## Indian Institute of Information Technology Surat

****

# Lab Report on

# Machine Learning (CS 601) Practical

**Submitted by**

### [RAHUL KUMAR SINGH] (UI21CS44)

**Course Faculty**

### Dr. Pradeep Kumar Roy

### Dr. Rajesh K. Ahir

## Department of Computer Science and Engineering

## Indian Institute of Information Technology Surat

## Gujarat-394190, India

**Jan-2024**

## Lab No: 5

**Aim:**

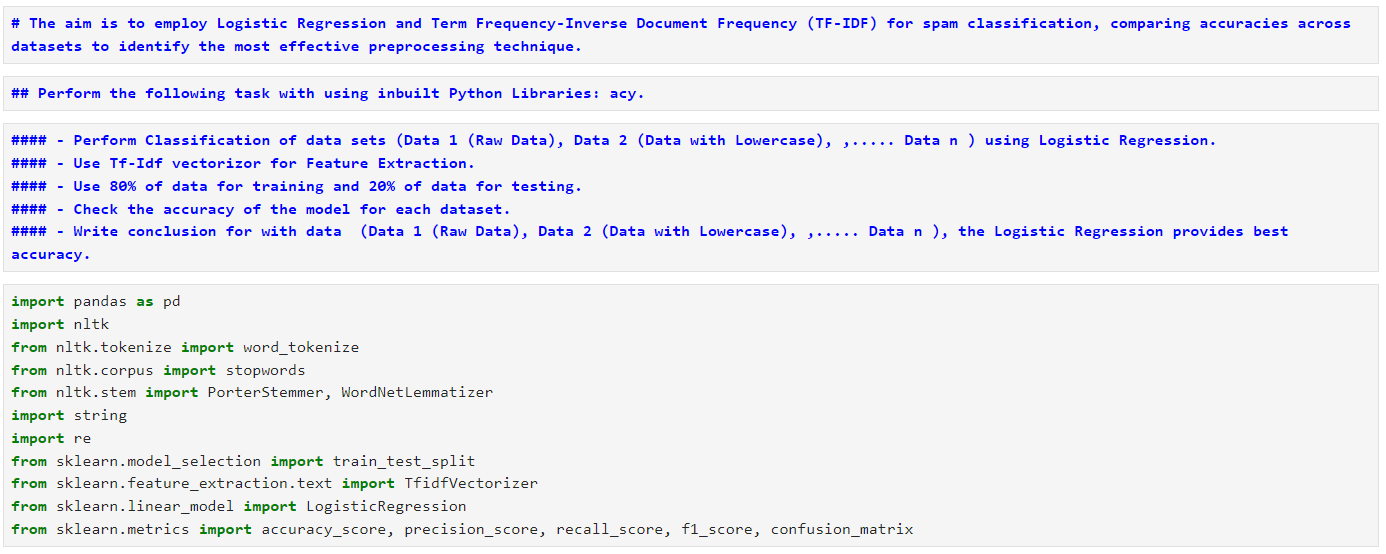
The aim is to employ Logistic Regression and Term Frequency-Inverse Document Frequency (TF-IDF) for spam classification, comparing accuracies across datasets to identify the most effective preprocessing technique.

**Description:**

Perform the following task with using inbuilt Python Libraries:

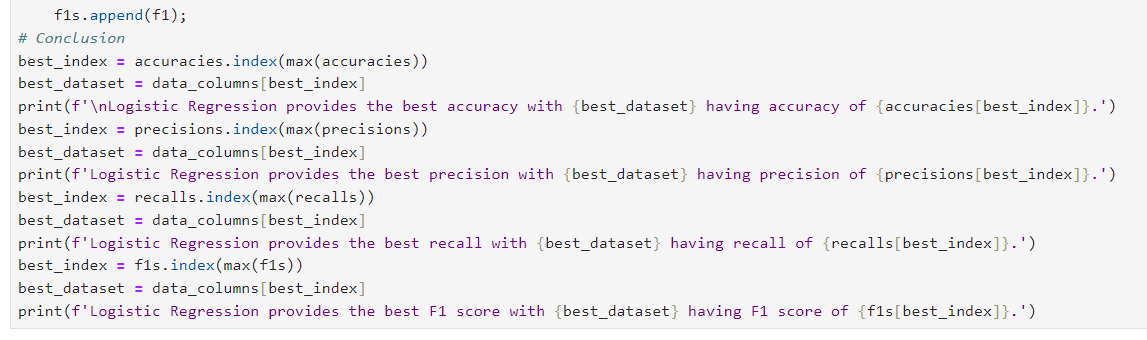
* Feature Extraction: Utilize TF-IDF vectorization to convert text data into numerical features.
* Data Splitting: Divide the datasets into 80% training and 20% testing subsets.
* Model Training: Train Logistic Regression models on the training data for each dataset.
* Prediction: Evaluate model performance by predicting labels on the testing sets.
* Accuracy Assessment: Calculate and compare accuracies to identify the most effective preprocessing technique among datasets.
* End Result: Determine which dataset, whether raw or preprocessed, yields the highest accuracy with Logistic Regression and TF-IDF.

## Source Code:

****

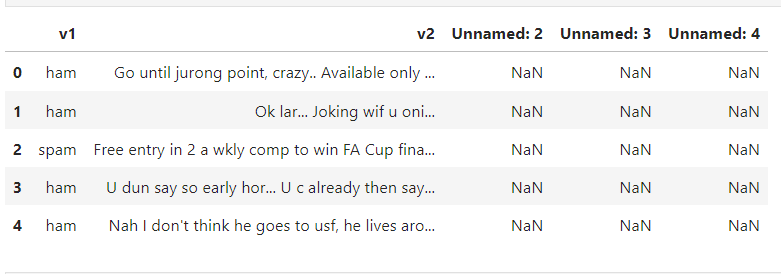
****

****

****

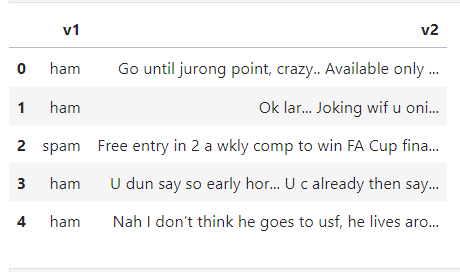
## Output:

**Spam Data:**



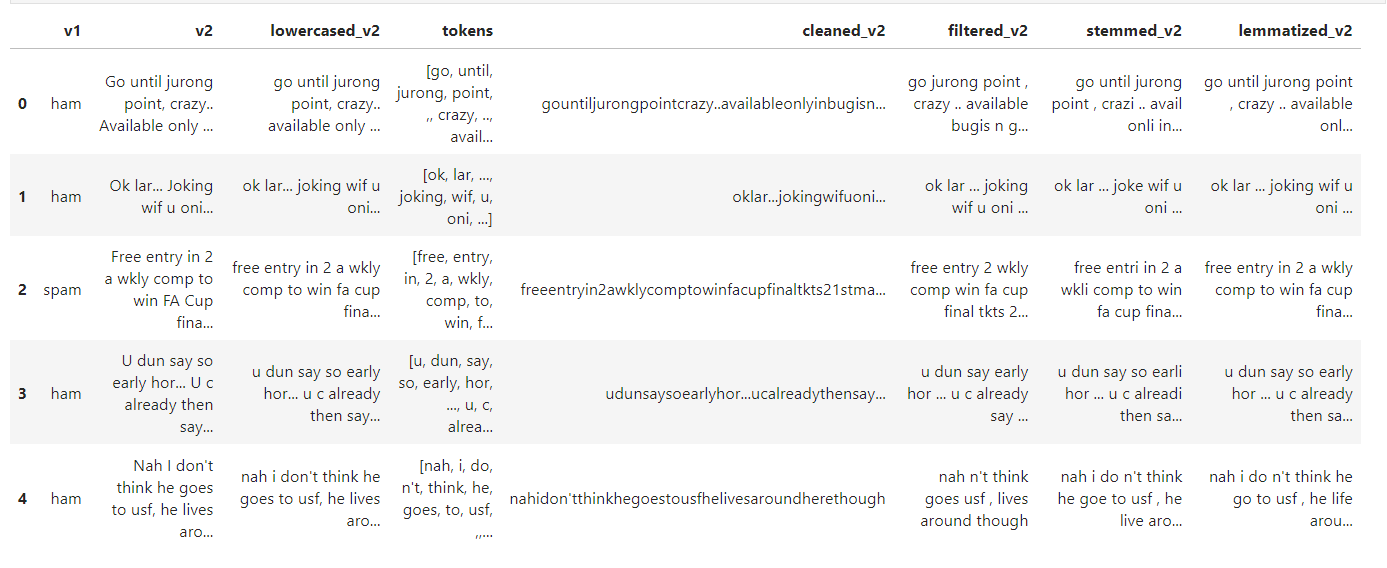


**After Dropping Unnecessary Columns:**

****

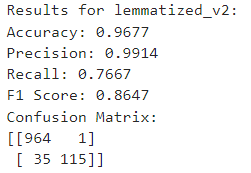
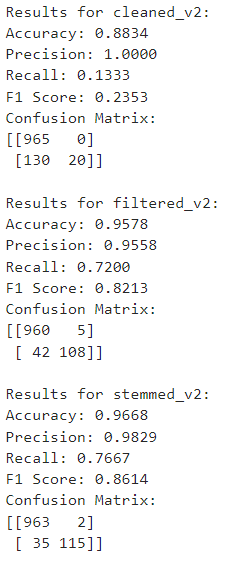
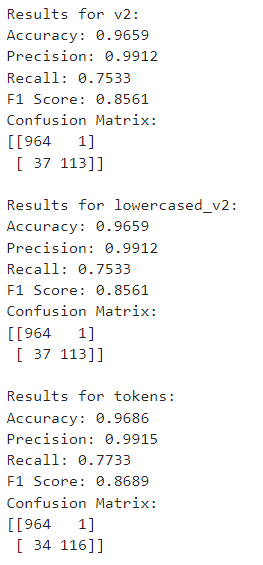


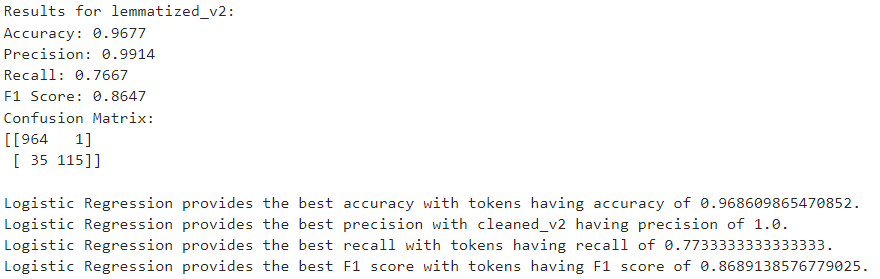
**After Applying Preprocessing Techniques:**

****



**Output:**

****

****



## Conclusion:

* Applied preprocessing steps including lowercasing, tokenization, cleaning, filtering, stemming, and lemmatization to enhance text data quality.
* Utilized Tf-Idf vectorization for feature extraction, capturing term importance in each dataset.
* Trained Logistic Regression models on each preprocessed dataset using 80% of data for training and 20% for testing.
* Evaluated model accuracy for each dataset, measuring performance on spam classification.
* Identified the dataset with the highest accuracy, indicating that Logistic Regression performs best on token generation approach.