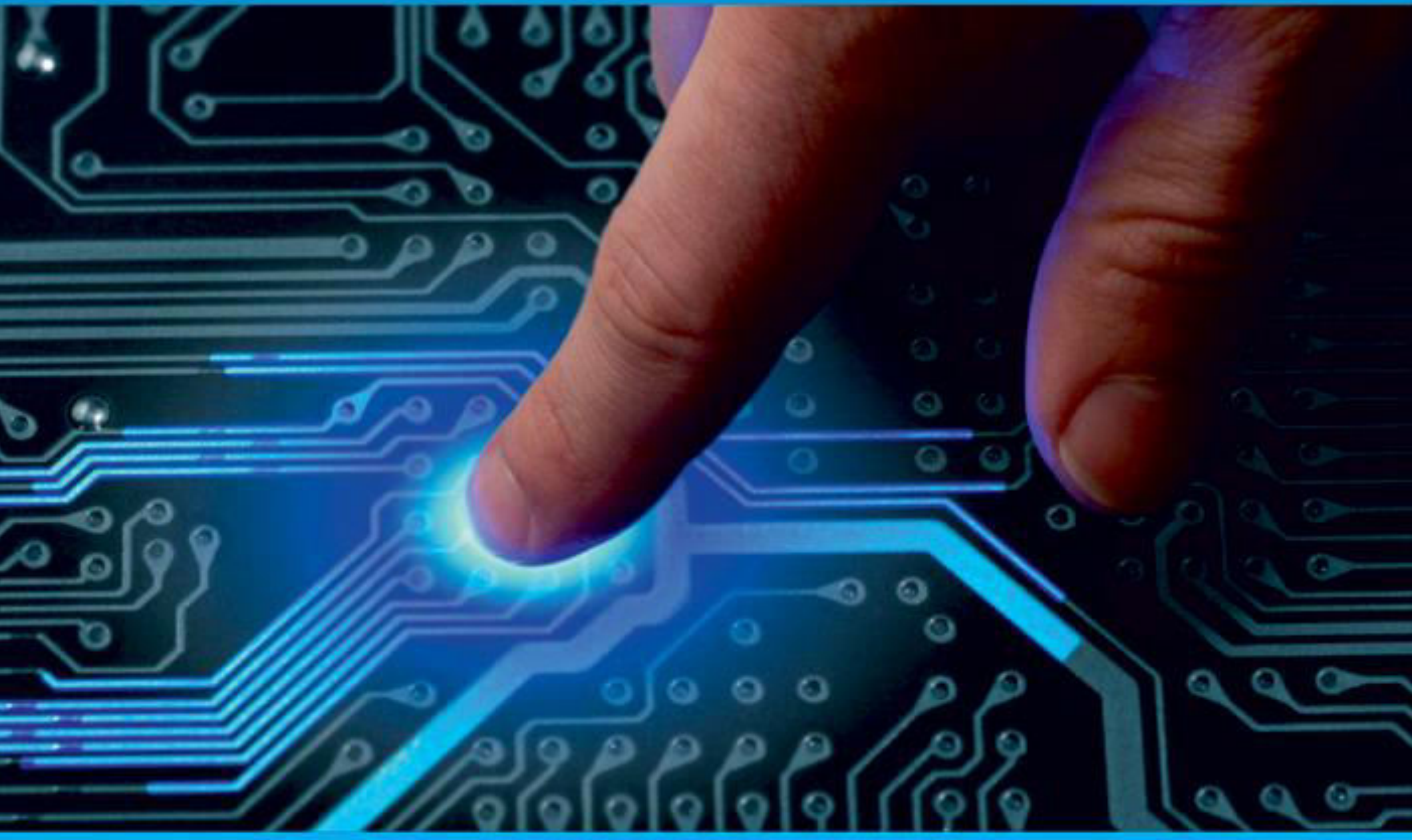
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 3, March 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

| DOI: 10.15680/IJIRCCE.2023.1103049 |

Machine Learning Based Spam Detection System

Samiksha Patade, Komal Shewale, Nandita Suryawanshi, SuchitaGirase, Prof. Dr. D. D. Bage

UG Students, Dept. of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, India

Assistant Professor, Dept. of Computer Engineering, K. K. Wagh Institute of Engineering Education &Research, Nashik, India

ABSTRACT: Nowadays email has become commonly used social media platform for information exchange for businesses, education and in various sectors. Everyone is believing on the emails or messages which are send by unauthorized persons and coming from any website therefore there is huge chances of frauds, hacking, information stalking is happening with the help of emails. With the increasing rate of internet users there is major problem of email spam is introduced. To prevent these frauds, hacking, detecting spam emails and messages is became need of time nowadays. In this paper, we are using voting classifier which is made up of three main classification models like Multinomial Naïve Bayes, Support Vector Machine, Extra Tree Classifier, a content-based technique for identifying spam email from the body of email. This system will identify spam email and covey the result to user.

KEYWORDS: Multinomial Naïve Bayes, Support Vector machine, Extra Tree Classifier, spam

I. INTRODUCTION

The most popular form of communication is email. It is due to the simplicity of use and speed compared to other communication applications. However, its performance is diminished by its inability to determine if the email content is spam or ham. Nowadays, numerous incidents involving the theft of personal information from users or phishing scams have been documented. We will talk about how machine learning aids in spam detection in this project. Machine learning is a form of artificial intelligence that enables the automatic learning and improvement of data without explicit programming. Using a binary classifier, the text will be divided into two groups which are spam and ham. The score will be predicted more precisely by the algorithm.

For solving these problems, there are various techniques available for identification of spam messages. In this paper we are using content-based approach for spam detection. Various machine learning algorithms such as Naïve Bayes algorithm and it's variants, Support Vector Machine, K-Nearest Neighbour, Decision Tree Classifier, Logistic Regression, Random Forest Classifier, Bagging, AdaBoost Classifier, Extra Tree Classifier, Gradient Boosting Classifier and XGB Classifier are used to detect spam emails.

Since the result shows that Multinomial Naïve Bayes, Support Vector Machine and Extra Tree Classifier gives good performance based on the accuracy and precision. This paper represent voting classifier using this three algorithms for detecting spam messages and convey the result to the user.

II. LITERATURE SURVEY

In [1] for detecting the spam messages they use Multilayer perceptron, Decision Tree Classifier, simplistic classification by Naïve Bayes be carried out on a Mail Server and Mail supporter as a method. They were comparing their proposed model with the existing one. When compared to the current model, the classification accuracy of proposed model is better and high, then it was working properly. Then they were concluded from the foregoing that the existing model's classification accuracy is lower than that of the suggested model.

In [2]have proposed paper on detection of spam using Supervised Machine Learning. They have used Naïve Bayes, neural Networks, K-Nearest Neighbour, Decision Trees, and Support Vector machine learning algorithms for detection of spam. Among which they have got best accuracy for Naive Bayes Classifier.

In [3] have proposed paper on email spam detection using ML algorithms. They have found that Multinomial Naive Bayes gives the best accuracy for spam detection. Along with that they have proposed an ensemble learning model

which make use of algorithms include "Naïve Bayes, Support Vector Machines, Neural Networks, K-nearest neighbour, Random Forest, etc".
In [4] have proposed a paper on email spam detection using integrated approach. They have used a integrated approach of Naive Bayes and Particle Swarm Optimization. When tested with results given by individual Naive Bayes algorithm this integrated approach gives the best accuracy.

III. PROBLEM DEFINITION

Spam has become a most critical and important issue on social media. Due to spam messages systems performance is degraded. Using these spam messages various attacks, frauds and illegal access of information is happened. Global spam volume as percentage of total e-mail traffic till 2021, there were 45.56% of spam messages occurs.

IV. METHODOLOGY

A. Data Pre-processing:

As Machines can only understand the language of 0s and 1s, we need to pre-process the data. As the raw data is in the text format it is not understand by the machine. Therefore, various steps of data pre-processing should be carried out on the text.
Following steps are performed during Data Pre-processing:
Data Cleaning: In this step we filling out the "missing values", "smoothing of noisy data", "identifying or removing outliers ", and "resolving of inconsistencies" are carried out.
Data Integration: The process of merging of data from several sources into a single, cohesive perspective is known as data integration.
Data Transformation: This technique must be performed before actual data mining. It changes format, structure of the data and convert it into usable data.
Data Reduction: In this step the data is reduced, typically the volume of data.

1. *Tokenization:*
   Tokenization is the process of splitting the text into the words called as tokens.
   Ex: *"I went to school."*
   tokens = [" I ", "went "," to "," school. "]
2. *Stop wordsRemoval:*
   Stop words are the most commonly used English words that are not useful in the given sentence. To reduced processing of this, Stop words should be ignored or removed.
   'I', 'me', 'The', 'Am', 'Is', 'Are' and so on these are nothing but the stop words.
3. *Stemming:*
   Stemming is the process of reducing the word into its base/root form.
   Ex: The word boat is stem for [boat, boater, boating, boats].
4. *TF-IDF Vectorizer:*
   TF is nothing but Term Frequency and IDF is Inverse Document Frequency.
   TF-IDF is used for extracting the main features of data.

$$TF\ (word) = (Number\ of\ times\ the\ word\ appeared\ in\ a\ Document)/\ (Total\ Number\ of\ words\ in\ a\ Document)$$

$$IDF\ (word) = loge\ (Total\ Documents/\ Number\ of\ Documents\ with\ term\ `word'\ in\ it)$$

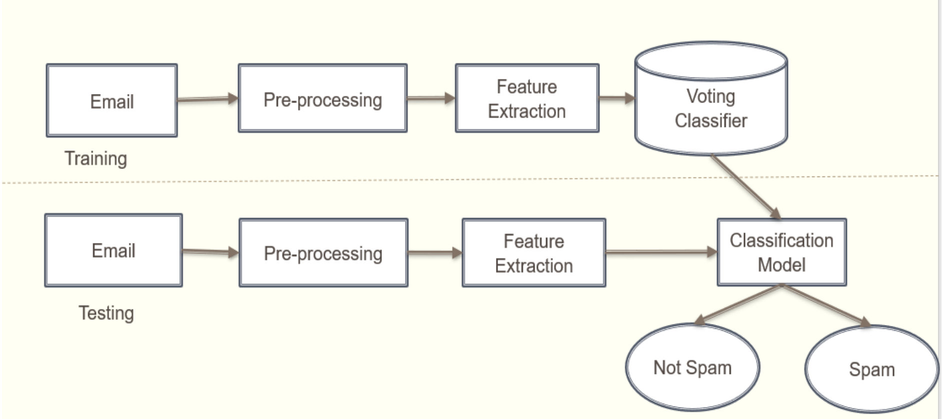


Fig 1. Block Diagram for Spam Detection System

B. Algorithms used:

1. Naïve Bayes:

Naïve Bayes classifier algorithm becomes the best technique for email filtering. Naïve Bayes algorithm is based on bayes theorem and naïve bayes is a supervised machine learning algorithm used for solving classification problems. Bayes Theorem is based on the concept of conditional probability. Conditional Probability is probability of an event happening, based on the previous event. The Naïve Bayes Classifier is used to categorized spam emails since word likelihood is the key factor in this process. Any term that frequently occurs in spam but not in ham indicates that an email is spam. Simple bayes classifier algorithm has emerged as the most effective method of email classification. For this, the model is extremely well trained using naïve bayes filter. Each class probability is always calculated by the naïve bayes algorithm, and the class with the highest probability is then selected as the output. The results of naïve bayes are always reliable.

$$P(A/B) = P(B/A).\ P(A) / P(B)$$

There are three variants of naïve bayes algorithm which are Multinomial, Gaussian and Bernoulli. Among them this model uses Multinomial Naïve Bayes. Because the Multinomial Naïve Bayes is good at handling the discrete data.

2. Support Vector Machine:

Support Vector Machine (SVM) is supervised machine learning algorithm used for classification purpose. The goal of SVM is to create best decision boundary that can segregate n-dimensional data into classes. The best decision boundary or the best line that can segregate the given data into classes is called hyperplane. Support Vector Machine algorithm gives the output as a hyperplane which classifies new sample into appropriate classes.

3. Extra Tree Classifier:

Extra Tree is an extremely randomized tree is an ensemble machine learning technique, like random forest, that trains a large number of decision trees and combines the output of the group of decision trees to provide a forecast. There are smaller differences between Random Forest algorithm and Extra Trees classifier. Random Forest algorithm chooses the optimum split when creating the decision trees while Extra Tree classifier chooses the split randomly. Extra Trees classifier would be a better option than other ensemble tree-based models when developing models. When developing models that need significant feature engineering or feature selection processes where computational cost is high. Using Extra Tree classifier computational cost is reduced.

4. Voting Classifier:

Voting Classifier is mainly used for improvement of performance of model. Voting Classifier is an ensemble machine learning method that trains various based models and predicts the output on the basis of aggregate result of each base model. The voting criteria used in voting classifier is of two types.

I. Hard Voting:

In hard voting classifier majority voting is taken as outcome. Suppose three classifiers generates a prediction as (1,1,0) then final output generate from hard voting classifier is 1.

II.Soft Voting:

Soft Voting classifier classifies the input data based on the probabilities of prediction made by different models. In soft voting classifier mean of all the probabilities of every classifier is taken and based upon this average probability final prediction is done. Suppose for email to be spam the probability is 0.8 and for ham the probability is 0.3 and if our voting classifier generates the probability 0.8 then the message is spam.

In our proposed model we have created soft voting classifier using Multinomial Naïve Bayes, Support Vector Machine and Extra Tree Classifier.

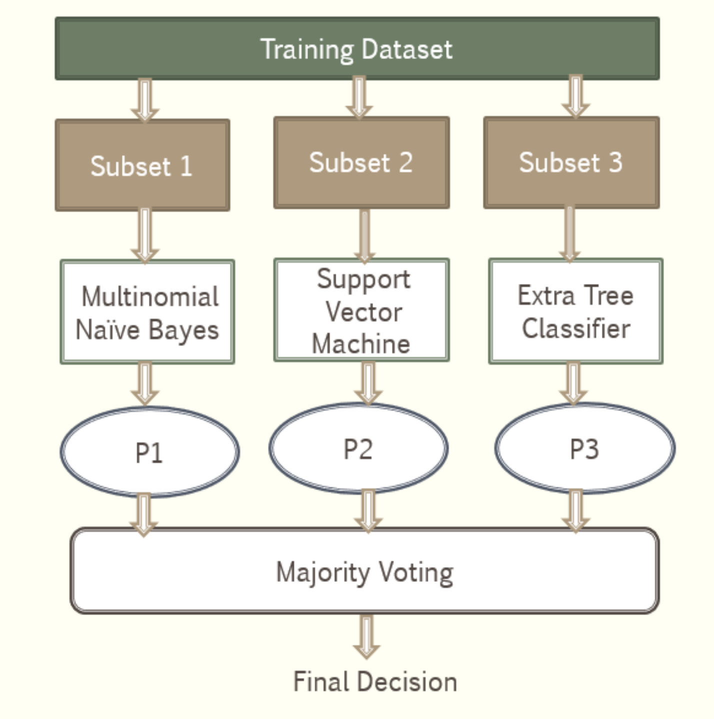


Fig 2. Voting Classifier

V. PERFORMANCE METRICS

Various performance metrics are used to check the accuracy of model. The performance is calculated in the terms of precision and classification accuracy. For this, first confusion matrix is drawn using the results given by model. Confusion matrix gives true positive (TP), true negative(TN), false positive (FP) and false negative (FN) tuples.

True Positive (TP) is the number of spam messages correctly identified as spam.
True Negative (TN) is the number of ham messages correctly identified as ham.
False Positive (FP) is the number of ham messages incorrectly identified as spam.
False Negative (FN) is the number of spam messages incorrectly identified as ham.

I. Precision:

precision calculates the effectiveness of classifier. Precision is the percentage of correctly labeled spam emails out of all the results that shows spam email.

Precision (P) = (TP) / (TP+FP)

II. Accuracy:

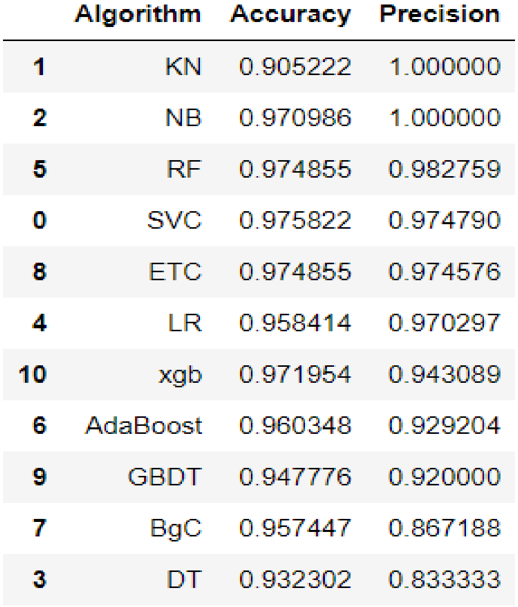
Accuracy is the ratio of correct predictions to the total predictions made by classifier.

Accuracy = (TP+TN) / (TP+TN+FP+FN)

VI. RESULTS

In this paper, the proposed system is created for identification of spam and ham messages using a voting classifier is done.

Firstly, we have checked the accuracy of algorithms such as Naïve Bayes algorithm and its variants, Support Vector Machine, K-Nearest Neighbour, Decision Tree Classifier, Logistic Regression, Random Forest Classifier, Bagging, AdaBoost Classifier, Extra Tree Classifier, Gradient Boosting Classifier and XGB Classifier.



	Algorithm	Accuracy	Precision
1	KN	0.905222	1.000000
2	NB	0.970986	1.000000
5	RF	0.974855	0.982759
0	SVC	0.975822	0.974790
8	ETC	0.974855	0.974576
4	LR	0.958414	0.970297
10	xgb	0.971954	0.943089
6	AdaBoost	0.960348	0.929204
9	GBDT	0.947776	0.920000
7	BgC	0.957447	0.867188
3	DT	0.932302	0.833333

Fig 3. Comparison chart

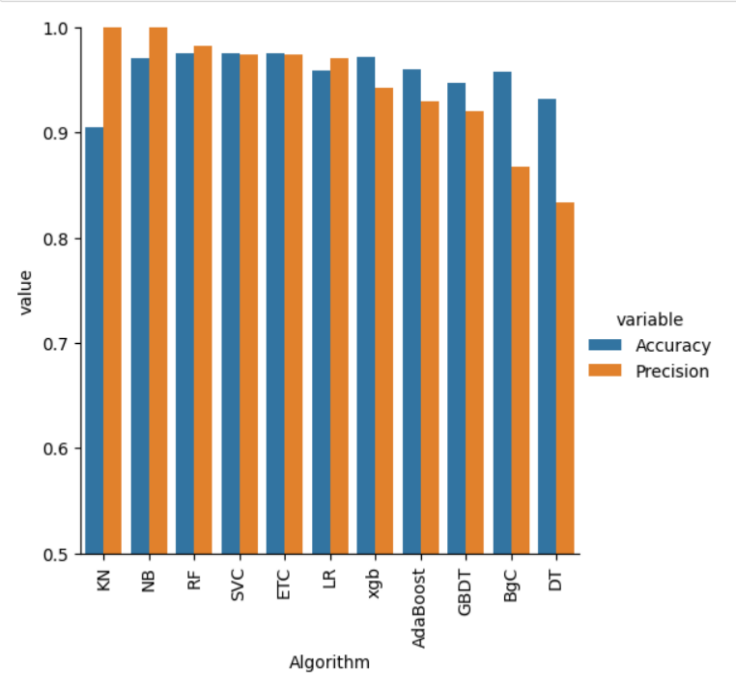
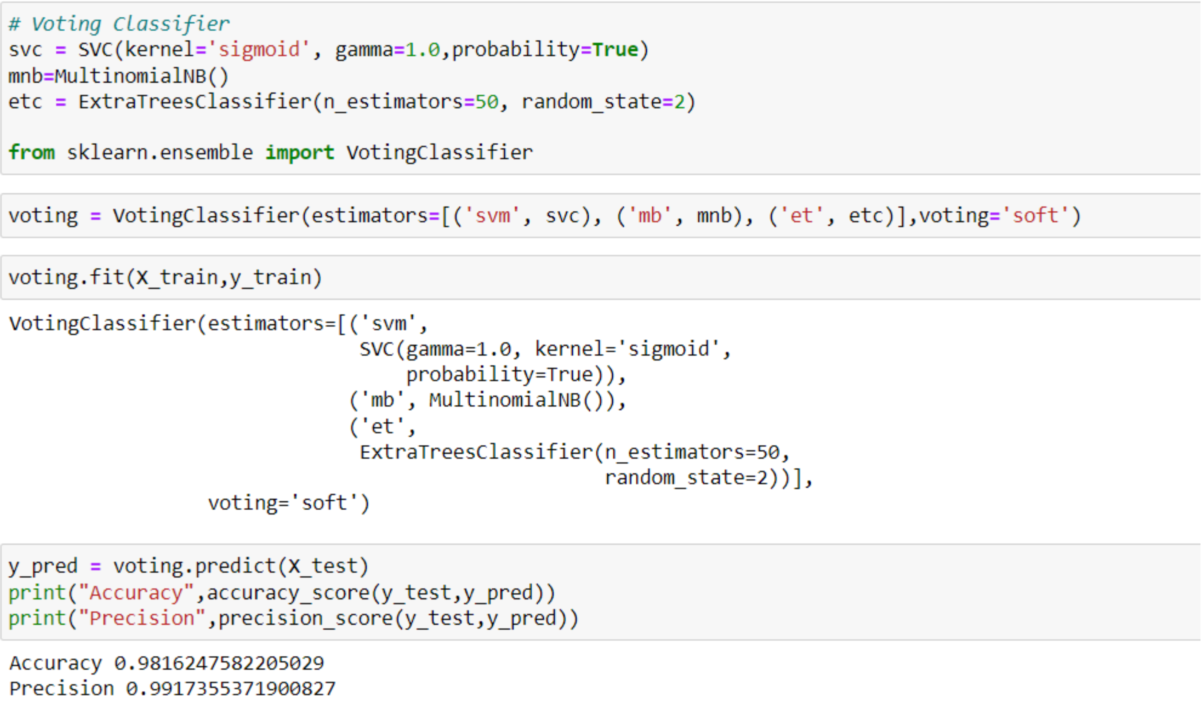


Fig 4. Visualized Comparison

From the comparison chart we can interpret that Multinomial Naïve Bayes, Support Vector Machine and Extra Tree Classifier algorithms gives the good accuracy along with precision. Therefore, voting classifier made up of these three algorithms gives the best accuracy as well as precision.



```
# Voting Classifier
svc = SVC(kernel='sigmoid', gamma=1.0,probability=True)
mnb=MultinomialNB()
etc = ExtraTreesClassifier(n_estimators=50, random_state=2)

from sklearn.ensemble import VotingClassifier
```

```
voting = VotingClassifier(estimators=[('svm', svc), ('mb', mnb), ('et', etc)],voting='soft')
```

```
voting.fit(X_train,y_train)
```

```
VotingClassifier(estimators=[('svm',
                              SVC(gamma=1.0, kernel='sigmoid',
                                  probability=True)),
                             ('mb', MultinomialNB()),
                             ('et',
                              ExtraTreesClassifier(n_estimators=50,
                                                   random_state=2))],
                 voting='soft')
```

```
y_pred = voting.predict(X_test)
print("Accuracy",accuracy_score(y_test,y_pred))
print("Precision",precision_score(y_test,y_pred))
```

```
Accuracy 0.9816247582205029
Precision 0.9917355371900827
```

VII. Conclusion and Future Work

Nowadays, email has become a popular platform for communication. An email spam is the most demanding research topic due to the cybercrimes are happening with the help of emails.

Use of different machine learning algorithms to improve the performance and accuracy of spam detection is the primary topic of our study. We have highlighted a data mining approach for spam detection using a voting classifier which is made up of Multinomial Naïve Bayes, Support Vector Machine and Extra Tree Classifier. In our proposed system, we have detected spam by extracting the content of email body. For further enhancement, this proposed system can be deployed with different email techniques by adding some features in future along with that spam filtering using domain names, header based and based on IP address can be done.

References

[1] V. Sasikala, K. Mounika, Y. SravyaTulasi, D. Gayathri, M. Anjani,” Performance evaluation of Spam and Non-Spam E-mail detection using Machine Learning algorithms”, Proceeding of the International Conference on Electronics and Renewable Systems, ISBN:978-1-6654- 8425-1,2022.

[2] Abhila B1, Delphin periyanayagi M1, Koushika M1,” Spam Detection System Using Supervised ML”, International Conference on System, Computation, Automation and Networking(ISSCAN), ISBN:978- 1-6654-3986-2,2021.

[3] Nikhil Govil, Kunal Agarwal, Ashi Bansal, Astha Varshney,” A Machine Learning Based Spam Detection Mechanism”, International Conference of Computing Methodologies and Communication(ICCMC), ISBN:978- 1-7281-4889-2, 2020.

[4] Nikhil Kumar, SanketSonowal, Nishant,” Email Spam detection using machine learning algorithm”, Proceedings of the Second International Conference on Inventive Research in Computing Applications(ICIRCA), ISBN:978-1-7281-5374-2, 2020.

[5] Thashina Sultana, K A Sapnaz, Fathima Sana,,”Email Spam Detection using integrated approach of Naive Bayes and Particle Swarm Optimization”, International Journal of Engineering Research Technology(IJERT), ISSN:2278-0181, 2020.

[6] Kriti Agarwal, Tarun Kumar," Email based Spam Detection", Proceedings of the Second International Conference on Intelligent Computing and Control Systems(ICICCS), 2018.
[7] Aakash Atul Alurkar, Sourabh Bharat Ranade, "A Proposed Data Science Approach for Email Spam Classification using Machine Learning Techniques",ISBN:978-1-5386-3197-3, 2017.
[8] Asif Karim, Sami Azam, "A Comprehensive Survey for Intelligent Spam Email Detection", 2017.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.379

doi® crossref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

📱 9940 572 462 ☎ 6381 907 438 ✉ ijircce@gmail.com

Scan to save the contact details

www.ijircce.com