**ANÁLISE PREDITIVA E DATA MINING**

**Turma 41BDT**

Trabalho Individual

**Parte 1) Árvore de Decisão**

# Técnica de Discriminação: árvore.

# limpar memória do R

rm(list=ls(all=TRUE))  
install.packages("rpart")   
install.packages("rpart.plot")

# importar os arquivos

Titanic\_train <- read.csv("C:/Users/regin/FIAP/R/titanic")  
Titanic\_test <- read.csv("C:/Users/regin/FIAP/R/titanic")

View(Titanic\_train)

str(Titanic\_train)

> str(Titanic\_train)

'data.frame': 891 obs. of 12 variables:

$ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...

$ Survived : int 0 1 1 1 0 0 0 0 1 1 ...

$ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...

$ Name : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...

$ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...

$ Age : num 22 38 26 35 35 NA 54 2 27 14 ...

$ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...

$ Parch : int 0 0 0 0 0 0 0 1 2 0 ...

$ Ticket : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...

$ Fare : num 7.25 71.28 7.92 53.1 8.05 ...

$ Cabin : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...

$ Embarked : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...

attach(Titanic\_train)

summary(Titanic\_train)

> summary(Titanic\_train)

PassengerId Survived Pclass

Min. : 1.0 Min. :0.0000 Min. :1.000

1st Qu.:223.5 1st Qu.:0.0000 1st Qu.:2.000

Median :446.0 Median :0.0000 Median :3.000

Mean :446.0 Mean :0.3838 Mean :2.309

3rd Qu.:668.5 3rd Qu.:1.0000 3rd Qu.:3.000

Max. :891.0 Max. :1.0000 Max. :3.000

Name Sex Age

Abbing, Mr. Anthony : 1 female:314 Min. : 0.42

Abbott, Mr. Rossmore Edward : 1 male :577 1st Qu.:20.12

Abbott, Mrs. Stanton (Rosa Hunt) : 1 Median :28.00

Abelson, Mr. Samuel : 1 Mean :29.70

Abelson, Mrs. Samuel (Hannah Wizosky): 1 3rd Qu.:38.00

Adahl, Mr. Mauritz Nils Martin : 1 Max. :80.00

(Other) :885 NA's :177

SibSp Parch Ticket Fare Cabin

Min. :0.000 Min. :0.0000 1601 : 7 Min. : 0.00 :687

1st Qu.:0.000 1st Qu.:0.0000 347082 : 7 1st Qu.: 7.91 B96 B98 : 4

Median :0.000 Median :0.0000 CA. 2343: 7 Median : 14.45 C23 C25 C27: 4

Mean :0.523 Mean :0.3816 3101295 : 6 Mean : 32.20 G6 : 4

3rd Qu.:1.000 3rd Qu.:0.0000 347088 : 6 3rd Qu.: 31.00 C22 C26 : 3

Max. :8.000 Max. :6.0000 CA 2144 : 6 Max. :512.33 D : 3

(Other) :852 (Other) :186

Embarked

: 2

C:168

Q: 77

S:644

Survived<-as.factor(Survived)

Pclass<-as.factor(Pclass)

attach(Titanic\_train)

summary(Titanic\_train)

> summary(Titanic\_train)

PassengerId Survived Pclass Name

Min. : 1.0 0:549 1:216 Abbing, Mr. Anthony : 1

1st Qu.:223.5 1:342 2:184 Abbott, Mr. Rossmore Edward : 1

Median :446.0 3:491 Abbott, Mrs. Stanton (Rosa Hunt) : 1

Mean :446.0 Abelson, Mr. Samuel : 1

3rd Qu.:668.5 Abelson, Mrs. Samuel (Hannah Wizosky): 1

Max. :891.0 Adahl, Mr. Mauritz Nils Martin : 1

(Other) :885

Sex Age SibSp Parch Ticket

female:314 Min. : 0.42 Min. :0.000 Min. :0.0000 1601 : 7

male :577 1st Qu.:20.12 1st Qu.:0.000 1st Qu.:0.0000 347082 : 7

Median :28.00 Median :0.000 Median :0.0000 CA. 2343: 7

Mean :29.70 Mean :0.523 Mean :0.3816 3101295 : 6

3rd Qu.:38.00 3rd Qu.:1.000 3rd Qu.:0.0000 347088 : 6

Max. :80.00 Max. :8.000 Max. :6.0000 CA 2144 : 6

NA's :177 (Other) :852

Fare Cabin Embarked

Min. : 0.00 :687 : 2

1st Qu.: 7.91 B96 B98 : 4 C:168

Median : 14.45 C23 C25 C27: 4 Q: 77

Mean : 32.20 G6 : 4 S:644

3rd Qu.: 31.00 C22 C26 : 3

Max. :512.33 D : 3

(Other) :186

Titanic\_train$Survived <- factor(Titanic\_train$Survived)

Titanic\_train$Pclass <- factor(Titanic\_train$Pclass)

# sibsp Number of Siblings/Spouses Aboard ( Número de irmãos / cônjuges a bordo)

> Titanic\_train$SibSp <- factor(Titanic\_train$SibSp)

# parch Number of Parents/Children Aboard (Número de pais / filhos a bordo)

> Titanic\_train$Parch <- factor(Titanic\_train$Parch)

> summary(Titanic\_train)

PassengerId Survived Pclass Name Sex

Min. : 1.0 0:549 1:216 Abbing, Mr. Anthony : 1 female:314

1st Qu.:223.5 1:342 2:184 Abbott, Mr. Rossmore Edward : 1 male :577

Median :446.0 3:491 Abbott, Mrs. Stanton (Rosa Hunt) : 1

Mean :446.0 Abelson, Mr. Samuel : 1

3rd Qu.:668.5 Abelson, Mrs. Samuel (Hannah Wizosky): 1

Max. :891.0 Adahl, Mr. Mauritz Nils Martin : 1

(Other) :885

Age SibSp Parch Ticket Fare Cabin Embarked

Min. : 0.42 0:608 0:678 1601 : 7 Min. : 0.00 :687 : 2

1st Qu.:20.12 1:209 1:118 347082 : 7 1st Qu.: 7.91 B96 B98 : 4 C:168

Median :28.00 2: 28 2: 80 CA. 2343: 7 Median : 14.45 C23 C25 C27: 4 Q: 77

Mean :29.70 3: 16 3: 5 3101295 : 6 Mean : 32.20 G6 : 4 S:644

3rd Qu.:38.00 4: 18 4: 4 347088 : 6 3rd Qu.: 31.00 C22 C26 : 3

Max. :80.00 5: 5 5: 5 CA 2144 : 6 Max. :512.33 D : 3

NA's :177 8: 7 6: 1 (Other) :852 (Other) :186

table(Titanic\_train$Survived)

prop.table(table(Titanic\_train$Survived))

> prop.table(table(Titanic\_train$Survived))

0 1

0.6161616 0.3838384

#comando para gerar em 2 linhas e 3 colunas os plots

par (mfrow=c(2,3))

plot(as.factor(Titanic\_train$Pclass), as.factor(Titanic\_train$Survived),main='Pclass')

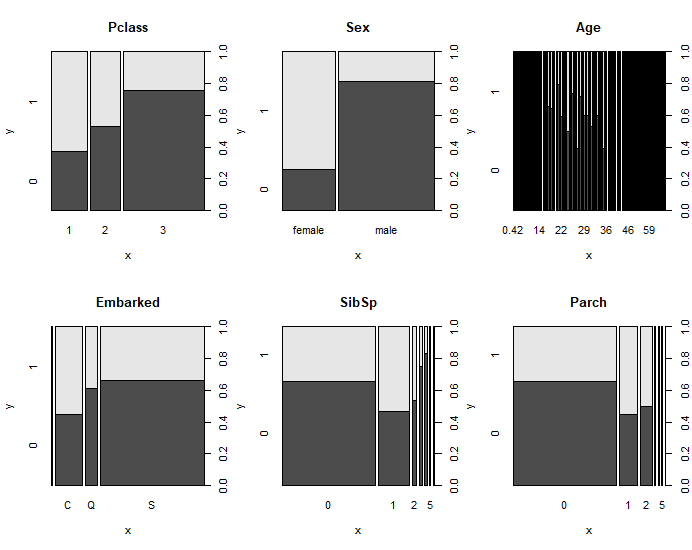
plot(Titanic\_train$Sex, as.factor(Titanic\_train$Survived),main='Sex')

plot(as.factor(Titanic\_train$Age), as.factor(Titanic\_train$Survived),main='Age')

plot(Titanic\_train$Embarked, as.factor(Titanic\_train$Survived),main='Embarked')

plot(as.factor(Titanic\_train$SibSp), as.factor(Titanic\_train$Survived),main='SibSp')

plot(as.factor(Titanic\_train$Parch), as.factor(Titanic\_train$Survived), main='Parch')



par (mfrow=c(1,1))

# column percentages

prop.table(table(as.factor(Titanic\_train$Pclass),as.factor(Titanic\_train$Survived)),2)

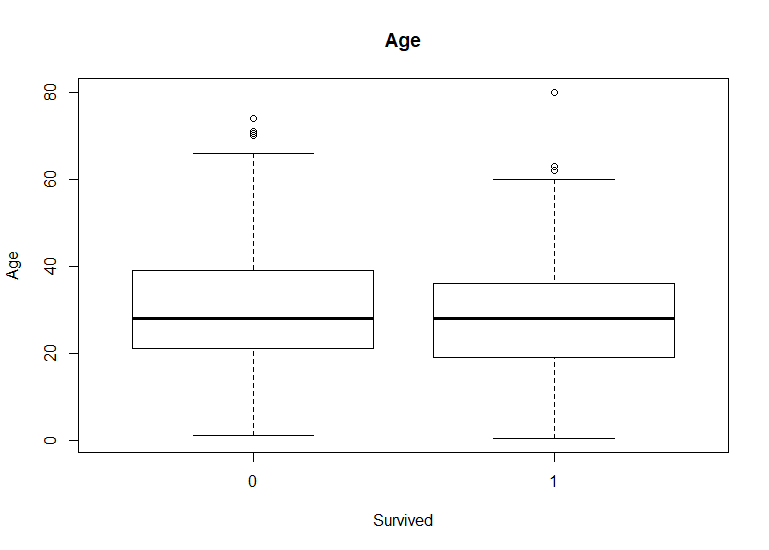
> prop.table(table(as.factor(Titanic\_train$Pclass),as.factor(Titanic\_train$Survived)),2)

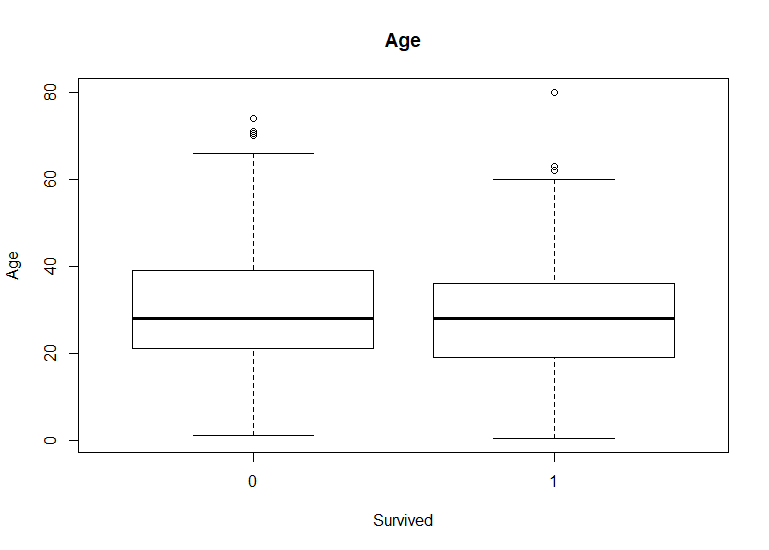
0 1

1 0.1457195 0.3976608

2 0.1766849 0.2543860

3 0.6775956 0.3479532

attach(Titanic\_train)  
table(as.factor(Age), Survived, useNA = "ifany")  
  
boxplot(Age ~ Survived, main='Age')

boxplot(Fare ~ Survived , main='Fare')

# Carrega o pacote: árvore de decisão

library(rpart)

library(rpart.plot)

attach(Titanic\_train)

# informaçoes dos Parãmetros do Modelo

rpart.model01 <- rpart (Survived ~ Pclass+Sex+Age+Embarked+Parch+Fare, maxdepth=10, Titanic\_train)

# Faz o Gráfico (type=0 a 4) (extra=0 a 9)

rpart.plot(rpart.model01, type=4, extra=104, under=FALSE, clip.right.labs=TRUE,

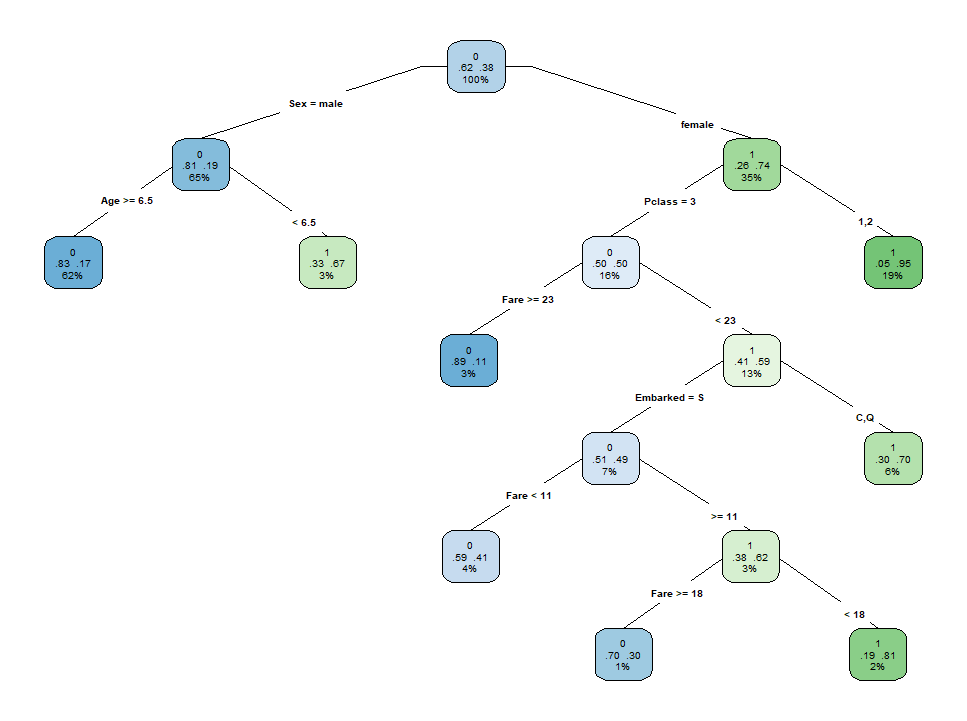
fallen.leaves=FALSE,

digits=2, varlen=-8, faclen=10,

cex=0.6, tweak=1,

compress=TRUE,

snip=FALSE)



# veja as opções

rpart.plot(rpart.model01, type=5, extra=104, under=FALSE, clip.right.labs=TRUE,

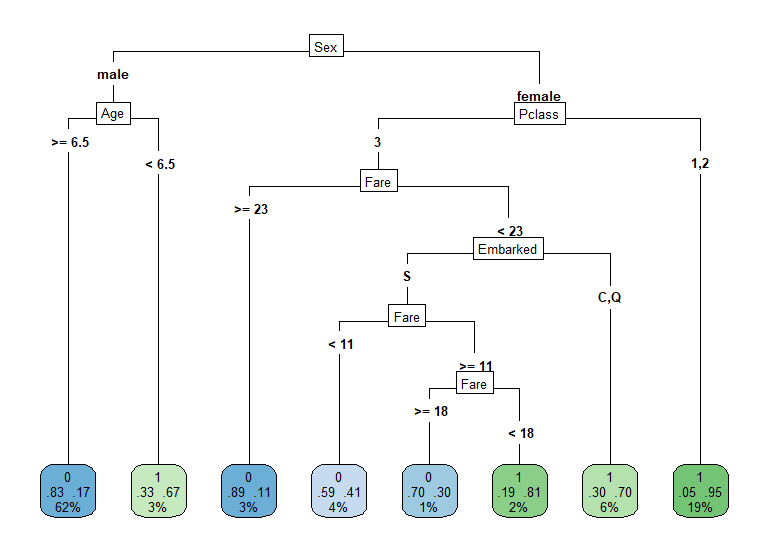
fallen.leaves=TRUE,

digits=2, varlen=-8, faclen=10,

cex=0.8, tweak=1,

compress=TRUE,

snip=FALSE)



# printcp(rpart.model01) # display the results

> print(rpart.model01)

n= 891

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 891 342 0 (0.61616162 0.38383838)

2) Sex=male 577 109 0 (0.81109185 0.18890815)

4) Age>=6.5 553 93 0 (0.83182640 0.16817360) \*

5) Age< 6.5 24 8 1 (0.33333333 0.66666667) \*

3) Sex=female 314 81 1 (0.25796178 0.74203822)

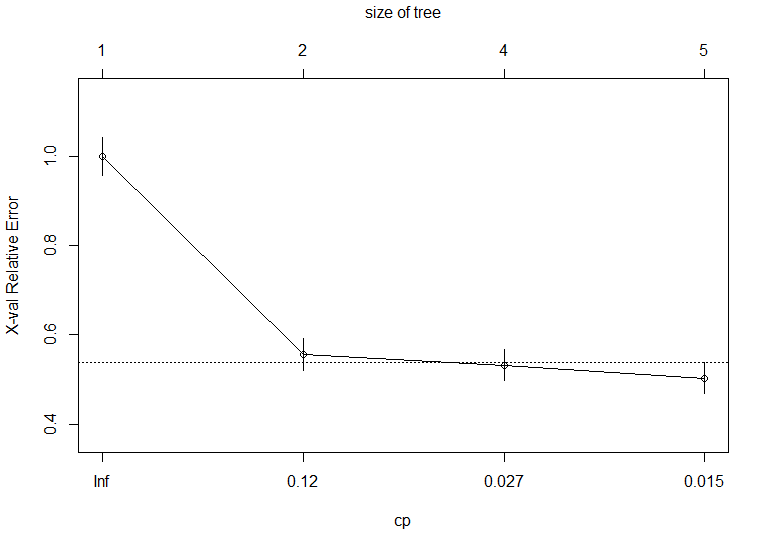
6) Pclass=3 144 72 0 (0.50000000 0.50000000)

12) Fare>=23.35 27 3 0 (0.88888889 0.11111111) \*

13) Fare< 23.35 117 48 1 (0.41025641 0.58974359) \*

7) Pclass=1,2 170 9 1 (0.05294118 0.94705882) \*

# plotcp(rpart.model01) # visualize cross-validation results



summary(rpart.model01) # detailed summary of splits

> summary(rpart.model01) # detailed summary of splits

Call:

rpart(formula = Survived ~ Pclass + Sex + Age + Embarked + Parch +

Fare, data = Titanic\_train, maxdepth = 3)

n= 891

CP nsplit rel error xerror xstd

1 0.44444444 0 1.0000000 1.0000000 0.04244576

2 0.03070175 1 0.5555556 0.5555556 0.03574957

3 0.02339181 3 0.4941520 0.5321637 0.03518794

4 0.01000000 4 0.4707602 0.5029240 0.03444798

Variable importance

Sex Fare Pclass Parch Age Embarked

54 17 14 7 5 3

Node number 1: 891 observations, complexity param=0.4444444

predicted class=0 expected loss=0.3838384 P(node) =1

class counts: 549 342

probabilities: 0.616 0.384

left son=2 (577 obs) right son=3 (314 obs)

Primary splits:

Sex splits as RL, improve=124.42630, (0 missing)

Pclass splits as RRL, improve= 43.78183, (0 missing)

Fare < 10.48125 to the left, improve= 37.94194, (0 missing)

Embarked splits as RRLL, improve= 12.86541, (0 missing)

Parch splits as LRRRLLL, improve= 11.54522, (0 missing)

Surrogate splits:

Fare < 77.6229 to the left, agree=0.679, adj=0.089, (0 split)

Parch splits as LRRRLRR, agree=0.678, adj=0.086, (0 split)

Embarked splits as RLLL, agree=0.650, adj=0.006, (0 split)

Node number 2: 577 observations, complexity param=0.02339181

predicted class=0 expected loss=0.1889081 P(node) =0.647587

class counts: 468 109

probabilities: 0.811 0.189

left son=4 (553 obs) right son=5 (24 obs)

Primary splits:

Age < 6.5 to the right, improve=10.788930, (124 missing)

Fare < 26.26875 to the left, improve=10.216720, (0 missing)

Pclass splits as RLL, improve=10.019140, (0 missing)

Parch splits as LRRLLL-, improve= 3.946409, (0 missing)

Embarked splits as -RLL, improve= 3.079304, (0 missing)

Node number 3: 314 observations, complexity param=0.03070175

predicted class=1 expected loss=0.2579618 P(node) =0.352413

class counts: 81 233

probabilities: 0.258 0.742

left son=6 (144 obs) right son=7 (170 obs)

Primary splits:

Pclass splits as RRL, improve=31.163130, (0 missing)

Fare < 48.2 to the left, improve=10.114210, (0 missing)

Parch splits as RRRRLLL, improve= 5.140857, (0 missing)

Embarked splits as RRLL, improve= 3.750944, (0 missing)

Age < 12 to the left, improve= 1.891684, (53 missing)

Surrogate splits:

Fare < 25.69795 to the left, agree=0.799, adj=0.562, (0 split)

Embarked splits as RRLR, agree=0.637, adj=0.208, (0 split)

Parch splits as RRLRLLL, agree=0.567, adj=0.056, (0 split)

Age < 18.5 to the left, agree=0.564, adj=0.049, (0 split)

Node number 4: 553 observations

predicted class=0 expected loss=0.1681736 P(node) =0.620651

class counts: 460 93

probabilities: 0.832 0.168

Node number 5: 24 observations

predicted class=1 expected loss=0.3333333 P(node) =0.02693603

class counts: 8 16

probabilities: 0.333 0.667

Node number 6: 144 observations, complexity param=0.03070175

predicted class=0 expected loss=0.5 P(node) =0.1616162

class counts: 72 72

probabilities: 0.500 0.500

left son=12 (27 obs) right son=13 (117 obs)

Primary splits:

Fare < 23.35 to the right, improve=10.051280, (0 missing)

Embarked splits as -RRL, improve= 7.071429, (0 missing)

Parch splits as RRLRLLL, improve= 3.937500, (0 missing)

Age < 38.5 to the right, improve= 3.875163, (42 missing)

Surrogate splits:

Parch splits as RRLRRLL, agree=0.882, adj=0.37, (0 split)

Node number 7: 170 observations

predicted class=1 expected loss=0.05294118 P(node) =0.1907969

class counts: 9 161

probabilities: 0.053 0.947

Node number 12: 27 observations

predicted class=0 expected loss=0.1111111 P(node) =0.03030303

class counts: 24 3

probabilities: 0.889 0.111

Node number 13: 117 observations

predicted class=1 expected loss=0.4102564 P(node) =0.1313131

class counts: 48 69

probabilities: 0.410 0.590

## Predict com tipo 'classe' retorna se sobreviveu ou não.

previsto.com.modelo<-predict(rpart.model01,Titanic\_train,type='class')

matriz.de.confusão<-table(Titanic\_train$Survived, previsto.com.modelo)

matriz.de.confusão

> matriz.de.confusao

previsto.com.modelo

0 1

0 513 36

1 114 228

diagonal <- diag(matriz.de.confusão)

Acc <- sum(diagonal)/sum(matriz.de.confusão)

Acc

> Acc

[1] 0.8316498

## Passando na base de teste

Titanic\_test$Pclass <- factor(Titanic\_test$Pclass)  
previsto.valid<-predict(rpart.model01,Titanic\_test,type='class')  
#nesta base não tem a variável resposta - pode avaliar no kaggle

Aplicação do modelo: A partir do modelo ajustado calcule a probabilidade do passageiro sobreviver ao naufrágio.

Obs: os Ids abaixo estão na base de treino e não na base de teste. Então o modelo de predição foi feito em cima da base de treino para predizer estes passageiros.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PassengerId | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked | Probabilidade de sobreviver |
| 12 | 1 | Bonnell, Miss. Elizabeth | female | 58.00 | 0 | 0 | 113783 | 26.5500 | C103 | S | 0.83642527 |
| 13 | 3 | Saundercock, Mr. William Henry | male | 20.00 | 0 | 0 | A/5. 2151 | 8.0500 |  | S | 0.11243576 |
| 14 | 3 | Andersson, Mr. Anders Johan | male | 39.00 | 1 | 5 | 347082 | 31.2750 |  | S | 0.05903258 |
| 15 | 3 | Vestrom, Miss. Hulda Amanda Adolfina | female | 14.00 | 0 | 0 | 350406 | 7.8542 |  | S | 0.66342474 |
| 16 | 2 | Hewlett, Mrs. (Mary D Kingcome) | female | 55.00 | 0 | 0 | 248706 | 16.0000 |  | S | 0.60659549 |
| 17 | 3 | Rice, Master. Eugene | male | 2.00 | 4 | 1 | 382652 | 29.1250 |  | Q | 0.19775834 |
| 18 | 2 | Williams, Mr. Charles Eugene | male | #NULO! | 0 | 0 | 244373 | 13.0000 |  | S | NA  (não tem a informação da idade) |

**Parte 2)**

**Utilize o texto 1 para completar os exercícios a seguir.**

1. **Preencher o quadro conceitual estatístico abaixo:**

Estudo Estatístico – Quadro Conceitual

|  |  |
| --- | --- |
| COMPONENTES | DESCRIÇÃO |
| 1. Tema | Análise de crédito |
| 1. Problema | Melhor modelor que classifique corretamente a minha resposta (bom ou mal pagador) |
| 1. Hipóteses conceituais | Existe um modelo matemático que tenha melhor desempenho que o modelo aual |
| 1. Objetivo Principal | Identificar o melhor modelo que classifique corretamente o cliente em mau ou bom pagador |
| 1. População de Estudo | Todos os clients ativos do banco |
| 1. Plano Básico de Análise   Usar grupo de comparação???  Amostra ou universo???  Identificar os tipos de variáveis | Amostra de 20.000 clientes, divididas em:  8.000 para treinamento 6.000 para validação  6.000 para teste  Levantar as variáveis para o modelo  Variável resposta: 0 (mau) ou 1 (bom) |
| 1. Técnica Estatística | Competição de modelos:  Regressão logística  Redes neurais  Árvore de decisão |
| 1. Resultados Estatístico Principal | Equação matemática |

1. **Use o resultado da Regressão Logística Múltipla apresentado na Tabela 3 da página 112 do artigo para calcular o escore das pessoas a seguir:**

|  |  |  |
| --- | --- | --- |
| **ID** | **Características** | **Escore de crédito** |
| **1** | **Mulher, casada, primeira faixa de tempo de emprego, primeira faixa de valor da parcela, tipo de crédito carnê, primeira faixa de idade, primeira categoria de cep, primeira categoria de profissão, primeira faixa de empréstimo e primeira aquisição de empréstimo.** | 0,5868  + 0 + 0 - 0,4848  + 0,9633  - 1,34 - 0,7429  - 0,3549  + 0,3033  + 0,2481  - 0,6513  **Score = 1,4724** |
| **2** | **Homem, casado, primeira faixa de tempo de emprego, primeira faixa de valor da parcela, tipo de crédito carnê, primeira faixa de idade, primeira categoria de cep, primeira categoria de profissão, primeira faixa de empréstimo e primeira aquisição de empréstimo.** | 0,5868  + 0 - 0,314  - 0,4848  + 0,9633  -1,34  - 0,7429  - 0,3549  + 0,3033  + 0,2481  - 0,6513  **Score = -1,7864** |
| **3** | **Homem, solteiro, primeira faixa de tempo de emprego, segunda faixa de tempo de residência, primeira faixa de valor da parcela, tipo de crédito carnê, primeira faixa de idade, primeira categoria de cep e primeira aquisição de empréstimo** | 0,5868  - 0,1707  - 0,4848  - 0,3363  + 0,9633  - 1,34  -0,7429  - 0,3549  - 0,6513  **Score = -2,5308** |

1. **Utilize os dados da página 116 para calcular o acerto do modelo da Regressão Logística na amostra de treinamento. Preencha o quadro a seguir e responda**

|  |  |  |  |
| --- | --- | --- | --- |
| Observado | Predito | | Total |
| Mau | Bom |
| Mau | 2913 | 1087 | 4000 |
| Bom | 1184 | 2816 | 4000 |
| Total | 4097 | 3903 | 8000 |

1. **Qual o percentual de acerto da resposta Mau pagador?**2913 / 4000 = **72,8 %**
2. **Qual o percentual de acerto da resposta Bom pagador?**2816 / 4000 = **70,4 %**
3. **Qual o percentual de acerto médio do modelo?**(2913 + 2816) / 8000 = **71,6 %**
4. **Utilize os dados da página 116 para calcular o acerto do modelo de Redes Neurais na amostra de treinamento. Preencha o quadro a seguir e responda**

|  |  |  |  |
| --- | --- | --- | --- |
| Observado | Predito | | Total |
| Mau | Bom |
| Mau | 3000 | 1000 | 4000 |
| Bom | 1280 | 2720 | 4000 |
| Total | 4280 | 3720 | 8000 |

1. **Qual modelo você escolheria para implementar na empresa?**

Percentual de acerto Mau = 3000 / 4000 = **75,0 %**

Percentual de acerto Bom = 2720 / 4000 = **68,0 %**

Percentual de acerto médio = (3000 + 2720) / 8000 = **71,5 %**

**O modelo de rede neural teve percentual maior de acerto para “bom”, enquanto o de regressão logística teve melhor desempenho para “mau”. Porém os resultados totais estão bem próximos, com variação de apenas 1 décimo percentual. Ambos os modelos estão com taxas boas de acertos, qualquer um dos modelos poderia ser implementado na empresa.**

**Data de entrega: 31/03/2020**

**Formato: Word**

**Regina Bernal**

**17/03/2020**