# ADSO-GA04-Actividad de Aprendizaje Seis Análisis de Datos

**Juan Manuel Torres**

**Cesar Andrés Acosta**

**Luis Fernando Sánchez**

**Ciencia de Datos**

**2828523**

**SENA**

**Medellín, Antioquia**

**4 de Abril de 2025**

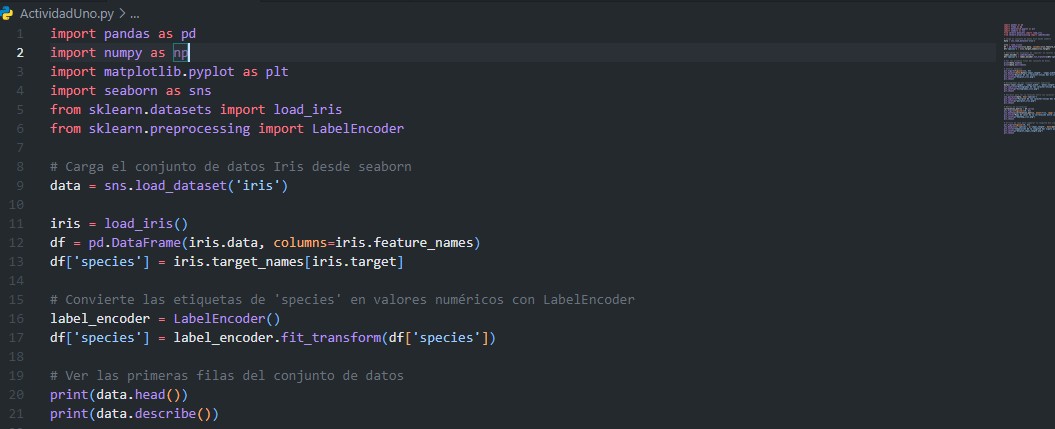
**Análisis Exploratorio de Datos (EDA) - Conjunto de Datos Iris**

# 1. Introducción

Este documento presenta un Análisis Exploratorio de Datos (EDA) utilizando el conjunto de datos Iris de la librería Seaborn. El objetivo es identificar patrones, valores atípicos y realizar visualizaciones básicas. Se usan las bibliotecas Pandas, Numpy, Matplotlib y Seaborn para el análisis y generación de gráficos.

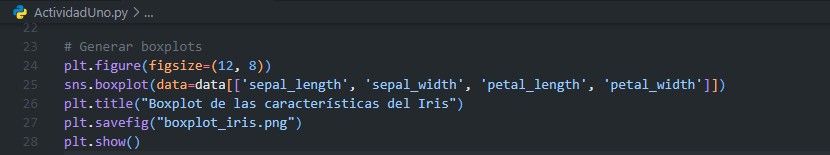
# 2. Carga y descripción de los datos

Código para cargar y describir el conjunto de datos.



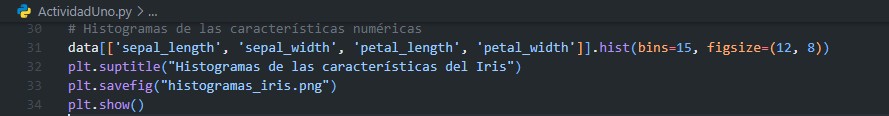
# 3. Identificación de valores atípicos

Los valores atípicos pueden afectar el análisis de datos. Para identificarlos, utilizamos diagramas de caja, mejor conocidos como boxplots.



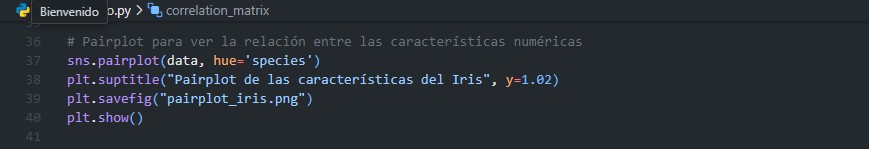
# 4. Distribución de los datos

Se analizan las distribuciones de las variables mediante histogramas.



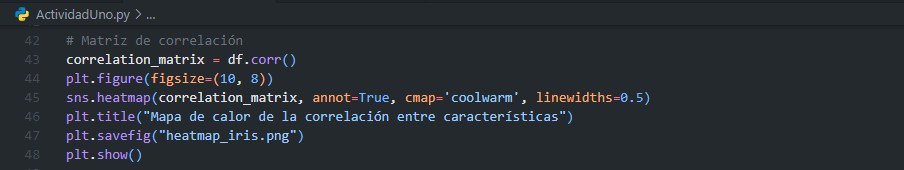
# 5. Relación entre características

Se analizan relaciones entre características con gráficos de dispersión.



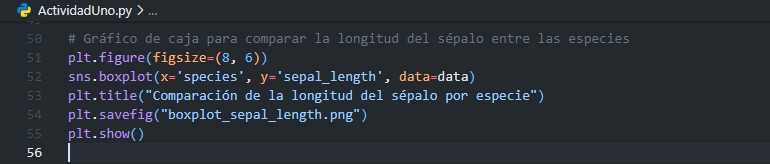
# 6. Correlación entre características

Se genera un mapa de calor para visualizar la correlación entre variables.



# 7. Comparación entre especies

Se comparan las longitudes del sépalo entre especies con un gráfico de caja.



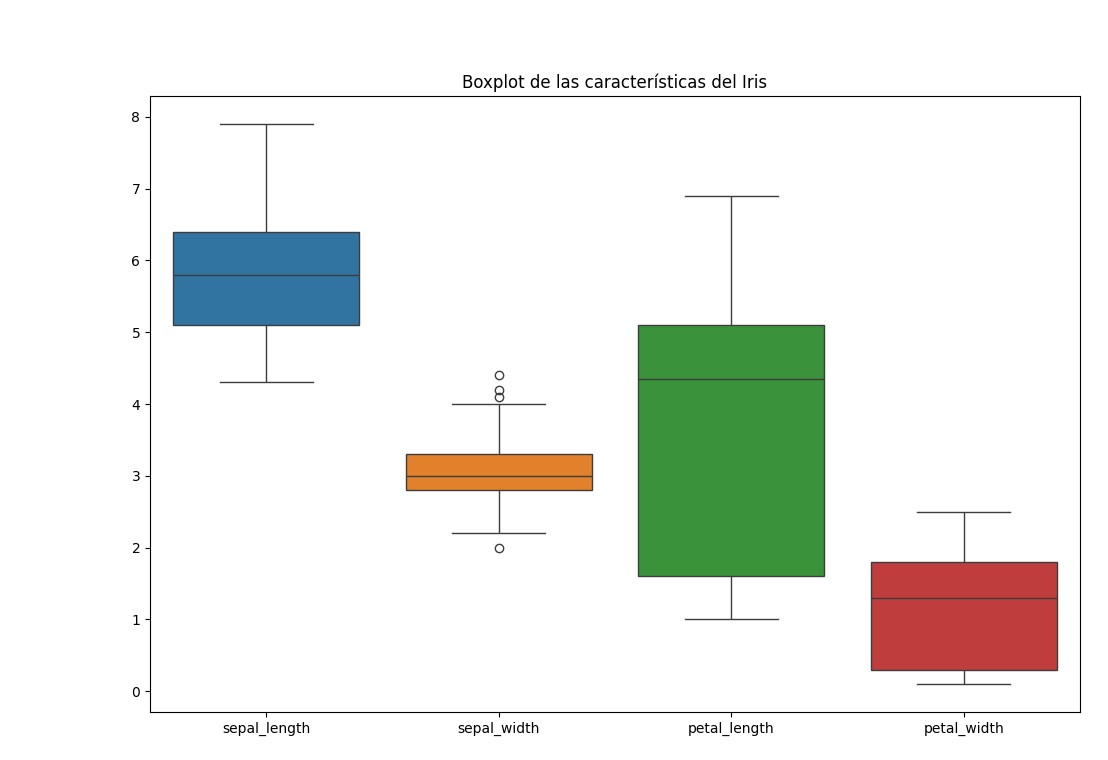
# 8. Conclusiones

El análisis del conjunto de datos "Iris" reveló diferencias significativas entre las especies en características como la longitud y el ancho del sépalo y pétalo, con Iris Setosa mostrando dimensiones más pequeñas. La visualización de datos a través de histogramas y scatter plots facilitó la identificación de patrones y correlaciones, anotando una alta correlación entre longitud del pétalo y longitud del sépalo. Se identificaron valores atípicos que requieren atención antes de implementar modelos predictivos.

# 9. Evidencias y Entrega

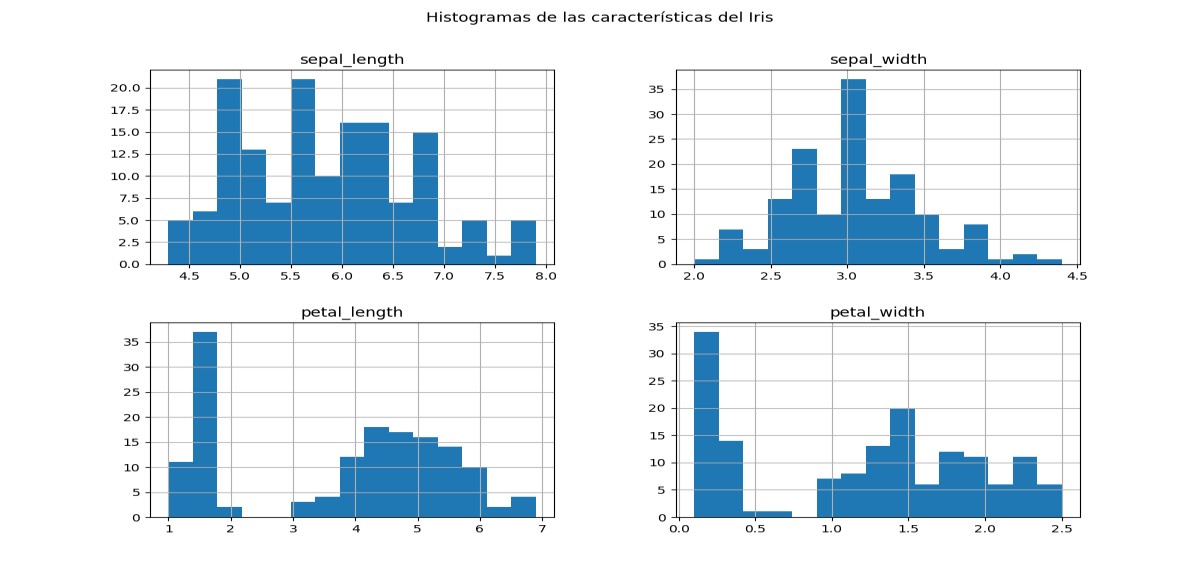
* Boxplot de las características del Iris

Este gráfico muestra la dispersión y posibles valores atípicos de las variables del conjunto de datos.



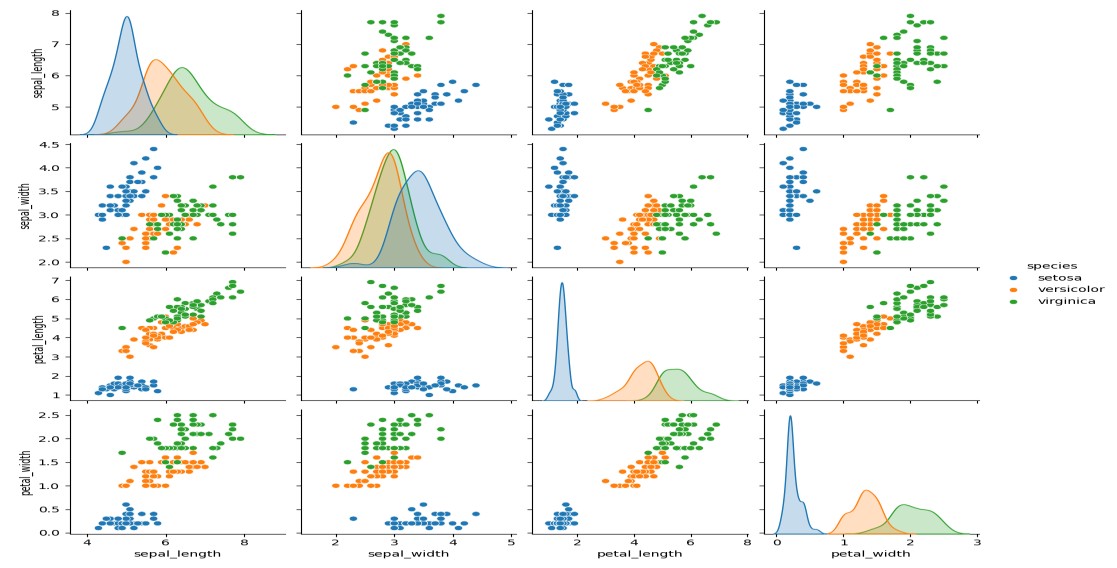
* Histogramas de las características del Iris

Representación de la distribución de las variables numéricas mediante histogramas.



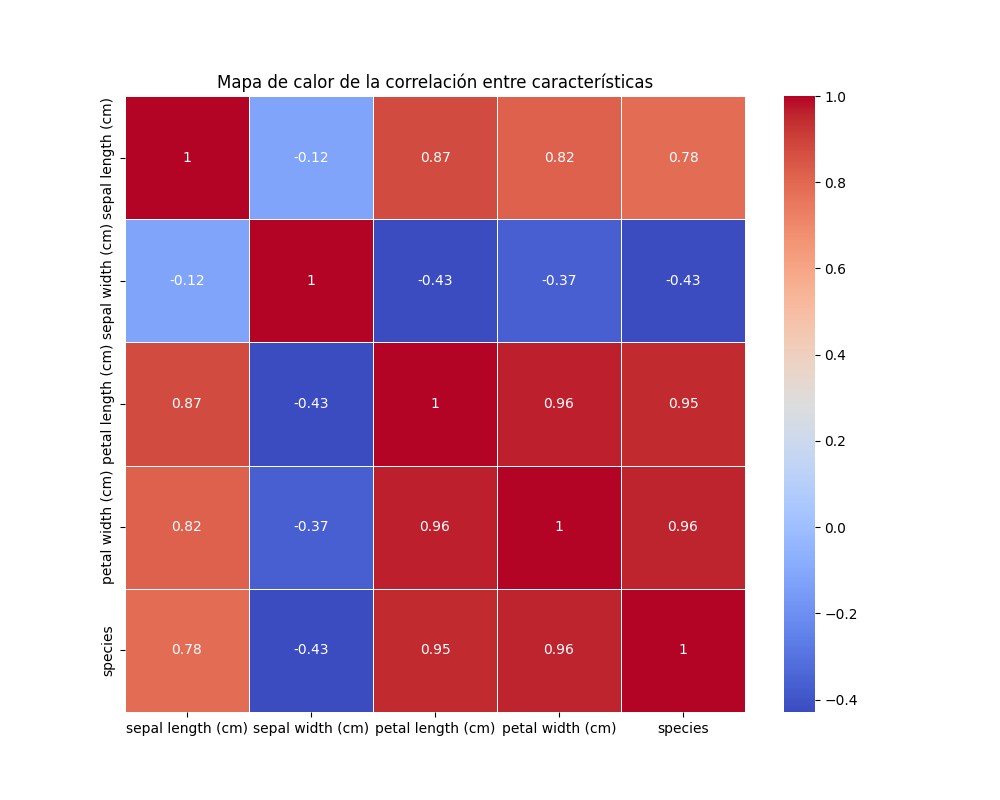
* Pairplot de las características del Iris

Visualización de las relaciones entre las características del conjunto de datos, agrupadas por especie.



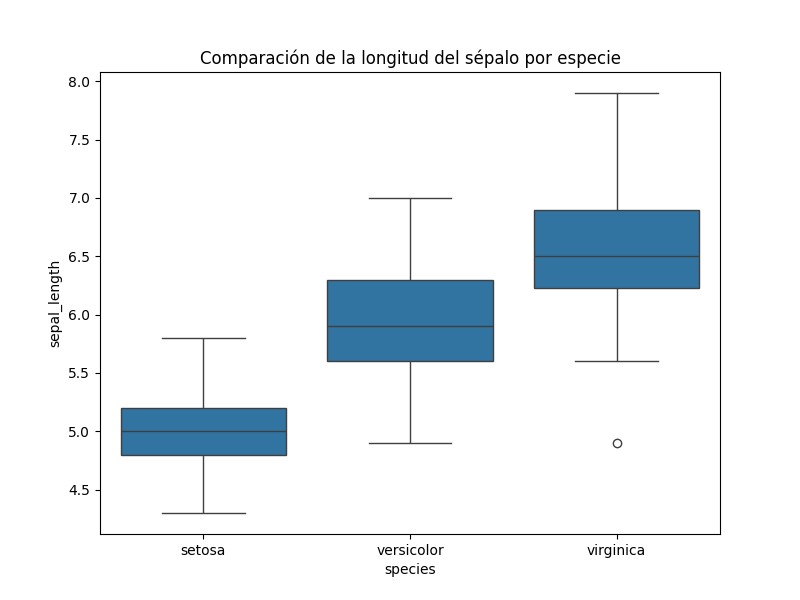
* Mapa de calor de la correlación entre características

Matriz de correlación entre las variables del conjunto de datos para analizar relaciones.



* Comparación de la longitud del sépalo por especie

Boxplot comparando la longitud del sépalo entre las tres especies de Iris.



**Actividad 2: Implementación de Modelos de Machine Learning**

# 1. Introducción

En esta actividad, se implementan diversos modelos de Machine Learning para la clasificación y regresión, utilizando el conjunto de datos Iris. Se exploraron modelos como la Regresión Logística, Árboles de Decisión, SVM, KNN, Redes Neuronales, Random Forest y Regresión Lineal.

# 2. Objetivos

* Implementar diferentes modelos de Machine Learning.
* Evaluar su desempeño en la clasificación de especies de flores.
* Realizar una regresión para predecir una variable continua del conjunto de datos.

# 3. Desarrollo

## 3.1. Carga de Datos y Preprocesamiento

Se utilizó la librería seaborn para cargar el conjunto de datos Iris. Posteriormente, se seleccionaron las variables independientes (sepal\_length, sepal\_width, petal\_width) y la variable objetivo (petal\_length). Además, se dividió el conjunto de datos en entrenamiento y prueba usando train\_test\_split.

## 3.2. Implementación del Modelo de Regresión Lineal

Se utilizó LinearRegression de scikit-learn para entrenar un modelo predictivo de petal\_length. Por otra parte se calcularon los coeficientes y la intersección del modelo, realizando predicciones sobre el conjunto de prueba.

Se evaluó el desempeño mediante el error cuadrático medio (MSE) y el coeficiente de determinación (R²) además de generar una gráfica comparativa entre los valores reales y las predicciones.

## 3.3. Evaluación de Resultados

Los valores obtenidos de MSE y R² indicaron el nivel de ajuste del modelo.

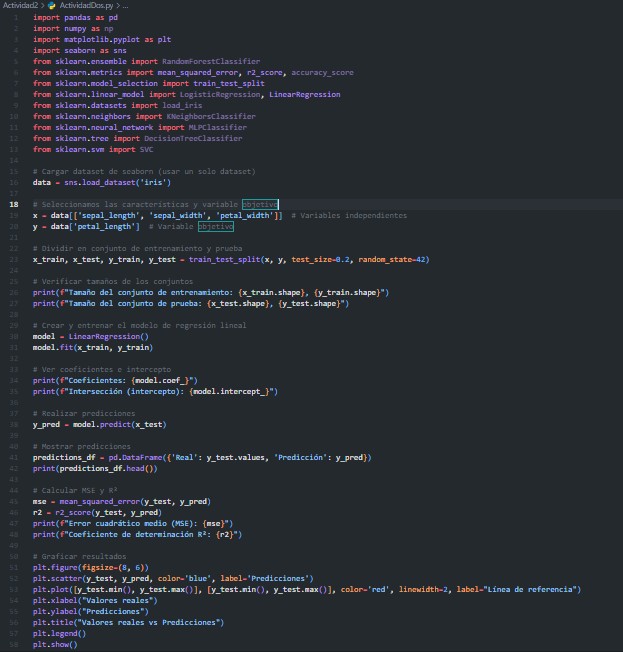
Además, la gráfica de dispersión mostró la relación entre valores reales y predichos.

# 4. Conclusiones

* Se implementó exitosamente un modelo de regresión lineal para predecir petal\_length.
* Se observó la importancia de seleccionar adecuadamente las variables predictoras.
* Se identificó la precisión del modelo y su aplicabilidad en problemas de predicción continua.

# 5. Anexos

**Código utilizado para la implementación.**



# Gráfica generada

