Neteja i anàlisis de les dades

Autor: Noé Sanchez Soldevila, Gemma Garcia de la Fuente

Desembre 2019

Introducció

Les dades que utilitzarem per a la segona pràctica de l'assignatura: Tipologia i cicle de vida de les dades, són les dades del dataset Titanic de Kaggle (https://www.kaggle.com/c/titanic/data).

L'objectiu del present treball serà la neteja i l'anàlisi del dataset per arribar a conclusions sobre la influència de les diferents característiques descrites en ell sobre la supervivència dels individus en l'accident del Titànic, així com la creació d'un model predictiu que d'acord amb aquestes característiques sigui capaç de preveure si l'individu sobreviuria o no a l'accident.

Descripció del dataset.

Perquè és important i quina pregunta/problema pretén respondre?

El data set recull un conjunt de variables referents als passatgers del titànic, a continuació podem veure

```
# Guardem el joc de dades test i train en un únic dataset
test <- read.csv('titanic-test.csv',stringsAsFactors = FALSE)
train <- read.csv('titanic-train.csv', stringsAsFactors = FALSE)
# Unim els dos jocs de dades en un només
# Verifiquem l'estructura del joc de dades
pander::pander(str(train))
```

'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 Thayer)" "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" ...

Veiem que tenim un total de 12 variables, aquestes tenen diferents formats: enters, numèrics i caràcter. Les variables codificades om enteres algunes en realitat representen factors, aplicarem les transformacions necessàries perquè cada variable pugui ser analitzada de la forma més adient:

```
train$Survived <- as.factor(train$Survived)
train$Pclass <- as.factor(train$Pclass)
train$Sex <- as.factor(train$Sex)
train$SibSp <- as.factor(train$SibSp)
train$Parch <- as.factor(train$Parch)
train$Embarked <- as.factor(train$Embarked)
summary(train)
```

```
##   PassengerId    Survived Pclass      Name              Sex
##  Min.   :  1.0   0:549    1:216   Length:891         female:314
##  1st Qu.:223.5   1:342    2:184   Class :character   male  :577
##  Median :446.0            3:491   Mode  :character
##  Mean   :446.0
##  3rd Qu.:668.5
```

```
##  Max.   :891.0
##
##       Age        SibSp   Parch      Ticket              Fare
##  Min.   : 0.42   0:608   0:678   Length:891         Min.   :  0.00
##  1st Qu.:20.12   1:209   1:118   Class :character   1st Qu.:  7.91
##  Median :28.00   2: 28   2: 80   Mode  :character   Median : 14.45
##  Mean   :29.70   3: 16   3:  5                      Mean   : 32.20
##  3rd Qu.:38.00   4: 18   4:  4                      3rd Qu.: 31.00
##  Max.   :80.00   5:  5   5:  5                      Max.   :512.33
##  NA's   :177     8:  7   6:  1
##     Cabin           Embarked
##  Length:891          :  2
##  Class :character   C:168
##  Mode  :character   Q: 77
##                     S:644
##
##
##
```

Podem comprovar els diferents valors estadístics de les variables contínues i els recomptes dels valors categòrics, a més de la quantitat de valors nuls que conté cada variable. Trobem valors nuls en la variable objectiu, Survived, pel que a continuació haurem d'eliminar-los dela nostra anàlisi.

Integració i selecció de les dades d'interès a analitzar.

En ser una base de dades no molt amplia, decidim mantenir tots els atributs disponibles. La variable resposta la qual volem intentar predir amb les anàlisis és la variable Survived, així doncs el que volem semblar és si amb les altres variables som capaços de predir la supervivència dels subjectes.

Neteja de les dades.

En primer lloc mirem si tenim alguna variable amb valors mancants:

```
colSums(is.na(train))
```

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

Veiem que la variable Agete valors mancants. L'únic que considerem necessari respecte la depuració de dades és tractar amb els valors mancants, ja que això pot influir en la nostra anàlisi. Per fer-ho adoptem dos criteris diferents en funció de si el valor mancant està en la variable resposta o la variable explicativa.

- Variable Survived és la variable resposta, si hi hagues missings, no els tractaríem, eliminem tupla.
- Missings vairbales explicatives: com sabem els missings poden tenir diversos orígens. En aquest cas considerem que es tracte d'un cas MAR: valors mancants que depenen de coses que hem observat. Per tant aplicarem la tècnica de Múltiple imputació a través de la funció `mice()`

```
train<- mice(train)
```

```
##
##  iter imp variable
##   1   1  Age
##   1   2  Age
```

```
##   1   3  Age
##   1   4  Age
##   1   5  Age
##   2   1  Age
##   2   2  Age
##   2   3  Age
##   2   4  Age
##   2   5  Age
##   3   1  Age
##   3   2  Age
##   3   3  Age
##   3   4  Age
##   3   5  Age
##   4   1  Age
##   4   2  Age
##   4   3  Age
##   4   4  Age
##   4   5  Age
##   5   1  Age
##   5   2  Age
##   5   3  Age
##   5   4  Age
##   5   5  Age
```

```
train <- complete(train)
```

*Embarked: Omplim la variable Embarked amb el valor C

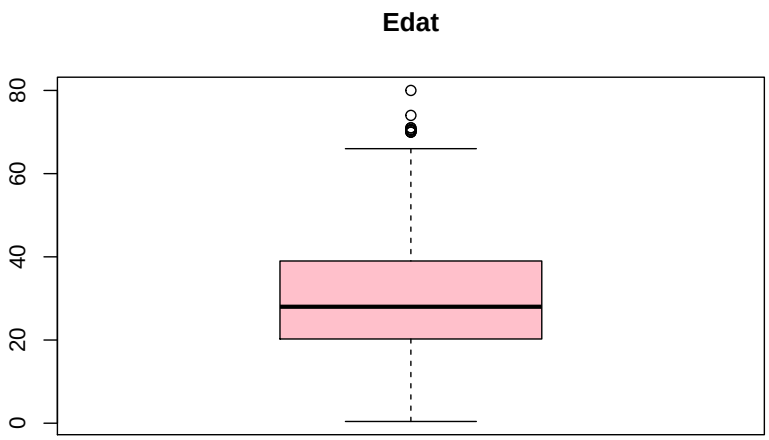
```
train$Embarked[train$Embarked==""]="C"
```

Identificació i tractament de valors extrems.

En aquesta secció analitzem la presència de valors extrems, per fer-ho primer fem una representació gràfica mitjançant un boxplot de les variables numèriques.

- AGE

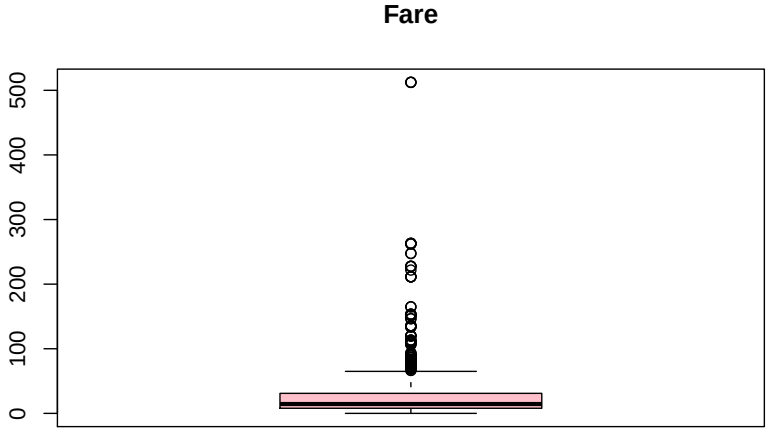
```
boxplot(train$Age, main="Edat", col="pink")
```



Podem veure clarament la presència de varius punts fora de la caixa, aquests serien considerats valors atípics. En aquest cas no es pot aplicar cap transformació, ja que no és traca de valors entrats malament, ni de mal codificats, encara que hi hagues poques persones amb edats superiors als 80 anys, la presència d'aquest era probable així que no fem res amb els valors atípics d'aquesta variable.

- FARE

```
boxplot(train$Fare, main="Fare", col="pink")
```



Amb la representació gràfica veiem que hi ha clarament un valor que sí que hauríem de considerar com anòmal. Aquest valor, molt fora de la resta, podria tractar-se d'un outlier. Encara així, el mantindrem, ja que el valor no manté tanta diferència amb el grup de valors superiors, i és podria tractar d'una taxa molt superior a la resta per la seva exclusivitat.

Anàlisi de les dades.

Selecció dels grups

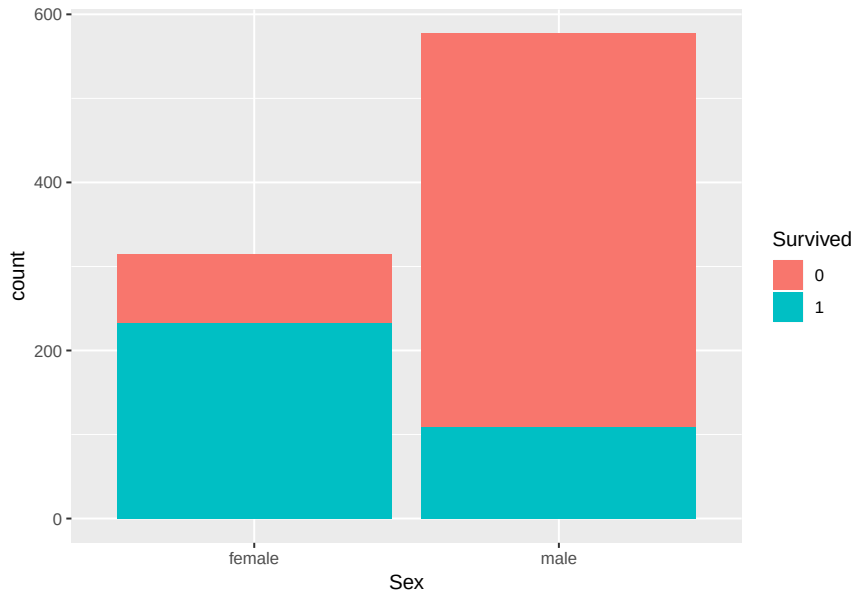
Seleccionem dos grups per a ser analitzat, la partició la fem en funció de si sobreviuen o no:

```
viu <- train[train$Survived == 1 ,]
mort <- train[train$Survived == 0,]
```

Fem una breu descripció gràfica bivariant amb la variable "survival" en enfront de algunes altres:

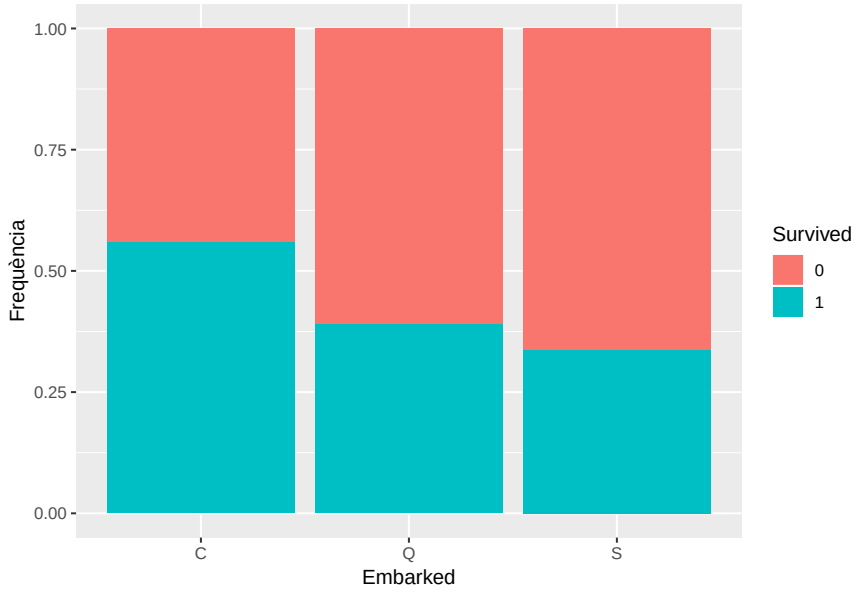
- Visualitzem la relació entre les variables "sex" i "survival":

```
filas=dim(train)[1]
ggplot(data=train[1:filas,],aes(x=Sex,fill=Survived))+geom_bar()
```



- Survival com a funció de Embarked:

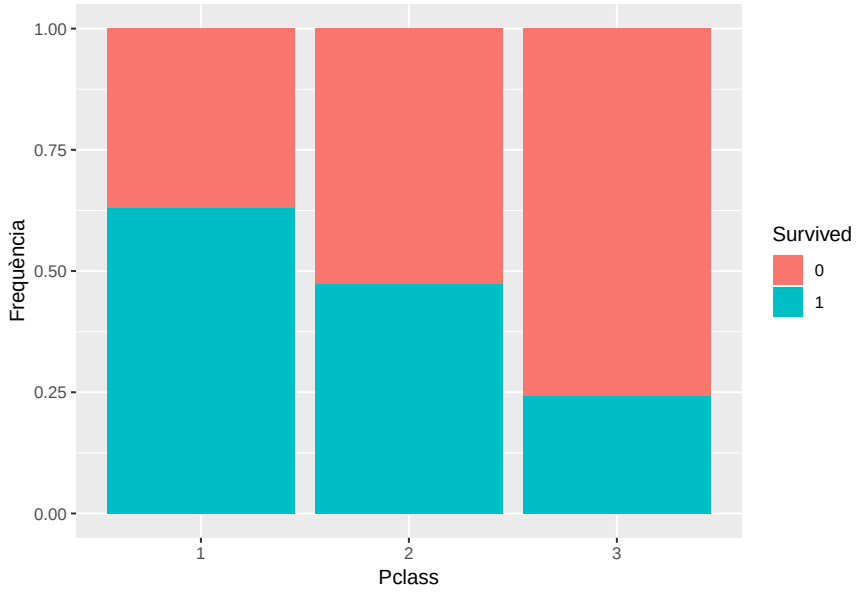
```
ggplot(data = train[1:filas,],aes(x=Embarked,fill=Survived))+geom_bar(position="fill")+ylab("Frequència
```



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%

- Survival com a funció de clase:

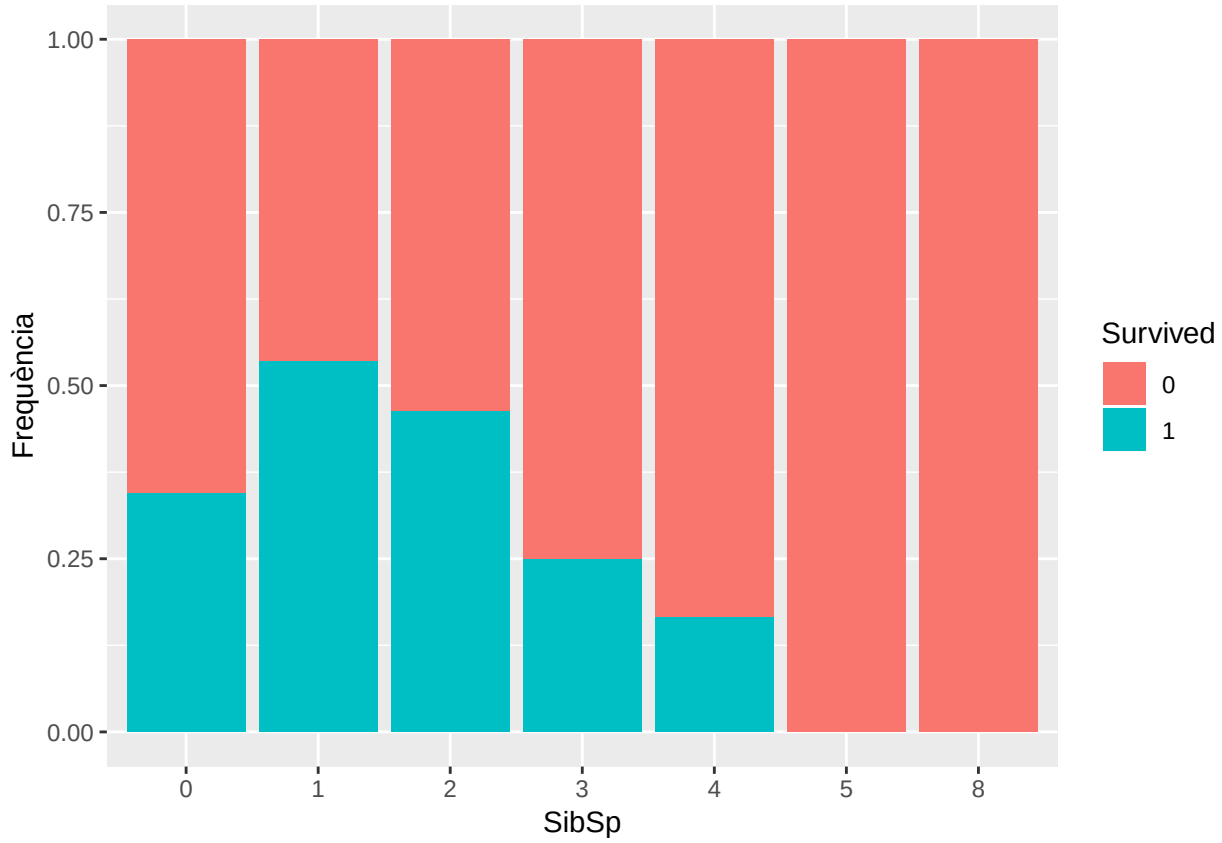
```
ggplot(data = train[1:filas,],aes(x=Pclass,fill=Survived))+geom_bar(position="fill")+ylab("Frequència")
```



Sembla que sovreviuen més els Subejctes de primera classe.

- suopervivencia en funció del nombre de germans

```
ggplot(data = train[1:filas,],aes(x=SibSp,fill=Survived))+geom_bar(position="fill")+ylab("Frequència")
```



Tot i que es curios com a relació semblaria que a ménys germans més possibilitats de sobreviure.

Comprovació de la normalitat i homogeneïtat de la variància.

Per analitzar la normalitat de les dades tenim diversos testos estadístics els quals busquen respondre el següent contrast d'hipòtesis:

Ho: normalitat H1: no normalitat

```
alpha = 0.05
col.names = colnames(train)
for (i in 2:ncol(train)) {
  if (i == 2) cat("Variables NO normals:\n")
    if (is.integer(train[,i]) | is.numeric(train[,i])) {
      p_val = shapiro.test(train[,i])$p.value
        if (p_val < alpha) {
            cat(col.names[i])
# Format output
if (i < ncol(train) - 1) cat(", ")
    if (i %% 3 == 0) cat("\n")
      }
  }
}
```

```
## Variables NO normals:
## Age,
## Fare,
```

Pel que fa a la homogeneitat de la variància:

Ho: hi ha homogeneïtat de variàncies H1: les variables són heteroblàstiques

Apliquem aquest test per a les variables numèriques, indicant que els grups que volem testar són els de la variables sobre vivència:

- Age:

```
library(car)
leveneTest(train$Age,train$Survived)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.9644 0.3263
##       889
```

Amb el p-valor > 0.05 viem que no tenim raons per rebutjar la hipòtesi nul · la i per tant considerem que la variància del grup de gent que va morir és igual a la de grup de gent que va sobreviure.

- Fare:

```
library(car)
leveneTest(train$Fare,train$Survived)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   1    45.1 3.337e-11 ***
##       889
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Amb el p-valor < 0.05 rebutgem la hipòtesi nul · la i per tant direm que no hi ha homogeneïtat de variàncies entre el grup de supervivents i els que moren en funció de la tarifa que van pagar.

Aplicació de proves estadístiques

Farem una comparació de mitjanes per comprovar si els dos grups definits són o no diferents:

- Edat, en aquest cas teníem homogeneïtat de variàncies per tant podem indicar-ho en el test, a més a més les dades no eren normals per tant hem d'aplicar un test no paramètric: Wilcoxon Rank sum Test. El contrast:

H0: igualtat de mitjanes m1: mitjanes diferents

```
wilcox.test(viu$Age,mort$Age)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  viu$Age and mort$Age
## W = 84952, p-value = 0.01685
## alternative hypothesis: true location shift is not equal to 0
```

Com que el p-valor és molt elevat direm que ambdós grups tenen mitjanes que es poden considerar iguals.

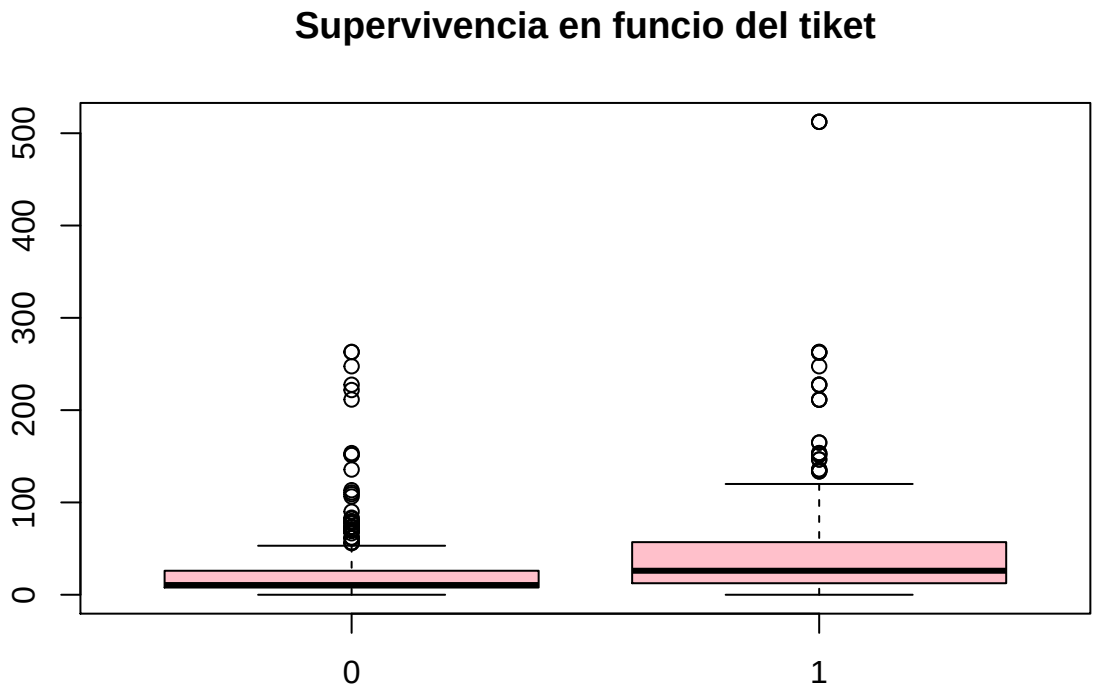
- Preu tiquet, la variable tampoc es considerava normal per tant apliquem el mateix test, i el mateix contrast:

```
wilcox.test(viu$Fare,mort$Fare)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  viu$Fare and mort$Fare
## W = 129950, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

En aquest cas les mitjanes d'ambdós grups no es poden considerar iguals, de fet si mirem una descriptiva bivariant, veurem que els que sobreviuen pagaven tarifes més altres:

```
boxplot(Fare~Survived,data=train, main="Supervivencia en funcio del tiket", col="pink")
```



Veiem en aquests boxplots que la població que sobreviu (1) són gent que va pagar un preu més elevat.

També podem fer les mateixes proves per variables categóriques del dataset:

En aquest cas:

H0: Les variables no tenen un efecte significatiu en el resultat. H1: Les variables sí el tenen.

- Comparem la supervivència amb el sexe:

```
chisq.test(train$Survived, train$Sex)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  train$Survived and train$Sex
## X-squared = 260.72, df = 1, p-value < 2.2e-16
```

Podem observar, amb un p valor molt inferior a 0.05, que el sexe va tenir una gran influència en la supervivència dels individus.

- Comparem la supervivència amb la classe:

```
chisq.test(train$Survived, train$Pclass)
```

```
##
##  Pearson's Chi-squared test
##
## data:  train$Survived and train$Pclass
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

També comprovem que la variable classe està molt relacionada amb la supervivència dels individus.

- Comparem la supervivència amb embarked:

```
chisq.test(train$Survived, train$Embarked)
```

```
##
##  Pearson's Chi-squared test
##
## data:  train$Survived and train$Embarked
## X-squared = 28.005, df = 2, p-value = 8.294e-07
```

També comprovem que la variable Embarked està molt relacionada amb la supervivència dels individus.

- Comparem la supervivència amb Parch:

```
chisq.test(train$Survived, train$Parch)
```

```
##
##  Pearson's Chi-squared test
##
## data:  train$Survived and train$Parch
## X-squared = 27.926, df = 6, p-value = 9.704e-05
```

També comprovem que la variable Parch està molt relacionada amb la supervivència dels individus.

Així, hem comprovat que totes les variables categòriques están molt relacionades amb la supervivència de l'individu.

Per últim, generarem un model per tal de predir els futurs resultats de supervivència per altres mostres amb els mateixos atributs:

```
library(randomForest)
modelo <- randomForest(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare +Embarked , data=train,  imp
```

Representació dels resultats a partir de taules i gràfiques.

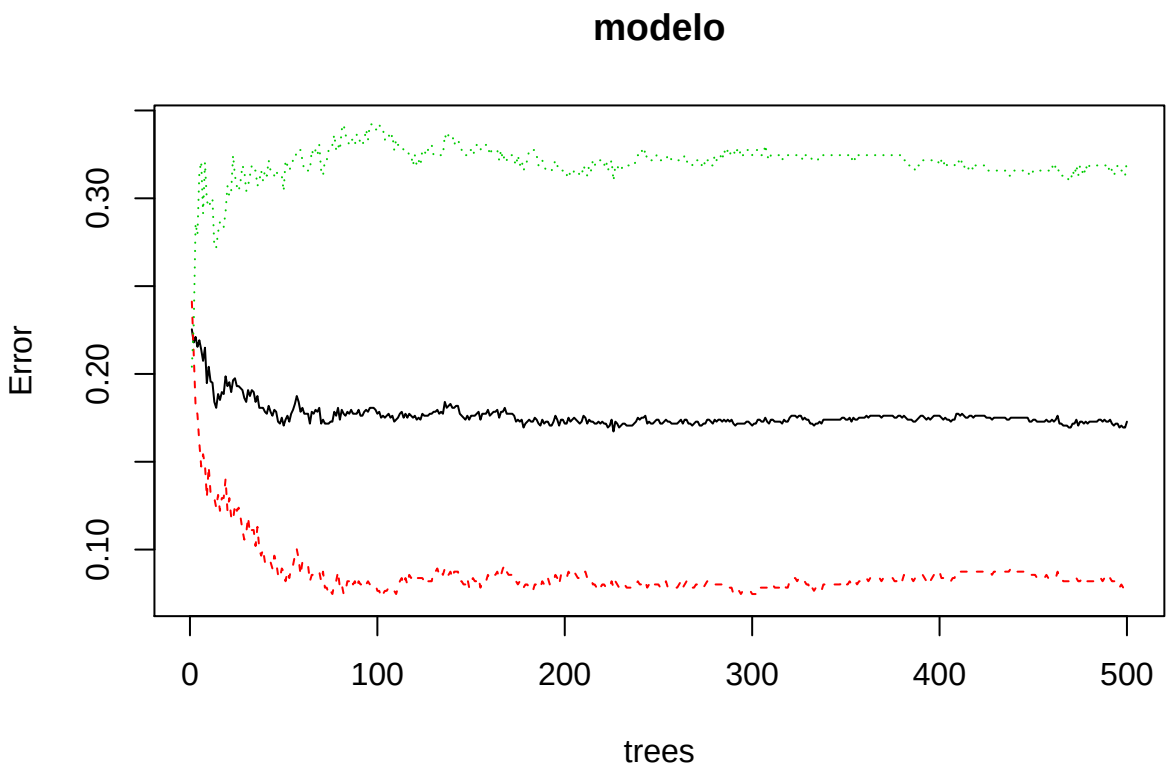
A continuació extreurem del model les mesures "MeanDecreaseAccuracy" y "MeanDecreaseGini", que mesuren la importància de les variables en la classificació de les mostres:

```
varImpPlot(modelo)
```

Podem veure que ambdúes mesures ponderen com a l'element més primordial el sexe, seguit de la tarifa, la clase i l'edad.

Aquest model també ens permet veure l'error del model en funció dels arbres que s'utilitzen per crearlos:

```
plot(modelo)
```



Per últim, veurem el model en sí, en el qual podrem trobar el error del model creat i la matriu de confusió en les dades de test, amb un resum de l'error per element resultant:

```
modelo
```

```
##
## Call:
##  randomForest(formula = Survived ~ Pclass + Sex + Age + SibSp +      Parch + Fare + Embarked, data =
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 17.28%
## Confusion matrix:
##     0   1 class.error
## 0 505  44  0.08014572
## 1 110 232  0.32163743
```

Resolució del problema.

A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Com a conclusió, podem dir que en aquest dataset totes les varibables estàn relacionades amb el resultat final, per tant totes deuen ser presses en consideració per qualsevol model estadístic o predictiu que pretenga resoldre la qüestió de si un subjecte sobreviurà al accident del Titànic o no.

A mes a mes, aquest dataset ens permet construir un model predictiu de tipus "RandomForest", el qual ens permet observar quines son les variables més influents en un resultat de predicció, confirmant les nostres

sospites al graficar les relacions entre variables en el dataset, y la precissió que pot tenir aquest model amb les dades facilitades, que es situa sobre el 16.5%.