**International Journal of Advanced Education and Research**
**www.multidisciplinaryjournals.net**
**ISSN: 2455-5746**
**Received: 02-03-2024, Accepted: 25-03-2024, Published: 16-04-2024**
**Volume 9, Issue 2, 2024, Page No. 1-4**

# End to end spam classifier with Kaggle dataset and Heroku deployment

**C Radha[1], L Lokesh[2], N Prabhu[2], G Pavithra[2], R Prathibha[2]**
[1] Associate Professor, Department of Engineering, Muthayammal Engineering College, Namakkal, Tamilnadu, India
[2] Department of Engineering, Muthayammal Engineering College, Namakkal, Tamilnadu, India

**Abstract**
Email Spam has become a major problem nowadays, with Rapid growth of internet users, Email spams is also increasing. People are using them for illegal and unethical conducts, phishing and fraud. Sending malicious link through spam emails which can harm our system and can also seek in into your system. Creating a fake profile and email account is much easy for the spammers, they pretend like a genuine person in their spam emails, these spammers target those peoples who are not aware about these frauds. So, it is needed to Identify those spam mails which are fraud, this project will identify those spam by using techniques of machine learning, this paper will discuss the machine learning algorithms and apply all these algorithm on our data sets and best algorithm is selected for the email spam detection having best precision and accuracy.

**Keywords:** Spam, dataset, classifier

## Introduction

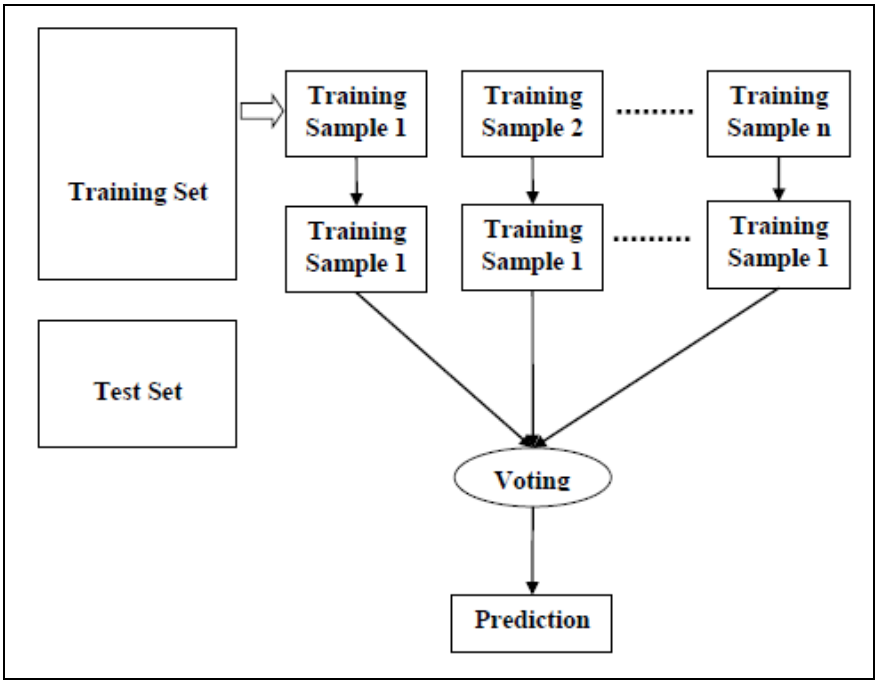
Recently unsolicited commercial bulk e-mail also known as spam, become a big trouble over the internet. Spam is waste of time, storage space and communication bandwidth. The problem of spam e-mail has been increasing for years. In recent statistics, 40% of all emails are spam which about 15.4 billion email per day and that cost internet users about $355 million per year. Automatic e-mail filtering seems to be the most effective method for countering spam at the moment and a tight competition between spammers and spam-filtering methods is going on. Spammers began to use several tricky methods to overcome the filtering methods like using random sender addresses and/or append random characters to the beginning or the end of the message subject line. Knowledge engineering and machine learning are the two general approaches used in e-mail filtering. In knowledge engineering approach a set of rules has to be specified according to which emails are categorized as spam or ham. A set of such rules should be created either by the user of the filter, or by some other authority (e.g. the software company that provides a particular rule-based spam-filtering tool). By applying this method, no promising results shows because the rules must be constantly updated and maintained, which is a waste of time and it is not convenient for most users. Machine learning approach is more efficient than knowledge engineering approach it does not require specifying any rules. Instead, a set of training samples, these samples is a set of pre classified e-mail messages. A specific algorithm is then used to learn the classification rules from these e-mail messages. Machine learning approach has been widely studied and there are lots of algorithms can be used in e-mail filtering. Machine learning field is a subfield from the broad field of artificial intelligence, this aims to make machines able to learn like human. Learning here means understood, observe and represent information about some statistical phenomenon. In unsupervised learning one tries to uncover hidden regularities (clusters) or to detect anomalies in the data like spam messages or network intrusion. In e-mail filtering task some features could be the bag of words or the subject line analysis. E-mail classification tasks are often divided into several sub-tasks. Firstly, Data collection and finally, the e-mail classification phase of the process finds the actual mapping between training set and testing set. Thealgorithm used to detect the accuracy are Naive Bayes, Support vector machine, Random forest, Decision tree, Kneighbors.

## Objectives

The main objective of the project is to find accuracy level of the spam message by using some of the algorithms. In, that algorithm which of the two algorithm gives a more accuracy level and that algorithm is named as a hybrid algorithm. For finding that we can collect some of the datasets and check whether it is spam or not. The highest level of accuracy is shows up to 99% accuracy level.

## Methodology

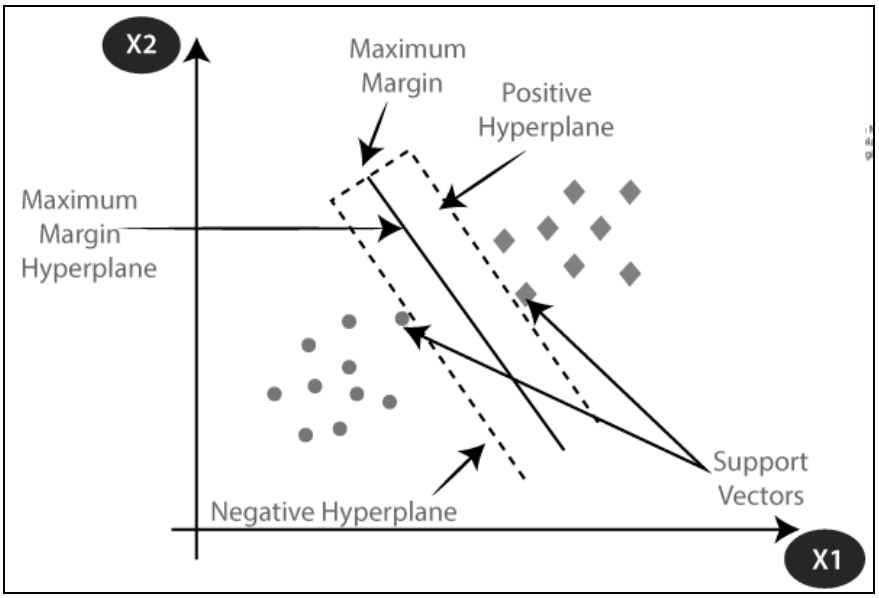
The methodology is done by using five types of algorithms such as Random Forest, Naive bayes, Decision tree, Kneighbors and Support vector machine. In, Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemblelearning**,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. It takes less training time as compared to other algorithms and it predicts output with high accuracy, even for the large dataset it runs efficiently and also maintain accuracy when a large proportion of data is missing.




In, Naive bayes it is called Naive because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other. Bayes is a functionality of getting a probability. The steps to implement the naive bayes is,

- Data Pre-processing step
- Fitting Naive Bayes to the Training set
- Predicting the test result
- Test accuracy of the result
- Visualizing the test set result.

In, support vector machine the main goal is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

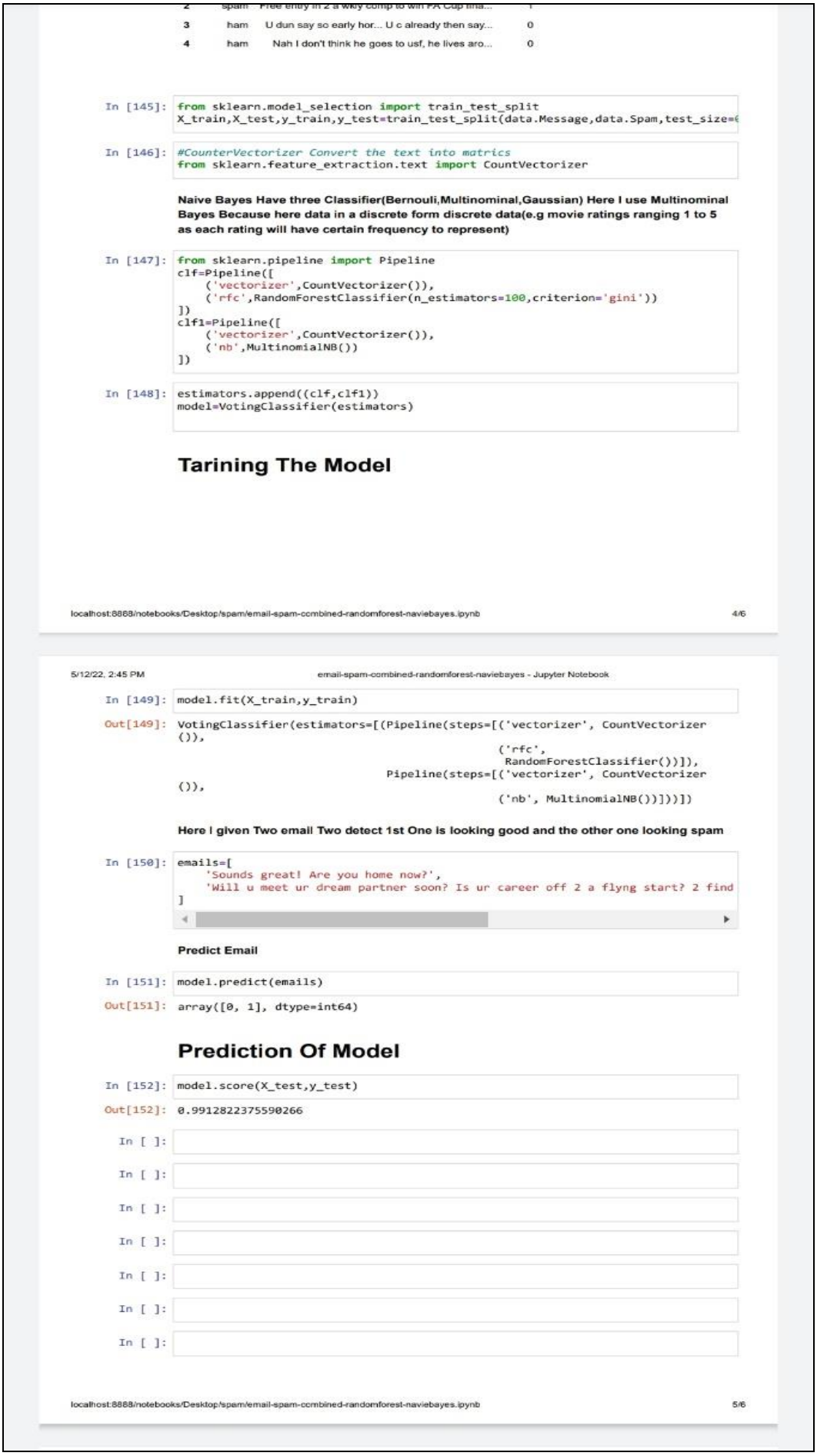



In, Kneighbors algorithm there is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5. A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.Large values for K are good, but it may find some difficulties.

In, Decision tree algorithm There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

**Result**



| | | | |
|---|---|---|---|
| 3 | ham | U dun say so early hor... U c already then say... | 0 |
| 4 | ham | Nah I don't think he goes to usf, he lives aro... | 0 |

In [145]:
```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(data.Message,data.Spam,test_size=
```

In [146]:
```
#CounterVectorizer Convert the text into matrics
from sklearn.feature_extraction.text import CountVectorizer
```

**Naive Bayes Have three Classifier(Bernouli,Multinominal,Gaussian) Here I use Multinominal Bayes Because here data in a discrete form discrete data(e.g movie ratings ranging 1 to 5 as each rating will have certain frequency to represent)**

In [147]:
```
from sklearn.pipeline import Pipeline
clf=Pipeline([
    ('vectorizer',CountVectorizer()),
    ('rfc',RandomForestClassifier(n_estimators=100,criterion='gini'))
])
clf1=Pipeline([
    ('vectorizer',CountVectorizer()),
    ('nb',MultinomialNB())
])
```

In [148]:
```
estimators.append((clf,clf1))
model=VotingClassifier(estimators)
```

## Tarining The Model





In [149]:
```
model.fit(X_train,y_train)
```

Out[149]:
```
VotingClassifier(estimators=[('Pipeline(steps=[('vectorizer', CountVectorizer
())),
                                               ('rfc',
                                                RandomForestClassifier())])',
                              Pipeline(steps=[('vectorizer', CountVectorizer
())),
                                               ('nb', MultinomialNB())])')])
```

**Here I given Two email Two detect 1st One is looking good and the other one looking spam**

In [150]:
```
emails=[
    'Sounds great! Are you home now?',
    'Will u meet ur dream partner soon? Is ur career off 2 a flyng start? 2 find
]
```

**Predict Email**

In [151]:
```
model.predict(emails)
```

Out[151]:
```
array([0, 1], dtype=int64)
```

## Prediction Of Model

In [152]:
```
model.score(X_test,y_test)
```

Out[152]:
```
0.9912822375590266
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

## Conclusion

Security is a major issue nowadays. In this spam plays a major role in network which creates a serious damage to authenticated users. We conclude that to predict the accuracy rate with the help of some algorithms and import the minimum amount of datasets to find the accuracy. Because, Machine learning does not support large amount of datasets.

## References

1. Balakrishnan D. Machine Learning to Improve the Accuracy of a Period Prediction App. In: International Conference on Recent Trends in Emerging Technologies and Engineering, 2022.
2. Balakrishnan D. Appliance of Artificial Intelligence Techniques to Develop a Chatterbot. Pramana Research Journal, 2018, 8(9). [Page range if applicable].
3. Balakrishnan D. Implementation of Text and Voice Enabled Artificial Intelligence Chatter Bot. International Journal of Research in Computer Science, 2018, 5(2). [Page range if applicable].
4. Balakrishnan D, Gowri T. Protocol for Location Privacy Using K- Anonymity with Hla in Wireless Sensor Networks. International Journal of Science and Engineering Research, 2016, 4(1). [Page range if applicable].
5. Qureshi F. Dataset of spam email. Kaggle dataset, 2020.