# Title: Building a Naive Bayes Classifier for Email Spam Detection

# Introduction:

This document presents a Python script for building a Naive Bayes classifier to detect spam emails. The script utilizes the Multinomial Naive Bayes algorithm implemented in the scikit-learn library. The goal is to accurately classify emails as spam or not spam based on their content.

# Code Overview:

The Python script follows a structured workflow to achieve the spam classification task:

# Data Loading and Analysis:

The script loads a dataset containing email messages using the Pandas library.

Basic analysis of the dataset, including examining the schema and descriptive statistics, is performed.

## Preprocessing:

A new column is added to the dataset to represent whether an email is spam or not.

The dataset is split into training and testing sets using scikit-learn's train\_test\_split function.

## Vectorization:

The email text is converted into numerical features using scikit-learn's CountVectorizer.

## Model Training:

A Multinomial Naive Bayes classifier is trained on the training data using scikit-learn's MultinomialNB class.

## Prediction:

The trained model is used to predict whether new emails are spam or not spam.

## Evaluation:

The accuracy of the model is evaluated on the test set to assess its performance.

## Pipeline Approach:

An alternative approach using scikit-learn's pipeline is demonstrated, combining vectorization and model training steps into a single estimator.

# Key Features:

Utilizes scikit-learn for data preprocessing, model training, and evaluation.

Implements Multinomial Naive Bayes algorithm for text classification.

Demonstrates both manual workflow and pipeline approach for building the classifier.

# Screenshots:

## Data collection

A screenshot of a computer

Description automatically generated

## Vectorization and training

A screenshot of a computer

Description automatically generated

## Prediction

A screenshot of a computer

Description automatically generated

## Pipeline Approach

A screenshot of a computer

Description automatically generated

# Conclusion:

The provided Python script offers a practical implementation of a Naive Bayes classifier for email spam detection. By leveraging scikit-learn's powerful tools, it provides a robust and efficient solution for classifying emails as spam or not spam.

Ref –

[Out-of-core classification of text documents — scikit-learn 1.4.2 documentation](https://scikit-learn.org/1.4/auto_examples/applications/plot_out_of_core_classification.html#sphx-glr-auto-examples-applications-plot-out-of-core-classification-py)

<https://uhlibraries.pressbooks.pub/buildingskillsfordatascience/chapter/naive-bayes/>