Clasificación de email de spam

Por Juan Pablo Bernal Lafarga, A01742342

La clasificación de correos electrónicos como spam es una de las aplicaciones más comunes del Procesamiento del Lenguaje Natural (NLP). Esta tarea consiste en identificar automáticamente si un correo pertenece a la categoría de "spam" (correo no deseado) o "no spam" (correo legítimo) utilizando técnicas de aprendizaje automático y procesamiento de texto.

A continuación se muestra la evaluación de 5 modelos para clasificación de corres de SPAM:

```
DEBUG::El accuracy score de regresión logística es::
0.9933333333333333
```

El modelo clasificador de regresión logística obtuvo un 99.33% de exactitud en la clasificación de SPAM.

```
DEBUG::El accuracy score del Clasificador SVC es::
0.93
```

El modelo clasificador de Support Vector Machine obtuvo un 93% de exactitud en la clasificación de SPAM.

```
DEBUG::El RF testing accuracy score es::
0.99
```

El modelo clasificador de Random Forest obtuvo un 99% de exactitud en la clasificación de SPAM.

```
Los mejores parámetros encontrados:
{'min_samples_leaf': 3, 'min_samples_split': 2, 'n_estimators': 100}
La accuracy estimada es:
0.9866666666666667
```

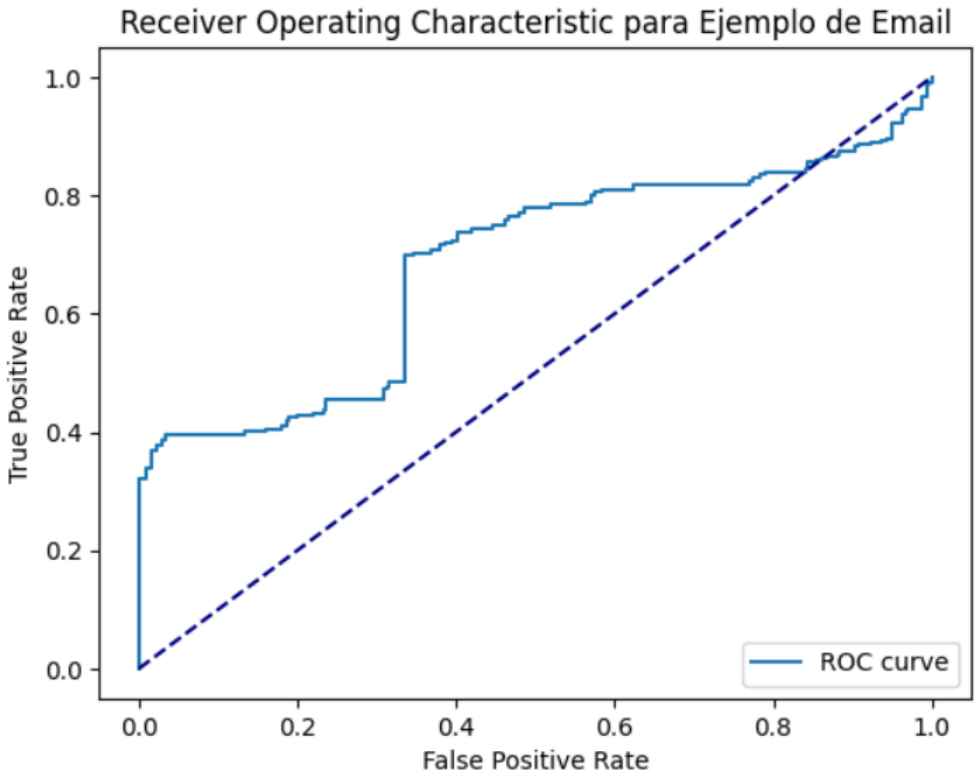
El modelo clasificador de Random Forest con hiperparámetros encontrados por GridSearch obtuvo un 98.67% de exactitud en la clasificación de SPAM.

```
DEBUG::El testing accuracy score de Gradient Boosting es::
0.9816666666666667
```

El modelo clasificador de Gradient Boosting obtuvo un 98.16% de exactitud en la clasificación de SPAM.

Dados los resultados, podemos notar que el ranking de resultados de los mejores modelos es:

Rank	Modelos
1	Regresión Log
2	Random Forest
3	Random Forest con GridSearch
4	Gradient Boosting
5	SVC



La curva ROC (en azul) muestra el rendimiento del modelo. Cuanto más cerca esté de la esquina superior izquierda, mejor será el rendimiento del modelo. La línea punteada (línea base) representa un clasificador aleatorio, y cualquier área por encima de esta línea indica un modelo mejor que el azar.

En esta curva, el modelo muestra un rendimiento aceptable ya que la curva se mantiene por encima de la diagonal durante la mayor parte de su recorrido, lo que indica una capacidad de clasificación superior al azar.