JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Phishing Website Prediction using Machine Learning

S. Irin sherly [1], Renuka shri [2], Swethashree [3], Yashika [4]

[1]*Associate Professor, Department of Information Technology, Panimalar Institute of Technology, Chennai, India.*

[2,3,4]*UG Scholars, Department of Information Technology, Panimalar Institute of Technology, Chennai, India*

Abstract— The Internet has forge to be essential part of our life. It also gives openings to anonymously performed mischievous activities like phishing. Phishers steal information like personal detail, account information by creating fake website. Although numerous methods have been proposed to find phishing website, Phishers evolved different procedure to escape from these detection methods. One of the most successful techniques for detecting these mischievous activities is by using machine learning approaches. The aim of this paper is to predict the URL data whether it is legitimate, suspicious or phishing website. We also compared the results of multiple machine learning Algorithms like Logistic Regression, Random Forest Classifier, Decision Tree Classifier, Navies Bayes for finding the best accuracy. This work can be enhanced to implement in AI environment and connect to cloud.

Keywords— Phishing Website, Machine Learning, Phishing attack, Detection, Fake Website.

I. Introduction

Phishing becomes main area of concern for Security because it is not delicate to design the fake website which look so close to legal website. Internet has various services available like e-mail, online shopping, multimedia, etc. In which it has more possible way to have both legitimate and phishing web application. Cyber Threat Intelligence report that the phishing attack have been increased in past few years. For preprocessing the time series problems, TNN (Traditional neural network) and RNN (Recent Neural Network) was used [4]. Prediction of phishing website was carried out by CNN-LSTM algorithm and used web2vec feature to automatically include corpus content [6]. MFPD (Multidimensional feature Phishing detection) includes all features together and differentiate it by XG-Boost [3]. Online learning is difficult unless it is outlined for quick training a data As Phishing corpus increases it is difficult to train a classifier [9]. The capsule network and bidirectional recurrent network are used to detect the spam website with high dimensional feature to classify the model. The web2vec approach is used to train the URL content and features [15]. The Heuristics approach reduces false positives where the legitimate website is labelling as phishing website. Traditional phishing webpage prediction methodologies are substantially on the analysis of multi-source features [7].

II. Related Work

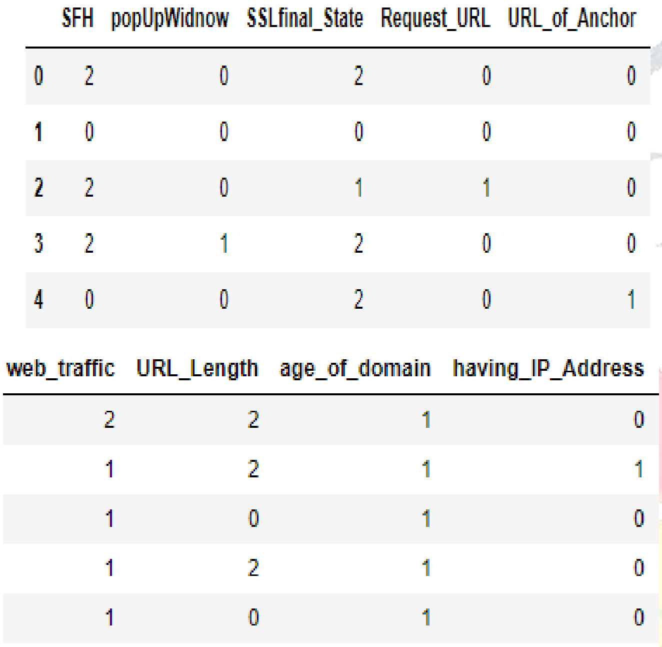
M.A Adebowale, et.al [5] have proposed the concept of hybrid LSTM and CNN method in deep learning (i.e Intelligent Phishing Detection System (IPDS)) in which using image and text content they detect the phishing website. Peng Yang, et.al [3] have proposed the concept of Multidimensional feature phishing detection (MFPD) approach in deep learning in which they combine all feature into multidimensional feature and it will be classified by XG-Boost to find the phishing website. S. Sheng, et.al [7] they measure the Quantitative length of phishing blacklist URL and tested it with Anti phishing testbed in which it has Client -Server architecture. To evaluate the Anti phishing testbed, they used eight toolbar which is used to find various blacklist and heuristic data. Jian Feng, et.al [6] have proposed web2vec to extract webpage features automatically in multiple-aspects.They use CNN and Bi-directional LSTM to extract local features and to obtain content semantics respectively and also it classifies the differnence between phishing and bengin webpages. finally it predicts by classifier algorithm. C. N. Gutierrez et al [9] have proposed a methodology called SAFE-PC for detecting new kind of phishing campaigns in legitim ate email dataset. The National Language Processing (NLP) is used to detect phishing or spam in the actual dataset. P. Prakash, et.al [2] have described the architecture of Phish net and to evaluate the URL prediction component by generating new URL and validating the spam by matching host-name of blacklist dataset. A. Aljofey, et.al [12] have proposed deep learning-based solution for URL detection and they used stacked non- linear projection to represent multiple level of abstraction. Convolutional neural Network (CNN) applies URL string as input and evaluates the sequence of character that are converted into matrix representation and finally generates the output of phishing website or not. Jian Ting Yuan, et.al [15] have proposed Joint Neural Network Algorithm Model for combining two Methods (i.e

Bidirectional independent recurrent network and capsule network) to identify the spam URL with high dimensional features to classify the webpage.

In our study, we have used phishing website network dataset which contains 10 attributes. Data Preprocessing and cleaning is done to remove missing values and noisy data. Machine learning classifiers such as Logistic Regression, Decision Tree, Random Forest and Naive Bayes are used to train and test the model.

III.DATA DESCRIPTION

Phishing Website Network Dataset are used in this study in which features take numerical values of -1, 0, 1 is shown in Figure 1. This dataset contains 10 attributes and 1353 records which has target values of legitimate, suspicious and phishing.



	SFH	popUpWidnow	SSLfinal_State	Request_URL	URL_of_Anchor	web_traffic	URL_Length	age_of_domain	having_IP_Address
0	2	0	2	0	0	2	2	1	0
1	0	0	0	0	0	1	2	1	1
2	2	0	1	1	0	1	0	1	0
3	2	1	2	0	0	1	2	1	0
4	0	0	2	0	1	1	0	1	0

Fig. 1. Phishing Website Network Dataset Sample

IV. METHODOLOGY

A.Working

The proposed idea is to build a machine learning model for anomaly detection in phishing website as shown in figure 2. Anomaly detection is a technique in which it finds outliner in the dataset. The Process of finding out the abnormal behavior in the dataset in which it predicts the suspicious values of rest of normal observation The machine learning model is built by applying proper data science techniques like variable identification that is the dependent and independent variables. Then the visualization of the data is done to insights of the data. The model is build based on the previous dataset where the algorithm learn data and get trained different algorithms are used for better comparisons. The dataset is processed and cleaned to eradicated duplicate, missing values after that the 70% of dataset will be trained and 30% of data will be tested. A trained dataset is further compared with different machine learning algorithm of supervised learning model in which it finds the accuracy of each model and generates the best accuracy model as Output. The user interface display whether the webpage is phishing or not. The machine learning model is built by applying proper data science techniques like variable identification (i.e dependent and independent variable).

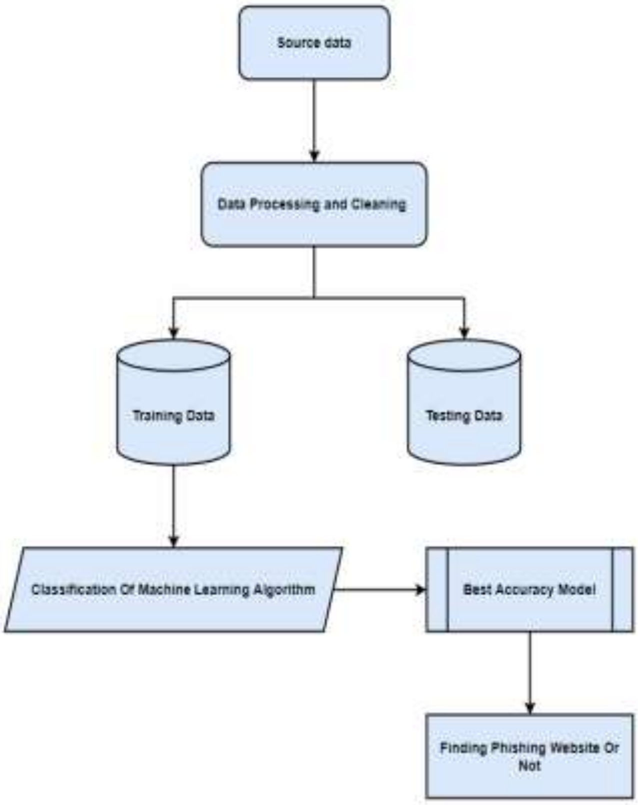

Source data
Data Processing and Cleaning
Training Data
Testing Data
Classification Of Machine Learning Algorithm
Best Accuracy Model
Finding Phishing Website Or Not

Fig.2. Process flow of the working model

The visualization of the data is done to insights of the data. Different machine learning algorithms are used for better comparisons. The performance metrics are calculated and compared. The anomaly detection is done as an automated process using the machine learning classifiers. The accuracy level of machine learning algorithm model is calculated and compared in order to get a better model.

B. Proposed model

In the system architecture shown in figure 3, the user gives input through the user interface where the inputs are checked with the past dataset. It undergoes pre-processing technique where it converts raw data into clean dataset and also validates the data which finds duplicate value and missing value. It also finds the error rate of machine learning model. After this process it visualizes the data in plots and graphs which demonstrate the key relationship of the data. It is helpful to explore and learn a dataset and also identifies the pattern, duplicate values, false data, layouts, etc. Next it compares the performance of multiple machine learning algorithms in which we can get accuracy value using different algorithms. It then undergoes accuracy selection process in which we can find the best accuracy of machine learning algorithm. Next using flask, the user interface will show whether the website is phishing website or not.

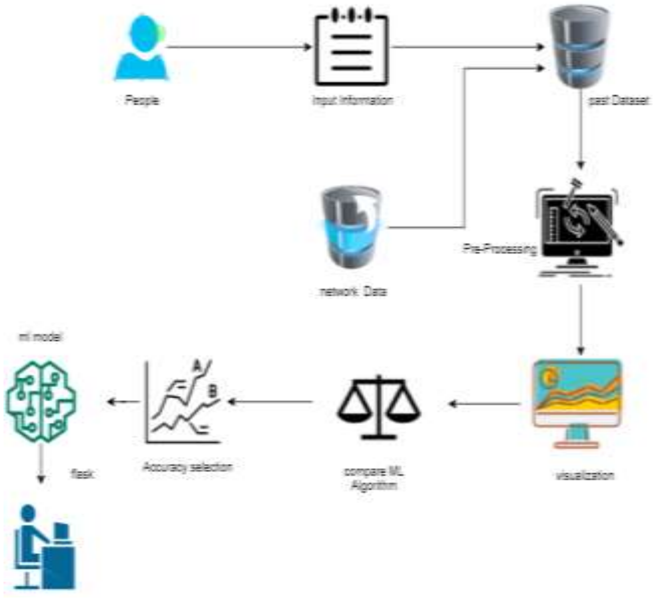


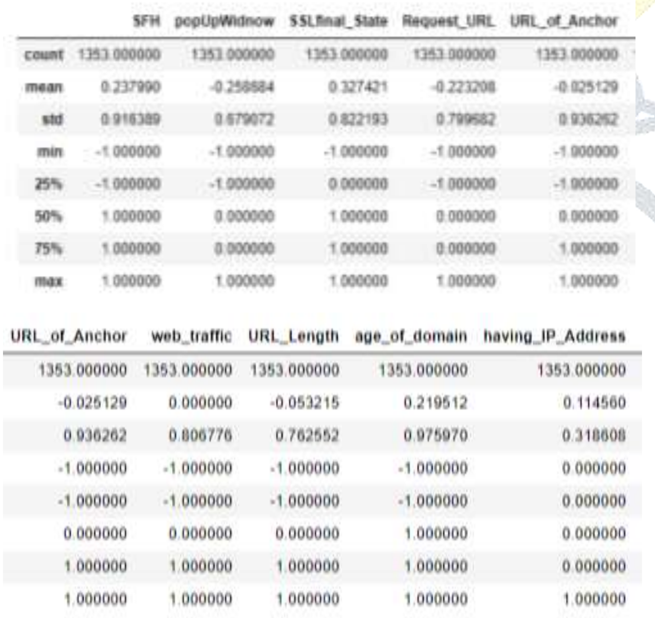
Fig. 3. System Architecture Model

VI.EXPERIMENTAL DISCUSSION

The main motive of our system is to find the best accuracy by comparing the different algorithms. It is necessary to compare the performance metrics of different machine learning algorithms constantly and it produce a analyse harness to compare multiple different machine learning algorithms in Python with scikit- learn. Each model will have dissimilar performance characteristics, calculating matrix we can get the value of precision, recall, F- measure, F1 Score, so that we can get accurate performance metrics.

6.1 DATA PREPROCESSING

Data Pre-Processing is a methodology to process raw facts into a clean data. For better accomplishment of results in machine learning, the data are processed in applied with pre-processing techniques. For instance, Random Forest algorithm does not support null values, so it executes all null value and controls the original data.



	SFH	popUpWidnow	SSLfinal_State	Request_URL	URL_of_Anchor
count	1353.000000	1353.000000	1353.000000	1353.000000	1353.000000
mean	0.237990	-0.258684	0.327421	-0.223208	-0.025129
std	0.916389	0.679072	0.822193	0.799682	0.936262
min	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
25%	-1.000000	-1.000000	0.000000	-1.000000	-1.000000
50%	1.000000	0.000000	1.000000	0.000000	0.000000
75%	1.000000	0.000000	1.000000	0.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000

URL_of_Anchor	web_traffic	URL_Length	age_of_domain	having_IP_Address
1353.000000	1353.000000	1353.000000	1353.000000	1353.000000
-0.025129	0.000000	-0.053215	0.219512	0.114560
0.936262	0.806776	0.762552	0.975970	0.318608
-1.000000	-1.000000	-1.000000	-1.000000	0.000000
-1.000000	-1.000000	-1.000000	-1.000000	0.000000
0.000000	0.000000	0.000000	1.000000	0.000000
1.000000	1.000000	1.000000	1.000000	0.000000
1.000000	1.000000	1.000000	1.000000	1.000000

Fig. 4. Processed dataset of phishing website

6.2 DATA VISUALIZATION

Data visualization is a significant technique applied in statistics and machine learning. Data visualization is key factor of gaining a contextual learning. It can be used to express and demonstrate vital connections in plots and maps.

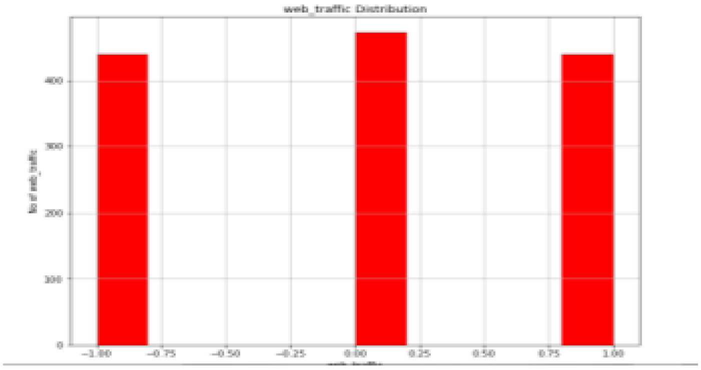


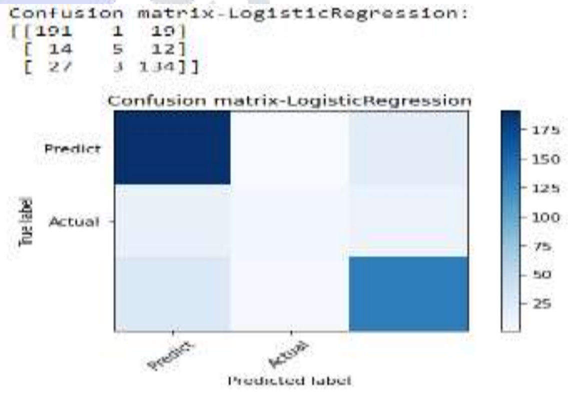
Fig.5. Web Traffic Distribution

6.3 COMPARING ALGORITHM

6.3.1 LOGISTIC REGRESSION

Logistic Regression provides the ability to find and classify n-number of independent variables. The dichotomous variable is measured as the result with only two outcomes of possibility. The aim of logistic regression is to find the appropriate model to describe the relationship between the dichotomous characteristic of interest. In Logistic regression, the machine learning algorithm are used to predict the probability of dependent variable(categorial). The dependent variable is binary value (i.e. It holds only 0 and 1), 0 indicates yes, success.1 indicate

No, Failure.

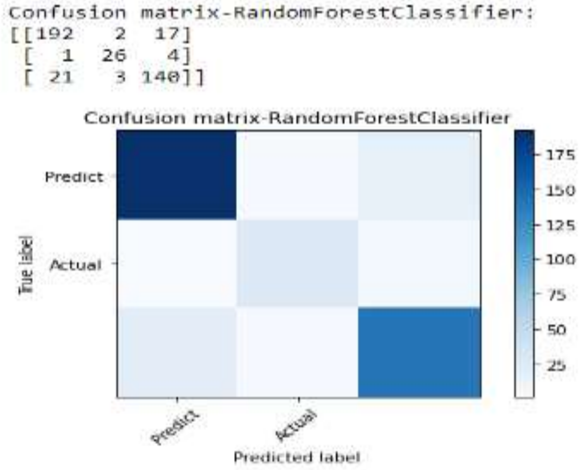


```
Confusion matrix-LogisticRegression:
[[191    1   19]
 [ 14    5   12]
 [ 27    3  134]]
```

Fig. 6. Confusion matrix of Logistic Regression

6.3.2 RANDOM FOREST CLASSIFIER

The output is that the mode of class (Classification) or regression (mean prediction) of each tree. Random decision forests over- fit the training set decision tree which is correct to the subset of classes. The output is that the mode of class (Classification) or regression (mean prediction) of each tree. The Random Forest Algorithm can combine multiple or different algorithm of the same type i.e. It results in multiple decision tree of forest, hence the name "Random Forest". Both regression and classification tasks can be used in Random Forest Algorithm.

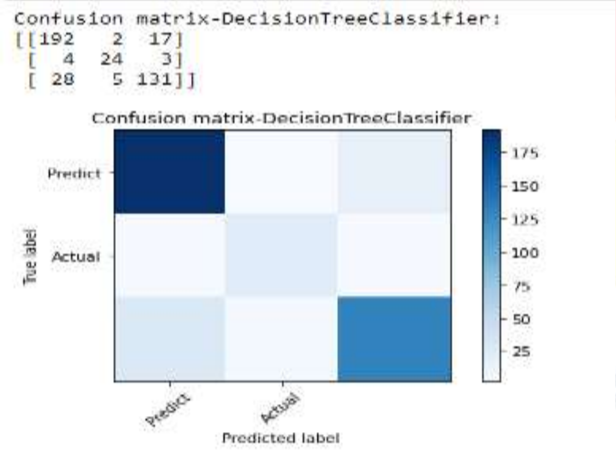


```
Confusion matrix-RandomForestClassifier:
[[192   2  17]
 [  1  26   4]
 [ 21   3 140]]
```

Fig. 7. Cofusion Matrix Of Random Forest Classifier

6.3.3 DECISION TREE CLASSIFIER

Decision-tree algorithm falls under the classification of supervised learning algorithms. Decision tree is created by classification or regression models in the form of a tree structure. The dataset is divided into smaller subset and simultaneously it develops an associated decision tree. A decision node is divided into two or more branches. Decision trees can hold both numerical and categorial data. The Procedure are trained sequentially one at a time. Each time a procedure is learned, the tuples covered by the rules are removed. This process is continued until it reaches a terminating condition.

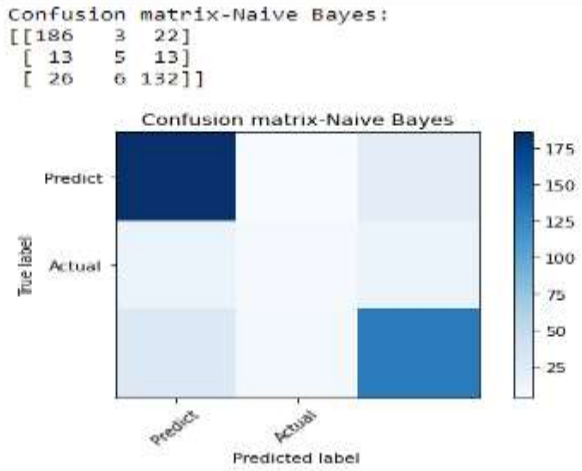


```
Confusion matrix-DecisionTreeClassifier:
[[192   2  17]
 [  4  24   3]
 [ 28   5 131]]
```

Fig. 8 .Confusion Matrix Of Decision Tree Classifier

6.3.4 NAVIES BAYES ALGORITHM

The Naive Bayes algorithm is an intuitive system that uses the chances of each feature belonging to each class to make a prediction. Naive bayes simplifies the calculation of chances by assuming that the probability of each feature belonging to a given class value is independent of all other attributes. This is a strong premise but result in a fast and an effective method.



```
Confusion matrix-Naive Bayes:
[[186   3  22]
 [ 13   5  13]
 [ 26   6 132]]
```

Fig. 9. Confusion Matrix Of Navies Bayes Algorithm

6.4 DEPLOYMENT USING FLASK

Python Flask Framework Web application are developed to induce content based on collected data, where if user interaction is high on the website. The server is accountable for querying, retrieving, and updating data. These operations make web applications to become very slower and more complicated for deploying the static website. Flask is the best web development framework for REST API creation. It is created by top of python in which it can accomplish all python features. Flask is a backend software, where it uses jinja 2 templating language which incorporate HTML, XML and other markup formats to create and use HTTP request user via.

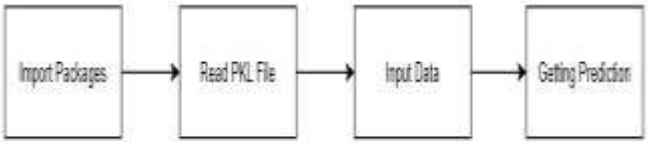


Fig. 10. Workflow of Flask

VII. Results And Discussion

The phishing Dataset contain 10 attributes (SFH, Popup window, final state, request URL, URL of anchor, Web Traffic, URL length, age of domain, IP address and results) and it contains 1353 data of (-1,0,1). The user interface is built on python which is high- level interpreted programming language. The software used Anaconda with Jupyter Notebook which is a open source distribution of python. Anaconda Navigator allows conda packages to launch and common Python Programming Packages. The hardware required are processor – Pentium (IV/III) and the hard disk needed is minimum 80GB and RAM- minimum 2 GB. The main goal of this project is to find out the best accuracy by comparing the different algorithm. The dataset is processed and cleaned by data pre processing technique to remove duplicate values and missing values and visualize the values of web traffic, URL length, IP address in charts and plots. The machine learning model compare different algorithm to find best accuracy and deploys the best model through the user interface and predicts whether the website is phishing or not.

7.1 ACCURACY

Accuracy is one of the metrics to find the best prediction from different algorithmic value. It can be measured by number of correct predictions to the total number of predictions.

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

Algorithm	Accuracy Value
Logistic Regression	82.8
Random Forest Classifier	89.4
Decision Tree Classifier	87.6
Navies Bayes Algorithm	81.8

Table. 1. Accuracy value

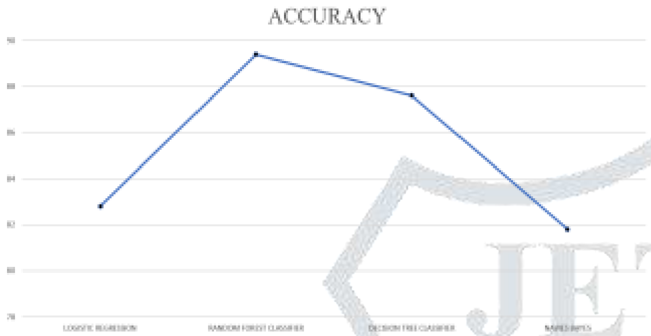


Fig. 11. Accuracy Of Algorithms

7.2 PRECISION

It can be measured by the proportion of appropriate positive prediction. It can be calculated by the ratio of actual classified instance to the total number of classified instances. A Precision we got in this paper is 0.78

PRECISION = TP/ (TP + FP)

Algorithm	-1	0	1
Logistic Regression	0.82	0.56	0.81
Random Forest Classifier	0.91	0.80	0.87
Decision Tree Classifier	0.86	0.83	0.88
Navies Bayes Algorithm	0.83	0.36	0.79

Table. 2. Precision value

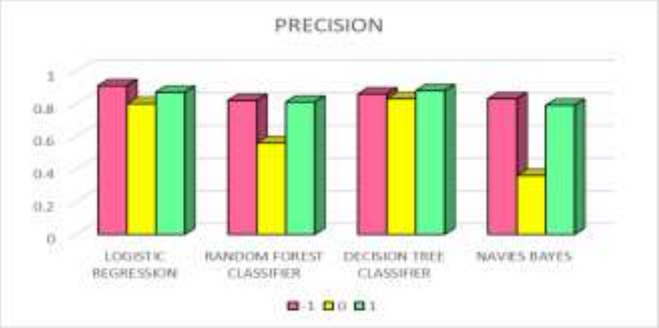


Fig. 12. Precision value Of Algorithms

7.3 RECALL

It can be measured by the proportion of definite positive label identified by the model. Each algorithm has different sensitive value of observation. It can be calculated by the ratio of definite positive observation to the total observation. RECALL = TP/ (TP + FN)

Algorithm	-1	0	1
Logistic Regression	0.91	0.16	0.82
Random Forest Classifier	0.91	0.77	0.87
Decision Tree Classifier	0.91	0.77	0.82
Navies Bayes Algorithm	0.88	0.16	0.80

Table. 3. Recall value

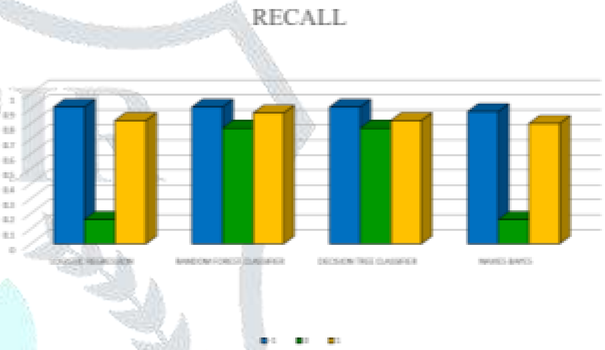


Fig. 13. Recall Value of algorithm

7.4 F1 SCORE

F1 score is the combination of precision and recall which is relatively more specific to positive class. When you have uneven class distribution of Predicting fraud data the majority score is legitimate and minor values is phishing. The F1 score is calculated by the weighted average of the precision and recall. The best value of F1 score is 1 and the worst value is 0.

F-MEASURE = 2TP/ (2TP + FP + FN)

F1 SCORE = 2*(RECALL*PRECISION) / (RECALL + PRECISION)

Algorithm	-1	0	1
Logistic Regression	0.86	0.25	0.81
Random Forest Classifier	0.91	0.79	0.87
Decision Tree Classifier	0.89	0.80	0.85
Navies Bayes Algorithm	0.85	0.22	0.80

Table. 4. F1- Score value

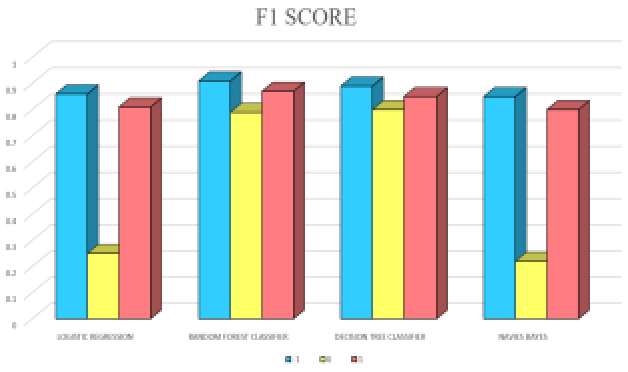


Fig. 13. F1-Score value of algorithm

7.5 SCREENSHOT

PREDICTION OF PHISHING NETWORK
PREDICTION OF PHISHING NETWORK

Fig. 1. Accuracy Of Algorithms

VIII. Conclusion And Future Work

As a result, A good Phishing Prediction website must have high performance metrics and best accuracy. Our Proposed methodology Anomaly detection which predicts the abnormal data and action performed by the phishing network. We conduct series of experiment consistently on a dataset to predicts the outcome of unusual data given to the user interface. The Analytical process starts from Data Processing and cleaning, visualization, exploratory analysis of different algorithms and finally the model is built to find the website is legitimate, suspicious or phishing. The best accuracy on common test case is high. This web application helps us to Predict the website is Phishing website or not. The machine learning model is built by applying proper data science techniques like variable identification that is the dependent and independent variables. Then the visualization of the data is done to insights of the data. In this paper we are using four Algorithm to detect the accuracy from which the best one will be chose or displayed as the result. From the experimental result, we found that the proposed model has high accuracy and performance metrics than the existing model. In Future development of this project can be implemented to connect with cloud model. To enhance this works it can be executed in Artificial Intelligence environment.

References

[1] Yidong Chai, Yonghang Zhou, Weifeng Li, and Yuanchun Jiang "An Explainable Multi-Modal Hierarchical Attention Model for Developing Phishing Threat Intelligence "2021.

[2] P. Prakash, M. Kumar, R. Kompella and M. Gupta, "Phish Net: Predictive Blacklisting to Detect Phishing Attacks," 2010.

[3] P. Wang, G. Zhao, and P. Zhang, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning. 2019.

[4] Z. C. Lipton, J. Berkowitz and C. Elkan, "A critical review of recurrent neural networks for sequence learning", Oct. 2015.

[5] M. A. Adebowale, K. T. Lwin, and M. A. Hossain, "Intelligent phishing detection scheme using deep learning algorithms scheme," J. Enterprise. Info. Management, 2020.

[6] Jian Feng, P. Wang, G. Zhao, and P. Zhang "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning", 2020.

[7] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An Empirical Analysis of Phishing Blacklists," in Proc. 6th Conf. Email Anti-Spam (CEAS'09), Jul. 2019.

[8] J. Zhang, P. A. Porras, and J. Ullrich, "Highly predictive blacklisting." in USENIX Security Symposium, P. C. van Oorschot, Ed. USENIX Association, 2014.

[9] C. N. Gutierrez et al., "Learning from the ones that got away: Detecting new forms of phishing attacks," IEEE Trans. Dependable. Secure. Computer, vol. 15, 2018.

[10] B. Wei, R. A. Hamad, L. Yang, X. He, H. Wang, B. Gao, and W. L. Woo, "A deep-learning-driven light-weight phishing detection sensor," Sensors (Switzerland), vol. 19, no. 19, pp. 1–13, 2019.

[11] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," IEEE Commun. Surv. Tutorials, vol. 15, no. 4, pp. 2091– 2121, 2013.

[12] A. Aljofey, Q. Jiang, Q. Qu, and M. Huang, "An Effective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL," 2020.

[13] A. Abbasi, Z. Zhang, D. Zimbra, H. Chen, and J. F. Nunamaker, "Detecting fake websites: The contribution of statistical learning theory," MIS Q., vol. 34, no. SPEC. ISSUE 3, pp. 435–461, 2010.

[14] R. M. Mohammad, "An Assessment of Features Related to Phishing Websites using an Automated Technique," pp. 492–497, 2012.

[15] JianTing Yuan , YiPeng Liu , and Long Yu , "A Novel Approach for Malicious URL Detection Based on the Joint Model", 13 December 2021