Proposal: Classifying Spam Emails

Team Members:
Daniel Jouran
Ryan Smith
Morgan Smith-Wood
Shiyunyang Zhao

Problem of Interest:
We would like to predict whether emails, based on their content, should be classified as spam emails.

Dataset Description:
There are two datasets available to us. The first is one provided in the R package 'kernlab'. This dataset contains data on 4601 emails classified as spam or not spam. For each email, the frequency of select words or select characters (represented as a proportion of total words or total characters, respectively) is given for a total 54 variables. A total of Information of capital letters is also represented in three ways; the average, the longest, and the total run-length of capital letters. The second dataset contains a subset of the original variables with only 6 predictive variables. This dataset can be found on the TidyTuesday Github repository.

R package: https://search.r-project.org/CRAN/refmans/kernlab/html/spam.html
Github Repository:
https://github.com/rfordatascience/tidytuesday/tree/master/data/2023/2023-08-15

ML Techniques:
We plan to employ supervised machine learning and use techniques such as logistic regression, LDA, QDA, and Naive Bayes to build and evaluate our predictive models. Logistic regression is useful to predict the probability of an email being spam, ensuring that the predicted values lie between 0 and 1. LDA and QDA will be used to distinguish between spam and non-spam emails, and Naive Bayes can evaluate the likelihood of an email being spam based on the frequencies of words and phrases in the emails.

Potential Challenges:
We found that some challenges could include the interpretability of the variables. For example, the proportions of words or strings could be greater than one which could be confusing to an individual who has not thoroughly understood the subject matter. Additionally, selecting variables to use for our analysis could be difficult since their significance in predicting spam emails could vary. Additionally, we expect the usage of some words and characters to be correlated.