Tipologia i cicle de dades. Pràctica 2

Marcos F. Vilaboa & Joaquim Salomon

22 de mayo de 2019

Índex

---

1 Introducció

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes.

1.1 Competències

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science: - Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l. - Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi.

1.2 Objectius

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.

- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

2 Resolució

2.1 Descripció del *dataset*

El conjunt de dades utilitzat en el present anàlisi s'ha extret de la web kaggle.com. Concretament s'ha utilitzat el *set* d'entrenament (train.csv) que forma part del total de dades de Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic/data).

2.1.1 Càrrega inicial de dades

Per tal de descriure el conjunt, realitzarem una càrrega inicial de les dades amb R:

```
titanic.original <- read.csv("~/R/TCVD/Titanic/data/titanic_train.csv", header=TRUE)
str(titanic.original)
```

```
## '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       : Factor w/ 891 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/ 681 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/ 148 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 ...
```

Inicialment, el *dataset* es composa de 12 variables (columnes) amb un total de 891 observacions (registreS).

2.1.2 Descripció de les variables

La definició de cada camp és la següent:

- ***PassengerId*** (*int*): identificador únic del passatger (i de cada registre).
- ***Survived*** (*int*): si el passatger va sobreviure o no. "0" = No i "1" = Si
- ***Pclass*** (*int*): classe del bitllet d'embarcament. "1" = primera classe, "2" = segona i "3" = tercera.
- ***Name*** (*int*): nom del passatger. Inclou el títol com "Mr.", "Mrs.", "Dr.", ...
- ***Sex*** (*Factor*): gènere del passatger. "female" = dona i "male" = home.
- ***Age*** (*num*): edat.
- ***SibSp*** (*Factor*): nombre de germans i cònjuges a bord.
- ***Parch*** (*int*): nombre de pares i fills a bord.

- ***Ticket*** (*Factor*): número de tiquet.
- ***Fare*** (*num*): tarifa del passatger.
- ***Cabin*** (*Factor*): número de camarot. Consta d'una lletra que significa la coberta i el número de camarot: "A10", "C85",...
- ***Embarked*** (*Factor*): port a on el passatger va embarcar: "C" = Cherbourg, "S" = Southampton i "Q" = Queenstown

2.1.3 Importància i objectius

El Titanic es va enfonsar, durant el seu viatge inaugural el 15 d'abril de 1912, xocant amb un iceberg. Van morir 1502 passatgers i tripulants d'un total de 2224.

La raó principal d'aquest número tan important de víctimes de la tragèdia va ser la quantitat escassa de bots salvavides envers el nombre de vides a bord. Es diu que, per preferència, els nens, les dones i la classe alta tenien més possibilitats de sobreviure.

L'objectiu principal d'aquest estudi és el de conèixer si aquesta afirmació és certa. Es pretén doncs, respondre a la pregunta de quin grup de persones va tenir mes possibilitats de sobreviure i quin tipus de característiques té.

2.2 Pre-processament

2.2.1 Integració i selecció de les dades

TO-DO: Integració i selecció de les dades d'interès a analitzar La integració de les dades consisteix a combinar les dades de diferents fonts de dades. En aquest cas, com que ens basem en un *dataset* concret, no serà necessari integrar més fonts.

En canvi,

```
titanic <- titanic.original
```

2.2.2 Neteja de les dades

TO-DO

2.2.2.1 Zeros y elements buits

TO-DO: Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

2.2.2.2 Valors extrems

TO-DO: Identificació i tractament

2.2.3 Exportació de les dades preprocessades

TO-DO

2.3 Anàlisi de les dades

TO-DO

2.3.1 Selecció dels grups de dades

TO-DO: Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

2.3.2 Comprovació de la normalitat i homogeneïtat de la variància

TO-DO

2.3.3 Aplicació de proves estadístiques

TO-DO: 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.

2.4 Representació dels resultats

TO-DO: Taules i gràfiques

2.5 Resolució del problema

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