Daniel Estrada Ocaña A01369854

Undertanding the IRIS data set

The notebook aims to undertand the content of the iris data set.

IRIS data set

1. IRIS classes -PGN file
2. IRIS measurements 1 -PGN file
3. IRIS measurements 2 -JPEG file
4. IRIS parts -PGN file
5. Iris.csv The dataset description
    - The dataset used here is an extension from the original IRIS dataset
    - Total numer of observations: 150
    - Many observations/measurements/recordings of the characteristics/attributes/variables of iris plant
    - Variables: measurements petal 1, measurements petal 2, measurements sepal 1, measurements sepal 2, Type of iris plant

Importing and inspecting the data

```
# Define where you are running the code: colab or local
RunInColab          = True     # (False: no  | True: yes)

# If running in colab:
if RunInColab:
    # Mount your google drive in google colab
    from google.colab import drive
    drive.mount('/content/drive')


    # Find location
    #!pwd
    #!ls
    #!ls "/content/drive/My Drive/Colab Notebooks/MachineLearningWithPython/"

    # Define path del proyecto
    Ruta            = "/content/drive/MyDrive/Sistemas/4to_semestre/semanaTec/TC1002S"

else:
    # Define path del proyecto
    Ruta            = ""
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
# Import the packages that we will be using
import matplotlib.pyplot as plt
import pandas as pd

# Dataset url
url = Ruta + "/NotebooksProfessor/datasets/iris/iris.csv"

# Load the dataset
dataset = pd.read_csv(url )

```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
# Print the dataset
dataset
```

	Ms 1	Ms 2	Ms 3	Ms 4	Type
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 5 columns

```
# Print the number of rows
Nrows = dataset.shape[0]
Nrows
```

```
150
```

```
# Print the number of columns
Ncols = dataset.shape[1]
Ncols
```

```
5
```

Data types

[ ] *1 cell hidden*

Activity: work with the iris dataset

1. Load the iris.csv file in your computer and understand the dataset

```
dataset
```

	Ms 1	Ms 2	Ms 3	Ms 4	Type
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 5 columns

2. How many observations (rows) are in total?

- **There area 150 rows.**

```
1 Nrows
```

```
150
```

3. How many variables (columns) are in total? What do they represent?

- **There are 5 columns, this ones represent the measurements of the plants leafs, including the sepal and petal. Also the last column is for the type of iris plant.**

```
1 Ncols
```

```
5
```

4. How many observations are for each type of flower?

- **There are 50 observations of each type (setosa, versicolor and virginica)**

```
1 dataset["Type"].value_counts()
```

```
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
Name: Type, dtype: int64
```

5. What is the type of data for each variable?

- **The first 4 columns contain float variables, the last one string variables.**

```
1 dataset.dtypes
```

```
Ms 1    float64
Ms 2    float64
Ms 3    float64
Ms 4    float64
Type     object
dtype: object
```

6. What are the units of each variable?

- **Sepal Length: measured in centimeters (cm)**
- **Sepal Width: measured in cm**
- **Petal Length: measured in cm**
- **Petal Width: measured in cm**
- **Species: This column represents the species of the iris flower, and it doesn't have units since it's a categorical variable.**