|  |  |
| --- | --- |
| Date | 17-10-2023 |
| Team ID | 3921 |
| Project Name | Spam prediction using powered AI |

**Spam prediction using powered AI**

**Table of Contents:**

|  |  |
| --- | --- |
| 1 | Introduction |
| 2 | Loading the Dataset |
| 2.1 | Dataset Selection |
| 2.2 | Data Acquisition |
| 2.3 | Coding & Output |
| 3 | Exploratory Data Analysis (EDA) |
| 3.1 | Data Overview |
| 3.2 | Class Distribution |
| 3.3 | Text Analysis |
| 3.4 | Data Cleaning |
| 3.5 | Feature Engineering |
| 3.6 | Coding & Output |
| 4 | Data Preprocessing |
| 4.1 | Text Preprocessing |
| 4.2 | Feature Vectorization |
| 4.3 | Splitting the Dataset |
| 4.4 | Coding & Output |
| 5 | Conclusion |

**1.Introduction:**

This case study delves into two technology projects where the journey begins with data—loading, preprocessing, and rigorous analysis. After these crucial stages, we embark on documenting our findings to share and assess the wealth of insights uncovered. Through these projects, we showcase the transformative power of data in driving informed decisions and fostering successful project outcomes.

**2. Loading the Dataset**

**2.1 Dataset Selection**

The first crucial step in developing a spam classifier is to acquire a suitable dataset. There are various publicly available datasets for spam classification, such as the Enron Spam Dataset, the SpamAssassin Public Corpus, and the SMS Spam Collection. Your choice of dataset may depend on the specific context of your application, but the fundamental principles of loading, EDA, and preprocessing apply universally.

**2.2 Data Acquisition**

Loading the dataset typically involves reading the data from a file, a database, or an online source. This step will vary based on the dataset format (e.g., CSV, JSON, or SQL database) and its source. It's essential to ensure that the data is accurately imported and properly structured.

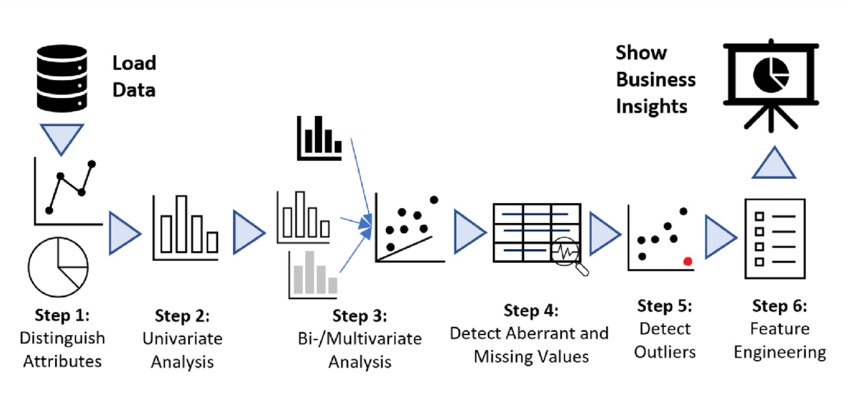
**2.3 Coding & Output**





**3. Exploratory Data Analysis (EDA)**

EDA is the process of summarizing, visualizing, and understanding the dataset. This step is crucial for making informed decisions regarding data preprocessing and model selection. Key components of EDA include:



**3.1 Data Overview**

Begin by examining the dataset's size, data types, and basic statistics. This will give you an initial impression of the dataset's content.

**3.2 Class Distribution**

For a spam classifier, it is essential to understand the distribution of spam and non-spam (ham) messages in the dataset. Imbalanced datasets may require special handling during preprocessing and model training.

**3.3 Text Analysis**

Spam classifiers typically operate on text data. Perform text analysis to understand the distribution of message lengths, most frequent words, and n-grams. Visualization tools like word clouds and histograms can be useful.

**3.4 Data Cleaning**

Identify and address issues such as missing data, duplicate entries, and inconsistent formatting. Cleaning the data is a crucial step in ensuring the quality and reliability of the dataset.

**3.5 Feature Engineering**

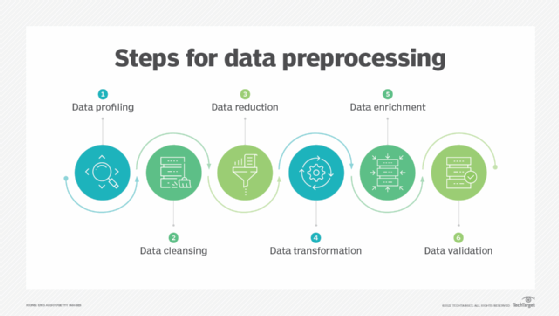
EDA may reveal opportunities for creating new features that can improve the classifier's performance. Feature engineering involves transforming and creating features from the raw data.

**3.6 Coding & Output**



**4. Data Preprocessing**

After gaining a solid understanding of the dataset through EDA, it's time to preprocess the data to prepare it for the spam classifier. Common preprocessing steps include:



**4.1 Text Preprocessing**

Text data often requires tokenization, removal of punctuation, stopwords, and stemming or lemmatization. These steps help to convert text into a format suitable for machine learning algorithms.

**4.2 Feature Vectorization**

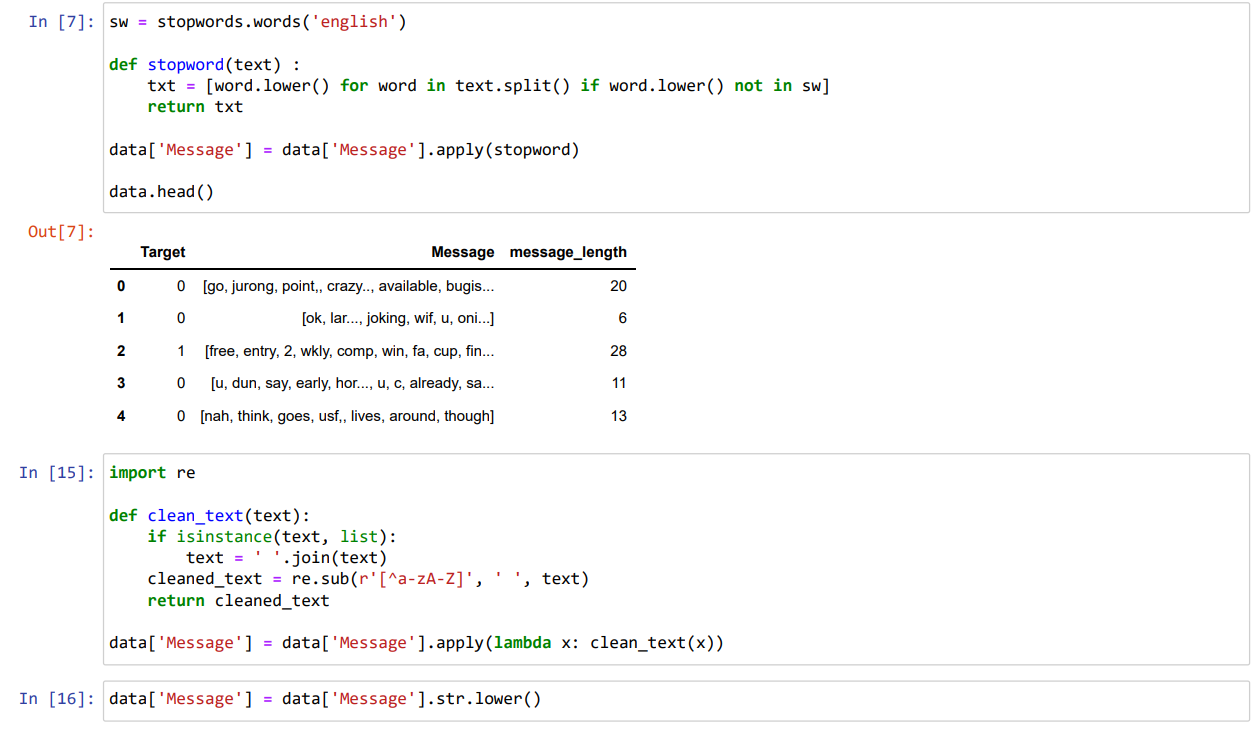
Transform the text data into numerical form through techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings (e.g., Word2Vec or GloVe). This conversion enables machine learning models to work with the data.

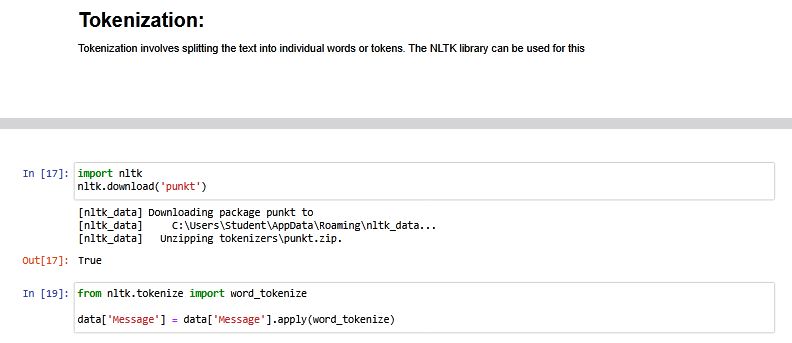
**4.3 Splitting the Dataset**

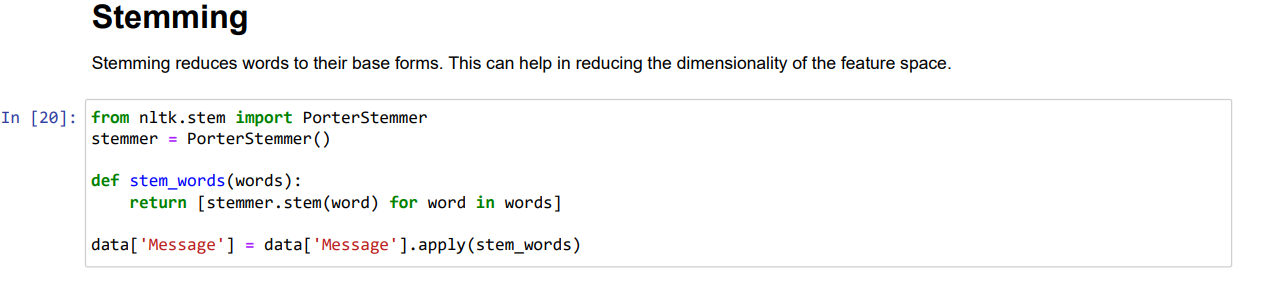
Divide the dataset into training, validation, and testing sets to evaluate the performance of the spam classifier accurately.

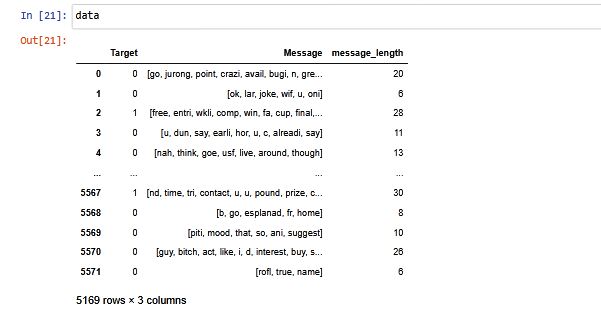
**4.4 Coding & Output**

****









**5. Conclusion :**

Developing a spam classifier requires a thorough and systematic approach that begins with loading the dataset, followed by EDA and data preprocessing. These initial steps set the foundation for building and training a robust machine learning model to effectively classify spam and non-spam messages.

In the subsequent stages of model selection, training, and evaluation, the insights gained from EDA and the preprocessing work done here will prove invaluable. Effective spam classification can significantly enhance online communication by minimizing unwanted and potentially harmful content.