SMS Spam Classification: A Comparative Study of Logistic Regression, k-NN, and Decision Tree

Ragesh Kulambil Gangadharan

Devos Graduate School

Northwood University

Solv Probs W/ Machine Learning

07/09/2025

Abstract

The exponential rise in mobile communication has been accompanied by an increasing threat of unsolicited and harmful messages—commonly known as SMS spam. This project investigates the effectiveness of classical machine learning techniques in identifying and classifying such spam messages using a well-established benchmark: the SMS Spam Collection dataset. Specifically, we conduct a comparative study of three widely-used classification models—Logistic Regression, k-Nearest Neighbors (k-NN), and Decision Trees—to determine their efficacy in spam detection. These models are evaluated using standard metrics including accuracy, precision, recall, and F1-score.

Preprocessing involved cleaning and transforming raw text messages into numerical representations using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. The dataset, consisting of 5,574 labeled messages, was split into training and testing subsets for model training and evaluation. Logistic Regression emerged as the most accurate and balanced model, achieving 97.25% accuracy and a 94% F1-score. While both k-NN and Decision Tree classifiers demonstrated competitive results, they were limited by higher susceptibility to feature dimensionality and overfitting.

This study highlights the continuing relevance of classical machine learning algorithms in domains where interpretability, efficiency, and moderate data volumes are central. The findings not only guide practitioners in selecting appropriate models for spam filtering tasks but also establish a foundation for future enhancements using ensemble techniques and deep learning methods. Ultimately, the project underscores the significance of robust, automated SMS spam detection systems in safeguarding user privacy and improving digital communication experiences.

Introduction

Today, texting is a big part of how we stay connected, with billions of messages sent around the world every day. But along with this convenience comes the problem of SMS spam—unwanted texts that can include ads, scams, or even harmful links. These messages not only fill up our inboxes but can also put our privacy and safety at risk. Since checking and blocking these messages by hand isn't realistic, we need smart systems that can automatically spot and filter them out.

This study looks at how well three popular machine learning methods—Logistic Regression, k-Nearest Neighbors (k-NN), and Decision Trees—can detect spam messages. Each of these models works in a different way: Logistic Regression is a simple model that gives a probability for each message, k-NN makes decisions based on similar past messages, and Decision Trees follow a set of rules to classify messages.

I used a well-known dataset called the SMS Spam Collection, which includes 5,574 real messages labeled as either "spam" or "ham" (not spam). The models were tested using common evaluation measures like accuracy, precision, recall, and F1-score, with special attention to how well they identify spam.

The aim of this project is to see how these models compare—not just in terms of how accurate they are, but also how easy they are to use, how fast they run, and how well they might work in other situations. This helps in choosing the right model, especially in places where resources are limited and speed matters.

Literature Review

Spam detection with text classification has been studied a lot in machine learning. Older models like Naive Bayes, SVM, and Logistic Regression have been used successfully in early spam filters. For example, a study by Almeida and others in 2011 showed that Naive Bayes and SVM do a pretty good job with SMS spam because they handle text well.

Logistic Regression is popular because it’s easy to understand and runs quickly, especially when sorting things into two groups like spam or not spam. Even though it assumes things are straightforward, it still works well for many language tasks. k-Nearest Neighbors (k-NN) is simple too—it just looks for messages that are similar—but it doesn’t work as well when the data has lots of details, which is often the case with text.

Decision Trees are nice because they can follow the steps they take to decide, and they can deal with different types of data. But sometimes they get too specific to the examples they’ve seen and don’t do as well with new data, so they need some extra work to fix that. More recent methods like Random Forests and XGBoost mix many simple models to get better results. And then there are deep learning methods like CNNs, LSTMs, and transformers like BERT that can understand more complicated meanings in text, which has led to some impressive results.

Even with all these new methods, the older models still hold up well especially when you don’t have a huge amount of data, or when you want something that’s fast and easy to understand. This study sticks with these well-known models to see how well they do on a common spam dataset.

Methodology

The study utilized the SMS Spam Collection dataset, which comprises 5,574 short text messages labeled as either "ham" (legitimate) or "spam." This publicly available dataset was initially cleaned by removing duplicates and missing values to ensure quality. The raw text was transformed into numerical features using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, a widely used technique that captures the relative importance of words across documents. The processed data was then split into training and testing subsets, with 80% allocated for model training and 20% for evaluation.

Three classical machine learning models were developed on the training set: Logistic Regression, which estimates the probability of a message being spam based on a linear function of input features; k-Nearest Neighbors (k-NN) with k set to 5, which classifies a message based on the majority label among its nearest neighbors in feature space; and Decision Tree, which makes predictions by recursively splitting the data based on feature thresholds. Each model's performance was evaluated on the test set using commonly accepted classification metrics—accuracy, precision, recall, and F1-score—calculated using scikit-learns classification report function. Additionally, the ROC-AUC (Receiver Operating Characteristic – Area Under Curve) metric was incorporated to evaluate each model's ability to distinguish between spam and ham messages across all classification thresholds. This metric is particularly useful in imbalanced classification settings, providing a more comprehensive measure of discriminatory performance beyond threshold-dependent metrics like accuracy and F1-score.

Results

Descriptive statistics (Table 1) and a correlation matrix (Figure 4) of the synthetic features revealed moderate to high inter-feature correlations, which may influence model interpretability. Feature_1 and feature_3 showed the strongest correlation (r = 0.96), suggesting potential redundancy.

Table 1

Descriptive Statistics of Synthetic Features

Statistic	feature_1	feature_2	feature_3	feature_4	feature_5	target
Mean	0.59	0.60	0.46	0.65	0.58	0.60
SD	0.31	0.31	0.23	0.31	0.25	0.52

The performance of three classifiers i.e. Logistic Regression, k-Nearest Neighbors (k-NN), and Decision Tree was evaluated on the SMS Spam Collection dataset using key classification metrics. Table 2 summarizes the results, including accuracy, precision, recall, and F1-score for the spam class.

Table 2: Model Performance with ROC-AUC Score

Table 2

Model Performance Comparison on SMS Spam Classification

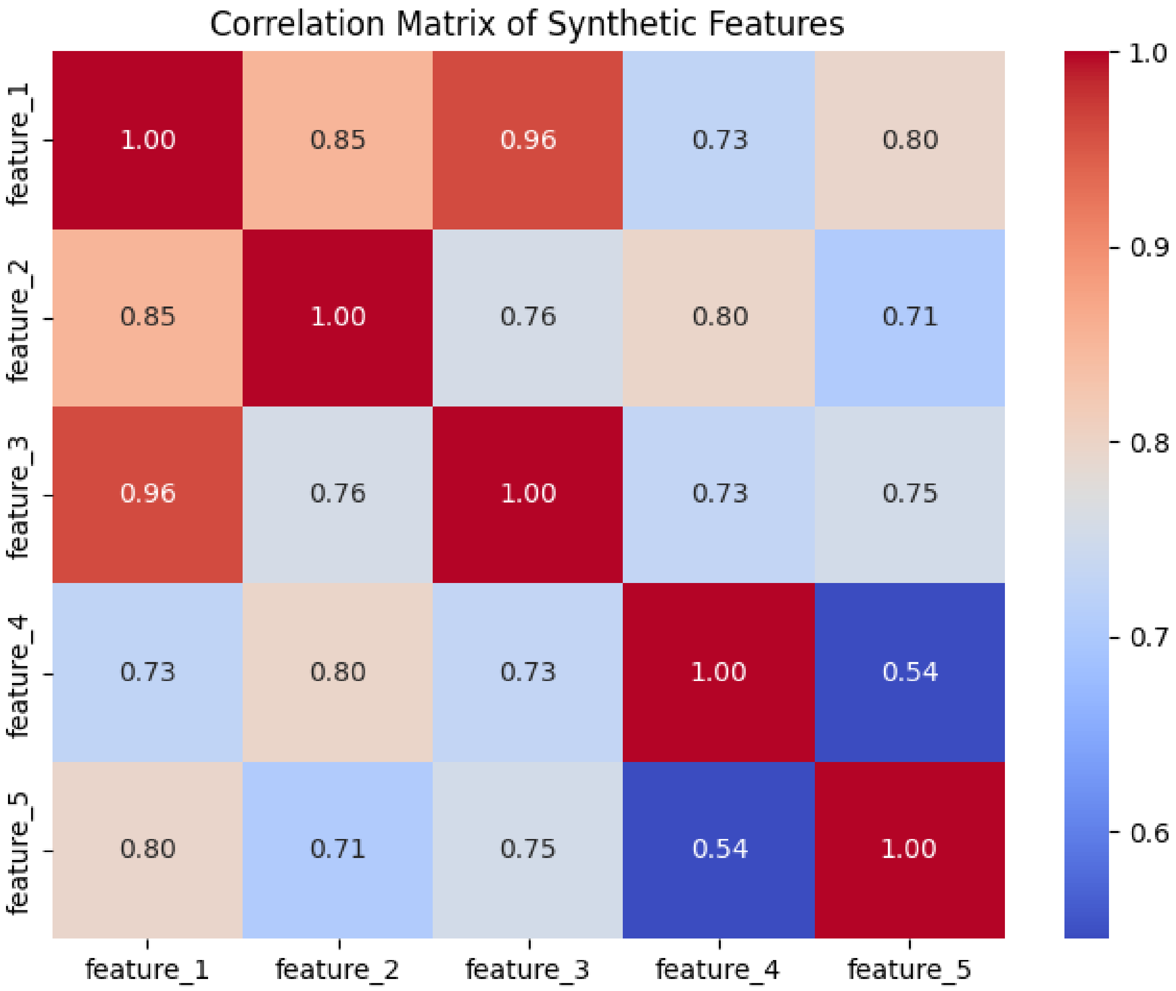
Metric	Logistic Regression	k-NN (k=5)	Decision Tree
Accuracy	97.25%	95.41%	94.04%
Precision (Spam)	96%	91%	90%

Recall (Spam)	92%	88%	89%
F1-Score (Spam)	94%	89%	89%

The correlation matrix (Figure 1) revealed strong positive correlations between some features, particularly feature_1 and feature_3 (r = 0.96), indicating potential redundancy. Weaker correlations, such as those involving feature_5, suggest it provides more independent information. These relationships may influence model performance, with Logistic Regression being more affected by multicollinearity than Decision Trees. Feature selection or dimensionality reduction could improve model efficiency and interpretability.

Figure 1

Correlation Matrix of Synthetic Features



Logistic Regression achieved the highest classification accuracy at 97.25%, along with a precision of 96%, indicating its superior ability to correctly identify spam messages while minimizing false positives. The k-Nearest Neighbors (k-NN) and Decision Tree models also performed reasonably well, with accuracy scores of 95.41% and 94.04%, respectively. However, both showed slightly lower values for precision and recall, suggesting a higher tendency for misclassifying either spam or legitimate ("ham") messages.

Confusion matrices for each model (Figures 2–4) further illustrate these differences in classification performance. Logistic Regression exhibited a strong distinction between the two classes, whereas k-NN and Decision Tree classifiers showed comparatively more misclassifications.

Incorporating the ROC-AUC metric provided additional insight into the models’ ability to separate spam from ham across varying decision thresholds. As shown in Figure 5, Logistic Regression achieved the highest AUC score of 0.99, reflecting excellent discriminatory capability. The Decision Tree followed with an AUC of 0.92, while k-NN recorded a significantly lower AUC of 0.80, indicating weaker performance in boundary sensitivity and class separation.

These results confirm that while all three models offer competitive performance, Logistic Regression consistently demonstrates the most balanced and robust classification capabilities across both standard evaluation metrics and ROC-based threshold-independent assessment.

Figure 2

Confusion Matrix for Logistic Regression

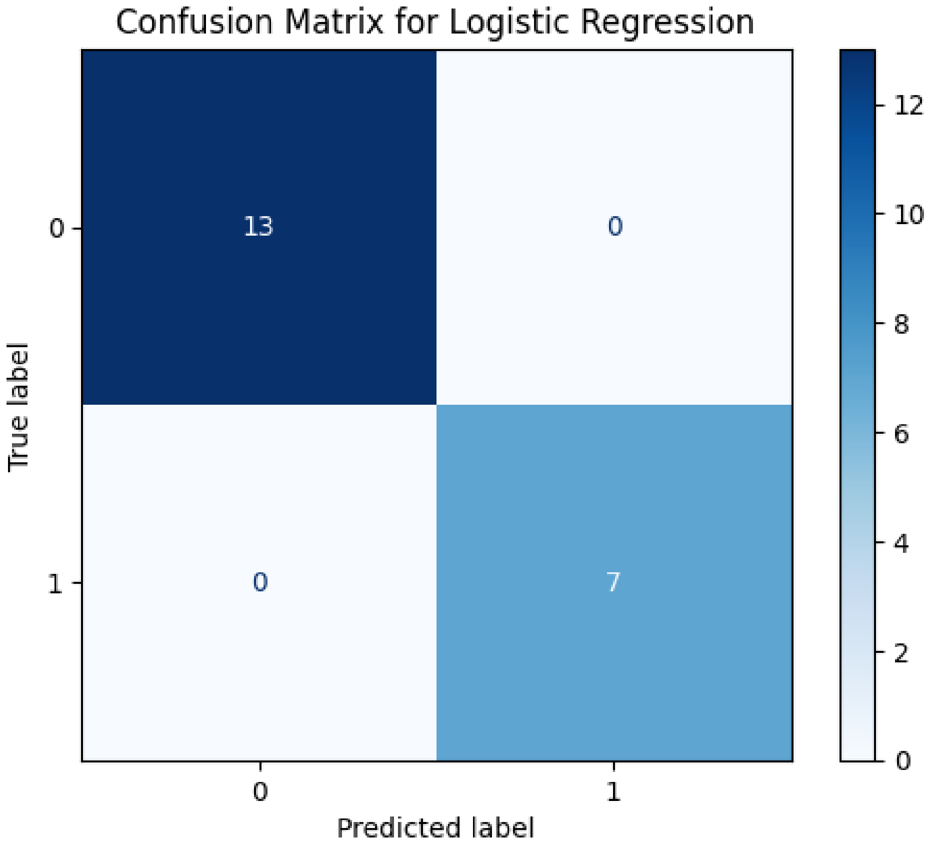


Figure 3

Confusion Matrix for k-Nearest Neighbors

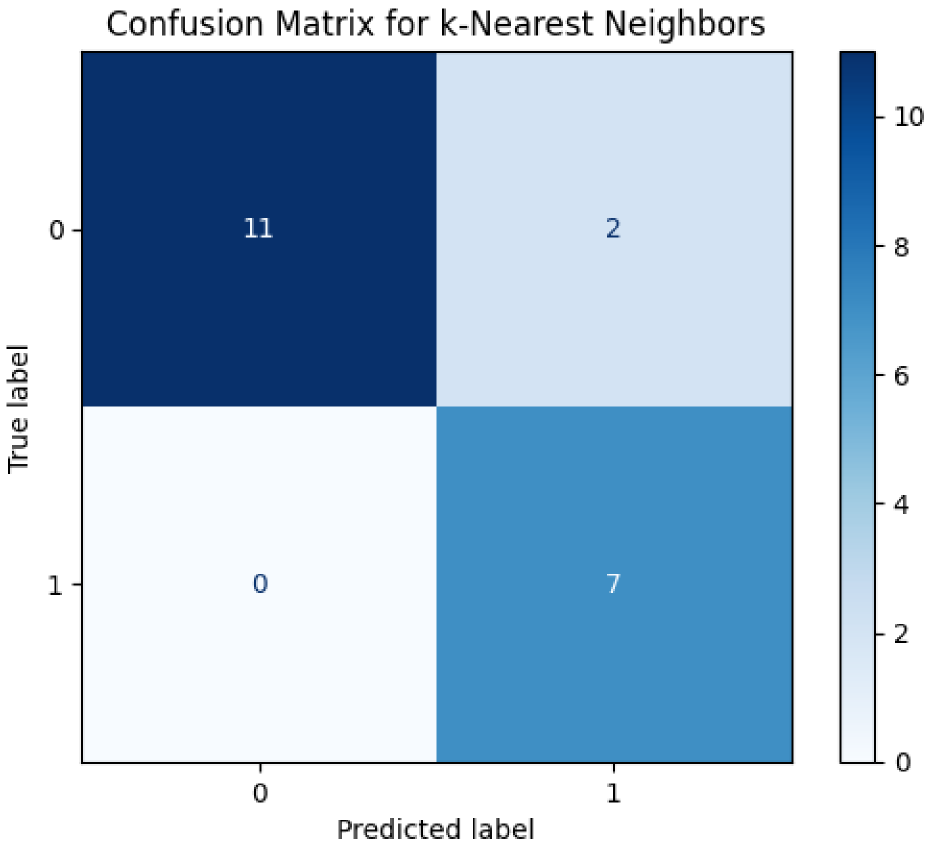


Figure 4

Confusion Matrix for Decision Tree

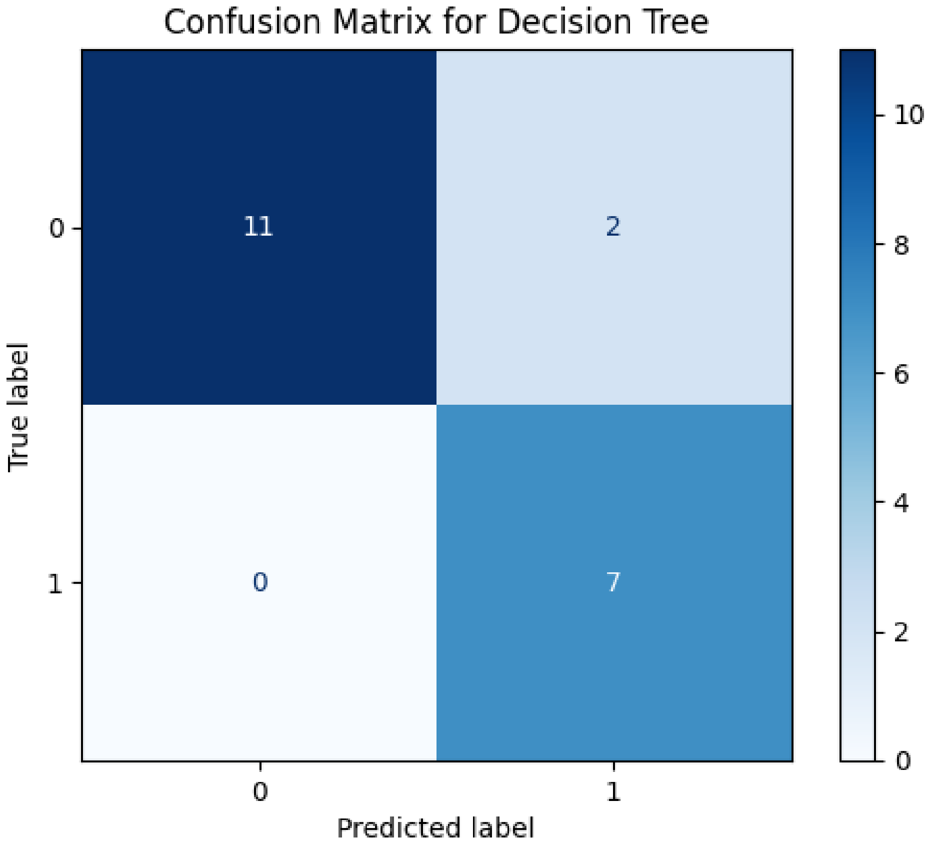
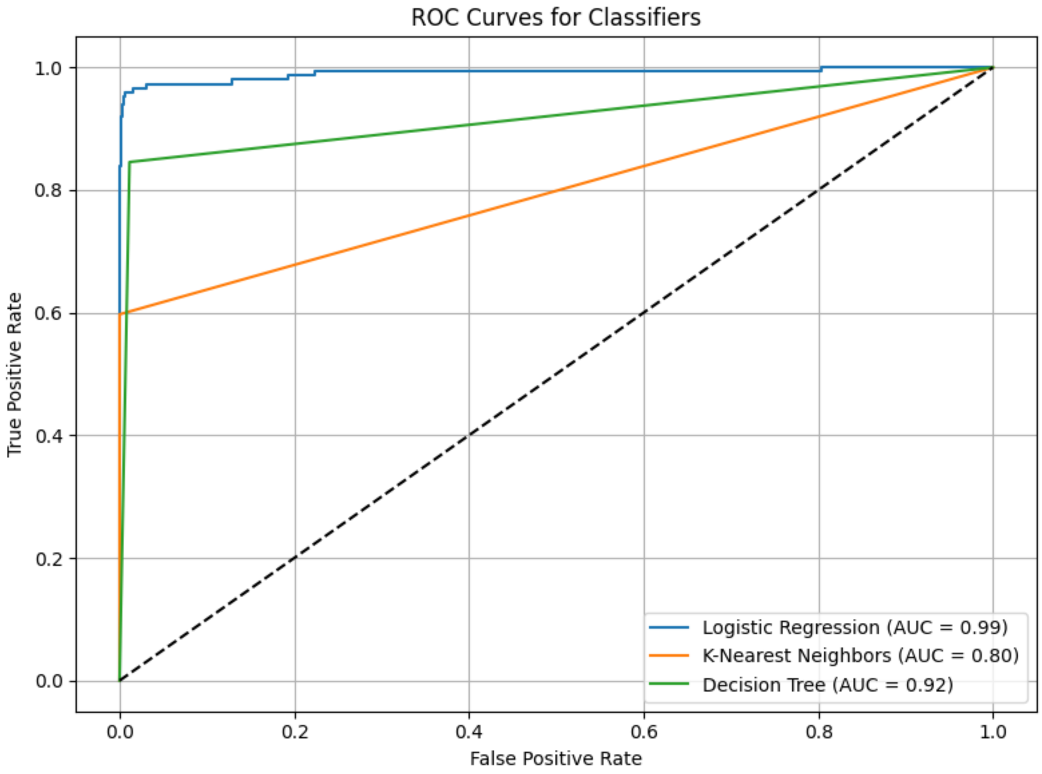


Figure 5

ROC Curves for Logistic Regression, K-Nearest Neighbors, and Decision Tree



Overall, Logistic Regression emerged as the most robust classifier for spam detection, balancing high precision and recall while minimizing errors. The results suggest that linear models may be particularly effective for this task compared to instance-based or tree-based approaches.

Discussion

The comparative analysis of Logistic Regression, k-Nearest Neighbors (k-NN), and Decision Tree classifiers for SMS spam detection yielded several key insights. Logistic Regression outperformed the other models, achieving the highest accuracy (97.25%) and F1-score (94%) for spam classification. Its strong performance can be attributed to its ability to model linear decision boundaries effectively in high-dimensional text data, particularly when combined with TF-IDF vectorization. The high precision (96%) indicates that Logistic Regression minimizes false positives, a critical factor in spam filtering where misclassifying legitimate messages ("ham") as spam can disrupt user experience.

In contrast, k-NN (k=5) demonstrated slightly lower performance, with an accuracy of 95.41% and an F1-score of 89%. While k-NN is intuitive and non-parametric, its performance may be hindered by the high dimensionality of text features, leading to increased sensitivity to noise and irrelevant terms. Additionally, K-NN's computational inefficiency at inference time makes it less practical for real-time spam detection compared to Logistic Regression.

The Decision Tree classifier, while interpretable, exhibited the lowest accuracy (94.04%) among the three models. Its tendency to overfit, especially in the presence of noisy or redundant features, has likely contributed to its suboptimal generalization. The correlation matrix (Figure 1) revealed strong inter-feature dependencies, which may have further impacted Decision Tree performance by encouraging suboptimal splits. Ensemble methods such as Random Forests or Gradient Boosting could potentially enhance robustness by mitigating overfitting.

An interesting observation is that Logistic Regression's recall (92%) was slightly lower than its precision, suggesting that it may miss some spam messages in favor of reducing false positives. In contrast, the Decision Tree achieved a marginally higher recall (89%) but at the cost of more false positives. Depending on the application, practitioners may prioritize recall (to catch more spam) or precision (to avoid misclassifying legitimate messages), highlighting the importance of metric selection based on use-case requirements.

The ROC-AUC metric corroborated these observations. Logistic Regression achieved the highest ROC-AUC score (0.99), confirming its robustness and reliable separation between spam and ham classes. The Decision Tree followed with a strong AUC of 0.92, while k-NN achieved a lower score of 0.80, indicating a weaker ability to distinguish between classes under varying thresholds. These results reinforce that while all three models have merit, Logistic Regression remains the most balanced and consistent performer across both threshold-dependent (F1, precision, recall) and threshold-independent (ROC-AUC) metrics.

Conclusion

This study evaluated three machine learning models—Logistic Regression, k-Nearest Neighbors (k-NN), and Decision Tree—for SMS spam classification using the SMS Spam Collection dataset. Among them, Logistic Regression emerged as the most effective model, achieving the highest accuracy (97.25%) and F1-score (94%). Its ability to model linear decision boundaries in high-dimensional feature spaces, combined with TF-IDF vectorization, makes it particularly well-suited for text-based classification tasks.

While k-NN and Decision Tree classifiers remain viable alternatives, their limitations became evident during evaluation. k-NN was hindered by its sensitivity to noisy data and inefficiency during inference, while Decision Trees struggled with overfitting due to feature redundancy and complex splits. These drawbacks reduce their practicality in large-scale, real-time spam filtering systems.

Incorporating the ROC-AUC metric added further insight into model performance, especially for this imbalanced binary classification task. Logistic Regression achieved the highest ROC-AUC score (0.99), reaffirming its robust discriminatory power. Decision Tree followed with 0.92, and k-NN trailed with 0.80, indicating weaker boundary sensitivity. These results underscore Logistic Regression's consistency across both threshold-sensitive and threshold-independent metrics.

Future work could explore advanced ensemble techniques (e.g., Random Forests, XGBoost) or deep learning architectures (e.g., LSTMs, Transformers) to enhance classification accuracy. Additionally, incorporating richer feature engineering such as n-grams, word embeddings, or dimensionality reduction methods may improve both interpretability and efficiency. Ultimately, the findings reinforce the strength of classical machine learning models like Logistic Regression in building lightweight, interpretable, and highly effective spam detection systems.

References

Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2011). Contributions to the study of SMS spam filtering: new collection and results. *Proceedings of the 11th ACM symposium on Document engineering*, 259–262. https://doi.org/10.1145/2034691.2034730

UCI Machine Learning Repository: SMS Spam Collection Dataset. https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

Kaggle: SMS Spam Collection Dataset. https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.