va

Inteligencia Artificial y Machine Learning
Certificación Avanzada en Ciencia de Datos

Trabajo Práctico 1

Análisis Exploratorio de Datos

Profesora: Dra Marcela Riccillo

Alumno: Federico José Zarategui

Dni: 27.691.042

Ejercicio 1 – Parte teórica

Responda la pregunta asignada en Campus. Copie aquí la pregunta y la respuesta.

¿Por qué es importante testear un modelo de Machine Learning?

Es muy importante testearlo para saber que esta funcionando correctamente. Se debe probar con todas las categorías. Y además probar con casos que no sean del conjunto de entrenamiento para ver si el modelo realmente entendió la lógica.

Ejercicio 2 – Regresión

Los casos de Regresión se caracterizan por tener una variable cuantitativa para
predecir.
Seleccione un dataset con un caso de Regresión. El dataset debe ser obtenido
de alguna librería de R o de una página web pública (no incluir datos
confidenciales).
Por ejemplo, se podría utilizar:

➢ Datasets de R como: mtcars de base, iris de base,

cheddar de faraway, etc.

➢ datasets de UCI (Universidad de California) https://archive.ics.uci.edu

➢ datasets de Kaggle https://www.kaggle.com/

➢ datasets de ISLR https://www.statlearning.com/resources-second-edition

El dataset debe contener al menos 3 variables y una de ellas debe ser
numérica. (Nota: este dataset es solamente para este ejercicio y no se espera
ser utilizado en otros ejercicios).

1) Indique el nombre del dataset, y la librería de R o la página web fuente del mismo.

```
Nombre del dataset: mtcars
De: Librería Base de R
```

2) Identifique la variable a predecir (indique el nombre textual de la variable) y de qué trata el caso a predecir.

La variable a predecir será "am". Según las cualidades que me gustarían en un auto, quiero saber si me conviene que sea manual o automático.

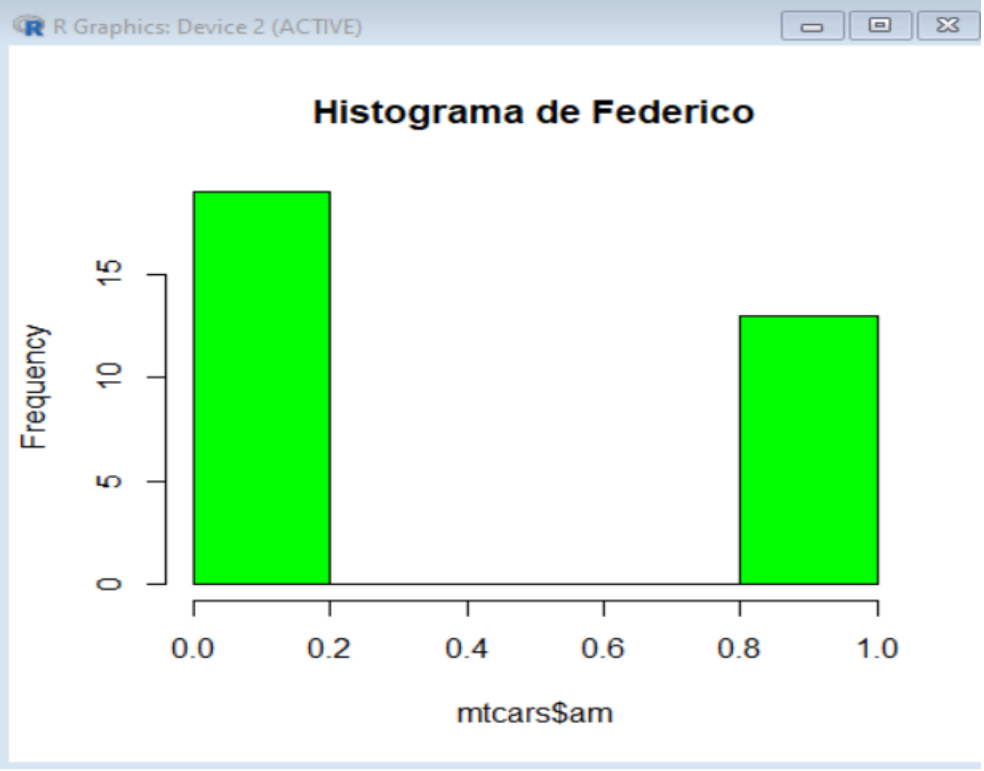
3) Muestre un dim y un summary de la base.

```
> dim(mtcars)
[1] 32 11
> summary(mtcars)
      mpg             cyl             disp             hp             drat             wt             qsec      
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760   Min.   :1.513   Min.   :14.50  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695   Median :3.325   Median :17.71  
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597   Mean   :3.217   Mean   :17.85  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930   Max.   :5.424   Max.   :22.90  
       vs              am              gear            carb      
 Min.   :0.0000   Min.   :0.0000   Min.   :3.000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
 Median :0.0000   Median :0.0000   Median :4.000   Median :2.000  
 Mean   :0.4375   Mean   :0.4062   Mean   :3.688   Mean   :2.812  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :1.0000   Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

4) ¿Cuántos registros tiene la base? ¿Cuántas variables? ¿De qué tipo son las variables?

La base tiene 32 registros y 11 variables (todas numéricas).

5) Realice un histograma de la variable a predecir. ¿En qué rango se encuentran los valores?
hist(base$variable,main="Título",col="color")



Los valores únicamente pueden ser 0 (si es automático) o 1 (si es manual)

a) Para el título ingrese su nombre, como “Histograma de Marcela”.

```
Se indico el nombre “Histograma de Federico”
```

b) Elija un color para el gráfico. Tenga en cuenta que ingresa colors() en R verá que hay +500 colores posibles.

```
Color elegido: “green1”
```

c) Indique el código R utilizado.

```
hist(mtcars$am,main="Histograma de
Federico",col=”green1”)
```

Parte B – Conjuntos

1) Considere su DNI para el seteo de la semilla y particione la base en un conjunto de entrenamiento y uno de testeo con la librería caret.
Además, si su DNI termina en 0, 1, 2 ó 3
Setee p=0.70

Si su DNI termina en 4, 5, 6 ó 7
Setee p=0.75
Si su DNI termina en 8 ó 9
Setee p=0.80
set.seed(DNI);particion=createDataPartition(y=BASE$Variable APred,p=asignado,list=FALSE)
entreno=BASE[particion,]
testeo=BASE[-particion,]
Indique cómo quedó el código R utilizado.

```
set.seed(27691042);particion=createDataPartition
(y=mtcars$am,p=0.7,list=F)
entreno=mtcars[partición,]
testeo=mtcars[-particion,]
```

2) Muestre un head y un summary del conjunto de entrenamiento y del conjunto de testeo.
head(entreno)
summary(entreno)
head(testeo)
summary(testeo)

```
> head(entreno)
                mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4      21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag  21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710     22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive 21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Valiant        18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Merc 240D      24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
> summary(entreno)
      mpg             cyl             disp             hp             drat             wt             qsec      
 Min.   :10.40   Min.   :4   Min.   : 71.1   Min.   : 52.0   Min.   :2.760   Min.   :1.513   Min.   :14.50  
 1st Qu.:16.10   1st Qu.:4   1st Qu.:130.6   1st Qu.: 94.0   1st Qu.:3.115   1st Qu.:2.470   1st Qu.:16.96  
 Median :19.70   Median :6   Median :167.6   Median :113.0   Median :3.770   Median :3.190   Median :17.82  
 Mean   :20.58   Mean   :6   Mean   :222.8   Mean   :141.2   Mean   :3.661   Mean   :3.201   Mean   :17.93  
 3rd Qu.:23.60   3rd Qu.:8   3rd Qu.:302.5   3rd Qu.:177.5   3rd Qu.:3.920   3rd Qu.:3.515   3rd Qu.:18.90  
 Max.   :33.90   Max.   :8   Max.   :472.0   Max.   :335.0   Max.   :4.930   Max.   :5.424   Max.   :22.90  
       vs               am              gear            carb      
 Min.   :0.0000   Min.   :0.0000   Min.   :3.00   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.00   1st Qu.:2.000  
 Median :0.0000   Median :0.0000   Median :4.00   Median :2.000  
 Mean   :0.4783   Mean   :0.4783   Mean   :3.87   Mean   :2.957  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.00   3rd Qu.:4.000  
 Max.   :1.0000   Max.   :1.0000   Max.   :5.00   Max.   :8.000  
```

```
> head(testeo)
                  mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Fiat 128          32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Toyota Corona     21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
> summary(testeo)
      mpg             cyl             disp             hp             drat             wt             qsec      
 Min.   :13.30   Min.   :4.000   Min.   : 78.7   Min.   : 66.0   Min.   :2.760   Min.   :2.200   Min.   :15.41  
 1st Qu.:15.20   1st Qu.:4.000   1st Qu.:121.0   1st Qu.:109.0   1st Qu.:3.070   1st Qu.:2.780   1st Qu.:16.87  
 Median :17.30   Median :8.000   Median :275.8   Median :175.0   Median :3.210   Median :3.520   Median :17.60  
 Mean   :18.84   Mean   :6.667   Mean   :251.0   Mean   :160.8   Mean   :3.431   Mean   :3.258   Mean   :17.65  
 3rd Qu.:21.40   3rd Qu.:8.000   3rd Qu.:350.0   3rd Qu.:180.0   3rd Qu.:3.730   3rd Qu.:3.730   3rd Qu.:18.60  
 Max.   :32.40   Max.   :8.000   Max.   :360.0   Max.   :245.0   Max.   :4.110   Max.   :3.840   Max.   :20.01  
       vs               am              gear            carb      
 Min.   :0.0000   Min.   :0.0000   Min.   :3.000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
 Median :0.0000   Median :0.0000   Median :3.000   Median :2.000  
 Mean   :0.3333   Mean   :0.2222   Mean   :3.222   Mean   :2.444  
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:3.000   3rd Qu.:3.000  
 Max.   :1.0000   Max.   :1.0000   Max.   :4.000   Max.   :4.000  
```

3) ¿Cuántos registros quedaron en cada conjunto (entrenamiento y testeo) en total?

Quedaron 23 registros en entreno y 9 en testeo.

Anexo con el código en R utilizado:

```
mtcars

dim(mtcars)

summary(mtcars)

hist(mtcars$am,main="Histograma de
Federico",col=”green1”)

set.seed(27691042);particion=createDataPartition
(y=mtcars$am,p=0.7,list=F)

entreno=mtcars[partición,]

testeo=mtcars[-particion,]

head(entreno)
summary(entreno)
head(testeo)
summary(testeo)
```