

# PRACTICA APRENDIZAJE SUPERVISADO

**AUTOR:** Guillermo Criado Morato

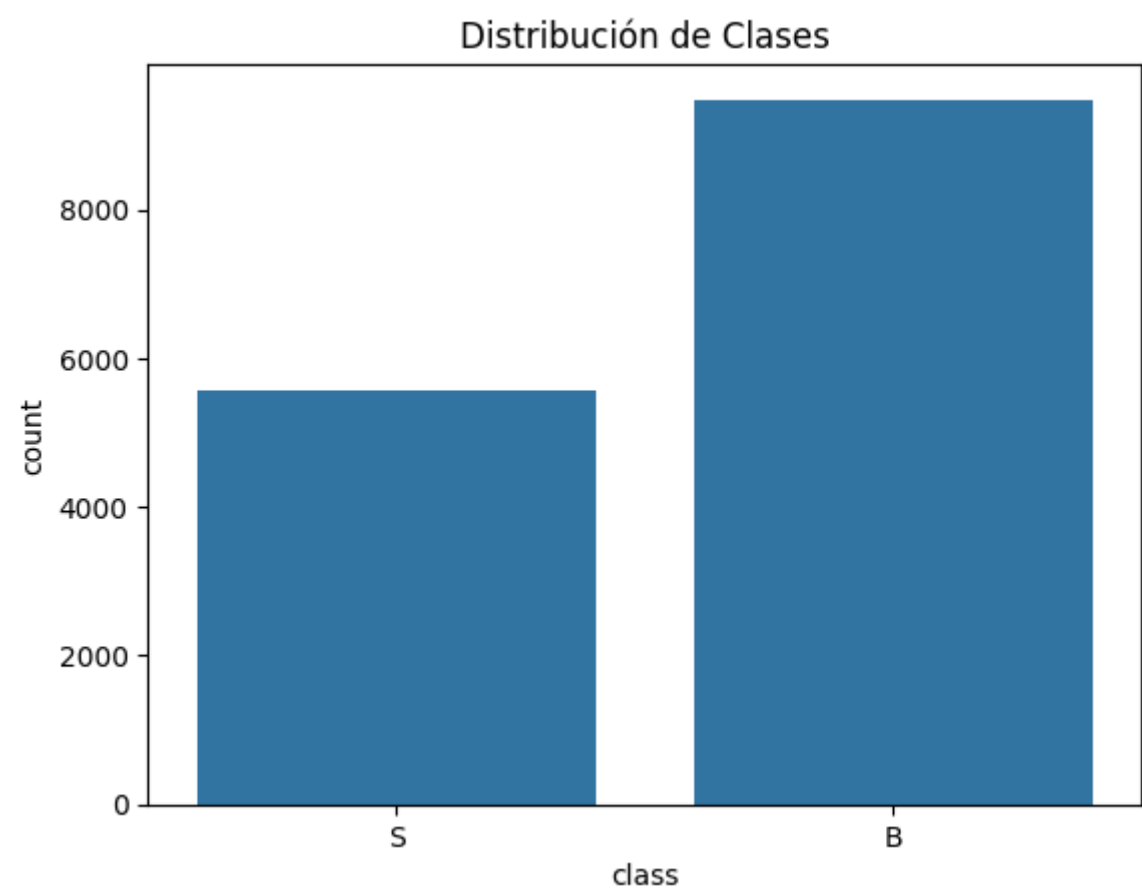
## Ejercicio 1: Detección de Malware

Nos encontramos ante dos dataset uno con las features y otro con los datos.

Tarea 1a: Realiza una descripción del dataset

Analizamos ambos dataset y nos damos cuenta de que es un dataset que en función de que acciones se den se clasifica como malware o no.

Aquí tenemos su distribución:



Tarea 1b: Realiza las tareas de pre-procesamiento

Para las labores de pre-procesamiento hemos tenido que reemplazar los valores no numéricos de la columna "class", también hemos eliminado los valores nulos que están como "?" y para ver la correlación de características hemos hecho la matriz de correlación pero al haber tantas características el dataset la información que nos proporciona es nula así que para la selección de características se ha utilizado SelectKBest y para escalarlas StandarScaler.

Tarea 1c: Entrega y evalúa al menos 3 algoritmos de clasificación

Los algoritmos seleccionados han sido:

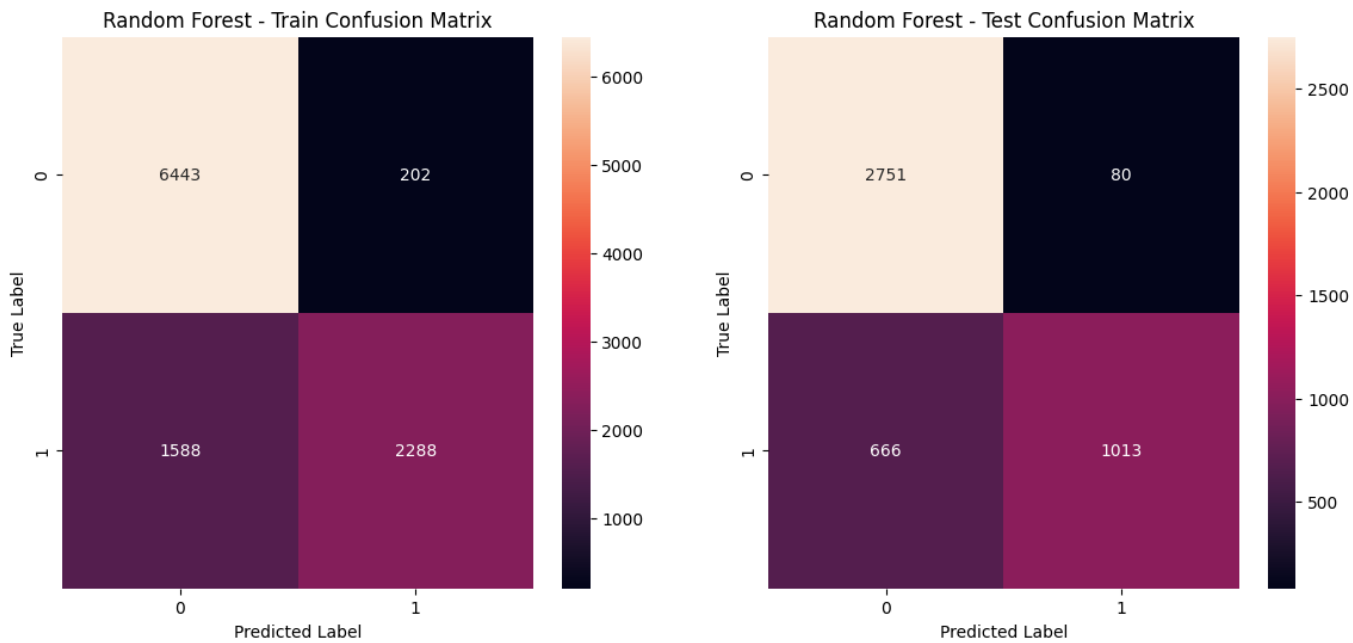
- Random Forest
- Máquina de Soporte Vectorial
- KNN

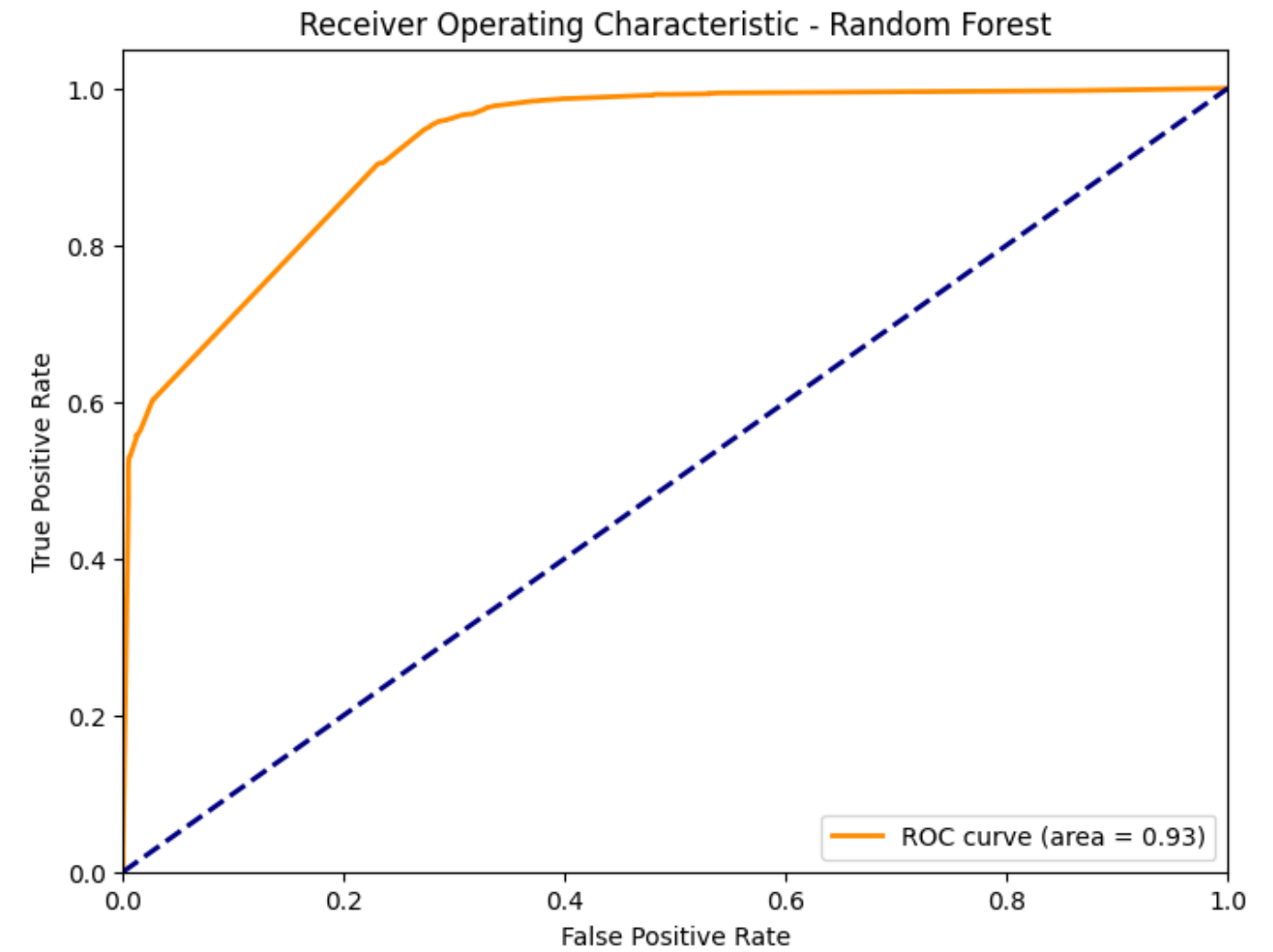
Ya que esytos son los mejores para la selección de características binaria. En el caso de la Máquina de Soporte Vectorial hemos elegido un SVC con kernel lineal.

Además sospechamos de que haya o underfitting u overfitting por lo que mediremos los resultados con el conjunto train y test para cada modelo

Para Random Forest:

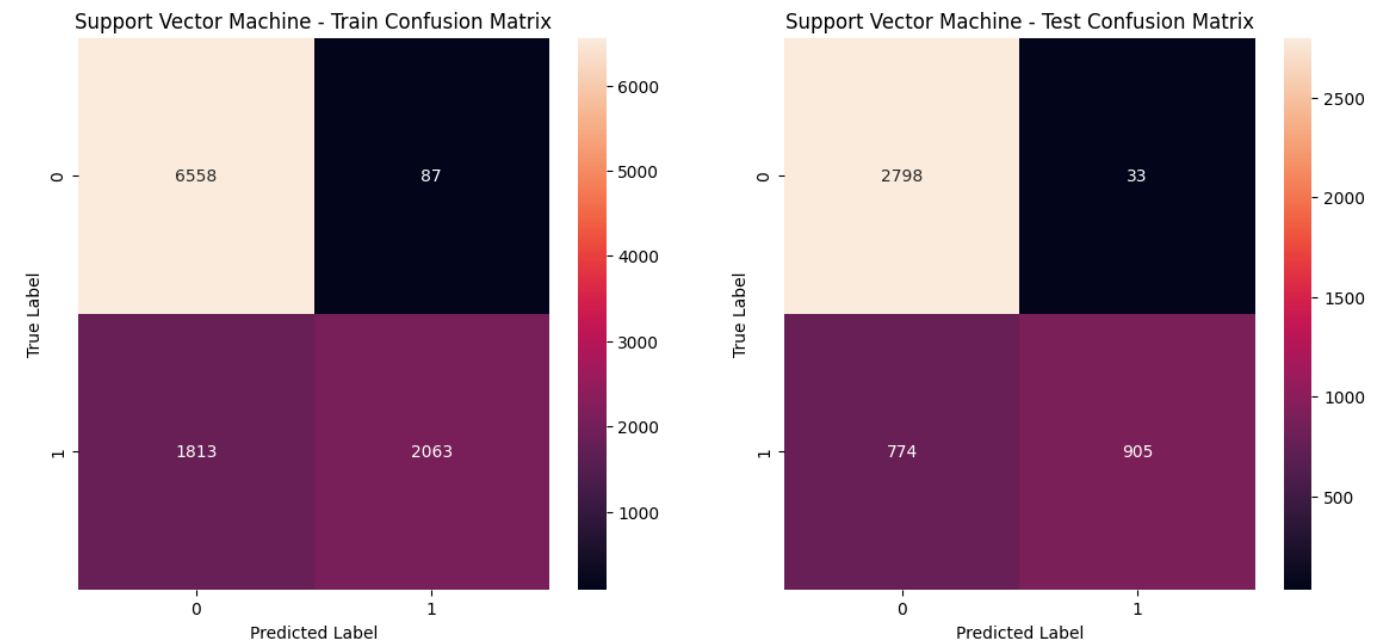
- Train Accuracy: 0.8298640813610874
- Test Accuracy: 0.834589800443459

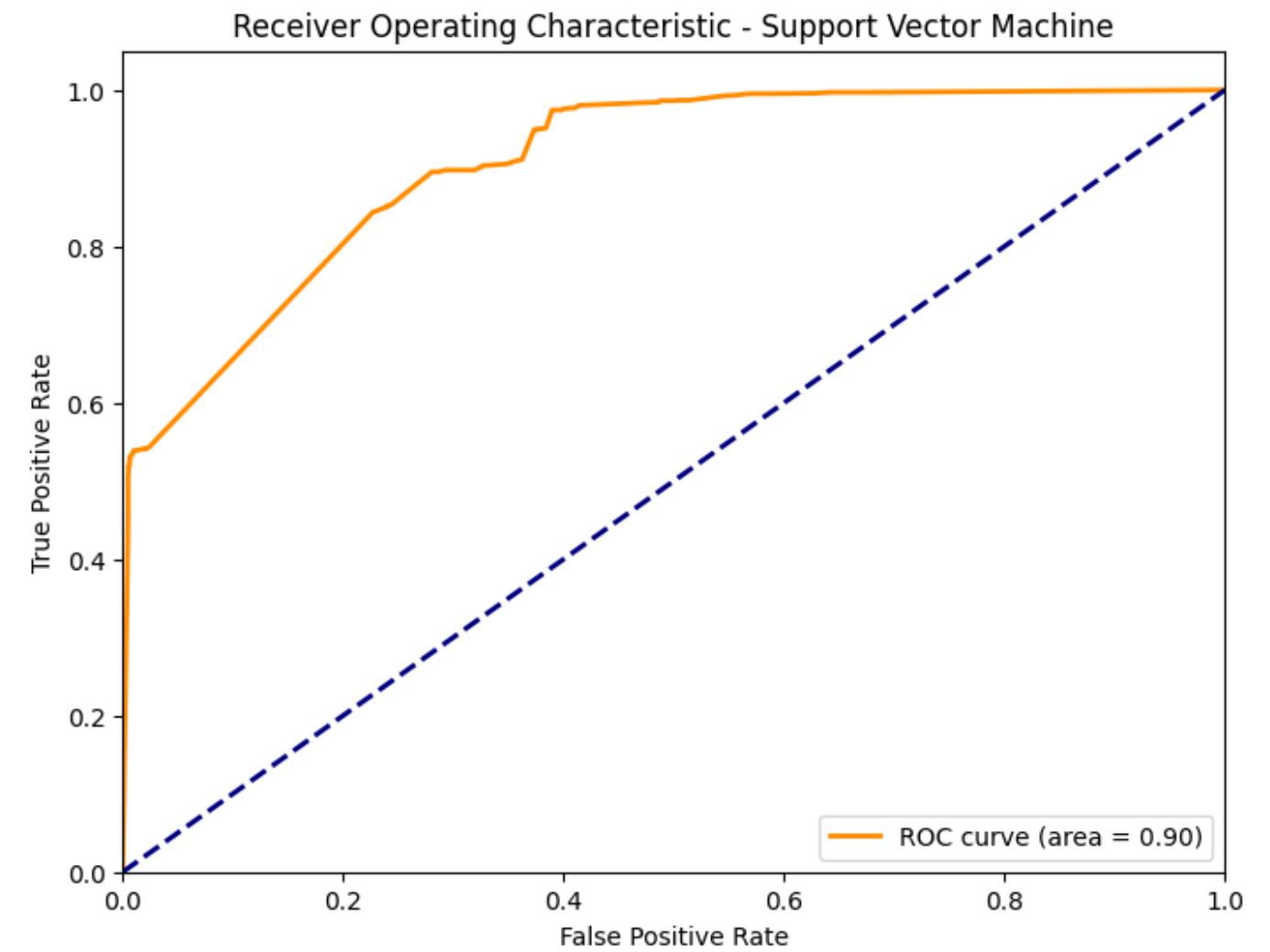




Para Máquina de Soporte Vectorial:

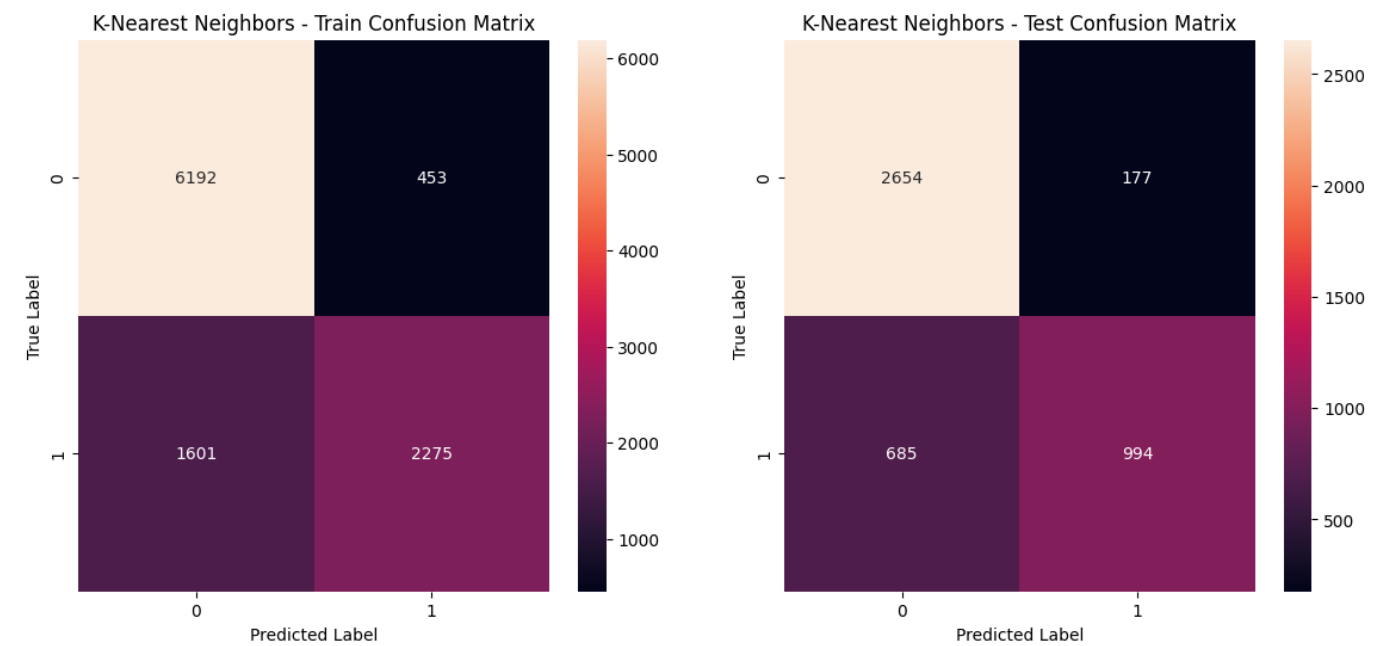
- Train Accuracy: 0.8194088014447296
- Test Accuracy: 0.8210643015521064

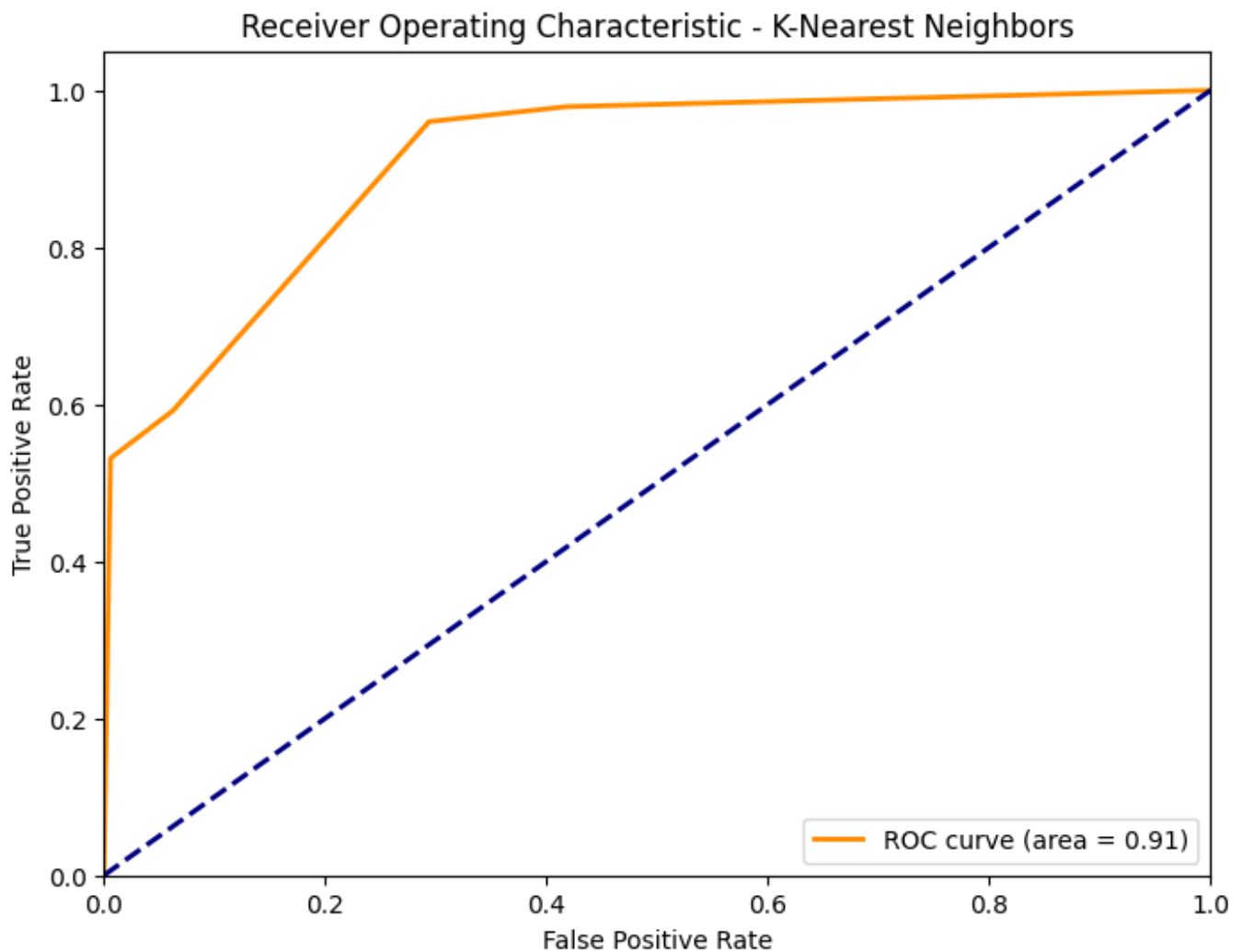




Y para KNN:

- Train Accuracy: 0.8194088014447296
- Test Accuracy: 0.8210643015521064





Siendo el mejor Random Forest y eliminando así las sospechas de underfitting u overfitting ya que tienen resultados similares ambos conjuntos.

## Ejercicio 2 - Clasificación de textos

Tarea 2a: Realiza una descripción del dataset

Nos encontramos frente a un dataset donde se tiene una etiqueta y texto asociado a esa etiquetas. Dichas etiquetas son. 'Emotion' 'Financial' 'Politics' 'Sport' 'Health' 'Science'.

Seguidamente hacemos el preprocesamiento del texto:

- Eliminar caracteres no ASCII
- Convertir a minúsculas
- Eliminar puntuación
- Eliminar espacios adicionales

Y visualizamos las palabras más comunes por tópico:

[illegible][illegible]

A word cloud visualization of terms related to Mendelian diseases. The words are arranged in a circular pattern, with 'people' and 'symptom' being the largest and most central. Other prominent words include 'condition', 'sign', 'patient', 'often', 'diagnosis', 'inheritance', 'disorder', 'found', 'frequency', 'treatment', 'clinical trial', 'fraction', 'genetic testing', 'autosomal dominant', 'autosomal recessive', 'information', 'doctor', 'may', 'life', 'time', 'percentage', 'risk', 'management', 'affect', 'change', 'age', 'mean', 'type', 'form', 'rare disease', 'individual definition', 'feature', 'many people', 'look used', 'body help', 'person', 'phenotype', 'human', 'cause', 'part', 'case', 'table', 'examined', 'may', 'common', 'one study', 'children', 'result', 'ontology', 'phenotype', 'study', 'include', 'protein', 'diagnosis', 'health care', 'caused', 'mutation', 'usually seen skin', 'make', 'listed', 'example', 'brain called', 'use'.

A word cloud visualization of tweets from January 19, 2017. The most prominent words are "donald trump", "president", "trump", "say", "democrat", "republican", "will", "state", "new", "gop", "said", "one", "campaign", "report", "know", "year", "obama", "people", "american", "want", "election", "law", "going", "keep", "group", "claim", "look", "help", "called", "many", "way", "think", "women", "supreme", "court", "voter", "hillary", "clinton", "may", "two", "make", "congress", "first", "use", "still", "much", "right", "country", "case", "win", "don't", "face", "war", "million", "white", "house", "call", "health", "care", "bill", "plan", "give", "work", "official", "see", "hes", "senate", "political", "stop", "even", "day", "former", "question", "now", "debate", "take", "america", "time", "fight", "job", "back", "bernie", "sanders", "support", "party", "attack", "end", "week", "senator", "need", "change", "u", "govern", "ment". The words are arranged in a dense, overlapping manner, with colors ranging from dark blue to light green.



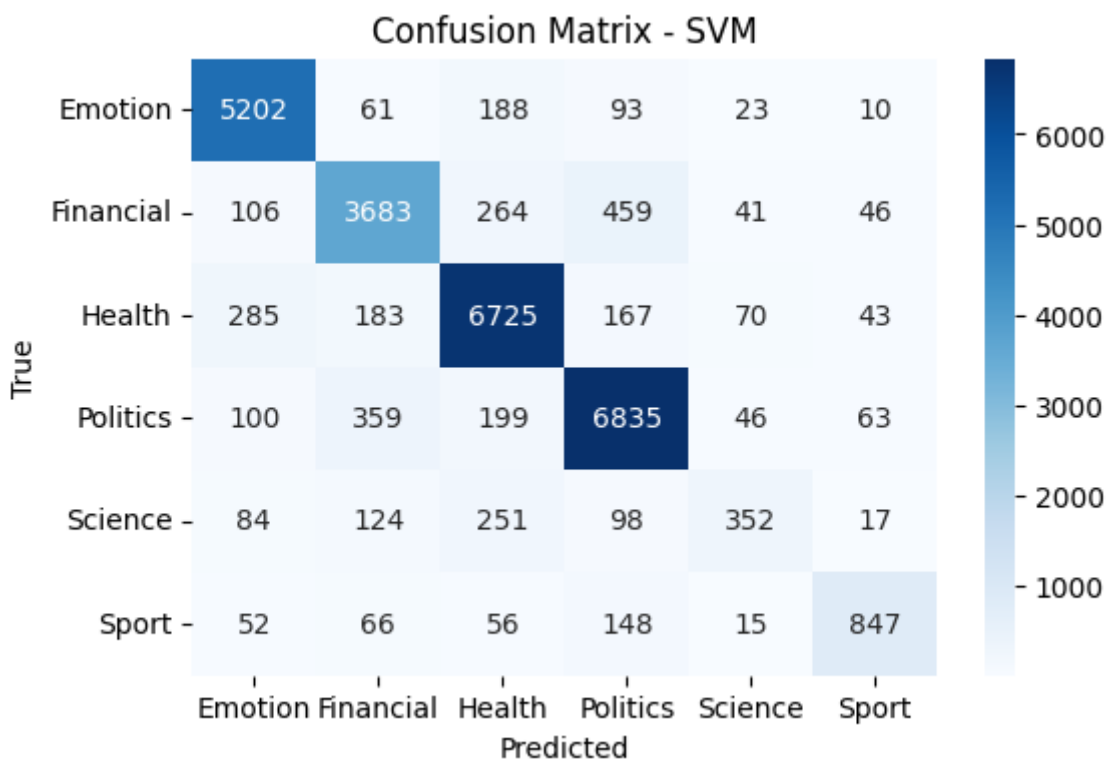
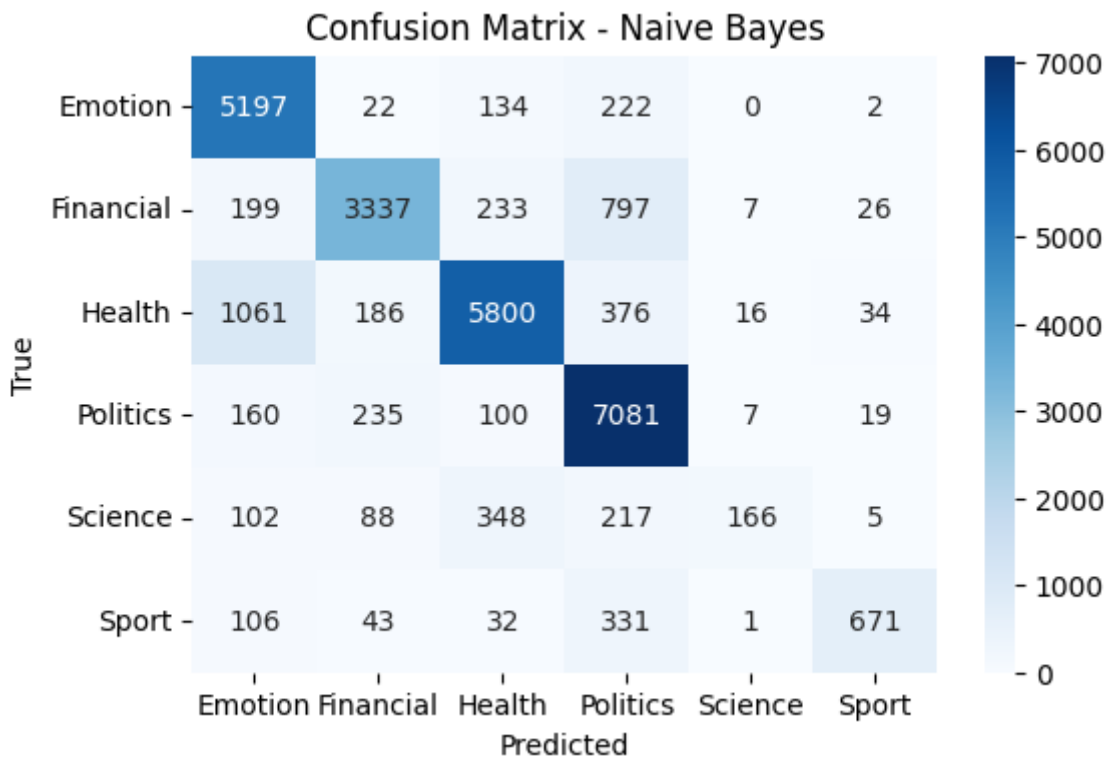
[illegible][illegible]

Se hace la representación de textos mediante TF\_IDF seleccionando solo las 4000 palabras más significativas

Se eligen Navie Bayes y LinearSVC obteniendo los siguientes resultados:

- 8 / 15



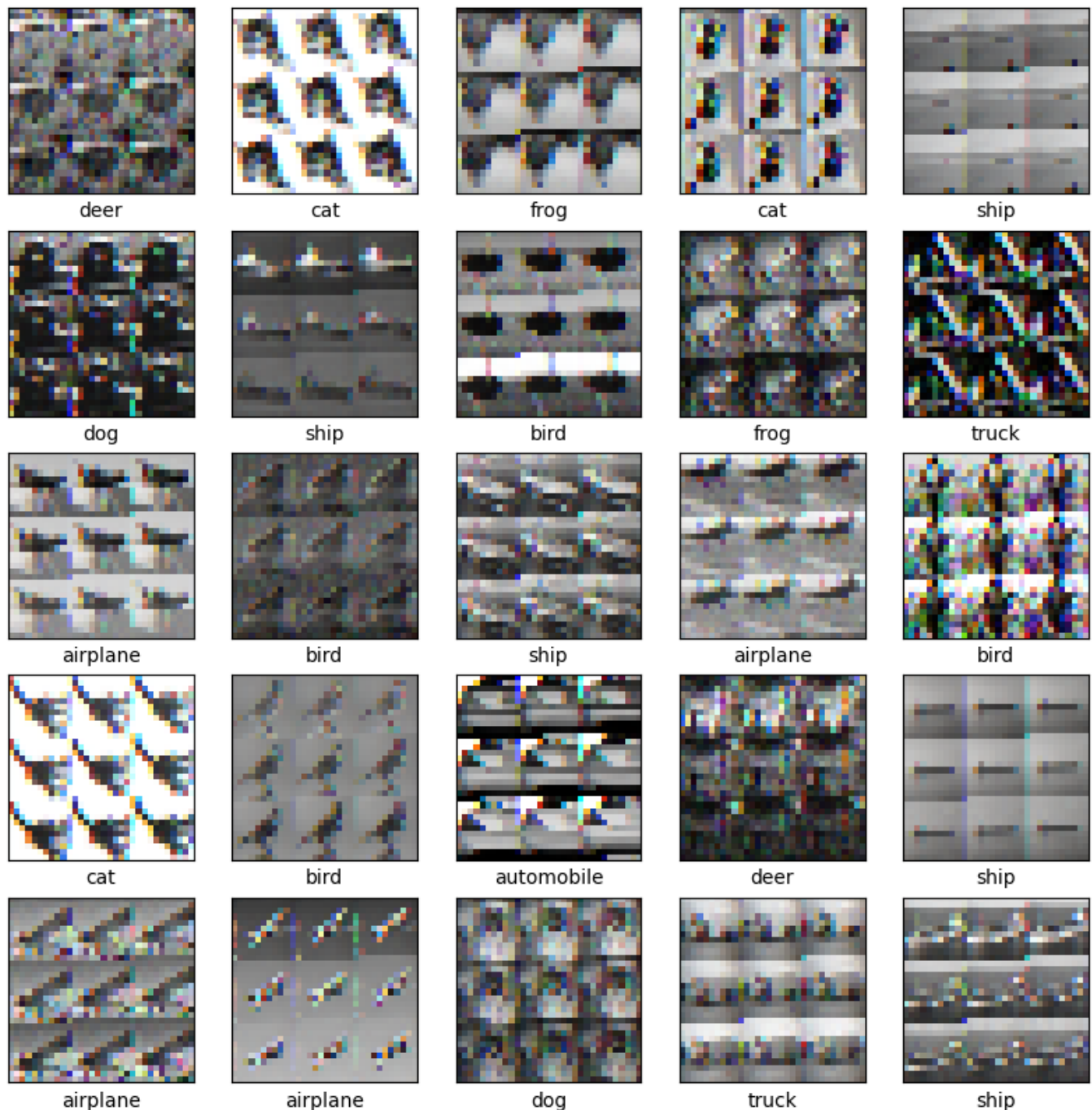


Siendo LinearSVC la mejor de los algoritmos propuestos

### Ejercicio 3 – Clasificación de imágenes

Tarea 3a: Realiza una descripción del dataset.

Nos encontramos ante un dataset donde se tienen imágenes de diferentes objetos. Se normaliza los valores de los píxeles de las imágenes y se visualiza las diferentes imágenes. Al dividir en conjuntos se hace más pequeño para la SVM.



### Tarea 3b: Entrega y evalúa al menos 3 algoritmos de clasificación

Los algoritmos elegidos son:

- KNN
- SVM con PCA
- Redes Neuronales Convolucionales (CNN)

Con sus resultados:

- KNN accuracy: 0.30
- SVM con PCA accuracy: 0.31
- Redes Neuronales Convolucionales (CNN) accuracy: 0.5440000295639038

Esto se debe a que las imágenes están muy distorsionadas al normalizarse.

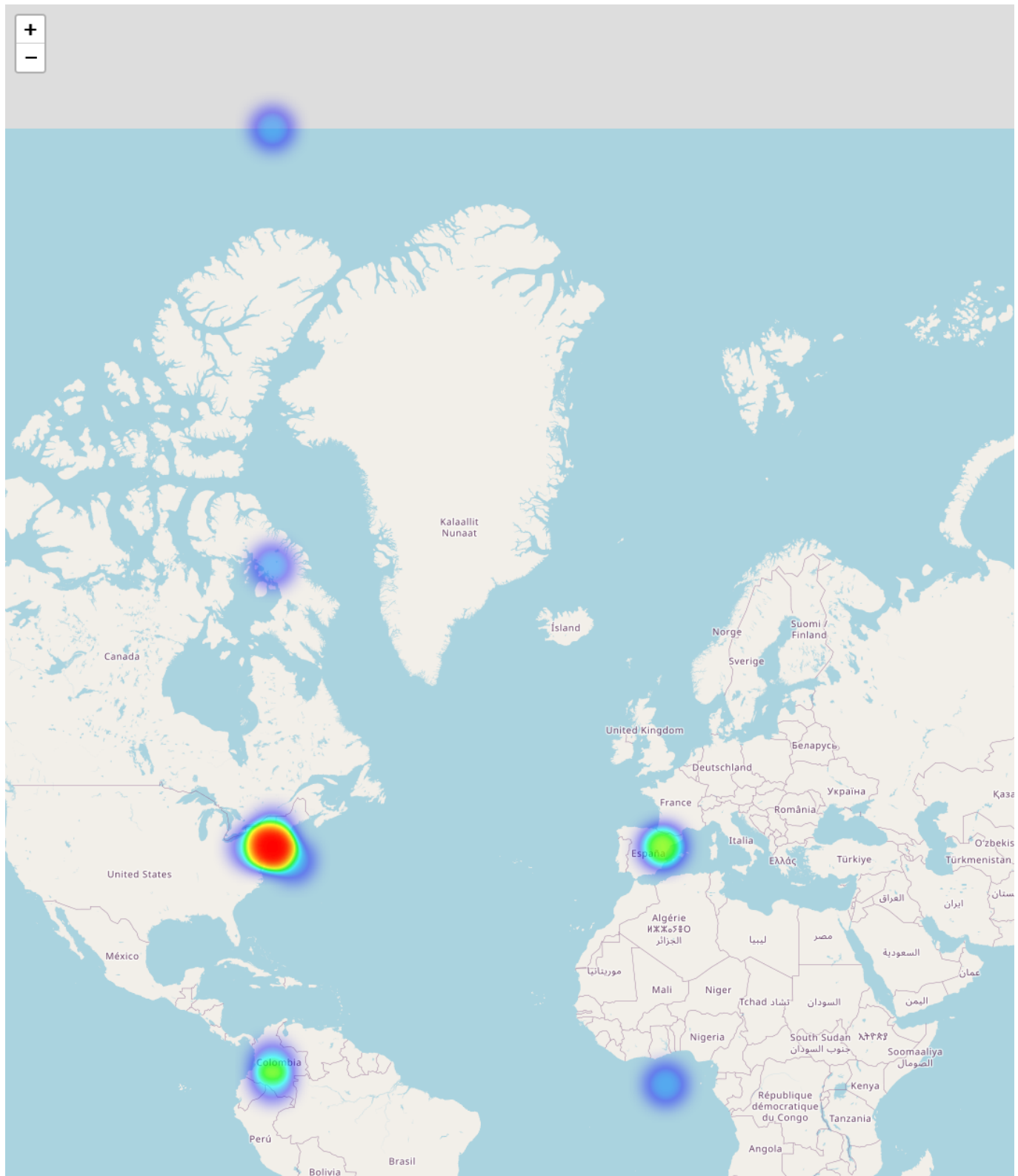
## Ejercicio 4 : Predicción

### Tarea 4a: Realiza una descripción del dataset

Nos encontramos ante un dataset donde se guarda lo que cobra el taxi, si lugar de recogida y su lugar de destino además del día y hora de recogida en Nueva York

### Tarea 4b: Realiza las tareas de pre-procesamiento

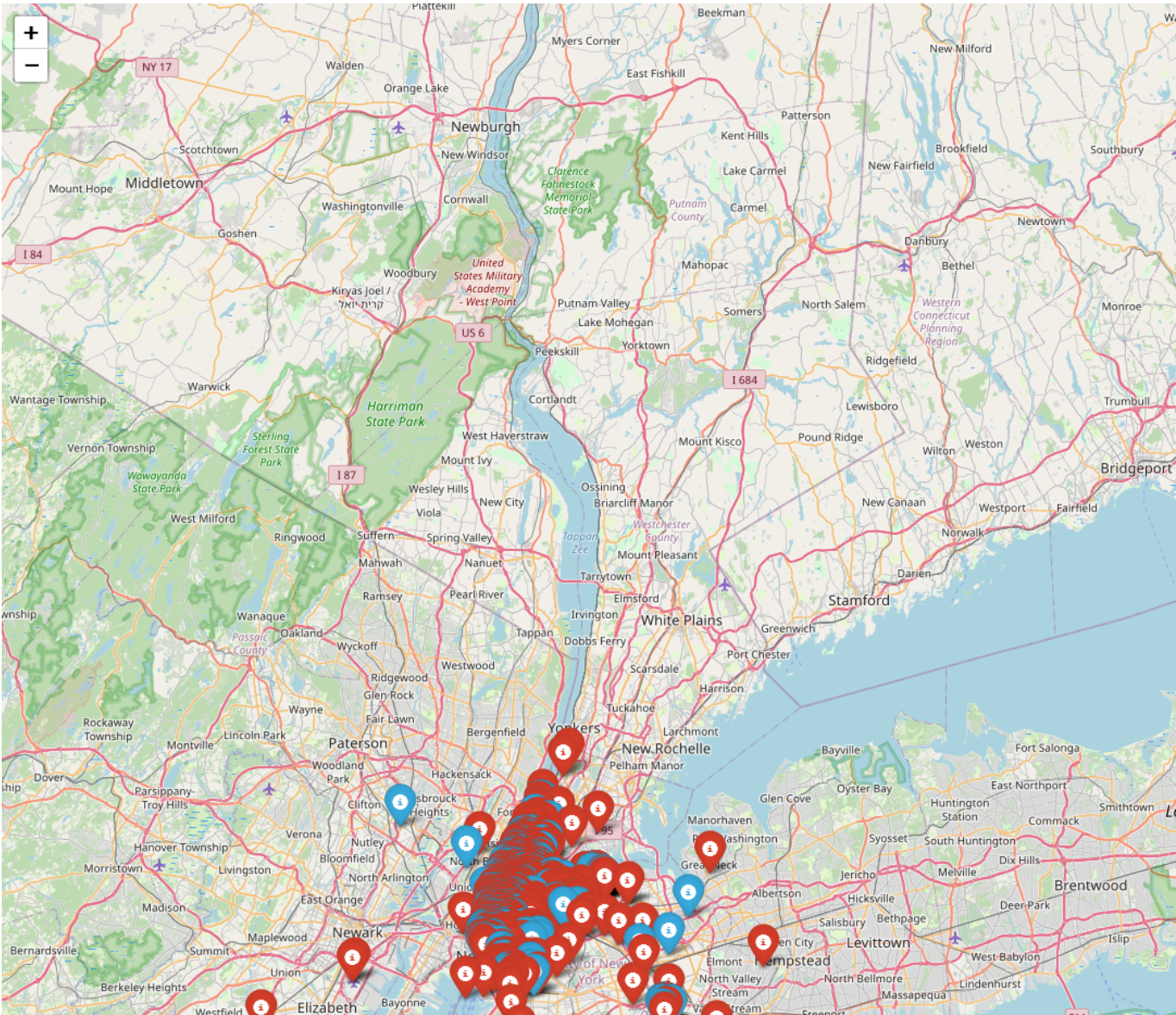
Primero se elimina el valor "key" ya que no sirve para el entrenamiento. Luego visualizamos los valores de las coordenadas de recogida y de destino para detectar valores erróneos.

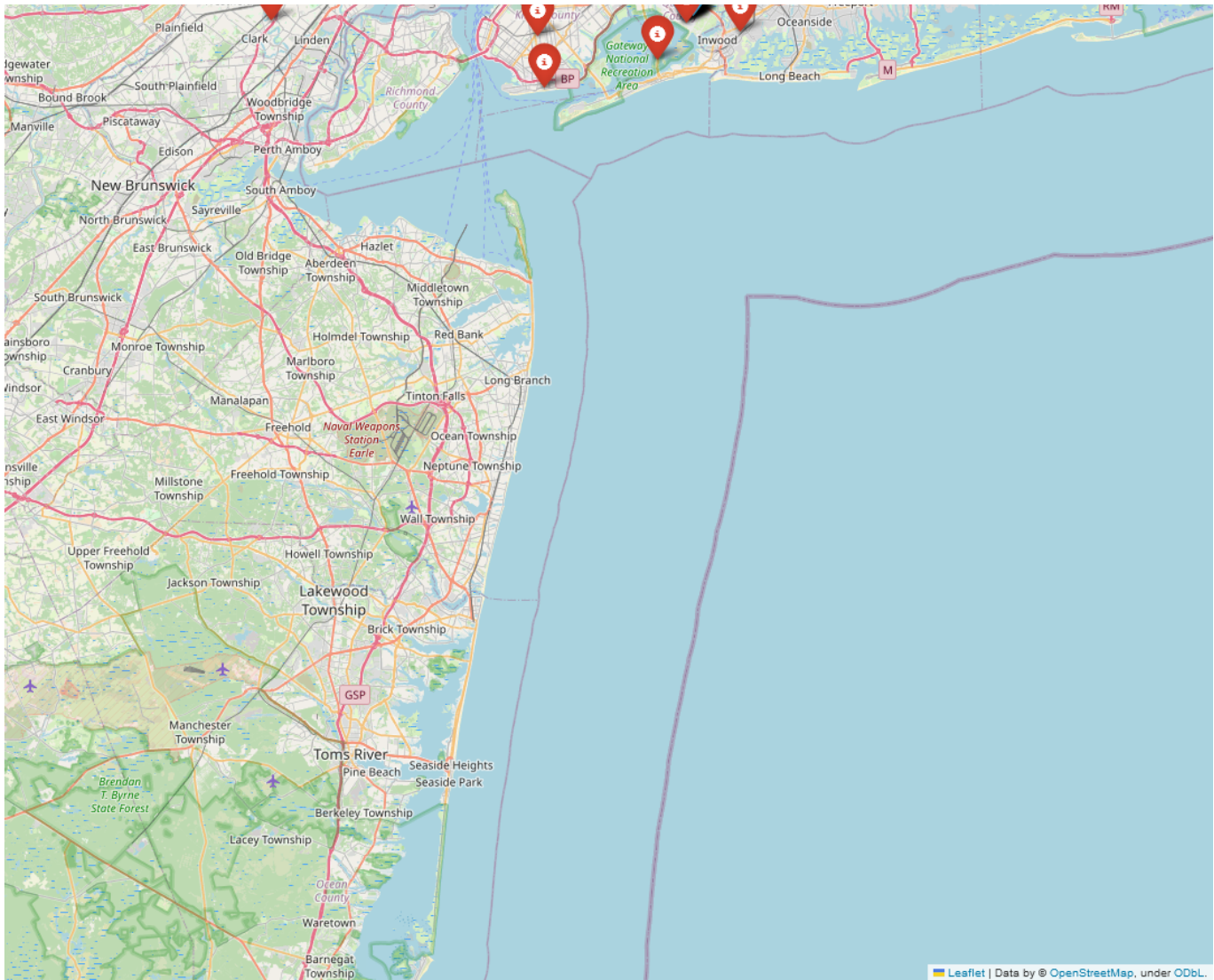




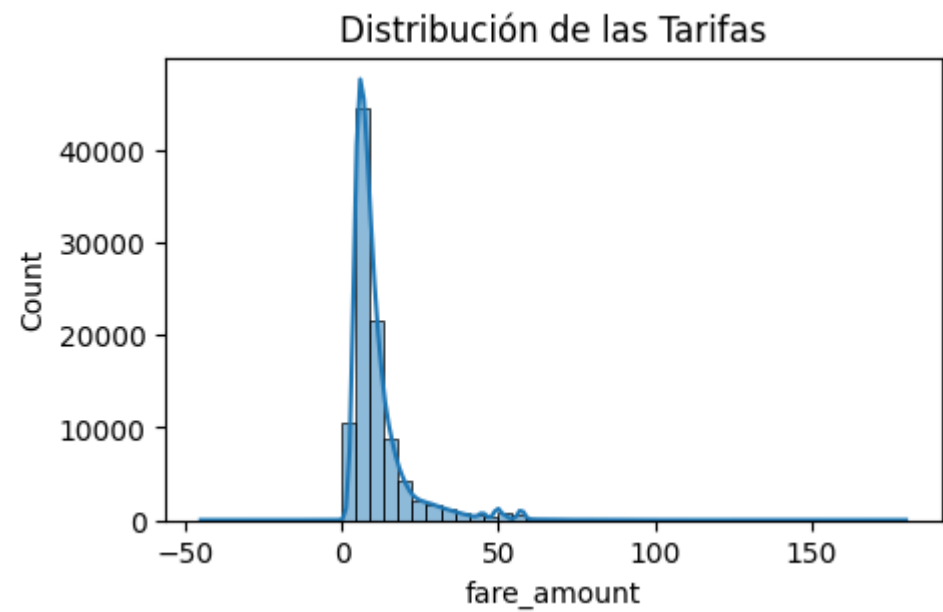


Como se aprecia en la imagen se tienen coordenadas de localizaciones diferentes a Nueva York, para ello establecemos como centro el centro de Nueva York y eliminamos las localizaciones que estén fuera de un radio de 50km.

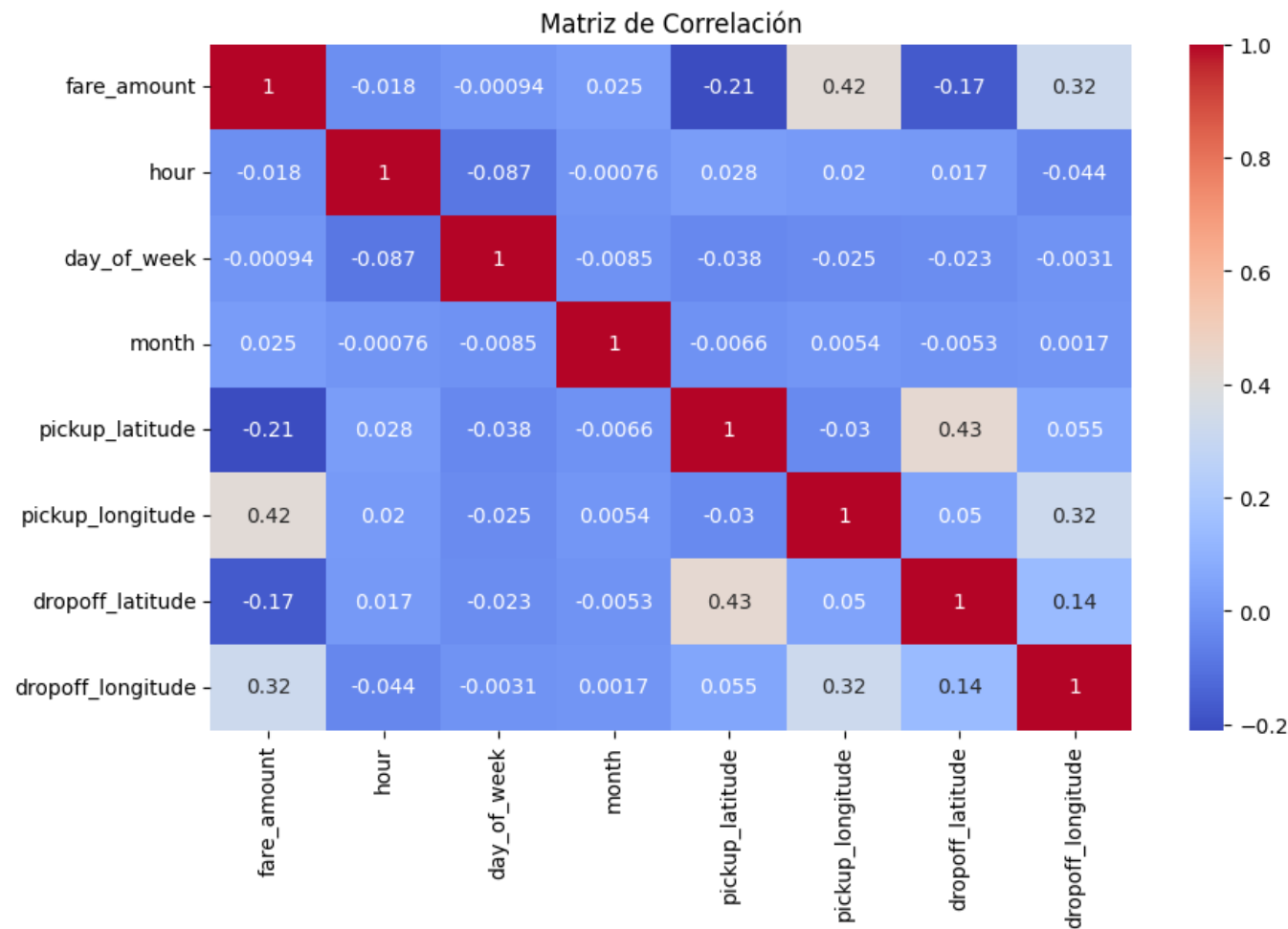




Seguidamente, convertimos la fecha y hora a valores numéricos y vemos la distribución de la tarifa que tiene outliers para ello a la hora de entrenar pondremos como máximo 100\$ e iremos banajando para evitar el efecto de los outliers.



Finalmente vemos la correlación entre las variables:



Las coordenadas son las que tienen mejor correlación con la tarifa.

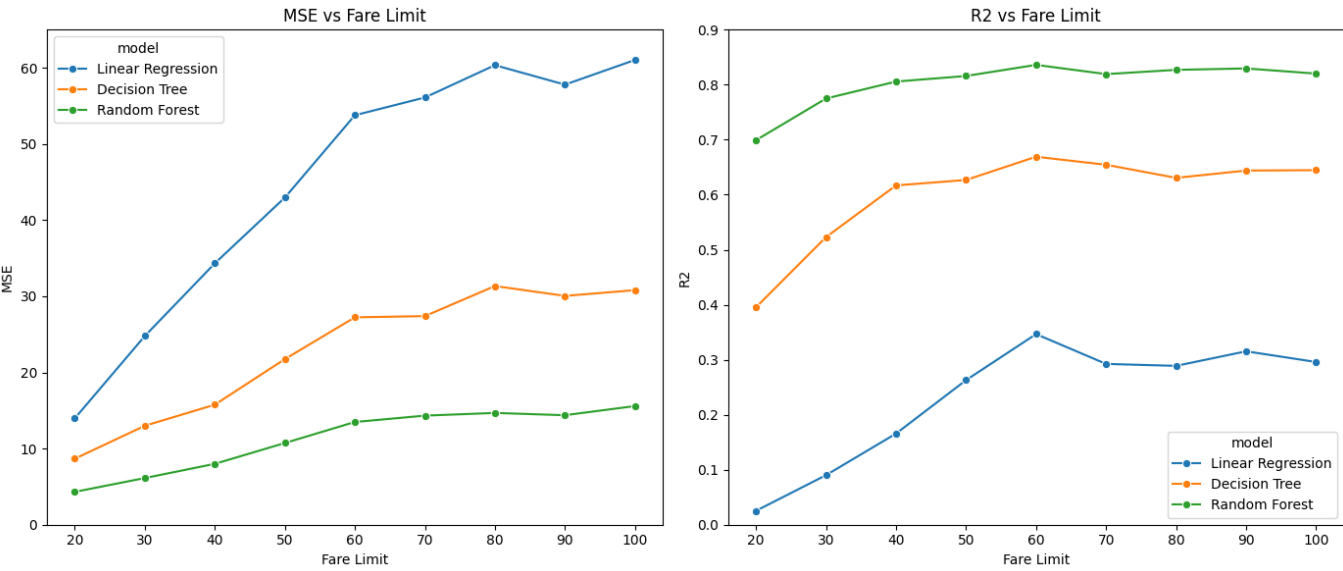
Tarea 4c: Entrega y evalúa al menos 3 algoritmos para predicción

Los algoritmos de predicción son:

- "Linear Regression"
- "Decision Tree"
- "Random Forest":

Obtenniendo los siguientes resultados:





Se puede observar que el mejor modelo es Random Forest ya que tiene menos porcentaje de error (MSE)