International Journal of Engineering Research and Advanced Technology (IJERAT)

DOI: 10.31695/IJERAT.2021.3732

E-ISSN : 2454-6135

Volume.7, No. 9
September -2021

Detect Phishing Website by using Machine Learning

Ali Aljaberi[1], Osman Ucan[2]

[1]Department of Information Technology, Istanbul, Turkey

[2]Department of Electrical and Computer Engineering

Altinbas University Istanbul, Turkey

ABSTRACT

Phishing is one of the types of electronic crimes, where the attacker uses what is called social engineering to deceive users of Internet networks and this is done by sending messages via e-mail, phone call or text messages by the attacker who pretends to the victim that he is a real and legitimate company or institution that provides a specific service, And thus luring people to write their personal data in addition to important and sensitive information such as bank accounts, credit cards and passwords used by individuals. Then the attacker simply uses this data and information to obtain the property and accounts of the victim, and on the other hand, the attacker can monitor everything related to the victim during his entry and movement on the sites, that is why we developed a model that detects phishing by using the random forest algorithm.

Key Words: *Phishing, Random Forest Algorithm, Machine Learning, URL*

1. INTRODUCTION

The Internet has an important and distinctive role in all life and daily activities in various parts of the world, such as commercial activities, business management, job applications, communications, shopping ... etc It is therefore necessary for those who carry out these activities to have access to the websites of these institutions on the Internet, but the internet does not provide strong security which makes this an opportunity for the weak of souls to get financial gain or steal important data , that is why phishing has become an incentive for many attackers to continue committing these crimes, since cybercrime has risks and many rewards. Anti-fraud and phishing group reported in 2006 that cyber-attacks are constantly increasing, causing a lot of financial losses to individuals and organizations[1]. most attacks are done by sending e-mail messages to Internet users, which makes the recipient of the message believe that it is from a trusted institution or entity and asks him to re-evaluate his personal information and confirm this, perhaps by password and sensitive information, so the phishing achieves his goals and the victim falls into the phishing trap The most important sites that a phishing uses to steal information and deceive the victim are: Facebook, Twitter or Instagram[2].

2. PROBLEM STATEMENT

Suspicious sites have increased, especially after the development of technology and the multiplicity of social networking sites, which led to an increase in the risks of phishing, as Internet users suffer from the problem of theft of bank accounts as well as their passwords, as scammers penetrate accounts through their malicious programs, and due to the increase in breaches and frauds, they must There is a way to reduce these malicious programs. In this thesis, we focus on developing a model that detects phishing links to reduce phishing problems as much as possible.

3. LITERATURE REVIEW

E-phishing is designed to attack individuals, institutions and companies that have money stores or important databases to provide a specific service to many customers in various businesses and in large areas around the world. This is done by sending e-mail messages containing malicious links. When you click on these suspicious and sent links By the attackers, pages are opened that are completely identical to the pages of well-known legitimate entities and organizations[3]. Through these pages, which he designed specifically to trap the victim, the attacker writes some sensitive data, such as typing personal information, bank account number, or passwords of the victim, and slandered by using methods of enticement that the user will receive a prize after completing the writing of the data, or he uses the method of intimidation where the user is deceived into having to Retype his data and confirm the password, and if he does not do so, his account will be closed and he will be prevented from entering the site[4]. Therefore, scientists and researchers launched electronic crimes of this type with phishing.

Then the attacker uses that data he obtained by phishing to steal the victim's property or use the data in the espionage process and monitor the victim when he moves in various websites. After the continuous increase in electronic attacks through phishing with the advancement of technology that was launched over the Internet, new and innovative methods appear for users of phishing crimes. For this reason, blacklists are used to control and overcome fraudulent attacks and detect harmful links[5]. These lists need to be constantly updated and for this reason it has been suggested A new structure for blacklists, where this structure works by comparing the name of the page or website that was sent by the attacker via e-mail, which impersonates a well-known legitimate entity with the original name of the page of that entity, and this comparison is done through search engines such as Google In the event that the fake website name matches the real website name of the legitimate entity, this is considered a link to a real page, and if no match occurs and the real page is not found, then this link is considered to be a deceptive fake web page and this message is classified as spam phishing.

4. *DATA COLLECTION*

Bengin URL was compiled from two sources :Open Directory project (DMOZ) and Yahoo!'s directory Malicious URLs .The following sources are the approved ones, where the spam was obtained from jwSpam Spy It is called spam We have tested approximately 421436 random links, and 76,453 of them were classified as unsafe links. The number of safe links was 326133 .

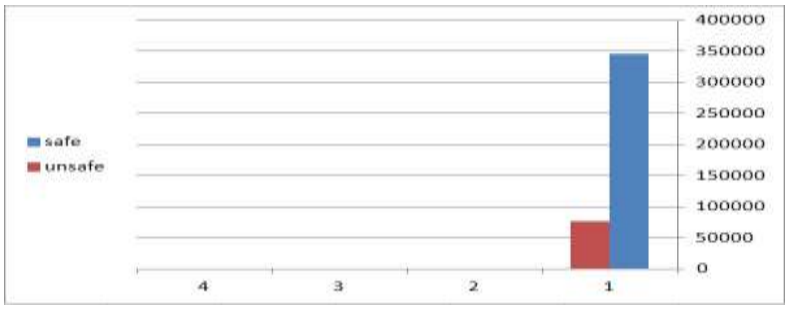


Figure1.1: Graph Based on Dataset Features

5. RESULTS

In Machine learning arrays are an important part, it using for testing of performance of algorithm, matrix consists of the table that includes information with detail of the current (human predicted) with categoration of predictive all columns in a matrix represents (Predictive group), and each row represents (Actual class). And by using the data contained in the matrix, the worker's performance is evaluated. The size of the matrix depends on the number of classes .

The figure below shows the shape of the matrix and the meaning of each column and row in it : In Machine learning arrays are an important part, it using for testing of performance of algorithm, matrix consists of the table that includes information with detail of the current (human predicted) with categoration of predictive all columns in a matrix represents (Predictive group), and each row represents (Actual class). And by using the data contained in the matrix, the worker's performance is evaluated[5]. The size of the matrix depends on the number of classes .The figure below shows the shape of the matrix and the meaning of each column and row in it.

DOI : 10.31695/IJERAT.2021.3732

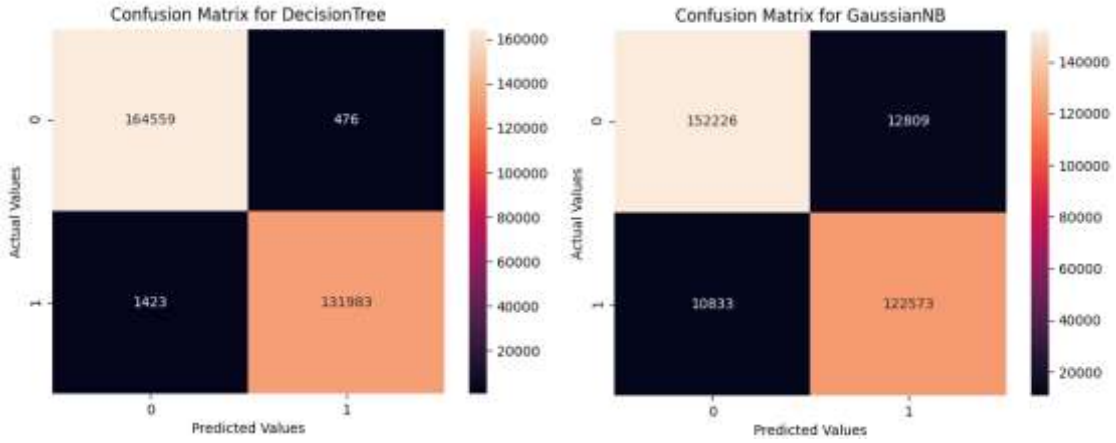
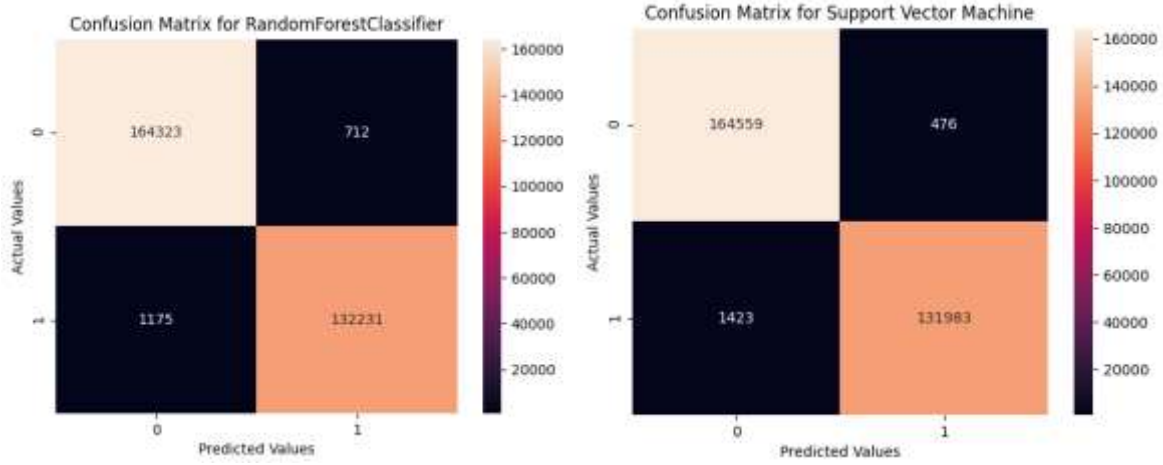


Figure 2 . The shape of the matrix



TP	FP
FN	TN

Figure 3. The shape of the matrix

True Positive (TP): It represents the number of cases that the classifier correctly predicts of the p-class and actually belonging to the p-class. The negative false (FN): an instance number that a classifier incorrectly predicts represents class p and actually belongs to class n. The negative true (TN): cases of numbers that a classifier correctly predicts represents class n and actually belongs to class p. The Positive false (FP): instances of number that a classifier incorrectly predicts represents class n and actually belongs to class p.

6. CONCLUSION

Phishing is attack to steal sensitive data from victims' and important information of internet users , so we must be reduced this risk by Developing model that detect phishing URLS and prevent them from steal password account or financial accounts , Therefore, in this study we presented a model using the Machine learning with python to detect phishing by using random forest classifier and we compared it with another three methods of classification, Decision tree matrix, Gaussian naïve Bayes and support vector machine, we found that random forest classifier is the best one of them and it gave us a high accuracy than the results of the rest of the other classifiers.

REFERENCES

[1] J. Jang-Jaccard and S. Nepal, 'A survey of emerging threats in cybersecurity', *Journal of Computer and System Sciences*, vol. 80, no. 5, pp. 973–993, Aug. 2014, doi: 10.1016/j.jcss.2014.02.005

DOI : 10.31695/IJERAT.2021.3732

[2] M. Selvakumari, M. Sowjanya, S. Das, and S. Padmavathi, 'Phishing website detection using machine learning and deep learning techniques', *J. Phys.: Conf. Ser.*, vol. 1916, no. 1, p. 012169, May 2021, doi: 10.1088/1742-6596/1916/1/012169.

[3] A. Alswailem, B. Alabdullah, N. Alrumayh, and A. Alsedrani, 'Detecting Phishing Websites Using Machine Learning', in *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, Riyadh, Saudi Arabia, May 2019, pp. 1–6. doi: 10.1109/CAIS.2019.8769571.

[4] M. H. Alkawaz, S. J. Steven, and A. I. Hajamydeen, 'Detecting Phishing Website Using Machine Learning', in *2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*, Langkawi, Malaysia, Feb. 2020, pp. 111–114. doi: **10.1109/CSPA48992.2020.9068728**

[5] T. Nathezhtha, D. Sangeetha, and V. Vaidehi, 'WC-PAD: Web Crawling based Phishing Attack Detection', in *2019 International Carnahan Conference on Security Technology (ICCST)*, CHENNAI, India, Oct. 2019, pp. 1–6. doi: 10.1109/CCST.2019.8888416.

DOI : 10.31695/IJERAT.2021.3732