IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, December 2022
Impact Factor: 6.252

Applications of Machine Learning Techniques to Detect Phishing Websites

Abhignya Tayi
Vellore Institute of Technology, Chennai, Tamil Nadu, India
abhignya15@gmail.com

Abstract: *Phishing attacks are a common way for hackers to obtain sensitive and valuable information from unsuspecting users. These attacks often target critical data such as passwords and financial details. To combat this threat, cybersecurity professionals are constantly searching for reliable and effective techniques for detecting phishing websites. This project investigates the use of machine learning algorithms to identify phishing URLs by extracting and analyzing various features of both legitimate and phishing URLs. The goal is to create a blacklist of known phishing websites that can alert individuals when they browse or access a potentially dangerous site. The project will compare the performance of four machine learning algorithms such as Ensemble Adaboost Classifier, Multi-Layer Perceptron Classifier, Stochastic Gradient Descent classifier, and XGBoost - based on their accuracy, speed, and other factors.*

Keywords: Phishing, Ada boost, Multilayer perceptron, Stochastic gradient descent, Kernel approximation, Machine Learning, Accuracy, Comparison

I. INTRODUCTION

Phishing is a major concern for cybersecurity professionals because it is becoming increasingly easy for attackers to create fake websites that mimic the appearance of legitimate ones. As a result, it is becoming more difficult for users to identify and avoid these traps. The primary goal of phishers is often to steal bank account credentials and other sensitive information. In the United States alone, businesses lose an estimated $2 billion per year due to phishing attacks, and the worldwide impact is believed to be as high as $5 billion according to the 3rd Microsoft Computing Safer Index Report. A lack of awareness among internet users is a major contributing factor to the success of phishing attacks, which exploit vulnerabilities in human behavior. Although it is difficult to completely eliminate the threat of phishing, improving detection techniques is crucial in the fight against these attacks.

One common approach to detecting phishing websites is to maintain a blacklist of known malicious URLs and IP addresses, which is then used to update antivirus databases. However, attackers can use various techniques, such as obfuscation and fast-flux, to evade these blacklists and make their phishing websites appear legitimate. The main drawback of this method is that it is not effective against zero-hour phishing attacks, which are launched without warning. Heuristic-based detection involves identifying characteristics that are commonly found in phishing attacks and using them to detect zero-hour attacks. However, these characteristics are not always present in all attacks, leading to a high rate of false positives. As a result, this method may not be reliable for accurately detecting phishing attacks.

To address the limitations of blacklisting and heuristic-based methods, many security researchers are turning to machine learning techniques as an alternative. Machine learning algorithms use past data to predict future outcomes, and can be trained on a dataset of both legitimate and malicious URLs to accurately detect phishing websites, including zero-hour attacks. By analyzing the features of these URLs, the algorithm is able to identify patterns and characteristics that are indicative of phishing websites.

II. RELATED WORK

Maher Aburrous et al (2013) proposed a new fuzzy classification technique for detecting phishing websites. Most of the algorithms used in other research were not able to make the decision dynamically to reduce the problem a new fuzzy classification algorithm was created to detect phishing websites. In this algorithm, a fuzzy logic model is incorporated with classification algorithms. The classification model proposed provided effective help for real-time phishing websites with an accuracy score of 86% by avoiding fake alarmed phishing alarms. An intelligent heuristic approach was used to

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, December 2022
Impact Factor: 6.252

implement the phishing detection toolbar; this technique has been used to increase the performance of the detection rate. This kind of algorithm is not used in many research papers to take it to the next level. This phishing detection system can be made as a desktop application by which most people will be safe from fraudulent alarms and viruses from the website and it will also increase the awareness among people to be safe from phishing websites.

Mustafa Al-Fayoumi et al (2020) proposed a new Association classification algorithm as an AI-generated tool to increase the performance for detecting phishing websites. The IAC algorithm known as the Intelligent association algorithm is a newly derived algorithm that has discovered the hidden patterns of the information which have not been discovered by the existing system. Compared to the other algorithms like the random forest, decision tree, and other classification models, and provided better precision, recall, accuracy, and f-measure values. After comparing with the existing models IAC outperformed them with the vast difference in the scores related to the performance of the model. The performance was increased in the IAC model using the efficient statistical measure. This new model is based on harmonic Mexican measures in the phases of classification. Working on different types of statistical measures will have an impact on the performance at a better rate than before.

Muhammad Rayhan Natadimadja et al (2020) proposed a new framework developed by google known as Hadoop and MapReduce with the incorporation of PhishTank service which is an online service that can be used to report malicious activity. The project aims to develop new techniques and provide better accuracy to detect malicious or phishing websites. To increase the performance for detecting malicious websites, hand loops and MapReduce frameworks are used to be run in a distributed environment. This system is very advanced and has high-performance results compared to other classification models. The algorithms such as Neural network and optimal feature selection with accuracy 99.9% and TVD algorithm with 99.5% accuracy.

Syukrina Kamilia Binti Jasmi et al (2019) proposed an algorithm known as a two-class locally deep support vector for increasing the performance of the model. The research focuses on the evaluation of the website as a normal or malicious one. To test for phishing websites, a feature selection algorithm is combined with an ensemble learning methodology. which helps the dataset to check if the model is accurate or not. This model gives an accuracy of around 95% which is better than the existing algorithms like random forest and decision tree which gave accuracy less than the proposed method. classification methods are used to identify phishing URLs for which many machine learning models are used in that two-class boosted decision tree has the highest accuracy of around 97.3%. It is important to educate the user to be careful of these fraudulent attacks.

Rishikesh Mahajan, and Irfan Siddavatam (2018) proposed to detect phishing URLs as well as find the best machine learning algorithm by comparing of accuracy rate, false positive, and false negative rate of each algorithm. The collection of benign URLs was done from www.alexa.com and phishing websites from www.phishtank.com. The data set consisted of a total of 36,711 URLs which included 17058 benign URLs and 19653 phishing URLs. Benign URLs were classified as “0” and phishing URLs were classified as “1”. Accuracy of 97.14% was achieved using a random forest algorithm with the lowest false positive rate. Also, classifiers gave better performance when more data was used as training data.

Vahid Shahrivari et al compared various machine learning techniques in the detection of phishing websites. Implementation and Evaluation of twelve classifiers were done on the dataset consisting of 6157 valid websites and 4898 phishing websites. The list of tested classifiers included Logistic Regression, Decision Tree, Support Vector Machine, Ada Boost, Random Forest, Neural Networks, KNN, Gradient Boosting, and XGBoost. Excellent performance was shown by ensemble classifiers like Random Forest and XGBoost. Their accuracy and computation duration were very competitive. Several weak learners were together into strong ensemble algorithms.

Ashit Kumar Dutta 2021 proposed a recurrent neural network method to classify phishing URLs. The method was tested with 7900 malicious and 5800 valid websites. RNN was used because of its capability to deal with large amounts of data. LSTM technique was used to identify malicious and legitimate websites. A new crawler was developed to crawl 7900 URLs from the AlexaRank portal and the Phishtank dataset was used to measure the efficiency of the proposed URL detector. This method provided better results than deep learning methods that exist currently.

Atharva Deshpande et al in 2021 explained the domain of phishing characteristics, the differing features from valid domains, significance of determining these domains, and the use of machine learning and natural language processing methods for their detection. Traditional Methods to detect phishing websites- blacklist and heuristic evaluation were analyzed and their disadvantages were ascertained. Two machine learning algorithms were tested on a dataset. The best

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, December 2022
Impact Factor: 6.252

classifier was selected among the two and a chrome extension was built out of the same. Random Forest algorithm was found to have an accuracy of 97.31%

III. PROPOSED WORK

Phishing websites are a major threat to internet users, and it is essential to be able to identify and remove them. To predict which websites are likely to be phishing sites, we use classification algorithms that can segregate URLs into phishing and legitimate categories. Traditional machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, and SVM have been used for this purpose, but they are not always accurate enough to detect malicious content. To improve the accuracy of phishing website detection, newer and more sophisticated algorithms have been developed, including the Ensemble Adaboost Classifier, Multi-Layer Perceptron Classifier, Stochastic Gradient Descent classifier, and XGBoost. These algorithms are able to achieve higher levels of performance and are more effective at detecting phishing websites.

3.1 Data Collection

The dataset used in this research was obtained from Kaggle and includes data from 10,000 web pages (5,000 phishing and 5,000 legitimate) accessed between January 2015 and June 2017. The dataset includes 48 characteristics that were collected using the Selenium WebDriver browser automation framework, which is a more accurate and reliable method for extracting features than using regular expressions. This dataset may be useful for researchers and experts in the field of anti-phishing, as it can be used for analyzing the characteristics of phishing attacks, conducting rapid proof-of-concept tests, or comparing the performance of different phishing classification algorithms.

3.2 Data Cleaning

The dataset obtained from Kaggle includes null values and special characters that may affect the accuracy of the predictions. To ensure that the model is able to make accurate predictions, these random values must be removed through a process called data cleaning. In this process, null and special character values are eliminated, and any unnecessary columns that do not contribute to the prediction are dropped. The dataset has a total of 10,000 rows and 50 columns, but only 26 columns are required for the prediction. The dataset is shuffled to mix the legitimate and phishing URLs, as the first 5,000 rows contain legitimate URLs and the next 5,000 rows contain phishing URLs. This shuffling is necessary to ensure that the training and testing datasets are properly split.

3.3 Exploratory Data Analysis

In the data analysis module, different types of plots such as line plots for univariate terms, scatter plots for bivariate terms, Histograms, and correlation maps are visualized.

3.4 Data Preprocessing

Data preprocessing is the process of selecting and preparing the attributes that are needed for the prediction process. In this case, the target column "CLASS_LABEL" is used to determine whether a website is a phishing site. The parameters x and y are defined, with x representing all the columns in the dataset except for the target column, and y representing only the target column. After the parameters are split, the dataset is divided into training and testing datasets in a ratio of 70:30 using the train_test_split function from the sklearn library. The training dataset is used to train the model, while the testing dataset is used to evaluate the model's performance.

3.5 Model Training

Machine learning models such as Ensemble Adaboost Classifier, Multi-Layer Perceptron Classifier, Stochastic Gradient Descent classifier, and XGBoost are used.

3.5.1 Stochastic Gradient Descent Classifier

The Stochastic Gradient Descent (SGD) classifier is an optimization algorithm that adjusts the parameters of the model based on the training data in order to minimize the error function or cost variable. It is particularly useful for large datasets

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, December 2022
Impact Factor: 6.252

because it updates its values for each training example, making it more accurate than other models such as Logistic Regression, Random Forest, and Decision Tree. The SGD classifier can be imported from the Sklearn module. In this study, the SGD classifier was found to have an accuracy of 99.8%.

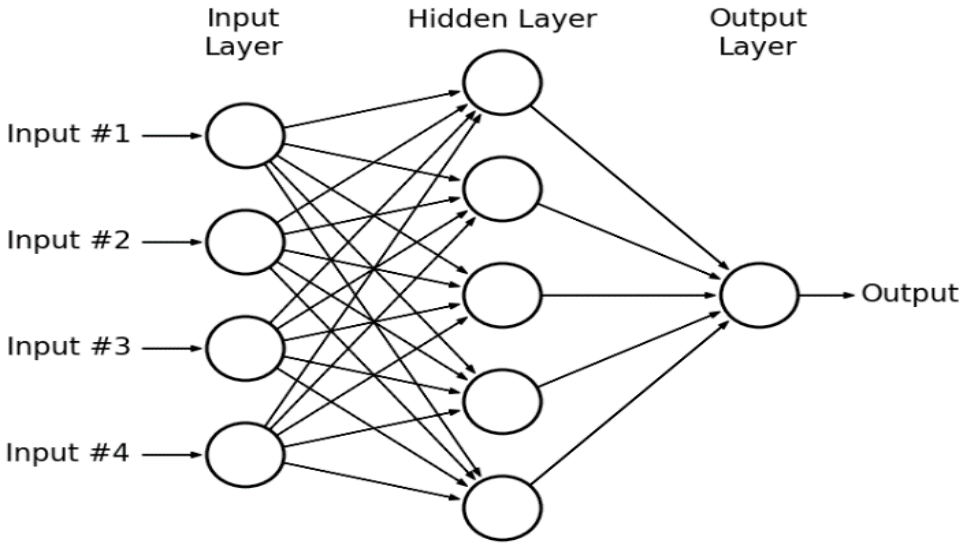
Stochastic gradient descent pseudo-code:

1. Initialize the wi = 0, where i = 1,2,3,.....d
2. Find the value of N (generally the value of n lies between 0 and 1).
3. Repeat these steps until convergence

```
For t=1 to n
{
For i = 0 to d
{
 Wi = Wi - N([g(x^t) - r^t]) * Xi^t
}
}
```

3.5.2 Multilayer Perceptron Classifier

The Multi-Layer Perceptron Classifier is a type of feed-forward neural network that has intermediate layers, also known as hidden layers, between the input and output layers. This model can be used for either classification or regression tasks, depending on the specific application. In this study, the model was implemented as a non-discriminant classifier and used the sigmoid function for the hidden layers. The accuracy of the Multi-Layer Perceptron Classifier was found to be 99.96%.



Multi-layer perceptron model pseudo-code for k outputs:

1. Initialize all the Uih and Wjh to rand
2. Repeat for all (x^t , r^t) belongs to X in random order

```
For h = 1,2,3,.....H
 Zh <= sigmod(Wh^t * x^t)
For i = 1,2,3,....k
 Yi = Vi^t*Z
For i = 1,3,.....k
 delta(Vi) = D(Vi^t - Yi^t)*Z
For h = 1,2,3…..H
 delta(Wh) = n* Summation( (ri^t - yi^t) * Uih *Zh*(1-Zh)*x^t)
For i = 1,2,3….k
```

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, December 2022
Impact Factor: 6.252

Ui = Ui + delta(Ui)
For h = 1,2,3….H
Wh = Wh + delta(Wh)
convergence
3. Repeat the steps until convergence

3.5.3 Kernel Approximation Classifier

Kernelized algorithms, such as Support Vector Machines (SVMs), have a runtime and space complexity that is largely independent of the input space's dimensionality. Instead, they scale with the number of data points used for training. There are various techniques to accelerate the training process for SVMs, but in general, the runtime is proportional to the number of samples. In order to save on computational resources, excellent implementations often avoid calculating the kernel values for all pairs of training points, although this can come at the cost of increased complexity. In some cases, it may be desirable to save the entire kernel matrix, which stores the kernel value for all pairs of training points, but this can be quadratic in the number of samples and may not be feasible for large datasets.

Algorithm for kernel approximation:

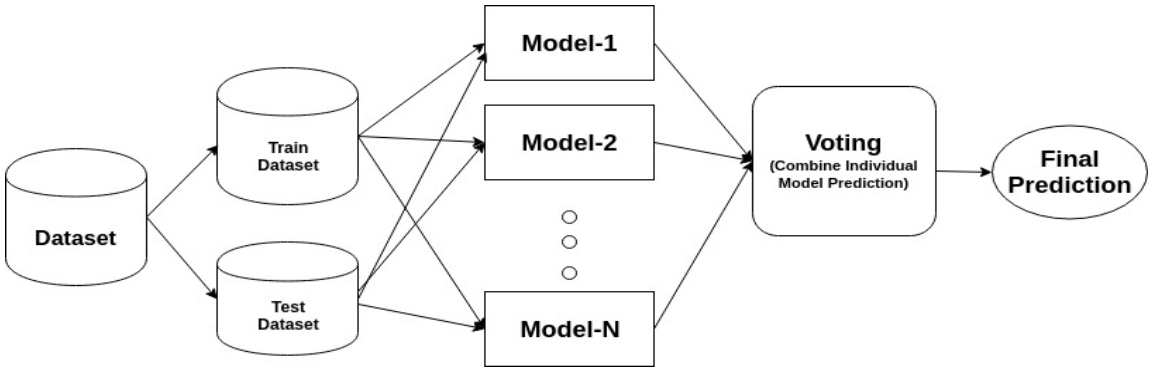
1. A positively definite shifting invariant kernel is used as an input. K(x, y) = K(x − y)
2. Calculate the kernel k's Fourier transform p.

P(w) = 1/2π* Integral Rd (e^(−jwTδ) * K (δ)* dΔ)

3. Draw L iid w1, w2, ..., wL ∈ R d samples from p
4. Return: z(x) = 1 /√ L [cos(Wi^ T* x),sin(Wi^T *x)] L i=1 ∈ R 2L

3.5.4 Adaboost Ensemble Classifier

Boosting is a method of building a strong classifier from a collection of weak classifiers by iteratively improving upon an initial model. This is done by training the first model on the training data, then using the second model to correct any errors made by the first model. Additional models are added until the training set is accurately predicted or the maximum number of models is reached. AdaBoost was the first boosting algorithm to achieve widespread success, particularly for binary classification tasks. It is considered a good starting point for learning about boosting, and many modern boosting algorithms, such as stochastic gradient boosting machines, are based on AdaBoost. In this study, the tree depth ensemble technique was used.



3.5.5 Xgboost Classification Model

XGBoost is a powerful library for building gradient boosting models, and it can be used in the scikit-learn framework through the XGBoost wrapper class. This class allows XGBoost models to be used as classifiers or regressors, and they can be integrated with the rest of the scikit-learn library. The XGBClassifier is the XGBoost model for classification tasks. To build and fit the model to the training data, you can use the XGBClassifier object and the fit() method from the scikit-learn API. You can also specify model parameters using the object() function during training.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, December 2022
Impact Factor: 6.252

IV. RESULT AND DISCUSSION

4.1 Experimental Results for Data Collection

	id	NumDots	SubdomainLevel	PathLevel	UrlLength	NumDash	NumDashInHostname	AtSymbol	TildeSymbol	NumUnderscore	Nu
0	1	3	1	5	72	0	0	0	0	0	
1	2	3	1	3	144	0	0	0	0	2	
2	3	3	1	2	58	0	0	0	0	0	
3	4	3	1	6	79	1	0	0	0	0	
4	5	3	0	4	46	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	
9995	9996	3	1	1	50	0	0	0	0	0	
9996	9997	2	1	4	59	1	0	0	0	0	
9997	9998	2	1	4	57	0	0	0	0	0	
9998	9999	3	1	1	49	0	0	0	0	0	
9999	10000	3	1	2	52	3	0	0	0	0	

10000 rows × 50 columns

Fig 1: Dataset collection

4.2 Experimental Results of Data Analysis

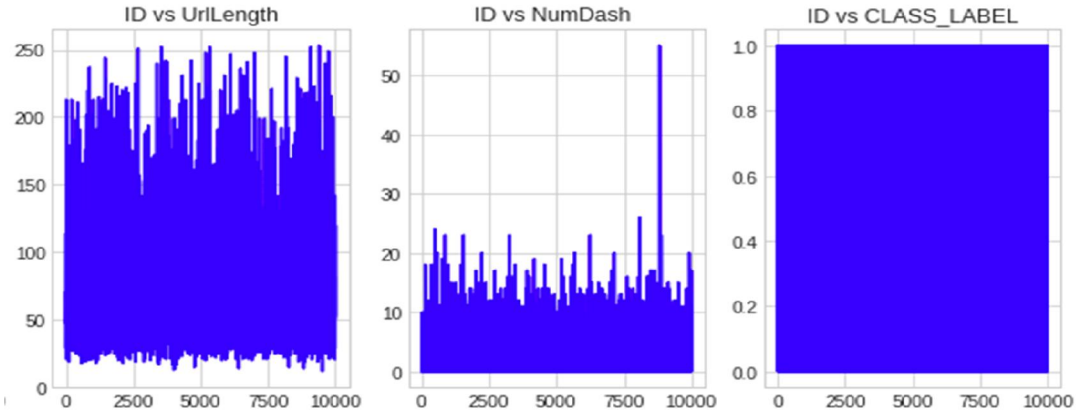


Fig 2. line plot for url_length, numdah, and class_label against id

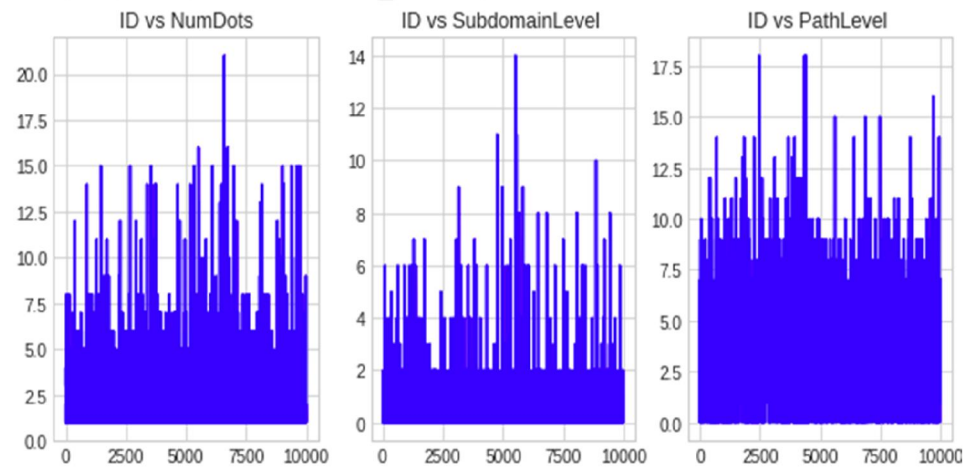


Fig 3. line plot for num_dots, sub_domain_level, and path_level against id

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, December 2022
Impact Factor: 6.252

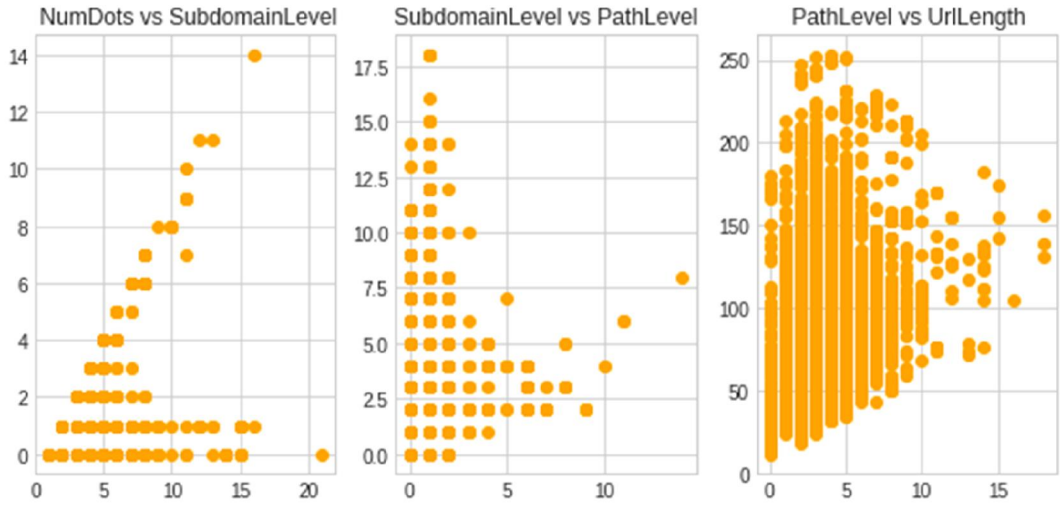


Fig 4. Scatter plot for numdots vs subdomain, sub_domain_level vs path_level
Path_level vs url_length

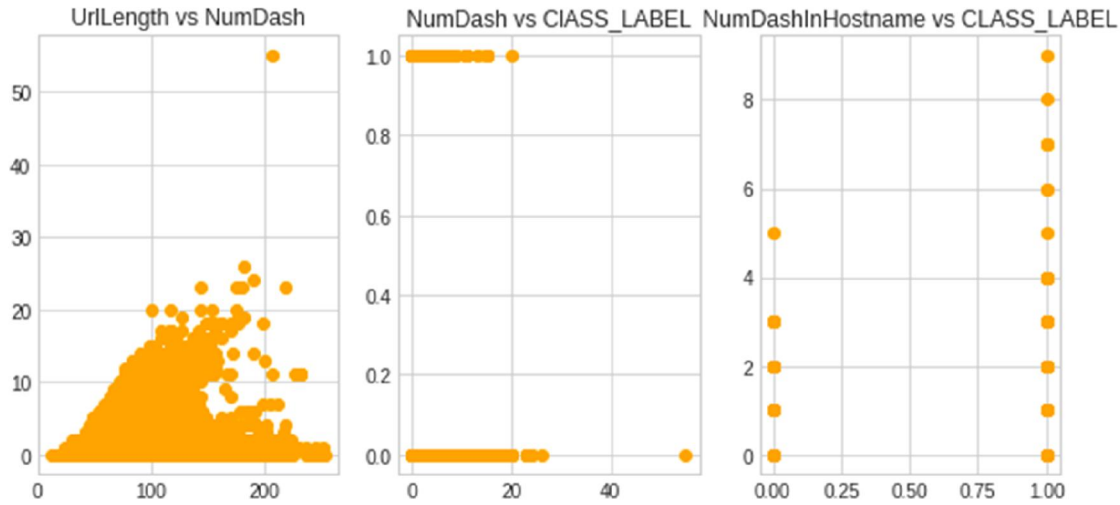


Fig 5. Scatter plot for Url_length vs Num_dash , Num_dash vs class_label ,
num_dash_hostname vs class_label

DOI: 10.48175/IJARSC-7822

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, December 2022
Impact Factor: 6.252

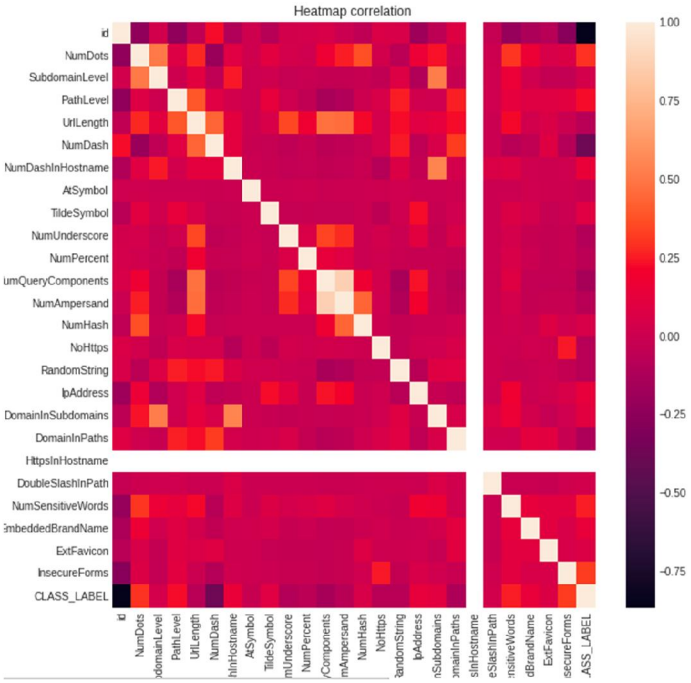


Fig 6. Correlation map of the dataset

4.3 Experimental Results of Data Preprocessing

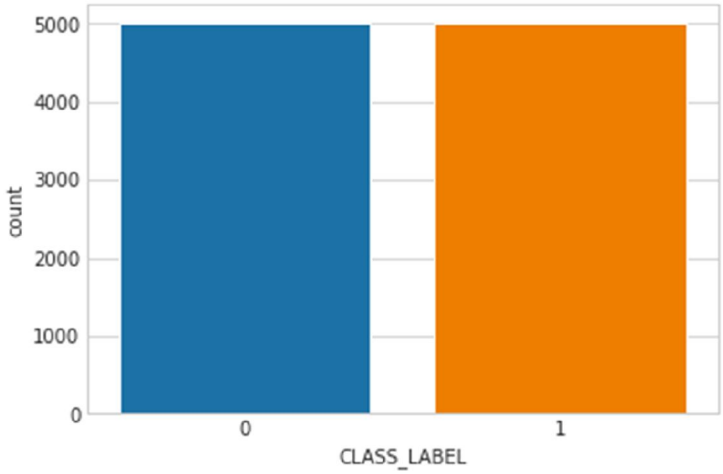


Fig 7 . count of phishing and legitimate URLs

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, December 2022
Impact Factor: 6.252

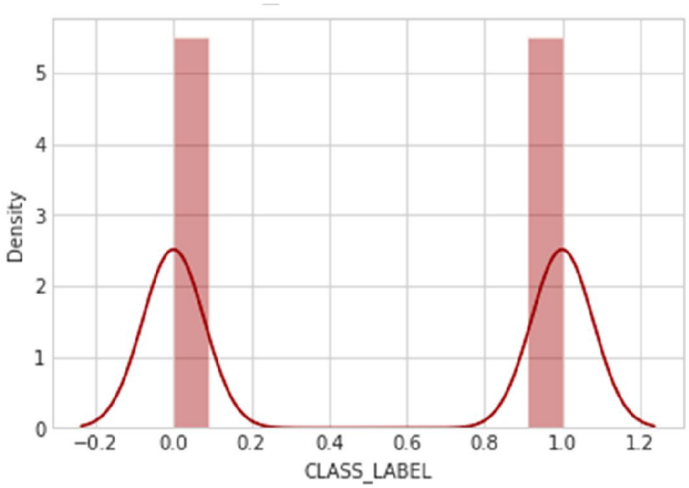


Fig 8. The density of class_label consists of phishing and legitimate URLs

4.4 Experimental Results of Model Training

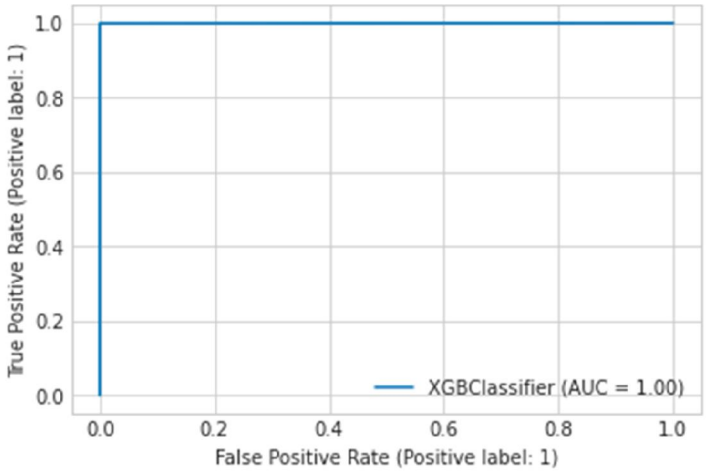


Fig 9: XGboost plot

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, December 2022
Impact Factor: 6.252

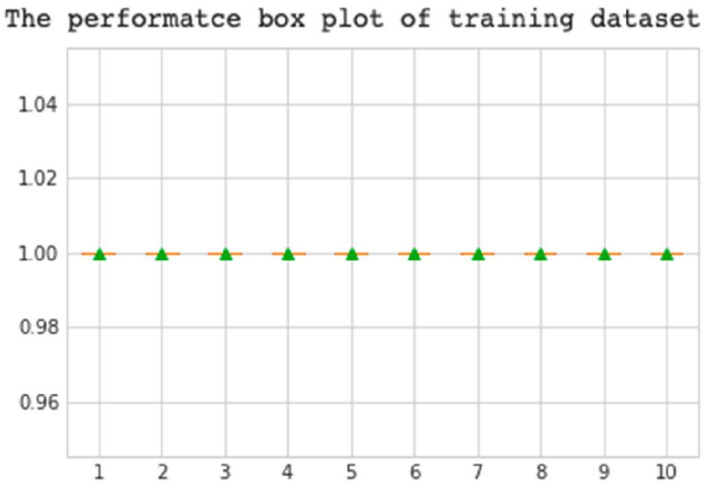


Fig 10 Performance plot of Adaboost ensemble depth method

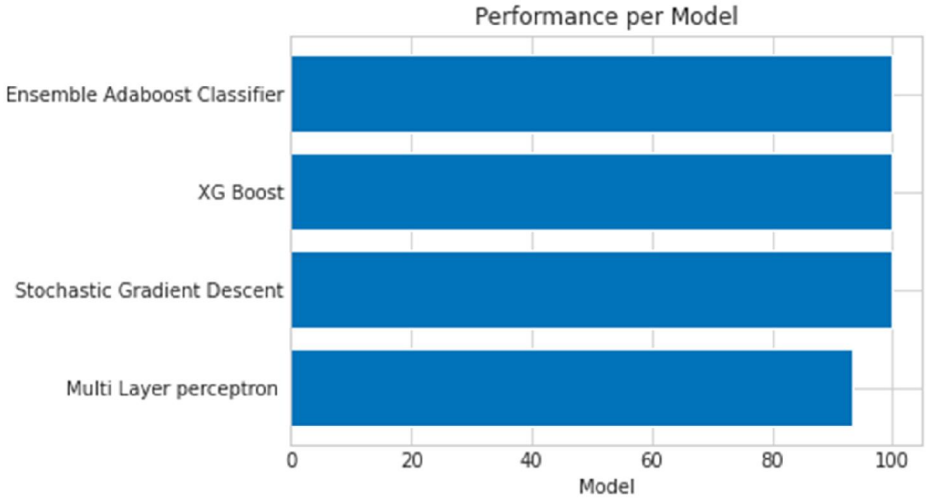


Fig 11 Performance per model analysis

V. CONCLUSION

Traditionally, classification machine learning models such as Support Vector Machines, Random Forest Classifiers, Decision Trees, and Logistic Regression have been used for detecting phishing websites, but these models are not always accurate enough. To improve the performance of phishing detection, advanced machine learning models such as the Multi-Layer Perceptron Classifier, AdaBoost, Ensemble Tree Depth Models, Stochastic Gradient Descent Classifier, and XGBoost have been developed. The proposed Ensemble Adaboost Classifier model uses feature selection and ensemble learning techniques to address the issue of overfitting and increase the accuracy of the model. As a future direction, a website could be developed to allow users to check for potentially malicious activity by detecting phishing links.

REFERENCES

[1]. Alyssa Anne Ubing , Syukrina Kamilia Binti Jasmi , Azween Abdullah, NZ Jhanjhi, Mahadevan Supramaniam, "Phishing Website Detection: An Improved Accuracy through Feature Selection and Ensemble Learning", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 1, 2019

[2]. Choon Lin Tan, Kng, Leng Chiew, Nah Sze, "Phishing website detection using URL-assisted brand name weighting system", International Symposium on Intelligent signal processing and communication

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, December 2022
Impact Factor: 6.252

systems,14885974, Dec 2014

[3]. Muhammad Rayhan Natadimadja, Maman Abdurohman, Hilal Hudan Nuha ,"A Survey on Phishing Website Detection Using Hadoop",Jurnal Informatika Universitas Pamulang,Vol. 5, No. 3, September 2020

[4]. Maher Aburrous, Adel Khelifi," Phishing Detection Plug-In Toolbar Using Intelligent Fuzzy-Classification Mining Techniques," The International Journal of Soft Computing and Software Engineering, San Francisco State University, CA, U.S.A., March 2013

[5]. Mustafa Al-Fayoumi, Jaber Al Widyan, and Mohammad Abusaif," Intelligent Association Classification Technique for Phishing Website Detection", The International Arab Journal of Information Technology, Vol. 17, No. 4, July 2020.

[6]. Atharva Deshpande , Omkar Pedamkar , Nachiket Chaudhary , Dr. Swapna Borde, "Detection of Phishing Websites using Machine Learning" he International Arab Journal of Information Technology,Vol. 10 ,No. 5, May 2021

[7]. Ashit kumar dutta,"Detecting phishing websites using machine learning technique" International Journal of Advanced Computer Science and Applications,Vol. 15, No. 9, 2021 october

[8]. Rishikesh Mahajan ,Irfan Siddavatam ,"Phishing Website Detection using Machine Learning Algorithms",International Journal of Computer Applications ,Volume 181 – No. 23, October 2018

[9]. Malaika Rastogi , Anmol chhetri , Divyanshu kumar singh, "Survey on Detection and Prevention of Phishing Websites using Machine Learning",International Conference on Computing Communication Control and automation ,2016

[10]. Jian mao , Jingdong Bian ,Shishi Zuh , "International conference of identification,information and knowledge in internet of things",2018

[11]. Vahid Shahrivari, Mohammad mahdi darabi , Mohammad Izadi ,"Phishing Detection Using Machine Learning Techniques",International Symposium on Intelligent signal processing and communication systems, 20220

[12]. R. M. Mohammad, F. Thabtah, and L. McCluskey, "phishing websites features", School of Computing and Engineering, University of Huddersfield, 2015.