Pràctica 2: Neteja i anàlisi de les dades

Mireia Olivella i Gabriel Izquierdo

31 de maig de 2020

Índex

1 Descripció del dataset

El conjunt de dades escollit recull informació dels passatgers del titanic, en el que es pot analitzar la supervivència i les característiques d'aquests. Aquest conjunt de dades s'ha obtingut de la web de Kaggle. S'hi pot accedir a partir de l'enllaç que es mostra a continuació:

https://www.kaggle.com/c/titanic

EL conjunt de dades utilitzat està format per 1309 registres amb 12 atributs. Els atributs d'aquest conjunt de dades són els següents:

- **passengerId**: identificador dels registres del dataset.
- **survived**: indica si el passatger va sobreviure (0=No, 1=Sí).
- **pclass**: indica la classe en la que viatjava el passatger (1=1a, 2=2a, 3=3a).
- **name**: nom del passatger.
- **sex**: gènere del passatger (*female* o *male*).
- **age**: edat del passatger.
- **sibsp**: número de germans i cònjugues a bord del Titànic.
- **parch**: número de pares i fills a bord del Titànic.
- **ticket**: número del bitllet
- **fare**: preu de compra del bitllet.
- **cabin**: número de cabina on viatjava el passatger.
- **embarked**: port on va embarcar el passatger (C=Cherbourg, Q=Queenstown, S=Southampton).

2 Neteja de les dades

2.1 Selecció de les variables

2.2 Dades amb elements buits (valors perduts)

2.3 Identificació de valors extrems

3 Anàlisi de les dades

3.1 Comprovació de la normalitat i homogeneïtat de la variància

3.2 Aplicació de proves estadístiques

3.2.1 Matriu de correlació

3.2.2 X^2

3.2.3 Regressió logística

4 Representació dels resultats

5 Resolució del problema