**### Installing Libraries**

Libraries used in project

!pip install xgboost

import pandas as pd

import numpy as np

import nltk

from sklearn.preprocessing import LabelEncoder

import seaborn as sns

from sklearn import metrics

from sklearn.metrics import confusion\_matrix, f1\_score, precision\_score, recall\_score, ConfusionMatrixDisplay, accuracy\_score

import matplotlib.pyplot as plt

from matplotlib.pyplot import figure

from nltk.corpus import stopwords

import string

from nltk.stem.porter import PorterStemmer

import re

from nltk.tokenize import word\_tokenize

from nltk.stem.wordnet import WordNetLemmatizer

from wordcloud import WordCloud

from collections import Counter

from sklearn.feature\_extraction.text import CountVectorizer,TfidfVectorizer

from sklearn.naive\_bayes import GaussianNB,MultinomialNB,BernoulliNB

from sklearn.metrics import accuracy\_score,confusion\_matrix,precision\_score

from sklearn.metrics import plot\_confusion\_matrix

from sklearn.linear\_model import LogisticRegression

from sklearn.svm import SVC

from sklearn.naive\_bayes import MultinomialNB

from sklearn.tree import DecisionTreeClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.ensemble import AdaBoostClassifier

from sklearn.ensemble import BaggingClassifier

from sklearn.ensemble import ExtraTreesClassifier

from sklearn.ensemble import GradientBoostingClassifier

from xgboost import XGBClassifier

from sklearn.ensemble import VotingClassifier

from sklearn.ensemble import StackingClassifier

from imblearn.over\_sampling import SMOTE

from gensim.matutils import Sparse2Corpus

from gensim.models import LdaModel

from gensim.models import TfidfModel, LsiModel

from gensim.models.ldamodel import LdaModel

from gensim import matutils

from sklearn.cluster import KMeans

from gensim import corpora

from collections import defaultdict

**# Usage**

The Spam\_detection\_ipynb files contains the entire model for spam detection and topic modeling.

Following are the multiple stages covered in the model: Exploratory Data analysis, Data Preprocessing, Model Building, Topic modeling.

This notebook reads the SMSSpamcollection file and perfoms the following operations on data :

1. EDA:

Exploratory data analysis is performed in this step to analyze the text data. In this following steps are performed -

1. Missing values check

2. Duplicate values check

3. Label Distribution

4. Total number of character in data is check and add it to df

5. Total number of word in data is check and add it to df

6. Total number of sentence in data is check and add it to df

7. Realation between features

2. Data Preprocessing:

#def data\_preprocessing(sms):

fuction created for applying text data preprocessing; Convert into lower case, Stop word removal, Special Charater removal, Number removal, Puntuation removal and Stemming using poterstemmer.

#"df['preprocessed\_text']=df['sms'].apply(lambda x: data\_preprocessing(x))" is used to tranform sms data.

3. Model Building:

This is used to train all classification models under different settings.

#clfs={} is a is a dictionary which contains all the classification model used in model buidling.

#def train\_classifer() is created to train the clfs ={} for predictions

4. Topic Modeling:

\*Note: Topic modeling has been added after professors' suggestion to get more insight form sms data.

Topic modeling is used to extract all relavent topic from the preprocessed data to get more insight form the SMS message data.

**# Running the code**

The entire nootebook file can be run in jupiternotebook. Run the each cell until Model building.

1. Model building needs be run 3 time each time with different setting.

1. In first attempt, use count vectorizer and run the model.

2. In second attempt, use TF-IDF vectorizer and run the model.

3. In third attemot, use MinMax scaler and run the model.

Follow instruction in each cell while model building.

2.Run the Oversampled model once and no change is required for model training.

3.Run the Topic modeling from once without any change.

4.The libraries used in the project are python3.9 compliant and can be installed using pip.

Installation code: pip install library-name-version

*\*Note - Running all the files may take time concerning the resources of the machine\**