Diplomatura en Ciencia de Datos

Módulo 2: Machine Learning

TP Nro 1°

Profesor: Marcela Leticia Riccillo

Nombre: Ivan Kim

DNI: 37.687.040

EJERCICIO 1 - TEORÍA

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

Como se ha dicho en clase siempre hay que testear un modelo de machine learning primero para comprobar que funcionen y luego poder compararlo con otros modelos para elegir cuál se ajusta mejor a los datos.

Por otro lado, para evitar overfitting o que aprende de memoria sino que generalice y pueda predecir y generar resultados en base a lo aprendido.

Esto hace que el modelo tenga confiabilidad para la toma de decisiones de casos reales por ejemplo en salud, donde es importante la precisión de los resultados y reducir riesgos potenciales.

EJERCICIO 2 - REGRESIÓN

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

El dataset que vamos a usar es de la librería R **"EuStockMakerts"**

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 sería SMI (Swiss Market Index) donde su objetivo muestra el valor del índice de bolsa de Suiza en relación con los otros índices dentro de la base que son Alemania (DAX), Francia (CAC) e Inglaterra (FTSE).

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

Aca se muestra el summary de la base de datos.
La mínima, la máxima, los rangos intercuartiles, la media y la mediana.
summary(EuStockMarkets)

```
      DAX             SMI             CAC             FTSE
 Min.   :1402   Min.   :1587   Min.   :1611   Min.   :2281
 1st Qu.:1744   1st Qu.:2166   1st Qu.:1875   1st Qu.:2843
 Median :2141   Median :2796   Median :1992   Median :3247
 Mean   :2531   Mean   :3376   Mean   :2228   Mean   :3566
 3rd Qu.:2722   3rd Qu.:3812   3rd Qu.:2274   3rd Qu.:3994
 Max.   :6186   Max.   :8412   Max.   :4388   Max.   :6179
```

El dim(EuStockMarkets) muestra la cantidad de registros y columnas de la base.

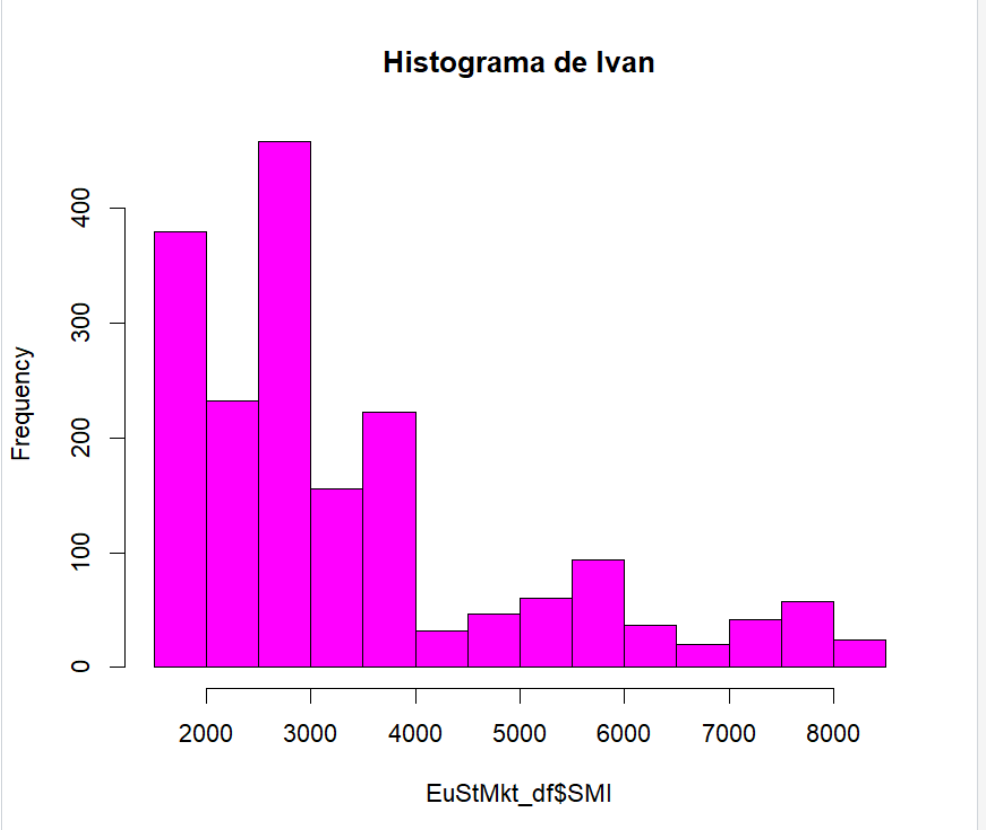
```
[1] 1860    4
```

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

4 variables y 1860 registros.
Variable del tipo numérico continuo.

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

El rango de los valores es entre 2000 a 8000 (valor de mercado)
y la frecuencia es de 0 a 400 (dias donde ese valor se mantuvo dentro de ese rango)



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

El título utilizado fue "Histograma de Iván"

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

El color usado para el histograma fue "Magenta"

c) Indique el código R utilizado.

```
data(EuStockMarkets)
```

```
EuStMkt_df <- as.data.frame(EuStockMarkets)

view(EuStMkt_df)

str(EuStockMarkets)

dim(EuStockMarkets)

summary(EuStockMarkets)

hist(EuStMkt_df$SMI, main = "Histograma de Ivan", col = "magenta")
```

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$VariableAPred,p=asignad
o,list=FALSE)
entreno=BASE[particion,]
 testeo=BASE[-particion,]
```

Indique cómo quedó el código R utilizado.

```
library(caret)
set.seed(37687040)
particion <- createDataPartition(y = EuStMkt_df$SMI, p = 0.70, list = FALSE)
entreno <- EuStMkt_df[particion, ]
testeo  <- EuStMkt_df[-particion, ]

dim(entreno) - dio 1304 registros
dim(testeo)  - dio 556 registros
```

```
> dim(entreno)
[1] 1304    4
> dim(testeo)
[1] 556   4
```

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

head(entreno) - muestra los valores de la base de entreno.

```
      DAX    SMI    CAC   FTSE
1 1628.75 1678.1 1772.8 2443.6
4 1621.04 1684.1 1708.1 2470.4
5 1618.16 1686.6 1723.1 2484.7
6 1610.61 1671.6 1714.3 2466.8
7 1630.75 1682.9 1734.5 2487.9
8 1640.17 1703.6 1757.4 2508.4
```

summary(entreno) - La mínima, la máxima, los rangos intercuartiles, la media y la mediana de la base entreno.

```
      DAX             SMI             CAC             FTSE
 Min.   :1402   Min.   :1596   Min.   :1611   Min.   :2285
 1st Qu.:1742   1st Qu.:2166   1st Qu.:1876   1st Qu.:2839
 Median :2143   Median :2796   Median :1995   Median :3234
 Mean   :2536   Mean   :3384   Mean   :2230   Mean   :3569
 3rd Qu.:2722   3rd Qu.:3812   3rd Qu.:2275   3rd Qu.:3992
 Max.   :6186   Max.   :8401   Max.   :4388   Max.   :6179
>
```

head(testeo) muestra los valores de la base de testeo.

```
       DAX    SMI    CAC   FTSE
2  1613.63 1688.5 1750.5 2460.2
3  1606.51 1678.6 1718.0 2448.2
9  1635.47 1697.5 1754.0 2510.5
11 1647.84 1723.8 1759.8 2532.5
14 1621.49 1733.3 1757.5 2547.3
16 1627.63 1728.3 1762.8 2558.5
```

summary(testeo) - La mínima, la máxima, los rangos intercuartiles, la media y la mediana de la base testeo.

```
      DAX              SMI              CAC              FTSE
 Min.   :1454   Min.   :1587   Min.   :1646   Min.   :2281
 1st Qu.:1745   1st Qu.:2178   1st Qu.:1873   1st Qu.:2847
 Median :2134   Median :2798   Median :1985   Median :3265
 Mean   :2517   Mean   :3358   Mean   :2222   Mean   :3559
 3rd Qu.:2711   3rd Qu.:3806   3rd Qu.:2259   3rd Qu.:3994
 Max.   :6184   Max.   :8412   Max.   :4340   Max.   :6133
```

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

Seteando el p.valor a 0.70
Se usa la base en 70% de entrenamiento y 30% de testeo.

nrow(entreno) → 1304 Registros
nrow(testeo) → 556 Registros

```
> nrow(testeo)
[1] 556
> nrow(entreno)
[1] 1304
>
```

ANEXO CODIGO R

```
# PARTE A

data(EuStockMarkets)
EuStMkt_df <- as.data.frame(EuStockMarkets)

view(EuStMkt_df)

str(EuStockMarkets)

dim(EuStockMarkets)

summary(EuStockMarkets)

hist(EuStMkt_df$SMI, main = "Histograma de Ivan", col = "magenta")

# PARTE B
library(caret)
set.seed(37687040)
particion <- createDataPartition(y = EuStMkt_df$SMI, p = 0.70, list = FALSE)
entreno <- EuStMkt_df[particion, ]
testeo  <- EuStMkt_df[-particion, ]

dim(entreno)
dim(testeo)

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

nrow(testeo)
nrow(entreno)
```