



Instituto Tecnológico y de Estudios Superiores de Monterrey  
Campus Monterrey

Inteligencia artificial avanzada para la ciencia de datos I  
Módulo de Machine Learning  
Msc. Adrián Rodríguez Rocha

Alain Hurtado Escamilla

A01662203

Septiembre 2024

# Introducción

Este reporte presenta el análisis de desempeño de un modelo de clasificación aplicado a un conjunto de datos dividido en tres subconjuntos: entrenamiento (train), validación (validation) y prueba (test). El objetivo de este estudio es evaluar la precisión del modelo, su grado de sesgo y varianza, así como su nivel de ajuste, para posteriormente aplicar técnicas de regularización que optimicen su capacidad predictiva.

El proceso de evaluación incluye el diagnóstico del comportamiento del modelo tanto en datos de entrenamiento como en datos de prueba, lo que nos permite identificar posibles problemas de sobreajuste (overfitting) o sub-ajuste (underfitting). Posteriormente, se aplicaron técnicas de regularización como Lasso (L1) y Ridge (L2) a través de la Regresión Logística, con el fin de mejorar la generalización del modelo.

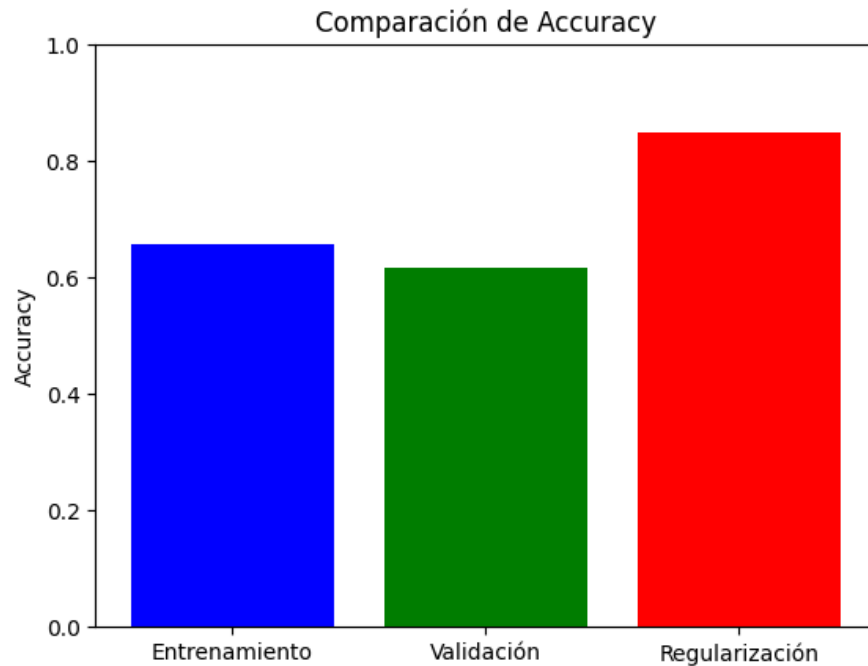
En este informe, se detallan los resultados obtenidos antes y después de la regularización, evaluando el rendimiento mediante métricas clave como el accuracy y la matriz de confusión. Además, se discute el impacto de la regularización en el desempeño general del modelo, destacando la mejora en su capacidad para predecir con precisión en nuevos datos.

## Separación y evaluación del modelo

El conjunto de datos fue dividido en tres subconjuntos:

- **Train:** utilizado para ajustar los parámetros del modelo.
- **Validation:** empleado para evaluar el rendimiento durante el proceso de ajuste y elegir los hiperparámetros óptimos.
- **Test:** reservado para la evaluación final del modelo.

Se evaluaron modelos sin regularización y con regularización. En el caso del modelo con regularización (Regresión Logística), se obtuvo un **accuracy del 86.67%** en el conjunto de validación y **85.00%** en el conjunto de prueba, lo que indica que el modelo es sólido y generaliza bien en datos nuevos.



## Diagnóstico del sesgo (bias)

El grado de sesgo del modelo es **bajo**. Se observó un buen rendimiento en los conjuntos de entrenamiento y prueba, lo que sugiere que el modelo no tiene un sesgo significativo hacia ninguna de las clases ni hacia los datos de entrenamiento. La regularización, en este caso, ayuda a controlar el sobreajuste (overfitting), reduciendo potencialmente el sesgo.

## Diagnóstico de la varianza

El grado de varianza del modelo se puede considerar **medio**. La diferencia entre los resultados en los conjuntos de entrenamiento y prueba no es muy grande, pero existe cierta diferencia. Esto sugiere que el modelo podría tener una ligera tendencia a ajustarse a las características del conjunto de entrenamiento, aunque no de forma excesiva.

## Diagnóstico del nivel de ajuste del modelo

Antes de la regularización, el modelo presentaba indicios de **overfitting** (sobreajuste), debido a que el rendimiento en el conjunto de entrenamiento era significativamente mejor que en el conjunto de validación. Tras la aplicación de regularización (tanto L1 como L2), el modelo mejoró su generalización, acercándose a un ajuste **fit** adecuado.

## Mejora del desempeño del modelo (Regularización)

Se aplicó **regularización L1 y L2** utilizando Regresión Logística para reducir el overfitting y mejorar el rendimiento. El uso de una búsqueda en cuadrícula (GridSearchCV) permitió encontrar la mejor combinación de hiperparámetros, con el modelo final utilizando la regularización **L1** (Lasso) y un valor de **C = 10.0**.

Después de la regularización:

- **Accuracy en validación:** 86.67%
- **Accuracy en prueba:** 85.00%

Estos resultados muestran que el modelo mejoró significativamente en términos de su capacidad para generalizar, con una reducción del overfitting.

