titanic_dataset_project

Manuel Herrera Lara y Anahí Berumen Murillo

15/9/2020

The data domain

Usaremos Machine Learning para crear un modelo que predice cuáles pasajeros sobrevivieron al naufragio del titanic y/o qué tipo de personas tenían mas probabilidades de sobrevivir, usando información de los pasajeros que viajaban en el titanic; como su nombre, edad, sexo, clase socioeconómica, etc. Como breve descripción podemos decir que el hundimiento del titanic fue uno de los naufragios mas infames y recordados de la historia. El RMS Titanic fue un crucero de pasajeros británico que se hundió en el Océano Atlántico Norte y esto sucedió el 15 de abril de 1912, durante su viaje inaugural; y el RMS Titanic, considerado "insumergible", se hundió tras chocar con un iceberg. Desafortunadamente, no había suficientes botes salvavidas para todos a bordo, lo que resultó en la muerte de 1502 de los 2224 pasajeros y la tripulación. Al parecer algunos grupos de personas tenían más probabilidades de sobrevivir que otros. Y por último destacamos que el Titanic era el barco más grande a flote en el momento y fue construido por el astillero Harland and Wolff en Belfast.

Describe each of the variables

» Variable, definición y/o posibles valores

PassengerId Número de identificación al pasajero.
Survived Indica si el pasajero sobrevivió. 0 = No, 1 = Yes.
Pclass Define la clase socioeconómica del pasajero. 1 = Baja, 2 = Media y 3 = Alta.
Name Nombre del pasajero.
Sex Género del pasajero. Masculino y/o Femenino.
Age Edad del pasajero.
SibSp Número de hermanos y/o cónyuges a bordo.
Parch Número de padres y/o niños a bordo.
Ticket Número de boleto.
Fare Tarifa de pasajero.
Cabin Número de cabina.
Embarked Porción de embarcación. (C = Cherbourg; Q = Queenstown; S = Southampton).

Notas adicionales para algunas variables

pclass Indica el status o clase socioeconómica del pasajero.
1 = Baja
2 = Media
3 = Alta
sibsp El dataset define las relaciones familiares de esta forma:
sibling = hermano, hermana, hermanastro, hermanastra.
spouse = esposo y/o esposa.
parch El dataset define las relaciones familiares de esta forma:
parent: mamá o papá.
child: hijo, hija, hermanastro y/o hermanastra.

Basic summary statics

```
# path of the dataset
setwd("/home/chino/Documentos/17_materias_IS/1_mineria_de_datos/4_semana_miniproyecto1/1_titanic_dataset")

# read the dataset
titanic <- read.csv("titanic.csv", stringsAsFactors = FALSE)
```

```
head(titanic)
```

```
##   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
```

```
summary(titanic$PassengerId)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   223.5   446.0   446.0   668.5   891.0
```

```
summary(titanic$Pclass)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.309   3.000   3.000
```

```
summary(titanic$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.42   20.12   28.00   29.70   38.00   80.00     177
```

```
summary(titanic$SibSp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.523   1.000   8.000
```

```
summary(titanic$Parch)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3816  0.0000  6.0000
```

```
summary(titanic$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.91   14.45   32.20   31.00  512.33
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

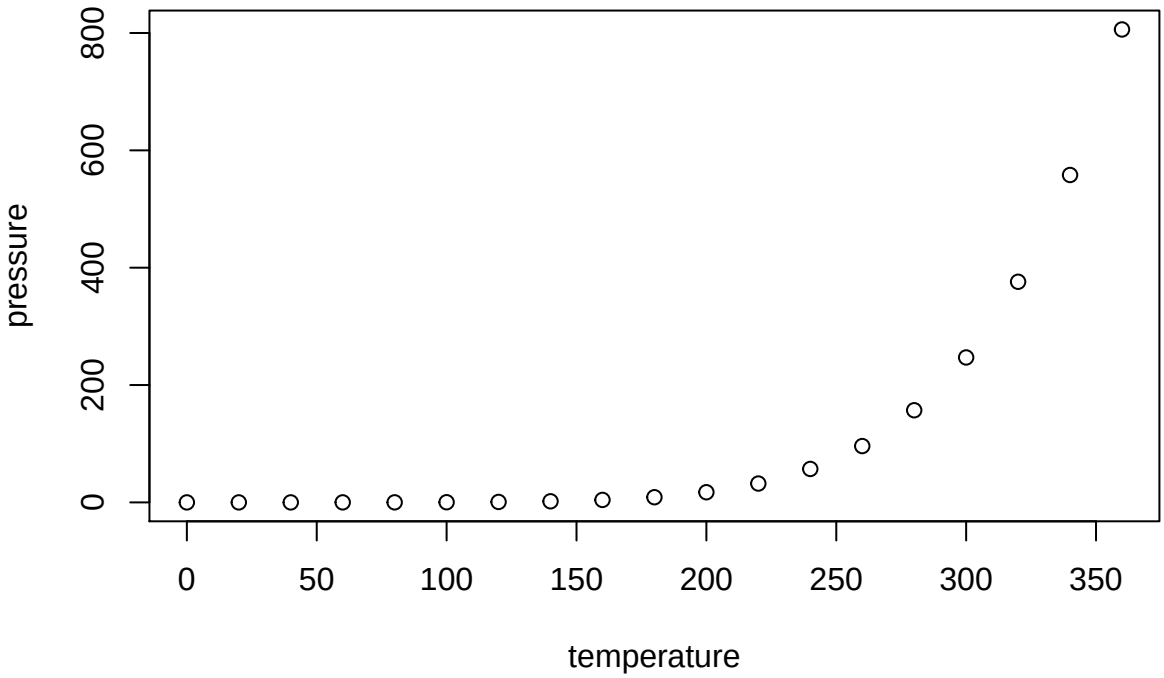
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.