

# PRA2

Autor: Omar Brid Roufak i Montse Sanahuja Mateu

Decembre 2021

## Descripció del dataset

El dataset escollit, recull un conjunt de dades dels passatgers que van pujar al Titanic. Recull dades demogràfiques com el nom, l'edat, el gènere i els familiars que tenien a bord i altres dades com: si van sobreviure o no, de quina classe eren, quina quantitat van pagar pel bitllet i en quin lloc van pujar al vaixell.

El que volem estudiar en aquest conjunt de dades és la correlació dels diferents atributs amb el fet de si van sobreviure o no.

## Integració i selecció del dataset

Primer de tot carregarem el dataset per poder analitzar els atributs que conté.

```
totalData <- read.csv('train.csv', stringsAsFactors = FALSE)
```

```
summary(totalData)
```

```
##   PassengerId      Survived       Pclass         Name
##   Min.   : 1.0   Min.   :0.0000   Min.   :1.000   Length:891
##   1st Qu.:223.5  1st Qu.:0.0000  1st Qu.:2.000   Class  :character
##   Median :446.0   Median :0.0000  Median :3.000   Mode   :character
##   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
##
##           Sex          Age       SibSp       Parch
##   Length:891      Min.   : 0.42  Min.   :0.000   Min.   :0.0000
##   Class  :character 1st Qu.:20.12  1st Qu.:0.000   1st Qu.:0.0000
##   Mode   :character  Median :28.00  Median :0.000   Median :0.0000
##                   Mean   :29.70  Mean   :0.523   Mean   :0.3816
##                   3rd Qu.:38.00  3rd Qu.:1.000   3rd Qu.:0.0000
##                   Max.   :80.00  Max.   :8.000   Max.   :6.0000
##                   NA's   :177
##
##           Ticket        Fare       Cabin       Embarked
##   Length:891      Min.   : 0.00  Length:891      Length:891
##   Class  :character 1st Qu.: 7.91  Class  :character  Class  :character
##   Mode   :character  Median :14.45  Mode   :character  Mode   :character
##                   Mean   :32.20
##                   3rd Qu.:31.00
##                   Max.   :512.33
##
```

```
str(totalData)
```

```
## '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       : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs T
hayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
## $ Sex        : chr "male" "female" "female" "female" ...
## $ 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     : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
## $ Fare        : num 7.25 71.28 7.92 53.1 8.05 ...
## $ Cabin      : chr "" "C85" "" "C123" ...
## $ Embarked   : chr "S" "C" "S" "S" ...
```

Tenim 12 atributs i 891 observacions.

A continuació detallarem cadascun dels atributs:

**PassengerId** integer amb el identificador de cada passatger.

**Survived** integer podem trobar dos valors: 0 i 1 segons si van sobreviure o no.

**Pclass** integer amb el número de classe en que viatjaven, 1a, 2a o 3a classe.

**Name** string amb el nom dels passatgers.

**Sex** string amb dos categories, male i female.

**Age** Numèric, edat dels passatgers.

**SibSp** numèric, nombre de germans o cònjuges a bord.

**Parch** numèric, nombre de pares o fills a bord.

**Ticket** string, amb el nom del tiquet.

**Fare** numèric, preu del tiquet.

**Cabin** string, nom de la cabina dels passatgers.

**Embarked** string, lloc de l'embarcament amb tres opcions C, S i Q.

## Neteja de les dades

Abans de passar a l'anàlisis de les dades les hem de netejar.

Primer de tot decidirem amb quins atributs ens volem quedar. Anem a revisar quants valors nuls o buits conté cada atribut:

```
colSums(is.na(totalData))
```

```
## PassengerId     Survived     Pclass      Name       Sex     Age
##          0         0          0          0         0    177
##      SibSp     Parch     Ticket     Fare     Cabin Embarked
##          0         0          0          0         0         0
```

Tenim 177 nuls en el camp Age, com no podem saber l'edat d'aquests passatgers els hi podem assignar la mitjana.

```
totalData$Age[is.na(totalData$Age)] <- mean(totalData$Age,na.rm=T)
```

```
colSums(totalData=="")
```

```
## PassengerId     Survived     Pclass      Name       Sex     Age
##          0         0          0          0         0     0
##      SibSp     Parch     Ticket     Fare     Cabin Embarked
##          0         0          0          0         0    687     2
```

De 891 observacions, en l'atribut Cabin que mostra el nom de la cabina dels tripulants, tenim 687 valors buits, com no podem saber el nom de la cabina, assignarem “Desconegut” en aquests valors buits.

```
totalData$Cabin[(totalData$Cabin)=="" ] <- "Desconegut"
```

En Embarked tenim també dos valors buits. Anem a visualitzar aquests dos passatgers:

```
totalData[totalData$Embarked=="" , ]
```

```
##   PassengerId Survived Pclass          Name
## 62        62        1     1   Icard, Miss. Amelie
## 830       830        1     1 Stone, Mrs. George Nelson (Martha Evelyn)
##   Sex Age SibSp Parch Ticket Fare Cabin Embarked
## 62 female 38     0     0 113572   80    B28
## 830 female 62     0     0 113572   80    B28
```

És curiós que tots dos passatgers tenen el mateix número de bitllet, de preu i de cabina. En aquest cas, com no podem saber des d'on van embarcar aquests dos individus, ho omplirem com a “Desconegut”.

```
totalData$Embarked[(totalData$Embarked)=="" ] <- "Desconegut"
```

Finalment, ja no tenim cap atribut amb valors buits.

```
colSums(totalData=="")
```

```
## PassengerId     Survived     Pclass      Name       Sex     Age
##          0         0          0          0         0     0
##      SibSp     Parch     Ticket     Fare     Cabin Embarked
##          0         0          0          0         0         0
```

Si ens tornem a fixar amb l'atribut Age, veiem que hi ha valors que contenen decimals.

```
totalData$Age
```

```

## [1] 22.00000 38.00000 26.00000 35.00000 35.00000 29.69912 54.00000 2.00000
## [9] 27.00000 14.00000 4.00000 58.00000 20.00000 39.00000 14.00000 55.00000
## [17] 2.00000 29.69912 31.00000 29.69912 35.00000 34.00000 15.00000 28.00000
## [25] 8.00000 38.00000 29.69912 19.00000 29.69912 29.69912 40.00000 29.69912
## [33] 29.69912 66.00000 28.00000 42.00000 29.69912 21.00000 18.00000 14.00000
## [41] 40.00000 27.00000 29.69912 3.00000 19.00000 29.69912 29.69912 29.69912
## [49] 29.69912 18.00000 7.00000 21.00000 49.00000 29.00000 65.00000 29.69912
## [57] 21.00000 28.50000 5.00000 11.00000 22.00000 38.00000 45.00000 4.00000
## [65] 29.69912 29.69912 29.00000 19.00000 17.00000 26.00000 32.00000 16.00000
## [73] 21.00000 26.00000 32.00000 25.00000 29.69912 29.69912 0.83000 30.00000
## [81] 22.00000 29.00000 29.69912 28.00000 17.00000 33.00000 16.00000 29.69912
## [89] 23.00000 24.00000 29.00000 20.00000 46.00000 26.00000 59.00000 29.69912
## [97] 71.00000 23.00000 34.00000 34.00000 28.00000 29.69912 21.00000 33.00000
## [105] 37.00000 28.00000 21.00000 29.69912 38.00000 29.69912 47.00000 14.50000
## [113] 22.00000 20.00000 17.00000 21.00000 70.50000 29.00000 24.00000 2.00000
## [121] 21.00000 29.69912 32.50000 32.50000 54.00000 12.00000 29.69912 24.00000
## [129] 29.69912 45.00000 33.00000 20.00000 47.00000 29.00000 25.00000 23.00000
## [137] 19.00000 37.00000 16.00000 24.00000 29.69912 22.00000 24.00000 19.00000
## [145] 18.00000 19.00000 27.00000 9.00000 36.50000 42.00000 51.00000 22.00000
## [153] 55.50000 40.50000 29.69912 51.00000 16.00000 30.00000 29.69912 29.69912
## [161] 44.00000 40.00000 26.00000 17.00000 1.00000 9.00000 29.69912 45.00000
## [169] 29.69912 28.00000 61.00000 4.00000 1.00000 21.00000 56.00000 18.00000
## [177] 29.69912 50.00000 30.00000 36.00000 29.69912 29.69912 9.00000 1.00000
## [185] 4.00000 29.69912 29.69912 45.00000 40.00000 36.00000 32.00000 19.00000
## [193] 19.00000 3.00000 44.00000 58.00000 29.69912 42.00000 29.69912 24.00000
## [201] 28.00000 29.69912 34.00000 45.50000 18.00000 2.00000 32.00000 26.00000
## [209] 16.00000 40.00000 24.00000 35.00000 22.00000 30.00000 29.69912 31.00000
## [217] 27.00000 42.00000 32.00000 30.00000 16.00000 27.00000 51.00000 29.69912
## [225] 38.00000 22.00000 19.00000 20.50000 18.00000 29.69912 35.00000 29.00000
## [233] 59.00000 5.00000 24.00000 29.69912 44.00000 8.00000 19.00000 33.00000
## [241] 29.69912 29.69912 29.00000 22.00000 30.00000 44.00000 25.00000 24.00000
## [249] 37.00000 54.00000 29.69912 29.00000 62.00000 30.00000 41.00000 29.00000
## [257] 29.69912 30.00000 35.00000 50.00000 29.69912 3.00000 52.00000 40.00000
## [265] 29.69912 36.00000 16.00000 25.00000 58.00000 35.00000 29.69912 25.00000
## [273] 41.00000 37.00000 29.69912 63.00000 45.00000 29.69912 7.00000 35.00000
## [281] 65.00000 28.00000 16.00000 19.00000 29.69912 33.00000 30.00000 22.00000
## [289] 42.00000 22.00000 26.00000 19.00000 36.00000 24.00000 24.00000 29.69912
## [297] 23.50000 2.00000 29.69912 50.00000 29.69912 29.69912 19.00000 29.69912
## [305] 29.69912 0.92000 29.69912 17.00000 30.00000 30.00000 24.00000 18.00000
## [313] 26.00000 28.00000 43.00000 26.00000 24.00000 54.00000 31.00000 40.00000
## [321] 22.00000 27.00000 30.00000 22.00000 29.69912 36.00000 61.00000 36.00000
## [329] 31.00000 16.00000 29.69912 45.50000 38.00000 16.00000 29.69912 29.69912
## [337] 29.00000 41.00000 45.00000 45.00000 2.00000 24.00000 28.00000 25.00000
## [345] 36.00000 24.00000 40.00000 29.69912 3.00000 42.00000 23.00000 29.69912
## [353] 15.00000 25.00000 29.69912 28.00000 22.00000 38.00000 29.69912 29.69912
## [361] 40.00000 29.00000 45.00000 35.00000 29.69912 30.00000 60.00000 29.69912
## [369] 29.69912 24.00000 25.00000 18.00000 19.00000 22.00000 3.00000 29.69912
## [377] 22.00000 27.00000 20.00000 19.00000 42.00000 1.00000 32.00000 35.00000
## [385] 29.69912 18.00000 1.00000 36.00000 29.69912 17.00000 36.00000 21.00000
## [393] 28.00000 23.00000 24.00000 22.00000 31.00000 46.00000 23.00000 28.00000
## [401] 39.00000 26.00000 21.00000 28.00000 20.00000 34.00000 51.00000 3.00000
## [409] 21.00000 29.69912 29.69912 29.69912 33.00000 29.69912 44.00000 29.69912

```

```

## [417] 34.00000 18.00000 30.00000 10.00000 29.69912 21.00000 29.00000 28.00000
## [425] 18.00000 29.69912 28.00000 19.00000 29.69912 32.00000 28.00000 29.69912
## [433] 42.00000 17.00000 50.00000 14.00000 21.00000 24.00000 64.00000 31.00000
## [441] 45.00000 20.00000 25.00000 28.00000 29.69912 4.00000 13.00000 34.00000
## [449] 5.00000 52.00000 36.00000 29.69912 30.00000 49.00000 29.69912 29.00000
## [457] 65.00000 29.69912 50.00000 29.69912 48.00000 34.00000 47.00000 48.00000
## [465] 29.69912 38.00000 29.69912 56.00000 29.69912 0.75000 29.69912 38.00000
## [473] 33.00000 23.00000 22.00000 29.69912 34.00000 29.00000 22.00000 2.00000
## [481] 9.00000 29.69912 50.00000 63.00000 25.00000 29.69912 35.00000 58.00000
## [489] 30.00000 9.00000 29.69912 21.00000 55.00000 71.00000 21.00000 29.69912
## [497] 54.00000 29.69912 25.00000 24.00000 17.00000 21.00000 29.69912 37.00000
## [505] 16.00000 18.00000 33.00000 29.69912 28.00000 26.00000 29.00000 29.69912
## [513] 36.00000 54.00000 24.00000 47.00000 34.00000 29.69912 36.00000 32.00000
## [521] 30.00000 22.00000 29.69912 44.00000 29.69912 40.50000 50.00000 29.69912
## [529] 39.00000 23.00000 2.00000 29.69912 17.00000 29.69912 30.00000 7.00000
## [537] 45.00000 30.00000 29.69912 22.00000 36.00000 9.00000 11.00000 32.00000
## [545] 50.00000 64.00000 19.00000 29.69912 33.00000 8.00000 17.00000 27.00000
## [553] 29.69912 22.00000 22.00000 62.00000 48.00000 29.69912 39.00000 36.00000
## [561] 29.69912 40.00000 28.00000 29.69912 29.69912 24.00000 19.00000 29.00000
## [569] 29.69912 32.00000 62.00000 53.00000 36.00000 29.69912 16.00000 19.00000
## [577] 34.00000 39.00000 29.69912 32.00000 25.00000 39.00000 54.00000 36.00000
## [585] 29.69912 18.00000 47.00000 60.00000 22.00000 29.69912 35.00000 52.00000
## [593] 47.00000 29.69912 37.00000 36.00000 29.69912 49.00000 29.69912 49.00000
## [601] 24.00000 29.69912 29.69912 44.00000 35.00000 36.00000 30.00000 27.00000
## [609] 22.00000 40.00000 39.00000 29.69912 29.69912 29.69912 35.00000 24.00000
## [617] 34.00000 26.00000 4.00000 26.00000 27.00000 42.00000 20.00000 21.00000
## [625] 21.00000 61.00000 57.00000 21.00000 26.00000 29.69912 80.00000 51.00000
## [633] 32.00000 29.69912 9.00000 28.00000 32.00000 31.00000 41.00000 29.69912
## [641] 20.00000 24.00000 2.00000 29.69912 0.75000 48.00000 19.00000 56.00000
## [649] 29.69912 23.00000 29.69912 18.00000 21.00000 29.69912 18.00000 24.00000
## [657] 29.69912 32.00000 23.00000 58.00000 50.00000 40.00000 47.00000 36.00000
## [665] 20.00000 32.00000 25.00000 29.69912 43.00000 29.69912 40.00000 31.00000
## [673] 70.00000 31.00000 29.69912 18.00000 24.50000 18.00000 43.00000 36.00000
## [681] 29.69912 27.00000 20.00000 14.00000 60.00000 25.00000 14.00000 19.00000
## [689] 18.00000 15.00000 31.00000 4.00000 29.69912 25.00000 60.00000 52.00000
## [697] 44.00000 29.69912 49.00000 42.00000 18.00000 35.00000 18.00000 25.00000
## [705] 26.00000 39.00000 45.00000 42.00000 22.00000 29.69912 24.00000 29.69912
## [713] 48.00000 29.00000 52.00000 19.00000 38.00000 27.00000 29.69912 33.00000
## [721] 6.00000 17.00000 34.00000 50.00000 27.00000 20.00000 30.00000 29.69912
## [729] 25.00000 25.00000 29.00000 11.00000 29.69912 23.00000 23.00000 28.50000
## [737] 48.00000 35.00000 29.69912 29.69912 29.69912 36.00000 21.00000 24.00000
## [745] 31.00000 70.00000 16.00000 30.00000 19.00000 31.00000 4.00000 6.00000
## [753] 33.00000 23.00000 48.00000 0.67000 28.00000 18.00000 34.00000 33.00000
## [761] 29.69912 41.00000 20.00000 36.00000 16.00000 51.00000 29.69912 30.50000
## [769] 29.69912 32.00000 24.00000 48.00000 57.00000 29.69912 54.00000 18.00000
## [777] 29.69912 5.00000 29.69912 43.00000 13.00000 17.00000 29.00000 29.69912
## [785] 25.00000 25.00000 18.00000 8.00000 1.00000 46.00000 29.69912 16.00000
## [793] 29.69912 29.69912 25.00000 39.00000 49.00000 31.00000 30.00000 30.00000
## [801] 34.00000 31.00000 11.00000 0.42000 27.00000 31.00000 39.00000 18.00000
## [809] 39.00000 33.00000 26.00000 39.00000 35.00000 6.00000 30.50000 29.69912
## [817] 23.00000 31.00000 43.00000 10.00000 52.00000 27.00000 38.00000 27.00000
## [825] 2.00000 29.69912 29.69912 1.00000 29.69912 62.00000 15.00000 0.83000

```

```
## [833] 29.69912 23.00000 18.00000 39.00000 21.00000 29.69912 32.00000 29.69912
## [841] 20.00000 16.00000 30.00000 34.50000 17.00000 42.00000 29.69912 35.00000
## [849] 28.00000 29.69912 4.00000 74.00000 9.00000 16.00000 44.00000 18.00000
## [857] 45.00000 51.00000 24.00000 29.69912 41.00000 21.00000 48.00000 29.69912
## [865] 24.00000 42.00000 27.00000 31.00000 29.69912 4.00000 26.00000 47.00000
## [873] 33.00000 47.00000 28.00000 15.00000 20.00000 19.00000 29.69912 56.00000
## [881] 25.00000 33.00000 22.00000 28.00000 25.00000 39.00000 27.00000 19.00000
## [889] 29.69912 26.00000 32.00000
```

Els menors d'1 significa que l'individu és un nadó de mesos, però els superiors a 1 no queda molt clar. El que farem serà discretitzar aquesta columna per formar grups d'edat.

```
totalData["Age"] <- cut(totalData$Age, breaks = c(0,10,20,30,40,50,60,70,100), labels = c("0-9",
"10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-79"))

head(totalData)
```

```
##   PassengerId Survived Pclass
## 1           1         0     3
## 2           2         1     1
## 3           3         1     3
## 4           4         1     1
## 5           5         0     3
## 6           6         0     3
##
##                                     Name   Sex Age SibSp Parch
## 1             Braund, Mr. Owen Harris male 20-29    1    0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 30-39    1    0
## 3           Heikkinen, Miss. Laina female 20-29    0    0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female 30-39    1    0
## 5           Allen, Mr. William Henry male 30-39    0    0
## 6           Moran, Mr. James   male 20-29    0    0
##
##       Ticket   Fare Cabin Embarked
## 1 A/5 21171 7.2500   Desconegut      S
## 2   PC 17599 71.2833        C85      C
## 3 STON/O2. 3101282 7.9250   Desconegut      S
## 4      113803 53.1000        C123      S
## 5      373450  8.0500   Desconegut      S
## 6      330877  8.4583   Desconegut      Q
```

Un cop hem netejat les dades i aplicat les transformacions pertinents, procedim a analitzar-les.

## Anàlisis de les dades

En primer lloc fem un anàlisi visual de les dades de les que disposem. Per tal de fer-ho, construïm la següent visualització.

```
if (!require('grid')) install.packages('grid')

## Loading required package: grid
```

```
if (!require('gridExtra')) install.packages('gridExtra')
```

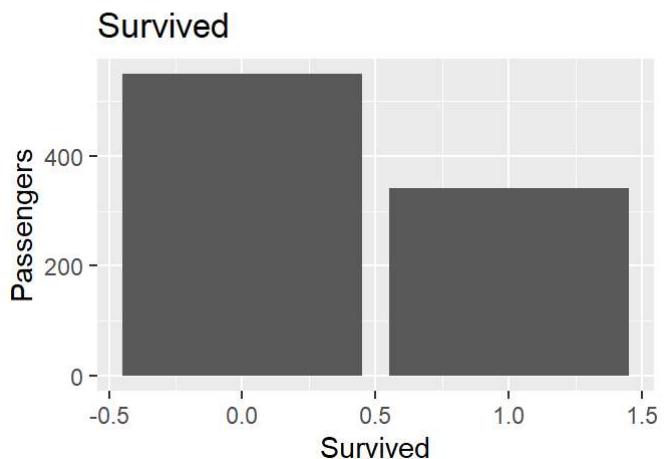
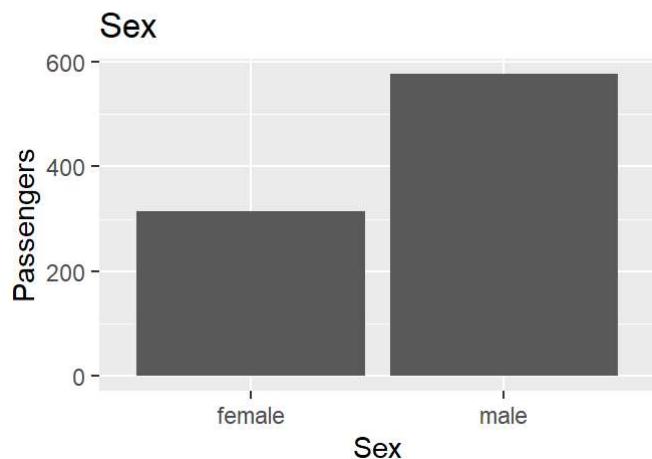
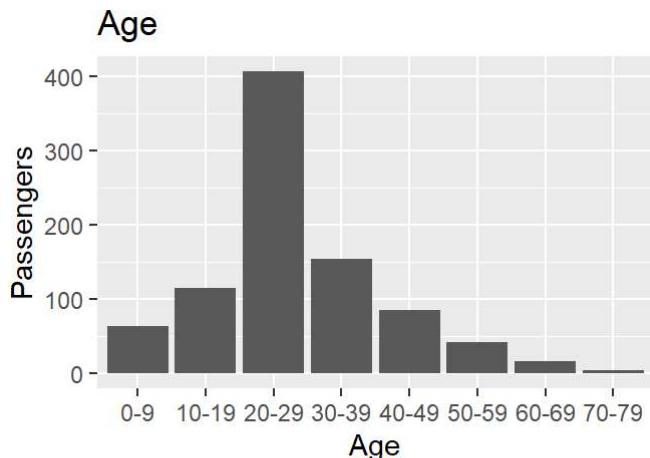
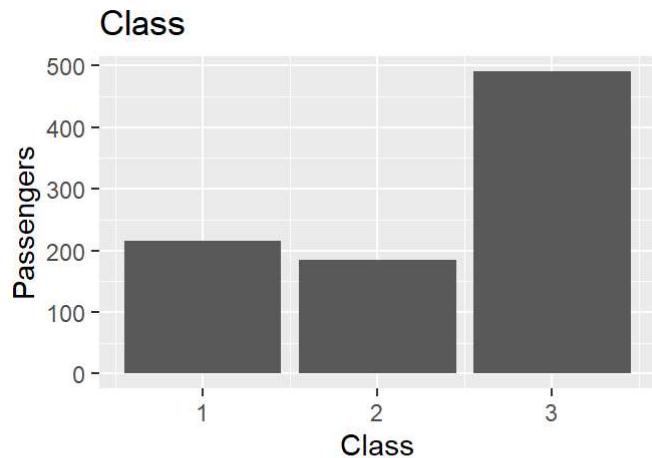
```
## Loading required package: gridExtra
```

```
if (!require('ggplot2')) install.packages('ggplot2')
```

```
## Loading required package: ggplot2
```

```
library(gridExtra)
library(ggplot2)

grid.newpage()
plotbyClass<-ggplot(totalData,aes(Pclass))+geom_bar() +labs(x="Class", y="Passengers")+
guides(fill=guide_legend(title=""))+ scale_fill_manual(values=c("blue","#008000"))+ggtitle("Class")
plotbyAge<-ggplot(totalData,aes(Age))+geom_bar() +labs(x="Age", y="Passengers")+
guides(fill=guide_legend(title=""))+ scale_fill_manual(values=c("blue","#008000"))+ggtitle("Age")
plotbySex<-ggplot(totalData,aes(Sex))+geom_bar() +labs(x="Sex", y="Passengers")+
guides(fill=guide_legend(title=""))+ scale_fill_manual(values=c("blue","#008000"))+ggtitle("Sex")
plotbySurvived<-ggplot(totalData,aes(Survived))+geom_bar() +labs(x="Survived", y="Passengers")+
guides(fill=guide_legend(title=""))+ scale_fill_manual(values=c("blue","#008000"))+ggtitle("Survived")
grid.arrange(plotbyClass,plotbyAge,plotbySex,plotbySurvived,ncol=2)
```



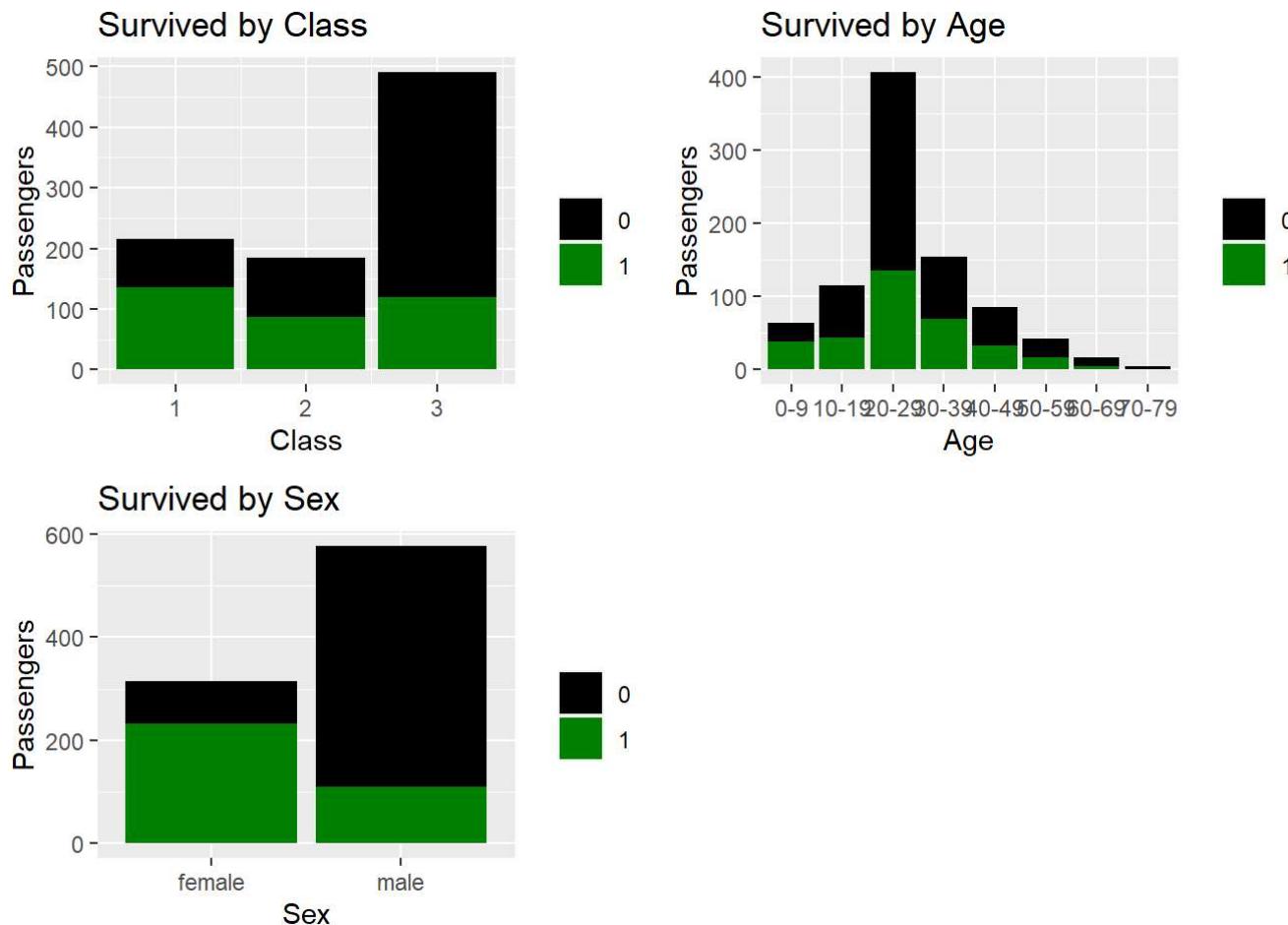
Com podem observar, en el dataset amb el que treballarem hi ha una majoria de persones que viatjaven en 3a classe. A més, comptem amb més passatgers de sexe masculí que de sexe femení. Per altra banda, tenim més mostres de passatgers no-supervivents que de supervivents. Tots aquests fets s'han de tenir en compte a l'hora de construir un model per tal de no introduir biaixos.

L'anàlisi que es vol fer amb el dataset proposat és la relació entre els diferents atributs amb si els passatgers han sobreviscut o no i extreure'n conclusions. És per això que mitjançant la següent visualització veurem quina proporció de passatgers han sobreviscut segons l'atribut observat.

Per tal de fer això, haurem de convertir l'atribut "Survived" a categòric (ja que ara és numèric).

```
totalData$Survived = as.factor(totalData$Survived)
```

```
grid.newpage()
plotbyClassbySurv<-ggplot(totalData,aes(Pclass,fill=Survived))+geom_bar() +labs(x="Class", y="Passengers")+
guides(fill=guide_legend(title=""))+ scale_fill_manual(values=c("black","#008000"))+
ggtitle("Survived by Class")
plotbyAgebySurv<-ggplot(totalData,aes(Age,fill=Survived))+geom_bar() +labs(x="Age", y="Passenger s")+
guides(fill=guide_legend(title=""))+ scale_fill_manual(values=c("black","#008000"))+ggtitle ("Survived by Age")
plotbySexbySurv<-ggplot(totalData,aes(Sex,fill=Survived))+geom_bar() +labs(x="Sex", y="Passenger s")+
guides(fill=guide_legend(title=""))+ scale_fill_manual(values=c("black","#008000"))+ggtitle ("Survived by Sex")
grid.arrange(plotbyClassbySurv,plotbyAgebySurv,plotbySexbySurv,ncol=2)
```



A simple vista podem observar com la majoria dels passatgers que no van sobreviure viatjaven en tercera classe. També podem observar que, en proporció, el percentatge de dones sobrevivents és major que el d'homes. Anem a veure els resultats de forma numèrica.

```
taula_SS <- table(totalData$Sex, totalData$Survived)
taula_SS
```

```
##
##          0   1
## female  81 233
## male    468 109
```

```
prop.table(taula_SS, margin = 1)
```

```
##
##          0         1
## female 0.2579618 0.7420382
## male   0.8110919 0.1889081
```

Com podem observar a la taula anterior, tenim un 74,2% de dones que sobreviuen front a un 18,8% d'homes sobrevivents.

```
taula_CS <- table(totalData$Pclass, totalData$Survived)
taula_CS
```

```
##
##          0   1
## 1  80 136
## 2  97  87
## 3 372 119
```

```
prop.table(taula_CS, margin = 1)
```

```
##
##          0         1
## 1 0.3703704 0.6296296
## 2 0.5271739 0.4728261
## 3 0.7576375 0.2423625
```

Veiem que el percentatge de gent que sobreviu en primera classe és pràcticament el doble que de gent no-sobrevivent. Per contra, en tercera classe el percentatge de passatgers que van morir és del triple que de passatgers que van sobreviure.

```
taula_AS <- table(totalData$Age, totalData$Survived)
taula_AS
```

```
##          0   1
##  0-9     26  38
## 10-19    71  44
## 20-29   271 136
## 30-39    86  69
## 40-49    53  33
## 50-59    25  17
## 60-69    13   4
## 70-79     4   1
```

```
prop.table(taula_AS, margin = 1)
```

```
##          0           1
##  0-9  0.4062500 0.5937500
## 10-19 0.6173913 0.3826087
## 20-29 0.6658477 0.3341523
## 30-39 0.5548387 0.4451613
## 40-49 0.6162791 0.3837209
## 50-59 0.5952381 0.4047619
## 60-69 0.7647059 0.2352941
## 70-79 0.8000000 0.2000000
```

Com veiem a la taula anterior, la franja d'edat que té en proporció més supervivents és la franja 0-9. La que menys la trobem a l'extrem, a la franja 70-79.

```
taula_ASC <- table(totalData$Age, totalData$Survived, totalData$Pclass)
taula_ASC
```

```
## , , = 1
##
##
##          0   1
## 0-9      1   2
## 10-19    3   15
## 20-29    27  43
## 30-39    12  37
## 40-49    16  21
## 50-59    10  15
## 60-69    9   2
## 70-79    2   1
##
## , , = 2
##
##
##          0   1
## 0-9      0   17
## 10-19    9   9
## 20-29    43  29
## 30-39    24  19
## 40-49    9   10
## 50-59    10  2
## 60-69    2   1
## 70-79    0   0
##
## , , = 3
##
##
##          0   1
## 0-9      25  19
## 10-19    59  20
## 20-29    201 64
## 30-39    50  13
## 40-49    28  2
## 50-59    5   0
## 60-69    2   1
## 70-79    2   0
```

```
prop.table(taula_ASC, margin = 1)
```

```

## , , = 1
##
##
##          0      1
## 0-9    0.01562500 0.03125000
## 10-19   0.02608696 0.13043478
## 20-29   0.06633907 0.10565111
## 30-39   0.07741935 0.23870968
## 40-49   0.18604651 0.24418605
## 50-59   0.23809524 0.35714286
## 60-69   0.52941176 0.11764706
## 70-79   0.40000000 0.20000000
##
## , , = 2
##
##
##          0      1
## 0-9    0.00000000 0.26562500
## 10-19   0.07826087 0.07826087
## 20-29   0.10565111 0.07125307
## 30-39   0.15483871 0.12258065
## 40-49   0.10465116 0.11627907
## 50-59   0.23809524 0.04761905
## 60-69   0.11764706 0.05882353
## 70-79   0.00000000 0.00000000
##
## , , = 3
##
##
##          0      1
## 0-9    0.39062500 0.29687500
## 10-19   0.51304348 0.17391304
## 20-29   0.49385749 0.15724816
## 30-39   0.32258065 0.08387097
## 40-49   0.32558140 0.02325581
## 50-59   0.11904762 0.00000000
## 60-69   0.11764706 0.05882353
## 70-79   0.40000000 0.00000000

```

De la taula anterior veiem fets interessants. Per exemple, a tercera classe no va sobreviure cap persona de la franja 70-79. També veiem que d'aquesta mateixa franja, cap persona viatjava en segona classe. També podem observar que dels passatgers de la franja 0-9 anys, pràcticament tots els que van morir viatjaven en tercera classe.

Procedim a fer tests estadístics per tal de veure el grau de significança de la relació entre els atributs.

```

if(!require(DescTools)){
  install.packages('DescTools', repos='http://cran.us.r-project.org')
  library(DescTools)
}

```

```
## Loading required package: DescTools
```

```
Phi(taula_SS)
```

```
## [1] 0.5433514
```

```
CramerV(taula_SS)
```

```
## [1] 0.5433514
```

La funció Phi() ens retorna el coeficient de correlació entre dues variables. El coeficient és un valor entre 0 i 1, que quan major és, indica major correlació entre les variables. Per altra banda, la funció CramerV() ens retorna un estadístic que indica la mesura d'associació entre dues variables nominals.

Cal destacar que per a una taula 2x2 (el qual és el nostre cas), el valor absolut de l'estadístic Phi equival a la V de Crammer. És per això que els valors obtinguts coincideixen.

Pel valor obtingut, el gènere està mitjanament correlacionat amb haver sobreviscut. Anem a veure els estadístics per a les altres variables.

```
Phi(taula_CS)
```

```
## [1] 0.3398174
```

```
Phi(taula_AS)
```

```
## [1] 0.153579
```

Observem que de les variables analitzades, la que més correlació té amb haver sobreviscut és 'Sex'. Tot i així, el valor obtingut per al coeficient de correlació no ens indica una relació estadística significativa.

A continuació revisarem també la correlació entre la classe i la quantitat de familiars a bord. Ho farem amb les correlacions de Pearson i Spearman.

```
cor.test(totalData$Pclass, totalData$SibSp)
```

```
## 
## Pearson's product-moment correlation
## 
## data: totalData$Pclass and totalData$SibSp
## t = 2.4858, df = 889, p-value = 0.01311
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.01749944 0.14795146
## sample estimates:
##      cor
## 0.08308136
```

```
cor.test(totalData$Pclass, totalData$SibSp, method = "spearman")
```

```
## Warning in cor.test.default(totalData$Pclass, totalData$SibSp, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
## 
## Spearman's rank correlation rho
## 
## data: totalData$Pclass and totalData$SibSp
## S = 122962713, p-value = 0.1995
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## -0.04301877
```

En el cas de la correlació entre la classe i la quantitat de germans o cònjuges a bord veiem que és compleix la condició de normalitat, ja que p-valor és major a 0.05. Per tant, ens basarem amb la correlació de Pearson. No obstant, el valor de la correlació és molt petit. No podem confirmar que hi hagi correlació entre aquestes dues variables.

```
cor.test(totalData$Pclass, totalData$Parch)
```

```
## 
## Pearson's product-moment correlation
## 
## data: totalData$Pclass and totalData$Parch
## t = 0.54998, df = 889, p-value = 0.5825
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.04729202 0.08401831
## sample estimates:
##      cor
## 0.01844267
```

```
cor.test(totalData$Pclass, totalData$Parch, method = "spearman")
```

```
## Warning in cor.test.default(totalData$Pclass, totalData$Parch, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
## Spearman's rank correlation rho
##
## data: totalData$Pclass and totalData$Parch
## S = 120579257, p-value = 0.4967
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##          rho
## -0.02280134
```

Si ho revisem amb la quantitat de pares o fills a bord veiem que compleix també la condició de normalitat, però la correlació també és molt petita. En aquest cas tampoc podem afirmar que hi hagi correlació entre aquestes dues variables.

Finalment, per acabar les analisis utilitzarem el mètode de classificació. Farem un arbre de decisió amb els atributs Sex, Age i PClass.

Primer de tot crearem un conjunt de dades nou amb els atributs que necessitem.

```
if (!require('dplyr')) install.packages('dplyr'); library(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
data <- select( totalData, Pclass, Age , Sex , Survived)
head(data)
```

```
##   Pclass    Age     Sex Survived
## 1      3 20-29 male      0
## 2      1 30-39 female     1
## 3      3 20-29 female     1
## 4      1 30-39 female     1
## 5      3 30-39 male      0
## 6      3 20-29 male      0
```

Passarem tots aquests atributs a categòrics.

```
str(data)
```

```
## 'data.frame': 891 obs. of 4 variables:
## $ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...
## $ Age    : Factor w/ 8 levels "0-9","10-19",...: 3 4 3 4 4 3 6 1 3 2 ...
## $ Sex    : chr "male" "female" "female" "female" ...
## $ Survived: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
```

```
data$Sex <- as.factor(data$Sex)
data$Pclass <- as.factor(data$Pclass)
```

```
str(data)
```

```
## 'data.frame': 891 obs. of 4 variables:
## $ Pclass : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
## $ Age    : Factor w/ 8 levels "0-9","10-19",...: 3 4 3 4 4 3 6 1 3 2 ...
## $ Sex    : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
## $ Survived: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
```

A continuació dividirem el conjunt de dades en Entrenament i Test, 2 terços seran entrenament i el terç restant el test.

```
set.seed(666)
y <- data[,4]
X <- data[,1:3]

split_prop <- 3
indexes = sample(1:nrow(data), size=floor(((split_prop-1)/split_prop)*nrow(data)))
trainX<-X[indexes,]
trainy<-y[indexes]
testX<-X[-indexes,]
testy<-y[-indexes]
```

```
summary(trainX)
```

```
## Pclass      Age       Sex
## 1:152    20-29 :278   female:194
## 2:119    30-39 :101   male  :400
## 3:323    10-19 : 75
##          40-49 : 56
##          0-9  : 45
##          50-59 : 22
##          (Other): 17
```

```
str(trainy)
```

```
## Factor w/ 2 levels "0","1": 2 1 2 1 2 2 1 2 1 1 ...
```

```
summary(testX)
```

```
## Pclass      Age       Sex
## 1: 64     20-29 :129   female:120
## 2: 65     30-39 : 54   male  :177
## 3:168    10-19 : 40
##          40-49 : 30
##          50-59 : 20
##          0-9  : 19
##          (Other): 5
```

```
summary(testy)
```

```
## 0 1
## 173 124
```

Un cop tenim les dades separades en dos conjunts ja podem crear l'arbre de decisió.

```
if(!require(C50)){
  install.packages('C50', repos='http://cran.us.r-project.org')
  library(C50)
}
```

```
## Loading required package: C50
```

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

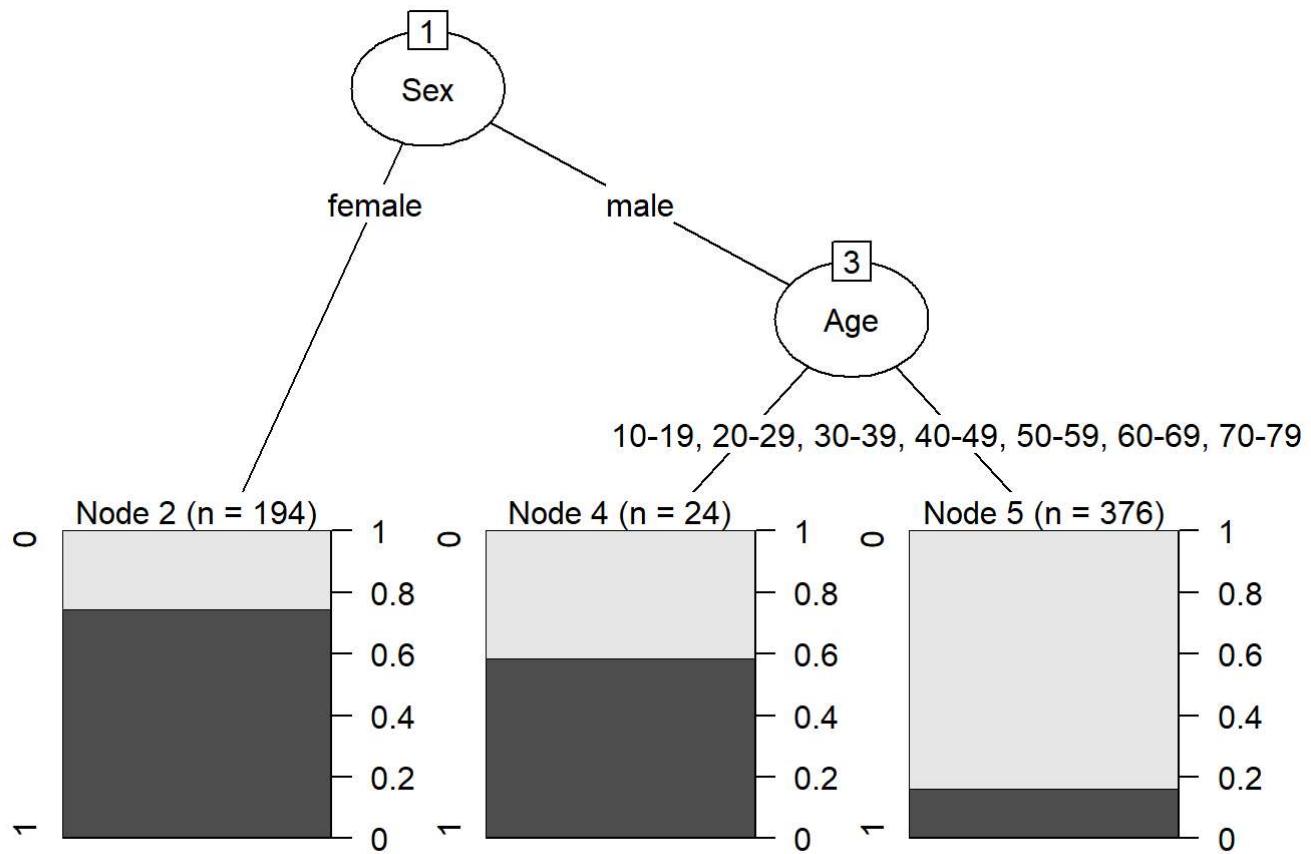
```
##  
## Call:  
## C5.0.default(x = trainX, y = trainy, trials = 10)  
##  
##  
## C5.0 [Release 2.07 GPL Edition]      Tue Jan 04 19:06:36 2022  
## -----  
##  
## Class specified by attribute `outcome'  
##  
## Read 594 cases (4 attributes) from undefined.data  
##  
## ----- Trial 0: -----  
##  
## Decision tree:  
##  
## Sex = female: 1 (194/50)  
## Sex = male:  
## :...Age = 0-9: 1 (24/10)  
##     Age in {10-19,20-29,30-39,40-49,50-59,60-69,70-79}: 0 (376/60)  
##  
## ----- Trial 1: -----  
##  
## Decision tree:  
##  
## Pclass in {1,2}: 1 (257.4/107.1)  
## Pclass = 3: 0 (336.6/82.5)  
##  
## ----- Trial 2: -----  
##  
## Decision tree:  
##  
## Sex = female: 1 (195.7/77.8)  
## Sex = male: 0 (398.3/114.7)  
##  
## ----- Trial 3: -----  
##  
## Decision tree:  
##  
## Pclass = 1: 1 (155.9/63.3)  
## Pclass in {2,3}: 0 (438.1/155.7)  
##  
## ----- Trial 4: -----  
##  
## Decision tree:  
##  
## Sex = female: 1 (205.6/89.5)  
## Sex = male: 0 (388.4/148.8)  
##  
## ----- Trial 5: -----  
##  
## Decision tree:
```

```
##  
## Pclass = 3: 0 (324/126.5)  
## Pclass in {1,2}:  
## ....Sex = female: 1 (74.4/13)  
##      Sex = male: 0 (195.7/84.9)  
##  
## ----- Trial 6: -----  
##  
## Decision tree:  
##  
## Pclass in {1,2}: 1 (296.5/120.1)  
## Pclass = 3: 0 (272.5/88.9)  
##  
## ----- Trial 7: -----  
##  
## Decision tree:  
##  
## Sex = female: 1 (217/74.6)  
## Sex = male: 0 (342/108.3)  
##  
## ----- Trial 8: -----  
##  
## Decision tree:  
##  
## Pclass in {1,2}: 1 (287.3/124.4)  
## Pclass = 3: 0 (265.7/79.6)  
##  
## ----- Trial 9: -----  
##  
## Decision tree:  
##  
## Sex = female: 1 (257.3/113.7)  
## Sex = male: 0 (266.7/26.8)  
##  
##  
## Evaluation on training data (594 cases):  
##  
## Trial      Decision Tree  
## -----  
##      Size      Errors  
##  
##      0        3  120(20.2%)  
##      1        2  199(33.5%)  
##      2        2  124(20.9%)  
##      3        2  194(32.7%)  
##      4        2  124(20.9%)  
##      5        3  122(20.5%)  
##      6        2  199(33.5%)  
##      7        2  124(20.9%)  
##      8        2  199(33.5%)  
##      9        2  124(20.9%)  
## boost          122(20.5%)    <<
```

```
##
##      (a)      (b)    <-classified as
##      ----  -----
##      371      5      (a): class 0
##      117     101     (b): class 1
##
##
## Attribute usage:
##
## 100.00% Pclass
## 100.00% Sex
## 67.34% Age
##
##
## Time: 0.0 secs
```

Pel que podem observar, en les 10 observacions que s'ha fet ha utilitzat el 100% de l'atribut Pclass, el 100% de Sex i el 67.34% de Age.

```
plot(model)
```



Utilitzem el conjunt de test per calcular la precisió.

```

predicted_model <- predict( model, testX, type="class" )
print(sprintf("La precisió de l'arbre és: %.4f %",100*sum(predicted_model == testy) / length(predicted_model)))

## [1] "La precisió de l'arbre és: 77.1044 %"

```

Una precisió del 77.1044% és una bona precisió, però s'hauria d'estudiar si seria possible millorant-la utilitzant altres atributs o creant-ne de nous.

Per acabar, mostrarem també la matriu de confusió.

```

mat_conf<-table(testy,Predicted=predicted_model)
mat_conf

##      Predicted
## testy    0   1
##      0 169   4
##      1   64  60

```

## Conclusions

En aquest treball s'ha analitzat un conjunt de dades sobre els passatgers del Titanic. El que s'ha volgut estudiar és si els diferents atributs que conté tenen una relació amb l'atribut de Survived que és un atribut categòric que compren dos valors 0 i 1, segons si els passatgers van sobreviure al Titanic o no.

Després de netejar i fer una primera anàlisi de les dades, s'ha arribat a la conclusió que els atributs que més relació tenien amb l'atribut Survived han sigut: Age, Sex i Pclass. A Age tenim l'edat dels passatgers, a Sex tenim el gènere dels passatgers i a Pclass, la classe en la qual viatjaven.

Amb les diferents anàlisis hem pogut observar que van viatjar molts més homes que dones, però un 74,2% de dones van sobreviure enfront d'un 18,8% d'homes.

També s'ha pogut observar que el percentatge de gent que va sobreviure en primera classe és pràcticament el doble que de gent no-sobrevivent. Per contra, en tercera classe el percentatge de passatgers que van morir és del triple que de passatgers que van sobreviure.

La majoria de passatgers tenien entre 20 i 29 anys. Els nens i nenes de primera i segona classe van sobreviure gairebé tots, en canvi, més del 50% dels nens i nenes que viatjaven en tercera classe van morir.

Finalment, hem creat un model de classificació utilitzant un arbre de decisió, per revisar si amb els atributs Sex, Age i Pclass seria possible predir l'atribut Survived. Hem obtingut un 77.1044 % de precisió.

Per tant, i per concloure, podem confirmar que els atributs Sex, Age i Pclass són atributs fonamentals per poder predir si un pacient va poder sobreviure o no al Titanic.

## Contribucions

**Investigació prèvia** MSM, OBR

**Redacció de les respostes** MSM, OBR

**Desenvolupament del codi MSM, OBR**

# Recursos

# Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.