CLASIFICACIÓN DE FLORES CON REGRESIÓN LOGÍSTICA

CESAR AGUIRRE HURTADO

UNIVERSIDAD DE CUNDINAMARCA
ALEXANDER ESPINOSA
MACHINE LEARNING
SEPTIEMBRE

Introducción

El presente trabajo tiene como propósito aplicar un modelo de aprendizaje supervisado para la clasificación de especies de flores del conjunto de datos Iris. Este dataset contiene medidas de sépalos y pétalos correspondientes a tres especies distintas: Iris setosa, Iris versicolor y Iris virginica. La motivación principal es comprobar cómo, a través de técnicas estadísticas como la Regresión Logística, es posible predecir la especie de una flor a partir de sus características físicas. Además, se busca interpretar el comportamiento de las variables y determinar cuáles son más relevantes para la clasificación.

Diseño del modelo

El diseño partió de un enfoque clásico de clasificación multiclase. Se definió la Regresión Logística como modelo base por su simplicidad, eficiencia y facilidad de interpretación. El conjunto de datos Iris fue tomado directamente de la librería *scikit-learn*. Este dataset incluye 150 observaciones, divididas equitativamente en las tres especies. Cada observación contiene cuatro variables de entrada: longitud del sépalo, ancho del sépalo, longitud del pétalo y ancho del pétalo. El modelo se estructuró bajo la idea de entrenar un clasificador con el 70% de los datos, mientras que el 30% restante se utilizó para validación. De esta manera, fue posible evaluar la capacidad del modelo para generalizar a nuevos casos.

Procedimiento

El procedimiento seguido para el desarrollo del trabajo se puede resumir en las siguientes etapas:

Primero se realizó la carga y exploración inicial del dataset, identificando sus variables y ajustando los nombres de las características al español para facilitar la interpretación.

Posteriormente, se llevó a cabo la división del conjunto de datos en entrenamiento y prueba. Esta etapa fue fundamental para evitar el sobreajuste y garantizar que el modelo fuera evaluado de manera justa.

En la tercera etapa se entrenó el modelo de Regresión Logística multiclase, ajustando los parámetros necesarios para que pudiera reconocer patrones en los datos de entrenamiento.

Luego se realizó la evaluación del desempeño mediante la matriz de confusión y el reporte de clasificación. Esto permitió analizar los aciertos y errores del modelo en la predicción de cada clase. Por último se elaboraron las gráficas de importancia de las características, con el fin de determinar qué variables influyen en mayor medida en la clasificación de las flores.

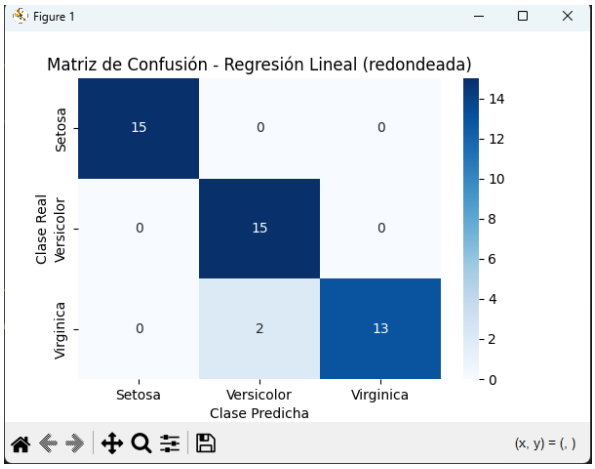
Descripción del algoritmo

El algoritmo de Regresión Logística pertenece a los métodos de aprendizaje supervisado. Su funcionamiento se basa en calcular probabilidades de pertenencia a una clase a través de la función logística (sigmoide). En el caso de problemas multiclase, como el dataset Iris, el algoritmo utiliza una extensión conocida como Regresión Logística Multinomial. Aquí, el modelo estima una probabilidad para cada clase y asigna la observación a aquella con el valor más alto.

De manera simplificada, el procedimiento matemático consiste en aplicar una transformación lineal a los datos de entrada y posteriormente pasarlos por la función logística, lo cual entrega valores entre 0 y 1 que representan las probabilidades de clasificación.

Resultados e interpretación

Matriz de confusión.



La matriz de confusión muestra el desempeño del modelo de clasificación comparando los valores reales contra las predicciones.

Clase Setosa

- 15 observaciones reales de *Setosa* fueron clasificadas correctamente como *Setosa*.
- No se presentó ningún error en esta categoría, lo que significa que el modelo reconoce a la perfección esta especie.

Clase Versicolor

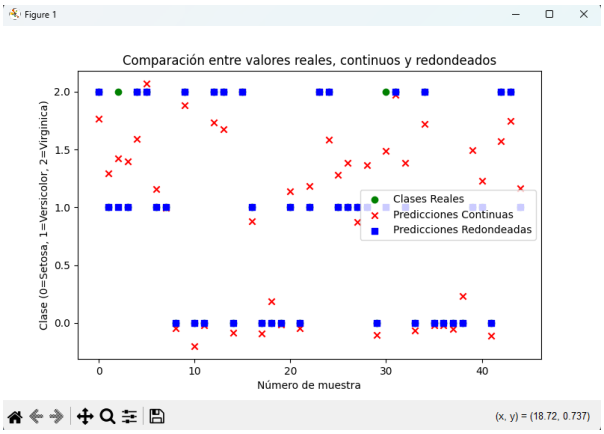
- 15 observaciones reales de *Versicolor* fueron clasificadas correctamente.
- Tampoco hubo errores en esta clase, lo que refleja una gran capacidad del modelo para diferenciar a esta especie.
-

Clase Virginica

- De las 15 observaciones reales de *Virginica*, 13 fueron correctamente clasificadas.
- Sin embargo, 2 fueron confundidas con *Versicolor*.
- Este error es entendible porque *Versicolor* y *Virginica* presentan medidas de pétalos muy similares, lo cual dificulta al modelo distinguirlas.

El modelo logró clasificar de manera correcta 43 de 45 casos en el conjunto de prueba. Esto equivale a una exactitud del 95,6%, lo cual demuestra un desempeño excelente. Los errores están concentrados únicamente en la confusión entre Versicolor y Virginica. Esto confirma que la especie Setosa es fácilmente separable en el espacio de características, mientras que Versicolor y Virginica tienen fronteras más difusas.

Comparación entre valores.



En el eje X se tiene el número de muestra y en el eje Y la clase de la flor (0 = Setosa, 1 = Versicolor, 2 = Virginica).

Elementos principales:

Clases reales (verde, círculos)
Representan la categoría verdadera de cada muestra. Se ubican únicamente en los valores 0, 1 o 2, ya que son clases discretas.

Predicciones continuas (rojo, equis)
Aquí se observa cómo funciona la regresión lineal: el modelo no entrega una clase directamente, sino un valor numérico continuo que puede estar entre 0 y 2.

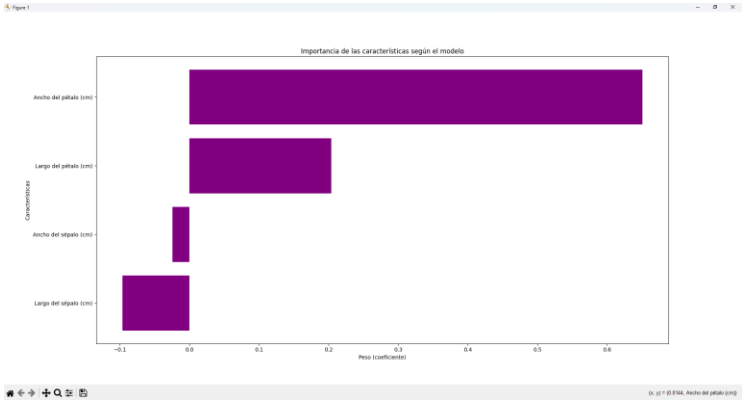
- Por ejemplo, en vez de dar la clase 1, puede dar un valor como 1.3 o 0.8.
- Esto explica por qué algunos puntos rojos aparecen entre los valores enteros del eje Y.

Predicciones redondeadas (azul, cuadrados)
Para convertir las salidas continuas en clases válidas (0, 1, 2), se aplica un redondeo.

- Con este ajuste, muchas de las predicciones se alinean con los valores reales.
- No obstante, en algunos casos se genera error, como cuando una predicción continua cercana a 1.6 se redondea a 2, clasificando mal una flor que en realidad era clase 1 (Versicolor).

La gráfica evidencia que la regresión lineal no es un modelo nativo de clasificación, ya que produce valores continuos. Sin embargo, al redondearlos se logra una aproximación bastante buena. Los círculos verdes y los cuadrados azules coinciden en la mayoría de los casos, lo que refleja el buen rendimiento del modelo. Los errores ocurren cuando las predicciones continuas quedan demasiado alejadas del valor correcto y, tras el redondeo, caen en la clase equivocada. Este comportamiento confirma lo observado en la matriz de confusión: Setosa es siempre bien clasificada, pero hay confusión entre Versicolor y Virginica.

Importancia de las características.



Se muestra el peso coeficiente que la Regresión Lineal asigna a cada variable del dataset Iris, lo que permite identificar cuáles características tienen más influencia en la clasificación de las flores.

Ancho del pétalo (cm)
Es la característica con mayor peso positivo en el modelo. Esto significa que es el rasgo más determinante para diferenciar entre las especies. En la práctica, cuando el ancho del pétalo aumenta, la probabilidad de que la flor pertenezca a Virginica se incrementa notablemente.

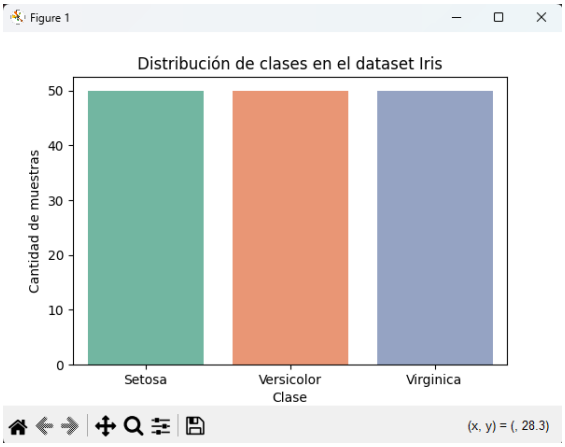
Largo del pétalo (cm)
También tiene un peso positivo importante, aunque menor que el ancho del pétalo. Es otra variable fundamental, sobre todo para distinguir entre Setosa y las otras dos especies.

Ancho del sépalo (cm)
Presenta un peso pequeño y positivo, lo que indica que su influencia en la clasificación es limitada, pero de todas formas aporta algo de información.

Largo del sépalo (cm)
Tiene un peso negativo en el modelo. Esto significa que valores más altos en el largo del sépalo no ayudan a distinguir las clases y, en algunos casos, incluso generan confusión en la predicción.

La gráfica confirma lo que se sabe del dataset Iris: las características más relevantes para separar las especies son las medidas del pétalo (ancho y largo), mientras que las características del sépalo tienen un aporte menor. Esto explica por qué el modelo logra clasificar tan bien a Setosa, que tiene pétalos muy diferentes respecto a las otras clases, mientras que Versicolor y Virginica, que son más parecidas en esas medidas, generan mayor confusión.

Distribución de clases.



Aquí se representa la cantidad de muestras por cada clase de flor dentro del conjunto de datos Iris. Se observa que: Setosa, Versicolor y Virginica tienen exactamente el mismo número de registros: 50 muestras cada una. Ademas esto genera una distribución equilibrada y uniforme entre las clases. Este equilibrio es muy importante porque garantiza que el modelo de clasificación no se entrene con sesgo hacia una clase mayoritaria, permite que las métricas de rendimiento (precisión, recall, F1-score) sean más representativas y justas y reduce la necesidad de aplicar técnicas de balanceo de datos (como sobremuestreo o submuestreo), ya que no hay desbalance.

Conclusiones.

El modelo de clasificación entrenado con el dataset Iris logró una exactitud del 95.6%, lo que significa que de las 45 muestras de prueba, solo se equivocó en 2 predicciones. Los coeficientes del modelo muestran la importancia de cada característica: el largo del sépalo (-0.0963) y el ancho del sépalo (-0.0245) aportan muy poco a la clasificación, mientras que el largo del pétalo (0.2037) y especialmente el ancho del pétalo (0.6510) son los factores más relevantes para diferenciar las especies. La matriz de confusión confirma estos resultados: para Setosa (15/15) y Versicolor (15/15) todas las muestras se clasificaron correctamente, mientras que en Virginica, 13 fueron clasificadas bien y 2 se confundieron con Versicolor. Esto se refleja en el reporte de clasificación, donde Setosa alcanzó valores perfectos de precisión, recall y f1-score (1.00), Versicolor obtuvo precisión de 0.88, recall de 1.00 y f1-score de 0.94, y Virginica mostró una precisión de 1.00, recall de 0.87 y f1-score de 0.93. A nivel global, los promedios ponderados se mantuvieron en 0.96 para todas las métricas. En cuanto a la validación cruzada, los valores de R^2 por fold fueron [0.00, 0.85, 0.00, 0.76, 0.00], con un promedio de 0.32, lo que refleja que esta métrica no es la más adecuada para clasificación pero igualmente muestra que en algunos subconjuntos el modelo mantiene un buen desempeño. Finalmente, en las predicciones de prueba con nuevos datos, el modelo identificó correctamente a una muestra como Setosa y otra como Virginica, lo cual confirma su capacidad de generalizar más allá de los datos de entrenamiento.

El trabajo permitió comprobar la utilidad de la Regresión Logística en un problema clásico de clasificación multiclase. El modelo alcanzó un desempeño sobresaliente, logrando clasificar de manera precisa a la mayoría de las observaciones. La especie Iris setosa fue distinguida de forma perfecta, mientras que los errores se dieron únicamente en la confusión entre versicolor y virginica. Además, se observó que las características relacionadas con los pétalos tienen mayor peso en la clasificación que las medidas de los sépalos.