Project Summary

- The purpose is to predict whether a text message is spam or not.
- Lemmatizing is used to complete tokenization.
- Term Frequency-Inverse Document Frequency is used in vectorization.
- By comparing between Random Forest and Gradient Boosting models, it appears that Gradient Boosting has slightly lower precision score and ROC-AUC but slightly higher recall score and overall accuracy score. However, the most significant difference is the fit time where Gradient Boosting is about 20 times slower than Random Boosting. If runtime performance is a bottle neck, Random Forest may be a better choice in this case.

Model Name	Fit Time	Predict Time	Precision Score	Recall Score	Accuracy Score	ROC-AUC
Random Forest	50.349 Seconds	0.716 Seconds	1.0	0.855	0.979	0.997
Gradient Boosting	1017.748 Seconds	0.25 Seconds	0.979	0.893	0.982	0.984