**Project Report: Email Spam Detection**

**Introduction**

This project aims to build a predictive system for email spam detection using a dataset obtained from Kaggle. The dataset contains various emails labeled as spam or not spam. The following steps outline the process taken to achieve an accurate and reliable spam detection model.

**Dataset**

* **Source**: [Kaggle Email Spam Dataset](https://www.kaggle.com/datasets/esraaaabdelrazek/mail-data)
* **Content**: Emails labeled as spam or not spam.

**Data Collection and Pre-processing:**

**Importing the Dataset**:



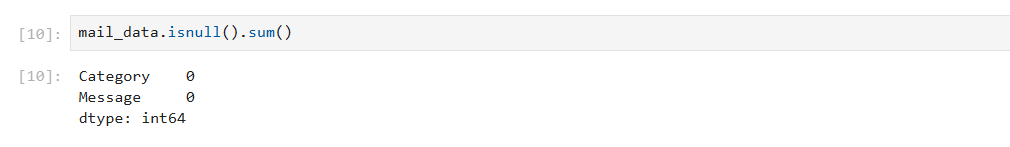
**Data Cleaning**:

* **Label Encoding**:
  + Converted categorical labels (spam/not spam) into numerical format.



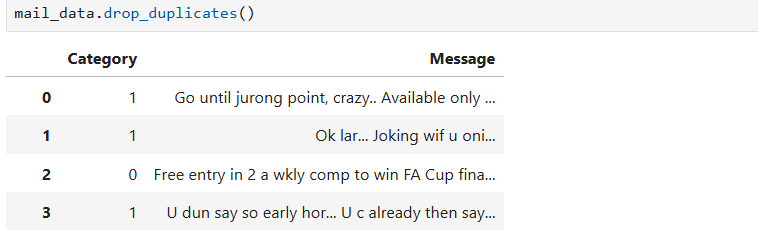
**Handling Missing Values**:

* + Checked for and handled any missing values in the dataset.



**Removing Duplicates:**

* + Identified and removed any duplicate records.

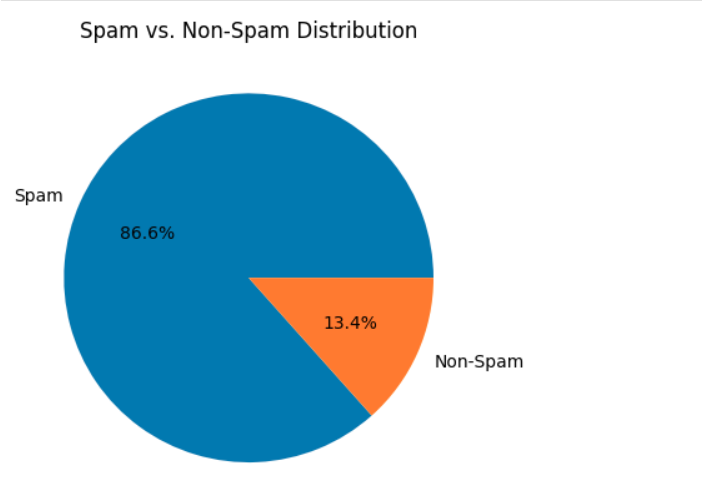


**Exploratory Data Analysis (EDA)**

**Pie Chart of Spam vs. Not Spam:**

* + Visualized the distribution of spam and not spam emails.





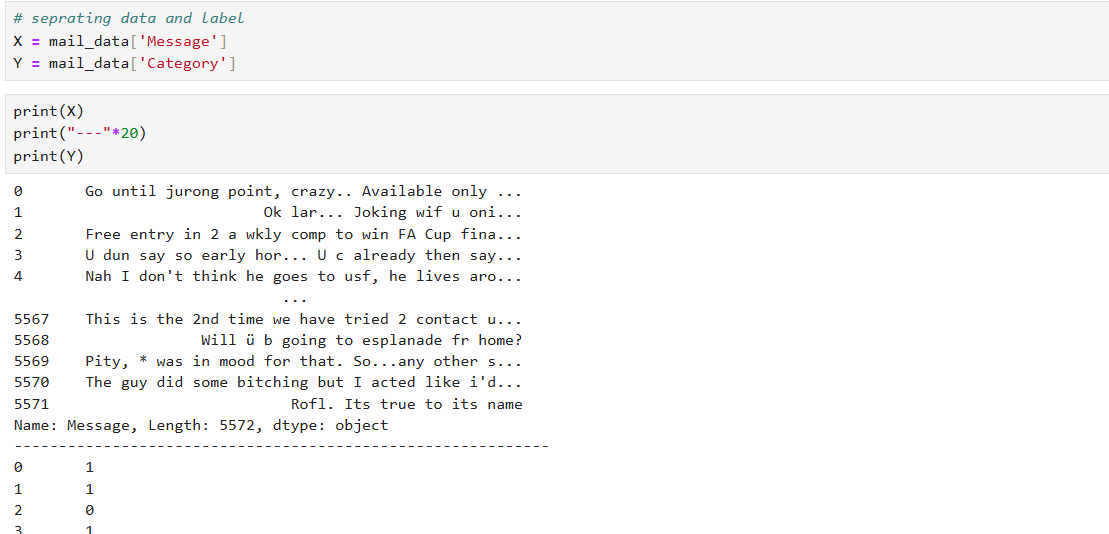
**Character, Word, and Sentence Count:**

* + Analysed the text data to understand its structure.



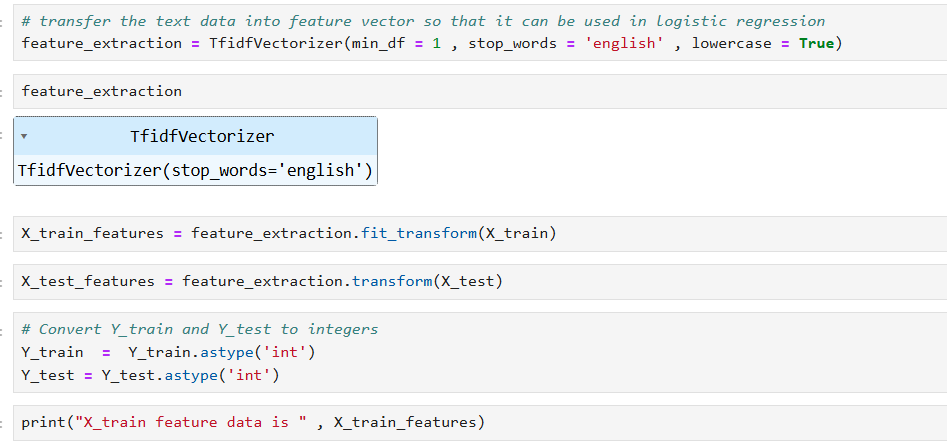
**Data Splitting**

* Split the dataset into training and testing sets.



**Feature Extraction**

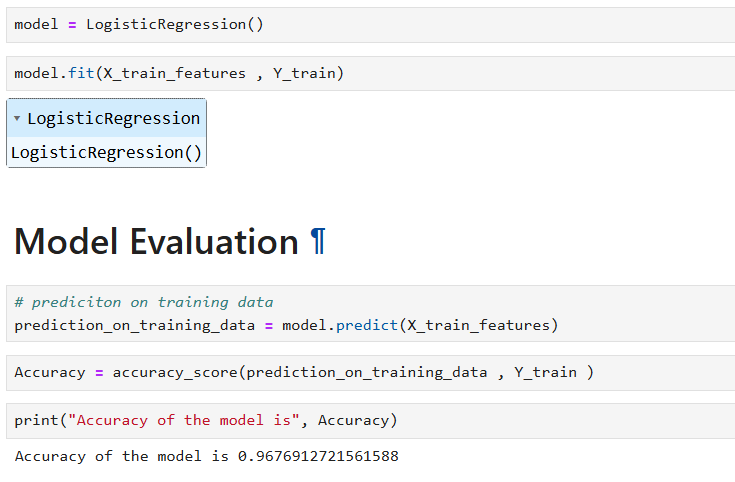
* Used TF-IDF (Term Frequency-Inverse Document Frequency) for feature extraction.

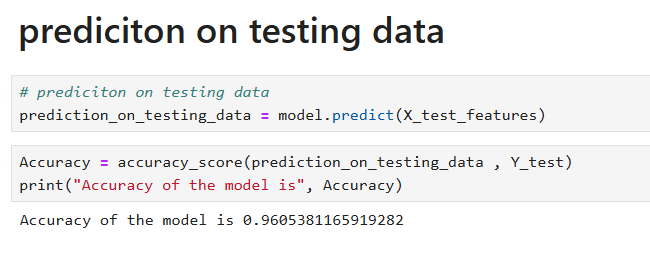


**Model Training and Evaluation**

**Logistic Regression**:

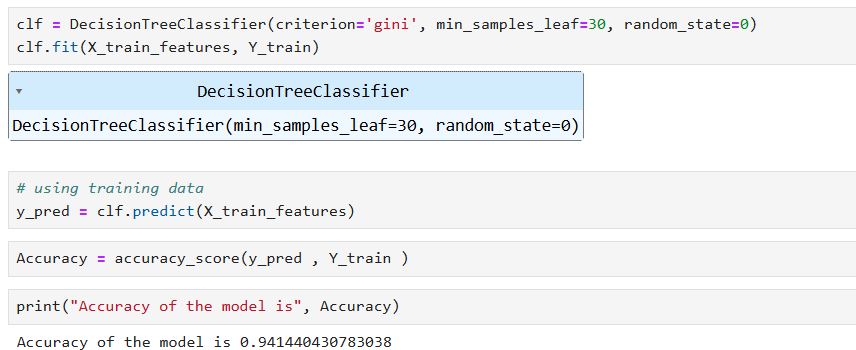
* + Trained a logistic regression model.

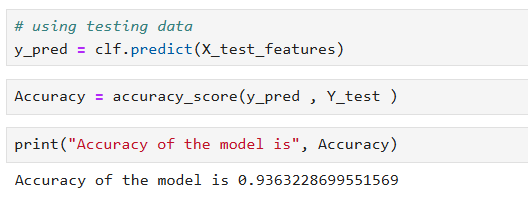




**Decision Tree Classifier**:

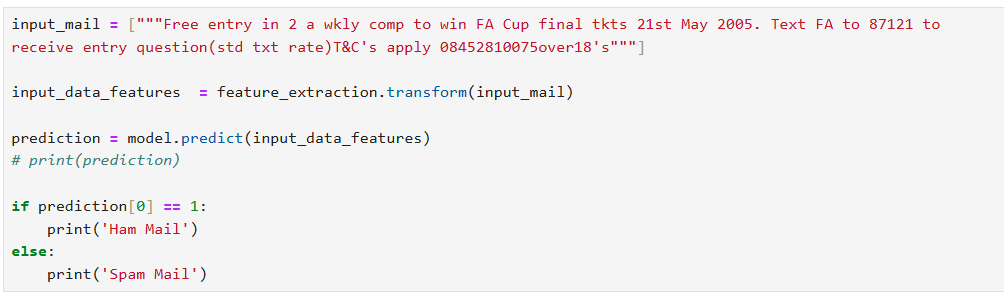
* Trained a decision tree model.





**Predictive System:**

* Built a simple predictive system using the logistic regression model.



**Conclusion:**

The logistic regression model achieved a high accuracy of 96%, making it a reliable choice for the spam detection system.

The project successfully implemented data cleaning, feature extraction, model training, and evaluation to build an effective email spam detection system. Further improvements can be explored by testing other algorithms and fine-tuning the existing models.