Universitat Oberta de Catalunya uoc.edu

Tipologia i cicle de vida de les dades

Pràctica 2: Neteja i anàlisis de les dades

Marc Alemany (malemanys@uoc.edu) i Josep Garcia (josepgarcia@uoc.edu)

Introducció

Per a la realització de la pràctica s'ha escollit el famós dataset del Titanic, un dels recomanats a l'aula. Comprèn una base de dades suficient per posar en pràctica l'après aquest semestre. Al llarg del document es mostrarà una descripció de les dades i les eines d'integració, neteja, validació i anàlisi d'aquestes.

Les competències desenvolupades han estat les següents:

- 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 la seva posterior anàlisi.

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.

1. Descripció del dataset

A la xarxa hi ha disponibles diferents conjunts de dades del Titanic, no totes coincideixen amb el nombre de registres i atributs. S'ha optat per descarregar el conjunt recomanat a l'aula en el repositori de Kaggle.
Aquest conjunt anomenat train comprèn 891 registres amb 12 atributs cadascun que resumim a continuació:

- PassengerId: int que indica l'id del passatger.
- Survived: int que indica si el passatger ha sobreviscut o no, és el valor a predir.
- Pclass: int que indica la classe del passatger.
- Name: string que indica el nom del passatger.
- Sex: string que indica el sexe del passatger.
- Age: int que indica l'edat del passatger.
- SibSp: int que indica el nombre de germans i esposes
- Parch: int que indica el nombre de pares i fills
- Ticket: string que indica el número de tiquet del passatger.
- Fare: numeric que indica la tarifa del passatge.
- Cabin: string que indica la cabina on s'ubica el passatger.
- Embarked: string que indica el port d'embarcament.

```
 PassengerId       Survived          Pclass          Name               Sex                 Age        
Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891         Length:891         Min.   : 0.42  
1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character   1st Qu.:20.12  
Median :446.0   Median :0.0000   Median :3.000   Mode  :character   Mode  :character   Median :28.00  
Mean   :446.0   Mean   :0.3838   Mean   :2.309                                         Mean   :29.70  
3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000                                         3rd Qu.:38.00  
Max.   :891.0   Max.   :1.0000   Max.   :3.000                                         Max.   :80.00  
                                                                                       NA's   :177    
     SibSp           Parch           Ticket               Fare            Cabin              Embarked        
Min.   :0.000   Min.   :0.0000   Length:891         Min.   :  0.00   Length:891         Length:891        
1st Qu.:0.000   1st Qu.:0.0000   Class :character   1st Qu.:  7.91   Class :character   Class :character  
Median :0.000   Median :0.0000   Mode  :character   Median : 14.45   Mode  :character   Mode  :character  
Mean   :0.523   Mean   :0.3816                      Mean   : 32.20                                        
3rd Qu.:1.000   3rd Qu.:0.0000                      3rd Qu.: 31.00                                        
Max.   :8.000   Max.   :6.0000                      Max.   :512.33                                        
```

Hipòtesi a contrastar

Sobreviure o no era una qüestió d'atzar, o va haver-hi factors determinants que van fer decantar la balança cap a una banda o l'altra? L'anàlisi del dataset ens ajudarà a esbrinar si la distribució de morts i supervivents a bord del Titanic respon a característiques dels passatgers com l'edat, sexe o la classe del passatge que van pagar.

2. Transformació de les dades d'interès a analitzar

La variable Survived serà la variable dependent a analitzar i al voltant de la qual girarà tota l'anàlisi i els models construïts. Juntament amb la classe del passatge Pclass i el sexe Sex les convertim a categòriques.

Survived, Pclass, Sex → categòrica

2.1 Valors buits

Quan analitzem els **valors buits** observem que manquen registres pels camps Cabin, Embarked i Age. Pels dos primers no hi ha problema perquè no formaran part dels models predictius, l'edat en canvi, és una variable a tenir en compte i que probablement resulta determinant a l'hora de predir si un passatger sobreviu o no.

Valors buits → Cabin 687, Embarked 2, Age 177

Quan analitzem en més detall el camp Age hi trobem registres erronis que no corresponen a edats reals. Es decideix assignar-hi NA i **eliminar tots els registres sense l'edat del passatger**.

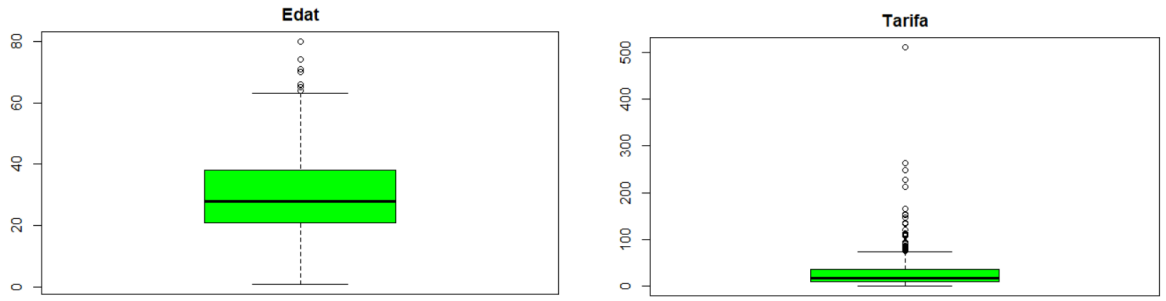
```
head(sort(totalData$Age),10) # hi ha registres decimals erronis
```

```
[1] 0.42 0.67 0.75 0.75 0.83 0.83 0.92 1.00 1.00 1.00
```

S'ha optat per eliminar els registres i no fer servir altres mètodes per a camps buits com assignar la mitjana poblacional, per exemple, ja que l'edat és una variable amb una forta dispersió i per no perdre la informació de la resta de camps acabaríem afegint soroll en aquest atribut, dels favorits a priori per al disseny de models.

2.2 Valors extrems

Pel que fa als **valors extrems,** observem com hi ha alguns outlets tant en l'edat com en la tarifa pagada. En ambdós casos mantenim els registres, ja que semblen valors reals i eliminar-los treuria informació rellevant al conjunt.



Valors extrems → Age, Fare

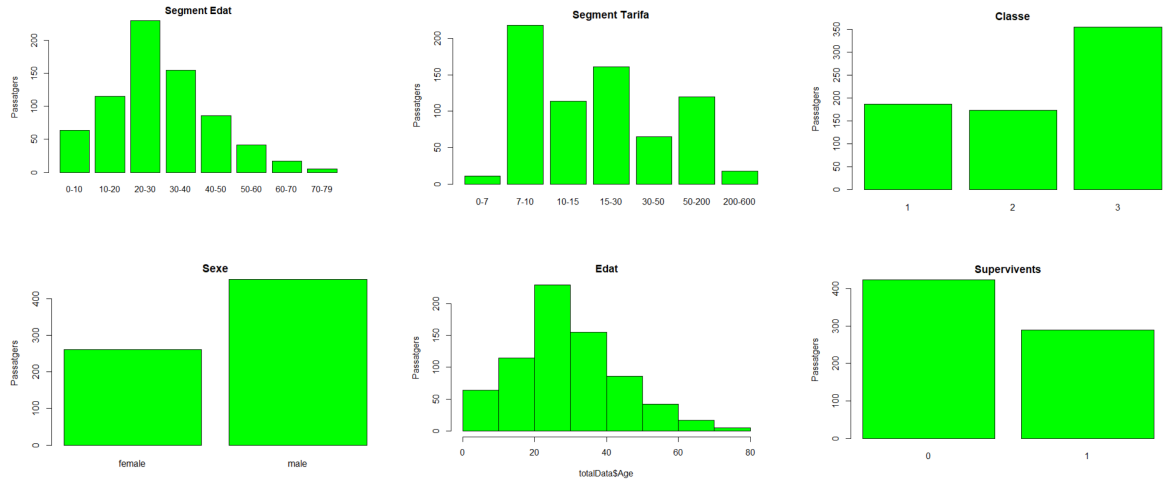
Continuem amb l'edat; discretitzem la variable en segments de 10 anys i afegim un camp nou de majors i menors d'edat. També discretitzem la tarifa en els trams mostrats a continuació:

```r
# Discretitzem
totalData["Segment_edat"] <- cut(totalData$Age, breaks = c(0,10,20,30,40,50,60,70,100), labels =
c("0-10", "10-20", "20-30", "30-40","40-50","50-60","60-70","70-79"))
totalData["Segment_tarifa"] <- cut(totalData$Fare, breaks = c(0,7,10,15,30,50,200,600), labels =
c("0-7", "7-10", "10-15","15-30","30-50","50-200","200-600"))
totalData["Segment_majors_menors"] <- cut(totalData$Age, breaks = c(0,18,100), labels = c("Menor
d'edat", "Major d'edat"))
```

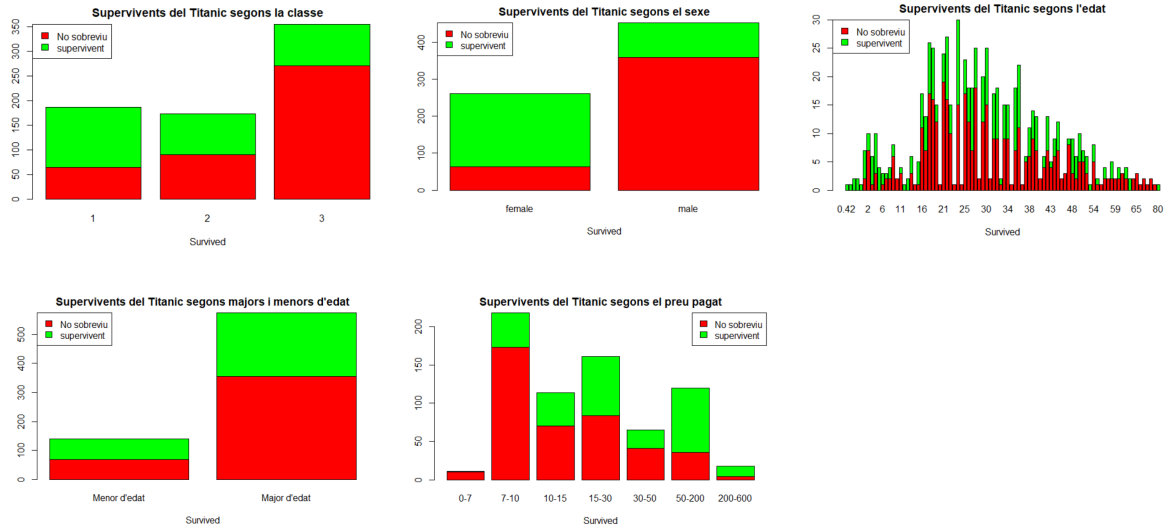
L'objectiu és comprovar si hi ha alguna relació de supervivència en funció de l'edat, el preu pagat i una més general que seria si són majors o menors d'edat.

Discretització Age, Fare → Segment_edat, Segment_tarifa, Segment_majors_menors

La distribució de les dades es representen amb els següents histogrames:



Creuant-les amb Survived:



A primer cop d'ull s'observa com pertànyer a la tercera classe i ser home no augurava un bon pronòstic a bord del vaixell.

3. Selecció i anàlisi de les dades escollides

Dels atributs originals acabarem treballant amb els següents:

Survived	Pclass	Sex	Age	Fare	Segment_edat	Segment_tarifa	Segment_majors_menors
0	3	male	22	7.2500	20-30	7-10	Major d'edat
1	1	female	38	71.2833	30-40	50-200	Major d'edat
1	3	female	26	7.9250	20-30	7-10	Major d'edat
1	1	female	35	53.1000	30-40	50-200	Major d'edat
0	3	male	35	8.0500	30-40	7-10	Major d'edat
0	1	male	54	51.8625	50-60	50-200	Major d'edat
0	3	male	2	21.0750	0-10	15-30	Menor d'edat

3.1 Normalitat i homocedasticitat

Per a comprovar la normalitat apliquem el **test de Kolmogorov-Smirnov** i el de **Shapiro-Wilk**.

```
        Shapiro-Wilk normality test

data:  totalData$Age
W = 0.98146, p-value = 7.337e-08
```

```
        Shapiro-Wilk normality test

data:  totalData$Fare
W = 0.52809, p-value < 2.2e-16
```

```
        Shapiro-Wilk normality test

data:  totalData$Pclass
W = 0.74616, p-value < 2.2e-16
```

```
        Shapiro-Wilk normality test

data:  totalData$Survived
W = 0.62364, p-value < 2.2e-16
```

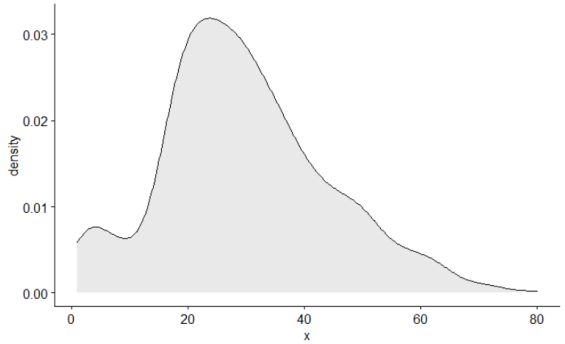
```
        Fligner-Killeen test of homogeneity of variances

data:  totalData$Age and totalData$Survived
Fligner-Killeen:med chi-squared = 0.28528, df = 1, p-value = 0.5933
```

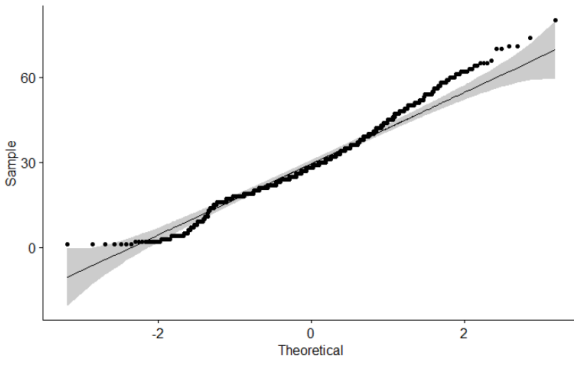
```
        Fligner-Killeen test of homogeneity of variances

data:  totalData$Fare and totalData$Survived
Fligner-Killeen:med chi-squared = 66.827, df = 1, p-value = 2.965e-16
```

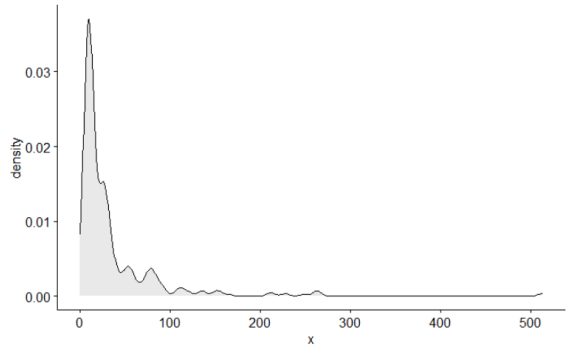
```
        Fligner-Killeen test of homogeneity of variances

data:  as.numeric(totalData$Pclass) and totalData$Survived
Fligner-Killeen:med chi-squared = 6.505, df = 1, p-value = 0.01076
```

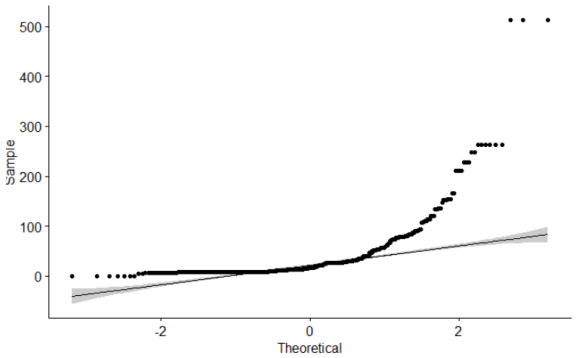
Els resultats d'ambdós tests rebutgen la normalitat de les dades, amb p-value inferior a 0.05.

Distribució i normalitat d'Age i Fare gràficament:









El resultat del **test de Fligner-Killeen** per a l'homogeneïtat de la variància, test adequat a les variables que no compleixen la condició de normalitat, rebutja l'homoscedasticitat per Fare (<0.05), és a dir, presenta variàncies estadísticament diferents per als grups Survived. En canvi, existeixen variàncies homogènies pels grups d'Age i Pclass.

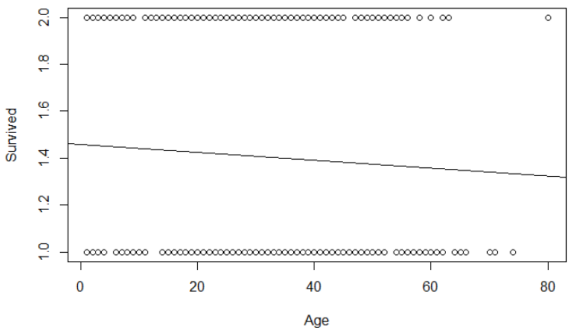
4. Models predictius

4.1 Regressions lineals

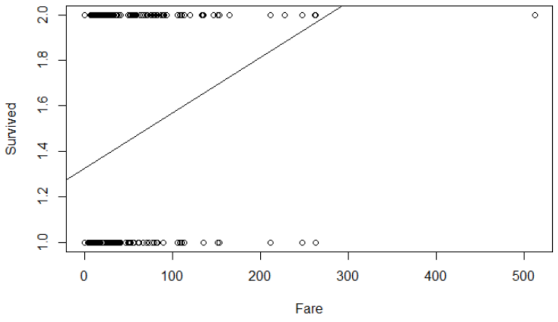
Començarem amb regressions lineals simples univariants per analitzar la dependència de Survived amb la resta de variables:

- Age: no és una variable estadísticament significativa i R2 = 0,23%
- Fare: és una variable significativa amb R2 = 7,07%
- Pclass: variable significativa amb R2 = 12,94%

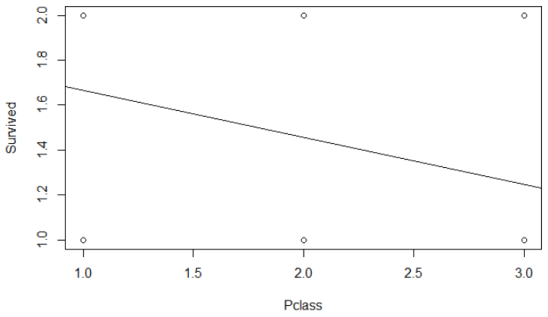
```
Call:
lm(formula = Survived ~ Age, data = totalData)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4574 -0.4174 -0.3840  0.5810  0.6743

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.459027   0.043349  33.658   <2e-16 ***
Age         -0.001666   0.001310  -1.272    0.204
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4918 on 687 degrees of freedom
Multiple R-squared:  0.002351,  Adjusted R-squared:  0.0008989
F-statistic: 1.619 on 1 and 687 DF,  p-value: 0.2037
```




```
Call:
lm(formula = Survived ~ Fare, data = totalData)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9663 -0.3583 -0.3421  0.5901  0.6769

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.3230804  0.0216571  61.092  < 2e-16 ***
Fare        0.0024458  0.0003381   7.234 1.26e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4747 on 687 degrees of freedom
Multiple R-squared:  0.07078,   Adjusted R-squared:  0.06943
F-statistic: 52.33 on 1 and 687 DF,  p-value: 1.256e-12
```



```
Call:
lm(formula = Survived ~ Pclass, data = totalData)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6653 -0.2468 -0.2468  0.3347  0.7532

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.87458    0.04950   37.87   <2e-16 ***
Pclass      -0.20926    0.02082  -10.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4598 on 687 degrees of freedom
Multiple R-squared:  0.1282,    Adjusted R-squared:  0.1269
F-statistic:   101 on 1 and 687 DF,  p-value: < 2.2e-16
```



Els models multivariants ens mostren els següents resultats:

- Pclass, Fare: les dues són significatives i pugen R2 a 13,62%
- Pclass^2 (relació quadràtica): R2 = 13,01%, però el regressor no és significatiu.

4.2 Regressió logística

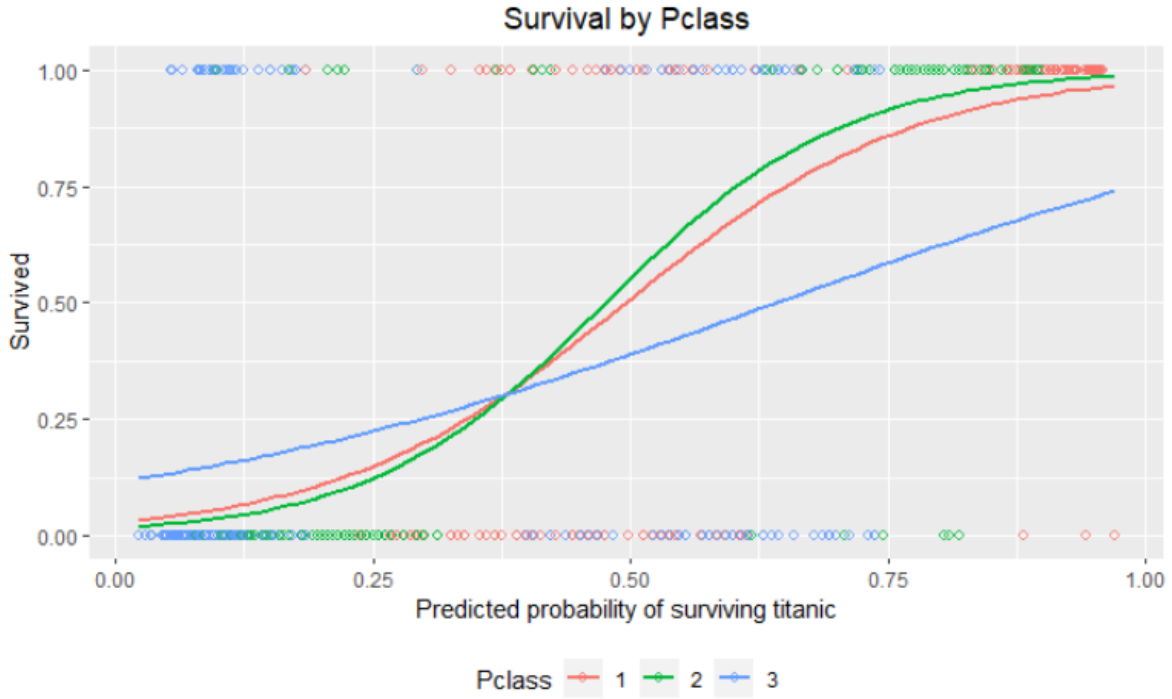
Tenint com a dependent una variable categòrica, resulta més adient una regressió logística que ens predigui la probabilitat de sobreviure o morir. El model analitzat és el següent:

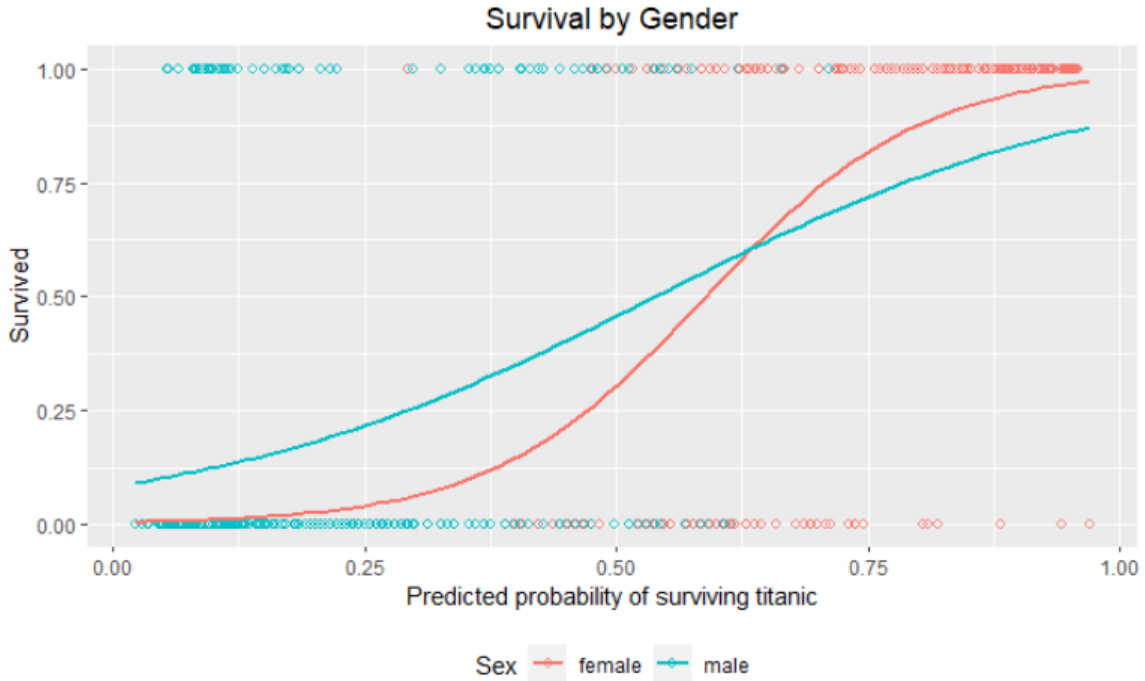
```
Call:
glm(formula = Survived ~ Pclass + Age + Sex, family = binomial(link = logit),
    data = totalData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7303  -0.6780  -0.3953   0.6485   2.4657

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.777013   0.401123   9.416  < 2e-16 ***
Pclass2     -1.309799   0.278066  -4.710 2.47e-06 ***
Pclass3     -2.580625   0.281442  -9.169  < 2e-16 ***
Age         -0.036985   0.007656  -4.831 1.36e-06 ***
Sexmale     -2.522781   0.207391 -12.164  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 647.28  on 709  degrees of freedom
AIC: 657.28

Number of Fisher Scoring iterations: 5


                 2.5 %     97.5 %
(Intercept) 20.37890724 98.3863313
Pclass2      0.15515074  0.4621731
Pclass3      0.04299250  0.1297997
Age          0.94905346  0.9780124
Sexmale      0.05293643  0.1194848
```

L'odds ratio ens indica l'augment o disminució de probabilitat de supervivència en funció de les variables predictives. Tots els regressors que apareixen tenen valors inferiors a la unitat, això vol dir que la seva presència disminueix la probabilitat de supervivència.

Amb un interval de confiança del 95%, ser de la segona classe fa que la probabilitat de sobreviure augmenti de 0,15 a 0,46 vegades. És a dir, la probabilitat de morir és d'1/0,15 a 1/0,46, o sigui del 2,17% al 6,67%. En el cas de ser de tercera classe, la probabilitat de no sobreviure era del 6,94% al 21,27%, i si eres home la probabilitat pujava del 8,52% al 19,41%. L'edat, tot i ser estadísticament significativa, no afecta massa a la probabilitat, ser un any més gran disminueix la probabilitat de sobreviure de l'1,01% a l'1,04%.

El model prediu per exemple, que la probabilitat de sobreviure per a un home de 40 anys de la tercera classe era del 5,7%. En canvi, una nena de 5 anys de primera classe tenia una probabilitat de sobreviure del 97,3%.





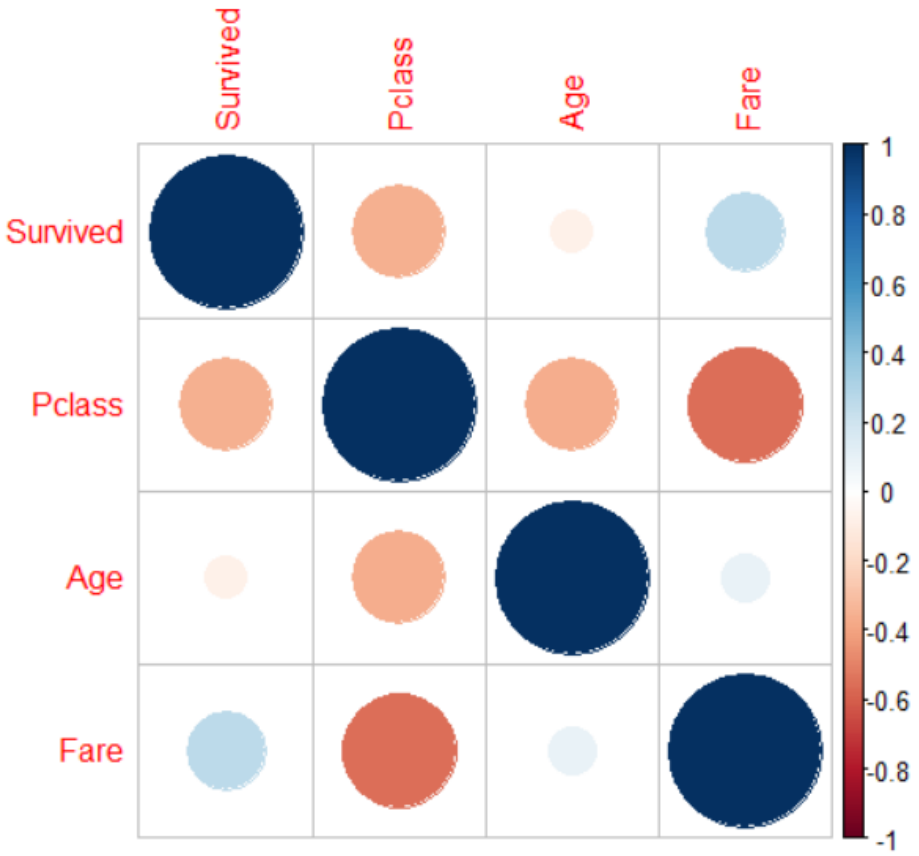
El model mostra la següent matriu de confusió →

```
Confusion Matrix and Statistics

       FALSE TRUE
FALSE    342   78
TRUE      65  204

               Accuracy : 0.7925
                 95% CI : (0.7602, 0.8222)
    No Information Rate : 0.5907
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5677

 Mcnemar's Test P-Value : 0.3156

            Sensitivity : 0.8403
            Specificity : 0.7234
         Pos Pred Value : 0.8143
         Neg Pred Value : 0.7584
             Prevalence : 0.5907
         Detection Rate : 0.4964
   Detection Prevalence : 0.6096
      Balanced Accuracy : 0.7818
```

El model de regressió logística mostra doncs bons resultats, una precisió que gairebé arriba al 80% amb la informació de bàsicament tres atributs de cada passatger: la classe, l'edat i el sexe.

4.3 Matriu de correlacions

La correlació més gran amb la supervivència és amb el preu pagat. A continuació la correlació és en funció de l'edat i finalment amb la classe, amb un resultat menor.



4.4 Arbre de decisió

Modelem un arbre de decisió que permeti analitzar quines persones sobreviuran segons les anàlisis efectuades anteriorment, és a dir, la variable per la qual classificarem és el camp Survived.

Per a la futura avaluació de l'arbre de decisió, és necessari dividir el conjunt de dades en un conjunt d'entrenament i un conjunt de prova. El conjunt d'entrenament és el subconjunt del conjunt original de dades utilitzat per a construir un primer model; i el conjunt de prova, el subconjunt del conjunt original de dades utilitzat per a avaluar la qualitat del model.
El més correcte serà utilitzar un conjunt de dades diferent del que utilitzem per a construir l'arbre, és a dir, un conjunt diferent del d'entrenament. No hi ha cap proporció fixada respecte al nombre relatiu de components de cada subconjunt, però la més utilitzada acostuma a ser 2/3 per al conjunt d'entrenament i 1/3, per al conjunt de prova.

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

```
Call:
C5.0.default(x = trainX, y = trainy, rules = TRUE)


C5.0 [Release 2.07 GPL Edition]         Wed Jun 02 00:45:31 2021
-------------------------------

Class specified by attribute `outcome'

Read 459 cases (8 attributes) from undefined.data

Rules:

Rule 1: (13, lift 1.6)
        Pclass = 3
        Sex = female
        Fare > 20.575
        ->  class 0  [0.933]

Rule 2: (10, lift 1.5)
        Pclass = 3
        Age > 16
        Segment_tarifa = 10-15
        ->  class 0  [0.917]

Rule 3: (166/25, lift 1.4)
        Pclass = 3
        Sex = male
        ->  class 0  [0.845]

Rule 4: (278/51, lift 1.4)
        Sex = male
        Age > 10
        ->  class 0  [0.814]

Rule 5: (8/1, lift 1.3)
        Segment_tarifa = 0-7
        ->  class 0  [0.800]
```

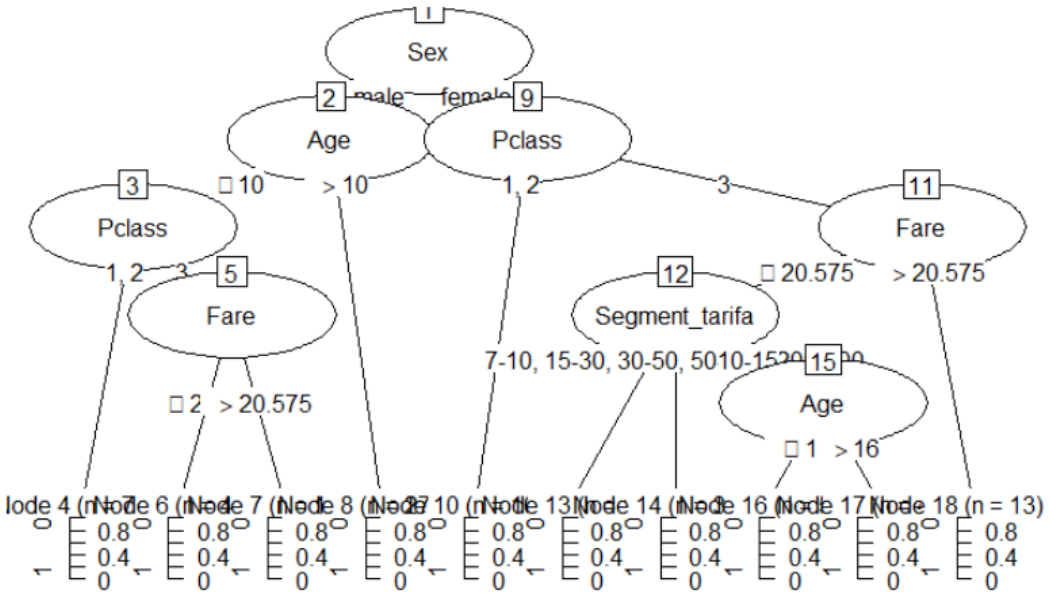
```
Rule 6: (12, lift 2.3)
        Pclass in {1, 2}
        Age <= 10
        ->  class 1  [0.929]

Rule 7: (5, lift 2.1)
        Sex = male
        Age <= 10
        Fare <= 20.575
        ->  class 1  [0.857]

Rule 8: (159/37, lift 1.9)
        Sex = female
        ->  class 1  [0.764]

Default class: 0


Evaluation on training data (459 cases):

             Rules
           ----------------
             No      Errors

              8   71(15.5%)   <<


            (a)   (b)    <-classified as
           ----  ----
            255    19    (a): class 0
             52   133    (b): class 1


        Attribute usage:

         98.69% Sex
         64.92% Age
         42.48% Pclass
          3.92% Fare
          3.92% Segment_tarifa
```

Es crea l'arbre de decisió usant les dades d'entrenament (cal no oblidar que la variable outcome és de tipus factor). Aquest model ens generarà unes regles que ens determinaran les probabilitats de supervivència.

El model ens ha creat 8 regles per poder crear l'arbre de decisió, regles que ens donen el percentatge de probabilitats que el passatger sobrevisqui o no, segons el valor class sigui classificat en 0 o 1.
Una vegada tenim el model, comprovem la qualitat predient la classe per a les dades de prova que ens hem reservat al principi, amb una precisió del 80,87%.
Quan hi ha poques classes, la qualitat de la predicció es pot analitzar mitjançant una matriu de confusió que identifica els tipus d'errors comesos.

```
      Predicted
testy   0   1
    0 116  34
    1  16  72
```

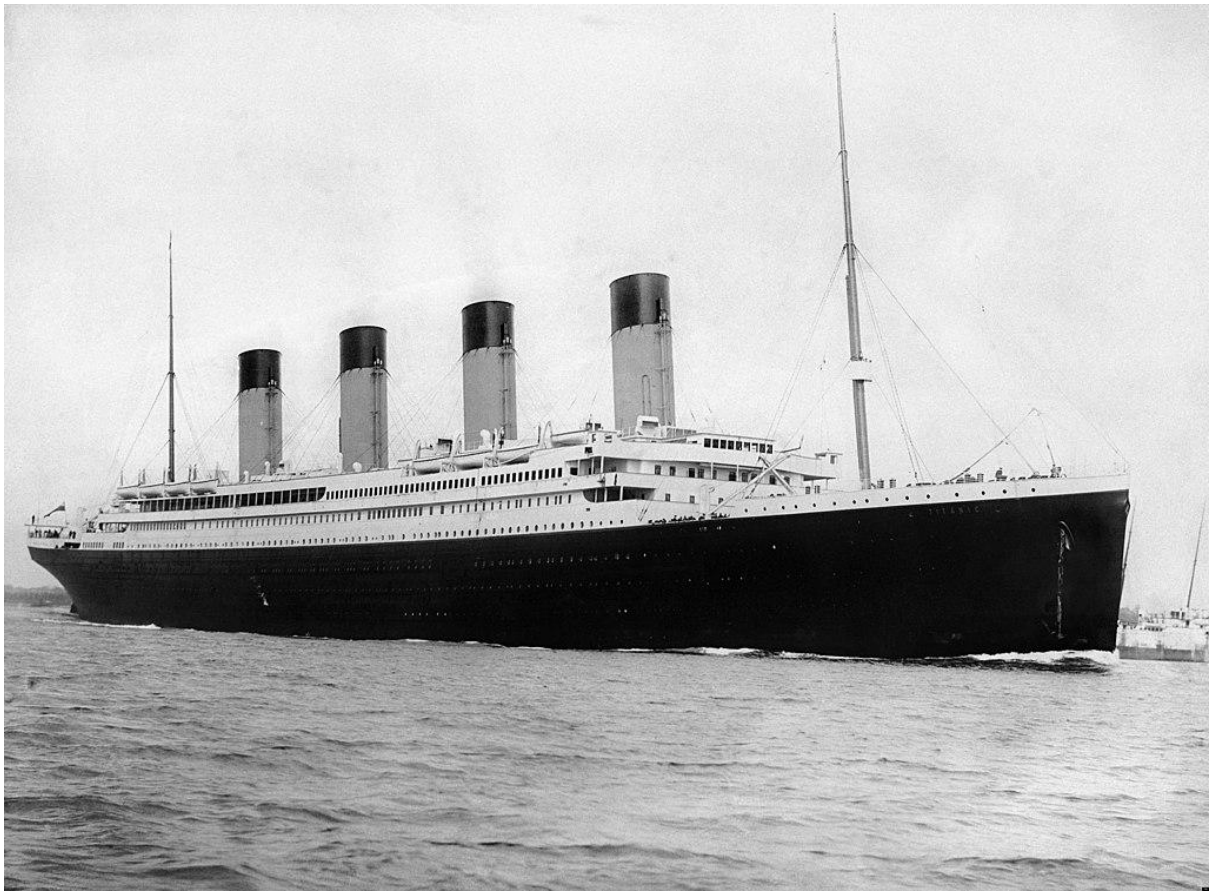
El gràfic de l'arbre de decisió resultant:



5. Conclusions

El resultat de les anàlisis de l'informe presentat mostren com sobreviure a la catàstrofe del Titanic no fou un tema d'atzar. Com s'ha calculat, no tenia la mateixa probabilitat un home de 40 anys de tercera classe que una nena de 5 anys de primera classe, per posar dos exemples extrems.
De fet, els models desenvolupats prediuen amb una precisió al voltant del 80% el resultat. I amb bàsicament tres variables, la classe en la qual viatjava el passatger, el seu gènere i l'edat.
Com s'ha anat destacant al llarg de l'anàlisi hi ha dos factors clars que jugaven en contra a bord, la primera és no ser a primera classe i l'altre ser un home. Els models més complexos com la regressió logística i l'arbre de decisió donen molta importància a ambdós factors, i en cas del segon també discrimina de forma important per l'edat a partir dels 10 anys.

Contribucions	Signa
Investigació prèvia	Marc Alemany, Josep Garcia
Redacció de les respostes	Marc Alemany, Josep Garcia
Desenvolupament codi	Marc Alemany, Josep Garcia



Font: Viquipèdia