SMS Classifier Project Report

1. Project Overview

The objective of this project was to develop a text classification model to classify SMS messages as either spam or non-spam. Utilizing various machine learning techniques, we aimed to build and evaluate models to determine the most effective approach for this classification task.

2. Dataset Description

- **Dataset**: SMS messages labeled as "spam" or "ham" (non-spam).
- **Features**:
  - `Message`: The text content of the SMS.
  - `Label:` classification label ("spam" or "ham").

3. Data Preprocessing

1. **Text Cleaning**:
   - Removed punctuation, special characters, and numbers.
   - Converted text to lowercase.
2. **Tokenization**:
   - Split messages into tokens (words).
3. **Stop Word Removal**:
   - I removed common, non-informative words.
4. **Stemming/Lemmatization**:
   - Reduce words to their root form.
5. **Feature Extraction**:
   - **Bag of Words (BoW)**: Converted text into token counts.
   - **TF-IDF**: Weighted tokens based on their importance.

4. Model Selection

I evaluated the following models:

1. **Logistic Regression**
2. **Random Forest**
3. **Support Vector Machine (SVM)**
4. **Gradient Boosting**
5. **Naive Bayes**
6. **K-Nearest Neighbors (KNN)**

5. Model Training and Evaluation

1. **Splitting Data**:
   - Training set: 80%
   - Testing set: 20%
2. **Evaluation Metrics**:
   - **Accuracy**: Proportion of correctly classified messages.
   - **Precision**: Proportion of true positives out of predicted positives.
   - **Recall:** proportion of true positives out of actual positives.
   - **F1 Score:** harmonic mean of precision and recall.
3. **Performance Results**:

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9704	0.9685	0.9982	0.9831
Random Forest	0.9500	0.9457	0.9995	0.9719
SVM	0.9708	0.9688	0.9984	0.9834
Gradient Boosting	0.9538	0.9530	0.9956	0.9738
Naive Bayes	0.9805	0.9813	0.9964	0.9888
KNN	0.8712	0.8703	1.0000	0.9307

4. **6. Analysis**

- **Naive Bayes**:
  - **Best Performance**: Achieved the highest accuracy (0.9805), precision (0.9813), recall (0.9964), and F1 score (0.9888).
  - **Strengths**: Excellent overall performance, particularly in precision and recall.
- **SVM**:
  - **High Accuracy**: Slightly lower than Naive Bayes but competitive with Logistic Regression (0.9708).
  - **Strong F1 Score**: Comparable to Logistic Regression (0.9834).
- **Logistic Regression**:
  - **Solid Performance**: High accuracy (0.9704) and F1 score (0.9831), close to SVM.
  - **Balanced:** good precision and recall.
- **Random Forest**:
  - **Moderate Accuracy** slightly lower accuracy (0.9500) but very high recall (0.9995).
  - **Balanced Precision and Recall**: Decent F1 score (0.9719).
- **Gradient Boosting**:
  - **Good Performance**: Accuracy (0.9538) and F1 score (0.9738) are good but not the best.
  - **Strong Recall**: Slightly lower in precision but balanced overall.
- **KNN**:

    - **Lowest Accuracy**: Significant drop in accuracy (0.8712) compared to others.
    - **Perfect Recall**: High recall (1.0000) but lower precision and F1 score (0.9307).

7. Best Parameters

For the Logistic Regression model, we fine-tuned the hyperparameters and found the following best configuration:

- **Best Parameters**: `{'C': 10, 'max_iter': 100}`
- **Best Score**: 0.9722

These parameters were selected based on their performance on the validation set, achieving a high accuracy of 0.9722.

8. Recommendations

- **Best Overall Model**: **Naive Bayes** is recommended as it outperforms other models across all evaluation metrics.
- **Alternative High Performers**: **SVM** and **Logistic Regression** also provide strong performance and are suitable alternatives based on their balanced precision and recall.
- **Considerations**: While **KNN** offers perfect recall, its lower accuracy and precision make it less ideal for this classification task.

9. Conclusion

The Naive Bayes model proved to be the most effective in classifying SMS messages as spam or non-spam. Its superior performance in terms of accuracy, precision, recall, and F1 score makes it the preferred choice. The SVM and Logistic Regression models also demonstrated strong performance and are valuable alternatives. Future work could involve exploring more advanced models or techniques to further enhance performance and deploy the model for practical use.

10. References

- **Dataset**: mail_data
- **Libraries Used**: `pandas`, `numpy`, `scikit-learn`, `nltk`, `re`