Clasificación de email de spam: pre-procesamiento y baselines

Eryk Elizondo González - A01284899

Resultados:

Modelo	Tiempo de Entrenamiento	Accuracy Score
Clasificador de Regresión Logística	2 segundos	0.9883
Clasificador de Support Vector Machine	64 segundos	0.8716
Random Forests	2 segundos	0.98333
Random Forests With Tuned Hyper-parameters	739.1 segundos	0.986666
Máquinas Gradient Boosting	245 segundos	0.98

Conclusión

Con base en la exactitud y tiempo de entrenamiento, se observa que el mejor modelo fue el de Clasificador de Regresión Logística ya que obtuvo la puntuación más alta en el menor tiempo posible. El modelo que le sigue considerando el tiempo es el de Random Forests con un ligero declive en la exactitud pero con el mismo tiempo, es posible aumentar esta exactitud con el refinamiento de los hiper parámetros pero eso solo hace una mejora ligera que no supera a la regresión logística. Es importante notar que con Máquinas Gradient Boosting no se obtuvieron más decimales en el puntaje de exactitud y es posible que los decimales superen a aquellos de la regresión logística, sin embargo la regresión logística sigue ganando ya que solo se tienen 0.0117 área de mejora que con la diferencia de tiempo de 243 segundos no se compensa. Entonces, recapitulando, el modelo de Clasificador de Regresión Logística es el más óptimo para este problema debido a su alta exactitud y bajo tiempo de entrenamiento.