Spam Email Classification Using Machine Learning

This project explores using machine learning algorithms to classify emails as spam or not spam (ham). Spam classification is a common natural language processing (NLP) task, and we compare multiple models using standard evaluation metrics.

- Input**:** A dataset of SMS messages labelled as "spam" or "ham".
- Output: A binary prediction indicating whether a new SMS message is spam or not.
- Dataset**:** The SMS Spam Collection Dataset on kaggle, consisting of 5,574 labelled SMS messages. The dataset is publicly available and widely used for spam detection models.
- Preprocessing**:** Text was cleaned, lowercased, tokenized, stopwords removed, and converted to numerical features using TF-IDF vectorization.

Three machine learning algorithms that were implemented:

1. **Logistic Regression**
    - A linear model for binary classification.
    - Hyperparameter tuned**:** `C` (inverse of regularization strength)
    - Best value**:** `C = 10`

2. **Multinomial Naive Bayes**
    - A probabilistic model well-suited for text classification.
    - Hyperparameter tuned**:** `alpha` (smoothing parameter)
    - Best value**:** `alpha = 0.1`

3. **Random Forest**
    - An ensemble model based on decision trees.
    - Hyperparameters tuned:
        - `n_estimators` (number of trees)
        - `max_depth` (tree depth)
        - `min_samples_split` (minimum samples to split)
    - Best values**:** `n_estimators=100`, `max_depth=None`, `min_samples_split=5`

The following metrics were used to evaluate model performance :

- Accuracy**:** Measures the overall correctness of the model.
- Precision**:** Indicates how many of the predicted spam messages are actually spam
- Recall**:** Measures how many of the actual spam messages the model correctly identified.
- F1 Score**:** Combination of precision and recall.

- ROC AUC**:** Summarizes the model's ability to discriminate between spam and non-spam messages across all classification thresholds. An AUC closer to 1.0 indicates excellent performance.

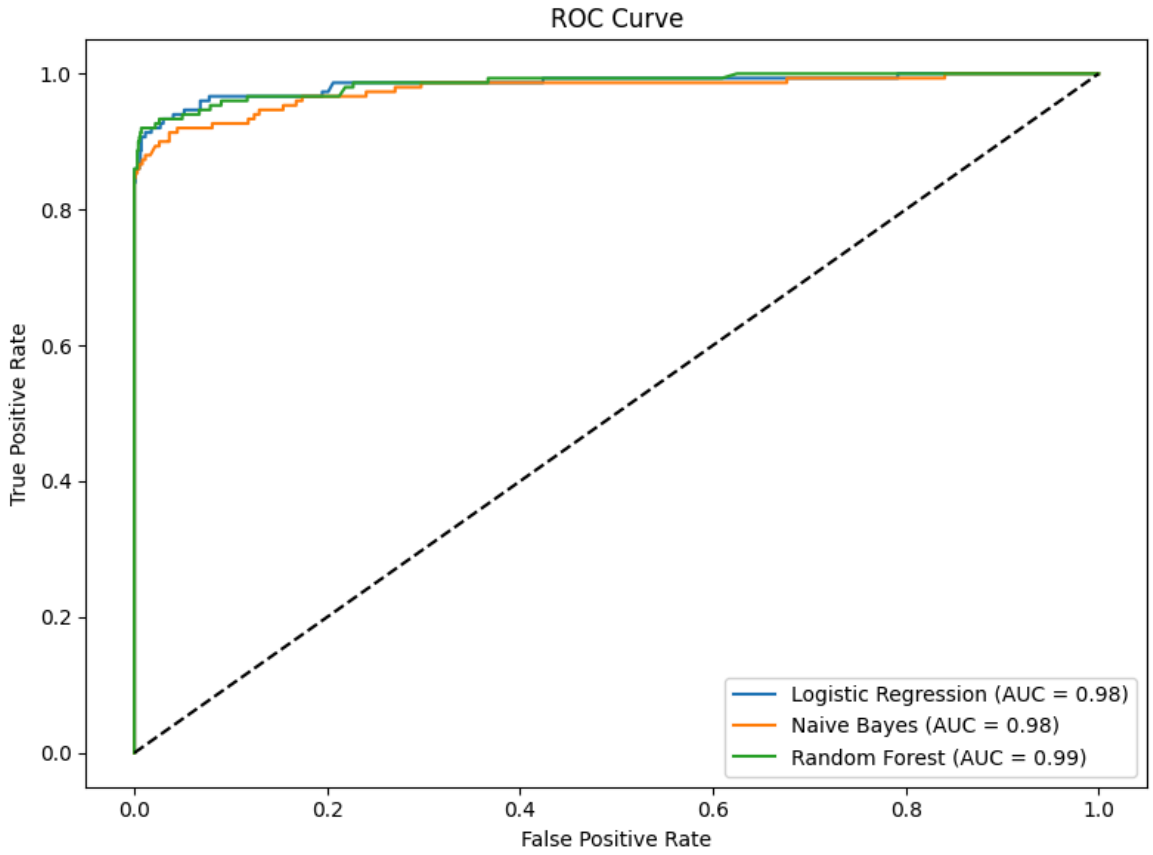
While the real goal is to accurately classify spam vs. ham, there is often a trade-off between precision and recall depending on the real-world consequences: If we care more about not missing spam, we prioritize recall. If we care more about not mislabeling good emails, we prioritize precision.
F1 score ensures balance between false positives and false negatives, and ROC AUC reflects the trade-off between sensitivity and specificity, making them good approximations of the real task goal.

Results

	Accuracy	**Precision**	**Recall**	**F1 Score**	**ROC AUC**	**Confusion matrix**
Logistic Regression	0.9794	0.9847	0.86	0.9181	0.9849	[963 2 21 129]
Naives Bayes	0.9794	0.9922	0.8533	0.9176	0.9768	[964 1] 22 128]
Random Forest	0.9740	1.0	0.8067	0.8930	0.9858	[965 0] 29 121]

Interpretation:

- Logistic Regression and Naive Bayes performed similarly, with Logistic Regression slightly edging out in recall and AUC.
- Random Forest achieved perfect precision but had lower recall, meaning it made no false spam predictions but missed some actual spam.



The ROC (Receiver Operating Characteristic) curve demonstrates how well our spam classifier distinguishes between spam and ham across various threshold settings

All three models are excellent at distinguishing spam from ham (AUC close to 1), especially Logistic Regression and Random Forest, which both nearly achieve perfect discrimination.