ejercicio 6

Javier

16/11/2020

```
library(C50)

## Warning: package 'C50' was built under R version 4.0.3

library(gmodels)

## Warning: package 'gmodels' was built under R version 4.0.3

library(lattice)
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.3

library(rpart)
library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 4.0.3

library(mlbench)

## Warning: package 'mlbench' was built under R version 4.0.3

library(ROCR)

## Warning: package 'ROCR' was built under R version 4.0.3

library(e1071)

## Warning: package 'e1071' was built under R version 4.0.3

library(caret)

## Warning: package 'caret' was built under R version 4.0.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```
# Extraemos la información y limpiamos
data <- read.csv2("german_credit.csv")
colSums(is.na(data))
```

```
##     checking_balance months_loan_duration       credit_history 
##                    0                    0                    0 
##              purpose               amount      savings_balance 
##                    0                    0                    0 
##    employment_length     installment_rate      personal_status 
##                    0                    0                    0 
##        other_debtors    residence_history             property 
##                    0                    0                    0 
##                  age     installment_plan              housing 
##                    0                    0                    0 
##     existing_credits              default 
##                    0                    0
```

```
str(data)
```

```
## 'data.frame':    1000 obs. of  17 variables:
##  $ checking_balance    : chr  "< 0 DM" "1 - 200 DM" "unknown" "< 0 DM"
...
##  $ months_loan_duration: int  6 48 12 42 24 36 24 36 12 30 ...
##  $ credit_history      : chr  "critical" "repaid" "critical" "repaid"
...
##  $ purpose             : chr  "radio/tv" "radio/tv" "education"
"furniture" ...
##  $ amount              : int  1169 5951 2096 7882 4870 9055 2835 6948
3059 5234 ...
##  $ savings_balance     : chr  "unknown" "< 100 DM" "< 100 DM" "< 100
DM" ...
##  $ employment_length   : chr  "> 7 yrs" "1 - 4 yrs" "4 - 7 yrs" "4 - 7
yrs" ...
##  $ installment_rate    : int  4 2 2 2 3 2 3 2 2 4 ...
##  $ personal_status     : chr  "single male" "female" "single male"
"single male" ...
##  $ other_debtors       : chr  "none" "none" "none" "guarantor" ...
##  $ residence_history   : int  4 2 3 4 4 4 4 2 4 2 ...
##  $ property            : chr  "real estate" "real estate" "real
estate" "building society savings" ...
##  $ age                 : int  67 22 49 45 53 35 53 35 61 28 ...
##  $ installment_plan    : chr  "none" "none" "none" "none" ...
##  $ housing             : chr  "own" "own" "own" "for free" ...
##  $ existing_credits    : int  2 1 1 1 2 1 1 1 1 2 ...
##  $ default             : int  1 2 1 1 2 1 1 1 1 2 ...
```

```
# Generamos una semilla aleatoria y seleccionamos un sample
set.seed(123)
# Dataset formado por 1000 observaciones y 17 variables
train_sample <- sample(1000,800)

str(train_sample)

##  int [1:800] 415 463 179 526 195 938 818 118 299 229 ...

# Preparamos Train y Test
train <- data[train_sample,]
train$default <- as.factor(train$default)
test <- data[-train_sample,]

prop.table(table(train$default))

##
##      1      2
## 0.7125 0.2875

prop.table(table(test$default))

##
##    1    2
## 0.65 0.35

# Usamos el algoritmos C5.0 para el modelo
model <- C5.0(x=train[-17],train$default)
model

##
## Call:
## C5.0.default(x = train[-17], y = train$default)
##
## Classification Tree
## Number of samples: 800
## Number of predictors: 16
##
## Tree size: 37
##
## Non-standard options: attempt to group attributes

pred <- predict(model,test)
CrossTable(test$default,pred,
           prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
           dnn=c('Actual','Predicción'))

##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  200 
## 
## 
##              | Predicción 
##       Actual |         1 |         2 | Row Total | 
## -------------|-----------|-----------|-----------|
##            1 |       118 |        12 |       130 | 
##              |     0.590 |     0.060 |           | 
## -------------|-----------|-----------|-----------|
##            2 |        41 |        29 |        70 | 
##              |     0.205 |     0.145 |           | 
## -------------|-----------|-----------|-----------|
## Column Total |       159 |        41 |       200 | 
## -------------|-----------|-----------|-----------|
## 
## 
```

```r
# Usamos el algoritmos C5.0 para el modelo (añadimos los trials)
model1 <- C5.0(x=train[-17],train$default,trials = 10)
model1
```

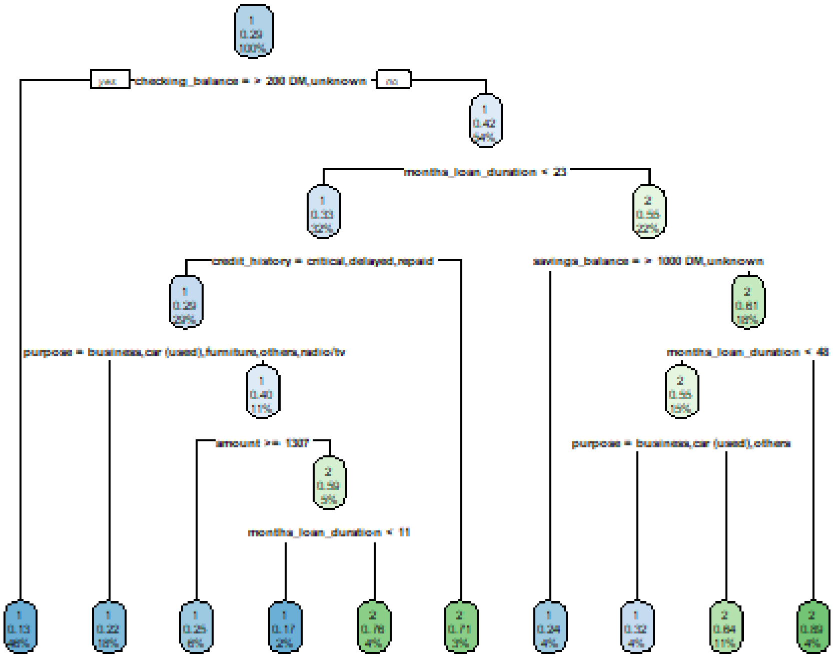
```
## 
## Call:
## C5.0.default(x = train[-17], y = train$default, trials = 10)
## 
## Classification Tree
## Number of samples: 800 
## Number of predictors: 16 
## 
## Number of boosting iterations: 10 
## Average tree size: 26.9 
## 
## Non-standard options: attempt to group attributes
```

```r
pred1 <- predict(model1,test)
CrossTable(test$default,pred1,
           prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
           dnn=c('Actual','Predicción'))
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
## 
```

```
##
## Total Observations in Table:  200
##
##
##              | Predicción
##       Actual |         1 |         2 | Row Total |
## -------------|-----------|-----------|-----------|
##            1 |       113 |        17 |       130 |
##              |     0.565 |     0.085 |           |
## -------------|-----------|-----------|-----------|
##            2 |        39 |        31 |        70 |
##              |     0.195 |     0.155 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       152 |        48 |       200 |
## -------------|-----------|-----------|-----------|
##
##

# Creación del árbol
tree <- rpart(default ~ ., data=train)
rpart.plot(tree)
```


1
0.29
100%
yes
checking_balance = > 200 DM,unknown
no
1
0.42
64%
months_loan_duration < 23
1
0.33
32%
2
0.55
22%
credit_history = critical,delayed,repaid
savings_balance = > 1000 DM,unknown
1
0.29
20%
2
0.61
18%
purpose = business,car (used),furniture,others,radio/tv
months_loan_duration < 48
1
0.40
11%
2
0.55
15%
amount >= 1307
purpose = business,car (used),others
2
0.59
5%
months_loan_duration < 11
1
0.13
36%
1
0.22
18%
1
0.25
6%
1
0.17
2%
2
0.78
4%
2
0.71
3%
1
0.24
4%
1
0.32
4%
2
0.64
11%
2
0.89
4%