



**INFORME**

**ANALISIS EXPLORATORIO CON IRIS**

**NICKY ALEXANDER FLOREZ BUSTAMANTE**

**JUAN ANDRES MENENDEZ VILLARRAGA**

**DAVID FERNANDO GOMEZ ARISTIZABAL**

**LUIS FERNANDO SANCHEZ**

**2828523**

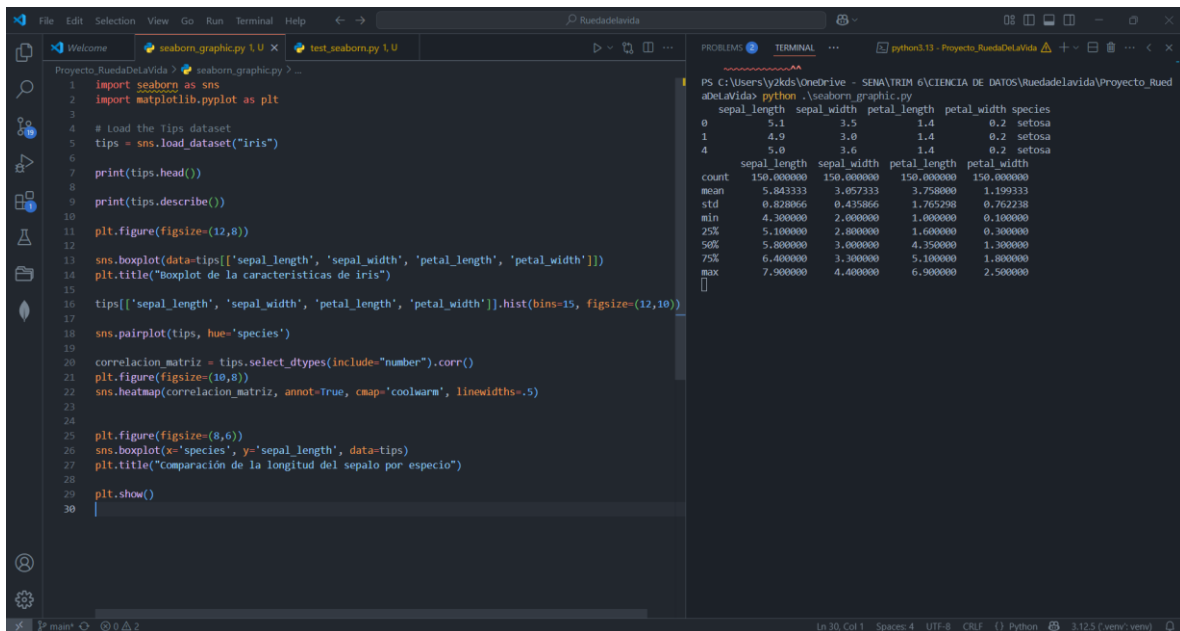
**CTMA – SERVICIO NACIONAL DE APRENDIZAJE**

**2025**

# INFORME DE ANÁLISIS EXPLORATORIO CON EL DATASET IRIS

Inicialmente importamos el dataset iris, el cual contiene 150 muestras de flores y cada flor tiene 4 características numéricas: `sepal_length` (largo del sépalo), `sepal_width` (ancho del sépalo), `petal_length` (largo del pétalo) y `petal_width` (ancho del pétalo). Además, cada muestra está clasificada en una de tres especies: `setosa`, `versicolor` y `virginica`.

Este fue el código final de nuestro análisis



```
1 import seaborn as sns
2 import matplotlib.pyplot as plt
3
4 # Load the tips dataset
5 tips = sns.load_dataset("iris")
6
7 print(tips.head())
8
9 print(tips.describe())
10
11 plt.figure(figsize=(12,8))
12
13 sns.boxplot(data=tips[['sepal_length', 'sepal_width', 'petal_length', 'petal_width']])
14 plt.title("boxplot de la características de iris")
15
16 tips[['sepal_length', 'sepal_width', 'petal_length', 'petal_width']].hist(bins=15, figsize=(12,10))
17
18 sns.pairplot(tips, hue='species')
19
20 correlacion_matriz = tips.select_dtypes(include="number").corr()
21 plt.figure(figsize=(10,8))
22 sns.heatmap(correlacion_matriz, annot=True, cmap='coolwarm', linewidths=.5)
23
24
25 plt.figure(figsize=(8,6))
26 sns.boxplot(x='species', y='sepal_length', data=tips)
27 plt.title("comparación de la longitud del sepal por especie")
28
29 plt.show()
30
```

	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	2.800000	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

## 1. Histograma

Realizamos un histograma para la variable `petal_length`.

En este gráfico se observó que la especie `setosa` tiene pétalos significativamente más cortos (entre 1.0 y 2.0 cm), `virginica` tiene los más largos (entre 4.5 y 6.9 cm), y `versicolor` se encuentra en un punto intermedio (entre 3.0 y 5.0 cm), a lo que podemos concluir `petal_length` es una variable muy útil para diferenciar especies.

## 2. Pairplot

En este apartado realizamos un grafico pairplot permitió visualizar las relaciones entre todas las variables numéricas, diferenciando las especies por color. Pudimos ver que las combinaciones `petal_length` vs `petal_width` y `sepal_length` vs `petal_length` separan claramente las especies

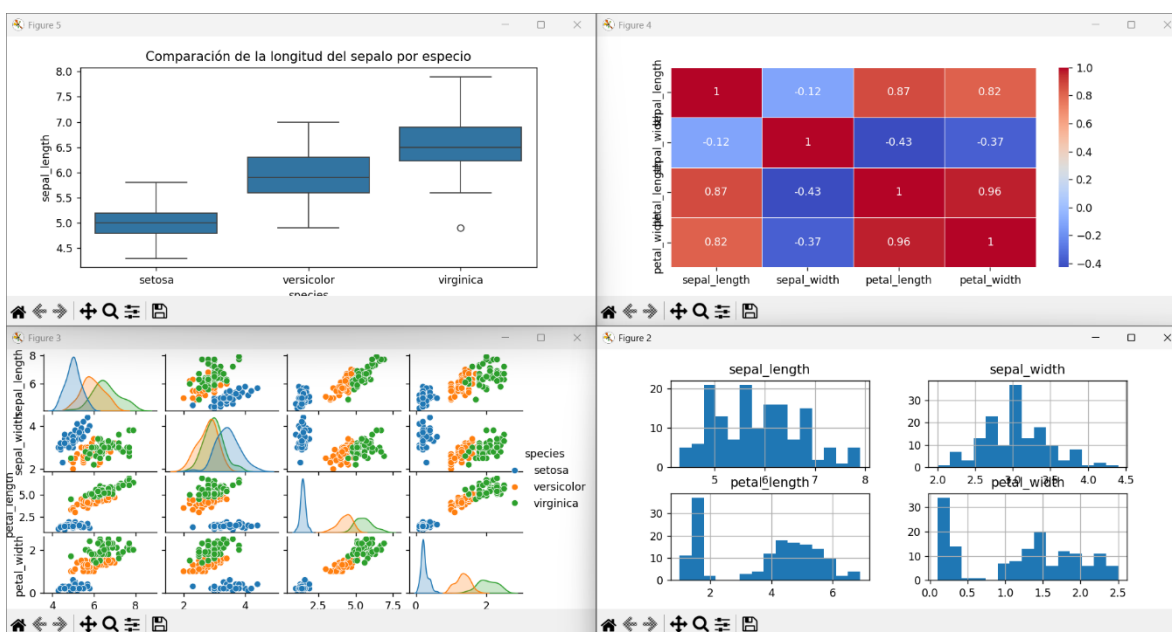
### 3. Scatter plot

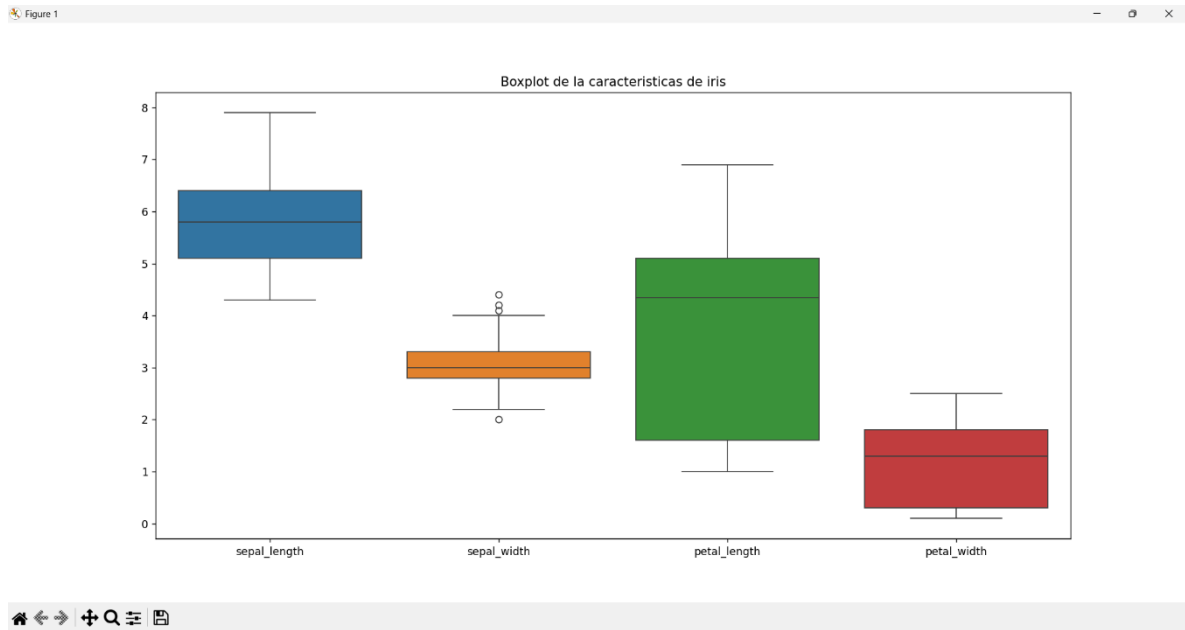
Este gráfico nos muestra cómo el largo y ancho del pétalo están fuertemente relacionados. Los puntos se agrupan por especie, formando tres grupos bien definidos. Esta visualización por sí sola permite identificar la especie con alta precisión.

### 4. Matriz de correlación

Se calculó la matriz de correlación entre las variables numéricas, y, mediante la visualización podemos observar que las variables del pétalo son las más correlacionadas entre sí y las más útiles para clasificación.

A continuación están las capturas de las graficas resultantes:





### Conclusión general

El dataset iris es limpio, balanceado y se puede separar fácilmente por especie.

Este dataset es ideal para practicar análisis exploratorio, visualización de datos y algoritmos de clasificación, debido a la claridad de sus patrones y la simplicidad de su estructura.