URL’S Phishing Detection Based on Machine Learning Approach

Prof. Vindhya L[1], Anusha D[2], Deekshitha M[3], Keerthana V[4], Manasa[5]
Assistant Professor[1] and Students[2,3,4,5]
S. J. C Institute of Technology, Chikkaballapur, India
anushad1819@gmail.com, mdeekshitha0@gamil.com, keerthanav0502@gmil.com, manasakris2000@gmail.com

Abstract: *Phishing detection is a challenging problem, and many different solutions are proposed in a market as a blacklist, rule-based detection, anomaly-based detection etc. Phishing Websites are duplicate webpages created to mimic real websites in-order to deceive people to get their personal information. Because of the adaptability of their tactics with little cost detecting and identifying phishing websites is really a obscure and dynamic problem.*

Keywords: Blacklist, rule-based detection, mimic, deceive, tactics, obscure, dynamic.

I. INTRODUCTION

In our everyday life, we used to carry outmost of our work on computerized platforms using a desktops or laptops with internet in rural and urban areas facilities our business and private life. It permits us to complete the transaction and operations swiftly in areas such as Education, health, banking, research, engineering and other public services etc. The users who need to access internet connectivity easily at anywhere and anytime with the development of wireless technologies. Although this provides a great assurance. Thus the need for users in cyberattack on websites to compute against possible phishing URL’s has appeared. Cyberattacks can be carried out by peoples such as freebooter, cybercriminals, campaigner etc. The main aim to reach the content or to steal the personal information in many other ways.

These cyberattacks are targeted in different areas like fraud, malware applications, social engineering, castrate. The main aim of attackers is to reach the users to steal the users personal information.

II. LITERATURE SURVEY

- Protecting computer and network information of an organizations and individuals become an important task, because compromised information can cause huge loss.
- Hence, intrusion detection system is used to prevent this damage. To enrich the function of IDS, different machine learning approaches get developed.

The main objective is to address the problem of adaptability of Intrusion Detection System (IDS).

- The proposed IDS has the proficiency to recognize the well-known attacks as well as unknown attacks.
- The proposed IDS consist of three major mechanisms: Clustering Manager (CM), Decision Maker (DM), Update Manager (UM). Dataset is applied to estimate the working of the proposed IDS.
- Both supervised techniques were accompanied.
- The information received to the system is grounded on the education of an agent who disregards the correction proposals presented by IDS. This technique is applied on supervised mode. Both known and unknown traffics can be detected by the system, when they work under unsupervised mode.

III. EXISTING SYSTEM

The performance of the proposed model is evaluated by the datasets. In order to train classifiers like SVM and training dataset is taken which contains large number of instances. Dataset is taken rather than entire dataset, because applying entire dataset will cause several problems. Symbolic attributes like protocol services and flag get changed are removed .Finally, the instances get labelled under four categories: Normal, DoS, Probe, and R2L.They have trained SVM with the dataset.

DOI: 10.48175/IJARSCT-5043

IV. PROPOSED APPROACH METHODS

Dataset pre-processing, classification and result evaluation are the vital phases in the proposed model. In proposed system each phase is essential and enhances important influence on its performance. To examine the function of SVM and Naïve Bayes classifiers are the essential steps of this work. A. Pre-Processing Dataset contains symbolic features; these features are unable to process by the classifier.

V. MACHINE LEARNING BASED ALGORITHMS

5.1 Support Vector Machine: Support Vector Machine (SVM) comes under supervised learning method, in which various types of data from different subjects get trained. In a high-dimensional space, SVM creates hyperplane or multiple hyperplanes. The hyperplane which optimally separates the given data into various classes with the major partition, consider as a best hyperplane. For evaluate the margins between hyperplanes, a non-linear classifier applies various kernel functions. Maximizing margins between hyperplanes is the main aim of these kernel functions like linear, polynomial, radial basis, and sigmoid. Due to the growing attention in SVMs, the eminent applications have been established by the developers and researchers. SVM deals a main role in image processing and pattern recognition applications. Usually a classification task mainly involves dividing data into two sets namely, training datasets and testing datasets. In that class label will be defined as “target variables” and attributes will be defined as features or “observed variables.

5.2 XGBOOST: Xgboost is a decision-tree based ensemble machine learning algorithm that uses a gradient boosting framework**.** In prediction problems involving unstructured data (images, text,etc..) artificial neural networks tend to outperform all other algorithms or framework. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

5.3 Decision Tree: Decision tree (DT) is one of the most popular algorithms in machine learning for binary classification. It results from the decision very fast by creating a small tree and can predict upon training dataset. As its name implies tree, it holds nodes and attribute denotes a test. The branch is the consequence of the test, and each terminal or end node, which is called leaves are the labels of the classification.

Determining the best attribute is the most important in this algorithm. Ross Quinlan developed the decision tree ID3 algorithm. It was primarily used in data mining and information theory. Now it is used in machine learning and natural language processing. The proposed model has used the ID3 algorithm in this paper to classify the website, whether it was an official or phishing website. The following steps are followed to get the outcome of the classification of this algorithm:

1. Start with the training data set. Give it the name 'S' and it should have attributes and classification
2. Determine the best attribute of the data sets
3. Divide the 'S' which each have a value of the best attributes
4. Build a decision tree node which holds the best attribute
5. Use iteration from step 3 and construct a new decision tree unless you cannot classify any more. Represent the leaf node as the outcome of the classification. Information gain can be obtained using entropy. Split information and gain ratio were used to select the alternative attributes which contained numerical values.

5.4 Random Forest Classifier: Random Forest (RF) is one of the robust algorithms in machine learning. It is a supervised learning algorithm used for classification and regression. It uses the bagging method to combine the learning model and average the overall result for better prediction. RF is used to classify the website between legitimate and phishing. As the random forest is a combination of many single trees, it can produce high.

Firstly it selected samples randomly. Then a decision tree on each sample of the dataset was built where the results are obtained from each decision tree. Then the method was applied to predict the result and to select the highest voted result for final prediction.

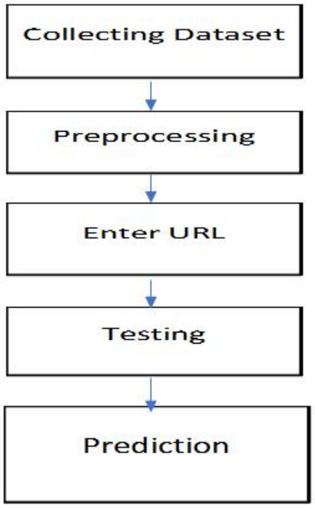
RF produced high accuracy over a single decision tree, even the data was missing. It could overcome the over-fitting problem.

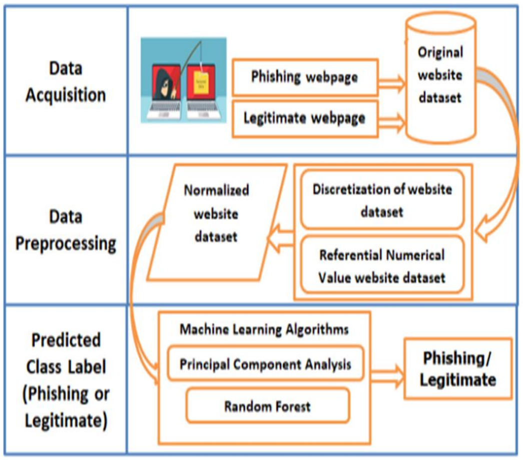
DOI: 10.48175/IJARSCT-5043

5.5 K-Nearest Neighbour: It is one of the simplest machine learning algorithms based on supervised learning technique K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
K-NN algorithms decide a number k which is the nearest neighbor to the data point that is to be classified. If the value of k is 6 it will look for 6 nearest neighbors to the data point. In this example if we assume k=8 K-NN finds out about the 8 nearest neighbors.

5.6. Logistic Regression: Logistic regression is a supervised learning classification algorithm used to predict the probability of a target value.
Logistic regression predicts the output of categorical dependent variable. Therefore the outcome must be a categorical value. It can be either Yes or NO, 0 or 1, True or False etc.. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lies between 0 and 1.
Logistic regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
In Logistic regression instead of fitting a regression line, we fit an "s" shaped logistic function, which predicts two maximum values 0 or 1.
The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
Logistic Regression is a significant machine learning because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

VI. SYSTEM DESIGN

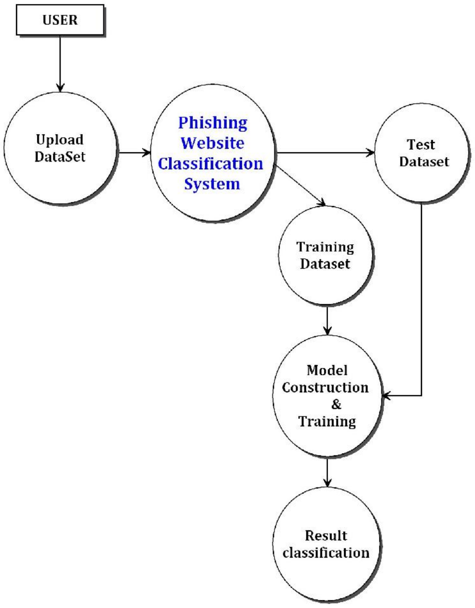



Data Acquisition
Phishing webpage
Legitimate webpage
Original website dataset
Data Preprocessing
Normalized website dataset
Discretization of website dataset
Referential Numerical Value website dataset
Predicted Class Label (Phishing or Legitimate)
Machine Learning Algorithms
Principal Component Analysis
Random Forest
Phishing/ Legitimate

- This is the Usage of finding the entered website is Phishing or not.
- For that used machine learning approach. In that carry with various algorithms like RF, KNN and Logistic Regression, support vector, decision tree etc.
- Here will first collect the dataset, for that dataset can carry with preprocessing like training and testing process.
- Train our model to recognize fake vs real URLs.
- Evaluate our model to see how it performs.
- Check for phishing or not.

VII. DATA FLOW DIAGRAM

Data Flow Diagram Phising Website


USER
Upload DataSet
Phishing Website Classification System
Test Dataset
Training Dataset
Model Construction & Training
Result classification

In this above diagram User will first upload the dataset and the same dataset given to the classification system and that dataset is divided into testing dataset, train dataset. The training and testing dataset are giving to the model construction & training that is used to detect to results Whether predict phishing or not.

VIII. IMPLEMENTATION

8.1 Import Required Libraries

```
import pandas as pd
import NumPy as np
import seaborn as sns
import os
import matplotlib. pyplot as plt
```

8.2 Split Dataset into Train and Test

```
From sklearn.model_selection
import train_test_split
x_train,x_test,y_train,y_test=train_test_split(dataset.iloc[:,-1]dataset.iloc[:,-1],test_size=20, random_state=6)
```

8.3 Building the ModelA

A. Logistic Regression

```
importing of logistic regression
Instantiate the model
regression= logisticregression()
Fit the model
regression.fit(x_train, y_train)
predicting the target value from the model
y_test_regression = regression.predict(x_test)
y_train_regression = regression.predict(x_train)
computing the accuracy of the model performance.
```

DOI: 10.48175/IJARSCT-5043

8.4 Decision Tree

```
importing of  decisiontreeclassifier
   Instantiate the model
         tree= decisiontreeclassifier()
   Fit the model
          forest.fit(x_train, y_train)
  predicting the target value from the model
         y_test_tree = tree.predict(x_test)
         y_train_tree = tree.predict(x_train)
  computing the accuracy of the model performance.
```

8.5 Random Forest Classifier

```
importing of Random forest Classifier
     Instantiate the model
         forest= randomforestclassifier()
Fit the model
          forest.fit(x_train, y_train)
     predicting the target value from the model
         y_test_forest= forest.predict(x_test)
         y_train_forest= forest.predict(x_train)
   computing the accuracy of the model performance.
```

8.6. k-Nearest Neighbor

```
    importing of k-Nearest Neighbor
    Instantiate the model
       knn= k-Nearest Neighbor() = knn.predict(x_test)
       y_train_ knn = knn.predict(x_train)
   computing the accuracy of the model performance.
```

8.7 XG Boost:

```
importing of XG-Boost
Instantiate the model
              xgb= XG-Boost()
Fit the model
             knn.fit(x_train, y_train)
     predicting the target value from the model
            y_test_ knn
Fit the model
             xgb.fit(x_train, y_train)
        predicting the target value from the model
            y_test_xgb= xgb.predict(x_test)
            y_train_xgb= xgb.predict(x_train)
   computing the accuracy of the model performance.
```

8.8 Artificial Neural Network:

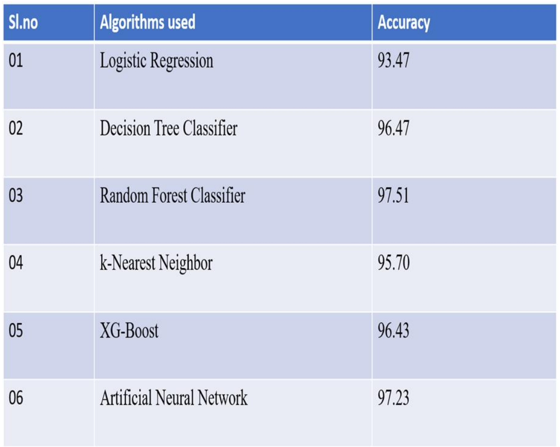
```
importing of Artificial Neural Network
    Instantiate the model
            ann= Artificial Neural Network()
     Fit the model
    ann.fit(x_train, y_train)
```

predicting the target value from the model

```
    y_test_ ann  = ann.predict(x_test)
    y_train_ ann = ann.predict(x_train)
```

computing the accuracy of the model performance.

IX. RESULT



Sl.no	Algorithms used	Accuracy
01	Logistic Regression	93.47
02	Decision Tree Classifier	96.47
03	Random Forest Classifier	97.51
04	k-Nearest Neighbor	95.70
05	XG-Boost	96.43
06	Artificial Neural Network	97.23

X. CONCLUSION

Phishing is a serious security concern which may leads to loss sensitive personal information by the attackers. This project mainly focused on identifying features useful for detecting the phishing websites based on URL of the website and applying machine learning algorithms to classify the website is legitimate or phishing. It involves comparison of accuracy of six machine learning algorithms and then finalized Random forest algorithm emerging as the most accurate. RF algorithm is used to find whether website is phishing or not.

ACKNOWLEDGEMENTS

We would like to express to gratitude to our college, SJC Institute of Technology our guide Prof. VINDYA L and our project coordinator Prof. ARAVIND TEJAS CHANDRA to have provides as with the opportunity to work on this project and given us support with guidance to make this project a success. We would also like to thank our teammates for their contribution and continued support and zeal towards this project. This project wouldn't be a success without them.

REFERENCES

[1]. Rishikesh Mahajan, Irfan Tidivate "Phishing Website Detection using Machine Learning Algorithms". In October 2018 https://www.researchgate.net/publication/32854178_Phishing_Website_Detection_using_Machine_Learning_Algorithms.

[2]. Arun Kulkarni, Leonard L. Brown " Phishing Websites Detection using Machine Learning" In July 2019, International Journal of Advanced computer science and applications. https://thesai.org/Publications/ViewPaper?Volume=10&Issue=7&Code=IJACSA&SerialNo=2 .

[3]. Sushma Joshi, Dr S.M Joshi "Phishing URL's Detection Using Machine Learning Techniques", In 25thth- -th3 June 2019, International Journal Of computer engineering in research trends multidisciplinary,open,access. https://ijcert.org/ems/ijcert_papers/V6I602.pdf.

[4]. Liz hen Tang, Qusayr H. Mahmoud "A Survey of Machine Learning-Based Solutions for Phishing Website Detection", In 20 August 2021,department of electrical, computer, and software engineering, Ontario tech university, Oshawa, on LIG 0C5,Canada. https://www.mdpi.com/2504-4990/3/3/34.

[5]. Naga Sundar Rao Pawar Babu Rao Pawar "Detection Of Phishing URL using machine learning", In 16th Aug 2021 MSc research projects**.** http://norma.ncirl.ie/5100/1/nagasunderraopawarbaburaopawar.pdf.

[6]. Ali A. Ghorbani, Wei Lu and M. Tavallaee, Network intrusion detection and prevention: Concepts and Techniques, Advances in Information security, Springer, 2010.

[7]. A. O. Adetunmbi, S.O. Falaki, O. S. Adewale, and B. K. Alese, Network Intrusion Detection based on rough set and k-nearest neighbour, Intl. Journal of computing and ICT research, 2(1) (2008), 60-66.

[8]. C. Krugel and T. Toth, Using decision tree to improve signature based intrusion detection, in: Proceedings of RAID, 2003, G. Vigna, E. Jonsson, and C. Kruegel, eds, Lecture Notes in Computer Science, Vol. 2820, 173-191.

[9]. D. E. Denning and P. G. Neumann, Audit trail analysis and usage data collection and processing, Technical report project 5910,SRI International.

[10]. G. Wang, J. Hao, J. Ma and L. Huang, A new approach to intrusion detection using artificial neural networks and fuzzy clustering, Expert system with applications, 37 (2010), 6225-6232, Elsevier.

DOI: 10.48175/IJARSCT-5043