Algoritmos de classificação: Indução de Árvores de Decisão

Autor: Luciano Moraes Da Luz Brum

Sumário

- **Introdução**
- **Objetivo**
- **Metodologia**
- **Passo a passo dos resultados**
- **Conclusões**

Introdução

- Para a aplicação de algoritmos de classificação, foi recebida uma base de dados dos anos de 2014, 2015 e 2016 sobre dados de produtores de leite do município de Derrubadas – RS.

- Os dados desta planilha eram numéricos, com 167 amostras e 37 atributos.

Introdução

Município: Derrubadas - RS

		Produtor		ID-UF	S1	S2	S3	S4	S5	S6	E1	E2	E3	E4	E5	E6	E7	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
4	1	Produtor 1	2014	0,62	0,50	0,67	0,60	1,00	1,00	1,00	0,34	0,00	0,36	0,10	0,60	0,51	1,00	1,00	1,00	0,00	1,00	0,50	0,00	0,67	1,00	0,50	1,00	1,00
4	1	Produtor 1	2015	0,54	0,82	0,67	0,60	0,00	0,00	1,00	0,49	0,00	0,36	0,10	0,80	0,29	1,00	1,00	1,00	0,00	1,00	0,50	0,00	0,67	1,00	0,50	0,00	1,00
4	1	Produtor 1	2016	0,54	0,82	0,67	0,60	0,00	0,00	1,00	0,49	0,00	0,36	0,10	0,80	0,29	1,00	1,00	1,00	0,00	1,00	0,50	0,00	0,67	1,00	0,50	0,00	1,00
47	2	Produtor 2	2014	0,64	0,80	0,67	0,60	1,00	0,00	1,00	0,37	0,00	0,52	0,14	0,80	0,29	1,00	1,00	1,00	0,00	0,00	0,50	0,00	1,00	0,90	0,50	1,00	1,00
47	2	Produtor 2	2015	0,60	0,50	0,67	0,60	1,00	0,00	1,00	0,35	0,00	0,40	0,14	0,85	0,26	0,93	1,00	1,00	0,00	0,00	0,50	0,00	1,00	0,90	0,50	1,00	1,00
47	2	Produtor 2	2016	0,60	0,50	0,67	0,60	1,00	0,00	1,00	0,35	0,00	0,40	0,14	0,85	0,26	0,93	1,00	1,00	0,00	0,00	0,50	0,00	1,00	0,90	0,50	1,00	1,00
37	3	Produtor 3	2014	0,45	0,25	0,67	0,60	1,00	1,00	1,00	0,18	0,00	0,31	0,18	0,80	1,00	0,10	1,00	1,00	0,00	1,00	0,50	0,00	0,67	1,00	0,50	0,00	0,00
15	4	Produtor 4	2015	0,63	0,83	0,67	0,60	1,00	0,00	1,00	0,52	0,00	0,18	0,10	0,85	0,50	1,00	1,00	1,00	0,00	1,00	1,00	0,50	0,67	0,90	0,50	1,00	1,00
15	4	Produtor 4	2016	0,60	0,80	0,67	0,60	1,00	0,00	1,00	0,31	0,00	0,18	0,10	0,85	0,00	1,00	1,00	1,00	0,00	1,00	1,00	0,50	0,67	0,90	0,50	1,00	1,00
15	4	Produtor 4	2014	0,59	0,50	0,67	0,60	1,00	0,00	1,00	0,52	0,00	0,18	0,18	0,60	0,50	1,00	1,00	1,00	0,00	1,00	1,00	0,50	0,47	0,90	0,50	0,50	1,00
52	5	Produtor 5	2015	0,51	0,00	0,67	0,60	1,00	0,00	1,00	0,69	0,00	0,22	0,07	0,60	0,17	0,83	0,60	1,00	0,00	0,00	0,50	0,50	0,80	0,65	0,50	1,00	0,00
52	5	Produtor 5	2016	0,51	0,00	0,67	0,60	1,00	0,00	1,00	0,71	0,00	0,18	0,07	0,60	0,19	0,84	0,60	1,00	0,00	0,00	0,50	0,50	0,80	0,65	0,50	1,00	0,00
52	5	Produtor 5	2014	0,45	0,00	0,67	0,60	1,00	1,00	1,00	0,12	0,00	0,22	0,07	0,60	1,00	0,00	0,60	1,00	0,00	0,00	0,50	0,50	0,80	0,65	0,50	1,00	0,00
21	6	Produtor 6	2016	0,60	0,25	0,67	0,60	1,00	0,00	0,00	0,60	0,00	0,09	0,17	0,85	0,29	1,00	1,00	1,00	0,00	0,00	1,00	0,50	0,73	0,90	0,50	1,00	1,00
21	6	Produtor 6	2015	0,59	0,25	0,67	0,60	1,00	0,00	0,00	0,55	0,00	0,09	0,09	0,85	0,21	1,00	1,00	1,00	0,00	0,00	1,00	0,50	0,73	0,90	0,50	1,00	1,00
21	6	Produtor 6	2014	0,55	0,25	0,67	0,60	1,00	0,00	0,00	0,41	0,00	0,09	0,09	0,85	0,18	0,81	1,00	1,00	0,00	0,00	0,50	0,50	0,40	0,90	0,50	1,00	1,00
3	7	Produtor 7	2015	0,55	0,00	0,33	0,90	0,00	0,00	1,00	0,28	1,00	0,60	0,18	0,80	0,30	1,00	1,00	0,00	0,00	1,00	1,00	0,50	0,53	0,90	1,00	1,00	1,00
3	7	Produtor 7	2016	0,55	0,00	0,33	0,90	0,00	0,00	1,00	0,28	1,00	0,60	0,18	0,80	0,30	1,00	1,00	0,00	0,00	1,00	1,00	0,50	0,53	0,90	1,00	1,00	1,00

Figura 1: planilha com dados do município de Derrubadas – RS. Fonte: Elaborada pelo autor, 2017

Objetivo

- Aplicar algoritmos de classificação para efetuar a predição do indicador ID.UPF do ano de 2014, com base em indicadores de dimensões:
  - Sociais;
  - Econômicas;
  - Ambientais;
  - Produtivas;

Metodologia Adotada

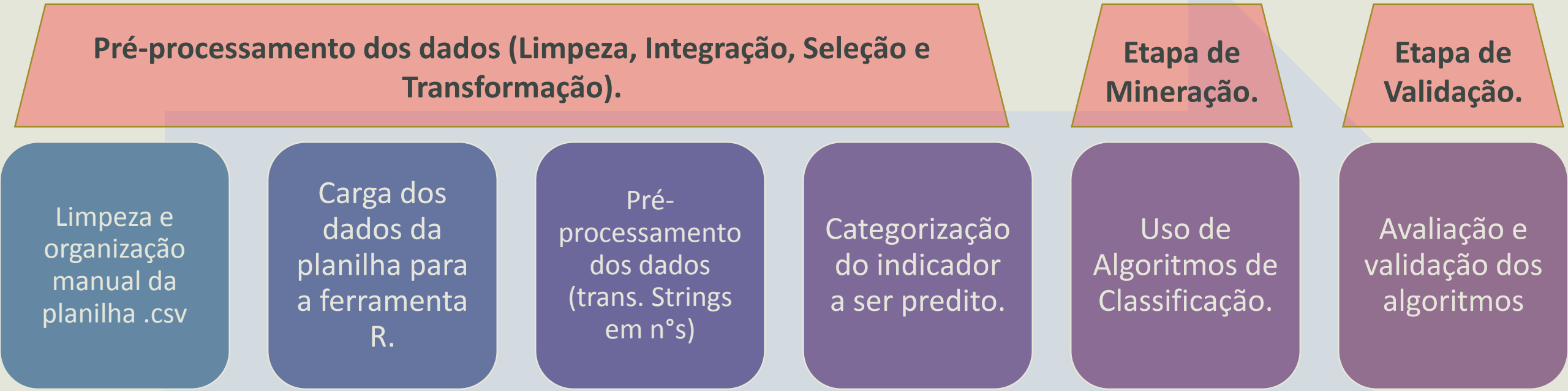


Figura 2: Metodologia adotada. Fonte: Elaborada pelo autor, 2017

Passo a Passo

- Foi utilizada a ferramenta R para o processo de limpeza, categorização e mineração de dados (regras de associação).

- Das instruções utilizadas no R:
  - library(rpart)
  - library(caret)
  - library(rpart.plot)
  - install.packages("C50")
  - library(C50)
  - library(nnet)
  - x = read.csv("training.csv", stringsAsFactors=FALSE)
  - x2=x

Passo a Passo

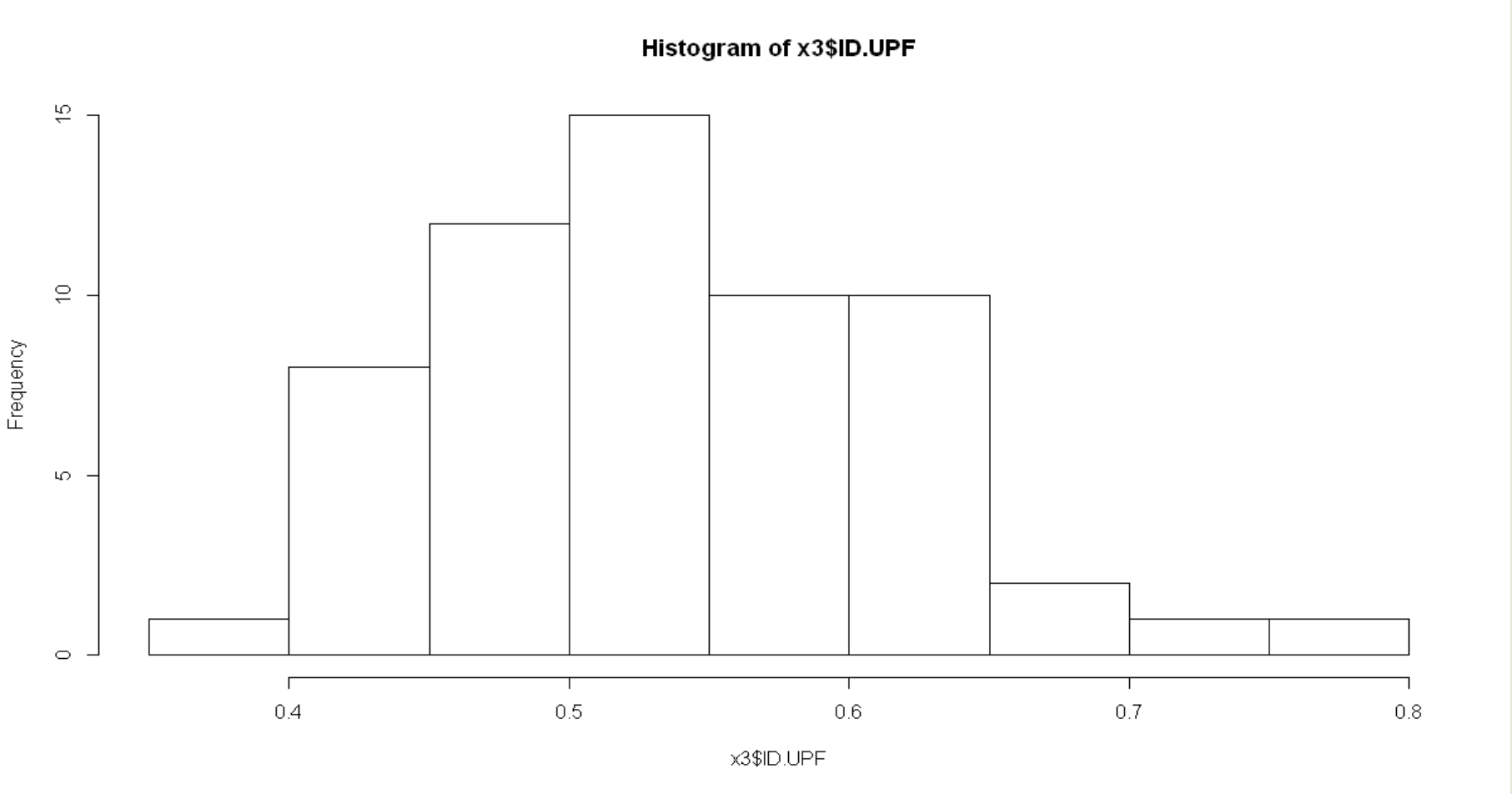


Figura 3: Histograma do indicador analisado Fonte: Elaborada pelo autor, 2017

Passo a Passo

- for(i in 1:nrow(x2)){
- for(j in 1:ncol(x2)){
- x2[i,j] = as.numeric(gsub("[,]",".",x[i,j]))
- x2[,j] = as.numeric(x2[,j])}}
- x2$ID.UPF<-cut(x2$ID.UPF, c(0,0.49,0.59,0.80),labels=c("BAIXO","MEDIO","ALTO"))

- Para o algoritmo da biblioteca rpart:
- x3=x2
- fit3 <- caret::train(x3[,-1], x3$ID.UPF,trControl=trainControl(method="repeatedcv",number=10,repeats=3),method=“rpart")
- fit3pred <- predict(fit3,newdata=x3[,-1])
- mc3<-table(fit3pred,x3$ID.UPF)
- ac3 <-(mc3[1,1]+mc3[2,2]+mc3[3,3])/sum(mc3)

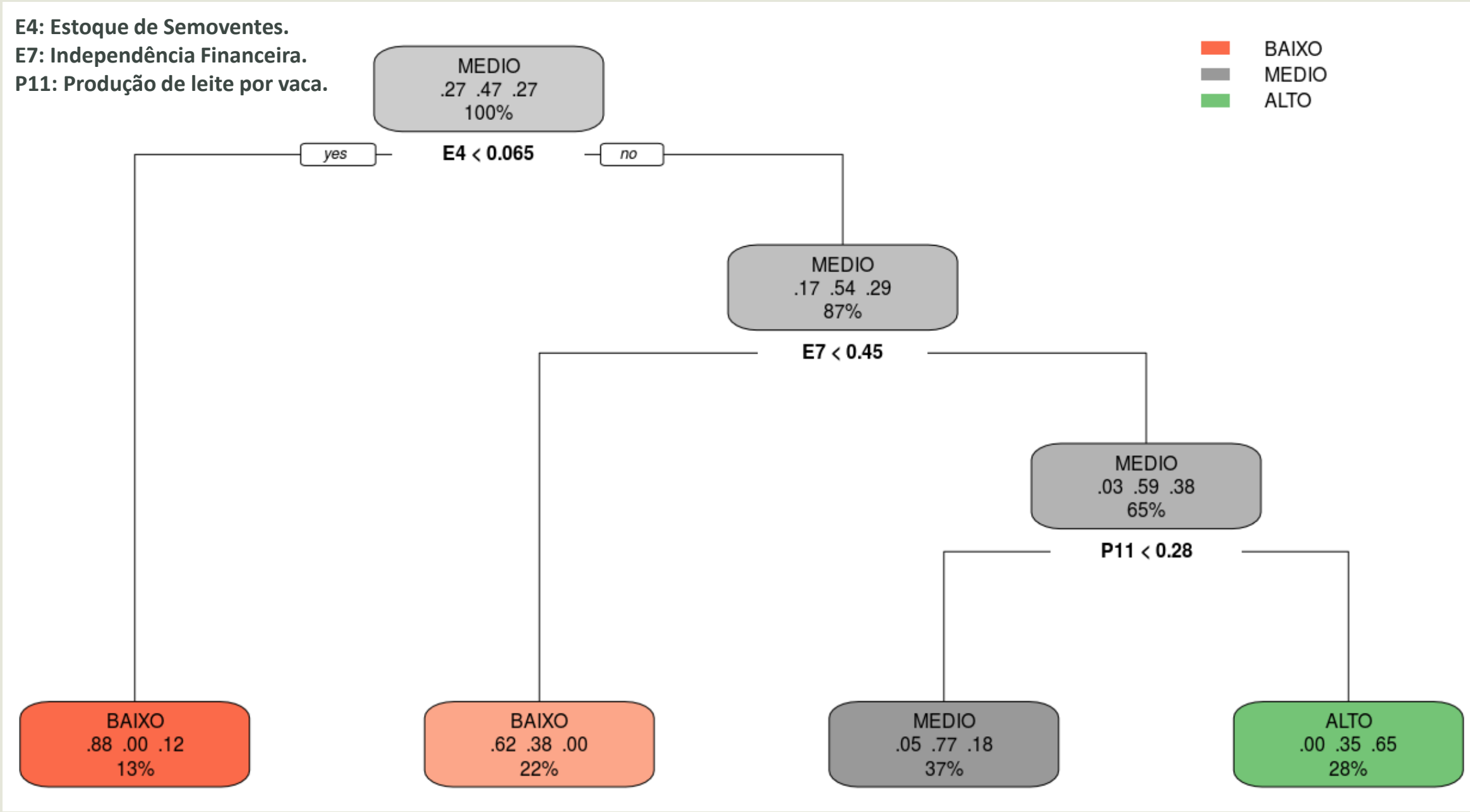


Figura 4: Árvore de decisão gerada usando o *rpart*. Fonte: Elaborado pelo autor, 2017.

Passo a Passo

Para o algoritmo da biblioteca C50:

- x3=x2
- fit3 <- caret::train(x3[,-1], x3$ID.UPF,trControl=trainControl(method="repeatedcv",number=10,repeats=3),method="C5.0")
- fit3pred <- predict(fit3,newdata=x3[,-1])
- mc3<-table(fit3pred,x3$ID.UPF)
- ac3 <-(mc3[1,1]+mc3[2,2]+mc3[3,3])/sum(mc3)

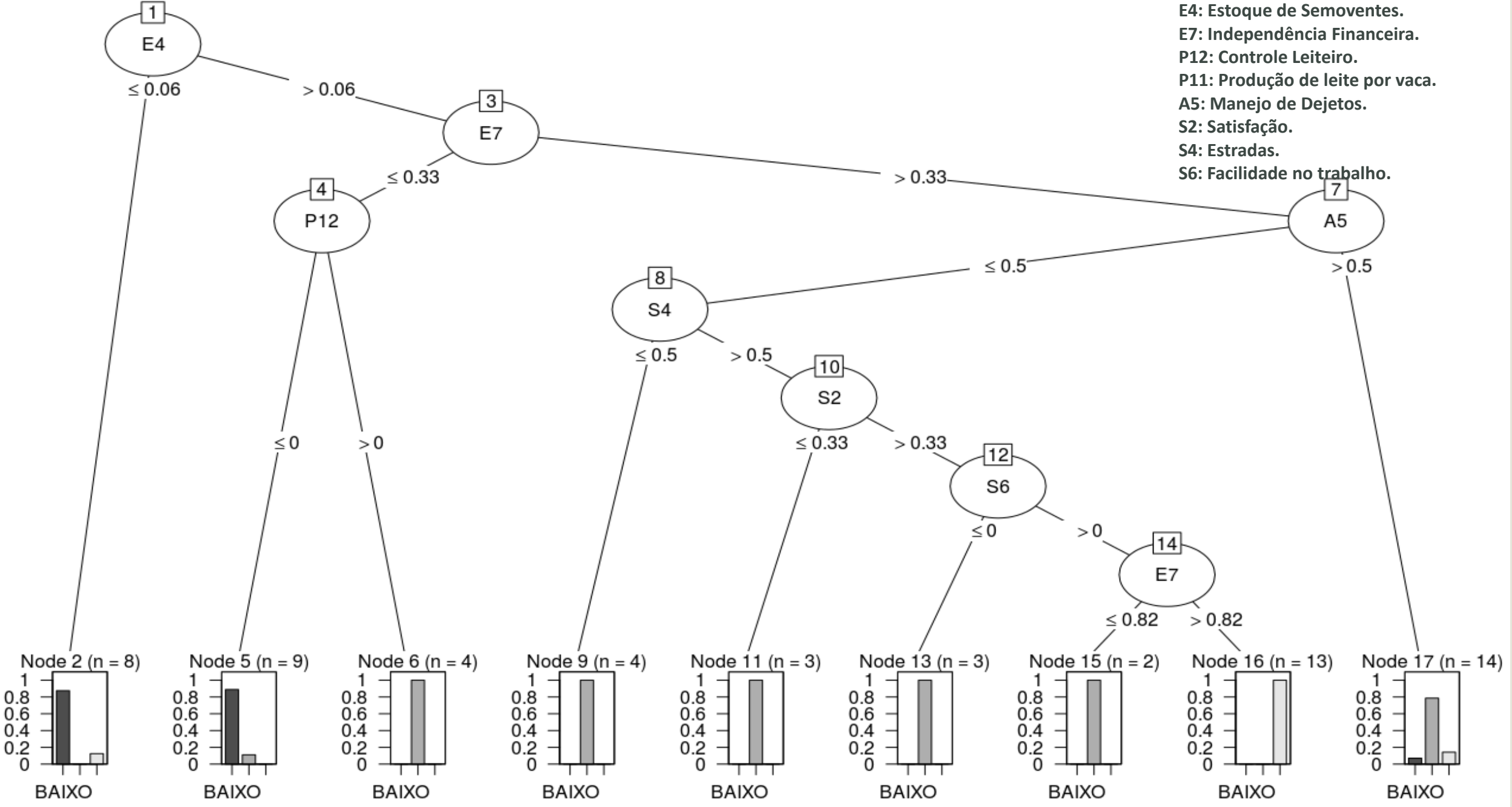


Figura 5: Árvore de decisão gerada usando o *C5.0*. Fonte: Elaborado pelo autor, 2017.

Passo a Passo

- Precisões obtidas com os algoritmos, efetuando o *10-fold cross-validation*:
  - Para o rpart: 71,66%
  - Para o C5.0: 91,66%

- Porém, essa não é a acurácia real. É preciso utilizar um método de avaliação e validação.
- Foi feito o *10-fold cross-validation* usando estruturas de repetição para avaliação dos modelos.

Passo a Passo

Mudanças:

- set.seed(12345)
- x3<-x2[sample(nrow(x2)),]
- folds <- cvFolds(NROW(x3), K=10)
- x3$holdoutpred <- rep(0,nrow(x3))
- for(i in 1:10){
  - trainData <- x3[folds$subsets[folds$which != i],]
  - testData <- x3[folds$subsets[folds$which == i],]
  - fit <- rpart(ID.UPF ~ ., method="class", data=trainData)
  - fit <- caret::train(trainData[,-1], trainData$ID.UPF,trControl=trainControl(method="cv",number=10),method="rpart")
  - newpred <- predict(fit,newdata=testData)
  - mc<-table(newpred,testData$ID.UPF) #monta tabela de confusão
  - ac3[i] <- (mc[1,1]+mc[2,2]+mc[3,3])/sum(mc)
  - x3[folds$subsets[folds$which == i],]$holdoutpred <- newpred}
- ac3

Passo a Passo

- Precisões obtidas com os algoritmos, efetuando o *10-fold Cross-Validation* :
  - Para o rpart: 40,00%
  - Para o C5.0: 64,12%
- Precisões obtidas com os algoritmos, efetuando o *LOOCV (Leave-One-Out Cross-Validation)*:
  - Para o rpart: 35,00%
  - Para o C5.0: 61,00%

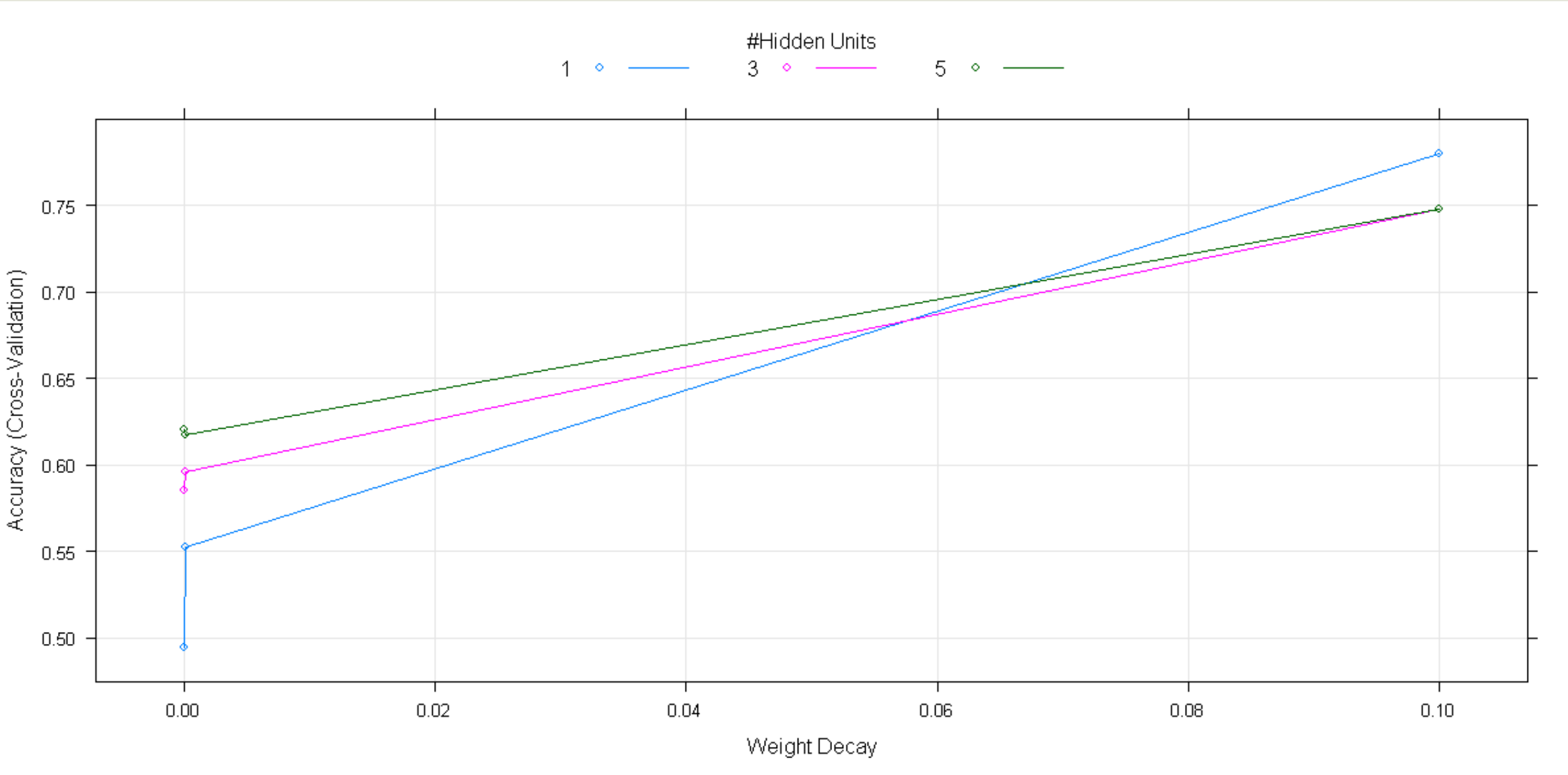


Figura 6: Acurácia em função do decaimento do peso. Fonte: Elaborado pelo autor, 2017.

Passo a Passo

- Precisão obtida com o nnet, efetuando o *10-fold cross-validation*:
  - 72,57%
  - size decay Accuracy Kappa
  - 1 0e+00 0.4916667 0.2175280
  - 1 1e-04 0.5392857 0.3232178
  - 1 1e-01 0.7257143 0.5355373
  - 3 0e+00 0.6295238 0.3975643
  - 3 1e-04 0.5535714 0.3080018
  - 3 1e-01 0.7023810 0.5112600
  - 5 0e+00 0.7114286 0.5194526
  - 5 1e-04 0.6307143 0.4294253
  - 5 1e-01 0.7023810 0.5127226

Passo a Passo

- Foi iniciada a análise do indicador “Renda Familiar Per Capita”, devido a sua variabilidade.

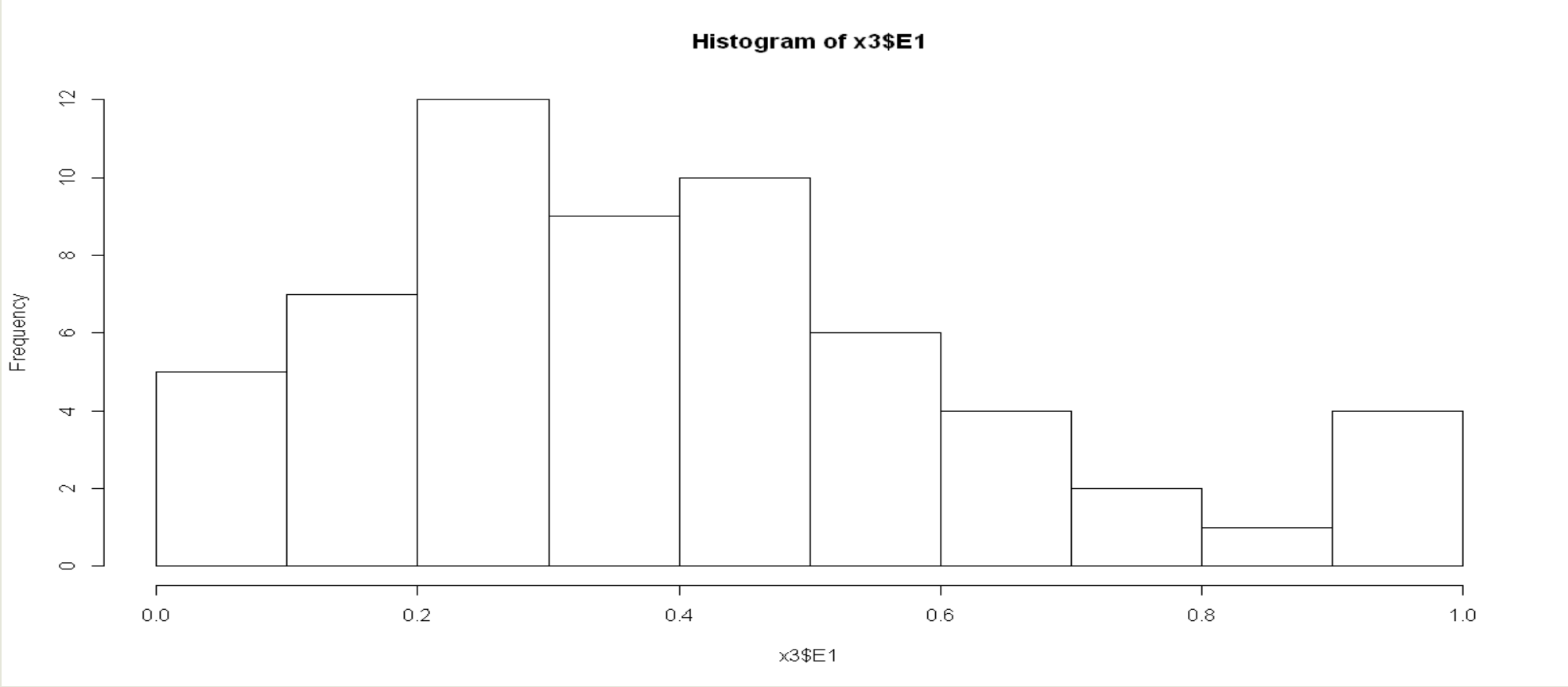


Figura 7: Histograma da Renda Familiar Per Capita. Fonte: Elaborado pelo autor, 2017.

Passo a Passo

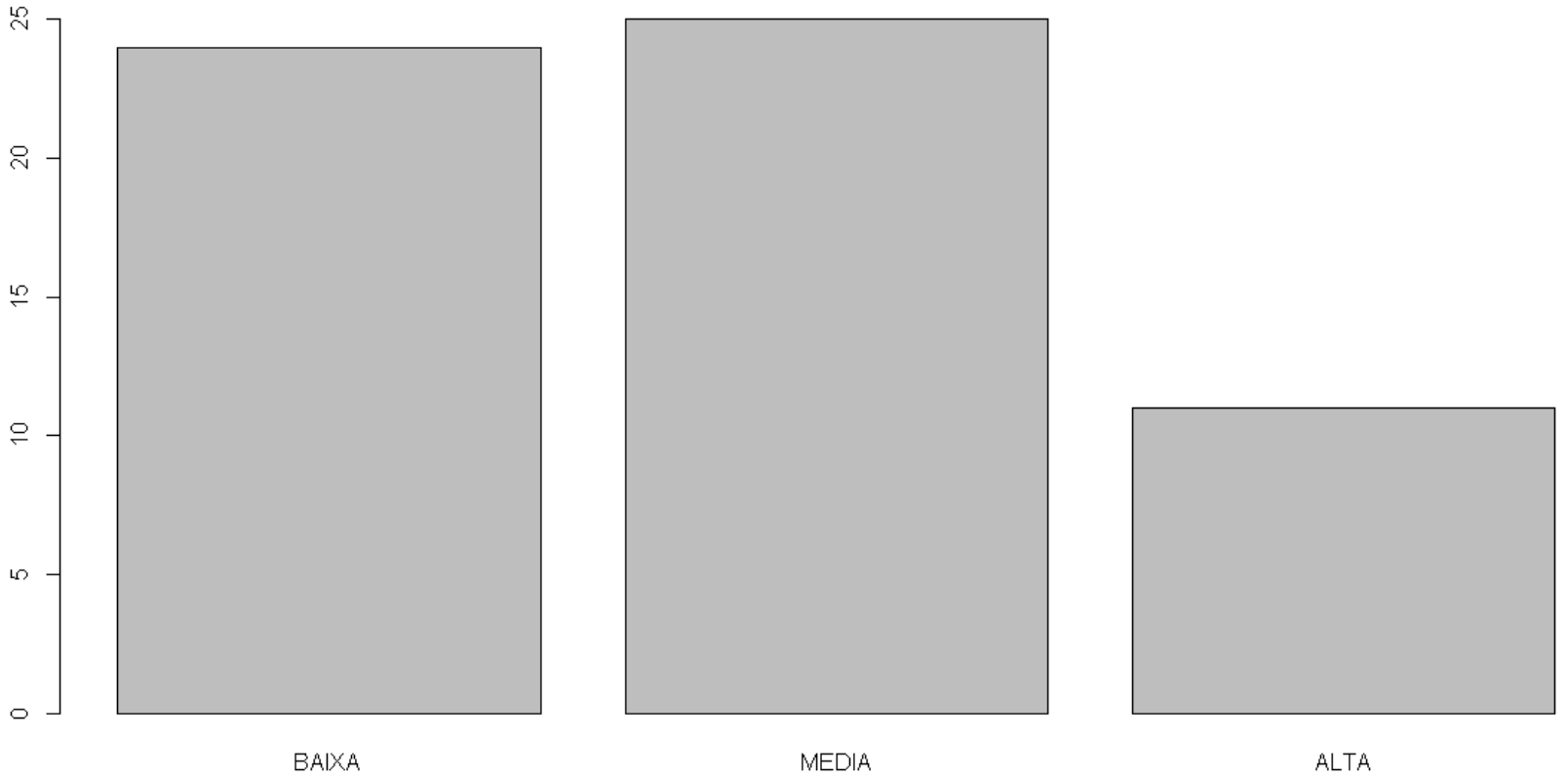


Figura 8: Histograma da Renda Familiar Per Capita após categorização. Fonte: Elaborado pelo autor, 2017.

Passo a Passo

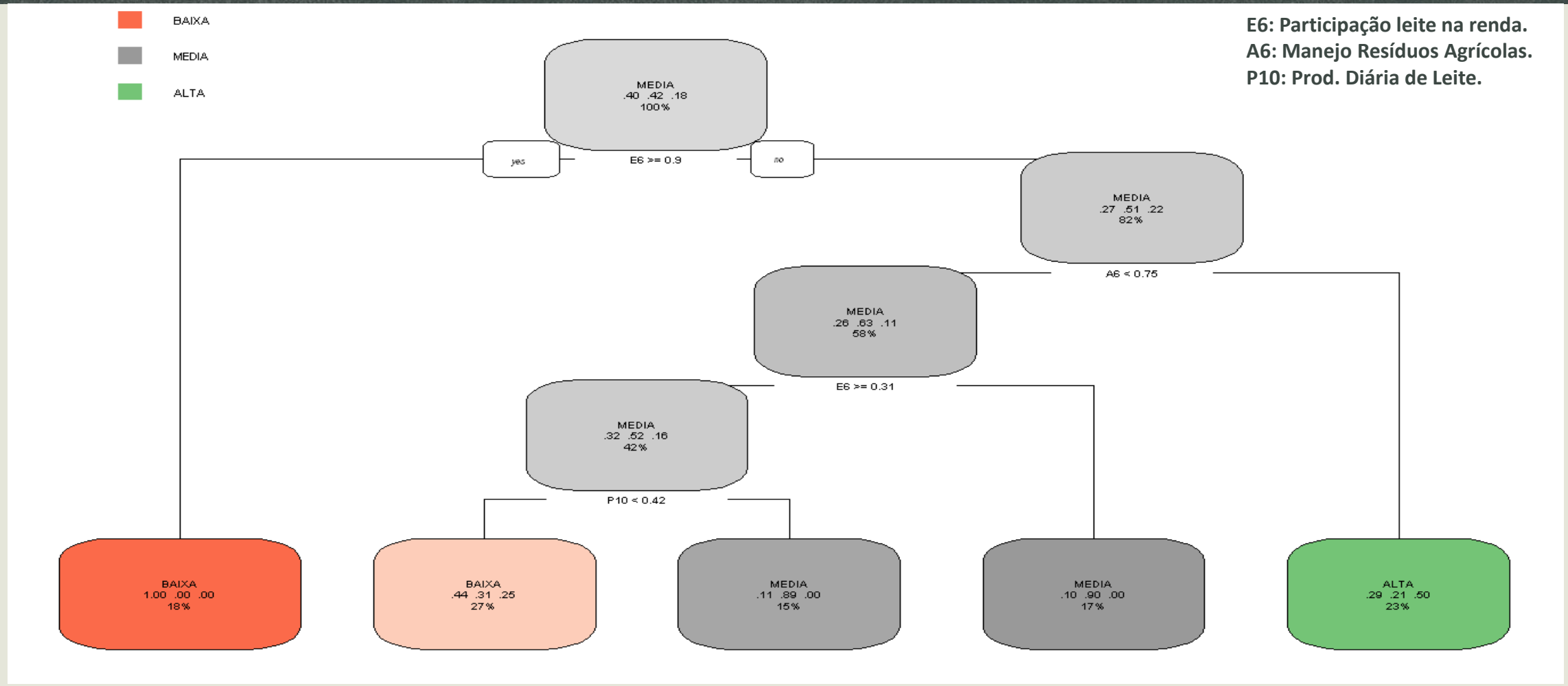


Figura 9: Árvore gerada pelo Rpart. Fonte: Elaborado pelo autor, 2017.

Passo a Passo

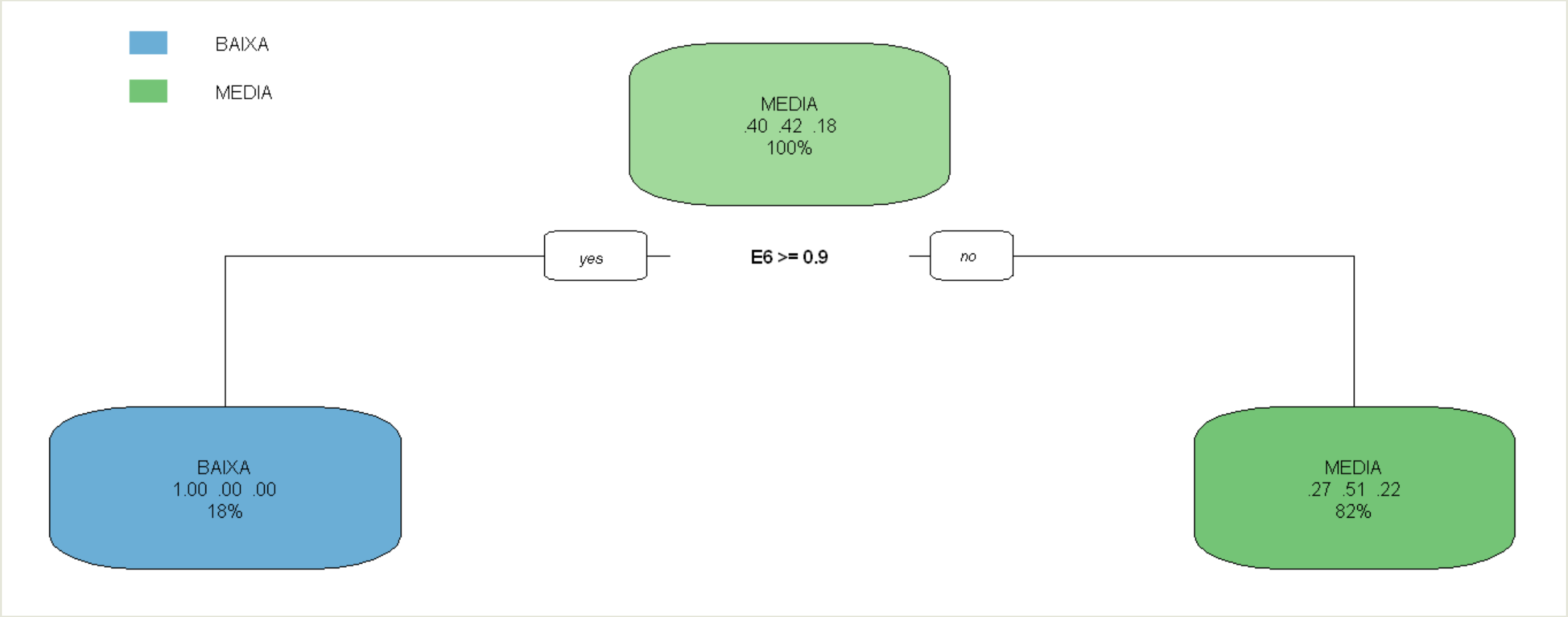
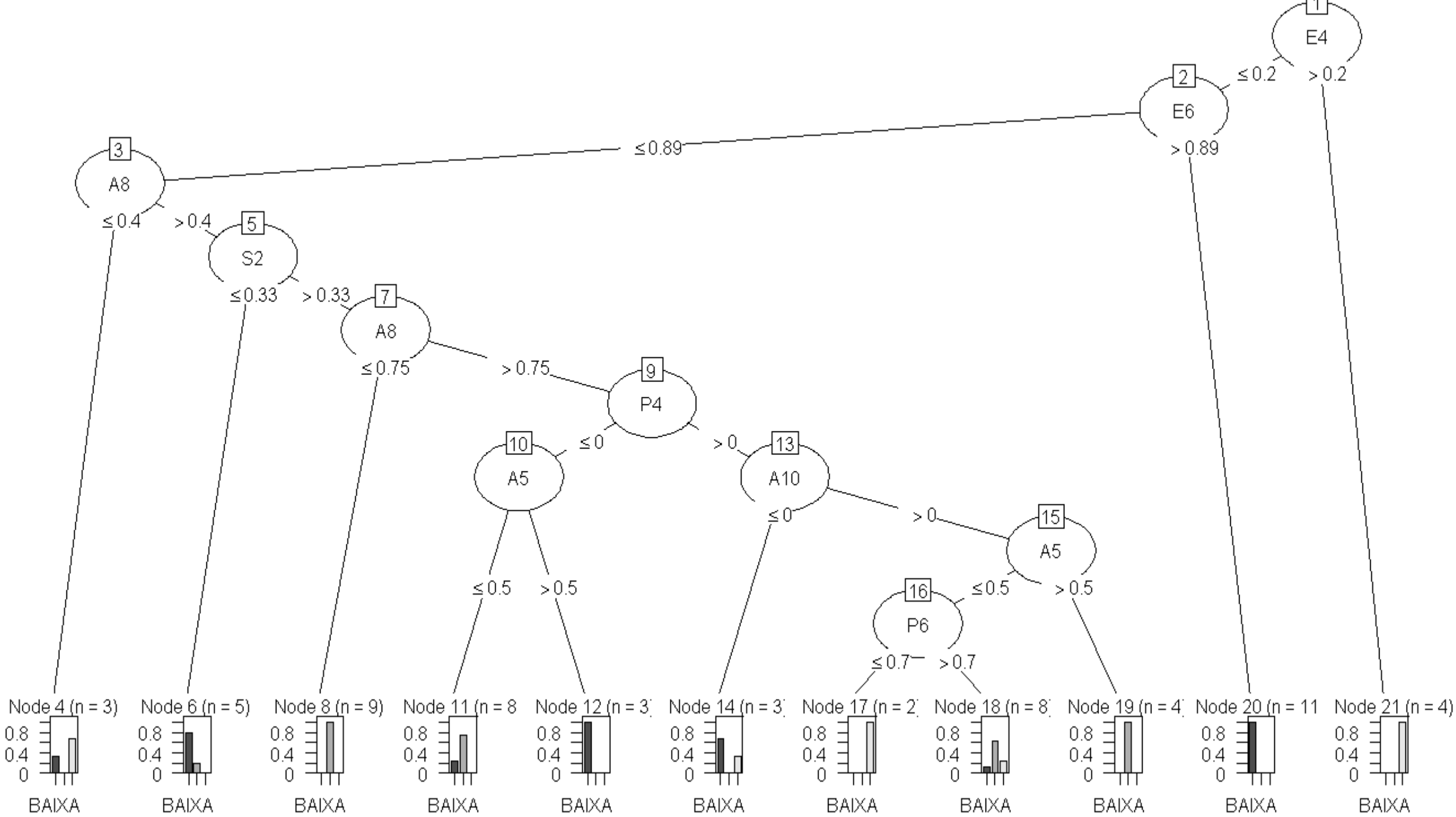


Figura 10: Árvore pós-poda gerada pelo Rpart. Fonte: Elaborado pelo autor, 2017.


1
E4
≤ 0.2
> 0.2
2
E6
> 0.89
≤ 0.89
3
A8
≤ 0.4
> 0.4
5
S2
≤ 0.33
> 0.33
7
A8
≤ 0.75
> 0.75
9
P4
≤ 0
> 0
10
A5
≤ 0.5
> 0.5
13
A10
≤ 0
> 0
15
A5
≤ 0.5
> 0.5
16
P6
≤ 0.7
> 0.7
Node 4 (n = 3)
Node 6 (n = 5)
Node 8 (n = 9)
Node 11 (n = 8
Node 12 (n = 3
Node 14 (n = 3
Node 17 (n = 2
Node 18 (n = 8
Node 19 (n = 4
Node 20 (n = 11
Node 21 (n = 4)
0.8
0.4
0
BAIXA

Conclusões

- A aplicação de algoritmos de classificação, especificamente árvores de decisão, para efetuar a predição de informações categóricas se mostrou como uma técnica efetiva.

- No mundo dos negócios, pode ser uma importante ferramenta para ser utilizada na orientação de decisões.

Referências Bibliográficas

- HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. 2° ed. Morgan Kauf-mann Publishers, p. 5–7, 2006.

- RDocumentation. Disponível em: <https://www.rdocumentation.org/packages/caret/versions/6.0-76/topics/train>. Acesso em: 21/05/2017.