

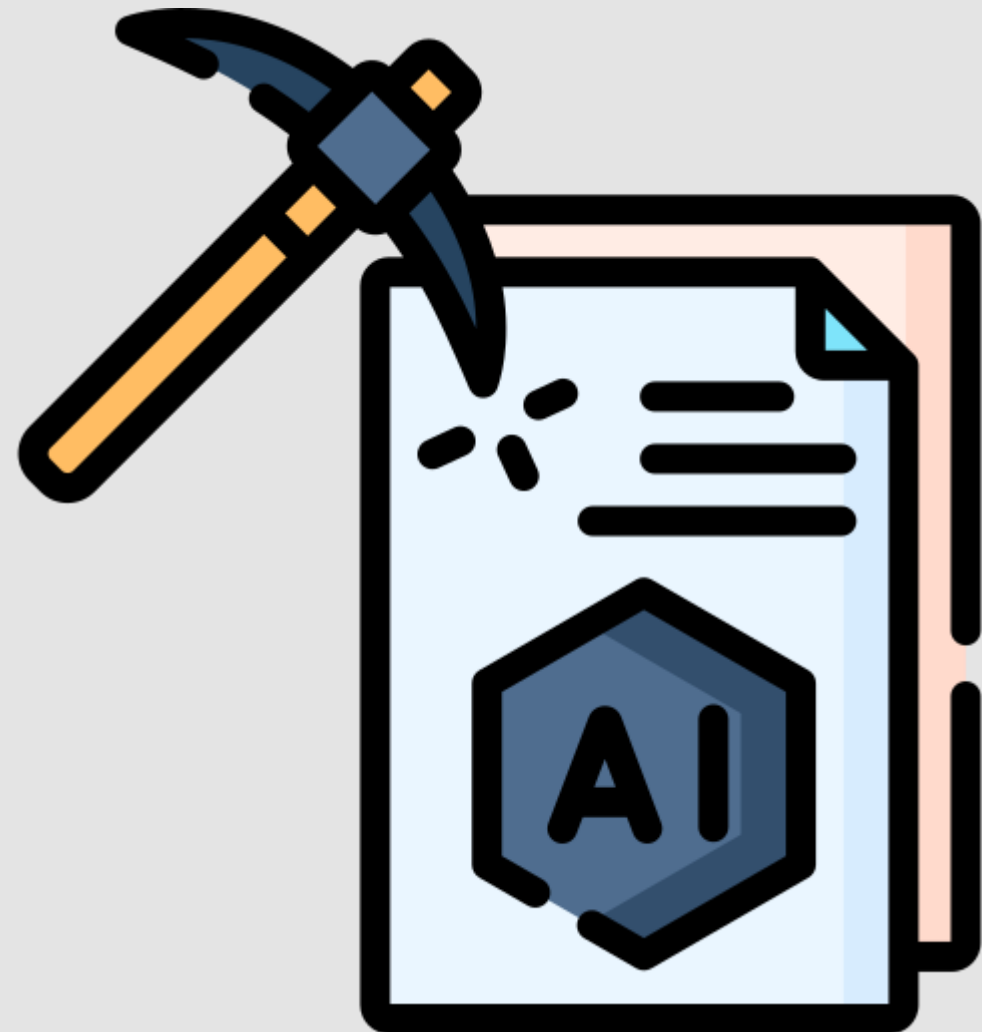
# PROYECTO FINAL

## IRIS DATASET

JUAN PABLO BORRERO  
CARLOS LEAL MEDINA



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA  
"Diseño y prestación de servicios de docencia, investigación  
y extensión de programas de pregrado, aplicando todos los  
requisitos de las normas ISO implementadas en sus sedes  
Neiva y Pitalito"



# ¿Cómo se pueden predecir las clases de las flores Iris (Iris-setosa, Iris-versicolor, Iris-virginica) a partir de las características medidas de los sépalos y pétalos?

**número de registros:** 150

**Número de variables (columnas):** 5

**tipos de datos:**

- sepal\_length: Longitud del sépalo (tipo de dato: float64)
- sepal\_width: Ancho del sépalo (tipo de dato: float64)
- petal\_length: Longitud del pétalo (tipo de dato: float64)
- petal\_width: Ancho del pétalo (tipo de dato: float64)
- class: Clase de la flor (tipo de dato: object)

**Estadísticos descriptivos Valores no nulos:**

Sepal Length: 99 no nulos

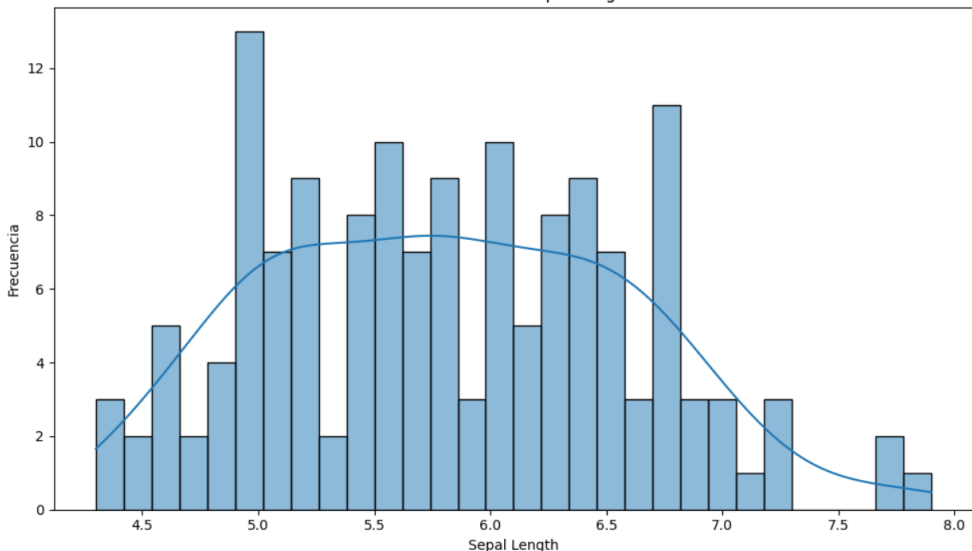
Sepal Width: 99 no nulos

Petal Length: 108 no nulos

Petal Width: 121 no nulos

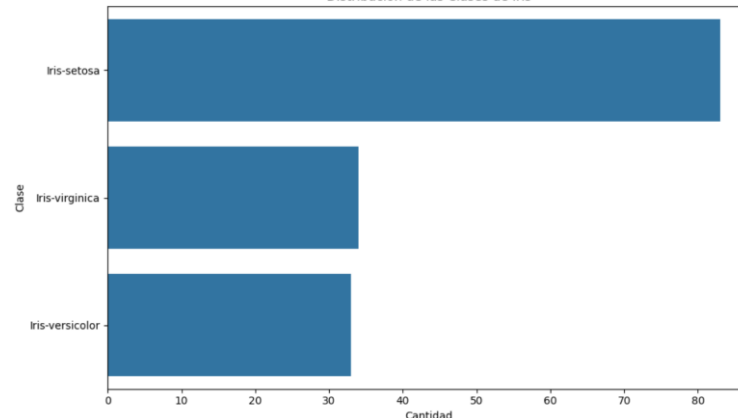
Class: 107 no nulos

Distribución de Sepal Length

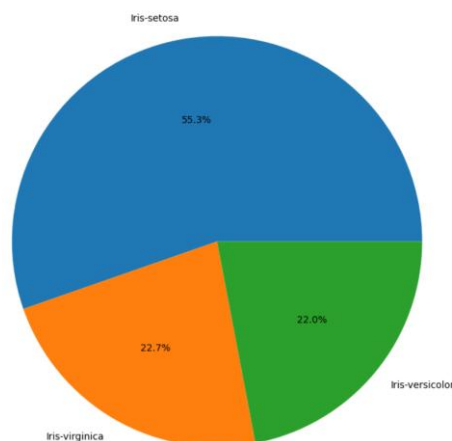


La distribución de la longitud del sépalo muestra una ligera asimetría hacia la derecha, con una mayor concentración de datos entre 5.0 y 6.5. La frecuencia más alta se encuentra alrededor de 6.0, y la curva de densidad indica que los valores más bajos a medios son los más comunes, mientras que los valores altos tienen una frecuencia menor. Aunque existen algunos valores atípicos en los extremos, la mayoría de las observaciones se agrupan en el rango medio, lo que sugiere una tendencia natural en la longitud del sépalo de las flores del dataset.

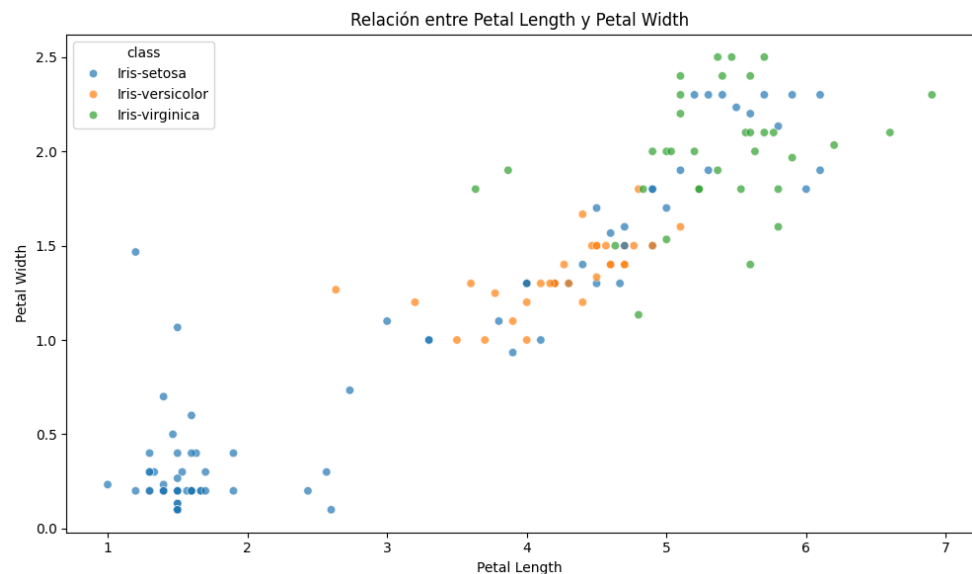
Distribución de las Clases de Iris



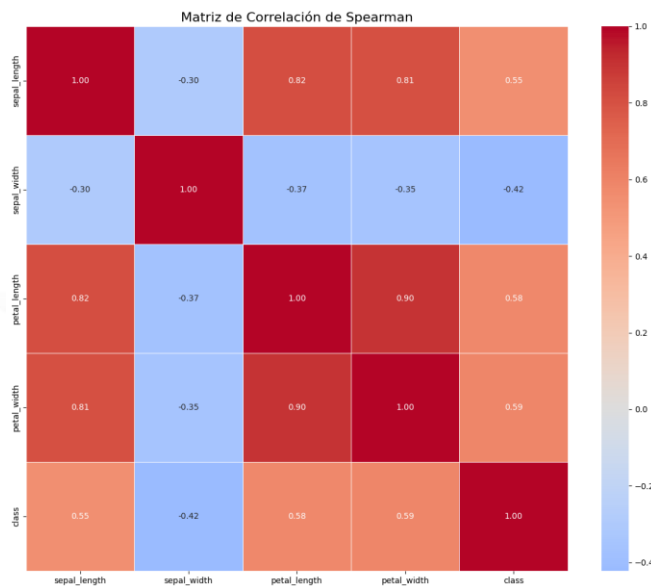
La gráfica muestra que la clase Iris-setosa es la más representada, con más de 80 registros, seguida por Iris-virginica y Iris-versicolor, que tiene la menor cantidad de muestras. Esta distribución desbalanceada podría afectar los modelos predictivos, ya que los algoritmos podrían estar sesgados hacia la clase más numerosa, por lo que sería importante considerar técnicas como el muestreo balanceado o el ajuste de pesos para mitigar este efecto.



La gráfica muestra que Iris-setosa representa el 55.3% de las observaciones, siendo la clase más predominante, mientras que Iris-virginica y Iris-versicolor tienen una representación similar, con 22.7% y 22.0% respectivamente. Este desbalance en la distribución podría influir en los modelos predictivos, favoreciendo a la clase más frecuente, Iris-setosa.



La gráfica muestra una relación positiva clara entre Petal Length y Petal Width, donde a medida que la longitud del pétalo aumenta, también lo hace su ancho. Iris-setosa (puntos azules) tiene valores más bajos en ambas dimensiones, mientras que Iris-versicolor (naranjas) y Iris-virginica (verdes) presentan valores más altos, especialmente Iris-virginica, que tiene los valores más grandes en ambas características. Esto indica que estas dos variables son clave para diferenciar las clases de flores Iris.



La matriz de correlación muestra que las variables relacionadas con los pétalos tienen fuertes correlaciones entre sí, especialmente entre petal\_length y petal\_width (0.90). Además, sepal\_length y sepal\_width están positivamente correlacionados con el tamaño de los pétalos, aunque sepal\_width tiene correlaciones negativas con las otras variables de tamaño. La class muestra una relación moderada con las variables de tamaño de los pétalos y el sépalo, sugiriendo que la clase de la flor depende en parte de estas características, pero no de manera tan fuerte.



### Patrones Identificados:

Existe una correlación positiva clara entre la longitud y el ancho del pétalo, con una fuerte relación entre Petal Length y Petal Width (0.90).

Iris-setosa muestra tamaños más pequeños de sépalos y pétalos en comparación con Iris-versicolor y Iris-virginica, los cuales tienen tamaños mayores, especialmente Iris-virginica.

### Correlaciones Importantes:

Petal Length y Petal Width están altamente correlacionados (0.90), lo que sugiere que el tamaño de los pétalos se mide de forma consistente en ambas dimensiones.

Sepal Length tiene una fuerte correlación con Petal Length (0.82) y Petal Width (0.81), indicando que las características del sépalo también influyen en las del pétalo.

### Outliers:

Se observan outliers en el sepal\_length y petal\_length en los registros de Iris-setosa, especialmente en valores cercanos a 7, que son menos comunes.

Algunos registros de Iris-virginica muestran valores muy altos para petal\_width, con algunos puntos fuera de la tendencia general.

### Distribuciones Relevantes:

La distribución de Sepal Length muestra una ligera asimetría hacia la derecha, mientras que Petal Length y Petal Width tienen distribuciones más simétricas.

Iris-setosa es la clase más frecuente, con un 55.3% de las observaciones, mientras que Iris-versicolor y Iris-virginica tienen proporciones más equilibradas entre sí, pero con menos muestras que Iris-setosa.



**Resultados del Modelo Predictivo:**

R² (Coeficiente de determinación): 0.9347, lo que indica que el modelo explica el 93.47% de la variabilidad en los datos, lo cual es un excelente ajuste.

MSE (Error cuadrático medio): 0.0054, que muestra que el modelo tiene un error pequeño en las predicciones.

RMSE (Raíz del error cuadrático medio): 0.0732, lo que también sugiere una baja diferencia entre las predicciones y los valores reales.

**Interpretación de Coeficientes:**

Petal Width (0.6115): La variable más importante en el modelo. A medida que el ancho del pétalo aumenta, también lo hace la variable objetivo (posiblemente el tipo de flor).

Sepal Length (0.2993): La longitud del sépalo también tiene una influencia significativa en la variable objetivo, pero su impacto es menor que el del ancho del pétalo.

Sepal Width (-0.2550): El ancho del sépalo tiene un impacto negativo, lo que significa que a medida que aumenta el ancho del sépalo, la variable objetivo tiende a disminuir.

Class (0.0207): La clase tiene una baja influencia en el modelo, lo que sugiere que el tipo de flor no es un factor determinante por sí solo para la predicción.

**Respuesta a la Pregunta de Investigación:**

El modelo predictivo confirma que las características de los pétalos (especialmente el ancho del pétalo) son las más relevantes para predecir el tipo de flor, respondiendo efectivamente a la pregunta de investigación sobre cómo se pueden predecir las clases de flores a partir de sus características.

=== EVALUACIÓN DEL MODELO ===

MSE: 0.0054

RMSE: 0.0732

R²: 0.9347

El modelo explica el 93.47% de la variabilidad.

=== IMPORTANCIA DE VARIABLES ===

	Variable	Coeficiente
2	petal_width	0.611505
0	sepal_length	0.299347
1	sepal_width	-0.255007
3	class	0.020790



### **Limitaciones del Dataset:**

**Desbalance de clases:** La clase Iris-setosa está sobre-representada (55.3%), lo que podría influir en los resultados del modelo, favoreciendo la clase mayoritaria.

**Datos faltantes:** Algunas variables presentan registros con valores faltantes (por ejemplo, sepal\_width tiene 99 registros no nulos), lo que podría afectar la calidad de las predicciones si no se manejan adecuadamente.

**Tamaño limitado del dataset:** El conjunto de datos solo contiene 150 registros, lo que podría no ser suficiente para generalizar conclusiones a poblaciones más grandes o para entrenar modelos más complejos.

### **Restricciones del Análisis:**

**Modelo simple:** El modelo utilizado es lineal y puede no capturar completamente las relaciones no lineales o complejas entre las variables, lo que limita la capacidad de predicción en situaciones más complejas.

**Variables limitadas:** El análisis se basa en solo unas pocas características (longitud y ancho de sépalos y pétalos), y podría haber otras variables relevantes no incluidas en el dataset que afecten la clasificación de las flores.

### **Posibles Sesgos:**

**Sesgo de muestreo:** Al ser un dataset relativamente pequeño y específico (solo tres tipos de flores Iris), los resultados podrían no ser representativos de otras especies de flores o de situaciones fuera del entorno controlado.

**Sesgo de clases:** El desbalance en la distribución de clases podría llevar a un modelo que favorezca incorrectamente la clase Iris-setosa en las predicciones.

### **Recomendaciones para Futuros Análisis:**

**Recolección de más datos:** Ampliar el conjunto de datos con más muestras y posiblemente más clases de flores para mejorar la generalización y precisión del modelo.

**Manejo de datos faltantes:** Implementar técnicas avanzadas para manejar los datos faltantes, como imputación o recolección de datos completos.

**Explorar modelos más complejos:** Evaluar el uso de modelos no lineales (como árboles de decisión, SVM, o redes neuronales) para mejorar la predicción y capturar patrones más complejos.

**Balanceo de clases:** Implementar técnicas como SMOTE o ajuste de pesos en las clases para mitigar el desbalance y mejorar la precisión en la predicción de clases menos representadas.

