# Spam Detection with Naive Bayes

This project demonstrates a spam detection system using a Naive Bayes classifier. The dataset consists of email messages labeled as either spam or ham (not spam). The project includes data exploration, visualization, feature extraction using CountVectorizer, and model training and evaluation.

## Table of Contents

* [Installation](https://chatgpt.com/c/f2d83a64-4e65-4ae6-9938-58afba3711c5#installation)
* [Usage](https://chatgpt.com/c/f2d83a64-4e65-4ae6-9938-58afba3711c5#usage)
* [Data Exploration](https://chatgpt.com/c/f2d83a64-4e65-4ae6-9938-58afba3711c5#data-exploration)
* [Data Visualization](https://chatgpt.com/c/f2d83a64-4e65-4ae6-9938-58afba3711c5#data-visualization)
* [Feature Extraction](https://chatgpt.com/c/f2d83a64-4e65-4ae6-9938-58afba3711c5#feature-extraction)
* [Model Training](https://chatgpt.com/c/f2d83a64-4e65-4ae6-9938-58afba3711c5#model-training)
* [Model Evaluation](https://chatgpt.com/c/f2d83a64-4e65-4ae6-9938-58afba3711c5#model-evaluation)
* [Results](https://chatgpt.com/c/f2d83a64-4e65-4ae6-9938-58afba3711c5#results)

## Installation

To run this project, ensure you have Python installed along with the following libraries:

* pandas
* numpy
* seaborn
* matplotlib
* scikit-learn

## Data Exploration

The data exploration phase involves loading the dataset, displaying the first and last 10 records, providing a statistical summary, and obtaining dataframe information.

## Data Visualization

Visualize the distribution of spam and ham emails to understand the dataset's balance. A count plot is used to show the number of spam versus ham emails.

## Feature Extraction

Transform the text data into numerical data using CountVectorizer, which converts the email text into a matrix of token counts.

## Model Training

The dataset is split into training and testing sets. A Naive Bayes classifier is then trained on the training data.

## Model Evaluation

Evaluate the model's performance using a confusion matrix and a classification report. These tools help assess the model's accuracy, precision, recall, and F1-score on both training and testing data.

## Results

* The model's performance is evaluated using metrics such as precision, recall, and F1-score.
* Confusion matrices provide a visual representation of the model's performance on training and testing data.