17-5-2022

Diplomatura en BIG DATA

Data Mining & Machine Learning

ALUMNO: Escalada Christian, DNI: 33549575
PROFESORA: MARCELA LETICIA RICCILLO

Trabajo Práctico N°1

Análisis Exploratorio de Datos – Iris Dataset

Enunciados:

Ejercicio 1 – *Parte teórica*

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

Ejercicio 2 – *Clasificación*

Parte A: *Análisis Exploratorio de Datos*

1. Abra el dataset iris:
    a. data(iris) .
    b. Escriba en R ?iris y copie el párrafo de Description.
    c. ***Optativo:*** busque en Internet alguna imagen de flores iris: *setosa*, *versicolor* y/o *virginica*. Indique la página web origen de la imagen.
2. Muestre dim(iris) y str(iris):
    a. ¿Cuántas variables tiene el dataset?
    b. ¿Qué significa cada variable?
3. Muestre summary(iris$Species)
    a. ¿Cuántas flores por cada especie hay en el dataset?
4. Realice un gráfico de barras de la variable a predecir:
    a. Elija un título y color.
    b. Indique cómo quedó el código R.
5. Muestre un summary(iris):
    a. ¿En qué rango se encuentran los largos de los sépalos de las flores?
6. Elija 2 variables cuantitativas (que **no sean *Species***)
    a. Realice un gráfico de dispersión con esas 2 variables.
    b. Coloréelo según la variable a predecir ***Species***.
    c. Elija un pch y con main=”” .
    d. Puede agregar un título al gráfico.
    e. Indique cómo quedó el código R.
7. Con la instrucción iris[numFlor,] se puede obtener los datos de una flor de la base. Considere los 2 últimos números de su DNI.
    a. Muestre los datos de la flor de esa posición.
    b. ¿De qué especie es esa flor?
    c. ***Optativo:*** realice un esquema o dibujo de las medidas de la flor seleccionada.

Parte B: *Conjuntos*

1. Considere los 3 últimos dígitos de su DNI para el seteo de semilla.
   a. Particione la base en un conjunto de entrenamiento y uno de testeo.
   b. Además:
      i. Si su DNI termina en **0, 1, 2, 3**: Setee p=0.70
      ii. Si su DNI termina en **4, 5, 6, 7**: Setee p=0.75
      iii. Si su DNI termina en **8, 9**: Setee p=0.80
   c. Indique el código R utilizado.
2. ¿Cuántos elementos quedaron en total en el conjunto de entrenamiento y de testeo?
3. Muestre:
   a. head(entreno)
   b. head(testeo)
   c. str(entreno)
   d. str(testeo)
4. Muestre:
   a. summary(iris$Species)
   b. summary(entrenar$Species)
   c. summary(testeo$Species)
   d. Verifique que en entreno y testeo haya quedado el porcentaje esperado de elementos según la partición con createDataPartition.
5. Escriba entreno – Enter y luego testeo – Enter.
   a. La flor que obtuvo antes, ¿quedó en entrenamiento o en testeo?

Ejercicio 1 – Parte Teórica

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

- Es importante testear un modelo de Machine Learning porque es justamente en **Test** donde se puede verificar que el algoritmo realmente aprendió lo que se le trataba de enseñar en la parte de **Train**, además de que en este apartado es donde realmente se le puede medir su verdadera performance, para luego así **deployar** el algoritmo en proyectos de la vida real.

Ejercicio 2 – Parte Práctica

Parte A: Análisis Exploratorio de Datos

Respuestas 1:

- Abra el `dataset iris`.

```
data(iris)
```

- Escriba en R `?iris` y copie el párrafo de Description.

```
?iris

## starting httpd help server ... done

Description
This famous (Fisher's or Anderson's) iris data set gives the measurements in centimete
rs of the variables sepal length and width and petal length and width, respectively, f
or 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor,
and virginica.
```

- **Optativo:**

https://jdvelasq.github.io/courses/_images/iris.png

Especies de Flores Iris

Respuestas 2:

- Muestre `dim(iris)` y `str(iris)`.

```
dim(iris)

## [1] 150   5

str(iris)

## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 .
..
```

- El Dataset de Iris tiene 5 variables, de las cuales **Species** es la variable ***categórica cualitativa***, y **las restantes** son las variables ***numéricas cuantitativas***.
- La variable **Species** indica la ***especie a la que pertenece la flor***, mientras que **sepal.length & sepal width** indican el ***largo y ancho del sépalo de la flor*** y por último **Petal.Length & Petal.Width**, indican el ***largo y ancho del pétalo de la flor***.

Respuestas 3:

- Muestre `summary(iris$Species)`.

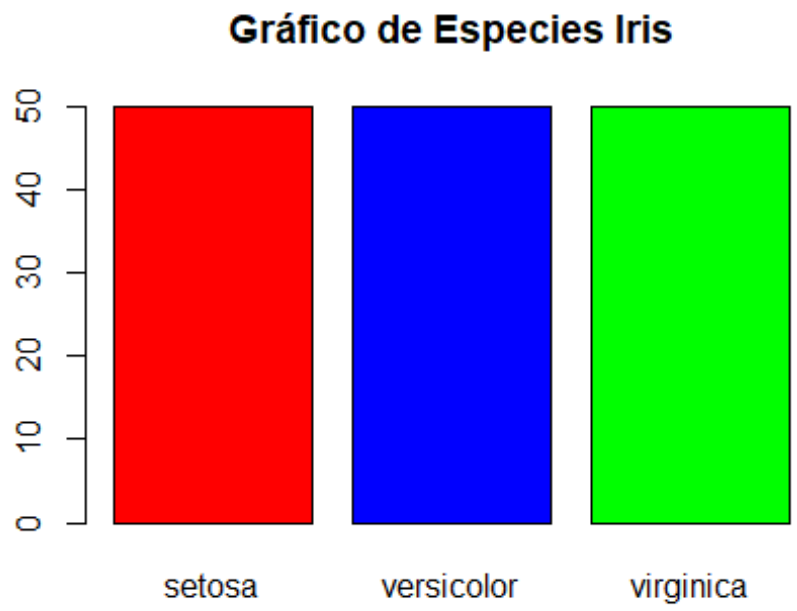
```
summary(iris$Species)

##     setosa versicolor  virginica
##         50         50         50
```

- Por cada especie hay ***50 flores*** en el dataset.

Respuestas 4:

- Gráfico de barras de la **variable Species**:

```
plot(iris$Species, main="Gráfico de Especies Iris",col=c("red", "blue", "green"))
```



Respuestas 5:

- Muestre un summary(iris).

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

- El largo de los sépalos de las flores se encuentra en el rango de los 4.3cm - 7.9cm.

Respuestas 6:

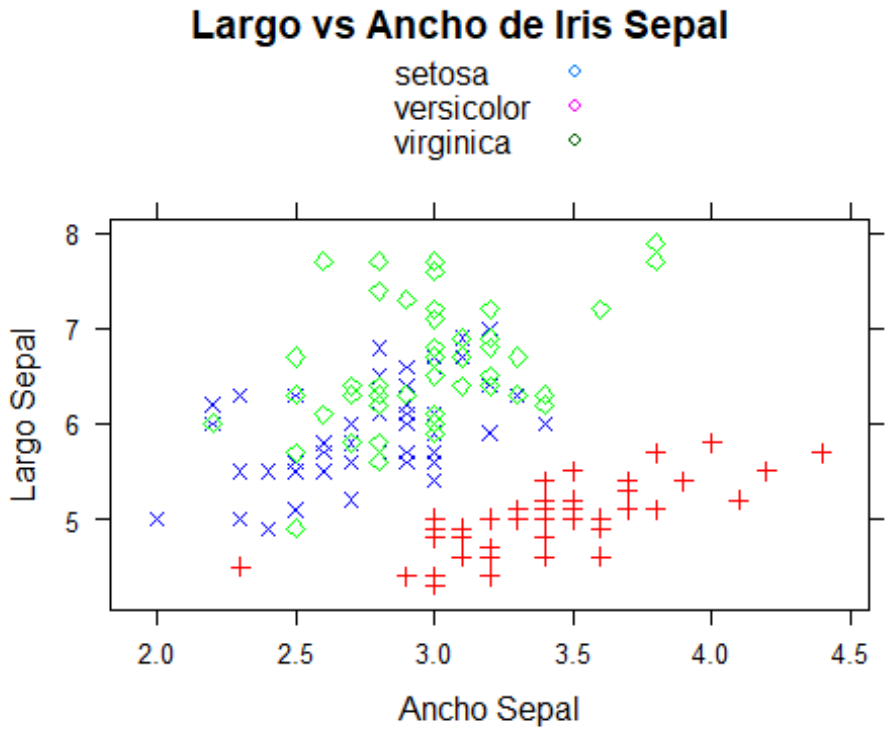
- Elija 2 variables y grafique.

```
# Primero importamos librerias
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
# Graficamos:
xyplot(iris$Sepal.Length~iris$Sepal.Width, groups=iris$Species, pch=c(3,4,5),col=c("red", "blue", "green"),auto.key=TRUE, main="Largo vs Ancho de Iris Sepal",xlab="Ancho Sepal", ylab="Largo Sepal")
```



- Elegí las variables cuantitativas: **Sepal.Length & Sepal.Width** para realizar el gráfico de dispersión requerido, además para representar a cada una de las flores elegí un ***pch de 3, 4, 5***, que representan ***un rombo, una cruz y un signo suma***, los ***colores*** que utilicé para cada especie son el ***rojo, azul, verde***, por último a mi gráfico le agregue un ***título: "Largo vs Ancho de Iris Sepal"***.

Respuestas 7:

- Muestre los datos de la flor de la posición: `iris[2numDNI,]`

```
flor= iris[75,]
flor

##    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 75          6.4         2.9          4.3         1.3 versicolor
```

- La Flor que se muestra según mis dos últimos números de mi DNI es: una ***Especie Versicolor***
- **Optativo:**

Especie Versicolor: Click Aquí

Especie Versicolor

Parte B: Conjuntos

Respuestas 1:

- Mis 3 últimos números de mi DNI son: 575.
- Mi último número de DNI es 5.
- p va a ser igual a 0.75.
- El código va a quedar:

```
set.seed(575); particion=createDataPartition(y=iris$Species,p=0.75,list=FALSE)
entreno=iris[particion,]
testeo=iris[-particion,]
```

Respuestas 2:

- En el conjunto de Entrenamiento quedaron 114 elementos.

```
dim(entreno)
```

```
## [1] 114   5
```

- En el conjunto de Testeo quedaron 36 elementos.

```
dim(testeo)
```

```
## [1] 36  5
```

Respuestas 3:

- Muestre:

```
# Observamos las primeras filas de Train.
head(entreno)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 8          5.0         3.4          1.5         0.2  setosa
```

```
# Observamos las primeras filas de Test.
head(testeo)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 6           5.4         3.9          1.7         0.4  setosa
## 7           4.6         3.4          1.4         0.3  setosa
## 13          4.8         3.0          1.4         0.1  setosa
## 17          5.4         3.9          1.3         0.4  setosa
## 20          5.1         3.8          1.5         0.3  setosa
## 28          5.2         3.5          1.5         0.2  setosa
```

```
# Observamos la estructura interna de Entreno.
str(entreno)
```

```
## 'data.frame':    114 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5 4.4 4.9 5.4 4.8 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.4 2.9 3.1 3.7 3.4 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.5 1.4 1.5 1.5 1.6 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.1 0.2 0.2 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 .
..
```

```
# Observamos la estructura interna de Testeo.
str(testeo)
```

```
## 'data.frame':    36 obs. of  5 variables:
##  $ Sepal.Length: num  5.4 4.6 4.8 5.4 5.1 5.2 5.2 5.5 5 4.5 ...
##  $ Sepal.Width : num  3.9 3.4 3 3.9 3.8 3.5 3.4 4.2 3.2 2.3 ...
##  $ Petal.Length: num  1.7 1.4 1.4 1.3 1.5 1.5 1.4 1.4 1.2 1.3 ...
##  $ Petal.Width : num  0.4 0.3 0.1 0.4 0.3 0.2 0.2 0.2 0.2 0.3 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 .
..
```

Respuestas 4:

- Muestre:

```
summary(iris$Species)

##     setosa versicolor  virginica
##         50         50         50

summary(entreno$Species)

##     setosa versicolor  virginica
##         38         38         38

summary(testeo$Species)

##     setosa versicolor  virginica
##         12         12         12
```

- Verifique que en entreno y testeo haya quedado el porcentaje esperado de elementos según la partición con createDataPartition.

```
# Comprobamos si el % de Train coincide con la partición creada.

porcentaje= (38*3*100)/150
cat("El Porcentaje obtenido es:",porcentaje,"%; Coincide con la partición Hecha con cr
eateDataPartition.")

## El Porcentaje obtenido es: 76 %; Coincide con la partición Hecha con createDataPart
ition.

# Comprobamos si el % de Test coincide con la partición creada.

porcentaje= (12*3*100)/150 #%
cat("El Porcentaje obtenido es:",porcentaje,"%; Coincide con la partición Hecha con cr
eateDataPartition.")

## El Porcentaje obtenido es: 24 %; Coincide con la partición Hecha con createDataPart
ition.
```

Respuestas 5:

- Escriba `entreno - Enter` y luego `testeo - Enter`.

```
entreno

##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1            5.1         3.5          1.4         0.2     setosa
## 2            4.9         3.0          1.4         0.2     setosa
## 3            4.7         3.2          1.3         0.2     setosa
## 4            4.6         3.1          1.5         0.2     setosa
## 5            5.0         3.6          1.4         0.2     setosa
## 8            5.0         3.4          1.5         0.2     setosa
## 9            4.4         2.9          1.4         0.2     setosa
## 10           4.9         3.1          1.5         0.1     setosa
## 11           5.4         3.7          1.5         0.2     setosa
## 12           4.8         3.4          1.6         0.2     setosa
## 14           4.3         3.0          1.1         0.1     setosa
## 15           5.8         4.0          1.2         0.2     setosa
## 16           5.7         4.4          1.5         0.4     setosa
## 18           5.1         3.5          1.4         0.3     setosa
## 19           5.7         3.8          1.7         0.3     setosa
## 21           5.4         3.4          1.7         0.2     setosa
## 22           5.1         3.7          1.5         0.4     setosa
## 23           4.6         3.6          1.0         0.2     setosa
```

```
## 24          5.1         3.3          1.7         0.5     setosa
## 25          4.8         3.4          1.9         0.2     setosa
## 26          5.0         3.0          1.6         0.2     setosa
## 27          5.0         3.4          1.6         0.4     setosa
## 30          4.7         3.2          1.6         0.2     setosa
## 31          4.8         3.1          1.6         0.2     setosa
## 32          5.4         3.4          1.5         0.4     setosa
## 33          5.2         4.1          1.5         0.1     setosa
## 35          4.9         3.1          1.5         0.2     setosa
## 37          5.5         3.5          1.3         0.2     setosa
## 38          4.9         3.6          1.4         0.1     setosa
## 39          4.4         3.0          1.3         0.2     setosa
## 40          5.1         3.4          1.5         0.2     setosa
## 41          5.0         3.5          1.3         0.3     setosa
## 43          4.4         3.2          1.3         0.2     setosa
## 44          5.0         3.5          1.6         0.6     setosa
## 46          4.8         3.0          1.4         0.3     setosa
## 47          5.1         3.8          1.6         0.2     setosa
## 48          4.6         3.2          1.4         0.2     setosa
## 49          5.3         3.7          1.5         0.2     setosa
## 51          7.0         3.2          4.7         1.4 versicolor
## 52          6.4         3.2          4.5         1.5 versicolor
## 54          5.5         2.3          4.0         1.3 versicolor
## 55          6.5         2.8          4.6         1.5 versicolor
## 56          5.7         2.8          4.5         1.3 versicolor
## 57          6.3         3.3          4.7         1.6 versicolor
## 59          6.6         2.9          4.6         1.3 versicolor
## 60          5.2         2.7          3.9         1.4 versicolor
## 61          5.0         2.0          3.5         1.0 versicolor
## 64          6.1         2.9          4.7         1.4 versicolor
## 65          5.6         2.9          3.6         1.3 versicolor
## 66          6.7         3.1          4.4         1.4 versicolor
## 67          5.6         3.0          4.5         1.5 versicolor
## 68          5.8         2.7          4.1         1.0 versicolor
## 69          6.2         2.2          4.5         1.5 versicolor
## 71          5.9         3.2          4.8         1.8 versicolor
## 73          6.3         2.5          4.9         1.5 versicolor
## 74          6.1         2.8          4.7         1.2 versicolor
## 75          6.4         2.9          4.3         1.3 versicolor
## 76          6.6         3.0          4.4         1.4 versicolor
## 77          6.8         2.8          4.8         1.4 versicolor
## 78          6.7         3.0          5.0         1.7 versicolor
## 79          6.0         2.9          4.5         1.5 versicolor
## 80          5.7         2.6          3.5         1.0 versicolor
## 81          5.5         2.4          3.8         1.1 versicolor
## 82          5.5         2.4          3.7         1.0 versicolor
## 83          5.8         2.7          3.9         1.2 versicolor
## 84          6.0         2.7          5.1         1.6 versicolor
## 85          5.4         3.0          4.5         1.5 versicolor
## 86          6.0         3.4          4.5         1.6 versicolor
## 87          6.7         3.1          4.7         1.5 versicolor
## 88          6.3         2.3          4.4         1.3 versicolor
## 92          6.1         3.0          4.6         1.4 versicolor
## 94          5.0         2.3          3.3         1.0 versicolor
## 95          5.6         2.7          4.2         1.3 versicolor
## 96          5.7         3.0          4.2         1.2 versicolor
## 97          5.7         2.9          4.2         1.3 versicolor
## 100         5.7         2.8          4.1         1.3 versicolor
## 101         6.3         3.3          6.0         2.5  virginica
```

```
## 102          5.8         2.7          5.1         1.9  virginica
## 105          6.5         3.0          5.8         2.2  virginica
## 106          7.6         3.0          6.6         2.1  virginica
## 108          7.3         2.9          6.3         1.8  virginica
## 110          7.2         3.6          6.1         2.5  virginica
## 111          6.5         3.2          5.1         2.0  virginica
## 112          6.4         2.7          5.3         1.9  virginica
## 114          5.7         2.5          5.0         2.0  virginica
## 115          5.8         2.8          5.1         2.4  virginica
## 116          6.4         3.2          5.3         2.3  virginica
## 117          6.5         3.0          5.5         1.8  virginica
## 118          7.7         3.8          6.7         2.2  virginica
## 119          7.7         2.6          6.9         2.3  virginica
## 120          6.0         2.2          5.0         1.5  virginica
## 121          6.9         3.2          5.7         2.3  virginica
## 122          5.6         2.8          4.9         2.0  virginica
## 124          6.3         2.7          4.9         1.8  virginica
## 126          7.2         3.2          6.0         1.8  virginica
## 127          6.2         2.8          4.8         1.8  virginica
## 128          6.1         3.0          4.9         1.8  virginica
## 129          6.4         2.8          5.6         2.1  virginica
## 130          7.2         3.0          5.8         1.6  virginica
## 131          7.4         2.8          6.1         1.9  virginica
## 133          6.4         2.8          5.6         2.2  virginica
## 134          6.3         2.8          5.1         1.5  virginica
## 136          7.7         3.0          6.1         2.3  virginica
## 137          6.3         3.4          5.6         2.4  virginica
## 138          6.4         3.1          5.5         1.8  virginica
## 140          6.9         3.1          5.4         2.1  virginica
## 141          6.7         3.1          5.6         2.4  virginica
## 143          5.8         2.7          5.1         1.9  virginica
## 144          6.8         3.2          5.9         2.3  virginica
## 145          6.7         3.3          5.7         2.5  virginica
## 146          6.7         3.0          5.2         2.3  virginica
## 147          6.3         2.5          5.0         1.9  virginica
## 148          6.5         3.0          5.2         2.0  virginica
## 150          5.9         3.0          5.1         1.8  virginica

testeo

##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 6            5.4         3.9          1.7         0.4     setosa
## 7            4.6         3.4          1.4         0.3     setosa
## 13           4.8         3.0          1.4         0.1     setosa
## 17           5.4         3.9          1.3         0.4     setosa
## 20           5.1         3.8          1.5         0.3     setosa
## 28           5.2         3.5          1.5         0.2     setosa
## 29           5.2         3.4          1.4         0.2     setosa
## 34           5.5         4.2          1.4         0.2     setosa
## 36           5.0         3.2          1.2         0.2     setosa
## 42           4.5         2.3          1.3         0.3     setosa
## 45           5.1         3.8          1.9         0.4     setosa
## 50           5.0         3.3          1.4         0.2     setosa
## 53           6.9         3.1          4.9         1.5 versicolor
## 58           4.9         2.4          3.3         1.0 versicolor
## 62           5.9         3.0          4.2         1.5 versicolor
## 63           6.0         2.2          4.0         1.0 versicolor
## 70           5.6         2.5          3.9         1.1 versicolor
## 72           6.1         2.8          4.0         1.3 versicolor
## 89           5.6         3.0          4.1         1.3 versicolor
```

```
## 90           5.5         2.5          4.0         1.3 versicolor
## 91           5.5         2.6          4.4         1.2 versicolor
## 93           5.8         2.6          4.0         1.2 versicolor
## 98           6.2         2.9          4.3         1.3 versicolor
## 99           5.1         2.5          3.0         1.1 versicolor
## 103          7.1         3.0          5.9         2.1  virginica
## 104          6.3         2.9          5.6         1.8  virginica
## 107          4.9         2.5          4.5         1.7  virginica
## 109          6.7         2.5          5.8         1.8  virginica
## 113          6.8         3.0          5.5         2.1  virginica
## 123          7.7         2.8          6.7         2.0  virginica
## 125          6.7         3.3          5.7         2.1  virginica
## 132          7.9         3.8          6.4         2.0  virginica
## 135          6.1         2.6          5.6         1.4  virginica
## 139          6.0         3.0          4.8         1.8  virginica
## 142          6.9         3.1          5.1         2.3  virginica
## 149          6.2         3.4          5.4         2.3  virginica
```

- La flor que obtuvo antes, ¿quedó en entrenamiento o en testeo?

```
# Verificamos si está en Train:
entreno["75",]

##    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 75          6.4         2.9          4.3         1.3 versicolor

# Verificamos si está en Test:
testeo["75",]

##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## NA           NA          NA           NA          NA    <NA>
```

Podemos observar en base a los datos, que la **Flor** obtenida según mis ***2 últimos números de DNI***, quedó en el Dataset de ***Entrenamiento***.

Bibliografia:

- R plot pch symbols
- xyplot-xlab-ylab
- R-Graphics
- R Markdown Cheat Sheet I
- R Markdown Cheat Sheet II
- Flores Iris
- R Chunks
- R cat
- R print
- Ejemplo Iris I
- Ejemplo Iris II

Anexo Código:

Parte A: Análisis Exploratorio de Datos

```
data(iris)
?iris
dim(iris)
str(iris)
summary(iris$Species)
plot(iris$Species, main="Gráfico de Especies Iris",col=c("red", "blue", "green"))
summary(iris)
library(caret)
xyplot(iris$Sepal.Length~iris$Sepal.Width, groups=iris$Species, pch=c(3,4,5),col=c("re
d", "blue", "green"),auto.key=TRUE, main="Largo vs Ancho de Iris Sepal",xlab="Ancho Se
pal", ylab="Largo Sepal")
flor= iris[75,]
flor
```

Parte B: Conjuntos:

```
set.seed(575); particion=createDataPartition(y=iris$Species,p=0.75,list=FALSE)
entreno=iris[particion,]
testeo=iris[-particion,]
dim(entreno)
dim(testeo)
head(entreno)
head(testeo)
str(entreno)
str(testeo)
summary(iris$Species)
summary(entreno$Species)
summary(testeo$Species)
(38*3*100)/150 #%
(12*3*100)/150 #%
entreno["75",] # esta en train
testeo["75",]  # no está en test
```

¡Muchas Gracias!