PACDII: QUIZ II

grupo 4

2023-09-28

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())

# Incluir as libraries de que necessita

library(caret)

library(plyr)
library(ggplot2)
library(dplyr)

library(psych)

library(tree)
library(VIM)

library(plyr)
library(lubridate)

library(tidyverse)

library(rpart)
```

Questão 1 [5 valores]

Leitura dos dados condutores.csv.

```
condutores <-read.csv("condutores.csv", header=TRUE, stringsAsFactors =
T, sep=",",dec=",",check.names=F,na.strings=c("NA","NÃO
DEFINIDO"),fileEncoding = "UTF-8")
```

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

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

```
## Rows: 37,071
## Columns: 40
## $ `Id. Acidente`                    <dbl> 20201824120, 20201824546,
202018245…
## $ Datahora                          <fct> 2020:01:01 00:20:00,
2020:01:01 01:…
## $ Sexo                              <fct> Masculino, Masculino,
Feminino, Mas…
## $ `Lesões a 30 dias`                <fct> Ferido leve, Ileso, Ileso,
Ferido l…
## $ `Licença Condução`                <fct> Com licença/ carta adequada
ao veic…
## $ `Teste Alcool`                    <fct> Submetido ao teste do
álcoolemia, S…
## $ `Acções Condutores`               <fct> Em marcha normal, Em marcha
normal,…
## $ `Inf. Comp. a Acções e Manobras`  <fct> Não identificada, Não
identificada,…
## $ Nomeoutrosfactores                <fct> Normal, Normal, Normal,
Sono/sonolê…
## $ `Tempo Condução Continuada`       <fct> Menos de 1 hora, Ignorada,
Ignorada…
## $ `Acessórios Condutores`           <fct> Cinto de segurança, Cinto
de segura…
## $ `Categoria Veículos`              <fct> Automóvel ligeiro,
Automóvel ligeir…
## $ `Tipo Veiculo`                    <fct> Passageiros, Passageiros,
Passageir…
## $ `Tipo Serviço`                    <fct> Particular, Particular,
Particular,…
## $ `Veiculo Especial`                <fct> NA, NA, NA, NA, NA, NA, NA,
NA, NA,…
## $ `Ano matricula`                   <int> 2014, 2016, 2015, 2018,
1996, 1999,…
## $ `Inspecção Periódica`             <fct> Válida, Não obrigatória,
```

```
Válida, Nã…
## $ `Certificado Adr`                  <fct> NA, NA, NA, NA, NA, NA, NA,
NA, NA,…
## $ `Carga Lotação`                    <fct> Sem carga, Sem carga, Sem
carga, Se…
## $ Pneus                              <fct> Sem deficiência, Sem
deficiência, S…
## $ Seguros                            <fct> Com seguro, Com seguro, Com
seguro,…
## $ Distrito                           <fct> Porto, Faro, Faro, Lisboa,
Faro, Li…
## $ Concelho                           <fct> Maia, Albufeira, Albufeira,
Cascais…
## $ `Condutor Gr.Etario(<=5) SUM`      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0,…
## $ `Condutor Gr.Etario(6-9) SUM`      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0,…
## $ `Condutor Gr.Etario(10-14) SUM`    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0,…
## $ `Condutor Gr.Etario(15-17) SUM`    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0,…
## $ `Condutor Gr.Etario(18-20) SUM`    <int> 0, 0, 0, 1, 0, 0, 0, 0, 0,
1, 0, 0,…
## $ `Condutor Gr.Etario(21-24) SUM`    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0,…
## $ `Condutor Gr.Etario(25-29) SUM`    <int> 0, 0, 1, 0, 0, 1, 0, 0, 0,
0, 1, 0,…
## $ `Condutor Gr.Etario(30-34) SUM`    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0,…
## $ `Condutor Gr.Etario(35-39) SUM`    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0,…
## $ `Condutor Gr.Etario(40-44) SUM`    <int> 0, 0, 0, 0, 1, 0, 0, 0, 1,
0, 0, 0,…
## $ `Condutor Gr.Etario(45-49) SUM`    <int> 0, 1, 0, 0, 0, 0, 1, 0, 0,
0, 0, 0,…
## $ `Condutor Gr.Etario(50-54) SUM`    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0,
0, 0, 0,…
## $ `Condutor Gr.Etario(55-59) SUM`    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 1,…
## $ `Condutor Gr.Etario(65-69) SUM`    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0,…
## $ `Condutor Gr.Etario(70-74) SUM`    <int> 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0,…
## $ `Condutor Gr.Etario(>=75) SUM`     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0,…
## $ `Condutor Gr.Etario(Não Def.) SUM` <int> 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0,…
```

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

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".

Remova os valores omissos da variável Idade.Condutor

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.

```
condutores$`Tempo Condução Continuada` -> condutores$TempoContinuada
condutores$`Tempo Condução Continuada`<- NULL

condutores$Idade.veiculo<-rep(NA, 37071)
condutores$Idade.veiculo<- 2020 - condutores$`Ano matricula`
condutores$Idade.condutor<-rep(NA, 37071)
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(<=5) SUM`
>= 1)]<- "< 15"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(6-9)
SUM`==1)]<- "< 15"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(10-14)
SUM`==1)]<- "< 15"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(15-17)
SUM`==1)]<- "15 - 17"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(18-20)
SUM`==1)]<- "18-20"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(21-24)
SUM`==1)]<- "21-29"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(25-29)
SUM`==1)]<- "21-29"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(30-34)
SUM`==1)]<- "30-39"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(35-39)
SUM`==1)]<- "30-39"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(40-44)
SUM`==1)]<- "40-49"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(45-49)
SUM`==1)]<- "40-49"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(50-54)
SUM`==1)]<- "50-59"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(55-59)
SUM`==1)]<- "50-59"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(65-69)
SUM`==1)]<- "65-69"
```

```
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(70-74)
SUM`==1)]<- ">=70"
condutores$Idade.condutor[which(condutores$`Condutor Gr.Etario(>=75)
SUM`==1)]<- ">=70"
condutores <- condutores[!(is.na(condutores$Idade.condutor)), ]


set.seed(500)
index<-sample(1:nrow(condutores),round(nrow(condutores)*0.7))
train<-condutores[index,]
test<-condutores[-index,]

describe(train$Idade.veiculo)[,c(3,4,5,10,11,12)]

##     mean   sd median range skew kurtosis
## X1 12.95 8.52     13   106 0.44     0.39

describe(test$Idade.veiculo)[,c(3,4,5,10,11,12)]

##     mean  sd median range skew kurtosis
## X1 13.06 8.7     13    76 0.45     0.12
```

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_arvore <- tree(Idade.veiculo ~ TempoContinuada + Idade.condutor,
data = train , mincut = 5, minsize = 10, mindev = 0.001,split =
"deviance")

## Warning in tree(Idade.veiculo ~ TempoContinuada + Idade.condutor, data
= train,
## : NAs introduced by coercion

pred.rtree.condutores<-predict(modelo_arvore, test)

## Warning in pred1.tree(object, tree.matrix(newdata)): NAs introduced by
coercion
```

```r
round(pred.rtree.condutores[1:10], 0)
```

```
##  3  4  6  7 21 22 27 29 30 35
## 12 14 12 14 14 14 12 12 14 14
```

```r
summary(modelo_arvore)
```

```
##
## Regression tree:
## tree(formula = Idade.veiculo ~ TempoContinuada + Idade.condutor,
##     data = train, split = "deviance", mincut = 5, minsize = 10,
##     mindev = 0.001)
## Variables actually used in tree construction:
## [1] "TempoContinuada"
## Number of terminal nodes:  2
## Residual mean deviance:  71.76 = 1749000 / 24370
## Distribution of residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -13.78000  -7.78400  -0.03236   0.00000   6.21600  92.22000
```

```r
modelo_arvore
```

```
## node), split, n, deviance, yval
##       * denotes terminal node
##
## 1) root 24370 1767000 12.95
##   2) TempoContinuada: De 1 a 3 horas,De 3 a 5 horas,Ignorada,Mais de 5
horas 11557  795100 12.03 *
##   3) TempoContinuada: Menos de 1 hora 12813  953600 13.78 *
```

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.

```r
RMSE(obs = test$Idade.veiculo, pred = pred.rtree.condutores)
```

```
## [1] 8.643033
```

```r
MAE(pred.rtree.condutores, test$Idade.veiculo)
```

```
## [1] 7.245458
```

```r
RMSE(obs = test$Idade.veiculo, pred = pred.rtree.condutores)^2
```

```
## [1] 74.70202
```

```r
RMSE(obs = train$Idade.veiculo, pred = pred.rtree.condutores)
```

```
## Warning in pred - obs: longer object length is not a multiple of
shorter object
## length

## [1] 8.560966

MAE(pred.rtree.condutores, train$Idade.veiculo)

## Warning in pred - obs: longer object length is not a multiple of
shorter object
## length

## [1] 7.164569

RMSE(obs = train$Idade.veiculo, pred = pred.rtree.condutores)^2

## Warning in pred - obs: longer object length is not a multiple of
shorter object
## length

## [1] 73.29014
```

Apesar de haver um erro no conjunto de teste superior ao conjunto de treino não é considerado overfitting pois não existe muita diferença

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

Script auxiliar:

```
summary(modelo_arvore)

##
## Regression tree:
## tree(formula = Idade.veiculo ~ TempoContinuada + Idade.condutor,
##     data = train, split = "deviance", mincut = 5, minsize = 10,
##     mindev = 0.001)
## Variables actually used in tree construction:
## [1] "TempoContinuada"
## Number of terminal nodes:  2
## Residual mean deviance:  71.76 = 1749000 / 24370
## Distribution of residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -13.78000  -7.78400  -0.03236   0.00000   6.21600  92.22000

sum((test$Idade.veiculo - pred.rtree.condutores)^2)

## [1] 780187.9

(pred.rtree.condutores[1]-test$Idade.veiculo[1])^2

##        3
## 49.45411
```

```
#A Árvore de Regressão é constituída por  2 nós folha;  a Residual
Deviance associada ao modelo sobre a amostra de teste é 780187.9; o erro
quadrático de previsão, relativo a Idade.Veiculo, para a primeira
observação do conjunto teste é 49.45411. 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).
```

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.