

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Querétaro



**Tecnológico
de Monterrey**

TC3006C. Inteligencia artificial avanzada para la ciencia de datos I

Grupo 101

Momento de Retroalimentación :

“Módulo 2 Análisis y Reporte
sobre el desempeño del modelo.
(Portafolio Análisis)”

Evidencia presentada por:

Emiliano Mendoza Nieto

A01706083

Profesor :

Benjamín Valdés Aguirre

Fecha de entrega :

Domingo 10 de Septiembre de 2023

I. Introducción

El objetivo de este análisis es evaluar el desempeño de un modelo de Regresión Logística en el conjunto de datos 'Iris.csv' tomado de Kaggle(<https://www.kaggle.com/datasets/uciml/iris?select=Iris.csv>)'. A través de diferentes métricas, se buscará entender el nivel de ajuste del modelo.

II. Descripción Base de Datos:

Clasifique las plantas de iris en tres especies en este conjunto de datos clásico.

Acerca del conjunto de datos

El conjunto de datos Iris se utilizó en R.A. El artículo clásico de Fisher de 1936, El uso de medidas múltiples en problemas taxonómicos, también se puede encontrar en el Repositorio de aprendizaje automático de la UCI.

Incluye tres especies de iris con 50 muestras cada una, así como algunas propiedades de cada flor. Una especie de flor es linealmente separable de las otras dos, pero las otras dos no son linealmente separables entre sí.

Las columnas de este conjunto de datos son:

- Id
- SepalLengthCm
- SepalWidthCm
- PetalLengthCm
- PetalWidthCm
- Species

III. Evaluación del Modelo:

- ➔ Underfitting (Subajuste): Si el modelo tiene alto bias y baja varianza.

- ➔ Ajuste Correcto: Si el modelo tiene un balance entre bias y varianza.
- ➔ Overfitting (Sobreaajuste): Si el modelo tiene bajo bias y alta varianza.

A continuación se muestran ejemplos (Figura 1) de los diferentes ajustes que se podrían obtener.

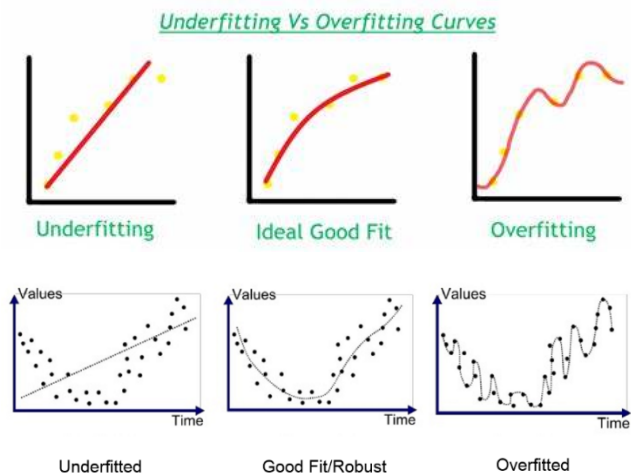


Figura 1. Ejemplos de comportamiento

V. Mejoras y Regularización:

Para mejorar el modelo, se puede considerar:

- ★ Regularización: En el modelo se utiliza la regularización L2 y L1. Se puede experimentar con diferentes valores de la constante de regularización para mejorar el desempeño.
- ★ Ajuste de Parámetros: Aumentar el max_iter o cambiar el solver puede ayudar en la convergencia y desempeño del modelo. Como veremos a continuación con los 3 diferentes solver utilizados.

VI. ANALISIS:

➤ Regresión Logística con Liblinear

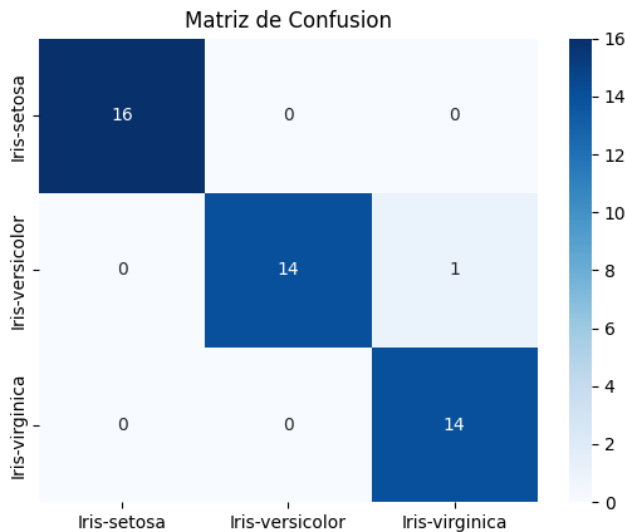


Figura 2. Matriz de Confusion de Liblinear

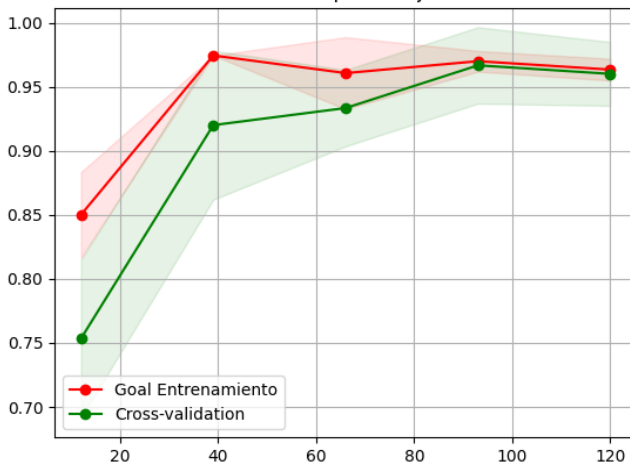


Figura 3. Curva de Aprendizaje Liblinear

Accuracy (TRAIN)	0.971
Error (TRAIN)	0.0285
Accuracy (TEST)	0.9777
Error (TEST)	0.02222

Figura 4. Tabla de resultados de Liblinear

El modelo ya está utilizando la regularización L1 (penalización 'l1'). La regularización L1 tiende a hacer que algunos coeficientes sean exactamente cero, lo que se evidencia en los coeficientes impresos (`model.coef_`).

Las métricas indican que el modelo tiene un alto rendimiento, y el error en los conjuntos de entrenamiento y prueba es muy similar.

Dado que la exactitud es alta y el error es bajo en ambos conjuntos, el sesgo es bajo.

La diferencia entre la precisión de entrenamiento y prueba es mínima, por lo que podemos asumir que el modelo no sufre de alta varianza.

En este caso, dado que ambos, sesgo y varianza, son bajos, **podemos decir que el modelo está bien ajustado.**

El modelo actual ya tiene un rendimiento impresionante. Sin embargo, para mejorar aún más, podría experimentar con:

Uso de técnicas de validación cruzada para una evaluación más robusta.

➤ Regresión Logística Newton-Conjugate Gradient

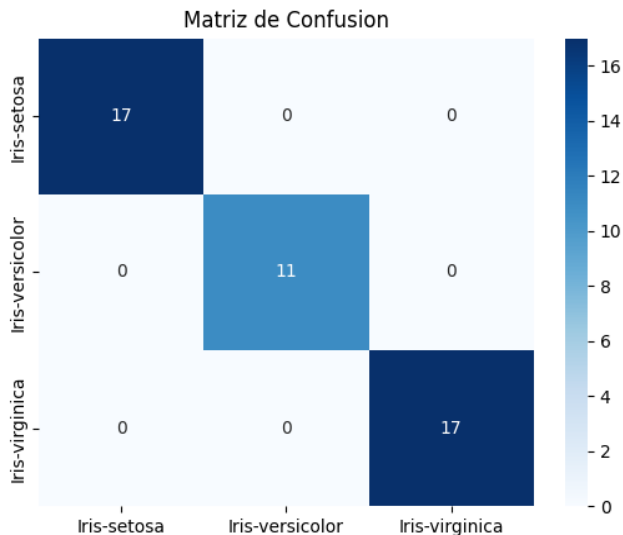


Figura 5. Matriz de Confusion de Newton-CG

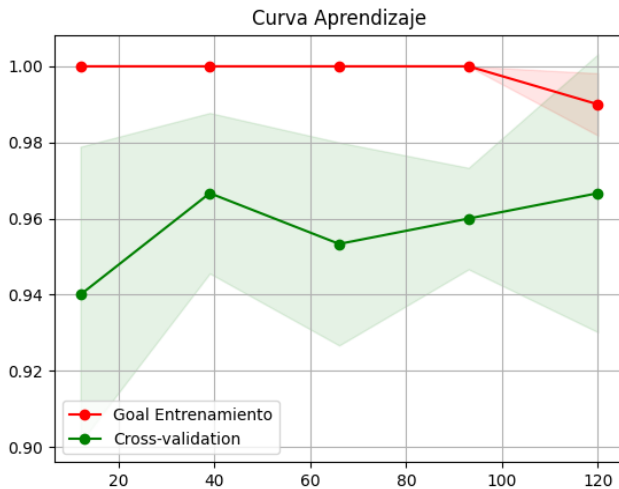


Figura 6. Curva de Aprendizaje n de Newton-CG

Accuracy (TRAIN)	1.0
Error (TRAIN)	0.019
Accuracy (TEST)	1.0
Error (TEST)	0.0

Figura 7. Tabla de resultados Newton-CG

La diferencia entre la precisión de entrenamiento y prueba es mínima. Aunque el modelo tiene una perfecta exactitud en el conjunto de prueba, hay un pequeño error en el conjunto de entrenamiento.

Dado que ambos, sesgo y varianza, son bajos, podemos decir que el modelo está bien ajustado,

La regresión logística con el método de Newton-Conjugate Gradient ha funcionado bien en este conjunto de datos. Y aunque el modelo no está regularizado (`penalty='none'`), en este caso, el modelo parece generalizar muy bien los datos no vistos.

Es importante notar que el conjunto de datos Iris es relativamente simple

➤ Regresión Logística con Stochastic Gradient Descent

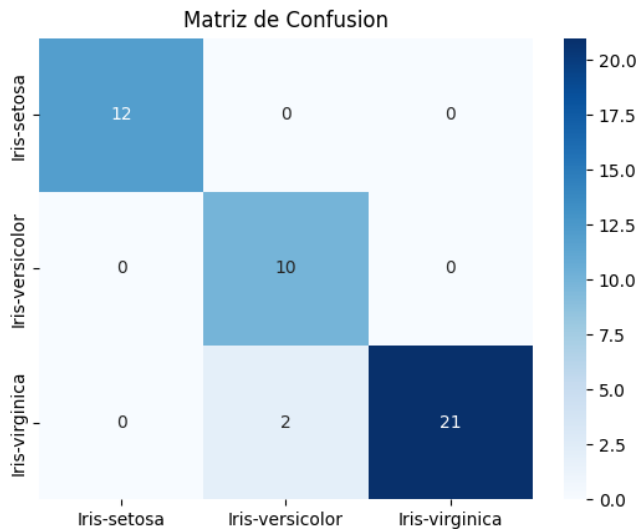


Figura 8. Matriz de Confusion de SAG

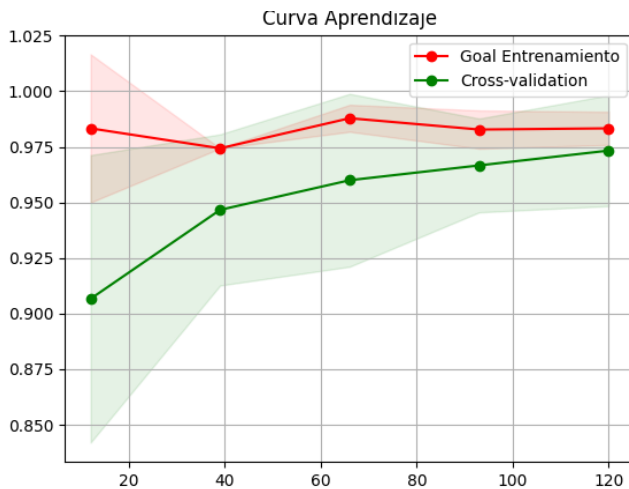


Figura 9. Curva de Aprendizaje de SAG

Accuracy (TRAIN)	0.981
Error (TRAIN)	0.019
Accuracy (TEST)	0.9556
Error (TEST)	0.0444

Figura 10. Tabla de resultados de SAG

La diferencia entre la precisión de entrenamiento y prueba no es muy grande, aunque es un poco más notable en comparación con el modelo anterior (Newton-Conjugate Gradient). El modelo sigue teniendo una exactitud alta en el conjunto de prueba

El modelo tiene un bajo sesgo y una varianza moderada. **Esto sugiere que el modelo está bien ajustado, pero quizás no tan perfectamente.** Por lo que podríamos considerar que existe un **poco de overfitting.**

VII. Conclusión

La Regresión Logística con SGD demostró ser una opción robusta y viable, aunque no logró el rendimiento perfecto del modelo basado en Newton-Conjugate Gradient. Sin embargo, con una adecuada optimización de hiperparámetros, este modelo podría mejorar su rendimiento.