MVA Homework 5 - Decision Trees and Association Rules

1 Read the file CREDSCO.TXT

```
setwd("/Users/diego/Documents/UPC-MIRI/semester2/MultiVariate-Analysis/")
set.seed(1111)
credit <- read.delim("assignment5/credsco.txt", header = TRUE, sep="\t")
dim(credit)
#4455   14

#We select a sample at random of 1000 individuals
x <- sample(1:nrow(credit),size=1000,replace=FALSE)
credit_thousand <- credit[x,]
dim(credit_thousand)
[1] 1000   14
```

2 Declare as factor all categorical variables

```
#using categorical levels found in credsco.info
names(credit)
%"Dictamen"             "Ant..Trabajo"              "Vivienda"       "Plazo"
%"Edad"                 "Estado.civil"              "Registros"      "Tipo.trabajo"
%"Gastos"               "Ingresos"                  "Patrimonio"     "Cargas.patrimoniales"
%"Importe.solicitado"   "Precio.del.bien.financiado"

table(credit_thousand$Dictamen)
%   1   2
%   748 252
credit_thousand$Dictamen <- as.factor(credit_thousand$Dictamen)
levels(credit_thousand$Dictamen) <- c("positivo","negativo")
table(credit_thousand$Dictamen)
%   positivo negativo
%   748      252

table(credit_thousand$Vivienda)
%  0   1   2   3   4   5   6
%  1 210 495  45   2 179  68
credit_thousand$Vivienda <- as.factor(credit_thousand$Vivienda)
levels(credit_thousand$Vivienda) <-
        c("NA","alquiler","escritura publica","contrato privado","ignora contrato","padres","otros")
table(credit_thousand$Vivienda)
% NA   alquiler escritura publica  contrato privado   ignora contrato    padres    otros
% 1         210               495                45                 2       179       68

table(credit_thousand$Estado.civil)
%   0   1   2   3   4   5
%   1 214 731  13  38   3
credit_thousand$Estado.civil <- as.factor(credit_thousand$Estado.civil)
levels(credit_thousand$Estado.civil) <- c("NA","soltero","casado","viudo","separado","divorciado")
table(credit_thousand$Estado.civil)
%     NA    soltero     casado      viudo   separado divorciado
%      1        214        731         13         38          3

table(credit_thousand$Registros)
```

```
%   1   2
% 832 168
credit_thousand$Registros <- as.factor(credit_thousand$Registros)
levels(credit_thousand$Registros) <- c("no","si")
table(credit_thousand$Registros)
%   no  si
%  832 168


table(credit_thousand$Tipo.trabajo)
%   1   2   3   4
% 625 105 229  41
credit_thousand$Tipo.trabajo <- as.factor(credit_thousand$Tipo.trabajo)
levels(credit_thousand$Tipo.trabajo) <- c("empleado fijo","empleado temporal","autonomo","otros")
table(credit_thousand$Tipo.trabajo)
% empleado fijo empleado temporal          autonomo             otros
%           625               105               229                41
```

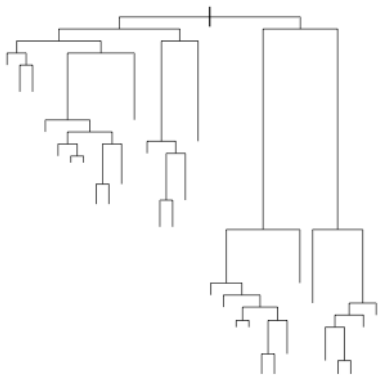
3 Impute the missing values of the continuous variables with the mice function

```
summary(credit_thousand)
%      Dictamen    Ant..Trabajo                 Vivienda       Plazo            Edad
%  positivo:748   Min.   : 0.000   NA                :  1   Min.   : 6.00   Min.   :19.00
%  negativo:252   1st Qu.: 1.000   alquiler          :210   1st Qu.:36.00   1st Qu.:28.75
%                 Median : 5.000   escritura publica:495   Median :48.00   Median :37.00
%                 Mean   : 7.889   contrato privado  : 45   Mean   :46.91   Mean   :37.71
%                 3rd Qu.:12.000   ignora contrato   :  2   3rd Qu.:60.00   3rd Qu.:46.00
%                 Max.   :47.000   padres            :179   Max.   :60.00   Max.   :66.00
%                                  otros             : 68
%      Estado.civil Registros           Tipo.trabajo     Gastos          Ingresos
%  NA        :  1   no:832    empleado fijo    :625   Min.   : 35.00   Min.   :       0
%  soltero   :214   si:168    empleado temporal:105   1st Qu.: 35.00   1st Qu.:      81
%  casado    :731             autonomo         :229   Median : 51.00   Median :     120
%  viudo     : 13             otros            : 41   Mean   : 55.02   Mean   :  700132
%  separado  : 38                                     3rd Qu.: 71.00   3rd Qu.:     170
%  divorciado:  3                                     Max.   :131.00   Max.   :99999999
%
%    Patrimonio       Cargas.patrimoniales Importe.solicitado Precio.del.bien.financiado
%  Min.   :       0   Min.   :       0     Min.   : 107       Min.   : 125
%  1st Qu.:       0   1st Qu.:       0     1st Qu.: 750       1st Qu.:1130
%  Median :    3500   Median :       0     Median :1000       Median :1408
%  Mean   : 1205177   Mean   :  400331     Mean   :1057       Mean   :1471
%  3rd Qu.:    6500   3rd Qu.:       0     3rd Qu.:1302       3rd Qu.:1704
%  Max.   :99999999   Max.   :99999999     Max.   :5000       Max.   :6500

library(mice)
# We see Ant..Trabajo has 0 values but the unit of measurement is years and its reasonable
# to assume people have worked 0 years on a job, especially since the value is numerable.
length(which(credit_thousand$Ant..Trabajo == 0))
# 123

#In Ingresos (column 10) we see values of 0 and values of 99999999 which are both clearly wrong
credit_thousand[which(credit_thousand$Ingresos == 0),10] <- NA
credit_thousand[which(credit_thousand$Ingresos == 99999999),10] <- NA
length(which(is.na(credit_thousand$Ingresos)))
#76 missing values now!
```

```
# this seems feasible because from the original set there are 347 0's and 34 9999999's
# which is 381 total missing.  Since our sample is 22 percent of the whole dataset,
# we'd expect around 85 missing values ( which is close to 76)

#We see the same behavior in Patrimonio (col 11) and Cargas Patrimonales (col 12)
#so we substitute in NA for those values
credit_thousand[which(credit_thousand$Patrimonio == 0),11] <- NA
credit_thousand[which(credit_thousand$Patrimonio== 99999999),11] <- NA

credit_thousand[which(credit_thousand$Cargas.patrimoniales == 0),12] <- NA
credit_thousand[which(credit_thousand$Cargas.patrimoniales== 99999999),12] <- NA

#However the amount of 0's in Cargas Patrimonales is 82 percent (822 out of 1000)!
#so we are pay attention to the values generated from mice.

#now we impute those missing values
credit_thousand.imp <- mice(credit_thousand,m=1)
credit_thousand.cleaned <- complete(credit_thousand.imp)
summary(credit_thousand.cleaned)
%      Dictamen    Ant..Trabajo                 Vivienda       Plazo            Edad
%  positivo:748   Min.   : 0.000   NA                :  1   Min.   : 6.00   Min.   :19.00
%  negativo:252   1st Qu.: 1.000   alquiler          :210   1st Qu.:36.00   1st Qu.:28.75
%                 Median : 5.000   escritura publica:495   Median :48.00   Median :37.00
%                 Mean   : 7.889   contrato privado : 45   Mean   :46.91   Mean   :37.71
%                 3rd Qu.:12.000   ignora contrato  :  2   3rd Qu.:60.00   3rd Qu.:46.00
%                 Max.   :47.000   padres            :179   Max.   :60.00   Max.   :66.00
%                                  otros             : 68
%      Estado.civil Registros            Tipo.trabajo     Gastos          Ingresos
%  NA        :  1   no:832    empleado fijo    :625   Min.   : 35.00   Min.   : 19.0
%  soltero   :214   si:168    empleado temporal:105   1st Qu.: 35.00   1st Qu.: 90.0
%  casado    :731             autonomo         :229   Median : 51.00   Median :125.0
%  viudo     : 13             otros            : 41   Mean   : 55.02   Mean   :142.3
%  separado  : 38                                     3rd Qu.: 71.00   3rd Qu.:172.2
%  divorciado:  3                                     Max.   :131.00   Max.   :715.0
%
%    Patrimonio     Cargas.patrimoniales Importe.solicitado Precio.del.bien.financiado
%  Min.   :   500   Min.   :   12        Min.   : 107       Min.   : 125
%  1st Qu.:  3500   1st Qu.:  800        1st Qu.: 750       1st Qu.:1130
%  Median :  5000   Median : 1947        Median :1000       Median :1408
%  Mean   :  7834   Mean   : 2450        Mean   :1057       Mean   :1471
%  3rd Qu.:  9000   3rd Qu.: 3000        3rd Qu.:1302       3rd Qu.:1704
%  Max.   :100000   Max.   :21400        Max.   :5000       Max.   :6500
```

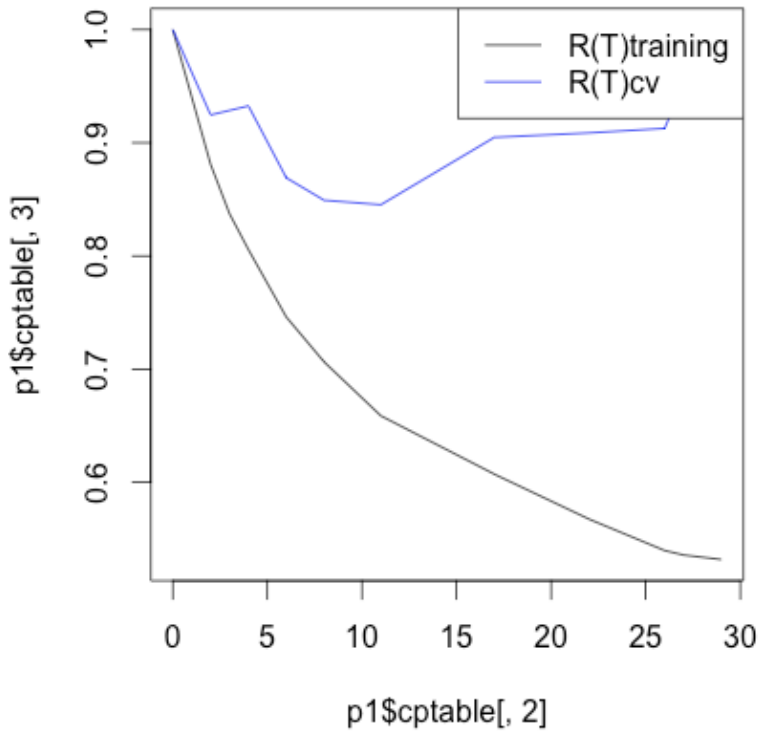
4 Produce useful rules to predict the granting of credits by using rpart() with crossvalidation to obtain a decision tree

```
library(rpart)
attach(credit_thousand.cleaned)
#grow a decision tree with 10 fold cross validation
p1 <- rpart(Dictamen ~ Ant..Trabajo + Vivienda + Plazo + Edad + Estado.civil
      + Registros + Tipo.trabajo + Gastos + Ingresos + Patrimonio + Cargas.patrimoniales
      + Importe.solicitado + Precio.del.bien.financiado, data= credit_thousand.cleaned,
      control=rpart.control(cp=0.001,xval=10))

plot(p1)


p1$cptable
```



```
%             CP nsplit rel error    xerror       xstd
% 1  0.059523810      0 1.0000000 1.0000000 0.05448168
% 2  0.043650794      2 0.8809524 0.9246032 0.05304874
% 3  0.031746032      3 0.8373016 0.9285714 0.05312779
% 4  0.029761905      4 0.8055556 0.9325397 0.05320643
% 5  0.019841270      6 0.7460317 0.8690476 0.05189757
% 6  0.015873016      8 0.7063492 0.8492063 0.05146567
% 7  0.007936508     11 0.6587302 0.8452381 0.05137793
% 8  0.006944444     17 0.6071429 0.9047619 0.05264723
% 9  0.006613757     22 0.5674603 0.9087302 0.05272838
% 10 0.003968254     26 0.5396825 0.9126984 0.05280910
% 11 0.001984127     27 0.5357143 0.9523810 0.05359347
% 12 0.001000000     29 0.5317460 0.9642857 0.05382085
```

```
plot(p1$cptable[,2],p1$cptable[,3],type="l")
lines(p1$cptable[,2],p1$cptable[,4],col="blue")
legend("topright",c("R(T)training","R(T)cv"),col=c("black","blue"),lty=1)
```

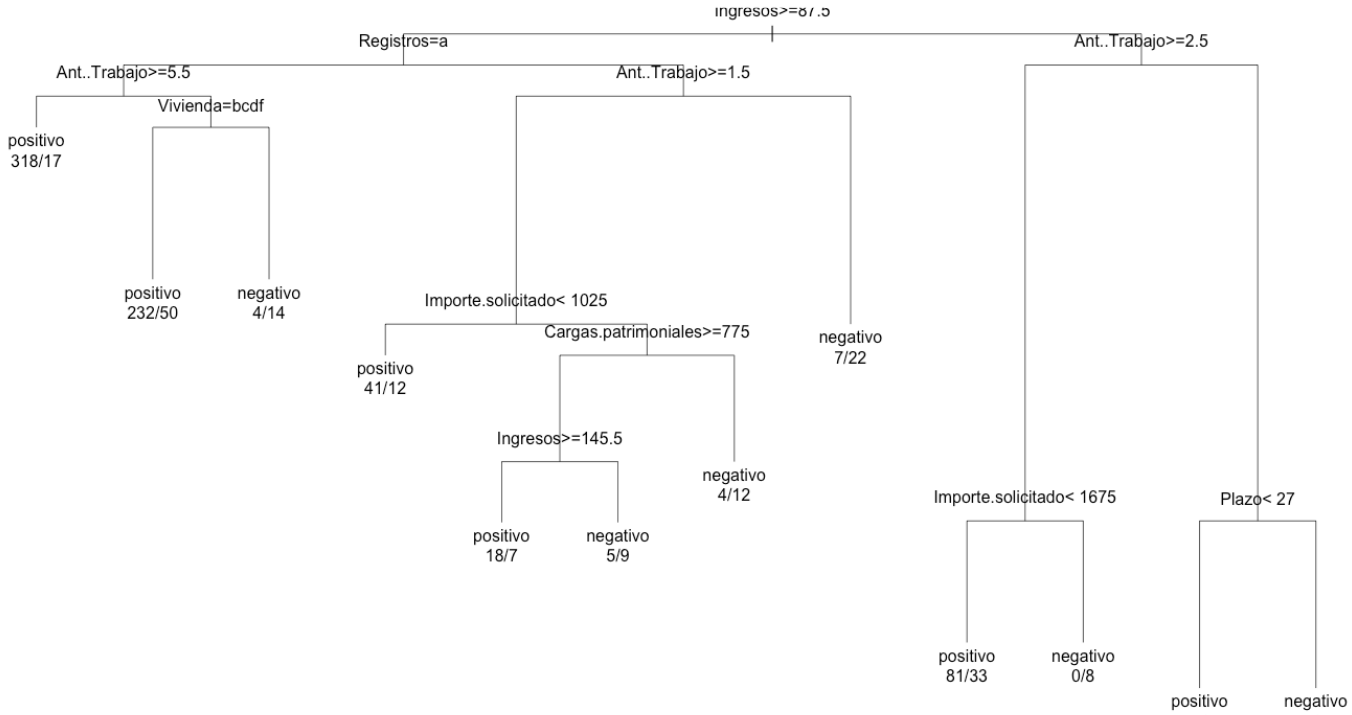



```
#We see that the complexity parameter (alpha) which gives us the lowest
```

```
#cross validation error ( of 0.8452) is 0.007936508.

#Thus we prune the tree by setting alpha to 0.00794
alfa = 0.00794
p1.pruned <- prune(p1,cp=alfa)
p1.pruned
% n= 1000
% node), split, n, loss, yval, (yprob)
%       * denotes terminal node
%  1) root 1000 252 positivo (0.74800000 0.25200000)
%    2) Ingresos>=87.5 772 143 positivo (0.81476684 0.18523316)
%      4) Registros=no 635  81 positivo (0.87244094 0.12755906)
%        8) Ant..Trabajo>=5.5 335  17 positivo (0.94925373 0.05074627) *
%        9) Ant..Trabajo< 5.5 300  64 positivo (0.78666667 0.21333333)
%         18) Vivienda=alquiler,escritura publica,contrato privado,padres 282  50 positivo (0.82269504 0.17730
%         19) Vivienda=otros 18   4 negativo (0.22222222 0.77777778) *
%      5) Registros=si 137  62 positivo (0.54744526 0.45255474)
%       10) Ant..Trabajo>=1.5 108  40 positivo (0.62962963 0.37037037)
%         20) Importe.solicitado< 1025 53  12 positivo (0.77358491 0.22641509) *
%         21) Importe.solicitado>=1025 55  27 negativo (0.49090909 0.50909091)
%           42) Cargas.patrimoniales>=775 39  16 positivo (0.58974359 0.41025641)
%             84) Ingresos>=145.5 25   7 positivo (0.72000000 0.28000000) *
%             85) Ingresos< 145.5 14   5 negativo (0.35714286 0.64285714) *
%           43) Cargas.patrimoniales< 775 16   4 negativo (0.25000000 0.75000000) *
%       11) Ant..Trabajo< 1.5 29   7 negativo (0.24137931 0.75862069) *
%    3) Ingresos< 87.5 228 109 positivo (0.52192982 0.47807018)
%      6) Ant..Trabajo>=2.5 122  41 positivo (0.66393443 0.33606557)
%       12) Importe.solicitado< 1675 114  33 positivo (0.71052632 0.28947368) *
%       13) Importe.solicitado>=1675 8   0 negativo (0.00000000 1.00000000) *
%      7) Ant..Trabajo< 2.5 106  38 negativo (0.35849057 0.64150943)
%       14) Plazo< 27 19   4 positivo (0.78947368 0.21052632) *
%       15) Plazo>=27 87  23 negativo (0.26436782 0.73563218) *

par(mfrow=c(1,1))
plot(p1.pruned)
text(p1.pruned, use.n=T, cex=0.8)


#You can then use the decision tree as a predictor as follows
test.data <- credit[-x,]
#we would need to change this test data to be categorical and handle for missing data as we did before,
#but the general idea would be as follows.
p1l=predict(p1.pruned,data=test.data)
```

5 Transform the dataframe as a transactions file

```
#5. Transform the dataframe as a transactions file (after having recoded all variables as categorical).
# We need to handle the continous variables which are Ant..Trabajo, Plazo, Edad, Gastos, Ingresos
# Patrimonio, Cargas.patrimoniales, Importe.solicitado and Precio.del.bien.financiado

table(credit_thousand.cleaned$Ant..Trabajo)
%   0   1   2   3   4   5   6   7   8   9  10  11  12   ... 33  35  36  37  38  40  47
% 123 132  98  80  31  61  34  28  37  18  59  26  37   ... 1   2   1   1   3   3   1

#too many to use each value as modality so split into 10 roughly equal bins
trabajo_levels <- cut(credit_thousand.cleaned$Ant..Trabajo,breaks=c(-1,0,1,2,3,5,8,12,15,20,45))
table(trabajo_levels)
%  (-1,0]   (0,1]   (1,2]   (2,3]   (3,5]   (5,8]  (8,12] (12,15] (15,20] (20,45]
```




```
%     123     132      98      80      92      99     140      75      87      73
credit_thousand.cleaned$Ant..Trabajo <- trabajo_levels
levels(credit_thousand.cleaned$Ant..Trabajo) <-
    c("less than 1 year", "1 to 2 years","2 to 3 years", "3 to 4 years","4 to 6 years", "6 to 9 years",
      "9 to 13 years", "13 to 16 years", "16 to 21 years", "more than 21 years")
table(credit_thousand.cleaned$Ant..Trabajo)
%   less than 1 year       1 to 2 years       2 to 3 years       3 to 4 years       4 to 6 years
%                123                132                 98                 80                 92
%       6 to 9 years      9 to 13 years     13 to 16 years     16 to 21 years more than 21 years
%                 99                140                 75                 87                 73

table(credit_thousand.cleaned$Plazo)
%   6  12  18  24  30  36  42  48  54  60
%   7  29  21  76  14 201   5 195   5 447
credit_thousand.cleaned$Plazo <- as.factor(credit_thousand.cleaned$Plazo)

table(credit_thousand.cleaned$Edad)
% 19 20 21 22 23 24 25 26 27 28 29 30 31 ... 56 57 58 59 60 61 62 63 64 65 66
%  6 11 18 28 32 20 36 33 24 42 35 28 34 ... 13  7 10  5  9  6  5  7  2  5  4
#There are too many levels so we split this up into reasonable bins
edad_levels <- cut(credit_thousand.cleaned$Edad,breaks=c(0,25,30,35,40,45,50,55,100))
credit_thousand.cleaned$Edad <- edad_levels
levels(credit_thousand.cleaned$Edad) <- c("25 and under","26 to 30","31 to 35", "36 to 40",
                                          "41 to 45","45 to 50","51 to 55","56 and up" )
table(credit_thousand.cleaned$Edad)
% 25 and under     26 to 30     31 to 35     36 to 40     41 to 45     45 to 50     51 to 55     56 and up
%           151          162          157          160          107          109           81           73

table(credit_thousand.cleaned$Gastos) #too many values so split into groups
gasto_lvls <- cut(credit_thousand.cleaned$Gastos, breaks=c(0,35,59,74,132))
credit_thousand.cleaned$Gastos <- gasto_lvls
levels(credit_thousand.cleaned$Gastos) <- c("35 and less","between 36 and 59","between 60 and 74","75 and up")
table(credit_thousand.cleaned$Gastos)
%       35 and less between 36 and 59 between 60 and 74         75 and up
```

```
%              292               246               222               240

table(credit_thousand.cleaned$Ingresos)  #again too many so split into groups
ingreso_lvls <- cut(credit_thousand.cleaned$Ingresos, breaks=c(0,67,80,100,115,130,150,175,210,716))
credit_thousand.cleaned$Ingresos <- ingreso_lvls
levels(credit_thousand.cleaned$Ingresos) <- c("67 and less","68 to 80","81 to 100","101 to 115","116 to 130",
                                              "131 to 150","151 to 175","176 to 210","210 and up")
table(credit_thousand.cleaned$Ingresos)
% 67 and less    68 to 80   81 to 100  101 to 115  116 to 130  131 to 150  151 to 175  176 to 210  210 and up
%         105          93         132         108          94         130         100         103         135

cumsum(table(credit_thousand.cleaned$Patrimonio)) # too many values
patrimonio_lvls <- cut(credit_thousand.cleaned$Patrimonio,
                       breaks=c(0,2500,3000,3500,4000,5000,6000,8000,11000,15000,100000))
credit_thousand.cleaned$Patrimonio <- patrimonio_lvls
levels(credit_thousand.cleaned$Patrimonio) <- c("2500 and under","2501 to 3000","3001 to 3500","3501 to 4000",
                "4001 to 5000","5001 to 6000","6001 to 8000","8001 to 11000","11001 to 15000", "over 15000")
table(credit_thousand.cleaned$Patrimonio)
% 2500 and under   2501 to 3000   3001 to 3500   3501 to 4000   4001 to 5000   5001 to 6000   6001 to 8000
%             107            106             69            124            151             79            101
%  8001 to 11000 11001 to 15000     over 15000
%              94             85             84

cumsum(table(credit_thousand.cleaned$Cargas.patrimoniales))
cargas.patrimoniales_lvls <- cut(credit_thousand.cleaned$Cargas.patrimoniales,
                       breaks=c(0,200,500,900,1500,1999,2500,3000,4000,21400))
credit_thousand.cleaned$Cargas.patrimoniales <- cargas.patrimoniales_lvls
levels(credit_thousand.cleaned$Cargas.patrimoniales) <- c("200 and less","201 to 500","501 to 900",
                "901 to 1500","1501 to 1999","2000 to 2500","2501 to 3000","3001 to 4000","over 4000")
table(credit_thousand.cleaned$Cargas.patrimoniales)
% 200 and less   201 to 500   501 to 900  901 to 1500 1501 to 1999 2000 to 2500 2501 to 3000 3001 to 4000
%           99          105           97          138           86           95          131          145
%    over 4000
%           104

cumsum(table(credit_thousand.cleaned$Importe.solicitado))
importe.solicitado_lvls <- cut(credit_thousand.cleaned$Importe.solicitado,
                breaks=c(0,650,950,1125,1400,5000))
credit_thousand.cleaned$Importe.solicitado <- importe.solicitado_lvls
levels(credit_thousand.cleaned$Importe.solicitado) <-
               c("650 and below","651 to 950","951 to 1125","1126 to 1400","over 1400")
table(credit_thousand.cleaned$Importe.solicitado)
% 650 and below    651 to 950   951 to 1125  1126 to 1400     over 1400
%           205           200           195           222           178

cumsum(table(credit_thousand.cleaned$Precio.del.bien.financiado))
precio.del.bien.financiado_lvls <- cut(credit_thousand.cleaned$Precio.del.bien.financiado,
                breaks=c(0,1050,1300,1515,1800,6500))
credit_thousand.cleaned$Precio.del.bien.financiado <- precio.del.bien.financiado_lvls
levels(credit_thousand.cleaned$Precio.del.bien.financiado) <-
               c("1050 and below","1051 to 1300","1301 to 1515","1516 to 1800","over 1800")
table(credit_thousand.cleaned$Precio.del.bien.financiado)
% 1050 and below   1051 to 1300   1301 to 1515   1516 to 1800      over 1800
%            195            201            200            208            196

summary(credit_thousand.cleaned)
%      Dictamen             Ant..Trabajo              Vivienda       Plazo               Edad
%  positivo:748   9 to 13 years   :140   NA               :  1   60     :447   26 to 30    :162
%  negativo:252   1 to 2 years    :132   alquiler         :210   36     :201   36 to 40    :160
```

```
%                 less than 1 year:123   escritura publica:495   48     :195   31 to 35    :157
%                 6 to 9 years    : 99   contrato privado : 45   24     : 76   25 and under:151
%                 2 to 3 years    : 98   ignora contrato  :  2   12     : 29   45 to 50    :109
%                 (Other)         :407   padres           :179   18     : 21   41 to 45    :107
%                 NA's            :  1   otros            : 68   (Other): 31   (Other)     :154

%      Estado.civil Registros            Tipo.trabajo               Gastos           Ingresos
%  NA        :  1   no:832    empleado fijo    :625   35 and less      :292   210 and up :135
%  soltero   :214   si:168    empleado temporal:105   between 36 and 59:246   81 to 100  :132
%  casado    :731             autonomo         :229   between 60 and 74:222   131 to 150 :130
%  viudo     : 13             otros            : 41   75 and up        :240   101 to 115 :108
%  separado  : 38                                                             67 and less:105
%  divorciado:  3                                                             176 to 210 :103
%                                                                             (Other)    :287

%           Patrimonio    Cargas.patrimoniales     Importe.solicitado  Precio.del.bien.financiado
%  4001 to 5000  :151   3001 to 4000:145       650 and below:205      1050 and below:195
%  3501 to 4000  :124   901 to 1500 :138       651 to 950   :200      1051 to 1300  :201
%  2500 and under:107   2501 to 3000:131       951 to 1125  :195      1301 to 1515  :200
%  2501 to 3000  :106   201 to 500  :105       1126 to 1400 :222      1516 to 1800  :208
%  6001 to 8000  :101   over 4000   :104       over 1400    :178      over 1800     :196
%  8001 to 11000 : 94   200 and less: 99
%  (Other)       :317   (Other)     :278
```

```
#Now transform dataframe to a transaction file
library (arules)
Credit.transactions <- as(credit_thousand.cleaned,"transactions")
dim(Credit.transactions)
#1000   91
```

6 Obtain the 20 more interesting association rules according the lift criterion

```
#6. Obtain the 20 more interesting association rules according the lift criterion and
#   compare the results with the previously obtained rules with the decision tree.
rules = apriori(Credit.transactions, parameter=list(support=0.35, confidence=0.75))
rules # set of 25 rules
top20rules <- sort(rules,by="lift")[1:20]
inspect(top20rules)
%    lhs                              rhs                    support confidence     lift
% 1  {Dictamen=positivo,
%     Vivienda=escritura publica} => {Estado.civil=casado}   0.369  0.8891566 1.216357
% 2  {Vivienda=escritura publica,
%     Registros=no}               => {Estado.civil=casado}   0.359  0.8777506 1.200753
% 3  {Vivienda=escritura publica} => {Estado.civil=casado}   0.433  0.8747475 1.196645
% 4  {Vivienda=escritura publica,
%     Registros=no}               => {Dictamen=positivo}     0.362  0.8850856 1.183269
% 5  {Registros=no,
%     Tipo.trabajo=empleado fijo} => {Dictamen=positivo}     0.456  0.8620038 1.152411
% 6  {Vivienda=escritura publica,
%     Estado.civil=casado}        => {Dictamen=positivo}     0.369  0.8521940 1.139297
% 7  {Vivienda=escritura publica} => {Dictamen=positivo}     0.415  0.8383838 1.120834
% 8  {Estado.civil=casado,
%     Tipo.trabajo=empleado fijo} => {Dictamen=positivo}     0.387  0.8340517 1.115042
% 9  {Estado.civil=casado,
%     Registros=no}               => {Dictamen=positivo}     0.493  0.8230384 1.100319
% 10 {Tipo.trabajo=empleado fijo} => {Dictamen=positivo}     0.511  0.8176000 1.093048
% 11 {Dictamen=positivo,
%     Tipo.trabajo=empleado fijo} => {Registros=no}          0.456  0.8923679 1.072558
```

```
% 12 {Dictamen=positivo}          => {Registros=no}          0.660  0.8823529 1.060520
% 13 {Registros=no}               => {Dictamen=positivo}     0.660  0.7932692 1.060520
% 14 {Dictamen=positivo,
%     Estado.civil=casado}        => {Registros=no}          0.493  0.8741135 1.050617
% 15 {Dictamen=positivo,
%     Vivienda=escritura publica} => {Registros=no}          0.362  0.8722892 1.048424
% 16 {Dictamen=positivo,
%     Tipo.trabajo=empleado fijo} => {Estado.civil=casado}   0.387  0.7573386 1.036031
% 17 {Estado.civil=casado}        => {Dictamen=positivo}     0.564  0.7715458 1.031478
% 18 {Dictamen=positivo}          => {Estado.civil=casado}   0.564  0.7540107 1.031478
% 19 {Tipo.trabajo=empleado fijo} => {Registros=no}          0.529  0.8464000 1.017308
% 20 {Estado.civil=casado,
%     Tipo.trabajo=empleado fijo} => {Registros=no}          0.391  0.8426724 1.012827

# The lift obtain by require such high support and confidence is actually quite low overall
# so we look at what sort of rules we would obtain by lowering the support and confidence
# requirements.

rules2 = apriori(Credit.transactions, parameter=list(support=0.01, confidence=0.6))
rules2 # set of 66705 rules
top20rules2 <- sort(rules2,by="lift")[1:20]
inspect(top20rules2)
%    lhs                                rhs                      support confidence      lift
% 1  {Estado.civil=casado,
%     Tipo.trabajo=autonomo,
%     Ingresos=210 and up,
%     Cargas.patrimoniales=over 4000}  => {Patrimonio=over 15000}   0.010  1.0000000 11.904762
% 2  {Tipo.trabajo=autonomo,
%     Ingresos=210 and up,
%     Cargas.patrimoniales=over 4000}  => {Patrimonio=over 15000}   0.011  0.9166667 10.912698
% 3  {Dictamen=positivo,
%     Estado.civil=casado,
%     Tipo.trabajo=otros}              => {Edad=56 and up}          0.014  0.7368421 10.093727
% 4  {Vivienda=escritura publica,
%     Estado.civil=casado,
%     Ingresos=210 and up,
%     Cargas.patrimoniales=over 4000}  => {Patrimonio=over 15000}   0.011  0.8461538 10.073260
% 5  {Vivienda=escritura publica,
%     Tipo.trabajo=autonomo,
%     Cargas.patrimoniales=over 4000}  => {Patrimonio=over 15000}   0.010  0.8333333  9.920635
% 6  {Dictamen=positivo,
%     Vivienda=escritura publica,
%     Ingresos=210 and up,
%     Cargas.patrimoniales=over 4000}  => {Patrimonio=over 15000}   0.010  0.8333333  9.920635
% 7  {Vivienda=escritura publica,
%     Estado.civil=casado,
%     Registros=no,
%     Ingresos=210 and up,
%     Cargas.patrimoniales=over 4000}  => {Patrimonio=over 15000}   0.010  0.8333333  9.920635
% 8  {Dictamen=positivo,
%     Vivienda=escritura publica,
%     Registros=no,
%     Ingresos=210 and up,
%     Cargas.patrimoniales=over 4000}  => {Patrimonio=over 15000}   0.010  0.8333333  9.920635
% 9  {Dictamen=positivo,
%     Estado.civil=casado,
%     Registros=no,
%     Tipo.trabajo=otros}              => {Edad=56 and up}          0.013  0.7222222  9.893455
% 10 {Dictamen=positivo,
```

```
%     Vivienda=escritura publica,
%     Estado.civil=casado,
%     Tipo.trabajo=otros}              => {Edad=56 and up}          0.010  0.7142857  9.784736
% 11 {Dictamen=positivo,
%     Vivienda=escritura publica,
%     Estado.civil=casado,
%     Registros=no,
%     Tipo.trabajo=otros}              => {Edad=56 and up}          0.010  0.7142857  9.784736
% 12 {Vivienda=escritura publica,
%     Ingresos=210 and up,
%     Cargas.patrimoniales=over 4000}  => {Patrimonio=over 15000}   0.013  0.8125000  9.672619
% 13 {Vivienda=escritura publica,
%     Estado.civil=casado,
%     Tipo.trabajo=otros}              => {Edad=56 and up}          0.013  0.6842105  9.372747
% 14 {Vivienda=escritura publica,
%     Registros=no,
%     Ingresos=210 and up,
%     Cargas.patrimoniales=over 4000}  => {Patrimonio=over 15000}   0.011  0.7857143  9.353741
% 15 {Dictamen=negativo,
%     Ant..Trabajo=less than 1 year,
%     Importe.solicitado=1126 to 1400} => {Vivienda=otros}          0.010  0.6250000  9.191176
% 16 {Dictamen=positivo,
%     Estado.civil=casado,
%     Tipo.trabajo=autonomo,
%     Cargas.patrimoniales=over 4000}  => {Patrimonio=over 15000}   0.010  0.7692308  9.157509
% 17 {Estado.civil=casado,
%     Registros=no,
%     Tipo.trabajo=autonomo,
%     Cargas.patrimoniales=over 4000}  => {Patrimonio=over 15000}   0.010  0.7692308  9.157509
% 18 {Dictamen=positivo,
%     Registros=no,
%     Tipo.trabajo=autonomo,
%     Cargas.patrimoniales=over 4000}  => {Patrimonio=over 15000}   0.010  0.7692308  9.157509
% 19 {Vivienda=escritura publica,
%     Estado.civil=casado,
%     Registros=no,
%     Tipo.trabajo=otros}              => {Edad=56 and up}          0.012  0.6666667  9.132420
% 20 {Dictamen=positivo,
%     Tipo.trabajo=autonomo,
%     Cargas.patrimoniales=over 4000}  => {Patrimonio=over 15000}   0.012  0.7500000  8.928571

#These rules have much greater lifts than before, but the rules are a little more complex
#(ie, having three or four clauses in the antecedent) than the prior rules generated.

# In comparing these twenty association against the rules obtained with the decision tree prior,
# we need to be aware that the decision tree rules were derived by how to best split
# based on the response variable of Dictamen (postive/negative)  whereas the association
# rules ones are based on just finding any association rules with a certain level of
# support and confidence, and taking those most interesting (via their lift value)
# and hence as such can not be used as a predictor for Dictamen.

# To do that though, we could limit our association rules to those which have Dictamen
# in the consequent (having rules for postivo and negativo) as such:

rulesDictamenPositivo <- subset(rules2, subset= rhs %in% "Dictamen=positivo")
rulesDictamenNegativo <- subset(rules2, subset= rhs %in% "Dictamen=negativo")

rulesDictamenPositivo # set of 14179 rules
top20rules2DictaPos <- sort(rulesDictamenPositivo,by="lift")[1:20]
```

```
inspect(top20rules2DictaPos)
%    lhs                                          rhs                 support confidence     lift
% 1  {Plazo=30,
%     Estado.civil=casado}                     => {Dictamen=positivo}   0.010          1 1.336898
% 2  {Plazo=18,
%     Importe.solicitado=650 and below}        => {Dictamen=positivo}   0.014          1 1.336898
% 3  {Plazo=18,
%     Registros=no}                            => {Dictamen=positivo}   0.016          1 1.336898
% 4  {Vivienda=escritura publica,
%     Plazo=12}                                => {Dictamen=positivo}   0.015          1 1.336898
% 5  {Ant..Trabajo=more than 21 years,
%     Edad=56 and up}                          => {Dictamen=positivo}   0.020          1 1.336898
% 6  {Edad=56 and up,
%     Precio.del.bien.financiado=1301 to 1515} => {Dictamen=positivo}   0.013          1 1.336898
% 7  {Ant..Trabajo=more than 21 years,
%     Ingresos=151 to 175}                     => {Dictamen=positivo}   0.010          1 1.336898
% 8  {Ant..Trabajo=more than 21 years,
%     Cargas.patrimoniales=201 to 500}         => {Dictamen=positivo}   0.017          1 1.336898
% 9  {Ant..Trabajo=more than 21 years,
%     Patrimonio=2501 to 3000}                 => {Dictamen=positivo}   0.010          1 1.336898
% 10 {Ant..Trabajo=more than 21 years,
%     Edad=36 to 40}                           => {Dictamen=positivo}   0.011          1 1.336898
% 11 {Ant..Trabajo=more than 21 years,
%     Plazo=48}                                => {Dictamen=positivo}   0.014          1 1.336898
% 12 {Ant..Trabajo=more than 21 years,
%     Importe.solicitado=951 to 1125}          => {Dictamen=positivo}   0.016          1 1.336898
% 13 {Ant..Trabajo=more than 21 years,
%     Precio.del.bien.financiado=1051 to 1300} => {Dictamen=positivo}   0.014          1 1.336898
% 14 {Ant..Trabajo=more than 21 years,
%     Precio.del.bien.financiado=1516 to 1800} => {Dictamen=positivo}   0.013          1 1.336898
% 15 {Ant..Trabajo=more than 21 years,
%     Gastos=between 60 and 74}                => {Dictamen=positivo}   0.020          1 1.336898
% 16 {Ant..Trabajo=13 to 16 years,
%     Cargas.patrimoniales=3001 to 4000}       => {Dictamen=positivo}   0.010          1 1.336898
% 17 {Ant..Trabajo=13 to 16 years,
%     Plazo=48}                                => {Dictamen=positivo}   0.011          1 1.336898
% 18 {Ant..Trabajo=13 to 16 years,
%     Precio.del.bien.financiado=1051 to 1300} => {Dictamen=positivo}   0.013          1 1.336898
% 19 {Plazo=24,
%     Cargas.patrimoniales=501 to 900}         => {Dictamen=positivo}   0.010          1 1.336898
% 20 {Plazo=24,
%     Patrimonio=2500 and under}               => {Dictamen=positivo}   0.011          1 1.336898

rulesDictamenNegativo # set of 116 rules
top20rules2DictaNeg <- sort(rulesDictamenNegativo,by="lift")[1:20]
inspect(top20rules2DictaNeg)
%    lhs                                      rhs                 support confidence     lift
% 1  {Ant..Trabajo=less than 1 year,
%     Vivienda=otros,
%     Importe.solicitado=1126 to 1400}      => {Dictamen=negativo}   0.010  1.0000000 3.968254
% 2  {Ant..Trabajo=less than 1 year,
%     Plazo=60,
%     Edad=36 to 40}                        => {Dictamen=negativo}   0.010  1.0000000 3.968254
% 3  {Ant..Trabajo=less than 1 year,
%     Vivienda=otros}                       => {Dictamen=negativo}   0.016  0.9411765 3.734827
% 4  {Ant..Trabajo=less than 1 year,
%     Vivienda=otros,
%     Registros=no}                         => {Dictamen=negativo}   0.013  0.9285714 3.684807
% 5  {Ant..Trabajo=less than 1 year,
```

```
%     Vivienda=otros,
%     Estado.civil=casado}                  => {Dictamen=negativo}   0.010  0.9090909 3.607504
% 6  {Vivienda=alquiler,
%     Registros=si,
%     Importe.solicitado=over 1400}         => {Dictamen=negativo}   0.011  0.8461538 3.357753
% 7  {Ant..Trabajo=less than 1 year,
%     Estado.civil=casado,
%     Cargas.patrimoniales=3001 to 4000}    => {Dictamen=negativo}   0.010  0.8333333 3.306878
% 8  {Ant..Trabajo=less than 1 year,
%     Edad=36 to 40}                        => {Dictamen=negativo}   0.014  0.8235294 3.267974
% 9  {Ant..Trabajo=less than 1 year,
%     Edad=36 to 40,
%     Estado.civil=casado}                  => {Dictamen=negativo}   0.012  0.8000000 3.174603
% 10 {Ant..Trabajo=less than 1 year,
%     Vivienda=alquiler,
%     Estado.civil=casado}                  => {Dictamen=negativo}   0.011  0.7857143 3.117914
% 11 {Ant..Trabajo=less than 1 year,
%     Registros=si}                         => {Dictamen=negativo}   0.014  0.7777778 3.086420
% 12 {Ant..Trabajo=1 to 2 years,
%     Registros=si}                         => {Dictamen=negativo}   0.014  0.7777778 3.086420
% 13 {Ant..Trabajo=less than 1 year,
%     Edad=36 to 40,
%     Registros=no}                         => {Dictamen=negativo}   0.010  0.7692308 3.052503
% 14 {Ant..Trabajo=1 to 2 years,
%     Vivienda=alquiler,
%     Plazo=36}                             => {Dictamen=negativo}   0.010  0.7692308 3.052503
% 15 {Plazo=36,
%     Registros=si,
%     Tipo.trabajo=autonomo}                => {Dictamen=negativo}   0.010  0.7692308 3.052503
% 16 {Ant..Trabajo=1 to 2 years,
%     Ingresos=68 to 80}                    => {Dictamen=negativo}   0.012  0.7500000 2.976190
% 17 {Edad=45 to 50,
%     Estado.civil=casado,
%     Registros=si}                         => {Dictamen=negativo}   0.012  0.7500000 2.976190
% 18 {Ant..Trabajo=less than 1 year,
%     Estado.civil=casado,
%     Registros=si}                         => {Dictamen=negativo}   0.012  0.7500000 2.976190
% 19 {Edad=45 to 50,
%     Registros=si}                         => {Dictamen=negativo}   0.014  0.7368421 2.923977
% 20 {Registros=si,
%     Tipo.trabajo=empleado fijo,
%     Importe.solicitado=over 1400,
%     Precio.del.bien.financiado=over 1800} => {Dictamen=negativo}   0.011  0.7333333 2.910053

# The negative dictamen rules which have higher lift, show people working for less than 1 year
# (or two to a lesser extent) are more correlated with the negative dictamen.
# We see this rule in our decision tree (when Ant Trabajo > 1.5 is false) though support is low
# in both the association rules and decision tree case.

# For the positive dictamen rules above we see that having more "Anteguidad en el trabajo"
# is more correlated with positive dictament (though with lower support).  This is also
# reflected in the decision tree. Interesting while Ingresos is which splits the highest node
# in the decision tree, it only appears in one of the top twenty interesting positive association rules
# though its possibly due to our subdividing Ingresos into too many categorical groups
# when converting the dataframe to be able to be made into a transactions one.
```