Brian Stoss DS4002
4/28/25

Using Data Science to Identify and Classify Spam Emails

In today's digital world, where billions of emails are exchanged every day, the ability to quickly and accurately detect spam emails has never been more critical. Spam not only clutters inboxes but also poses serious security threats through phishing attacks and malware distribution. As email remains one of the most common forms of communication for individuals and businesses alike, ensuring its safety and efficiency is paramount. A data scientist like yourself plays a vital role in this mission - using machine learning techniques to build intelligent models that can distinguish legitimate emails from harmful ones. In doing so, you not only help enhance cybersecurity but also contribute to a more streamlined, trustworthy digital communication system.

This project focuses on the creation and evaluation of three machine learning models - K-Nearest Neighbor (KNN), Logistic Regression, and Decision Trees - each trained on the same email dataset to predict whether emails are spam ("Spam") or legitimate ("Not Spam"). Your analysis and technical skill in model building and evaluation will directly impact the ability to detect and filter unwanted emails, ultimately improving email security infrastructures across platforms. By comparing model performance based on precision and accuracy, you will identify which algorithm is most effective at correctly classifying spam, paving the way for smarter, more secure systems in the future.

The Deliverable:

In response to the rising need for efficient spam detection, your task is to develop and compare three machine learning models - K-Nearest Neighbor, Logistic Regression, and Decision Trees - trained on a labeled dataset of emails. You will preprocess the text data by removing stopwords, tokenizing words, stemming terms, and cleaning the dataset for clarity and consistency.

After preprocessing, you will train each model on 80% of the data and evaluate it on the remaining 20%. To ensure robustness, you will also apply techniques such as k-fold cross-validation and parameter fine-tuning (e.g., adjusting the k-value for KNN and the max_features parameter for TF-IDF vectorization). Your goal is to achieve a model accuracy of at least 95%, selecting the model that demonstrates the highest performance based on both precision and accuracy.

The deliverable you provide will not only highlight the most effective machine learning approach for classifying spam emails but could also serve as a foundation for future cybersecurity tools. Through this work, you will illustrate the profound impact that thoughtful data analysis and machine learning can have on real-world digital safety - making inboxes safer and communication more reliable for millions of users.