Detecting Phishing Websites Using Machine Learning

Aniket Garje[1], Namrata Tanwani[1], Sammed Kandale[1], Twinkle Zope[1], Prof. Sandeep Gore[2]

[1] *UG Students,*[2]*Assistant Professors, Computer Engineering Department,*
G H Raisoni College of Engineering and Management, Pune

ABSTRACT

Phishing is a type of cybersecurity attack that involves stealing personal information such as passwords, credit card numbers, etc. To avoid phishing scams, we have used Machine learning techniques to detect Phishing Websites. Therefore, in this paper, we are trying to find the total number of ways to find Machine Learning techniques and algorithms that will be used to detect these phishing websites. We are using different Machine Learning algorithms such as KNN, Naive Bayes, Gradient boosting, and Decision Tree to detect these malicious websites. The research is divided into the following parts. The introduction represents the focused zone, techniques, and tools used. The Preliminaries section has details of the preparation of the information that is required to move further. Later the paper emphasizes the detailed discussion of the sources of information.

Keywords— Algorithms, Cybersecurity, Machine Learning, Phishing

I. INTRODUCTION

As mentioned in [1], a very high amount of data is generated each day, IBM has concluded that each day 2.5 quintillion bytes of data are produced. That's why we need to pay attention to the problems and difficulties related to the private and secure transfer of data. Cybersecurity is now one of the important fields of computer science that involves avoiding and protecting user data from threats and attacks. It makes sure that users should not fall for such attacks and become a victim of Cyber-crimes [2].

The term 'crime' means an illegal act that is offensive and is punishable by the state [3]. This is harmful to a person or society, per se [4]. When a thing, person, or idea belongs to a part of computer and information technology they are termed as 'cyber' [5]. This said, Cyber-crime hardly has anything to do with the law [6]. At its core, it involves a network of computers, mobile phones, laptops that are also referred to as communicating nodes that are involved in the transferring of data from and to the target nodes [7].

The attacks or threats which involve cybercrime consist of committing frauds, trafficking child pornography, stealing identities, violating the privacy, etc. [8]. Amongst them, Phishing has become the most organized crime of the 21 st century. Since more than 60% of commercial transactions are done online that's why cybersecurity attacks are most effective. It is a cybersecurity attack in which attackers get hold of confidential information of the user (unknowingly to the user). This information involves login credentials, bank-related credentials, credit card or debit card details, etc. To get access to sensitive information of user emails, messages, and clone websites are used[9].

Different methods have been invented to tackle this issue of social engineering called "Phishing" but, most of them were unsuccessful. Machine learning is one of those methods which is successful in detecting phishing websites. URLs have certain patterns and behaviors which are used by machine learning to find whether a website is phishing or not[10]. Hence we intend to find a solution to detect phishing websites using Machine Learning techniques.

II. LITERATURE SURVEY

This section of the literature survey eventually reveals some facts based on thoughtful analysis of many authors' work as follows.

[11] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. "Machine Learning-Based Phishing Detection from URLs," Expert Systems with Applications, 117:345-357, January 2019:The dataset used is self-constructed. Where phishing websites belong to PhishTank and legitimate URLS are from Yandex Search API. The main purpose was to detect the word which is similar to brand names, to detect keywords, the words, which are formed from random characters. Various classification algorithms such as Naive Bayes, Random Forest, kNN(n=3), Adaboost, K-star, SMO, and Decision Tree including some feature extraction types such as NLP-based features, Word Vectors and Hybrid are used. This system gets high accuracy throughout the test.

[12] J. James, Sandhya L. and C. Thomas, "Detection of phishing URLs using machine learning techniques," International Conference on Control Communication and Computing (ICCC), December 2013:The system proposed used a method based on lexical features, host properties, and properties related to the page for the detection of phishing websites. For getting a proper understanding of the pattern of URLs, various data mining algorithms are used. The classification algorithms such as Naïve Bayes, J48 Decision Tree, K-NN, and SVM were considered for the detection of phishing websites. Decision Tree had better accuracy of 91.08% compared to other algorithms. So Tree-based classifiers are best suited for phishing URL classification.

[13] Pradeepthi, K. V., & Kannan, A. "Performance study of classification techniques for phishing URL detection," Sixth

International Conference on Advanced Computing (IcoAC), December 2014. : The system recognizes Phishing URLs, by examining the URL structure without attending the Phishing URL using classification algorithms. The data collected is first passed through the training state where it undergoes feature selection and classification. The dataset used here contains 4500 URL records, on which classification is performed. Out of which 2500 URLs are genuine and the other 2000 are the phishing ones. The 2500 URLs were collected from the DMOZ repository. The 2000 phishing URLs have been picked from PHISHTANK. Data classification after extraction of the relevant features was performed by Naive Bayes, Random Forest, Random Tree, Multi-layer Perceptron, C-RT, J 48 Tree, LMT, C 4.5, ID 3, and K-Nearest Neighbour. The Random Forest Algorithm had the highest classification accuracy.

[14] Dipayan Sinha, Dr. Minal Moharir, Prof. Anitha Sandeep, "Phishing Website URL Detection using Machine Learning," International Journal of Advanced Science and Technology, 29(3):2495-2504, 2020. : Detection of phishing websites is performed by using machine learning techniques like Logistic Regression, Decision tree, Random Forest, Adaboost, Gradient Boosting, Gaussian NB, and Fuzzy pattern tree classifier. Data collection involves phishing and legitimate websites. Extracting useful features has two steps: URL-based involves IP Address, '@' symbol in URL, dashes in URL, long URL, presence of unusual number, dot count, sub-domains in URL, etc. Domain-based includes Page Rank of the website, age of the Domain, and Validity of the Website. The dataset is split into training and testing set in the ratio 80:20. The Random Forest algorithm shows 96% of precision and recalls along with the highest F1 score of 95%.

[15] R. Kiruthiga, D. Akila, "Phishing Websites Detection Using Machine Learning," International Journal of Recent Technology and Engineering (IJRTE), 8(2S11):11-114, September 2019:2019 A total of 15 research papers have been studied in this research paper. In this research, one method was discussed which uses five different algorithms that are Decision Tree, Generalized Linear Model, Gradient Boosting, Generalized Additive Model, and Random Forest. On comparing the results, the Random Forest algorithm had the highest accuracy of 98.4%, 98.59% recall, and precision of 97.70%. Dataset used is from the UCI machine learning repository.

.

III. PROPOSED SYSTEM

The proposed system uses Machine Learning to combat the issue of Phishing. Following are the algorithms used in this approach:

1) Nearest Neighbour: It is a supervised machine learning algorithm used for classification and regression problems. This technique is used here with following hyperparameters:

'leaf_size' = list(range(25,40)),
'n_neighbors' = list(range(5,20)),
'p' = [1,2]

Knn - Confusion Matrix :
[[3186 284]
[247 4021]]

2) Naive Bayes: Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. It is a simple technique also called Gaussian Naive Bayes that is used with default hyperparameters.

Naive Bayes - Confusion Matrix :
[[3470 0]
[3097 1171]]

3) Decision Tree: A decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. The algorithm is used with the following hyperparameters:

'max_leaf_nodes' = list(range(2, 100)),
' min_samples_split' = [2, 3, 4]

Decision Tree - Confusion Matrix :
[[3193 277]
[210 4058]]

4) Gradient Boosting: Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Following are the hyperparameters used:

'min_samples_leaf' = np.linspace(0.1, 0.5, 12), 'max_depth' = [3,5,8],
'max_features' = ["log2","sqrt"],
'n_estimators' = [10]

Gradient Boosting - Confusion Matrix :
[[2880 590]
[107 4161]]

IV. RESULT

1) ROC curve of KNN:

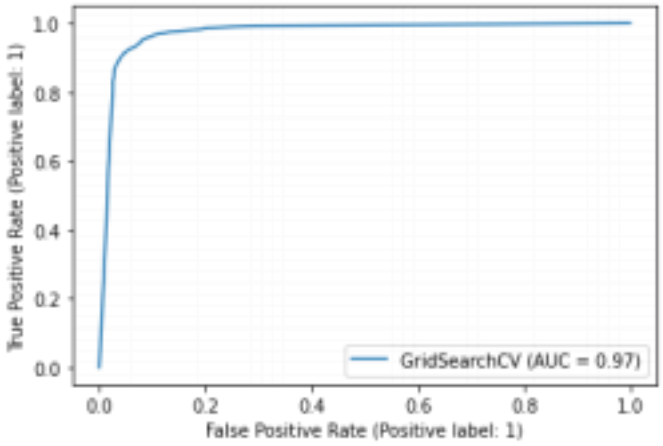


Fig. 3

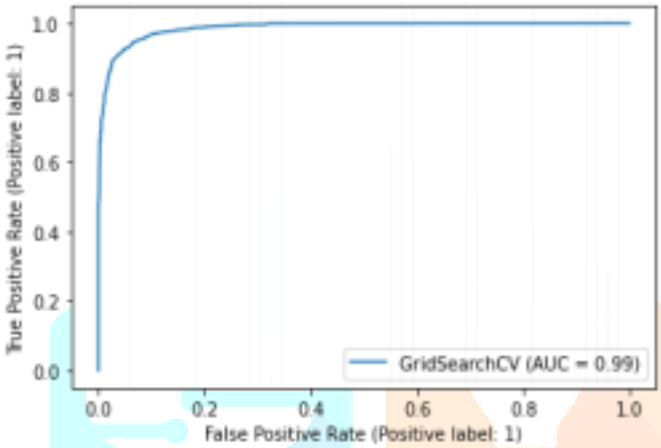


Fig. 1

2) ROC curve fo Naive Bayes:

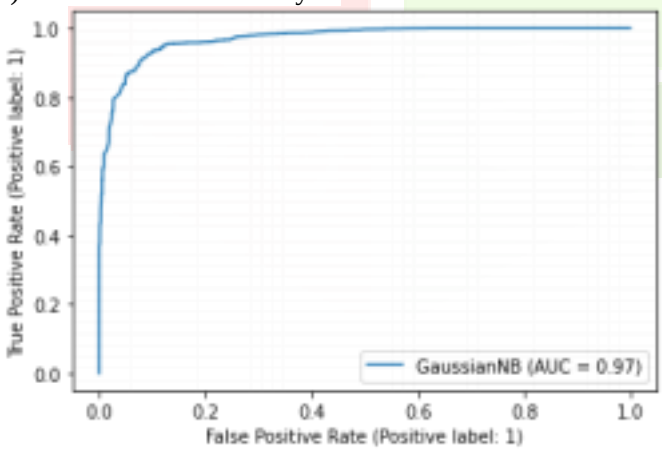
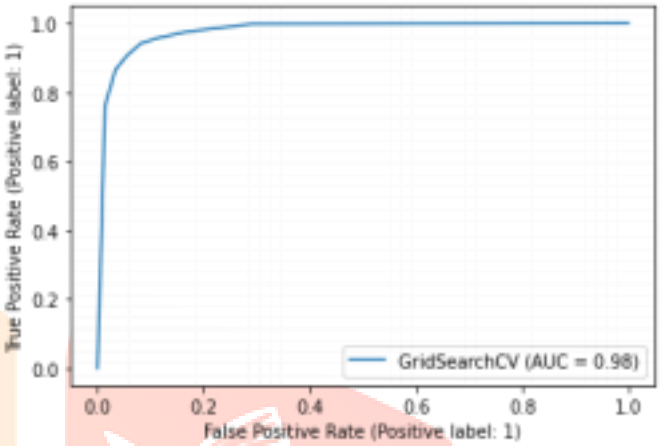


Fig. 2

3) ROC curve of Decision Tree:

Fig. 4

4) ROC curve of Gradient Boosting:



URL, Dashes in URL, Long URL, presence of unusual number, Dot Count, Sub-domains in URL, etc. Domain-based feature selection includes Page Rank of the Website, Age of the Domain, and Validity of the Website. In Implementation, the first step is to process the data. Dataset Exploration is done along with selecting useful attributes for further process. The dataset was split into training and testing set in the ratio 80:20. The training set with the extracted features was given as input to different machine learning classification algorithms. The accuracy of classification along with precision, recall, and F1 score was determined using
different algorithms mentioned above. Area Under the Curve of ROC curve was also determined. The results were then compared and analyzed for the different models.

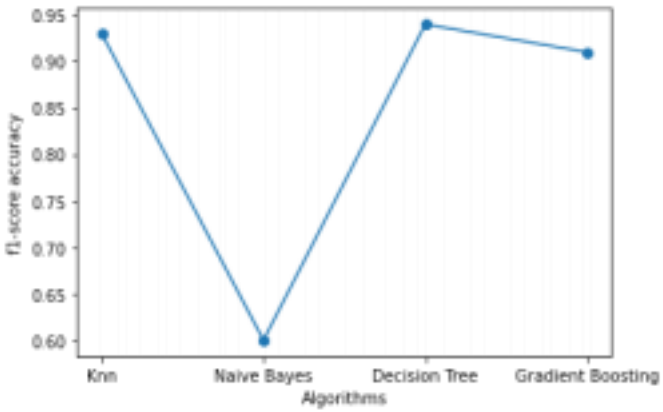


Fig. 5 Comparison of f1-Score accuracy of all algorithms

V. CONCLUSION:

Detection of phishing websites is performed using machine learning algorithms like KNN, Naive Bayes, Decision tree, Gradient Boosting. In the data collection phase, the data is collected on both phishing and legitimate websites. Then comes extracting the useful features of the given dataset. It involves two steps: URL-based features and Domain-based Features. URL-based feature selection involves IP Address, '@' symbol in The above figure shows that Decision trees give the best f1

score, that is, 0.94, a good balance between recall and precision, and has a pretty decent area under the curve of ROC. Hence, Decision trees are preferred over others.

References

[1] R. Devakunchari, "Analysis on big data over the years," International Journal of Scientific and Research Publications (IJSRP), vol. 04, no. 01, January 2014.

[2] Nikhita Reddy, G.J. Ugander Reddy, "A Study Of Cyber Security Challenges And Its Emerging Trends On Latest Technologies," International Journal of Engineering and Technology, vol. 4, no.1, January 2014.

[3] Larry Sanger, "Crime",en.wikipedia.org/wiki/Crime, September 20, 2001.

[4] Esther Ramdinmawii, Seema Ghisingh, Usha Mary Sharma, "A Study on the Cyber-Crime and Cyber Criminals: A Global Problem," International Journal of Web Technology, vol 04, pp. 53-57, June 2015.

[5] "Cyber",merriam-webster.com/dictionary/cyber, January 2021.

[6] TechTarget Contributor, "Cyber", searchsoa.techtarget.com/definition/cyber, April 05, 2005.

[7] Sharma, Ushamary and Ghisingh, Seema and Ramdinmawii, Esther, "A Study on the Cyber - Crime and Cyber Criminals: A Global Problem," International Journal of Web Technology, vol 03, pp. 172-179, June 2014.

[8] Andrewa, "Cybercrime", http://en.wikipedia.org/wiki/Computer_crime, October 15, 2003.

[9] Vayansky, I. and Kumar, S., "Phishing – challenges and solutions.", Computer Fraud & Security, vol 2018, no. 1, pp. 15-20, January 2018.

[10] Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi, "Phishing Detection Using Machine Learning Techniques," unpublished.

[11] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. "Machine Learning-Based Phishing Detection from URLs," Expert Systems with Applications, vol. 117, pp. 345-357, January 2019.

[12] J. James, Sandhya L. and C. Thomas, "Detection of phishing URLs using machine learning techniques," International Conference on Control Communication and Computing (ICCC), December 2013.

[13] Pradeepthi, K. V., & Kannan, A. "Performance study of classification techniques for phishing URL detection," Sixth International Conference on Advanced Computing (IcoAC), December 2014.

[14] Dipayan Sinha, Dr. Minal Moharir, Prof. Anitha Sandeep, "Phishing Website URL Detection using Machine Learning," International Journal of Advanced Science and Technology, vol. 29, no. 3, pp. 2495-2504, 2020.

[15] R. Kiruthiga, D. Akila, "Phishing Websites Detection Using Machine Learning," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2S11, pp. 11-114, September 2019