# EMAIL SPAM
# CLASSIFICATION

Heramb Devarajan
2021A7PS0033U

# WHAT IS EMAIL SPAM?


**Email spam** refers to unsolicited, irrelevant, or inappropriate messages sent over email. These messages are typically sent in bulk to a large number of recipients, often with commercial or malicious intent. The primary purpose of email spam is to promote products, services, or fraudulent activities, posing a nuisance and security risk to users.

# WHY IS EMAIL SPAM CLASSIFICATION REQUIRED?

1) **Phishing and Fraud:** Spam emails serve as a vehicle for cybercriminals to execute phishing attacks, deceiving recipients into divulging sensitive information.

2) **Malware Distribution:** Malicious attachments or links in spam emails are used to install malware on recipients' devices, enabling unauthorized access or data theft.

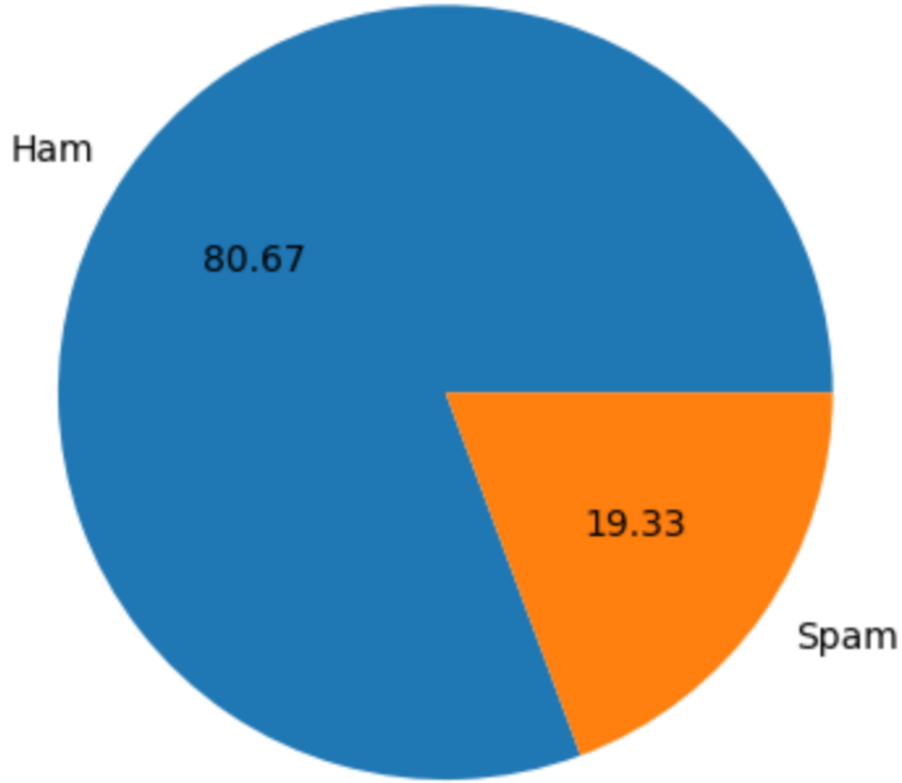
3) **Monetary Gain through Scams:** Some spam campaigns aim to trick individuals with scams, seeking financial or asset gains through deceptive schemes.

4) **Advertisement and Marketing:** Unethical entities resort to spam emails for mass advertising, promoting products, services, or offers without recipients' consent.

5) **Distributing Unwanted Content:** Spam emails may be utilized to disseminate legal but unwanted content, such as adult material, gambling promotions, or fake prescription drug advertisements, to a broad audience.

# DATASET DESCRIPTION

| | Body | Label |
|---|---|---|
| 6850 | Hmm .. Bits and pieces lol ... *sighs* ... | 0 |
| 563 | HTML>\nÂ Â Â Â Â Â Â Â Â Â Â Â Â \nÄ°NTERNE... | 1 |
| 6214 | Hmmm.. Thk sure got time to hop ard... Ya, can... | 0 |
| 3067 | Date: Wed, 11 Sep 2002 20:15:00 -07... | 0 |
| 7422 | Congratulations ur awarded either å£500 of CD ... | 1 |
| 10059 | G.W.R | 0 |
| 7565 | Ur cash-balance is currently 500 pounds - to m... | 1 |
| 9539 | Oh, i will get paid. The most outstanding one ... | 0 |



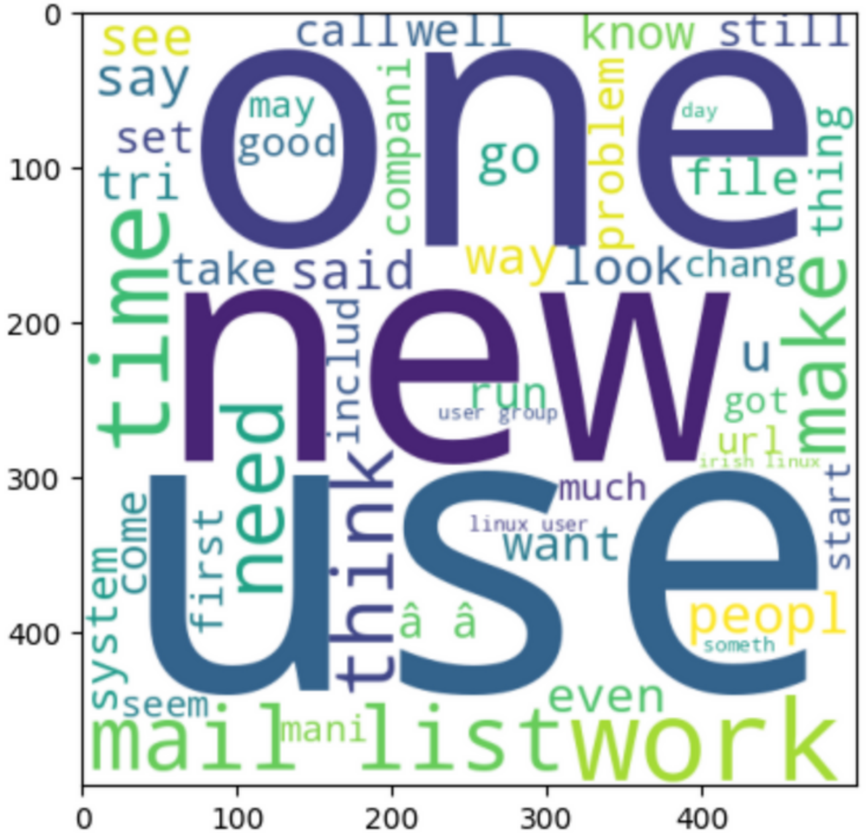

Ham emails: 8431

Spam emails: 2020

The dataset consists of 5572 emails from Spam emails UCL dataset and 6045 emails from SpamAssasin dataset merged into single dataset.

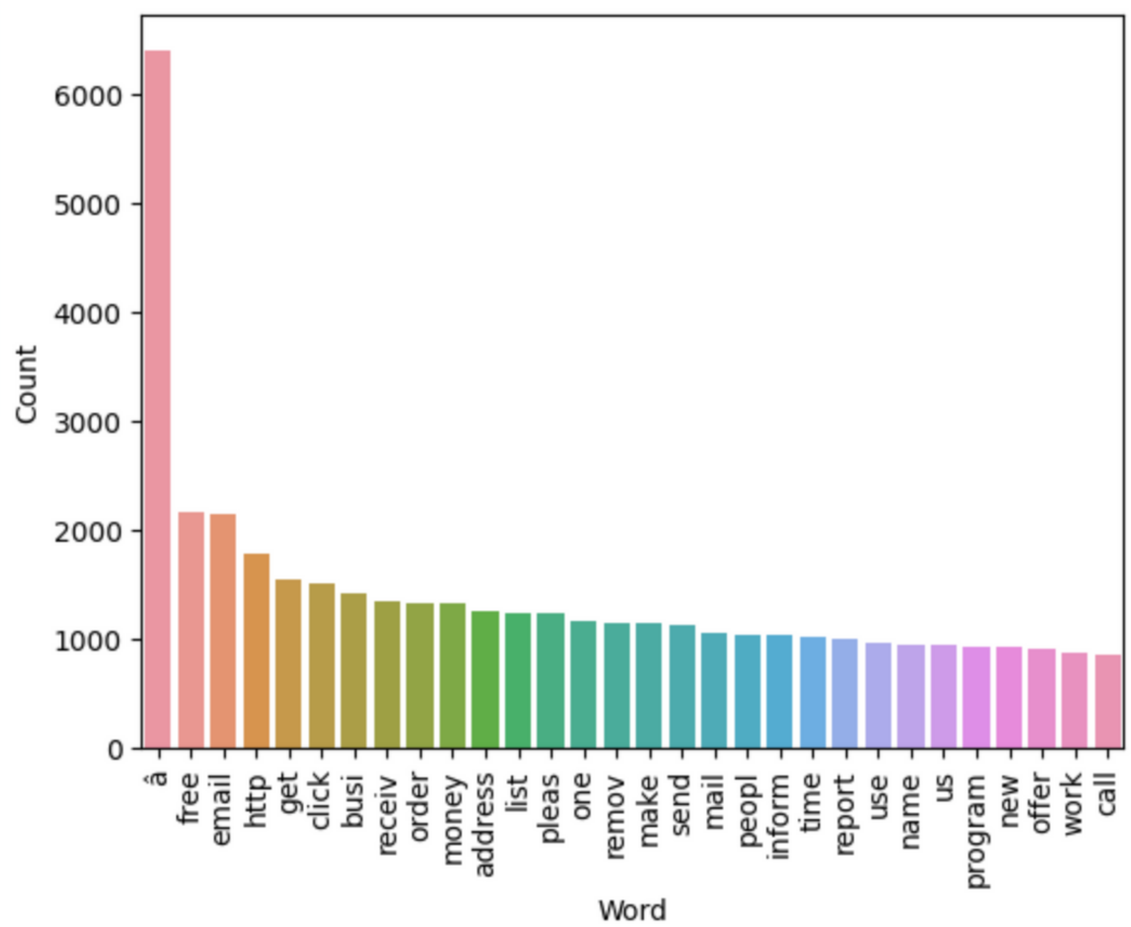
# EDA ON DATASET


WORD CLOUD OF THE SPAM EMAILS

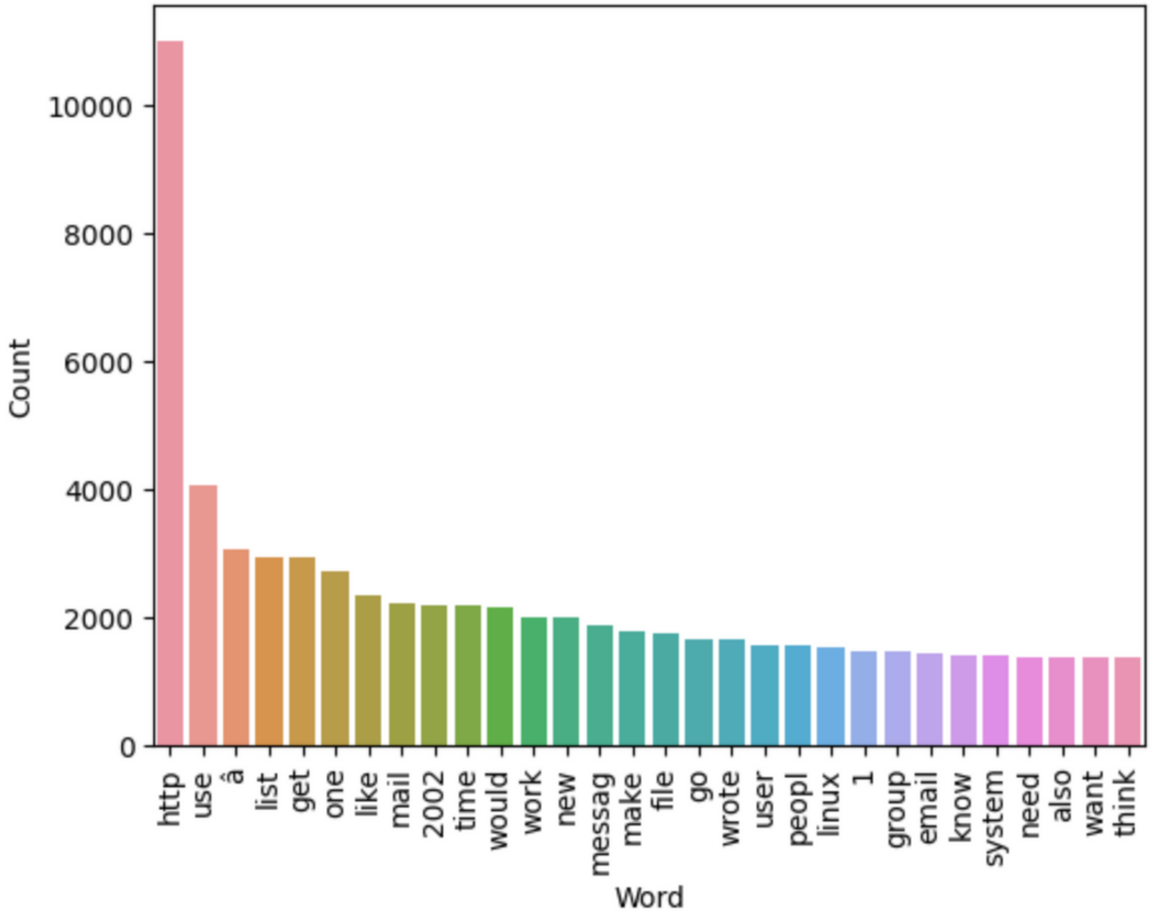



WORD CLOUD OF THE HAM EMAILS

# EDA ON DATASET




MOST COMMON WORDS IN SPAM EMAILS




MOST COMMON WORDS IN HAM EMAILS

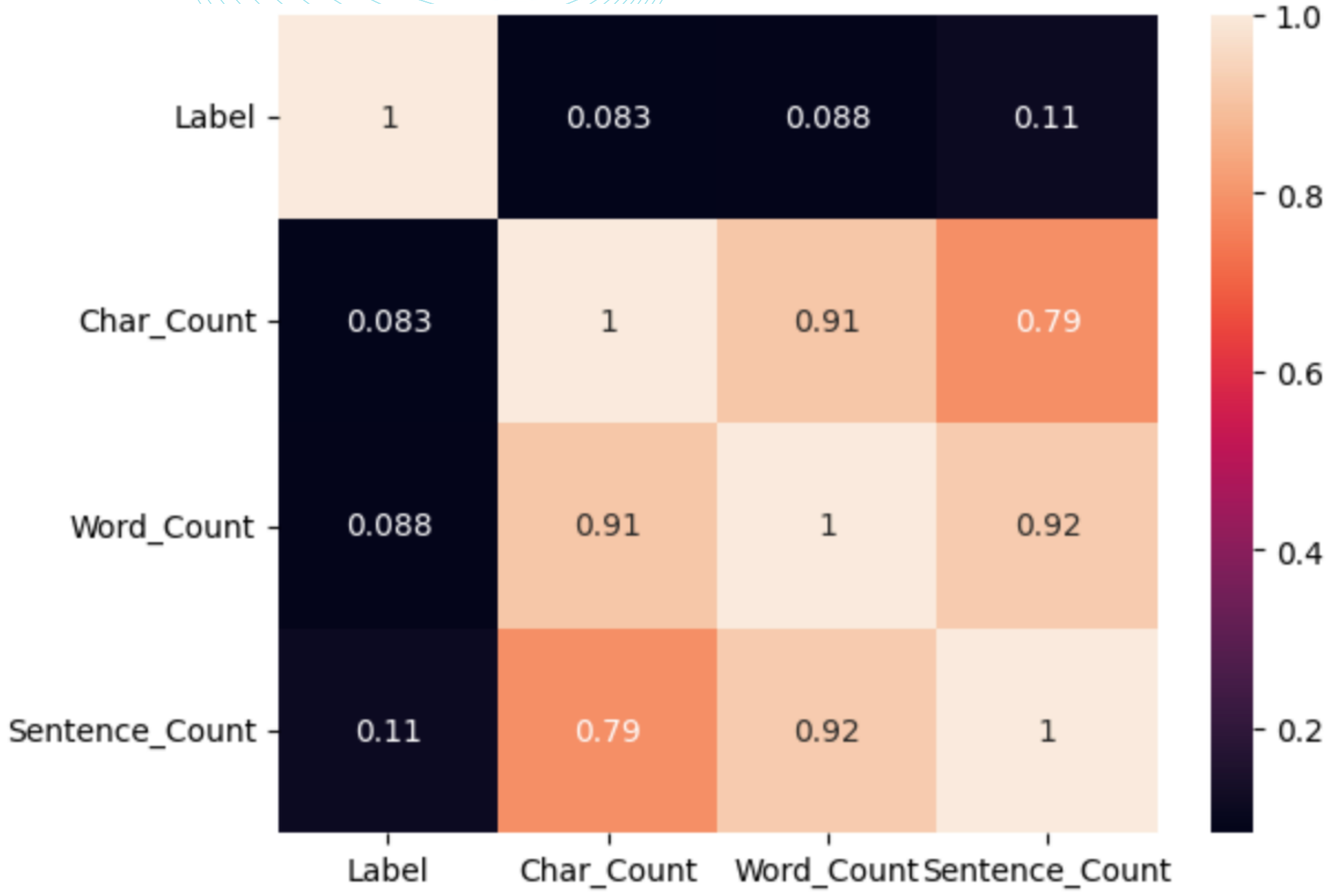
# EDA ON DATASET

| | Char_Count | Word_Count | Sentence_Count |
|---|---|---|---|
| count | 2020.000000 | 2020.000000 | 2020.000000 |
| mean | 1672.962376 | 301.119307 | 12.305446 |
| std | 4859.606811 | 747.006606 | 27.762316 |
| min | 1.000000 | 0.000000 | 0.000000 |
| 25% | 156.000000 | 32.000000 | 3.000000 |
| 50% | 705.000000 | 123.500000 | 6.000000 |
| 75% | 1507.500000 | 268.000000 | 11.000000 |
| max | 129635.000000 | 13288.000000 | 406.000000 |

TEXT ANALYSIS OF SPAM EMAILS

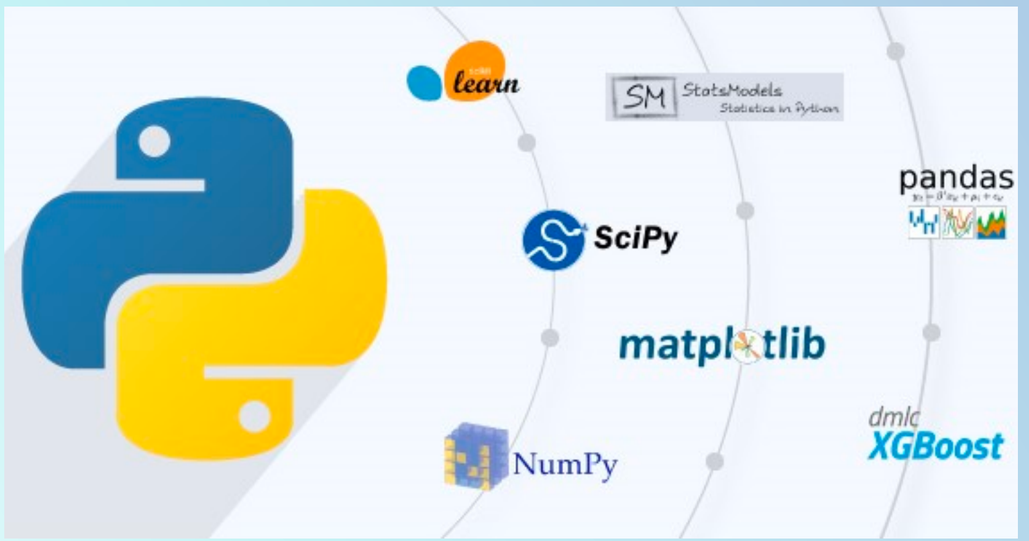
| | Char_Count | Word_Count | Sentence_Count |
|---|---|---|---|
| count | 8431.000000 | 8431.000000 | 8431.000000 |
| mean | 870.170205 | 167.282766 | 6.294983 |
| std | 3530.808360 | 552.987072 | 18.419597 |
| min | 2.000000 | 1.000000 | 1.000000 |
| 25% | 49.000000 | 12.000000 | 1.000000 |
| 50% | 148.000000 | 31.000000 | 2.000000 |
| 75% | 877.000000 | 173.500000 | 6.000000 |
| max | 194978.000000 | 18564.000000 | 808.000000 |

TEXT ANALYSIS OF HAM EMAILS




CORRELATION MATRIX

# LIBRARIES USED:




1) **Pandas:** Pandas is one of the tools in Machine Learning which is used for data cleaning and analysis. It has features which are used for exploring, cleaning, transforming and visualizing from data.

2) **NLTK:** NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning.

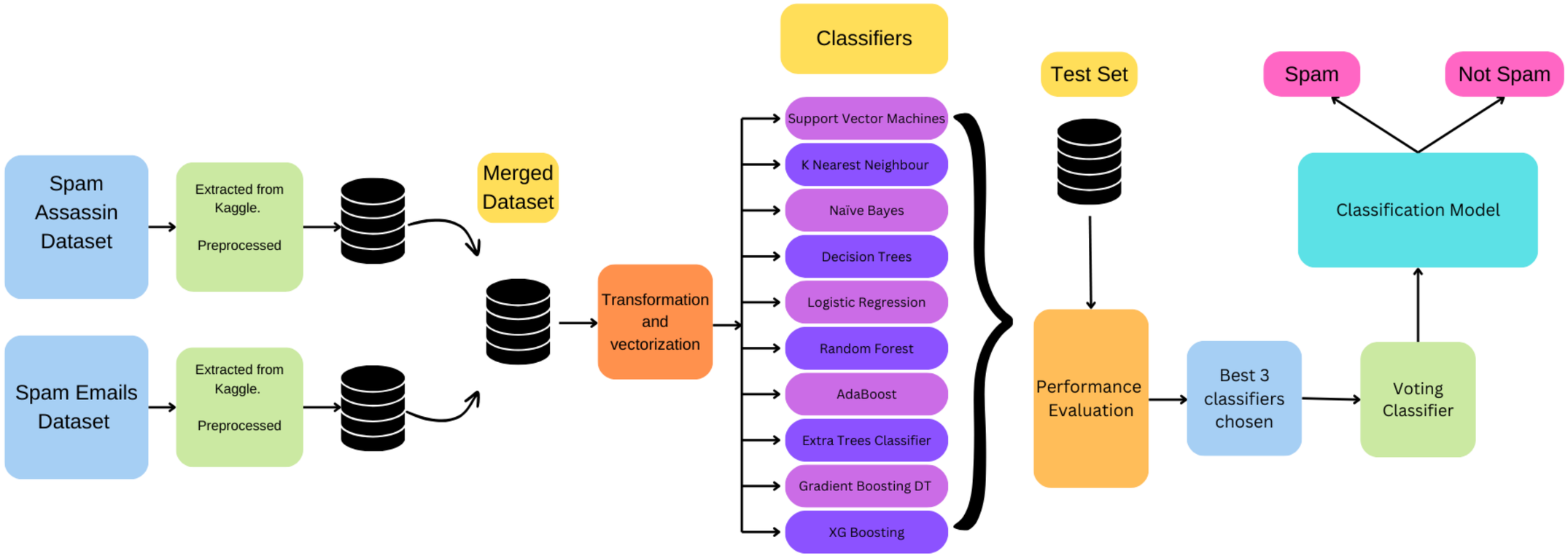
3) **NumPy:** NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. Moreover, NumPy forms the foundation of the Machine Learning stack.

4) **Matplotlib:** Matplotlib is a low-level library of Python which is used for data visualization. It is easy to use and emulates MATLAB like graphs and visualization. This library is built on the top of NumPy arrays and consist of several plots like line chart, bar chart, histogram, etc.

5) **Scikit-learn:** Scikit-learn is a comprehensive machine learning library for Python, offering a user-friendly interface and a rich set of tools for various tasks such as classification, regression, and clustering.

# METHODOLOGY

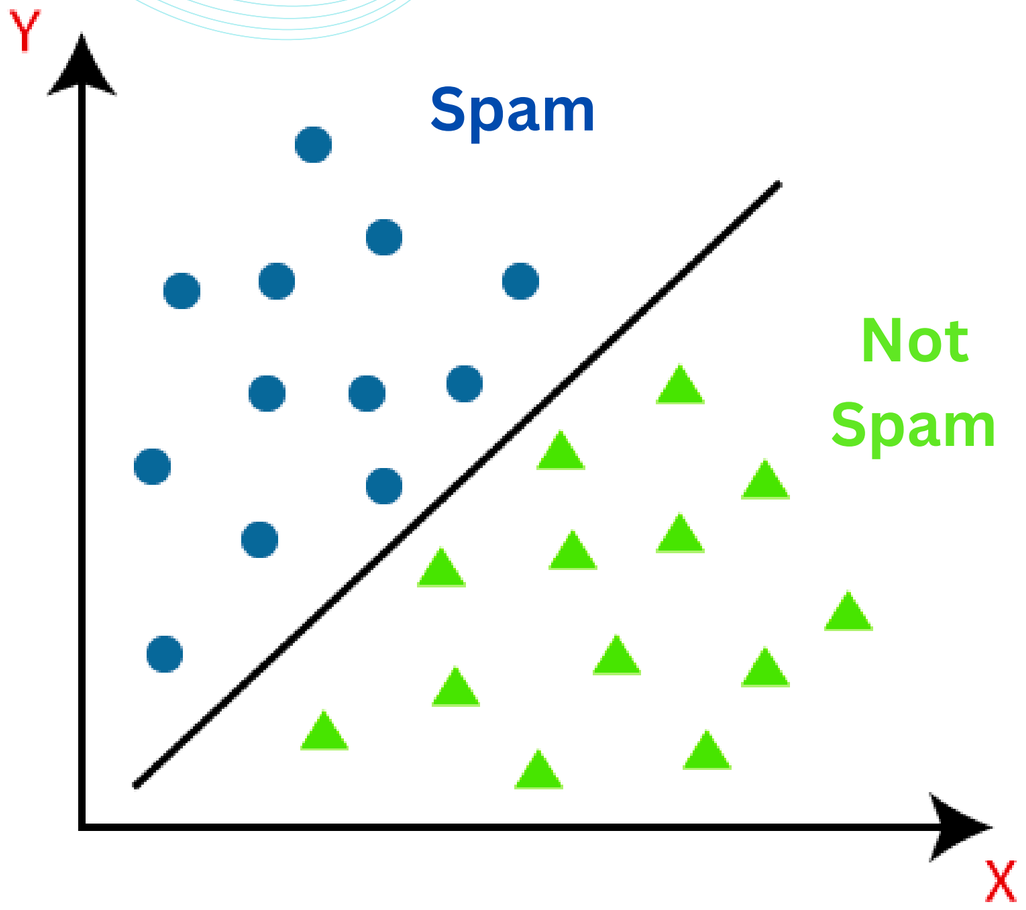
## Architecture Diagram:



# CLASSIFICATION ALGORITHMS

**A variety of algorithms were used to train the classification model:**

1. Naive Bayes
2. Extra Trees Classification
3. Random Forests
4. K - Nearest Neighbors
5. Support Vector Machines
6. Gradient Boosting Decision Tree
7. XGBoost Classifier
8. Logistic Regression Classifier
9. AdaBoost Classifier
10. Decision Tree Classifier



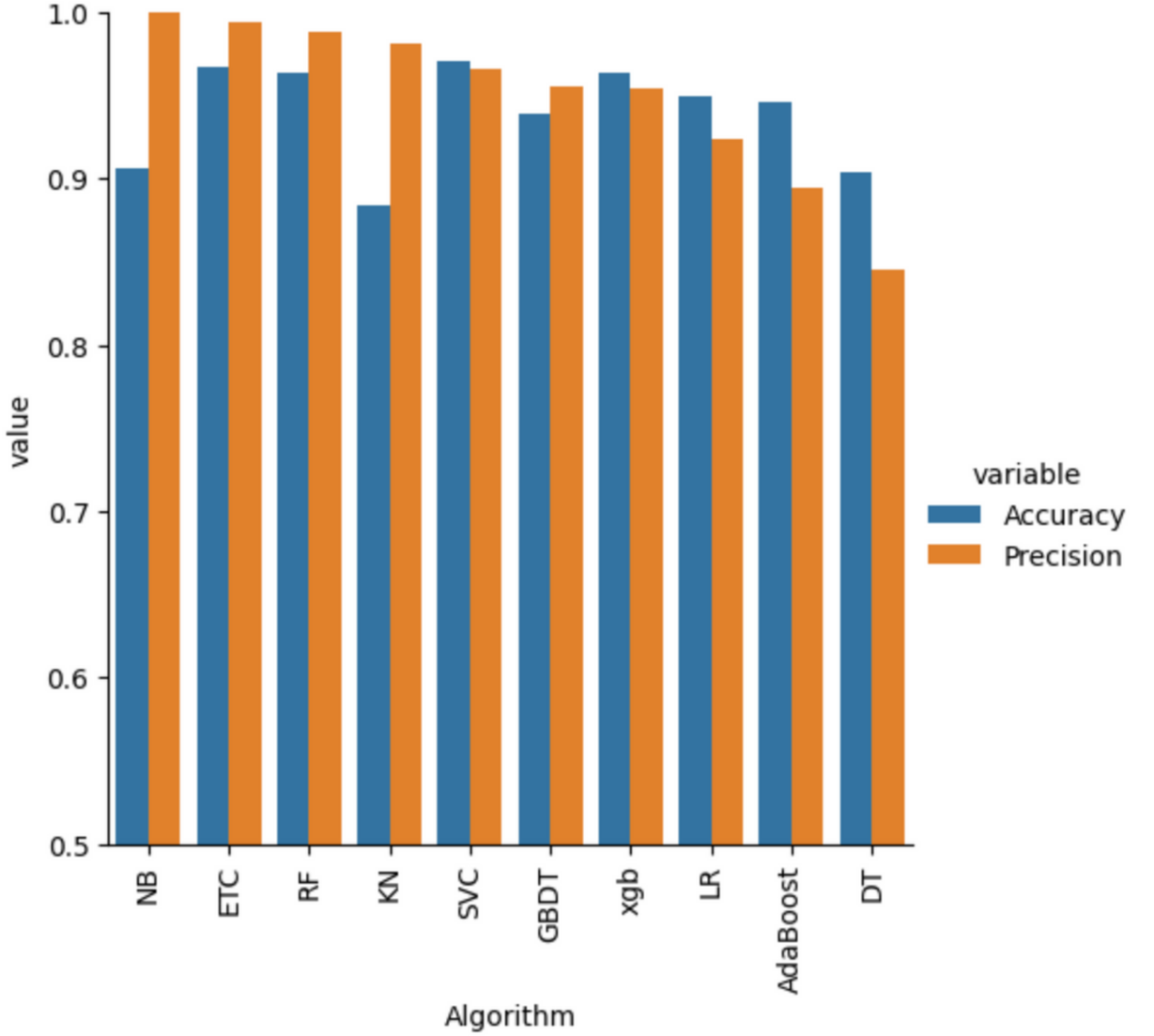
# ALGORITHMS PERFORMANCE

| | Algorithm | Accuracy | Precision |
|---|---|---|---|
| 2 | NB | 0.906743 | 1.000000 |
| 7 | ETC | 0.967480 | 0.993976 |
| 5 | RF | 0.963654 | 0.987805 |
| 1 | KN | 0.883788 | 0.981132 |
| 0 | SVC | 0.970827 | 0.966574 |
| 8 | GBDT | 0.938785 | 0.955782 |
| 9 | xgb | 0.963654 | 0.954545 |
| 4 | LR | 0.949785 | 0.924198 |
| 6 | AdaBoost | 0.946437 | 0.894444 |
| 3 | DT | 0.904352 | 0.845070 |

| | Algorithm | Accuracy | Precision |
|---|---|---|---|
| 0 | SVC | 0.970827 | 0.966574 |
| 7 | ETC | 0.967480 | 0.993976 |
| 5 | RF | 0.963654 | 0.987805 |
| 9 | xgb | 0.963654 | 0.954545 |
| 4 | LR | 0.949785 | 0.924198 |
| 6 | AdaBoost | 0.946437 | 0.894444 |
| 8 | GBDT | 0.938785 | 0.955782 |
| 2 | NB | 0.906743 | 1.000000 |
| 3 | DT | 0.904352 | 0.845070 |
| 1 | KN | 0.883788 | 0.981132 |

**The evaluation of the algorithms in descending order of precision and accuracy respectively.**

**The graph visualises the evaluation of the algorithms**



# BEST 3 PERFORMING ALGORITHMS

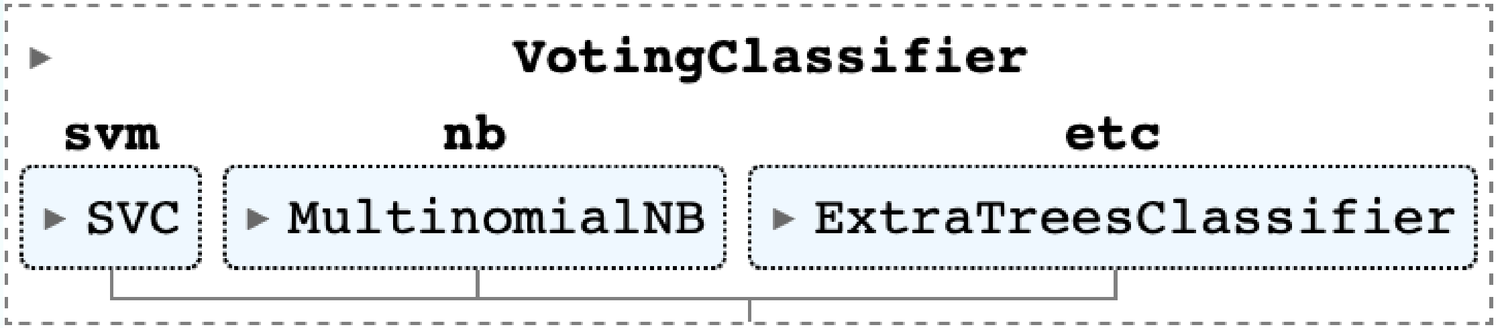
- **Support Vector Classifier(SVM):** Support Vector Classifier is a supervised machine learning algorithm that separates classes in a dataset by finding the hyperplane that maximizes the margin between them, making it effective for both linear and non-linear classification tasks.

- **Extra Trees Classifier (ETC):** Extra Trees Classifier is an ensemble learning method that builds multiple decision trees and combines their predictions, introducing additional randomness in the tree-building process to enhance robustness and reduce overfitting.

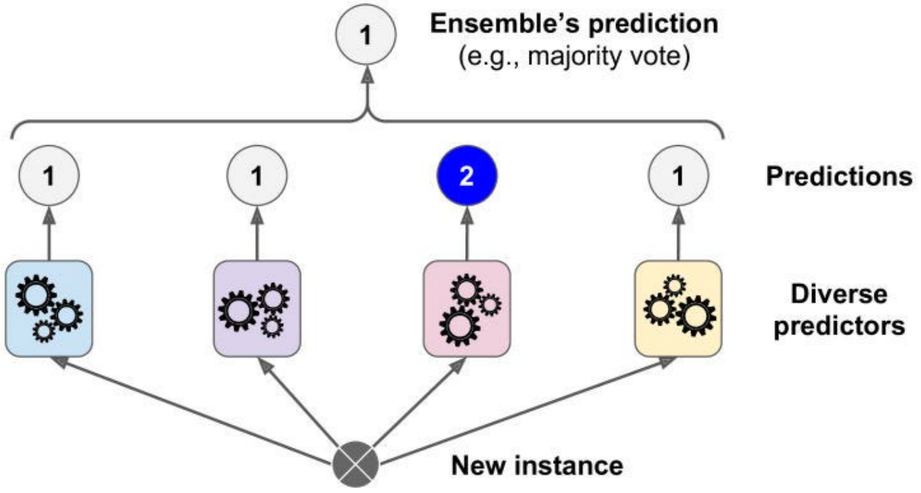
- **MNB (Multinomial Naive Bayes):** Multinomial Naive Bayes is a probabilistic classification algorithm suitable for text classification tasks. It's based on Bayes' theorem and assumes that features are conditionally independent given the class, making it efficient for high-dimensional datasets like word frequencies in documents.

# VOTING CLASSIFIER

- **Ensemble Approach:** A voting classifier is an ensemble learning technique that combines the predictions of multiple individual classifiers to make a final prediction. It leverages the diversity of different models to improve overall performance.



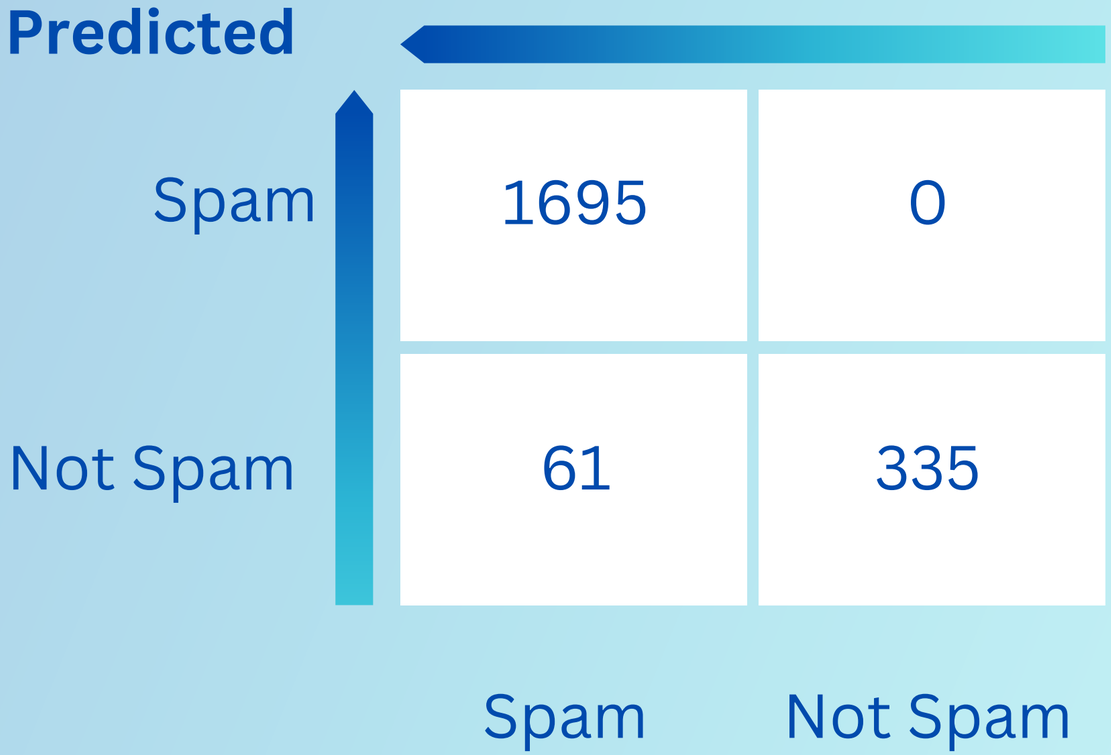

- **Voting Strategies:** The voting classifier can operate in two main modes: hard voting and soft voting. In hard voting, the majority vote from all classifiers determines the final prediction. In soft voting, the weighted average probability from all classifiers contributes to the final decision, providing a more nuanced approach.




- **Model Diversity:** The strength of a voting classifier lies in the diversity of its constituent models. By combining classifiers with different algorithms, architectures, or training datasets, the ensemble can better generalize to various patterns in the data, leading to improved overall performance and robustness.

# VOTING CLASSIFIER PERFORMANCE

**Confusion Matrix of model**



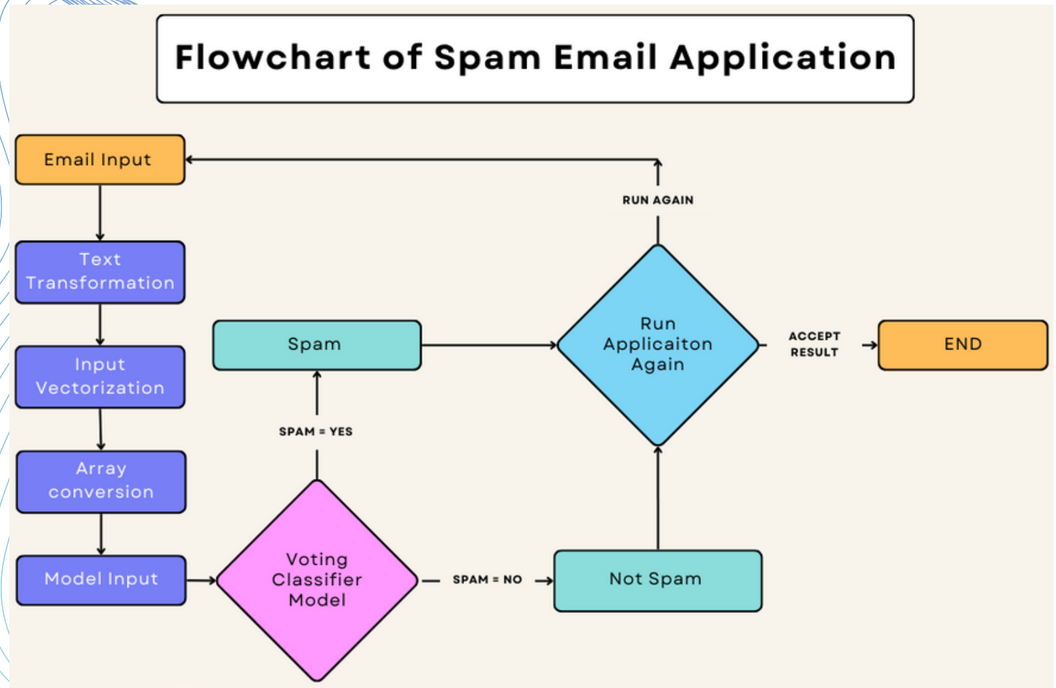

**Accuracy:** 97.08%

**Precision:** 1.00

This accuracy and prediction evaluation outperforms the previous algorithms and gives a better model. So, we will use this model for our application.

# APPLICATION GUI

Streamlit

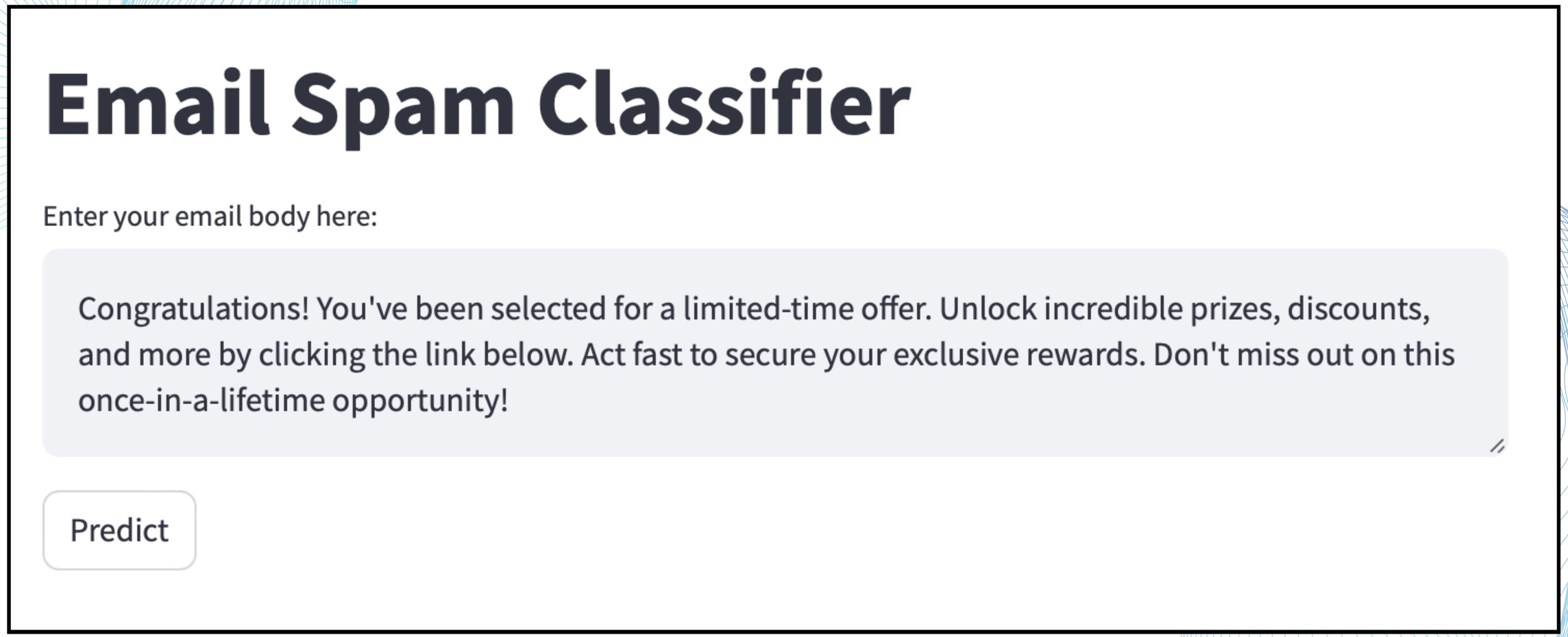



- This Email Spam Classification GUI built on Streamlit provides an intuitive interface for users to interact with and assess email content.
- Users can input email body, and the Streamlit app employs the machine learning algorithm to classify them as spam or ham.
- The GUI enhances accessibility and user experience in navigating and understanding the results of the spam classification model.

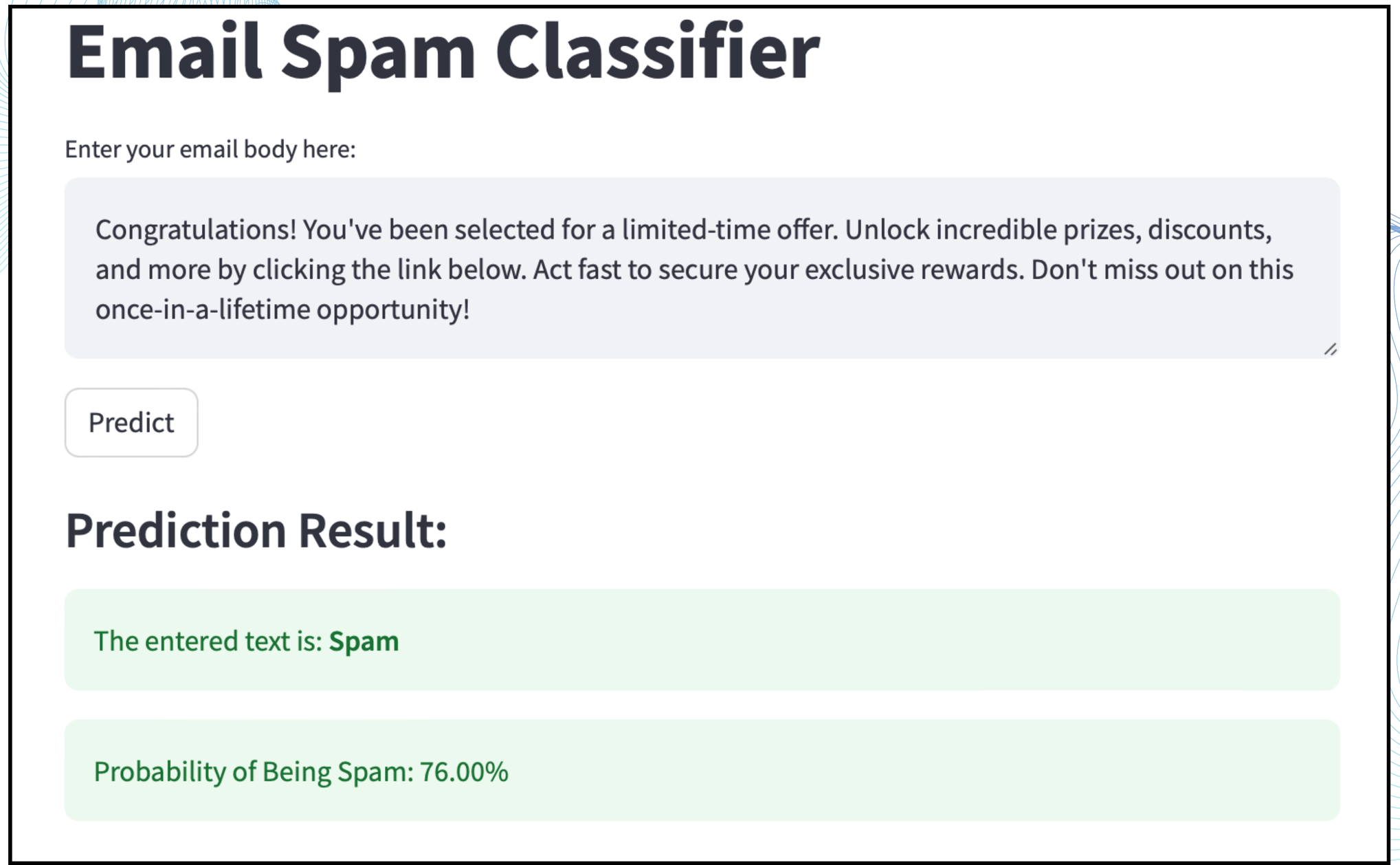
Click the link below to classify your own emails as spam and ham using our application:

https://email-spam-classification-ml-project.streamlit.app

# TEST CASE FOR SPAM EMAIL



# TEST CASE FOR SPAM EMAIL



# TEST CASE FOR HAM EMAIL

## Email Spam Classifier

Enter your email body here:

Hello, this is Heramb and I wanted to inform you that I have finished my machine learning project.
You can find the projects details below.

Press ⌘+Enter to apply

Predict

# TEST CASE FOR HAM EMAIL

## Email Spam Classifier

Enter your email body here:

Hello, this is Heramb and I wanted to inform you that I have finished my machine learning project.
You can find the projects details below.

Predict

## Prediction Result:

The entered text is: **Not Spam**

Probability of Being Spam: 0.00%

# Thank You