Email Spam Classification Report

Introduction:

This report presents an analysis of a spam email classification model using logistic regression. The model is trained on a dataset containing labeled emails as spam or ham (non-spam). We'll discuss the dataset, model training, evaluation metrics, and demonstrate how to classify new emails using the trained model.

Dataset:

The dataset used for this analysis is stored in a CSV file named mail_data.csv. It contains two columns: "Message" (email content) and "Category" (spam or ham).

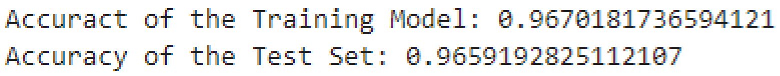
Data Preprocessing:

We have load the dataset using pandas and handling missing values and converted the 'Category' column to binary labels: 1 for ham and 0 for spam.

Model Training:

We have split the dataset into training and testing sets using an 80-20 split ratio where text data is transformed into numerical features using TF-IDF vectorization. Furthermore, Logistic Regression model is trained using the transformed features where training and testing accuracies are computed.

Evaluation:



```
Accuract of the Training Model: 0.9670181736594121
Accuracy of the Test Set: 0.9659192825112107
```

1. **Training Accuracy**: The accuracy of the model on the training set is approximately 96.70%. This means that the model correctly classifies around **96.70%** of the emails in the training dataset as either spam or ham.
2. **Testing Accuracy**: The accuracy of the model on the test set is approximately 96.59%. This indicates that the model's performance is consistent when applied to unseen data, correctly classifying approximately **96.59%** of the emails in the test dataset.

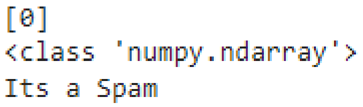
- **Model Performance:**

The training and testing accuracies obtained in this model are very close, with the testing accuracy being slightly lower than the training accuracy. This suggests that the model generalizes well to unseen data, as the performance on the test set is comparable to that on the training set. The high accuracies achieved on both the training and test sets indicate that the logistic regression model, coupled with TF-IDF vectorization, effectively discriminates between spam and ham emails.

Overall, these results validate the effectiveness of the model in accurately classifying emails and suggest it potential utility in real-world applications for email spam detection.

- **Email Classification:**

We test the model with a sample email text: "Congratulations! You have been selected as the winner...” The email is vectorized using the same TF-IDF vectorizer where the model predicts whether the email is spam or ham. Therefore, the model predicted that the above mail is a spam (It’s a Spam).



```
[0]
<class 'numpy.ndarray'>
Its a Spam
```

- **Conclusion:**

The logistic regression model demonstrates promising performance in classifying spam and ham emails. With a testing accuracy of 96.59%, the model effectively distinguishes between spam and legitimate emails. Further optimization and fine-tuning of the model could potentially enhance its performance in real-world scenarios.