**Project Title: Spam Detection Using Random Forest Classifier**

**1. Introduction**

Email spam filtering is essential for managing inbox clutter, security, and productivity. This project uses a Random Forest Classifier to classify emails as spam or non-spam based on word frequency data, providing an adaptable machine-learning solution.

**2. Objective**

To build a machine learning model that accurately distinguishes spam from non-spam emails, allowing real-time predictions for new input emails.

**3. Dataset Overview**

* **Features**: Word count columns representing the frequency of each word.
* **Labels**: Prediction column with 1 (spam) and 0 (non-spam).
* **Size**: 5,172 emails with over 3,000 word features.

**4. Data Preprocessing and Feature Engineering**

The word counts for each email serve as features directly. Data was split 80/20 for training and testing.

**5. Model Selection and Training**

A Random Forest Classifier was selected for its robustness with high-dimensional data. The model was trained with 100 trees, achieving an accuracy of **97.78%** on the test set, with balanced precision, recall, and F1-scores across spam and non-spam categories.

**6. Prediction Pipeline**

A pipeline was created for real-time predictions by transforming input email text into word count vectors aligned with the trained feature set.

**7. Comparison with Other Spam Filters**

This project stands out by:

* **Using Raw Word Counts**: A simple and interpretable approach compared to complex text embeddings like TF-IDF.
* **Random Forest over Naive Bayes**: Offers flexibility and avoids Naive Bayes’s assumption of feature independence.
* **Real-Time Prediction Compatibility**: The model is specifically designed for real-time use, aligned with a custom vocabulary from training data, ensuring efficient predictions.

**8. Challenges and Limitations**

High dimensionality and vocabulary limitations were addressed but may impact accuracy with entirely new language patterns. Future work could involve adding semantic features and experimenting with other ensemble models.

**9. Conclusion**

This project effectively demonstrates a machine-learning approach to spam filtering, achieving high accuracy and practicality. It offers a robust solution for static datasets, with potential for future adaptations to dynamic data sources and broader applications.