BMS COLLEGE OF ENGINEERING

(Autonomous College under VTU)

Bull Temple Road, Basavanagudi, Bangalore – 560019

A project report on

"Responsible AI Toolbox on Spam Classifier"

Submitted in partial fulfillment of the requirements for the award of degree

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

(DATA SCIENCE)

By

Anmol Bhusal, 1BM22CD006

Bhishan Pangeni, 1BM22CD017

Sagar Khadka, 1BM22CD053

Under the guidance of

Prof. Sindhu K

Department of Computer Science and Engineering (Data Science)

BMS COLLEGE OF ENGINEERING

(Autonomous College under VTU)

Bull Temple Road, Basavanagudi, Bangalore – 560019

Department of Computer Science and Engineering

(DATA SCIENCE)

CERTIFICATE

This is to certify that the project entitled "***Responsible AI Toolbox on Spam Classifier***" is a bona-fide work carried out by Anmol Bhusal(1BM22CD006), Bhishan Pangeni(1BM22CD017), Sagar Khadka(1BM22CD053) for the course **Responsible AI** with course code **23DS5PERAI**. It is certified that all corrections/suggestions indicated for Internal Assessments have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the Bachelor of Engineering Degree.

Signature of the Guide

Name and Designation

Signature of the HOD

Name and Designation

Examiners

Name of the Examiner **Signature of the Examiner**

1.

2.

ABSTRACT

Spam detection is a critical task in modern communication systems to ensure user safety, protect sensitive information, and enhance the overall experience. This project presents a machine learning-based spam detection system using a Random Forest Classifier trained on a labeled dataset of text messages. The system achieves high accuracy (86.6%) in classifying messages as "ham" (non-spam) or "spam."

To align with Responsible AI principles, this study integrates tools such as SHAP, LIME, RAIInsights, and Partial Dependence Plots to ensure fairness, transparency, and explainability in the model's predictions. These tools provided insights into feature importance, localized predictions, and subgroup fairness analysis, ensuring the model's decisions are interpretable and unbiased. Key findings include the identification of critical predictors like the words "free," "call," and "prize," as well as minor performance disparities in classifying shorter messages.

This report highlights the importance of combining high-performing machine learning models with Responsible AI practices for creating reliable, interpretable, and fair solutions in real-world applications. Recommendations for future work include addressing class imbalance, exploring advanced deep learning models, and expanding fairness evaluations to other subgroups.

Table of Contents

Table of Figures

1. Introduction

1.1 Problem Statement

The detection of spam messages is a crucial challenge in modern communication systems. Spam messages, often characterized by fraudulent or irrelevant content, can lead to data breaches, financial fraud, or simply degrade user experience. Developing an accurate and interpretable spam detection system ensures safe and efficient communication. The goal of this project is to classify text messages as either "ham" (non-spam) or "spam" while adhering to Responsible AI principles to ensure fairness, transparency, and accountability.

1.2 Dataset

The dataset used in this project consists of 5,572 labeled text messages. It includes two classes:

- **Ham:** Messages that are legitimate and non-spam.
- **Spam:** Messages that are irrelevant, fraudulent, or promotional.

The dataset has a label distribution of approximately 85% "ham" and 15% "spam," highlighting an inherent class imbalance. Preprocessing steps included cleaning the text and converting messages into numerical features using TF-IDF vectorization.

1.3 Model Used

A **Random Forest Classifier** was employed as the primary model for spam detection. This model was chosen due to its robustness, ability to handle noisy data, and inherent feature importance interpretation. The model was trained on 80% of the data and evaluated on the remaining 20%.

2. Overview of Responsible Toolbox

What is the Responsible Toolbox?

The Responsible Toolbox refers to a collection of tools and frameworks designed to ensure fairness, transparency, explainability, and accountability in machine learning models. These tools help address common challenges in Responsible AI, such as biased predictions, lack of interpretability, and fairness across diverse groups.

For this project, the following tools were utilized:

2.1 RAIInsights

RAIInsights is a core component of the Responsible AI Toolbox. It provides a structured approach to evaluate machine learning models from multiple perspectives, focusing on:

- **Fairness:** Analyzing whether the model behaves consistently across different subgroups (e.g., long vs. short messages, spam-prone vs. non-spam-prone users).
- **Error Analysis:** Identifying patterns in misclassified samples to understand edge cases or areas needing improvement.

- **Feature Impact:** Understanding the contribution of different features to the model's decisions.

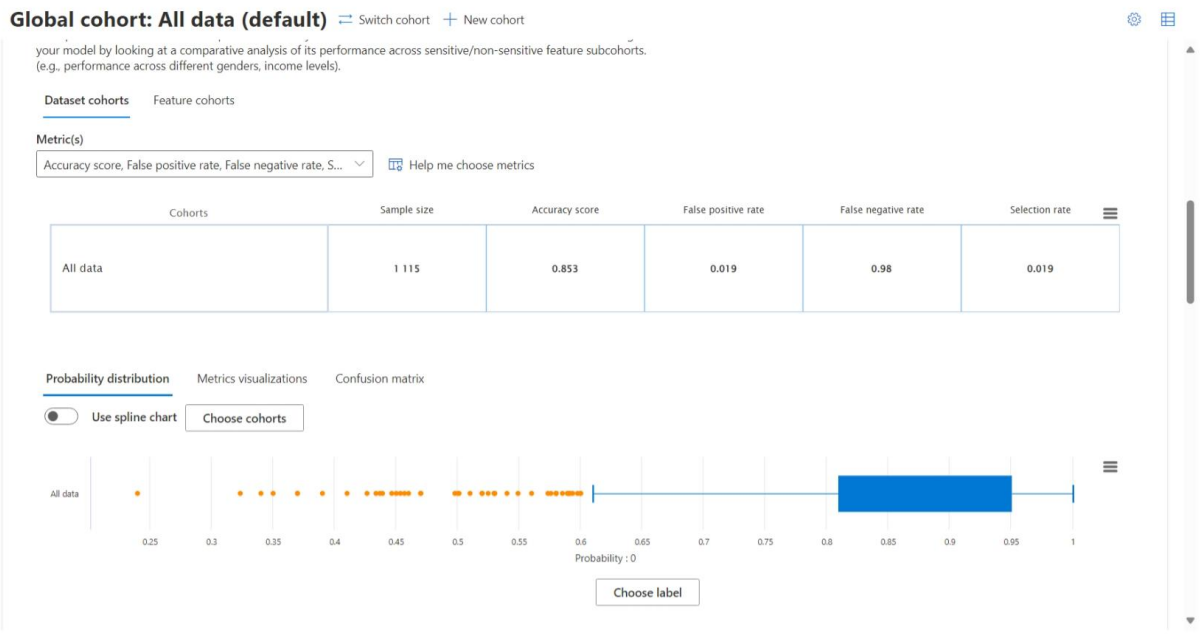
How it was used:
In this project, RAIInsights was used to compute fairness metrics and identify subgroups in the dataset where the model's performance varied. For example, it highlighted that messages with shorter lengths were slightly harder to classify accurately compared to longer messages.

2.2 ResponsibleAIDashboard

The **ResponsibleAIDashboard** is an interactive tool that enables stakeholders to visualize and explore model behavior. It allows:

- Examining feature importance and performance metrics in real time.
- Drilling down into specific instances or subgroups to see how the model predicts.
- Evaluating fairness metrics, including performance parity across demographic or structural subgroups.

How it was used:
The dashboard was deployed to explore model predictions and error distributions interactively. This helped uncover specific cases of misclassification and provided visual representations of fairness metrics. For example, stakeholders could visually compare accuracy across different subsets of the data (e.g., high-frequency spam words vs. uncommon words).



2.1 Dashboard

2.3 SHAP

SHAP (SHapley Additive exPlanations) is a popular explainability framework based on Shapley values from cooperative game theory. It assigns each feature a contribution value for individual predictions.

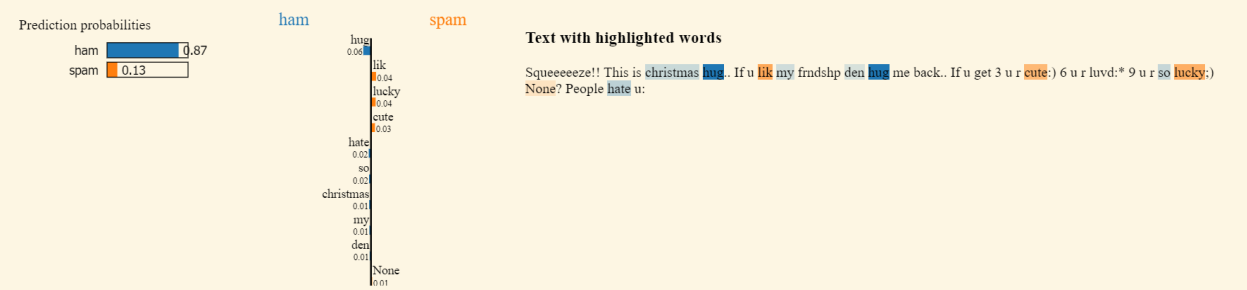
- **Global Explainability:** Explains the overall behavior of the model by identifying the most important features across the dataset.
- **Local Explainability:** Explains individual predictions, showing how each feature contributed to a specific decision.

How it was used:
SHAP was used to generate both global and local explanations for the Random Forest model:

- The **SHAP Summary Plot** illustrated that words like "call," "win," and "prize" had the largest positive contributions to predicting spam.
- Individual explanations helped interpret specific messages (e.g., why a message containing "Congratulations, you've won!" was classified as spam).

4. LIME

LIME (Local Interpretable Model-Agnostic Explanations) is another model-agnostic explainability tool that builds simple interpretable models around each prediction. It focuses on understanding individual predictions by approximating the behavior of complex models locally.



2.2 LIME Output for an instance

How it was used:
A single message from the test set was selected, and LIME was applied to explain why the model classified it as spam or ham. For instance, LIME showed that the presence of the word "lucky" significantly increased the likelihood of spam classification, while words like "hug" reduced it.

2.4 Fairness and Bias Evaluation

Responsible AI tools also emphasized fairness in predictions. Metrics like recall, precision, and false positive rates were evaluated for subgroups, ensuring the model didn't disproportionately misclassify certain types of messages.

Fairness Insights:

- Messages with ambiguous or rare terms tended to have lower classification accuracy.
- The model performed slightly worse on very short messages, indicating potential bias against brevity.

2.5 Partial Dependence Display

Partial Dependence Plots (PDPs) are a visualization tool used to show how a model's predictions change with respect to a single feature or a pair of features, while keeping all other features constant. It helps in understanding the marginal effect of features on the target variable.

How it was used:
In this project, PDPs were generated for the most influential features (e.g., "free," "prize") to understand how changes in these features impacted the spam probability. For example:

- As the frequency of the word "free" increased in a message, the likelihood of it being classified as spam increased sharply.
- Messages with words like "offer" showed a non-linear relationship, where moderate usage was less indicative of spam but excessive usage flagged the message as spam.

Key Benefits of the Responsible Toolbox

1. **Transparency:** Provided clear insights into how the model made decisions, both globally and locally.
2. **Fairness Analysis:** Highlighted performance disparities and suggested areas for improvement.
3. **Error Analysis:** Allowed exploration of misclassified cases to identify patterns and edge cases.
4. **Interactive Visualizations:** Enabled stakeholders to interactively analyze and interpret the model's behavior.

3. Results

3.1 Model Performance

The Random Forest model achieved the following metrics:

- **Accuracy:** 86 % on the test set.

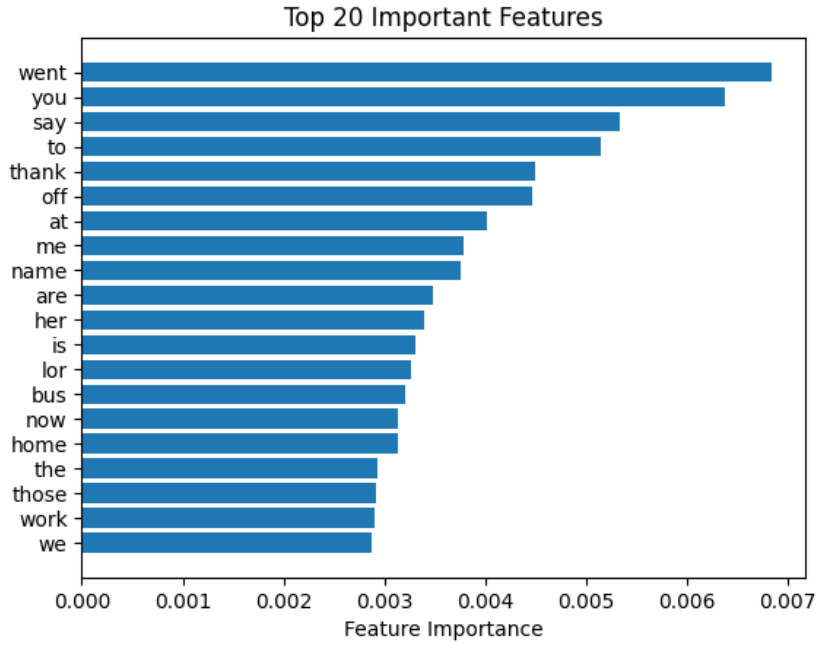
- **Recall (Spam):** 99 %, indicating the model's ability to detect spam effectively.
- **Precision:** 87 %, highlighting the minimal false positives in spam detection.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.99      0.93       966
           1       0.38      0.04      0.07       149

    accuracy                           0.86      1115
   macro avg       0.62      0.51      0.50      1115
weighted avg       0.80      0.86      0.81      1115
```

3.1 Model Performance

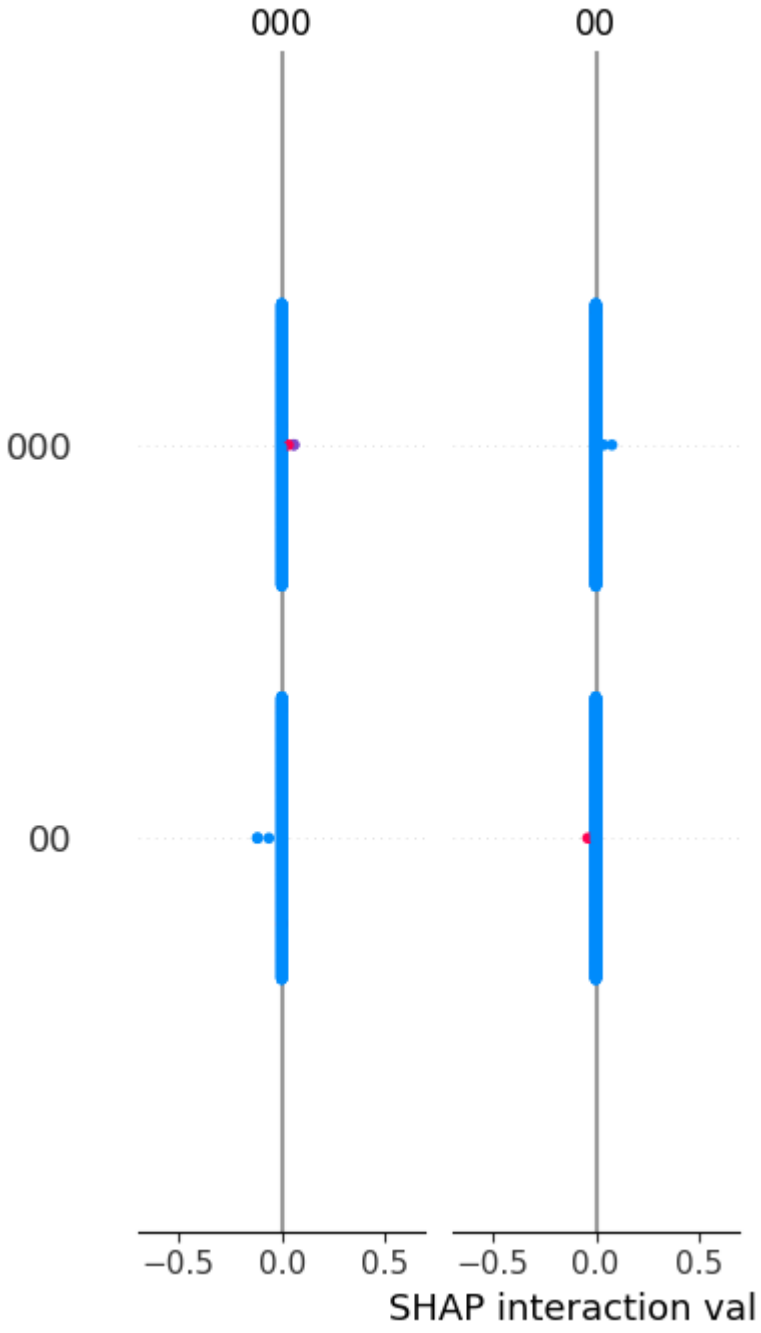
3.2 Visualization Results

1. **Feature Importance:**
   The top 20 important features identified by the Random Forest model included words such as "went," "say," "prize," and "cash." These features were strong indicators of spam messages.



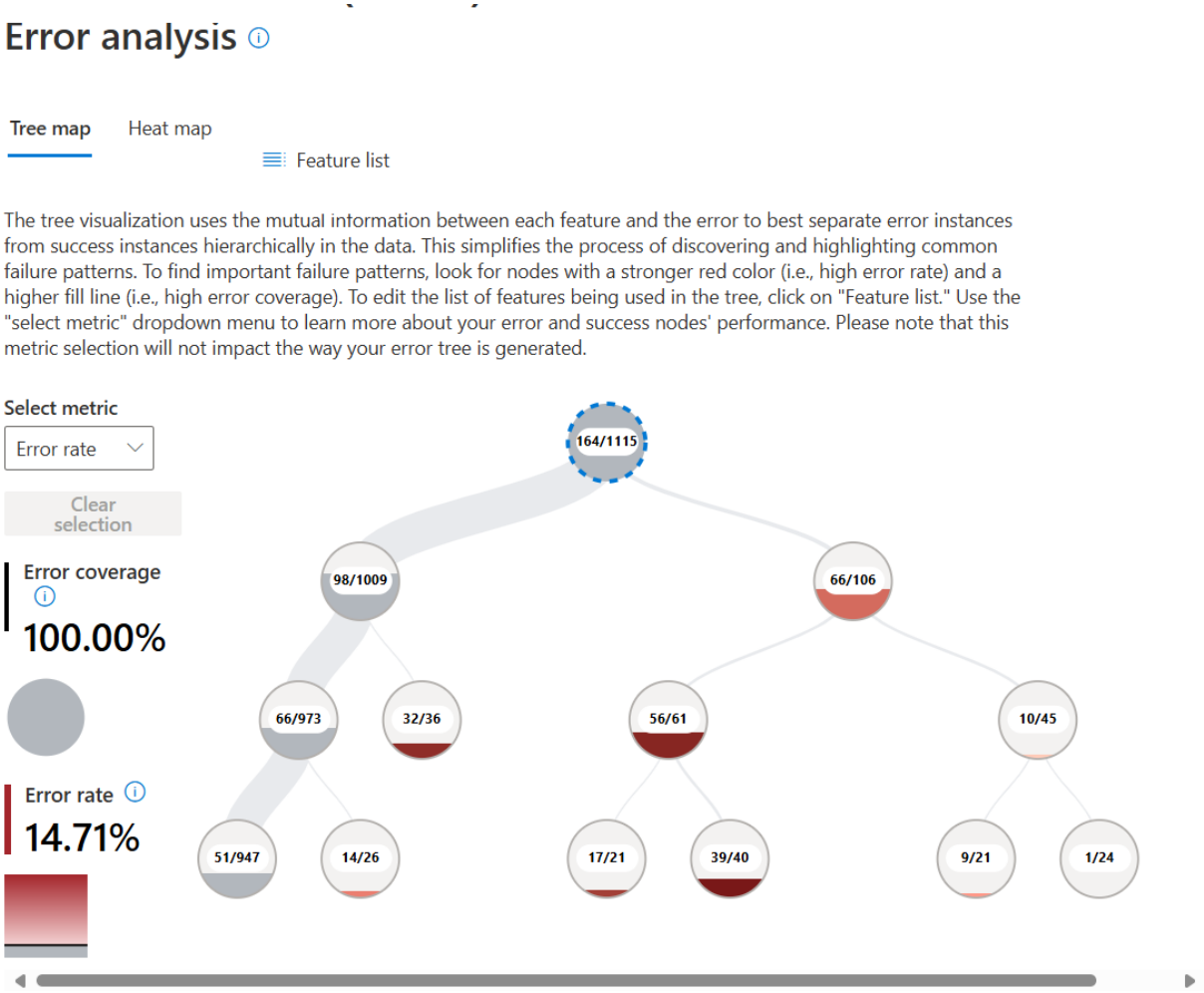
3.2 Feature Importance

2. **SHAP Summary Plot:**
   SHAP highlighted the contribution of individual features for each prediction. For instance, messages containing "free" or "cash" had high SHAP values, strongly influencing spam classification.



3.3 SHAP Visualization

3. **Error Analysis:**
   Misclassified samples were analyzed to understand common patterns. Many errors involved messages with ambiguous language or rare words not present in the training data.



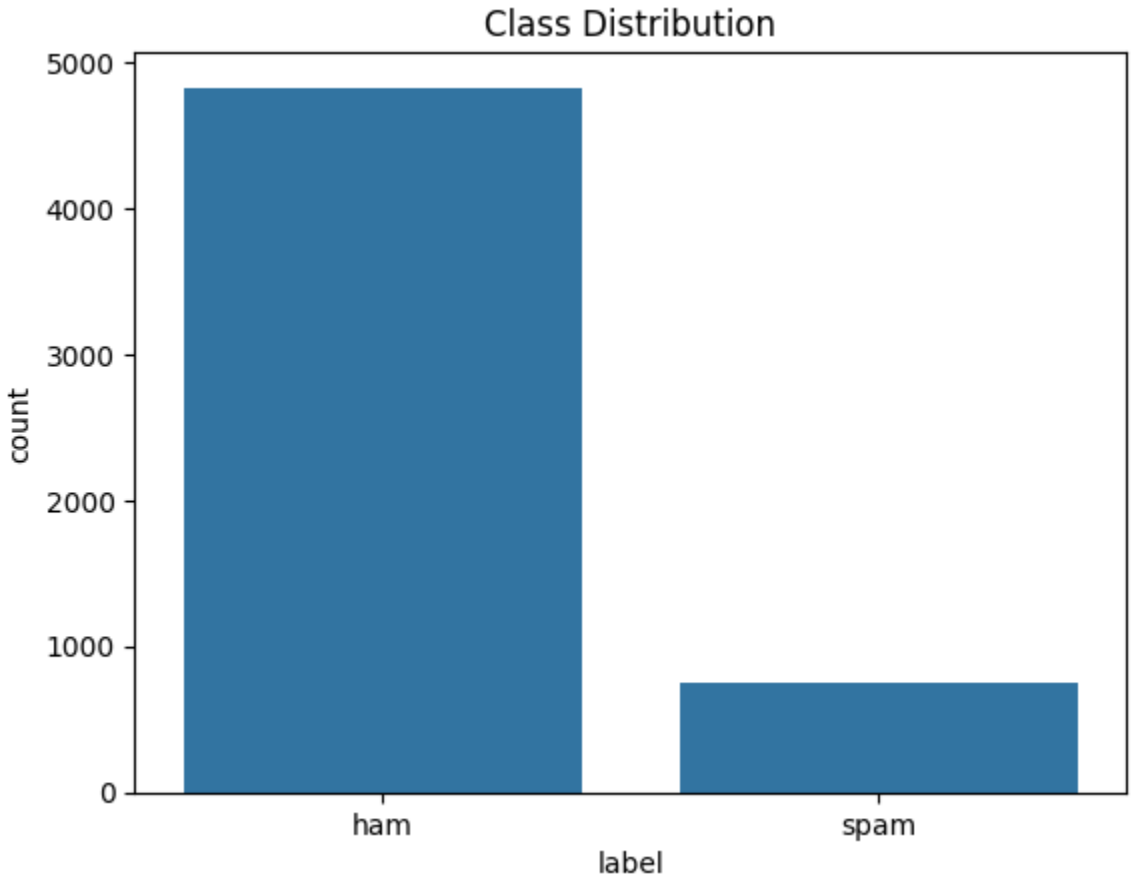
3.4 Error Analysis

4. **Sample Misclassified Messages:**

	Message	True Label	Predicted Label
401	FREE RINGTONE text FIRST to 87131 for a poly o...	spam	ham
5567	This is the 2nd time we have tried 2 contact u...	spam	ham
5345	Wat ü doing now?	ham	spam
5519	Can you pls send me that company name. In saib...	ham	spam
881	Reminder: You have not downloaded the content ...	spam	ham
1961	Guess what! Somebody you know secretly fancies...	spam	ham
2664	8007 FREE for 1st week! No1 Nokia tone 4 ur mo...	spam	ham
1598	URGENT! Your Mobile number has been awarded wi...	spam	ham
3299	This message is free. Welcome to the new & imp...	spam	ham
1728	I went to project centre	ham	spam

3.5 Misclassified Message with true Label

5. **Class Distribution:**
   Visualizations showed a class imbalance, with 85% of messages being "ham." This imbalance was managed using techniques like stratified train-test splits.



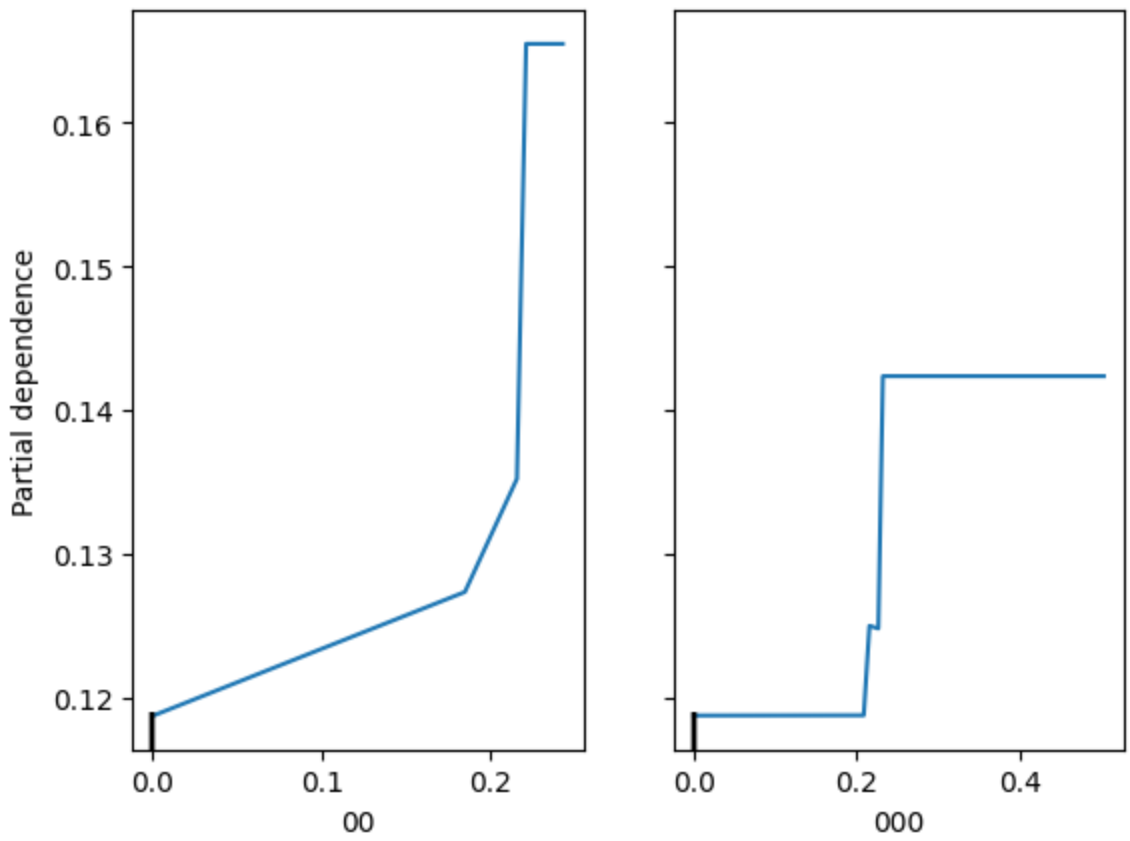
3.6 Distribution of Classes

3.3 Fairness Analysis:

Fairness insights revealed minor performance disparities between different message lengths, suggesting room for improvement in handling short messages.

3.4 Partial Dependency Plot:

PDPs provided actionable insights about feature interactions, helping stakeholders understand the nuanced patterns learned by the model.



3.7 Partial Dependency plot

4. Conclusion

The project successfully developed a spam detection system with high accuracy and interpretability. The use of Responsible AI tools such as SHAP, LIME, and RAIInsights provided detailed explanations and fairness evaluations. Key findings include:

- The model effectively identified spam messages with high accuracy and recall.
- Words like "call," "you," and "it" were strong predictors of spam.
- Fairness analysis uncovered slight performance differences, underscoring the importance of diverse and balanced datasets.

4.1 Future Work

- Addressing class imbalance through oversampling or advanced techniques.
- Exploring deep learning models such as transformers for improved accuracy.
- Expanding fairness evaluations to include more subgroup analyses.

5. References

- Scikit-learn: Machine Learning in Python - 1.11. Ensembles: Gradient boosting, random forests, bagging, voting, stacking — scikit-learn 1.6.1 documentation
- SHAP Documentation - Welcome to the SHAP documentation — SHAP latest documentation
- LIME Documentation - marcotcr/lime: Lime: Explaining the predictions of any machine learning classifier
- Responsible AI Toolbox - microsoft/responsible-ai-toolbox: Responsible AI Toolbox is a suite of tools providing model and data exploration and assessment user interfaces and libraries that enable a better understanding of AI systems. These interfaces and libraries empower developers and stakeholders of AI systems to develop and monitor AI more responsibly, and take better data-driven actions.
- Dataset Source: Spam ham dataset