María Fernanda Pérez Ruiz
A01742102

Clasificación de email de spam: pre-procesamiento y baselines

Contraste del desempeño de todos los clasificadores entrenados

En el código de clasificación de correos electrónicos para lograr identificar spam, hice la evaluación de cuatro clasificadores distintos: Clasificador de Regresión Logística, Clasificador de Support Vector Machine , Random Forests , Máquinas Gradient Boosting.
Así que a continuación haré un breve y conciso análisis de de los resultados que se obtuvieron el código de acuerdo al accuracy de los modelos.

1. Clasificador de Regresión Logística

```
DEBUG::El accuracy score de regresión logística es::
0.98
```

Vemos que el modelo de regresión logística se entrenó como un primer enfoque de clasificación. Tiene un muy desempeño de un accuracy de 0.98, es decir el modelo fue capaz de clasificar correctamente el 98% de los correos electrónicos en el conjunto de prueba. Este modelo es un modelo sencillo por lo cual no le llevó mucho tiempo correr el código . Así que tiene una precisión muy alta.

2. Clasificador de Support Vector Machine (SVM)

```
Entrenar el Clasificador SVC tomó  68 segundos

DEBUG::El accuracy score del Clasificador SVC es::
0.715
```

El modelo de support vector machine sabemos que tiene la capacidad de maximizar los márgenes entre las clases, en especial en conjuntos de datos con muchas características. Sin embargo para esta actividad el svm no obtuvo un accuracy muy alto, es aceptable pero no el ,mejor, el accuracy fue de 0.715. Vemos que es un valor más bajo que el de la regresión logística, y tambien le tomó más tiempo, o sea

68 segundos, sin embargo vuelvo a mencionar que el resultado y el tiempo que le tomó son aceptables.

3. Clasificador de Random Forest

```
Entrenar el Random Forest Classifier tomó   4 segundos
DEBUG::El RF testing accuracy score es::
0.9766666666666667
```

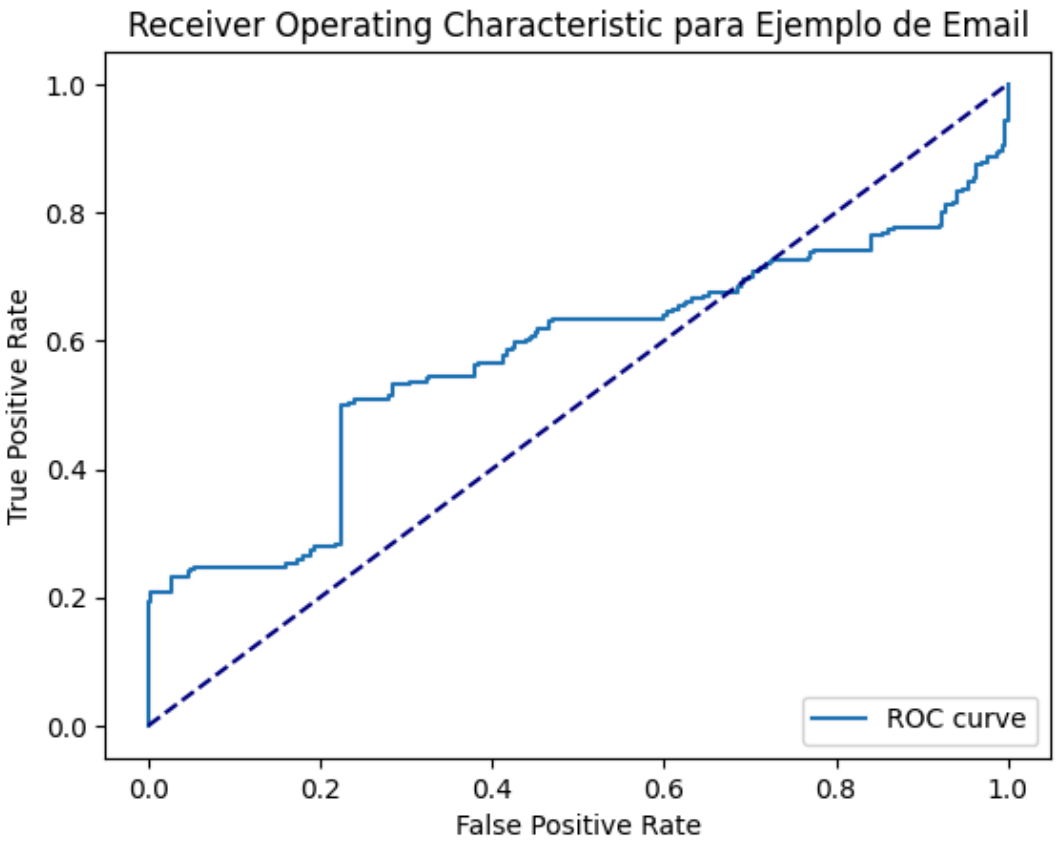
Sabemos que el modelo de random forest es una técnica de árboles de decisión para mejorar el rendimiento, en este caso vemos un muy buen resultado de accuracy para esta actividad siendo de 0.97666 , y además solamente le tomó 4 segundos, o sea que fue extremadamente bueno el resultado y extremadamente bueno el tiempo que le tomó.

4. Máquinas Gradient Boosting

```
El entrenamiento del Gradient Boosting Classifier tomó 345 segundos
DEBUG::El testing accuracy score de Gradient Boosting es::
0.955
```

Por último, chequee el modelo de máquinas gradient boosting obteniendo un 0.955 en el conjunto de prueba y un 0.9957 en el conjunto de entrenamiento. Y un AUC de 0.999706 o sea que tienen una excelente capacidad de discriminación entre las clases. Sin embargo fue el que mayor tiempo de entrenamiento tuvo, de 345 segundo , o sea muchísimo más alto que los otros modelos.

ROC



Se hizo un análisis adicional de la curva ROC , y vemos que el modelo tiene una buena capacidad de discriminar entre correos spam y no spam. Sin embargo vemos que en algunos puntos en la gráfica en true positive rate cae por debajo de lo esperado, pero a pesar de eso el modelo sigue mostrando un desempeño adecuado.

En conclusión vemos que Random Forest y Gradient Boosting se destacan como los mejores clasificadores en términos de precisión, Random Forest fue el más eficiente en tiempo y Gradient Boosting tiene la mayor capacidad predictiva sin embargo tienen un mayor costo computacional y lo vimos en el tiempo que le tomó. La Regresión Logística mostró un muy buen desempeño general, mientras que el SVM tuvo el peor rendimiento en este conjunto de datos.