K-NEAREST NEIGHBOUR CLASSIFIER FOR URL-BASED PHISHING DETECTION MECHANISM

Subba Reddy Borra[1], B Gayathri[2], B Rekha[2], B Akshitha[2], B. Hafeeza[2]

[1,2]Department of Information Technology

[1,2]Malla Reddy Engineering College for Women (A), Maisammaguda, Medchal, Telangana.

ABSTRACT

Phishing attack is the simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password, and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This project deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. In addition, the main motive of this research is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy of each algorithm.

Keywords: Phishing attack, K-nearest neighbour, Machine learning.

1. INTRODUCTION

1.1 Overview

Phishing is a fraudulent technique that uses social and technological tricks to steal customer identification and financial credentials. Social media systems use spoofed e-mails from legitimate companies and agencies to enable users to use fake websites to divulge financial details like usernames and passwords [1]. Hackers install malicious software on computers to steal credentials, often using systems to intercept username and passwords of consumers' online accounts. Phishers use multiple methods, including email, Uniform Resource Locators (URL), instant messages, forum postings, telephone calls, and text messages to steal user information. The structure of phishing content is similar to the original content and trick users to access the content in order to obtain their sensitive data. The primary objective of phishing is to gain certain personal information for financial gain or use of identity theft. Phishing attacks are causing severe economic damage around the world. Moreover, most phishing attacks target financial/payment institutions and webmail, according to the Anti-Phishing Working Group (APWG) latest Phishing pattern studies [1]. In order to receive confidential data, criminals develop unauthorized replicas of a real website and email, typically from a financial institution or other organization dealing with financial data. This e-mail is rendered using a legitimate company's logos and slogans. The design and structure of HTML allow copying of images or an entire website. Also, it is one of the factors for the rapid growth of Internet as a communication medium, and enables the misuse of brands, trademarks, and other company identifiers that customers rely on as authentication mechanisms. To trap users, Phisher sends "spooled" mails to as many people as possible. When these e-mails are opened, the customers tend to be diverted from the legitimate entity to a spoofed website.

There is a significant chance of exploitation of user information. For these reasons, phishing in modern society is highly urgent, challenging, and overly critical. There have been several recent studies against phishing based on the characteristics of a domain, such as website URLs, website content, incorporating both the website URLs and content, the source code of the website and the screenshot of the website. However, there is a lack of useful anti-phishing tools to detect malicious URL in an organization to protect its users. In the event of malicious code being implanted on the website, hackers may steal user information and install malware, which poses a serious risk to cybersecurity and user privacy.

1.2 Problem Statement

Phishing assault is a most straightforward approach to get delicate data from honest clients. Point of the phishers is to obtain basic data like username, secret key, and ledger subtleties. Network safety people are currently searching for dependable and consistent location methods for phishing sites recognition. To overcome the drawbacks of blacklist and heuristics-based method, many security researchers now focused on machine learning techniques. Machine learning technology consists of many algorithms which requires past data to decide or prediction on future data. Using this technique, algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites including zero- hour phishing websites.

2. LITERATURE SURVEY

Phishing attacks are categorized according to Phisher's mechanism for trapping alleged users. Several forms of these attacks are keyloggers, DNS toxicity, Etc., [2]. The initiation processes in social engineering include online blogs, short message services (SMS), social media platforms that use web 2.0 services, such as Facebook and Twitter, file-sharing services for peers, Voice over IP (VoIP) systems where the attackers use caller spoofing IDs [3, 4]. Each form of phishing has a little difference in how the process is carried out in order to defraud the unsuspecting consumer. E-mail phishing attacks occur when an attacker sends an e-mail with a link to potential users to direct them to phishing websites.

Phishing websites are challenging to an organization and individual due to its similarities with the legitimate websites [5]. There are multiple forms of phishing attacks. Technical subterfuge refers to the attacks include Keylogging, DNS poisoning, and Malwares. In these attacks, attacker intends to gain the access through a tool/technique. On the one hand, users believe the network and on the other hand, the network is compromised by the attackers. Social engineering attacks include Spear phishing, Whaling, SMS, Vishing, and mobile applications. In these attacks, attackers focus on the group of people or an organization and trick them to use the phishing URL [6, 7]. Apart from these attacks, many new attacks are emerging exponentially as the technology evolves constantly. Phishing detection schemes which detect phishing on the server side are better than phishing prevention strategies and user training systems. These systems can be used either via a web browser on the client or through specific host-site software [8, 9].

3. PROPOSED SYSTEM

3.1 K-Nearest Neighbor (KNN) Algorithm

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

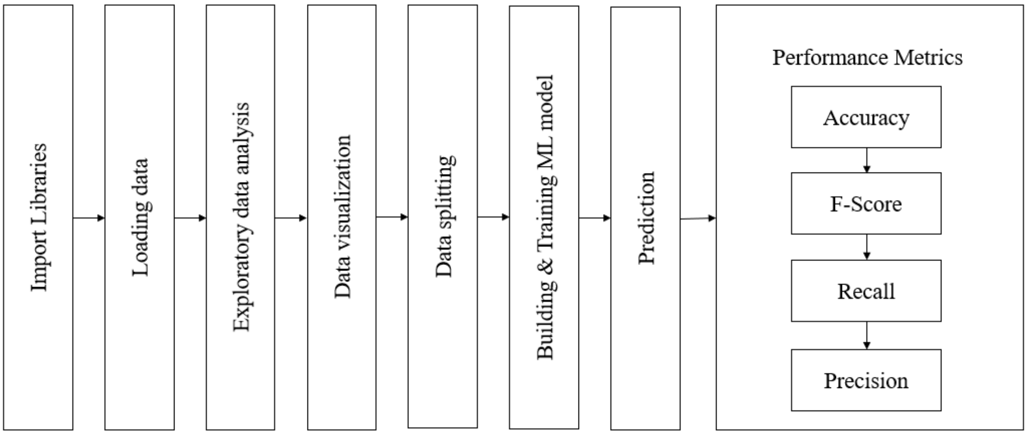


Fig. 1: Block diagram of proposed system.

Steps

1. Get labeled data: The labeled data consists of features and labels. Features are the characteristics or the property of the object whereas labels are the class of the object with those features.
2. Convert labeled data to encoded data: Usually computations are based on numerical form so we convert the data to numeric form by encoding them.
3. Create feature set: Creating a set of features by packing the features.
4. Split the data for train and test: The data are split training and testing. Usually, 80% for training and 20% for testing but can select based on need.
5. Train the classifier: Training the classifier with the training data by specifying the value of k. Use k =3 for binary classification, i.e., two labels classification. If used k =1 then it is simply a nearest neighbor classifier.
6. Test the classifier: Testing the classifier with the testing data.
7. Evaluate: Evaluating the classifier using confusion matrix and its evaluation metrics i.e., accuracy, precision, recall, et cetera.

3.2 Advantages of proposed system

- Simple and easy to implement.
- Training phase is faster due to lazy learning.
- Suitable for multi class problem.
- Works better for continuously changing data due to instance-based learning.

4. RESULTS AND DISCUSSION

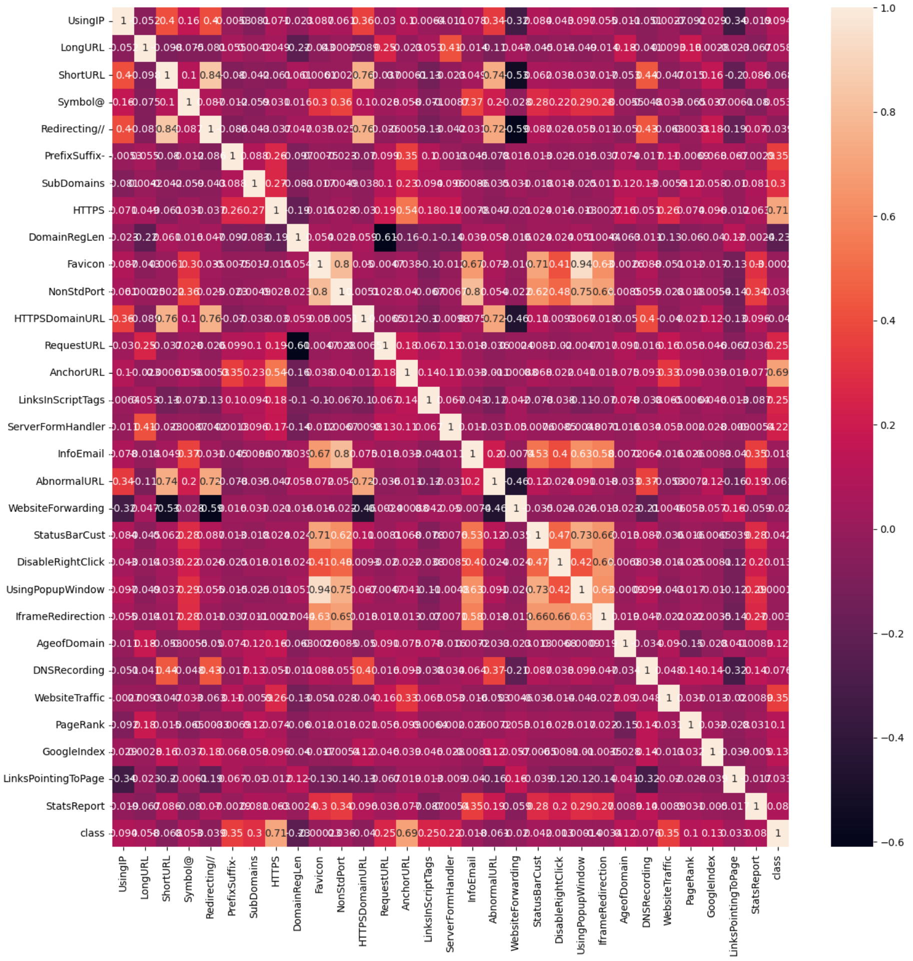
1. Loading Data
   - The dataset is collected from open-source platform
   - A collection of website URLs for 11000+ websites. Each sample has 30 website parameters and a class label identifying it as a phishing website or not (1 or -1).
   - The overview of this dataset is, it has 11054 samples with 32 features.
2. EDA
   - In this step, few dataframe methods are used to look into the data and its features.
3. Data Visualization¶
   - Few plots and graphs are displayed to find how the data is distributed and the how features are related to each other.
4. Data Splitting
   - The data is split into train & test sets, 80-20 split.
5. Building and Training ML Model

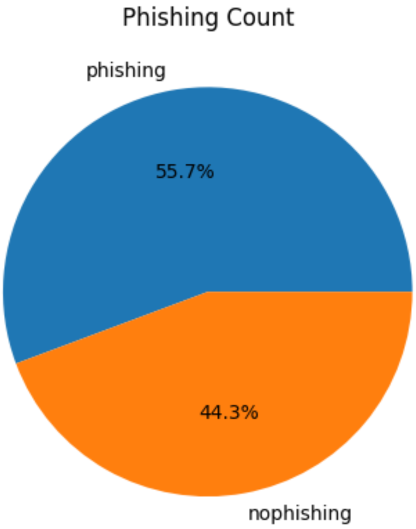
Supervised machine learning is one of the most commonly used and successful types of machine learning. Supervised learning is used whenever we want to predict a certain outcome/label from a given set of features, and we have examples of features-label pairs. We build a machine learning model from these features-label pairs, which comprise our training set. Our goal is to make accurate predictions for new, never-before-seen data.

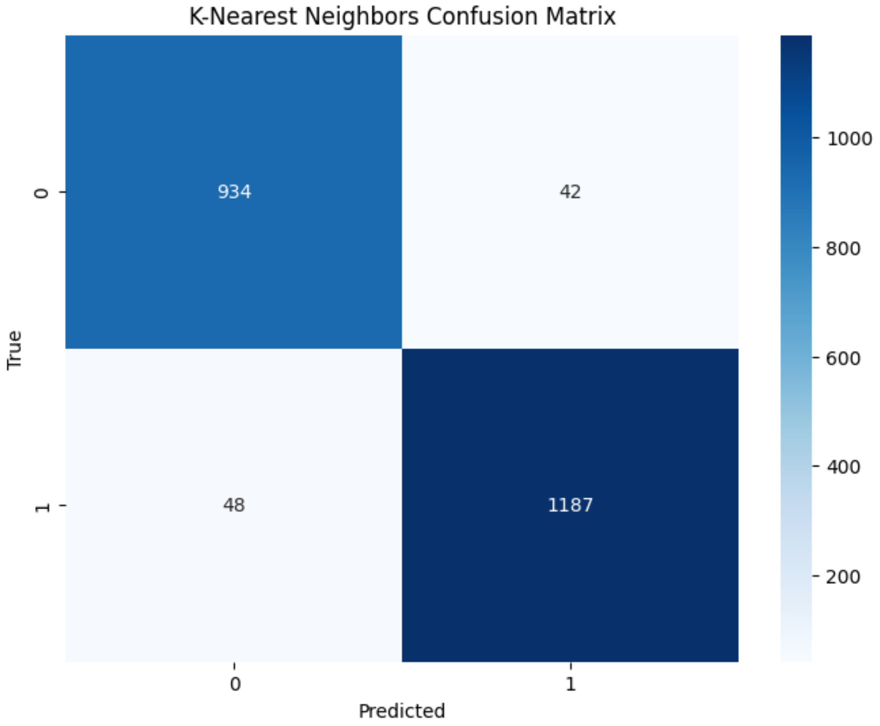
There are two major types of supervised machine learning problems, called classification and regression. Our data set comes under regression problem, as the prediction of suicide rate is a continuous number, or a floating-point number in programming terms. The supervised machine learning models (regression) considered to train the dataset in this notebook are:

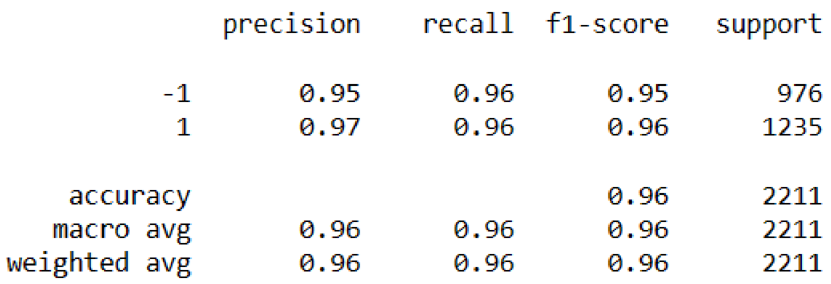
- Logistic Regression
- k-Nearest Neighbors

The metrics considered to evaluate the model performance are Accuracy & F1 score.


UsingIP
LongURL
ShortURL
Symbol@
Redirecting//
PrefixSuffix-
SubDomains
HTTPS
DomainRegLen
Favicon
NonStdPort
HTTPSDomainURL
RequestURL
AnchorURL
LinksInScriptTags
ServerFormHandler
InfoEmail
AbnormalURL
WebsiteForwarding
StatusBarCust
DisableRightClick
UsingPopupWindow
IframeRedirection
AgeofDomain
DNSRecording
WebsiteTraffic
PageRank
GoogleIndex
LinksPointingToPage
StatsReport
class
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Phishing Count
phishing
55.7%
44.3%
nophishing





```
              precision    recall  f1-score   support

          -1       0.95      0.96      0.95       976
           1       0.97      0.96      0.96      1235

    accuracy                           0.96      2211
   macro avg       0.96      0.96      0.96      2211
weighted avg       0.96      0.96      0.96      2211
```

5. CONCLUSION AND FUTURE SCOPE

Conclusion

Machine learning (ML) based phishing intrusion detection was proposed in this paper. The investigation utilizes many strategies to identify phishing intrusion detection. Standard datasets of phishing intrusion detection from kaggle.com were used as input for the ML algorithms. The machine learning algorithm KNN are implemented to analyze and select datasets for classification and detection. Principal component analysis (PCA) was applied to identify and classify the components of the datasets. KNN was used to both classify the website and classification. Finally, the confusion matrix was drawn to evaluate the performance of KNN algorithms. The random KNN achieved a high accuracy.

Future scope

In our future work, fishing attacks will be predicted from the logged dataset of attacks by using a convolution neural network (CNN). It will be added as a tool for intrusion detection system (IDS) and we plan to implement these solutions and develop a robust and generalized intrusion detection model.

REFERENCES

[1] Anti-Phishing Working Group (APWG), https://docs.apwg.org//reports/apwg_trends_report_q4_2019.pdf

[2] Jain A.K., Gupta B.B. "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning", Cyber Security. Advances in Intelligent Systems and Computing, vol. 729, 2018,

[3] Purbay M., Kumar D, "Split Behavior of Supervised Machine Learning Algorithms for Phishing URL Detection", Lecture Notes in Electrical Engineering, vol. 683, 2021,

[4] Gandotra E., Gupta D, "An Efficient Approach for Phishing Detection using Machine Learning", Algorithms for Intelligent Systems, Springer, Singapore, 2021, https://doi.org/10.1007/978-981-15-8711-5_12.

[5] Hung Le, Quang Pham, Doyen Sahoo, and Steven C.H. Hoi, "URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection", Conference'17, Washington, DC, USA, arXiv:1802.03162, July 2017.

[6] Hong J., Kim T., Liu J., Park N., Kim SW, "Phishing URL Detection with Lexical Features and Blacklisted Domains", Autonomous Secure Cyber Systems. Springer, https://doi.org/10.1007/978-3-030-33432-1_12.

[7] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran and B. S. Bindhumadhava, "Phishing Website Classification and Detection Using Machine Learning," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1–6, 10.1109/ICCCI48352.2020.9104161.

[8] Hassan Y.A. and Abdelfettah B, "Using case- based reasoning for phishing detection", Procedia Computer Science, vol. 109, 2017, pp. 281–288.

[9] Rao RS, Pais AR. Jail-Phish: An improved search engine-based phishing detection system. Computers & Security. 2019 Jun 1; 83:246–67.