



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

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.

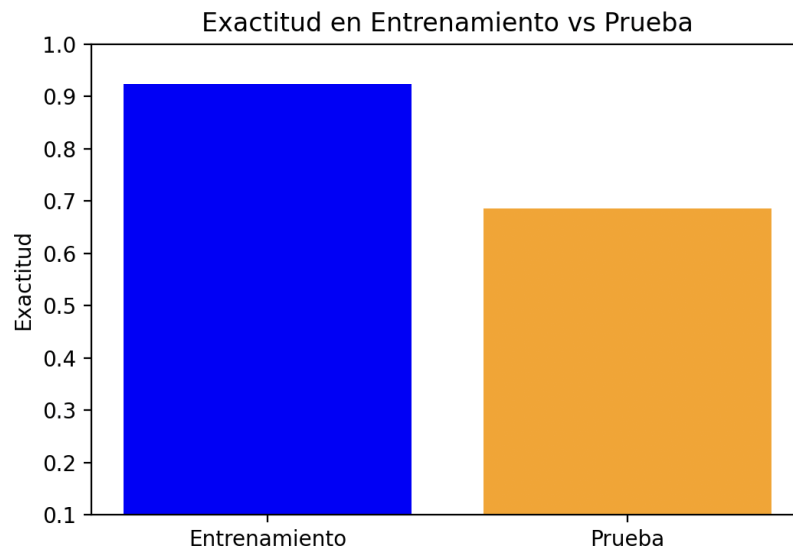


Imagen 1. Gráfica sobre la exactitud del dataset (dividido en entrenamiento y prueba).

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).

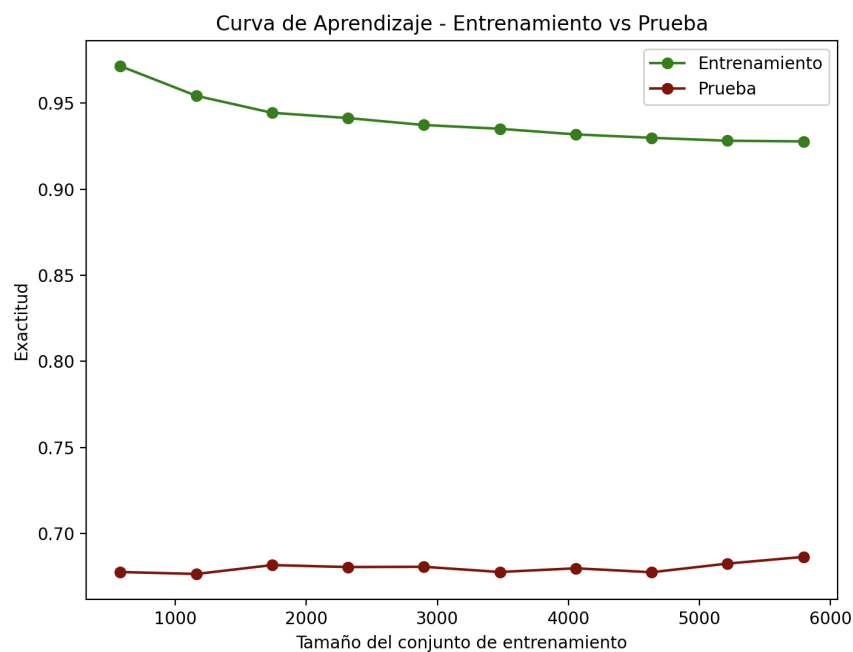


Imagen 2. Gráfica sobre la curva de aprendizaje del entrenamiento y prueba del dataset.

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

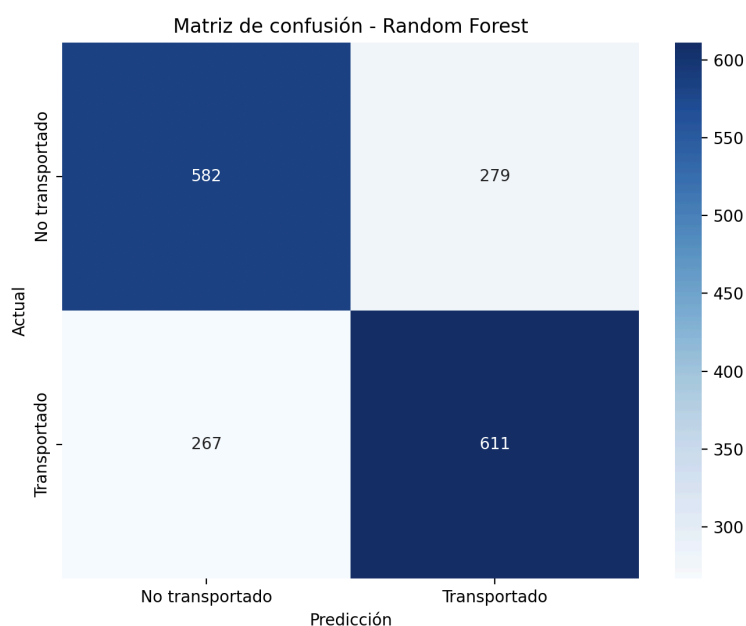


Imagen 3. Matriz de confusión sobre el conjunto de pruebas.

La matriz de confusión (*imagen 3*) nos presenta que el modelo predijo correctamente 582 casos como entre “No transportado” y 611 como “Transportado”. Podemos observar que el modelo presenta un número considerable de falsos negativos y falsos positivos, lo que indica que, el modelo es capaz de predecir correctamente muchas instancias pero comete errores significativos.

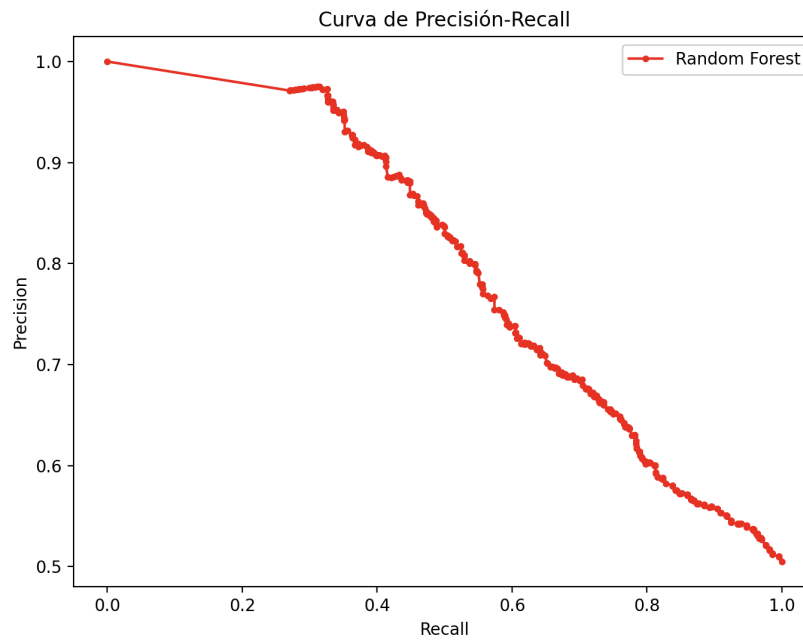


Imagen 4. Gráfica sobre la curva de precisión-recall.

Para nuestra gráfica de curva de precisión-recall (*imagen 4*), se observa que el modelo en un principio tiene una alta precisión (bajo recall), pero conforme el recall aumenta, la precisión empieza a disminuir, llegando a un valor aproximado de 0.5. Este comportamiento refleja un *trade-off* entre la precisión y el recall, donde optimizar uno afecta negativamente al otro.

Se podría decir que el nivel de vías respecto a nuestro modelo es de nivel medio. Ya que en la *imagen 3*, los falsos negativos y falsos positivos sugieren que el modelo tiende a clasificar erróneamente algunas instancias; está sesgo también se observa en la *imagen 4*, donde vemos que la precisión cae rápidamente conforme aumenta el recall. Esto indica que el modelo no tiene la capacidad suficiente para distinguir correctamente entre las clases cuando predice una cantidad mayor de datos positivos.

Diagnóstico del grado de Varianza

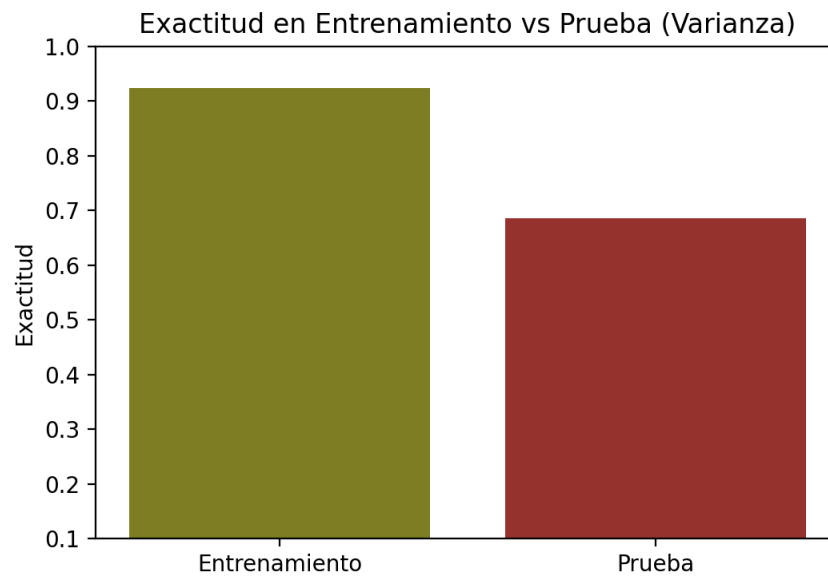


Imagen 5. Gráfica sobre la exactitud del dataset (dividido en entrenamiento y prueba).

En la *imagen 5* se compara la exactitud del modelo para el conjunto de entrenamiento y de prueba, esta métrica mide el porcentaje de predicciones correctas realizadas por el modelo. Se puede observar que el valor de la exactitud en el entrenamiento es alta (indica que el modelo predice correctamente las instancias), pero cuando lo comparamos con el valor de la exactitud que es cercano a 0.65 muestra que el modelo tiene un rendimiento menor; esto puede implicar que modelos se está ajustando a los datos.

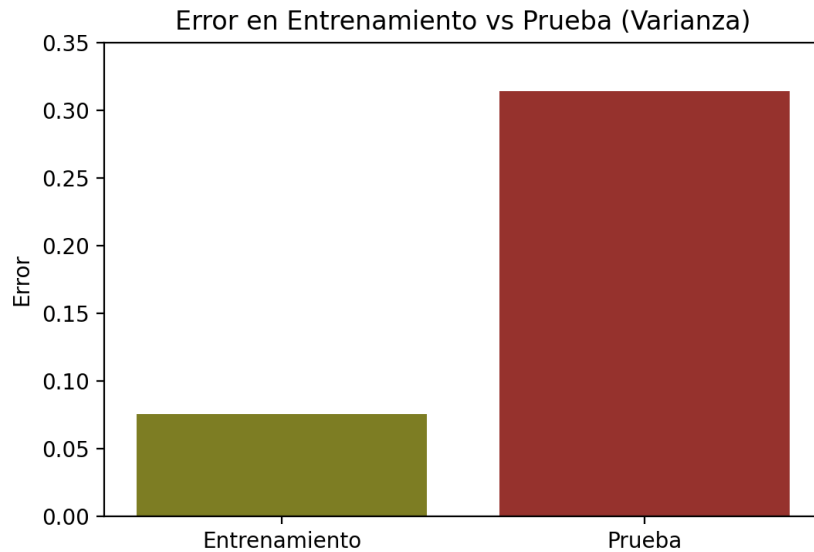


Imagen 6. Gráfica sobre el error obtenido en entrenamiento y prueba.

El objetivo de esta gráfica es representar el error cometido por el modelo, este es complementario a la exactitud. Para el error de entrenamiento podemos observar que es muy bajo (valor de 0.05), sin embargo, el error aumenta significativamente en el conjunto de prueba (llegando a valores de 0.3). Esto nos indica que el modelo no es capaz de predecir correctamente muchas instancias.

Ambas gráficas nos demuestran que el modelo sufre de una varianza alta, la diferencia observada en ambas gráficas entre las métricas de entrenamiento y pruebas sugiere que el modelo es muy sensible a los datos de entrenamiento, lo que lleva a una falta de capacidad de generalización.

Un error bajo en el conjunto de entrenamiento combinado con un error alto en el conjunto de prueba es característico de un modelo que ha memorizado los datos de entrenamiento, pero que tiene dificultades para generalizar.

Diagnóstico del nivel de ajuste del modelo

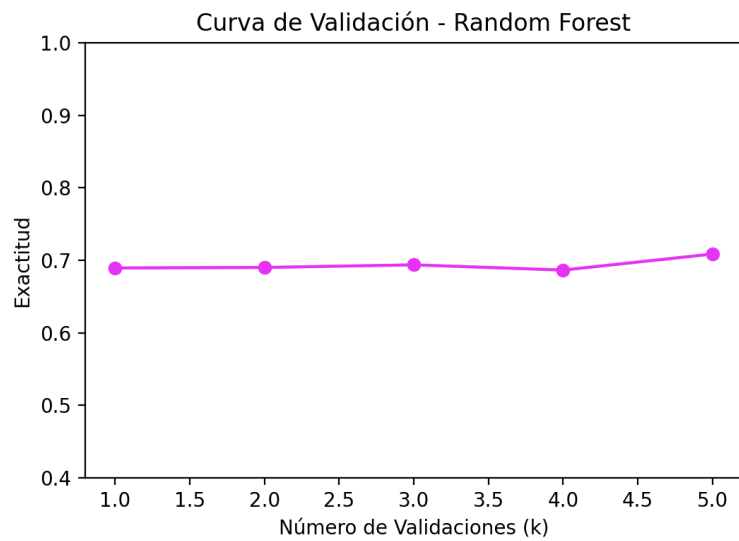


Imagen 7. Gráfica sobre la curva de validación

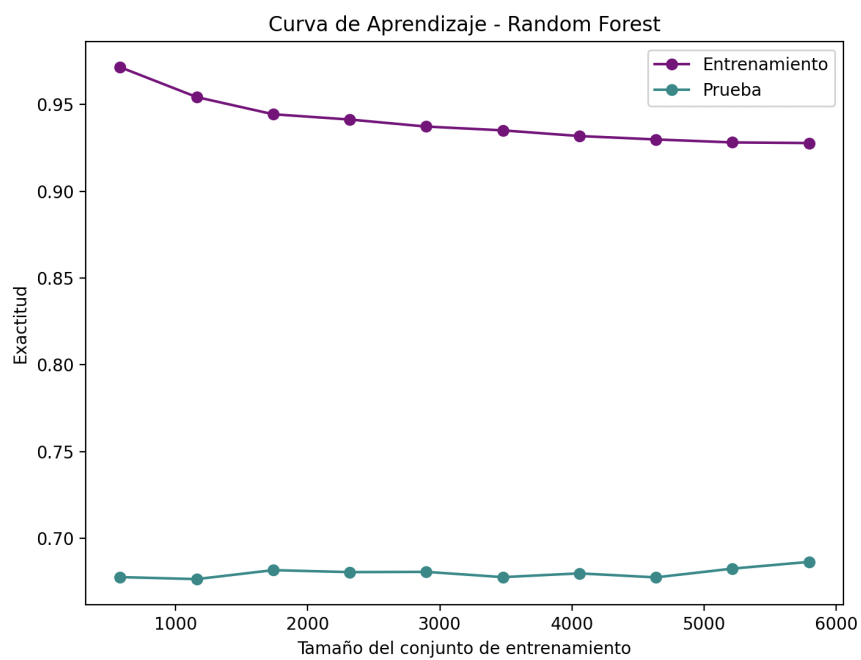


Imagen 8. Gráfica sobre la curva de aprendizaje.

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 Ensemble*. 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>

Analitika. (2023). *Accuracy vs Recall: Definiendo la Calidad de los Algoritmos con ROC y Métricas*. Recuperado de: <https://analitikacentroamerica.com/accuracy-vs-recall-definiendo-la-calidad-de-los-algoritmos-con-roc-y-metricas/>

Amat, J. (2020). *Árboles de predicción: bagging, random forest, boosting y C5.0*. Recuperado de: https://rpubs.com/Joaquin_AR/255596