Sesión 1 - Diplomado Data Science Duoc UC

Módulo: Machine Learning

A continuación, aprenderás a cargar un dataset con el uso de las librerías de Python desde fuentes de datos que se encuentran en la web.

Recuerda almacenar una copia de este Jupyter Notebook en tu Google Drive para poder ejecutar los bloques de código.

Comencemos con un clásico dentro del mundo de data science. Trabajaremos con **Iris**. Un dataset muy conocido en la academia, que sirve bastante para comenzar dentro del mundo del Data Science. Antes de comenzar, demos contexto a todo esto. Iris es un dataset que nos permite clasificar flores a partir de características que poseen las plantas.

Las características o features son las siguientes:

*sepal lenght o largo del sépalo en cm,

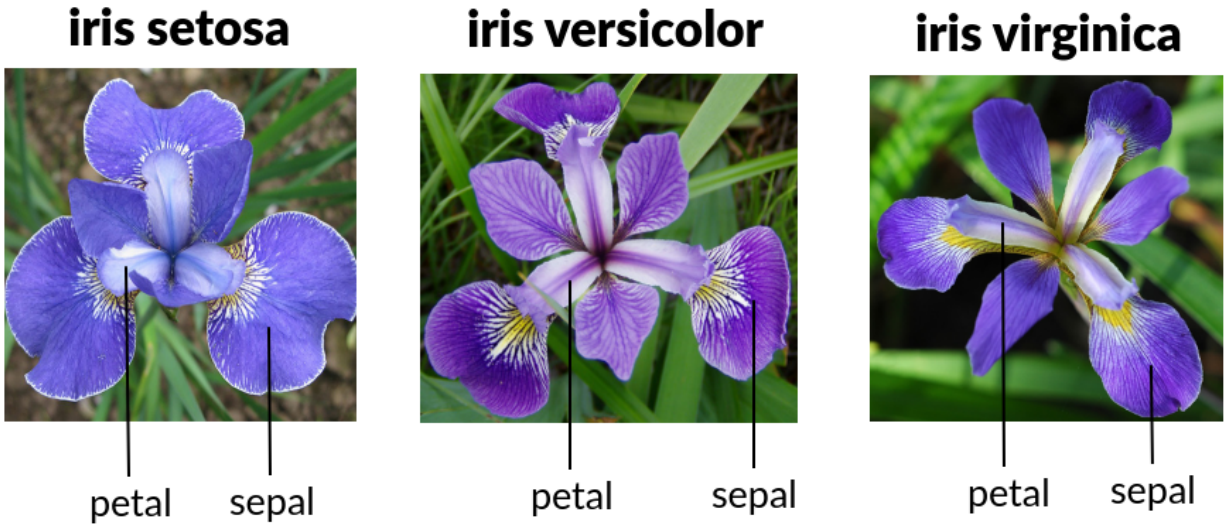
*sepal width o ancho del sépalo en cm,

*pelal lenth o largo del petalo in cm,

*pedal width o ancho del pétalo en cm.

Las diferentes especies a las que puede pertenecer una planta son:

- Setosa
- Virginica
- Versicolor




```
#Cargamos las librerías de Python que nos servirán para todo esto
import numpy as np
import pandas as pd


#Cargamos el dataset de Iris para comenzar a trabajarlo
#Puedes cargarlo desde la librería Scikit-learn
from sklearn.datasets import load_iris

data = load_iris()
df = pd.DataFrame(data.data, columns=data.feature_names)
df['target'] = data['target']


#Si el recurso lo tienes de forma local y trabajas desde Anaconda entonces añade la ruta del

########### data = pd.read_csv("your downloaded dataset location ")



#Si el recurso lo tienes en drive, puedes montar tu google Drive y hacer referencia a la rut

#from google.colab import drive
#drive.flush_and_unmount()
#drive.mount('/content/drive',force_remount=True)

#path = "/content/drive/My Drive/"  #Esta es la ruta
#!ls /content/drive/My\ Drive/
```

Utilizamos nuestras primeras instrucciones de python

```
#Utilizamos head
df.head()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

Añadir qué observaciones hay respecto del comando utilizado

```
#Utilizamos tail
df.tail()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
145	6.7	3.0	5.2	2.3	2
146	6.3	2.5	5.0	1.9	2
147	6.5	3.0	5.2	2.0	2
148	6.2	3.4	5.4	2.3	2
149	5.9	3.0	5.1	1.8	2

Añadir qué observaciones hay respecto del comando utilizado

```
#Utilizamos shape
df.shape
```

```
(150, 5)
```

Añadir qué observaciones hay respecto del comando utilizado

```
df.sample(10)
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
131	7.9	3.8	6.4	2.0	2
55	5.7	2.8	4.5	1.3	1
15	5.7	4.4	1.5	0.4	0
36	5.5	3.5	1.3	0.2	0
127	6.1	3.0	4.9	1.8	2

Añadir qué observaciones hay respecto del comando utilizado

```
#Utilizamos value_counts()
df.value_counts()
```

```
sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)  target
5.8                2.7               5.1                1.9               2         2
7.9                3.8               6.4                2.0               2         1
5.4                3.0               4.5                1.5               1         1
5.5                2.4               3.7                1.0               1         1
                   2.3               4.0                1.3               1         1
                                                                                     ..
6.3                2.5               4.9                1.5               1         1
                   2.3               4.4                1.3               1         1
6.2                3.4               5.4                2.3               2         1
                   2.9               4.3                1.3               1         1
4.3                3.0               1.1                0.1               0         1
Length: 149, dtype: int64
```

Añadir qué observaciones hay respecto del comando utilizado

```
#Utilizamos describe()
df.describe()
```

Añadir qué observaciones hay respecto del comando utilizado

```
count  150.000000  150.000000  150.000000  150.000000  150.000000
```

```
#Utilizamos info()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   sepal length (cm)  150 non-null    float64
 1   sepal width (cm)   150 non-null    float64
 2   petal length (cm)  150 non-null    float64
 3   petal width (cm)   150 non-null    float64
 4   target             150 non-null    int64  
dtypes: float64(4), int64(1)
memory usage: 6.0 KB
```

Responda las siguientes preguntas respecto al dataset y al modelo que podríamos aplicar

¿Cuál es la cantidad de registros que tiene el dataset?

Has doble clic e ingresa tu respuesta aquí.

¿Cuál es el promedio de los anchos y los largos de pétalo que hay en la muestra?

Has doble clic e ingresa tu respuesta aquí.

¿Hay registros nulls dentro de la data? ¿Cómo se dió cuenta?

Has doble clic e ingresa tu respuesta aquí.

En Machine Learning se habla mucho de aprendizaje supervisado y no supervisado. Uno tiene etiqueta y el otro no tiene etiqueta respectivamente.

¿Qué tipo de aprendizaje correspondería utilizar?¿Por qué?

Has doble clic e ingresa tu respuesta aquí.

¿Cómo están distribuídos los datos dentro de la muestra respecto de las especies?. En caso de tener más ejemplos de un tipo de flor, y menos de otro tipo ¿crees que influiría en el algoritmo final? ¿por qué?

Has doble clic e ingresa tu respuesta aquí.

¿Qué procesamiento podríamos hacer como rutina de limpieza?. Propone 3 acciones que se podrían efectuar a este dataset.

Has doble clic e ingresa tu respuesta aquí.

Una vez que hayamos ejecutado la rutina de limpieza que propones, ¿qué crees que se podría hacer con el dataset?

Has doble clic e ingresa tu respuesta aquí.