R Notebook

Comenzaremos leyendo el conjunto de Datos para realizar un modelo acerca de la prediccion de supervivencia de los fallecidos del titanic.

EDA

Lo primero que hare sera leer el archivo de entrenamiento. Verificare los tipos de datos de que corresponden para cada columna. Este tipo de problema es un tipo de problema de aprendizaje supervisaado, dado que en el conjunto de entrenamiento se nos provee de una clasificacion binario de 1 y 0, donde especifican si sobrevivio o no.

Los pasos a seguir que haria para resolver este tipo de problema antes de constuir un modelo seria analizar las variables que se incluyen en el conjunto de datos, por ejemplo, si existen relaciones entre ellas, tendencias entre generos, quiza algun factor socioeconomico, encontrar algun tipo de patron.

Lo siguiente seria considerar varios tipos de modelos de aprendizaje supervisados, una buena practica que en experiencia uno adquiere y a base de consejos es comenzar de lo mas simple a lo mas complejo. Por ejemplo antes de utilizar algun tipo de algortimo super complejo como una red nueronal (dado que quiza estemos matando hormigas con una basuca y solo estemos desperdiciando recursos) comenzar un algirtmo sencillo, quiza una regression, luego quiza algun algortimo de clustering como K-NN, y asi sucevivamente.

Explorando la estructura del conjunto de datos.

Es imprecindible, analizar con que tipos de datos estamos lidiando. De un punto de vista Informatico, saber si estamos trabajando con enteros, flotantes, cadena de carecteres entre otros. La razon, para ver cuales son las operaciones mas optimas que podemos realizar, ahora, hablando de un punto de vista estadistico, es importante saber si las variables son nominales, ordinales, factores, etc.

Siguiendo con el sombre de informatico, veamos la estrucutra del conjunto de datos. Para esto, antes de comenzar aunque R nos provee diversos comandos nativos de la biblioteca estandar existen otro tipo de bibliotecas creadas por otros desarrolladores para agilizar el trabajo. A medida que vayamos trabajando iremos agregando las librerias que vayamos necesitando.

Las primeras librerias que utilizaremos eran

- ggplot2 Una libreria para graficos
- dplyr Una libreria para la manipulacion de datos.

```
setwd(dir = "/home/marck/R_coursera/titanic_project/titanic/")
train_set <- read.csv("train.csv")
str(train_set)
```

```
## '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 16 559 520 629 417 581
##  $ 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 ...
```

Aqui lo que hecho me he posicionado en el directorio donde tengo mi base de datos (si, una base de datos puede ser cualquier tipo de archivo que contenga informacion, un excel, un csv, un par de txt, etc) luego, guardamos en una variable el archivo train.csv y mostramos la estructura del archivo con el comando str().

> **Atencion: Toda variable que se crea se guarda en ram. Recuerda que la memoria es finita**

Dentro del conjunto de datos se tienen 12 variables con 819 observaciones. Se puede observar como el comando str() provee el nombre de las clumnas, los tipos de datos y una breve muestra del contenido.

Es posible obtener rapidamente informacion relevante como la media, la mediana, los quartiles, maximos y minimos con un solo comando. Pero esto solo seria significativo en este caso para aquellas variables con valor de tipo numerico.