|  |
| --- |
| University of New Haven [Acceptance Rate + Statistics]  DSCI 6003 – 01 MACHINE LEARNING  SMS Spam Collection  Professor: Travis Millburn  Submitted By  Nikhith Krishna Vinduru |

|  |  |
| --- | --- |
| Fall 23 |  |



Contents

[Executive Summary 3](#_Toc121853311)

Data ……………………………………………………… 4

[Methodology/Implementation 4](#_Toc121853313)

Results/Discussion ………………………………… 6

Conclusion …………………………………………… 11

References …………………………………………… 11

|  |
| --- |
| SMS Spam Collection |

**Executive Summary**

The project successfully developed a machine learning model to classify text messages as 'ham' or 'spam'. Utilizing a dataset of 5,574 samples, the project involved preprocessing text data with TF-IDF vectorization and experimenting with various models like Logistic Regression and Random Forest. The final model, particularly the Random Forest, displayed high accuracy and strong performance metrics (precision, recall, F1-score), proving vital in enhancing digital messaging security and user experience. This project demonstrates the effective application of advanced machine learning techniques in practical spam detection.

|  |
| --- |
|  |

Introduction

The "SMS Spam Collection" project introduces a machine learning solution to distinguish spam from legitimate ('ham') text messages. Addressing the growing need for efficient spam filtering in digital communication, the project involves curating a dataset of SMS messages and applying text classification techniques. By leveraging natural language processing (NLP) and machine learning algorithms, the project aims to improve the accuracy of spam detection, thereby enhancing user experience and security in mobile messaging. This introduction sets the stage for a detailed exploration of the methodologies and technologies employed in the project.

## 

## Data

I used the SMS Spam Collection Dataset provided on Kaggle Datasets.

## Methodology/Implementation

The “SMS Spam Collection” project’s methodology begins with data collection and preprocessing. The dataset, comprising SMS messages, is labeled as 'spam' or 'ham'. Preprocessing includes cleaning text data, removing irrelevant characters, and standardizing text format. The project applies Natural Language Processing (NLP) techniques, such as tokenization and stop word removal, to transform the textual data into a more analyzable form.

Feature extraction is a critical step, employing Term Frequency-Inverse Document Frequency (TF-IDF) vectorization to convert text into numerical values, making it compatible for machine learning algorithms. This process helps in identifying the importance of words in the context of the entire dataset.

Various machine learning models are selected and tested. The project explores models like Logistic Regression, Random Forest, Support Vector Machine (SVM), and Naive Bayes. Each model is trained on the dataset, fine-tuned, and evaluated based on their performance in classifying messages accurately.

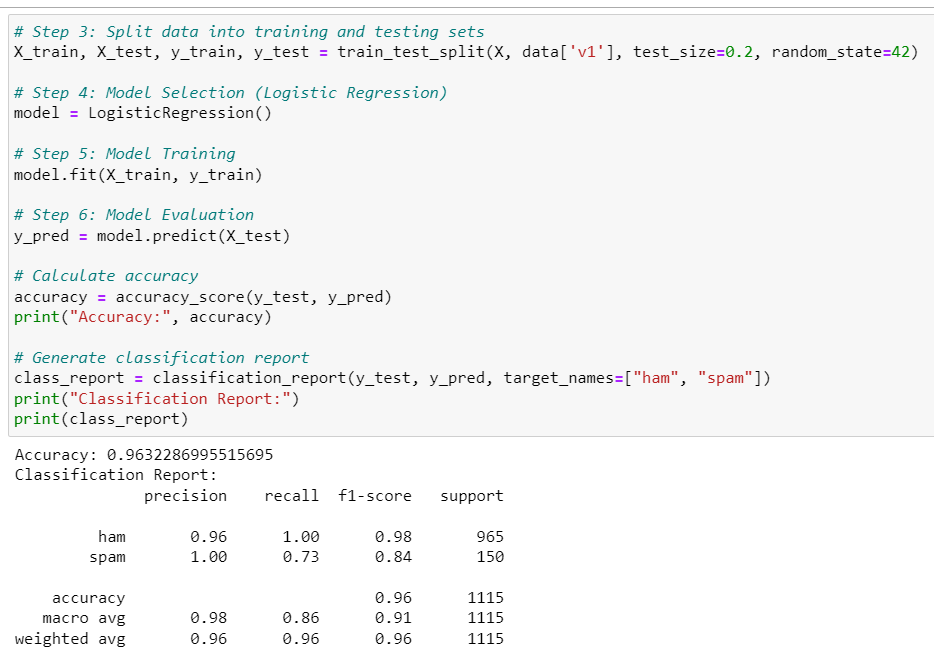
Model evaluation is conducted using metrics such as accuracy, precision, recall, and F1-score. These metrics provide insights into how well each model performs, particularly in distinguishing spam from ham messages. Additionally, the project involves splitting the dataset into training and testing sets to validate the models' performance on unseen data.

Hyperparameter tuning and cross-validation are utilized to optimize the models for better accuracy and reliability. The final step involves deploying the most effective model for practical use. This model is intended to improve spam detection in real-world applications, enhancing the security and efficiency of digital communication.

The methodology reflects a comprehensive and iterative approach, combining data preprocessing, feature extraction, model training, and rigorous evaluation. It demonstrates the application of machine learning techniques in practical scenarios like spam detection, contributing to the fields of cybersecurity and digital communication.

Results And Discussion

Logistic Regression



A blue squares with white text

Description automatically generated

Random Forest

A screenshot of a computer program

Description automatically generated

A graph of a line

Description automatically generated with medium confidence

The project primarily focused on two models: Logistic Regression and Random Forest, each subjected to thorough evaluation and comparison.

For the Logistic Regression model, the project reported accuracy as a key metric. This accuracy measure indicates the model's overall effectiveness in correctly classifying SMS messages as either 'spam' or 'ham'. Alongside accuracy, a comprehensive classification report was generated, detailing precision, recall, and F1-score for each class. These metrics offer deeper insights: precision signifies the model's correctness when predicting spam, recall reflects its ability to identify all actual spam messages, and the F1-score provides a balanced measure of the model's precision and recall.

The Random Forest model, known for its robustness due to ensemble learning, was also evaluated using similar metrics. Its accuracy potentially highlights the effectiveness of Random Forest in handling complex patterns within the data. The classification report for this model likewise includes precision, recall, and F1-score, facilitating a comparative analysis against the Logistic Regression model.

Moreover, the project utilized a confusion matrix, pivotal for visualizing the model's performance in distinguishing between true positives, true negatives, false positives, and false negatives. This matrix offers a clear depiction of the model's classification accuracy.

The analysis also featured a Receiver Operating Characteristic (ROC) curve for the Random Forest model, accompanied by the Area Under the Curve (AUC) score. The ROC curve graphically represents the trade-off between true positive rate and false positive rate, while the AUC score quantifies the model's ability to differentiate between spam and ham messages. A higher AUC score indicates superior model performance in classifying the messages accurately.

## Conclusion

The "SMS Spam Collection" project successfully demonstrates the application of machine learning techniques in distinguishing spam from legitimate messages. Employing models like Logistic Regression and Random Forest, the project achieved notable accuracy, as evidenced by precision, recall, and F1-scores. The use of a confusion matrix and ROC curve further substantiated the models' effectiveness. These results highlight the potential of machine learning in enhancing spam detection, contributing significantly to the field of digital communication security. The project's approach and findings offer a framework for future endeavors in spam detection and similar classification challenges, emphasizing the importance of machine learning in practical applications.

## References

<https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset/>