International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Phishing Websites Detection using Machine Learning Techniques

Mr. Kondeti Prem Sai Swaroop[1], Ms. Konka Renuka Chowdary[2], Ms. S. Kavishree[3]

[1]Student, Dept. Of CSE, SCSVMV (Deemed to be University), Kanchipuram, TamilNadu, India
[2]Student, Dept. Of CSE, SCSVMV (Deemed to be University), Kanchipuram, TamilNadu, India
[3]Assistant Professor, Dept. Of CSE, SCSVMV (Deemed to be University), Kanchipuram, TamilNadu, India

---

Abstract *- In general, usage of websites is the most common things lately, it may be for e-commerce purposes or entertainment, whatever it may be. In this project, our main factor is a website, whether it is a fraudulent one or a legit one. Detection of this quality of a website is the main theme of the project. Conventionally, a website can be detected whether it is harmful or not by the browser protection service, if it is redirecting to unusual or malicious sites, such sites are marked as harmful with a symbol before the URL. Even though, the browser's firewall is enabled, it can never detect a phishing website. Because, Phishing site is not malicious site it steals data without the user even knowing it. So, to detect such sites we are training an ML model using different algorithms to determine the phishing site based on URL feature extraction. Based on different features of URL, such as like Domain length, character length etc, we will train the model with one algorithm at a time store their results and compare to find the more accurate one and display the results using the approved algorithm.*

Keywords: **Phishing Website; Machine Learning Model (ML Model); URL (Uniform Resource Locator).**

1. INTRODUCTION

Now-a-days usage of internet and surfing through the browser has become a primary requirement for everyone. This could help them from gaining knowledge to fulfil requirements for their own needs. But the problem arises at security. Is it really healthy to surf through every website we see on the internet? Will it be safe and secure for the data present in the device? So, to resolve this problem we are going to train an ML model through various algorithm and let it study the websites to find fraudulent ones.

There are many kinds of algorithms and techniques in the market that could help through to find fraudulent websites, but they aren't that accurate and precise in judging the site. One of such technique is based on using Antiphishing via black listing. Although it is a detection technique this isn't that accurate for judging the correctness of a website.

This technique of ours gives the accurate and precise knowledge about the website rather than the normal procedures of detections. This system can be open-sourced to every running search engine or any other sustainable environment that works on various websites so that it could become easier for the work force to detect the legitimacy of the site.

1.1 Objective

The objective of this project is to develop a ML model that to detect the phishing websites with accuracy. Expected steps and procedures to be followed to fulfil the objectives are:

- To Load the dataset in to the model after feature extracting all the URLs.
- Sort out the dataset without null values and unwanted values.
- Visualize the data to know the frequent factor among the considered features from the dataset.
- Train the ML model with different algorithms and store their results for further comparison.
- Compare all the modules to find out the most accurate algorithm for the detection process.
- Sort out the dataset into phishing and legitimate to get the required output when a website is searched.

1.2 Scope of the Project

We are setting up to design a system comprising of three modules. The first module is data preparation, pre-processing and visualization. The second module is URL feature extraction and setting up the dataset. The last module is training and evaluation of the ML model and store the result.

The major insight behind this project is the raise of phishing attacks. In 2012, total phishing attacks is increased by 160% over than the previous year. The total numbers of attacks recorder in 2013 is 450000 and leading to a loss of $5.9 billion. The total number of attacks in the first quarter of 2014 alone is 125,215 i.e., 10.7 percent surge when compared to the whole cases and in the fourth quarter alone is more than 55% of 2013s overall cases.

Below is the display of financial losses due to phishing attacks from 2011 to 2015 in Figure 1.1 and growth of phishing attacks from 2005 to 2015 in Figure 1.2

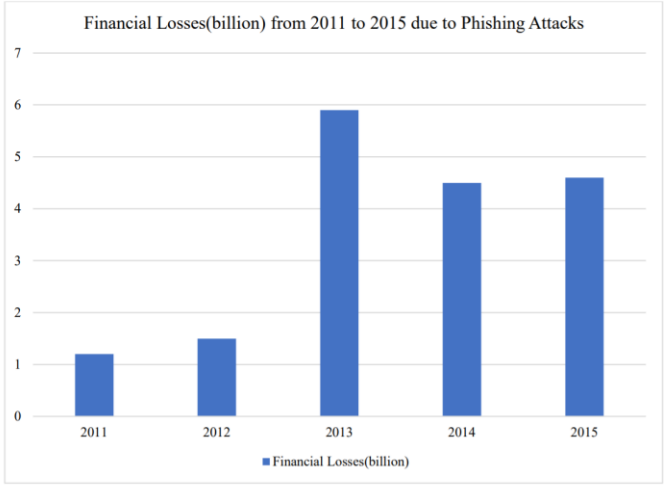

Figure 1.2.1

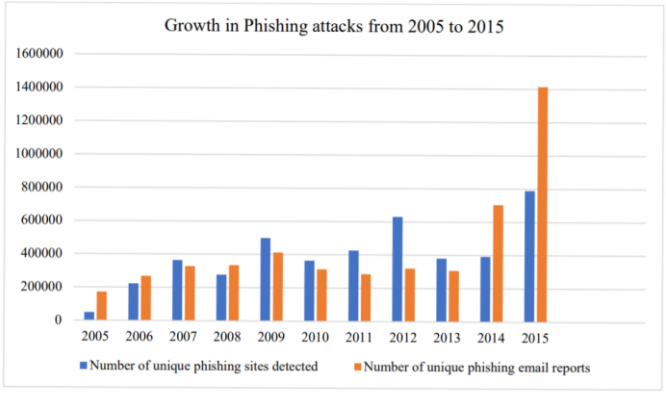

Figure 1.2.2

2. PROPOSED SYSTEM

The main objective of this project is to determine an optimal way for detecting phishing websites which has become a major threat lately. This can be obtained in multiple ways but they are not adaptive to the latest growing technology. In order to overcome the adaptiveness, feature we are use Machine Learning, Deep learning and Neural networks to find an optimal approach. The main requirement of this model is a perfect dataset. We are gathering a total of 10000 URLs from two different sources to form a huge dataset of phishing and legitimate sites shuffled together to determine the correctness of the trained model. In the dataset we used two values Phishing(1) and Legitimate(0) to know their existence difference.

2.1 Workflow

Initially, the dataset is loaded, visualized, familiarized and split in to two different part to train the model from two different viewpoint simultaneously. The brief descriptive flowchartof training our ML model is displayed below.

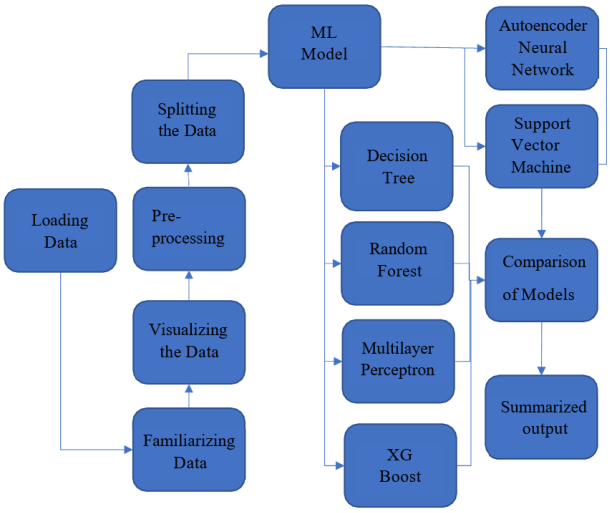

Fig-2. Project Workflow

2.2. Models and Training

From the dataset created we can learn that this is a supervised machine learning task. There are two types of supervised machine learning tasks. They are Classification and Regression. Our project comes under classification as the input URL is classified as Phishing (1) and Legitimate (0).

So, below are the bet classification supervised machine learning models which we are going to use to use to train our ML model.

- Decision Tree Classifier
- Random Forest Model
- Multilayer Perceptrons
- XGBoost Classifier
- Auto Encoder
- Support Vector Machines

2.3.1. Decision Tree Classifier:

Decision Tree Classifiers are widely used in classification and regression tasks which involvea decision task such as if/else question. This is an optimal decision maker that could give us the best decision much quicker.

In ML models, these decisions are named as tests. Tests in the sense, to find out whether our model is generalizable or not. This algorithm can be used to sequenciate all the tests from the dataset to find the information about the target variable.

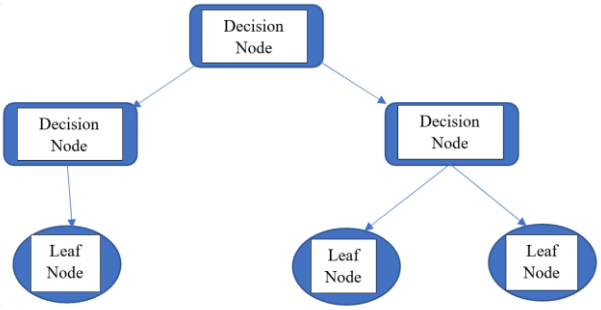

Fig.2.3.1-Decision tree classifier

2.3.2. Random Forest Classifiers:

It can be defined as a collection of decision trees. Here, multiple number of decision trees are collected together and worked simultaneously to get the best of the average of the result. This will be very robust and all you need to know no of trees required before building a random forest. This does not any parameters or scaling of data, all you need is n_estimators.

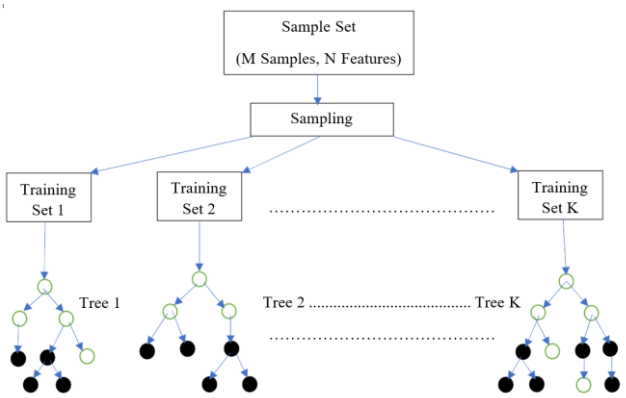

Fig-2.3.2. random forest clarifier

2.3.3. Multilayer Preceptrons:

Multilayer Perceptrons are known as feed forward neural networks. They are used to process multiple stages simultaneously and result in an optiomal decision for the processed stage.

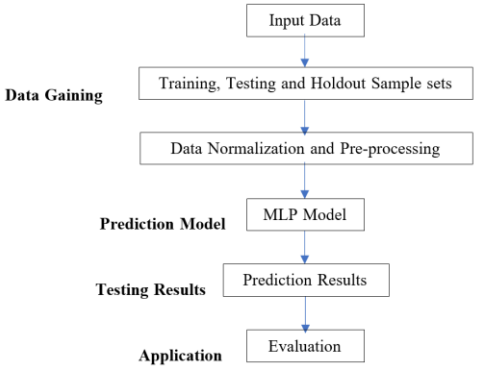

Fig-2.3.3. multilayer preceptrons

2.3.4. XGBoost Classifier:

XGBoost is not any different for classification or regression process, it is meant for speed andperformance. It will add gredient boosting to decision trees.

Fig-2.3.4.XGBoost classifier

2.3.5. Auto Encoder Neural Network:

It is like a neural network that has same no. of input neurons that of output neurons. It has fewer neurons in the hidden layers of the network that are called as predictors. The input neurons pass information to the predictors and process the output.

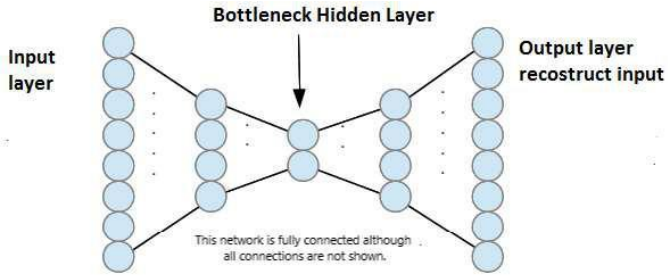

Fig-2.3.5. auto encoder neural network

2.3.6. Support Vector Machines:

Support vector machines also known as support vector networks analyse the data used for classification or regression task. The training data set is loaded and when analysed will be sorted out in to two different categories for echo new output appeared

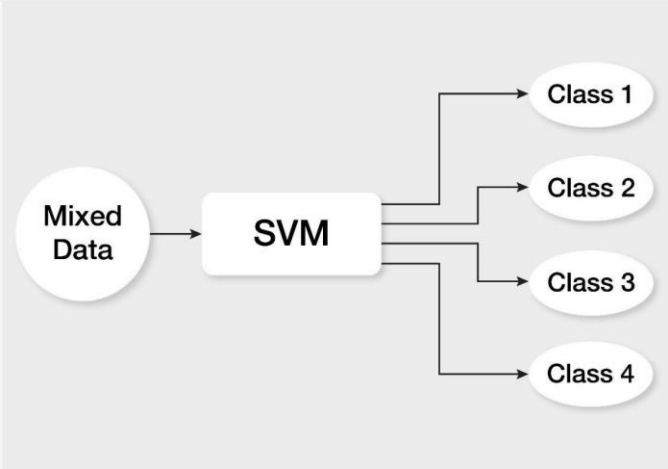

Fig-2.3.6. support vector machines

3. RESULTS

	ML Model	Train Accuracy	Test Accuracy
0	Decision Tree	0.810	0.826
1	Random Forest	0.814	0.834
2	Multilayer Perceptrons	0.858	0.863
3	XGBoost	0.866	0.864
4	Auto Encoder	0.819	0.818
5	Support Vector Machines	0.798	0.818

Table-3.1. Comparison of Training and Testing Accuracy of Modules

	ML Model	Train Accuracy	Test Accuracy
3	XGBoost	0.866	0.864
2	Multilayer Perceptrons	0.858	0.863
1	Random Forest	0.814	0.834
0	Decision Tree	0.810	0.826
4	Auto Encoder	0.819	0.818
5	Support Vector Machines	0.798	0.818

Table-3.2. Sorted out Values of Training and Testing

```
if tab:
    print(green +"Legitimate")
else:
    print(red + "Phishing")
```

```
icicibank.com
Legitimate
```

```
    print(green +"Legitimate")
else:
    print(red + "Phishing")
```

```
kinemind.com
Phishing
```

Figs-3.3 &3.4 outputs

(above picture shows if it is legitimate)

(below piture shows if it is phishing

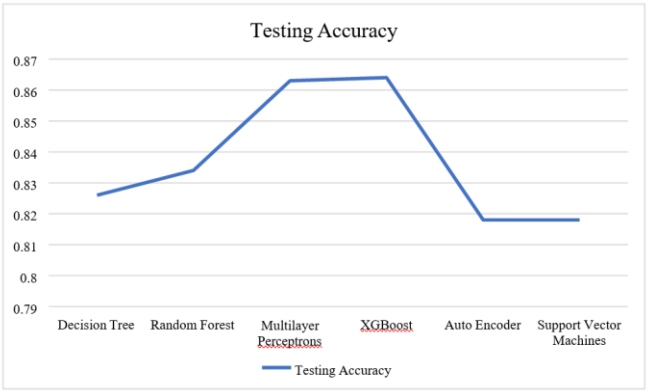

Figure 3.5

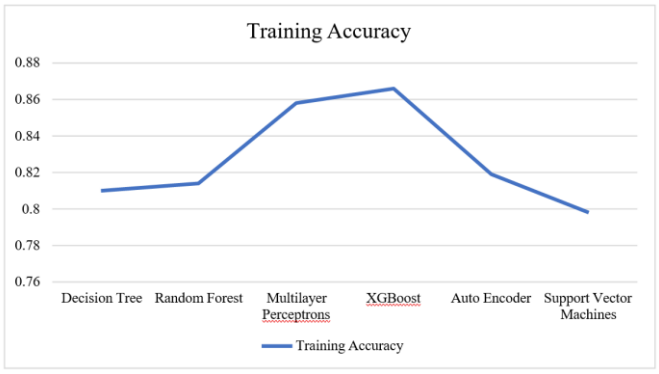

Figure 3.6

Fig-3.5 & 3.6 reprsents graphical presntation of table 3.1 & 3.2

4. Conclusion

The aim of our project is to make an ML model that can detect phishing websites and also can be adaptive to future upgrades in technology. All the models considered are successfullytrained and tested and their metric accuracy is recorded that could be useful for further enhancements of the project. As per the project the project our model can only detect the sites whether they are phishing or legitimate. So, from our defined dataset any URL from them is detected by the model based on URL feature extracted dataset.

REFERENCES

[1] Sahar Abdelnabi, Katharina Krombholz, Mario Fritz. Visual PhishNet: Zero-Day Phishing Website Detection by Visual Similarity: CISPA Helmholtz Centre for Information Security, Published in October 2020.

[2] Xiuwen Liu, Jianming Fu. SPWalk: Similar Property Oriented Feature Learning for Phishing Detection: Key Laboratory of Aerospace Information Security and Trusted Computing of Ministry of Education, School of Cyber Science and Engineering, Published on7th May 2020.

[3] J. Rajaram, M. Dhasaratham. Scope of Visual-Based Similarity Approach Using Convolutional Neural Network on Phishing Website Detection: Department of Computer Science and Engineering, Teegala Krishna Reddy Engineering College, Published on 11thAugust 2020.

[4] M Somesha, Alwyn Roshan Pais, Routhu Srinivasa Rao, Vikram Singh Rathour. EfficientDeep Learning Techniques for the Detection of Phishing Websites: Information Security Research Lab, National Institute of Technology Karnataka, Published on 27th June 2020.

[5] Yan Chen, Fatmeh Mariam Zahedi, Ahmed Abbasi, David Dobloyi. Trust calibration of automated security IT artifacts: A multi-domain study of phishing-website detection tools: College of Business, Florida International University, Sheldon B. Lubar School of Business, University of Wisconsin, Mendoza College of Business, University of Notre Dame, Published on 25th November 2020.

[6] Abdulhamit Subasi, Emir Kremic. Comparing AdaBoost with MultiBoosting for PhishingWebsite Detection: Effat University, College of Engineering, Published on 13th May 2020.

[7] Edwin Donald Frauenstein, Stephen Flowerday. Susceptibility to phishing on social network sites: A personality information processing model: Department of Information Systems, Rhodes University, Published on 1st My 2020.

[8] Ella Glikson, Anita Williams Woolley. Human Trust in Artificial Intelligence: Review of Empirical Research: Bar Ilan University, Published on 10th August 2020.

BIOGRAPHIES

Kondeti Prem sai swaroop is pursuing Computer Science and Engineering in SCSVMV (Deemed to be University).

Konka Renuka Chowdary is pursuing Computer Science and Engineering in SCSVMV (Deemed to be University).

Ms. S. Kavishree is Assistant Professor in Computer science and Engineering department in SCSVMV (Deemed to be University).