Tecnológico de Monterrey

Momento de Retroalimentación: Módulo 2

Análisis y Reporte del modelo Random Forest

Materia:

Inteligencia artificial avanzada para la ciencia de datos I

(Gpo 101)

Alumno:

Alfredo Azamar López - A01798100

Profesor:

Jorge Adolfo Ramírez Uresti

Tec de Monterrey Septiembre 2024

INTRO

Dataset

Nota:
Para la entrega pasada el código utiliza un dataset relacionado al tema de cáncer de mama (la base de datos se extrae de una librería de *sklearn*), por motivos del análisis del desempeño se cambia el dataset al de "Spaceship Titanic" (obtenido de Kaggle)

Este nuevo dataset tiene información relacionada a pasajeros en un entorno espacial ficticio, el cual incluye datos como su planeta destino, edad, gastos en varios servicios, si son clientes VIP, etc. La variable a predecir es si el pasajero fue transportado o no, esto lo convierte en un problema de clasificación binaria, el cual se adapta perfectamente a nuestro modelo de Random Forest debido a que crean múltiples árboles de decisión y se utiliza el voto mayoritario de estos para tomar una decisión final obteniendo un resultado robusto y preciso.

Las propiedades de los valores dentro del dataset son diversas ya que se incluyen variables numéricas como categóricas que influyen en la clasificación, Random Forest aprovecha esta característica y se adapta a la combinación de estas variables, siendo capaz de capturar las relaciones complejas entre las variables sin necesidad de que se especifiquen explícitamente las interacciones.

Separación y evaluación del modelo

La separación de datos es una etapa crucial en la construcción de un modelo ya que nos ayuda a checar la capacidad de generalización que tiene un modelo. Para este problema se utilizará la función `train_test_split()` de `sklearn` que utiliza el parámetro `test_size=0.2`, este nos indica que el 80% de los datos se utilizan para entrenar el modelo (conjunto de entrenamiento), y el 20% se utiliza para evaluarlo (conjunto de prueba). Para la evaluación del modelo se utilizaron las métricas de exactitud, reporte de clasificación, matriz de confusión y curvas de aprendizajes.

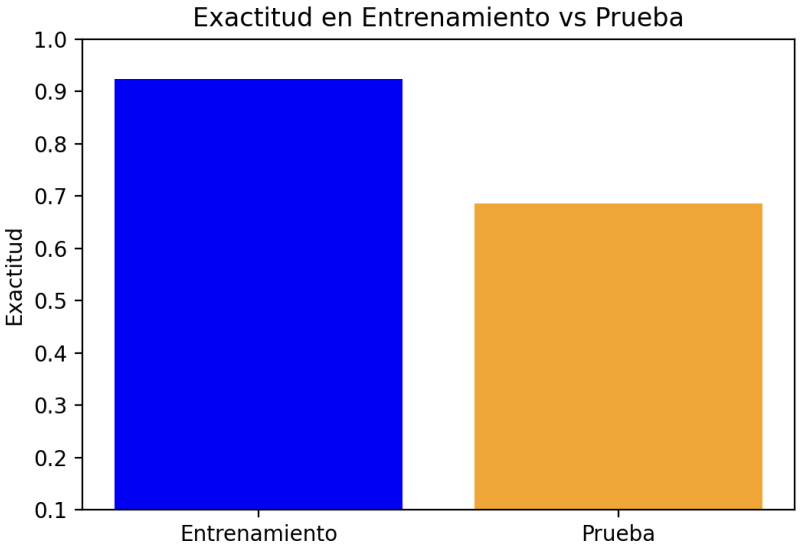

Imagen 1. Gráfica sobre la exactitud del dataset (dividido en entrenamiento y prueba).

En la imagen 1 podemos observar la exactitud obtenida para el conjunto de entrenamiento que representa un 90% (barra azul) y en el conjunto de pruebas la exactitud es un poco baja, aproximadamente un 70% (barra naranja). La diferencia entre las barras nos indica que el modelo tiene un rendimiento significativamente mejor en la parte de entrenamiento, esto puede indicar señales que nuestro modelo está realizando un sobreajuste (overfitting).

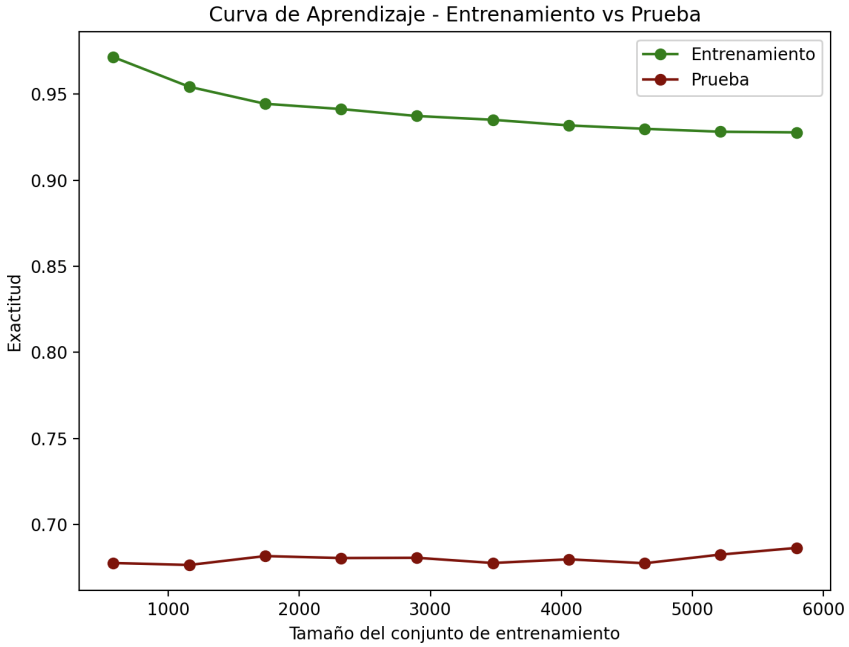

Imagen 2. Gráfica sobre la curva de aprendizaje del entrenamiento y prueba del dataset.

La segunda imagen nos muestra la evolución de la exactitud. Para el tamaño del conjunto de entrenamiento (línea verde), al inicio se puede observar que con pocos datos, la exactitud es muy alta (un 95%), pero conforme aumenta el tamaño del conjunto la exactitud baja ligeramente alcanzando un 90%. Para los datos de prueba (línea roja) la exactitud se mantiene constante alrededor de 65-70% y se muestra independiente al conjunto de entrenamiento.

Diagnóstico del grado de Bias

MEDIO

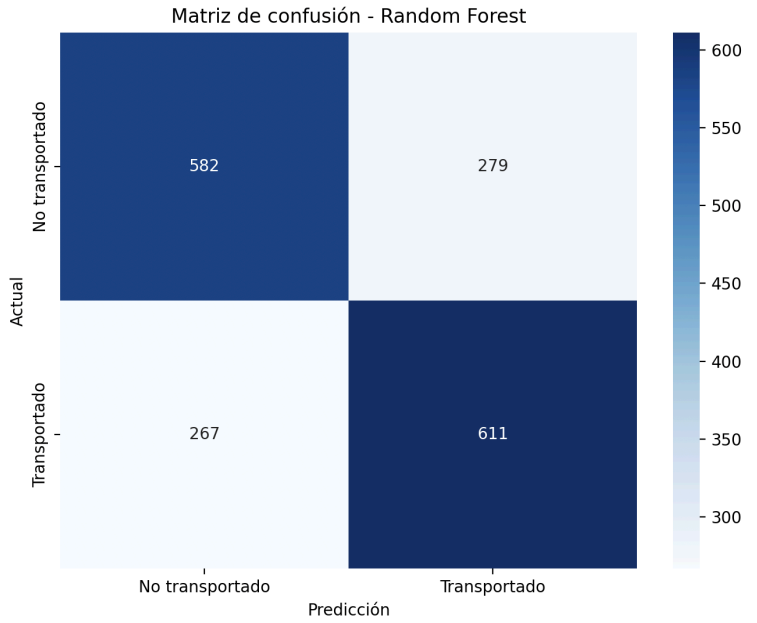

Imagen 3. Matriz de confusión sobre el conjunto de pruebas.

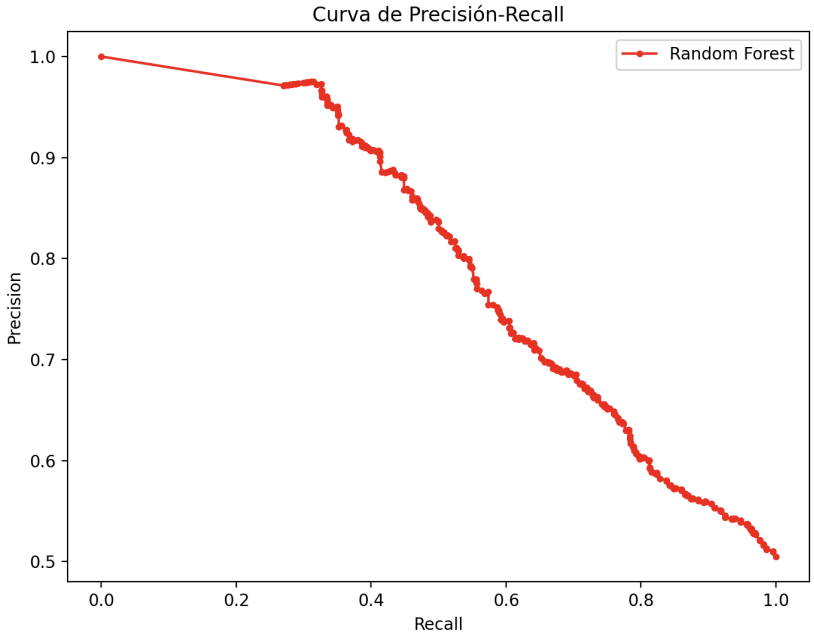

Imagen 4. Gráfica sobre la curva de precisión-recall.

Diagnóstico del grado de Varianza

BAJO

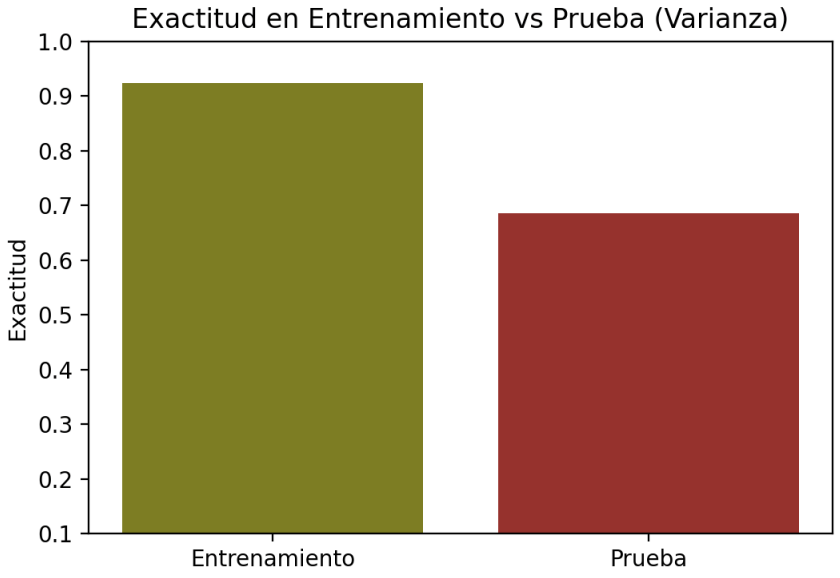

Imagen 5. Gráfica sobre la exactitud del dataset (dividido en entrenamiento y prueba).

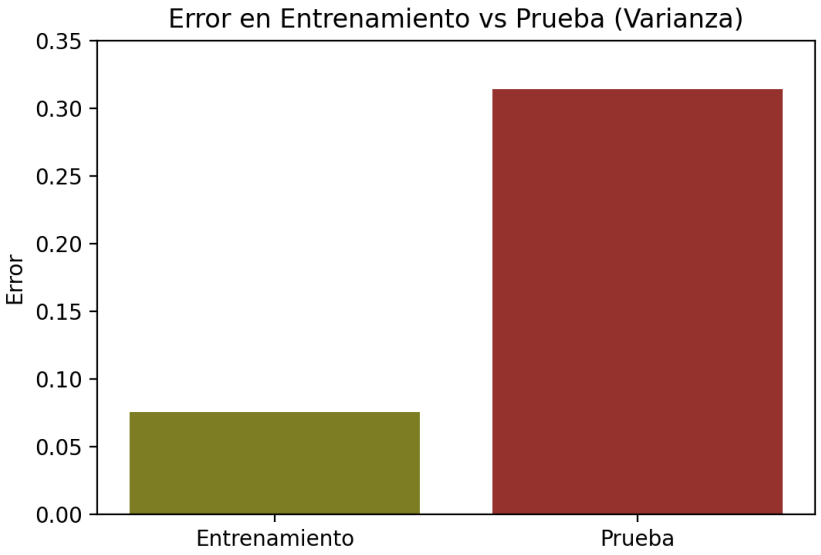

Imagen 6. Gráfica sobre el error obtenido en entrenamiento y prueba.

Diagnóstico del nivel de ajuste del modelo

FIT

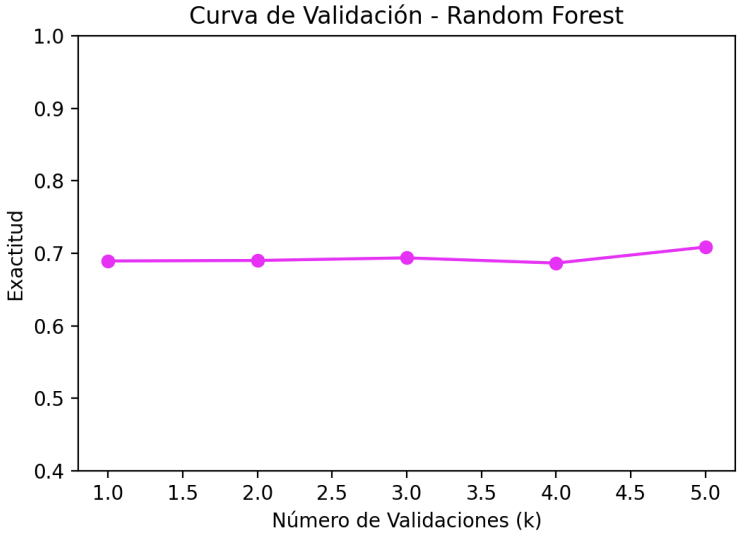

Imagen 7. Gráfica sobre la curva de validación

.

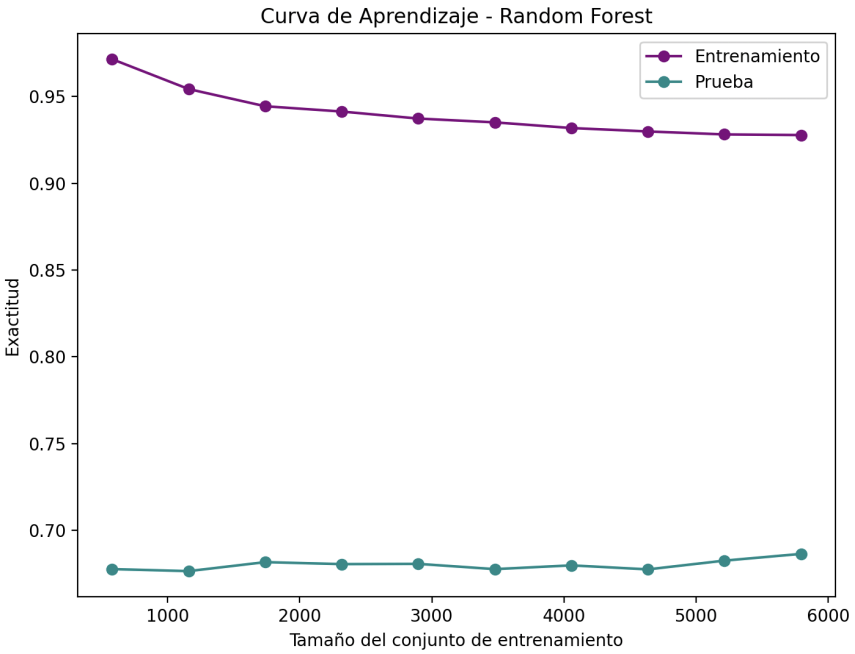

Imagen 8. Gráfica sobre la curva de aprendizaje.

Técnicas de regularización

Referencias

Kaggle.com. (2024). *Spaceship Titanic*. Recuperado de: https://www.kaggle.com/competitions/spaceship-titanic

Inesdi. (2023). *Random forest, la gran técnica de Machine Learning.* Recuperado de: *https://www.inesdi.com/blog/random-forest-que-es/*

Aprende Machine Learning. (2019). *Random Forest, el poder del Ensamble.* Recuperado de: https://www.aprendemachinelearning.com/random-forest-el-poder-del-ensamble/

KeepCoding. (2022). *La división de datos en Deep Learning.* Recuperado de: https://keepcoding.io/blog/division-datos-deep-learning/

datos.gob.es. (2023). *Cómo preparar un conjunto de datos para machine learning y análisis.* Recuperado de: https://datos.gob.es/es/blog/como-preparar-un-conjunto-de-datos-para-machine-learning-y-analisis