**FRAUD EMAIL DETECTOR**

Nitya Nimmagadda

**INTRODUCTION:**  
  
Emails have gained popularity as the most cost-effective, versatile, and practical means of online communication. However, the issue of spam has arisen as some individuals send unsolicited emails, also known as email spam. Such emails may be commercial in nature or contain fraudulent links, malware, or viruses, posing significant risks to the recipients. To ensure effective communication, it is essential to prevent the receipt of such emails. Thus, a text-based spam classifier can help detect and filter out such unwanted messages.  
  
Dataset was taken from Kaggle: <https://www.kaggle.com/uciml/sms-spam-collection-dataset>

PROCESS INVOLVED:

* 1. Data cleaning
* 2. EDA
* 3. Text normalization
* 4. Model building

**DATA CLEANING:**

Count the number of non-null values in each column to determine the completeness of the data.

* Remove the last three columns of the dataset.
* Rename the remaining columns to enhance readability.
* Check for missing values in the dataset to ensure data quality.
* Check for duplicate values in the dataset to avoid redundancy.
* Eliminate the duplicate values found in the dataset to ensure accuracy.

**EDA**

**Chart, pie chart

Description automatically generated**

* Our dataset was unbalanced
* We had 12.63% spam messages, and 87.36% ham messages

Correlation HeatmapGraphical user interface

Description automatically generated

**TEXT CLEANING:**Text pre-processing involves cleaning and normalizing textual data.

* The process aims to prepare the text for analysis by removing unwanted characters, punctuation, and special symbols.
* Text pre-processing can also involve converting text to lowercase, removing stop words, stemming, and lemmatization.
* The ultimate goal of text pre-processing is to improve the accuracy and effectiveness of the analysis of textual data.

****

**BUILDING THE MODEL**

TRAIN/TEST

The data was split into a 20:80 ratio for training and testing, respectively. The selection of the 80% of the training data was done randomly and uniformly. It is recommended to maintain a similar distribution of data between the training and testing sets, particularly for imbalanced datasets like this one, where the percentage of spam emails should be similar in both sets. To achieve this, we compared the performance of unsampled, up-sampled, and stratified cross-validation techniques, and the results are presented in subsequent slides.

Classifiers Used:

* SVC
* KNeighbors
* Naive Bayes
* Decision Trees
* Logistic regression
* Random Forest
* Gradient-Boosted Trees
* XGBoost

My Multinomial NB has the best precision and accuracy.

Multinomial NB

Text

Description automatically generated