PRAC2 - Neteja i validació de les dades

Rubén Vasallo González

4 de January 2019

Contents

Per continuar amb la practica, partim del dataset netejat de la part anterior de la PRAC.

```
# Eliminem l'aleatorietat a l'hora de executar els processos de calcul.
set.seed(5)
# Carregar el fitxer de dades en R
titanicdataset <- read.csv2(
  file = "titanic_data_clean.csv",
  header = TRUE,
  sep = ",",
  quote = "\"",
  fill = TRUE,
  encoding="UTF-8",
  na.strings=c("","NA")
)
# breu resumen dels atributs
summary(titanicdataset)
```

```
##     Survived          Pclass          Sex           Age
##  Min.   :0.0000   Min.   :1.000   female:314   Min.   : 1.00
##  1st Qu.:0.0000   1st Qu.:2.000   male  :577   1st Qu.:22.00
##  Median :0.0000   Median :3.000                Median :34.00
##  Mean   :0.3838   Mean   :2.309                Mean   :36.84
##  3rd Qu.:1.0000   3rd Qu.:3.000                3rd Qu.:48.00
##  Max.   :1.0000   Max.   :3.000                Max.   :88.00
##      SibSp           Parch        Embarked
##  Min.   :0.000   Min.   :0.0000   C:169
##  1st Qu.:0.000   1st Qu.:0.0000   Q: 77
##  Median :0.000   Median :0.0000   S:645
##  Mean   :0.523   Mean   :0.3816
##  3rd Qu.:1.000   3rd Qu.:0.0000
##  Max.   :8.000   Max.   :6.0000
```

Igual que en el cas anterior, R detecta alguns atributs com a numèrics, quan realment son atributs categòrics (amb valors acotats). Els transformem a atributs categòrics.

```
# Transformem els atributs necessaris a categòrics.
titanicdataset[,'Survived']<-factor(titanicdataset[,'Survived'])
titanicdataset[,'Pclass']<-factor(titanicdataset[,'Pclass'])
titanicdataset[,'SibSp']<-factor(titanicdataset[,'SibSp'])
titanicdataset[,'Parch']<-factor(titanicdataset[,'Parch'])
```

4 Analisis de les dades

4.1 Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

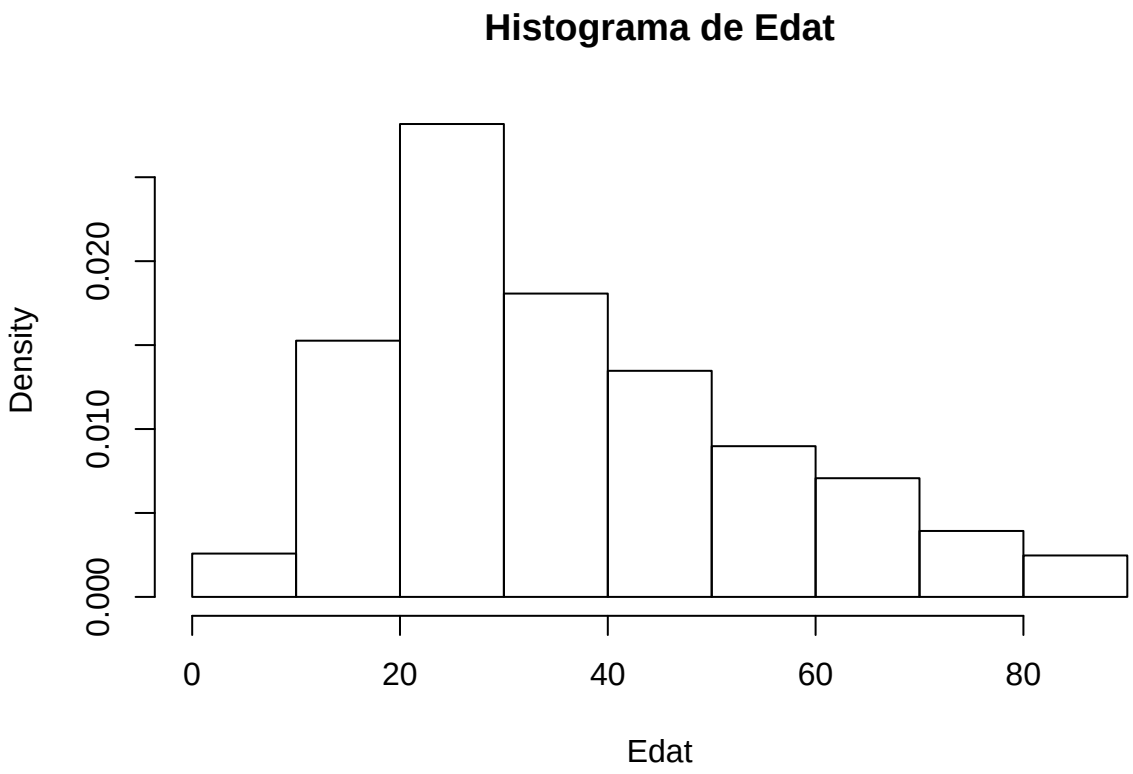
A continuació seleccionarem els grups que considerem poden ser interessants d'analitzar per detectar en les proves estadístiques si tenen relació amb si un passatger sobreviu o no.

```
# Agrupació per classe
titanicdataset.Pclass1 <- titanicdataset[titanicdataset$Pclass == "1",]
titanicdataset.Pclass2 <- titanicdataset[titanicdataset$Pclass == "2",]
titanicdataset.Pclass3 <- titanicdataset[titanicdataset$Pclass == "3",]

# Agrupació per Sexe
titanicdataset.male <- titanicdataset[titanicdataset$Sex == "male",]
titanicdataset.female <- titanicdataset[titanicdataset$Sex == "female",]

# Agrupació per origen del port
titanicdataset.Embarked_Cherbourg <- titanicdataset[titanicdataset$Embarked == "C",]
titanicdataset.Embarked_Queenstown <- titanicdataset[titanicdataset$Embarked == "Q",]
titanicdataset.Embarked_Southampton <- titanicdataset[titanicdataset$Embarked == "S",]

# Agrupació per Edat
titanicdataset.Edat_Menors <- titanicdataset[titanicdataset$Age < 18,]
titanicdataset.Edat_Adults <- titanicdataset[titanicdataset$Age >= 18 & titanicdataset$Age < 65,]
titanicdataset.Edat_Majors <- titanicdataset[titanicdataset$Age >= 65,]

# Agrupacio per # de germans / esposes a bord del Titanic
titanicdataset.SibSp_0 <- titanicdataset[titanicdataset$SibSp == "0",]
titanicdataset.SibSp_1 <- titanicdataset[titanicdataset$SibSp == "1",]
titanicdataset.SibSp_2 <- titanicdataset[titanicdataset$SibSp == "2",]
titanicdataset.SibSp_3 <- titanicdataset[titanicdataset$SibSp == "3",]
titanicdataset.SibSp_4 <- titanicdataset[titanicdataset$SibSp == "4",]
titanicdataset.SibSp_5 <- titanicdataset[titanicdataset$SibSp == "5",]
titanicdataset.SibSp_8 <- titanicdataset[titanicdataset$SibSp == "8",]

# Agrupacio per # de pares / fills a bord del Titanic
titanicdataset.Parch_0 <- titanicdataset[titanicdataset$Parch == "0",]
titanicdataset.Parch_1 <- titanicdataset[titanicdataset$Parch == "1",]
titanicdataset.Parch_2 <- titanicdataset[titanicdataset$Parch == "2",]
titanicdataset.Parch_3 <- titanicdataset[titanicdataset$Parch == "3",]
titanicdataset.Parch_4 <- titanicdataset[titanicdataset$Parch == "4",]
titanicdataset.Parch_5 <- titanicdataset[titanicdataset$Parch == "5",]
titanicdataset.Parch_6 <- titanicdataset[titanicdataset$Parch == "6",]
```

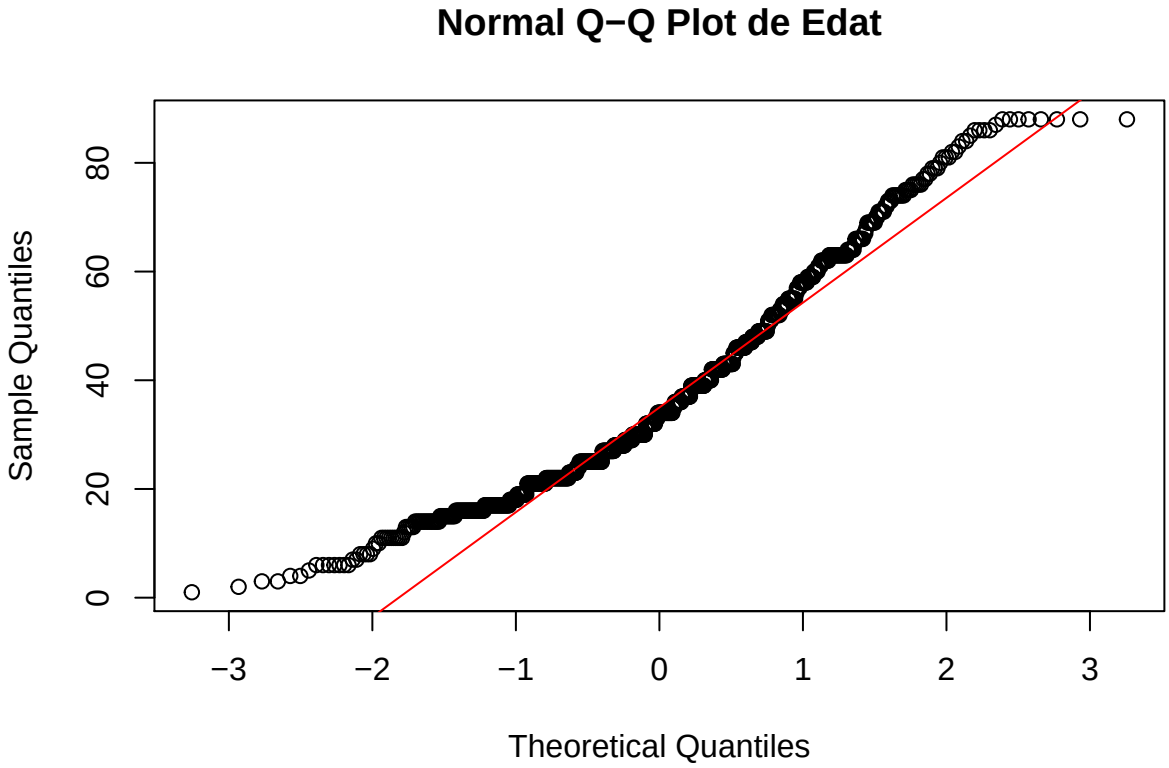
4.2 Comprovació de la normalitat i homogeneïtat de la variància.

A continuació revisarem si els atributs estan normalitzats o faria falta fer la normalització. Per detectar si els atributs tenen una distribució normal al fer, un plot o histograma d'aquests, deuríem de identificar una forma de campana des de el valor de les desviacions estàndard fins la mitjana d'aquest. Desafortunadament la majoria dels atributs que te aquest dataset son categòrics per el que solament podrem comprovar la normalitat del atribut Edat.

```
datanorm <- titanicdataset[,4]
hist(datanorm,
  main = "Histograma de Edat",
  xlab = "Edat",
  freq = FALSE
)
```




```
qqnorm(datanorm, main = "Normal Q-Q Plot de Edat")
qqline(datanorm, col="red")
```



Veiem en el gràfic normal Q-Q (gráfic quantile-quantile) que, tot i que el eix normal del estadístic k-enesim s'aproxima a la distribució, no acaba de alinear-se, per el que l'atribut es candidat per fer la normalització.

També podem fer el test de Shapiro per avaluar si l'atribut es candidat per ser normalitzat.

```
shapiro.test(titanicdataset$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanicdataset$Age
## W = 0.95232, p-value < 2.2e-16
```

Veiem que el p-value es inferior a 0.05, per el que el test ens confirma que el atribut no està normalitzat.

Finalment farem un test de Fligner per comprovar l'homogeneïtat de variàncies entre els atributs Edat i classe. El test de Fligner parteix de l'hipotens nul · la de que les variàncies dels dos atributs son iguales. Utilitzem el test de Fligner perquè sabem que l'atribut edat no està normalitzat.

```
fligner.test(Age ~ Pclass, data = titanicdataset)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Pclass
## Fligner-Killeen:med chi-squared = 14.992, df = 2, p-value =
## 0.0005552
```

Veiem que el valor del p-value es inferior a 0.05 per lo que no podem acceptar la hipòtesis de que les variàncies dels dos atributs son homogenis.

4.3 Aplicació de proves estadístiques

A continuació farem proves estadístiques al subconjunt de dades que hem estret anteriorment del dataset original. Aquestes proves consistiran en fer una prova de contrast d'hipòtesis sobre dos mostres per determinar si la probabilitat de sobreviure es major depenent d'un cert tipus d'atributs.

Per començar avaluarem si la probabilitat de sobreviure es major depenent de si viatjges en 1 o 2 classe. Per fer això utilitzarem el subconjunt de proves "titanicdataset.Pclass1" i "titanicdataset.Pclass2" utilitzant l'algoritme de Student's t-test, on passarem al test l'atribut "Survived" per comparar si la classe afecta directament a la probabilitat de sobreviure o no.

Val la pena comentar que, al ser variables categòriques hem de fer una primera transformació a valors numèrics (que en aquest cas la transformació es directe, ja que el dataset contempla com a valor 0 que el passatger no ha sobreviscut i 1 si aquest si que ha sobreviscut). A mes a mes com partim de un atribut categòric, aquest no te una distribució normal, però, com tenim mes de 30 observacions, podem donar el contrast d'hipòtesis com a valid.

Dons, començarem el següent contrast d'hipòtesis de dos mostres sobre la diferencia de mitjanes, on μ_1 es la mitjana de la població de la primera mostra i μ_2 es la mitjana de la població de la segona mostra. Si la resta de les dos mostres dona 0 podem acceptar la hipòtesis nul · la de que viatjar en primera o segona classe no afecta a la supervivència, mentre que si la resta de mitjanes no afecta, tendríem que descartar la hipòtesis nul · la i dir que viatjar en primera o segona classe si afecta a la supervivència. Utilitzarem un llindar de confiança del $\propto = 0.05$

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 < 0$

```
t.test(
  as.numeric(titanicdataset.Pclass1$Survived),
  as.numeric(titanicdataset.Pclass2$Survived),
  alternative = "less"
)
```

```
##
##  Welch Two Sample t-test
##
## data:  as.numeric(titanicdataset.Pclass1$Survived) and as.numeric(titanicdataset.Pclass2$Survived)
## t = 3.17, df = 383.5, p-value = 0.9992
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.2383621
## sample estimates:
## mean of x mean of y
##  1.629630  1.472826
```

Podem observar que el p-value es major a 0.05 per el que acceptem la hipòtesis nul · la i podem dir que viatjar en primera o segona classe no afecta a la supervivència amb un llindar de confiança de 95 %.

Farem la mateixa prova per avaluar si la probabilitat de sobreviure es major depenent de si viatges en 1 o 3 classe. Utilitzarem la mateixa hipòtesis de les mitjanes i l'algoritme de Student's t-test, on passarem al test l'atribut "Survived" per comparar si la classe afecta directament a la probabilitat de sobreviure o no.

```
t.test(
  as.numeric(titanicdataset.Pclass1$Survived),
  as.numeric(titanicdataset.Pclass3$Survived),
  alternative = "less"
)
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  as.numeric(titanicdataset.Pclass1$Survived) and as.numeric(titanicdataset.Pclass3$Survived)
## t = 10.137, df = 369.86, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.4502612
## sample estimates:
## mean of x mean of y
##  1.629630  1.242363
```

Podem observar que el p-value es major a 0.05 per el que acceptem la hipòtesis nul·la i podem dir que viatjar en primera o tercera classe no afecta a la supervivència amb un llindar de confiança de 95 %.

Nota: Veurem en el moment de crear el model predictiu, que l'atribut classe si serà rellevant a l'hora de decidir la supervivència del passatger.

Tot seguit avaluarem la mateixa prova però amb el sexe. Utilitzarem la mateixa hipòtesis de les mitjanes i l'algoritme de Student's t-test, on passarem al test l'atribut "Survived" per comparar si la classe afecta directament a la probabilitat de sobreviure o no.

```
t.test(
  as.numeric(titanicdataset.male$Survived),
  as.numeric(titanicdataset.female$Survived),
  alternative = "less"
)
```

```
##
##  Welch Two Sample t-test
##
## data:  as.numeric(titanicdataset.male$Survived) and as.numeric(titanicdataset.female$Survived)
## t = -18.672, df = 584.43, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.5043259
## sample estimates:
## mean of x mean of y
##  1.188908  1.742038
```

Podem observar que el p-value es menor a 0.05 per el que rebutgem la hipòtesis nul·la i podem dir que el sexe afecta a la supervivència amb un llindar de confiança de 95 %.

Tot seguit avaluarem la mateixa prova però amb el la edat. Utilitzarem els subconjunts de menors i adults per fer la prova i utilitzarem la mateixa hipòtesis de les mitjanes i l'algoritme de Student's t-test, on passarem al test l'atribut "Survived" per comparar si la classe afecta directament a la probabilitat de sobreviure o no.

```
t.test(
  as.numeric(titanicdataset.Edat_Adults$Survived),
  as.numeric(titanicdataset.Edat_Menors$Survived),
  alternative = "less"
)
```

```
##
##  Welch Two Sample t-test
##
## data:  as.numeric(titanicdataset.Edat_Adults$Survived) and as.numeric(titanicdataset.Edat_Menors$Sur
## t = -1.7892, df = 181.89, p-value = 0.03763
## alternative hypothesis: true difference in means is less than 0
```

```
## 95 percent confidence interval:
##         -Inf -0.00643574
## sample estimates:
## mean of x mean of y
##  1.377386  1.462121
```

Podem observar que el p-value es menor a 0.05 per el que rebutgem la hipòtesis nul · la i podem dir que l'edat afecta a la supervivència amb un llindar de confiança de 95 %.

Podríem continuar analitzant la resta d'atributs, però arribat a aquest punt, podem crear un model predictiu que ens permeti comprendre quin son els atributs rellevants i a mes a mes, ens permeti tenir un model predictiu per avaluar noves entrades. Per crear aquest model utilitzarem l'algoritme de arbre de decisió.

A continuació tenim que agafar un subset del dataset original per generar el model i un altre part per a comprovar que el model funciona. Normalment es sol utilitzar 2/3 per a fer el entrenament i 1/3 per a comprovar el test. No s'utilitza el dataset sencer per a fer el model, primer perquè no tindríem manera de assegurar-nos de que el model funciona i segon perquè tindríem un model massa ajustat nomes a les dades que tenim, per lo qual el model estaria sobre ajustat (overfitting) i no podríem garantir que funciona correctament amb noves observacions.

A continuació reordenem el dataset de manera aleatòria de cara a poder distribuir les observacions aleatòriament i així poder facilitar la divisió dels subsets d'una manera mes fàcil.

```
titanicdataset <- titanicdataset[sample(nrow(titanicdataset)), ]

head(titanicdataset, 16)
```

```
##     Survived Pclass    Sex Age SibSp Parch Embarked
## 179        0      2   male  34     0     0        S
## 610        1      1 female  49     0     0        S
## 816        0      1   male  46     0     0        S
## 253        0      1   male  76     0     0        S
## 93         0      1   male  57     1     0        S
## 622        1      1   male  52     1     0        S
## 468        0      1   male  69     0     0        S
## 715        0      2   male  64     0     0        S
## 845        0      3   male  15     0     0        S
## 98         1      1   male  23     0     1        C
## 241        0      3 female  28     1     0        C
## 432        1      3 female  25     1     0        S
## 280        1      3 female  42     1     1        S
## 491        0      3   male  30     1     0        S
## 231        1      1 female  42     1     0        S
## 177        0      3   male  18     3     1        S
```

Una vegada fet aixo, lo següent que fem es separem els atributs d'input amb el del classificador.

Després calculem fins a quint put fem la separació del dataset de entrenament al de test i creem els corresponents subsets.

```
X <- titanicdataset [ ,2:7]
y <- titanicdataset [ ,1]
# calculate split 1 / 3 to test . This number is where the dataset must split the values
split <- length(titanicdataset$Survived) - round(length(titanicdataset$Survived) / 3)


trainInputs <- X [1: split ,]
trainOutput <- y [1: split ]
```

```
testInputs <- X [( split + 1) : length(titanicdataset$Survived) ,]
testOutput <- y [( split + 1) : length(titanicdataset$Survived)]
```

A continuació utilitzarem el paquet C5.0 de R que te una implementació moderna de l'algorisme ID3 de Quinlan. Té els principis teòrics de l'ID3 més la poda automàtica.

```
model <- C50::C5.0(trainInputs, trainOutput)
summary(model)
```

```
##
## Call:
## C5.0.default(x = trainInputs, y = trainOutput)
##
##
## C5.0 [Release 2.07 GPL Edition]      Fri Jan  4 22:14:12 2019
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 594 cases (7 attributes) from undefined.data
##
## Decision tree:
##
## Sex = male: 0 (385/69)
## Sex = female:
## :...Pclass in {1,2}: 1 (111/7)
##     Pclass = 3:
##     :...Embarked in {C,Q}: 1 (39/12)
##         Embarked = S: 0 (59/22)
##
##
## Evaluation on training data (594 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##       4  110(18.5%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##     353    19    (a): class 0
##      91   131    (b): class 1
##
##
##  Attribute usage:
##
##  100.00% Sex
##   35.19% Pclass
##   16.50% Embarked
##
##
## Time: 0.0 secs
```

En el resum del model podem veure quins son els atributs decisius a l'hora de sobreviure al enfonsament del titànic:

```
Sex = male: 0 (385/69)
Sex = female:
:...Pclass in {1,2}: 1 (111/7)
    Pclass = 3:
    :...Embarked in {C,Q}: 1 (39/12)
        Embarked = S: 0 (59/22)
```

Noteu que, com a segon atribut important, el model predictiu ens diu que l'atribut rellevant es la classe on viatja el passatger, mentre que nosaltres anteriorment amb el algoritme de Student's t-test, hauriem arribat a la conclusió de que l'atribut de classe no era rellevant. Això es degut a que nosaltres només hauríem fet un estudi de comparació entre 1 i 2 classe o 1 i 3 classe, mentre que el model ha fet mes comparatives fins a arribar a la conclusió que viatjar en 1 i 2 classe si afecta a les probabilitats de sobreviure respecte a viatjar en 3 classe. Es important tenir en compte totes les alternatives abans de treure conclusions, ja que es molt fàcil descartar atributs que poden ser decisius a l'hora de fer els estudis. Afortunadament existeixen moltes eines (com ara el model predictiu que utilitzem) que ens ajuda a fer aquest anàlisis.

A continuació, executem el model amb el dataset de test generat anteriorment i calculem l'error.

```
prediction <- predict(model, testInputs, type = "class")

# Check the accuracy of the model
sum(prediction == testOutput)/length(prediction)
```

```
## [1] 0.8047138
```

Per comprovar l'exactitud del model, fem una comprovació senzilla en la que comprovem el output del model vrs el resultat esperat i el dividim per el numero d'observacions que te el dataset de test.

Ens dona gairebé un 80% d'exactitud. Sembla que esta prou be però podem tenir una millor observació si calculem la matriu de confusió.

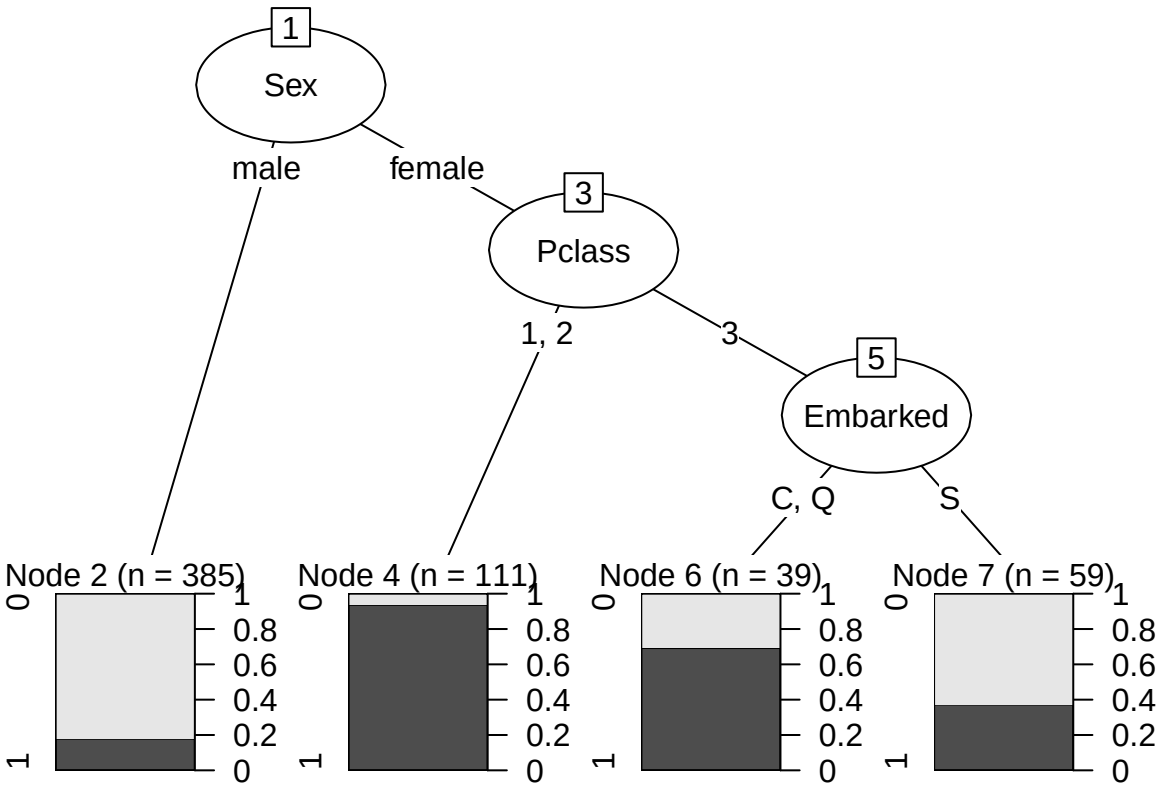
```
table(prediction, testOutput)
```

```
##           testOutput
## prediction   0   1
##          0 170  51
##          1   7  69
```

5. Representació dels resultats

Per tenir una millor visió de quins son els atributs determinants a l'hora de sobreviure, dibuixarem el model d'arbre de decisió generat al pas anterior.

```
plot(model)
```



Podem descriure aquest arbre com una serie de 4 regles determinants a l'hora de sobreviure al enfonsament del titànic.

- La primera regla que podem descriure es que si el sexe del passatger era home, la probabilitat de No sobreviure era aproximadament del 90%.
- La segona regla que podem descriure es que si el sexe del passatger era dona i viatjaba en 1 o 2 classe, la probabilitat de Si sobreviure era aproximadament del 98%.
- La tercera regla que podem descriure es que si el sexe del passatger era dona i viatjava en 3 classe i, a mes a mes, aquest havia embarcat en els ports de Cherbourg o Queenstown, la probabilitat de Si Sobreviure era aproximadament del 65%.
- I la quarta i ultima regla que podem descriure es que si el sexe del passatger era dona i viatjava en 3 classe i, a mes a mes, aquest havia embarcat en el port de Southampton, la probabilitat de No Sobreviure era aproximadament del 60%.

6. Conclusions

Com hem pogut veure en el transcurs de l'anàlisi, han sigut varis els factors que han fet que un passatger del Titànic sobrevisques al enfonsament.

Per fer aquest anàlisis hem obtingut un dataset amb un llistat de passatgers i atributs (característiques) relacionats amb ells. Després hem tingut que tractar aquests valors per avaluar atributs amb elements vuits i dades de dubtosa veracitat. Donat que no teníem cap manera de recuperar les dades originals dels dataset, em aconseguit aproximar els valors vuits utilitzant tècniques de mineria de dades (com es el cas del mètode dels veïns mes propers). En altres casos on els valors dels atributs tenien dubtosa veracitat (probablement per un error en la transcripció de les dades) no hem pogut tenir en compte aquest valors i s'han descartat. Es obvi fer menció a que si poguessin recuperar aquest valors, aquest podrien afectar al resultat del model.

Després hem fet el proves estadístiques per conèixer quin eren els atributs que eren mes rellevants per la

supervivència d'un passatger.

I finalment hem creat un model predictiu molt explicatiu, que te dos propòsits, el primer ens permet acabar de comprendre quins son els atributs més rellevants per la supervivència d'un passatger del Titànic i, a mes a mes ens permet introduir nous passatgers amb les seves característiques i fer una nova predicció.

Bibliografia

- Funcio Quantile-Quantile Plots => https://www.rdocumentation.org/packages/stats/versions/3.5.1/topics/qqnorm
- Anàlisis de l'homogeneïtat de la variància => https://rpubs.com/Joaquin_AR/218466
- Algoritme ID3 => https://ca.wikipedia.org/wiki/Algorisme_ID3