Pràctica 2: Neteja i anàlisi de les dades
Document de respostes

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

Seguint les principals etapes d'un projecte analític, les diferents tasques a realitzar (i justificar) són les següents:

1 DESCRIPCIÓ DEL DATASET (0,5 PUNTS)

Hem decidit acceptar una de les propostes de datasets que se'ns ha ofert, i treballarem les dades corresponents al Titanic: Machine Learning from Disaster.

L'obtenim accedint a la web: https://www.kaggle.com/c/titanic

Aquest dataset es presenta en format de fitxer .csv dividit en 3 parts:

- Train.csv:

Es tracta d'un fitxer de 891 files i 12 columnes.

Cada fila correspon a les dades d'un passatger del creuer. I de cada mostra, a cada fila el següent atribut, per ordre de columna:

- PassengerId: Identificador únic de cada passatger.
- Survived: indica si el passatger va sobreviure (valor =1) o si no va sobreviure (valor =0).
- Pclass: indica la classe a la que correspon el bitllet: El valor 1 correspon a la millor categoria (alta), el valor 2 a la classe mitja i el valor 3 correspon a la classe baixa.
- Name: Nom del passatger en format cognom, nom majoritàriament, però no sempre.
- Sex: Sexe del passatger (male / female)
- Age: Edat en anys, majoritàriament valors enters, però no tots. Són en valors decimals els valors menors de 1 i els que han estat aproximats es registren com a xx.5

- SibSp: nombre de germans/cunyats o cònjuges a bord.
- Parch: nombre d'ascendents (pare/mare) o descendents(fill/filla/gendre/nora) a bord.
- Ticket: nombre de passatge/bitllet
- Fare: tarifa del passatger
- Cabin: nombre de camarot assignat
- Embarked: port d'embarcament (C = Cherbourg, Q = Queenstown, S = Southampton)

- Test.csv:

Es tracta d'un fitxer de 418 files i 11 columnes. Les mateixes columnes que en traint, però sense la columna corresponent a la informació sobre la supervivència del passatger (Survived).
La descripció de tots els altres atributs és igual que per al fitxer train.csv

- Gender_submission.csv:

El dataset és usat per participar en un concurs. Aquest fitxer correspon a la predicció sobre la supervivència o no dels passatgers inclosos al fitxer de test. No és la informació real sinó un exemple del fitxer que cal entregar en cas de participar en el concurs. En concret es correspon a una predicció basada únicament en el sexe (si és dona sobreviu, si és home, no).

Tenint en compte aquesta descripció, hem de descartar el contingut del fitxer 'gender_submission.csv' donat que les dades que conté no són correctes.

Aleshores, ens trobem davant de la situació que les úniques dades complertes de les que disposem són les que hi ha en el fitxer train.csv, ja que les dades del fitxer test.csv no contenen informació sobre la supervivència.

Decidim, doncs, treballar només amb les dades del fitxer train i si s'arriba a alguna conclusió en base a les anàlisis que en farem, podrem usar les dades del fitxer test per fer la nostra pròpia predicció.

1.1 IMPORTÀNCIA I OBJECTIUS DE LES ANÀLISIS

Perquè és important i quina pregunta/problema pretén respondre?

Les dades contingudes en el data set permeten analitzar si existeixen variables que condicionin la supervivència dels passatgers del creuer.

Per això la pregunta que pretenem respondre amb l'anàlisi d'aquestes dades és:

L'edat, el sexe i la classe social tenen incidència en el grau de supervivència en el Titànic?

2 INTEGRACIÓ I SELECCIÓ (0,5 PUNTS)

Després d'analitzar els tres fitxers disponibles a l'apartat anterior, hem decidit fer l'anàlisi de les dades sobre el conjunt d'entrenament. Si el conjunt de test tingués la variable de classificació "Survived", l'haguéssim pogut integrar i construir un data set amb més registres, però no ha estat el cas.

Partim, doncs, d'un dataframe de 891 observacions de 12 atributs cadascun.

```
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen,
Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
 $ Sex        : chr  "male" "female" "female" "female" ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : chr  "" "C85" "" "C123" ...
 $ Embarked   : chr  "S" "C" "S" "S" ...
```

```
  PassengerId       Survived          Pclass          Name
Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
Median :446.0   Median :0.0000   Median :3.000   Mode  :character
Mean   :446.0   Mean   :0.3838   Mean   :2.309
3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
Max.   :891.0   Max.   :1.0000   Max.   :3.000

    Sex                 Age            SibSp           Parch
Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
Mode  :character   Median :28.00   Median :0.000   Median :0.0000
                   Mean   :29.70   Mean   :0.523   Mean   :0.3816
                   3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
                   Max.   :80.00   Max.   :8.000   Max.   :6.0000
                   NA's   :177
   Ticket               Fare          Cabin              Embarked
Length:891         Min.   :  0.00   Length:891         Length:891
Class :character   1st Qu.:  7.91   Class :character   Class :character
Mode  :character   Median : 14.45   Mode  :character   Mode  :character
                   Mean   : 32.20
                   3rd Qu.: 31.00
                   Max.   :512.33
```

A continuació, detallem la reducció que farem del conjunt de dades, tenint en compte quines són les variables útils per al nostre anàlisis. Encara que no seria necessari fer-ho, ja que no és un conjunt de dades amb un volum de registres elevats i el cost temporal de les operacions és molt baix, hem decidit deixar-ho preparat d'aquesta forma.

Les variables que eliminem, després de fer una inspecció del data set, són les següent: PassengerId, Name, SibSp, Parch, Ticket, Fare, Cabin i Embarked.

Les variables que seleccionem són aquelles que realment ens poden ajudar a respondre la nostra pregunta: Survived, Age, Sex i Pclass.

```
'data.frame':   891 obs. of  4 variables:
 $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass  : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Sex     : chr  "male" "female" "female" "female" ...
 $ Age     : num  22 38 26 35 35 NA 54 2 27 14 ...
```

És possible que, en el cas d'haver de fer una predicció de la classificació del conjunt de test, altres variables tinguin també un pes important en la decisió, però ens hem centrat en dades demogràfiques.

3 NETEJA DE LES DADES (2 PUNTS)

Després del procés de selecció el nostre dataset té només les variables principals d'interès, segons l'anàlisi efectuat:

- Edat
- Sexe
- Classe
- Supervivència

3.1 VALORS PERDUTS

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

Efectivament, després de comprovar-ho ens trobem que una de les variables, l'edat, té valors perduts.

Són 177 valors perduts. De fet, els valors perduts eren més, ja que la informació que tenim ens indica que totes les edats que tenen format xx.5 són aproximacions. Com que considerem que aquesta aproximació segur que es basa en informació que va portar a determinar aquella edat, aquests valors que no són reals però sí aproximats, es consideraran vàlids.

Els casos amb valors buits es poden gestionar simplement eliminant-los del data set. En aquest cas els valors perduts són gairebé un 20% de les mostres, per tant, eliminar-los suposaria una enorme pèrdua d'informació.

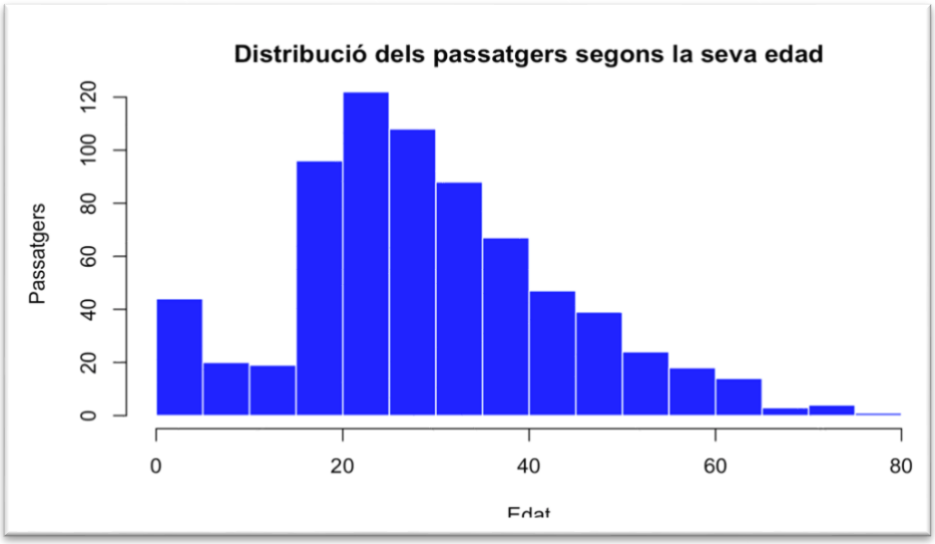
A més, en aquestes mostres la informació de l'edat és la única que no està disponible, per tant, eliminar-los significaria no poder fer ús de les altres dades sí disponibles.

Es decideix, per tant, que se substituiran per valors aproximats.

No considerem que sigui vàlida la opció de buscar valors propers de les altres mostres (per exemple amb un kNN) perquè no són valors relacionats i per tant, no es pot assumir que la resta de variables determini el valor d'aquest atribut.

S'opta per substituir els valors perduts per un valor central. Una opció hagués estat substituir-los pel valor "28" que és la mediana de totes les mostres.

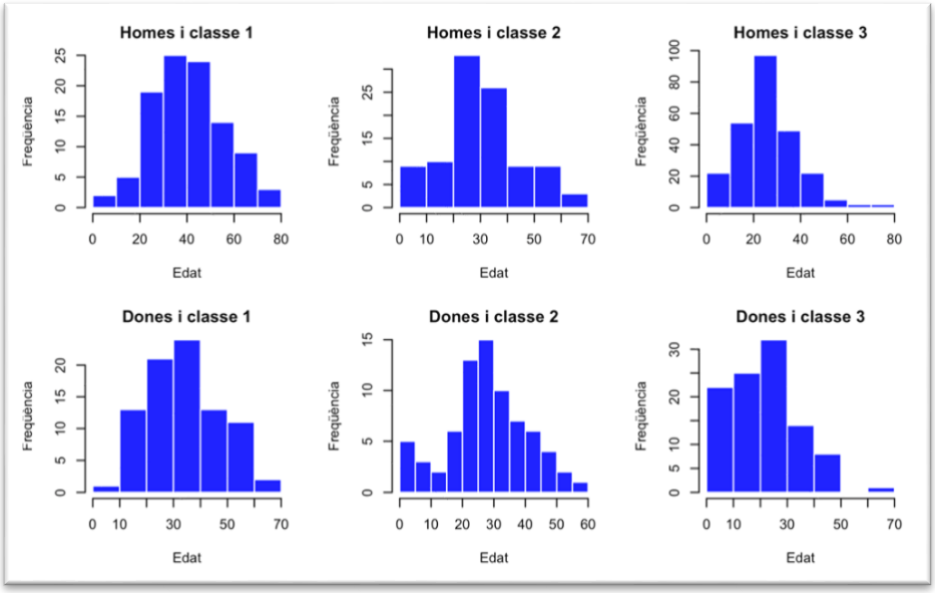
S'efectuen histogrames i matrius de distribució i s'observa que l'edat no es distribueix de forma normal.



A més, que no hi ha el mateix nombre d'individus en cada una de les combinacions sexe - classe.

```
        1   2   3
female 94  76 144
male  122 108 347
```

I cada combinació de sexe-classe té una distribució d'edats diferent.



Per això es decideix calcular la mediana de cada tipus d'individu i imputar a cada mostra el valor que li correspon segon el tipus d'individu.

- Mediana de les mostres de dones de classe 1: 35 anys
- Mediana de les mostres de dones de classe 2: 28 anys
- Mediana de les mostres de dones de classe 3: 21.5 anys
- Mediana de les mostres d'homes de classe 1: 40 anys
- Mediana de les mostres d'homes de classe 2: 30 anys
- Mediana de les mostres d'homes de classe 3: 25 anys

S'observa que, com més alta és la classe, més alta és l'edat mediana. I els homes tenen medianes majors, per cada classe.

Això reforça de decisió d'assignar a cada mostra l'edat en funció del sexe i classe que tingui

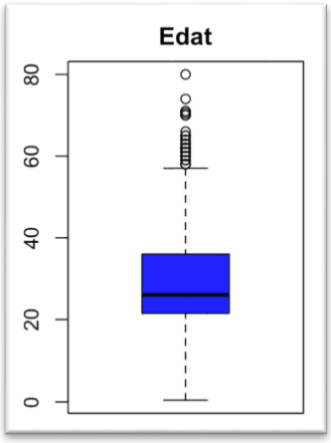
3.2 VALORS EXTREMS

Identificació i tractament de valors extrems

La presència de valors extrems només es podria donar en la variable 'edat'.

Els altres tres atributs són categories i tenen 2 o 3 valors disponibles per tant, no té sentit parlar de valors extrems.

Utilitzem un boxplot per determinar els valors extrems. Tant visualment com els seus resultats estadístics.



Descobrim que tenim 33 valors extrems, però que tots ells estan entre els 58 i els 80 anys.

Donat que aquests valors no són estranys i que probablement es consideren outliers perquè hi havia una majoria de passatge d'edat molt jove, es decideix tractar aquests valors com a vàlids i no fer-hi cap eliminació o modificació.

4 ANÀLISI DE LES DADES (2,5 PUNTS)

4.1 SELECCIÓ

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

Ja s'ha comentat abans que la selecció de les dades s'ha efectuat en una fase prèvia, abans de netejar-les i detectar blancs i extrems.

Ens quedem només amb 4 variables que puguin tenir a veure amb la supervivència del passatger, efectuant una reducció de la dimensionalitat.

En aquest cas és una reducció de la dimensionalitat simple, eliminant columnes innecessàries. No ens hem vist en la necessitat de reduir la dimensionalitat aplicant tècniques d'anàlisi de components principals.

No ens és necessari fer una reducció de la quantitat, ja que només son 891 mostres. No és una quantitat que requereixi mostrejar de cap manera.

4.2 COMPROVACIONS

Comprovació de la normalitat i homogeneïtat de la variància

Aquestes comprovacions només es poden fer per variables numèriques, no per a variables binàries o categòriques. Per tant, només hem pogut fer-ho sobre l'edat, un cop hem imputat valors als registres que tenien nuls.

En concret, les proves que hem efectuat sobre l'atribut de l'edat han estat:

4.2.1 SHAPIRO TEST

Amb això aconseguim comprovar si la variable té una distribució normal. Encara que ja havíem fet un histograma complert, aquest test és, clar, molt més precís.

```
        Shapiro-Wilk normality test

data:  data$Age
W = 0.96548, p-value = 1.118e-13
```

Amb un p-valor <0.05 rebutgem la Hipòtesi nul·la que establia la normalitat.

L'edat no es distribueix de forma normal.

4.2.2 TEST DE VARIÀNCIA (F-TEST)

Fem també un F-Test per analitzar si la variància de l'edat entre les mostres que sobreviuen li les que no sobreviuen es pot considerar igual.

La hipòtesi nul·la és que les dues variàncies són iguals. I pel que veiem al test, encara que la variància en la mostra té una relació no igual (0.83 enlloc de l'ideal 1), també veiem que el p-valor és major de 0.05 i, per tant, no podem rebutjar la hipòtesi. El valor 1 està dins de l'interval de confiança, encara que cal reconèixer que per poquet.

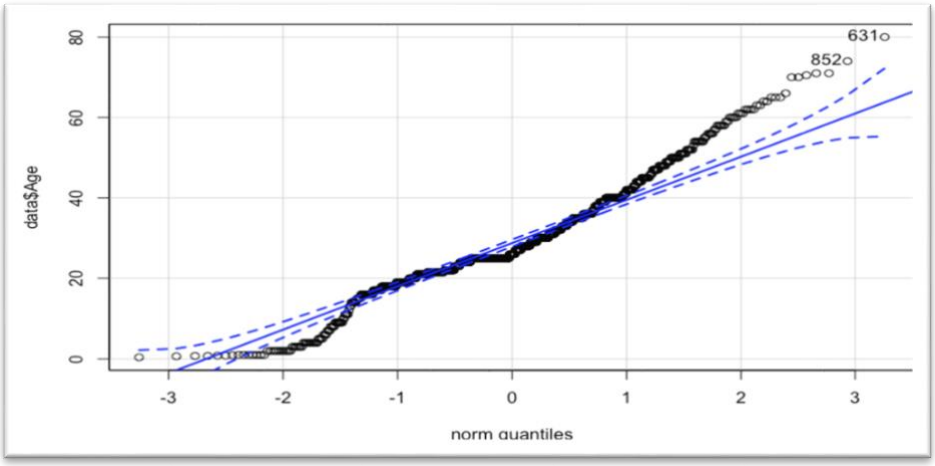
```
        F test to compare two variances

data:  data[data$Survived == "0", "Age"] and data[data$Survived == "1", "Age"]
F = 0.83704, num df = 548, denom df = 341, p-value = 0.06553
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6897026 1.0114736
sample estimates:
ratio of variances
         0.8370418
```

4.2.3 QQ PLOT

Per a comprovar la heterocedasticitat de la variable, efectuem un qqplot.

El gràfic QQplot compara la nostra distribució amb els quartils d'una distribució normal. De manera que com més s'aproximi a la recta, més podrem dir que s'assembla a una distribució normal.



Podem observar la confirmació del que ja ens feia intuir l'histograma que hem fet al començament. Que entre els 15 i els 40 anys la distribució sembla normal, però que els extrems (menors de 15 i majors de 40) no es comporten de cap manera com una distribució normal.

4.3 PROVES ESTADÍSTIQUES

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

En la comprovació de la normalitat i homogeneïtat de la variància ja hem fet servir algun contrast d'hipòtesis, per exemple, una prova estadística que ja hem resolt és si la variància de l'edat és igual entre la població supervivent que entre la població no supervivent.

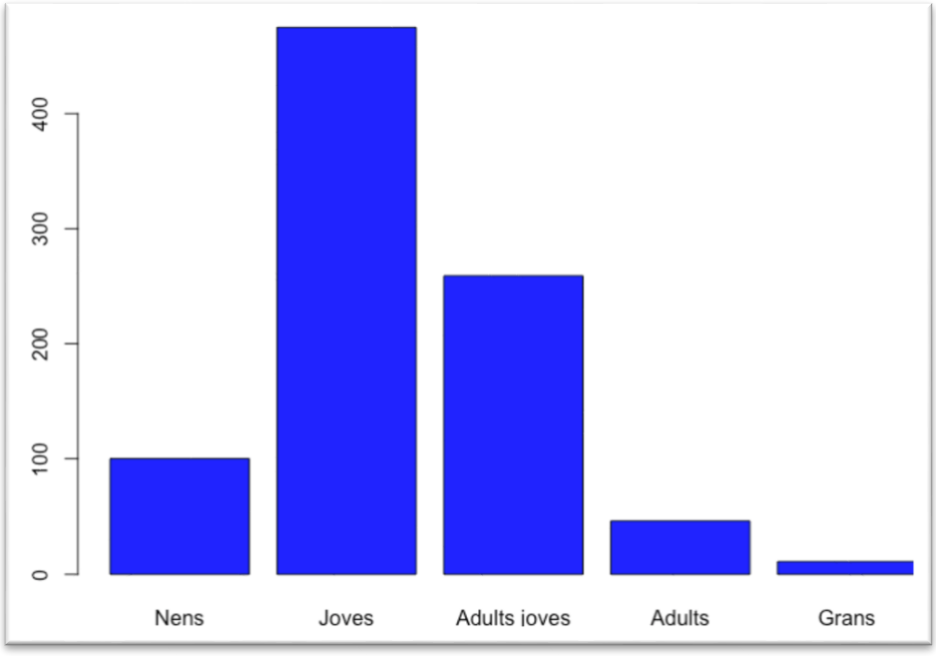
Però en aquest apartat efectuarem proves estadístiques addicionals:

4.3.1 PREPARACIÓ PRÈVIA DE LES VARIABLES

Per a dur a terme els estudis estadístics en molts casos no és imprescindible que l'edat sigui un factor, però donat que tant la variable dependent com les altres dues variables són factors, optem per factoritzar-la per fer un estudi complert amb variables qualitatives.

Factoritzem l'edat creant 5 categories:

- Nens: fins a 16 anys.
- Joves: de 17 a 31 anys
- Adults joves: de 32 a 51 anys
- Adults: de 52 a 64 anys
- Grans: més de 64 anys



4.3.2 RELACIONS DE DEPENDÈNCIA

El més habitual és dur a terme una taula de correlacions creuades i extreure d'aquesta manera la relació de dependència entre variables quantitatives. Però el nostre dataset ja no té cap variable quantitativa, són totes qualitatives, de manera que enlloc de fer un estudi de correlacions, el que fem és un estudi de relacions de dependència, per poder saber si les variables de les que disposem realment tenen afectació (relació de dependència) amb la variable categòrica que representa la supervivència.

L'avaluació s'efectua a través d'un test de Chi-quadrat (Pearson Chi-squared test). Es fan, per tant, tres test, un per cada variable.

En aquests tests la variable nul·la sempre és la mateixa: que són variables independents. D'aquesta manera si el p-valor d'algun dels tests surt per sobre del valor de significació (que establim en l'habitual 0,05) no podrem rebutjar aquesta hipòtesi.

Els resultats dels tests són els següents:

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  table(myData$Sex, myData$Survived)
X-squared = 260.72, df = 1, p-value < 2.2e-16
```

```
        Pearson's Chi-squared test

data:  (table(data$Pclass, data$Survived))
X-squared = 102.89, df = 2, p-value < 2.2e-16
```

```
        Pearson's Chi-squared test with simulated p-value (based on 2000
replicates)

data:  (table(data$Age.factor, data$Survived))
X-squared = 21.827, df = NA, p-value = 0.0004998
```

Dels resultats obtinguts hem de concloure que hem de rebutjar les tres hipòtesis nul·les i entendre que la supervivència no va ocórrer amb independència de cap d'aquestes tres variables.

4.3.3 LA MITJANA D'EDAT DELS SUPERVIVENTS ÉS MENOR DE 30 ANYS?

També hem formulat una hipòtesi sobre l'edat i la supervivència. Realitzem el test corresponent per a veure si la nostra hipòtesi és certa. En concret volem validar la nostra hipòtesi de que la mitjana d'edat dels supervivents és menor de 30 anys.

Per poder aplicar proves per contrast d'hipòtesis paramètriques, com la prova t de Student:

1) Les variables de les dades analitzades han d'estar normalment distribuïdes.

2) Les variàncies d'aquestes variables han de romandre constants al llarg del rang observat d'alguna altra variable.

Com que no es compleix per a la variable Age, fem servir una alternativa no paramètrica, com les proves de Wilcoxon o Mann-Whitney.

La diferència principal entre el test de suma de rangs o test U de Mann-Whitney i el test de rangs i signes de Wilcoxon és que en el primer les mostres són independents i en el segon que les dades es troben emparellades.

En aquest cas, les dades són independents, per tant, escollim el test U de Mann-Whitney (Mann–Whitney–Wilcoxon, Wilcoxon rank-sum test o Wilcoxon–Mann–Whitney).

```
        Wilcoxon rank sum test with continuity correction

data:  data_survived$Age and data_nosurvived$Age
W = 8469.5, p-value < 2.2e-16
alternative hypothesis: true location shift is less than 30
95 percent confidence interval:
         -Inf 3.052763e-05
sample estimates:
difference in location
            -0.9999991
```

Obtenim un p-value < 2.2e-16 i per tant podem assumir que és més probable sobreviure per al menors de 30 anys.

4.3.4 MODELS DE REGRESSIÓ LOGÍSTICA

Farem un model de regressió logística de la variable Survived en funció de tres variables descriptives: Sex, Age i Pclass.

Després de la preparació de dades i de la comprovació de relacions de l'apartat anterior, fem les regressions logarítmiques corresponents.

Per tal d'analitzar l'evolució de l'enriquiment del model, farem primer tres models de forma independent:

- Model de regressió logística de la variable "Survived" usant com a variable explicativa la classe: "Pclass"
- Model de regressió logística de la variable "Survived" usant com a variable explicativa el sexe: "Sex"

- Model de regressió logística de la variable "Survived" usant com a variable explicativa l'edat: "Age"

I posteriorment es fa un model múltiple usant les tres variables explicatives per analitzar-ne la millora.

Comencem, doncs analitzant els models univariables:

Model de regressió logística de la variable "Survived" usant com a variable explicativa la classe: "Pclass"

```
Call:
glm(formula = Survived ~ Pclass, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4094  -0.7450  -0.7450   0.9619   1.6836

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5306     0.1409   3.766 0.000166 ***
Pclass2      -0.6394     0.2041  -3.133 0.001731 **
Pclass3      -1.6704     0.1759  -9.496  < 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: 1186.7  on 890  degrees of freedom
Residual deviance: 1083.1  on 888  degrees of freedom
AIC: 1089.1

Number of Fisher Scoring iterations: 4
```

Llegint la taula observem la significació dels regressors que és:

- ***: significació p-valor <0.001
- **: significació p-valor <0. 01
- *: significació p-valor <0.1

Tots 3 regressors són significatius però la primera i la tercera classe tindrien significació fins i tot en valors extrems de nivell de significació (<0.001), la segona ho seria fins a un nivell inferior de significació (< 0.01, però tots tres compleixen p-valor< 0,05%.

Model de regressió logística de la variable "Survived" usant com a variable explicativa el sexe: "Sex"

```
Call:
glm(formula = Survived ~ Sex, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6462  -0.6471  -0.6471   0.7725   1.8256

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0566     0.1290   8.191 2.58e-16 ***
Sexmale      -2.5137     0.1672 -15.036  < 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: 1186.7  on 890  degrees of freedom
Residual deviance:  917.8  on 889  degrees of freedom
AIC: 921.8

Number of Fisher Scoring iterations: 4
```

En el cas del sexe, els dos regressors tenen influència molt significativa. Veiem que el p-valor és molt petit en els dos casos.

Model de regressió logística de la variable “Survived” usant com a variable explicativa l’edat: “Age”

```
Call:
glm(formula = Survived ~ Age.factor, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2637  -0.8958  -0.8958   1.3087   2.1899

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)              0.2007     0.2010   0.998   0.3181
Age.factorJoves         -0.9065     0.2234  -4.057 4.97e-05 ***
Age.factorAdults joves  -0.5041     0.2371  -2.126   0.0335 *
Age.factorAdults        -0.5521     0.3607  -1.531   0.1258
Age.factorGrans         -2.5033     1.0677  -2.345   0.0190 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.7  on 890  degrees of freedom
Residual deviance: 1162.7  on 886  degrees of freedom
AIC: 1172.7

Number of Fisher Scoring iterations: 4
```

En el cas de la regressió amb predictor edat, tenim dos factors amb p-valor>0,05% i que, per tant, no són significatius, Adults i Nens (que es troba a l’intercepte). Grans i adults joves són poc significatius. El factor més significatiu és joves. Coincideix que també és on hi ha més registres.

I un cop efectuats els tres models univariables, procedim amb el multivariable:

Model de regressió logística de la variable "Survived" usant "Pclass", "Sex" i "Age"

```
Call:
glm(formula = Survived ~ Age.factor + Pclass + Sex, family = binomial,
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6604  -0.6690  -0.4267   0.6471   2.3121

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)              3.5093     0.3700   9.485  < 2e-16 ***
Age.factorJoves         -0.9660     0.2798  -3.453 0.000555 ***
Age.factorAdults joves  -1.2166     0.3207  -3.794 0.000148 ***
Age.factorAdults        -1.7957     0.4782  -3.755 0.000173 ***
Age.factorGrans         -2.7258     1.1143  -2.446 0.014435 *
Pclass2                 -1.0860     0.2613  -4.155 3.25e-05 ***
Pclass3                 -2.3042     0.2553  -9.027  < 2e-16 ***
Sexmale                 -2.5897     0.1871 -13.844  < 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: 1186.66  on 890  degrees of freedom
Residual deviance:  804.42  on 883  degrees of freedom
AIC: 820.42

Number of Fisher Scoring iterations: 5
```

A la regressió múltiple lineal no tenim cap factor que no sigui significatiu. Tots els que es troben a l'intercepte són molt significatius per a la regressió. El que té un nivell p-valor mojor és el factor edat Grans.

Avaluar la bondat dels models

Una de les formes d'avaluar la bondat dels models regressius és comparant el paràmetre AIC (Akaike Information Criterion) que estima l'error de predicció. Per ser exactes, avalua la quantitat d'informació que es perd, per tant, a menor valor, millor qualitat del model.

- Amb el model de la classe tenim un AIC de 1089,1.
- Amb el del sexe, de 921,8.
- Amb el de l'edat de 1172,7.
- El model de regressió logística múltiple té un AIC de 820.42.

Per tant, el millor model és el múltiple.

4.3.5 ODDS RATIOS

Es fa el càlcul de les Odd Ràtio per a saber l'aportació de cadascuna de les variables per als models, amb especial interès per la lectura dels que corresponen al model multivariable.

Model de regressió logística de la variable "Survived" usant com a variable explicativa la classe: "Pclass"

predictor <chr>	oddsratio <dbl>	CI_low (2.5) <dbl>	CI_high (97.5) <dbl>	increment <chr>
Pclass2	0.528	0.353	0.786	Indicator variable
Pclass3	0.188	0.133	0.265	Indicator variable

El que el OR indica és com es modifica la ràtio de supervivència si canvia el predictor.

En aquest cas, tenim com a valor base la classe Pclass1. Per tant quan obtenim que la oddsratio de Pclass2 és 0.528 el que vol dir és que la ràtio de supervivents en la classe 2 en relació amb la classe 1 és de 0.528.

És a dir, la ràtio de supervivents en la classe 2 és la meitat que en la classe 1.

I per a la classe 3, la ràtio és de 0.188.

Més endavant calcularem en quant es quantifica aquesta caiguda de probabilitats.

Model de regressió logística de la variable "Survived" usant com a variable explicativa el sexe: "Sex"

predictor <chr>	oddsratio <dbl>	CI_low (2.5) <dbl>	CI_high (97.5) <dbl>	increment <chr>
Sexmale	0.081	0.058	0.112	Indicator variable

La proporció de supervivents entre els homes és 1:0.081 respecte les dones, és a dir, que la ràtio de supervivents entre les dones multiplica per més de 10 la dels homes.

Model de regressió logística de la variable "Survived" usant com a variable explicativa l'edat: "Age"

predictor <chr>	oddsratio <dbl>	CI_low (2.5) <dbl>	CI_high (97.5) <dbl>	increment <chr>
Age.factorJoves	0.404	0.260	0.625	Indicator variable
Age.factorAdults joves	0.604	0.378	0.960	Indicator variable
Age.factorAdults	0.576	0.281	1.162	Indicator variable
Age.factorGrans	0.082	0.004	0.451	Indicator variable

També en aquest cas, el nivell base (Nens) és el més favorable, ja que veiem que tots els oddsratios són menors que 1. Veiem a més que el pitjor ràtio es dona amb els més grans, seguit del joves, el adults i els

adults joves. És a dir, després dels nens, el rang d'edat que té un millor ràtio de supervivència són els adults joves. I el pitjor de tots és el de la gent més gran.

Model de regressió logística de la variable "Survived" usant "Pclass", "Sex" i "Age"

predictor <chr>	oddsratio <dbl>	CI_low (2.5) <dbl>	CI_high (97.5) <dbl>	increment <chr>
Age.factorJoves	0.381	0.219	0.657	Indicator variable
Age.factorAdults joves	0.296	0.157	0.553	Indicator variable
Age.factorAdults	0.166	0.064	0.418	Indicator variable
Age.factorGrans	0.065	0.003	0.411	Indicator variable
Pclass2	0.338	0.201	0.561	Indicator variable
Pclass3	0.100	0.060	0.163	Indicator variable
Sexmale	0.075	0.052	0.108	Indicator variable

Veient aquests oddsratios, es pot concloure que si l'intercepte era la millor combinació de paràmetres per sobreviure, la pitjor seria: home, edat Grans i Classe 3.

Si ho analitzem en percentatges, en quin percentatge es veu incrementada la probabilitat de sobreviure segons la classe?

La variació en la probabilitat de sobreviure es pot representar en una regressió logística com la que tenim com a 1-odds %, en aquest cas, on l'intercept recull els factors més favorables.

D'aquesta manera, el odds, en % representa el decrement de probabilitat de sobreviure i s'han calculat i representat en les taules següents:

Model de regressió logística de la variable "Survived" usant com a variable explicativa la classe: "Pclass"

predictor <chr>	oddsratio <dbl>	% <dbl>
Pclass2	0.528	47.2
Pclass3	0.188	81.2

El que s'observa aquí és que la probabilitat de sobreviure decau un 47.2% en la classe 2 i un 81,2% en la classe 3, respecte la probabilitat de supervivència de la classe 1.

Model de regressió logística de la variable "Survived" usant com a variable explicativa el sexe: "Sex"

predictor <chr>	oddsratio <dbl>	% <dbl>
Sexmale	0.081	91.9

Ser home, redueix la probabilitat de sobreviure en un 91,% respecte ser dona.

Model de regressió logística de la variable "Survived" usant com a variable explicativa l'edat: "Age"

predictor <chr>	oddsratio <dbl>	% <dbl>
Age.factorJoves	0.404	59.6
Age.factorAdults joves	0.604	39.6
Age.factorAdults	0.576	42.4
Age.factorGrans	0.082	91.8

Respecte l'edat infantil, els percentatges que hi ha a la taula representen la reducció de probabilitat de supervivència segons el rang d'edat.

Així, respecte a ser nen, ser:

- Jove: reducció del 59.6% de probabilitat de sobreviure
- AdultJove: reducció del 39.6% de probabilitat de sobreviure
- Adult: reducció del 42.4% de probabilitat de sobreviure
- Gran: reducció del 91,8% de probabilitat de sobreviure

Model de regressió logística de la variable "Survived" usant "Pclass", "Sex" i "Age"

predictor <chr>	oddsratio <dbl>	% <dbl>
Age.factorJoves	0.381	61.9
Age.factorAdults joves	0.296	70.4
Age.factorAdults	0.166	83.4
Age.factorGrans	0.065	93.5
Pclass2	0.338	66.2
Pclass3	0.100	90.0
Sexmale	0.075	92.5

Els predictors Nens, Classe 1 i sexe femení es troben a l'intercepte i són els casos més favorables per la supervivència (tots els odds, són menors que 1)

Per tant, els predictors empitjoren tots la possibilitat de supervivència en els percentatges següents:

- Joves: 61,9%
- Adults joves: 70,4%
- Adults: 83,4%
- Grans: 93,5%
- 2a Classe: 66,2%
- 3a Classe: 90%
- Sexe masculí: 92,5%

4.3.6 PREDICCIONS

Finalment, amb el millor model obtingut, que és el model de regressió logística multivariable fem la predicció d'un passatger del conjunt de test.

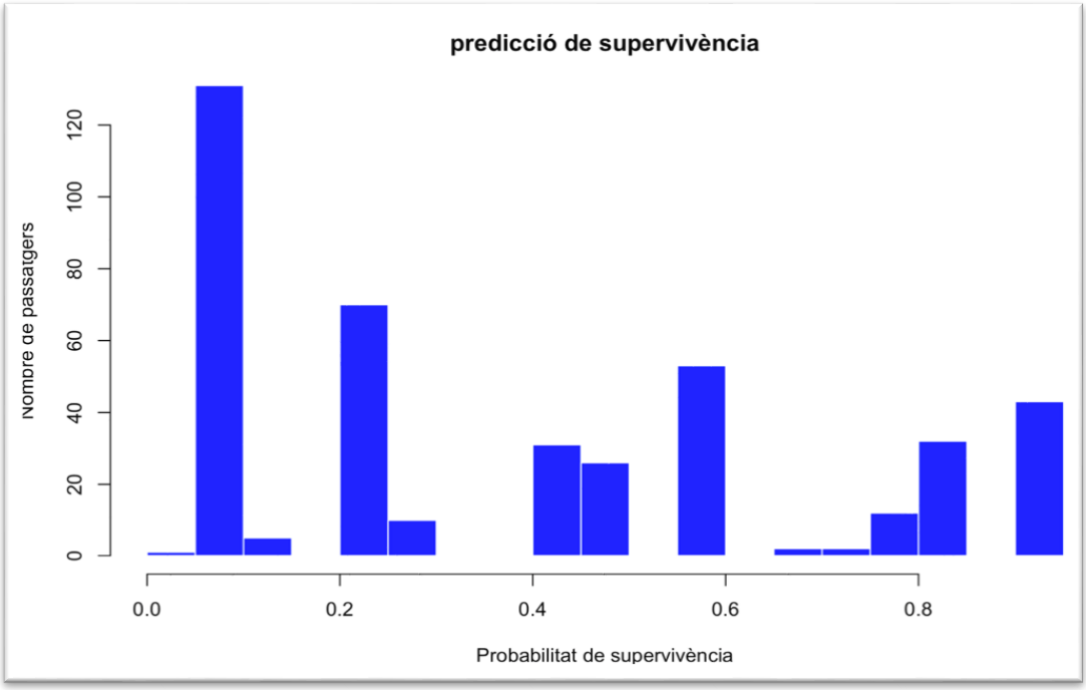
Com que hem fet també la predicció de tots els passatgers del dataset de test, no inclourem aquí detalls de la predicció feta per a un sol passatger.

Per a fer la predicció del conjunt de test, hem aplicat a les variables les mateixes operacions que hem realitzat anteriorment al conjunt d'entrenament. En concret, es fa la mateixa reducció, es factoritza i es tracten els valors nuls de la variable Age.

Un cop tenim un set de test amb les mateixes variables que el fitxer traint que hem fet servir per a generar el model, ja podem fer la predicció.

El resultat de la predicció, en ser una categoria binària, i usant el type = response, no ens dóna com a resultat una etiqueta o categoria sino un valor entre 0 i 1 que ens serveix a mode de percentatge de probabilitat de ser 1.

Si analitzem l'histograma de les prediccions obtingudes, tenim:



És a dir, del total de passatgers que s'han predit, el nombre més gran té entre 5% i 10% de probabilitats de ser un 0, és a dir de no sobreviure.

La dificultat, per tant, radica en decidir ara, amb aquests resultats quins passatgers considerem que hem predit que sobreviuen i quins no. Sense més elements de valoració, establirem el llindar en el 50%. És a dir considerarem que sobreviuen els que tinguin més de 0.5 en la predicció.

D'aquesta manera tenim la següent taula de prediccions (0 = no sobreviure; 1= sobreviure):

```
  0   1
274 144
```

Finalment, assignem a la columna 'Survived' del dataset de test, la predicció.

5 REPRESENTACIÓ DELS RESULTATS (2 PUNTS)

Representació dels resultats a partir de taules i gràfiques.

A mida que s'ha anat desenvolupant el codi s'han anat efectuant visualitzacions de les dades disponibles. Per tant, en el codi es poden consultar totes les gràfiques que s'han dut a terme al llarg de l'anàlisi.

A continuació es recullen les més significatives sobretot en relació amb les conclusions que es presentaran en el següent apartat.

També incloíem les gràfiques que formen part d'una anàlisi descriptiva del data set d'entrenament.

5.1 PERCENTATGES DE SUPERVIVENTS I NO SUPERVIVENTS

Són importants aquestes taules perquè, al cap i a la fi, són el principal element d'anàlisi, la supervivència.

5.1.1 CONJUNT D'ENTRENAMENT

Survived <fctr>	Percentatge <dbl>
No	61.61616
Sí	38.38384

S'observa que, de forma genèrica sobreviu poc menys del 40% del passatge.

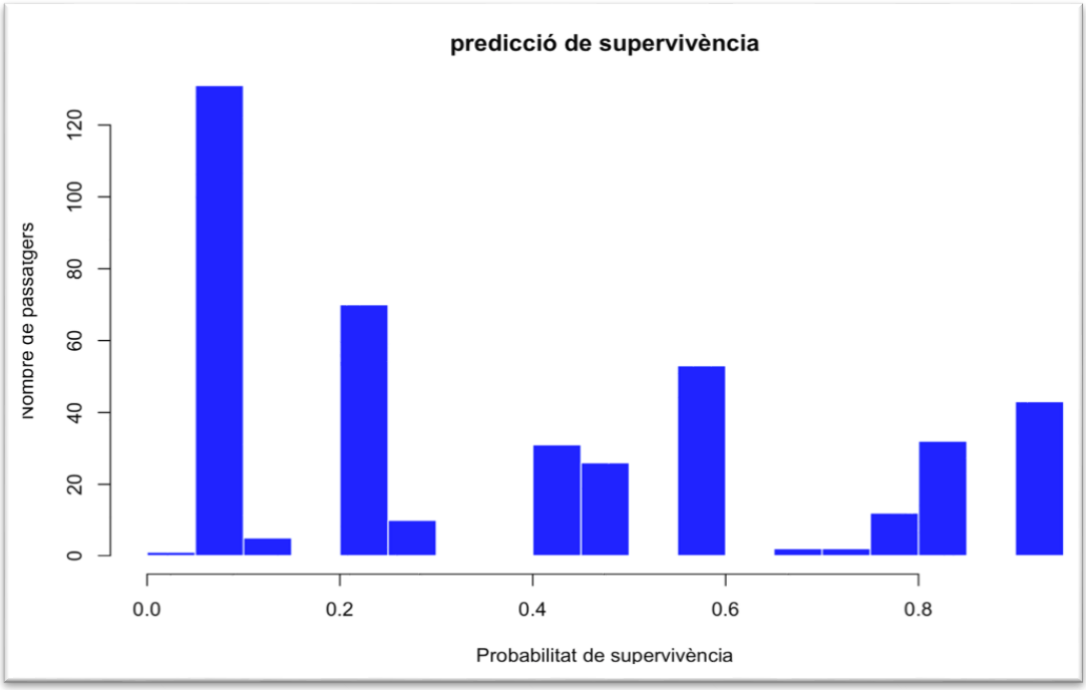
5.1.2 CONJUNT DE TEST O PREDICCIÓ

Survived <fctr>	Percentatge <dbl>
No	65.55024
Sí	34.44976

Un cop aplicat el model que hem desenvolupat i aplicant la predicció a les dades de test, ens dóna com a resultat que sobreviu poc menys del 35% del passatge.

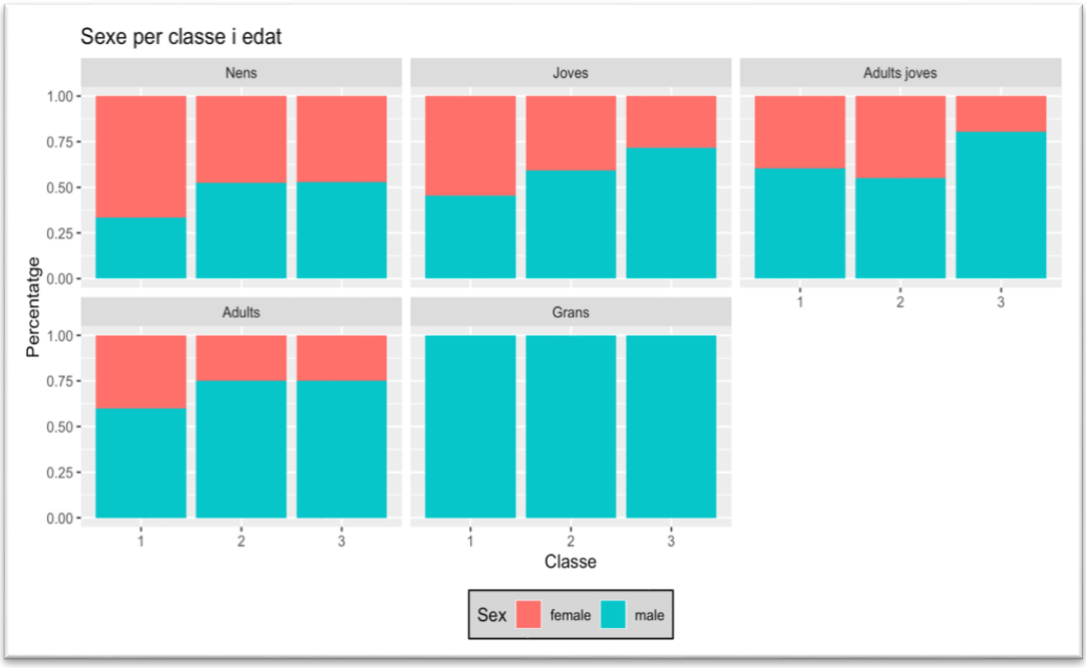
Tal com s'ha explicat al llarg del document, aquesta predicció no només ve condicionada pel model, sinó que un cop obtinguda la predicció en forma de valor de 0 a 1 que es podia aproximar al % de probabilitat de supervivència, hem decidit establir el punt de tall en el 0,5. Haver escollit un altre punt de tall hauria modificat els resultats.

Reproduïm aquí el resultat concret de la predicció, per la seva trascendència



5.2 GRÀFICS DEL CONJUNT D'ENTRENAMENT

Hem elaborat diferents visualitzacions per poder analitzar les dades d'entrenament.

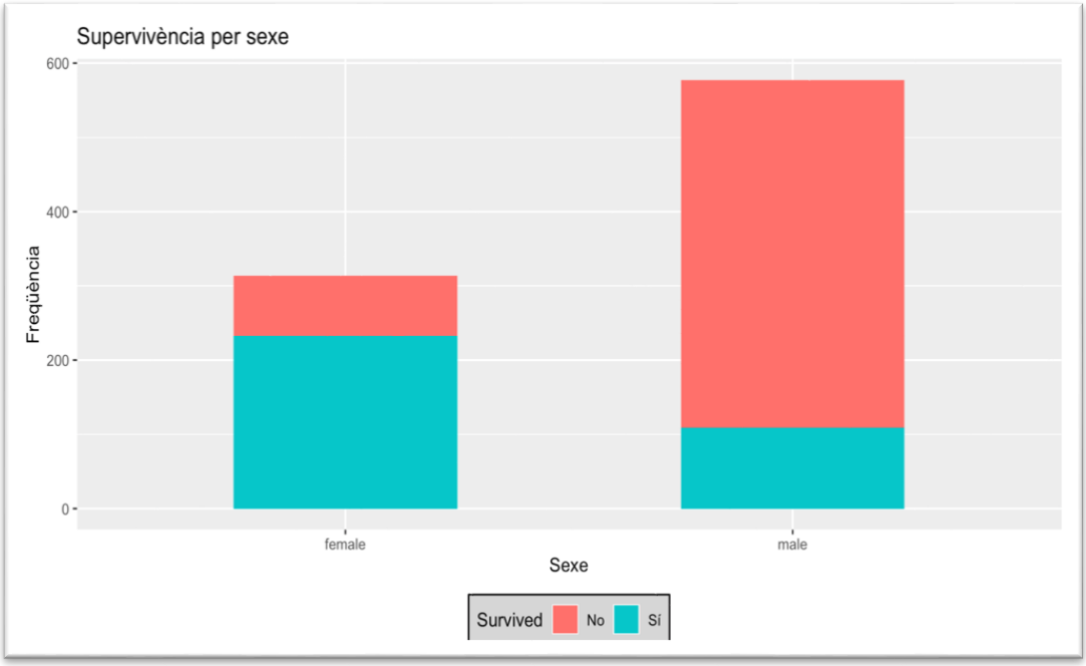


Observem que a mida que incrementem l'edat i disminuïm el nivell social hi ha una menor presència de dones.

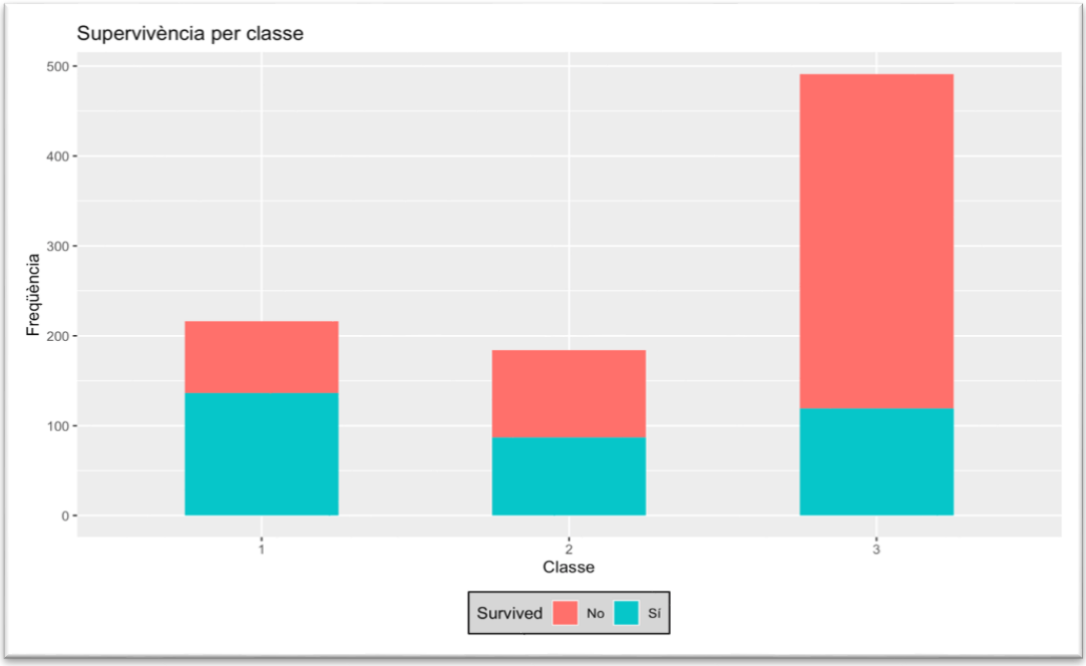
Destaquem que no hi ha dones en la franja major d'edat.

En canvi superen el 50% les dones en la classe alta i en en rang d'edat per sota del 32 anys.

En el rang d'edat dels nens, hi ha gairebé un equilibri de sexes.

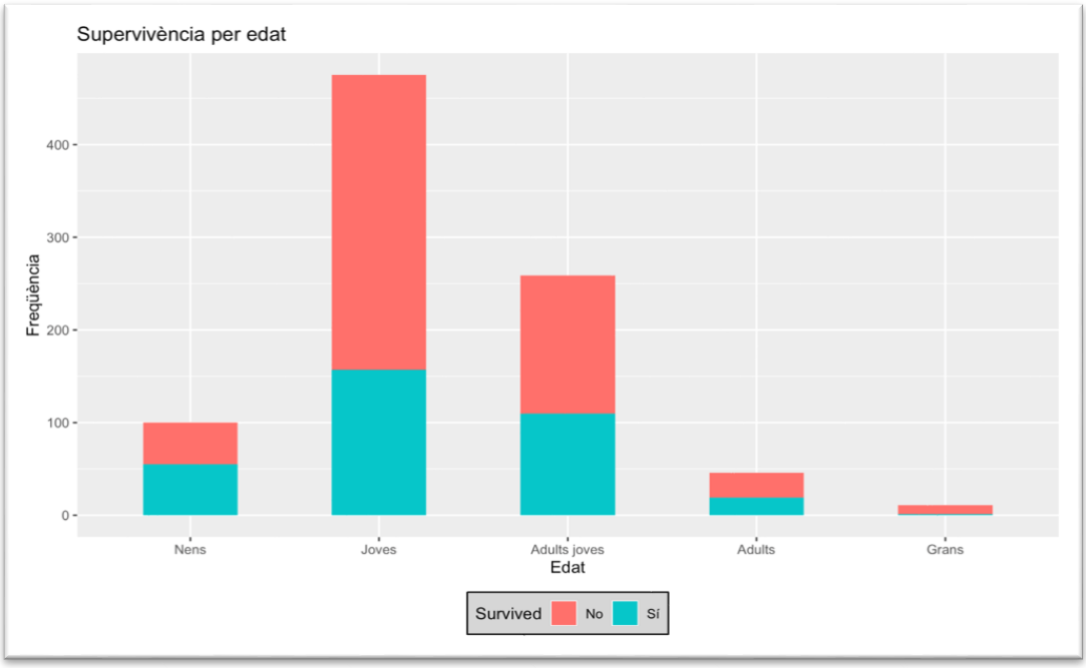


Aquesta gràfica ens permet observar que la supervivència entre les dones va ser molt més gran que entre els homes. La majoria de dones va sobreviure, la majoria d'homes hi ha perdre la vida.

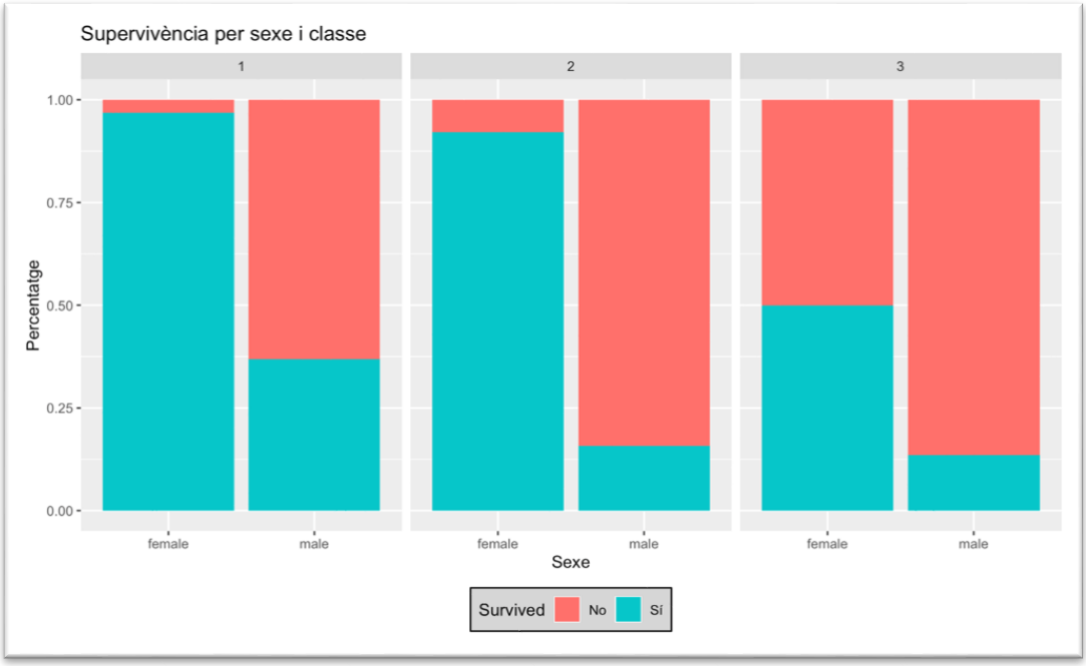


També és de destacar la distribució de la supervivència per classes. Mentre que en la classe 2 hi ha un equilibri de supervivent / no supervivents, en la classe alta la majoria sobreviuen i en la classe baixa la majoria no sobreviuen.

Té certa relació amb el gràfic anterior, ja que hem de recordar que en la classe alta és on hi ha més dones, per tant, d'alguna manera hi ha certa relació.



Tot i que numèricament els supervivents més nombrosos estan en el rang d'edat jove (17 a 31 anys), a nivell percentual la supervivència més alta es dona entre els nens.



Si analitzem els dos paràmetres en conjunt es pot veure que les dones sobreviuen majoritàriament, de forma gairebé total en la classe 1 i 2 i just en el 50% en la classe més baixa. En canvi els homes no superen en cap cas el 35% de supervivència, que es dóna en la classe alta, però decau a mida que decau la classe social.

En aquest sentit podem dir que ser dona i de classe alta, garantia altes probabilitats de supervivència. Ser home i de classe mitjana o baixa no permetia més del 15% d'esperança de supervivència.

Si enlloc del sexe, analitzem la classe en relació amb l'edat per observar la supervivència, veiem que es compleix sempre, per tots els rangs d'edat que a menor classe social, menor nivell de supervivència.

Havíem vist que els nens teníem alt nivell de supervivència i ara s'observa que aquesta supervivència està realment per sobre del 80% en les classes 1 i 2, mentre que decau dramàticament per a la classe 3.

En general també s'observa que a major edat, menor supervivència, excepte en el cas dels joves respecte els adults joves. Els adults joves sobreviuen més que els joves, en la classe 2.

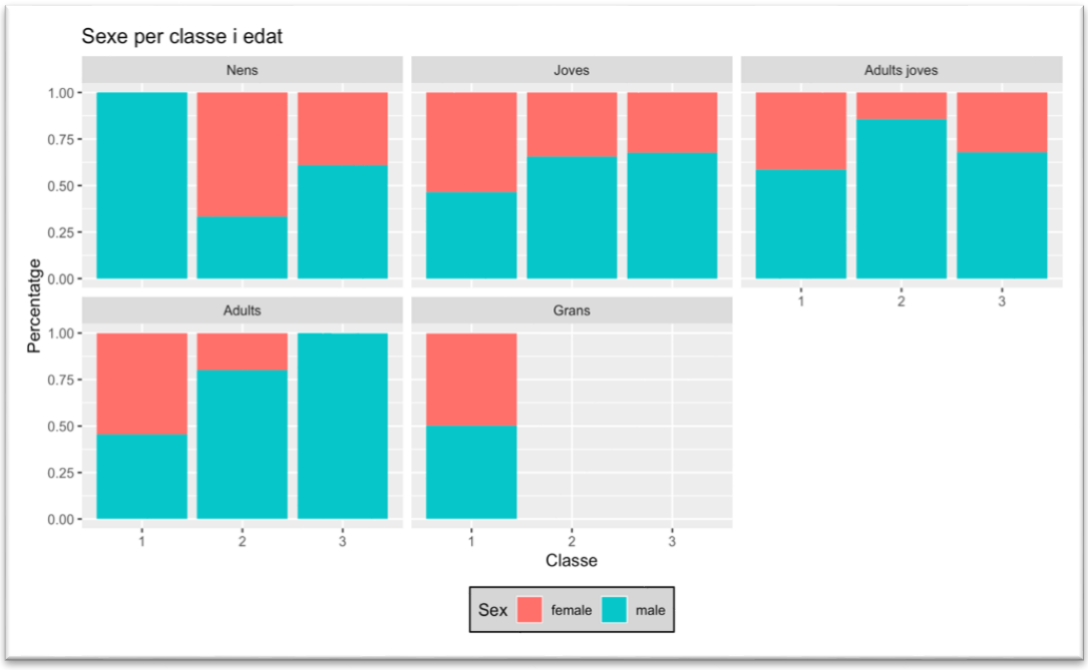
En aquest gràfic s'observa que a mida que creixem en edat, les dones sobreviuen més (no tenim dades de dones grans), en canvi amb els homes ocorre al revés. A mida que creix l'edat, decau el percentatge de supervivents. Excepte pels homes joves que sobreviuen menys que els adults joves.

5.3 GRÀFICS DEL CONJUNT DE TEST O PREDICCIÓ

Els gràfics a continuació es corresponen al grup de test. Menys observacions per tant.

Quan les gràfiques incorporen la variable supervivència, hem de tenir en compte que es refereix a la predicció efectuada.

Per tant, les dades d'observació de la relació entre sexe, edat i classe són reals, tot i que menys representatives del conjunt de passatgers del titànic, ja que són menys mostres, però les dades que inclouen supervivència, tenen cert marge d'error.

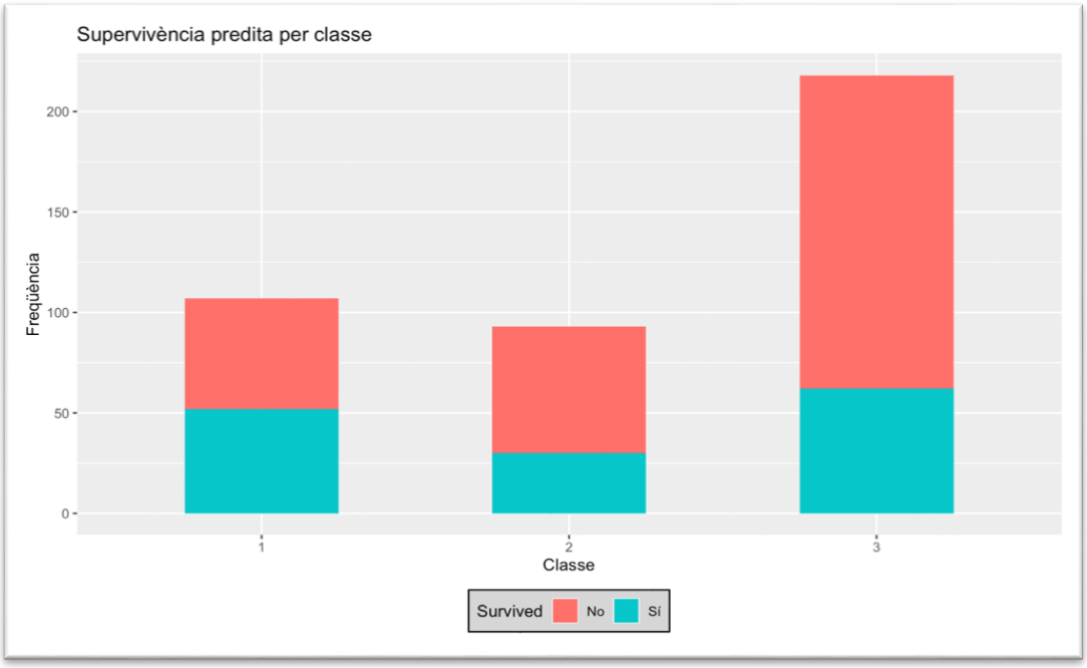


En el conjunt de test no hi ha passatgers grans de classe 2 i 3. I hi ha dones d'edat gran, que en el grup d'entrenament no n'hi havia. Per contra, en el grup de test no hi ha dones de classe 3 i edat adulta ni nenes de classe 1.

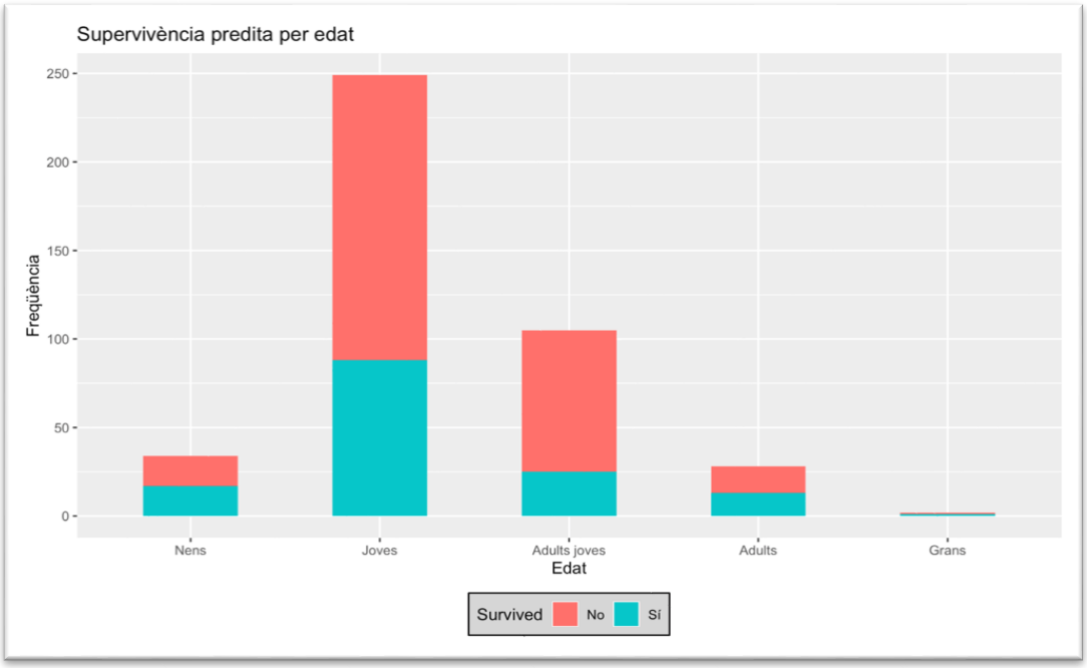
Les diferències existents entre els dos grups de dades, fan que les dades de supervivència per sexe sigui diferents en el conjunt de test que en el conjunt de dades.

En el conjunt de test, la supervivència en percentatge de les dones és encara major i en canvi la

supervivència dels homes és encara menor.



A diferència del que passa amb el conjunt d'entrenament, en el conjunt de test l'equilibri de supervivents/no supervivents es dóna en la classe 1. Però es manté la premissa de que a major classe social, millor percentatge de supervivència.



La supervivència per edat no ha patit moltes modificacions, amb les diferències normals degudes a la diferència de composició de la mostra de dades. Potser destacaria una major ràtio de supervivència entre

els adults joves.

De nou en la gràfica que relaciona la supervivència en funció del sexe i de la classe hi veiem l'efecte de la reducció de la mostra. Obtenim valors extrems en les dones de classe 1 i 2 i en els homes de classe 2 i 3, en part a causa de la predicció i del llindar del 0.5 aplicat en la definició final de la predicció.

De nou en la predicció de supervivència per sexe i edat ens trobem amb valors extrems, però que no es contradiuen amb les generalitzacions observades en el conjunt d'entrenament. Simplement s'han

extremat les distribucions.

6 RESOLUCIÓ DEL PROBLEMA (0,5 PUNTS)

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

Una explicació complerta dels resultat obtinguts s'ha anat efectuant a mida que s'han anat obtenint els resultats. Es detallen aquí, a mode de resum la recopilació de la informació que hem anat obtenint i les conclusions a les que ens fan arribar en conjunt.

- Cal revisar i netejar el conjunt de dades:

Encara que en el conjunt de dades que hem usat no hi havia dades extremes, cal sempre tenir en compte la seva presència i eliminar el seu efecte per poder obtenir conclusions fonamentades.
En el nostre cas, sí que teníem valors perduts que hem pout resoldre en aquest cas amb una substitució per un valor de mediana. Per afinar els resultats al màxim, hem usat valors mediana diferents segons les dades de cada observació que sí que teníem.
En aquest sentit, hem après que no hi ha una forma única de resoldre la neteja de dades.

- Cal treballar amb els registres necessaris: reducció de la dimensionalitat:

Un ús major de dades no implica uns millors resultats. En el nostre cas, hem reduït la dimensionalitat, eliminant els atributs que no ens interessaven per a l'anàlisi que volíem fer.
Donat que teníem un nombre reduït d'observacions, no ens ha calgut aplicar tècniques de mostreig.
El que hem fet, però, és factoritzar en categories la variable edat, per poder extreure conclusions raonades.
Hem après que cal tenir en compte quin tipus d'anàlisi es vol fer quan es seleccionen les dades.

- La supervivència no és independent de l'edat, el sexe o la classe social:

Després de fer els tres tests Chi-quadrat hem observat que en les tres variables hem de descartar la hipòtesi nul·la i descartar la independència de relació entre supervivència i les 3 variables d'estudi.
Hem après que aquest estudi ens permet arribar a fortes conclusions que es confirmaran gràficament

més endavant i que no es pot analitzar igual la dependència en variables qualitatives i quantitatives.

- Els supervivents són menors de 30 anys d'edat mitjana:

Hem pogut fer un contrast d'hipòtesis, en aquest cas per determinar si la mitjana d'edat dels supervivents era menor de 30 anys.

Hem arribat a la conclusió que és cert, usant un test de Mann-Whitney, que és una alternativa no paramètrica al no tenir totes les variables normalment distribuïdes ni de variància constant.

Hem après que l'estudi de la normalitat i de la variància són importants per saber quines eines són les més adients en els contrasts d'hipòtesis.

Hem aplicat el contrast a una pregunta concreta, però podríem haver aplicat contrasts d'hipòtesis per analitzar qualsevol aspecte que ens interessés.

- Models de regressió per modelar el problema

Hem creat models de regressió per intentar usar les variables (que sabem que tenen efecte la supervivència) i crear un model per explicar el problema i efectuar prediccions.

Hem creat 4 models, tots ells logístics perquè la variable d'estudi (supervivència) no és quantitativa sinó qualitativa.

Hem creat un model amb cada una de les variables (edat, sexe i classe) i després hem creat un model combinant les tres variables.

Hem pogut comprovar que totes 3 poden explicar bé el problema, però que és el model multifactorial el que millor s'hi aproxima. Per tant, es conclou que és amb el concurs de les tres variables que s'explica la supervivència de cada passatger.

Hem après a crear models de regressió i a comparar els resultats per escollir el millor model.

- La combinació perfecta per sobreviure és ser d'edat infantil, dona i de classe 1

Hem pres com a valor base ser nen, dona i de classe 1 i hem vist que aquesta és la combinació que té una major ràtio de supevivents.

Per a poder analitzar aquestes ràtios hem usat l'anàlisi de les OR (Odds Ratio) que mostra com varia el resultat quan canvia un predictor.

D'aquesta manera hem pogut veure com afecta cada variable, tant amb els models de regressió unifactorials, com amb el multifactorial.

En el cas del model multifactorial, hem pogut veure que, respecte el cas base (nen, sexe femení i classe 1), les probabilitats de sobreviure sempre empitjorem, sigui quina sigue la variable que es modifiqui. Les probabilitats de sobreviure decauen en els percentatges següents:

- Joves: 61,9%
- Adults joves: 70,4%
- Adults: 83,4%
- Grans: 93,5%
- 2a Classe: 66,2%
- 3a Classe: 90%
- Sexe masculí: 92,5%

Es conclou, també, per tant, que el pitjor cas és ser home, d'edat Gran i de classe 3.

Hem après que analitzant amb deteniment els paràmetres d'un model de regressió es pot entendre molt bé com afecta cada variable i cada valor de cada variable en el problema.

- Hem pogut preveure la supervivència o no d'un conjunt de passatgers

Si hem aconseguit conèixer bé el problema, el més normal era aplicar el coneixement per predir altres situacions. Gràcies a disposar d'un conjunt de test per fer prediccions hem pogut posar-ho en pràctica.

Primer ho hem aplicat a un sol subjecte, que ha resultat anecdòtic si tenim en compte que després hem predit tot el conjunt.

No tenim els resultats reals del conjunt de test, per tant, no podem valorar l'exactitud del nostre model. Però els percentatges de supervivents són molt similars als del conjunt d'entrenament.

Hem après a aplicar els models sobre un conjunt de noves dades i a valorar-ne els resultats.

Com a conclusió final del problema, hem generat un fitxer de sortida: titanic_final.csv on hi trobem les dades tant de test com d'entrenament tal com han quedat després de ser analitzades.

Conté 5 columnes:

- Survived: que conté un Sí o un No en funció si el passatger sobreviu o no
- Pclass: indica la classe (de 1 al 3, de millor a pitjor classe)
- Sex: male o female segons si és home o dona
- Age: edat en número
- Age.factor: rang d'edat en el que els hem classificat. Per a simplificar l'ús hi hem posat el rang d'edat (0-16 / 17-31/ 32-51 / 52-64 / 65+) enlloc dels noms (Nens, joves, adults,...) que seria indeterminat.

7 CODI (2 PUNTS)

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.

En la carpeta 'codi' hi ha disponible el codi en .rmd i en .pdf.

També incoem els fitxers d'origen de dades i el fitxer final generat.

8 CONTRIBUCIONS

CONTRIBUCIONS	INTEGRANT	SIGNATURA
Recerca prèvia	Meritxell Bosch	MB
	Marta Martínez	MM
Redacció de les respostes	Meritxell Bosch	MB
	Marta Martínez	MM
Desenvolupament codi	Meritxell Bosch	MB
	Marta Martínez	MM

9 FONTS CONSULTADES

- Chi-squared Test of Independence.

  Recurs en línea: http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-independence

- How do I interpret the AIC. 2018. [WEB] R-blogger.

  Recurs en línea: https://www.r-bloggers.com/how-do-i-interpret-the-aic/

- López Cano, Emilio. Ejemplo de Regresión Logística (modelo logit) con R.2017. [WEB] RPubs.

  Recurs en línea: https://rpubs.com/emilopezcano/logit

- Schratz, Patrick. Calculate Odds Ratios Of Generalized Linear (Mixed) Models. [Web] rdocumentation.

  Recurs en línea:
  https://www.rdocumentation.org/packages/oddsratio/versions/2.0.0/topics/or_glm