Brian Stoss DS4002
4/28/25

Using Data Science to Identify and Classify Spam Emails

Imagine being the digital security hero who prevents dangerous phishing attacks from ever reaching someone's inbox. In today's world where billions of emails are sent daily, detecting spam isn't just about cleaning up junk - it's about protecting people and businesses from real threats.

In this case study, you will step into the role of a data scientist tasked with building intelligent models to distinguish spam emails from legitimate ones. You'll work with real email data, train three machine learning models (K-Nearest Neighbors, Logistic Regression, and Decision Trees), and evaluate which model most effectively flags spam - helping to make communication faster, safer, and more reliable.

You will preprocess text data, build and test your models, and compare their performance to identify the most accurate spam detector. Your work will contribute to developing smarter, more secure email systems for the future.

Ready to make inboxes safer for everyone? Start here:

https://github.com/bmstoss13/CS3-DS4002

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.