Predicting Spam in Emails

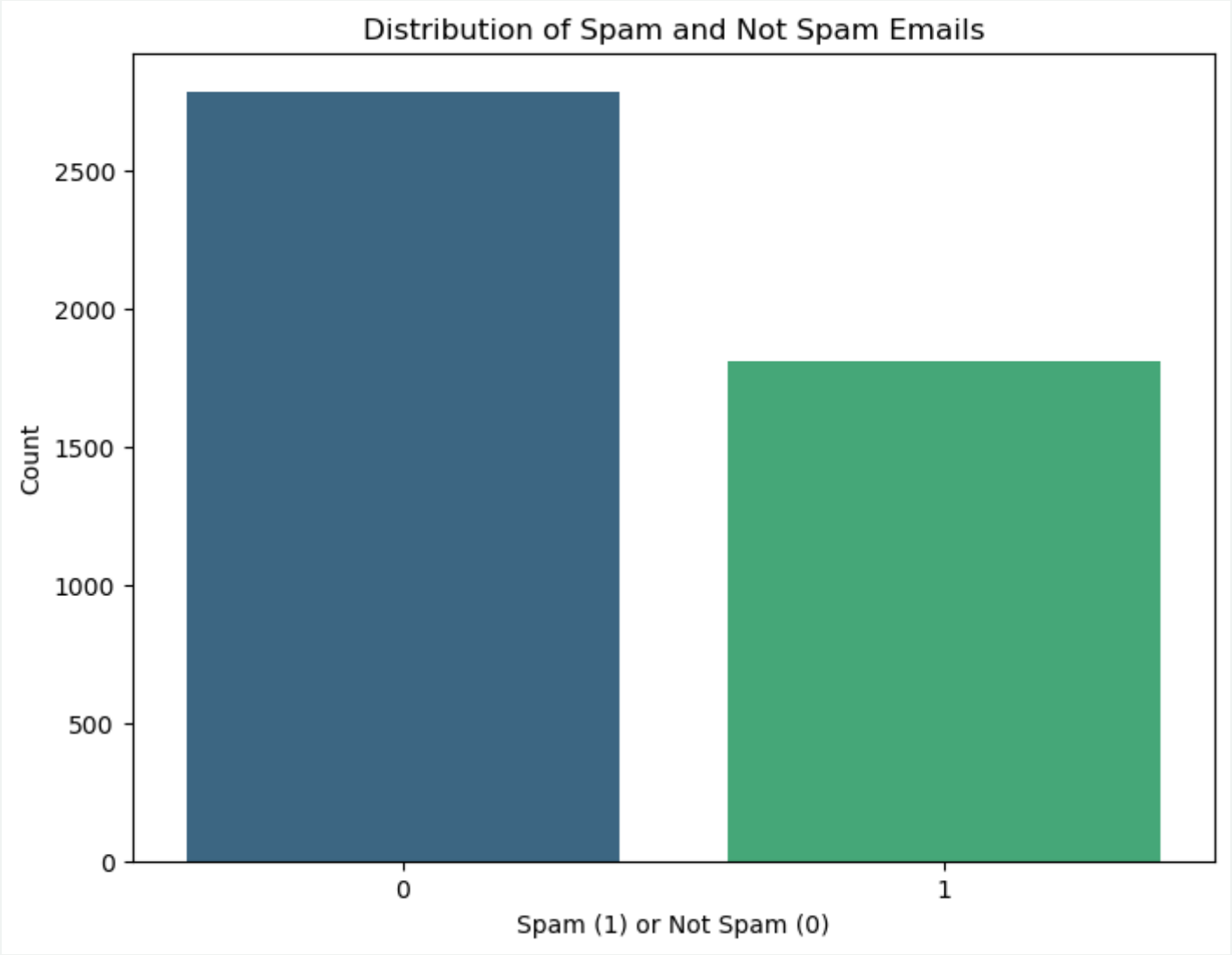
BY KAICHENG MAO, KRESHNAYOGI D.B., ALEXANDER HUANG, RADITYA AULIA, MARALEYSI CLAVIJO, KENNETH

Spam Dataset Overview

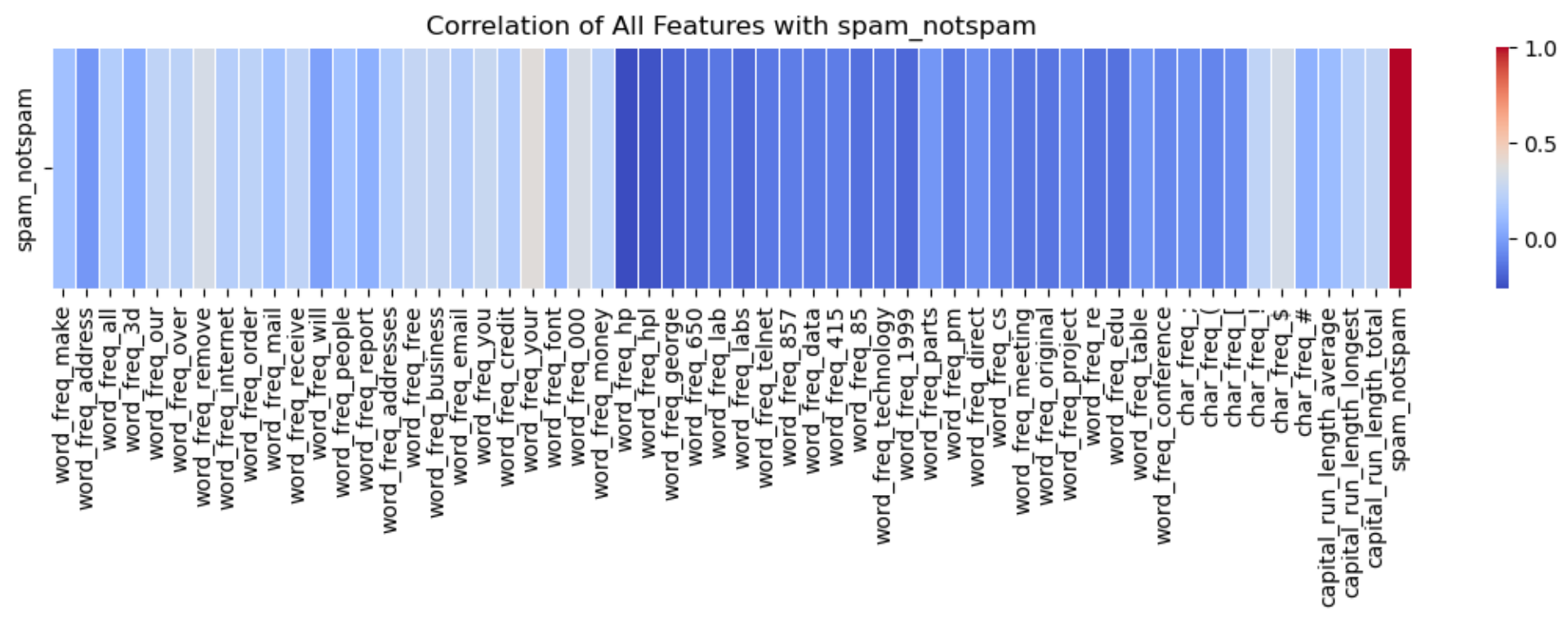
- Provided dataset with 4600 emails and 58 columns. There are no missing values.
- Most columns describe the frequency percentage of words (or characters) in the email that match the indicated word (or character).
- Formula: 100 * (number of times the WORD appears in the e-mail) / total number of words in the e-mail.
- Three columns describe the frequency and use of capital letters in an email. (“capital_run_length_average”, “capital_run_length_longest”, “capital_run_length_total”)
- The last column (“spam_notspam”) indicates whether the given email is spam or not spam. There are 2788 non-spam emails and 1813 spam emails.

Count Plot

There are 2788 non-spam emails and 1813 spam emails. About a 3:2 ratio.



Correlation of All Features



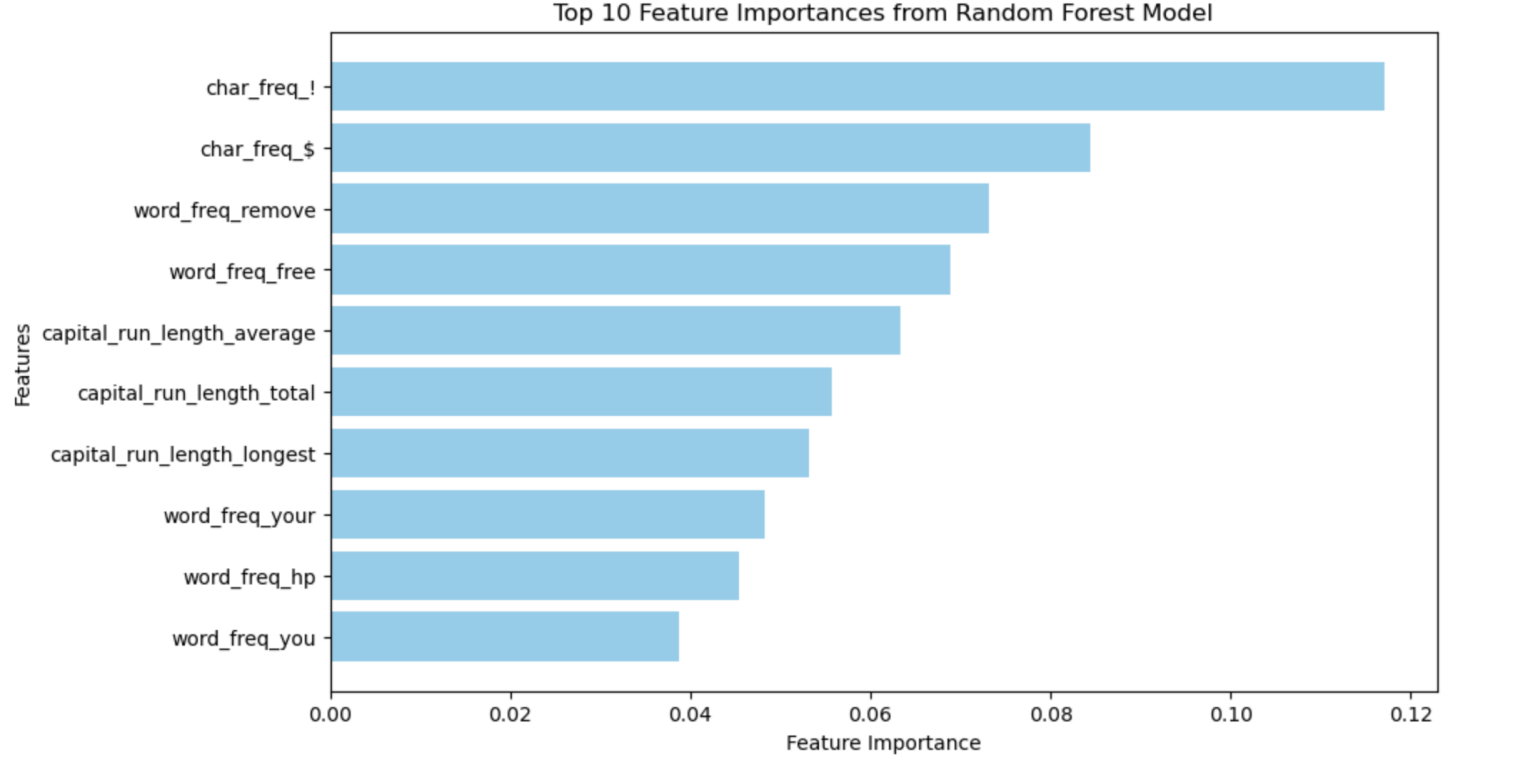
HOW CAN WE PREDICT
WHETHER AN EMAIL IS SPAM
OR NOT SPAM?

Problem Statement

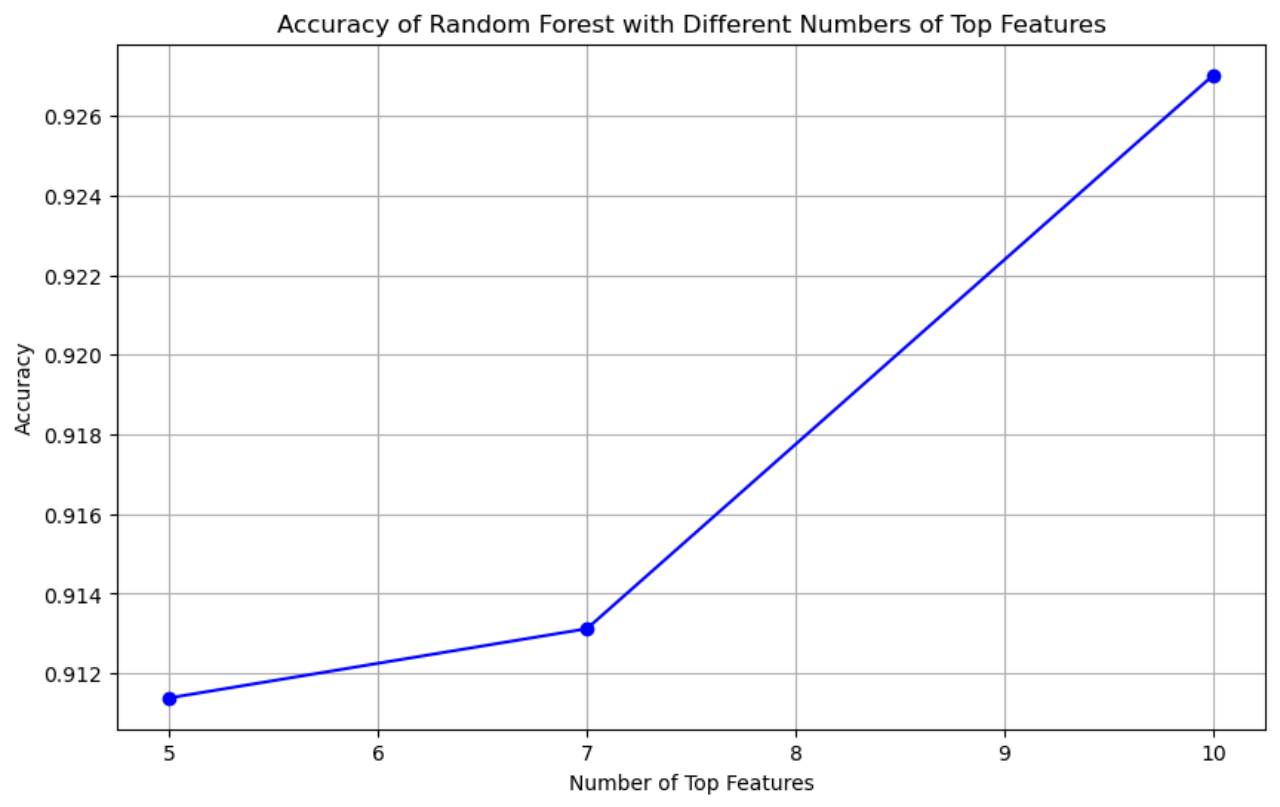
Our Solution

- We used classification to predict whether an email would be spam or not spam.
- For this we used both Random Forest and K Nearest Neighbors models.
- Grid Search CV were also used as a method to find the best parameters for each model.
- The models were created using the top 5, 7, 10 features chosen using ".feature_importances_" and compared them for accuracy.

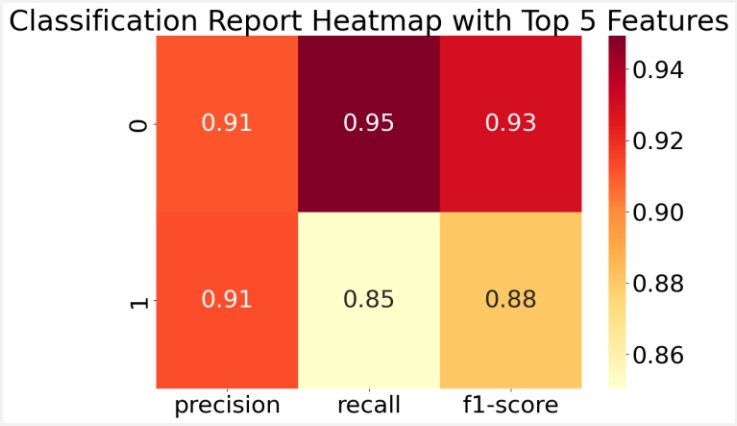
Top 10 features

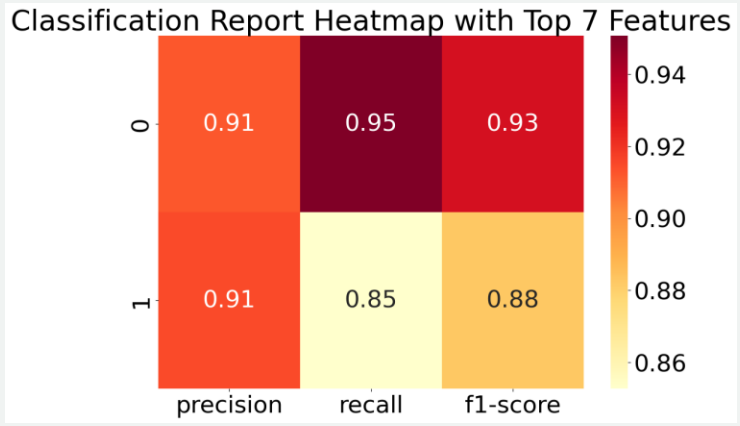


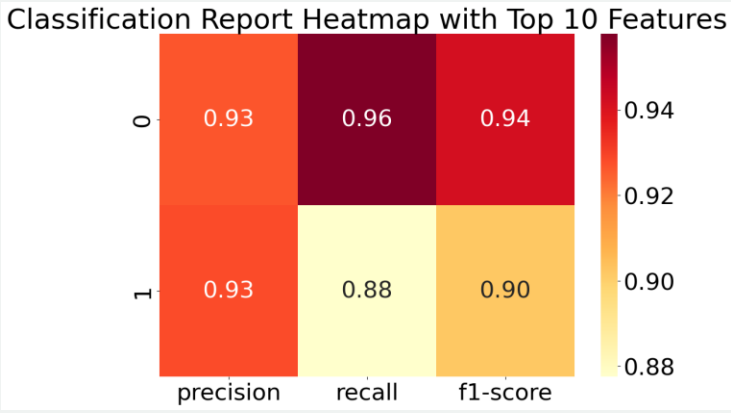
Random Forest Model Result



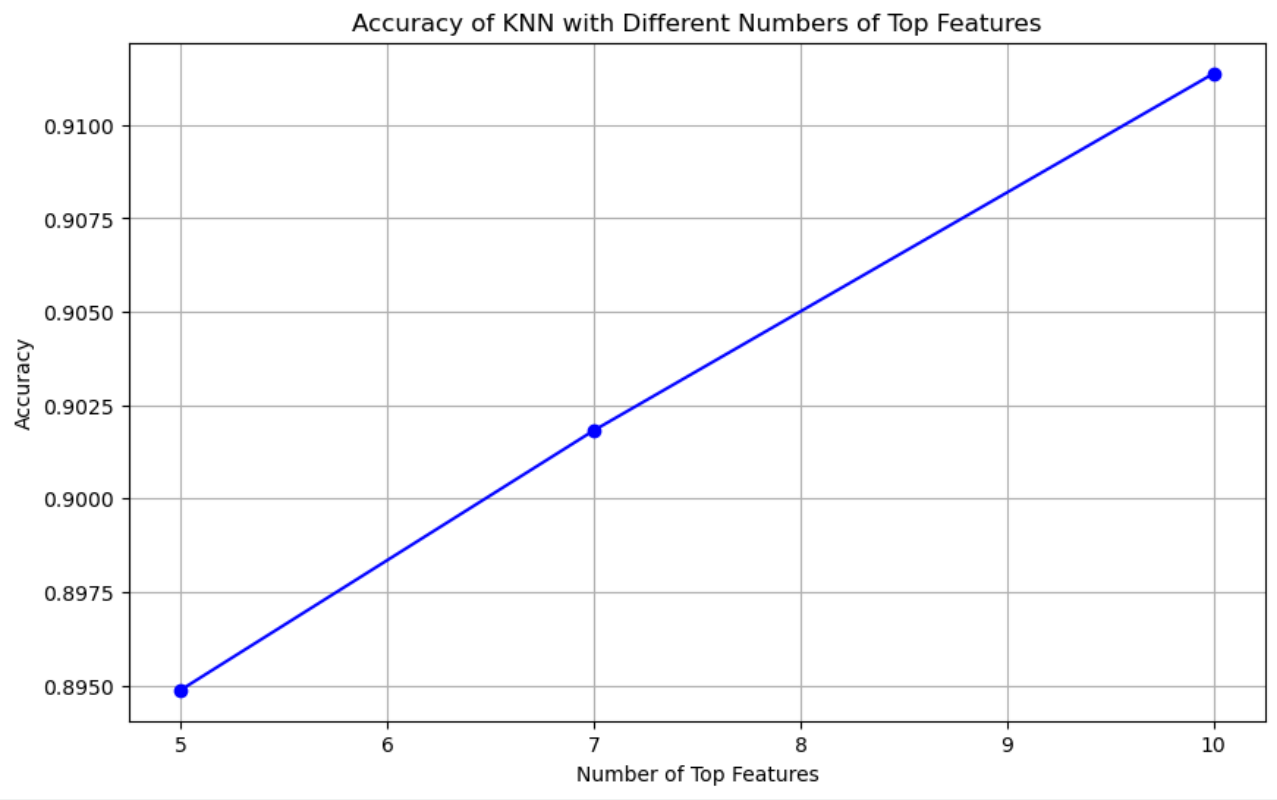
Random Forest Model Result



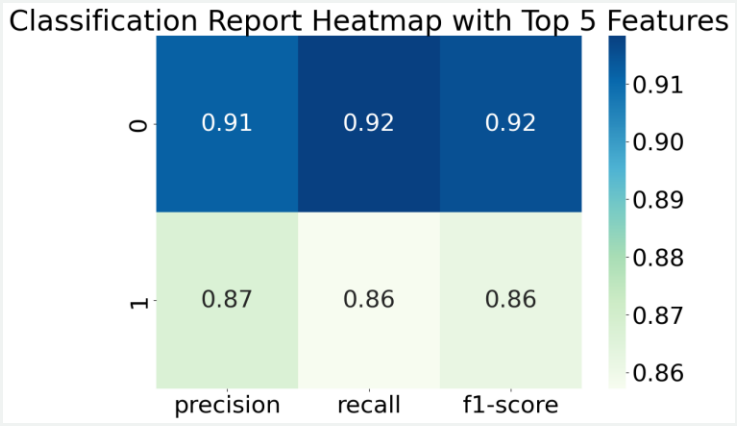


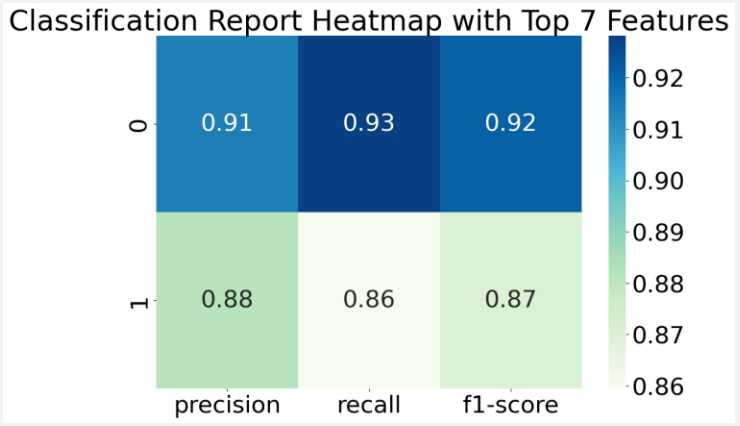


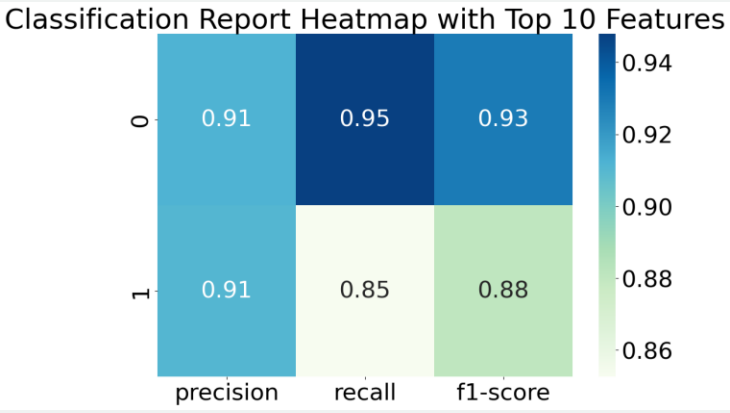
KNN Model Result



KNN Model Result







Conclusion

- The Random Forest model was more accurate in predicting spam emails than the KNN model.
- Reasons could be ensemble learning in the random forest model, accounting for importance of features in the random forest model, and KNN struggles with high dimensional data since distance metrics lose value with more dimensions.

- Alex and Radit worked on the EDA and visualization.
- Mao and Kreshnayogi worked on the ML models.
- Maraleysi and Kenneth worked on the presentation.