Equipo 4

Titanic - Machine Learning from Disaster

Volver al programa

Introducción y Objetivos

Contexto

El desafío "Titanic - Machine Learning from Disaster" es un problema el cual estamos tratando de superar, la cual busca predecir la supervivencia de los pasajeros a bordo del Titanic.

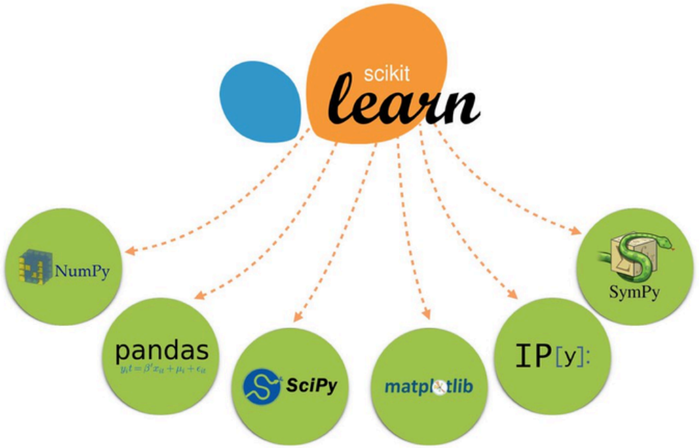
Problematica

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.

Recursos y herramientas

Google colab

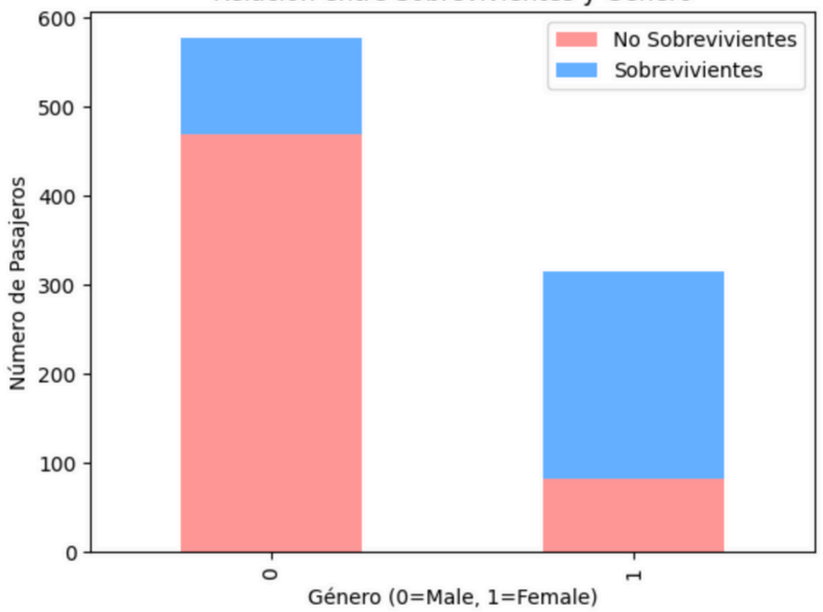
kaggle

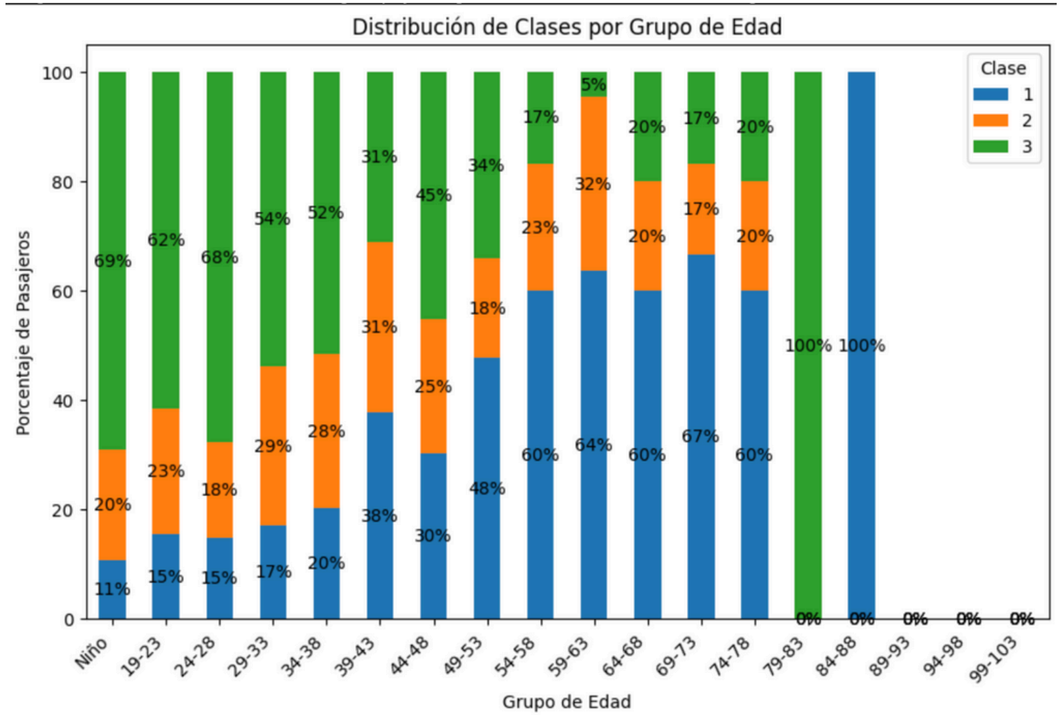


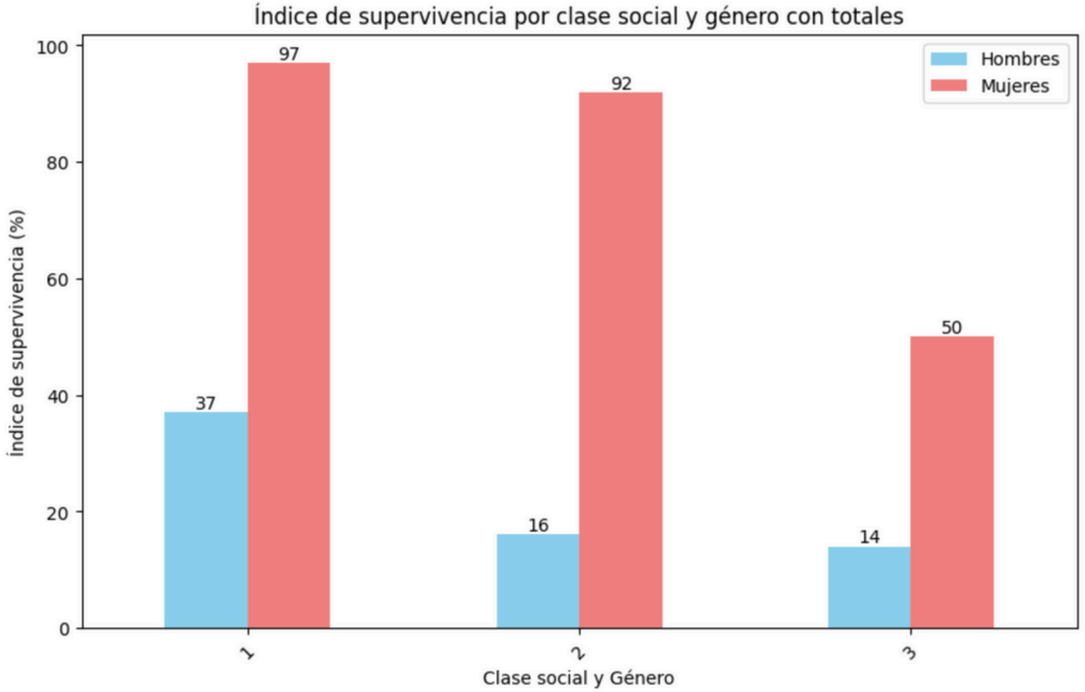
Base de Datos del Titanic disponible en Kaggle

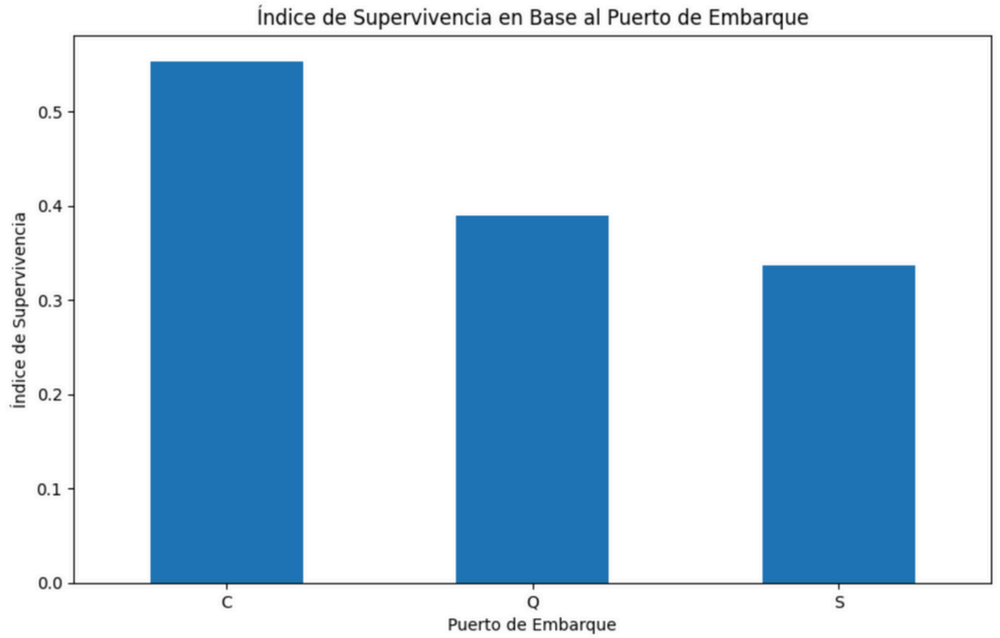
Volver al programa

Analisis de los datos


Relación entre Sobrevivientes y Género
No Sobrevivientes
Sobrevivientes
Número de Pasajeros
Género (0=Male, 1=Female)
0
100
200
300
400
500
600
0
1


Distribución de Clases por Grupo de Edad
Clase
1
2
3
Porcentaje de Pasajeros
Grupo de Edad
Niño
19-23
24-28
29-33
34-38
39-43
44-48
49-53
54-58
59-63
64-68
69-73
74-78
79-83
84-88
89-93
94-98
99-103


Índice de supervivencia por clase social y género con totales
Hombres
Mujeres
Índice de supervivencia (%)
Clase social y Género
37
97
16
92
14
50
1
2
3


Índice de Supervivencia en Base al Puerto de Embarque
Índice de Supervivencia
Puerto de Embarque
C
Q
S

Limpieza de datos

Durante el desarrollo del reto observaron que habían muchos datos faltantes y en blanco, por lo que se decidió hacer una limpieza de los mismos, para eliminar espacios en blanco, tales como “Cabin” y datos inecesarios como “Ticket” y “Fare”.

Para las columnas necesarias pero con datos en blanco tales como “Age” se decidio rellenar los datos usando vecinos y la columna de “embarked” decidimos llenar los datos faltantes con el valor de la embarcacion “S”.

Transformacion

Para mejorar la predicción decidimos transformar la columna de embarcación separándolo en 3 columnas, además de escalar la edad para que estuvieran en un rango más limitado de valores.

Survived	Pclass	Sex	Age	SibSp	Parch	C	Q	S
0	3	0	-0.891171	1	0	0.0	0.0	1.0
0	3	0	-0.386324	0	0	1.0	0.0	0.0
0	2	0	-0.450577	0	0	1.0	0.0	0.0
0	1	0	2.266420	0	0	0.0	0.0	1.0
0	2	0	0.650908	0	0	0.0	0.0	1.0

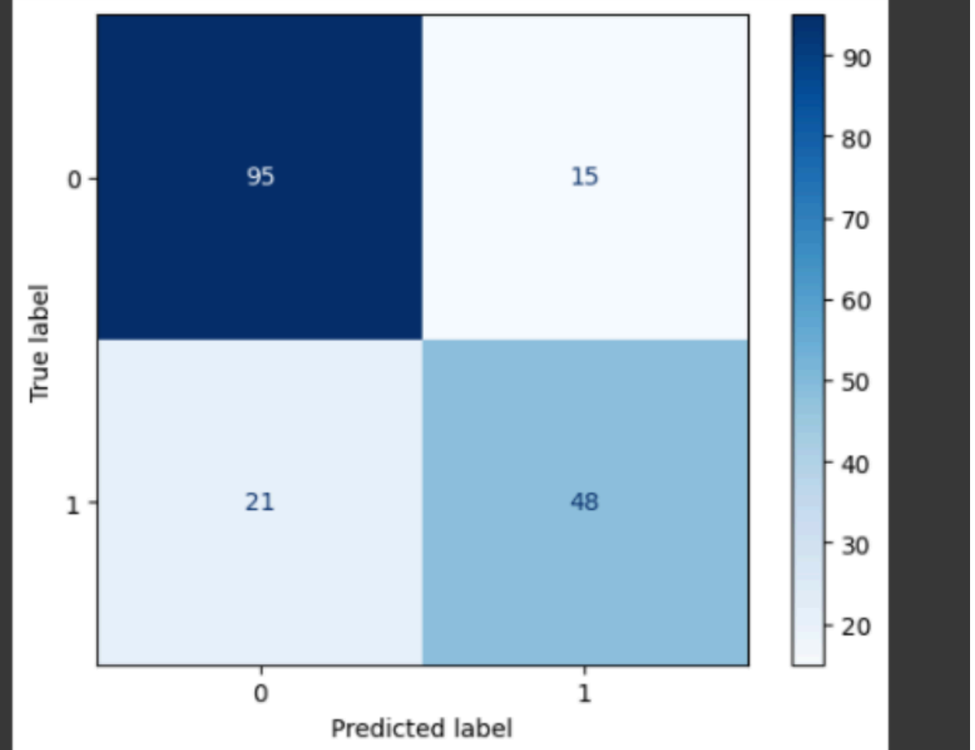
Volver al programa

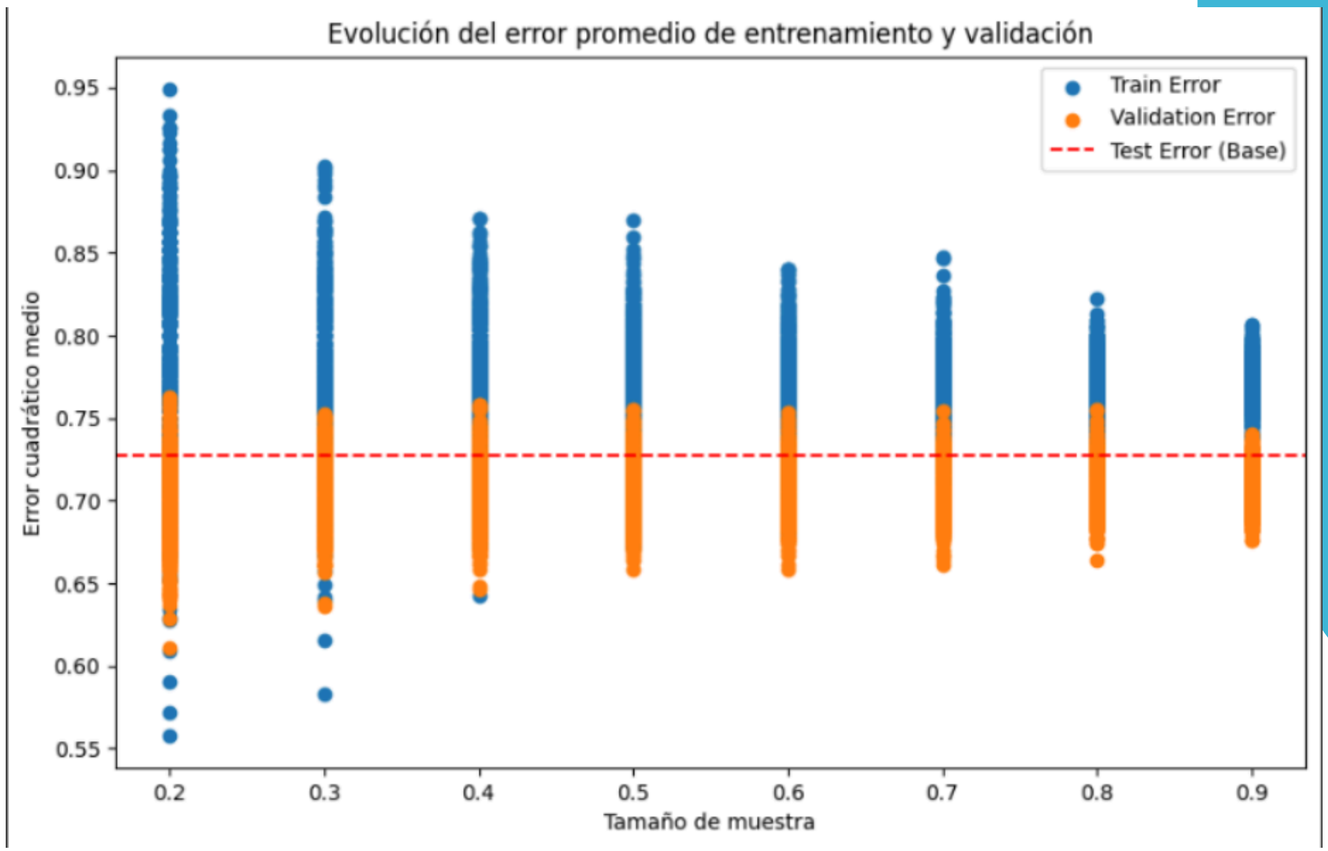
Entrenamiento de modelos

Regresión Logistica

Mejor calificación Kaggle: 77.9

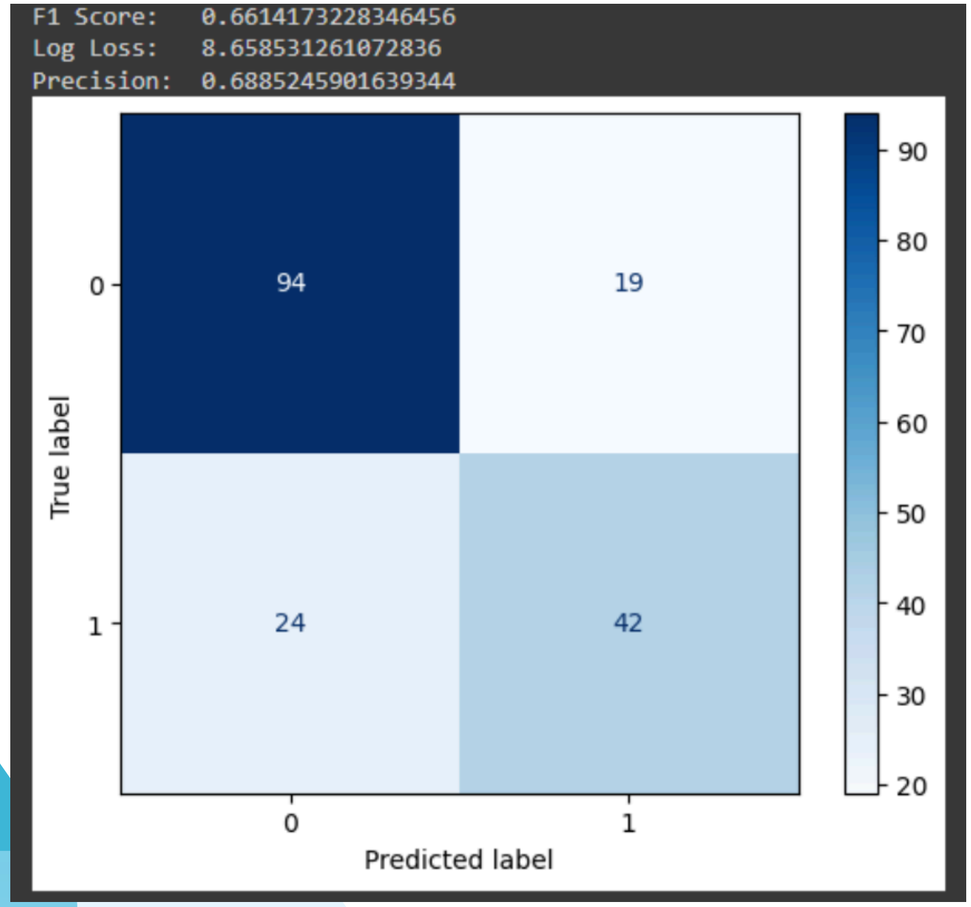
```
F1 Score:   0.7272727272727272
Log Loss:   7.249002916247025
Precision:  0.7619047619047619
```



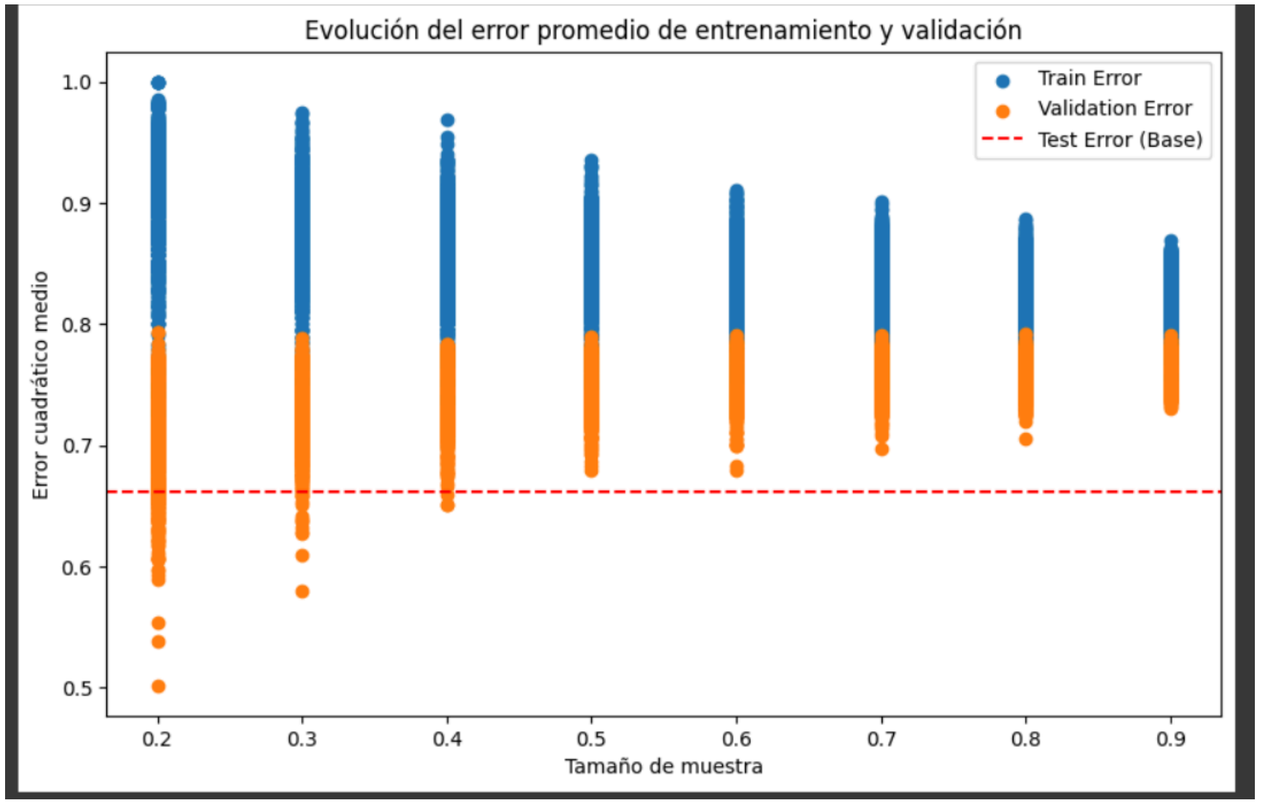


Bosque Aleatorio.

Parámetro	Valor
max_depth_list	[4]
n_estimators_list	[150]



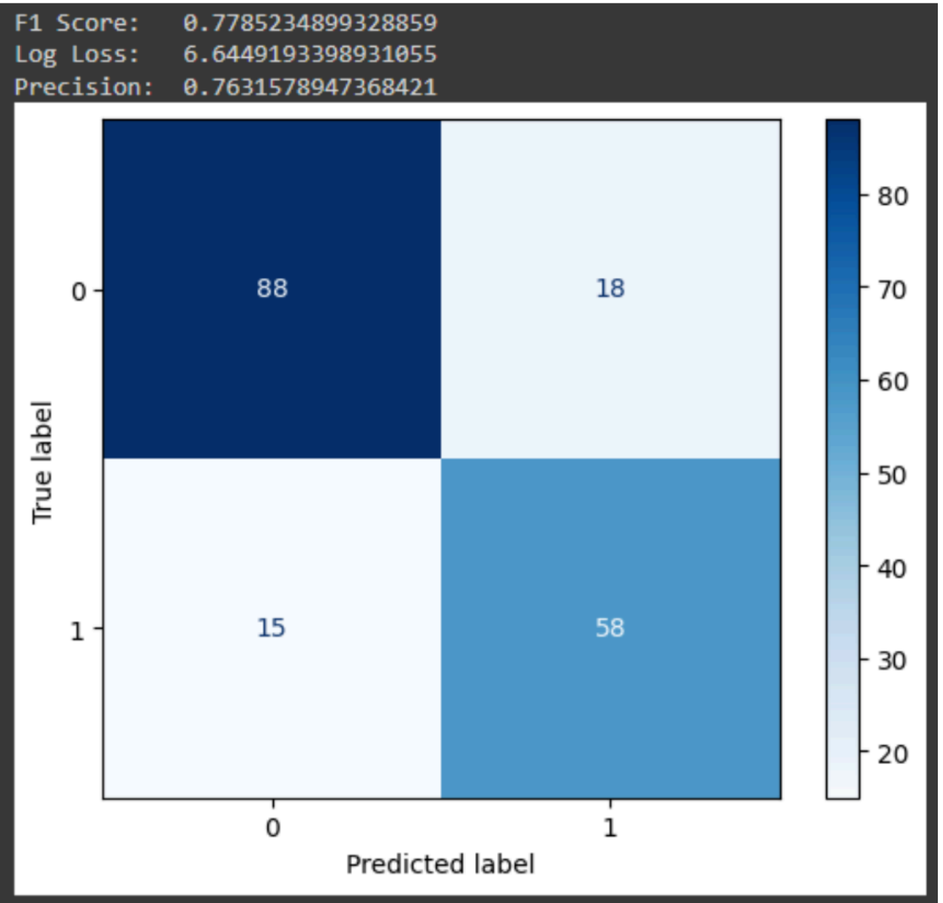
Mejor calificación Kaggle: 78.9

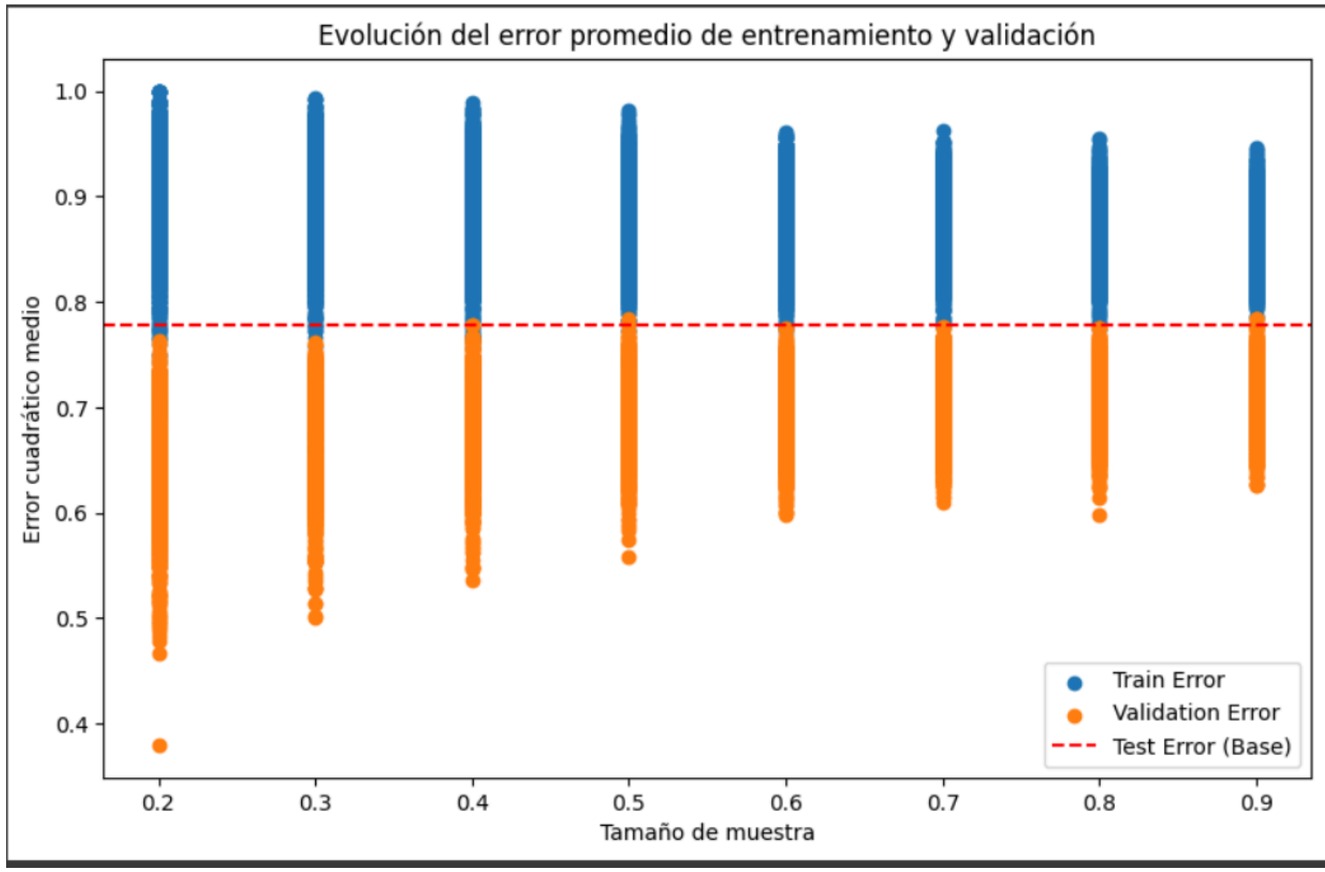


Árbol de Decisión

Parámetro	Valor
max_depth_list	[4]
min_samples_leaf_list	[1]

Mejor calificación Kaggle: 77.5

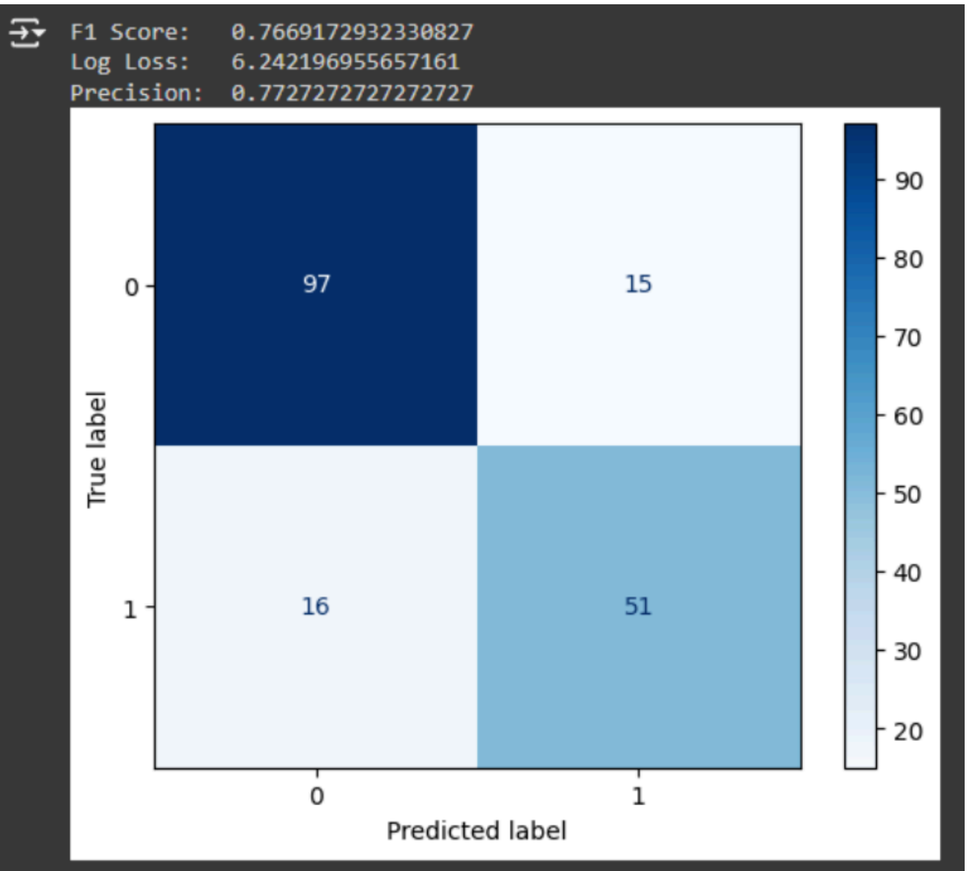




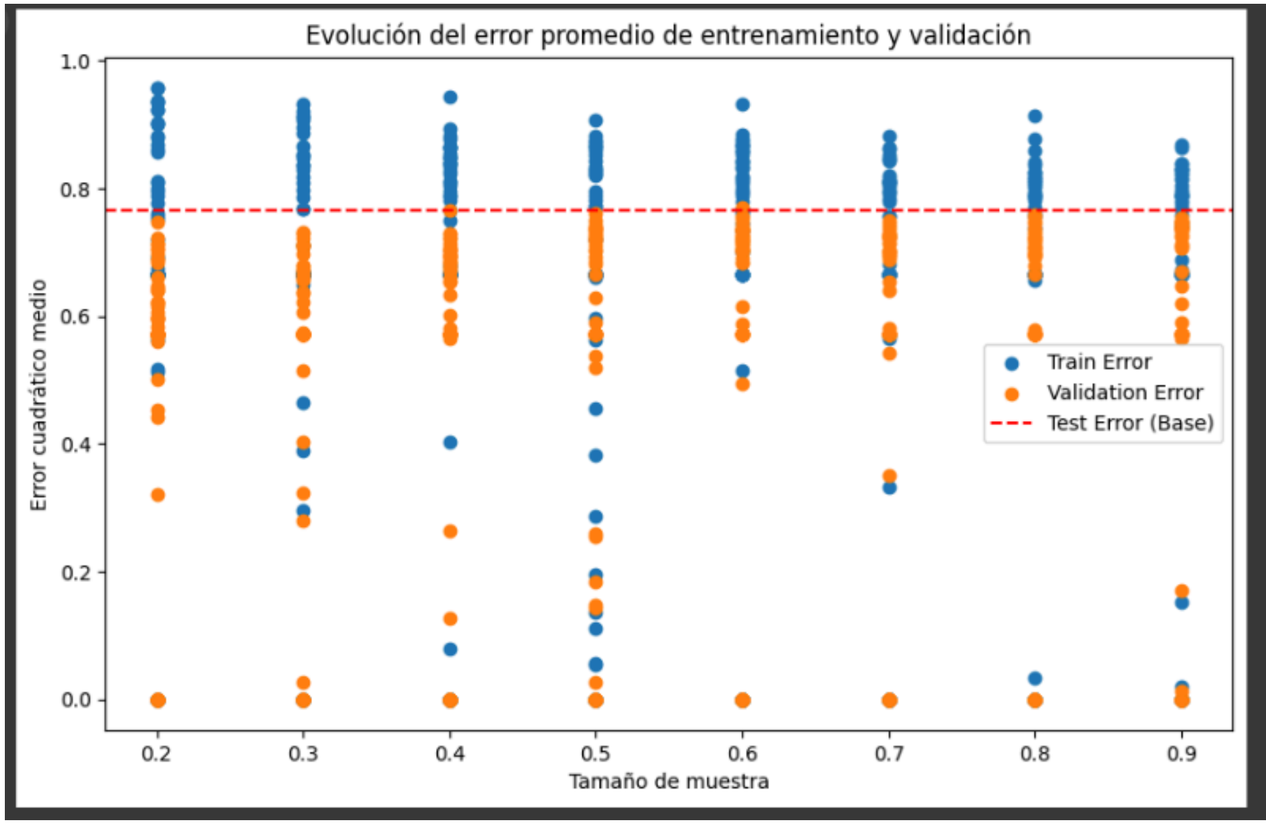
Redes Neuronales

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)



Mejor calificación Kaggle: 75.6



Resultados

El Bosque Aleatorio destacó como el modelo más prometedor. Se realizaron ajustes adicionales optimizando sus parámetros para encontrar la configuración ideal.

Conclusiones

El Bosque Aleatorio ha demostrado ser la opción más efectiva, combinando precisión y generalización. Aunque las redes neuronales podrían ofrecer mejoras en el futuro, el Bosque Aleatorio se posiciona actualmente como la solución más sólida para este contexto.

Gracias por su atencion!