



CIENCIA DE DATOS

# Iris Dataset – Datos sobre flores Iris

**BOGOTA DC – 2024**

**SANTIAGO CARVAJAL  
FERNANDEZ**



# Introducción

El Iris Dataset es uno de los conjuntos de datos más utilizados en el campo de la Ciencia de Datos y el aprendizaje automático. Creado por Ronald A. Fisher en 1936, este conjunto de datos contiene información sobre tres especies de la flor iris: Iris setosa, Iris versicolor e Iris virginica. Cada observación incluye cuatro características: la longitud y el ancho del sépalos, y la longitud y el ancho del pétalo, las cuales se utilizan para clasificar correctamente las especies de iris.


En el código presentado, el Iris Dataset se carga y se utiliza para realizar un análisis exploratorio inicial. Se comienza imprimiendo las primeras filas de las características ( $X$ ) y las etiquetas objetivo ( $y$ ), seguido por el cálculo de estadísticas descriptivas clave, como la media, la mediana y la desviación estándar para cada característica. Este análisis permite obtener una comprensión inicial de la distribución y el comportamiento de los datos, lo cual es un paso esencial en cualquier proceso de análisis en ciencia de datos.



# ENSAYO

El código implementado tiene como objetivo introducir al usuario en el manejo de datos utilizando la librería pandas y trabajar con el conjunto de datos Iris, disponible en la plataforma UCI Machine Learning Repository. Este dataset consta de 150 observaciones que describen tres especies de iris (setosa, versicolor y virginica), con cuatro características numéricas: longitud y ancho del sépalo, y longitud y ancho del pétalo. Estas características se utilizan para entrenar modelos de clasificación y análisis estadístico.

El script comienza imprimiendo la información básica del estudiante: su nombre, la materia y el salón de clases. A continuación, importa las bibliotecas necesarias (pandas y numpy), que son fundamentales para el manejo de datos en Python. La función `fetch_ucirepo` se emplea para descargar el dataset Iris, y los datos se dividen en dos conjuntos: `X`, que contiene las características del dataset, e `y`, que almacena los objetivos o etiquetas correspondientes a las especies de flores.



Después de la carga de los datos, el código imprime las primeras filas de las características y los objetivos, lo que permite visualizar rápidamente los valores del conjunto de datos. A continuación, se realizan varios cálculos estadísticos descriptivos que son cruciales para entender mejor la distribución de los datos:

1. **Media:** La media de cada característica proporciona una visión general de los valores promedio para cada variable. Esto es útil para identificar posibles tendencias o diferencias entre las especies.
2. **Mediana:** La mediana es un indicador clave de la tendencia central que es menos susceptible a valores atípicos, lo que la convierte en una medida robusta para el análisis exploratorio.
3. **Desviación estándar:** La desviación estándar mide la dispersión de los datos, es decir, cuánto varían los valores de cada característica con respecto a la media. Un valor alto indica una mayor variabilidad, mientras que un valor bajo sugiere que los datos están más concentrados cerca de la media.