Tipologia i cicle de vida de les dades: PAC 2

Autor: Albert Porteros

Juny 2020

Contents

Introducció **1**

Presentació 1

Competències 2

Objectius 2

Descripció de la PRA a realitzar 2

Recursos 3

Resolució **3**

Descripció del dataset 3

Objectius de l'analisi 4

Processos de neteja del joc de dades 4

Processos d'anàlisis del joc de dades 5

Selecció de variables i creació d'un arbre de decisions 13

Conclusions 15

Taula final 16

Link a GitHub 16

Introducció

Presentació

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes. Per fer aquesta pràctica haureu de treballar en grups de 2 persones.

Haureu de lliurar un sol fitxer amb l'enllaç Github (https://github.com) on hi hagi les solucions incloent els noms dels components de l'equip. Podeu utilitzar la Wiki de Github per descriure el vostre equip i els diferents arxius que corresponen la vostra entrega. Cada membre de l'equip haurà de contribuir amb el seu usuari Github. Podeu utilitzar aquests exemples com guia:

Competències

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi.

Objectius

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.
- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

Descripció de la PRA a realitzar

L'objectiu d'aquesta activitat serà el tractament d'un dataset, que pot ser el creat a la pràctica 1 o bé qualsevol dataset lliure disponible a Kaggle (https://www.kaggle.com). Alguns exemples de dataset amb els que podeu treballar són:

- Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009).
- Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic).

L'últim exemple correspon a una competició activa a Kaggle de manera que, opcionalment, podeu aprofitar el treball realitzat durant la pràctica per entrar en aquesta competició. Seguint les principals etapes d'un projecte analític, les diferents tasques a realitzar (i justificar) són les següents:

- Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?
- Integració i selecció de les dades d'interès a analitzar.
- Neteja de les dades.
  - Les dades contenen zeros o elements buits? Com gestionaries aquests casos?
  - Identificació i tractament de valors extrems.
- Anàlisi de les dades.
  - Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).
  - Comprovació de la normalitat i homogeneïtat de la variància.
  - Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

- Representació dels resultats a partir de taules i gràfiques.
- Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?
- Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

Recursos

Els següents recursos són d'utilitat per la realització de la pràctica:

- Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
- Megan Squire (2015). Clean Data. Packt Publishing Ltd.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.
- Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
- Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.
- Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.
- Tutorial de Github https://guides.github.com/activities/hello-world.

Resolució

Descripció del dataset

El dataset està compost per un total de 12 columnes (atributs) i 418 files en el cas del fitxer test i 891 en el fitxer de training (Persones en el Titanic). Les columnes són les següents:

- PassengerId: Type Integer. Definition: Es tracta d'un número únic per cada passatger.
- Survived: Type Binary. Definition: 0 significa que no ha sobreviscut, 1 significa que va sobreviure.
- Pclass: Type Factor. Definition: Significa la classe del ticket que s'ha comprat, sent 1 el més car i 3 el més barat.
- Sex: Type Char. Definition: Sexe del passatger, male o female.
- Age: Type num. Definition: Edat en format númeric del passatger.
- SibSp: Type num. Definition: Nombre de familiars del mateix rang d'edat (germans o marit/dona) que es trobaven en el vaixell.
- Parch: Type num. Definition: Nombre de fills/es i/o pares que es trobaven en el vaixell.
- Ticket: Type char. Definition: Número de billet.
- Fare: Type num. Definition: Tarifa.
- Cabin: Type char. Definition: Número de cabina.
- Embarked: Type char. Definition: Port en el qual va embarcar. C = Cherbourg, Q = Queenstown, S = Southampton.

Objectius de l'analisi

Amb aquest anàlisi volem respondre a una pregunta tant simple com si el tipus de ticket que van comprar, el sexe, l'edat o la cantitat de familiars al Titanic són factors que influencien en si la persona en qüestió va morir o no. Per tal de realitzar aqueust anàlisi juntarem el fitxer de test i training en un sol.

També crearem un arbre de decisions per tal de veure la contribució de cadascuna de les variables i poder preveure si una persona viura o no depenent de les seves condicions.

Processos de neteja del joc de dades

Primer contacte amb el joc de dades, visualitzem la seva estructura.

```
# Carreguem els paquets R que utilitzarem
library(ggplot2)
library(dplyr)

# Guardem el joc de dades test i train en un únic dataset
test <- read.csv('test.csv',stringsAsFactors = FALSE)
train <- read.csv('train.csv', stringsAsFactors = FALSE)

# Unim els dos jocs de dades en un només
totalData <- bind_rows(train,test)
filas=dim(train)[1]

# Verifiquem l'estructura del joc de dades
str(totalData)
```

```
## 'data.frame':    1309 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 Thayer)"
##  $ 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" ...
```

Treballem els atributs amb valors buits.

```
# Estadístiques de valors buits
colSums(is.na(totalData))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0         418           0           0           0         263
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           1           0           0
```

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

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0          NA           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0          NA        1014           2
```

```
# Prenem valor "C" per als valors buits de la variable "Embarked"
totalData$Embarked[totalData$Embarked==""]="C"

# Prenem la mitjana per a valors buits de la variable "Age"
totalData$Age[is.na(totalData$Age)] <- mean(totalData$Age,na.rm=T)
```

Discretitzem quan té sentit i en funció de cada variable.

```
# Per a quines variables tindria sentit un procés de discretització?
apply(totalData,2, function(x) length(unique(x)))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##        1309           3           3        1307           2          99
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           7           8         929         282         187           3
```

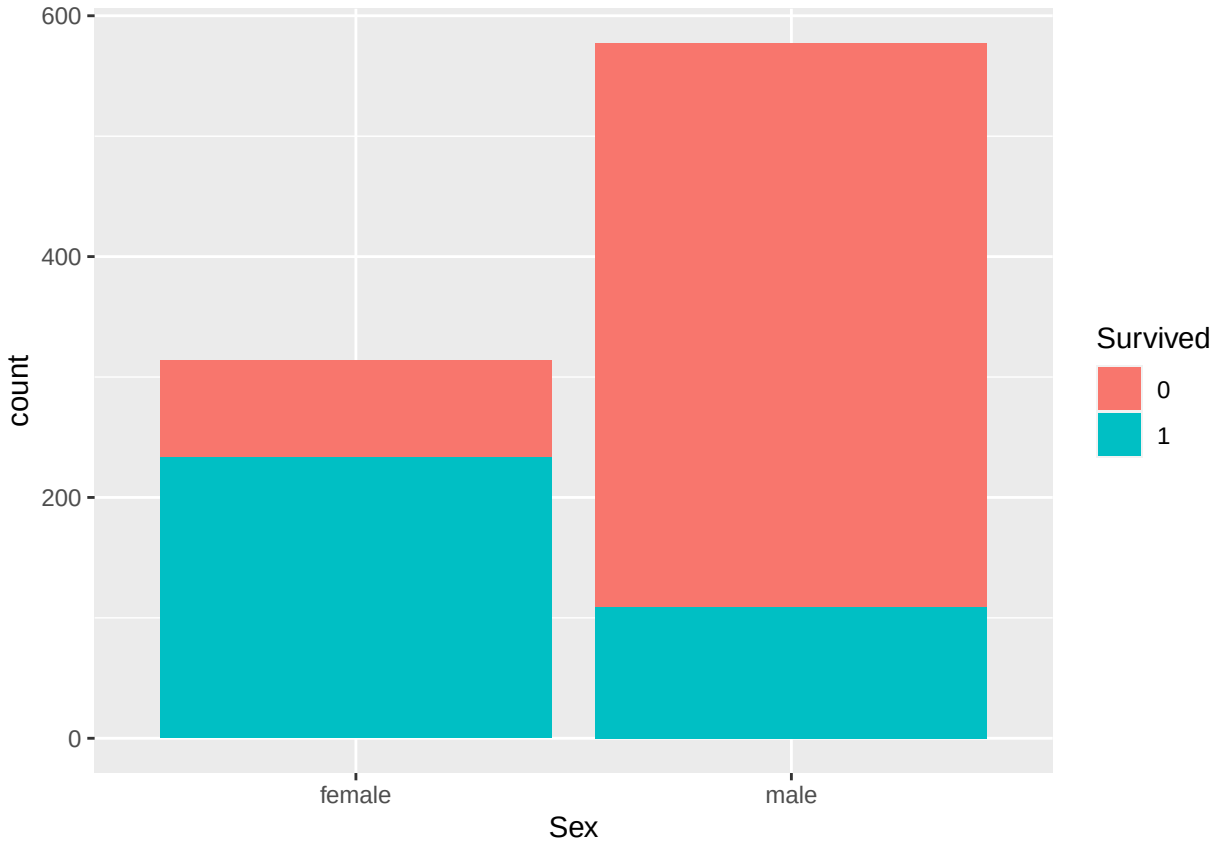
```
# Discretitzem les variables amb poques classes
cols<-c("Survived","Pclass","Sex","Embarked")
for (i in cols){
  totalData[,i] <- as.factor(totalData[,i])
}

# Després dels canvis, analitzem la nova estructura del joc de dades
str(totalData)
```

```
## 'data.frame':    1309 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 ...
##  $ 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   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
```

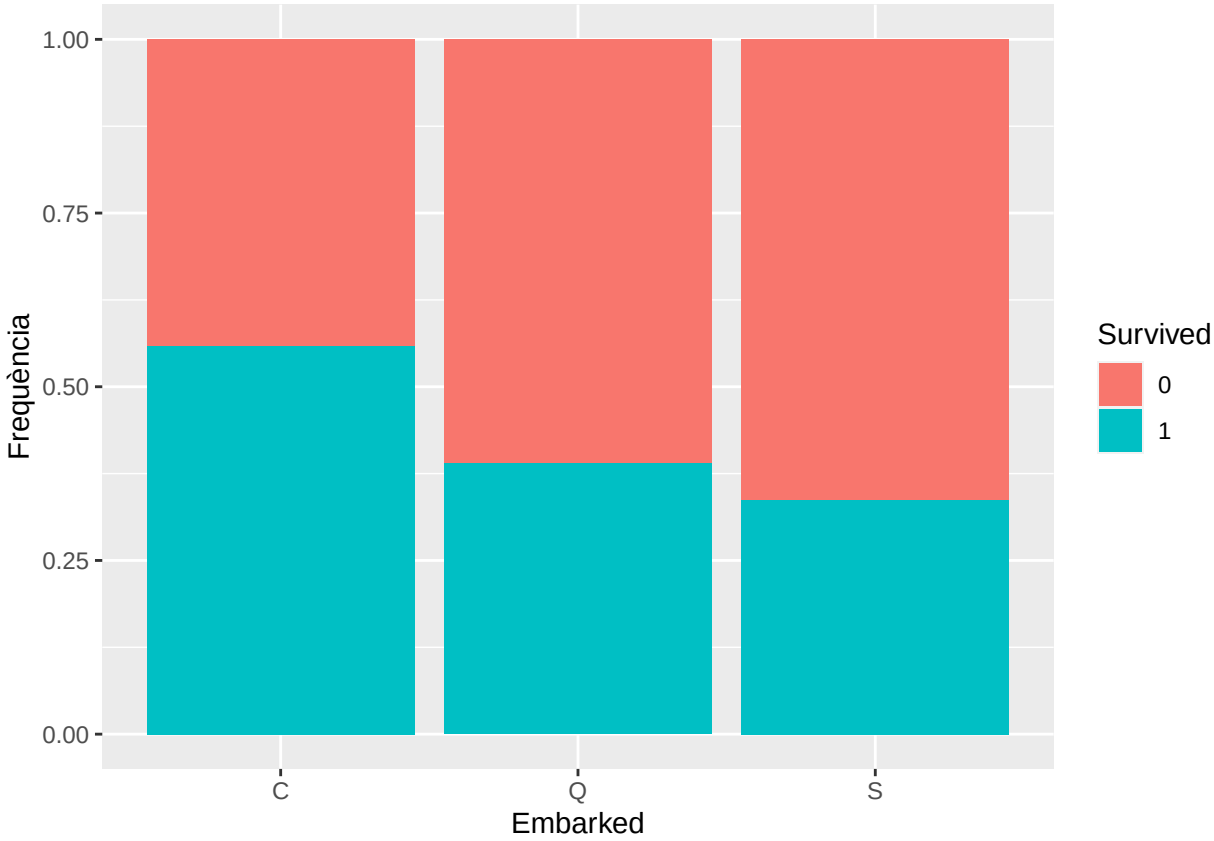
Processos d'anàlisis del joc de dades

Ens proposem analitzar les relacions entre les diferents variables del joc de dades.

```
# Visualitzem la relació entre les variables "sex" i "survival":
ggplot(data=totalData[1:filas,],aes(x=Sex,fill=Survived))+geom_bar()
```




```
# Un altre punt de vista. Survival com a funció de Embarked:
ggplot(data = totalData[1:filas,],aes(x=Embarked,fill=Survived))+geom_bar(position="fill")+ylab("Frequè
```



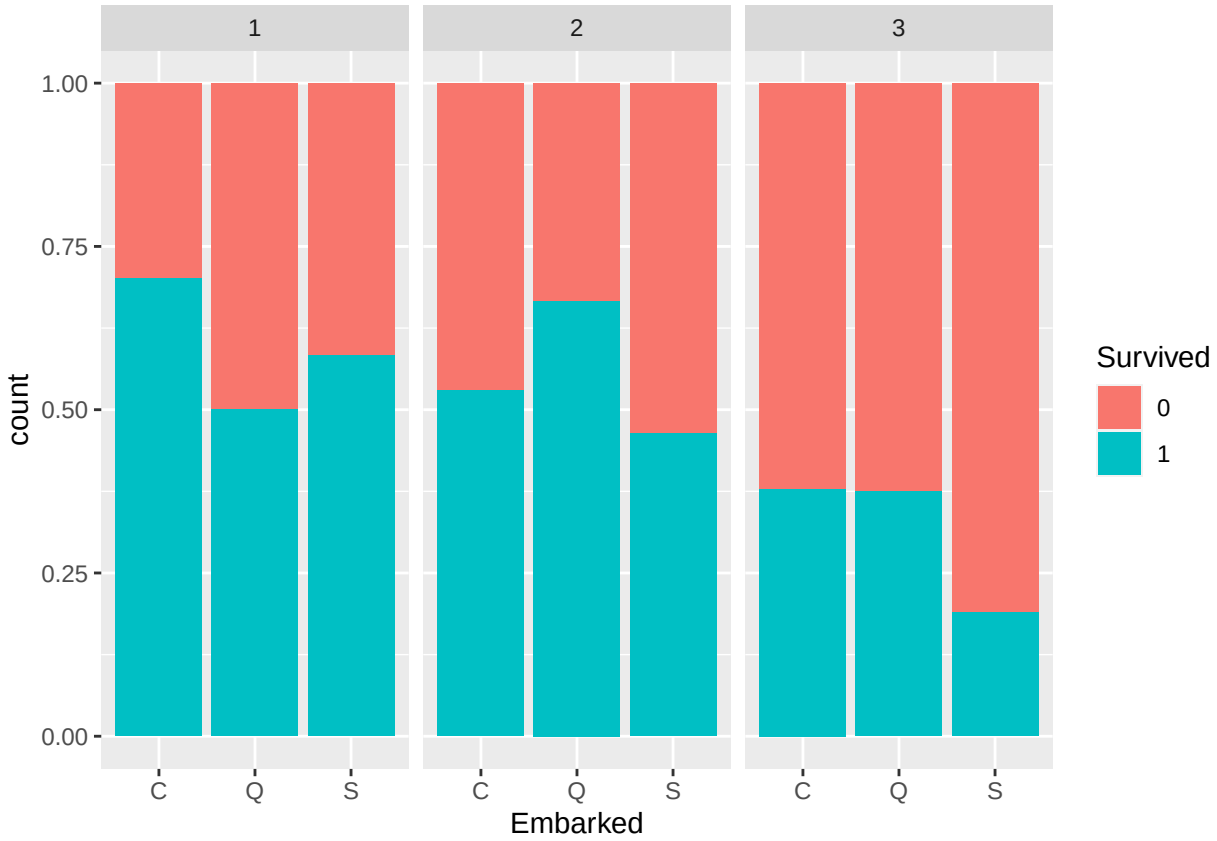
Obtenim una matriu de percentatges de freqüència. Veiem, per exemple que la probabilitat de sobreviure si es va embarcar en "C" és d'un 55,88%

```
t<-table(totalData[1:filas,]$Embarked,totalData[1:filas,]$Survived)
for (i in 1:dim(t)[1]){
    t[i,]<-t[i,]/sum(t[i,])*100
}
t
```

```
##
##            0        1
##   C 44.11765 55.88235
##   Q 61.03896 38.96104
##   S 66.30435 33.69565
```

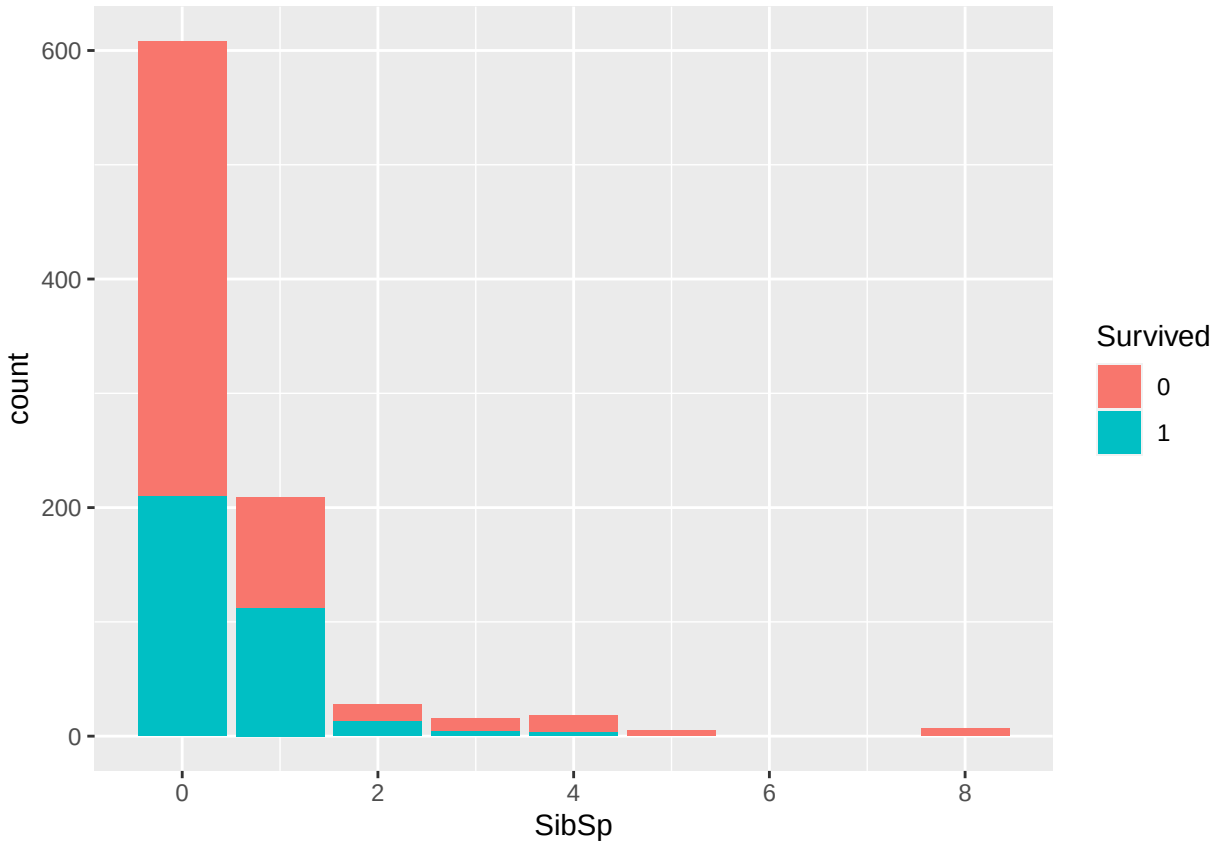
Vegem ara com en un mateix gràfic de freqüències podem treballar amb 3 variables: Embarked, Survived i Pclass.

```
# Mostrem el gràfic d'embarcats per Pclass:
ggplot(data = totalData[1:filas,],aes(x=Embarked,fill=Survived))+geom_bar(position="fill")+facet_wrap(~
```

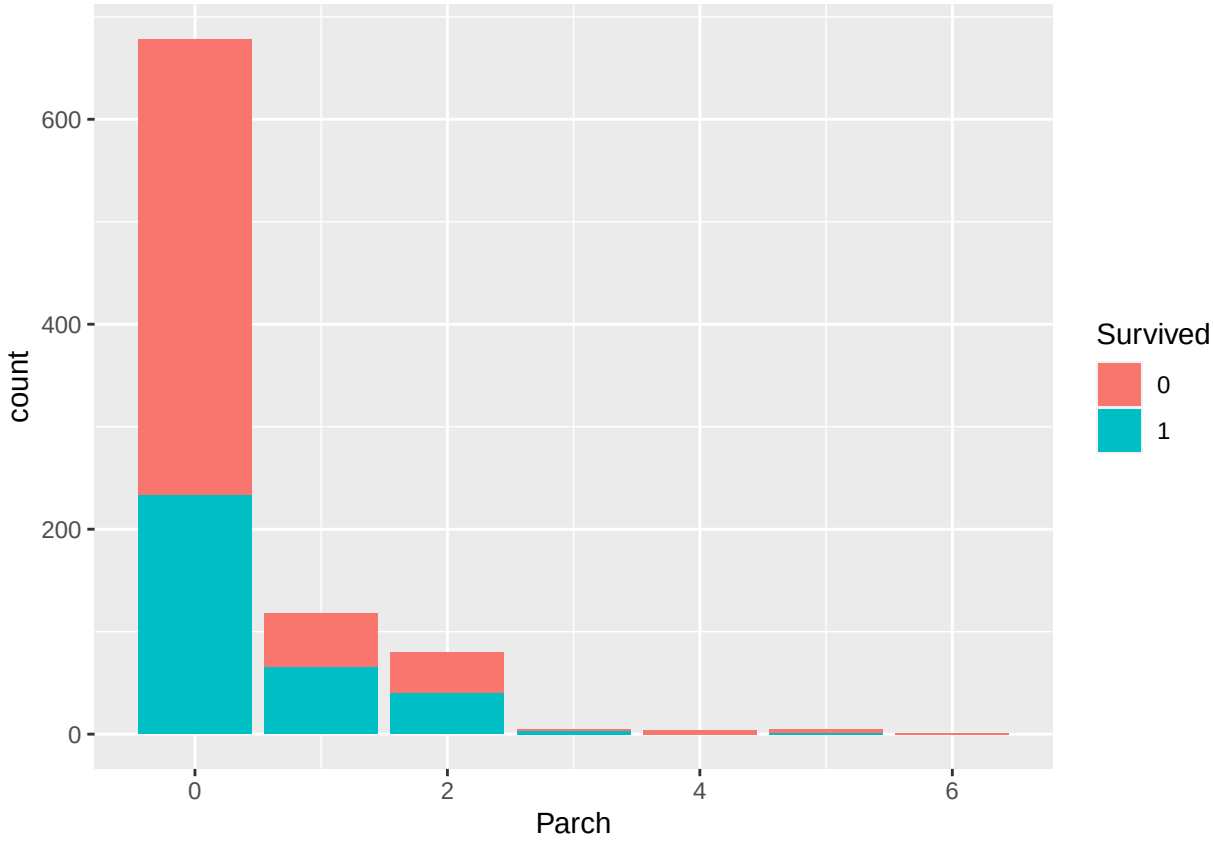


Comparem ara dos gràfics de freqüències: Survived-SibSp i Survived-Parch

```
# Survival com a funció de SibSp i Parch
ggplot(data = totalData[1:filas,],aes(x=SibSp,fill=Survived))+geom_bar()
```



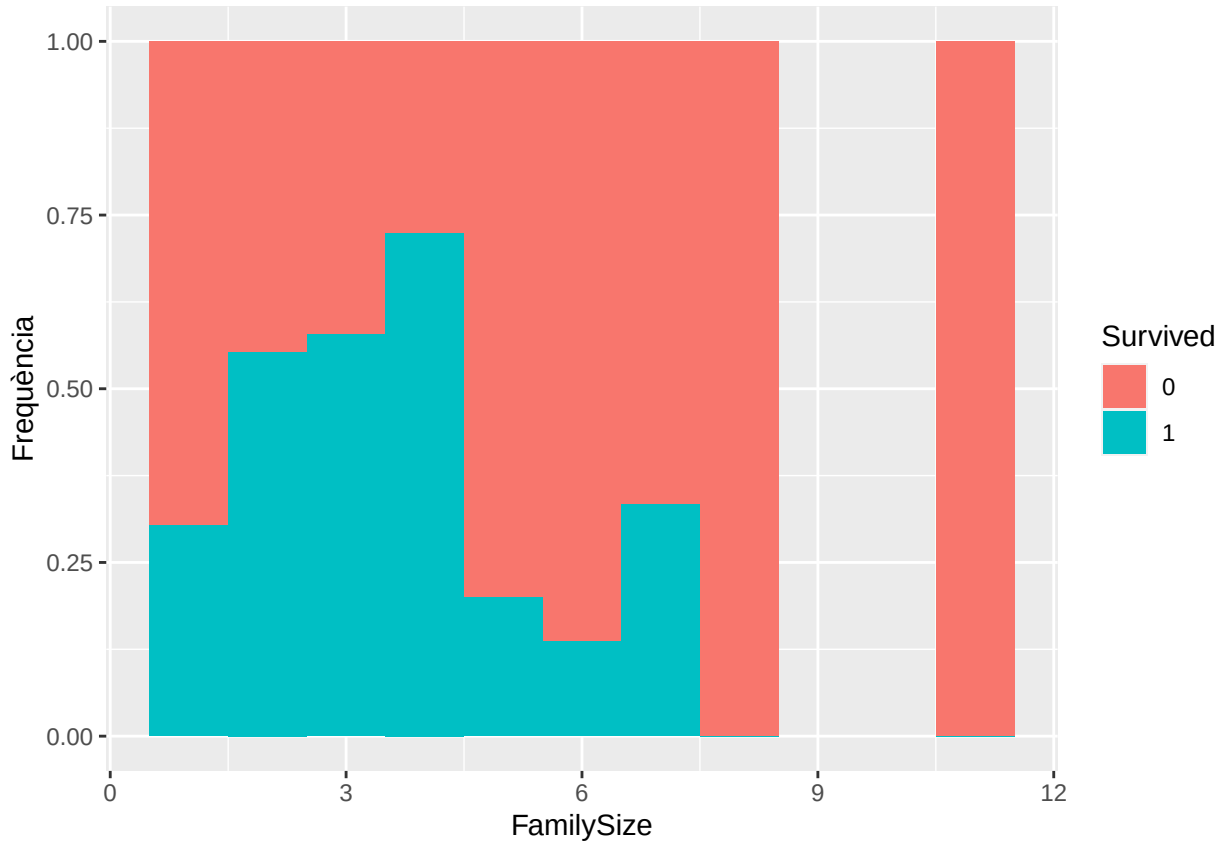

```
ggplot(data = totalData[1:filas,],aes(x=Parch,fill=Survived))+geom_bar()
```




```
# Veiem com la forma d'aquests dos gràfics és similar. Aquest fet ens pot indicar presència de correlaci
```

Vegem un exemple de construcció d'una variable nova: Grandària de família

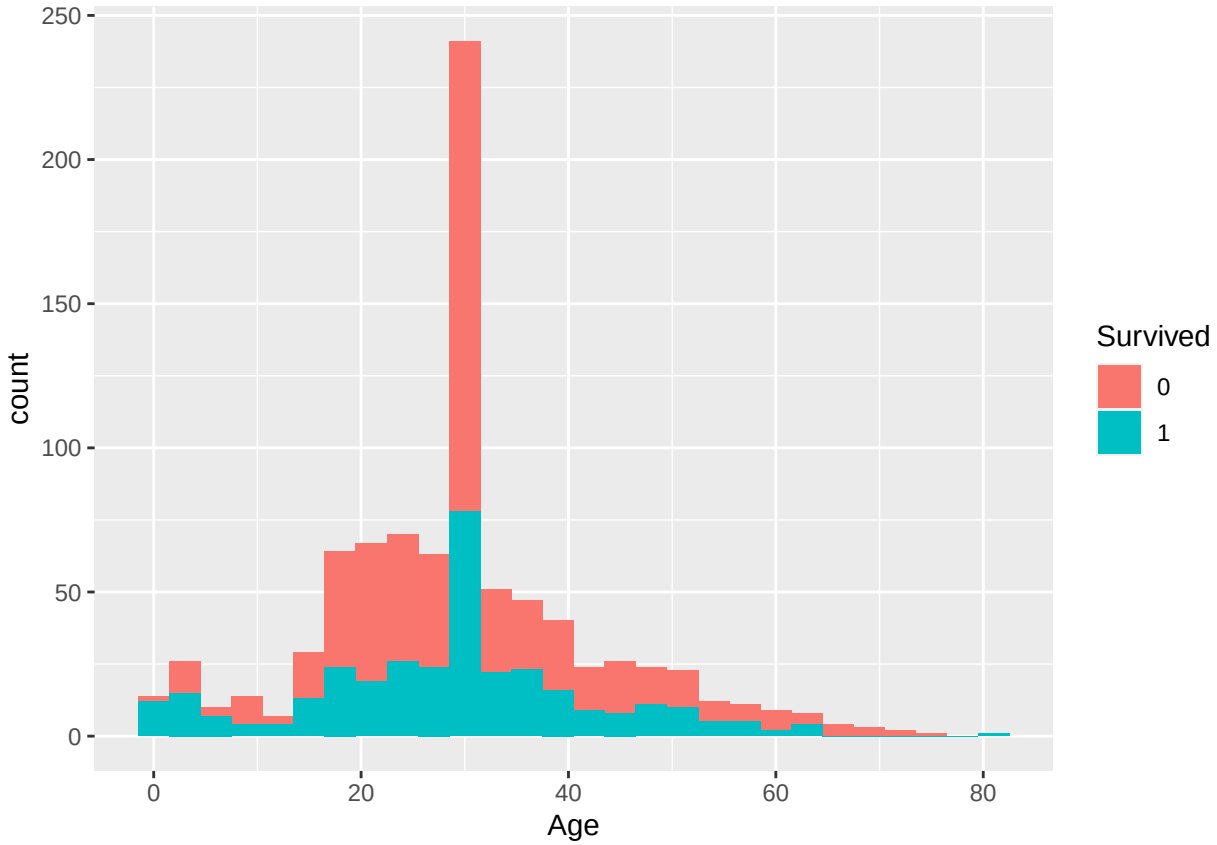
```
# Construïm un atribut nou: family size..
totalData$FamilySize <- totalData$SibSp + totalData$Parch +1;
totalData1<-totalData[1:filas,]
ggplot(data = totalData1[!is.na(totalData[1:filas,]$FamilySize),],aes(x=FamilySize,fill=Survived))+geom_
```




```
# Observem com les famílies de 2 a 4 membres tenen més del 50% de possibilitats de supervivència.
```

Vegem ara dos gràfics que ens compara els atributs Age i Survived. Observem com el paràmetre position="fill" ens dóna la proporció acumulada d'un atribut dins d'un altre

```
# Survival com a funció de age:
ggplot(data = totalData1[!(is.na(totalData[1:filas,]$Age)),],aes(x=Age,fill=Survived))+geom_histogram(b
```




```
ggplot(data = totalData1[!is.na(totalData[1:filas,]$Age),],aes(x=Age,fill=Survived))+geom_histogram(bin
```

Selecció de variables i creació d'un arbre de decisions

A continuació seleccionarem els 4 atributs que creiem més importants i decisius de cara a saber si un tripulant va morir o no, que son Pclass, Sex, Age i FamilySize.

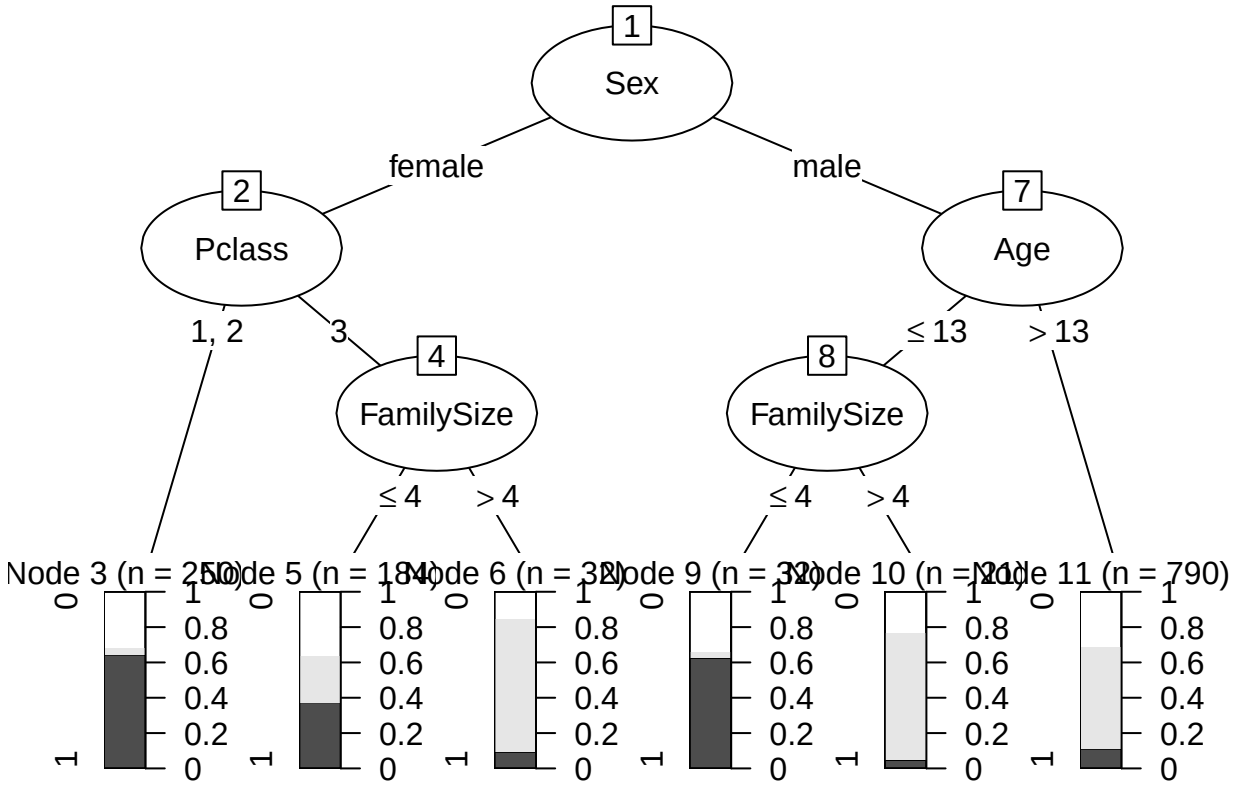
```
library(C50)
DecisionTree <- totalData %>% select(3,5,6,13,2)
y <- DecisionTree[,5]
X <- DecisionTree[,-5]
model <- C50::C5.0(X, y, rules=TRUE )
summary(model)
```

```
##
## Call:
## C5.0.default(x = X, y = y, rules = TRUE)
##
##
## C5.0 [Release 2.07 GPL Edition]      Tue Jun 09 21:11:55 2020
## -----------------------------
##
## Class specified by attribute 'outcome'
## *** ignoring cases with bad or unknown class
##
## Read 891 cases (5 attributes) from undefined.data
```

```
##
## Rules:
##
## Rule 1: (54/4, lift 1.5)
##  Pclass = 3
##  FamilySize > 4
##  ->  class 0  [0.911]
##
## Rule 2: (540/88, lift 1.4)
##  Sex = male
##  Age > 13
##  ->  class 0  [0.836]
##
## Rule 3: (170/9, lift 2.5)
##  Pclass in {1, 2}
##  Sex = female
##  ->  class 1  [0.942]
##
## Rule 4: (21/1, lift 2.4)
##  Sex = male
##  Age <= 13
##  FamilySize <= 4
##  ->  class 1  [0.913]
##
## Rule 5: (281/57, lift 2.1)
##  Sex = female
##  FamilySize <= 4
##  ->  class 1  [0.795]
##
## Default class: 0
##
##
## Evaluation on training data (891 cases):
##
##          Rules
##    ----------------
##      No      Errors
##
##       5  150(16.8%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##     491    58    (a): class 0
##      92   250    (b): class 1
##
##
##  Attribute usage:
##
##   95.17% Sex
##   62.96% Age
##   39.96% FamilySize
##   25.14% Pclass
##
```

```
##
## Time: 0.0 secs
```

```
model <- C50::C5.0(X, y)
plot(model)
```




```
write.csv(totalData, file = "ResultatPRA2.csv", row.names = FALSE)
```

Per tal d'entendre aquest decision tree cal tenir clar que el estat final de 0 vol dir que no sobreviu, mentre que el 1 vol dir que si que sobreviu.Aquest arbre de decisions té una precisió del 83.2%.

Conclusions

Tal i com observem en l'arbre de decisió, la conclusió a la qual arribem es la següent:

- Els homes van tenir més posibilitats de morir que les dones. La principal resposta que trobem aquí és que a les dones les debian de pujar als botes salvavidas per tal que cuidesin dels nens.
- Si la familia era menor a 4 integrants, tenien moltes més posibilitats de sobreviure que si la familia tenia més de 4 integrants. Suposem que això passava ja que no volien deixar a ningú de la familia enrere i aconseguir més de 4 places en un bot salvavides era molt difícil.
- Els infants tenien més posibilitat de sobreviure que els majors de 13 anys. Això es degut a que a la hora de decidir qui es salvaria prioritzaven infants i gent gran.
- Sorprenentment, la situació economica també tenia un impacte. El rang 1 i 2 haviem comentat que eren els més costosos i per tant representen la població amb més poder econòmic.

Taula final

```
Contribucions = c("Investigació prèvia", "Redacció de les respostes ", "Desenvolupament codi ")
Firma = c("Albert Porteros Villar", "Albert Porteros Villar","Albert Porteros Villar")
laTabla = data.frame (cbind(Contribucions,Firma))
knitr::kable(laTabla)
```

Contribucions	Firma
Investigació prèvia	Albert Porteros Villar
Redacció de les respostes	Albert Porteros Villar
Desenvolupament codi	Albert Porteros Villar

Link a GitHub

El treball final es pot trobar a https://github.com/Albert-Porteros/PRA2