Actividad 3: Clasificación con el dataset iris con regresión lineal

Brayan Rodríguez Joseph Valencia Cruz 16 de septiembre de 2025

# Introducción

En esta Actividad, desarrollamos para clasificar las especies del dataset iris utilizando un modelo de aprendizaje automático basado en regresión lineal. El objetivo fue clasificar automáticamente las flores en 3 categorías: iris setosa, iris versicolor e iris virginica, a partir de un conjunto de datos con 150 muestras. Este informe detalla el proceso seguido, con énfasis en cómo funciona l regresión lineal.

# Descripción del Proyecto

La actividad consistió en crear un programa en Python que procesa el dataset iris, lo estandariza, entrena un modelo de regresión lineal para clasificar las especies y evalua su rendimiento. A continuación, describimos los pasos principales:

* Carga de datos: Se cargo el dataset iris con 150 muestras y características como longitud y ancho de sepalos y pétalos. Y se mostraron estadísticas básicas.
* Preprocesamiento: Se estandarizaron los datos usando StandardScaler para mejorar la precisión del modelo.
* División de datos: Los datos se dividieron en 80% entrenamiento y 20% para prueba, manteniendo la proporción de clases.
* Entrenamiento: Se entreno un modelo de regresión lineal para clasificar especies.
* Evaluación: Se midió el rendimiento con métricas como precisión, sensibilidad y F1-score, además de una matriz de confusión.
* Análisis: Se identificaron las características mas influyentes para la clasificación.

# Regresión Lineal

La regresión lineal, es un algoritmo de aprendizaje automático utilizado para clasificar datos en múltiples categorías. Funciona ajustando una línea de decisión basada en las características para separa las clases. Es ideal en problemas de clasificación multiclase y da un enfoque simple pero efectivo cuando se combina con regularización.

# Resultados

El modelo fue evaluado con varias métricas:

 **Precisión general (Accuracy)**: Porcentaje de las flores clasificadas correctamente.

 **Precisión (Precisión)**: Proporción de clasificaciones correctas por especie.

 **Sensibilidad (Recall)**: Cantidad de flores de cada especie detectadas correctamente.

 **F1-Score**: Medida balanceada y precisión y sensibilidad.

Se genero una matriz de confusión para mostrar clasificaciones correctas e incorrectas. Las características mas importantes, como la longitud de pétalo, se destacaron como clave para distinguir entre especies.

# Conclusión

Desarrollamos un sistema efectivo para clasificar las especies del dataset Iris utilizando regresión lineal. El proceso incluyó cargar y estandarizar los datos, entrenar un modelo y evaluar su rendimiento. En este enfoque demostró ser el adecuado para problemas de clasificación multiclase con datos numéricos, proporcionando una base solida para entender como el aprendizaje automático puede resolver problemas botánicos.