Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Equipo 4

Elementos Reporte TC3006C.pdf

Integrantes:

Eliezer Cavazos Rochin A00835194
Facundo Colasurdo Caldironi A01198015
Saul Francisco Vázquez del Río A01198261
José Carlos Sänchez Gómez A01174050

Campus Monterrey **Fecha: 9/9/2024**

Índice

Introducción

El desafío "Titanic - Machine Learning from Disaster" es un problema el cual estamos tratando de superar, esta, busca predecir la supervivencia de los pasajeros a bordo del Titanic. Este desafío implica construir un modelo de clasificación binaria capaz de predecir si un pasajero específico sobrevivió al hundimiento del Titanic, utilizando un conjunto de datos que incluye diversas características de los pasajeros, como su edad, sexo, clase en el barco, y otras.

El problema se adentra en el aprendizaje supervisado, donde la meta es, entrenar un modelo a partir de datos etiquetados para que pueda hacer predicciones sobre datos no vistos previamente. Dado que el objetivo es predecir una de dos posibles categorías de supervivencia de los pasajeros, se requiere de un modelo de clasificación binaria.

Análisis exploratorio

Para abordar este problema, es crucial comprender la naturaleza de los datos disponibles, los cuales se dividen en variables categóricas y numéricas.

Estas variables incluyen información crítica como el sexo del pasajero, su edad, la clase en la que viajaba, el número de familiares a bordo, entre otras, tal como se puede ver en la siguiente tabla:

Variable	Definición	Tipo de Datos	Modelos Compatibles
survival	Supervivencia	Categórico (Binario)	Regresión logística, árboles de decisión, SVM, redes neuronales, KNN, Naive Bayes, Random Forest
pclass	Clase de ticket	Categórico (Ordinal)	Regresión logística, árboles de decisión, Random Forest, SVM, KNN
sex	Sexo	Categórico (Nominal)	Regresión logística, árboles de decisión, SVM, redes neuronales, KNN, Naive Bayes, Random Forest
age	Edad en años	Numérico (Continua)	Regresión lineal, regresión logística, árboles de decisión, SVM, redes neuronales, KNN, Random Forest
subsp	Número de	Numérico	Regresión lineal, regresión logística, árboles de

	hermanos/esposos a bordo	(Discreta)	decisión, SVM, redes neuronales, KNN, Random Forest
parch	Número de padres/hijos a bordo	Numérico (Discreta)	Regresión lineal, regresión logística, árboles de decisión, SVM, redes neuronales, KNN, Random Forest
ticket	Número de ticket	Categórico (Nominal)	Generalmente no se utiliza como predictor directo en modelos, se puede transformar o excluir
fare	Tarifa del pasajero	Numérico (Continua)	Regresión lineal, regresión logística, árboles de decisión, SVM, redes neuronales, KNN, Random Forest
cabin	Número de cabina	Categórico (Nominal)	Generalmente no se utiliza como predictor directo en modelos, se puede transformar o excluir
embarked	Puerto de embarque	Categórico (Nominal)	Regresión logística, árboles de decisión, SVM, redes neuronales, KNN, Naive Bayes, Random Forest

Objetivo

Una vez comprendido lo anterior, es necesario dejar en claro cuáles fueron los principales objetivos que tuvimos que tomar en cuenta en el desarrollo de nuestro proyecto:

El primero fue garantizar la limpieza y preparación de los datos. Lo cual no solo implicó la eliminación de espacios vacíos o innecesarios, también, la normalización de los datos pertinentes para que puedan ser útiles en el contexto de nuestro problema, logrando así que los modelos puedan operar con información de alta calidad, lo cual es vital para obtener datos precisos y confiables.

A continuación, fue la implementación de cinco algoritmos de clasificación, siendo estos; Regresión Logística, Regresión Lineal, Bosque Aleatorio, Árbol de Decisión y Redes Neuronales.

Cada uno de estos algoritmos fueron seleccionados debido a que permiten manejar distintos análisis de nuestros datos, con los cuales fue posible evaluar los resultados obtenidos por cada uno, para poder obtener un modelo más robusto, el cual permita generar predicciones más precisas.

Una vez que se obtuvieron los resultados iniciales con los distintos algoritmos, fue de vital importancia mejorar la optimización de los mismos, ajustando los parámetros dentro de los mismos para mejorar el rendimiento del modelo, lo cual en consiguiente permitió obtener una mayor precisión en las predicciones y un mejor resultado.

Finalmente, uno de los puntos más importantes fue la visualización de los resultados obtenidos, ya que no solo permite interpretar los resultados de manera intuitiva, sino también, identificar las características más relevantes las cuales afectan la supervivencia y la efectividad de los modelos.

Limpieza datos

Una vez comprendido lo anterior, fue posible avanzar hacia la ejecución del proyecto. El primer paso consistió en la obtención de datos relevantes, que incluían información sobre los pasajeros, como Edad, Clase y Sexo, elementos esenciales para predecir la supervivencia en el escenario planteado. Estos datos fueron extraídos con el objetivo de construir un modelo de clasificación que pudiera responder a la pregunta de si un pasajero sobrevivió o no.

Para mejorar la imputación de estos valores faltantes, se implementaron tres métodos:

Primero, se probaron varias regresiones lineales utilizando diferentes campos (como Clase, Número de hermanos/esposos y Número de padres/hijos) para identificar el mejor predictor de la Edad.

En segundo lugar, se desarrolló un método que asignaba una edad estimada a los pasajeros utilizando otros campos (PClass, SubSp, Parch).

Finalmente, también se consideró la imputación de los valores faltantes utilizando el promedio de edad.

Tras evaluar estos métodos, se concluyó que el último metodo proporcionaba una mejora significativa en la predicción, por lo que se implementó en la versión final.

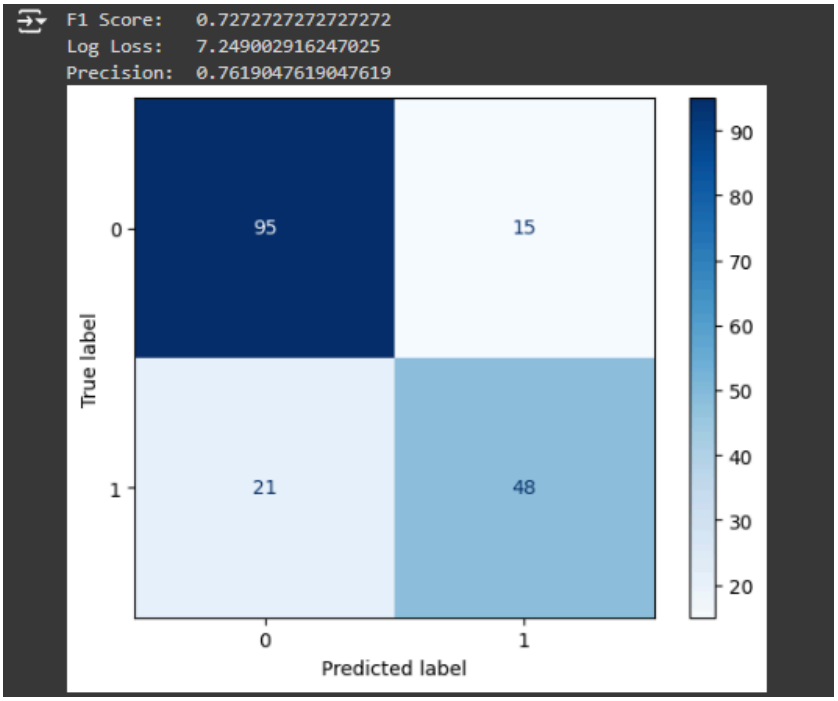
De igual manera, nuestros datos cuentan con la información de dónde fue que abordaron el Titanic, mediante el campo de Embarcación. Para poder hacer uso de esa información, decidimos emplear la técnica de One Hot Encoding, la cuál consiste en crear nuevas columnas con los valores originales de la columna de Embarcamiento, con el objetivo de marcar con 0s y 1s la categoría a la que pertenece el dato, esto lo usamos ya que identificamos que el sesgo era mayor cuando lo trabajamos en una sola columna.

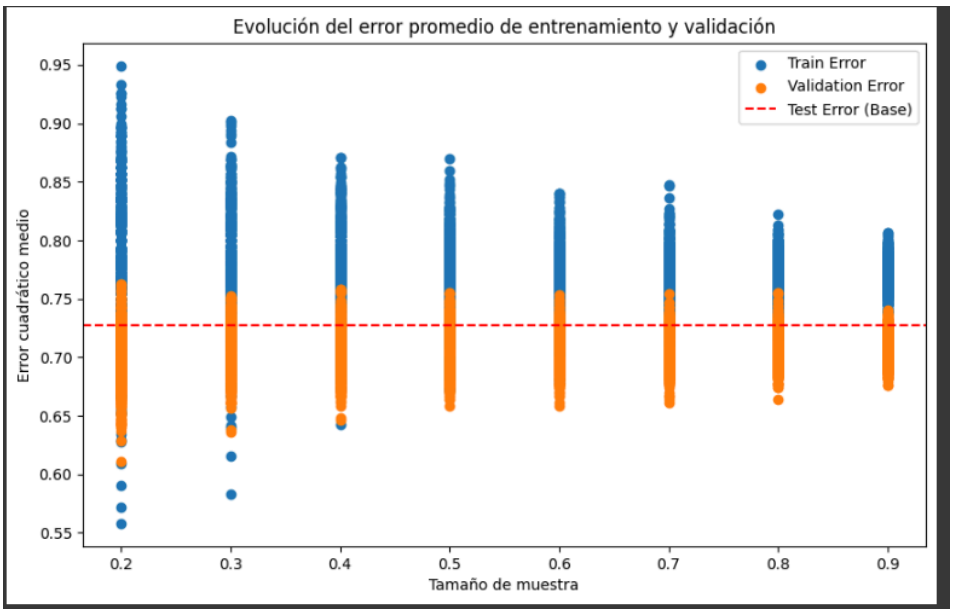
Modelos de predicción

Una vez que se tuvieron los datos filtrados y procesados, implementamos cinco métodos de modelamiento para predecir la supervivencia de los pasajeros:

Modelo de Regresión Logística

Utilizamos la regresión logística como primer modelo para clasificar la supervivencia. Este modelo fue elegido debido a la naturaleza binaria de la variable supervivencia. Los resultados iniciales mostraron una precisión aceptable, pero limitada en comparación con otros modelos más complejos.





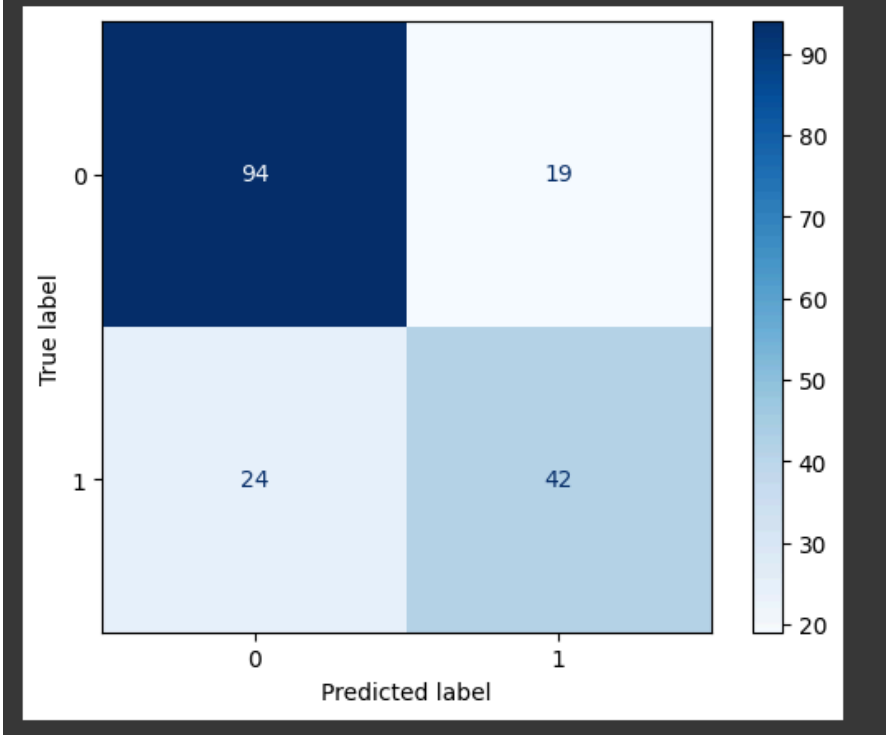
Bosque Aleatorio

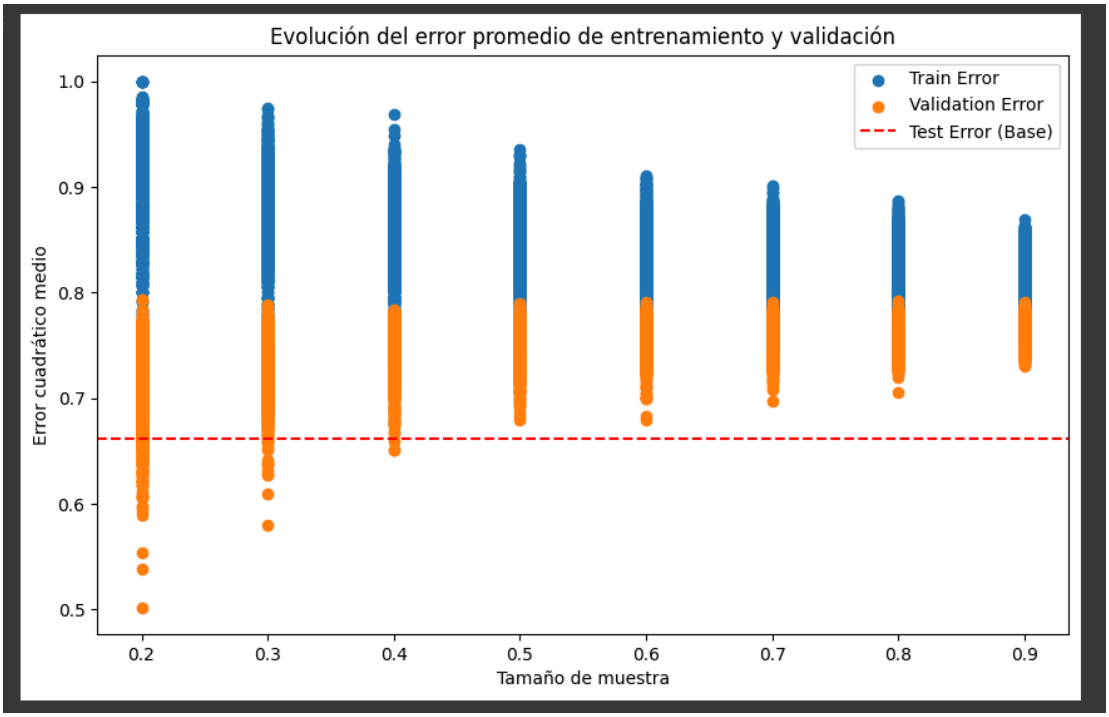
El tercer método que seleccionamos fue el modelo de Bosque Aleatorio resultó ser uno de los más efectivos. Este enfoque se centró en combinar múltiples árboles de decisión para mejorar la precisión y minimizar el sobreajuste. El modelo mostró un rendimiento superior y fue una de las mejores opciones para nuestro conjunto de datos.

Tras hacer varias iteraciones con los valores de los parámetros de entrenamiento de bosque aleatorio, obtuvimos un modelo, el cuál tiene un F1 Score de Entrenamiento de 0.95, y un F1 Score de Validación de 0.87. Los resultados obtenidos fueron obtenidos con estos parámetros:

Parámetro	Valor
max_depth_list	[4]
n_estimators_list	[150]

```
F1 Score:   0.6614173228346456
Log Loss:   8.658531261072836
Precision:  0.6885245901639344
```





best_forest_predictions_Kaggle (12).csv
Complete · now
0.78947

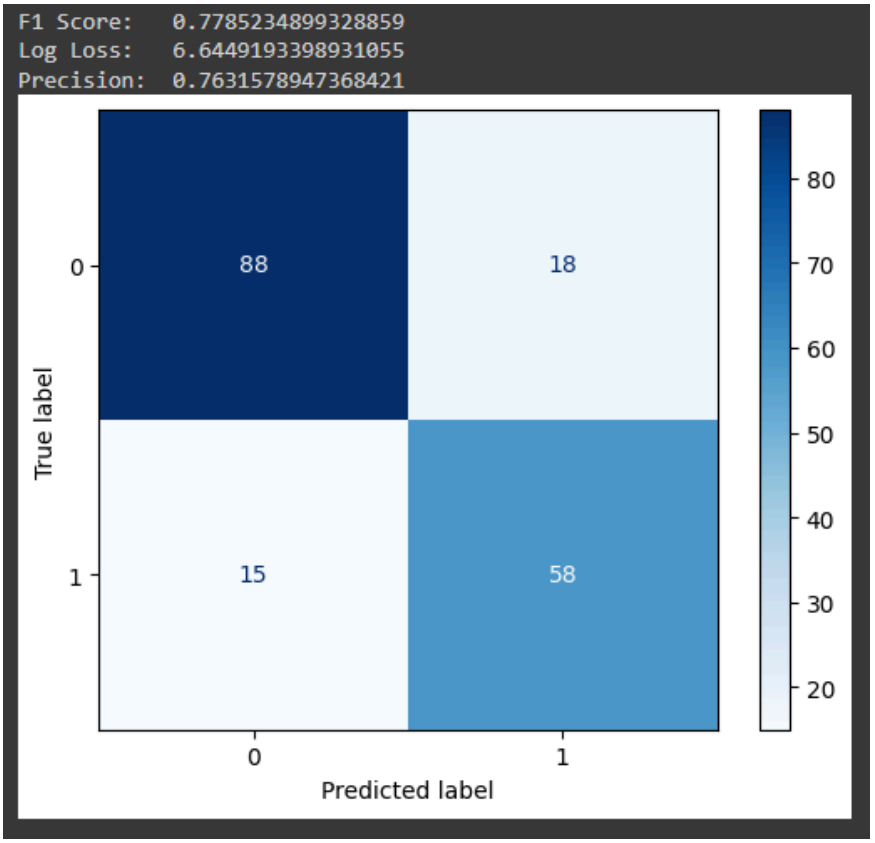
Árbol de Decisión

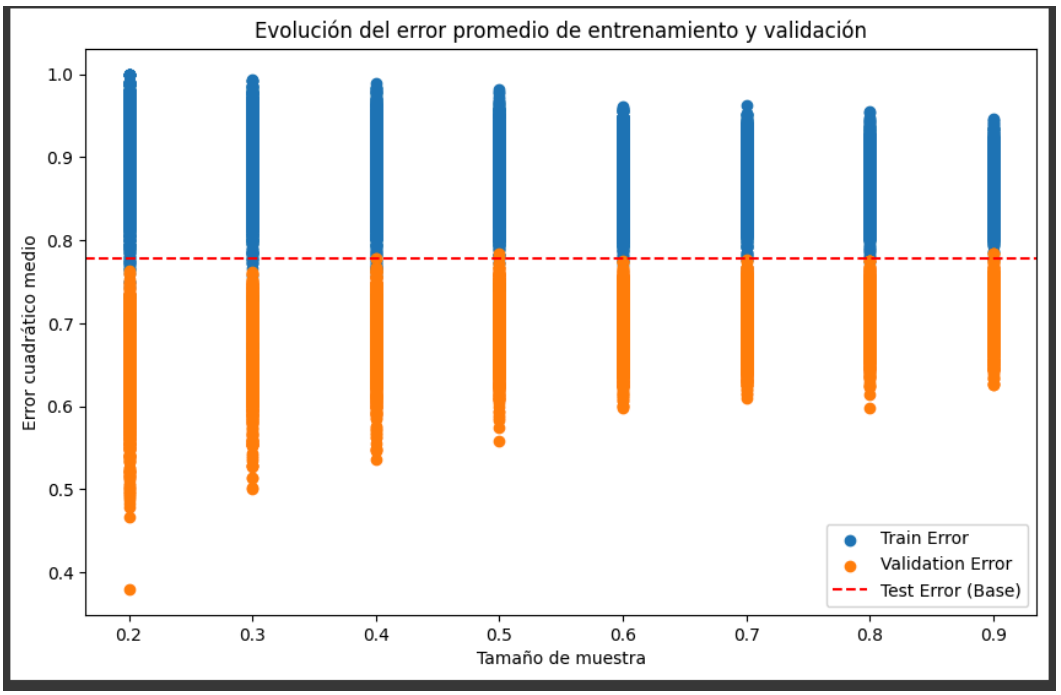
También probamos el modelo de Árbol de Decisión, el cual permitió visualizar de manera clara el proceso de toma de decisiones basado en las características de los pasajeros. Aunque el rendimiento del Árbol de Decisión no superó al del Bosque Aleatorio, presentó resultados aceptables.

Tras hacer varias iteraciones con los valores de los parámetros de entrenamiento del árbol de decisión, obtuvimos un modelo, el cuál tiene un F1 Score de Entrenamiento de 0.77, y un F1 Score de Validación de 0.86. Los resultados obtenidos fueron obtenidos con estos parámetros:

Parámetro	Valor
max_depth_list	[4]
min_samples_leaf_list	[1]

(Los demás parámetros de la función quedaron como default)

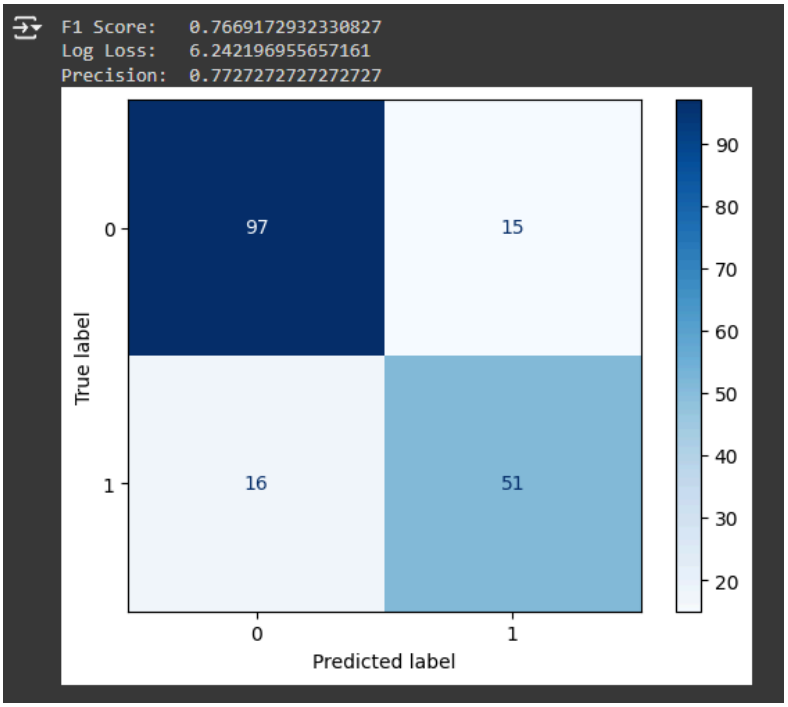




predictionAD.csv
Complete · now
0.77511

Redes Neuronales

También se probó el modelo de redes neuronales para probar si la predicción de supervivencia mejoraba, Aunque el rendimiento dejó mucho que desear, ya que fue el más lento de los 4 modelos, tampoco logró superar el resultado de Bosque Aleatorio.

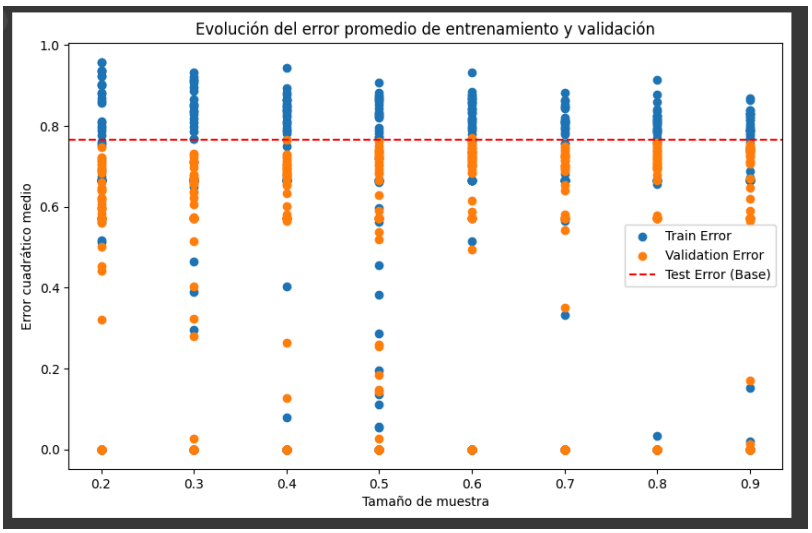


Tras hacer varias iteraciones con los valores de los parámetros de entrenamiento de la red neuronal, obtuvimos un modelo, el cuál tiene un F1 Score de Entrenamiento de 0.84, y un F1 Score de Validación de 0.77. Los resultados obtenidos fueron obtenidos con estos parámetros:

Parámetro	Valor
hidden_layer_sizes	(6, 2, 2, 1)
max_iter	2000
solver	lbfgs
activation	relu

(Los demás parámetros de la función quedaron como default)

Tras obtener el modelo hicimos las predicciones con el dataset de test.csv, sin embargo, los resultados obtenidos dejaron que desear, pues el score que se obtuvo en Kaggle, fue de 0.72.



Modelo Final

Después de evaluar los resultados iniciales, el modelo de Bosque Aleatorio se destacó como el más prometedor. Para mejorar aún más este modelo, se llevaron a cabo una serie de ajustes y pruebas adicionales. En primer lugar, se realizó una optimización de los parámetros, para encontrar la configuración óptima. Además, se empleó la validación cruzada para garantizar que el modelo generaliza bien a nuevos datos y no se sobre ajustará al conjunto de entrenamiento. También se llevaron a cabo técnicas de ingeniería de características, que consistieron en la creación de nuevas características o en la transformación de las existentes para mejorar el rendimiento del modelo.

Conclusión

El modelo de Bosque Aleatorio, tras las optimizaciones y pruebas adicionales, demostró ser el más efectivo. Los resultados finales indicaron un F1 Score de 0.8857 en Kaggle, una precisión de 0.9118 y un score total de 78.9. Estos resultados reflejan que el modelo de Bosque Aleatorio ofrece una alta precisión y una sólida capacidad de generalización, superando a otros modelos implementados y posicionándose como la mejor opción para la predicción de supervivencia en este conjunto de datos.

El modelo de Bosque Aleatorio ha demostrado ser la mejor opción para este conjunto de datos, combinando alta precisión y una sólida capacidad de generalización. Las pruebas adicionales y ajustes realizados han validado su superioridad frente a otros enfoques. Aunque las redes neuronales podrían ofrecer

mejoras adicionales, el Bosque Aleatorio actualmente representa la solución más robusta y efectiva para la predicción de supervivencia en este contexto.

Mejoras a futuro

La experiencia y los resultados obtenidos durante este proyecto subrayan la importancia de elegir y ajustar adecuadamente los modelos en función de las características específicas del problema y los datos disponibles.

Bibliografía:

Amazon (2023) ¿Qué es el modelado de datos?. Recuperado de https://aws.amazon.com/es/what-is/data-modeling/

Hillier, F. S., & Lieberman, G. J. (2006). Introducción a la investigación de operaciones (8.a ed.). McGraw-Hill.

Taha, H. A. (2012). Investigación de operaciones: una introducción (9.a ed.). Pearson Educación.

Winston, W. L. (2005). Investigación de operaciones (4.a ed.). Thomson.