KLE Society's

KLE Technological University



**Information Security Course Project Report**

**On**

**Phishing Website Detection Using Machine Learning**

**Bachelor of Engineering**

**In**

**Computer Science and Engineering**

SUBMITTED BY

| Name | USN |
| --- | --- |
| Laxman Bhajantri | 01FE19BCS095 |
| Vinayak Mane | 01FE19BCS107 |
| Suleman Navalur | 01FE19BCS114 |

**Under the guidance of**

**Pooja Shettar**

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

HUBLI–580031 (India).

The academic year 2022-23

**ABSTRACT**

Access control is important to protect network security. The increasing demand for secure interactions between network domains brings new challenges to access control technologies. With the rapid development of the informatization process, the demand to collaborate, share, and exchange information between different departments of information systems has become more and more imperative. By introducing the principle of need-to-share and the notion of multi-level objects, a multi-level security model of information systems based on the BLP model is proposed. Then, the formal description and security analysis of the multi-level security model is presented. The new security model allows information being securely shared with the right users and protected from the wrong users while maintaining the multi-level security of information systems.

***KEYWORDS:*** Phishing Websites, security, Machine Learning, Random Forest, Decision tree

Logistic Regression.

| **Chapter No.** | **TABLE OF CONTENTS** | | **Page No.** |
| --- | --- | --- | --- |
| **1.** | **INTRODUCTION** | | **4** |
|  | 1.1 | Introduction | **4** |
|  | 1.2 | Motivation | **5** |
|  | 1.3 | Objectives of the project | **5** |
|  | 1.4 | Problem Definition | **5** |
|  | 1.5 | Problems in the current system | **6** |
| **2.** | **LITERATURE SURVEY** | | **7-8** |
| **3.** | **PROPOSED MODELS** | | **9** |
|  | 3.1.1 | Decision Tree | **9** |
|  | 3.1.2 | Random Forest | **9** |
|  | 3.1.3 | Logistic Regression | **9** |
| **4.** | **IMPLEMENTATION** | | **10-11** |
| **5.** | **RESULT AND DISCUSSION** | | **12-13** |
| **6.** | **CONCLUSION** | | **14** |
| **7.** | **REFERENCES** | | **14** |

**1. INTRODUCTION**

* 1. **Introduction**

The Internet global connectivity is regarded as a crossroads where users can meet and share information. This is the primary reason phishers choose this method of data exchange as a point of contact to conduct widespread phishing attacks by infecting computers with spyware that directs people to bogus websites.

Phishing attacks are one of the most dangerous threats to online accounts and data because these attacks pose as a trustworthy firm or person, and they employ social engineering techniques to make victims more likely to fall for the scam. Phishing can be defined as impersonating a valid site to trick users by stealing their personal data comprising usernames, passwords, accounts numbers, national insurance numbers, etc.

Phishing frauds might be the most widespread cybercrime used today. There are countless domains where phishing attacks can occur like the online payment sector, webmail, and financial institutions, file hosting or cloud storage and many others. The webmail and online payment sector was embattled by phishing more than in any other industry sector. Phishing can be done through email phishing scams and spear phishing hence users should be aware of the consequences and should not give their 100 percent trust on common security applications.

Machine Learning is one of the efficient techniques to detect phishing as it removes drawbacks of existing approaches. The objective, which is the most vital thing in the proposed project, is to verify the validity of the website by capturing blacklisted URLs. To notify the user on a blacklisted website through pop-up while they are trying to access and to notify the user on a blacklisted website through email while they are trying to access. This proposed project will allow administrators to add blacklisted URLs in order to alert users during their inquiry.

The two scope of the project, which is well known as user scope and system scope. User has some responsibility towards the system. The system includes a few standards and policies that are required to be obliged in order to comply with the system. The user can be notified if a blacklisted website is being accessed. The admin can capture the blacklisted URLs to alert users.

* 1. **Motivation**

With the maturity of research on security models in stand-alone computer systems, it is only natural to attempt to extend these concepts to network security problems. Landwehr et al. proposed a more complete and conceptually more appealing development of a security model for a specific application of the military message system Liu et al. designed a type of secure Web server system based on the BLP model, which applies multi-level security mechanisms to the network environment. Si et al. developed an extension of BLP in the local area network, which controls communications between subjects by adding dynamic monitor units and resource servers to the system. Security modeling techniques have been widely used in the design of protection mechanisms for the operating systems of stand-alone computers.

Therefore, a security enforcement mechanism for information systems should be added to the existing protection mechanisms within individual computers. To meet the needs of information sharing and access control, we extend the classical BLP model of stand-alone computer systems to information systems, introduce the need-to-share category and multi-level object, and develop a multi-level security model of information systems for information sharing, called SBLP. The SBLP model can process information with different classification levels and provides a secure information-sharing service for users with different clearance levels.

* 1. **Objectives**
* To address fake URLs and domain names by identifying phishing website links.
* Reduce the rate of financial theft from individuals and organizations online.
* Educate Internet Users on the deception technique of phishers.
* Development of a web-based application for detection.
* Dataset collection and pre-processing.
* Machine learning model selection and development.
* Integration of the developed model to web application.
  1. **Problem Statement**

Phishing Website Detection using Machine learning.

The aim of this project is to detect phishing websites using machine learning and deep neural networks by developing a web application that allows users to check if a URL is phishing or legitimate and have access to resources to help tackle phishing attacks.

* 1. **Problems in the current system**

Phishing is a type of practice done on the Internet where individual data is obtained by illegal approaches. It provides the ability to obtain sensitive information, as an example, usernames, passwords, and positive identification points of interest, often for malignant reasons, by taking up the looks of an electronic correspondence. Phishing attacks will be enforced in varied kinds like Email phishing, website phishing, spear phishing, Whaling, Tab off his guard, Evil twin phishing etc.

Phishing is known as webpage violence. Phishing is often done by email spoofing or texting, and it typically guides users to enter points of interest at a fake website which look and feel the same. It tries to handle the increasing range of phishing that has to be met by clients in awareness and alternative efforts to ascertain protection of numerous anti-phishing tools.

The emerging technology industry which deeply influences today’s security problems has given a non-ease of mind to some employers and home users. Occurrences that exploit human vulnerabilities have been on the upsurge in recent years. In the new era there are many security systems being developed to ensure security is given the utmost priority and prevention to be taken from being hacked by those who are involved in cyber-criminal and essential prevention is also taken as high consideration in organization to ensure network security is not being breached.

Cyber security employees are currently searching for trustworthy and steady detection techniques for phishing websites detection. Due to wide usage of the internet to perform various activities such as online bill payment, banking transaction, online shopping, and, etc. Customers face numerous security threats like cybercrime. There are many cybercrime that are extensively executed for example spam, fraud, cyber terrorisms and phishing. Among this phishing is known as the popular cybercrime today. Phishing has become one amongst the highest 3 most current forms of law-breaking in line with recent reports, and both frequency of events and user susceptibility has enlarged in recent years, more combining the danger of economic damage.

**2.** **Literature Survey**

**In [1]** Phishing attacks target vulnerabilities that exist in systems due to the human factor. Many cyber attacks are spread via mechanisms that exploit weaknesses found in end-users, which makes users the weakest element in the security chain. The phishing problem is broad and no single silver-bullet solution exists to mitigate all the vulnerabilities effectively, thus multiple techniques are often implemented to mitigate speciﬁc attacks. This paper aims at surveying many of the recently proposed phishing mitigation techniques. A high-level overview of various categories of phishing mitigation techniques is also presented, such as: detection, offensive defense, correction, and prevention, which we believe is critical to present where the phishing detection techniques ﬁt in the overall mitigation process.

**In [2]** It is a crime to practice phishing by employing technical tricks and social engineering to exploit the innocence of unaware users. This methodology usually covers up a trustworthy entity so as to influence a consumer to execute an action if asked by the imitated entity. Most of the time, phishing attacks are being noticed by the practiced users but security is a main motive for the basic users as they are not aware of such circumstances. However, some methodologies are limited to look after the phishing attacks only and the delay in detection is mandatory. In this paper we emphasize the various techniques used for the detection of phishing attacks. We have also discovered various techniques for detection and prevention of phishing. Apart from that, we have introduced a new model for detection and prevention of phishing attacks.

**In [3]** The Internet dragged more than half of the world's population into the cyber world. Unfortunately, with the increase in internet transactions, cybercrimes also increase rapidly. With the anonymous structure of the internet, attackers attempt to deceive the end-users through different forms namely phishing, malware, SQL injection, man-in-the-middle, domain name system tunneling, ransomware, web trojan, and so on. Amongst them, phishing is the most deceiving attack, which exploits the vulnerabilities in the end-users. Phishing is often done through emails and malicious websites to lure the user by posing themselves as a trusted entity. Security experts have been proposing many anti-phishing techniques. Till today there is no single solution that is capable of mitigating all the vulnerabilities. A systematic review of current trends in web phishing detection techniques is carried out and a taxonomy of automated web phishing detection is presented. The objective of this study is to acknowledge the status of current research in automated web phishing detection and evaluate their performance. This study also discusses the research avenues for future investigation.

**In [4]** Phishing and fraud sites have been widespread on the internet in recent times, which's become a source of great concern and a serious cybersecurity problem, as internet fraudsters target sensitive data and personal information of users, especially the username and password. Numerous approaches have been proposed and used to prevent and reduce these phishing websites and attacks, and protect users and their privacy. In this paper, we categorized the present anti-phishing approaches into two main classes: Content-based and Non-content-based. The content-based approach is also classified into URL content analysis and webpage content analysis. This helps in finding out numerous anti-phishing techniques and algorithms to choose the best approach in future contributions.

**In [5]** URL based phishing attacks are one of the most common threats to internet users. In this type of attack, the attacker exploits human vulnerability rather than software flaws. It targets both individuals and organizations, induces them to click on URLs that look secure, and steal confidential information or inject malware on our system. Different machine learning algorithms are being used for the detection of phishing URLs, that is, to classify a URL as phishing or legitimate. Researchers are constantly trying to improve the performance of existing models and increase their accuracy. In this work we aim to review various machine learning methods used for this purpose, along with datasets and URL features used to train the machine learning models. The performance of different machine learning algorithms and the methods used to increase their accuracy measures are discussed and analyzed. The goal is to create a survey resource for researchers to learn the current developments in the field and contribute in making phishing detection models that yield more accurate results.

**3. PROPOSED MODELS**

**3.1 Decision Tree Algorithm**

One of the most widely used algorithms in machine learning technology. Decision tree algorithm is easy to understand and also easy to implement. Decision tree begins its work by choosing the best splitter from the available attributes for classification which is considered as a root of the tree. Algorithm continues to build the tree until it finds the leaf node. Decision tree creates a training model which is used to predict target value or class in tree representation each internal node of the tree belongs to attribute and each leaf node of the tree belongs to class label. In decision tree algorithms, gini index and information gain methods are used to calculate these nodes.

**3.2** **Random Forest Algorithm**

Random forest algorithm is one of the most powerful algorithms in machine learning technology and it is based on the concept of decision tree algorithm. Random forest algorithm creates the forest with a number of decision trees. High number of trees gives high detection accuracy. Creation of trees is based on the bootstrap method. In bootstrap method features and samples of the dataset are randomly selected with replacement to construct a single tree. Among randomly selected features, a random forest algorithm will choose the best splitter for the classification and like a decision tree algorithm.This process will continue until a random forest creates a number of trees. Each tree in the forest predicts the target value and then the algorithm will calculate the votes for each predicted target. Finally, the random forest algorithm considers a highly voted predicted target as a final prediction.

**3.3 Logistic Regression**

[Linear regression models](https://www.ibm.com/in-en/topics/linear-regression) are used to identify the relationship between a continuous dependent variable and one or more independent variables. When there is only one independent variable and one dependent variable, it is known as simple linear regression, but as the number of independent variables increases, it is referred to as multiple linear regression. For each type of linear regression, it seeks to plot a line of best fit through a set of data points, which is typically calculated using the least squares method.

**4. IMPLEMENTATION**

* Scikit-learn tool has been used to import Machine learning algorithms.
* Python language is used.
* Dataset is divided into training sets and testing sets in 70:30 ratio respectively.
* These Models consist of Decision Tree, Support Vector Machine and Random Forest.

**Benefits**

* Will be able to differentiate between phishing(0) and legitimate(1) URLs.
* Will help reduce phishing data breach for organizations.
* Will be helpful to individuals and organizations.
* Will be easy to use.

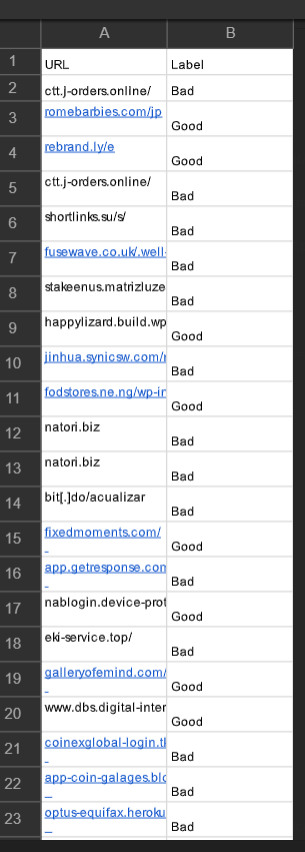
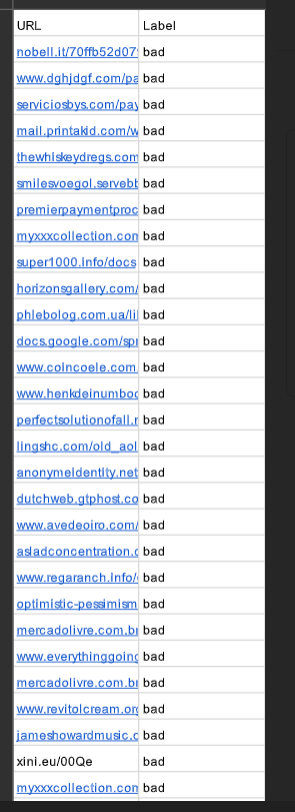
** **

Fig 4.1 Dataset

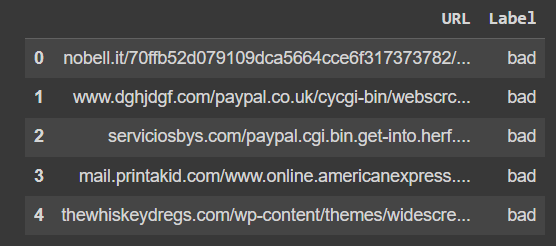
****

Fig 4.2 Loading Data in Colab

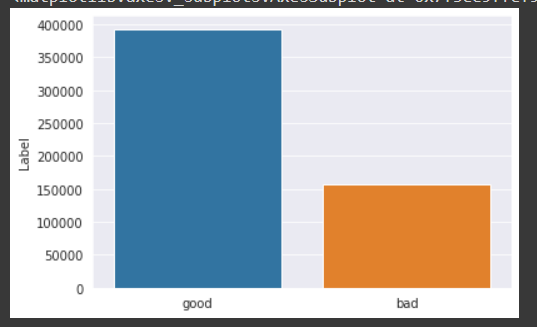


Fig 4.3 Visualizing the Classes.

**5. RESULT AND DISCUSSION**



Fig 5.1 Tokenized and Stemming the URL.

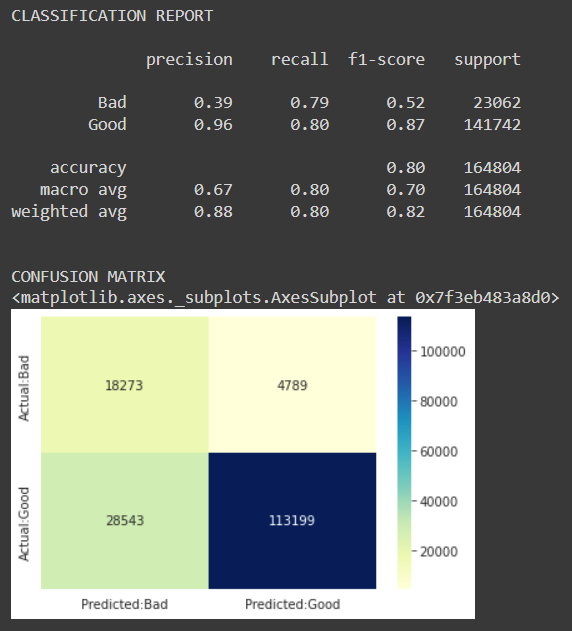


Fig 5.2 Confusion Matrix for Decision Tree.

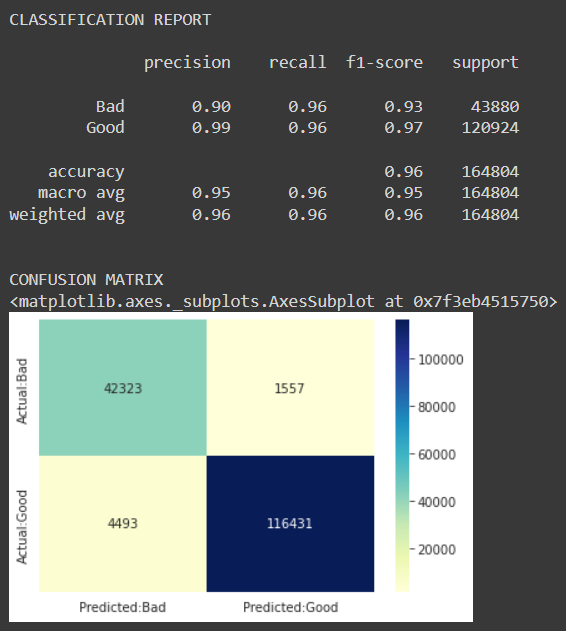
****

Fig 5.3 Confusion Matrix for Logistic Regression.

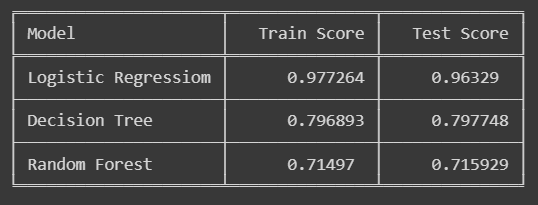
****

Fig 5.4 Accuracy of all Models.

**6. CONCLUSION**

The contribution of this research work is to help both individuals and organizations identify and understand phishing techniques used by phishers as well as help them detect phishing URL attacks effectively and efficiently by employing machine learning models instead of the previous method of detecting phishing websites.

**7. REFERENCES**

[1] Mahmoud Khonji, Youssef Iraqi, Senior Member, IEEE, and Andrew Jones - “Phishing Detection: A Literature Survey”.

[2] Pratik Patil , Prof. P.R. Devale “A Literature Survey of Phishing Attack”.

[3] [M. Vijayalakshmi](https://ietresearch.onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Vijayalakshmi%2C+M),[S. Mercy Shalinie](https://ietresearch.onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Mercy+Shalinie%2C+S),[Ming Hour Yang](https://ietresearch.onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Yang%2C+Ming+Hour),[Raja Meenakshi U.](https://ietresearch.onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=U.%2C+Raja+Meenakshi) “Web phishing detection techniques”.

[4] Rania Zaimi, Mohamed Hafidi, Mahnane Lamia “A literature survey on anti-phishing in websites”.

[5] Arathi Krishna V\*, Anusree A, Blessy Jose, Karthika Anilkumar, Ojus Thomas Lee - “Phishing Detection using Machine Learning based URL Analysis”.

[6] www.phishtank.com.