Actividad

- **Nombre:** Daira Adriana Chavarría Rodríguez
- **Matrícula:** A01274745

Entregar

Archivo PDF de la actividad y la liga de la actividad en su repostitorio.

Nota:

Todas las tareas entregadas fuera de la fecha limite se califican sobre 50 de los 100 puntos posibles.

In [1]:

```
# Si trabajamos en Google Colaboratory corremos las siguientes lineas de código
from google.colab import drive
drive.mount('/gdrive')
```

```
Mounted at /gdrive
```

In [2]:

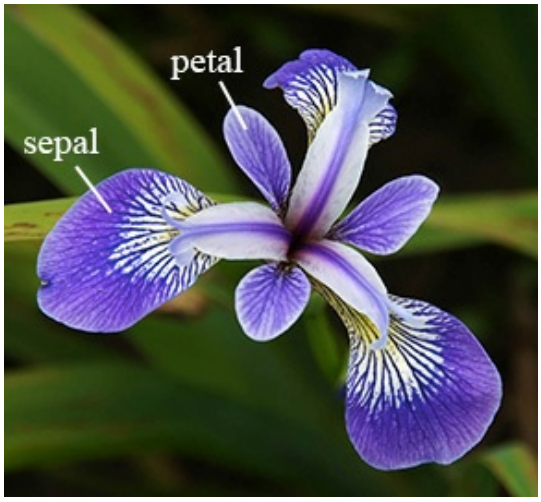
```
# Nos cambiamos a la carpeta donde tengamos el repositorio. Si es otra carpeta asegúrate de cambiar la ruta.
%cd /gdrive/MyDrive/semanaTec/arte-analitica
```

```
/gdrive/MyDrive/semanaTec/arte-analitica
```

Iris dataset

Este famoso conjunto de datos corresopnde a un análisis multivariado que el biólogo Ronald Fisher utilizó en su artículo de 1936 *The use of multiple measurements in taxonomic problems* para ejemplificar un análisis discriminante lineal.

Los datos consisten en 4 mediciones para tres especies de flor Iris (Iris setosa, Iris virginica and Iris versicolor). Las 4 mediciones corresponden a la longitud y altura de los sépalos y de los pétalos de cada especimen.



Actividad

Carga el dataset `iris.csv` de la carpeta de datos y haz un análisis estadístico de las variables.

In [7]:

```
# Carga los datos y muestra los primeros renglones
import numpy as np
import pandas as pd

# Importar los datos
iris_df = pd.read_csv('data/iris.csv')
```

In [12]:

```
# Crea una tabla resumen con los estadísticos generales de las variables numéricas
iris_df.describe()
```

Out[12]:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

In [11]:

```
# ¿Cuántas observaciones hay de cada tipo de flor (cuántas setosas hay, cuántas virginica, etc.)?
iris_df['Species'].value_counts()
```

Out[11]:

```
versicolor    50
virginica     50
setosa        50
Name: Species, dtype: int64
```

In [13]:

```
iris_df
```

Out[13]:

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

150 rows × 5 columns

In [20]:

```
# ¿Cómo se comparan los pétalos y sépalos para las distintas especies de flor?
# Calcula los estadísticos de tendencia central y dispersión para cada especie.
iris_df.groupby('Species').mean()
```

Out[20]:

Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
virginica	6.588	2.974	5.552	2.026

In [19]:

```
iris_df.groupby('Species').std()
```

Out[19]:

Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	0.352490	0.379064	0.173664	0.105386
versicolor	0.516171	0.313798	0.469911	0.197753
virginica	0.635880	0.322497	0.551895	0.274650

In [22]:

```
# ¿Qué relación hay entre las 4 variables? Obtén la matriz de correlación y escribe tus c
onclusiones.
iris_df.corr()
```

Out[22]:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.000000	-0.117570	0.871754	0.817941
Sepal.Width	-0.117570	1.000000	-0.428440	-0.366126
Petal.Length	0.871754	-0.428440	1.000000	0.962865
Petal.Width	0.817941	-0.366126	0.962865	1.000000

Sabemos que una correlación de 1 significa que ambas variables tienen una relación lineal perfecta, por tanto se podría afirmar que:

- **El ancho de cada pétalo está estrechamente relacionado con el largo del sépalo y del pétalo, pero no con el ancho del sépalo.**
- **El largo de cada pétalo está está estrechamente relacionado con el largo y ancho del pétalo, pero no con el ancho del sépalo.**