.8))

barras <- ggplot(df_avg, aes(x = TimeFormatted)) +
 geom_col(aes(y = avg_oil_rate1), fill = "blue", width = 0.8) +
 geom_col(aes(y = avg_oil_rate2), fill = "yellow", width = 0.4) +
 labs(title = "Produção di á ria de petr ó leo - Março 2014", x = "Data", y = "Produção di á ria de petr ó leo m é dia") +
 theme_minimal() +
 scale_x_date(date_labels = "%d/%m/%Y", date_breaks = "5 days") # Formatar eixo x como datas completas com intervalos de 5 dias
```

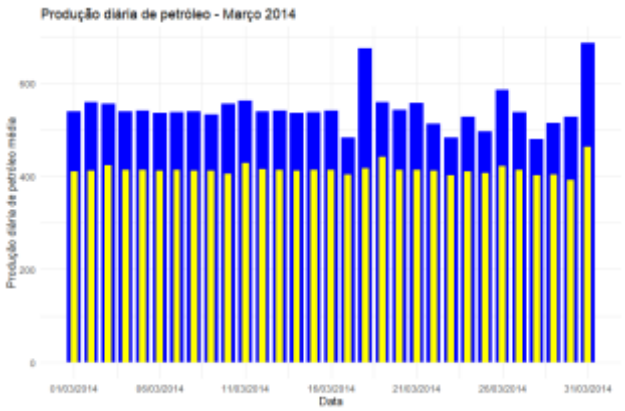

Figura 3 -Produção diária de petróleo; bomba 1 vs 2

É possível concluir que a produção diária de petróleo da bomba 1 em março é significativamente maior do que a bomba 2.

Para saber em que **mês a bomba 1 extraiu mais petróleo** procedeu-se à elaboração de um gráfico de barras com a média da produção de barris por mês:

Os dados compreendidos entre **Junho de 2013 e Maio de 2014** foram filtrados e as médias de produção de barris de óleo em função do mês foram agregadas. Para finalizar, editamos, calculámos a coluna com maior média e criamos o ***barplot*** necessário:

```
barris <- dados[dados$TimeFormatted >= as.Date("2013-06-01") & dados$TimeFormatted <= as.Date("2014-05-31"), ]
dados_mensais_bomba1 <- aggregate(as.numeric(barris$OilRateICO1.8) ~ format(as.Date(barris$TimeFormatted), "%Y-%m"), data = barris, FUN = mean)

colnames(dados_mensais_bomba1) <- c("M ê s", "M é dia de barris produzidos")
cores<-rep("#234F1E", length(dados_mensais_bomba1[,1]))

mes_max <- which(dados_mensais_bomba1$M ê s == dados_mensais_bomba1$M ê s[which.max(dados_mensais_bomba1$`M é dia de barris produzidos`)])
cores[mes_max] <- "#B2D3C2"

barplot(dados_mensais_bomba1$`M é dia de barris produzidos`, col = cores, names.arg = dados_mensais_bomba1$M ê s, xlab = "Meses", ylab = "M é dia da produção de barris de  ó leo", main = "Produção de barris de  ó leo entre Junho de 2013 e Maio de 2014 da Bomba 1")
```

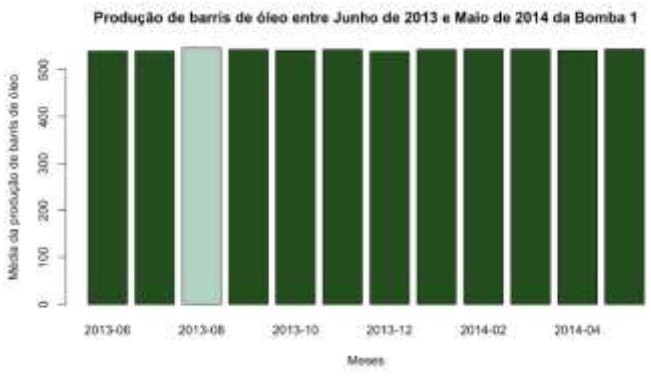

Figura 4 – mês em que a bomba 1 extraiu mais petróleo

A **Produção diária das bombas** 1 e 2 pode ser avaliada recorrendo a um ***boxplot.*** ***E***xtraindo-se uma **amostra aleatória de dias entre os dias 1-6-2013 e 31-5-2014:**

```
set.seed(300)
sample(1:365,10)
```

Começamos com o cálculo das médias da produção diária de barris de óleo para as bombas 1 e 2, para de seguida usar o código fornecido e gerar os 10 dias aleatórios.

```
dados_diarios_bomba1 <-
aggregate(as.numeric(barris$OilRateICO1.8) ~
format(as.Date(barris$TimeFormatted), "%Y-%m-%d"), data =
barris, FUN = mean)
dados_diarios_bomba2 <-
aggregate(as.numeric(barris$OilRateICO2.8) ~
format(as.Date(barris$TimeFormatted), "%Y-%m-%d"), data =
barris, FUN = mean)


set.seed(300)
dias_aleatorios <- sample(1:365, 10)
```

Passámos para a conversão dos índices aleatórios nas datas correspondentes e no formato correto , para de seguida filtrar as médias calculadas pelos dias aleatórios. Por fim juntámos as tabelas e criámos o boxplot para a bomba 1 e 2:

```
datas_iniciais <- as.Date("2013-06-01")
datas_aleatorias <- datas_iniciais + as.difftime(dias_aleatorios -
1, units = "days")
datas_aleatorias_formatadas <- format(datas_aleatorias, "%Y-
%m-%d")


medias_aleatorias_bomba1 <-
dados_diarios_bomba1[dados_diarios_bomba1$`format(as.Dat
e(barris$TimeFormatted), "%Y-%m-%d")` %in%
datas_aleatorias_formatadas, ]
medias_aleatorias_bomba2 <-
dados_diarios_bomba2[dados_diarios_bomba2$`format(as.Dat
e(barris$TimeFormatted), "%Y-%m-%d")` %in%
datas_aleatorias_formatadas, ]


medias_aleatorias <- cbind(medias_aleatorias_bomba1,
medias_aleatorias_bomba2)
colnames(medias_aleatorias) <- c("Dia", "M é dia Bomba
1","Dia", "M é dia Bomba 2" )


box <- boxplot(medias_aleatorias$`M é dia Bomba 1`,
medias_aleatorias$`M é dia Bomba 2`,
names = c("Bomba 1", "Bomba 2"),
main = "Produção Di á ria - Bomba 1 vs. Bomba 2",
ylab = "Produção Di á ria (barris)",
col = c("lightblue", "pink"))
```

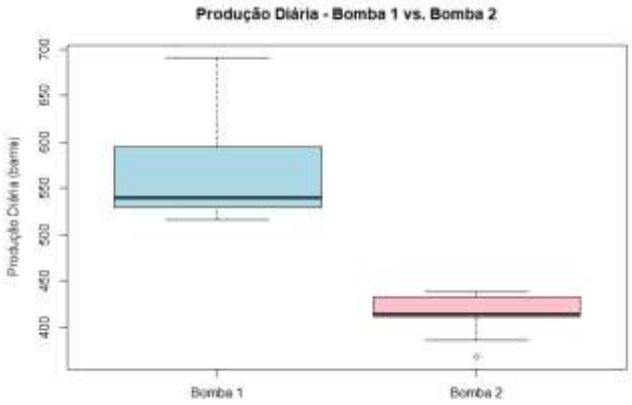

Figura 5 - ***Produção diária aleatória das bombas*** *1 e 2*

Utilizando as amostras aleatórias da anteriores para efetuámos um **teste de hipóteses** que permite 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.**

- **H0**: não há diferença significativa entre as médias das duas bombas
- **H1**: média da bomba 1 é superior à da bomba 2.

Extraimos os dados de produção diária para a amostra de dias na bomba 1 e bomba 2 e realizámos **um teste t pareado** (utilizado para comparar as médias de duas amostras pareadas ou dependentes):

```
producao_bomba1 <- c(medias_aleatorias$`M é dia Bomba 1`)
producao_bomba2 <- c(medias_aleatorias$`M é dia Bomba 2`)

resultado_teste <- t.test(producao_bomba1,
producao_bomba2, paired = TRUE)

valor_p <- resultado_teste$p.value
```

```
        Paired t-test

data:  producao_bomba1 and producao_bomba2
t = 7.5848, df = 8, p-value = 6.395e-05
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 109.8494 205.8227
sample estimates:
mean difference
       157.8361

>
```

Figura 6 - teste t pareado

- **Valor de t:** O valor de t calculado é 7.5848, indicando uma diferença significativa entre as médias das duas amostras pareadas (producao_bomba1 e producao_bomba2).

- **Valor de p:** O valor de p é 6.395e-05, menor que o nível de significância comum de 0.05,

indicando evidências estatísticas para rejeitar a hipótese nula de que não há diferença significativa entre as médias das duas amostras.

Como o valor do **p_value** obtido é menor que 0.05, então há evidência estatística de que a média da produção diária de petróleo da bomba 1 foi superior à da bomba 2 no período de 1-6-2013 a 31-5-2014, rejeitando-se a hipótese nula.

B. ***DADOS2.csv*** *- algoritmos de Machine Learning: SVM, DT, KN, RF, ML e GB*

A correlação entre a precisão de cada par de algoritmos foi avaliada desenvolvendo a matriz de correlações. Começamos por extrair, tanto os dados do ficheiro, como as colunas de precisão de cada algoritmo e criámos a matriz recorrendo à função ***rcorr:***

```
dados <- DADOS2
library(Hmisc)
precisao_svm <- dados$SVM
precisao_dt <- dados$DT
precisao_kn <- dados$KN
precisao_rf <- dados$RF
precisao_ml <- dados$ML
precisao_gb <- dados$GB
matriz_correlacoes <- rcorr(cbind(precisao_svm,
precisao_dt, precisao_kn, precisao_rf, precisao_ml,
precisao_gb))
```

```
            precisao_svm precisao_dt precisao_kn precisao_rf precisao_ml precisao_gb
precisao_svm    1.0000000   0.2619970   0.6374486   0.4659169   0.7111270   0.8629016
precisao_dt     0.2619970   1.0000000   0.4339395   0.8802559   0.6247459   0.2127059
precisao_kn     0.6374486   0.4339395   1.0000000   0.4834416   0.8539377   0.7502013
precisao_rf     0.4659169   0.8802559   0.4834416   1.0000000   0.5719541   0.3240401
precisao_ml     0.7111270   0.6247459   0.8539377   0.5719541   1.0000000   0.7211135
precisao_gb     0.8629016   0.2127059   0.7502013   0.3240401   0.7211135   1.0000000
>
```

Figura 7-matriz de correlação

Analisando a matriz existem cinco análises importantes a retirar:

- A precisão do algoritmo SVM possui uma correlação positiva moderada com a precisão dos algoritmos KN, RF, ML e GB, com coeficientes de correlação variando de 0,46 a 0,86. Isso sugere que quando a precisão do SVM aumenta, é provável que a precisão desses outros algoritmos também aumente, e vice-versa.

- A precisão do algoritmo DT possui correlação positiva moderada com a precisão dos algoritmos RF e ML, com coeficientes de correlação de 0,62 e 0,57, respectivamente. Isso indica que quando a precisão do DT aumenta, é provável que a precisão desses outros algoritmos também aumente.

- A precisão do algoritmo KN possui correlação positiva moderada com a precisão do algoritmo RF, com um coeficiente de correlação de 0,48. Isso sugere que quando a precisão do KN aumenta, é provável que a precisão do RF também aumente.

- A precisão do algoritmo RF possui correlação positiva moderada com a precisão do algoritmo ML, com um coeficiente de correlação de 0,57. Isso indica que quando a precisão do RF aumenta, é provável que a precisão do ML também aumente, e vice-versa.

- A precisão do algoritmo GB possui uma correlação positiva moderada com a precisão do algoritmo SVM, com um coeficiente de correlação de 0,86. Isso sugere que quando a precisão do GB aumenta, é provável que a precisão do SVM também aumente.

Para averiguar **se existem diferenças significativas entre a precisão** dos diferentes algoritmos recorremos a um **teste de hipóteses**:
Para isso usámos o ***Teste de Shapiro-Wilk,*** este é amplamente utilizado e recomendado para amostras de tamanho pequeno a moderado (n <= 50).

- **H0:** dados são normalmente distribuídos,
- **H1**: dados não são normalmente distribuídos.

- **Valor-p > 0,05**: aceitar a hipótese nula, dados podem seguir uma distribuição normal.
- **Valor-p < 0,05**: rejeitar a hipótese nula, dados não seguem uma distribuição normal.

```
shapiro.test(dados$SVM) #p-value=0.2687 não se rejeita
shapiro.test(dados$DT) #p-value=0.06772 não se rejeita
shapiro.test(dados$KN) #p-value=0.06926 não se rejeita
shapiro.test(dados$RF) #p-value=0.3138 não se rejeita
shapiro.test(dados$ML) #p-value=0.02138 rejeita-se H0
shapiro.test(dados$GB) # p-value=0.5125 não se rejeita
```

```
        Shapiro-Wilk normality test

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

> shapiro.test(dados$ML)  # p-value=0.02138

        Shapiro-Wilk normality test

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

> shapiro.test(dados$GB)  # p-value=0.5125 >

        Shapiro-Wilk normality test

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

Figura 8-teste de shapiro

Sendo que os dados não são normalmente distribuídos, recorre-se a um teste não paramétrico: ***Friedman***

```
friedman.test(cbind(dados$SVM, dados$DT, dados$KN,
dados$RF, dados$ML, dados$GB))
```

```
        Friedman rank sum test

data:  cbind(dados$SVM, dados$DT, dados$KN, dados$RF, dados$ML, dados$GB)
Friedman chi-squared = 8.7097, df = 5, p-value = 0.1212
```

Figura 9-teste de Friedman

O resultado mostrou:

- valor de chi-squared de 8.7097
- df de 5 graus de liberdade
- p-value= 0.1212 > 0.05
- não há evidências estatisticamente significativas para rejeitar a hipótese nula, o que significa que não há diferenças significativas entre a precisão dos diferentes algoritmos.

C. ***DADOS3.csv*** *- 4 variáveis (aceleração, número de cilindros, peso e potência) de 99 viaturas escolhidas aleatoriamente*

No que constam as 99 viaturas, se as mesmas fossem divididas em três grupos: **4, 6 e 8 cilindros**.
Existiram **diferenças significativas** na aceleração entre os três grupos?

Primeiramente abordámos a questão da normalização dos dados, se os mesmo não tiverem distribuição normal, a média não é uma representação dos dados.

Para isso usámos novamente o **Teste de Shapiro-Wilk onde:**

- **H0:** dados são normalmente distribuídos,
- **H1**: dados não são normalmente distribuídos.

- **Valor-p > 0,05**: aceitar a hipótese nula, dados podem seguir uma distribuição normal.
- **Valor-p < 0,05**: rejeitar a hipótese nula, dados não seguem uma distribuição normal.

Extraímos e armazenámos os dados de aceleração para cada grupo, realizámos o teste para cada, com a função ***shapiro.test*** e obtivemos ***o p-value*** para verificar a normalização**;**

```
dados <- DADOS3
cilindros4<- as.numeric(dados$Acceleration[dados$
Cylinders == 4])
cilindros6<- as.numeric(dados$Acceleration[dados$
Cylinders == 6])
cilindros8<- as.numeric(dados$Acceleration[dados$
Cylinders == 8])

res_shapiro_cilindros4 <- shapiro.test(cilindros4)
res_shapiro_cilindros6 <- shapiro.test(cilindros6)
res_shapiro_cilindros8 <- shapiro.test(cilindros8)

p_cilindros4 <- res_shapiro_cilindros4$p.value
p_cilindros6 <- res_shapiro_cilindros6$p.value
p_cilindros8 <- res_shapiro_cilindros8$p.value

if (res_shapiro_cilindros4$p.value > 0.05 &
  res_shapiro_cilindros6$p.value > 0.05 &
  res_shapiro_cilindros8$p.value > 0.05) {

print("Os grupos seguem uma distribuição normal.")
    cat("4_p-valor =", p_cilindros4, "\n")
    cat("6_p-valor =", p_cilindros6, "\n")
    cat("8_p-valor =", p_cilindros8, "\n")

    } else {
print("Os grupos não seguem uma distribuição normal.
Considere usar uma abordagem não param é trica.")
    cat("4_p-valor =", p_cilindros4, "\n")
    cat("6_p-valor =", p_cilindros6, "\n")
    cat("8_p-valor =", p_cilindros8, "\n")
    }
```

```
[1] "Os grupos não seguem uma distribuição normal.
4_p-valor = 0.1055678
6_p-valor = 0.03627768
8_p-valor = 0.2729467
```

Figura 10-teste de shapiro e p-values

Uma vez que os dados não passaram nos pressupostos de normalização, tivemos de usar um teste não paramétrico: ***Kruskal-Wallis*** para analisar as diferenças entre os grupos**;**

```
res_kruskal_wallis <- kruskal.test(list(cilindros4, cilindros6,
cilindros8))

p_kruskal_wallis <- resultado_kruskal_wallis$p.value

if (p_kruskal_wallis < 0.05) {
 cat("p-value =", p_kruskal_wallis, " < 0,05 -> H á diferenças
significativas entre os 3 grupos.")
} else {
cat("p-value =", p_kruskal_wallis, " > 0,05 -> Não h á diferenças
significativas entre os 3 grupos.")}
```

```
p-value = 2.795281e-11  < 0,05
```

Figura 11- p-value do teste Kruskal

```
Há diferenças significativas entre os 3 grupos.
```

Figura 12- conclusão do teste Kruskal

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

Modelo de **regressão linear** (usada para prever ou estimar o valor de uma variável dependente com base numa ou mais variáveis independentes):

No caso, a variável dependente é a aceleração e os cilindros (como uma variável *dummy*/*fator*), peso e potência as independentes. Um valor de p baixo < 0.05 indica uma relação estatisticamente significativa entre a variável independente e a variável dependente.

Primeiramente transformamos a variável "*Cylinders*" numa variável *dummy* (fator) . Aqui o objetivo é encontrar os coeficientes de regressão que melhor explicam a relação linear entre as variáveis.

```
dados$Cylinders <- as.factor(dados$Cylinders)

modelo <- lm(Acceleration ~ Cylinders + Weight + Horsepower, data = dados)
```

```
Call:
lm(formula = Acceleration ~ Cylinders + Weight + Horsepower,
    data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5617 -1.2154 -0.2788  0.9860  7.1221

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.7490563  1.4169610   9.703 8.68e-16 ***
Cylinders6  -1.5159977  0.7434813  -2.039 0.044283 *
Cylinders8  -4.8674507  1.2075112  -4.031 0.000114 ***
Weight       0.0031515  0.0006631   4.752 7.30e-06 ***
Horsepower  -0.0573811  0.0120787  -4.751 7.35e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.12 on 93 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6098,    Adjusted R-squared:  0.5931
F-statistic: 36.34 on 4 and 93 DF,  p-value: < 2.2e-16

> |
```

Figura 13- regressão linear

- ***Cylinders***: efeito significativo na aceleração, com veículos de 6 cilindros tendendo a ter uma aceleração reduzida em média de 1,52 unidades e veículos de 8 cilindros tendendo a ter uma aceleração reduzida em média de 4,87 unidades em comparação com os veículos de referência.

- ***Weight***: efeito significativo positivo na aceleração, com um aumento médio de 0,00315 unidades na aceleração para cada aumento de 1 unidade no peso do veículo.

- ***Horsepower***: efeito significativo negativo na aceleração, com uma redução média de 0,0574 unidades na aceleração para cada aumento de 1 unidade na potência do veículo.

- ***P-Value***: muito baixo (< 2.2e-16) indica que o modelo de regressão linear é globalmente significativo, ou seja, pelo menos uma das variáveis independentes é significativa na explicação da variabilidade na aceleração do veículo.

Tendo em conta esta última situação, como seria a **aceleração de uma viatura com um peso de 2950 kg, potência de 100 Hp e 4 cilindros.**

Criámos um novo data frame com os valores de entrada e transformamos o 4 em fator, pois estamos a falar dum valor específico (variável categórica, que representa características discretas).
O weight e horsepower são variáveis contínuas, ou seja, representam quantidades numéricas em uma escala contínua. Usámos também, a função *predict*() para estimar a aceleração com base no modelo ajustado e nos novos dados

```
novos_dados <- data.frame(Cylinders = as.factor(4), Weight = 2950, Horsepower = 100)

estimativa <- predict(modelo, newdata = novos_dados)

cat("Estimativa de aceleração de uma viatura com um peso de 2950 kg, potência de 100 Hp e 4 cilindros:", estimativa, "\n")
```

```
Estimativa de aceleração
s: 17.30784
```

Figura 14- regressão linear

Figura 15- regressão linear

- O modelo de regressão linear ajustado foi significativo, indicando que as variáveis independentes (*Weight, Horsepower e Cylinders*) têm uma relação estatisticamente significativa com a variável dependente (*Acceleration*).

- Os coeficientes estimados das variáveis independentes sugerem que o número de cilindros (*Cylinders*) tem uma relação negativa com a aceleração, enquanto o peso do veículo (*Weight*) e a potência do motor (*Horsepower*) têm uma relação positiva com a aceleração.

- A estimativa de aceleração para uma viatura com um peso de 2950 kg, potência de 100 Hp e 4 cilindros foi de aproximadamente 17.31, com base no modelo ajustado.

V. CONCLUSÕES

Na realização do trabalho foi possível salientar a importância que uma boa análise e um estudo intuitivo dos dados conferem na interpretação dos mesmos. Através da elaboração de gráficos dinâmicos que permitem por exemplo comparar a temperatura do motor em diferentes bombas e avaliar a produção de barris de petróleo produzida, é possível perceber em que aspetos pode ser viável a otimização do rendimento da produção de uma empresa de extração de petróleo . Por outro lado, a criação de uma matriz de correlações e a utilização de um teste de hipótese não paramétrico, auxiliam a comparação entre algoritmos de Machine Learning. Por fim, na tentativa de perceber se a quantidade de cilindros afetava a acelaração de um carro, foram uteis os testes de Shapiro e kruskal.

VI. REFERÊNCIAS

[4] "Linear Regression for Machine Learning." https://machinelearningmastery.com/linear-regression-for-machine-learning/ (accessed Jun. 15, 2022).

[5] "About Linear Regression | IBM." https://www.ibm.com/topics/linear-regression (accessed Jun. 15, 2022).

[6] "Coeficientes de correlação: Para que servem e como interpreta-los?" https://operdata.com.br/blog/coeficientes-de-correlacao/ (accessed Jun. 15, 2022).

[7] "Correlação de Pearson: de que trata esse coeficiente?" https://www.questionpro.com/blog/pt-br/correlacao-de-pearson/ (accessed Jun. 15, 2022).

[8] "Análise de regressão: Como interpretar o R-quadrado e avaliar a qualidade de ajuste?" https://blog.minitab.com/pt/analise-de-regressao-como-interpretar-o-r-quadrado-e-avaliar-a-qualidade-de-ajuste (accessed Jun. 16, 2022).

[9] "Análise de Regressão Simples x Análise de Regressão Múltipla - Oper." https://operdata.com.br/blog/analise-de-regressao-simples-x-analise-de-regressao-multipla/ (accessed Jun. 16, 2022)."