**Project 3 Proposal**  
**Date:** 06/23/2025

**Group Members:**  
Roger Leung  
Bryan Mejia  
Abul Hasan  
Ahmed Ali

**Detecting Spam Emails Using Machine Learning: A Comparative Study of Classification Models**

**Research Question:**  
How accurately can modern machine learning models detect spam emails, and which algorithm or ensemble approach provides the best trade-off between accuracy, precision, and recall in real-world email datasets?

Spam detection is a vital part of securing personal and enterprise communications. As spammers evolve their techniques, machine learning-based systems must adapt and stay ahead. With increasing access to labeled datasets and evolving algorithms, now is an ideal time to explore which models are most effective in spam filtering. This study focuses on evaluating individual models and ensemble methods to identify a reliable, accurate, and scalable solution.

**Research Papers (Brief Summaries):**

1. **Zhang (2024):** Compared Naive Bayes, Decision Tree, and SVM. Found that SVM achieved the highest accuracy. A hybrid model combining SVM and NB showed even better results.
2. **AIP Publishing (2025):** Analyzed RF, NB, MLP, and SVM. Random Forest (RF) achieved the highest accuracy (~98.8%) and demonstrated consistent, stable performance.
3. **Sevli & Keskin (2024):** Evaluated RF, LR, NB, SVM, and ANN. RF performed best. Simpler ensemble models like RF outperformed even neural networks for medium-sized datasets.
4. **Li (2024):** Compared NB and RF. NB had better recall; RF had better precision. The paper emphasized choosing a model based on spam filter goals (recall vs. precision).
5. **Adnan et al. (2023):** Used stacking ensembles (LR, DT, KNN, GNB, AdaBoost). The stacked model outperformed all individual models, with 98.8% accuracy and 98.9% F1 score

**Shared Findings Across Research Papers:**

All five studies agree on key points:

* **Ensemble models** (like stacking) outperform single classifiers.
* **Random Forest** offers strong accuracy and stability.
* **Naive Bayes** is fast and effective for simple use cases.
* **Model choice depends on the goal** (recall vs. precision).
* **TF-IDF and text cleaning** are essential for better results.

**Planned Experiment:**This project will replicate and extend findings from the five selected research papers using two public datasets. The models to be implemented are based on those used in prior studies, including **Naive Bayes (NB)**, **Support Vector Machine (SVM)**, **Logistic Regression (LR)**, **Random Forest (RF)**, **Artificial Neural Network (ANN)**, and ensemble methods like **AdaBoost** and **Stacking** (e.g., combining LR, RF, and Gaussian Naive Bayes).

Evaluation metrics will include **accuracy**, **precision**, **recall**, **F1 score**, and **ROC-AUC**, as these were used consistently across the reviewed literature.

We will use **5-fold cross-validation** to ensure reliable and fair performance comparisons and apply **GridSearchCV** for **hyperparameter tuning** where applicable.

Preprocessing steps will follow techniques outlined in the research papers, including **lowercasing**, **stop-word removal**, **punctuation cleaning**, and **TF-IDF vectorization** for converting email text to numerical features.

The implementation will be done using **Python**, with the help of libraries such as **Scikit-learn**, **NLTK**, and **Pandas** for model training, preprocessing, and evaluation.

**Datasets:**

1. **Email Spam Classification Dataset CSV**  
   <https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv>
2. **Spam Email Classification Dataset**  
   <https://www.kaggle.com/datasets/purusinghvi/email-spam-classification-dataset>

**Expected Contribution:**

This project aims to identify the best-performing model(s) for spam classification and evaluate the value of ensemble approaches. It will contribute a clear, comparative understanding of model performance across datasets and help guide future implementations of adaptive spam filters.

**References:**

1. Enhancing Spam Filtering: A Comparative Study of Modern ML Techniques (Chenwei Zhang, 2024):  
   <https://www.sciencedirect.com/science/article/pii/S2772662223002308>
2. Measuring the Efficiency of RF, NB, MLP & SVM in Email Spam Detection (AIP Publishing, 2025):  
   <https://pubs.aip.org/aip/acp/article-abstract/3270/1/020052/3343767/Measuring-the-efficiency-of-random-forest-naive>
3. Machine Learning Based Classification for Spam Detection (Sevli & Keskin, 2024):  
   <https://www.researchgate.net/publication/380000625_Machine_Learning_Based_Classification_for_Spam_Detection>
4. Analysis of Spam Classification Based on NB and RF (Li, 2024):  
   <https://www.ewadirect.com/proceedings/aemps/article/view/12660>
5. Improving Spam Email Classification Using Stacking Ensembles (Adnan et al., 2023):  
   <https://link.springer.com/article/10.1007/s10207-023-00756-1>