Clasificación de Flores Iris con Regresión Lineal

Alejandro Flórez Lesmes
Yeffersson Stiven Castro
Universidad de Cundinamarca

15 de septiembre de 2025

1. Introducción

El dataset de flores Iris es uno de los conjuntos de datos más famosos y utilizados en el campo del aprendizaje automático para problemas de clasificación. Generalmente, los modelos alcanzan precisiones muy altas, a menudo del 100 %, debido a la fuerte capacidad predictiva de las medidas del pétalo de las flores.

Sin embargo, un rendimiento perfecto en un entorno de laboratorio puede ser engañoso y no representa los desafíos de los problemas del mundo real, donde las clases a menudo se superponen. Por esta razón, este informe documenta un experimento donde se restringe el análisis a las características del **sépalo** (*Sepal Length* y *Sepal Width*), las cuales son conocidas por tener una menor capacidad de separación entre las especies *Iris-versicolor* e *Iris-virginica*.

El objetivo principal es doble: primero, comparar el rendimiento de un modelo de **Regresión Lineal** contra un modelo de **Regresión Logística** adaptado para clasificación; y segundo, analizar críticamente cómo la elección de características menos informativas crea un problema más realista y reduce la precisión del modelo, obligándonos a interpretar sus errores.

2. Metodología

El proceso de análisis se implementó en un script de Python utilizando librerías estándar del ecosistema de ciencia de datos. Los pasos metodológicos fueron los siguientes:

2.1. Preparación de los Datos

Se cargó el dataset `Iris.csv` utilizando la librería `pandas`. Las etiquetas categóricas de la columna `Species` ('Iris-setosa', 'Iris-versicolor', 'Iris-virginica') fueron transformadas a valores numéricos (0, 1, 2) mediante un `LabelEncoder` de `scikit-learn` para hacerlas compatibles con los modelos.

2.2. Selección de Variables y División del Conjunto

De acuerdo con el objetivo del estudio, se seleccionaron únicamente dos características para el entrenamiento:

- `SepalLengthCm` (Largo del Sépalo)
- `SepalWidthCm` (Ancho del Sépalo)

Posteriormente, el conjunto de datos fue dividido en un 80 % para entrenamiento y un 20 % para pruebas, utilizando una semilla aleatoria (`random_state=42`) para garantizar la reproducibilidad de los resultados.

2.3. Entrenamiento de Modelos

Se entrenaron dos modelos diferentes con los datos de entrenamiento:

1. **Regresión Lineal:** Aunque es un modelo para regresión, se utilizó para predecir un valor continuo que luego fue redondeado y acotado para simular una clasificación. Se incluye como punto de referencia de un modelo no ideal.

2. **Regresión Logística:** Es el modelo principal de estudio, diseñado específicamente para problemas de clasificación.

2.4. Evaluación

La métrica principal para evaluar el rendimiento fue la **exactitud (accuracy)**, que mide la proporción de predicciones correctas. Los resultados obtenidos en la consola fueron los siguientes:

Listing 1: Salida de Precisión de los Modelos

```
 Precision comparativa:
- Regresion Lineal: 86.67%
- Regresion Logistica: 90.00%
```

La Regresión Logística obtuvo un rendimiento superior, como era de esperar. Ambos modelos mostraron una precisión significativamente menor al 100 %, confirmando que las características del sépalo presentan un desafío de clasificación.

3. Resultados y Gráficos

El análisis visual es fundamental para comprender por qué los modelos no alcanzan una precisión perfecta.

3.1. Gráfico 1: Dispersión de Características del Sépalo

Este gráfico (Figura 1) visualiza la distribución de las tres especies de Iris en el espacio definido por las medidas del sépalo.

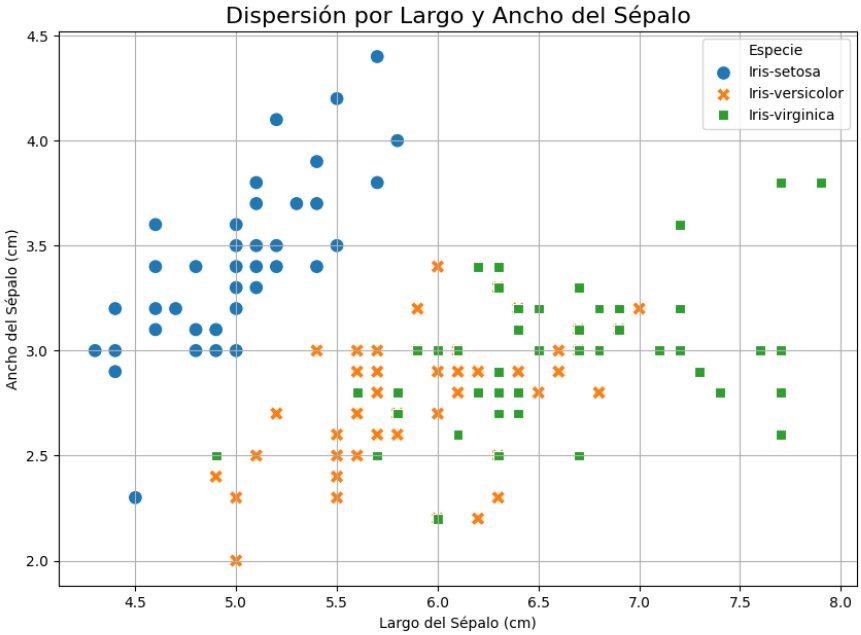

Figura 1: Distribución de las especies de Iris según las medidas del sépalo. Se observa un claro solapamiento entre Iris-versicolor e Iris-virginica.

Como se puede apreciar, mientras que *Iris-setosa* (en azul) forma un clúster claramente separable, los puntos correspondientes a *Iris-versicolor* (naranja) e *Iris-virginica* (verde) están considerablemente mezclados. Este solapamiento es la causa principal de los errores de clasificación.

3.2. Gráfico 2: Matriz de Confusión del Modelo Lineal

La matriz de confusión (Figura 2) detalla los aciertos y errores del modelo de Regresión Lineal adaptado sobre el conjunto de prueba.

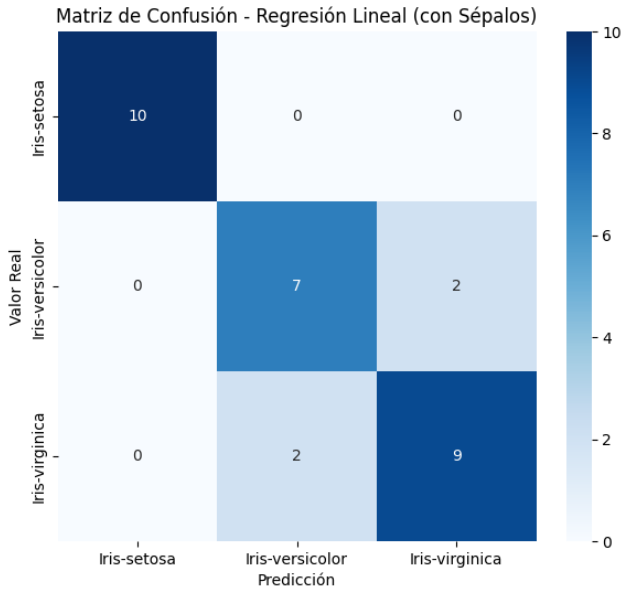

Figura 2: Matriz de confusión para el modelo de Regresión Lineal.

La diagonal principal muestra las predicciones correctas. Los valores fuera de la diagonal indican errores. En este caso, el modelo clasificó incorrectamente dos *Iris-versicolor* como *Iris-virginica*, lo cual es consistente con el solapamiento observado en la Figura 1.

4. Discusión Crítica

El experimento demuestra una lección fundamental en el aprendizaje automático: **la calidad de las características es a menudo más importante que la complejidad del modelo**. A pesar de utilizar algoritmos capaces, la precisión se vio limitada por la información inherente en los datos de entrada.

Al forzar a los modelos a usar solo las medidas del sépalo, simulamos un escenario común en problemas reales donde no existen características "perfectas"que separen las clases de manera trivial. La caída de la precisión del 100 % (obtenible con las características del pétalo) a un 90 % refleja esta dificultad añadida.

5. Conclusiones

Del presente análisis se desprenden las siguientes conclusiones:

1. La **Regresión Logística** es más adecuada para tareas de clasificación que la Regresión Lineal adaptada, obteniendo una mayor precisión en el problema planteado.

2. La **selección de características** es un paso crítico que define el límite superior del rendimiento de un modelo. Las características del sépalo del dataset Iris no permiten una separación lineal perfecta entre todas las especies.

3. El **análisis visual**es una herramienta indispensable para diagnosticar por qué un modelo comete errores y para entender la estructura intrínseca de los datos.

4. Al limitar las características, se logró transformar un problema de libro de texto en un **escenario de clasificación más realista**, donde la superposición de clases es la norma.

Repositorio del Proyecto: https://github.com/StivenCastro138/Machine_Learning-Modelo2-git