



# Titanic

Neteja i anàlisi de les dades



Jordi Dil i Giró i Carlota Font Castell  
TIPOLOGIA I CICLE DE VIDA DE LES DADES

## Taula de continguts

1	Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre? .....	4
2.	Integració i selecció de les dades d'interès a analitzar .....	5
3.	Neteja de les dades.....	6
3.1	Les dades contenen zeros o elements buits? Com gestionaries aquests casos?.....	6
3.2	Identificació i tractament de valors extrems .....	7
4.	Anàlisi de les dades.....	9
4.1	Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar). ....	9
4.2	Comprovació de la normalitat i homogeneïtat de la variància. ....	9
4.2.1	Normalitat .....	9
4.2.2	Homogeneïtat.....	11
4.2.3	Creació de noves variables .....	12
4.2.4	Discretització de variables.....	14
4.3	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. ....	16
4.3.1	Matriu de correlacions .....	16
4.3.2	Relacions de dependència .....	16
4.3.3	Contrast d'hipòtesis .....	17
4.3.4	Reducció de la dimensionalitat.....	18
4.3.5	Divisió del data set .....	18
4.3.6	Models: resolució i representació del resultat .....	19
4.3.7	Predictions .....	28
5.	Representació dels resultats a partir de taules i gràfiques .....	30
5.1	Percentatges de Supervivents i No Supervivents .....	30
5.2	Gràfics del Conjunt d'Entrenament .....	30
5.3	Gràfics del Conjunt de Test o Predicció .....	35
6.	Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?.....	39
7.	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. ....	41
8.	Participació dels integrants.....	41
9.	Bibliografia.....	41

## 1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

En aquesta segona pràctica es presenta una anàlisis sobre el data set de “Titanic”. Aquest data set conté informació sobre els passatges del RMS Titanic, el famós transatlàctic britànic que es va enfonsar a l’abril del 1912 durant el seu primer viatge des de Southampton fins a Nova York. En concret, hi trobem informació sobre l’edat, el gendre i les característiques socio-econòmiques dels passatgers, així com també si van sobreviure al naufragi o no.

Aquest data set es troba disponible amb llicència pública en el següent enllaç: <https://www.kaggle.com/c/titanic>

### Detalls sobre el data set

De forma més detallada, el data set es troba disponible en format “*comma separated values (csv)*” i conté les següents variables:

- PassengerId: es tracta d’una variable qualitativa ordinal, on cada valor representa el nom d’un passatge de manera numèrica.
- Survived: la segona variable indica si el passatger va sobreviure o no al naufragi i ho representa amb una variable qualitativa binària, on 0 indica que no ho va fer, mentre que 1 sí va sobreviure.
- Pclass: variable qualitativa ordinal, que pot prendre qualsevol valor enter entre 1 i 3. Cada valor indica el tipus de classe en la que viatjava el passatger.
- Name: variable qualitativa nominal que indica el nom del passatger.
- Sex: variable qualitativa binaria, que indica el gendre del passatger.
- Age: indica l’edat del passatger, pel que es tracta d’una variable quantitativa discreta.
- SibSp: indica el número de germans, germanes i/o germanastres de cada passatger en el vaixell. Es tracta d’una variable quantitativa discreta.
- Parch: Nombre de pares o fills en el vaixell, pel que també és una variable quantitativa discreta.
- Ticket: Número identificador del bitllet del passatger, pel que és una variable qualitativa ordinal.
- Fare: Preu pagat pel bitllet. Per tant, es tracta d’una variable quantitativa contínua.
- Cabin: Identificador de la cabina assignada a cada passatger. Es tracta d’una variable qualitativa ordinal.
- Embarked: Port en el que va embarcar el passatger, pel que és una variable qualitativa nominal.

En primer lloc, ajuntarem els dos data sets disponibles per tal de fer un tractament i neteja de les dades. Emprarem el conjunt de dades obtingut del “train.csv”, que és l’únic que conté les dades de supervivència, èr a entrenar i validar els models obtinguts. Dividirem aquest data set en dos: `training_set` i `test_set`, per a realitzar l’entrenament i posteriorvaluació de la bondat del model.

Finalment, emprarem les dades del data set “test.csv” per a realitzar prediccions amb el model més precís.

### Preguntes plantejades per a la resolució de l’activitat

- Quines variables tenen una influència en el grau de supervivència en el Titànic?
- Quina relació hi ha entre les diferents variables?
- Els passatgers més joves tenien una probabilitat de supervivència més alta que la resta?

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

El data set que emprarem principalment per dur a terme la anàlisis de les dades és la combinació dels dos data sets disponibles, "train.csv" i "test.csv". Amb la totalitat del data set, farem una anàlisi descriptiva de les dades. En aquest data set total hi afegirem una variable extra anomenada que ens permetrà traçar l'origen de la mostra: train.csv i test.csv.

Els motius de la combinació dels dos data sets són:

- Totes les dades provenen del mateix accident, per tant, al unificar-les, podrem:
  - o Tenir més dades en el cas de necessitar, més endavant, predir valors desconeguts.
  - o Unificar les accions a realitzar en únic fitxer. És a dir, el tractament de dades sigui el mateix.

```
## 'data.frame': 1309 obs. of 13 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" ...
## $ file_origin: chr "train" "train" "train" "train" ...
```

Il·lustració 1: Estructura del DataSet unificat

PassengerId	Survived	Pclass	Name
Min. : 1	Min. :0.0000	Min. :1.000	Length:1309
1st Qu.: 328	1st Qu.:0.0000	1st Qu.:2.000	Class :character
Median : 655	Median :0.0000	Median :3.000	Mode :character
Mean : 655	Mean :0.3838	Mean :2.295	
3rd Qu.: 982	3rd Qu.:1.0000	3rd Qu.:3.000	
Max. : 1309	Max. :1.0000	Max. :3.000	
NA's :418			
Sex	Age	SibSp	Parch
Length:1309	Min. : 0.17	Min. :0.0000	Min. :0.000
Class :character	1st Qu.:21.00	1st Qu.:0.0000	1st Qu.:0.000
Mode :character	Median :28.00	Median :0.0000	Median :0.000
	Mean :29.88	Mean :0.4989	Mean :0.385
	3rd Qu.:39.00	3rd Qu.:1.0000	3rd Qu.:0.000
	Max. :80.00	Max. :8.0000	Max. :9.000
	NA's :263		
Ticket	Fare	Cabin	Embarked
Length:1309	Min. : 0.000	Length:1309	Length:1309
Class :character	1st Qu.: 7.896	Class :character	Class :character
Mode :character	Median : 14.454	Mode :character	Mode :character
	Mean : 33.295		
	3rd Qu.: 31.275		
	Max. :512.329		
	NA's :1		

Il·lustració 2: Resum del contingut del DataSet unificat

Per altre costat, el data set "test.csv" s'utilitzarà només per a realitzar prediccions.

### 3. Neteja de les dades

#### Objectiu

En primer lloc, comprovarem si existeixen valors duplicats i si hi ha valors sense contestar o en blanc. En segon lloc, observarem la distribució de cada variable en funció de *Survived*.

A continuació, en el cas que existeixin valors buits o no contestats, analitzarem quina és la millor opció per a tractar-los. Per exemple, els podrem reemplaçar per algun criteri matemàtic o, bé, eliminar la totalitat de l'atribut.

Després, analitzarem i tractarem els valors extrems.

#### 3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

##### Valors buits o no contestats

Un cop hem fusionat els dos data sets, veiem que existeixen valors NA o en blanc per a les variables *Survived*, *Fare*, *Cabin*, *Embarked* i *Age*.

	missing	na_count
PassengerId	0	0
Survived	0	418
Pclass	0	0
Name	0	0
Sex	0	0
Age	0	263
SibSp	0	0
Parch	0	0
Ticket	0	0
Fare	0	1
Cabin	1014	0
Embarked	2	0
file_origin	0	0

Il·lustració 3: Obtenció dels valors perduts, na o en blanc

La variable *Age* conté 263 valors no contestats, els quals representen el 20.1% del total; la variable *Embarked* conté 2 valors buits, que representen el 0.15% del total; la variable *Fare* té un valor en blanc; la variable *Survived* en conté 468 i, finalment, la variable *Cabin* conté 1.014 valors buits, que representen el 77.46% dels casos.

Respecte la variable *Survived*, no és sorprenent que tinguem valors en blanc, ja que per a totes les observacions del data set “test.csv”, aquesta no existeix. Ja que és la nostra variable de predicció, no tractarem aquests valors NA.

Pel que fa a la variable *Cabin*, ja que la majoria de valors es troben en blanc, prescindirem d'ella per a la anàlisi.

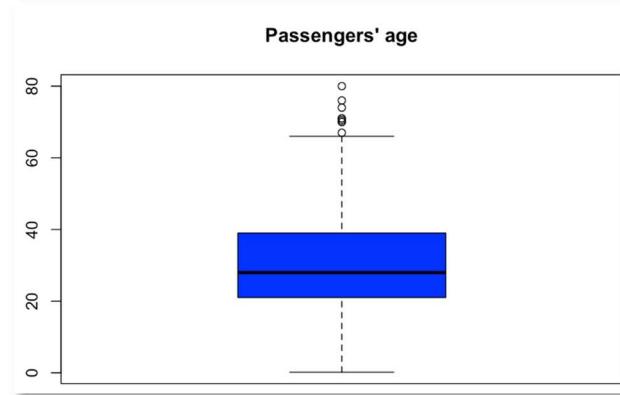
Respecte la variable *Embarked*, substituïm els valors buits pel valor més comú. En canvi, per a la variable *Fare*, agafarem la mitjana per l'únic valor que falta.

Finalment, per la variable edat, procedim d'una altra manera, ja que si ho substituïssim pel valor més comú o bé per la mitjana, podria fàcilment portar a anàlisis esbiaixats, ja que un nombre bastant elevat dels valors es troben no contestats. En aquest cas, substituirem els valors NA en funció dels passatgers semblants en funció de les altres variables.

### 3.2 Identificació i tractament de valors extrems

#### Valors extrems o outliers

Analitzem la variable numèrica Age, només en aquest cas se'n poden donar.

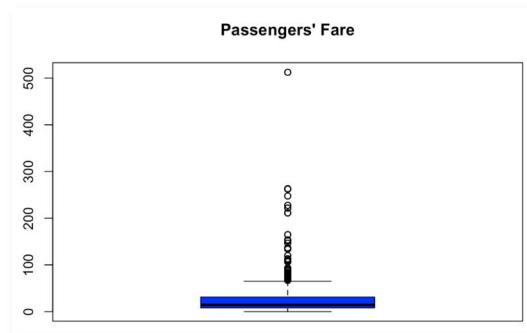


Il·lustració 4: Boxplot per a cercar valors extrems a l'atribut Age

A través d'un gràfic de caixes (box plot, Il·lustració 4) , observem que els outliers es donen únicament en el sector de la gent gran. Dins dels valors extrems, el més elevat és el 80 i, el més inferior és el 67.

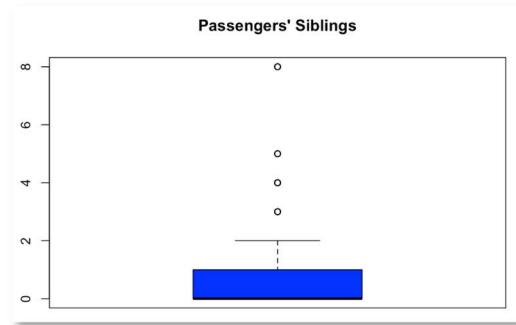
Donat que es tracta d'un interval d'edats totalment raonable, considerarem tots els valors dins de la variable Age com a vàlids.

A continuació s'analitzen la resta de variables numèriques com ara Fare, SibSp i Parch.

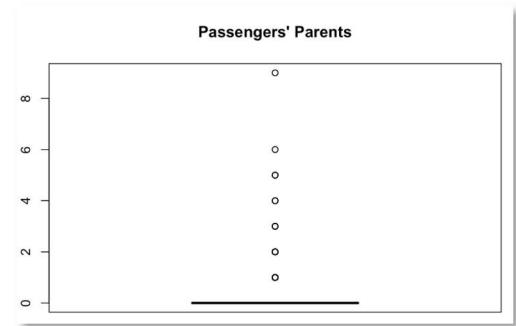


Il·lustració 5: Boxplot per a cercar valors extrems a l'atribut Fare

En el cas de *Fare* s'observen molts valors elevats. Segons hem consultat el preu del bitllet en cas de grups era englobat en un d'únic. En el cas de la compra de bitllets per un grup nombrós de persones tindria sentit aquesta mena de valors.



Il·lustració 6 : Boxplot per a cercar valors extrems a l'atribut Siblings



Il·lustració 7: Boxplot per a cercar valors extrems a l'atribut Parents

En aquestes altres 3 (Il·lustracions 5, 6 i 7) els possibles valors mostrats com a outliers, al igual que en el cas de Age, són possibles i, com a tal, són acceptats.

## 4. Anàlisi de les dades

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

#### Objectius

En aquest apartat, primer seleccionarem les variables rellevants per a la anàlisi, caldrà analitzar la distribució de les variables, la seva normalitat, així com la seva heterogeneïtat. D'aquesta manera, podrem decidir quins són els models que més s'ajusten a la nostre anàlisis.

A continuació, analitzarem si és convenient crear noves variables a partir dels atributs que venen per defecte. crearem dues variables noves, les quals generarem a partir d'atributs que venen per defecte.

I, finalment, analitzarem la possibilitat de discretitzar algunes variables, en el cas de ser convenient.

#### Resolució

##### Distribució Survived

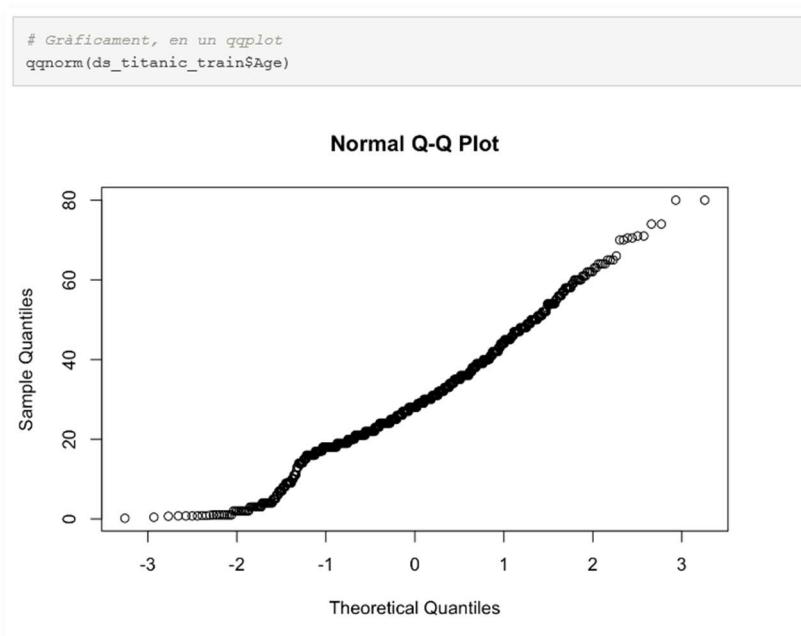
En primer lloc, hem analitzat la distribució de la variable de predicció. És a dir, del total d'observacions, hem comprovat que la majoria no van sobreviure al naufragi. De fet, només un 38.38% dels passatgers i tripulació ho van fer.

### 4.2 Comprovació de la normalitat i homogeneïtat de la variància.

#### 4.2.1 Normalitat

Comprovem les distribucions de la variables Age, Fare, SibSp i Parch, ja que són les úniques variables numèriques que tenim.

Com a exemple, el cas de la variable Age:



Il·lustració 8: Gràfica QQplot per l'atribut edat. avaluació de normalitat

Segons el gràfic (II·lustració 8), pot semblar que s'apropa a la normalitat en el centre, però no en els extrems.

Comprovem la normalitat a través d'un Shapiro test, amb un nivell de significació del 5%, on la hipòtesis nul·la assumeix normalitat

```
shapiro.test(ds_titanic_train$Age)

##
## Shapiro-Wilk normality test
##
## data: ds_titanic_train$Age
## W = 0.97945, p-value = 7.002e-10
```

II·lustració 9: resultat test shapiro aplicat a l'edat. amb el valor de p-value obtingut rebutjarem la hipòtesis nul·la concloent que edat no segueix una distribució normal.

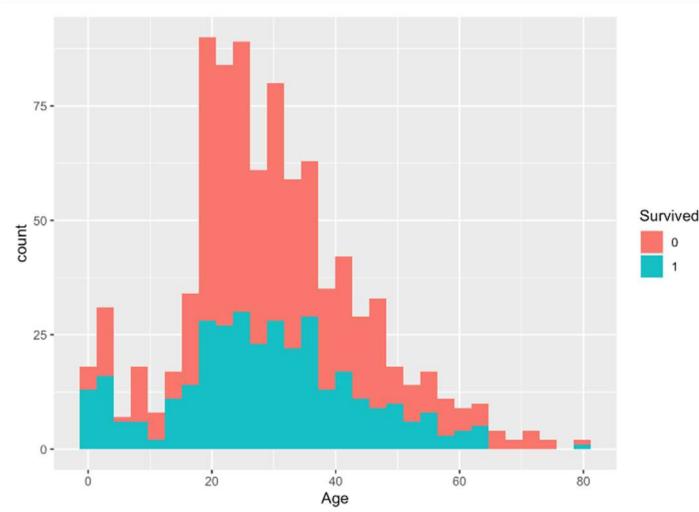
Tot i que pel teorema central del límit podríem considerar que segueix una distribució normal, el resultat obtingut pel Shapiro Test indica que la hipòtesis nul·la és rebutjada, pel que podem afirmar que la variable Age no segueix una distribució normal.

Observem la distribució en funció de la supervivència (II·lustració 11)

```
ggplot(data = ds_titanic_train) + aes(x = Age, fill = Survived) + geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

II·lustració 10: obtenim la distribució en funció de la supervivència.



II·lustració 11: Distribució de l'edat en funció de la supervivència

Per a cada variable, hem representat l'atribut en un qqplot. Per a la variable Age, malauradament, el gràfic no és prou precís, ja que els valors del centre podrien portar a la conclusió que la variable segueix una distribució normal. En canvi, els extrems indiquen el contrari. A més a més, pel teorema central del límit, podríem assumir normalitat. Ja que no és del tot clar, hem realitzat també un contrast d'hipòtesi a través del Shapiro-Wilk test.

Ja que la hipòtesi nul·la es rebutja amb un nivell de significació del 5%, afirmem que la variable no segueix una distribució normal.

Per a les altres tres variables, el gràfic QQplot és suficient per veure a cop d'ull que no segueixen una distribució normal.

El detall d'accions realitzades sobre les altres variables està en el propi codi amb el html generat per a ser consultat.

#### 4.2.2 Homogeneïtat

En el següent pas, hem comprovat si la variància de l'atribut Age és significativament diferent que la de la variable Survived, a través d'un test de variància (F-Test). Amb un nivell de significació del 5%, acceptem la hipòtesis nul·la, pel que podem concloure que les dues variables tenen variàncies significativament semblants. A més a més, l'Interval de Confiança es troba per sobre de 1, que reafirma que les variables són semblants.

En el cas d'Age:

```
var.test(x = ds_titanic_train[ds_titanic_train$Survived == "0", "Age"],
          y = ds_titanic_train[ds_titanic_train$Survived == "1", "Age"])
)

## 
## F test to compare two variances
##
## data: ds_titanic_train[ds_titanic_train$Survived == "0", "Age"] and ds_titanic_train[ds_titanic_train$Survived == "1", "Age"]
## F = 0.89666, num df = 548, denom df = 341, p-value = 0.2586
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7388307 1.0835216
## sample estimates:
## ratio of variances
## 0.8966649
```

Il·lustració 12: Test de variància f-test per analitzar si la variància de l'edat entre les mostres que sobreviuen i les que no poden ser considerades iguals

La hipòtesi nul·la planteja que les dues variàncies són iguals. Tenim com a resultat (com podem observar en la Il·lustració 12) un ratio de variàncies de 0.897 i un p-value de 0.2581, el qual és major que el nivell de significació del 5%. Per tant, podem concloure que no hi ha diferències significatives entre les dues variables.

```
var.test(x = ds_titanic_train[ds_titanic_train$Survived == "0", "Fare"],
          y = ds_titanic_train[ds_titanic_train$Survived == "1", "Fare"])
)

## 
## F test to compare two variances
##
## data: ds_titanic_train[ds_titanic_train$Survived == "0", "Fare"] and ds_titanic_train[ds_titanic_train$Survived == "1", "Fare"]
## F = 0.22214, num df = 548, denom df = 341, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.1830368 0.2684300
## sample estimates:
## ratio of variances
## 0.2221384
```

Il·lustració 13: Test de Variància F-Test per analitzar si la variància de Fare entre les mostres que sobreviuen i les que no poden ser considerades iguals

```

var.test(x = ds_titanic_train[ds_titanic_train$Survived == "0",'SibSp'],
         y = ds_titanic_train[ds_titanic_train$Survived == "1",'SibSp']
         )

##
## F test to compare two variances
##
## data: ds_titanic_train[ds_titanic_train$Survived == "0", "SibSp"] and ds_titanic_train[ds_titanic_train$Survived == "1", "SibSp"]
## F = 3.3052, num df = 548, denom df = 341, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 2.723366 3.993914
## sample estimates:
## ratio of variances
## 3.305151

```

Il·lustració 14: Test de Variància F-Test per analitzar si la variància de SibSp entre les mostres que sobreviuen i les que no poden ser considerades iguals

Pel que fa a les variables SibSp i Fare (Il·lustracions 13 i 14) , en els dos casos hem rebutjat la hipòtesis nul·la, pel que les variàncies són significativament diferents respecte Survived.

```

var.test(x = ds_titanic_train[ds_titanic_train$Survived == "0",'Parch'],
         y = ds_titanic_train[ds_titanic_train$Survived == "1",'Parch']
         )

##
## F test to compare two variances
##
## data: ds_titanic_train[ds_titanic_train$Survived == "0", "Parch"] and ds_titanic_train[ds_titanic_train$Survived == "1", "Parch"]
## F = 1.1378, num df = 548, denom df = 341, p-value = 0.1908
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.9375169 1.3749020
## sample estimates:
## ratio of variances
## 1.137796

```

Il·lustració 15: : Test de Variància F-Test per analitzar si la variància de Parch entre les mostres que sobreviuen i les que no poden ser considerades iguals

Per altra banda, per a la variable Parch (Il·lustració 15) , acceptem la hipòtesis nul·la i, per tant, concloem que les variàncies són similars.

#### 4.2.3 Creació de noves variables

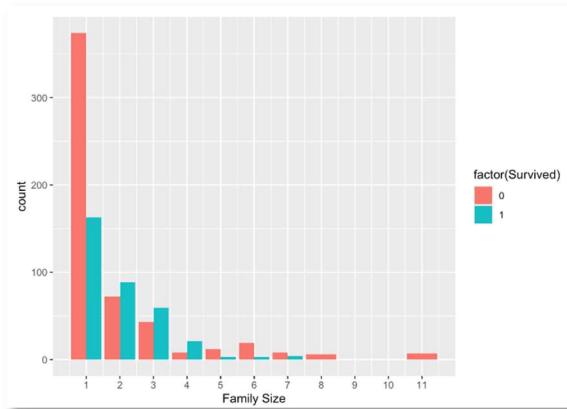
Primerament, hem creat una anomenada “FamilySize” (Il·lustració 16) , la qual indica la grandària de la família, on el valor mínim és 1, en el cas de que el passatger viatgés sol. Per tal realitzar-la, hem utilitzat les variables “SibSp” i “Parch”, les quals indiquen el número de germans i de pares que viatjaven amb el passatger, respectivament. Amb aquesta nova variable volem veure si la mida de la família podria tenir un impacte en la supervivència del passatger, així ajudant a respondre una de les preguntes plantejades.

```

# Creem la variables
ds_titanic$FamilySize<-ds_titanic$SibSp+ds_titanic$Parch+1

```

Il·lustració 16: Creació nova variable FamilySize



II·lustració 17: Visualització de la taxa de supervivència en funció de la mida de la Família

Per altre costat, hem creat una altra variable, anomenada "Title", que indica el títol del passatger o tripulant. Aquesta variable la crearem a partir de l'extracció de la primera part del nom ("Name") del passatger. Aquesta variable l'hem agrupat de 18 possibilitats a 5, per tal de reduir-ne la complexitat.

```
ds_titanic$Title<-gsub("^.*, (.*)\\..*$", "\\\\1", ds_titanic>Name)
unique(ds_titanic>Title)
```

```
## [1] "Mr"          "Mrs"         "Miss"        "Master"       "Don"
## [6] "Rev"         "Dr"          "Mme"         "Ms"          "Major"
## [11] "Lady"        "Sir"         "Mlle"        "Col"         "Capt"
## [16] "the Countess" "Jonkheer"    "Dona"
```

II·lustració 18: Creació Nova Variable Title

```
# Resumim els títols pel gendre
table(ds_titanic$Sex,ds_titanic>Title)
```

```
##          Capt Col Don Dona Dr Jonkheer Lady Major Master Miss Mlle Mme Mr Mrs
## female     0   0   0   1   1      0   1   0      0   260   2   1   0 197
## male       1   4   1   0   7      1   0   2      61   0   0   0 757   0
##
##          Ms Rev Sir the Countess
## female     2   0   0      1
## male       0   8   1      0
```

II·lustració 19: Obtenció de total de títols per tipus

Tenim un total de 18 Títols (II·lustració 19) diferents que agruparem (II·lustració 20).

```

ds_titanic$title[ds_titanic$title == 'Mlle'] <- 'Miss'
ds_titanic$title[ds_titanic$title == 'Ms'] <- 'Miss'
ds_titanic$title[ds_titanic$title == 'Mme'] <- 'Mrs'
ds_titanic$title[ds_titanic$title == 'Lady'] <- 'Miss'
ds_titanic$title[ds_titanic$title == 'Dona'] <- 'Miss'

ds_titanic$title[ds_titanic$title == 'Capt'] <- 'Officer'
ds_titanic$title[ds_titanic$title == 'Col'] <- 'Officer'
ds_titanic$title[ds_titanic$title == 'Major'] <- 'Officer'
ds_titanic$title[ds_titanic$title == 'Dr'] <- 'Officer'
ds_titanic$title[ds_titanic$title == 'Rev'] <- 'Officer'
ds_titanic$title[ds_titanic$title == 'Don'] <- 'Officer'
ds_titanic$title[ds_titanic$title == 'Sir'] <- 'Officer'
ds_titanic$title[ds_titanic$title == 'the Countess'] <- 'Officer'
ds_titanic$title[ds_titanic$title == 'Jonkheer'] <- 'Officer'

# Resum
table(ds_titanic$Sex,ds_titanic$title)

## 
##      Master Miss Mr Mrs Officer
##   female      0 266  0 198     2
##   male       61  0 757  0     25

```

Il·lustració 20: Agrupació dels Títols obtinguts

#### 4.2.4 Discretització de variables

Hem optat per discretitzar les variables Age i FamilySize per tal de reduir-ne la complexitat. Per tant, hem agrupat els grups d'edat en Children, Young, Adult, Advanced Adult i Elderly. Per altre costat, hem agrupat la mida de la família en Single, Small i Large.

```

summary(ds_titanic$Age)

##      Min. 1st Qu. Median    Mean 3rd Qu.    Max.
##      0.17   21.00  28.00  29.88  38.00  80.00

```

Il·lustració 21: Previ per analitzar Edat

Els punts de tall per l'edat han estat de 0,15,30,50, 64 i 100 com podem veure en la Il·lustració 22.

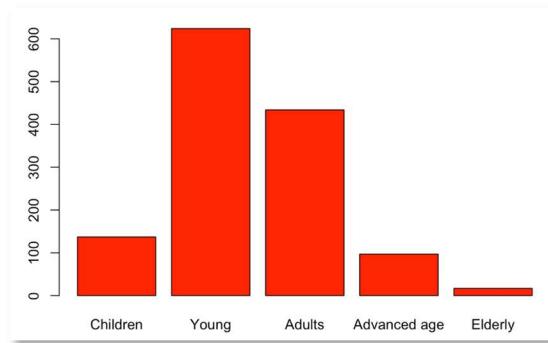
```

ds_titanic$Age.factor <- cut(ds_titanic$Age, breaks = c(0,15,30,50,64,100), labels=c('Children', 'Young', 'Adults', 'Advanced age', 'Elderly'))

plot(ds_titanic$Age.factor, col="red")

```

Il·lustració 22: Discretització de l'Edat



Il·lustració 23: Representació amb l'Edat discretitzada

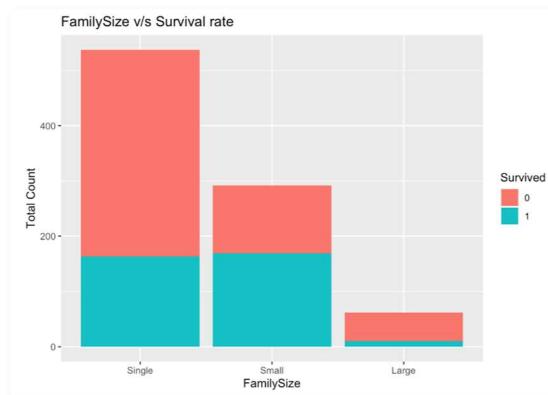
Observem (Il·lustració 23) que la majoria de passatgers són joves d'entre 15 i 29 anys.

#### Discretització de FamilySize

Els punts de tall per l'edat han estat de 0, 1, 4 i 90 com podem veure en la il·lustració 24.

```
ds_titanic$FamilySize <- as.numeric(ds_titanic$FamilySize)
ds_titanic$FamilySize.factor <- cut(ds_titanic$FamilySize, breaks = c(0,1,4,90), labels=c('Single', 'Small', 'Large'))
```

IL·LUSTRACIÓ 24: DISCRETITZACIÓ DE LA MIDA FAMILIA



Il·lustració 25: Representació amb la Mida Família discretitzada

Veiem (Il·lustració 25) que de entre els passatgers que no van sobreviure, destaquen aquells que pertanyien a unitats familiars amb més de 4 membres (Large).

D'entre els supervivents destaquen els que formaven part de petites unitats familiars.

**4.3 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.**

**4.3.1 Matriu de correlacions**

A partir d'una matriu de correlacions hem volgut comprovar quines variables tenen una correlació significativa entre elles.



Il·lustració 26: Matriu de Correlacions

Com a resultat (Il·lustració 26), podem veure que, de forma evident, la correlació més forta es troba entre les variables FamilySize amb Parch i SibSp. No és sorprenent ja que la mida de la família s'ha derivat a partir d'aquestes dues variables. Obviament PassengerId no presenta cap mena de correlació amb cap altra variable com és lògic.

A més a més, també existeix una correlació inversa entre Age i Survived. Això vol dir que a més jove el passatger, més probabilitats hauria de tenir de supervivència.

Alhora, es detecten correlacions entre Sex i Survived, així com Pclass i Survived. I, també, entre Edat i Fare.

A tenir en compte a efectes de comprensió: Valors elevats positius, propers a 1 i en aquest cas en color vermell intens, de correlació indiquen una correlació directa i valors elevats negatius, properes a -1 i en aquest cas en color blau intens , indiquen una correlació inversa.

**4.3.2 Relacions de dependència**

En aquest apartat, hem analitzat diferents relacions de dependència entre la variable Survived i totes les altres variables, sempre de forma individual. Per dur-ho a terme, hem emprat un Chi-Squared test, on la hipòtesis nul·la assumeix independència entre la variable Survived i cada variable.

En tots els casos, la hipòtesis nul·la es rebutja i s'accepta la alternativa. Per tant, la variable Survived té una relació de dependència o associació amb un nivell de significació del 5% amb cadascuna de les variables que estem utilitzant per a realitzar la anàlisis.

```
chisq.test(table(ds_titanic_train$Sex, ds_titanic_train$Survived))

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

II·lustració 27: Test Chi-Quadrat. Sexe amb Supervivència

Relació de dependència entre Sexe i Supervivència (II·lustració 27).

Ja que el p-value és més petit que 0.05, rebutgem la hipòtesis nul·la. Per tant, podem afirmar que les variables Sex i Survived són dependents.

```
chisq.test(table(ds_titanic_train$Age, ds_titanic_train$Survived))

## Warning in chisq.test(table(ds_titanic_train$Age, ds_titanic_train$Survived)) :
## Chi-squared approximation may be incorrect

##
## Pearson's Chi-squared test
##
## data: table(ds_titanic_train$Age, ds_titanic_train$Survived)
## X-squared = 113.24, df = 90, p-value = 0.04936
```

II·lustració 28: Test Chi-Quadrat. Age amb Supervivència

En el cas de la Relació de dependència entre Edat i Supervivència (II·lustració 28) rebutgem la hipòtesis nul·la i, per tant, les variables Age i Survived tenen una dependència significativa.

El mateix procediment l'apliquem per altres variables qualitatives i es pot consultar en el propi fitxer .rmd adjunt al projecte.

Sexe, Edat, PClass, Embarked i FamilySize totes elles tenen certa dependència significativa amb la de Supervivència.

#### 4.3.3 Contrast d'hipòtesis

Realitzem un contrast d'hipòtesis (II·lustració 29), per veure si els passatgers amb una edat inferior a 20 anys (alternative='less', mu=20) sobreviurien o no al naufragi.

El contrast d'hipòtesis l'hem realitzat a través del test U de Mann.Whitney, assumint que les mostres són independents. No podem utilitzar el test t-student, ja que necessitaríem que les dades es distribuïssin normalment i, a més a més, les variàncies de les variables haurien de mantenir-se constants en el rang observat d'alguna altra variable. En aquest test, la hipòtesis nul·la assumeix que els passatgers amb una edat inferior a 20 anys no tenen més probabilitat de supervivència.

```

data_alive = ds_titanic_train[ds_titanic_train$Survived == "1",]
data_notalive = ds_titanic_train[ds_titanic_train$Survived == "0",]

resultat_contrast <- wilcox.test(x=data_alive$Age, y=data_notalive$Age, alternative = "less", mu=20, pair
ed = FALSE, conf.int=0.95)
resultat_contrast

```

```

##
## Wilcoxon rank sum test with continuity correction
##
## data: data_alive$Age and data_notalive$Age
## W = 26430, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 20
## 95 percent confidence interval:
##       -Inf 2.242463e-05
## sample estimates:
## difference in location
##                  -1.000002

```

Il·lustració 29: Prova no paramètrica, Test U Mann-Whitney per avaluar si la mitjana d'edat dels supervivents es inferior als 20 anys.

Un cop realitzat el test, veiem que la hipòtesis nul·la es rebutja amb un nivell de significació del 5%. Per tant, les probabilitats de supervivència d'una persona amb una edat inferior a 20 anys és més elevada que la resta.

#### 4.3.4 Reducció de la dimensionalitat

Després d'haver analitzat totes les variables de forma detallada, hem decidit reduir la dimensionalitat del data set, així reduint el nombre d'atributs. Del total hem eliminat la variable Cabin, tal i com s'ha explicat en l'apartat de tractament de valors NA o nuls. A més a més, eliminem les variables SibSp i Parche, ja que a través de la mida de la família (FamilySize) ja agrupem la mateixa informació. Pel contrari, sinó, estaríem tenint en compte les mateixes variables de forma duplicada. Finalment, també hem eliminat PassengerId i Name, ja que no aporten cap valor afegit.

Per tant, en conclusió, ens hem quedat amb les següents variables: Age, Survived, PClass, Sex, Fare, FamilySize, Embarked i Title.

#### 4.3.5 Divisió del data set

Abans de crear qualsevol model, hem dividit (Il·lustració 30) el data set train en dos, un amb 75% dels casos i, l'altre, amb la resta. El primer l'hem utilitzat per a crear diferents models explicats en el següent apartat, mentre que l'altre per testejar la bondat dels models.

```

# Ens quedem amb les variables que volem pel model
# Age, Survived, Pclass, Sex, Fare, FamilySize, Embarked and Title
titanic_sub <- ds_titanic_train[,c(2,3,5,6,10,12,14,15,16,17)]


# Determinem un random seed
set.seed(123)
split = sample.split(titanic_sub$Survived, SplitRatio = 0.75)
training_set = subset(titanic_sub, split == TRUE)
test_set = subset(titanic_sub, split == FALSE)

# Observem el resultat
str(training_set)

```

```

## 'data.frame': 668 obs. of 10 variables:
## $ Survived : Factor w/ 2 levels "0","1": 1 2 2 2 1 2 2 2 1 1 ...
## $ Pclass   : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 3 1 3 3 ...
## $ Sex      : Factor w/ 2 levels "female","male": 2 1 1 1 2 1 1 1 2 2 ...
## $ Age      : num  22 38 26 35 31 27 4 58 20 39 ...
## $ Fare     : num  7.25 71.28 7.92 53.1 8.46 ...
## $ Embarked : Factor w/ 3 levels "C","Q","S": 3 1 3 3 2 3 3 3 3 3 ...
## $ Title    : Factor w/ 5 levels "Master","Miss",...: 3 4 2 4 3 4 2 2 3 3 ...
## $ FamilySize: Factor w/ 9 levels "1","2","3","4",...: 2 2 1 2 1 3 3 1 1 7 ...
## $ Age.factor: Factor w/ 5 levels "Children","Young",...: 2 3 2 3 3 2 1 4 2 3 ...
## $ FamilySize.factor: Factor w/ 3 levels "Single","Small",...: 2 2 1 2 1 2 2 1 1 3 ...

```

Il·lustració 30: Divisió del DataSet ds\_titanic\_train amb el training\_set per entrenar models i test\_set per evaluar-los. El criteri per formar-los, un SplitRatio=0.75

#### 4.3.6 Models: resolució i representació del resultat

##### Objectius

L'objectiu d'aquest apartat és crear el millor model que serà aquell que tingui una millor precisió. Aquest serà més tard utilitzat per a realitzar prediccions.

Hem realitzat models de regressió univariades, regressions logístiques multivariables, Arbres de decisió i Random Forest.

##### 4.3.6.1 Models de regressió univariades

En aquest punt, hem creat dos models de regressió per a testejar les variables Sex i Embarked en funció de Survived.

Hem escollit Sex i Embarked com una de les parelles interessants d'estudi i, per tant, hem creat un model amb la variable dependent Survived i cada atribut com a variable independent, on es tracta de trobar en quin percentatge es modifica la probabilitat de supervivència en funció de cada atribut.

### Anàlisi de Supervivència en funció del Sexe

```
model.glm <- glm(Survived ~ Sex, data = training_set, family = binomial )
summary (model.glm)

##
## Call:
## glm(formula = Survived ~ Sex, family = binomial, data = training_set)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.6651 -0.6253 -0.6253  0.7585  1.8592
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.0986   0.1491   7.37 1.71e-13 ***
## Sexmale     -2.6315   0.1955 -13.46 < 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: 889.27 on 667 degrees of freedom
## Residual deviance: 670.26 on 666 degrees of freedom
## AIC: 674.26
##
## Number of Fisher Scoring iterations: 4
```

Il·lustració 31: Regressió logística univariable entre Survived i Sexe

Dins la regressió (Il·lustració 31), la intersecció (“Intercept”) recull els factors més favorables. Els asteriscs (\*\*\* ) indiquen la rellevància de l’atribut en el model. Sexfemale estaria inclòs en el intercept.

Per altre banda, els odds (en %) (Il·lustració 32) ens indicaran la variació en probabilitat de supervivència.

```
odds <- or_glm(data = training_set, model = model.glm)
odds

##
## predictor oddsratio ci_low (2.5) ci_high (97.5)      increment
## 1  Sexmale      0.072        0.049       0.105 Indicator variable

p_odds <- odds[,1:2]
p_odds["%"] <- (1-p_odds[2])*100
p_odds

##
## predictor oddsratio %
## 1  Sexmale      0.072 92.8
```

Il·lustració 32: Obtenció Odds Ratio Sexe i Survived

Pel que fa al gènere, el fet que els homes sobrevisquin és inferior en un 92.8% que les dones.

### Anàlisi de Supervivència en funció d'Embarked

Apliquem el model per obtenir la regressió logística entre Survived i Embarked (II·lustració 33). Veiem com l'atribut Embarked S i en menys mesura el d'embarked Q tenen rellevància en el model (amb \*\*\* i \* respectivament). El port C està inclòs en l'Intercept.

```
model.glm <- glm(Survived ~ Embarked, data = training_set, family = binomial )
summary (model.glm)

##
## Call:
## glm(formula = Survived ~ Embarked, family = binomial, data = training_set)
##
## Deviance Residuals:
##    Min      1Q   Median      3Q     Max
## -1.2802 -0.9105 -0.9105  1.4702  1.4702
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.2384    0.1854   1.286  0.1985
## EmbarkedQ   -0.6279    0.3380  -1.857  0.0632 .
## EmbarkedS   -0.9046    0.2082  -4.345 1.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 889.27 on 667 degrees of freedom
## Residual deviance: 870.12 on 665 degrees of freedom
## AIC: 876.12
##
## Number of Fisher Scoring iterations: 4
```

II·lustració 33: Regressió Logística univariable entre Survived i Embarked

```
odds <- or_glm(data = training_set, model = model.glm)
odds

##
## predictor oddsratio ci_low (2.5) ci_high (97.5)      increment
## 1 EmbarkedQ      0.534       0.272        1.030 Indicator variable
## 2 EmbarkedS      0.405       0.268        0.608 Indicator variable

p_odds <- odds[,1:2]
p_odds["\$"] <- (1-p_odds[2])*100
p_odds

##
## predictor oddsratio %
## 1 EmbarkedQ      0.534 46.6
```

II·lustració 34: Obtenció Odds Ratio Sexe i Embarked

Pel que fa al port d'embarcació, els passatgers que ho van fer pel port Q tenen una probabilitat de supervivència inferior al 46.6% (II·lustració 34) respecte els del port C.

Si comparem aquests dos models veiem com en el cas d'aquest darrer model (II·lustració 33) obtenim un AIC de 876,12 mentre que en l'anterior (II·lustració 31) teníem un AIC de 674.26. El model amb AIC major el podem considerar millor predictor, més explicatiu.

Aquests dos models senzills ens serveixen per veure que seria beneficiós provar de construir un model de regressió amb totes les variables. Per tant, a continuació, hem creat un model de regressió logística multivariable.

#### 4.3.6.2 Model de regressió multivariable

Per tant, hem realitzat un únic model, on la variable dependent era la supervivència, mentre que les variables independents han estat la resta.

```
model_mult1 <- glm( Survived ~ Age + Pclass + Sex + Embarked + FamilySize, data = training_set, family = binomial)
summary(model_mult1)

##
## Call:
## glm(formula = Survived ~ Age + Pclass + Sex + Embarked + FamilySize,
##      family = binomial, data = training_set)
##
## Deviance Residuals:
##    Min      1Q   Median      3Q     Max 
## -2.8741 -0.5743 -0.3811  0.5466  2.5994 
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 4.132122  0.536988  7.695 1.42e-14 ***
## Age        -0.035737  0.009277 -3.852 0.000117 *** 
## Pclass2     -1.037481  0.327963 -3.163 0.001559 **  
## Pclass3     -2.159995  0.311184 -6.941 3.89e-12 *** 
## Sexmale     -2.926782  0.244842 -11.954 < 2e-16 *** 
## EmbarkedQ   -0.106787  0.476860 -0.224 0.822806 
## EmbarkedS   -0.452285  0.287558 -1.573 0.115754 
## FamilySize2 -0.257505  0.294294 -0.875 0.381578 
## FamilySize3  0.540466  0.333020  1.623 0.104606 
## FamilySize4  0.505674  0.605973  0.834 0.404009 
## FamilySize5 -2.337015  0.856940 -2.727 0.006388 ** 
## FamilySize6 -1.910107  0.836816 -2.283 0.022455 *  
## FamilySize7 -1.829604  0.905681 -2.020 0.043369 *  
## FamilySize8 -15.952346 852.217947 -0.019 0.985066 
## FamilySize9 -16.707423 789.572463 -0.021 0.983118 
## ---      
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 889.27 on 667 degrees of freedom
## Residual deviance: 553.57 on 653 degrees of freedom
## AIC: 583.57
##
## Number of Fisher Scoring iterations: 15
```

Il·lustració 35: Primer Model. Regressió logística multivariable de Survived amb Age, Pclass, Sex, Embarked i FamiliSize

En aquest model (Il·lustració 35), hem pogut observar que el valor AIC és inferior a 600, pel que la precisió del model no era gaire alta. A més a més, hem pogut observar que la variable Embarked no és significativa (no síndica amb cap tipus d'asterisc), és a dir, no ajuda a la predicción de la supervivència d'un passatger. Per tant, hem realitzat un segon model de regressió logística multivariable, on hem exclòs l'atribut Embarked.

```

model_mult2 <- glm( Survived ~ Age + Pclass + Sex + FamilySize, data = training_set, family = binomial)
summary(model_mult2)

##
## Call:
## glm(formula = Survived ~ Age + Pclass + Sex + FamilySize, family = binomial,
##      data = training_set)
##
## Deviance Residuals:
##    Min      1Q   Median      3Q     Max
## -2.9503 -0.5696 -0.3935  0.5668  2.6049
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.917381  0.504804  7.760 8.48e-15 ***
## Age        -0.036927  0.009249 -3.993 6.53e-05 ***
## Pclass2     -1.185784  0.315919 -3.753 0.000174 ***
## Pclass3     -2.227388  0.301089 -7.398 1.38e-13 ***
## Sexmale     -2.953632  0.241731 -12.219 < 2e-16 ***
## FamilySize2 -0.245676  0.291507 -0.843 0.399353
## FamilySize3  0.560684  0.330955  1.694 0.090239 .
## FamilySize4  0.495593  0.601026  0.825 0.409611
## FamilySize5 -2.382328  0.828140 -2.877 0.004018 **
## FamilySize6 -2.036635  0.844213 -2.412 0.015845 *
## FamilySize7 -1.983993  0.899596 -2.205 0.027424 *
## FamilySize8 -16.090139 850.901116 -0.019 0.984913
## FamilySize9 -16.862162 786.335428 -0.021 0.982891
##
## ---
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 889.27 on 667 degrees of freedom
## Residual deviance: 556.44 on 655 degrees of freedom
## AIC: 582.44
##
## Number of Fisher Scoring iterations: 15

```

Il·lustració 36: Segon Model. Regressió logística multivariable de Survived amb Age, Pclass, Sex i FamiliSize (Embarked exclòs)

En aquest segon model (Il·lustració 36), el valor AIC ha empitjorat lleugerament, amb un resultat no satisfactori. A més a més, hem observat que la variable FamilySize només és significativa en alguns nivells. Per tant, hem optat per realitzar un tercer model (Il·lustració 37), excloent també l'atribut FamilySize.

```

model_mult3 <- glm( Survived ~ Age + Pclass + Sex , data = training_set, family = binomial)
summary(model_mult3)

##
## Call:
## glm(formula = Survived ~ Age + Pclass + Sex, family = binomial,
##      data = training_set)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.5705 -0.6606 -0.4070  0.6034  2.4821
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.324747  0.416200 7.988 1.37e-15 ***
## Age         -0.029206  0.008144 -3.586 0.000336 ***
## Pclass2     -0.924989  0.297306 -3.111 0.001863 **
## Pclass3     -2.225656  0.285035 -7.808 5.79e-15 ***
## Sexmale     -2.672133  0.214859 -12.437 < 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: 889.27 on 667 degrees of freedom
## Residual deviance: 595.78 on 663 degrees of freedom
## AIC: 605.78
##
## Number of Fisher Scoring iterations: 5

```

II·lustració 37: Tercer Model. Regressió logística multivariable de Survived amb Age, Pclass, Sex (Embarked i FamilySize exclosos)

En aquest tercer model (II·lustració 37), el valor AIC millora per sobre dels 600 on totes les variables són significatives, al menys a un nivell del 1%. Per tant, dels tres models utilitzats, podem concloure que aquest és el més precís de tots.

Tot i així, hem provat de tornar a executar aquest tercer model però amb la variable Age factoritzada. Per a aquest quart model (II·lustració 38), el AIC ha estat el més alt dels quatre i totes les variables són significatives. Per tant, dins dels models de regressió logística multivariables, ens quedem amb aquest quart i últim model.

```

model_mult4 <- glm( Survived ~ Age.factor + Pclass + Sex , data = training_set, family = binomial)
summary(model_mult4)

##
## Call:
## glm(formula = Survived ~ Age.factor + Pclass + Sex, family = binomial,
##      data = training_set)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.5795 -0.6813 -0.4249  0.6173  2.3188
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.2902   0.4230  7.779 7.33e-15 ***
## Age.factorYoung -0.9082   0.3366 -2.699 0.00696 **
## Age.factorAdults -1.1668   0.3584 -3.256 0.00113 **
## Age.factorAdvanced age -1.2954   0.5285 -2.451 0.01424 *
## Age.factorElderly -2.0989   1.1491 -1.827 0.06776 .
## Pclass2     -0.8208   0.2923 -2.808 0.00499 **
## Pclass3     -2.0961   0.2780 -7.541 4.65e-14 ***
## Sexmale     -2.6451   0.2148 -12.314 < 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: 889.27 on 667 degrees of freedom
## Residual deviance: 596.82 on 660 degrees of freedom
## AIC: 612.82
##
## Number of Fisher Scoring iterations: 5

```

II·lustració 38: Quart Model. Regressió logística multivariable de Survived amb Age.factor , Pclass, Sex (Embarked i FamilySize exclosos)

Finalment, hem comprovat la bondat d'aquest quart model, així testejant la predicció a partir del 25% de les dades del data set. Per tal de veure-ho, hem dut a terme una matriu de confusió (Il·lustració 39), la qual ens indica que el model té un 76.2% de precisió.

És a dir, en el 76.2% dels casos, el model predirà de forma correcta si el passatger, donades les variables, sobrevisuria o no al naufragi. En 116 casos, predirà de forma correcte que el passatger no sobrevisuria i, en 54 casos, que sí sobrevisuria. Per altre costat, en 32 ocasions predirà que sí sobrevisrà quan de fet no ho va fer i, en 21 casos, predirà que no sobrevisuria, quan sí ho va fer.

```
# Matriu de confusió
confusionMatrix(as.factor(test_set$Survived2), as.factor(test_set$predict.logit))

## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0     1
##             0 116  21
##             1  32  54
##
##                  Accuracy : 0.7623
##                  95% CI : (0.7009, 0.8166)
##      No Information Rate : 0.6637
##      P-Value [Acc > NIR] : 0.0008825
##
##                  Kappa : 0.4862
##
## McNemar's Test P-Value : 0.1695641
##
##      Sensitivity : 0.7838
##      Specificity : 0.7200
##      Pos Pred Value : 0.8467
##      Neg Pred Value : 0.6279
##      Prevalence : 0.6637
##      Detection Rate : 0.5202
##      Detection Prevalence : 0.6143
##      Balanced Accuracy : 0.7519
##
##      'Positive' Class : 0
##
```

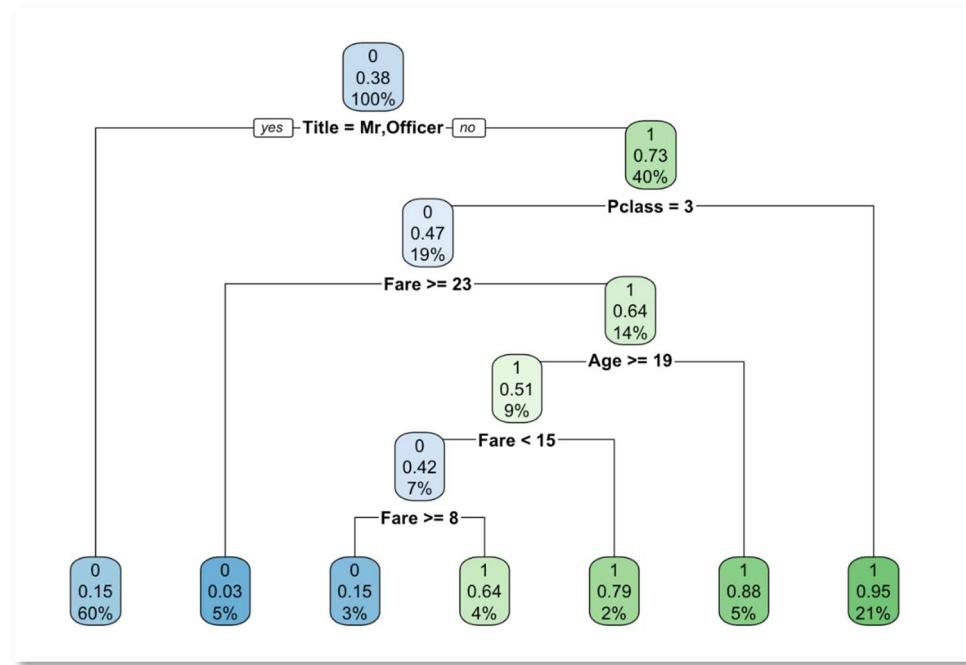
Il·lustració 39: Matriu de Confusió del Quart Model. Anàlisi de la bondat del model

#### 4.3.6.3 Decision Tree Classification

A continuació, hem entrenat un altre tipus de model, un arbre de decisió (Il·lustració 40), per comprovar si aquest tipus de model pot millorar la precisió de l'últim model entrenat.

```
# Entrenament del model
dec_tree_model <- rpart(Survived ~ Age + Sex + Pclass + FamilySize + Embarked + Title + Fare, data = training_set,
method = "class")
# Representació gràfica
rpart.plot(dec_tree_model)
```

Il·lustració 40. Arbre de Decisió



Il·lustració 41. Resultat obtingut de l'arbre de decisió

Aquest model obtingut (Il·lustració 41) prediu la probabilitat de supervivència de la següent manera:

- Comprova si el títol del passatger era Mr. Officer. En cas de ser-ho, només un 15% sobrevisuria al naufragi. Per tant, el 60% de tots els passatgers i tripulació no sobrevisuria.
- En cas de no tenir un títol Mr. Officer, comprova en quin tipus de classe viatjava el passatger. Si era ho feia en tercera classe, el 95% no sobrevisuria, que representa el 21% del total de passatgers.
- Si el passatger viatjava en primera o segona classe, tenia un 47% de supervivència, que representa un 19% del total de passatgers. A continuació, comprova si el passatger va pagar un bitllet igual o més car de 23GBP. En cas que sí, la probabilitat de supervivència era del 3% dins d'aquest tipus de passatgers, mentre que la probabilitat seria d'un 5% tenint en compte tots els passatgers.
- Si el passatger va pagar menys de 23GBP pel bitllet i l'edat era igual o superior a 19 anys, el 88% dels passatgers van morir, que representa un 5% del total.
- Per altre costat, si el passatger tenia una edat inferior a 19 anys i va pagar igual o més de 15GBP pel bitllet, tenia una probabilitat de supervivència del 79%, que representa un 2% del total.
- En cas que pagués menys de 15GBP però igual o més de 8GBP pel bitllet, la probabilitat de supervivència és del 64%, que representa un 4% del total. Per contra, si va pagar menys de 8GBP, tenia un 15% de probabilitats de no sobrevisuire.

Tot seguit per valorar-ne la bondat hem obtingut la matriu de confusió (Il·lustració 42) del mateix:

```

# Matriu de confusió
pred_survived <- predict(dec_tree_model, newdata = test_set[-1], type="class")
table(test_set$Survived,pred_survived)

##   pred_survived
##   0   1
##   0 121 16
##   1  33 53
  
```

Il·lustració 42: Matriu de Confusió de l'arbre de decisió

En 121 ocasions (com s'observa en la II·lustració 42), el model prediu que el passatger no sobrevisuria i no ho va fer. Per contra, en 16 ocasions prediu que no sobrevisuria quan sí ho va fer. En 33 ocasions prediu que sobrevisuria quan no ho va fer. I, en 53 ocasions, prediu de forma correcta que el passatger sobrevisuria.

```
# Avaluació de la precisió
mean(test_set$Survived==pred_survived)

## [1] 0.7802691
```

II·lustració 43: Avaluació de la precisió del model

En comprovar la bondat del model, veiem que té un nivell de precisió (la II·lustració 43) del 78.03%, el qual supera el del model multivariant anterior.

#### 4.3.6.4 Random Forest Classification

Finalment, hem executat un últim model, basat en Random Forest. Aquest model (II·lustració 44) prediu de forma correcta 378 no sobrevisurien al naufragi i 186 que sí ho farien. Per altra banda, en 34 ocasions prediu que no sobrevisuria quan sí ho faria i, en 70, prediu que sí sobrevisuria quan de fet no seria el cas.

```
set.seed(123)
rf_model <- randomForest(Survived~Age+Sex+Pclass+FamilySize+Embarked+Title+Fare, data = training_set)
print(rf_model)

##
## Call:
## randomForest(formula = Survived ~ Age + Sex + Pclass + FamilySize +      Embarked + Title + Fare, dat
a = training_set)
##          Type of random forest: classification
##                  Number of trees: 500
## No. of variables tried at each split: 2
##
##          OOB estimate of  error rate: 15.57%
## Confusion matrix:
##      0   1 class.error
## 0 378  34  0.08252427
## 1   70 186  0.27343750
```

II·lustració 44: Model Random Forest i Matriu de Confusió

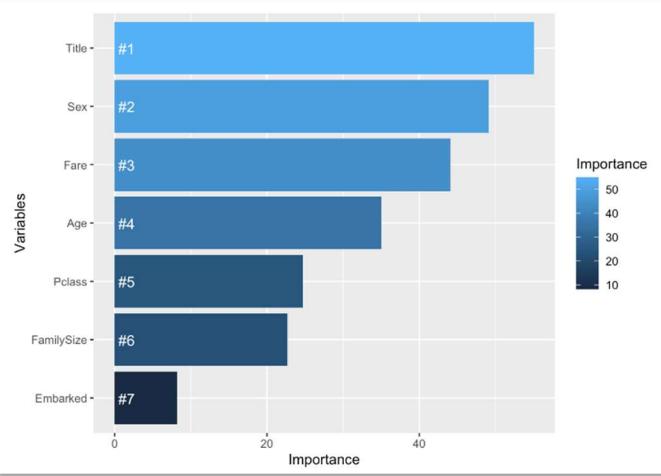
Quan hem comprovat al bondat del model, aquest té una precisió (II·lustració 45) del 80.3%. Per tant, podem concloure que aquest és el millor model de tots per a realitzar prediccions.

```
rf_pred<-predict(rf_model,test_set)
mean(test_set$Survived==rf_pred)

## [1] 0.8026906
```

II·lustració 45: Obtenció de la precisió del model Random Forest

Finalment, també ens ha interessat veure quines són les variables amb més influència (Il·lustració 46) i, per tant, més rellevants dins del model. Aquestes han estat Title, Sex i Fare.



Il·lustració 46: Gràfic amb les variables mes rellevants en el model per predir la supervivència

Recordar que tot el detall més extens de totes les accions realitzades es troben en el fitxer .rmd adjunt. En alguns casos, per no ser massa extensos en aquest document solució no hem inclòs accions repetides per diferents variables.

#### 4.3.7 Prediccions

##### Objectius

En aquest últim punt, l'objectiu és realitzar una predicció a partir del fitxer "test.csv".

##### Resolució

En primer lloc, hem realitzat la predicció del passatger número 71 (Il·lustració 47).

Aquest, tenia un número Id de 962, on el model prediu que el passatger sí que sobrevisuria, donades les diferents característiques.

El model escollit per a fer les prediccions ha estat aquell amb el que hem obtingut prèviament una major precisió: el Random Forest amb una precisió (Il·lustració 45) del 80.3%.

```

prediccio_base <- ds_titanic_test

passatger_individual <- prediccio_base[71,]

rf_kaggle <- predict(rf_model, passatger_individual)
passatger_individual$Survived <- rf_kaggle
head(passatger_individual)

##   PassengerId Survived Pclass          Name     Sex Age SibSp
## 962         962       1      3 Mulvihill, Miss. Bertha E female 24    0
##   Parch Ticket Cabin Embarked file_origin Title FamilySize Age.factor
## 962         0 382653 7.75 <NA>           Q        test  Miss           1     Young
##   FamilySize.factor
## 962             Single

```

Il·lustració 47: Predicció individual pel passatger número 71

Aplicant el model amb millor precisió, el Random Forest Classification, pel passatger 71 amb PassengerId = 962, del qual en desconeixem la supervivència, ens prediu que sobreviurà (Il·lustració 47) amb un valor retornat de Survived 1.

I, finalment, hem aplicat la predicció (Il·lustració 48) de tots els passatgers del fitxer "test.csv". El resultat es troba guardat en un fitxer anomenat "test\_amb\_prediccions.csv" (Il·lustració 49).

```

rf_kaggle<-predict(rf_model,ds_titanic_test)
ds_titanic_test$Survived<-rf_kaggle
head(ds_titanic_test)

##   PassengerId Survived Pclass          Name
## 892         892       0      3 Kelly, Mr. James
## 893         893       0      3 Wilkes, Mrs. James (Ellen Needs)
## 894         894       0      2 Myles, Mr. Thomas Francis
## 895         895       0      3 Wirz, Mr. Albert
## 896         896       1      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist)
## 897         897       0      3 Svensson, Mr. Johan Cervin
##   Sex Age SibSp Parch Ticket  Fare Cabin Embarked file_origin Title
## 892 male 34.5    0    0 330911 7.8292 <NA>      Q      test    Mr
## 893 female 47.0   1    0 363272 7.0000 <NA>      S      test    Mrs
## 894 male 62.0    0    0 240276 9.6875 <NA>      Q      test    Mr
## 895 male 27.0    0    0 315154 8.6625 <NA>      S      test    Mr
## 896 female 22.0   1    1 3101298 12.2875 <NA>      S      test    Mrs
## 897 male 14.0    0    0  7538 9.2250 <NA>      S      test    Mr
##   FamilySize Age.factor FamilySize.factor
## 892         1      Adults            Single
## 893         2      Adults            Small
## 894         1 Advanced age          Single
## 895         1      Young            Single
## 896         3      Young            Small
## 897         1 Children           Single

```

Il·lustració 48: Predicció per la totalitat del fitxer test.csv proporcionat pel repte Kaggle

```
write.csv(ds_titanic_test[,c(1,2)],file="../Fitxers/test_amb_prediccions.csv", row.names=FALSE)
```

Il·lustració 49: Creació del fitxer csv amb prediccions en format Kaggle

## 5. Representació dels resultats a partir de taules i gràfiques

### 5.1 Percentatges de Supervivents i No Supervivents

En el cas de l'anàlisi amb el conjunt de dades d'entrenament (II·lustració 50) hem obtingut:

	Survived	Percentatge
## 1:	No	61.61616
## 2:	Si	38.38384

II·lustració 50: Percentatge de Supervivència obtingut en el DataSet d'entrenament

S'observa com tan sòls el 38.38 % dels passatgers van aconseguir sobreuir.

Repetint el mateix anàlisi amb el conjunt de dades de test/predicció (II·lustració 51) hem obtingut:

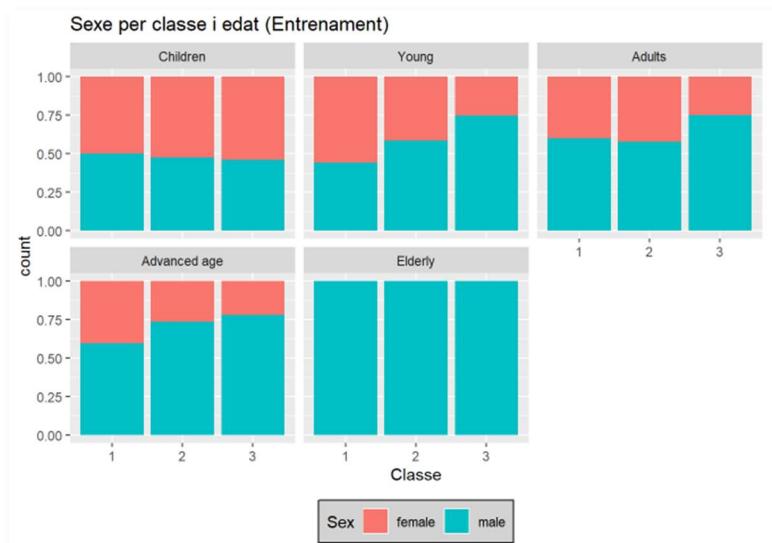
	Survived	Percentatge
## 1:	No	63.8756
## 2:	Si	36.1244

II·lustració 51: Percentatge de Supervivència obtingut en el DataSet de predicció

En aquest cas s'observa com tan sòls el 36.12 % dels passatgers van aconseguir sobreuir. Si bé el percentatge és un xic inferior en cap cas obtenim uns valors esbiaixats.

### 5.2 Gràfics del Conjunt d'Entrenament

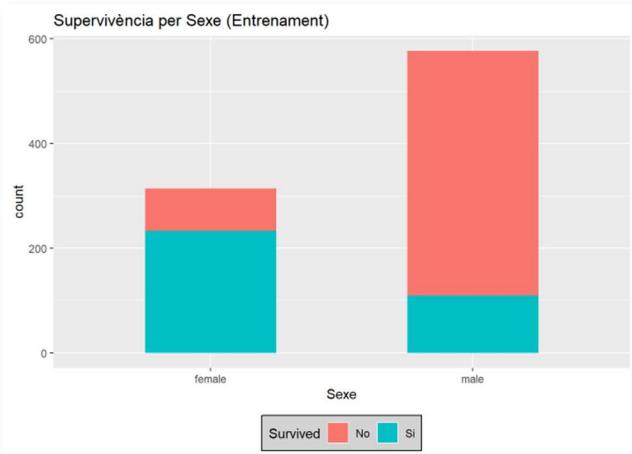
Distribució de la supervivència en relació al Sexe, Classe i Edat



II·lustració 52: Distribució de la supervivència en relació al Sexe, Classe i Edat (DataSet d'Entrenament)

En aquesta distribució (II·lustració 52) podem veure com a mida que l'edat és major i anem pujant de classe social cada cop es fa més predominant el gènere masculí.

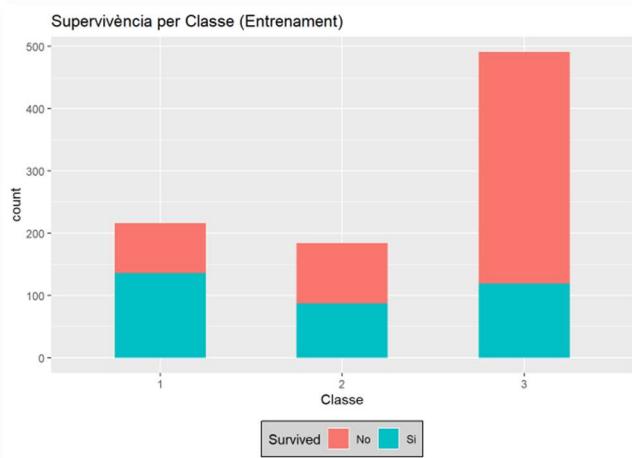
### Supervivència per Sexe



Il·lustració 53: Distribució de la Supervivència per Sexe (DataSet d'entrenament)

Per gènere s'observa (Il·lustració 53) amb claredat com el sexe femení és el que va tenir un major nombre de supervivents.

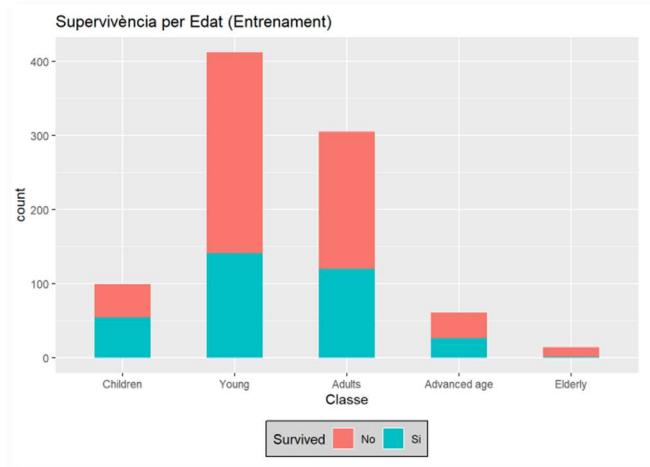
### Supervivència per Classe



Il·lustració 54: Distribució de la Supervivència per Classe (DataSet d'entrenament)

També s'observen (Il·lustració 54) diferències per classe. Veiem que mentre en segona classe la proporció entre els passatgers supervivents i els que van morir està força igualat en primera i tercera això canvia. En percentatge els passatgers amb una major proporció de supervivents van ser els de primera classe (classe alta) i els que menys, entre els de tercera classe (classe baixa).

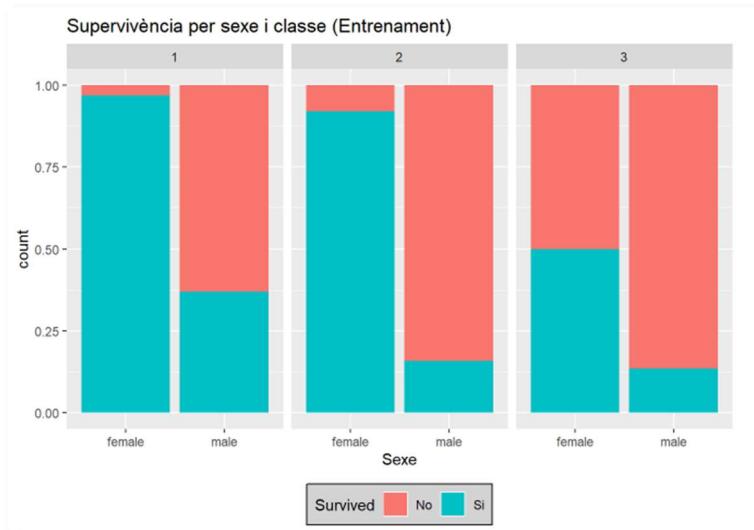
### Supervivència per Edat



II·lustració 55: Distribució de la Supervivència per Edat (DataSet d'entrenament)

La gràfica (II·lustració 55) mostra com la supervivència va ser superior entre els nens.

### Supervivència per Sexe i Classe



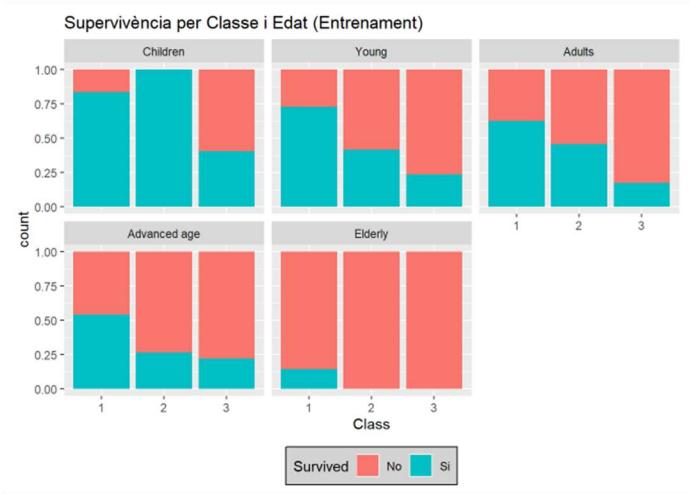
II·lustració 56: Distribució de la Supervivència per Sexe i Classe (DataSet d'entrenament)

Observem (II·lustració 56) com els passatgers de gènere femení són els que en major proporció van sobreviure si bé s'observa com a mida que la classe social augmenta, aquest percentatge decau. A pitjor classe social major mortalitat.

En el cas dels homes veiem com el percentatge de supervivència està molt per sota que el de les dones i també es fa evident com a mida que empitjora la classe social, menors són les probabilitats de sobreviure.

Veiem com entre els passatgers del classe alta i gènere femení són els que hi va haber més supervivents.

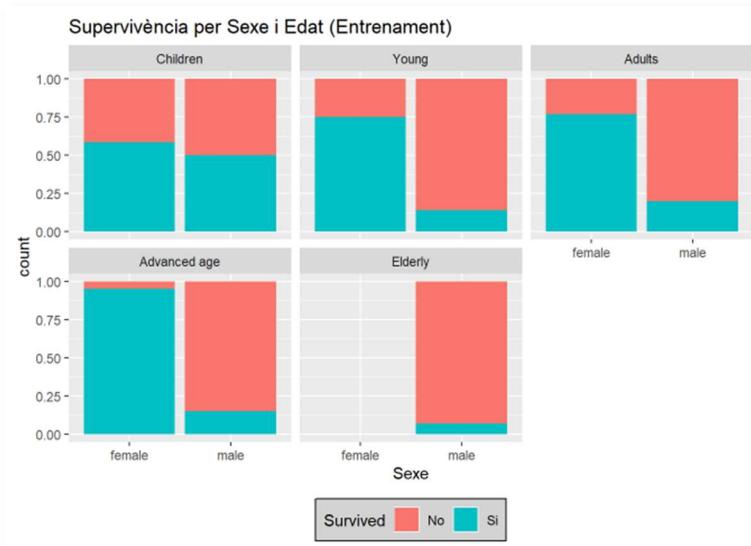
### Supervivència per Classe i Edat



Il·lustració 57: Distribució de la Supervivència per Classe i Edat (DataSet d'entrenament)

Analitzant per Classe i Edat (Il·lustració 57) es fa evident que a mida que empitjora la classe social a la que pertany el passatger, les seves opcions de supervivència són cada cop menors.

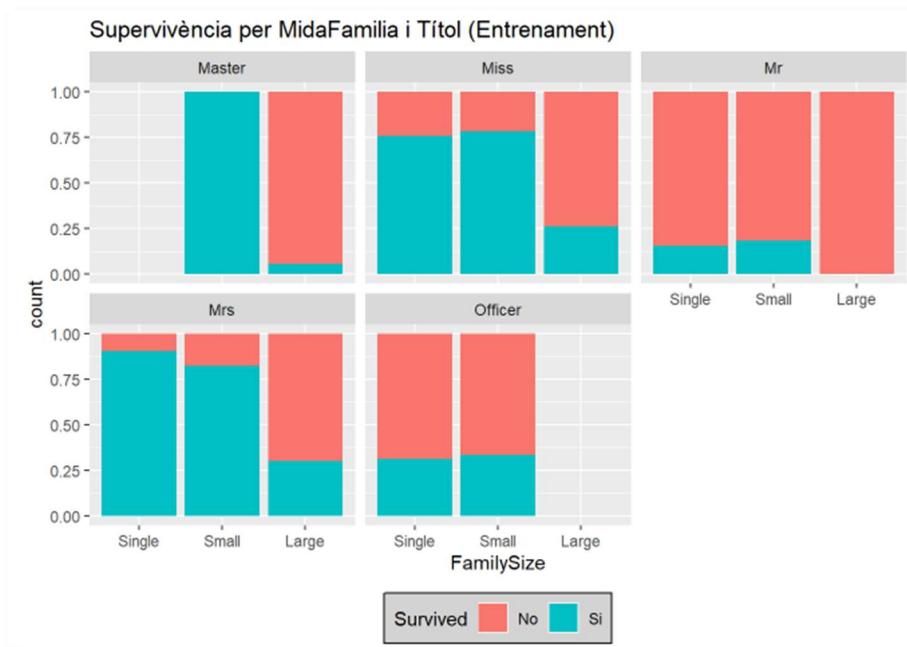
### Supervivència per Sexe i Edat



Il·lustració 58: Distribució de la Supervivència per Sexe i Edat (DataSet d'entrenament)

S'observa (Il·lustració 58) com a mida que les dones, que continuen apareixent com aquelles que amb major proporció sobrevenen, tenen major edat majors percentatge de supervivència obtenen.

En el cas dels homes veiem com la supervivència entre els nens de genere masculí està tot just una mica per sota del de les nenes. En la resta de franges d'edat veiem com la supervivència dels homes decau a mida que s'incrementa l'edat exceptuant aquelles que es troben en el grup d'adults en que està un xic millor.

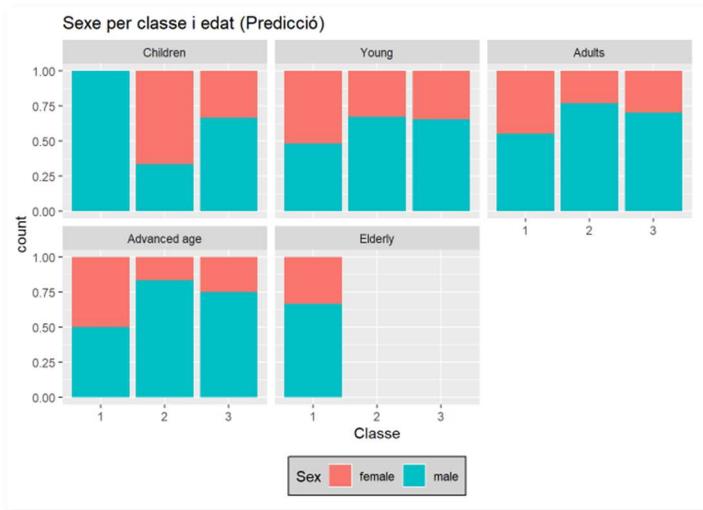
Supervivència per Mida Família i Títol

Il·lustració 59: Distribució de la Supervivència per Mida Família i Títol (DataSet d'entrenament)

Passatgers (Il·lustració 59) que formaven part de famílies amb més de 4 integrants són els que menys probabilitat de supervivència van tenir. Si mirem per Títol es repeteix el patró que mostra com infants i dones (Master, Miss i Mrs) són els que van tenir també major probablitat de sobreviure.

### 5.3 Gràfics del Conjunt de Test o Predicció

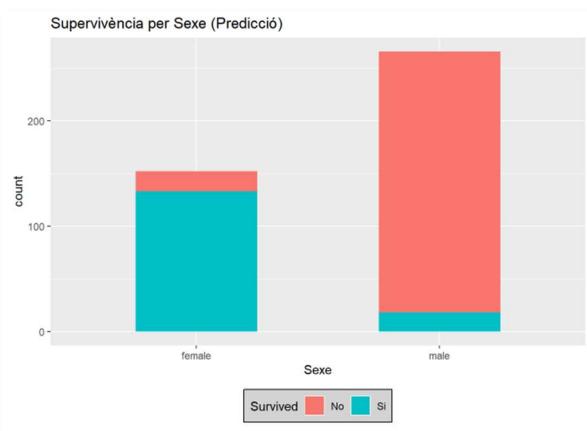
#### Distribució de la supervivència en relació al Sexe, Classe i Edat



Il·lustració 60: Distribució de la supervivència en relació al Sexe, Classe i Edat (DataSet de Predicció)

En aquesta distribució (Il·lustració 60), per la gent pertanyent al grup d'edat 'Elderly' veiem com no hi han persones en la segona i tercera classe. En canvi si que trobem en el grup d'edat 'Elderly' algun passatger del gènere femení. A banda podem observar com no hi ha infants de gènere femení en primera classe.

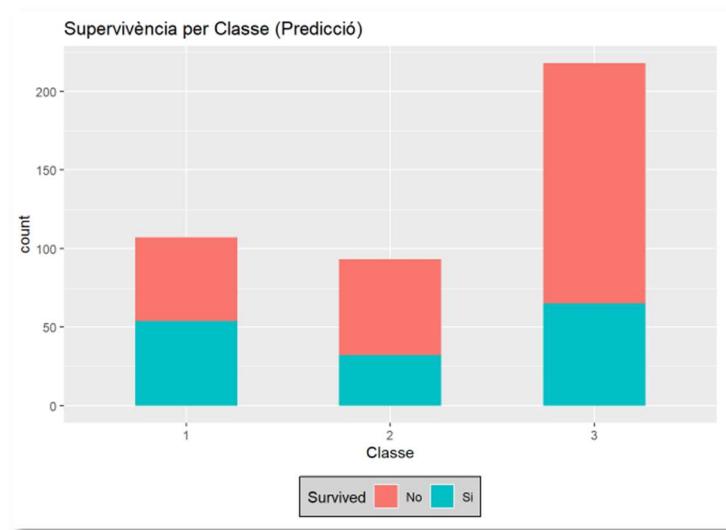
#### Supervivència per Sexe



Il·lustració 61: Distribució de la supervivència per Sexe (DataSet de Predicció)

De manera més accentuada (Il·lustració 61) veiem com el percentatge de supervivents entre els passatgers de sexe femení va ser més elevat.

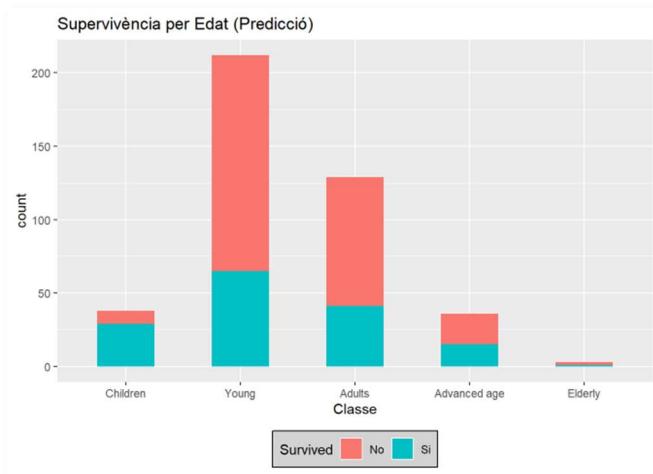
### Supervivència per Classe



Il·lustració 62: Distribució de la Supervivència per Classe (DataSet de Predicció)

Els resultats (Il·lustració 62) varien mínimament respecte als obtinguts amb les dades d'entrenament i es manté la cruel realitat: a millor classe social majors probabilitats de supervivència.

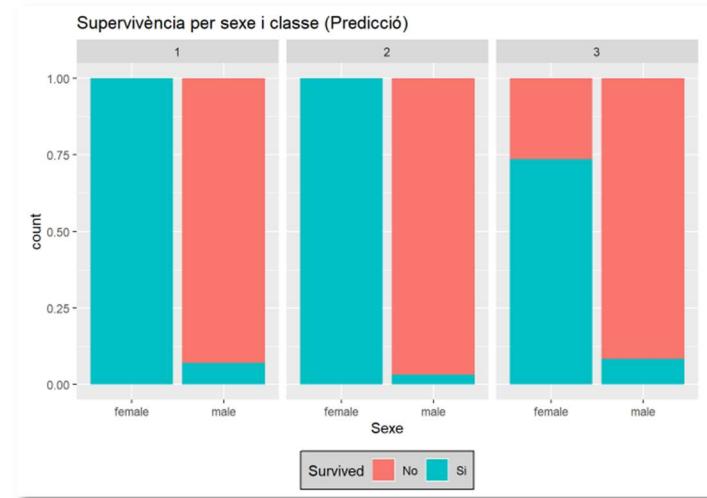
### Supervivència per Edat



Il·lustració 63: Distribució de la Supervivència per Edat (DataSet de Predicció)

De forma destacada (Il·lustració 63) s'observa com la supervivència va ser major entre els nens.

### Supervivència per Sexe i Classe



II·lustració 64: Distribució de la Supervivència per Sexe i Classe (DataSet de Predicció)

Les observacions fetes pel conjunt de dades d'entrenament (II·lustració 64) s'accentuen en aquest cas. Veiem com es manté el fet que els supervivents en proporció majoritàriament van ser dones. Com a diferència s'observa com al respecte del gràfic amb el conjunt de dades d'entrenament i per la Classe Mitjana sembla que les dones han incrementat lleugerament el percentatge de supervivència.

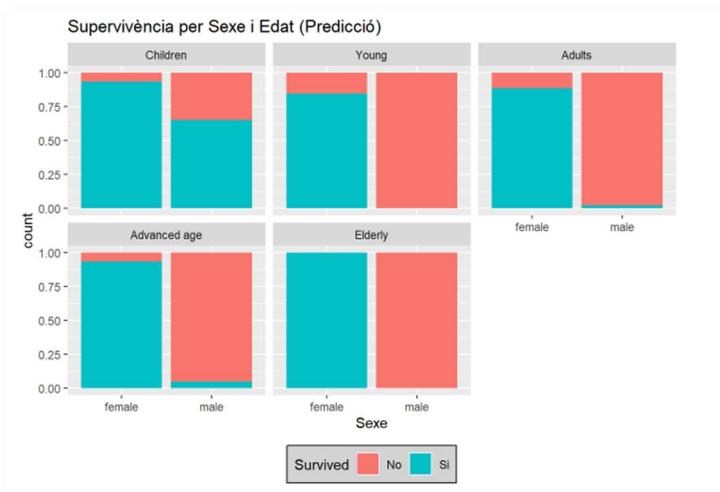
### Supervivència per Classe i Edat



II·lustració 65: Distribució de la Supervivència per Classe i Edat (DataSet de Predicció)

Analitzant per Classe i Edat (II·lustració 65) es fa evident que a mida que empitjora la classe social a la que pertany el passatger, les seves opcions de supervivència són cada cop menors.

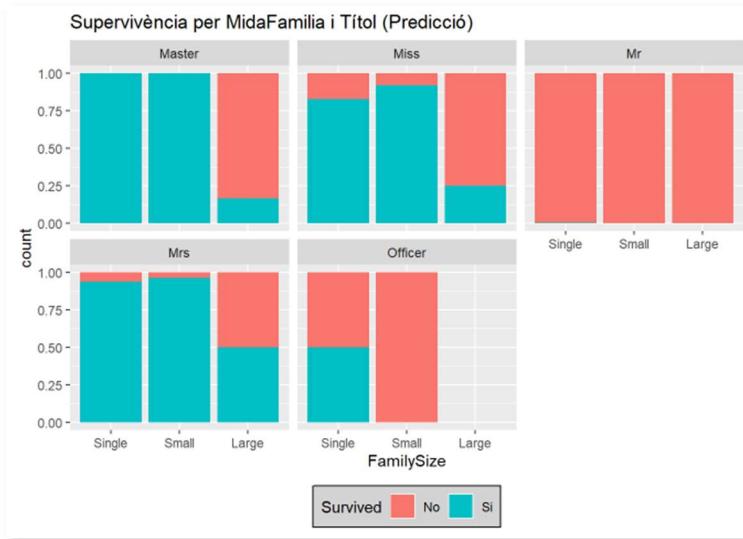
### Supervivència per Sexe i Edat



II·lustració 66: Distribució de la Supervivència per Sexe i Edat (DataSet de Predicció)

S'observa (II·lustració 66) com a mida que les dones, que continuen apareixent com aquelles que amb major proporció sobreviuen. En el cas dels homes veiem com les seves probabilitats de supervivència apareixen com a nul·les independentment del grup d'edat al que pertanyen exceptuant els nens.

### Supervivència per Mida Família i Títol



II·lustració 67: Distribució de la Supervivència per Mida Família i Títol (DataSet de Predicció)

Passatgers (II·lustració 67) que formaven part de famílies amb més de 4 integrants són els que menys probabilitat de supervivència van tenir. Si mirem per Títol es repeteix el patró que mostra com infants i dones (Master, Miss i Mrs) són els que van tenir també major probabilitat de sobreuir. Destacar el fet que sembla que els passatgers etiquetats com a Mr, és a dir homes adults, són els que pitjor probabilitat de supervivència tenien.

## 6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

L'explicació dels resultats s'ha anat detallant durant l'elaboració i obtenció resultats del projecte desenvolupat en R.

A nivell de resum s'exposen tot seguit les conclusions obtingudes.

- Ha estat necessari revisar i netejar el conjunt de dades:

Ha estat necessari analitzar la presència de valors extrems i plantejar solucions pels diferents valors perduts detectats. Sobre aquest darrer punt i investigant per internet hem vist que existeixen múltiples opcions i realment com a tal és tot un art el decidir o cercar el millor mecanisme per a informar-los.

- Reducció de dimensionalitat:

En aquest cas el nombre de mostres no era molt elevat però si que hem reduït la dimensionalitat del dataset original doncs s'ha analitzat camp a camp la seva utilitat i hem deixat de banda aquells que hem considerat que no ens aportaven informació interessant per l'estudi.

- Creació d'atributs addicionals:

Hem creut interessant crear un parell d'atributs nous 'Feature Engineering' a partir dels ja existents per mirar d'ajustar al màxim el model predictor. D'aquesta manera hem creat un atribut FamilySize i Title.

- S'observa com la supervivència va lligada a Edat, Sexe, Classe Social, Port d'embarcament o Mida de la Família.

Arran dels resultats obtinguts amb el tests Chi-quadrat veiem com amb aquests cinc atributs es descarta la hipòtesi nul·la i per tant, es descarta la independència entre la relació de supervivència i cadascun d'ells.

- A partir de les gràfiques obtingudes durant l'anàlisi de les dades:

- La probabilitat de supervivència en dones és major que en homes. Dins del genere homes, els que estaven per sota de 18 anys tenien una major probabilitat de supervivència.
- Les dones van tenir un 75% de ratio de supervivència mentre que en el cas dels homes el ratio va ser < 25%.
- S'observa com els passatgers de Primera Classe tenien un ratio de supervivència superior al 50%, mentre que els de segona tenen un ratio al voltant del 50%. Per altre costat, els de tercera classe tenien una probabilitat del 25%.
- Sembla que hi ha una forta correlació entre fare i taxa de supervivència. Aquells passatgers que van pagar menys de 50GBP tenien una taxa de supervivència inferior al 50%. A mesura que el passatger va pagar un ticket amb un preu més elevat, la taxa de supervivència augmenta.
- Hem observat que la majoria de passatgers eren joves d'entre 15 i 29 anys.
- Hem vist que dels passatgers que no van sobreviure, aquells que viatjaven en unitats familiars de més de 4 membres (Large) són els que amb menor percentatge van sobreviure al naufragi. Les probabilitats de supervivència van ser majors en el cas de passatgers que formaven part de famílies petites de fins a 4 membres.

- A la pregunta que ens hem fet: La mitjana d'edat dels supervivents és inferior als 20 anys?

A partir d'un contrast d'hipòtesis emprant el test U de Mann-Whitney hem arribat a la conclusió que aquesta és certa.

- Models de regressió creats per crear models:

A partir dels atributs que hem considerat més interessants hem creat tota una colla de models amb el propòsit de trobar aquell que sigui més precís en la tasca de predicción de supervivència dels passatgers.

Hem creat diferents models analitzant com milloraven (AIC) i veient quins eren els atributs més rellevants i descartant aquells que ens aportaven menys en la construcció del model.

- Model creat emprant un arbre de decisió:

Hem creat un model a partir d'un arbre de decisió que ens ha ajudat visualment a veure quins són els atributs que el model considerava més importants a mida que anava construint nodes i fulles.

- Comprova si el títol del passatger era Mr. Officer. En cas de ser-ho, només un 15% sobrevisuria al naufragi. Per tant, el 60% de tots els passatgers i tripulació no sobrevisuria.
- En cas de no tenir un títol Mr. Officer, comprova en quin tipus de classe viatjava el passatger. Si era ho feia en tercera classe, el 95% no sobrevisuria, que representa el 21% del total de passatgers.
- Si el passatger viatjava en primera o segona classe, tenia un 47% de supervivència, que representa un 19% del total de passatgers. A continuació, comprova si el passatger va pagar un bitllet igual o més car de 23GBP. En cas que sí, la probabilitat de supervivència era del 3% dins d'aquest tipus de passatgers, mentre que la probabilitat seria d'un 5% tenint en compte tots els passatgers.
- Si el passatger va pagar menys de 23GBP pel bitllet i l'edat era igual o superior a 19 anys, el 88% dels passatgers van morir, que representa un 5% del total.
- Per altre costat, si el passatger tenia una edat inferior a 19 anys i va pagar igual o més de 15GBP pel bitllet, tenia una probabilitat de supervivència del 79%, que representa un 2% del total.
- En cas que pagués menys de 15GBP però igual o més de 8GBP pel bitllet, la probabilitat de supervivència és del 64%, que representa un 4% del total. Per contra, si va pagar menys de 8GBP, tenia un 15% de probabilitats de no sobrevisuire.

- Model creat emprant un Random Forest

- Amb aquest model hem obtingut la millor precisió: un 80.3%. Mostrant el resum una estimació de l'error del 15.57%.
- A partir d'aquest model hem vist quins considera que són els atributs més rellevants que han estat de major a menor ordre: Title, Sex, Fare, Age, Pclass, FamilySize, Embarked.
- Amb aquest model hem realitzat la predicción del fitxer test a predir. Hem fet la predicción puntual per un passatger i posteriorment per la seva globalitat.

**7. 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.**

La manipulació de les dades s'ha efectuat en R, que es pot trobar en la carpeta “Codi” del repositori del Github del següent enllaç: <https://github.com/cfont03/Titanic-PRA2>

En aquesta mateixa carpeta hi ha disponible el codi en .rmd i en .html. Els fitxers originals, els transformats amb les dades netes i el fitxer final generat es troben en la carpeta Fitxers.

**8. Participació dels integrants**

Contribucions	Firma dels integrants
Investigació prèvia	Jordi Dil i Giró & Carlota Font Castell
Redacció de les respostes	Jordi Dil i Giró & Carlota Font Castell
Desenvolupament del codi	Jordi Dil i Giró & Carlota Font Castell

**9. Bibliografia**

1. Titanic: Machine Learning from Disaster | Kaggle [Internet]. Kaggle.com. 2020 [citat 7 de desembre 2020]. Disponible en l'enllaç: <https://www.kaggle.com/c/titanic/overview>