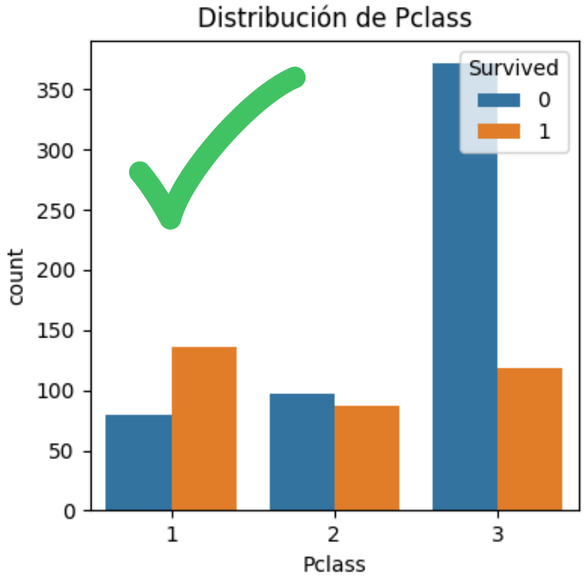
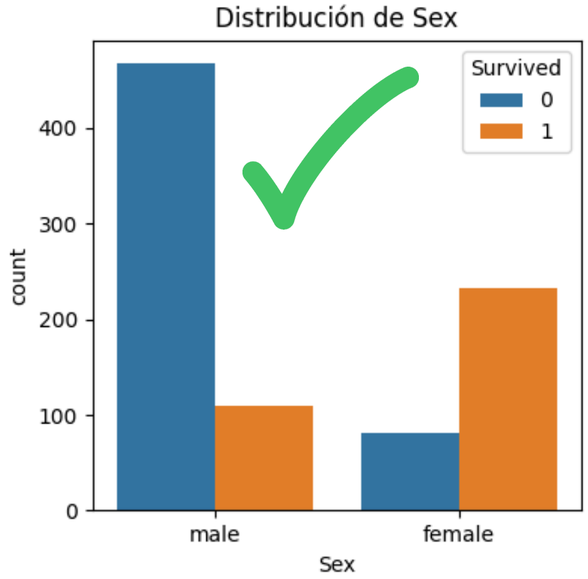
CONCENTRACIÓN IA

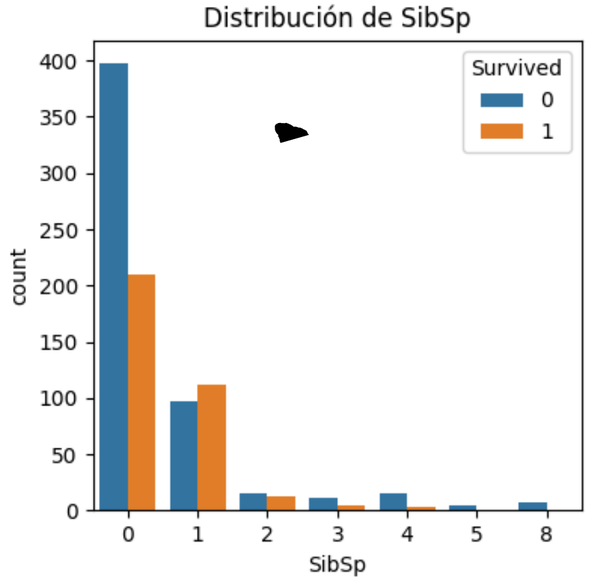
RETO TITANIC

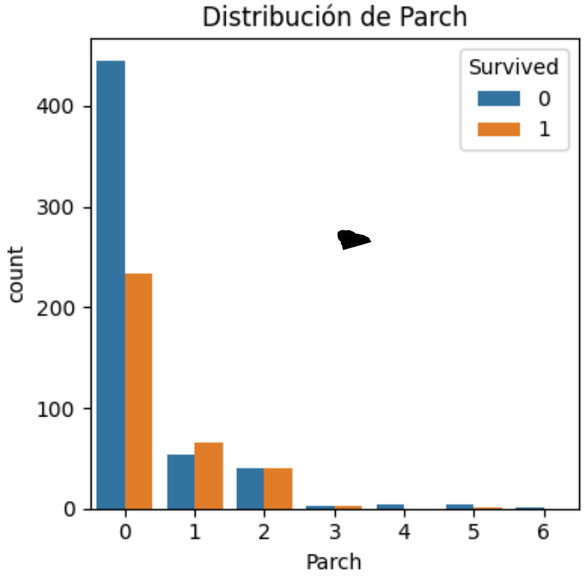
Análisis del dataset

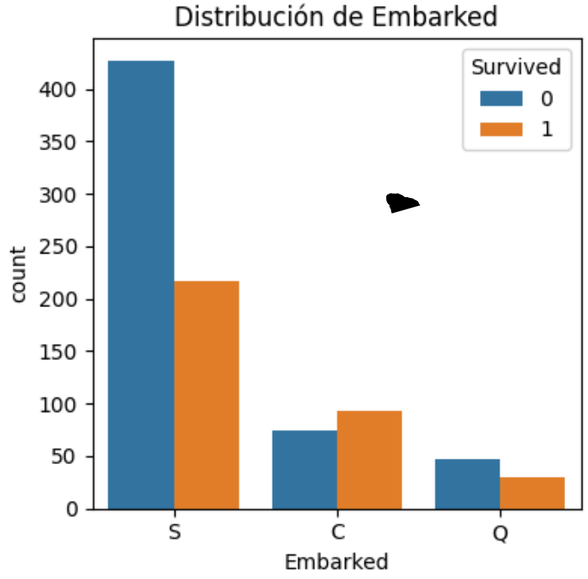
Clases Categóricas

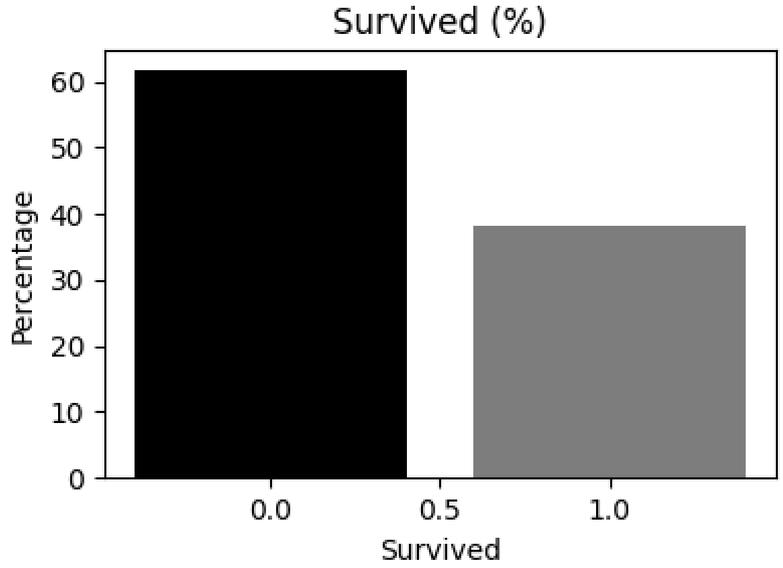






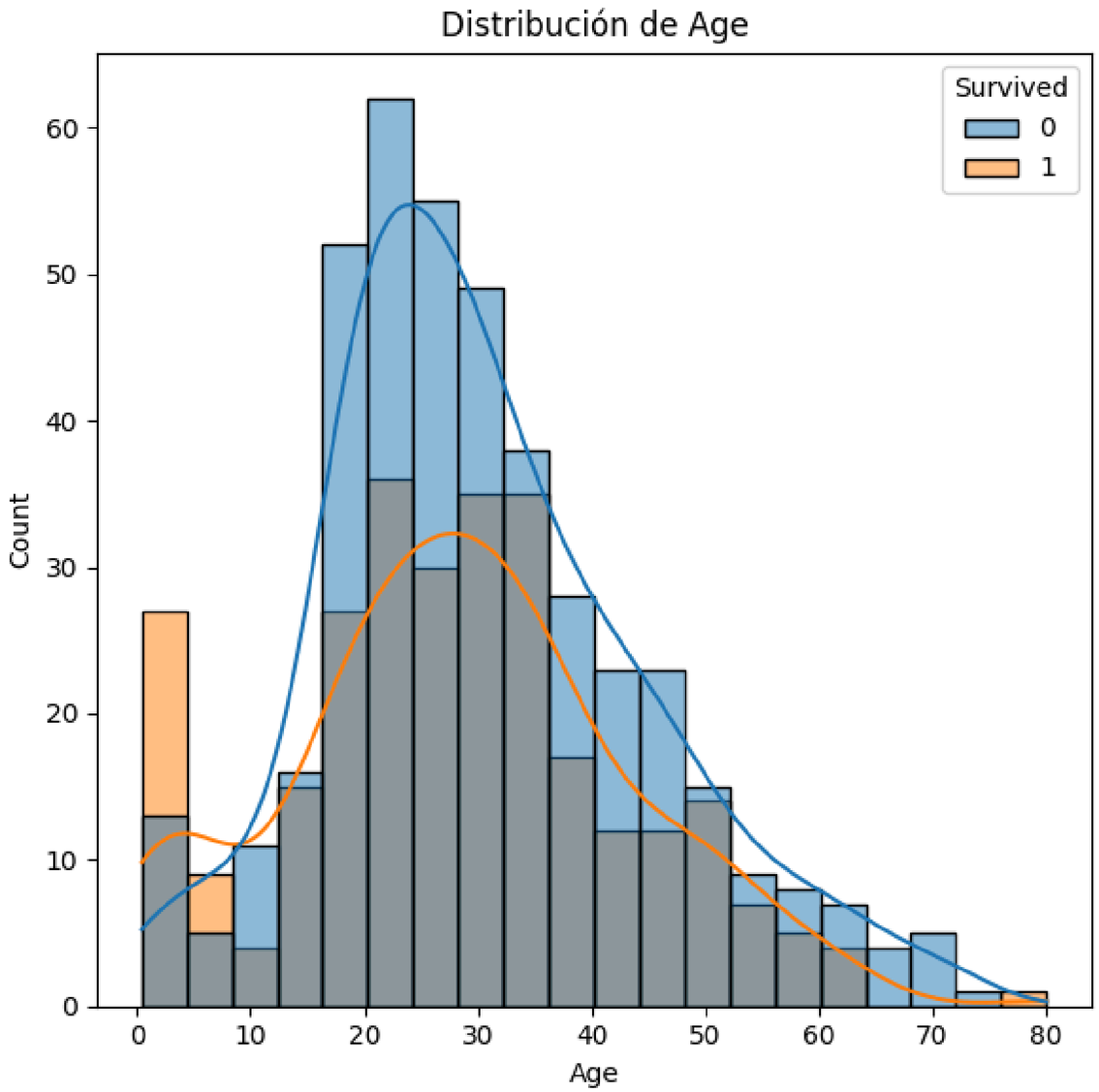


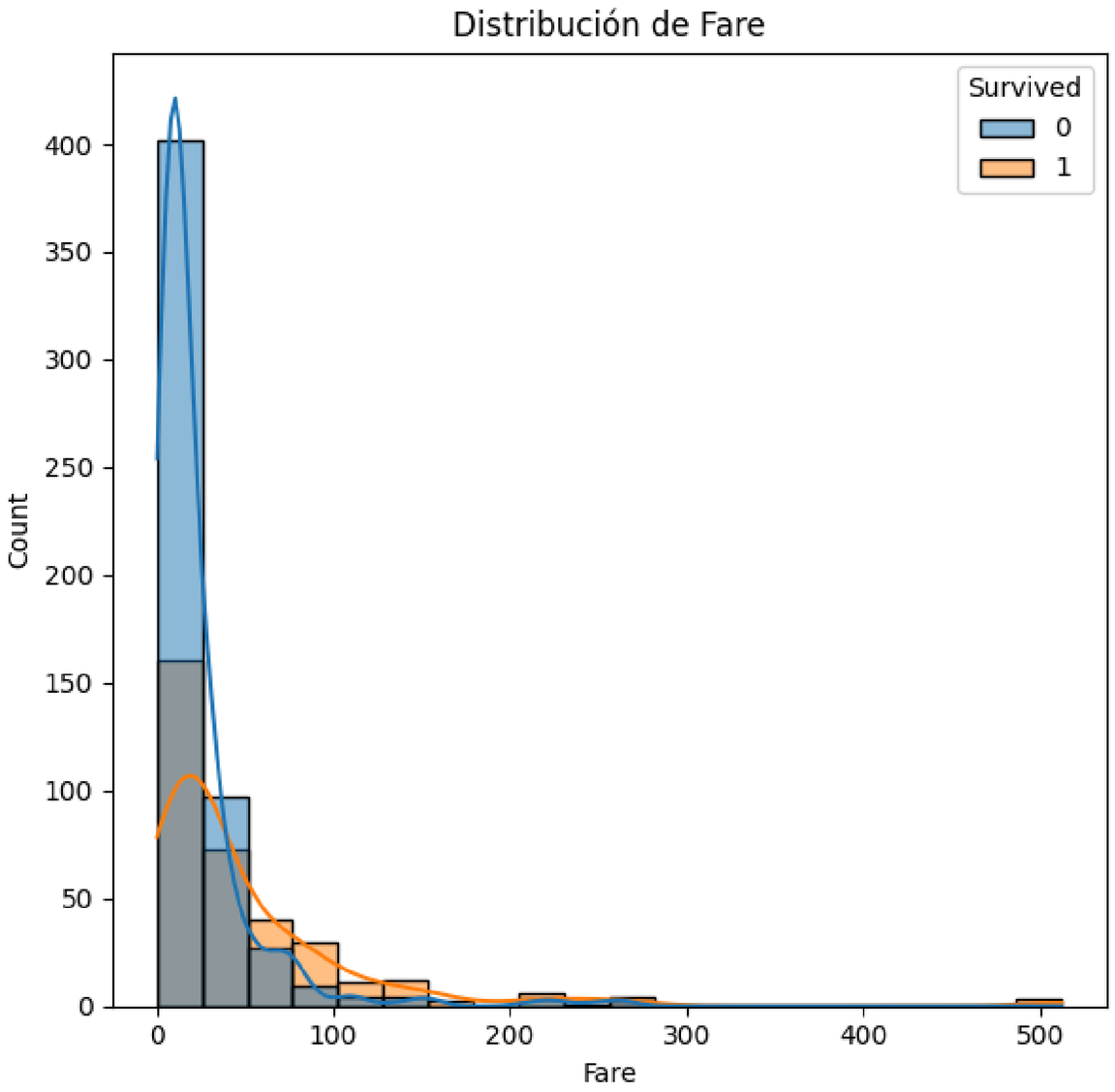




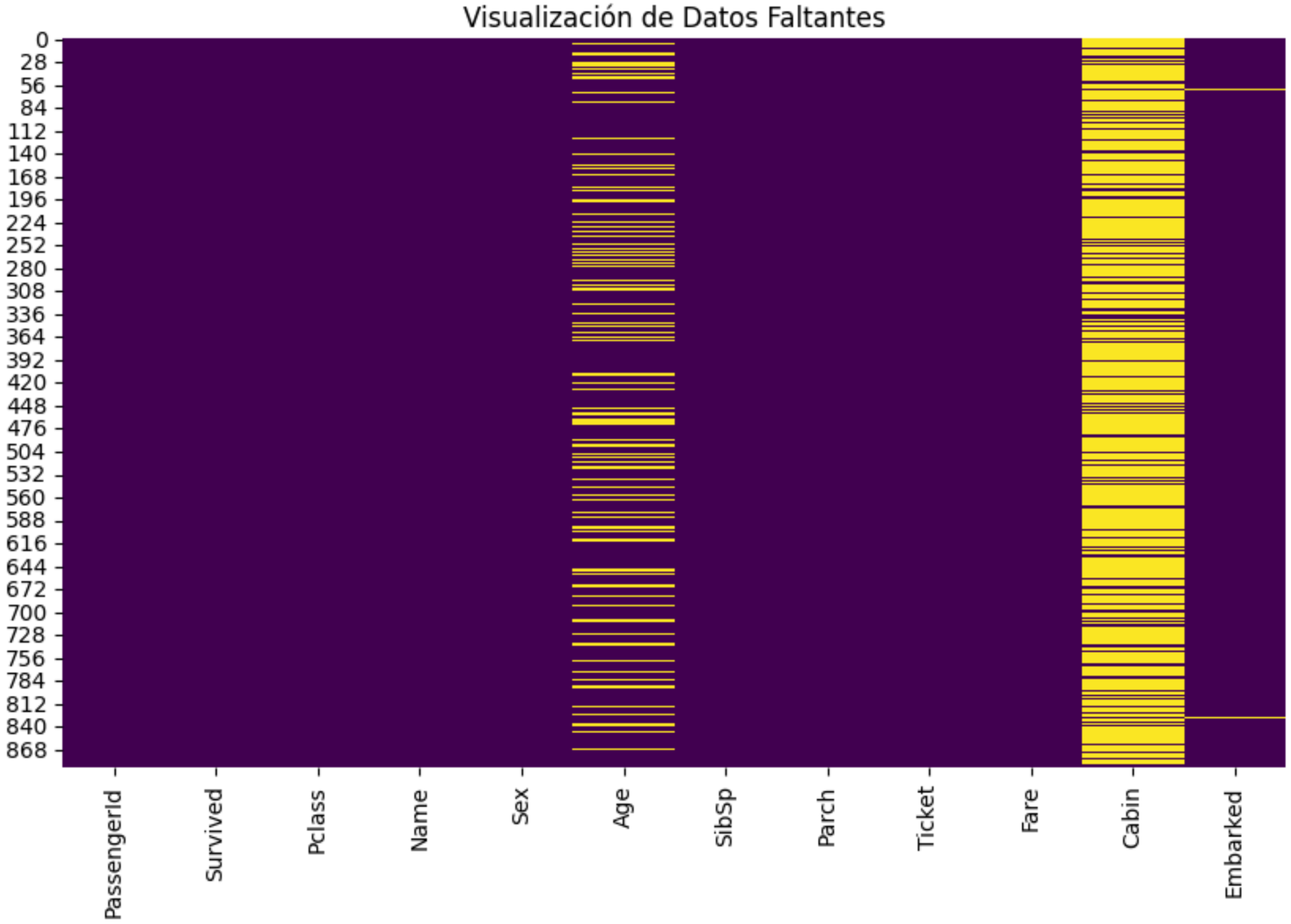
Análisis del dataset

Clases Numéricas

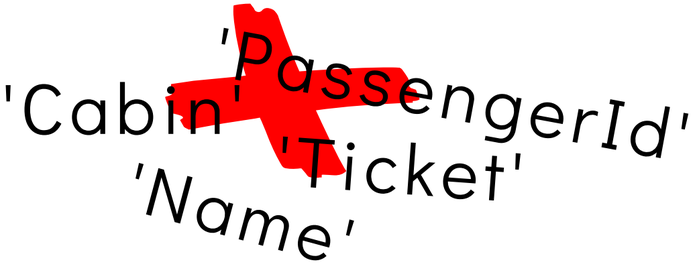




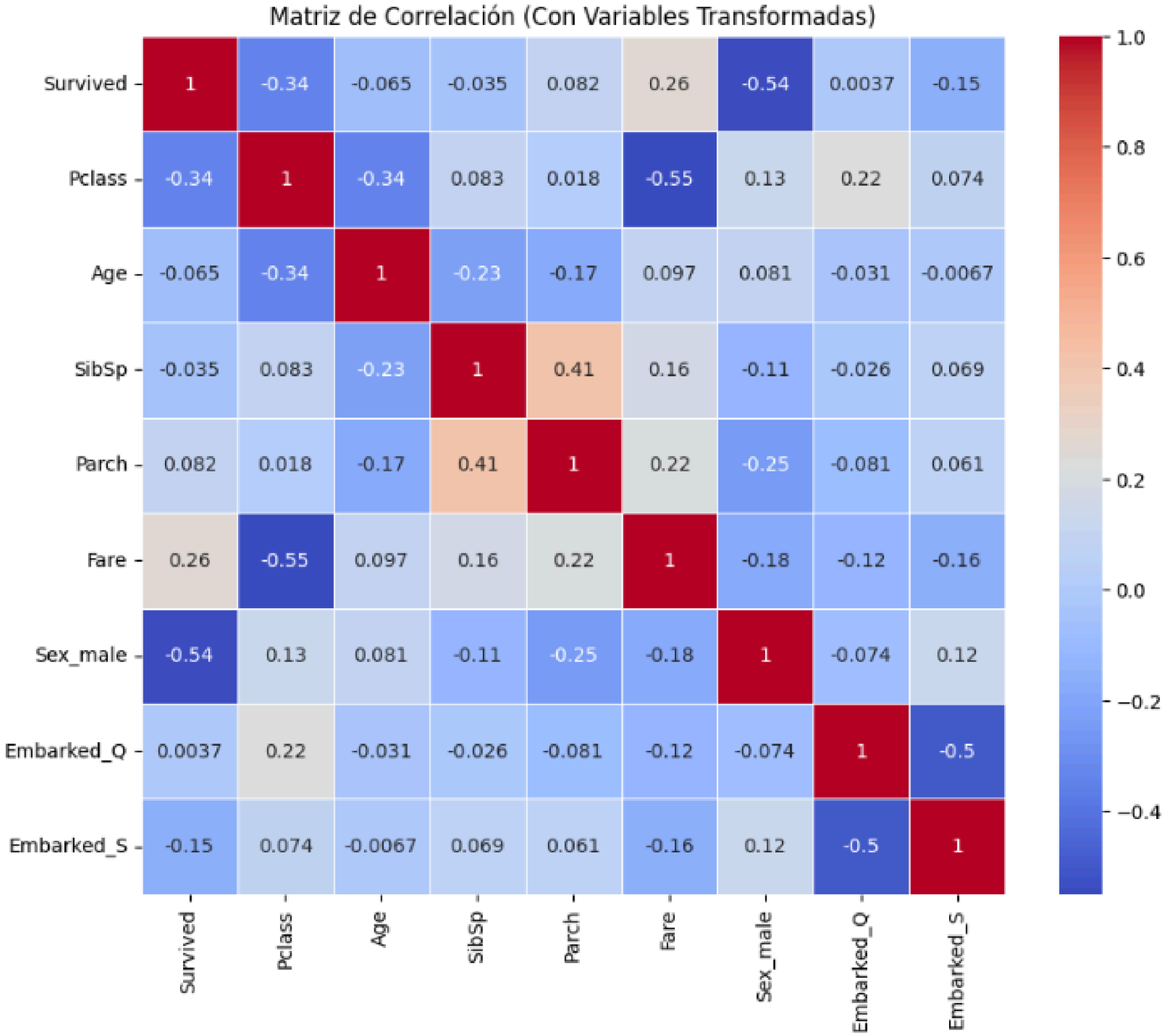
Eliminación de datos



01 Eliminación de las columnas ‘Cabin’, 'Name', 'Ticket' y ‘PassengerId’ 687/708



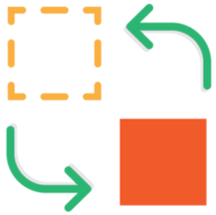
CORRELACIÓN DE DATOS



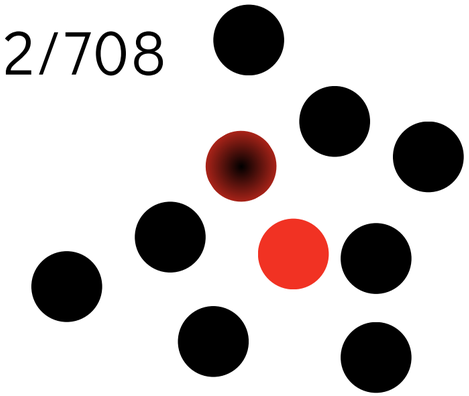
Imputación de datos

01 Eliminación de la columna ‘Embarked_Q’

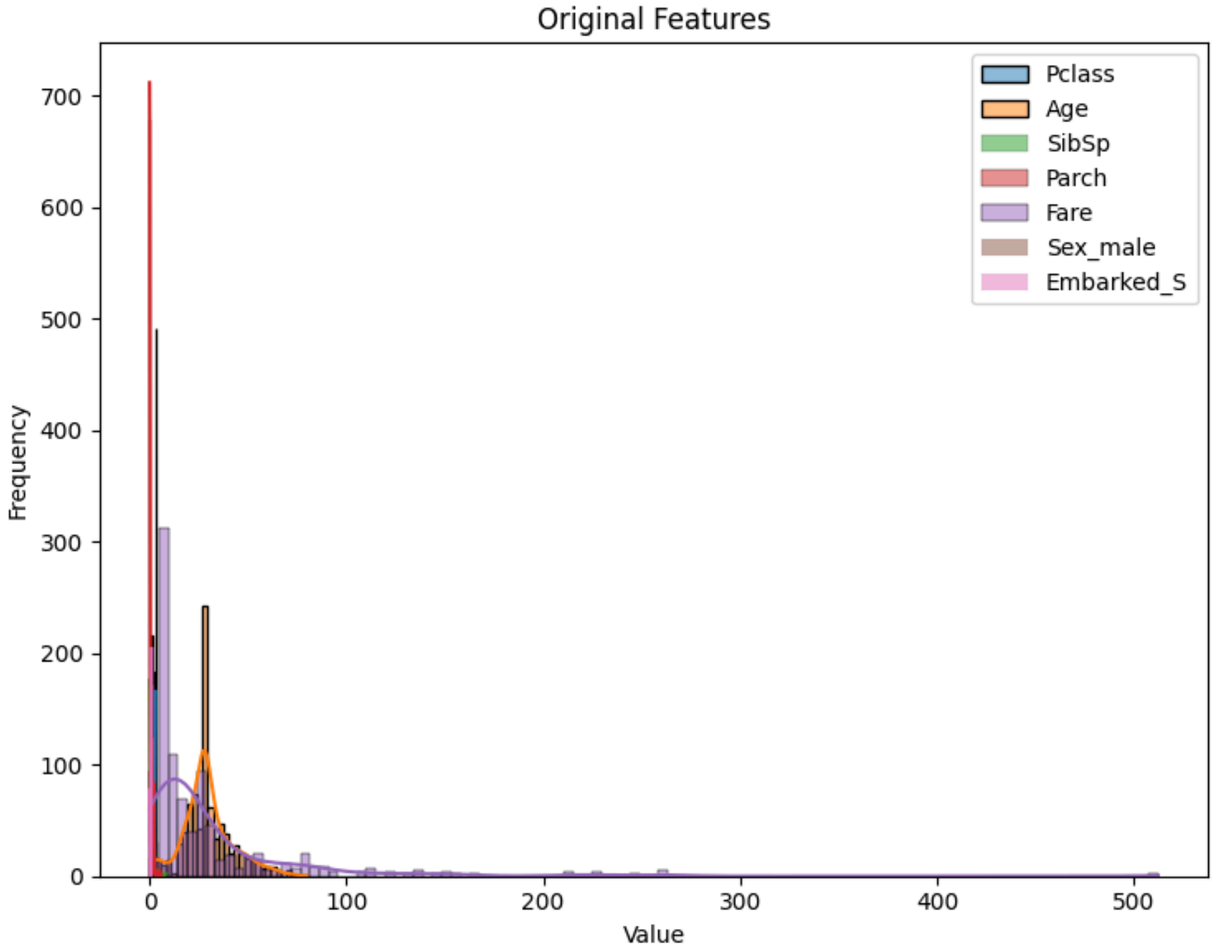
02 Reemplazo con mediana en ‘age’ 177/708



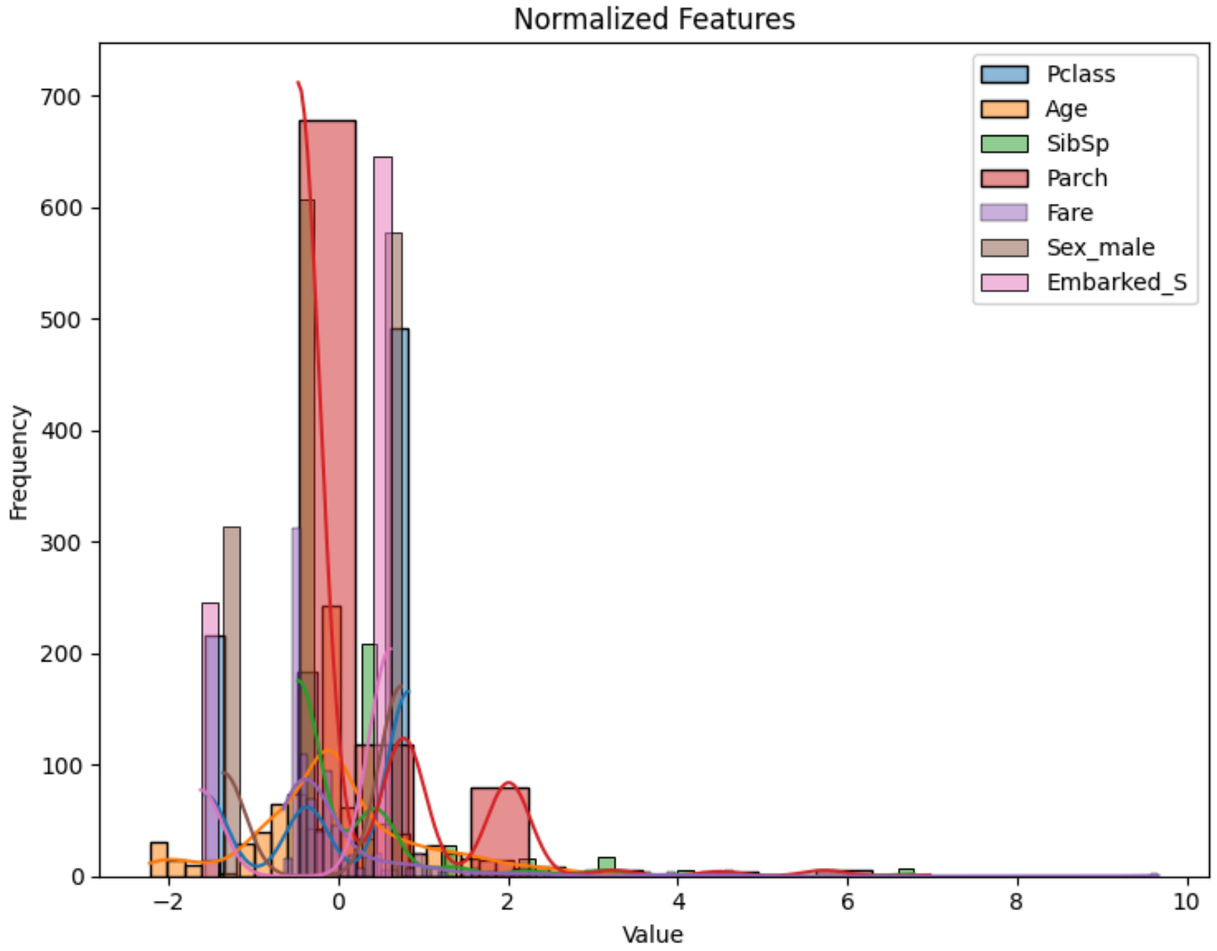
03 Reemplazo de valores faltantes con moda ‘Embarked_S’ 2/708



Normalización de datos



before



after

Conclusiones del preprocesamiento de datos

01 Las clases más relevantes en orden son: Sex_male, Pclass, Fare.

02 7 características identificadas para el modelo.

	Survived
Survived	1.000000
Sex_male	-0.543351
Pclass	-0.338481
Fare	0.257307
Embarked_S	-0.149683
Parch	0.081629
Age	-0.064910
SibSp	-0.035322

	Survived	Pclass	Age	SibSp	Parch	Fare	Sex_male	Embarked_S
0	0	3	22.0	1	0	7.2500	True	True
1	1	1	38.0	1	0	71.2833	False	False
2	1	3	26.0	0	0	7.9250	False	True
3	1	1	35.0	1	0	53.1000	False	True
4	0	3	35.0	0	0	8.0500	True	True

Selección del modelo

Máquina de vectores de soporte

- Puede manejar datos no lineales (kernel trick).
- Robusto a valores atípicos y puede encontrar hiperplanos óptimos para la separación de clases.
- Es útil para este problema ya que puede capturar relaciones complejas entre las características.

Regresión logística

- Clasificación binaria (sobrevivió/no sobrevivió)
- Es interpretable y proporciona información sobre la probabilidad de pertenencia a una clase.

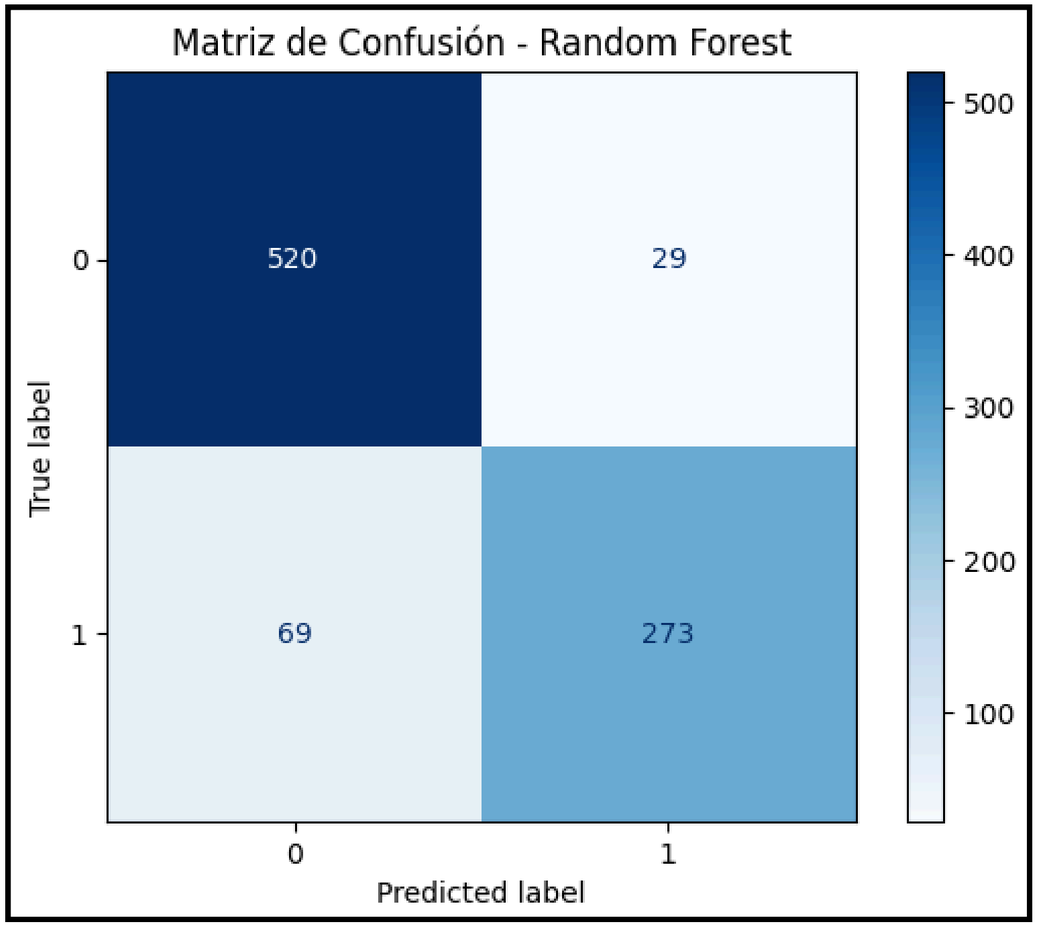
Bosque Aleatorio

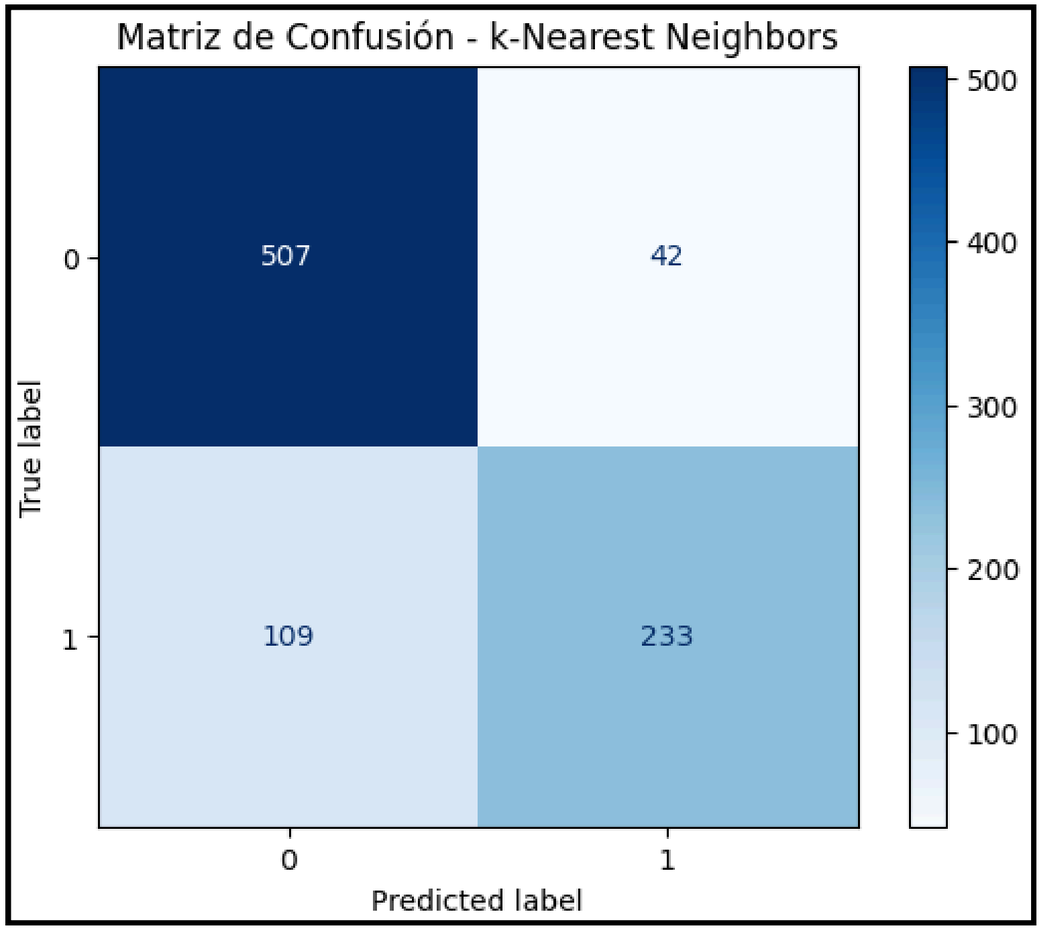
- Captura interacciones complejas entre las características.
- Es menos sensible a valores atípicos.
- Ajustable para manejar datasets desbalanceados.

K-NN

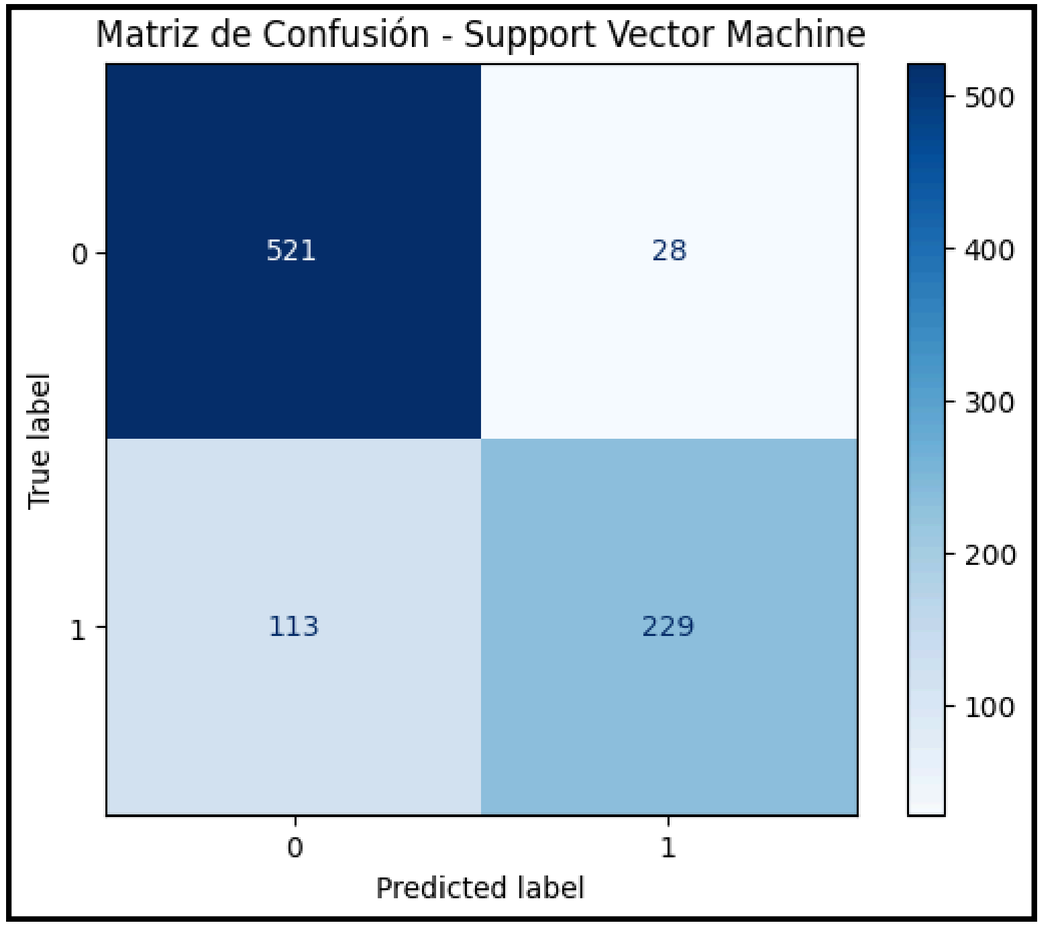
- Modelo no paramétrico.
- Requiere normalización.
- Clasificación con base en la similitud de carácteristicas de los pasajeros.

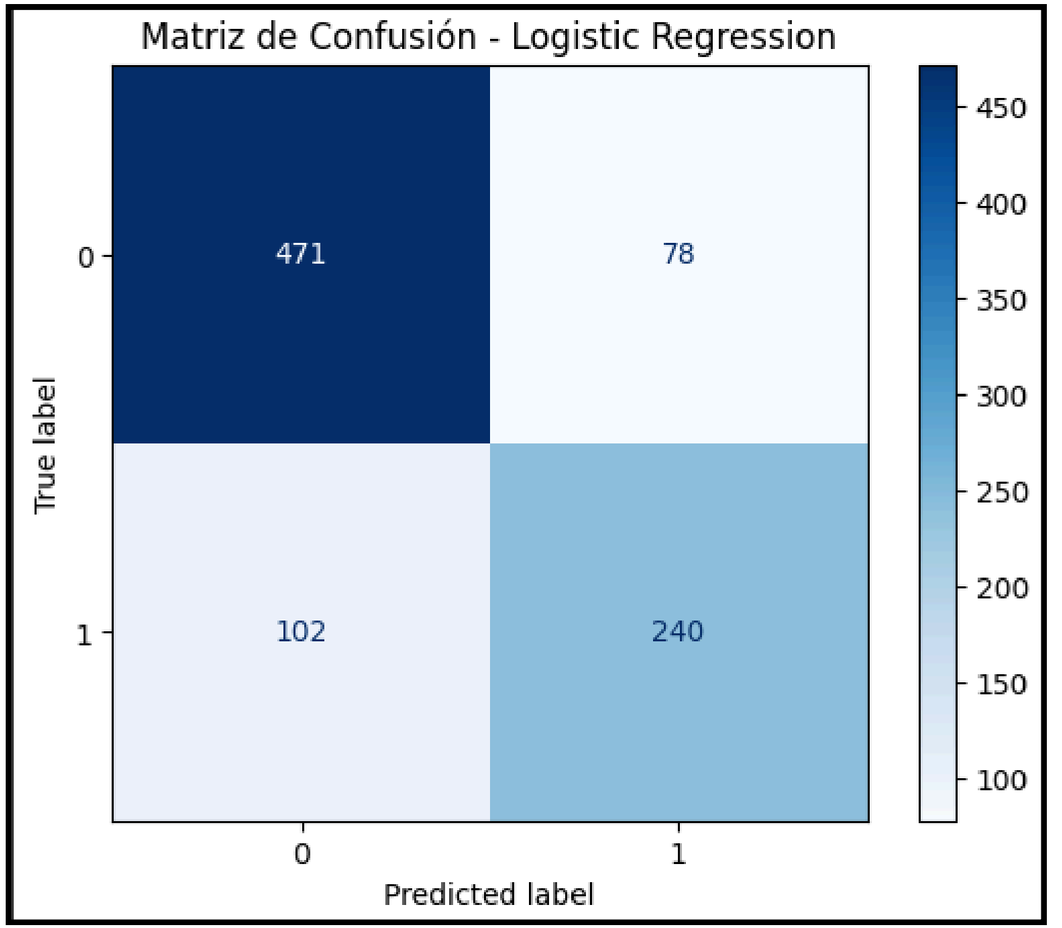
MATRICES DE CONFUSIÓN (RF VS KNN)



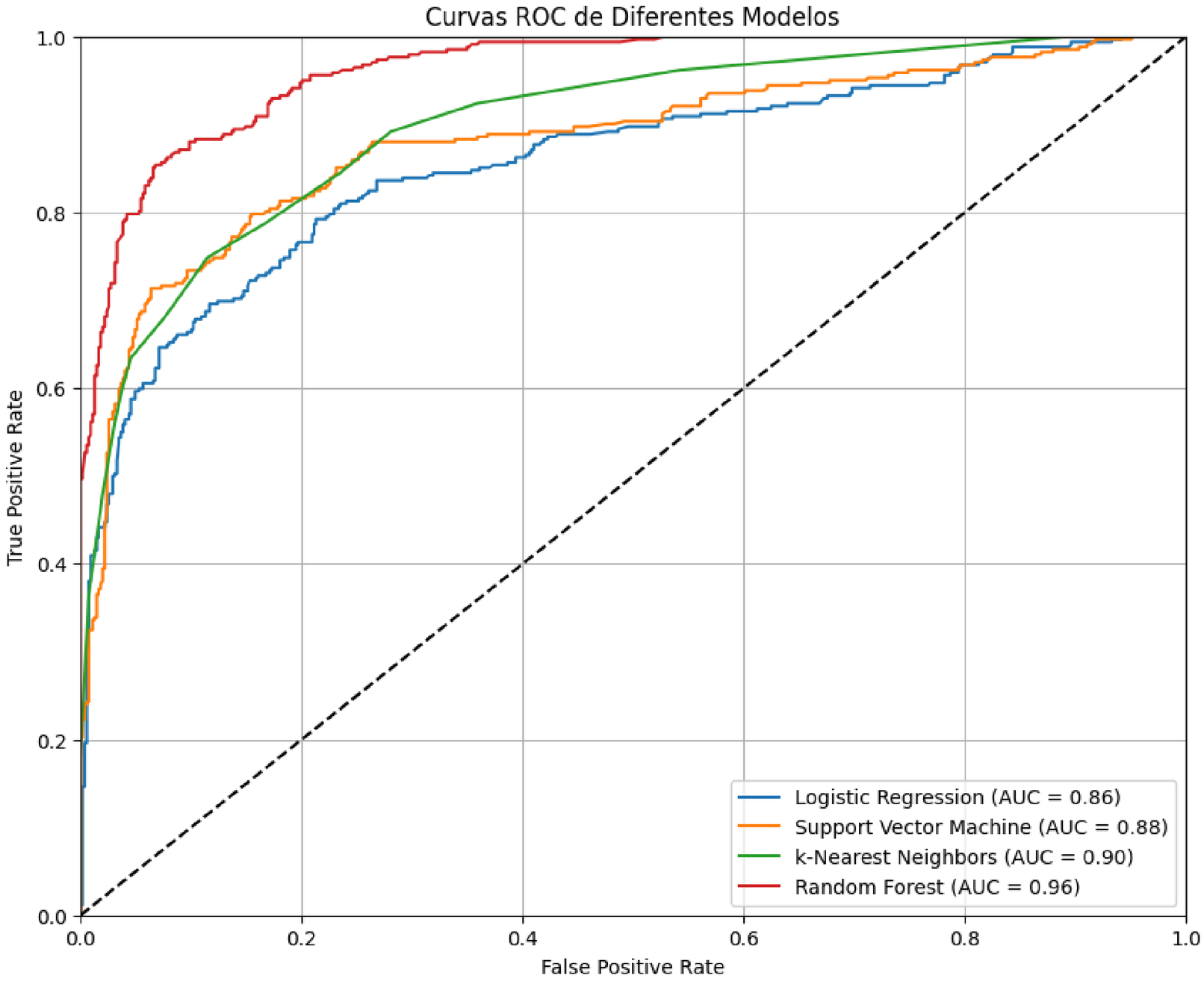


MATRICES DE CONFUSIÓN (SVM VS LR)





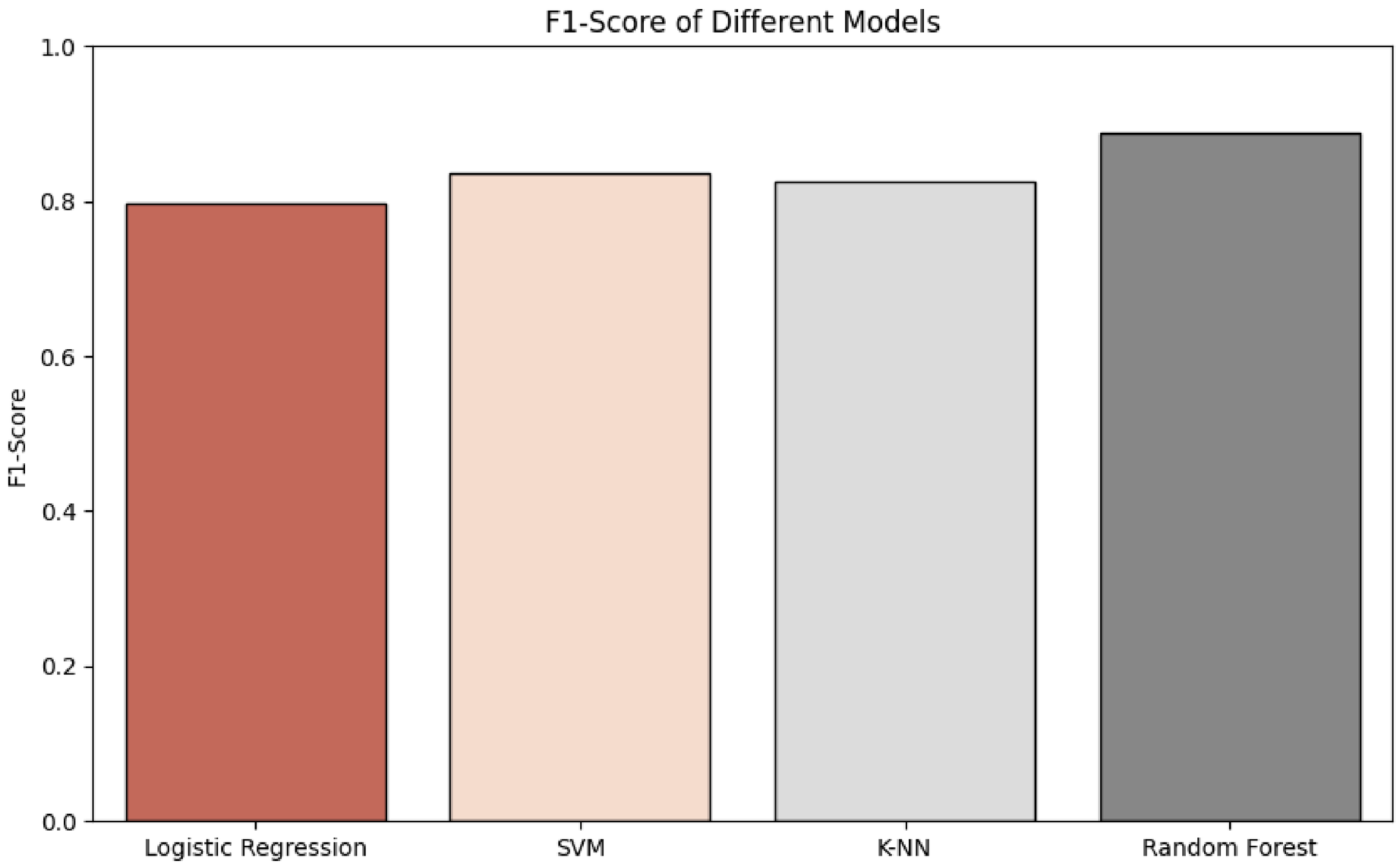
CURVA ROC


Curvas ROC de Diferentes Modelos
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.86)
Support Vector Machine (AUC = 0.88)
k-Nearest Neighbors (AUC = 0.90)
Random Forest (AUC = 0.96)

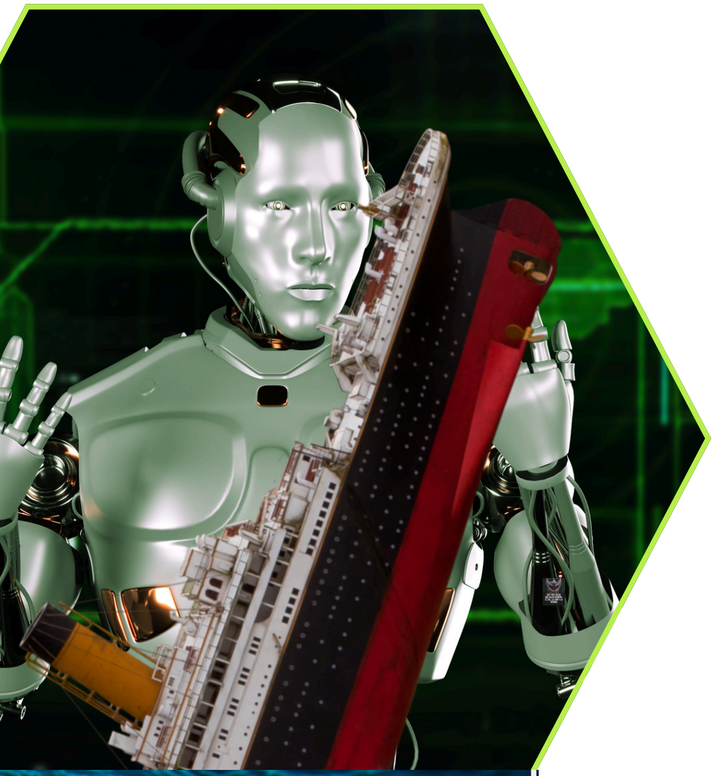
MODELOS ENTRENADOS Y RESULTADOS

		ACCURACY	PRECISION	F1 SCORE
01	BOSQUE ALEATORIO	0.89	0.88	0.91
02	MÁQUINA DE VECTORES DE SOPORTE	0.88	0.82	0.88
03	K-NN	0.82	0.82	0.87
04	REGRESION LOGÍSTICA	0.80	0.82	0.84

SELECCIÓN DEL MODELO



Conclusiones



Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Overview | Data | Code | Models | Discussion | Leaderboard | Rules

Dataset Description

Overview

The data has been split into two groups:

- training set (train.csv)
- test set (test.csv)

Files
3 files

Size
93.08 kB

Type
csv

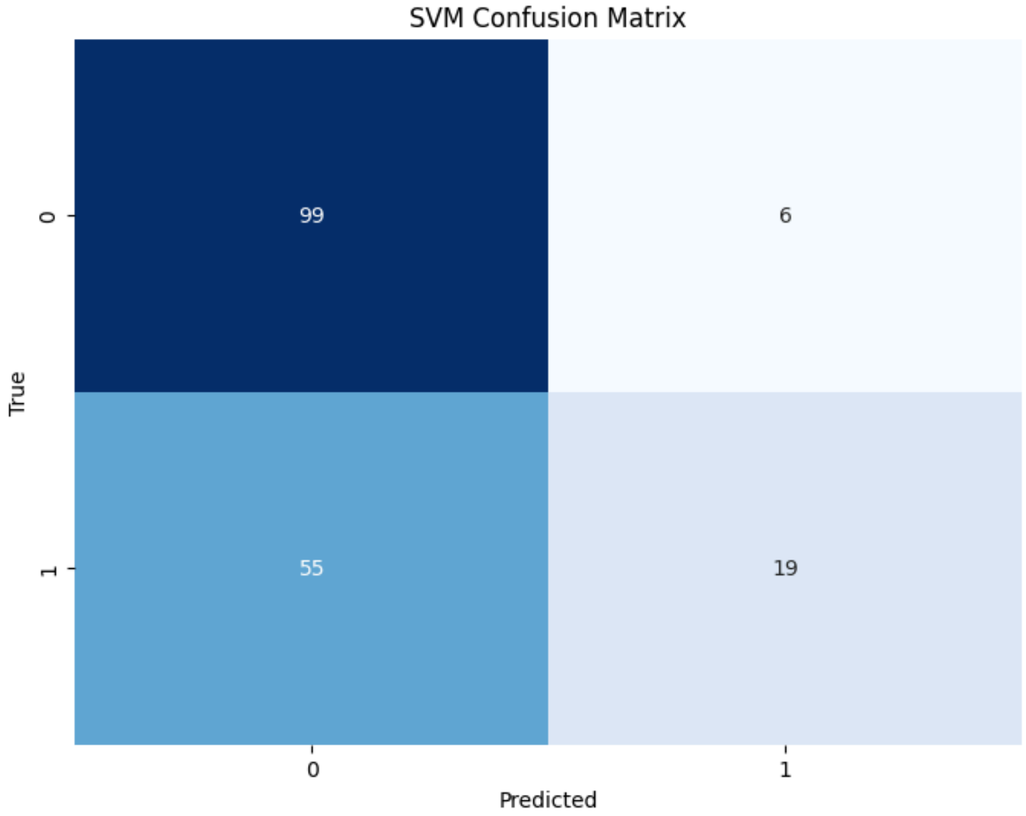
KAGGLE SCORE

Score: 0.77751

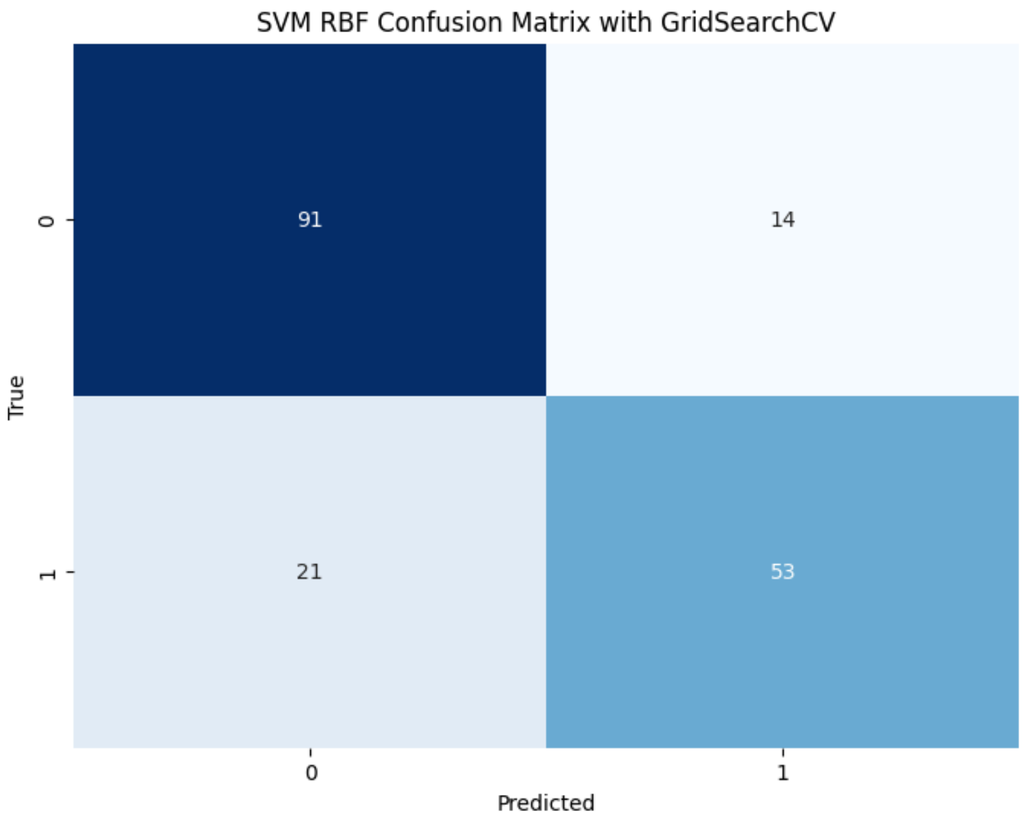
ANEXO

Gráficas que quizas no son relevantes pero para mostrarlas a la hora de preguntas

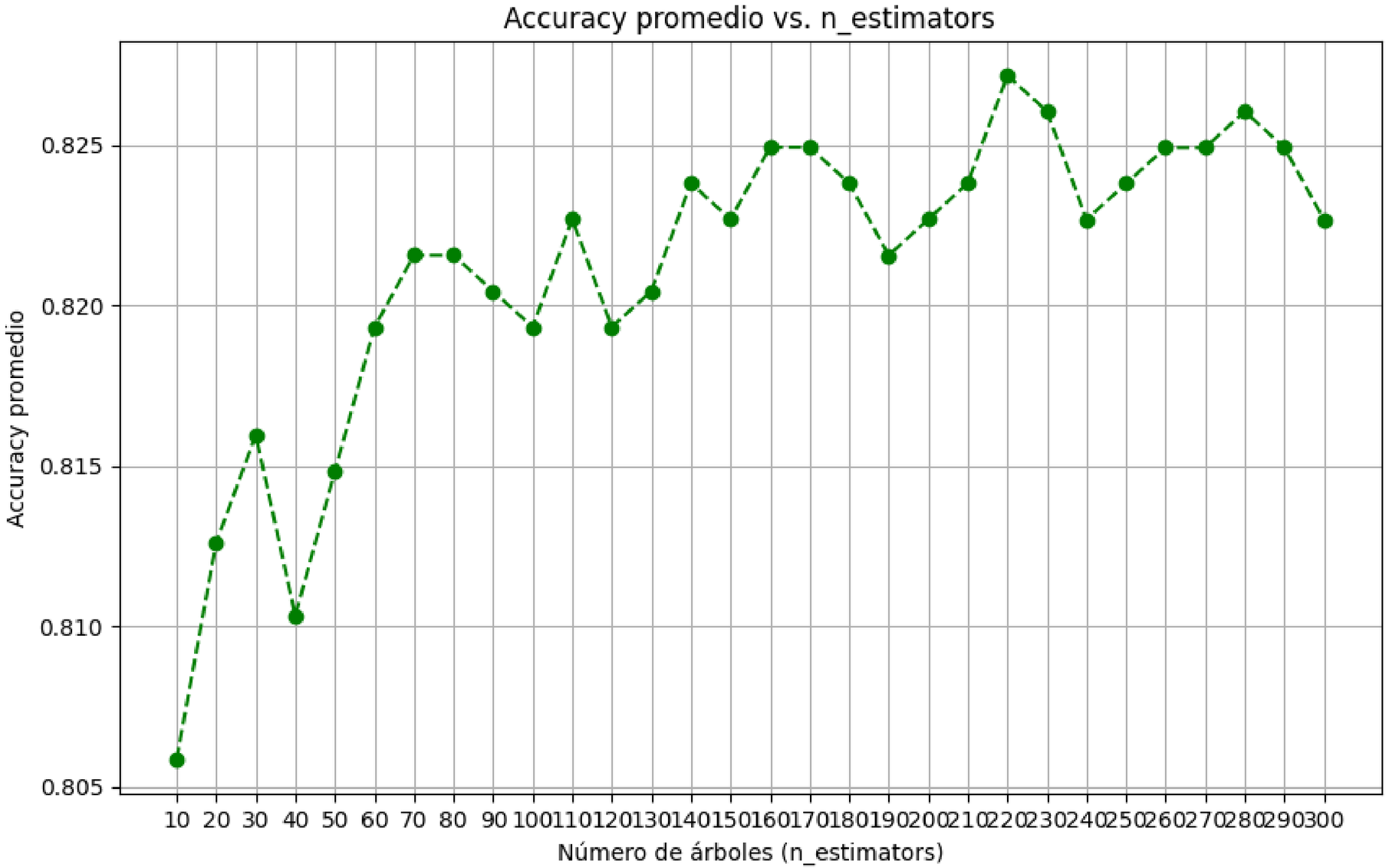
COMPORTAMIENTO DE LAS SVM CON Y SIN KERNEL TRICK



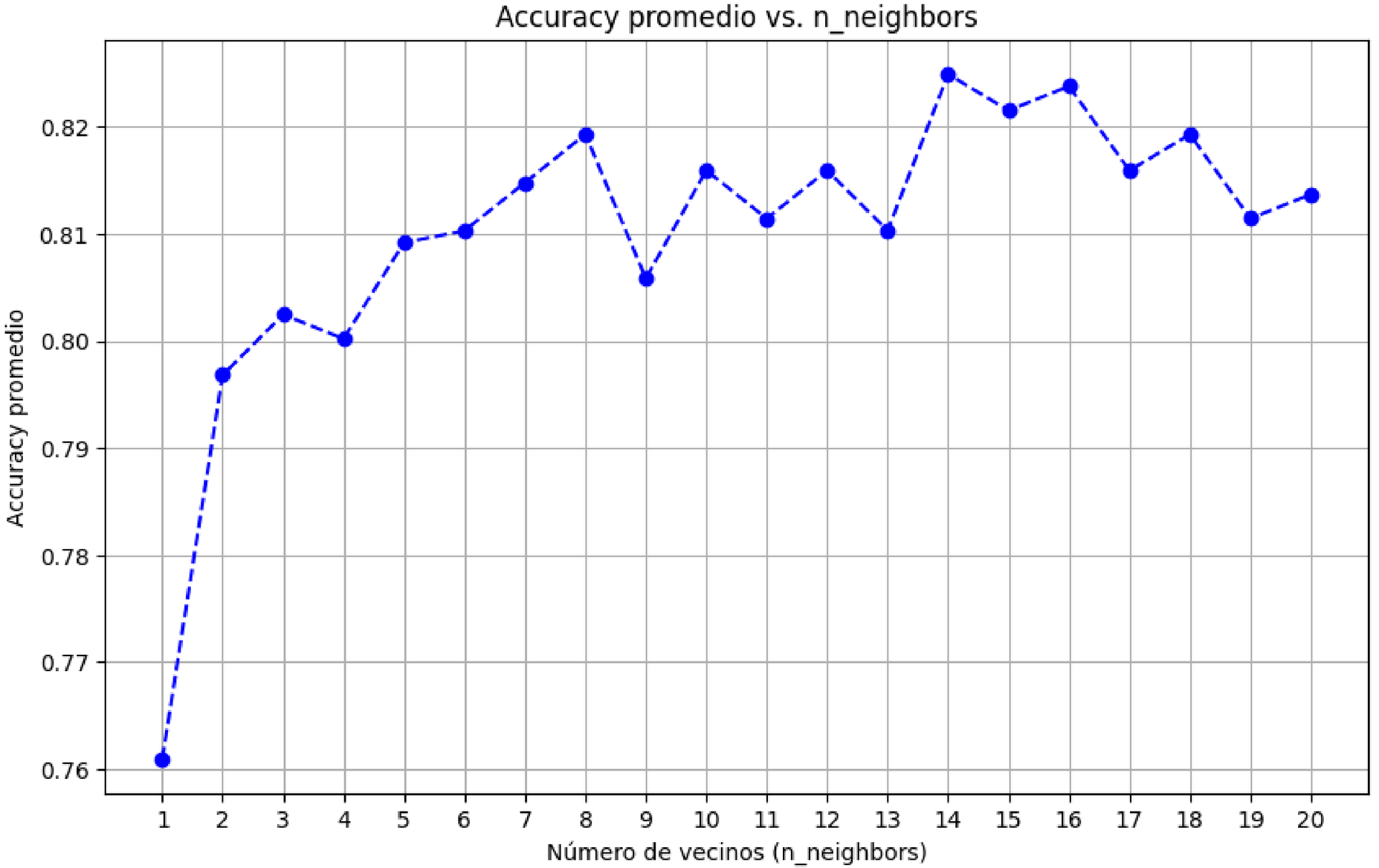
VS



TEST DE RANDOM FOREST



TEST DE KNN


Accuracy promedio vs. n_neighbors
Accuracy promedio
0.82
0.81
0.80
0.79
0.78
0.77
0.76
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Número de vecinos (n_neighbors)