M2.951 Tipologia i cicle de vida de les dades
PRAC 2: Neteja i validació de les dades

Estanislau Trepat

11 de juny, 2018

Índex

---

1 Introducció

L'enfonsament del vaixell *RMS Titanic* és un dels naufragis més infames de la història. El 15 d'Abril de 1912, durant el seu viatge inaugural, el *Titanic* es va enfonsar després de col · lisionar amb un iceberg, resultant en la mort de 1502 persones, entre passatge i tripulació, d'un total de 2224. Aquesta enorme tragèdia conmocionà la comunitat internacional i va conduïr a una millora en la regulació de la seguretat marítima i dels vaixells de passatgers.

Una de les raons per les que aquest naufragi va provocar tal pèrdua de vides fou que no hi havía suficients bots salvavides per als passatgers i la tripulació. Tot i que hi va haver algún element de sort involucrat en la supervivència a l'enfosament, alguns grups de persones tenien més probabilitat de sobreviure que d'altres, com les dones, els nens, i la classe alta.

Aquest document intenta respondre la pregunta "quin tipus de persones és més probable que sobrevisquin", proporcionant un anàlisi dels principals factors que podrien portar a la supervivència d'una persona segons les seves característiques. També s'utilitzaran tècniques de mineria de dades per intentar predir de forma automàtica si una persona hauria sobreviscut a la tragèdia o no.

Es pot consultar més informació sobre la tragèdia de l'enfonsament del *RMS Titanic* en la pàgina de la Viquipèdia o en la pàgina d'on s'ha extret el conjunt de dades que s'ha utilitzat a Kaggle: Titanic: Machine Learning from Disaster.

2 Descripció del conjunt de dades.

Tal com s'ha esmentat anteriorment, el conjunt de dades utilitzat en aquesta PRAC és l'utilitzat en la competició de Kaggle, Titanic: Machine Learning from Disaster i s'ha obtingut directament d'aquesta plataforma.

Aquest conjunt de dades conté diversa informació sobre els passatgers del vaixell *RMS Titanic*. Entre la informació emmagatzemada podem trobar la edat, el gènere, algunes característiques socio-econòmiques dels passatgers com la classe i/o el preu del bitllet i si el passatger va sobreviure o no al naufragi.

El conjunt de dades s'ens presenta separat en dos subconjunts: un d'entrenament i un de test. Això és així perque el primer és el que *Kaggle* ens proporciona per a realitzar l'anàlisi i entrenament del model de mineria de dades amb el que es realitzaran les prediccions sobre el segon subconjunt i així obtindre una mesura de la qualitat del model obtingut. Aquests dos subconjunts es troben en dos arxius en format *CSV*: `train.csv` i `test.csv` que contenen 891 observacions i 12 variables el primer, corresponent al conjunt d'entrenament, i 418 observacions i 11 atributs el segon, corresponent al conjunt de test. El subconjunt de test conté un atribut menys perque no disposa de la variable dependent, o classe, que és si el passatger va sobreviure o no.

En els primers processos de neteja i anàlisi exploratori de les dades, ajuntarem aquests subconjunts de dades i, per tant, considerarem que el nostre *dataset* disposa de 1309 observacions i 12 variables.

Els atributs disponibles en el conjunt de dades, i que proporcionen informació sobre les característiques socio-econòmiques de cada passatger, són els següents:

Atribut	Descripció	Valors
`PassengerId`	Identificador únic del passatger.	
`Survived`	Si el passatger ha sobreviscut al naufragi o no (atribut de predicció).	0 - No, 1 - Si
`Pclass`	Classe a la que pertany el passatger.	1 - 1a, 2 - 2a, 3 - 3a
`Name`	Nom del passatger.	
`Sex`	Sexe del passatger.	*male* o *female*
`Age`	Edat del passatger.	
`SibSp`	Nombre de germans/nes o marit/muller del passatger en el vaixell.	
`Parch`	Nombre de pares i/o fills del passatger en el vaixell.	
`Ticket`	Identificador del bitllet.	
`Fare`	Preu o tarifa del bitllet.	
`Cabin`	Identificador del camarot o cabina assignada al passatger.	
`Embarked`	Port d'embarcament.	C - Cherbourg, Q - Queenstown, S - Southampton

Tal com s'ha comentat, a partir de les dades disponibles en aquest conjunt de dades, l'objectiu principal de l'anàlisi que es preten realitzar es respondre a la pregunta "quin tipus de persones és més probable que sobrevisquin", a partir d'un anàlisi dels principals factors i caràcteristiques que es troben en les dades i que podrien afectar la supervivència d'una persona.

En l'anàlisi s'utilitzarà un model de mineria de dades per a realitzar la predicció automàtica de si una persona hauria sobreviscut a la tragèdia o no.

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

En primer lloc, realitzarem la lectura de les dades originals proporcionades per Kaggle. Com hem comentat en l'apartat anterior, aquestes es troben dividides en dos conjunts: un d'entrenament i un de test.

Per a facilitar l'anàlisi exploratori de les dades, així com el seu preprocess en ambdós conjunts, els ajuntarem en un de sol. Inicialment, tindrem la precaució de *"marcar"* les observacions que corresponen a cada fitxer i d'afegir la columna amb la variable dependent, o de predicció, `Survived` amb valors buits en el conjunt de test ja que aquesta no hi és present.

```
# Llegim les dades dels fitxers originals
train <- read.csv("../data/train.csv", sep=",", header=TRUE,
                  fill=TRUE, na.strings=c("", "NA"))

test <- read.csv("../data/test.csv", sep=",", header=TRUE,
                 fill=TRUE, na.strings=c("", "NA"))

# En l'atribut/columna `ds` ens referim a quin fitxer/conjunt correspon
train$ds <- "train"
test$ds <- "test"
# Omplim de valors buits l'atribut a predir `Survived` en
# el conjunt de test
test$Survived <- NA

# Ajuntem
titanic <- rbind(train, test)
titanic$ds <- as.factor(titanic$ds)
```

Encara que ajuntem els conjunts de dades, això només ho fem per la practicitat de no tenir que realitzar el preprocés de les dades en els dos conjunts de forma separada. Quan analitzem les variables o atributs presents i els relacionem amb si el passatger ha sobreviscut o no (per exemple: supervivència per classe o sexe), només utilitzarem les dades en el conjunt d'entrenament ja que, com hem comentat abans, la informació de la variable dependent de si un passatger ha sobreviscut o no, només es troba present en aquest conjunt i no en el de test.

Comencem, a continuació, una petita exploració prèvia de les dades que acabem de carregar per a fer-nos una idea de que tenim a les mans.

En primer lloc, mostrem un petit extracte amb les primeres observacions del conjunt de dades.

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	NA	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	NA	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	NA	S
6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583	NA	Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.0750	NA	S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	NA	S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	NA	C

Observem ara els atributs disponibles, així com els seus tipus de dades.

```
'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 : Factor w/ 1307 levels "Abbing, Mr. Anthony",..: 109 191 358 277
   16 559 520 629 417 581 ...
```

```
$ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
$ 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 : Factor w/ 929 levels "110152","110413",..: 524 597 670 50 473
   276 86 396 345 133 ...
$ Fare : num 7.25 71.28 7.92 53.1 8.05 ...
$ Cabin : Factor w/ 186 levels "A10","A14","A16",..: NA 82 NA 56 NA NA 130
   NA NA NA ...
$ Embarked : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
$ ds : Factor w/ 2 levels "test","train": 2 2 2 2 2 2 2 2 2 2 ...
```

A primera vista veiem els diferents atributs o variables que havíem descrit anteriorment. Ens adonem, però, que hi ha algunes variables que no tenen el tipus de dades correcte, com per exemple l'atribut `Pclass` que hauria de ser un factor de 3 nivells corresponents a la classe del passatger i, en canvi, ens ha estat reconegut com un valor enter.

El mateix ens passa amb l'atribut de predicció `Survived` que procedirem a convertir a factor.

```
titanic$Survived <- as.factor(titanic$Survived)
```

Entrarem en més detall en l'assignació correcta dels tipus de dades per als atributs en els pròxims apartats d'aquesta PRAC.

Inicialment ja podem pensar que, per exemple, l'atribut `PassengerId` no ens serà de gaira utilitat en l'anàlisi ja que simplement es tracta d'un identificador de passatger. La intuició ens diu que atributs com el sexe (`Sex`), l'edat (`Age`) i la classe (`Pclass`) poden ésser d'interés per a inferir les probabilitats de supervivència d'un passatger. Examinant els valors de l'atribut `Ticket` s'ens acut que el podem utilitzar per representar el nombre d'acompanyants d'un passatger. Sembla que pugui ser possible que aquelles persones que viatgessin acompanyades tinguessin més probabilitats de sobreviure.

Vegem, a continuació, un petit resum de la distribució de cada atribut on podem veure una aproximació preliminar de la distribució dels valors dels atributs continus i dels valors d'aquelles identificades com a factors. Recordem, però, que com que encara no hem realitzat cap tipus de procés de neteja i/o ajust en les dades, algunes es presenten incorrectament o amb un format de dades que no correspon. Si més no, ens serveix per obtenir una idea inicial sobre les dades que disposem.

```
summary(titanic)
```

```
  PassengerId   Survived      Pclass
 Min.   :   1   0   :549   Min.   :1.000
 1st Qu.: 328   1   :342   1st Qu.:2.000
 Median : 655   NA's:418   Median :3.000
 Mean   : 655              Mean   :2.295
 3rd Qu.: 982              3rd Qu.:3.000
 Max.   :1309              Max.   :3.000

                                Name          Sex           Age
 Connolly, Miss. Kate             :   2   female:466   Min.   : 0.17
 Kelly, Mr. James                 :   2   male  :843   1st Qu.:21.00
 Abbing, Mr. Anthony              :   1                Median :28.00
 Abbott, Mr. Rossmore Edward      :   1                Mean   :29.88
 Abbott, Mrs. Stanton (Rosa Hunt):   1                3rd Qu.:39.00
 Abelson, Mr. Samuel              :   1                Max.   :80.00
 (Other)                          :1301                NA's   :263
     SibSp            Parch            Ticket          Fare
 Min.   :0.0000   Min.   :0.000   CA. 2343:  11   Min.   :  0.000
```

```
1st Qu.:0.0000   1st Qu.:0.000   1601    :   8   1st Qu.:  7.896
Median :0.0000   Median :0.000   CA 2144 :   8   Median : 14.454
Mean   :0.4989   Mean   :0.385   3101295 :   7   Mean   : 33.295
3rd Qu.:1.0000   3rd Qu.:0.000   347077  :   7   3rd Qu.: 31.275
Max.   :8.0000   Max.   :9.000   347082  :   7   Max.   :512.329
                                 (Other) :1261   NA's   :1
            Cabin      Embarked       ds
C23 C25 C27    :   6   C   :270   test :418
B57 B59 B63 B66:   5   Q   :123   train:891
G6             :   5   S   :914
B96 B98        :   4   NA's:  2
C22 C26        :   4
(Other)        : 271
NA's           :1014
```

En aquest resum ja es pot observar alguns atributs que contenen valors buïts com: `Age`, `Cabin` i `Embarked`.

Com que el que volem predir és la classe resposta `Survived`, observem-ne la seva distribució.

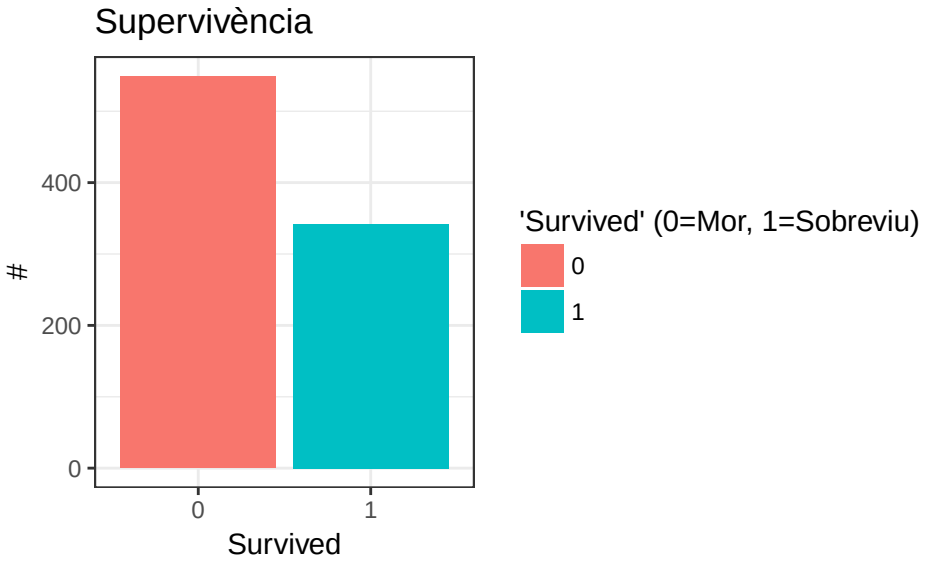


Figura 1: Passatgers que han mort/sobreviscut

En el conjunt de dades que s'ens suministra per a l'entrenament, moren un 62% dels passatgers.

Abans de procedir amb el procés de neteja i preparació de les dades, observem la distribució d'alguns atributs així com algunes relacions entre ells.

Un dels atributs que ens ha semblat que podria resultar significant és `Age`. Comencem analitzant de forma descriptiva la seva distribució de valors.

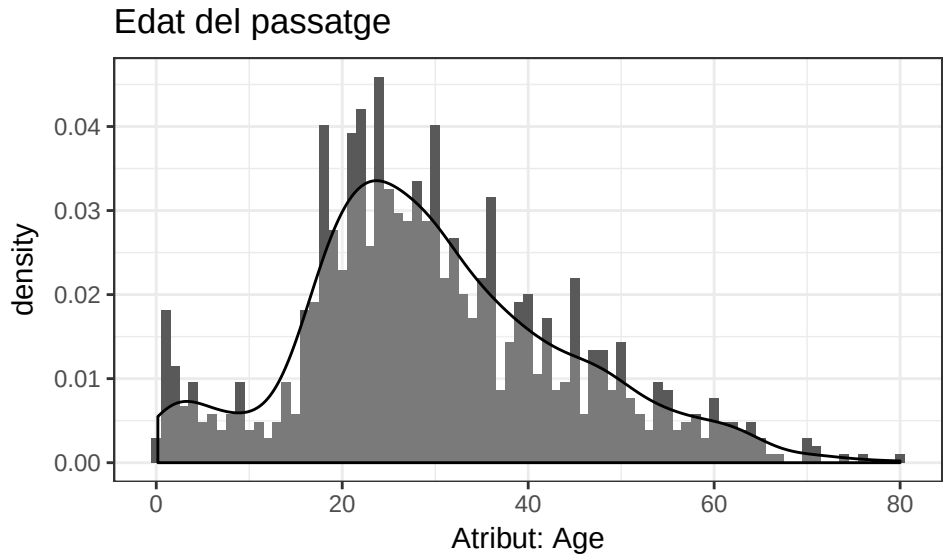


Figura 2: Distribució de l'atribut Age

La distribució de valors de l'edat del passatge té un rang de valors molt dispar, anant d'un mínim d'aproximadament 2 mesos fins a un màxim de 80 anys. S'observen algunes discrepàncies en la distribució en els extrems d'aquests valors, tal com mostra la gràfica Q-Q que segueix.

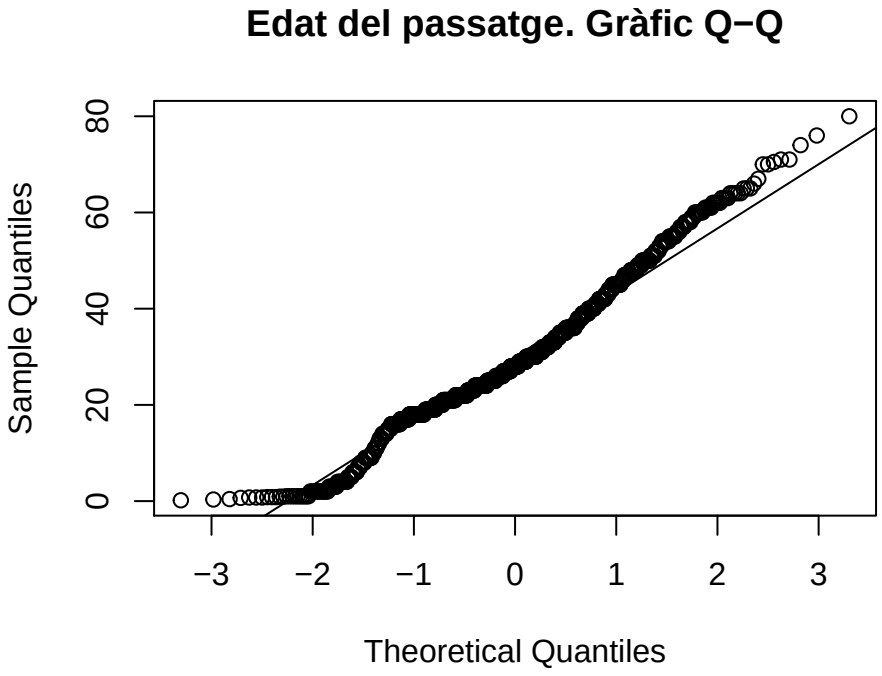


Figura 3: Gràfic Q-Q Normal

Observem-ho, també, en relació als passatgers que sobreviuen, contrastant-ho amb l'atribut `Survived`.

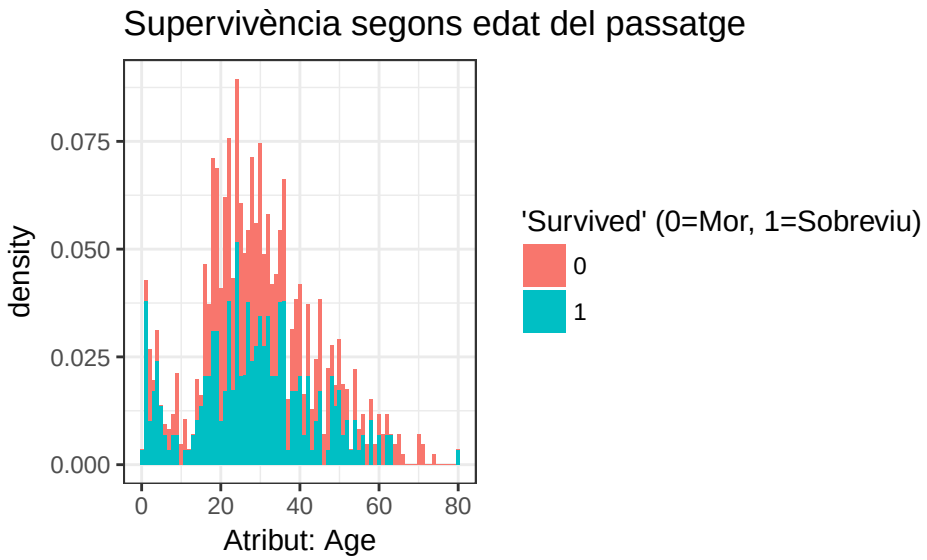


Figura 4: Distribució de Age segons supervivència

Sembla que l'atribut `Age` sí que ens donarà informació rellevant, on els trams baixos d'edat sembla que tindrán més possibilitats de sobreviure. Tot i això, segurament no utilitzarem aquest atribut directament sino que el discretitzarem en grups per tal d'evitar inconsistències en valors individuals del mateix. Això ho realitzarem en el proper apartat d'aquesta PRAC.

Com que hem comentat que el gènere i la classe del passatger també ens semblàva que podrien aportar-nos informació rellevant, vegem com es comporten en relació a la supervivència.

```
ggplot(titanic[titanic$ds=="train",], aes(x=factor(Pclass),fill=Survived)) +
  geom_bar() +
  labs(title="Supervivència segons classe", fill="'Survived' (0=Mor, 1=Sobreviu)") +
  xlab("Classe del passatger") +
  ylab("#") +
  theme_bw()
```

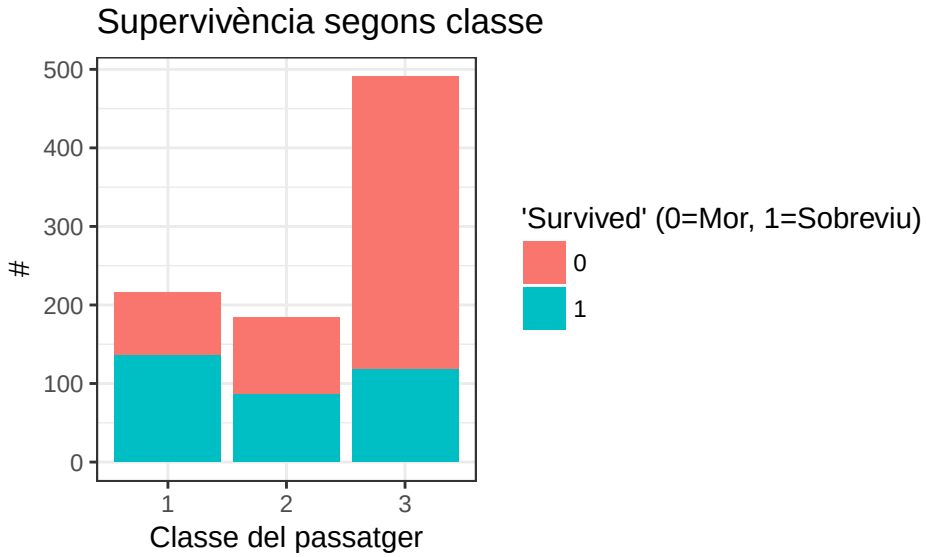


Figura 5: Supervivència segons classe del passatge

Sembla que els passatgers que viatjaven en primera classe, tenien més possibilitats de sobreviure. Dividim-ho, encara més utilitzant l'atribut `Sex`.

```
ggplot(titanic[titanic$ds=="train",], aes(x=factor(Pclass),fill=Survived)) +
  geom_bar() +
  facet_wrap(~Sex) +
  labs(title="Supervivència segons gènere/classe", fill="'Survived' (0=Mor, 1=Sobreviu)") +
  xlab("Classe del passatger") +
  ylab("#") +
  theme_bw()
```

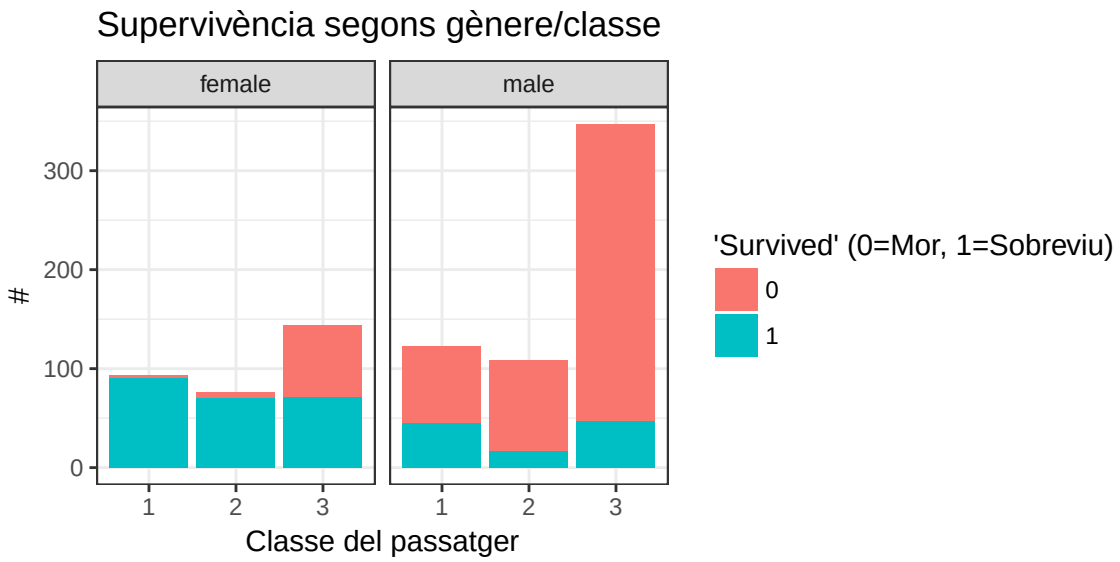


Figura 6: Supervivència segons classe/gènere del passatge

Sembla que tant la classe com el gènere serán bons *predictors* de la possibilitat de supervivència. Com més alta la classe més passatgers van sobreviure i, en general, les dones tenen més taxa de supervivència que els homes. Si recordem el que hem comentat per l'atribut `Age`, sembla que les dades fan honor a la dita *"les dones i els nens primer" (i els rics)*.

Per acabar, mirem la distribució de l'atribut `Fare`, que recordem que representa el preu del bitllet pagat pel passatger, en relació a la supervivència.

```
ggplot(data=titanic[titanic$ds=="train",], aes(x=Fare, fill=Survived)) +
      geom_density(alpha = 0.5) +
      xlab("Atribut: Fare") +
      labs(title="Supervivència segons tarifa", fill="'Survived' (0=Mor, 1=Sobreviu)") +
      theme_bw()
```

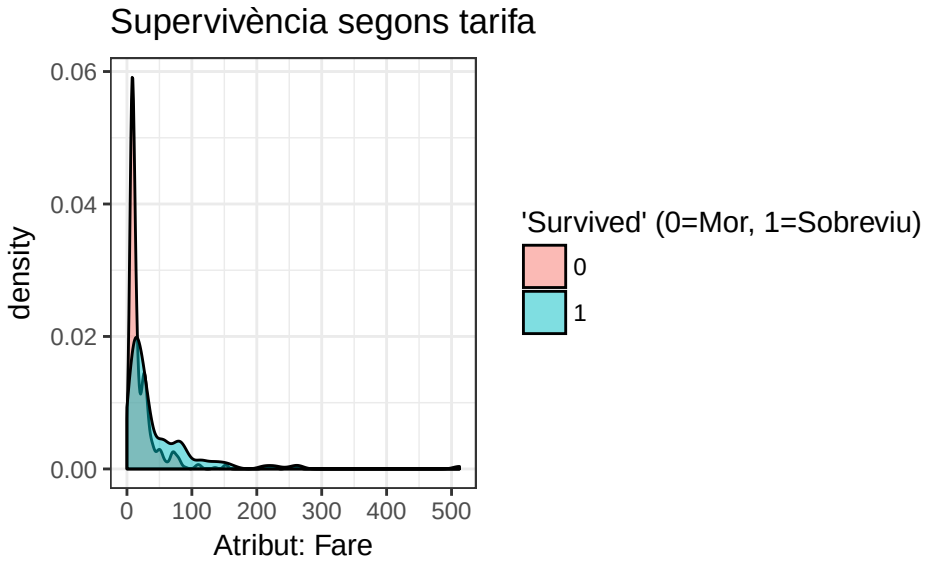


Figura 7: Supervivència contra tarifa

La variable `Fare` té una distribució força asimètrica. La major part de les tarifes pagades pel passatge era de cost baix i algunes de cost alt. Intentem tornar-ho a visualitzar aplicant una transformació logarítmica.

```
ggplot(data=titanic[titanic$ds=="train",], aes(x=log(Fare), fill=Survived)) +
      geom_density(alpha = 0.5, na.rm=TRUE) +
      xlab("Atribut: log(Fare)") +
      labs(title="Supervivència segons tarifa", fill="'Survived' (0=Mor, 1=Sobreviu)") +
      theme_bw()
```

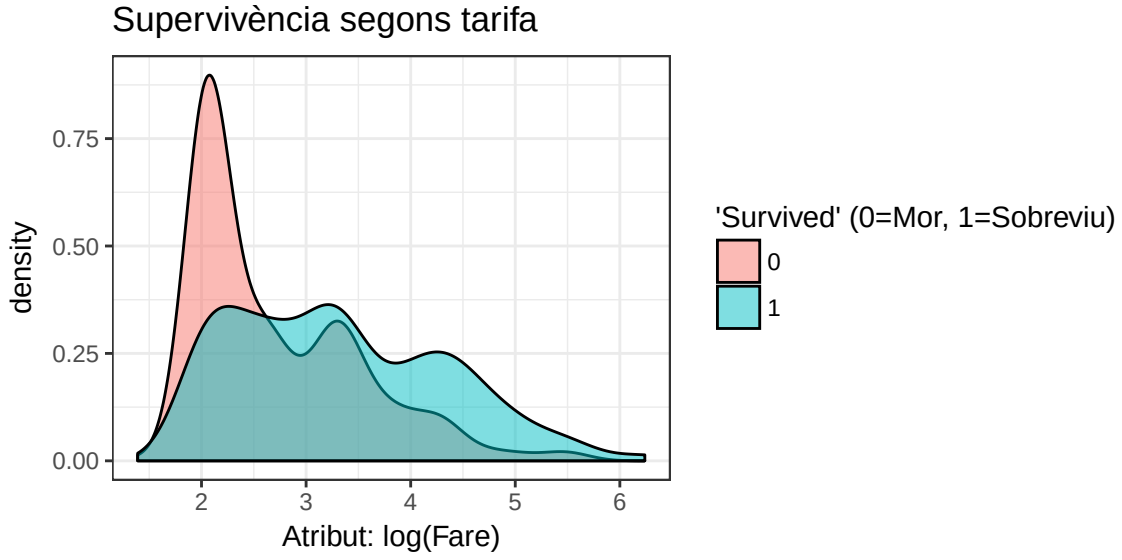


Figura 8: Supervivència contra tarifa

Sembla que el preu mig dels bitllets de les persones que van sobreviure era superior als de la resta. No sembla, però, que el fet d'haver pagat més per un bitllet garantitzi una probabilitat de supervivència major. Es probable que, com hem vist, els passatgers de primera classe tenien més probabilitats de supervivència, potser

perque tenien més facilitats per accedir a un bot salvavides. Entenem que els bitllets de primera classe costen més diners que els de les classes inferiors, d'aquí que el preu mig de la tarifa abonada entre els supervivents sigui superior.

De moment, seleccionarem totes les dades que s'ens proporcionen per a utilitzar-les en els processos de preparació i netejea de les mateixes que implementarem en l'aparatat que segueix. Tal com hem comentat, potser combinarem alguna d'elles en una de nova, en discretitzarem d'altres i, per tant, les acabarem eliminant del nostre conjunt de dades. L'única variable que en aquest punt sembla clar que no aporta informació per a l'anàlisi és la de l'atribut identificador `PassengerId`, pero també necessitem arrosegar-la per a realitzar les prediccions finals en el conjunt de test que ens proporciona *Kaggle*.

4 Preparació i neteja de les dades.

A continuació implementarem diversos processos de preparació i neteja de les dades. Convertirem el tipus d'alguns atributs, tractarem valors buïts i extrems segons el cas, construirem atributs nous que, potser, ens seran de més utilitat que els que s'ens donen, etc.

4.1 Eliminació de valors duplicats

El conjunt de dades representa els detalls de cada passatger i, per tant, no hauria de contenir registres o exemples duplicats. Comprovem-ho.

```
ifelse(length(unique(titanic[,1])) == nrow(titanic), "Sense duplicats.","Hi ha valors duplicats!")
```

```
[1] "Sense duplicats."
```

Tal com suposàvem no tenim valors duplicats en els conjunts de dades subministrats.

4.2 Tractament dels erros en els tipus de dades

En l'apartat anterior hem comentat que en la lectura dels conjunts de dades s'havien assignat malament alguns tipus de dades. Per exemple, l'atribut `Pclass` s'ha llegit com un enter quant en realitat sabem que es tracta d'un *factor* de 3 nivells. A continuació corregirem aquestes incongruències.

```
titanic$Pclass <- as.factor(titanic$Pclass)
```

4.3 Tractament de valors buïts

De la mateixa forma, anteriorment hem comentat que existeixen alguns atributs en el conjunt de dades que contenen valors buïts, o no emplenats. Analitzem de quines variables es tracten.

```
# Busquem NA (valors buits) en els atributs except en la columna
# ds que ens indica el tipus de subconjunt (train, test) i la de
# predicció Survived
colSums(is.na(titanic[, !names(titanic) %in% c("ds", "Survived")]))
```

```
PassengerId      Pclass        Name         Sex         Age       SibSp
          0           0           0           0         263           0
      Parch      Ticket        Fare       Cabin    Embarked
          0           0           1        1014           2
```

Sembla que els atributs en els que tenim valors buïts són: `Age`, `Fare`, `Cabin` i `Embarked`. Anirem tractant cada atribut individualment, decidint l'acció a realitzar per cada un ja que no ens sembla prudent realitzar la mateixa acció per tots.

4.3.1 Cabin

L'atribut amb més valors buits sembla que és `Cabin` amb més d'un 75% del seus valors sense dades.

En aquest cas ens sembla, però, que no estem tractant amb valors buits propiament dits, sino que simplement la gran majoria dels passatgers no tenien un camarot.

```
table(titanic[!is.na(titanic$Cabin),]$Pclass) %>% prop.table() * 100 %>% round(digits=2)
```

```
        1         2         3
86.779661  7.796610  5.423729
```

Sembla que gairebé el 87% dels passatgers que tenien un camarot corresponien a passatgers de primera classe amb el que ens sembla que la nostra hipòtesi pot ser correcta.

En comptes d'eliminar directament aquest atribut o d'imputar-li valors, l'utilitzarem per a crear un nou atribut discret `HasCabin`.

```
titanic$HasCabin <- as.factor(!is.na(titanic$Cabin))
```

Contrastem aquest nou atribut envers l'atribut predictiu i la classe del passatge.

```
ggplot(titanic[titanic$ds=="train",], aes(x=HasCabin,fill=Survived)) +
  geom_bar() +
  facet_wrap(~Pclass) +
  labs(title="Supervivència segons classe/camarot", fill="'Survived' (0=Mor, 1=Sobreviu)") +
  xlab("Classe del passatger") +
  ylab("#") +
  theme_bw()
```

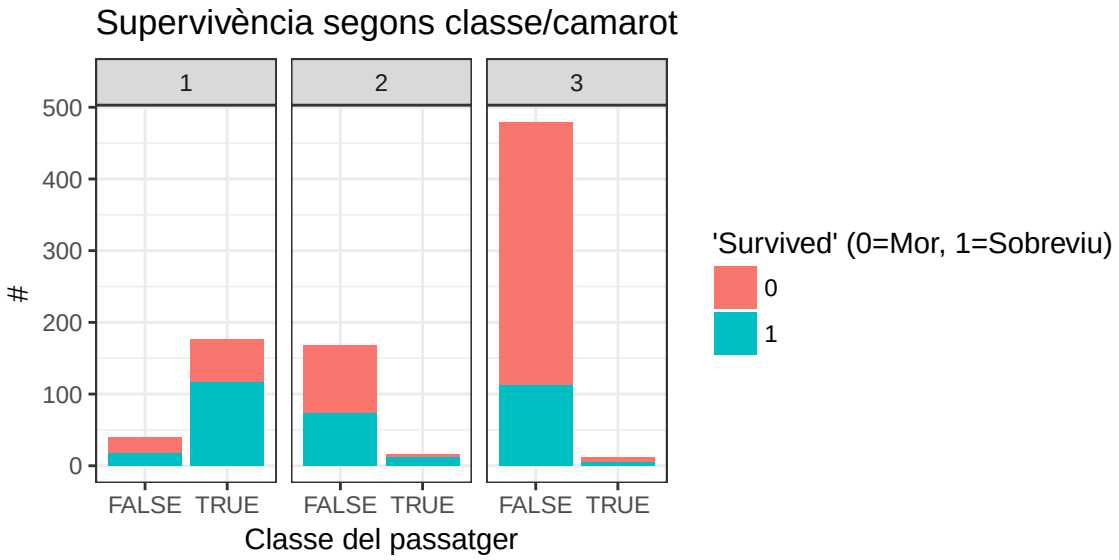


Figura 9: Supervivència segons classe/camarot

Tal com esperàvem, pocs passatgers en 2ª i 3ª classes disposàvem de camarot, però aquells que en tenien un

van tenir una taxa de supervivència més alta.

A partir d'ara només ens interessarà l'atribut `HasCabin` que acabem de crear i podem eliminar l'atribut `Cabin` original.

```
titanic$Cabin <- NULL
```

4.3.2 Embarked

Aquest atribut discret representa el port d'embarcament i hem detectat que tenim dos valors buits en el conjunt de dades.

Examinem quins registres són:

```
kable(titanic[is.na(titanic$Embarked),], format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down")
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	ds	HasCabin
62	62	1	1	Icard, Miss. Amelie	female	38	0	0	113572	80	NA	train	TRUE
830	830	1	1	Stone, Mrs. George Nelson (Martha Evelyn)	female	62	0	0	113572	80	NA	train	TRUE

Vegem que es tracta de dos passatgeres de 1ª classe que viatjàven juntes (tenen el mateix número de tíquet). Del fet que viatgèssin juntes podem suposar que totes dues van embarcar al mateix lloc.

```
table(titanic$Embarked)
```

```
  C   Q   S
270 123 914
```

Gairebé el 70% dels passatgers van embarcar a Southampton. Per al cas d'aquestes dues passatgeres, les imputarem a la classe majoritària dins l'atribut: `S`.

```
titanic$Embarked[c(62, 830)] <- "S"
titanic$Embarked <- as.factor(titanic$Embarked)
```

4.3.3 Fare

En l'atribut `Fare` només tenim un sol valor buit. Mirem quin és.

```
kable(titanic[is.na(titanic$Fare),], format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down")
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	ds	HasCabin
1044	1044	NA	3	Storey, Mr. Thomas	male	60.5	0	0	3701	NA	S	test	FALSE

Es tracta d'un passatger de 3ª classe, varó i d'uns 60 anys d'edat. Com que només tenim un sol valor buit, en aquest cas, imputarem al valor de `Fare` el valor mitjà dels registres similars. Per exemple la mitjana de preu de tarifa pagada en 3ª classe.

```
titanic$Fare[1044] <- mean(titanic[titanic$Pclass=="3",]$Fare, na.rm=TRUE)
```

4.3.4 Age

Finalment tractarem l'últim atribut que contenia valors buits, `Age`. Aquest atribut també conté molts elements amb valor buit. Aproximadament un 20% dels passatgers no tenen aquest atribut informat.

Com que volem aprofitar aquest atribut i són molts registres, emplearem un mètode d'imputació de valors predictiu per mitjà de la llibreria mice.

```
y <- titanic[,c("Pclass", "Sex", "Fare", "Embarked", "SibSp", "Parch", "Age")]
y <- data.frame(y)

ages.pred <- mice(y, method = 'rf')
```

```
 iter imp variable
  1   1  Age
  1   2  Age
  1   3  Age
  1   4  Age
  1   5  Age
  2   1  Age
  2   2  Age
  2   3  Age
  2   4  Age
  2   5  Age
  3   1  Age
  3   2  Age
  3   3  Age
  3   4  Age
  3   5  Age
  4   1  Age
  4   2  Age
  4   3  Age
  4   4  Age
  4   5  Age
  5   1  Age
  5   2  Age
  5   3  Age
  5   4  Age
  5   5  Age
```

```
z.ages <- complete(ages.pred)
```

Comparem ara la distribució de l'atribut `Age` amb les dades originals i les imputades per l'algorisme per veure si es manté.

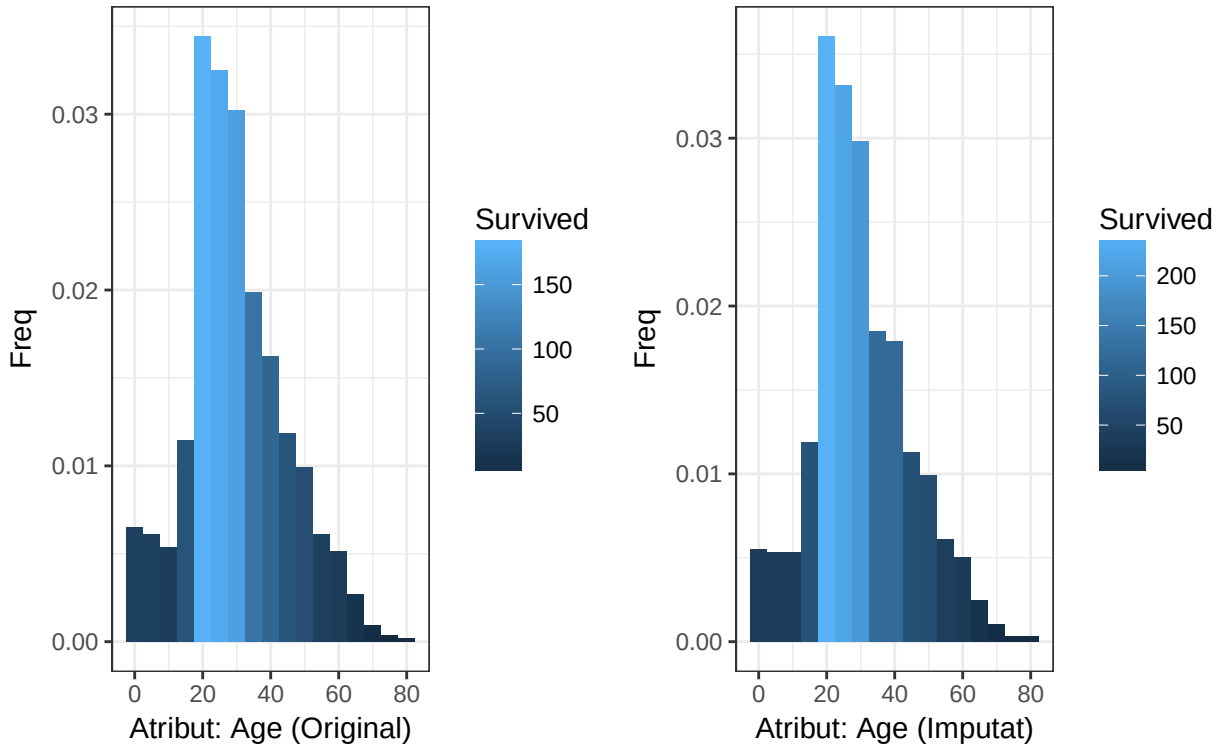


Figura 10: Distribució comparada d'Age (Original/Imputat)

Sembla que la distribució general de l'atribut s'ha mantingut. Finalment, assignem aquest nou atribut amb els valors imputats al conjunt de dades.

```
titanic$Age <- z.ages$Age
```

Finalement ja no tenim valors buits en les dades.

```
colSums(is.na(titanic[, !names(titanic) %in% c("ds", "Survived")]))
```

```
PassengerId      Pclass        Name         Sex         Age       SibSp
          0           0           0           0           0           0
      Parch      Ticket        Fare    Embarked    HasCabin
          0           0           0           0           0
```

4.4 Generació d'atributs nous

En aquest apartat generarem alguns atributs que ens semblen interessants i que, creiem, ens ajudaran en l'anàlisi.

4.4.1 Companions

A partir del nombre de tíquets iguals dels passatgers, establirem el nombre d'acompanyants del mateix. Siguin familiars, empleats de servei, etc. Creiem que és probable que un passatger que no viatjaba sol, tenia més possibilitats de sobreviure.

```
titanic$Companions <- 1;
```

```
for (i in 1:nrow(titanic)) {
  titanic$Companions[i] <- length(titanic$Ticket[titanic$Ticket == titanic$Ticket[i]]);
}
```

4.4.2 AgeRange

En comptes d'utilitzar l'edat directament en el model. Discretitzarem aquest atribut en rangs iguals per mitjà de la funció `quantile`.

```
# Creem els "talls"
age_cuts <- quantile(titanic[titanic$ds=="train", ]$Age, probs = seq(0,1,1/7))
# El primer tall comença a l'edat 0, encara que sigui inexistent...
age_cuts[1] <- 0
age_cuts
```

```
      0% 14.28571% 28.57143% 42.85714% 57.14286% 71.42857% 85.71429%
 0.00000  17.00000  22.00000  26.00000  30.00000  36.00000  45.42857
    100%
80.00000
```

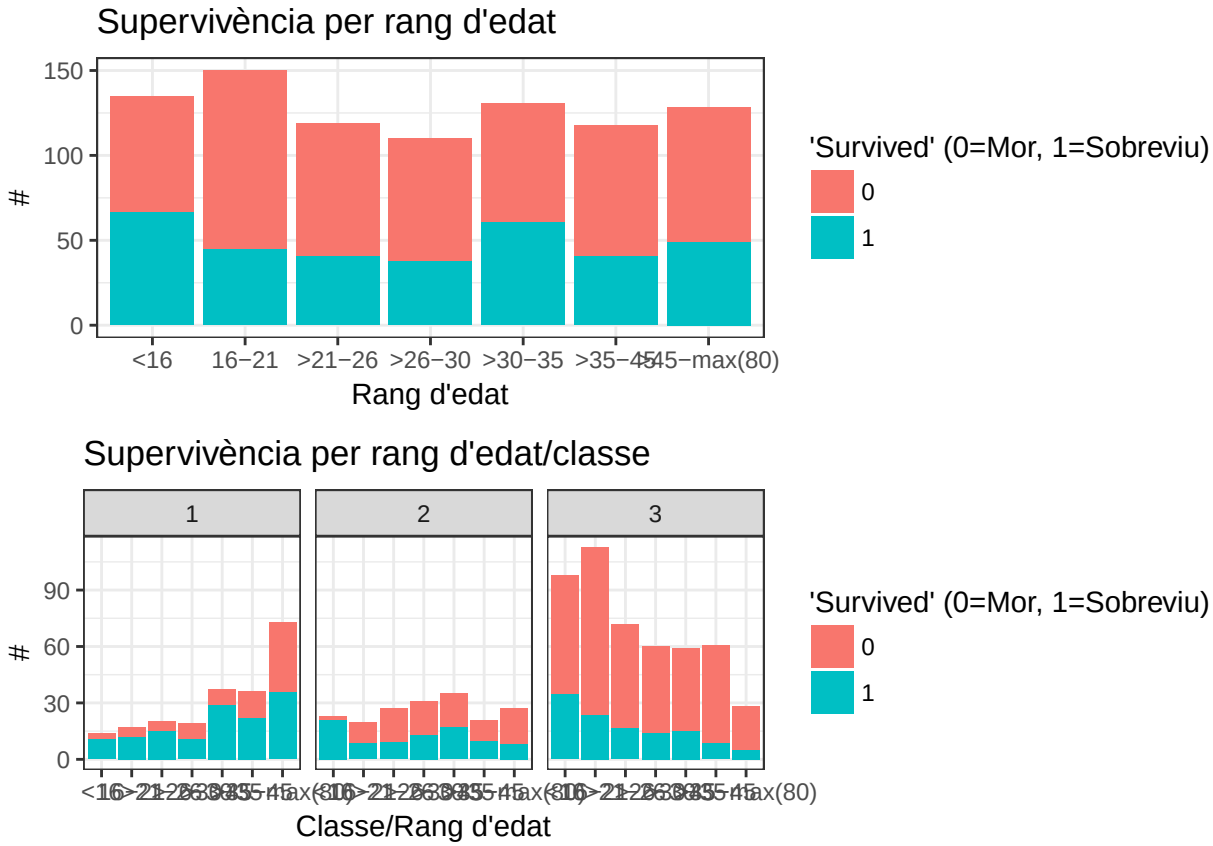
Un cop tenim els talls a partir de 7 quartils, els afegim al conjunt de dades, conjuntament amb les seves etiquetes corresponents.

```
# AgeRange
titanic$AgeRange <-
  factor(cut(titanic$Age, age_cuts, labels = seq(1,7,1), include.lowest = T, right = T))
titanic$AgeRange <- as.factor(titanic$AgeRange)

# Etiquetes AgeRange
ages.df <- data.frame(
  AgeRange=c(1,2,3,4,5,6,7),
  AgeRangeLabels=c("<16", "16-21", ">21-26", ">26-30", ">30-35", ">35-45", ">45-max(80)"))

titanic <- merge(titanic,ages.df,by="AgeRange")
titanic$AgeRangeLabels <- as.factor(titanic$AgeRangeLabels)
```

Un cop les tenim introduides de nou al conjunt de dades, examinem-ne la distribució contra la supervivència dels passatgers, analitzant-ho també per range classe.



4.4.3 Tractament de valors extrems.

En el conjunt de dades, no hem detectat valors *extrems* o *outliers*. Si que es veritat que hi havia valors que es consideràven apartats de la distribució de valors d'algun atribut, com en el cas d'`Age` en els seus extrems mínim i màxim, però hem considerat no tractar-los com a *outliers* i deixar-los en el conjunt de dades, doncs corresponen a dades dels passatgers.

5 Anàlisi de les dades.

5.1 Creació del conjunt de dades net.

Un cop hem aplicat la preparació i neteja de les dades, ara sí que centrarem l'anàlisi només en les variables que ens interessen. En el nostre cas: `Pclass`, `Sex`, `AgeRange`, `AgeRangeLabels`, `HasCabin`, `Companions` i `Embarked`.

Crearem, a continuació un conjunt de dades *net* amb aquests atributs.

```
columns = c("PassengerId", "Pclass", "Sex", "AgeRange", "AgeRangeLabels",
            "HasCabin", "Companions", "Embarked", "Survived")

titanic_train_clean <- titanic[titanic$ds=="train", columns]
write.csv(titanic_train_clean, file = "../data/titanic_train_clean.csv",
          row.names=TRUE)

titanic_test_clean <- titanic[titanic$ds=="test", columns[1:8]]
```

```
write.csv(titanic_test_clean, file = "../data/titanic_test_clean.csv",
          row.names=TRUE)
```

5.2 Predicció de la supervivència amb *Random Forest*

Tot seguit entrenarem un model *Random Forest* per a obtenir un model capaç de preveure la possibilitat de supervivència d'un passatger a partir de les seves dades.

En primer lloc, tornarem a dividir el conjunt d'entrenament en dos: un d'entrenament amb el 75% de les dades i un de test amb la resta. D'aquesta manera podrem comprobar la bondat del model.

```
set.seed(123)

xtr <- createDataPartition(y=titanic[titanic$ds=="train",]$Survived, p=0.75, list=FALSE)

titanic_train <- titanic[titanic$ds=="train",][xtr,]
titanic_test <- titanic[titanic$ds=="train",][-xtr,]
```

Entrenem el model.

```
library(randomForest)
```

```
randomForest 4.6-14
Type rfNews() to see new features/changes/bug fixes.


Attaching package: 'randomForest'
The following object is masked from 'package:dplyr':

    combine
The following object is masked from 'package:gridExtra':

    combine
The following object is masked from 'package:ggplot2':

    margin
```

```
frmla <- factor(Survived) ~
  Pclass + Sex + AgeRange + HasCabin + Companions + Embarked

rf <- randomForest(frmla, ntree=500, data=titanic_train, na.action=na.pass)
rf
```

```
Call:
 randomForest(formula = frmla, data = titanic_train, ntree = 500,      na.action = na.pass)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 20.03%
Confusion matrix:
    0   1 class.error
0 373  39  0.09466019
```

```
1  95 162  0.36964981
```

El model ja estima un error en l'*out-of-bag* (estimador pessimista de l'encert), d'un 0%.

Mostrem la matriu de confusió.

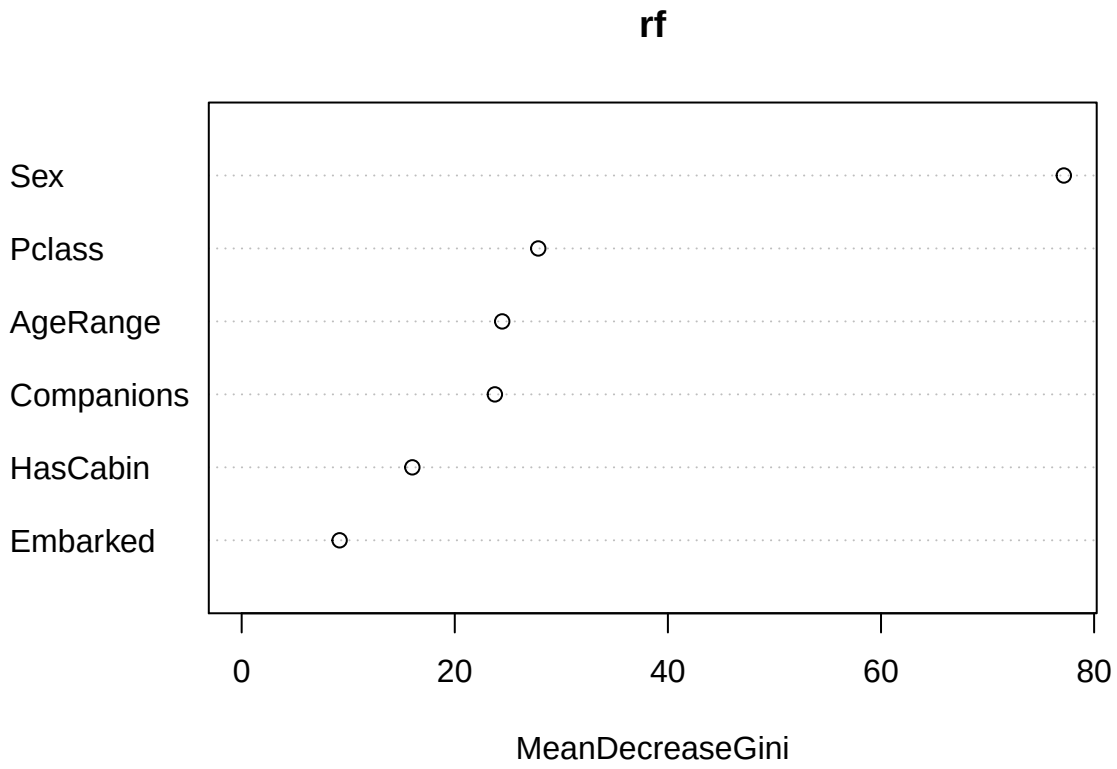
```
rf$confusion
```

```
    0   1 class.error
0 373  39  0.09466019
1  95 162  0.36964981
```

El model produeix molts més errors en la predicció de la supervivència que en la de la mort. Això sembla lògic doncs hi ha molts més passatgers que no ha sobreviscut.

Una qualitat interessant d'utilitzar un model de tipus *Random Forest* és que l'algorisme ens retorna la importància dels atributs que hem utilitzat. Vegem-ho.

```
varImpPlot(rf)
```



Es fàcil veure que el sexe i la classe són els atributs més importants considerats per l'algorisme. En canvi, el port d'embarcament, és el que menys i la variable que havíem construit `HasCabin` no ens està aportant els resultats que esperàvem.

Realitzem la predicció sobre el conjunt de test que hem creat anteriorment per mesurar la qualitat de la mateixa.

```
y_pred <- predict(rf, titanic_test)

y_cm <- confusionMatrix(y_pred, titanic_test$Survived)
```

L'algorisme obté una precisió d'un 83.3 en el conjunt de test.

6 Resultats i conclusions.

El model ha tingut un rendiment tirant a pobre en aquest exercici. Això és perque les variables que hem escollit per a entrenar-lo no l'hi han proporcionat la suficient capacitat de generalització.

L'atribut `Sex` ens ha marcat l'anàlisi i ha conduït la cerca dels arbres de decisió utilitzats per l'algorisme *Random Forest*. Hauria estat bé realitzar, potser, una mica més d'anàlisi i preparació d'atributs per ajudar a l'algorisme a generalitzar.

---

7 Referències

- Kaggle. Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic).
- Bruce, P.; Bruce, A. (2017). Practical Statistics for Data Scientists. O'Reilly.
- UC Business Analytics R Programming Guide. University of Cincinnati (http://uc-r.github.io/).