Titanic

Machine Learning from Disaster

Samuel Méndez - A01652277
Mariana Pérez - A01731813
Nancy Segura - A01734337
Paul García - A01750164
Iker Ledesma - A01653115

Agenda

01 Problemática

02 Análisis exploratorio de los datos

03 Procesamiento de los datos

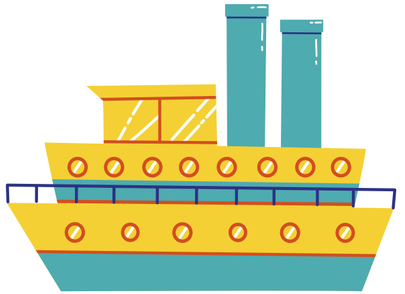
04 Implementación de modelos

05 Selección de un modelo

06 Conclusión

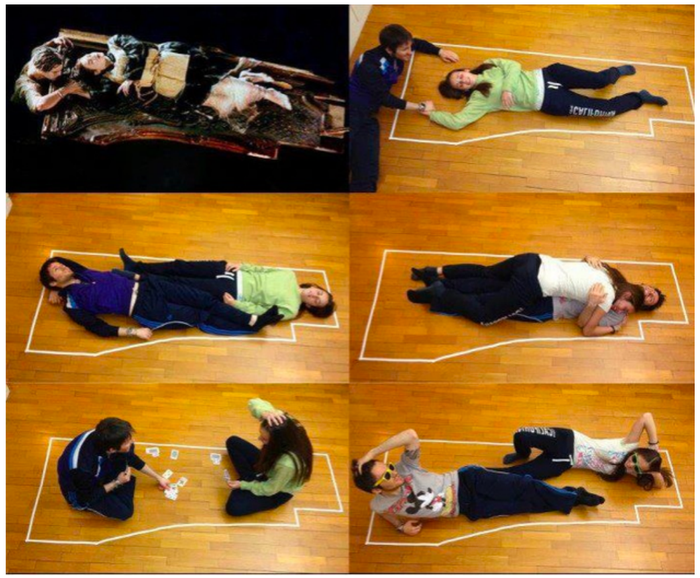
01

Problemática



Contexto histórico

El 14 de Abril de 1912 sucedió uno de los accidentes marítimos mas famosos de la historia, el hundimiento del Titanic. En el incidente murieron 1517 personas de las 2223 abordo.



Importancia

A través de la información recabada de algunos pasajeros, y de diferentes modelos de aprendizaje de máquina, es posible predecir si una persona murió o no en el siniestro, lo cual es de relevancia tener en cuenta para comportamientos similares en otros eventos.

02

Análisis exploratorio de los datos

Total de registros

891

Número de variables

12

Valores nulos

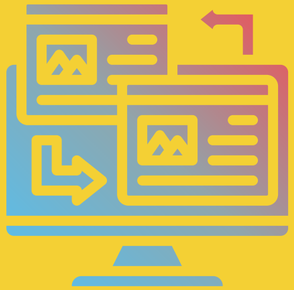
177

687

2

Registros duplicados

0

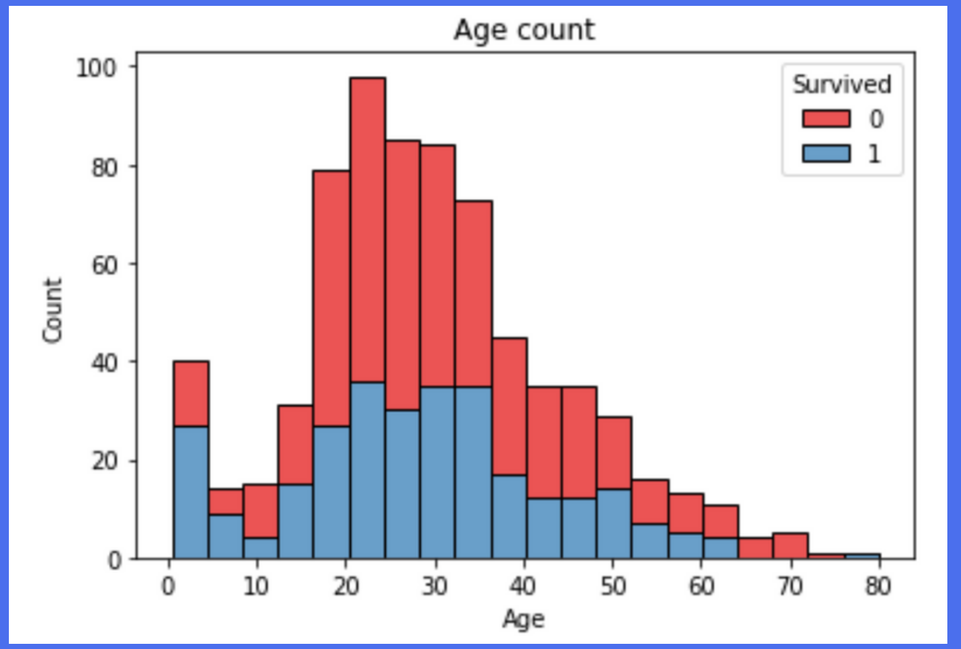


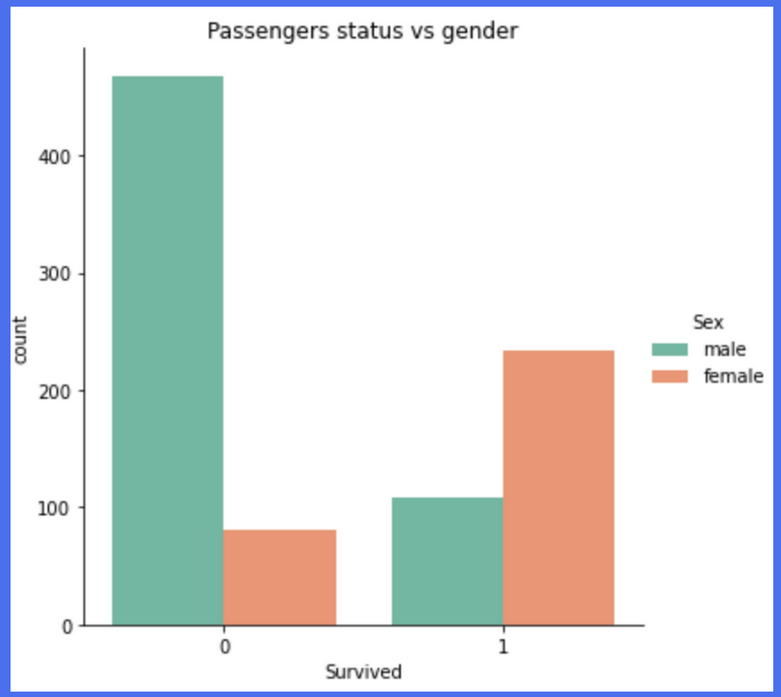
Sobrevivientes

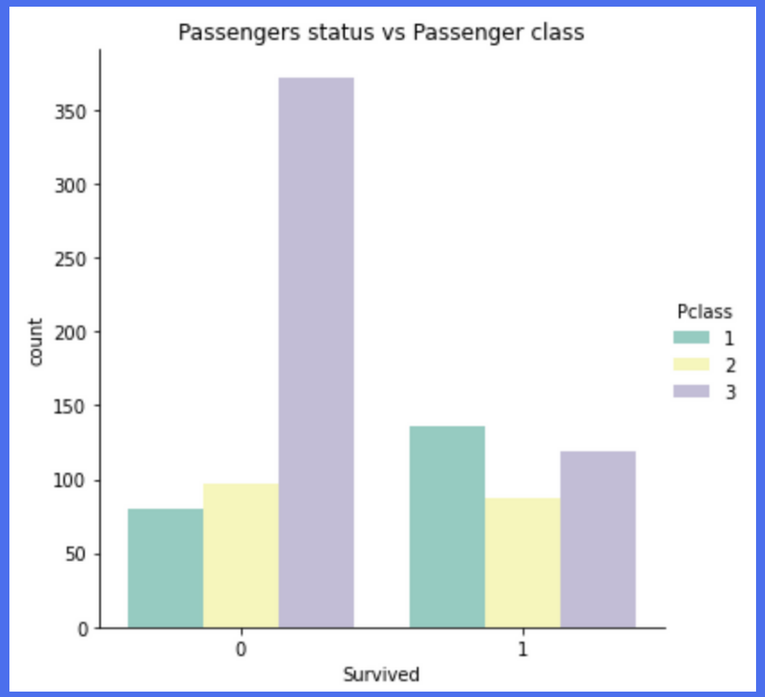
342

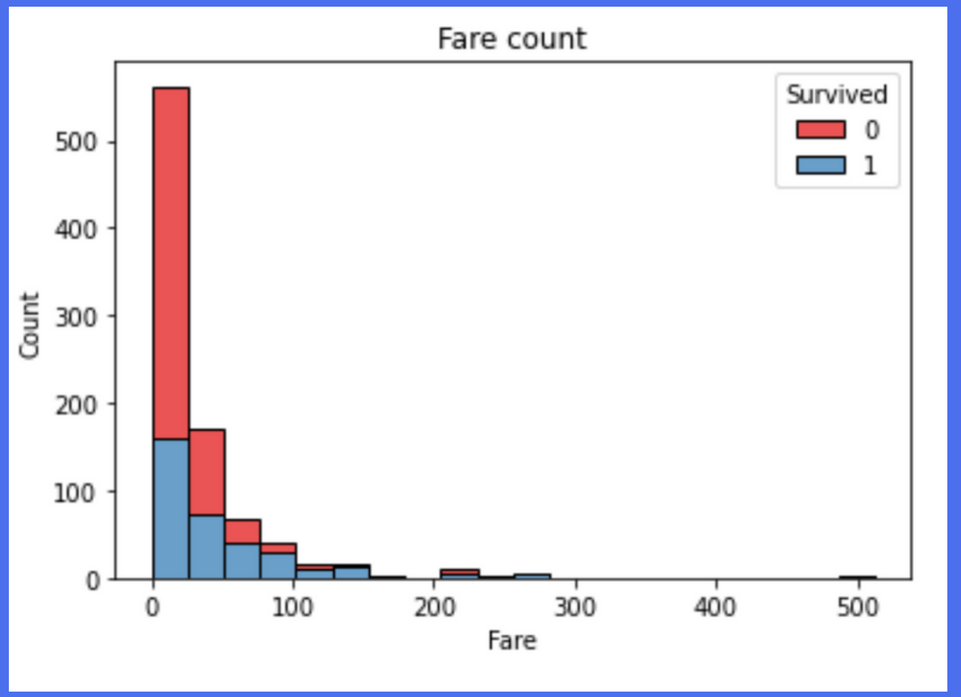
549

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200





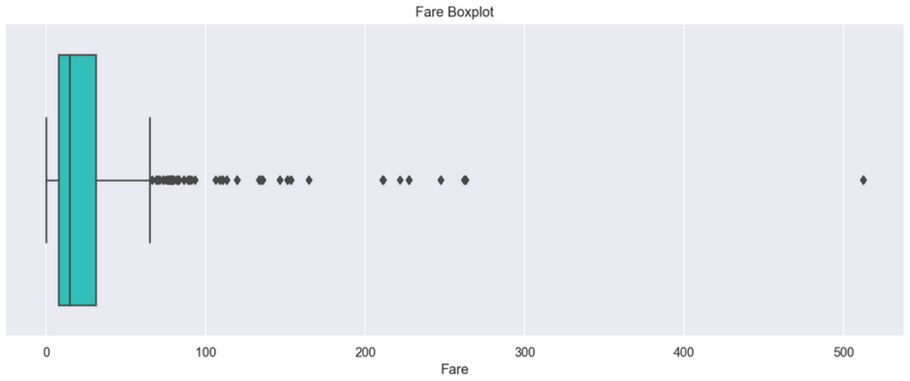




03 Procesamiento de los datos

- Transformación de los datos
- Técnicas de normalización
- Eliminación de datos
- Correlación entre variables
- División de los datos

Female	Male	Name_Pref
0.0	1.0	12
1.0	0.0	13
1.0	0.0	9
1.0	0.0	13
0.0	1.0	12
...	...	...
0.0	1.0	15
1.0	0.0	9
1.0	0.0	9
0.0	1.0	12
0.0	1.0	12

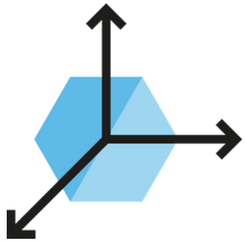


04

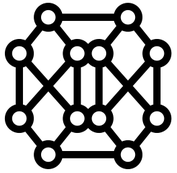
Implementación de modelos

Decision Tree Clasifier

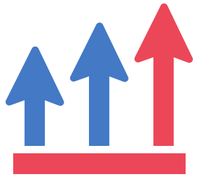
Random Forest Classifier



Support Vector Machine



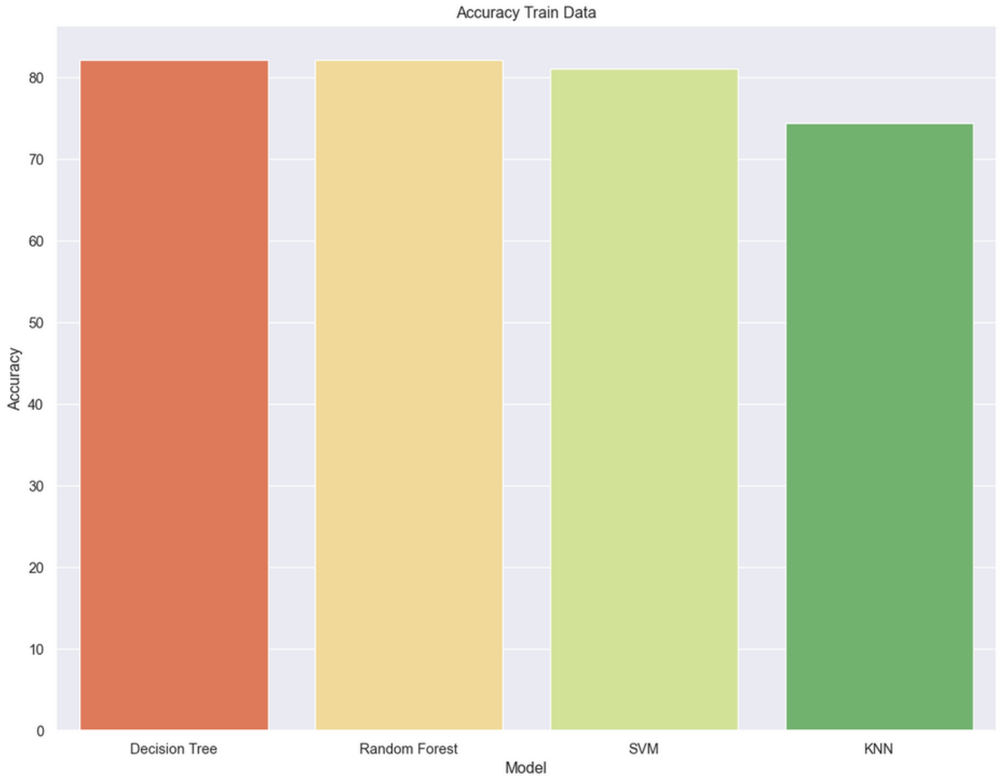
Neuronal Network (MLP)

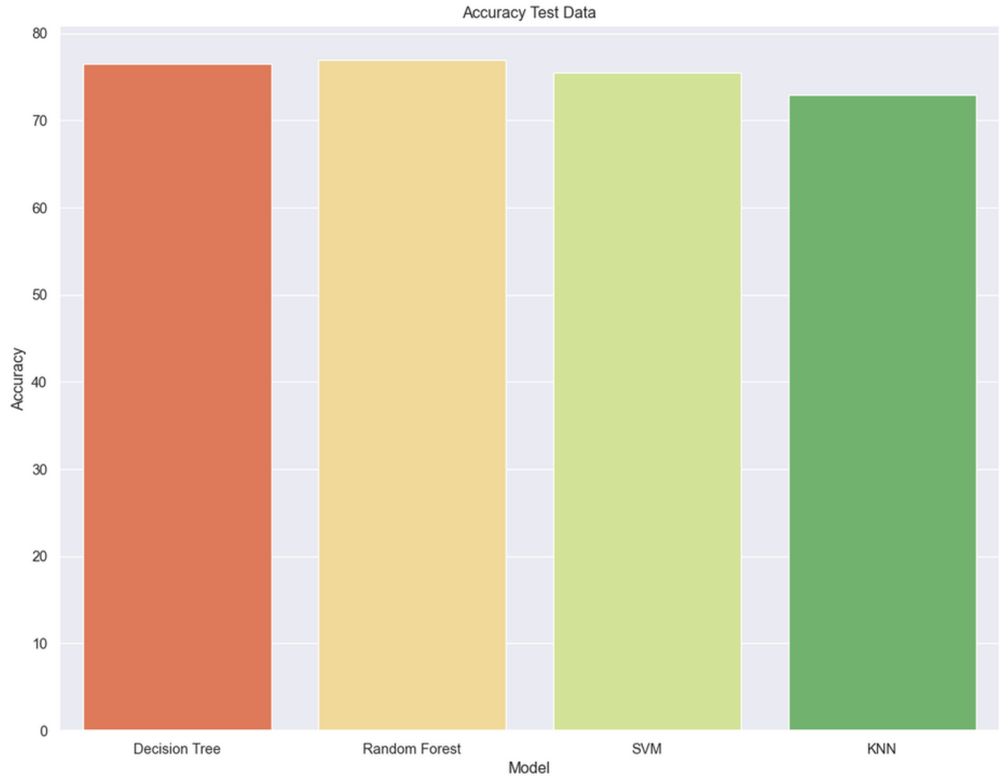


XGBoost

04

Implementación de modelos





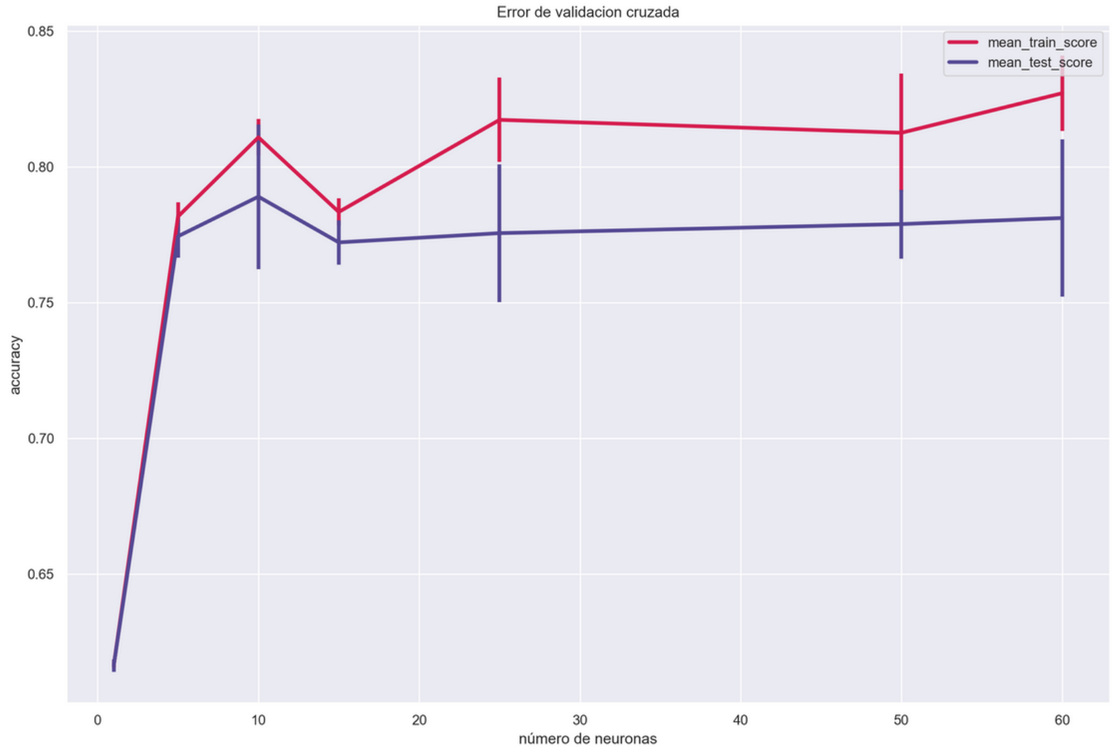
¿Overfitting?

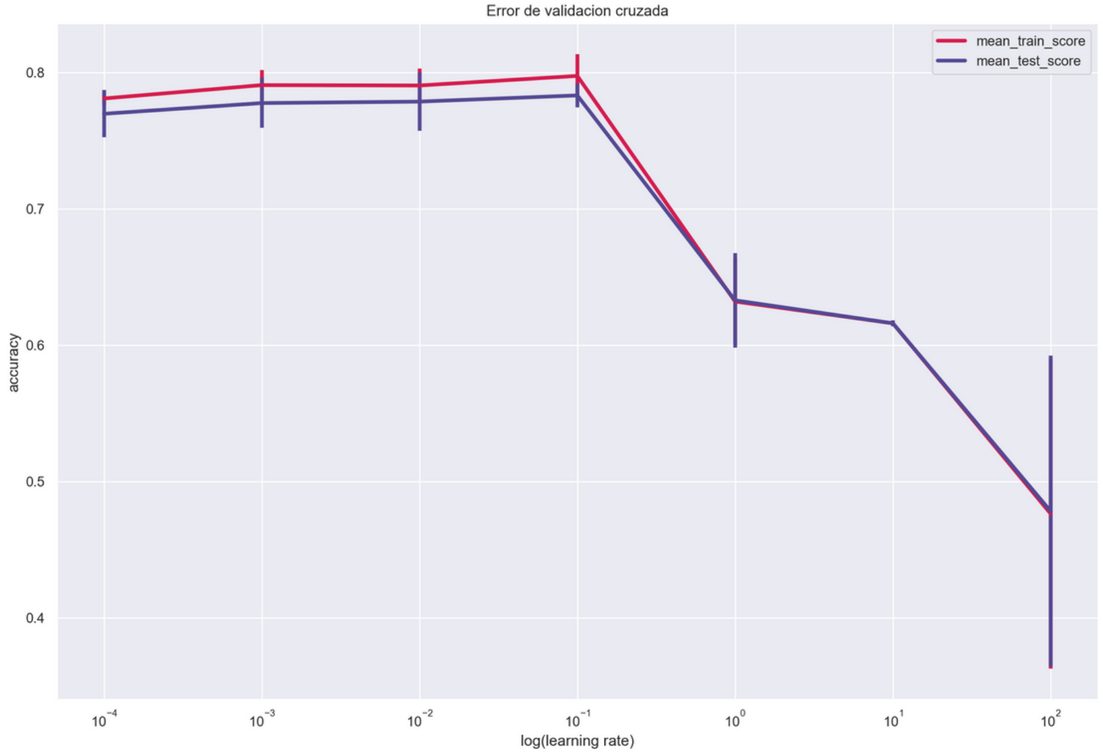
05

Selección de un modelo

Neuronal Network (MLP)

Hiperparámetros



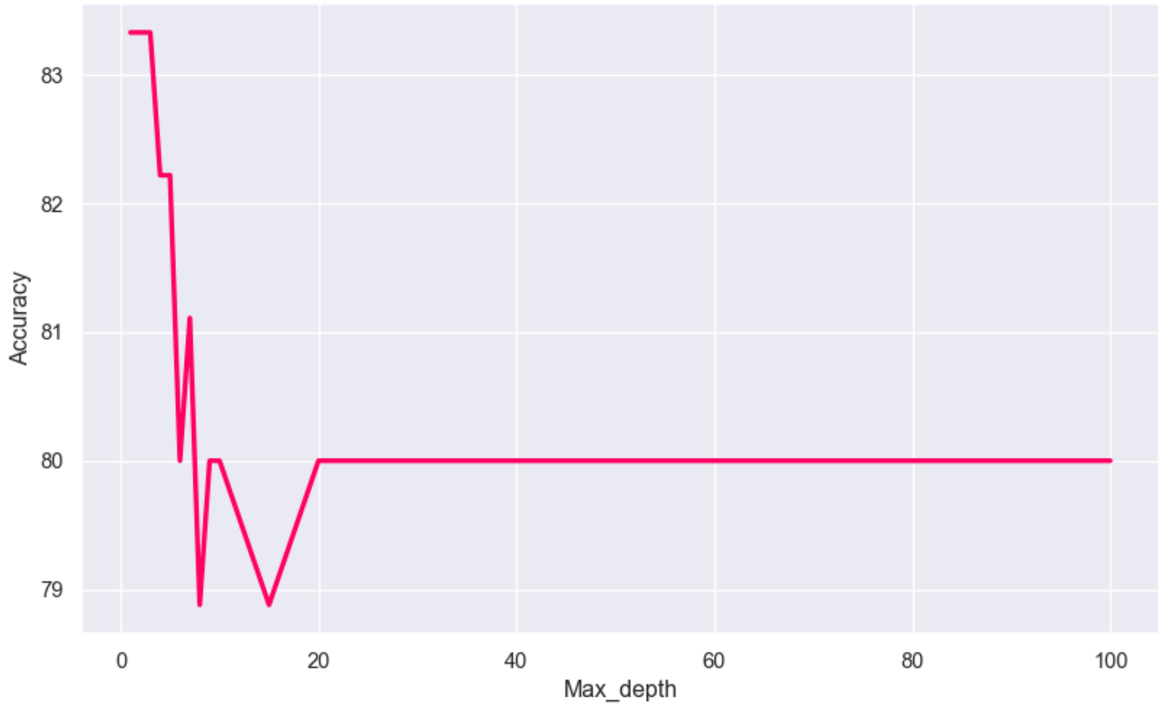


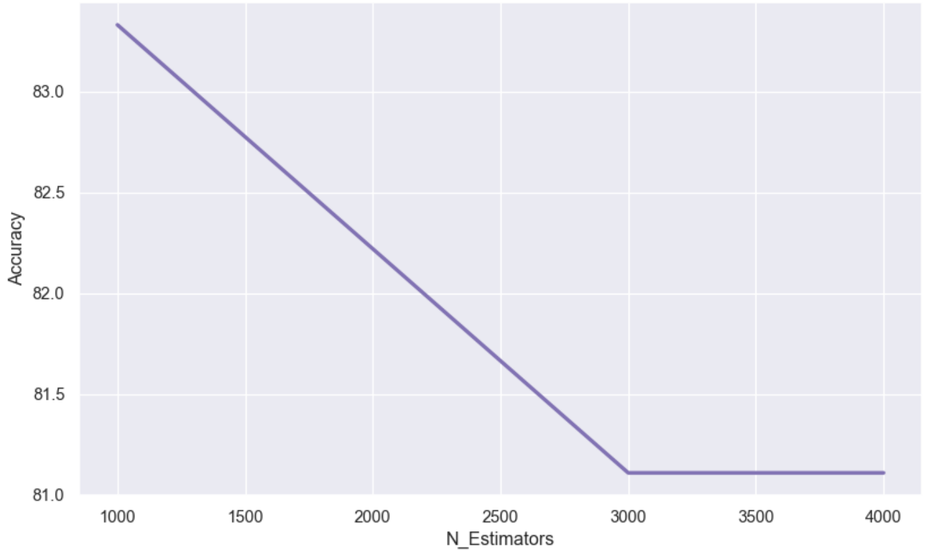
05

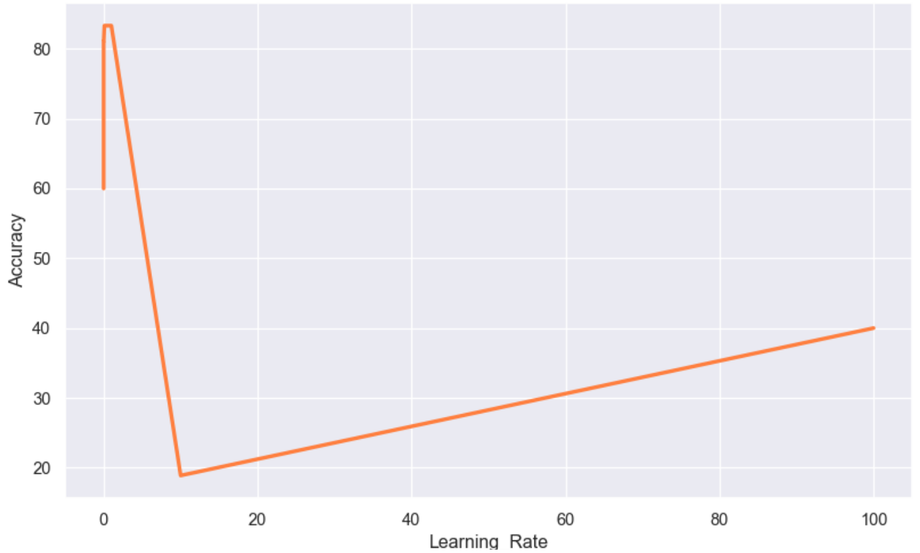
Selección de un modelo

XGBoost

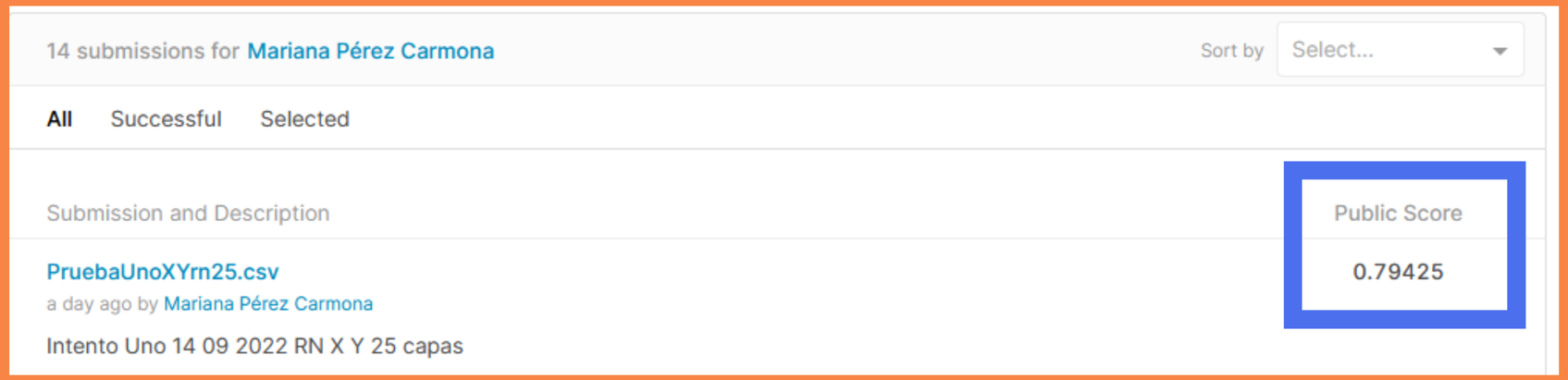
Hiperparámetros

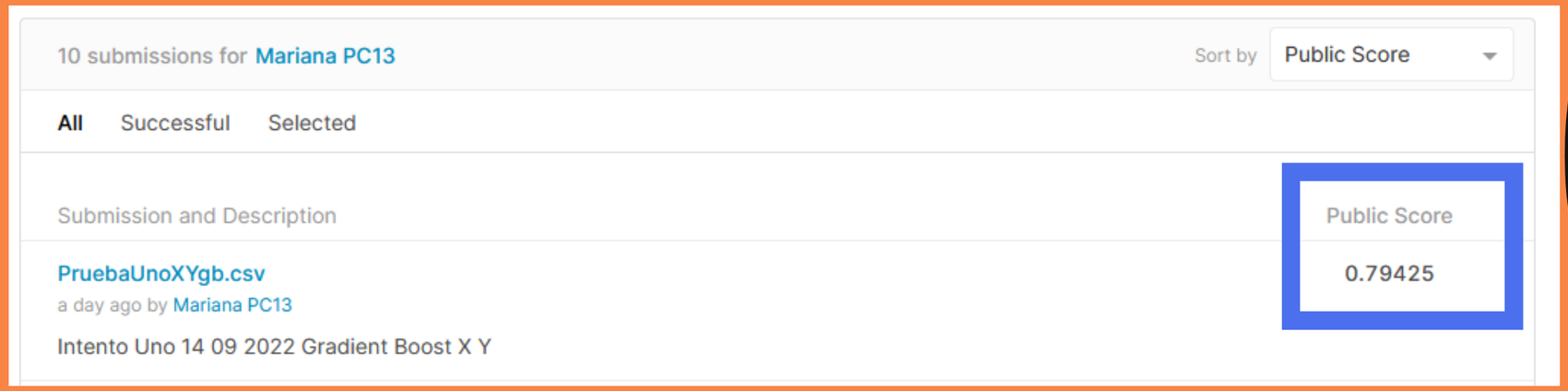






Predicciones en Kaggle





Evaluación del Reto

Se aprendió la importancia de realizar un buen estudio y manejo de base de datos.
Aunado a marcar un objetivo claro, realista y alcanzable para la realización de un proyecto.
Junto también a la definición del proceso a seguir para lograr el objetivo.

- Entender las variables.
- Analizar relación de las variables.
- Planteamiento de posibles modelos.
- Ajustes de hiperparámetros.
- Comparación de modelos.
- Regresar al primer punto.
- Selección.

Conclusiones

- Análisis y limpieza exitosa de la base de datos
- Aprendizaje sobre manejo de datos y variables
- Desarrollo efectivo de distintos modelos
- Áreas de oportunidad en la precisión