# Instituto Tecnológico y de Estudios Superiores de Monterrey

## Campus Querétaro


## 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. Descripcion 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. Separación de los Datos:

El conjunto de datos fue dividido en dos subconjuntos: entrenamiento (70%) y prueba (30%).

## IV. Evaluación del Modelo:

*a. Desempeño en el Conjunto de Prueba:*

El modelo mostró una exactitud de ***model.score(X_test, y_test)*** y un error cuadrático medio de ***mean_squared_error(y_pred, y_test)***. Estas métricas indican la capacidad del modelo para predecir correctamente las clases del conjunto de datos 'Iris' en datos no vistos.

*b. Diagnóstico de Bias y Varianza:*

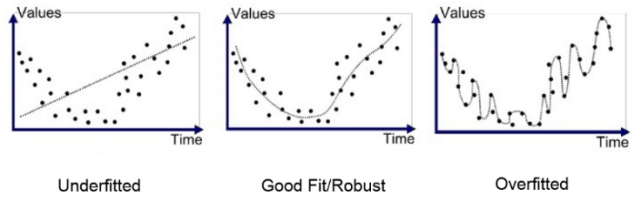
Observando la Curva de Aprendizaje, podemos diagnosticar el grado de bias y varianza:

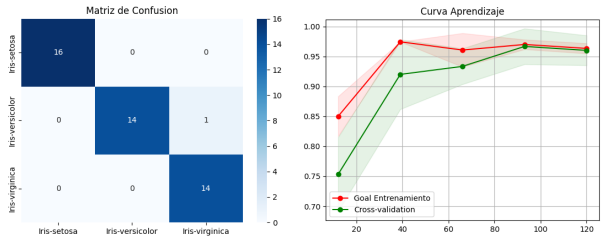
- Bias (Sesgo): Si la precisión en el conjunto de entrenamiento y prueba es baja, indica un alto bias.
- Varianza: Si hay una gran diferencia entre la precisión en el conjunto de entrenamiento y prueba, indica una alta varianza.

*c. Diagnóstico del Nivel de Ajuste 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 (Sobreajuste): Si el modelo tiene bajo bias y alta varianza.

A continuacion se muestran ejemplos de los diferentes ajustes que se podrian obtener.


- Exactitud en el conjunto de prueba: 0.9777
- Error del clasificador en el conjunto de prueba: 0.0222
- Error del clasificador en el conjunto de entrenamiento: 0.0285

## 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 continuacion con los 3 diferentes solver utilizados.

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.

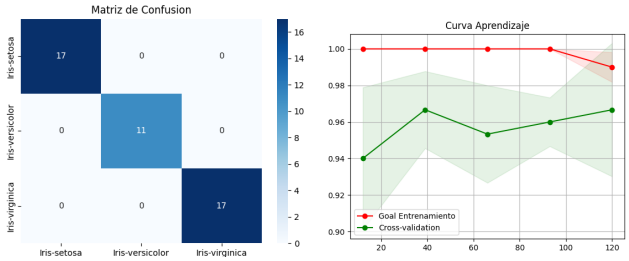
Dado que la exactitud es alta y el error es bajo en ambos conjuntos, el sesgo es bajo.

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.

## VI. ANALISIS:

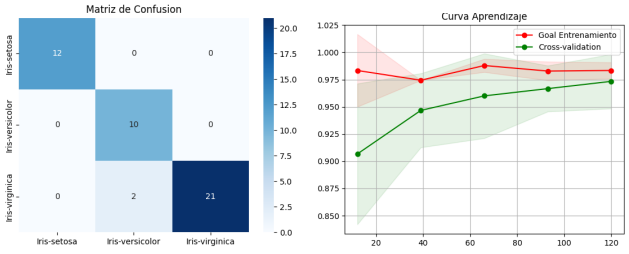
- ➢ Regresión Logística con Liblinear
- ➢ Regresión Logística Newton-Conjugate Gradient




- Exactitud en el conjunto de prueba: 1.0 (100%)
- Error del clasificador en el conjunto de prueba: 0.0
- Error del clasificador en el conjunto de entrenamiento: 0.019

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.

➢ Regresión Logística con Stochastic Gradient Descent




- Exactitud en el conjunto de prueba: 0.9556
- Error del clasificador en el conjunto de prueba: 0.0444
- Error del clasificador en el conjunto de entrenamiento: 0.019

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.