iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XII **Month of publication:** December 2022

DOI: https://doi.org/10.22214/ijraset.2022.48051

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

Spam Mail Classifier

Aryan Shirwadkar[1], Samuel Jacob[2]
Information Technology, Vidyalankar Institute of Technology, Mumbai

Abstract: Email is the worldwide use of communication application. It is because of the ease of use and faster than other communication application. However, its inability to detect whether the mail content is either spam or ham degrade its performance. Nowadays, lot of cases have been reported regarding stealing of personal information or phishing activities via email from the user. This project will discuss how machine learning help in spam detection. Machine learning is an artificial intelligence application that provides the ability to automatically learn and improve data without being explicitly programmed. Binary classifierwill be used to classify the text into two different categories: spam and ham. The algorithm will predict the score more accurately. The objectiveof developing this model is to detect and score word faster and accurately.
Keywords: Mail, Classifier, Detection

I. INTRODUCTION

In today's globalized world, email is a primary source of communication. This communication can vary from personal, business, corporate to government. With the rapid increase in email usage, there has also been increase in the SPAM emails. SPAM emails, also known as junk email, involves nearly identical messages sent to numerous recipients by email. Apart from being annoying, spam emails can also pose a security threat to computer system. We try to identify patterns using classification algorithms to enable us to classify the emails as HAM or SPAM.

In this project, a machine learning technique is used to detect the spam message of a mail. Machine learning is where computers can learn to do something without the need to explicitly program them for the task. It uses data and produce a program to perform a task such as classification. Machine learning techniques require messages that have been successfully pre-classified. The pre-classified messages make the training dataset which will be used to fit the learning algorithm to the model in machine learning studio.

A specific algorithm is used to learn the classification rules from these messages. Those algorithms are used for classification of objects of different classes. The algorithms are provided with input and output data and have a self-learning program to solve the given task.

Searching for the best algorithm and model can be time consuming. The Naive Bayes classifier is best used to classify the type of message either spam or ham. This algorithm is used to predict the probability and classification of data outcome.

II. LITERATURE REVIEW

A. Survey of Existing/Similar System

1) Google Spam Filter
2) Outlook Spam Filt

B. Project Contribution

1) This system is easy to use by the individual member due to flexibility of system.
2) It is user friendly which help the user to classify mail.
3) It provides sensitivity to the client and adapts well to the future spam techniques.
4) It considers a complete message instead of single words with respect to its organization.

III. PROBLEM STATEMENT

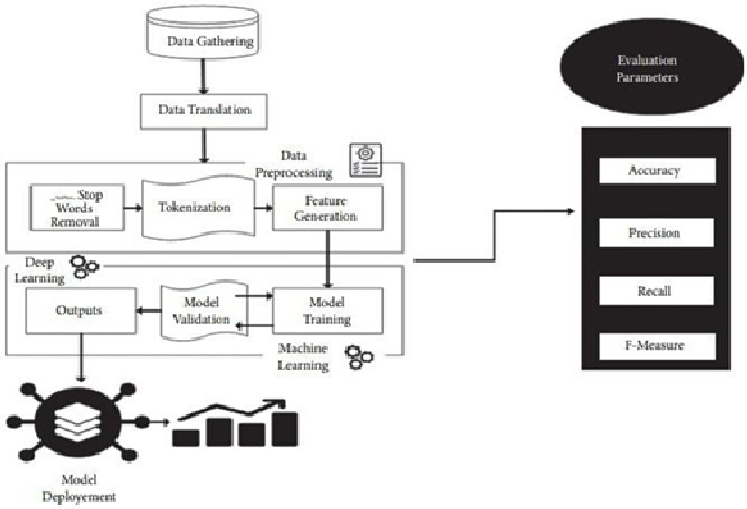
A tight competition between filtering method and spammers is going on per day, as spammers began to use tricky methods to overcome the spam filters like using random sender addresses or append random characters at the beginning or end of mails subject line. There is a lack of machine learning focuses on the model development that can predict the activity. Spam is a waste of time to the user since they must sort the unwanted junk mail and it consumed storage space and communication bandwidth. Rules in other existing must be constantly updated and maintained make it more burden to some user and it is hard to manually compare the accuracy of classified data.

IV. PROPOSED SYSTEM

A. Introduction

Naive Bayes work on dependent events and the probability of an event occurring in the future that can be detected from the previous occurring of the same event. This technique can be used to classify spam e-mails, words probabilities play the main rule here. If some words occur often in spam but not in ham, then this incoming e-mail is probably spam. Naive Bayes classifier technique has become a very popular method in mail filtering Email. Every word has certain probability of occurring in spam or ham email in its database. If the total of words probabilities exceeds a certain limit, the filter will mark the e-mail to either category. Here, only two categories are necessary: spam or ham.

B. Algorithm and Process Design



C. Naive Bayes

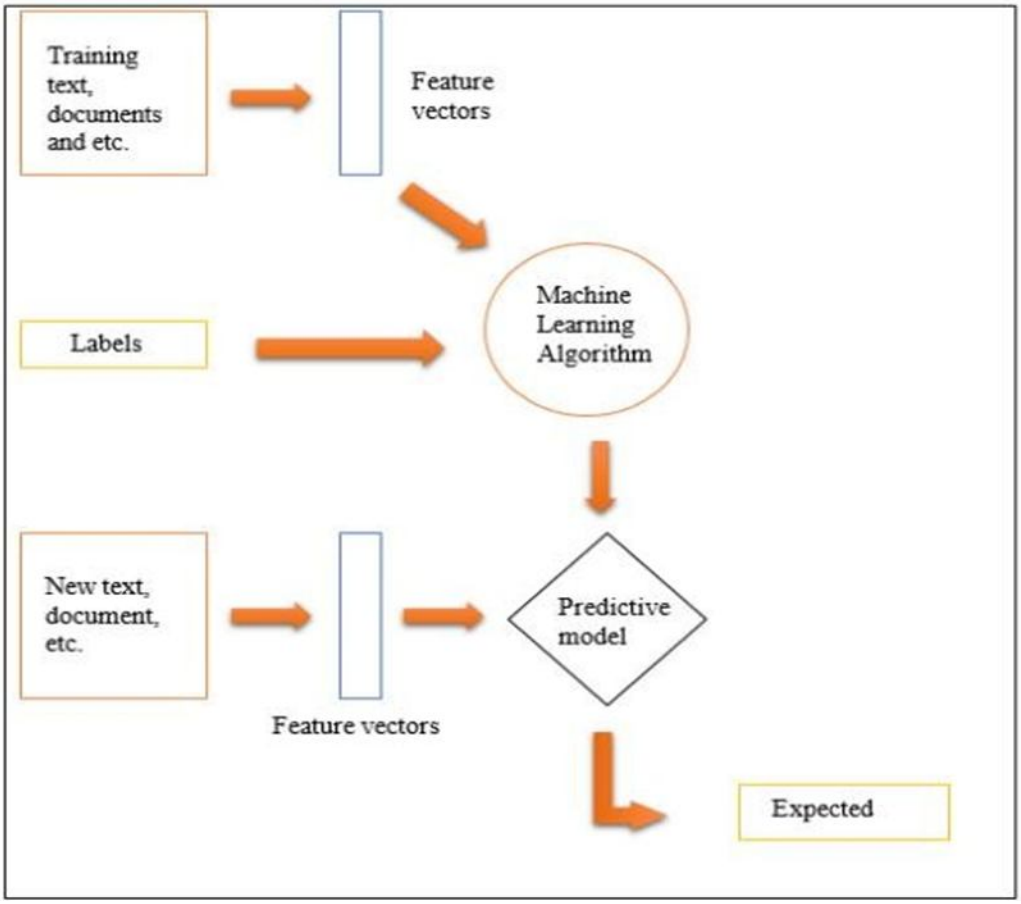
Naive Bayes classifier is based on Bayes theorem with an assumption of strong independence. The classifier is a probability-based classifier which computes the class probabilities of the given instances. The probability set is calculated by computing the combinational and frequency values of the data set. The class probability which is nearest to the rear end will be picked by the classifier. The Naïve Bayes classifier is a multiclass classifier and works efficiently with supervised learning approach. The concept of the Naïve Bayes classifier is explained with the help of Eq. (1). $P(y|x) = P(x|y)\ P(y)\ P(x)$

(1) Here, x is the set of feature vectors (x1, x2, x3,. xn) and y stands for the class variable with *m*possible outcomes (y1, y2, y3…. yn). P(y|x) is the posterior probability which depends on the likelihood of the feature set or attribute value belonging to class P(x|y), P(y) is the prior probability and P(x) is the evidence depending on the known feature variables.

D. SVM

SVM is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyperplane that differentiate the two classes very well [4].

Support Vectors are simply the coordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line). If the data requires non-linear classification, SVM can employee Kernels, which are functions which takes low dimensional input space and transform it to a higher dimensional space i.e., they convert non separable problem to separable problem.



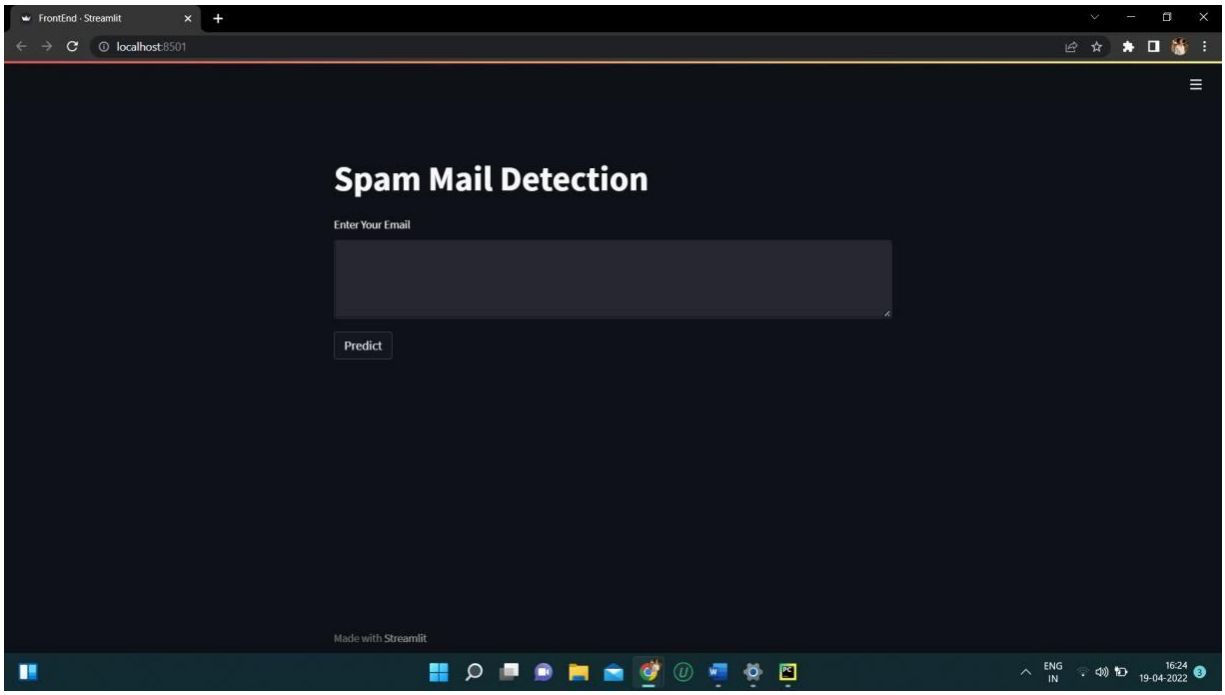
E. Working of Algorithm

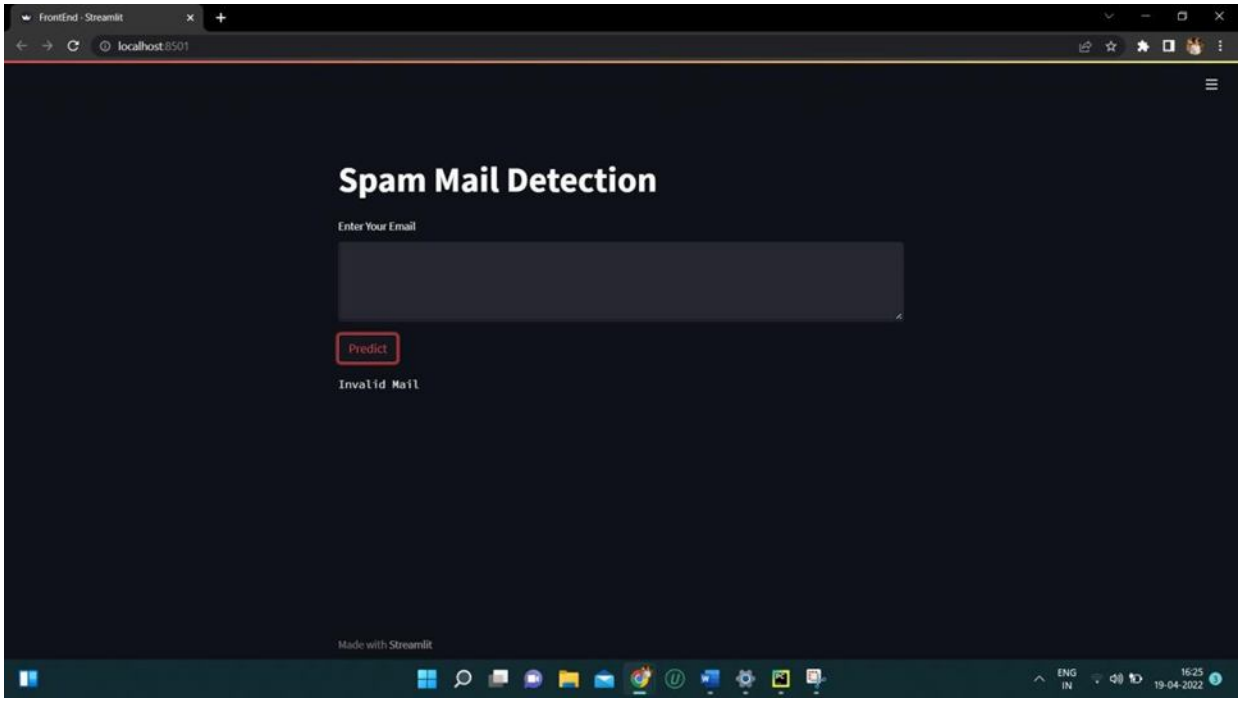
Input	Subject: A guaranteed return of 70% on investment . . . $300 million increase in the economic growth in a year. After breakup of the largest monopoly of America earn some profit. Over 70% of returns annually. For complete details just click here
Tokenization	'Subject'':' 'A' 'guaranteed' 'return' 'of' '70%' 'on' 'investment' '. . .' '$300' 'million' 'increase' 'in' 'the' 'economic' 'growth' 'in' 'a' 'year' '.' 'After' 'breakup' 'of' 'the' 'largest' 'monopoly' 'of' 'America' 'earn' 'some' 'profit' '.' 'Over' '70%' 'of' 'returns' 'annually' '.' 'For' 'complete' 'details' 'just' 'click' 'here'
Stop Word Removal	'guaranteed' 'return' '70%' 'investment' '$300' 'million' 'increase' 'economic' 'growth' 'year' 'After' 'breakup' 'largest' 'monopoly' 'America' 'earn' 'some' 'profit' 'Over' '70%' 'returns' 'annually' 'complete' 'details' 'just' 'click' 'here'
Stemming	'guarantee' 'return' '70%' 'investment' '$300' 'million' 'increase' 'economic' 'growth' 'year' 'After' 'breakup' 'largest' 'monopoly' 'America' 'earn' 'some' 'profit' 'Over' '70%' 'returns' 'annual' 'complete' 'details' 'just' 'click' 'here'
Classify Algorithm Output (NB or J48)	Email is Spam mail

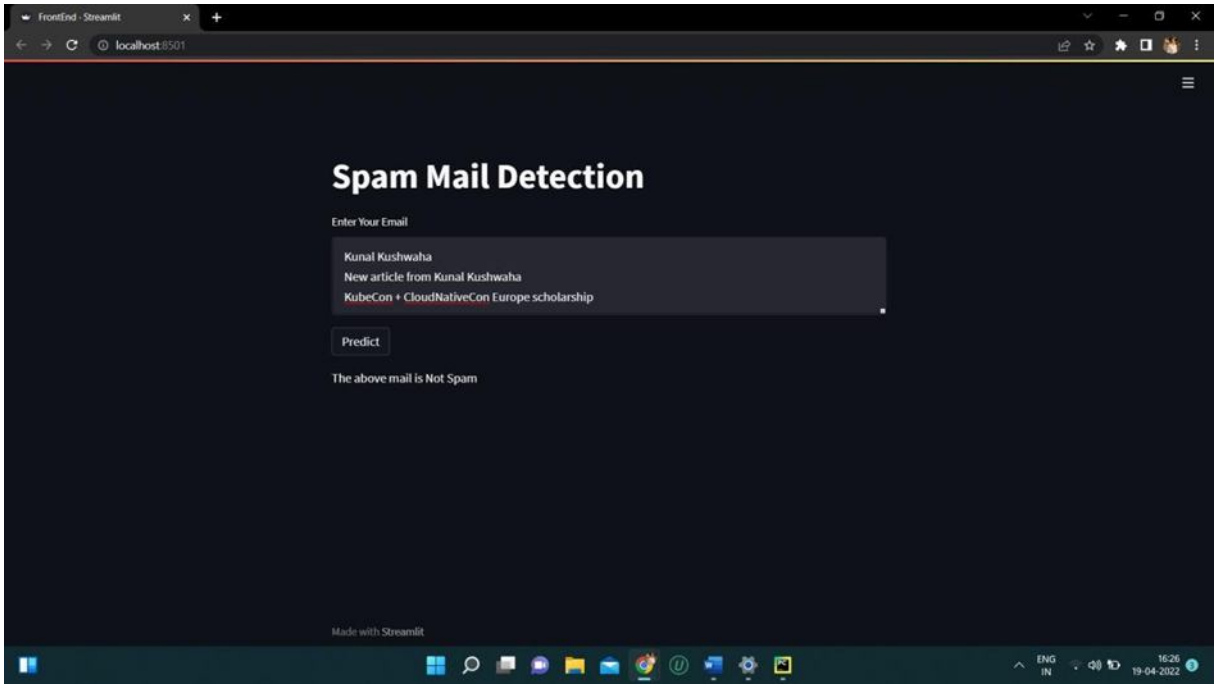
F. Details of Hardware & Software used Are

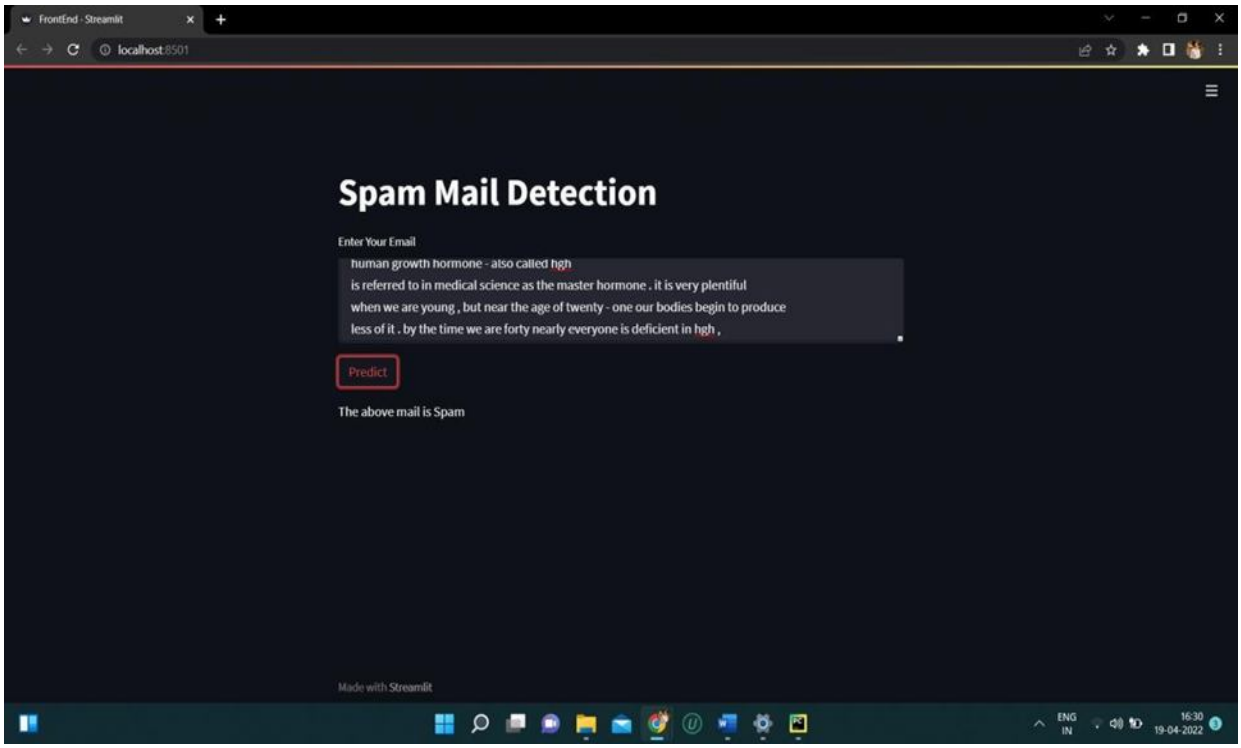
1) Python
2) Kaggle – Det dataset
3) Jupyter Notebook – Implementing model
4) Stream Lit - Front-end
5) PyCharm IDE

V. EXPERIMENT AND RESULTS









VI. CONCLUSIONS

We have successfully developed and implemented an intelligent crop recommendation system in this study that Indian farmers can utilize right away. The farmers would benefit from this system's help in making educated decisions about the types of diseases affecting their crops and possible treatments. Mandi Price, the ability to rent out equipment, and the ability to support many languages have also been included. Utilizing the methods outlined previously, the building of an electronic expert system for the identification of plant diseases affecting the leaves is accomplished with the inclusion of the other services outlined above.

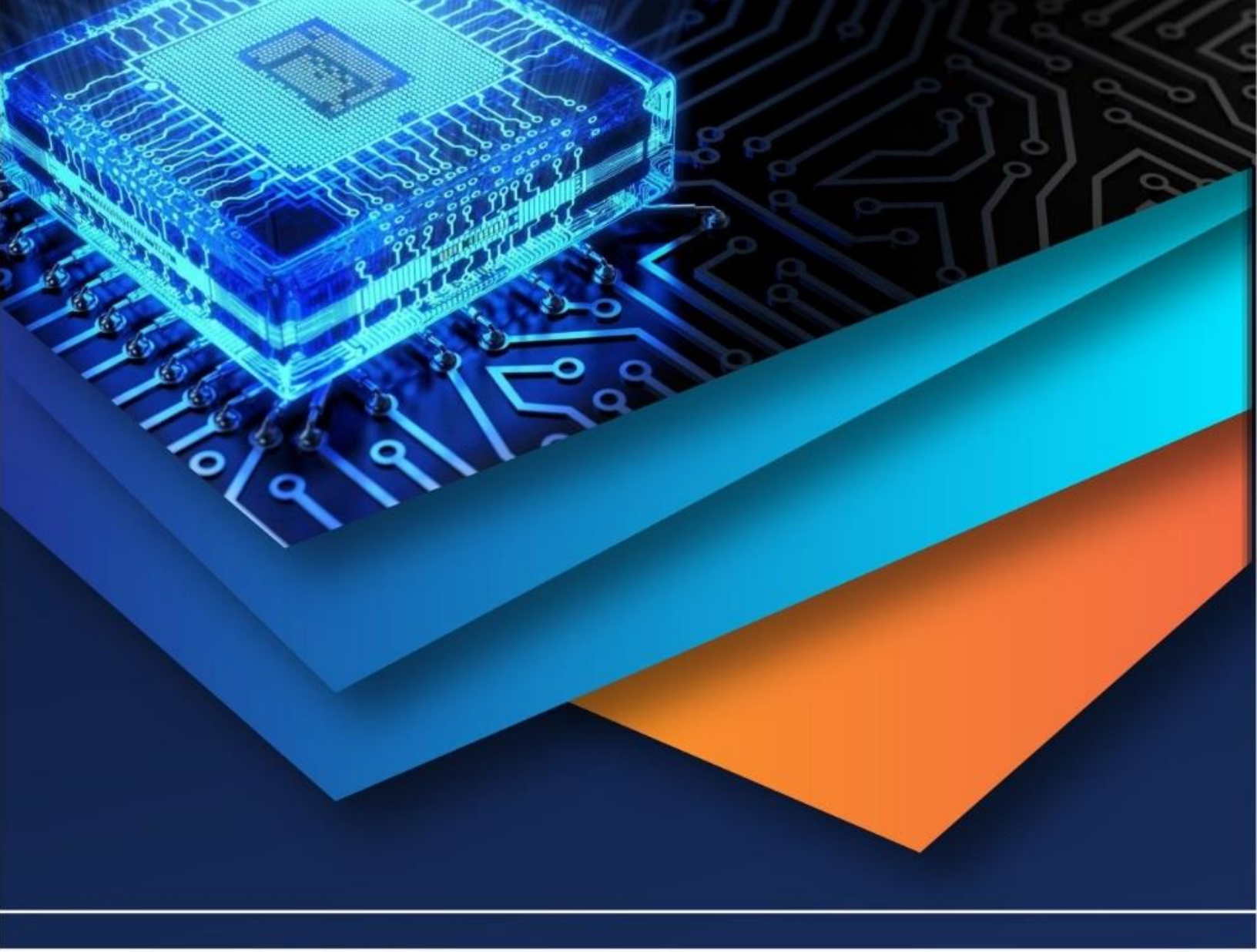
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

VII. ACKNOWLEDGMENT

We would like to express our special thanks of gratitude to our guide Prof. Samuel Jacob who gave us the golden opportunity to do this wonderful project on the topic Spam Mail Classifier which also helped us in doing a lot of Research and we came to know about so many new things we are thankful to him. Secondly, we would also like to thank our professors of review panel in finalizing and improving this project within the limited time frame. This project helped us in understanding the various parameters which are involved in the development of desktop application and the working and integration of front end along with the back end.

REFERENCES

[1] https://ieeexplore.ieee.org/document/7868411/
[2] https://ieeexplore.ieee.org/document/9291921
[3] https://ieeexplore.ieee.org/document/4403192



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY