PHISHING URL DETECTION

by

SISTA GOPALA KRISHNA	*421258*
PETETI RAM	*421240*

Under the guidance of

Dr. RAVEENDRA BABU PONNURU

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY ANDHRA PRADESH

TADEPALLIGUDEM-534101, INDIA

MAY 2023

PHISHING URL DETECTION

Thesis submitted to
National Institute of Technology Andhra Pradesh
for the award of the degree

of

Bachelor of Technology

by

SISTA GOPALA KRISHNA	*421258*
PETETI RAM	*421240*

Under the guidance of

Dr. RAVEENDRA BABU PONNURU

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY ANDHRA PRADESH

TADEPALLIGUDEM-534101, INDIA

MAY 2023

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signature
SISTA GOPALA KRISHNA

421258

Date:_____

Signature
PETETI RAM

421240

Date:_____

CERTIFICATE

It is certified that the work contained in the thesis titled "**PHISHING URL DETECTION**" by "SISTA GOPALA KRISHNA, bearing Roll No: 421258" and "PETETI RAM, bearing Roll No: 421240" has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

Signature
Dr. RAVEENDRA BABU PONNURU
CSE DEPARTMENT
N.I.T. Andhra Pradesh
MAY, 2023

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible, whose constant guidance and encouragement crowned our efforts with success. It is a pleasant aspect that I have now the opportunity to express my gratitude for all of them. We owe our sincere gratitude to our project guide Dr. RAVEENDRA BABU PONNURU, Department of Computer Science, National Institute of Technology, Andhra Pradesh, who took keen interest and guided us all along, till the completion of our project work by providing all the necessary information.

We avail ourselves of this proud privilege to express our gratitude to all the faculty of the department of Computer Science and Engineering at NIT Andhra Pradesh for emphasizing and providing us with all the necessary facilities throughout the work. We offer our sincere thanks to all our friends, fellow mates and other persons who knowingly or unknowingly helped us to complete this project.

LIST OF FIGURES

Abstract

In today's world,we live in the age of the internet. It has become an important part of our life such that we can't live without it. The internet is an invention of the high-end science and modern technology. Because of the internet, our lives have become more convenient as compared to the times when there is a world without internet. Internet has been used to do many things such as banking, shopping, social media, medicines, entertainment, etc.

As the rapid growth of technology occurs, some suspicious activities have been encountered in the internet. Phishing means the way or practice of sending emails or message purposely to get the data of an individual. The internet is connected with a url(unique resource locator). Most of hackers or companies which try to steal data of an individual do it by sending some fake url's to the individual, so that they can black mail or threatened them with the data they have collected from the url's.

Several anti phishing techniques emerge continuously but phishers come with new technique by breaking all the anti phishing mechanisms. Hence there is a need for efficient mechanism for the prediction of phishing website. A set of verified url's are considered, on the basis of it we can say whether it is a fake or original website. phishing websites may have url's that are misspelled or used a different domain names whereas legitimate website will have a url that matches the name of the company or organization.

Malicious websites help to promote the internet frauds and growth of internet criminal activities and hindering the growth of web services. As a result, the user is restricted from using the web services. We propose to classify the websites into three categories: Benign, Malware and malicious. Our project analyses the URL (Uniform Resource Locator)without exploring the contents of websites, hence, saves the time latency and thus eliminating the risk of exposing the user to browser based vulnerability. The learning algorithm is employed on the URL and it performs better on achieving the generality.

URL's are divided into three parts:
Benign: The website which are safe to use.
Malicious: The websites which flood the user's system with various advertising sites and disrupt the normal functioning of the web services.
Malware: The websites which intend to gather the sensitive information from the user

TABLE OF CONTENTS

Contents

1 Introduction

Phishing attacks are a type of cyber attack that targets individuals or organizations by tricking them into revealing sensitive information such as usernames, passwords,and credit card details. Phishing attacks are usually carried out through email, social media, or messaging apps. Detecting phishing attacks and preventing them from causing harm is crucial in today's digital landscape.

Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer's defense systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization's logos and other legitimate contents.

Typically a victim receives a message that appears to have been sent by a known contact or organization. The message contains malicious software targeting the user's computer or has links to direct victims to malicious websites in order to trick them into divulging personal and financial information, such as passwords, account IDs or credit card details.

2 Brief on data

As everything has its positive effect as well as negative effect, the internet has negative sides also. The internet has become the platform for committing frauds and other illegal activities. With the increasing activities on internet the fraudsters have used this platform to manipulate the hardware and software of the system. Phishing is one of most committed frauds across the internet. It is the practice of stealing the passwords or credit card information. The phishing usually takes two forms:
This attack deceives the victim to make them reveal their real identity and thus giving trustworthy information to the attacker.The attacker plants the malware into victim's computer.

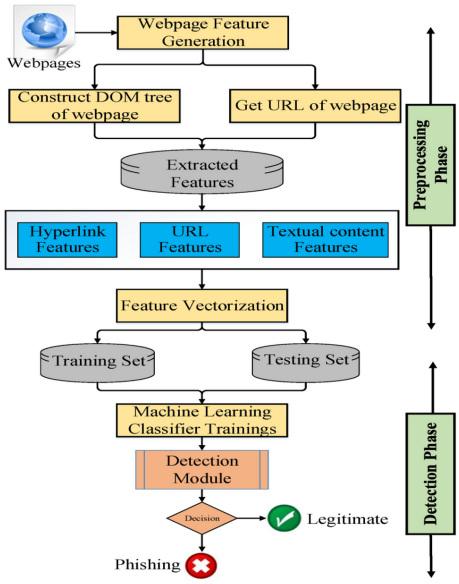

Fig-1: URL system architecture

2.1 Already present system or existing system

The existing model may cause the neural network or NN model to under fit the training data sets. On the other hand, the overfitting problem may occur in the model due to exaggeration of NN model to suit every data in the data sets. This overfitting problem maybe stopped by adding the new neurons to the hidden layers or adding new layer to the network. But the small numbers of hidden neurons in the data may cause problem as it increases the complexity of the model and if we exhibit the hiding of layers then the over fit problem will arise. So this a continuous problem. Hence an acceptable error rate is not specified and thus this system is not efficient as proposed system.
Disadvantages:

- It won't load all the dataset.
- Slow analysis.
- Process is not accurate

.

2.2 Developed system

The URL's of phishing websites look different from others. So analysing this we get to know that the URL from which we divide it into two parts: host name and path. Another one is that the malicious websites tend to act benign by containing popular brand names as tokens other than those in second level domain. Mostly the phishing websites tend to use the IP address more than the benign one and thus it is more easy to recognize the malicious website. The site popularity is considered as an important feature in the phishing detection and traffic rank feature is acquired from alexa.com. the malicious websites are always hosted from less popular hosting centres or regions.

Advantages:

- URLs are labelled
- We used the algorithm such as random forest to train using scikit-learn library.

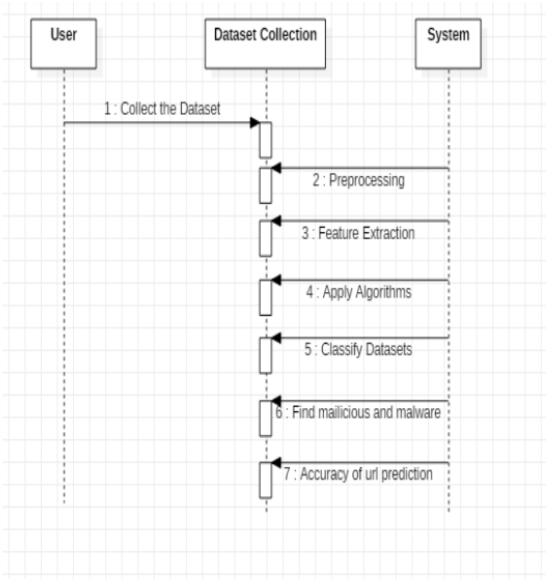

Fig-2:User case diagram

2.3 Background

The modern use of internet involves a lot of complications as the people on the internet tend to attract to several advertisements which are on malicious websites. These websites tend to attract lot of money to the fraudsters and other hackers who created these websites.
Most of the people visit these websites as the websites are related to them personally and it may harm them as these sites are made to gather personal information from the user and it may harm them. So in order to prevent the user from phishing websites our project tends to tell the user that the website is safe to use or not.

The web services are advanced nowadays and the fraudsters and hackers are also advanced due to that. The web services are to be prevented from that and thus some more advanced technologies are used and hence the new method is the URL analysis. The URL can tell us whether the website is malicious or not. So the motivation is that to prevent the users form visiting the malicious websites so that it can't harm the user.

2.4 Problem statement

The problem in existing method was that the old system was based on the NN model of the NLP. But the new system we are using has URL analyzing system which enables to find the benign or malicious website.The problem in the

existing system is the model which we are using. The NN model has several hidden layers which causes the increase in complexity and is less efficient.

Fig-3:Sequence or order flow

2.5 Objective

The main objective of this research is to provide user with a platform from which he can check the website whether it is safe to use or not. The research focus on the security of the users search and the user can search for a URL before visiting it.

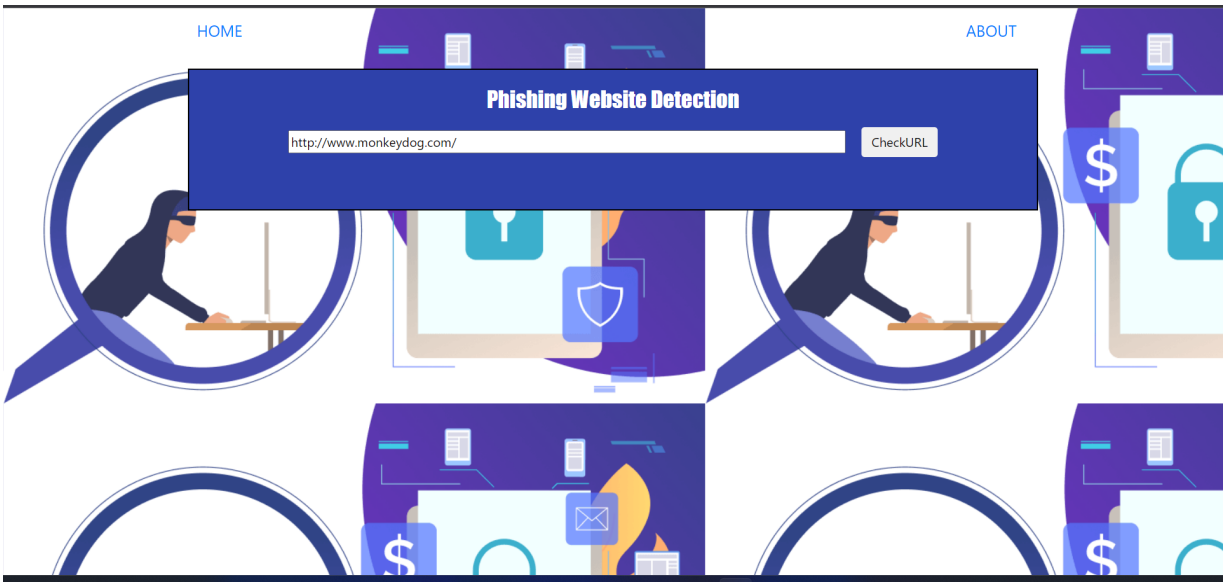

Fig-4:Home page

3 Types and Step wise working

Financial industry involves an extremely high volume of real time online transactions.This makes vulnerable to fraud. Fraud detection is using security measures to prevent third parties from obtaining funds. This process involves a manual check and automated verification of transaction details to spot any unusual activity that may be a sign of attack and block it.

Email Phishing:

Over the years there have been many attacks of Phishing and many people have lost huge sums of money by becoming a victim of phishing attack. In a phishing attack emails are sent to user claiming to be a legitimate organization, where in the email asks user to enter information like name, telephone, bank account number important passwords etc. Such emails direct the user to a website where in user enters these personal information. These websites also known as phishing website now steal the entered user information and carries out illegal transactions thus causing harm to the user. Phishing website and their mails are sent to millions of users daily and thus are still a big concern for cyber security.

3.1 Collection of data

The first step is to collect a dataset of legitimate and phishing websites. Legitimate websites can be collected from popular domains, such as .edu, .gov, or .com, while phishing websites can be collected from phishing repositories or by searching for common phishing tactics and patterns.

Fig-5:About page

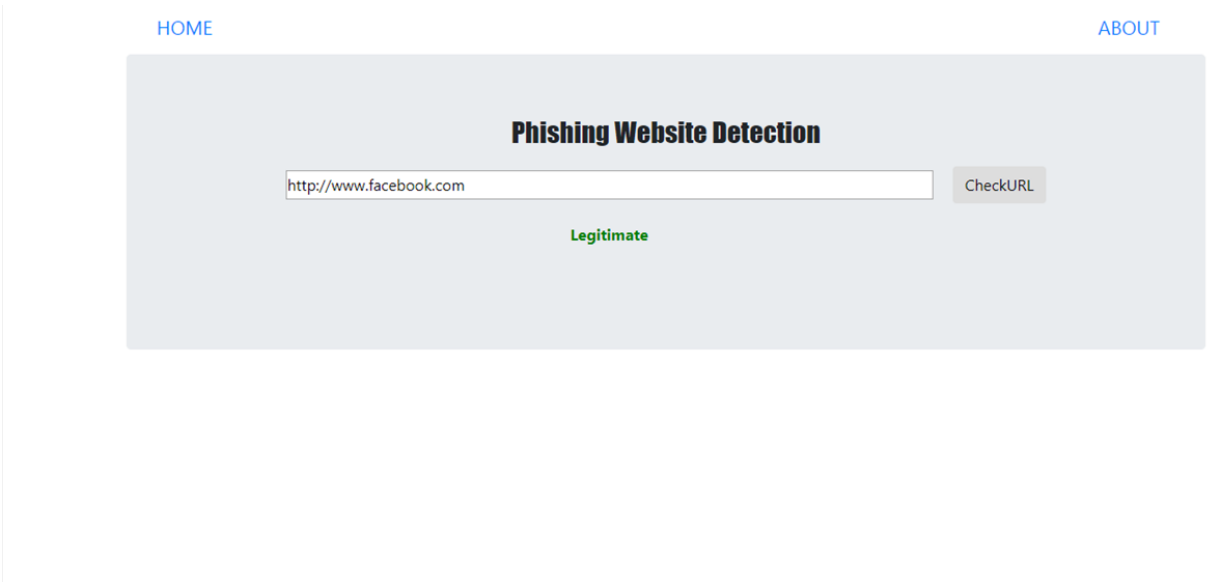

Fig-6:Legitimate

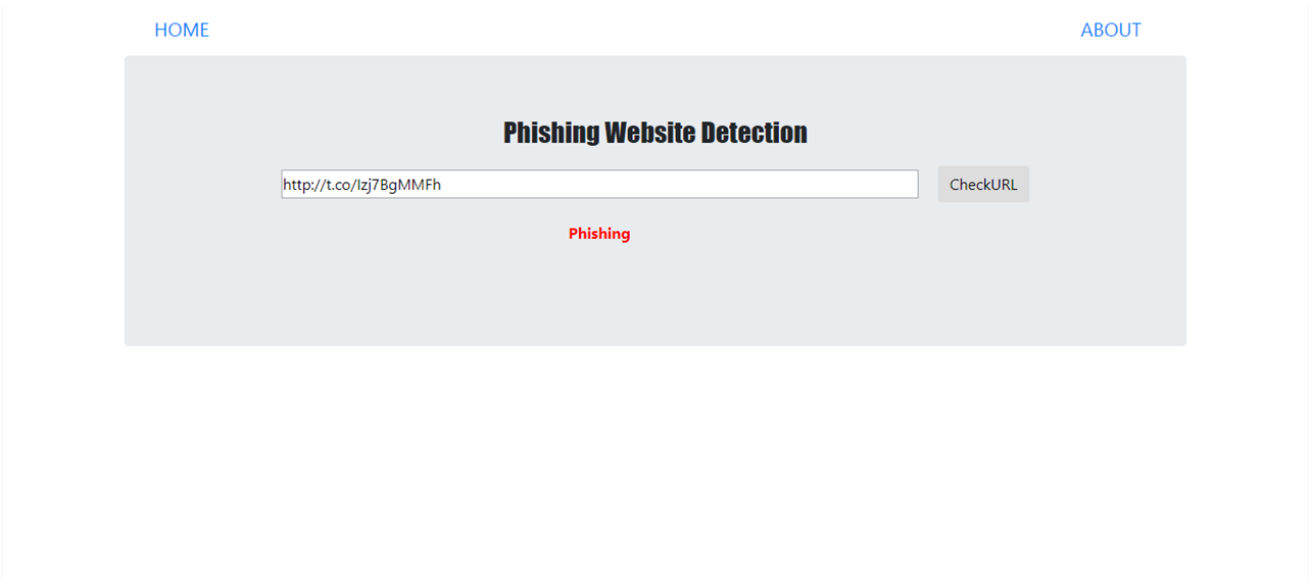

Fig-7:Phishing

3.2 Classifier used

In this project we are young mainly two classifiers. They are: 1.Random Forest classifier
2.Decision Tree classifier
Decision trees are very easy as compared to the random forest. A decision tree combines some decisions, whereas a random forest combines several decision trees. Thus, it is a long process, yet slow. Whereas, a decision tree is fast and operates easily on large data sets, especially the linear one.

3.3 Features Extraction

3.3.1 Feature 1

1.Long URL to Hide the Suspicious Part
If the length of the URL is greater than or equal 54 characters then the URL classified as phishing.

0 — Indicates legitimate

1 — Indicates Phishing

2 — Indicates Suspicious

3.3.2 Feature 2

2.URL's having "@" Symbol

Using "@" symbol in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol.

IF Url Having @ Symbol⟶ Phishing Otherwise⟶ Legitimate

0 — Indicates legitimate

1 — Indicates Phishing

3.3.3 Feature 3

3.Redirecting using "//"

The existence of "//" within the URL path means that the user will be redirected to another website. An example of such URL's is: "http://www.legitimate.com//http://www.phishing.com.We examine the location where the "//" appears.We find that if the URL starts with "HTTP", that means the "//" should appear in the sixth position. However,if the URL employs "HTTPS" then the "//" should appear in seventh position.

IF ThePosition of the Last Occurrence of "//" in the URL ⟶ Phishing

Otherwise⟶ Legitimate

0 — Indicates legitimate

1 — Indicates Phishing

3.3.4 Feature 4

4.Adding Prefix or Suffix Separated by (-) to the Domain

The dash symbol is rarely used in legitimate URLs.Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage.

For example http://www.Confirme-paypal.com/.

IF Domain Name Part Includes () Symbol ⟶ Phishing

Otherwise ⟶ Legitimate

1 ⟶ Indicates phishing

0 ⟶ Indicates legitimate

3.3.5 Feature 5

5. Sub-Domain and Multi Sub-Domains The legitimate URL link has two dots in the URL since we can ignore typing "www.". If the number of dots is equal to three then the URL is classified as "Suspicious" since it has one sub-domain. However, if the dots are greater than three it is classified as "Phishy" since it will have multiple sub-domains

0 — Indicates legitimate

1 — Indicates Phishing

2 — Indicates Suspicious

3.3.6 Feature 6

6.Using the IP Address

If an IP address is used as an alternative of the domain name in the URL,such as “http://125.98.3.123/fake.html

Rule: IFIf The Domain Part has an IP Address ⟶ Phishing Otherwise⟶ Legitimate

1 ⟶ Indicates phishing

0 ⟶ Indicates legitimate

3.3.7 Feature 7

7.Using URL Shortening Services “TinyURL”

URL shortening is a method on the “World Wide Web” in which a URL may be made considerably smaller in length and still lead to the required webpage. This is accomplished by means of an “HTTP Redirect” on a domain name that is short, which links to the webpage that has a long URL. For example, the URL “http://portal.hud.ac.uk/

Rule: IFTinyURL ⟶ Phishing

Otherwise⟶ Legitimate

1 ⟶ Indicates phishing

0 ⟶ Indicates legitimate

3.3.8 Feature 8

8.The Existence of “HTTPS” Token in the Domain Part of the URL

The phishers may add the “HTTPS” token to the domain part of a URL in order to trick users. For example, http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/.

Rule: IFUsing HTTP Token in Domain Part of The URL⟶ Phishing

Otherwise⟶ Legitimate.

3.3.9 Feature 9

9.Abnormalurl

This feature can be extracted from WHOIS database. For a legitimate website, identity is typically part of its URL.

Rule: IF The Host Name Is Not Included In URL ⟶ Phishing

Otherwise⟶ Legitimate

3.3.10 Feature 10

10.Google Index

This feature examines whether a website is in Google’s index or not. When a site is indexed by Google, it is displayed on search results (Webmaster resources, 2014). Usually, phishing webpages are merely accessible for a short period and as a result, many phishing webpages may not be found on the Google index.

Rule: IFWebpage Indexed by Google ⟶ Legitimate

Otherwise ⟶ Phishing

3.3.11 Feature 11

11.Website Traffic

This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit. However, since phishing websites live for a short period of time, they may not be recognized by the Alexa database (Alexa the Web Information Company., 1996). By reviewing our dataset, we find that in worst scenarios, legitimate websites ranked among the top 100,000. Furthermore, if the domain has no traffic or is not recognized by the Alexa database, it is classified as “Phishing”. Otherwise, it is classified as “Suspicious”.

Rule: IfWebsite Rank¡100,000 ⟶ LegitimateWebsite Rank¿100,
000 ⟶SuspiciousOtherwise ⟶ Phish

3.3.12 Feature 12

Domain Registration Length

Based on the fact that a phishing website lives for a short period of time, we believe that trustworthy domains are regularly paid for several years in advance. In our dataset, we find that the longest fraudulent domains have been used for one year only.

Rule: IfDomains Expires on 1 years ⟶ Phishing Otherwise⟶ Legitimate

3.3.13 Feature 13

This feature can be extracted from WHOIS database (Whois 2005). Most phishing websites live for a short period of time. By reviewing our dataset, we find that the minimum age of the legitimate domain is 6 months.

Rule: If Age Of Domain6 months ⟶ Legitimate
Otherwise ⟶ Phishing

3.3.14 Feature 14

DNS Record

For phishing websites, either the claimed identity is not recognized by the WHOIS database (Whois 2005) or no records founded for the hostname (Pan and Ding 2006). If the DNS record is empty or not found then the website is classified as “Phishing”, otherwise it is classified as “Legitimate”.

Rule: Ifno DNS Record For The Domain ⟶ Phishing
Otherwise⟶ Legitimate

3.3.15 Feature 15

Statistical-Reports Based Feature

Several parties such as PhishTank (PhishTank Stats, 2010-2012), and StopBadware (StopBadware, 2010-2012)formulate numerous statistical reports on phishing websites at every given period of time some are monthly and others are quarterly.

Rule: IfHost Belongs to Top Phishing IPs or Top Phishing Domains ⟶ Phishing
Otherwise ⟶ Legitimate

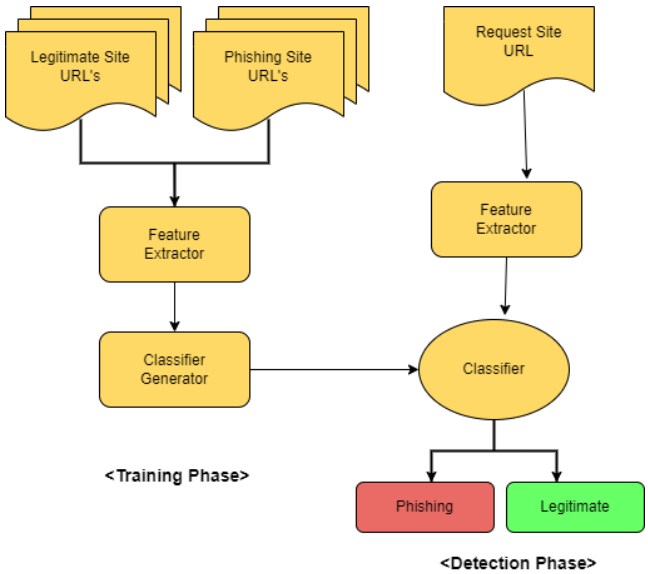

Fig-8:Features

3.4 Train and Split

After collecting data merged two data frames top 1000 rows will have legitimate urls and bottom 1000 rows will have phishing urls. So if we split the data now and create a model for it will over fit so we need to shuffle the rows before splitting the data into training set and test set splitting the data into train data and test data Dividing the data in the ratio of 70:30 or 80:20

checking the split of labels in train and test data The split should be in equal proportion for both classes
Phishing - 1
Legitimate - 0
Now count the no of phishing and legitimate website in test and train part. In result no of phishing and legitimate website count is nearly same. Now create a random Forest classifier model and fitting the data into that model and check the prediction of train and test values.

Creating confusion matrix and checking the accuracy. (Different between predicted values and original values)Similarly by using Decision tree classifier model predict the train and test values. Find the accuracy using confusion Matrix.

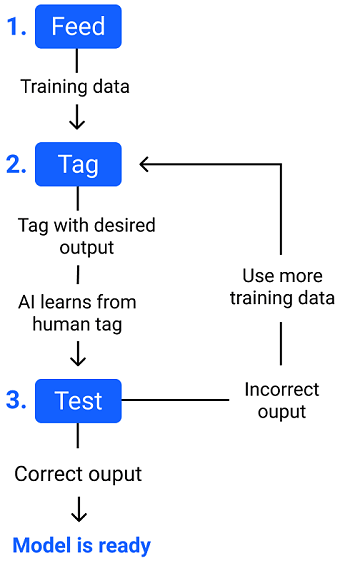

Fig-9:Train and test

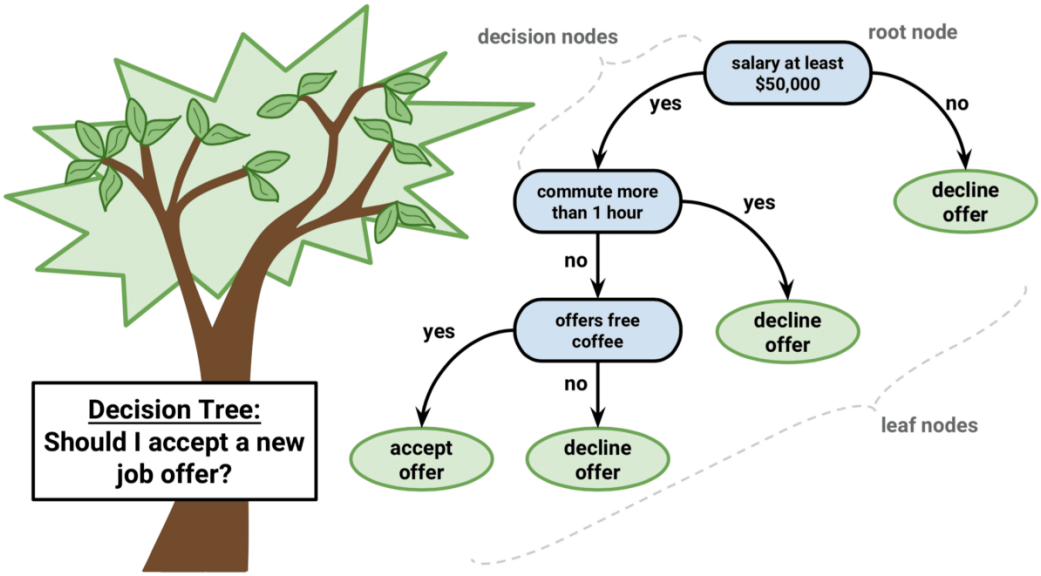

Fig-10:Decision tree

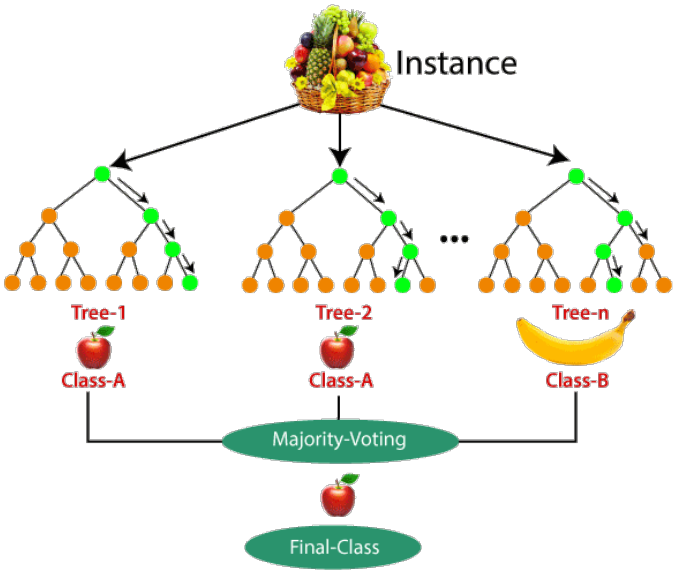

Fig-11:Random forest classifier

4 System design

We faced one of the most common challenges that almost every researcher in this field might come across, it was the availability of reliable datasets. Although we found some articles about predicting phishing websites with the use of data mining techniques but nothing exactly reliable enough for training. This made it difficult to shape a data set which would cover all the possible features. Here we have given a brief on a few important feature that have been effective enough for predicting phishing websites.

We also tried our hands on coming up with some feature that can be used to update some other feature or assigning new rule to the features already well known.It is the integral part of the model evaluation. It helps to find the error and hence rectifying it. The accuracy of the project is usually seen with evaluation models. There are usually the two methods for evaluation models: Hold-Out and Cross-Validation. To avoid this method of over fitting, each data set is taken and the mode is thus evaluated. The model's result will be based on averaged and the result will be in visualized form. The result is displayed in form of graphs. By the method we can calculate the accuracy of the model. The accuracy is calculated by dividing the predictions which are correct by the total prediction.

5 Conclusion

Trends are bound to happen. The new more enhanced things take up the place of the old dull ones. The new things are more enhanced and much more future proof. One such thing is ML or machine learning. the arrival of this has led to the generation of new use cases such as cyber security. As is happens usually – when any technology goes mainstream, a couple of things happen: we expect too much, we don't fully use the primary application of the tool. Cyber security is the latest field using ml to enhance its functionality and making the world a better place. There's a new generation of hackers prevailing now. Fame alone no longer fuels for this crop – they want to hold people hostage by capturing their data or threatening to release private information. Cyber security has come a long way as it was from before. A major part of it can be because of machine learning. Earlier there were methods like:

- captcha:the person had to manually enter a peice of randomly generated text by the website and the user had to enter it.
- otp(one time password):the person had to enter a 6 or 4 digit passcode sent to the registered phone number.
- Images:the person had to choose certain photos out of a grid satisfying a certain condition So seeing that all these methods were somewhat imperfect and had some or the other drawbacks and hence ml was introduced. So the crux of it is, we give "training" data to the machine learning model; we teach it to notice patterns across that data,and then we test and refine those classification processes. There are numerous ways to achieve this, from a variety of code libraries and languages to supervised and unsupervised training styles. The main highlight however, is that machine learning can enable much faster decision-making; it's just not necessarily in more "intelligent" ways than those used by a human.

So, ML does cyber security in the following ways:

Intrusion detection: Using old network logs, we can teach a machine learning algorithm to study IP addresses, timestamps, and connection IDs to detect anomalous behavior. We classify certain blocks of network activity as acceptable or normal; we then classify other blocks of network activity as unacceptable or unusual. All of this data gets stored into the machine learning data sets, which then examines the data for patterns – after which we test its understanding with fresh information. As time goes on and the model examines more of this data,it becomes more robust. Thus, once some percent accuracy and precision is achieved, we can integrate the ML model into an existing security system to make faster, real-time decisions about intrusion detection – increasing the difficulty for attackers to remotely enter and move within a system.

6 Future Scope and Conclusion

Cyber security along with the help of machine learning has come a long way but still has to go a long way. It is true that the algorithms cannot distinguish perfectly like the human touch.
The cyber attacks have enhanced in 3 way:

- MOTIVE: In the past viruses were introduced by curious programmers. Today cyber attacks are a result of well executed plan
- SPEED : The potential rate at which the attack spreads has increased.

7 References

- R. Heartfield and G. Loukas, “A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks,” ACM Computing Surveys (CSUR), vol. 48, no. 3, p. 37, 2015.
- .J. Hong,“The state of phishing attacks,” Communications of the ACM,vol. 55, no. 1,pp. 74–81, 2012