Phishing Website Detection Using Machine Learning Algorithms

Parvathy R
Department of MCA
Amal Jyothi College Of Engineering
Kanjirappally, Kerala
parvthyr2022@mca.ajce.in

Ms. Grace Joseph
Asst. Professor in MCA
Amal Jyothi College Of Engineering
Kanjirappally, Kerala
gracejoseph@amaljyothi.ac.in

***Abstract*— Phishing is a type of identity fraud that involves carrying sensitive information including usernames, watchwords, bank account figures, and credit card figures. People working in the field of cyber security are now looking for reliable and consistent phishing website detection solutions. The purpose of this research is to apply machine learning to detect phishing URLs by extracting and analysing different features of genuine URLs. Decision trees, KNN, logistic regression, random forest, and support vector machine algorithms are used to detect phishing websites.. The goal of the study is to find the optimal machine learning algorithm by comparing accuracy rates, false positives, and false negatives.**

***Keywords*— Phishing, machine learning algorithms**

I. INTRODUCTION

Phishing is a type of cybercrime that involves establishing a fake website that looks like a legitimate website in order to obtain vital or confidential information from consumersThe delicacy of relating a genuine website is advanced than that of relating a phoney website. Attackers' primary goal is to steal bank account passwords. Businesses lose $2 billion per year as a result of their clientele falling victim to phishing. The public is the target of a phishing assault. Phishing attacks are tough to defend against because they prey on victims' vulnerabilities. It's critical to improve phishing detection techniques.

The "blacklist" method is a broadway for detecting phishing websites by adding blacklisted URLs and IP addresses to the antivirus database. To get around blacklists, attackers utilise clever techniques like obfuscation and many more basic techniques like fast-flux, where proxies are automatically built to host the web page, algorithmic production of new URLs, and so on. The major drawback of this method is that it cannot detect zero-hour phishing attacks.

Heuristic-primarily based totally strategies extract one or greater attributes from a web site to pick out phishing on any list. The majority of those traits are derived from the web site's URL and HTML Document Object Model (DOM).The extracted features are compared with known features collected from phishing and validated pages to decide its legitimacy. To overcome this drawback, many researchers focused on machine learning techniques Machine learning is a collection of algorithms that use historical data to make decisions or forecast future data. The algorithm will uncover numerous banned and valid URLs, as well as features to accurately detect phishing websites, including 0-hour phishing websites, using these strategies.

II. MACHINE LEARNING ALGORITHMS

We are using a five machine learning classification model.

a. Decision tree

The decision tree is one of the most extensively used machine literacy algorithms. It's a supervised learning-based classification model. It's easy to grasp and put into practise. It's a tree-like structure with an internal node that contains the dataset's features, branches that reflect decision rules, and a leaf node that represents the outcome. The decision or test is based on the characteristics of the presented dataset.

b. KNN

The simplest machine learning algorithm based on supervised learning approaches is the K-Nearest Neighbour algorithm. The algorithm finds the nearest data points in the training set to predict a new data point. K is the positive number of the nearest neighbour. Euclidean distance is the distance between two nearest neighbours.

The Euclidean distance between two points A and B can be calculated using the formula: $d(A,B) = \sqrt{\sum_{i=1}^{n} (B_i - A_i)2}$

c. Logistic Regression

DOI: 10.5281/zenodo.6329727
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

One of the maximum substantial device studying algorithms is logistic regression, which falls below the device studying approach. It predicts the output of a specific established variable. 0 & 1 are the results usually noted with logistic regression. In logistic regression, the sigmoid function is used to calculate the likelihood of valid and invalid classes.

Sigmoid function ,P $\frac{1}{1+e-(a+bx)}$

d. Random Forest

It is a device getting to know set of rules used for class and regression data. Random forest algorithm creates the forest with a number of decision trees. A high number of trees gives high detection accuracy. Any type of dataset can be used in this algorithm. It functions by creating a host of decision trees while training and outputting the mode o the classes or the regression of discrete trees.

e. Support Vector Machine

SVM is a effective device studying set of rules used for each regression and type challenges. In SVM, every statistics set is plotted in n-dimensional space. We carry out type with the aid of using locating the hyperplane that differentiates specific classes.

III. DATA SET DESCRIPTION

We use a python program to extract the features of URLs.

1. Using IP

Using IP addresses in URLs indicates the bushwhacker is trying to steal sensitive information. IP address represented in URL is set to 1 or 0. *Long URL*

Links to the webpage that has a long URL. For example, the URL http://sharif.hud.ac.uk/ bit.ly/1sSEGTB.

2. Short URL

Short URL service allows phishers to hide long phishing URLs by making them short. The goal is to redirect users to phishing websites.

3. Symbol@

If @ symbol is present in the URL then the feature is set to 1 else set to 0.

4. Prefix Suffix-

Phishers try to add prefixes or suffixes to the domain name separated by (-) to make users believe they are dealing with a legitimate website.

5. Subdomains

There are subdomains in the URL.

6. HTTP

Having HTTPS in URLs. For eg, http://https-www-mellat-phish.ir

7. DomainRegLen

Phishing website lives for a short period.

8. Favicon

It is a graphic webpage associated with web pages. If the favicon is displayed in the address bar, the webpage is most certainly a Phishing attempt.

9. NonStdPort

To manipulate intrusions, it's miles tons higher to open ports which you need.

10. RequestURL

The Request URL test determines whether the external elements on a webpage, such as photos, videos, and music, are loaded from a different domain.

11. AnchorURL

An anchor is an element defined by the< a> label. This feature is treated exactly as a Request URL.

12. LinkInScriptTag

Meta tags are generally utilized by legit web sites to offer metadata approximately HTML documents, Script tags to generate a client-facet script, and Link tags to fetch different internet resources.

13. ServerFormHandler

If the domain name is different from the domain name of the webpage.

14. InfoEmail

The phisher redirects the information to his email.

15. AbnormalUrl

It is extracted from the WHOIS database.

16. StatusBarCust

Use JavaScript to show a fake URL in the status bar to users.

17. DisableRightClick

It's the same as when you use onMouseOver to hide the Link.

DOI: 10.5281/zenodo.6329727
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

18. *WebsiteTraffic*

By calculating the number of visitors, this function determines the popularity of the website.

19. *IFrameRedirection*

The IFrame is an HTML tag that allows you to insert another webpage into the one you're now viewing

20. *.AgeOfDomain*

If the age of the domain is less than a month.

21. *DNSRecording*

Have the DNS record.

22. *PageRank*

Page rank is a fee starting from zero to 1. PageRank ambitions to degree how essential a website is at the Internet.

23. *GoogleIndex*

It examines whether the website is a google indicator or not.

24. *LinksPonitingToPage*

Links pointing to the page.

25. *StatsReport*

If the IP belongs to top phishing IPs or not.

IV. IMPLEMENTATION AND RESULT

To import system studying algorithms, we use a jupyter notebook. The overall performance of professed classifiers is lessoning set and testing. The overall performance of classifiers is estimated in expressions of delicacy, perfection and recall.

Figure 1, Classifiers performance

Classifiers	Accuracy	Precision	Recall
LogisticRegression	0.93	0.90	0.94
K-NN	0.64	0.58	0.59
Decision tree	0.95	0.93	0.95
Random Forest	0.97	0.96	0.98
SVM	0.56	0	0

From the above figure, we can conclude that random forest has better detection accuracy rather than decision tree and support vector machine.

It also indicates that as more datasets are utilised as training datasets, the detection accuracy of phishing websites improves. All classifiers perform better performance when 90% are training datasets.

V. CONCLUSION

The thing of this exploration is to develop a machine literacy-grounded discovery system for phishing websites. Using the arbitrary timber fashion, we were suitable to attain a discovery delicacy of 97 per cent. In addition, the results demonstrate that classifiers perform better when more data is utilised as training data. In the future, hybrid technology will be used to use the random forest algorithm of machine learning technology to detect phishing websites more precisely.

VI. REFERENCES

[1] Nevon projects, Detecting phishing website using machine learning algorithms, https://nevonprojects.com/detecting-phishing-websites-using-machine-learning/

[2]Research Gate, Phishing website detection using machine learningalgorithm, https://www.researchgate.net/publication/328541785_Phishing_Website_Detection_using_Machine_Learning_Algorithms

[3] Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013

[4]Phishingwebsites https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016

[5]IJERT, Detecting of phishing websites using machine learning techniques, https://www.ijert.org/detection-of-phishing-websites-using-machine-learning

DOI: 10.5281/zenodo.6329727
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam