SPAM MAIL CHECKER

By Vineet Uthaiah-225890410

Section AI-C

ABOUT THE PROJECT

A program that can sort through a database of emails and mark them as potential spam. It uses Natural Language Processing to analyse the text in the email to determine whether an email is spam or ham

RESOURCES USED

- **Python: the main language to implement the program**
- **NLTK library: a library used for NLP**
- **pandas: a data library in python used for data analysis and manipulation**
- **Dataset of emails: includes the emails used for checking. The data used is in the .csv format**

WORKING OF THE MODEL

- First, we load the dataset. The dataset used is a .csv file which contains the emails
- Next, we stem the words and initialise the stemmer
- Next we define the preprocessing function. This includes:
  - Tokenisation
  - Removing punctuation
  - Setting all text to lower case
  - Removing stop words
  - Stemming
- Next, just in case, we fill any null fields with ' '(blank character). This was not necessary with the dataset I used
- Next, we preprocess the data with the function
- Next, we extract features using TF-IDF(feature extraction method)

THANK YOU