Phishing Email Detection

Tanisha Mittal
Electronics and Communication
Institute of Technology
Nirma University
Email: 19bec132@nirmauni.ac.in

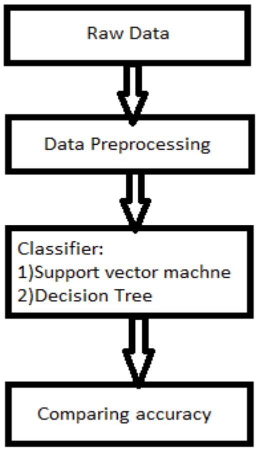
Yashasvi Patel
Electronics and Communication
Institute of Technology
Nirma University
Email: 19bec148@nirmauni.ac.in

***Abstract*—Phishing mails have become a very big problem these days. Frauds of very big magnitude are taking place because of these mails. This paper focuses on detecting such phishing emails based on different machine learning algorithms and the best algorithm is selected based on their accuracy.**

1. Introduction

Phishing emails are a type of scam emails that attempt to steal money by sending spam emails to targeted customers and appear to be coming from a well known source. For the detection of such emails techniques like support vector machine and decision tree is used. Before applying the algorithm the data is pre processed.

2. Flowchart

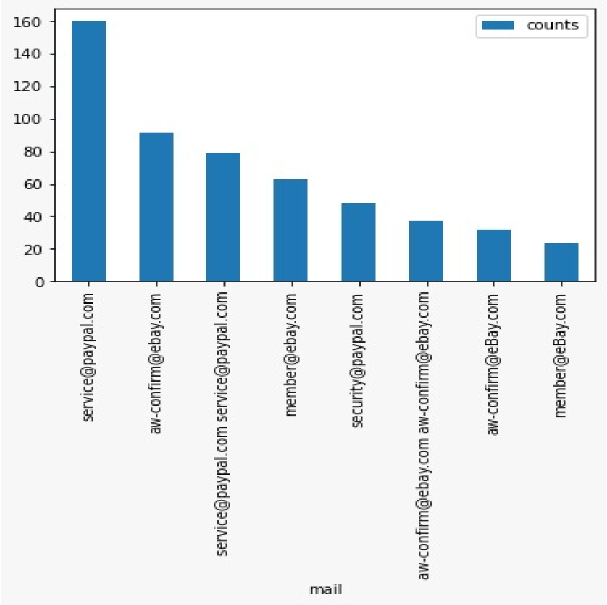


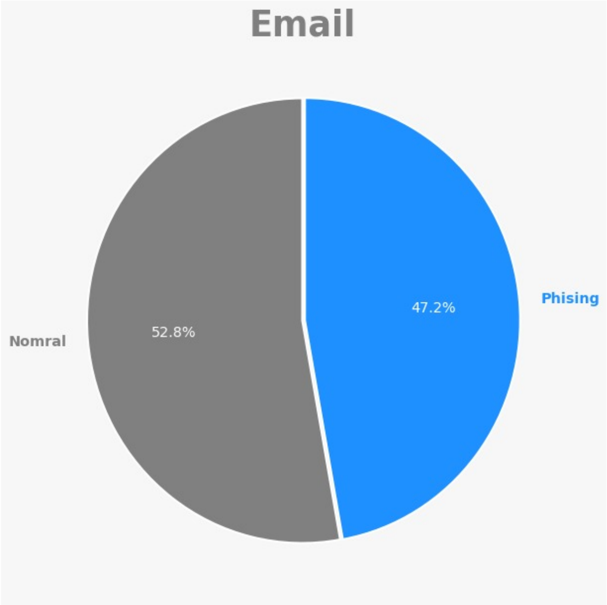
3. Working

3.1. Data Preprocessing

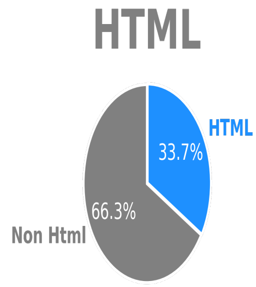
Data preprocessing is the procedure for preparing raw data for use in a machine learning model. It's the first and most important stage in building a machine learning model. It is not always the case that we come across clean and prepared data when working on a machine learning project. And, before doing any data-related activity, it is necessary to clean the data and format it. As a result, we use a data preprocessing activity for this. In this project :

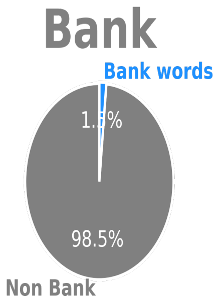
- All the mail bodies with only float values are converted to string by replacing with NAN
- All the stop words like is, the , i, you etc are removed
- Concaving phishing data and normal data

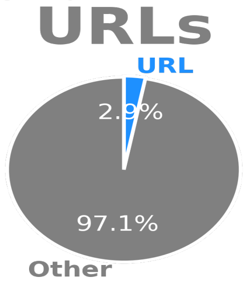




Common words like HTML tags, words related to banking and redirecting URLs were removed.

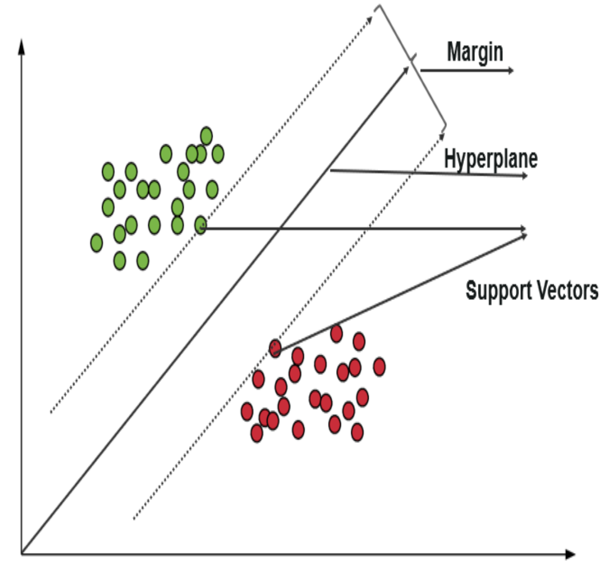




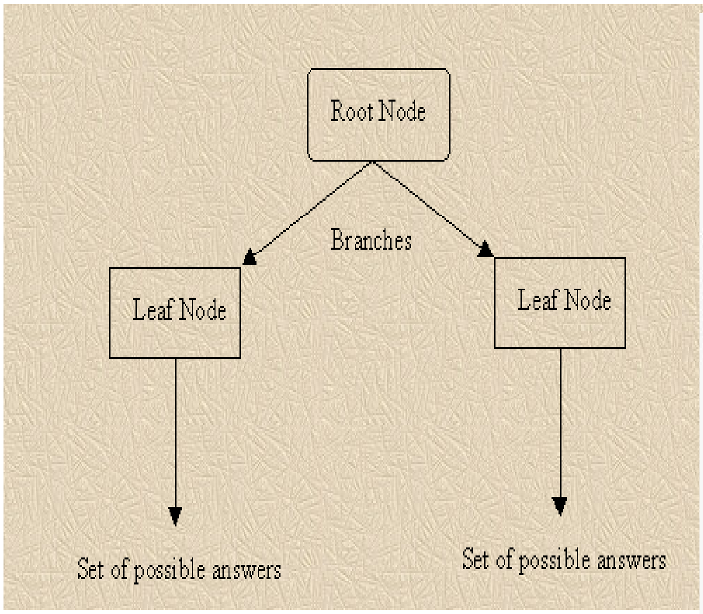


3.2. Model Training

3.2.1. Support Vector Machine. The Support Vector Machine, or SVM, is a popular Supervised Learning technique that may be used to solve both classification and regression issues. However, it is mostly utilised in Machine Learning for Classification difficulties. The SVM algorithm's purpose is to find the optimum line or decision boundary for categorising n-dimensional space into classes so that additional data points can be readily placed in the correct category in the future. A hyperplane is the name for the optimal choice boundary. The extreme points/vectors that assist create the hyperplane are chosen via SVM. Support vectors are the extreme instances, and the algorithm is called a Support Vector Machine.

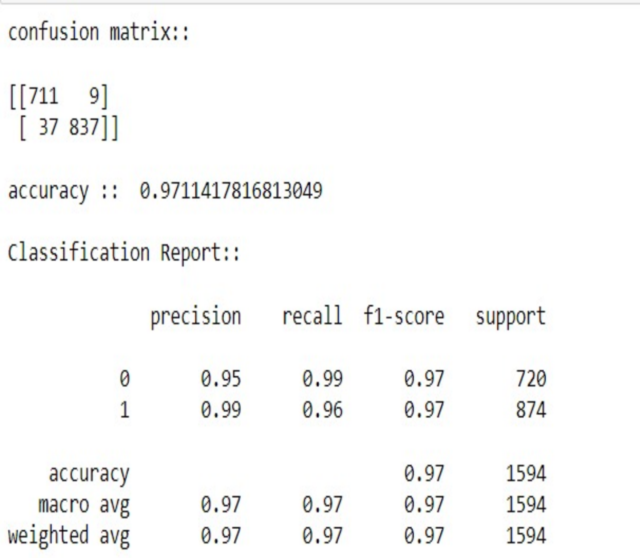


3.2.2. Decision Tree. Decision Tree is a supervised learning technique that may be used to solve both classification and regression problems, however it is most commonly employed to solve classification issues. Internal nodes represent dataset attributes, branches represent decision rules, and each leaf node provides the conclusion in this tree-structured classifier. The Decision Node and the Leaf Node are the two nodes of a Decision tree. Leaf nodes are the output of those decisions and do not contain any more branches, whereas Decision nodes are used to make any decision and have several branches. The decisions or tests are made based on the characteristics of the given dataset. It's a graphical depiction for obtaining all feasible solutions to a problem/decision depending on certain parameters. It's termed a decision tree because, like a tree, it starts with the root node and grows into a tree-like structure with additional branches. We utilise the CART algorithm, which stands for Classification and Regression Tree algorithm, to form a tree.



3.3. Model Testing and prediction

The data was then fit into different models like support vector machine and decision tree. After that their accuracy and precision of each model was compared.



```
confusion matrix::

[[711   9]
 [ 37 837]]

accuracy ::  0.9711417816813049

Classification Report::

              precision    recall  f1-score   support

           0       0.95      0.99      0.97       720
           1       0.99      0.96      0.97       874

    accuracy                           0.97      1594
   macro avg       0.97      0.97      0.97      1594
weighted avg       0.97      0.97      0.97      1594
```

Figure 1. Classification report for SVM

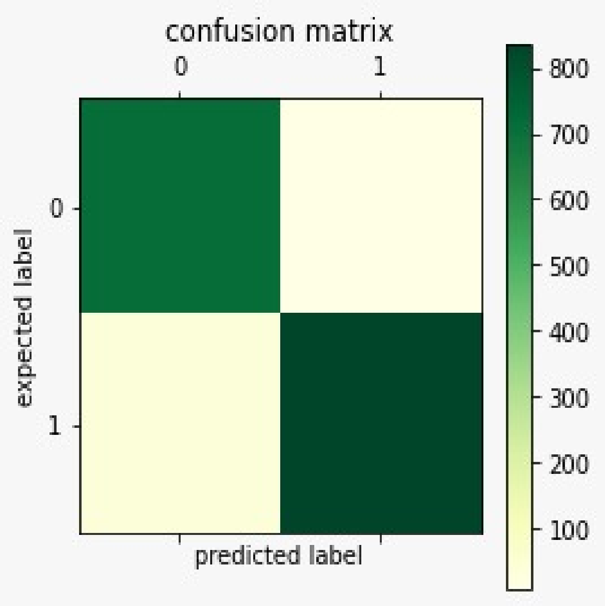
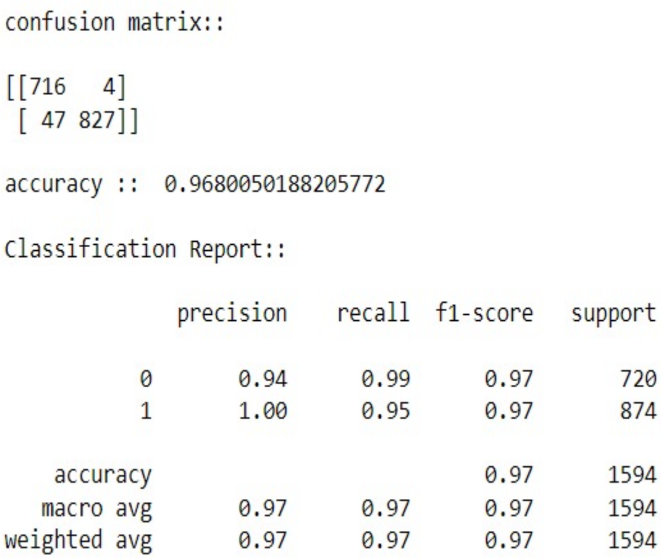
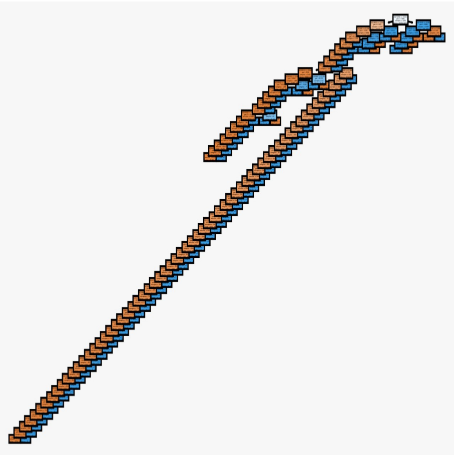


Figure 2. Confusion matrix for SVM



```
confusion matrix::

[[716   4]
 [ 47 827]]

accuracy ::  0.9680050188205772

Classification Report::

              precision    recall  f1-score   support

           0       0.94      0.99      0.97       720
           1       1.00      0.95      0.97       874

    accuracy                           0.97      1594
   macro avg       0.97      0.97      0.97      1594
weighted avg       0.97      0.97      0.97      1594
```

Figure 3. Classification report for Decision tree



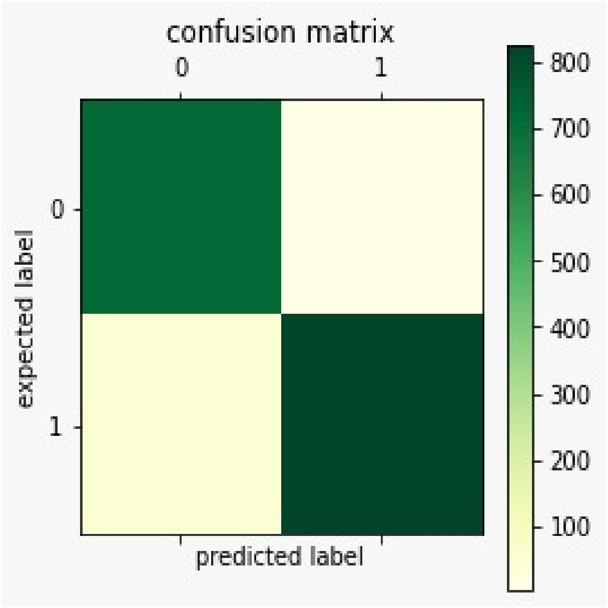


Figure 4. Confusion matrix for Decision Tree

4. Conclusion

After careful comparison between the two models, it can be concluded that the accuracy and precision for support vector machine is better.

Acknowledgments

This paper was performed under the guidance of Dr. Manish Patel and Dr. Ruchi Gajjar of Electronics and Communication Department,Institute of Technology Nirma University. The authors of the paper are thankful to the university and the professors for the guidance in completion of the term paper as a part of the teaching and learning process.

References

[1] Noor Ghazi M. Jamee , Loay E. George (2014), “Detection Phishing Emails Using Features Decisive Values”,257-259

[2] Madhuri Bhalekar,Rashmi Rane(2017)”Detecting spam and phishing mails using SVM and obfuscation URL detection algorithm”

[3] . Baykara and Z. Z. Gürel, ”Detection of phishing attacks,” 2018 6th International Symposium on Digital Forensic and Security (ISDFS), 2018, pp. 1-5, doi: 10.1109/ISDFS.2018.8355389.