Modelo Regresión Lineal Dataset Iris

Integrantes: Gabriel Martinez Roldan, Ivone López Cruz
Machine Learning 801

Link al Repositorio de Github

El código fuente completo y los archivos se encuentran disponibles en el siguiente repositorio de Github:
https://github.com/RFGRONA/Machine-Learning-801-IIPA-25/tree/128310ff9d95de6ae7d202e3e9be3611d73b3059/2.%20Regresi%C3%B3n%20Lineal%20-%20Iris

Descripción del Diseño del Programa

Este programa utiliza un modelo de Regresión Lineal para resolver un problema de clasificación, para el conjunto de datos Iris. El flujo de trabajo del programa se divide en los siguientes pasos:

➢ Carga y Preparación de Datos:

El script comienza cargando el conjunto de datos Iris.csv usando la librería pandas.
Para que el modelo de regresión lineal pueda procesar los datos, las etiquetas de texto de las especies (Iris-setosa, Iris-versicolor, Iris-virginica) se convierten a valores numéricos (0, 1 y 2 respectivamente).

➢ Entrenamiento del Modelo:

Se crea una instancia del modelo LinearRegression de la librería scikit-learn.
El modelo se entrena utilizando el método .fit(), pasándole las cuatro características de las flores (largo y ancho del sépalo y pétalo) como X y las especies numéricas como y. El modelo aprende a encontrar la relación matemática lineal entre las medidas de las flores y su categoría numérica.

➢ Predicción y Clasificación:

Una vez entrenado, el modelo utiliza el método .predict() para generar predicciones sobre los mismos datos. Estas predicciones son valores continuos (con decimales).
Para convertir estas predicciones en una clase específica, se aplica un redondeo al entero más cercano. Por ejemplo, una predicción de 1.98 se convierte en 2 (Iris-virginica). Se utiliza np.clip para asegurar que los resultados no se salgan del rango [0, 2].

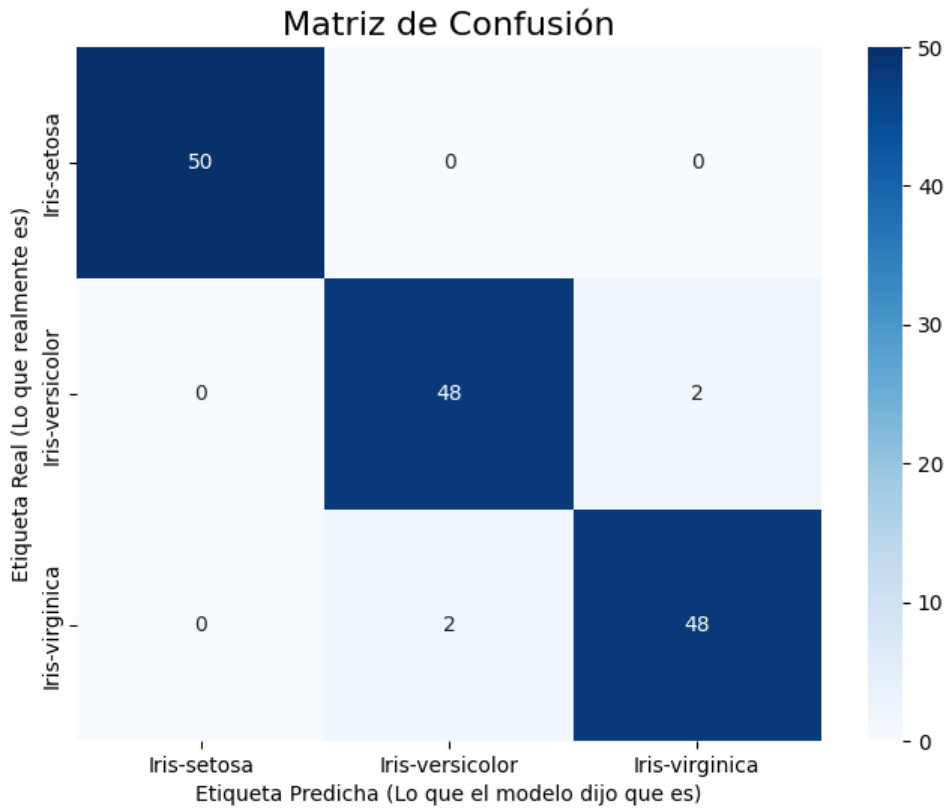
➢ Evaluación y Visualización:

Se calculan y se imprimen en la consola diversas métricas de rendimiento, como la precisión (accuracy) y el error cuadrático medio (MSE**)**, para evaluar la efectividad del modelo.
Finalmente, el script genera tres gráficas clave para visualizar y analizar el comportamiento y los resultados del modelo.

➢ Análisis de las Gráficas Generadas

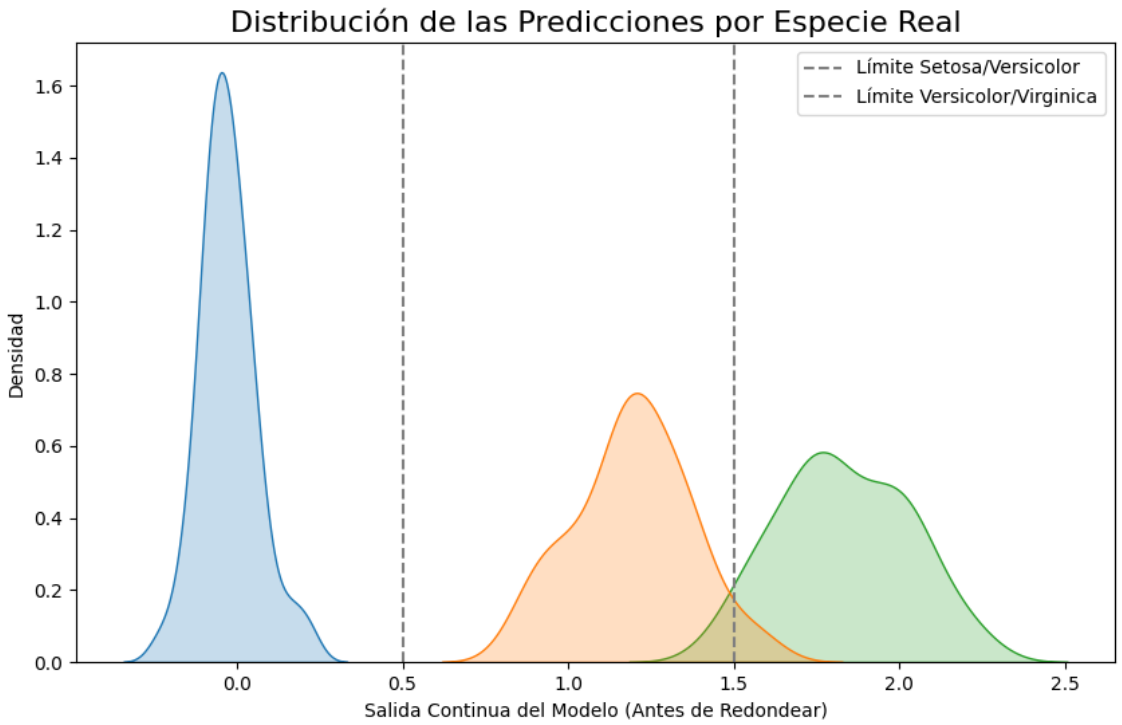
El programa genera las siguientes visualizaciones para interpretar el rendimiento del modelo:

✓ Gráfica 1: Matriz de Confusión



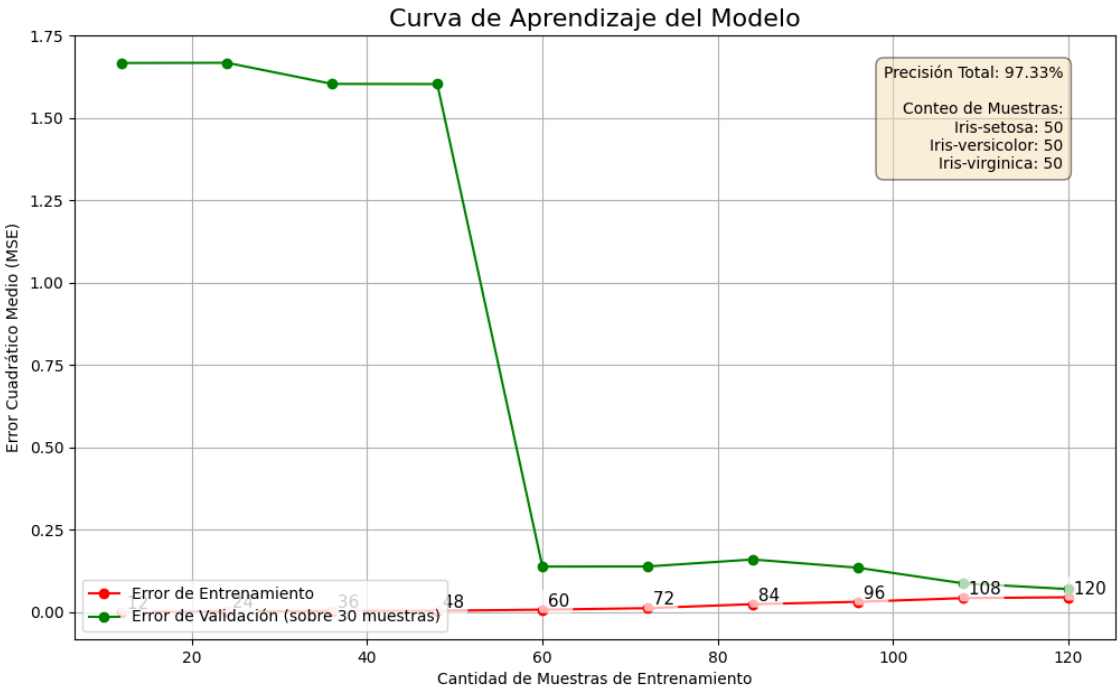
- **¿Qué Muestra?**: Esta matriz es el reporte visual de los aciertos y errores del modelo. Compara las etiquetas reales de las flores con las etiquetas que el modelo predijo.
- **Análisis**: Se observa que el modelo es perfecto para Iris-setosa, clasificando correctamente las 50 muestras. También tiene un alto rendimiento para las otras dos especies, acertando en 48 de 50 para Iris-versicolor y 48 de 50 para Iris-virginica.
  La matriz revela que la única confusión del modelo ocurre entre versicolor y virginica. Específicamente, clasificó 2 versicolor como virginica y 2 virginica como versicolor.

✓ **Gráfica 2: Distribución de las Predicciones**



- **¿Qué Muestra?**: Muestra cómo se agrupan las predicciones del modelo (los valores con decimales) para cada especie real.
- **Análisis**: La curva de Iris-setosa (azul) está completamente aislada y centrada en 0, lo que confirma por qué el modelo nunca falla con ella.
  Las curvas de versicolor (naranja) y virginica (verde) están bien centradas en 1 y 2. Sin embargo, sus bases se superponen ligeramente alrededor del valor 1.5. Esta pequeña área de cruce es la que causa los 4 errores de clasificación que se ven en la Matriz de Confusión.

✓ Gráfica 3: Curva de Aprendizaje del Modelo



- **¿Qué Muestra?**: Ilustra cómo mejora el error del modelo a medida que se le proporcionan más datos de entrenamiento. También muestra información clave sobre el rendimiento general.
- **Análisis**: Las líneas de error de entrenamiento (roja) y de validación (verde) convergen y se estabilizan en un valor de error muy bajo. Este es el comportamiento ideal, ya que indica que el modelo no sufre de sobreajuste (no está memorizando) ni de subajuste (no es demasiado simple). Aprende bien y generaliza correctamente.
  El hecho de que las curvas se aplanen al final sugiere que el modelo ha alcanzado su máximo rendimiento. Añadir más datos probablemente no mejoraría significativamente su precisión.
  El texto en la esquina superior derecha resume los datos clave: la precisión final del 97.33% y el conteo balanceado de 50 muestras por especie, proporcionando un contexto completo en una sola vista.