## Pregunta 2.

En aquest exercici aneu a seguir els passos del cicle de vida d'un projecte de mineria de dades pel cas d'un algorisme de classificació i més concretament un arbre de decisió. Ho fareu amb el fitxer titanic.csv. Que trobareu a la wiki. Aquest fitxer conté un registre per cada passatger que viatjava en el Titànic. En les variables es caracteritza si era home o dona, adult o nen, en quina categoria viatjava o si era membre de la tripulació.

### Estudieu les dades.

per exemple: Número de registres dels fitxer? Distribucions de valors per variables? Hi ha camps mal informats o buits?

Importació de l'arxiu, prèviament s'ha inspeccionat amb l'editor de text, es tracta d'un arxiu csv separat per $, conté les capçaleres i no pareix que tingui camps NA. L'arxiu conté 2201 observacions

passatgers<-read.csv("/Users/jmari/Dropbox/Master Ciencia de les dades/mineria de dades/PAC3/titanic.csv",header=TRUE,na.strings = "NA")

Examinem el contingut, totes les columnes s??n categ??riques, s??n completes i els valors que prenen s??n raonables.

str(passatgers)

## 'data.frame': 2201 obs. of 4 variables:  
## $ CLASS : Factor w/ 4 levels "1a","2a","3aa",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ AGE : Factor w/ 2 levels "Adult","Nen": 1 1 1 1 1 1 1 1 1 1 ...  
## $ SEX : Factor w/ 2 levels "Dona","Home": 2 2 2 2 2 2 2 2 2 2 ...  
## $ SURVIVED: Factor w/ 2 levels "Mor","Sobreviu": 2 2 2 2 2 2 2 2 2 2 ...

Examinem el sumari, a on podem veure que no hi ha cap columna amb valors NA i les categories.

summary(passatgers)

## CLASS AGE SEX SURVIVED   
## 1a :325 Adult:2092 Dona: 470 Mor :1490   
## 2a :285 Nen : 109 Home:1731 Sobreviu: 711   
## 3aa :706   
## crew:885

### Prepareu les dades.

En aquest cas ja estan en el format correcte i no cal discretitzar ni generar atributs nous. Cal triar quines s??n les variables que s'utilitzaran per construir el model i quina ??s la variable que classifica. En aquest cas la variable per la que classificarem ??s el camp de si el passatger va sobreviure o no.

Per a poder dur a terme aquest apartat, primer de tot instal·lem el paquet C50 i inclourem la llibreria C50 a la sessió fent:

install.packages("C50")  
install.packages("gmodels")  
library(C50)  
library(gmodels)

Partint d'aquest conjunt de dades crearem dos subconjunts, un per a la creació del model i altre per testà aquest model:

indexes = sample(2201,1000)  
passatgers.train = passatgers[indexes,]  
passatgers.test = passatgers[-indexes,]

### Creació del model

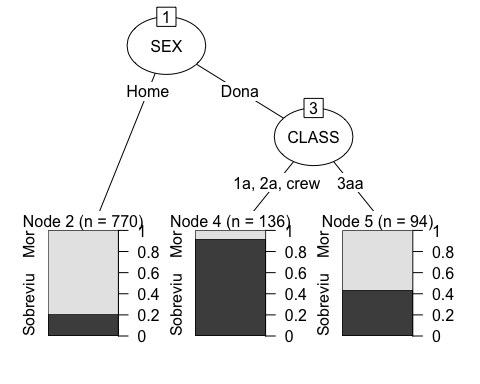
Per a crear el model, triem les columnes CLASS, AGE i SEX, i classifiquem per SURVIVED. crearem dues matrius, trainX i trainY per passar-les com argument a la funció C50::C5.0

trainX <-passatgers.train[,c("CLASS","AGE","SEX")]  
trainY <-passatgers.train[,c("SURVIVED")]  
passatgers.model <- C50::C5.0( trainX, trainY )  
summary( passatgers.model )

##   
## Call:  
## C5.0.default(x = trainX, y = trainY)  
##   
##   
## C5.0 [Release 2.07 GPL Edition] Sun Nov 5 09:42:43 2017  
## -------------------------------  
##   
## Class specified by attribute `outcome'  
##   
## Read 1000 cases (4 attributes) from undefined.data  
##   
## Decision tree:  
##   
## SEX = Home: Mor (770/163)  
## SEX = Dona:  
## :...CLASS in {1a,2a,crew}: Sobreviu (136/11)  
## CLASS = 3aa: Mor (94/41)  
##   
##   
## Evaluation on training data (1000 cases):  
##   
## Decision Tree   
## ----------------   
## Size Errors   
##   
## 3 215(21.5%) <<  
##   
##   
## (a) (b) <-classified as  
## ---- ----  
## 660 11 (a): class Mor  
## 204 125 (b): class Sobreviu  
##   
##   
## Attribute usage:  
##   
## 100.00% SEX  
## 23.00% CLASS  
##   
##   
## Time: 0.0 secs

L'error global del model que és del 22,2% (resultat donat al summary anterior).

plot(passatgers.model)



Les regles serien:

SEX = Home => Mor SEX = Dona & CLASS pertany[1a,2a,crew] => Sobreviu SEX = Dona & CLASS = 3aa => Mor

Es va provar de generar un model fent trials=10 i veure si resultava un error menor. Es va obtenir subconjunt amb un error menor, però amb les mateixes regles.

passatgers.model <- C50::C5.0( trainX, trainY, trials=10 )  
summary( passatgers.model )

### Avaluació del model

Provarem de fer la predicció amb el model obtingut, damunt del subconjunt de test.

passatgers.predict <- predict.C5.0(passatgers.model, passatgers.test)   
CrossTable(passatgers.predict,passatgers.test$SURVIVED)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1201   
##   
##   
## | passatgers.test$SURVIVED   
## passatgers.predict | Mor | Sobreviu | Row Total |   
## -------------------|-----------|-----------|-----------|  
## Mor | 810 | 253 | 1063 |   
## | 9.992 | 21.423 | |   
## | 0.762 | 0.238 | 0.885 |   
## | 0.989 | 0.662 | |   
## | 0.674 | 0.211 | |   
## -------------------|-----------|-----------|-----------|  
## Sobreviu | 9 | 129 | 138 |   
## | 76.967 | 165.016 | |   
## | 0.065 | 0.935 | 0.115 |   
## | 0.011 | 0.338 | |   
## | 0.007 | 0.107 | |   
## -------------------|-----------|-----------|-----------|  
## Column Total | 819 | 382 | 1201 |   
## | 0.682 | 0.318 | |   
## -------------------|-----------|-----------|-----------|  
##   
##

Comparant els subconjunts podem veure que dels passatgers testats que no van sobreviure s'han encertat un 80% de les prediccions. Dels que s?? que van sobreviure s'han encertat un 75%. De 1201 mostres tenim 952 encerts i 249 errats. En total tenim un 80% d'encerts.