Detection and Classification of Malicious Websites

Shubhankar[1], Siddhartha Bhaumik[2], Prakash Biswagar[3]

[1]Student, Dept of Electronics and Communication, R.V. College of Engineering,
Bangalore, INDIA -560059
[2]Student, Dept of Electronics and Communication, R.V. College of Engineering,
Bangalore, INDIA -560059

[3]Professor, Dept of Electronics and Communication, R.V. College of Engineering,
Bangalore, INDIA -560059

[1]shubhankarec.ec16@rvce.edu.in,[2]sidhharthab.ec16@rvce.edu.in

[3]prakashbiswagar@rvce.edu.in

Abstract: *Phishing is quite possibly the most appealing technique used by attackers in the point of taking the individual subtleties of unsuspected individuals. Phishing sites are essentially tricks which are used by data fraud hoodlums and fakes. They use spam, fake sites made to look like the first sites, email and direct messages to trick somebody into sharing significant information, like passwords and secret information. New enemies of phishing techniques are coming out each day, yet attackers think of new ways by focusing on all the new enemies of phishing techniques. So there is an earnest requirement for new strategies for the expectation of phishing sites. The paper portrays the correlation models in classification of phishing sites for expectation utilizing distinctive Machine learning models. Different models are used for predicting which model gives the best exactness in phishing sites classification. All the information is classified as either Benign for substantial Websites or Phish as Phishing Websites. Results are generated that show RF gives the best performance on this dataset for classification of phishing sites.*

Keywords: **Machine Learning,Phishing Websites,Spam**

1. INTRODUCTION

The presence of advances has shown a transcending impact in the turn of events and progression of associations, navigating transverse correspondences over various applications on the web going from running little e-organizations, web based banking, and individual to individual correspondence. Truth being, in the present age it is moderately mandatory to have an online presence to run the assistance which can keep up with the high demands of the clients, give the cus-tomers solid substance and keep an advanced web presence. In this way, the meaning of the World Wide Web(WWW) has been perseveringly expanding. Unfortunately, the movements in innovation come joined with complex strategies created over the a long time by attackers that snare the dumbfounded customers into getting what the attacker needs them to do.The uninformed client gets presented to different class of attacks where he may be diverted to a fake website page which mirrors the genuine site in looks or made to download a document from a source appearing completely authentic however contains a completely imperceptible vindictive connection. These strategies depicted are normally known as phishing and drive by download attacks separately. Experienced attackers have a skill set to execute attacks including various methods going from phishing, man in the center attacks and numerous more.Such attacks fuse destinations that trick clients to gather their touchy and secret information like passwords, Mastercard subtleties and so on at last prompting character thefts,bank cheats and in some different cases presenting malware in the customer's framework. The one thing basic to by far most attacks is redirection of the objective to the ideal page by the attacker.

2. Literature Review

The advancement in technology has given the internet platform for different levels of illegal activities starting from spam to financial frauds. These activities are carried out by embedding malware programs in these websites and URLs. Blacklisting servers can be used but the creation of newer websites pose a challenge. In the paper [1] algorithms such as Apriori, FPGrowth and Decision Tree Rules which are used to generate and establish relationships between the features.

The detection of any category of URLs or Websites are done through the method of blacklists. But blacklists are extensive and they cannot be updated for the newly generated malicious URLs or Websites. The writing [2] gives us the idea of complete understanding of Malicious URLs with help of machine learning models.

Large scale streaming URLs are generally used in with many commercial products that remove URLs and websites based on the risk level. In those cases, its filtering is done by classifying a web page content or its contextual information. The authors in the paper [3] show the effectiveness of the given approach with large scale data sets.

The amount of Malicious Websites have increased largely around us. It shows the increasing trends in attack vectors through Malicious Websites, which tells us the need to start building prediction models. Much research has been done to compare Machine Learning algorithms for detection. Here, the attribute selection is more important than any other aspects. So, there is need to compare and analyze the various attributes for different Malicious websites faster and better. The paper [5] mainly focuses on how fewer and smaller attributes can do a better job.

The paper [5] describes the accuracy of a classifier or predictor which is normally generated through the help of a confusion matrix, which in turn is used in estimating how a classifier recognizes tuples of separate classes. The calculation of the classification accuracy with a confusion matrix is simple for different attributes is done .

3. Machine Learning Models

1. Logistic Regression

Logistic Regression is a learning algorithm which predicts the probability of a given variable.The given variable which is dependent and is bilateral,which tells us that there are only two classes.In straightforward words, the dependent variable has binary data which is either 1 (success) or 0 (failure). A Logistic Regression model gives P(Y=1) as a function of X. Logistic Regression is one of the basic algorithms used in spam detection,Diabetes prediction and cancer prediction etc.

2. K-Nearest Neighbours

K-Nearest Neighbours is a type of Machine Learning Algorithm which is used in classification as well as regression for prediction models.Generally, it is used in prediction models. The given two properties define K-Nearest Neighbour as - Lazy learning algorithm:- K-Nearest Neighbours is a lazy learning algorithm as it does not have a training phase and uses the data during classification. Non-parametric learning algorithm:- K-nearest neighbors is a Non-parametric learning algorithm because it doesn’t assume anything about the data.

3. Classification and Regression Tree

A Classification and Regression Tree(CART) is a prediction algorithm used for machine learning models.It shows how a target dependent variable is predicted based on the different values.It is a decision tree where each fork is split in a predictor variable and each node at the end has a prediction for the target variable.

4. Support Vector Machine

Support vector machine is a useful and important machine learning algorithm used for classification and regression. But they are generally used for the classification models.Support Vector Machines implement the models differently as compared to other algorithms.They have been extremely made use of because of their ability to handle multiple continuous and categorical variables.

5. Decision Tree

Decision Trees are a type of Machine Learning algorithm where we take what the input is and then undertake the corresponding output data in which the data is split with a parameter.The decision tree is given with two parts: nodes and leaves. The leaves are the given decisions with the final outcomes.And with the decision nodes the data are split.An example of a decision tree can be given by a binary tree.Let us take that we want to predict given whether is fit with data like age,eating habit,physical activity etc.

6. Random Forest Classifier

Random Forest Classifiers are a type of Machine Learning Algorithm which is used in both Classification and Regression. As we have been told that, a forest is made up of trees and more trees make it more robust. So, a Random Forest Algorithm has decision trees for data samples and then from each of them prediction is done and the best solution is taken.It is a method which is better than the decision tree method because it prevents overfitting while averaging its result.

7. XGBoost Classifier

XGBoost is a decision-tree-based ensemble that has a gradient boost framework. In prediction models with different unstructured data artificial neural network does better than all the other algorithms. But for small and medium structured tabular data, decision trees are the best options.

4. Methodology

In the topic , the method needed for training of the model is given in which the seven ML are used to get the most efficient model for classification for Benign and Malicious Websites. The beginning stage and the data set are something very similar for every one of the models to contrast them and accuracy, Precision and Recall of the models.

The most popular models at that point are taken in for prediction where the models are isolated and then training is accomplished for every one of the models for the various steps. The preparation is prevailing by testing where the best chosen model gives us the most ideal result.

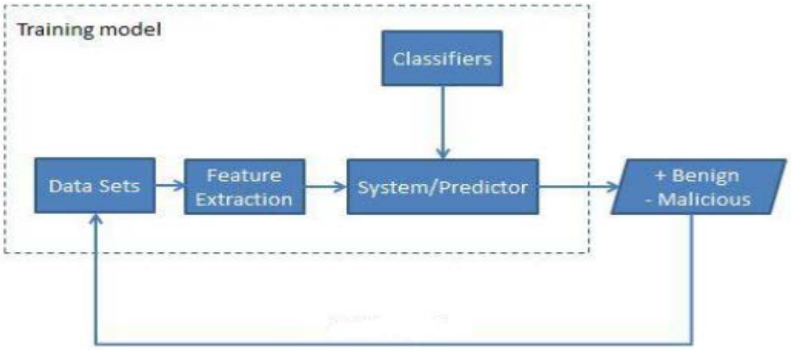


Figure1 Proposed Architecture

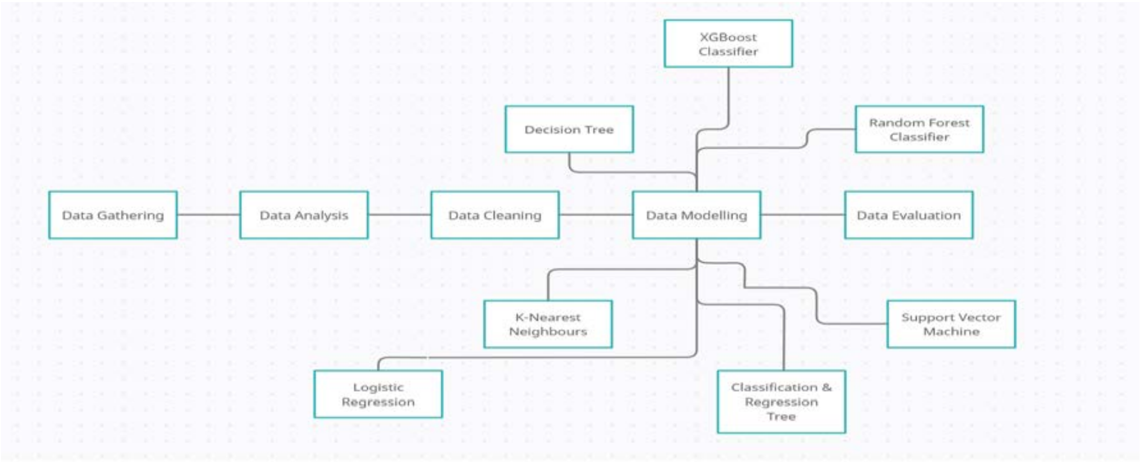


Figure 2 Procedure involved in training and testing the models for the best accuracy

Figure 2 shows the block diagram of steps in the procedure used for testing and training of all the different models which also gives out the best predicted techniques for prediction and classification.

The process has the training of all the different models used in the following procedure where Data Gathering, Data Analysis ,Data Cleaning ,Data Modelling and Data Evaluation is performed to find the best algorithm to start the procedure.

5. Results and Discussion

A.Attribute Selection

Feature and Attribute Selection is done through with the help of Decision Tree where the Tree branches give us the steps to undergo while doing Feature Extraction to provide us with the attributes required in further analysis in Prediction and Classification.

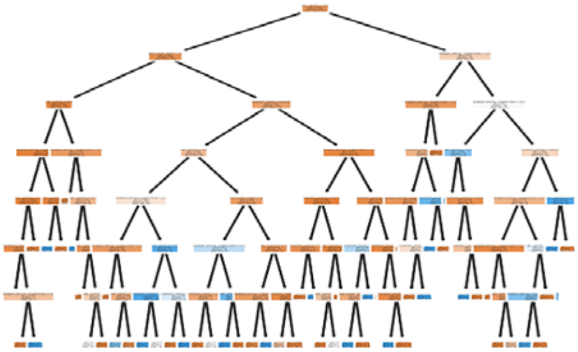


Figure 3 Decision Tree For Attribute Selection

B.Learning Rate

The given data set is prepared under a neural network where the training is done in various stages to diminish the Training loss and Validation Loss which thus expands the efficiency of the model. The neural network prepares the model where the epochs which are the quantity of passes of the training data set that the machine learning algorithm has completed.The level of epochs then furnish with different degree of Validation Loss ,and training loss with its precision and what amount of time each of the epochs require to wrap up.

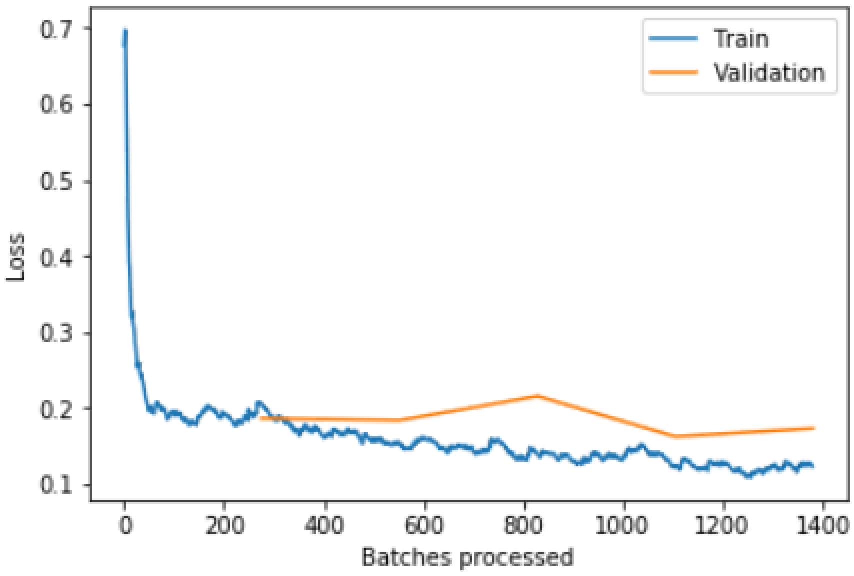
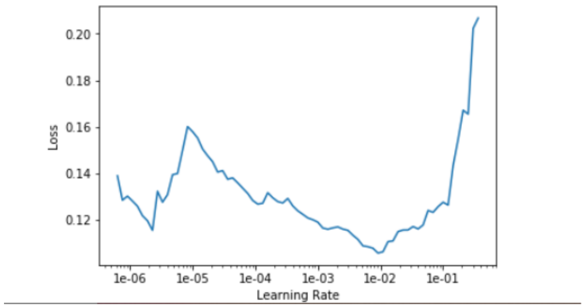


Figure 4 The Training and validation Loss for the Dataset at the start

Epochs	Training Loss	Validation Loss	Accuracy	Time (s)
0	0.112382	0.185257	0.923112	00:04
1	0.100710	0.180946	0.921755	00:04
2	0.103667	0.166621	0.927182	00:04
3	0.089619	0.152128	0.936680	00:04
4	0.088446	0.172709	0.933514	00:04

Table 1 Loss and validation at the start of training



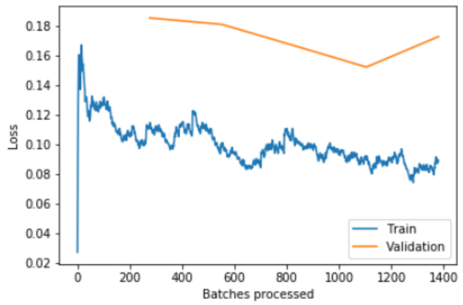


Figure 5 The learning rate in the model

Figure 6 Change of loss in the model

Epochs	Training Loss	Validation Loss	Accuracy	Times(S)
0	0.070832	0.142825	0.938942	00:02
1	0.067276	0.136672	0.040298	00:02
2	0.071374	0.139362	0.941203	00:02
3	0.068225	0.142303	0.939846	00:02
4	0.069698	0.137403	0.934871	00:02
5	0.061990	0.138733	0.938489	00:02
6	0.063320	0.143927	0.939394	00:02
7	0.064499	0.142344	0.942108	00:02
8	0.056307	0.144487	0.940298	00:02
9	0.062314	0.148076	0.943108	00:02

Table 2 Training Loss and Validation loss after improving the learning rate of the model

The Machine Learning algorithms are trained and tested on the set of predefined dataset to find the best accuracy for classification and prediction.After the process, all the Machine learning algorithms are compared and the one with best accuracy and a second model for comparison is considered for the given dataset.

Figure 7 shows the diagram of comparison of the ML algorithms through a box plot where the best model is taken into consideration for prediction and classification. The box plots basically depicts group of numeral values through their quartiles.Box plots have lines extending which indicates upper and lower quartiles for their variability.

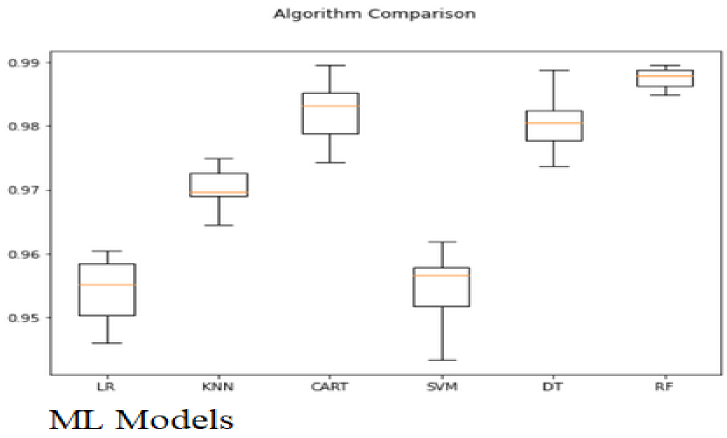


Figure 7 Accuracy of ML algorithms through a box plot

C.Confusion Matrix Realization

For Machine Learning and Statistical classification, the confusion matrix is also inferred as error matrix.It is also known as a specific table design which gives the visualization of the execution of an algorithm ,which is generally a learning algorithm under supervision.It describes the data of a classification model on a data set in which the true values are given.

Confusion Matrix Visualization for Types Of Malicious URL:- The Confusion matrix here infers about types of malicious URLs present in the data set. The type of Malicious URLs can be also given as defacement URLs.

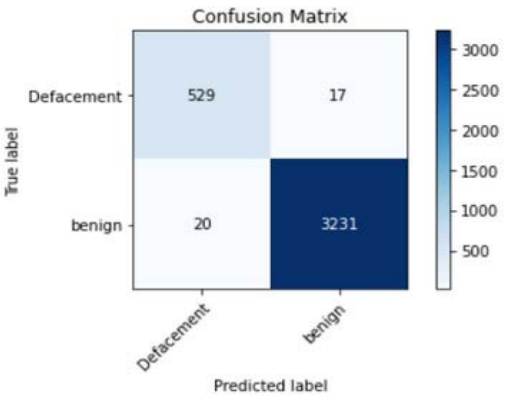


Figure 8 Confusion Matrix for types of Malicious Websites

Confusion Matrix Visualization For Benign and Malicious URLs:- The Confusion matrix infers about the classification between the benign and Malicious URLs present. The value of the Benign

URL and the Malicious URL can be predicted through the True Label and the Predicted Label present.

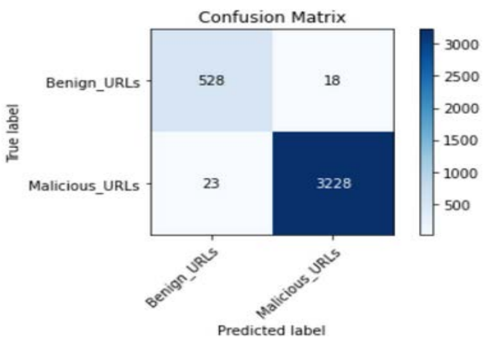


Figure 9 Confusion Matrix for Benign and Malicious Websites

We need to generate the accuracy,Precision,Recall and F1-score of the best scenario algorithm for our model.

For Accuracy:(TP+TN)/(TP+TN+FP+FN)

For Precision:TP/(TP+FP)

For Recall:TP/(TP+FN)

For F1-score/F-Measure:(2*Recall*Precision)/(Recall+Precision)

Performance Metric	Score
Accuracy	98.9%~99%
Precision	96.7%~97%
Recall	95.8%~96%
F1-score/F-Measure	96.24%

Table 3 Performance Metric Score for the best case algorithm

D. Performance Metric Score With Distribution Plots

The Performance metric score is found in the best model which is the most accurate in prediction and classification for the data set. The best model is then compared with a second model to confer about how the best model is then the better choice for prediction and classification in the Machine Learning Model.

1. Case 1:

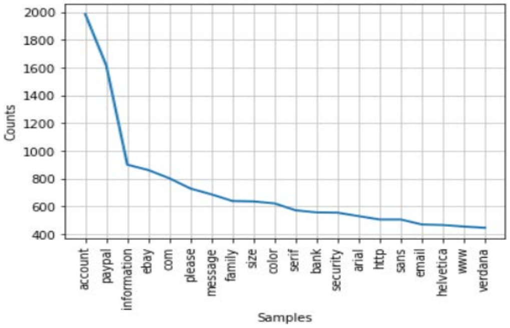
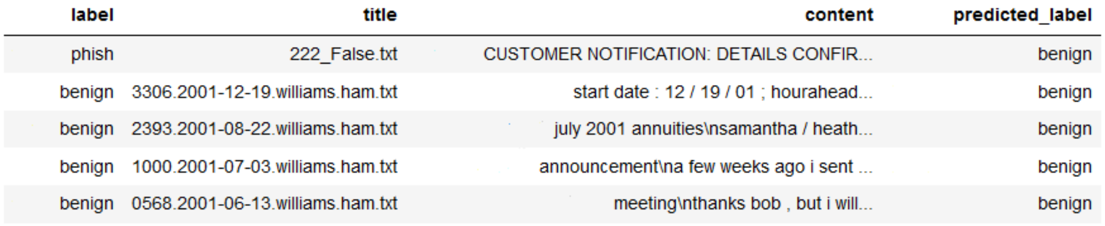
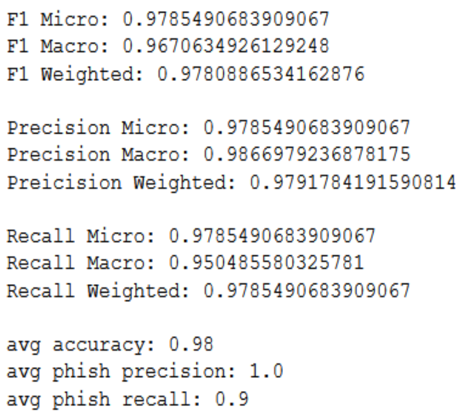


Figure9 Data Vs Contextual Information



label	title	content	predicted_label
phish	222_False.txt	CUSTOMER NOTIFICATION: DETAILS CONFIR...	benign
benign	3306.2001-12-19.williams.ham.txt	start date : 12 / 19 / 01 ; hourahead...	benign
benign	2393.2001-08-22.williams.ham.txt	july 2001 annuities\nsamantha / heath...	benign
benign	1000.2001-07-03.williams.ham.txt	announcement\na few weeks ago i sent ...	benign
benign	0568.2001-06-13.williams.ham.txt	meeting\nthanks bob , but i will...	benign

Figure 10 Prediction and Classification process to determine whether Website is benign or malicious



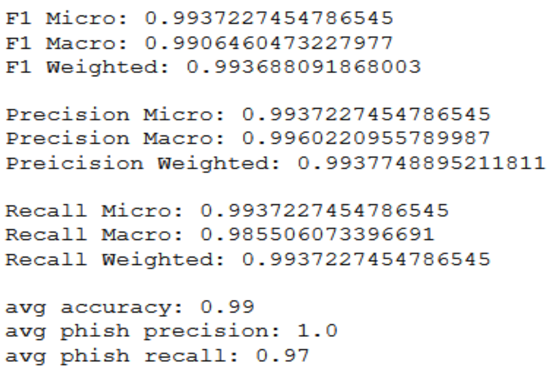
```
F1 Micro: 0.9785490683909067
F1 Macro: 0.9670634926129248
F1 Weighted: 0.9780886534162876

Precision Micro: 0.9785490683909067
Precision Macro: 0.9866979236878175
Preicision Weighted: 0.9791784191590814

Recall Micro: 0.9785490683909067
Recall Macro: 0.950485580325781
Recall Weighted: 0.9785490683909067

avg accuracy: 0.98
avg phish precision: 1.0
avg phish recall: 0.9
```

Figure 11 RF Classifier performance metric score for Phishing attack



```
F1 Micro: 0.9937227454786545
F1 Macro: 0.9906460473227977
F1 Weighted: 0.993688091868003

Precision Micro: 0.9937227454786545
Precision Macro: 0.9960220955789987
Preicision Weighted: 0.9937748895211811

Recall Micro: 0.9937227454786545
Recall Macro: 0.985506073396691
Recall Weighted: 0.9937227454786545

avg accuracy: 0.99
avg phish precision: 1.0
avg phish recall: 0.97
```

Figure 12 SVM performance metric score for Phishing attack

The Performance metric score is found for the best machine learning model which gives us the most accurate prediction and classification for the available data set. The machine learning model is compared with a second ML model to generate how the best accurate model is then the best for prediction and classification in the Machine Learning Model.

2. Case 2:

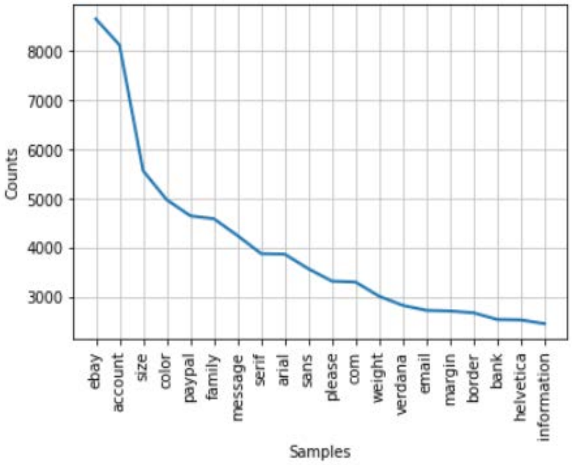
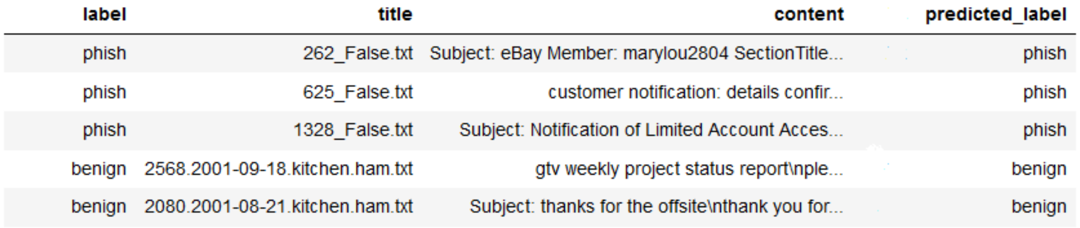
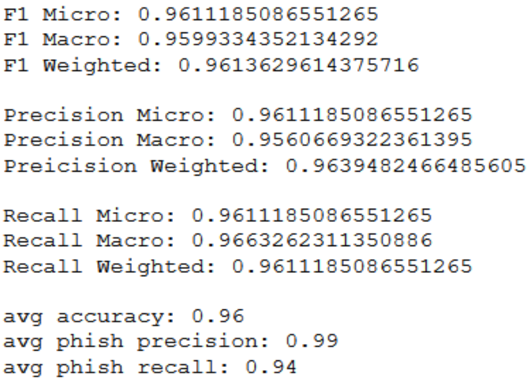


Figure 13 Data Vs Contextual Information



label	title	content	predicted_label
phish	262_False.txt	Subject: eBay Member: marylou2804 SectionTitle...	phish
phish	625_False.txt	customer notification: details confir...	phish
phish	1328_False.txt	Subject: Notification of Limited Account Acces...	phish
benign	2568.2001-09-18.kitchen.ham.txt	gtv weekly project status report\nple...	benign
benign	2080.2001-08-21.kitchen.ham.txt	Subject: thanks for the offsite\nthank you for...	benign

Figure 14 Prediction and Classification process to determine whether Website is benign or malicious (Case 2)



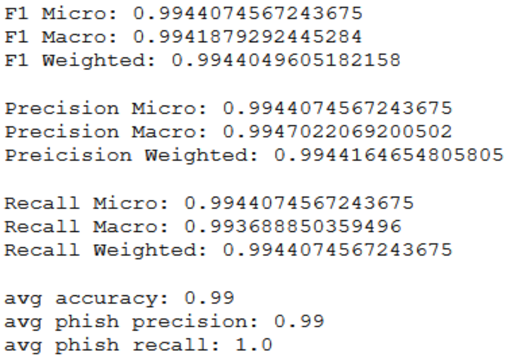
```
F1 Micro: 0.9611185086551265
F1 Macro: 0.9599334352134292
F1 Weighted: 0.9613629614375716

Precision Micro: 0.9611185086551265
Precision Macro: 0.9560669322361395
Preicision Weighted: 0.9639482466485605

Recall Micro: 0.9611185086551265
Recall Macro: 0.9663262311350886
Recall Weighted: 0.9611185086551265

avg accuracy: 0.96
avg phish precision: 0.99
avg phish recall: 0.94
```

Figure 15 RF Classifier performance metric score for Phishing attack (Case 2)



```
F1 Micro: 0.9944074567243675
F1 Macro: 0.9941879292445284
F1 Weighted: 0.9944049605182158

Precision Micro: 0.9944074567243675
Precision Macro: 0.9947022069200502
Preicision Weighted: 0.9944164654805805

Recall Micro: 0.9944074567243675
Recall Macro: 0.993688850359496
Recall Weighted: 0.9944074567243675

avg accuracy: 0.99
avg phish precision: 0.99
avg phish recall: 1.0
```

Figure 16 SVM performance metric score for Phishing attack

The above data shows the performance of the model for the given case where the Number of Data with contextual information is given through a plot for the different amount of information present for the data set.The model is trained and tested for the prediction and classification to work, where it checks whether the True Label and predicted label is same for the algorithm, The data set is trained with the best accurate model comparing it with a second model to infer about the best algorithm needed for a certain data set.

6. Conclusion

In the work,we have taken a distinctive machine learning algorithm that will give a superior outcome by utilizing the given data and list of capabilities. The data sets are prepared and afterward the best model is found with better precision and afterward is utilized in foreseeing malicious and benign websites in the accessible dataset. The test and train exactness is contemplated when contrasting and taking the best thought about algorithms for future work.

The Future work is to adjust the machine learning algorithm that will deliver the better outcome by using the given feature set. Adding to that the open inquiry is the ticket we can deal with the enormous number of websites whose features set will advance over time.Certain endeavors must be made in that course in order to concoct the more vigorous feature set which can affect the developing changes.

7. Conclusion

1. A. S. Manjeri, K. R., A. M.N.V. and P. C. Nair, ”A Machine Learning Approach for Detecting Malicious Websites using URL Features,” 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 555-561, doi: 10.1109/ICECA.2019.8821879
2. Sahoo, Doyen Liu, Chenghao Hoi, Steven. (2017). Malicious URL Detection using Machine Learning: A Survey.
3. N. Singh, N. S. Chaudhari and N. Singh, ”Online URL Classification for Large-Scale Streaming Environments,” in IEEE Intelligent Systems, vol. 32, no. 2, pp. 31-36, Mar.-Apr. 2019. doi: 10.1109/MIS.2019.39
4. A. Singh, N. Goyal, A comparison of machine learning attributes for detecting malicious websites, in 2019 11th International Conference on Communication Systems Networks (COMSNETS) (IEEE, 2019), pp. 352–358
5. V.M. Patro, M.R. Patra, Augmenting Weighted Average with Confusion matrix to Enhance classification accuracy. Trans. Mach,. Learn. Artif Intell. 2(4),77-91(2019)
6. H. Kumar, P. Gupta and R. P. Mahapatra, ”Protocol based Ensemble Classifier for Malicious URL Detection,”2018 3rd International Con- ference on Contemporary Computing and Informatics (IC3I), Gurgaon, India, 2018, pp. 331-336, doi: 10.1109/IC3I44769.2018.9007255.

7. J. Yoon, W. R. Zame and M. van der Schaar, ”ToPs: Ensemble Learning With Trees of Predictors,” in IEEE Transactions on Sig- nal Processing, vol. 66, no. 8, pp. 2141-2152, April15, 15 2018. doi:10.1109/TSP.2018.2807402
8. Nebali, Raj Wang, Yung Alshboul, Yazan. (2015). Detecting malicious short URLs on Twitter.