**Resolució de la Pràctica 2**

En aquesta segona pràctica es presenta una anàlisis sobre el data set de “Titanic”. Aquest data set conté informació sobre els passatges del RMS Titanic, el famós transatlàntic britànic que es va enfonsar a l’abril del 1912 durant el seu primer viatge des de Southampton fins a Nova York. En concret, hi trobem informació sobre l'edat, el gendre i les característiques socio-econòmiques dels passatgers, així com també si van sobreviure al naufragi o no.

Aquest data set es troba disponible amb llicència pública en el següent enllaç: <https://www.kaggle.com/c/titanic>

**Descripció del dataset**

**Detalls sobre el data set**

De forma més detallada, el data set es troba disponible en format “*comma separated values (csv)*” i conté les següents variables:

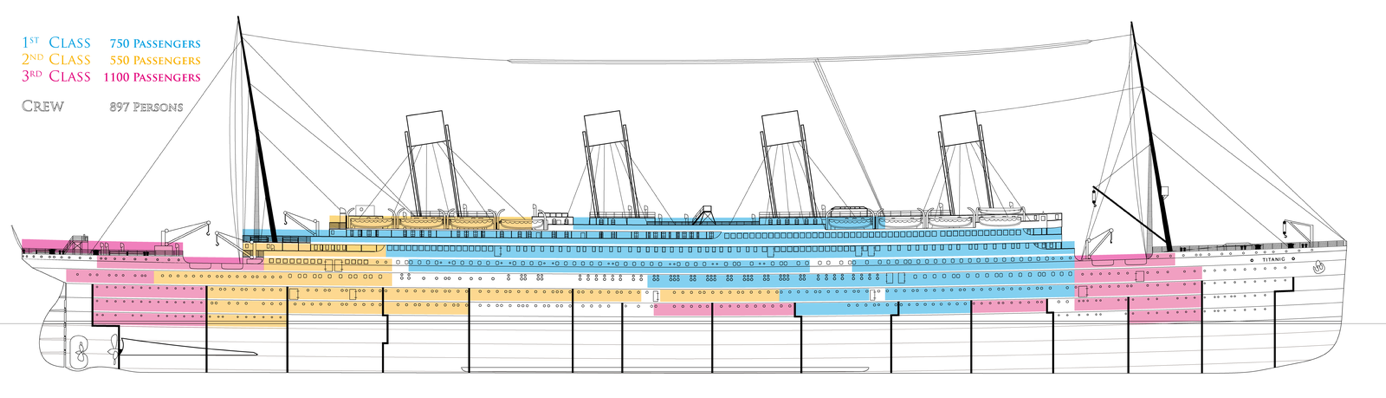
* PassengerId: es tracta d’una variable qualitativa ordinal, on cada valor representa el nom d’un passatge de manera numèrica.
* Survived: la segona variable indica si el passatger va sobreviure o no al naufragi i ho representa amb una variable qualitativa binària, on 0 indica que no ho va fer, mentre que 1 sí va sobreviure.
* Pclass: variable qualitativa ordinal, que pot prendre qualsevol valor enter entre 1 i 3. Cada valor indica el tipus de classe en la que viatjava el passatger.
* Name: variable qualitativa nominal que indica el nom del passatger.
* Sex: variable qualitativa binaria, que indica el gendre del passatger.
* Age: indica l’edat del passatger, pel que es tracta d’una variable quantitativa discreta.
* SibSp: indica el número de germans, germanes i/o germanastres de cada passatger en el vaixell. Es tracta d’una variable quantitativa discreta.
* Parch: Nombre de pares o fills en el vaixell, pel que també és una variable quantitativa discreta.
* Ticket: Número identificador del bitllet del passatger, pel que és una variable qualitativa ordinal.
* Fare**:** Preu pagat pel bitllet. Per tant, es tracta d’una variable quantitativa contínua.
* Cabin: Identificador de la cabina assignada a cada passatger. Es tracta d’una variable qualitativa ordinal.
* Embarked: Port en el que va embarcar el passatger, pel que és una variable qualitativa nominal.

El data set que utilitzarem serà l’anomenat “train.csv”, el qual conté 891 variables. El subdividrem en dos data sets, un per entrenar les dades amb el(s) model(s) escollit(s) per tal de que aquest aprengui i després poder-lo utilitzar per realitzar prediccions. El segon sub-data set s’utilitzarà per estudiar la bondat del model, mitjançant la comparació del valor predit pel model i el valor real de supervivència.

Nota: No és possible utilitzar el data set “test.csv” disponible a l’enllaç mencionat anteriorment, ja que no disposa dels valors reals de supervivència. Per tant, no seria possible analitzar la bondat de predicció del model entrenat.

**Preguntes plantejades per a la resolució de l’activitat**

* Quin és el grup de passatgers que va sobreviure més en percentatge del total? És a dir, hi ha alguna correlació entre l’edat (grup d’edat), el gendre, el fet de tenir fills i el tipus classe? (Fer matriu de correlació entre les diferents variables). Model predictiu: regressió logística. / random forest
* Dins del tipus de classe més alta, van sobreviure més els passatgers (en %) que van pagar un bitllet més car?
* Hi ha alguna diferència de supervivència si un passatger va embarcar en un port o un altre?
* Pel que fa a les dones, van sobreviure més les que estaven casades sense fills (el nom comença per Mrs.) o les que no estaven casades (nom comença per Miss.)?
* Van sobreviure més passatgers amb la cabina a popa que a proa (el vaixell es va començar a enfonsar per proa)? Si és així, quina classe és la que va patir més, donat que gran part de les cabines de la primera classe es trobava a proa?



**Integració i selecció de les dades d’interès a analitzar**

AA

**Neteja de les dades**

AA

**Anàlisi de les dades**

**Representació del resultat**

**Resolució del problema**

**Bibliografia**

1. Titanic: Machine Learning from Disaster | Kaggle [Internet]. Kaggle.com. 2020 [citat 7 de desembre 2020]. Disponible en l’enllaç: <https://www.kaggle.com/c/titanic/overview>