Abstract

Phishing attacks are 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 .machine learning technology is used for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, K-Means Clustering, Random forest and Naïve Bayes algorithms are used to detect phishing websites. Aim is to detect phishing URLs as well as narrow down to the best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

Keywords: *Machine learning, Phishing, Cyber Security, Cyber Awareness, Phishing Detection*

List of Abbreviations

1. ML: Machine Learning
2. DB: Database
3. URL: Uniform Resource Locator
4. HTTP: Hypertext Markup Transfer Protocol
5. HTTPS: Hypertext Markup Transfer Protocol Secure
6. DNS: Domain Name Server

List of Figures

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Background

Nowadays, Phishing has become a main area of concern for security researchers because it is not difficult to create fake websites which look so close to legitimate websites. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attacks. Main aim of the attacker is to steal bank account credentials. In United States businesses, there is a loss of US$2billion per year because their client become victims of phishing. In Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion. Phishing attacks are becoming successful because of lack of user awareness. Since phishing attacks exploit the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as the “blacklist" method. To evade blacklists attackers use creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including: fast-flux, in which proxies are automatically generated to host the web-page; algorithmic generation of new URLs; etc. Major drawback of this method is that it cannot detect zero-hour phishing attacks.

1.2 Motivation

Due to the rise of phishing attacks and phishing websites it is quite difficult to browse the web without any worries. This phishing attacks can be fatal as they try to steal sensitive information. Also it not so obvious to catch such websites as their UI/UX is exactly similar to original websites and manually detecting such websites can be time consuming. Hence, to avoid such phishing attacks we have developed phishing website detection system to help people detect such websites automatically.

Chapter 2

Literature Survey

2.1 Basic Terminologies

Phishing

Phishing is a type of cybersecurity attack during which malicious actors send messages pretending to be a trusted person or entity. Phishing messages manipulate a user, causing them to perform actions like installing a malicious file, clicking a malicious link, or divulging sensitive information such as access credentials. Phishing is the most common type of social engineering, which is a general term describing attempts to manipulate or trick computer users. Social engineering is an increasingly common threat vector used in almost all security incidents. Social engineering attacks, like phishing, are often combined with other threats, such as malware, code injection, and network attacks.

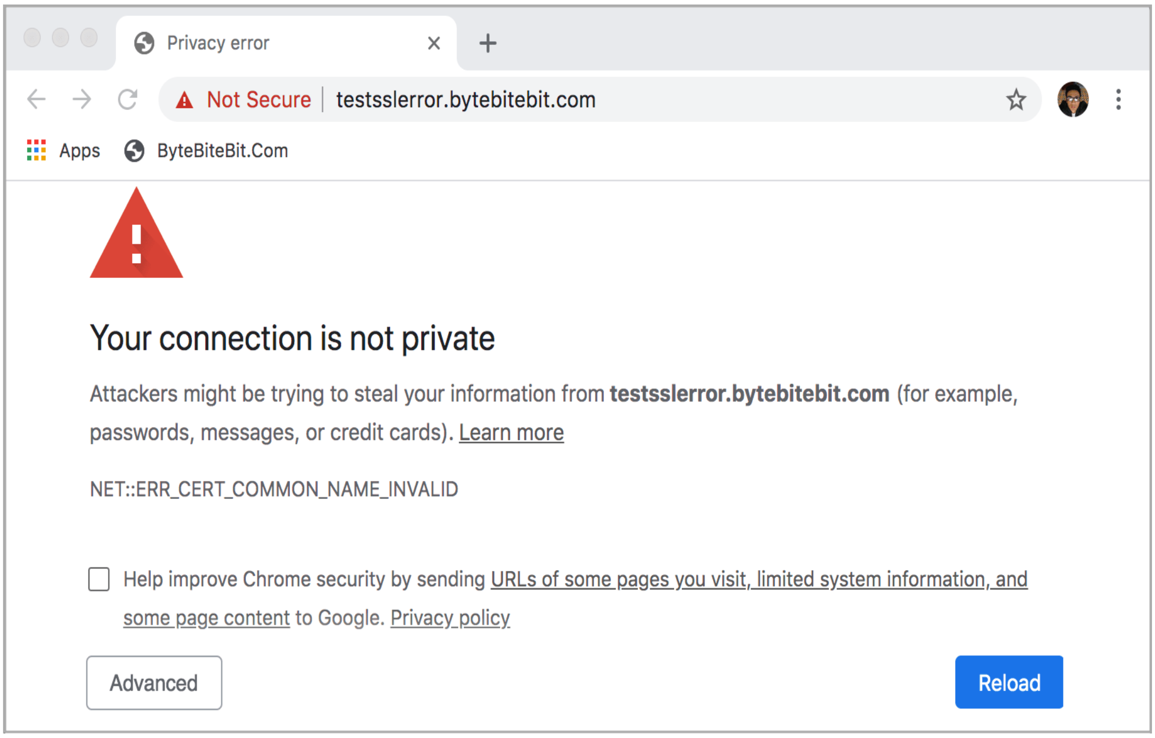
Machine Learning

Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

2.2 Existing System

In earlier systems users need to detect and block phishing websites manually, which is quite difficult nowadays because it very hard to tell the difference in legitimate website and the phishing website as they look exactly similar and it's not so obvious to tell the difference between them based on their UI.

In recent times most of the phishing attacks are done via emails as it is one of the most widely used technology and have become integral part of our life. Attackers sends various links to thousands of users everyday which leads users to phishing websites. Various software and spam filters can be used to avoid such type of attacks but it can be difficult to setup such software and they can also be costly sometimes. Also installation of anti-phishing and antivirus software can take excess of storage in system which will ultimately slow down the system.



Your connection is not private

Attackers might be trying to steal your information from **testsslerror.bytebitebit.com** (for example, passwords, messages, or credit cards). Learn more

NET::ERR_CERT_COMMON_NAME_INVALID

□ Help improve Chrome security by sending URLs of some pages you visit, limited system information, and some page content to Google. Privacy policy

Figure: 2.2.1 Existing System

2.3 Problem Statement

Phishing attacks are the simplest way to obtain them. Phishing is one of the techniques which are used by the intruders to get access to the user credentials or to gain access to the sensitive data. This type of accessing is done by creating the replica of the websites which looks the same as the original websites which we use on our daily basis but when a user clicks on the link he will see the website and think its original and try to provide his credentials. To prevent such attack creating a system for detection of phishing websites.

Chapter 3

Requirement Gathering

3.1 Details of Hardware & Software

1 Software Requirement

This project is built using Visual Studio Code and Google Collab.

- Language: Python
- Framework: Flask
- Library: sklearn, pandas

2 Hardware Requirement

We strongly recommend a computer fewer than 5 years old.

- Processor: Minimum 1 GHz; Recommended 2GHz or more
- Ethernet connection (LAN) OR a wireless adapter (Wi-Fi)
- Hard Drive: Minimum 32 GB; Recommended 64 GB or more
- Memory (RAM): Minimum 1 GB; Recommended 4 GB or above
- Rasberry-pi

Chapter 4

Plan of Project

4.1 Implemented System Architecture

To detect phishing website, we will collect the URL form user with help of web app and after the input of URL we will gather the required data of that website. After gathering of the data, we will proceed for the feature extraction step. In this step all 30 features of the websites are verified, Features which are packed in our program. Features like having IP Address, URL Lengthening, At symbol, double slash redirecting, prefix suffix, Request URL, pop up window, having sub domain, SSL certificate, Domain registration length, open ports, favicon, https token in url, Url of anchor, server from handler, submitting to email, abnormal url, site redirect, on mouse over change, right click disabled, iframe redirection, age of domain, DNS record, web traffic rank, page rank, google index, links pointing to page, and statistical reports etc. In next step various classification algorithms are used to train our module, algorithms such as Random forest, Naive Bayes etc. Among those algorithms which gives best accuracy is selected and used to train the module. After this process we will get an output of a Boolean value which will be 1 or 0. If output is 1 then website is phishing, if output is 0 then it is safe to visit the site without any worries. Based on this Boolean values, Output will be displayed to make user aware.

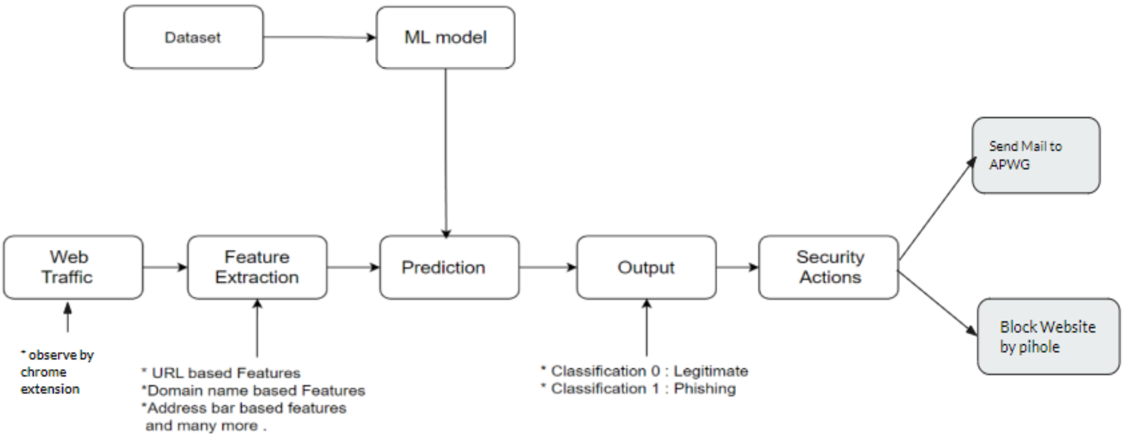
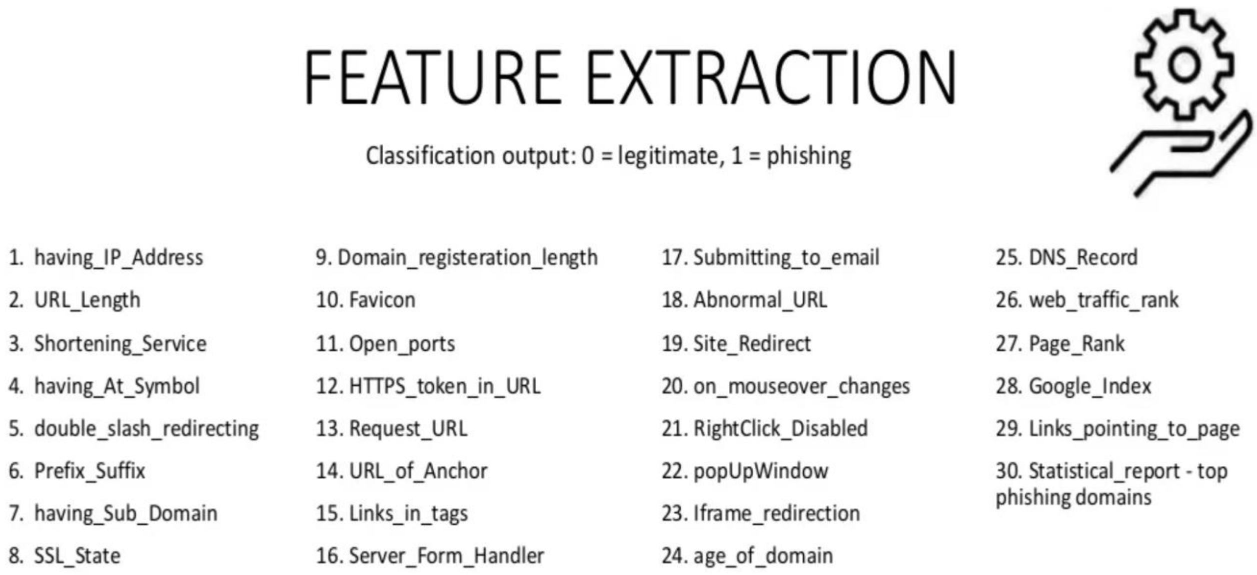

Dataset
ML model
Send Mail to APWG
Web Traffic
Feature Extraction
Prediction
Output
Security Actions
Block Website by pihole
* observe by chrome extension
* URL based Features
*Domain name based Features
*Address bar based features and many more .
* Classification 0 : Legitimate
* Classification 1 : Phishing

Figure: 4.1.1 System Architecture

4.2 Methodology.

Approach for building ML model for prediction system of phishing website start with feature extraction. In which 30 different features are consider for classification of phishing website these features play an important role in training the ML model.



FEATURE EXTRACTION

Classification output: 0 = legitimate, 1 = phishing

1. having_IP_Address
2. URL_Length
3. Shortening_Service
4. having_At_Symbol
5. double_slash_redirecting
6. Prefix_Suffix
7. having_Sub_Domain
8. SSL_State
9. Domain_registeration_length
10. Favicon
11. Open_ports
12. HTTPS_token_in_URL
13. Request_URL
14. URL_of_Anchor
15. Links_in_tags
16. Server_Form_Handler
17. Submitting_to_email
18. Abnormal_URL
19. Site_Redirect
20. on_mouseover_changes
21. RightClick_Disabled
22. popUpWindow
23. Iframe_redirection
24. age_of_domain
25. DNS_Record
26. web_traffic_rank
27. Page_Rank
28. Google_Index
29. Links_pointing_to_page
30. Statistical_report - top phishing domains

Figure: 4.2.1 Features Extraction

Data Flow diagram

- **Level 0**

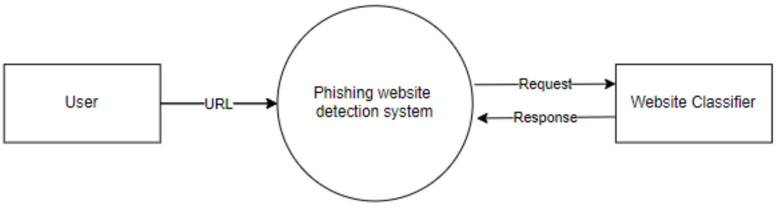


Figure: 4.2.2 DFD Level 0

- **Level 1**

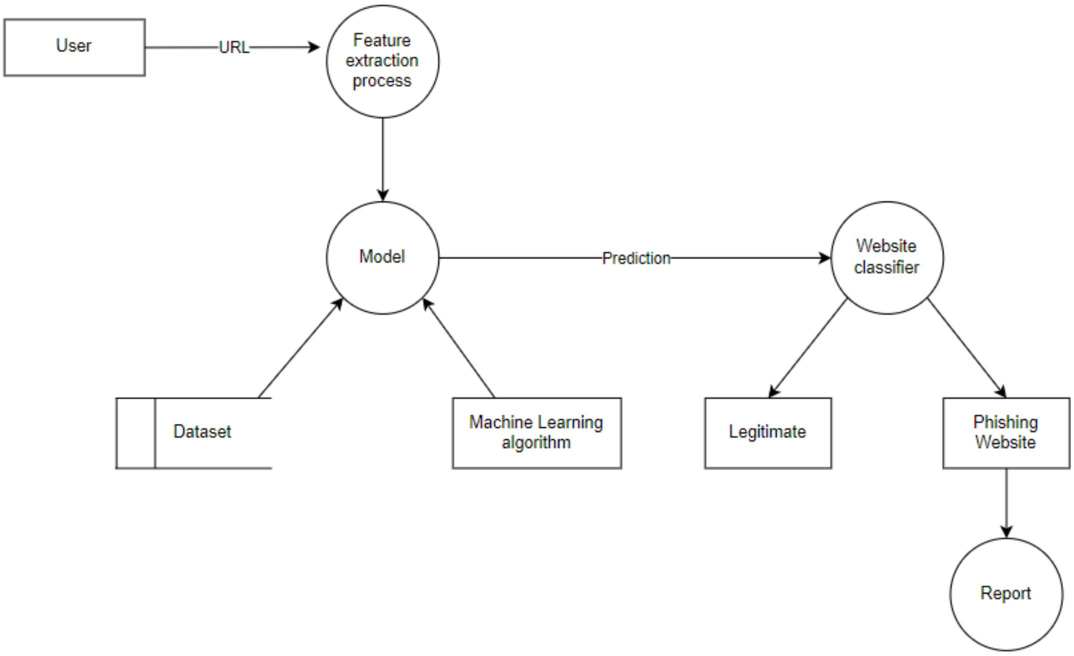


Figure: 4.2.3 DFD Level 1

Algorithm used to train model

• K-Means Clustering Algorithm

K-Means Clustering is an iterative algorithm that divides the unlabelled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

• Naïve Bayes Classifier Algorithm

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

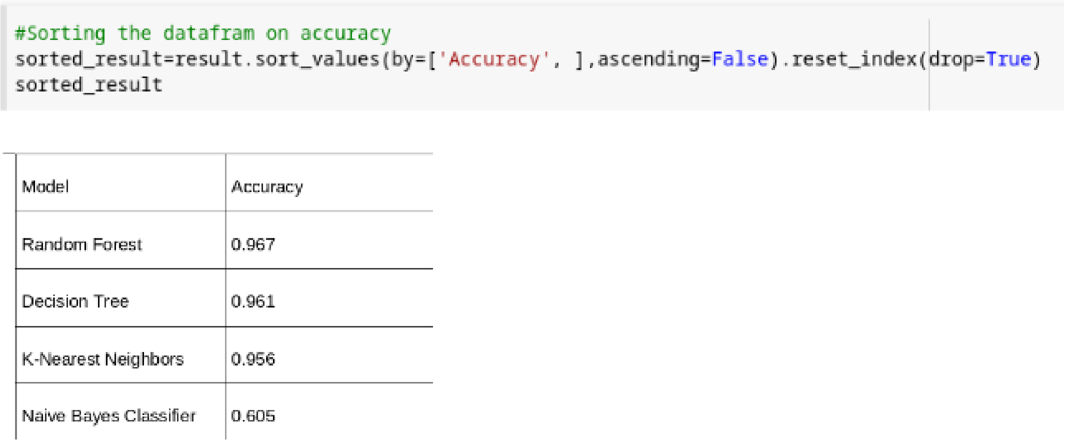
• Decision Tree Classification Algorithm

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

• Random Forest Algorithm

Random Forest is a machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Below is the accuracy result of the different ML models. The model random forest is giving the highest accuracy is selected to detect the phishing websites.



```
#Sorting the datafram on accuracy
sorted_result=result.sort_values(by=['Accuracy', ],ascending=False).reset_index(drop=True)
sorted_result
```

Model	Accuracy
Random Forest	0.967
Decision Tree	0.961
K-Nearest Neighbors	0.956
Naive Bayes Classifier	0.605

Figure: 4.2.4 Accuracy Result

Pi-hole

Pi-hole is a general purpose network-wide ad-blocker that protects your network from ads and trackers without requiring any setup on individual devices. It is able to block ads on any network device (e.g. smart appliances), and, unlike browser add-ons, Pi-hole blocks ads on any type of software.

The general setup works as follows (Fig. 1). You install Pi-hole on your server (in this case, we're using a Raspberry Pi) and assign it a static IP address. On your router, you set the DNS primary server to the Pi-hole IP address. When a device connects to your home network, it gets the Pi-hole IP address as its main DNS server from your router. When your device looks up the address for a hostname, it contacts the Pi-hole. If the host is an ad or tracker and present in the list used, the request is instantly blocked. Otherwise, the lookup is performed on some upstream server of your choice (e.g. OpenDNS, Cloudflare, GoogleDNS, your ISP).

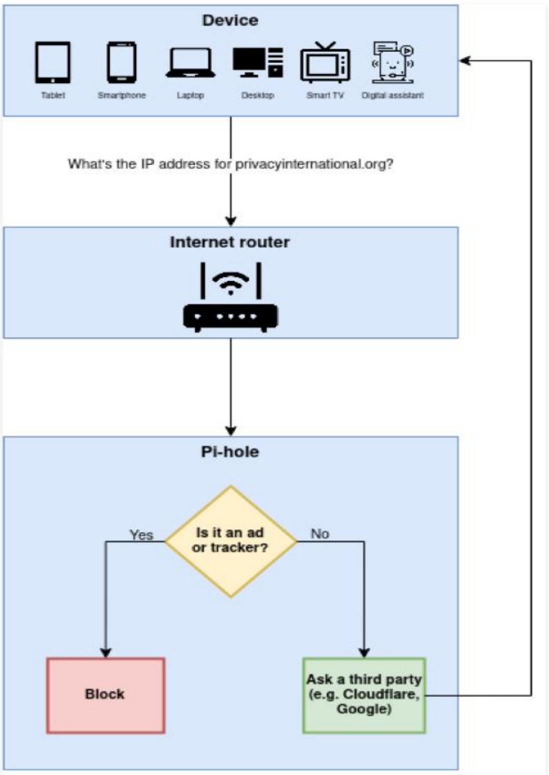


Figure: 4.2.5 Pi-hole Setup

After the successful connection of router an devices can see in home page of pi-hole which shows statistical information about number of DNS request serve, DNS request block, number of connected device and many more informational statistics.

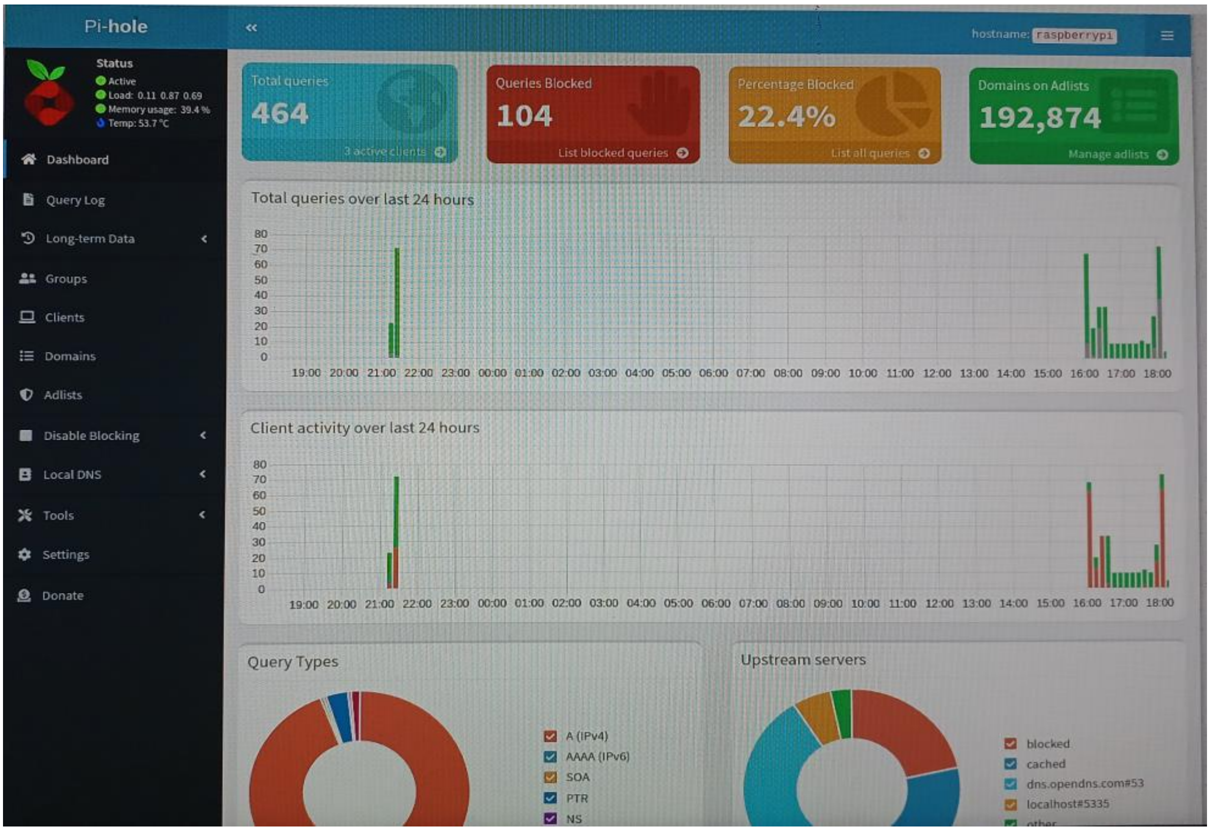


Figure: 4.2.6 Pi-hole Home page

Chapter 5

Result Analysis

5.1 Results and Discussion

Below is the home page of the phishing detection website. As given below user needs to enter the URL of the website in order to check whether that website is safe to use or not. After entering the URL user can simply click on check button.

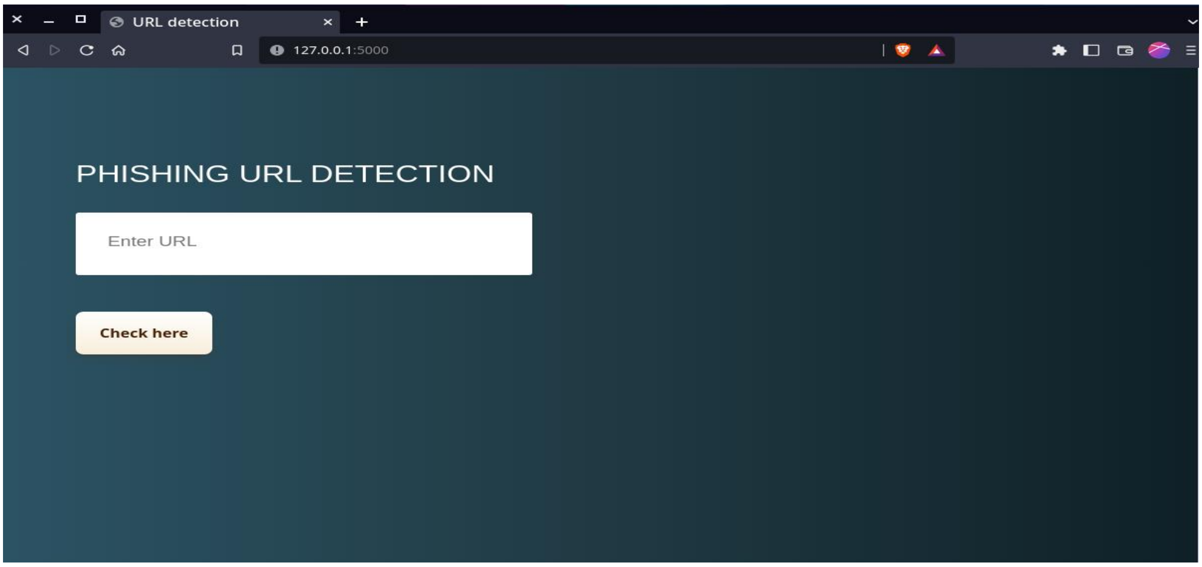
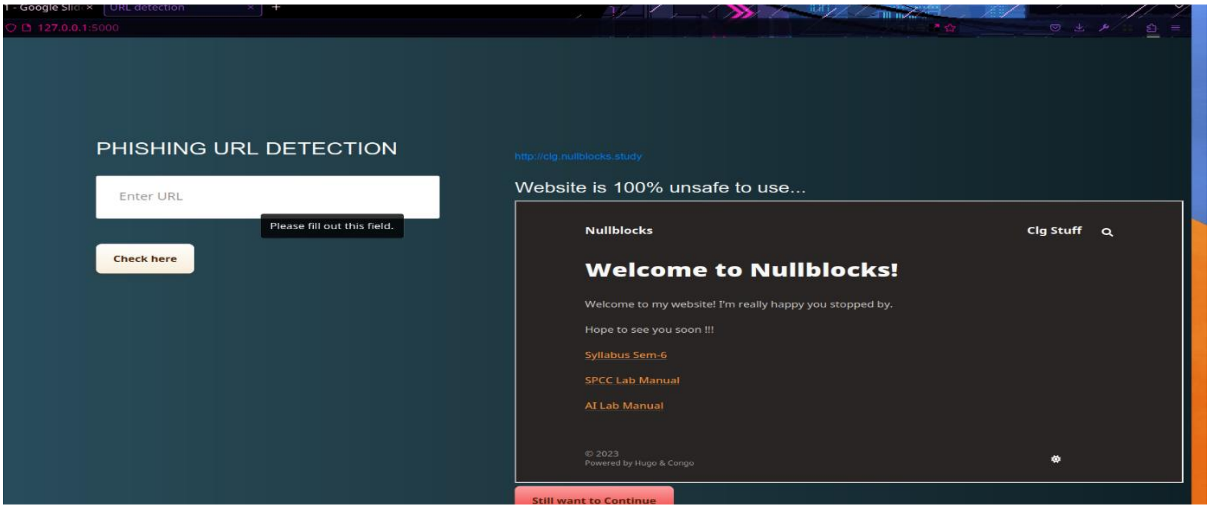


Figure: 5.1.1 Home Page

After clicking the check button, the ML model will classify whether the website is phishing or not based on the 30 features which were used to train the model.

As we can see below the website is classified as unsafe to use as it is using http and not https which makes the website vulnerable.



http://clg.nullblocks.study

Website is 100% unsafe to use...

Welcome to Nullblocks!

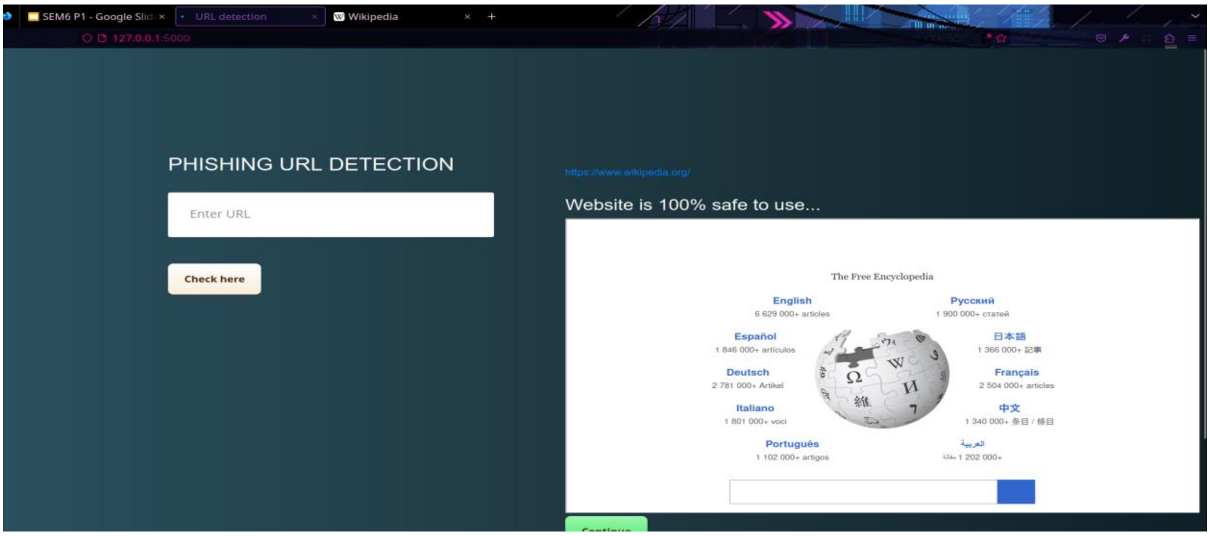
Welcome to my website! I'm really happy you stopped by.

Hope to see you soon !!!

© 2023
Powered by Hugo & Congo

Figure: 5.1.2 Search Results for Phishing Website

In this case the website is safe to use. Hence user can continue to use such websites. This result is also given on the basis of 30 features which were used to train the ML model.

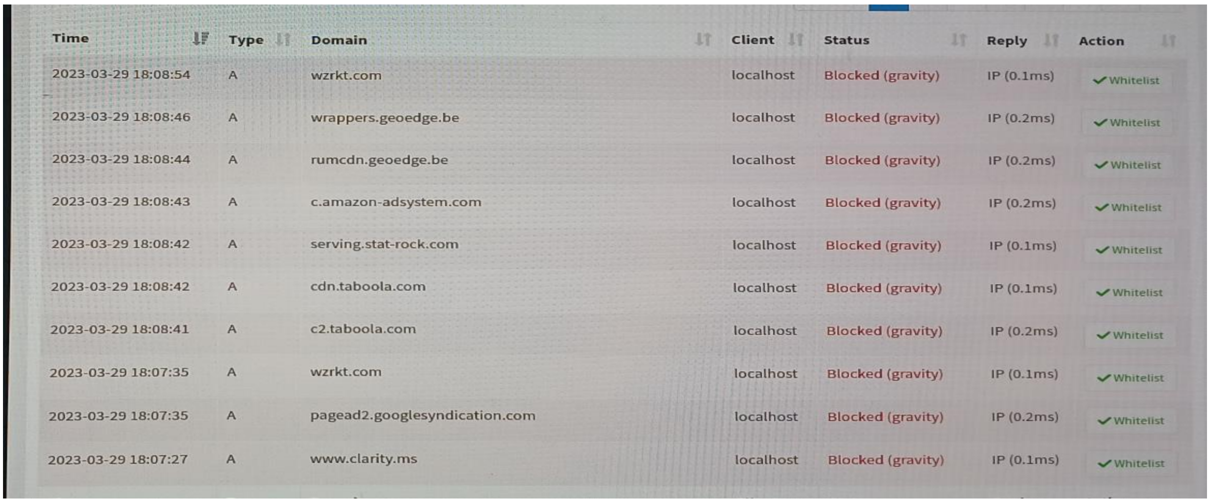


https://www.wikipedia.org/

Website is 100% safe to use...

Figure: 5.1.3 Search Result for Non-Phishing Website

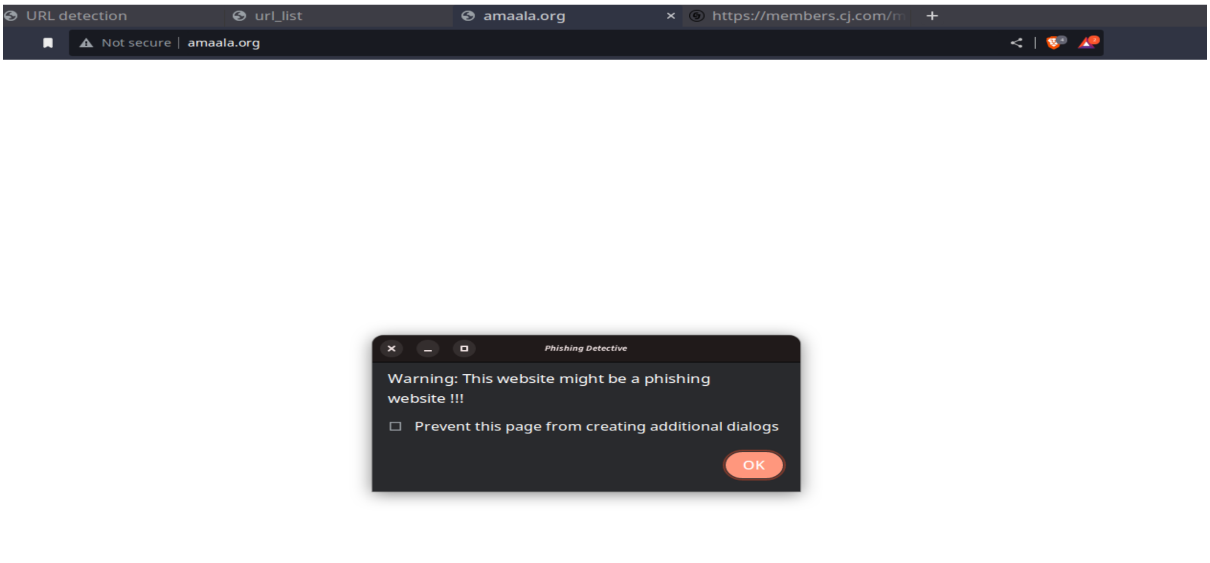
Here in below figure there are few request to ads server block by pi-hole to enhance user privacy and security.



Time	Type	Domain	Client	Status	Reply	Action
2023-03-29 18:08:54	A	wzrkt.com	localhost	Blocked (gravity)	IP (0.1ms)	✓ Whitelist
2023-03-29 18:08:46	A	wrappers.geoedge.be	localhost	Blocked (gravity)	IP (0.2ms)	✓ Whitelist
2023-03-29 18:08:44	A	rumcdn.geoedge.be	localhost	Blocked (gravity)	IP (0.2ms)	✓ Whitelist
2023-03-29 18:08:43	A	c.amazon-adsystem.com	localhost	Blocked (gravity)	IP (0.2ms)	✓ Whitelist
2023-03-29 18:08:42	A	serving.stat-rock.com	localhost	Blocked (gravity)	IP (0.1ms)	✓ Whitelist
2023-03-29 18:08:42	A	cdn.taboola.com	localhost	Blocked (gravity)	IP (0.1ms)	✓ Whitelist
2023-03-29 18:08:41	A	c2.taboola.com	localhost	Blocked (gravity)	IP (0.2ms)	✓ Whitelist
2023-03-29 18:07:35	A	wzrkt.com	localhost	Blocked (gravity)	IP (0.1ms)	✓ Whitelist
2023-03-29 18:07:35	A	pagead2.googlesyndication.com	localhost	Blocked (gravity)	IP (0.2ms)	✓ Whitelist
2023-03-29 18:07:27	A	www.clarity.ms	localhost	Blocked (gravity)	IP (0.1ms)	✓ Whitelist

Figure: 5.1.4 Pi-hole blocked DNS

When model detect any website is phishing website then extension will immediately block the website with alert box includes warning message about website.



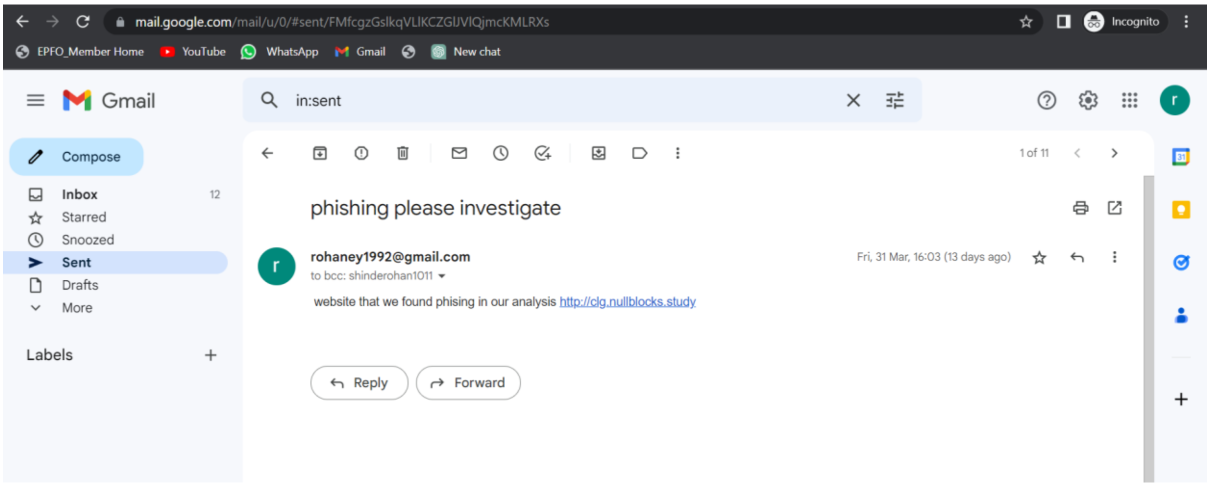
Warning: This website might be a phishing website !!!

□ Prevent this page from creating additional dialogs

OK

Figure: 5.1.5 Chrome Extension Warning the User

If phishing website is detected then it automatically send email to organization admin



phishing please investigate

rohaney1992@gmail.com

to bcc: shinderohan1011

Fri, 31 Mar, 16:03 (13 days ago)

website that we found phising in our analysis http://clg.nullblocks.study

Figure: 5.1.6 Sending g-mail from admin

Chapter 6

Conclusion

Conclusion

In conclusion, our project aimed to protect users from phishing websites by identifying and alerting them to potential threats, as well as blocking ads through the use of a tool such as pi-hole. Through our work, we found that our phishing website identification and alert system was effective in detecting and warning users about potential threats, while the use of pi-hole significantly reduced the risk of users being exposed to malicious ads that could lead to phishing attacks.

Our project highlights the importance of taking proactive steps to protect online users from phishing attacks, as well as the potential benefits of implementing similar systems or tools in other contexts to improve online safety and security. While there were some challenges and limitations encountered during the project, we believe that our work provides a valuable contribution to the ongoing efforts to enhance online safety and security.

In the future, further research could be done to explore the effectiveness of different approaches to phishing website detection and blocking, as well as the potential impact of other tools and technologies on reducing the risk of phishing attacks. Overall, our project serves as a reminder of the critical importance of protecting online users from phishing attacks, and the role that individuals and organizations can play in achieving this goal.

References

[1] Ankit Kumar Jain and B. B. Gupta. “Analysis of Visual Similarity Based Approaches” *IEEE* Security and Communication Networks, vol.2017, Issue: 10 Jan 2017.

[2] Wei King Tiong. “Utilisation of website logo for phishing detection” IEEE Computers & Security, vol 54, Issue: October 2015.

[3] Weili Han. “Anti-phishing based on automated individual white-list” IEEE Digital identity management, vol 12, Issue: October 2008

[4] Arathi Krishna , “Phishing Detection using Machine Learning based URL Analysis” International journal of engineering research & technology (ijert) , vol9 , Issue 02-08-2021