PACDII: QUIZ II

Grupo 7

30 de setembro, 2023

Elementos do grupo:

- Allan Kardec da Silva Rodrigues, nº 103380
- André Plancha Fernandes, nº 105289
- Diogo Alexandre Alonso de Freitas, nº 104841
- João Francisco Marques Gonçalves da Silva Botas, nº 104782
- Marco Delgado Esperança, nº 110451

Nota:

Deve efetuar todos os Save com "Save with encoding UTF-8" de modo a manter palavras acentuadas e caracteres especiais**

Base de dados:condutores.csv

```
# Remover tudo!
rm(list=ls(all=TRUE))

# # Incluir as libraries de que necessita
# pacman::p_load(VIM, tidyverse, conflicted, skimr, ggplot2, lsr,
# lubridate, nycflights13, tidyverse, dplyr, psych, tree)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tree)
library(conflicted)
conflicts_prefer(dplyr::filter)
```

```
## [conflicted] Will prefer dplyr::filter over any other package.
```

Questão 1 [5 valores]

Leitura dos dados condutores.csv.

```
condutores <-read.csv("condutores.csv", header=TRUE, stringsAsFactors = T, sep=",",dec=",",check.names="
```

Remoção dos valores omissos das variáveis Tempo.Condução.Continuada e Ano.matricula

```
condutoresLimpo <- condutores[!(is.na(condutores$Ano.matricula)), ]
condutoresLimpo <- condutoresLimpo[!(is.na(condutoresLimpo$Tempo.Condução.Continuada)), ]
```

Crie variável métrica Idade.Veiculo (2020-Ano.matricula).

```
condutoresLimpo$Idade.Veiculo <- 2020 - condutoresLimpo$Ano.matricula
# condutoresLimpo %>% select(Ano.matricula, Idade.Veiculo)
```

Crie a variável nominal Idade.Condutor com as classes "< 15", "15-17", "18-20", "21-29", "30-39", "40-49", "50-59", "65-69", ">= 70".

```
condutoresLimpoIdade <- condutoresLimpo %>%
  mutate(Idade.Condutor = case_when(
    Condutor.Gr.Etario...5..SUM      == 1 ~ "< 15",
    Condutor.Gr.Etario.6.9..SUM      == 1 ~ "< 15",
    Condutor.Gr.Etario.10.14..SUM    == 1 ~ "< 15",
    Condutor.Gr.Etario.15.17..SUM    == 1 ~ "15-17",
    Condutor.Gr.Etario.18.20..SUM    == 1 ~ "18-20",
    Condutor.Gr.Etario.21.24..SUM    == 1 ~ "21-29",
    Condutor.Gr.Etario.25.29..SUM    == 1 ~ "21-29",
    Condutor.Gr.Etario.30.34..SUM    == 1 ~ "30-39",
    Condutor.Gr.Etario.35.39..SUM    == 1 ~ "30-39",
    Condutor.Gr.Etario.40.44..SUM    == 1 ~ "40-49",
    Condutor.Gr.Etario.45.49..SUM    == 1 ~ "40-49",
    Condutor.Gr.Etario.50.54..SUM    == 1 ~ "50-59",
    Condutor.Gr.Etario.55.59..SUM    == 1 ~ "50-59",
    Condutor.Gr.Etario.65.69..SUM    == 1 ~ "65-69",
    Condutor.Gr.Etario.70.74..SUM    == 1 ~ ">= 70",
    Condutor.Gr.Etario...75..SUM     == 1 ~ ">= 70",
    Condutor.Gr.Etario.Não.Def...SUM == 1 ~ NA,
    TRUE ~ NA
  ) %>% forcats::as_factor()) %>% select(-c(Condutor.Gr.Etario...5..SUM:Condutor.Gr.Etario.Não.Def...SU
# condutoresLimpoIdade %>% select(Idade.Condutor)
```

Remova os valores omissos da variável Idade.Condutor

```
condutoresLimpoIdade %>% filter(!is.na(Idade.Condutor)) -> condutoresMaisLimpo
condutoresMaisLimpo %>% nrow()
```

```
## [1] 34814
```

Usando set.seed(500),efetue a divisão dos dados Data em amostra de treino (70%) e de teste (30%) e apresente uma tabela com a média, desvio padrão, mediana, amplitude, assimetria e curtose da variável Idade.Veiculo em cada amostra.

```
set.seed(500)
split <- sample(nrow(condutoresMaisLimpo), (nrow(condutoresMaisLimpo) * 0.7) %>% round())
treino <- condutoresMaisLimpo[split,]
teste <- condutoresMaisLimpo[-split,]

describeIt <- function(x) {
  psych::describe(x) %>%
    select(média = mean, desvio.padrao = sd, mediana = median, amplitude = range, assimetria = skew, cu
    t() %>% as.data.frame()
}
## Join them
describeIt(treino$Idade.Veiculo) -> treino.describe
describeIt(teste$Idade.Veiculo)   -> teste.describe
tibble(Estatisticas = rownames(treino.describe), treino.Idade.Veiculo = treino.describe$X1, teste.Idade
```

```
## # A tibble: 6 x 3
##   Estatisticas  treino.Idade.Veiculo teste.Idade.Veiculo
##   <chr>                       <dbl>               <dbl>
## 1 média                      13.0                13.1
## 2 desvio.padrao               8.52                8.70
## 3 mediana                    13                  13
## 4 amplitude                 106                  76
## 5 assimetria                  0.442               0.451
## 6 curtose                     0.395               0.124
```

Questão 2 [5 valores]

Obtenha um modelo em árvore, sobre a amostra de treino, sem utilizar poda, considerando as variáveis preditoras Tempo.Condução.Continuada, Idade.Condutor e a parametrização mincut = 5, minsize = 10, mindev = 0.001 e split = "deviance".

Estime Idade.Veiculo sobre amostra de teste, a partir da árvore obtida, e apresente as estimativas correspondentes às 10 primeiras observações desta amostra.

```
modelo <- tree::tree(
  Idade.Veiculo ~ Tempo.Condução.Continuada + Idade.Condutor,
  data = treino, control=tree.control(nrow(treino), mincut = 5,
  minsize = 10, mindev = 0.001), split = "deviance"
)
previsoes <- predict(modelo, newdata=teste)
## join tables
teste %>% mutate(previsao = previsoes) %>% select(Idade.Veiculo, previsao) %>% head(10)
```

```
##    Idade.Veiculo previsao
## 3              5 11.21787
## 4              2 14.08064
## 6             21 11.21787
## 7              5 12.93683
## 20            24 12.93683
```

```
## 21            24 12.93683
## 26            16 11.21787
## 27            10 11.21787
## 28            19 12.93683
## 32            22 12.93683
```

```
modelo %>% summary()
```

```
##
## Regression tree:
## tree::tree(formula = Idade.Veiculo ~ Tempo.Condução.Continuada +
##     Idade.Condutor, data = treino, control = tree.control(nrow(treino),
##     mincut = 5, minsize = 10, mindev = 0.001), split = "deviance")
## Number of terminal nodes:  7
## Residual mean deviance:  69.59 = 1695000 / 24360
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -17.0200  -7.2180  -0.2179   0.0000   6.0630  91.9200
```

Questão 3 [5 valores] Apresente os valores das métricas MSE (Mean Squared Error), RMSE (Root Mean Square Squared Error) e MAE (Mean Absolute Error) associados ao modelo aplicado sobre cada uma das amostras (Treino e Teste). Comente se há overfitting.

```
previsoes_treino <- predict(modelo, treino)
previsoes_teste <-  predict(modelo, teste)

tribble(
  ~Amostra, ~MSE, ~RMSE, ~MAE,
  "Treino", Metrics::mse(actual = treino$Idade.Veiculo, predicted = previsoes_treino), Metrics::rmse(ac
  "Teste", Metrics::mse(actual = teste$Idade.Veiculo,  predicted = previsoes_teste ), Metrics::rmse(actu
)
```

```
## # A tibble: 2 x 4
##   Amostra   MSE  RMSE   MAE
##   <chr>   <dbl> <dbl> <dbl>
## 1 Treino   69.6  8.34  6.97
## 2 Teste    72.3  8.50  7.13
```

As métricas do conjunto de treino e do teste são semelhantes, contudo verifica-se que no conjunto de teste o modelo prevê pior, sendo possível visualizar este contraste nos valores das métricas MSE, RMSE E MAE, que pode indiciar ligeiro overfitting. No entanto, esta conclusão é argumentável pois os valores são bastante próximos, como já referido.

Questão 4 [5 valores] Complete as frases seguintes em comentário do script:

A Árvore de Regressão é constituída por **7** nós folha; a Residual Deviance associada ao modelo sobre a amostra de teste é **755428**; o erro quadrático de previsão, relativo a Idade.Veiculo, para a primeira observação do conjunto teste é **38.66**. Para reduzir a complexidade do modelo em árvore o valor do argumento mindev da function tree deve ser alterado para **0.01**(selecione um dos seguintes valores: 0.01; 0.0001).

1

```
modelo
```

```
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 24370 1767000 12.95
##    2) Idade.Condutor: 40-49,21-29,18-20,50-59,30-39,15-17,< 15 20919 1458000 12.40
##      4) Tempo.Condução.Continuada: De 1 a 3 horas,De 3 a 5 horas,Ignorada,Mais de 5 horas 10128  675
##        8) Idade.Condutor: 40-49,21-29,30-39,15-17,< 15 7642  495100 11.22 *
##        9) Idade.Condutor: 18-20,50-59 2486  176300 12.68 *
##      5) Tempo.Condução.Continuada: Menos de 1 hora 10791  768800 13.18
##       10) Idade.Condutor: 15-17 176   14740  8.21 *
##       11) Idade.Condutor: 40-49,21-29,18-20,50-59,30-39,< 15 10615  749700 13.26
##         22) Idade.Condutor: 40-49,21-29,30-39,< 15 7614  517700 12.94 *
##         23) Idade.Condutor: 18-20,50-59 3001  229200 14.08 *
##    3) Idade.Condutor: >= 70,65-69 3451  265000 16.29
##      6) Tempo.Condução.Continuada: De 1 a 3 horas,De 3 a 5 horas,Ignorada,Mais de 5 horas 1429  1027
##      7) Tempo.Condução.Continuada: Menos de 1 hora 2022  159700 17.02 *
```

2

```
## Residual Deviance do teste
predicao <- predict(modelo, teste)
(residualDeviance <- Metrics::sse(teste$Idade.Veiculo, predicao))
```

```
## [1] 755427.7
```

3

```
## Erro quadrático de previsão
(predicao[1] - teste$Idade.Veiculo[1])^2
```

```
##        3
## 38.66197
```

Tarefa final: Submeta, no Moddle, um ficheiro pdf resultado da compilação do TEMPLATE_QUIZ2.

Caso os resultados apresentados não sejam coerentes com as respostas dadas, a classificação será penalizada.