Goal : To make a email classifier that can classify the mail as spam or not spam

Dataset:The dataset used is from kaggle “spam_ham_dataset.csv” , a brief description about dataset , the dataset consist of 4 columns serial no , label spam and ham(not spam) , the text data of the mail based on which we have to build our classification model ,and the fourth column label_num consist of numerical value for spam 1 and for non spam 0, its having 5171 mails of which 3572 mails are non spam and other 1599 mails are spam , it's a imbalanced dataset

Steps followed to achieve the goal

Data Cleaning :This consist of first converting all the text to lowercase and handling the text, this was done to remove some stop words the stop words were for english language taken from internet and kept in a list and any occurrence of them in the text data we remove them , next we remove the unwanted special characters from text like @,},/,\,# etc ,after that we remove extra spaces from the text if there are multiple spaces between text or after our previous manipulation if extra spaces are introduced we remove that.

Features, each unique word is a feature so we create a dictionary and each word is given a index but the number of unique words are too many even after data cleaning we got more that fifty thousand words , there are words which occur only once in the entire dataset so , we eliminate these words and based on frequency we have kept top 18000 higher frequency word That will be used to define the spam and non spam these 18000 words are given unique index

Model training
To achieve our goal of email classification we use bernoulli naive bayes , as this algorithm even after being simple it is taken as the base for the classification problem.

We train our model,we divide our dataset into 2 parts spam and non spam. After this we convert our every email into a vector of 0 and 1 , based on the dictionary we have created above. For each word present in the mail we have 1 corresponding to it and 0 if the word is not present in the mail. we add a dummy vector or mail to both the class spam and non spam having all the vector as 1, then we do the parameter estimation as each word is our feature we need to calculate the probabilities for all these words and for both spam and non spam so for X words we need 2X probabilities estimation these probabilities are calculated as number of occurrence of word in that class divide by total number of mails in that class this is done 2X times and latter 1 for the estimation of the spam or not , P is calculate which is the ratio of number of spam emails / (total no of emails) , after all the probabilities are calculated we store them in a dictionary as these will be used to predict if the mail we receive is spam or not

Testing

After we get our test email for which we have to make prediction, for all the words which are in the mail and which are also present in the dictionary for which we have calculated probabilities We calculate the probabilities P0 and P1.