PRACTICA APRENDIZAJE SUPERVISADO

AUTOR: Guillermo Criado Morato

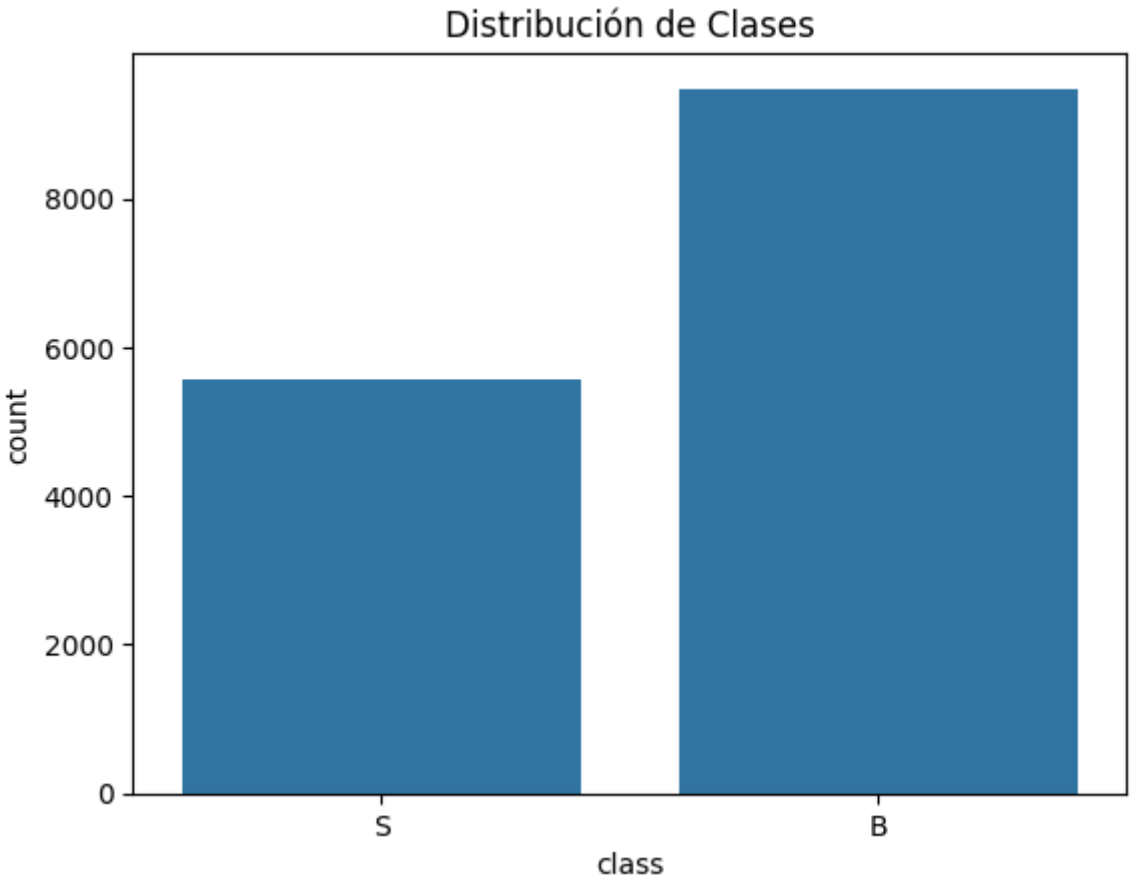
Ejercicio 1: Detección de Malware

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.

Aqui 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 estban 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:

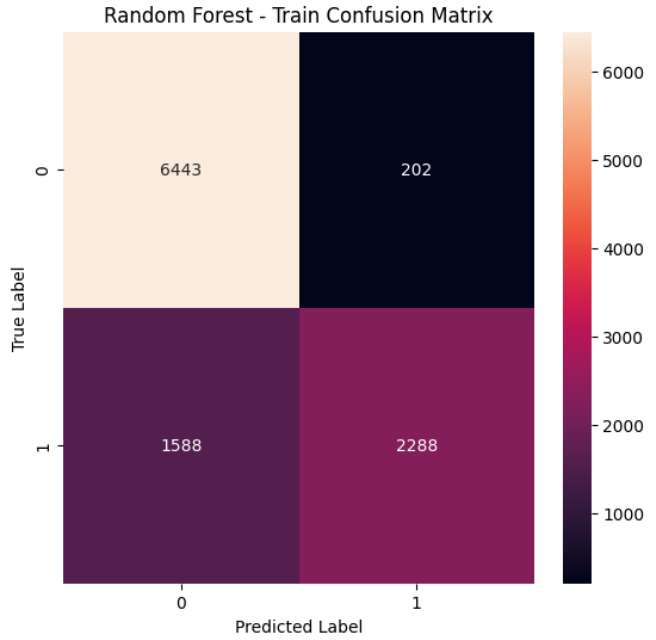
- Random Forest
- Máquina de Soporte Vectorial
- KNN

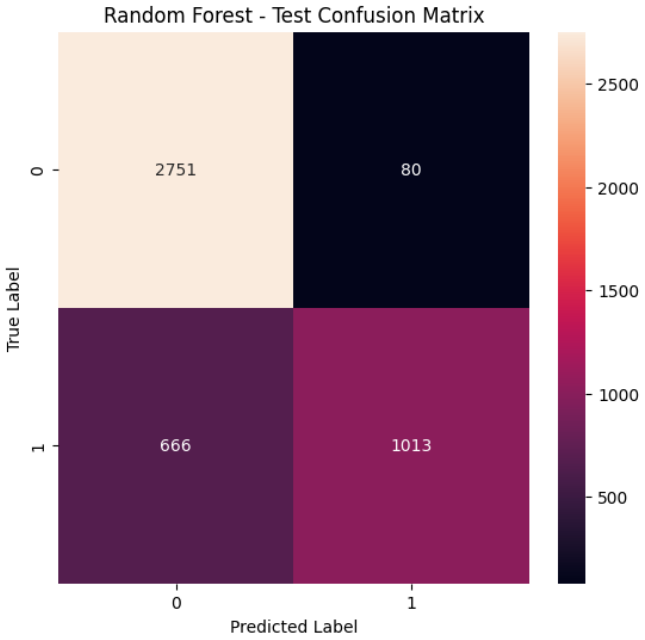
Ya que esytos son los mejores para la selección de características binaria. En el caso de la Máquina de Soporte Vectorial hemos elegido un SVC con kernel lineal.

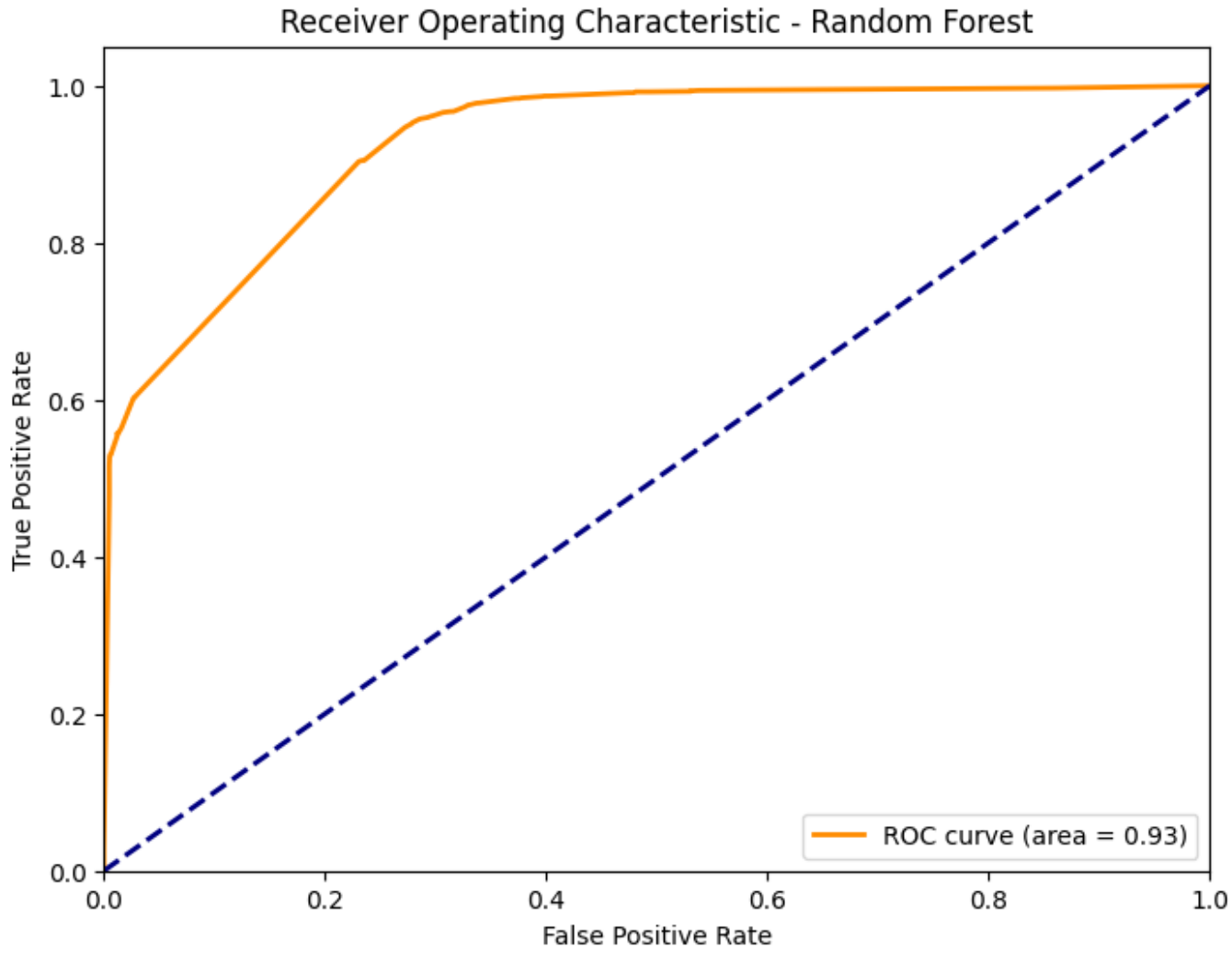
Además sospechamos de que haya o underfitting u overfitting por lo que mediremos los resultados con el conjunto train y test para cada modelo

Para Random Forest:

- Train Accuracy: 0.8298640813610874
- Test Accuracy: 0.834589800443459

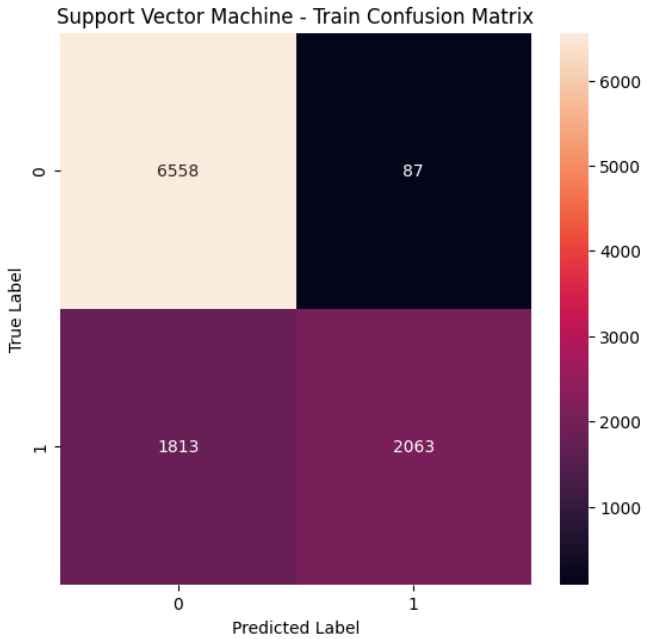


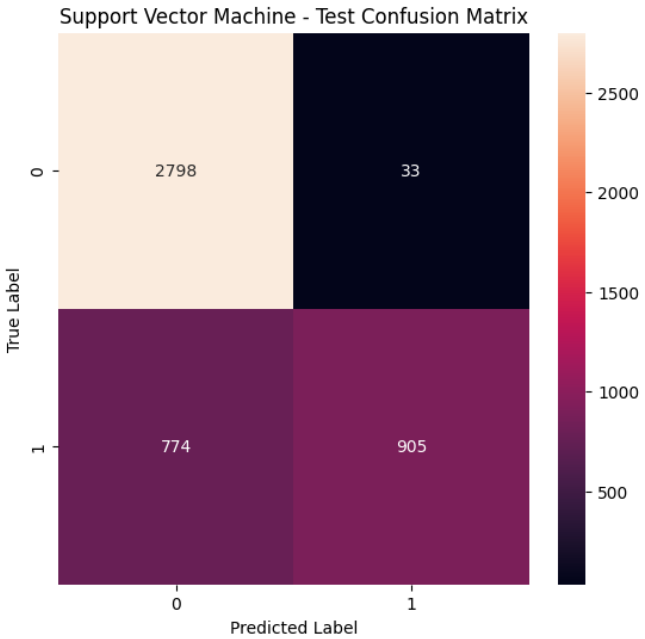


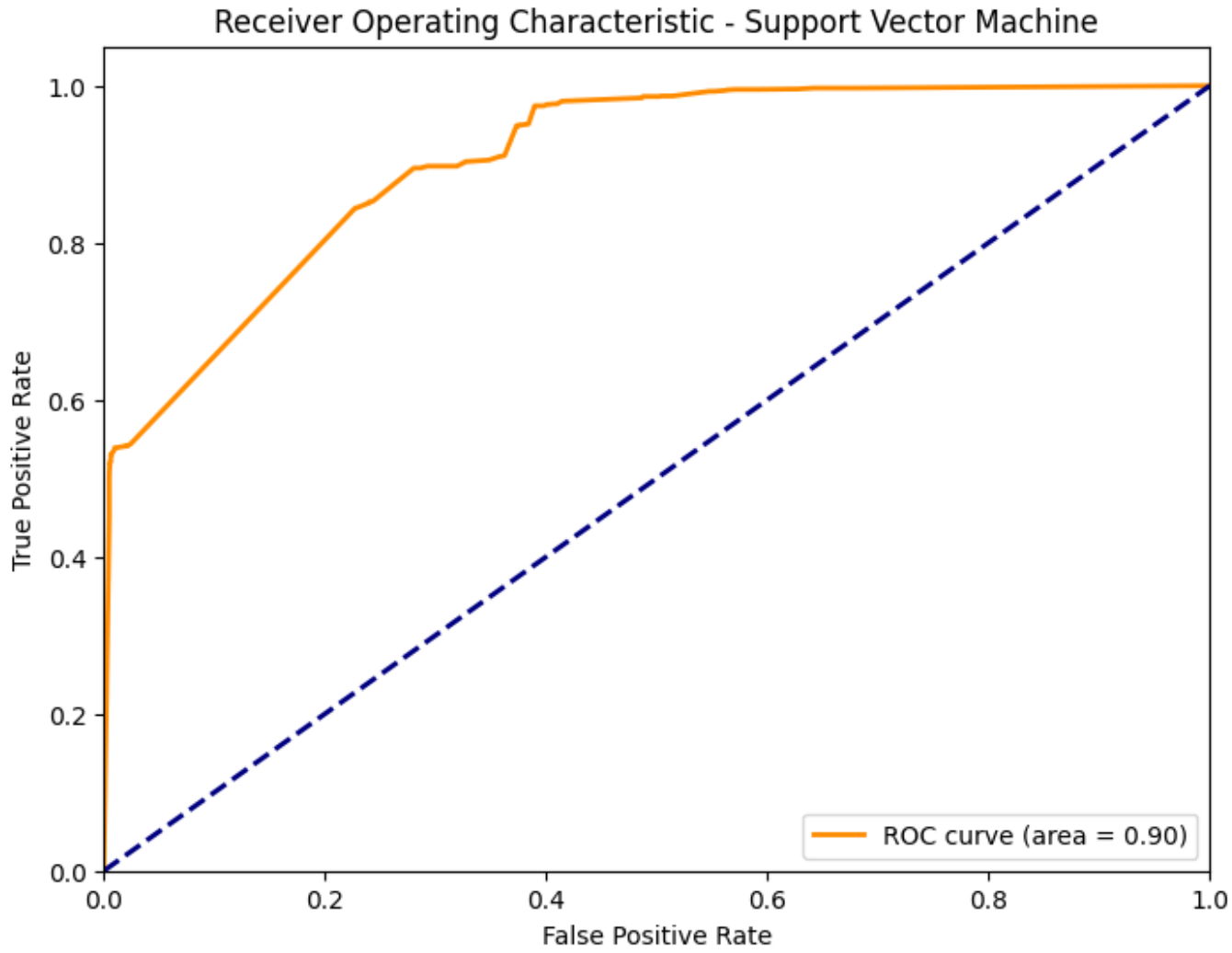


Para Máquina de Soporte Vectorial:

- Train Accuracy: 0.8194088014447296
- Test Accuracy: 0.8210643015521064

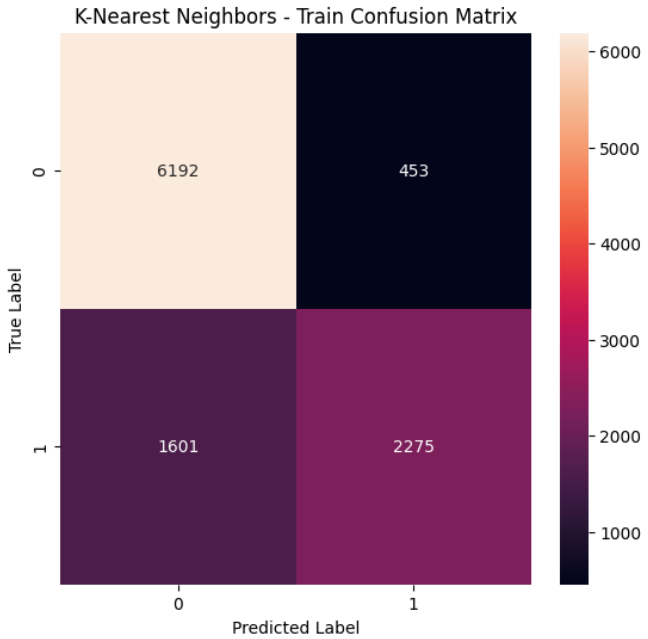


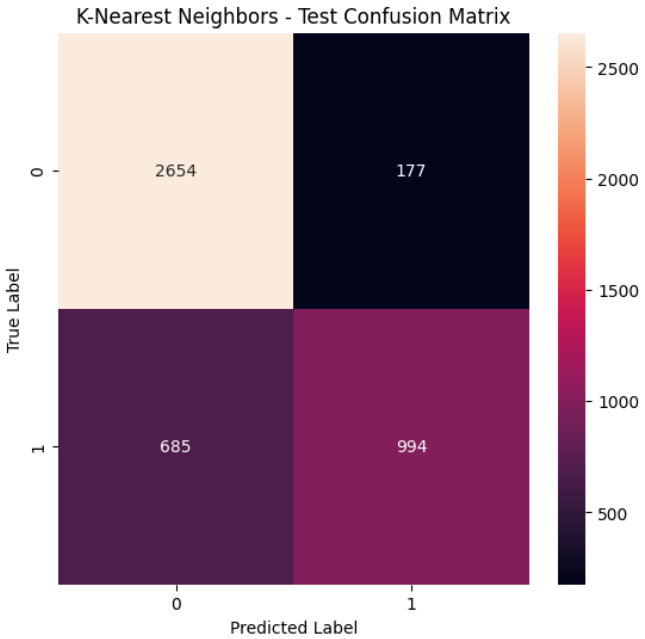


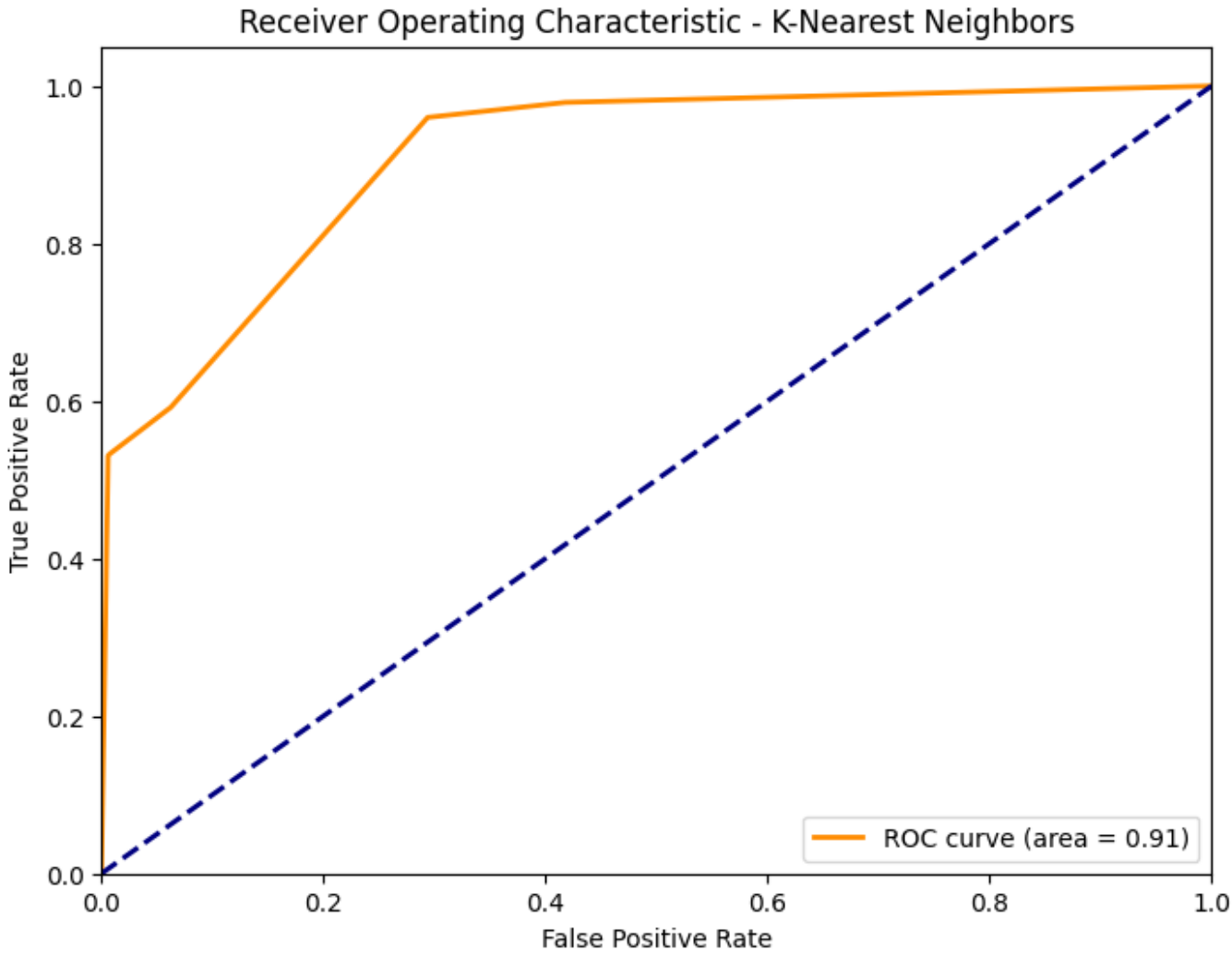


Y para KNN:

- Train Accuracy: 0.8194088014447296
- Test Accuracy: 0.8210643015521064







Siendo el mejor Random Forest y eliminando asi 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

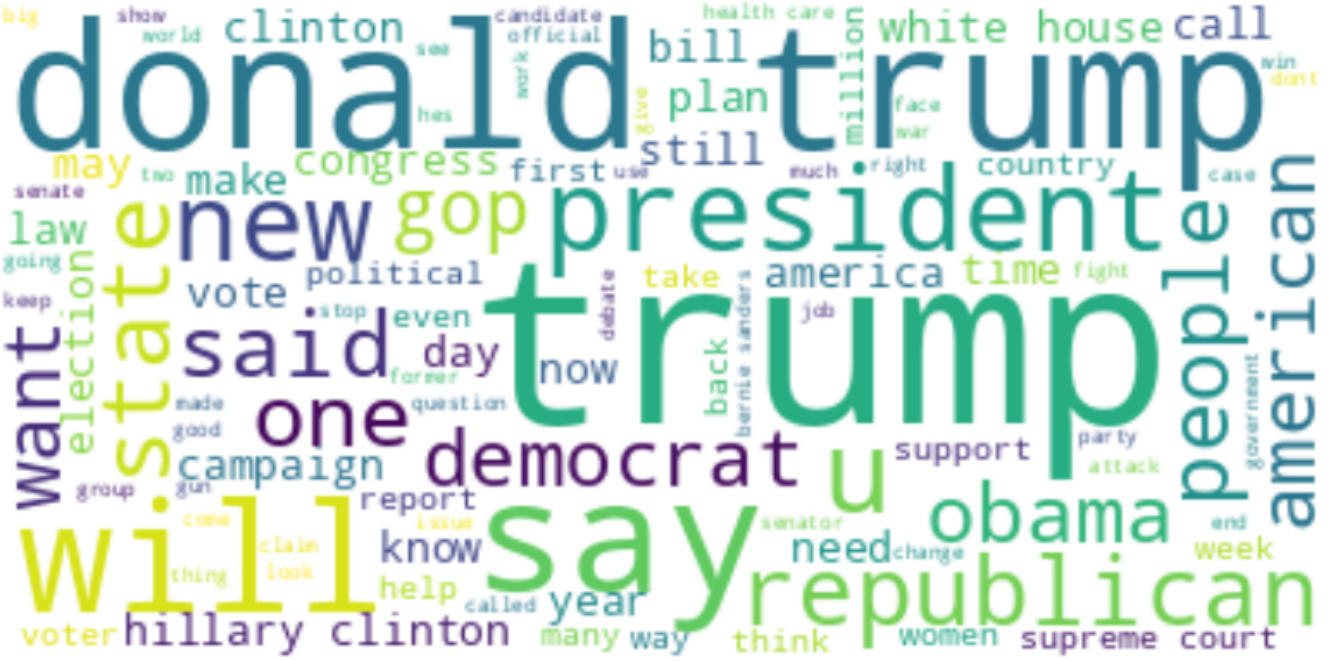
Y visualizamos las palabras más comunes por tópico:

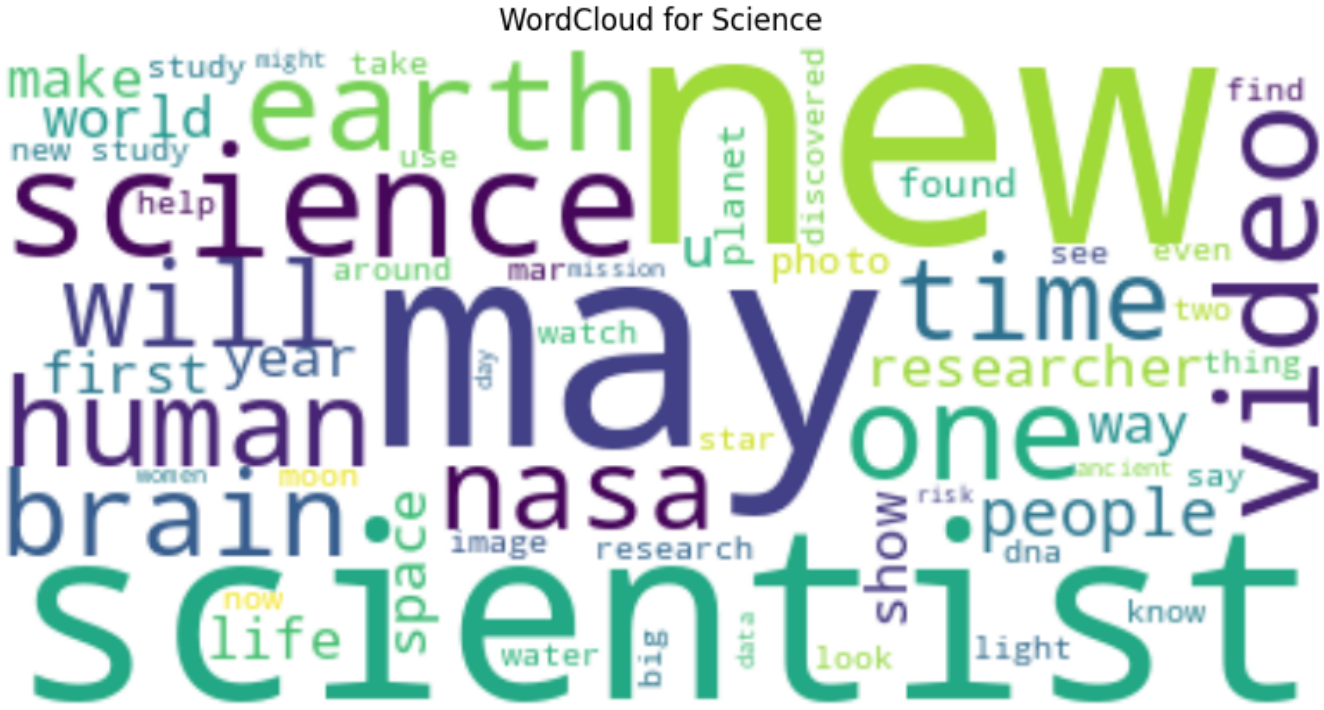
WordCloud for Emotion

WordCloud for Financial

WordCloud for Health

WordCloud for Politics





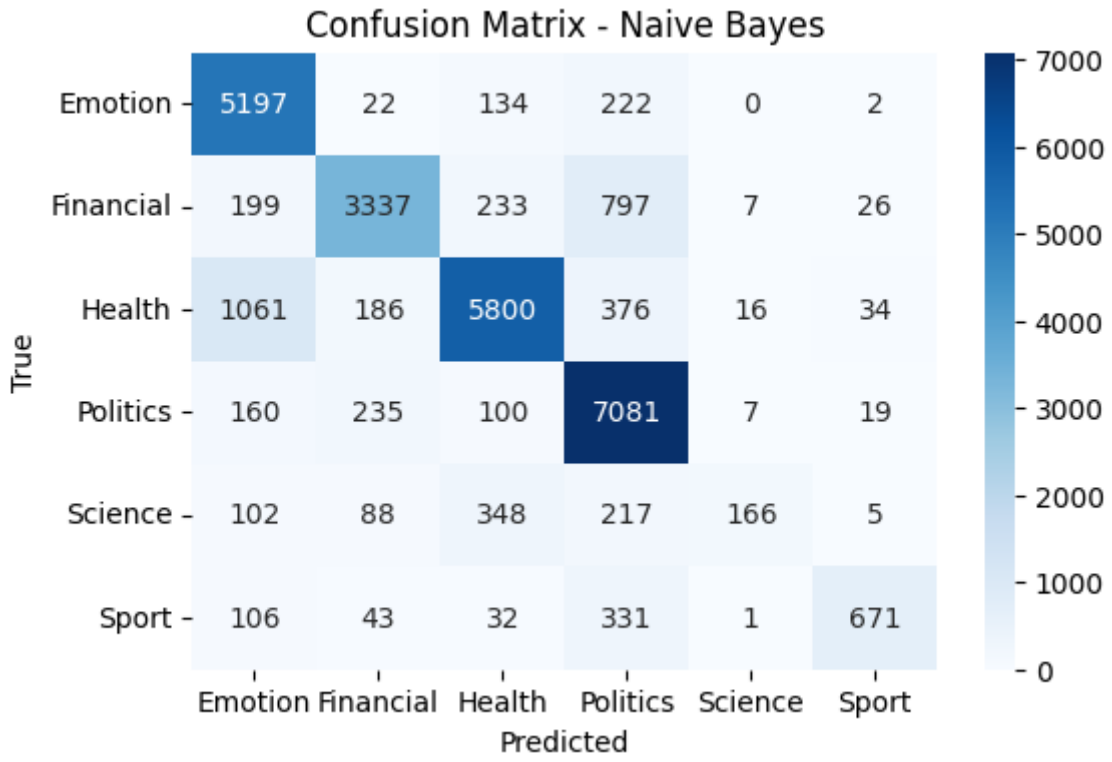
Tarea 2b. Realiza los pasos necesarios para generar una o varias representaciones de los textos

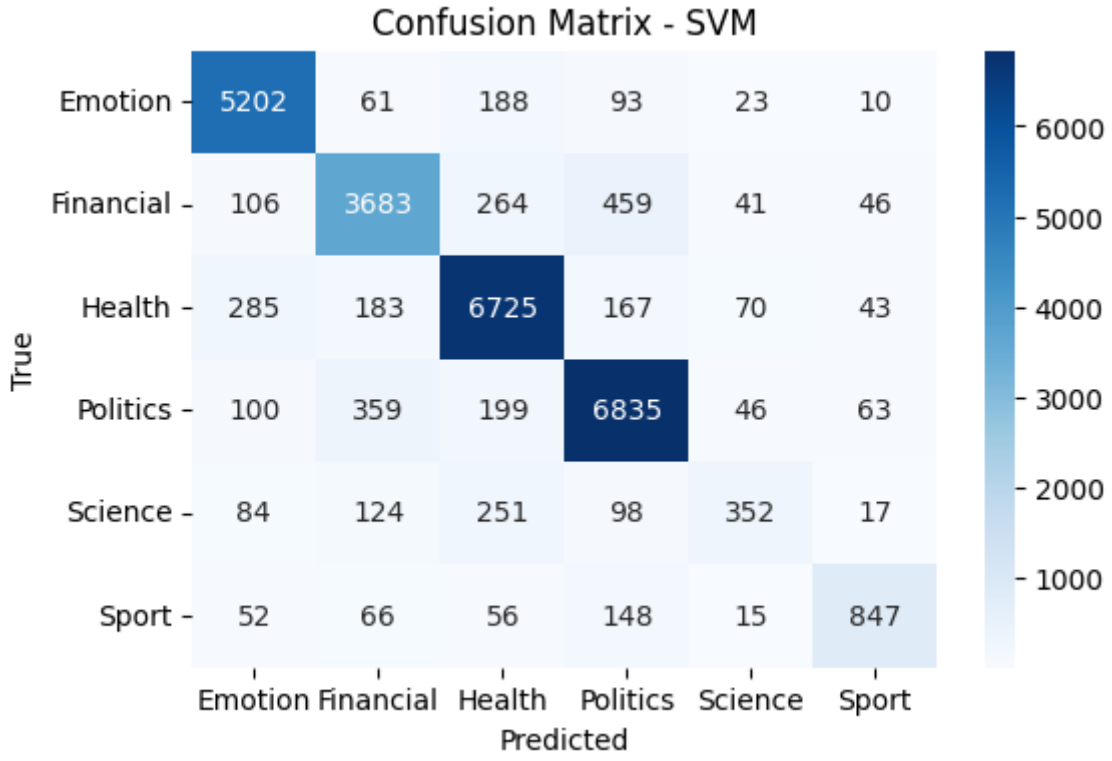
Se hace la representación de textos mediante TF_IDF seleccionando solo las 4000 palabrás más isgnificativas

Tarea 2c: Entrega y evalúa al menos 2 algoritmos de clasificación

Se eligen Navie Bayes y LinearSVC obteniendo los siguientes resultados:

- Navie Bayes accuracy: 0.8132743686268776
- LinearSVC accuracy: 0.86414970213077



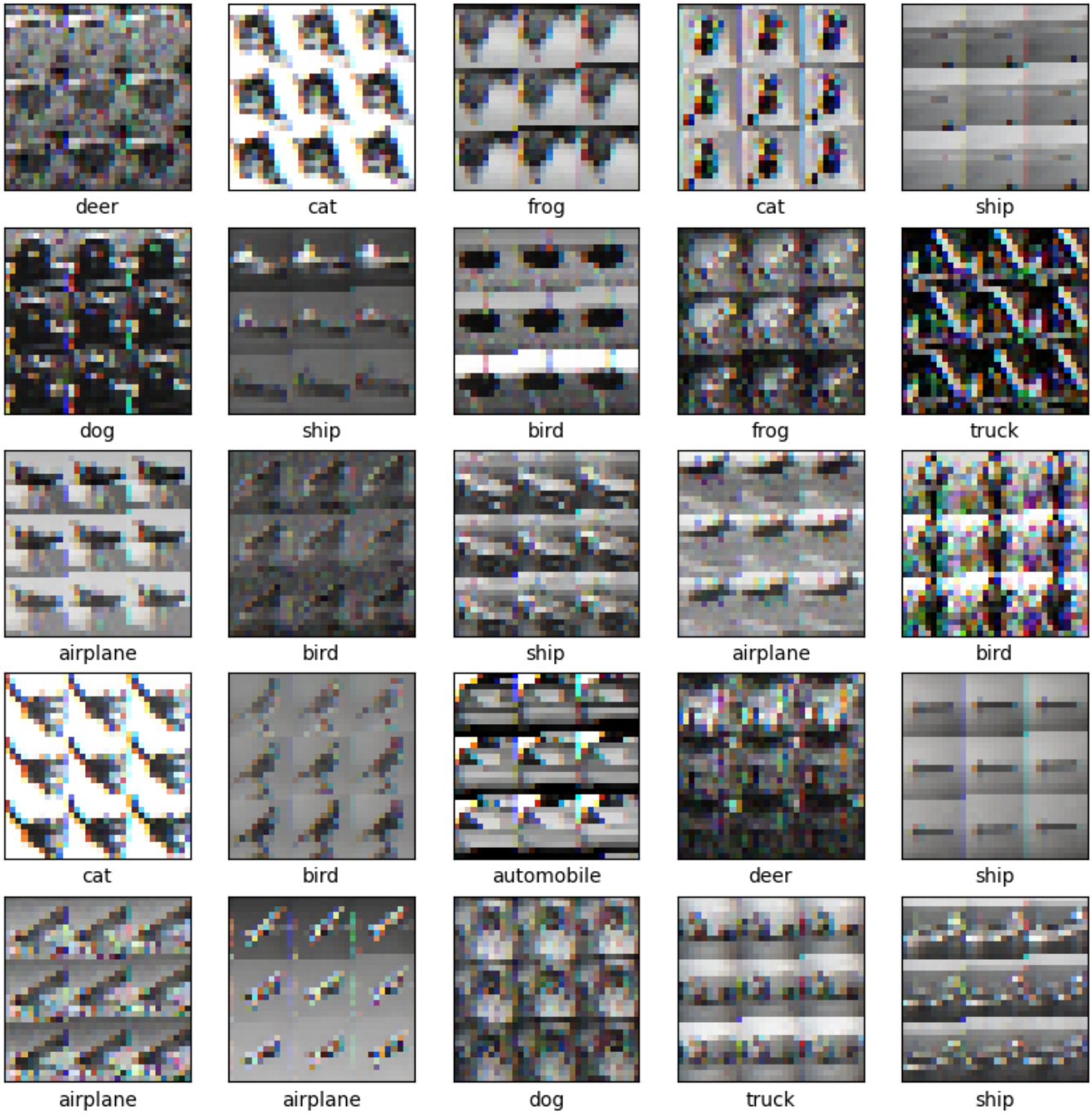


Siendo LinearSVC la mejor de los algoritmos propuestos

Ejercicio 3 – Clasificación de imágenes

Tarea 3a: Realiza una descripción del dataset.

Nos encontramos ente un dataset donde se tienen imágenes de diferentes objetos. Se normaliza los valores de los pixeles de las imágenes y se visualiza las diferentes imagenes. 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.

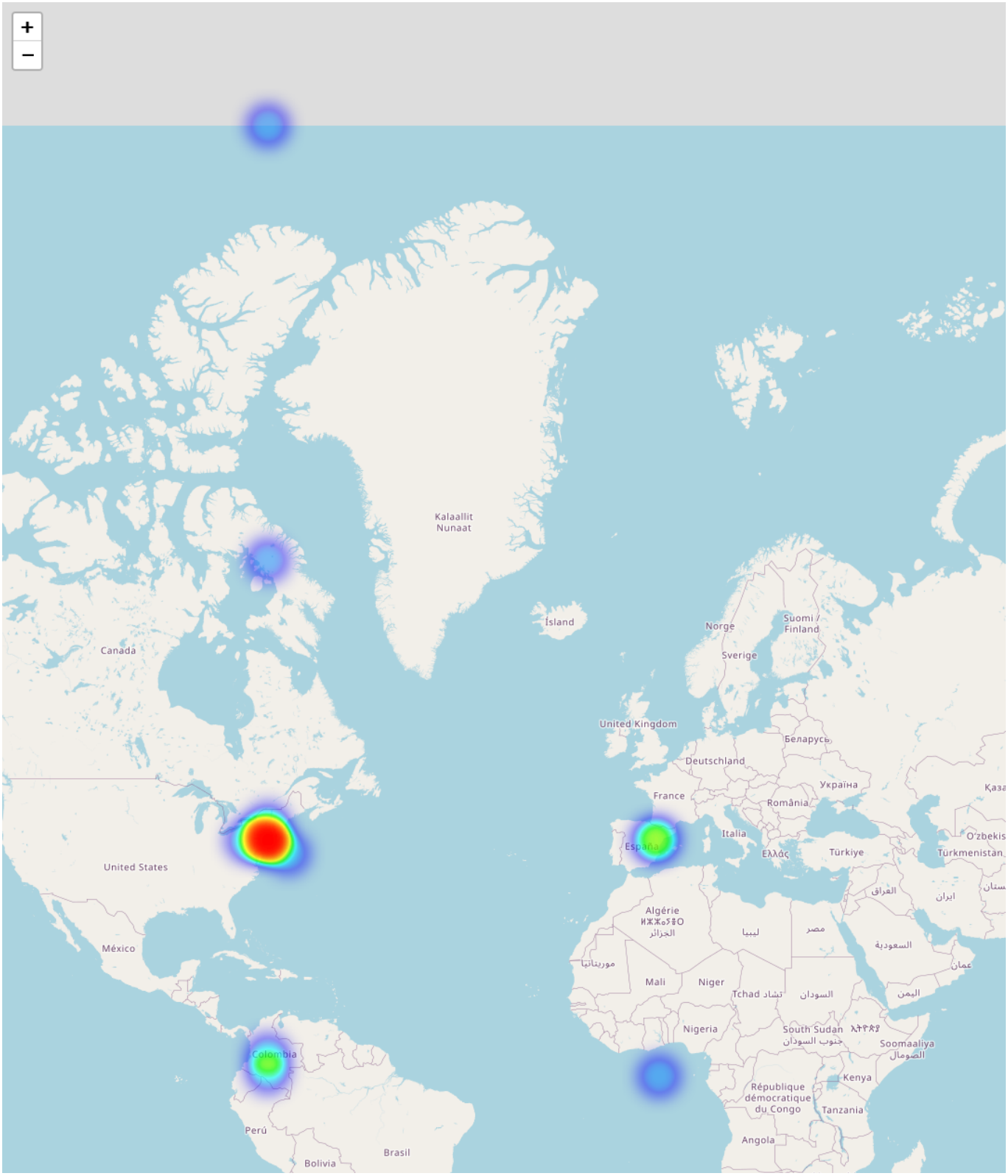
Ejercicio 4 : Predicción

Tarea 4a: Realiza una descripción del dataset

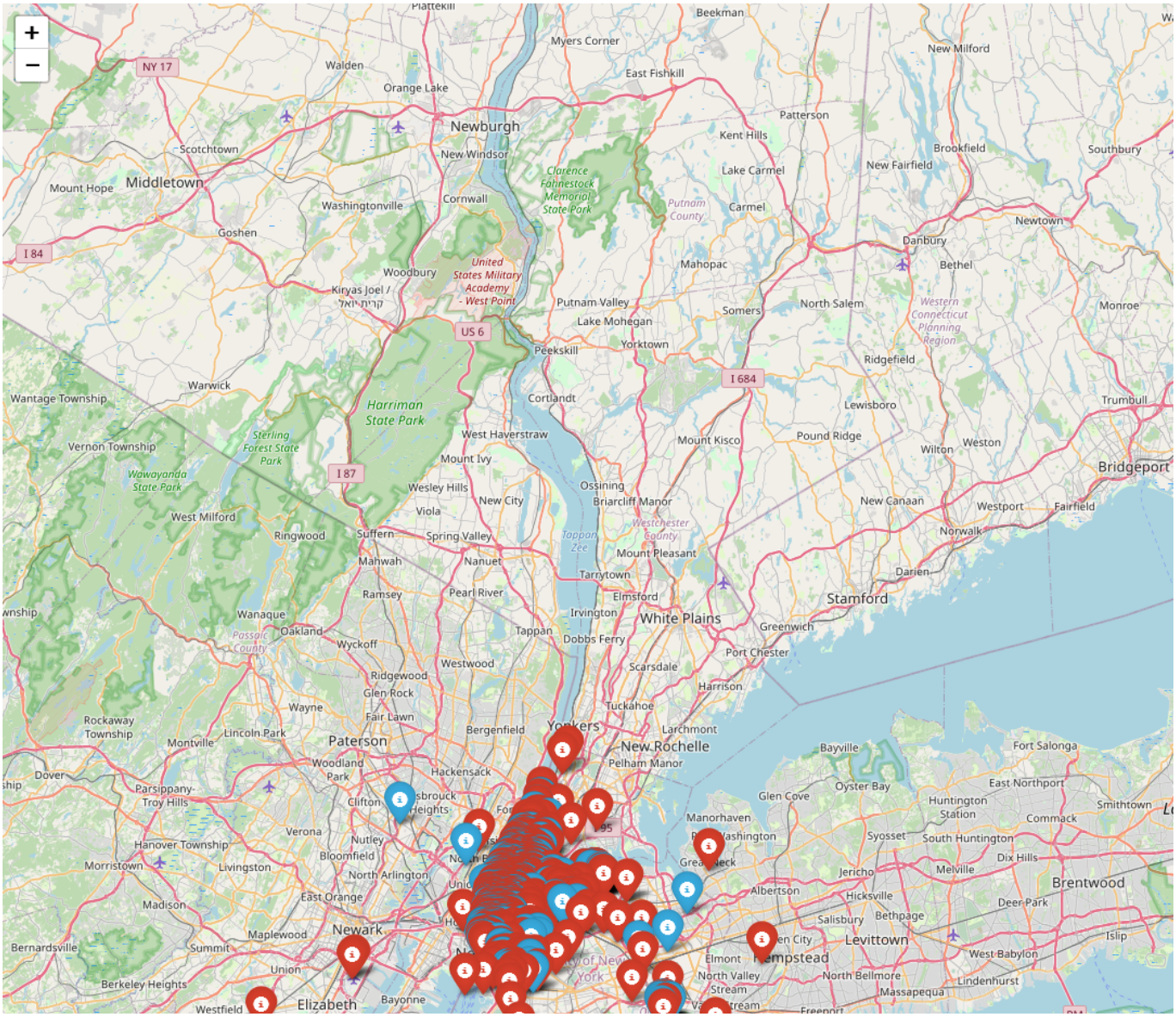
Nos encontramos ante un dataset donde se guarda lo que cobra el taxi, si lugar de recogida y su lugar de destino además del dia y hora de recogida en Nueva York

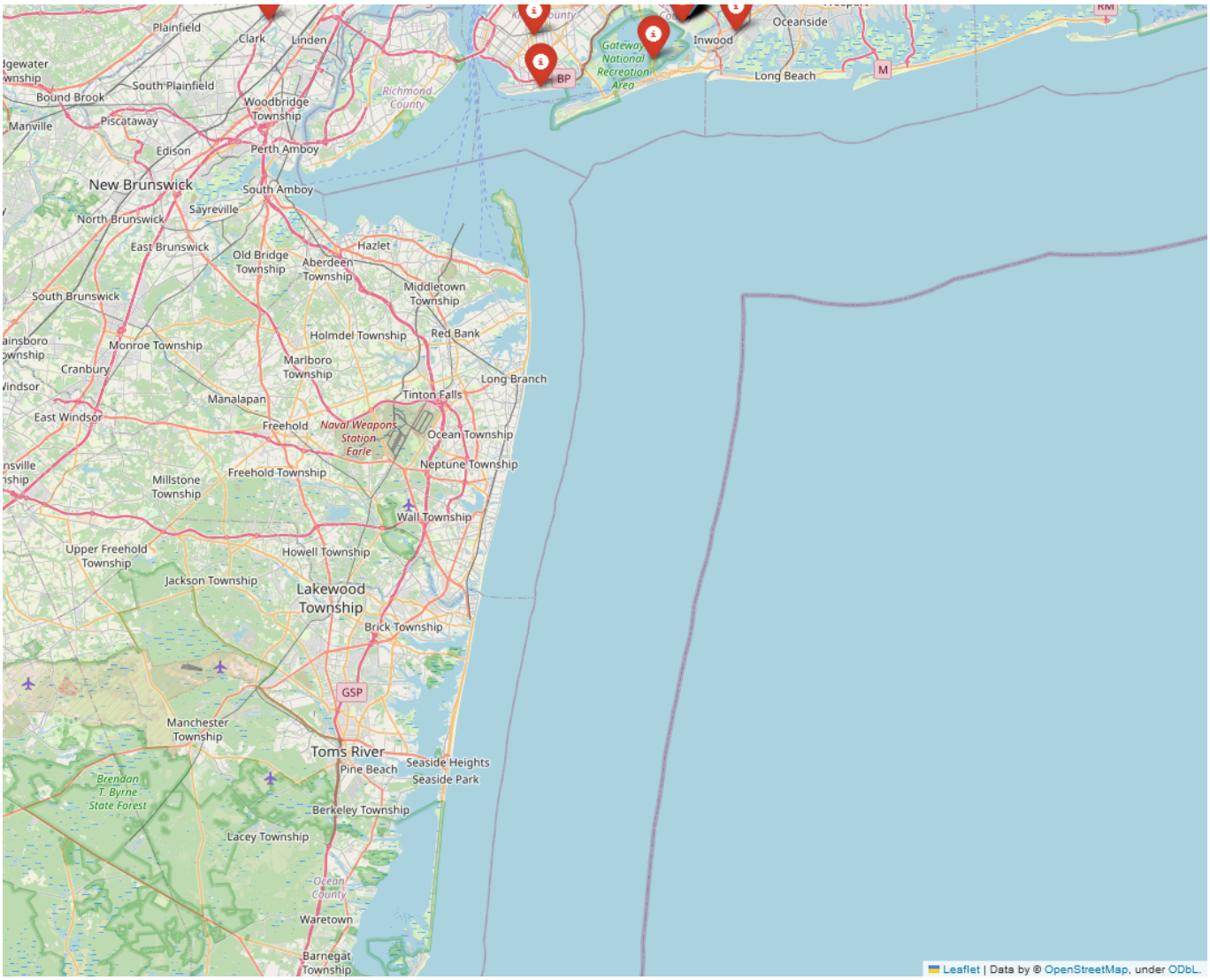
Tarea 4b: Realiza las tareas de pre-procesamiento

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.

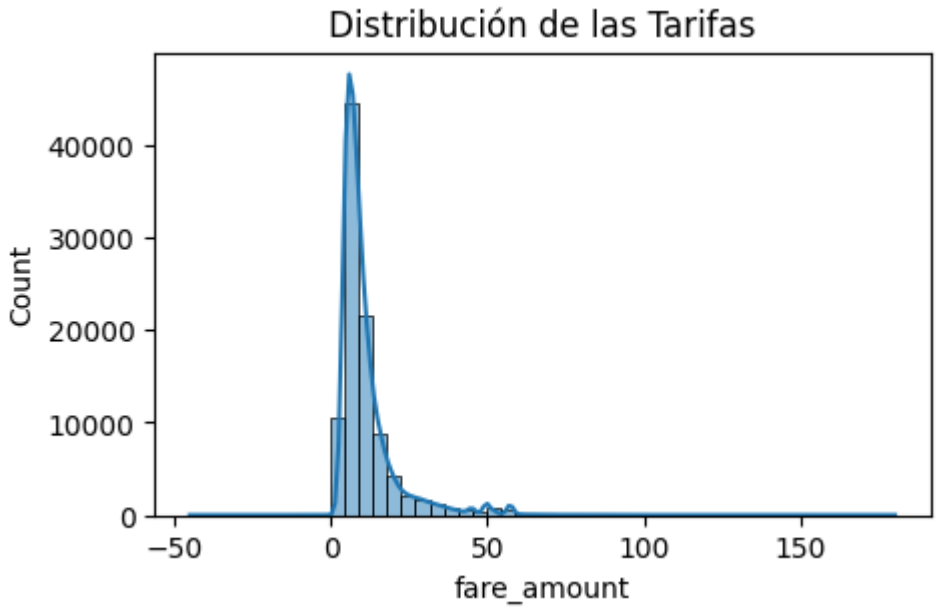


Como se aprecia en la imagen se tienen coordenadas de localizaciones diferentes a Nueva York, para ello establecemos como centro el centro de Nueva YOrk y eliminamos las localizaciones que estén fuera de un radio de 50km.

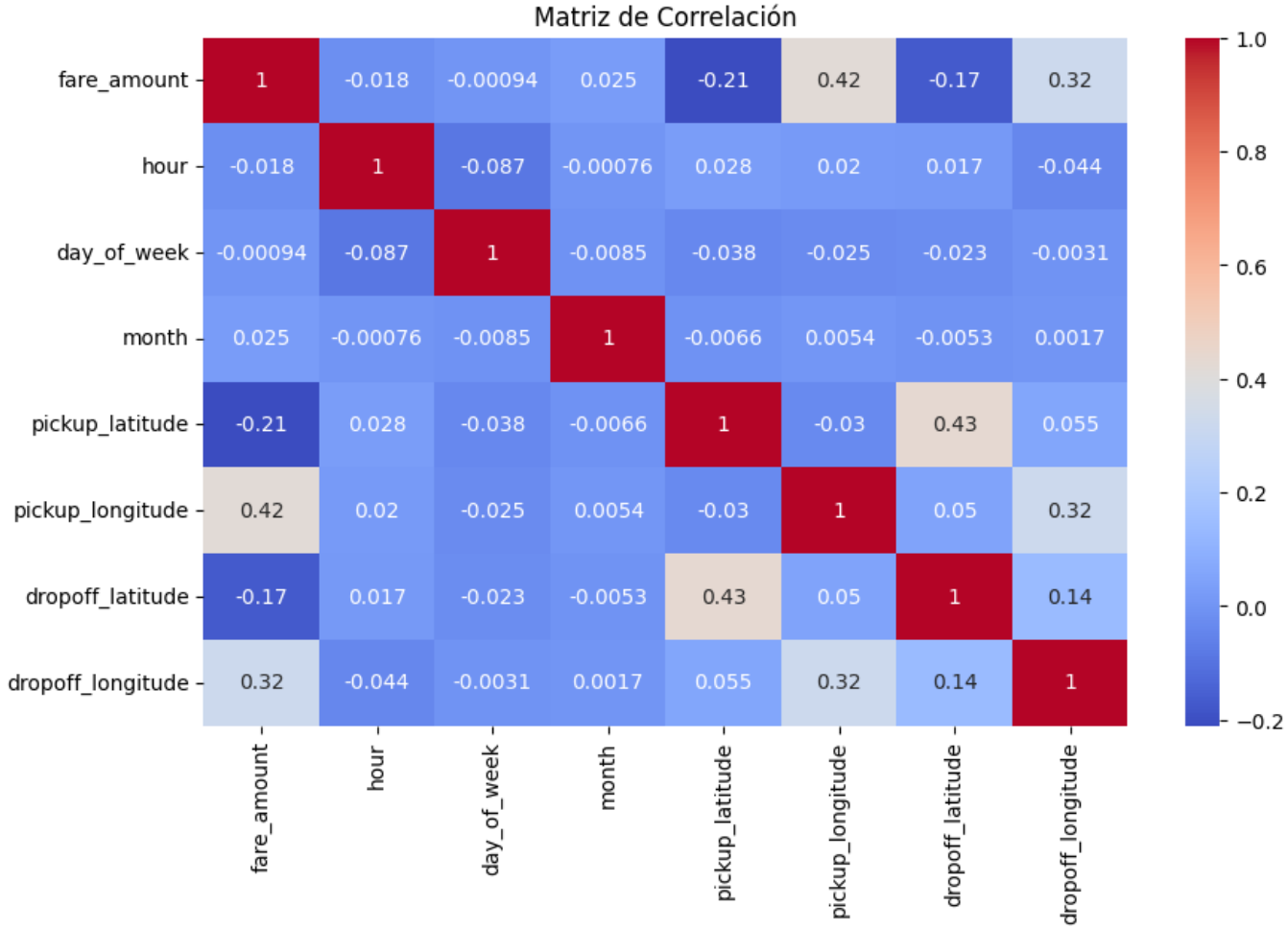




Seguidamente, convertimos la fecha y hora a valores numéricos y vemos la distribución de la tarifa que tiene outliers para ello a la hora de entrenar pondremos como máximo 100$ e iremos banajando para evitar el efecto de los outliers.



Finalmente vemos la correlación entre las variables:



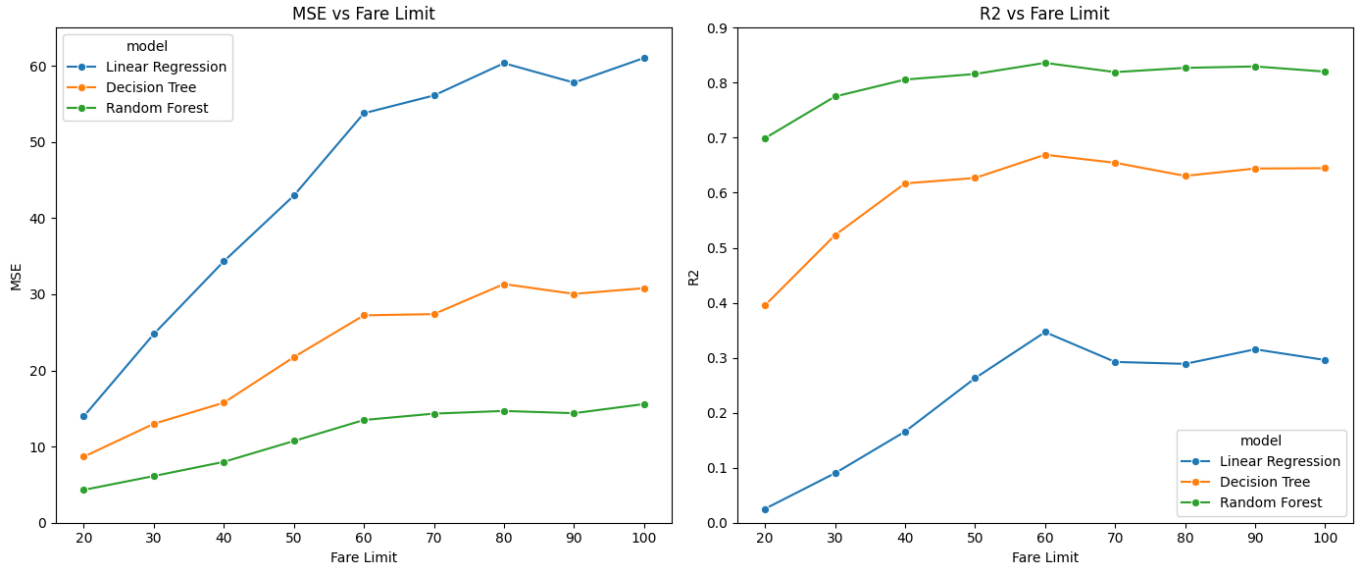
Las coordenadas son las que tienen mejor correlacón con la tarifa.

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)