Pràctica 2: Neteja i anàlisi de les dades

Adrián Alonso Gonzalo i Alexandre Vidal De Palol

Maig/Juny 2022

Contents

1 Descripció i càrrega del dataset.

El conjunt de dades train.csv s'ha obtingut del web https://www.kaggle.com/c/titanic.

Aquest conjunt de dades conté informació sobre la tripulació del Titanic amb 12 tipus de variables i un total de 891 registres (passatgers).

Les variables d'aquesta mostra son:

- PassengerId: Numero de passatger.
- Survived: Supervivència (0=No, 1=Si).
- Pclass: Classe de tiquet (1=Primera, 2=Segona, 3=Tercera) .
- Name: Nom.
- Sex: Sexe.
- Age: Edat.

- SibSp: Germans / Cónjugues a bord del Titanic.
- Parch: Pares / nens a bord del Titanic.
- Ticket: Número de ticket.
- Fare: Preu del ticket.
- Cabin: Numero de cabina.
- Embarked: Port de embarcament.

Lectura del fitxer.

```
# Carreguem els fitxers 'test.cvs' i 'train.csv' de la carpeta 'data/input' (indicant que volem els 'st
test_dataset <- read.csv("../data/input/test.csv", stringsAsFactors=TRUE)
train_dataset <- read.csv("../data/input/train.csv", stringsAsFactors=TRUE)

# Mostrem les primeres files dels jocs de dades
head(test_dataset)
```

```
##   PassengerId Pclass                                          Name    Sex  Age
## 1         892      3                              Kelly, Mr. James   male 34.5
## 2         893      3              Wilkes, Mrs. James (Ellen Needs) female 47.0
## 3         894      2                     Myles, Mr. Thomas Francis   male 62.0
## 4         895      3                              Wirz, Mr. Albert   male 27.0
## 5         896      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0
## 6         897      3                    Svensson, Mr. Johan Cervin   male 14.0
##   SibSp Parch  Ticket    Fare Cabin Embarked
## 1     0     0  330911  7.8292              Q
## 2     1     0  363272  7.0000              S
## 3     0     0  240276  9.6875              Q
## 4     0     0  315154  8.6625              S
## 5     1     1 3101298 12.2875              S
## 6     0     0    7538  9.2250              S
```

```
head(train_dataset)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                             Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                              Heikkinen, Miss. Laina female  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                            Allen, Mr. William Henry   male  35     0     0
## 6                                    Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

```r
# Creem una nova columna amb el nom del dataset del qual provenen les files
test_dataset['row_type'] <- as.factor('test')
train_dataset['row_type'] <- as.factor('train')

# Creem la columna 'Survived' ja que el joc de dades de test no la conté (introduïrem un valor dummy que
test_dataset['Survived'] <- 99

# Juntem els dos jocs de dades en un únic dataset per poder netejar els dos a la vegada
dataset <- rbind(train_dataset, test_dataset[,colnames(train_dataset)])

# Mostrem les primeres files del dataset unificat
head(dataset)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                             Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                              Heikkinen, Miss. Laina female  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                            Allen, Mr. William Henry   male  35     0     0
## 6                                    Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked row_type
## 1        A/5 21171  7.2500              S    train
## 2         PC 17599 71.2833   C85        C    train
## 3 STON/O2. 3101282  7.9250              S    train
## 4           113803 53.1000  C123        S    train
## 5           373450  8.0500              S    train
## 6           330877  8.4583              Q    train
```

```r
# Mostrem l'estructura del dataset
str(dataset)
```

```
## 'data.frame':    1309 obs. of  13 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : num  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 58
##  $ 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/ 187 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
##  $ row_type   : Factor w/ 2 levels "train","test": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
# Mostrem un resum de les estadistiques de les columnes del dataset
summary(dataset)
```

```
##   PassengerId      Survived         Pclass
##  Min.   :   1   Min.   : 0.00   Min.   :1.000
##  1st Qu.: 328   1st Qu.: 0.00   1st Qu.:2.000
##  Median : 655   Median : 1.00   Median :3.000
##  Mean   : 655   Mean   :31.87   Mean   :2.295
##  3rd Qu.: 982   3rd Qu.:99.00   3rd Qu.:3.000
##  Max.   :1309   Max.   :99.00   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  row_type
##                 :1014    :  2    train:891
##  C23 C25 C27    :   6   C:270    test :418
##  B57 B59 B63 B66:   5   Q:123
##  G6             :   5   S:914
##  B96 B98        :   4
##  C22 C26        :   4
##  (Other)        : 271
```

```r
# Mostrem el número de files del joc de dades unificat i per separat
dim(dataset)[1]
```

```
## [1] 1309
```

```r
dim(dataset[dataset$row_type=='test',])[1]
```

```
## [1] 418
```

```r
dim(dataset[dataset$row_type=='train',])[1]
```

```
## [1] 891
```

Observem que que el dataset conté 3 tipus de variables del quals caràcter, numèric i enter.

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

El procés d'integració i selecció de les dades es realitzarà al llarg del procés de neteja i anàlisi de les diverses variables del conjunt de dades d'entrenament del dataset.

En aquest procés es pretén anar analitzant les diferents variables en el procés de neteja i anàlisi, i en funció de les característiques que es vagin observant de les diverses variables es prendrà la decisió d'utilitzar un conjunt seleccionat el qual pugui ser útil per a la predicció del model i la comprovació amb el conjunt de test o validació.

El resultat del projecte pot respondre a possibles causes de mort dels tripulants que no van sobreviure a la tragèdia del Titanic, permetent establir models d'inferència sobre les causes relatives a la mortalitat entre diversos tipus de passatgers. Per altra banda la implementació de un model d'interès sobre quines han sigut variables que han influït més o menys en la supervivència del naufragi.

```
# Placeholder
```

3 Neteja de dades.

3.1 Valors buits (missing values).

```
# Placeholder
```

3.2 Valors extrems (outliers).

```
# Placeholder
```

3.3 Altres accions per a la neteja del joc de dades.

Primerament, observem que la primera variable "PassengerId" no és res més que un identificador, per tant procedim a eliminar-la del conjunt de dades ja que no ens interessa per l'estudi.

```
# eliminació de la primera columna PassengerId
# dataset <- dataset[,-1]
```

Observem que les variables 'Survived' i 'Pclass' són de tipus enter, però la seva funció es indicar una categoria. Per tant, procedim a convertir-les en tipu factor.

```
# transformació de les variables

# dataset$Survived <- as.factor(dataset$Survived)
# class(dataset$Survived)

# dataset$Pclass <- as.factor(dataset$Pclass)
# class(dataset$Pclass)
```

4 Anàlisi de les dades.

4.1 Selecció de grups a analitzar/comparar.

```
# Placeholder
```

4.2 Normalitat i homogeneïtat de la variància.

```
# Placeholder
```

4.3 Aplicació de proves estadístiques per a la comparació de grups.

```
# Placeholder
```

5 Representació de resultats (taules i gràfiques).

```
# Placeholder
```

6 Conclusions.