IBM – NAALAIYA THIRAN PROJECT

WEB PHISING DETECTION

INDUSTRY MENTOR : SANDESH.P

FACULTY MENTOR : S.MAMITHA

TEAM ID :PNT2022TMID52273

TEAM LEADER : RAJITHA ROJA.T

TEAM MEMBER : PRAVINSHA.V

TEAM MEMBER : VIBITHA T.V

TEAM MEMBER : PRIYA DHARSHINI A.S

1.INTRODUCTION

1.1 Project Overview

There are number of users who purchase products online and make payment through e-banking. There are e-banking website who ask user to provide sensitive data such as username, password or credit card details etc often for malicious reasons. This type of website is known as phishing website. In order to detect and predict e-banking phishing website, we proposed an intelligent, flexible and effective system and is based on using classification data mining algorithm. We implemented classification algorithm and technique to extracts the phishing data set criteria to classify their legitimacy. The e-banking phishing website can be detected based on some important characteristics like URL and Domain Identity, and security and encryption criteria in the final phishing rate. Once user makes transaction through online when he makes payment through e-banking website our system will use data mining algorithm to detect whether the e-banking website is phishing website or not.

1.2 Purpose

The purpose of web phishing detection is to safeguard online users from becoming victims of online fraud, divulging confidential information to an attacker among other effective users of phishing as an attacker’s tool; phishing detection tools play a vital role in ensuring a secure online experience for users.

2. LITERATURE SURVEY

Phishing website is a recent problem, nevertheless due to its huge impact on the financial and on-line retailing sectors and since preventing such attacks is an important step towards defending against e-banking phishing website attacks, there are several promising approaches to this problem and a comprehensive collection of related works. In this section, we briefly survey existing anti-phishing solutions and list of the related works. One approach is to stop phishing at the email level , since most current phishing attacks use broadcast email (spam) to lure victims to a phishing website .

Another approach is to use security tool bars. The phishing filter in IE7 is a tool bar approach with more features such as blocking the user's activity with a detected phishing site. Other approach is to visually differentiate the phishing sites from the spoofed legitimate sites. Dynamic Security Skins proposes to use a randomly generated visual hash to customize the browser window or web form elements to indicate the successfully authenticated sites. A fourth approach is two-factor authentication, which ensures that the user not only knows a secret but also presents a security token .

However, this approach is a server-side solution. Phishing can still happen at sites that do not support two-factor authentication. Sensitive information that is not related to a specific site, e.g., credit card information and SSN, cannot be protected by this approach either .

However, an automatic anti-phishing method is seldom reported. The typical technologies of anti phishing from the User Interface aspect are done. They proposed methods that need Web page creators to follow certain rules to create Web pages, either by adding dynamic skin to Web pages or adding sensitive information location attributes to HTML code. However, it is difficult to convince all Web page creators to follow the rules .

2.1 EXISTING PROBLEM:

If we can detect the phishing Web sites in time, we then can block the sites and prevent phishing attacks. It's relatively easy to (manually) determine whether a site is a phishing site or not, but it's difficult to find those phishing sites out in time. Here we list two methods for phishing site detection.

2.2 REFERENCE

[1] Androutsopoulos, J. Koutsias, K.V Chandrinos, and C.D. Spyropoulos. An Experimental Comparison of Naive Bayesian and KeywordBased Anti-Spam Filtering with Encrypted Personal E-mail Message.In Pro c. SIGIR 2000, 2000.

[2] The Anti-phishing working group. http://www.antiphishing.org/.Neil Chou, Robert Ledesma, Yuka Teraguchi, Dan Boneh, and John C.Mitchell. Client-side defence against web-based identity theft. In Pro cc.NDSS 2004, 2004.

[3] B. Aida, S. Heisenberg and R. Rives t, ―Lightweight Encryption for Email,‖ USENIX Steps to Reducing Unwanted Traffic on the Internet Workshop (SRUTI), 2005.

[4] Cynthia Dwork, Andrew Goldberg, and Moni Naor. On Memory-Bound.Functions for Fighting Spam. In Pro c. Crypto 2003, 2003.

[5] R. Mihijam and J.D. Garry, ―The Battle against Phishing: Dynamic Security Skins, I Pros. Symptom. Usable Privacy and Security, 2005.

[6] FDIC., ―Putting an End to Account-Hijacking Identity Theft,‖ http://www.fdic.gov/consumers/consumer/idtheftst Rudy/identity_theft.pdf, 2004.

[7] A. Y. Fu, L. Weeny and X. Dens, ― Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD) , I IEEE transactions on dependable and secure computing, volt. 3, no. 4, 2006.

[8]EarthLink.ScamBlocker. http://www.earthlink.net/software/free/toolbar/

[9] David Geer. Security Technologies Go Phishing. IEEE Computer, 38 (6):18-21, 2005.

[10]John Leyden. Trusted search software labels fraud site as safe‟. http://www.theregister.co.uk/2005/09/27/untrustedsearch/.

[11]Microsoft. Sender ID Framework. http://www.microsoft.com/

[12]mscorp/safety/technologies/sender-id/default.mspx.

[13]Net craft. Net craft tool-bar. http://toolbar.netcraft.com/.

[14]PhishGuard.com. Protect Against Internet Phishing Scams http://www.phishguard.com/.

[15]Jonathan B. Pastel. Simple Mail Transfer Protocol. RFC821:http:llwww.ietf. or/corf/rfcO82 1 .txt.

[16]Georgina Stanley. Internet Security - Gone phishing.http://www.cyota.com/news.asp?id=1 14.

[17]Meng Weng Wong. Sender ID SPF. http://www.openspf.org/whitepaper.pdf.

[18]T. Sharif, Phishing Filter in IE7, http://blogs.msdn.com/ie/archive/2005/09/09/4632 04.aspx, September 9, 2006.

[19]M. Wu, R. C. Miller and S. L. Garfield, —Do Security Tool-bars Actually Prevent Phishing Attacks?" CHI April 2006.

[20]M. Wu, R. C. Miller and G. Little, —Web Wallet: Preventing Phishing Attacks by Revealing User Intentions, I MIT Computer Science and Artificial Intelligence Lab, 2006.

2.3 PROBLEM AND STATEMENT

Attackers will steal information related to transaction details by using malicious links or by using any software or by sending emails. We need to be alert and not allow installing or downloading any unnecessary software or should not click on any unnecessary links. Now-a-days banks are sending mail or SMS for every transaction made on online. We need not to share any personal details related to bank. Though it an advantage, we need to keep transaction detail safe by having stronger algorithms. There are many security risks associated with web services on the Internet, including phishing websites. Online shopping and payments are popular among users. Some websites request sensitive information from users, such as username s, passwords, and credit card numbers, often for malicious purposes. This type of website is known as a phishing website. A proper solution is needed to detect and prevent phishing websites.

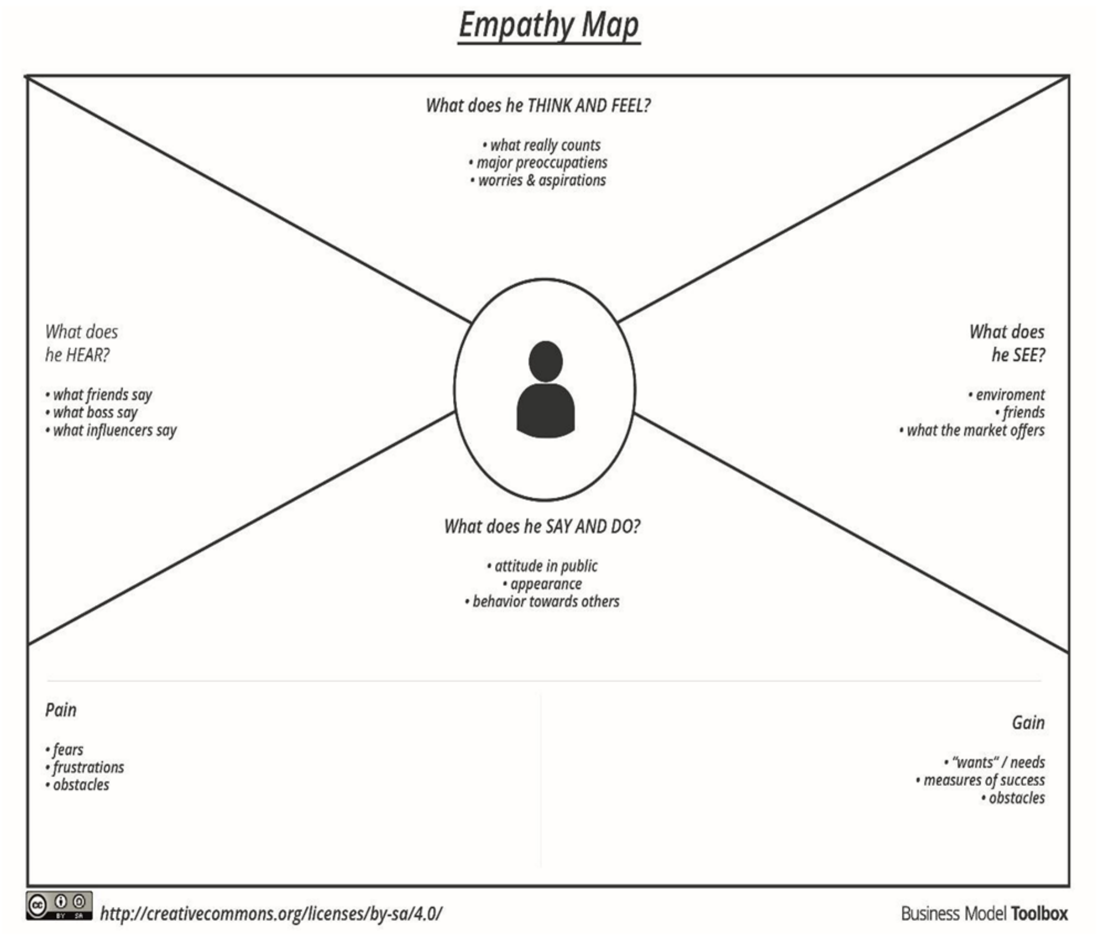
3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

An empathy map is a simple, easy-to-digest visual that captures knowledge about a user's behaviours and attitudes. It is a useful tool to help steams better understand their users.

Creating an effective solution requires understanding the true problem and the person who is experiencing it. The exercise of creating the map helps participants consider things from the user’s perspective along with his or her goals and challenges.

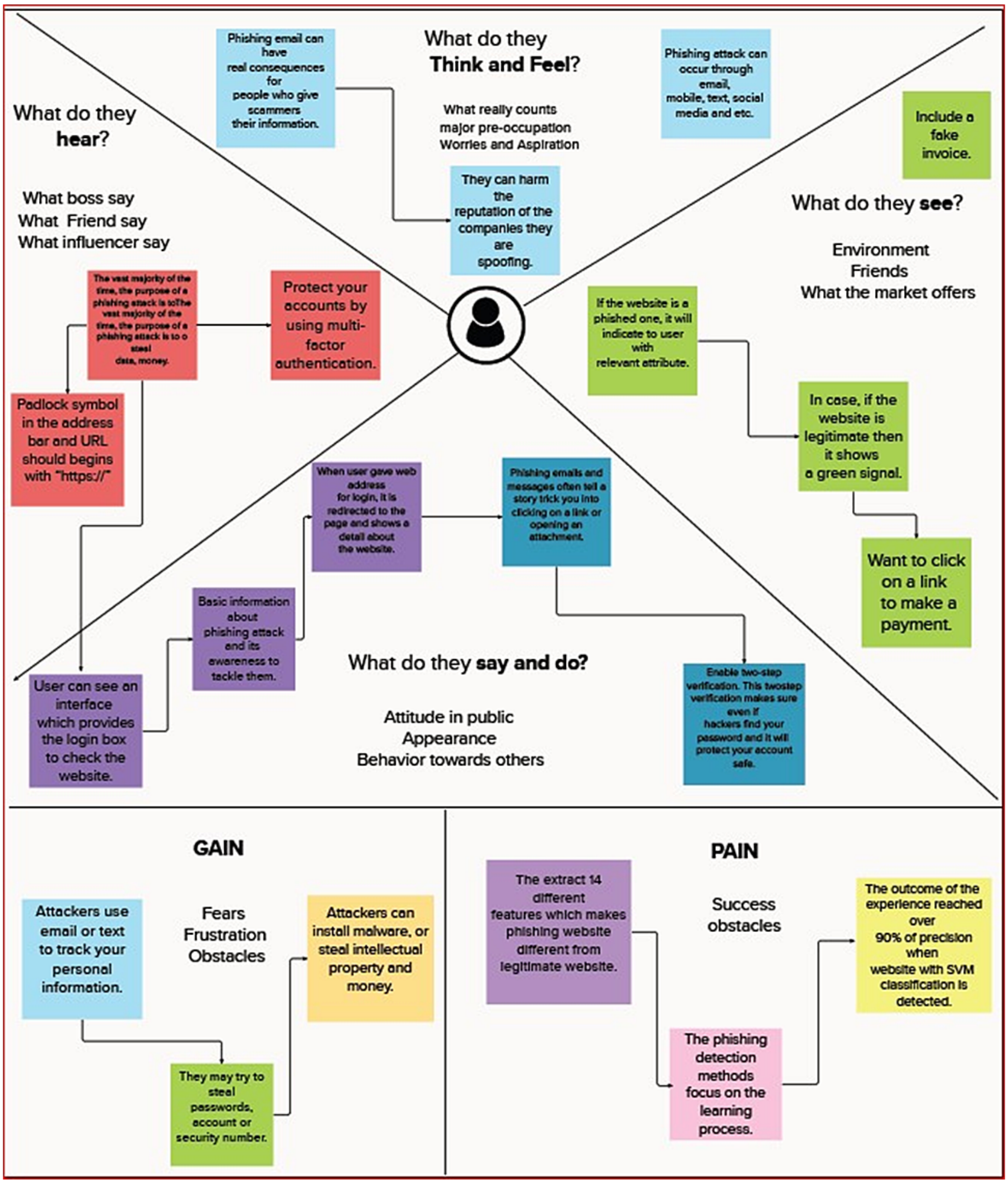
Example:



Reference:

https://app.mural.co/t/ibm7094/m/ibm7094/1665385347013/152cd495f9e73a31da69122e0ff8a3581128ce06?sender=u5aec96631bdad25b66f94634

Empathy Map:


What do they Think and Feel?
What really counts
major pre-occupation
Worries and Aspiration
Phishing email can have real consequences for people who give scammers their information.
Phishing attack can occur through email, mobile, text, social media and etc.
They can harm the reputation of the companies they are spoofing.
What do they hear?
What boss say
What Friend say
What influencer say
The vast majority of the time, the purpose of a phishing attack is toThe vast majority of the time, the purpose of a phishing attack is to o steal data, money.
Protect your accounts by using multi-factor authentication.
Padlock symbol in the address bar and URL should begins with "https://"
What do they see?
Environment
Friends
What the market offers
Include a fake invoice.
If the website is a phished one, it will indicate to user with relevant attribute.
In case, if the website is legitimate then it shows a green signal.
Want to click on a link to make a payment.
What do they say and do?
Attitude in public
Appearance
Behavior towards others
User can see an interface which provides the login box to check the website.
Basic information about phishing attack and its awareness to tackle them.
When user gave web address for login, it is redirected to the page and shows a detail about the website.
Phishing emails and messages often tell a story trick you into clicking on a link or opening an attachment.
Enable two-step verification. This twostep verification makes sure even if hackers find your password and it will protect your account safe.
GAIN
Fears
Frustration
Obstacles
Attackers use email or text to track your personal information.
They may try to steal passwords, account or security number.
Attackers can install malware, or steal intellectual property and money.
PAIN
Success
obstacles
The extract 14 different features which makes phishing website different from legitimate website.
The phishing detection methods focus on the learning process.
The outcome of the experience reached over 90% of precision when website with SVM classification is detected.

PROBLEM AND STATEMENT:

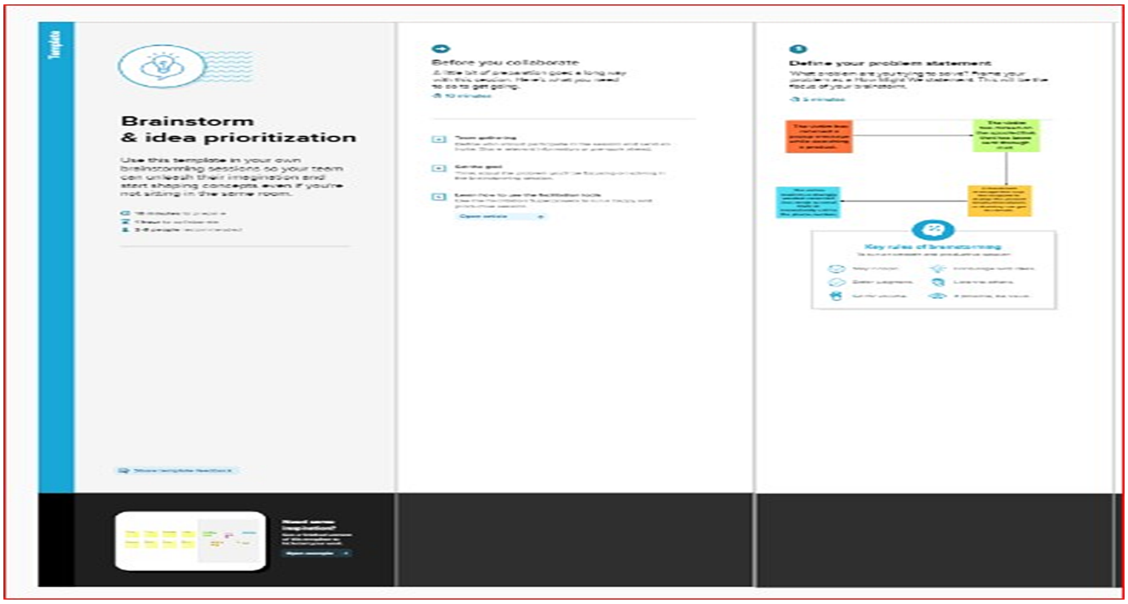
Attackers will steal information related to transaction details by using malicious links or by using any software or by sending emails. We need to be alert and not allow installing or downloading any unnecessary software or should not click on any unnecessary links. Now-a-days banks are sending mail or SMS for every transaction made on online. We need not to share any personal details related to bank. Though it an advantage, we need to keep transaction detail safe by having stronger algorithms.

3.2 Ideation & Brainstorming

Brainstorm & Idea Prioritization Template:

Brainstorming provides a free and open environment that encourages everyone within a team to participate in the creative thinking process that leads to problem solving.Prioritizing volume over value, out-of-the-box ideas are welcome and built upon, and all participants are encouraged to collaborate, helping each other develop a rich amount of creative solutions.

Team Gathering, Collaboration and Select the Problem Statement:



Template

Brainstorm & idea prioritization

Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.

Before you collaborate

A little bit of preparation goes a long way with this session. Here's what you need to do to get going.

Define your problem statement

Key rules of brainstorming

Fig) Team Gathering, Collaboration and Select the Problem Statement

Brainstorm, Idea Listing and Grouping:

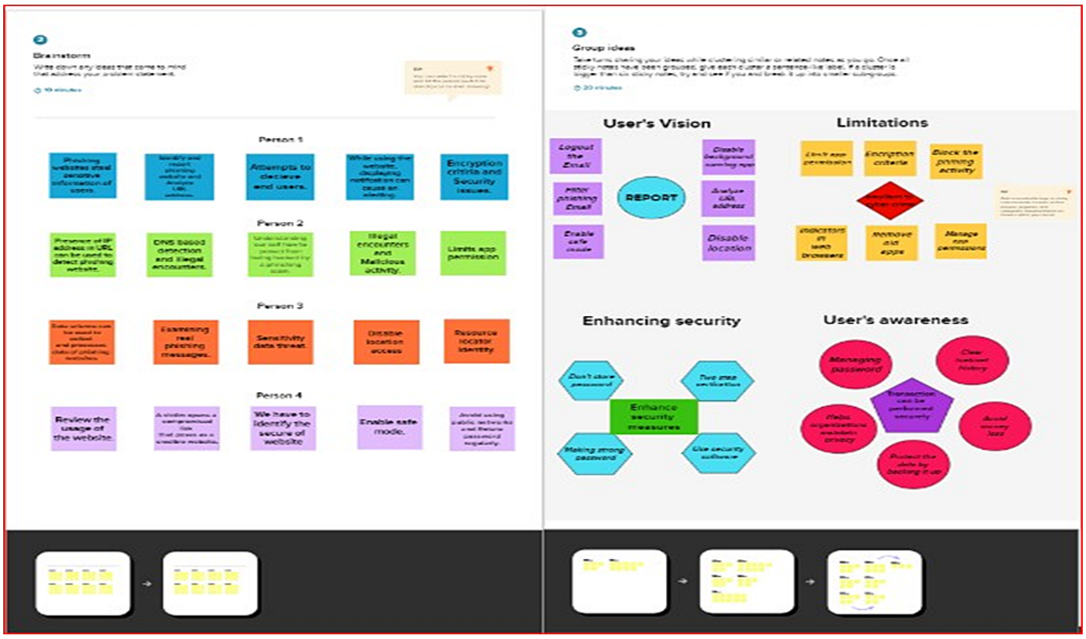

Brainstorm
Person 1
Phishing websites steal sensitive information of users.
Attempts to deceive end users.
Encryption critria and Security issues.
Person 2
Presence of IP address in URL can be used to detect phishing website.
DNS based detection and illegal encounters.
Illegal encounters and Malicious activity.
Limits app permission
Person 3
Examining real phishing messages.
Sensitivity data threat.
Disable location access
Resource locator identity
Person 4
Review the usage of the website.
We have to identify the secure of website
Enable safe mode.
Group ideas
User's Vision
Logout the Email
Filter phishing Email
Enable safe mode
REPORT
Disable background running apps
Analyse URL address
Disable location
Limitations
Limit app permission
Encryption criteria
Block the priming activity
Indicators in web browsers
Remove old apps
Manage app permissions
Enhancing security
Don't store password
Two step verification
Enhance security measures
Making strong password
Use security software
User's awareness
Managing password
Clear browser history
Transaction can be performed securely
Helps organizations maintain privacy
Avoid money loss
Protect the data by backing it up

Fig) Brainstorm, Idea Listing and Grouping

Idea Prioritization

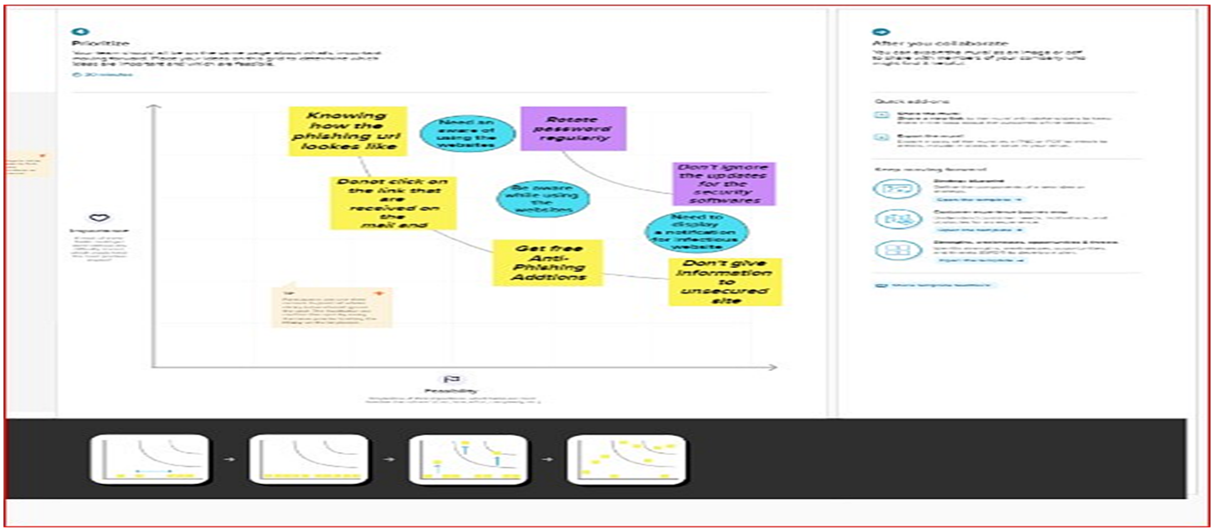

Prioritize
Knowing how the phishing url lookes like
Need an aware of using the websites
Rotate password regularly
Donot click on the link that are received on the mail and
Be aware while using the websites
Don't ignore the updates for the security softwares
Need to display a notification for infectious website
Get free Anti-Phishing Addtions
Don't give information to unsecured site
Importance
Feasibility
After you collaborate

Fig) Idea Prioritization

3.3 Proposed Solution

Sl.No	Parameter	Description

1.	Problem Statement (Problem to be solved)	1. Data and Assets may be stolen ordamaged. 2. Customers might be unable to access online services. 3. Malicious statement steals the login credentials or financial information like credit card numbers.
2.	Idea / Solution description	1. Detection of malicious websites 2. To detect the web phishing websites for providing secured e-banking transactions, we proposed an intelligent and effective system based on classification machine learning algorithm. 3. Classification algorithms helps to identify the phishing datasets based on their authorized information like URL, Domain identity and encryption criteria. 4. Once the user logs in to the e-banking websites, the proposed algorithm identifies the legitimate of the website and blocks the phishing site.

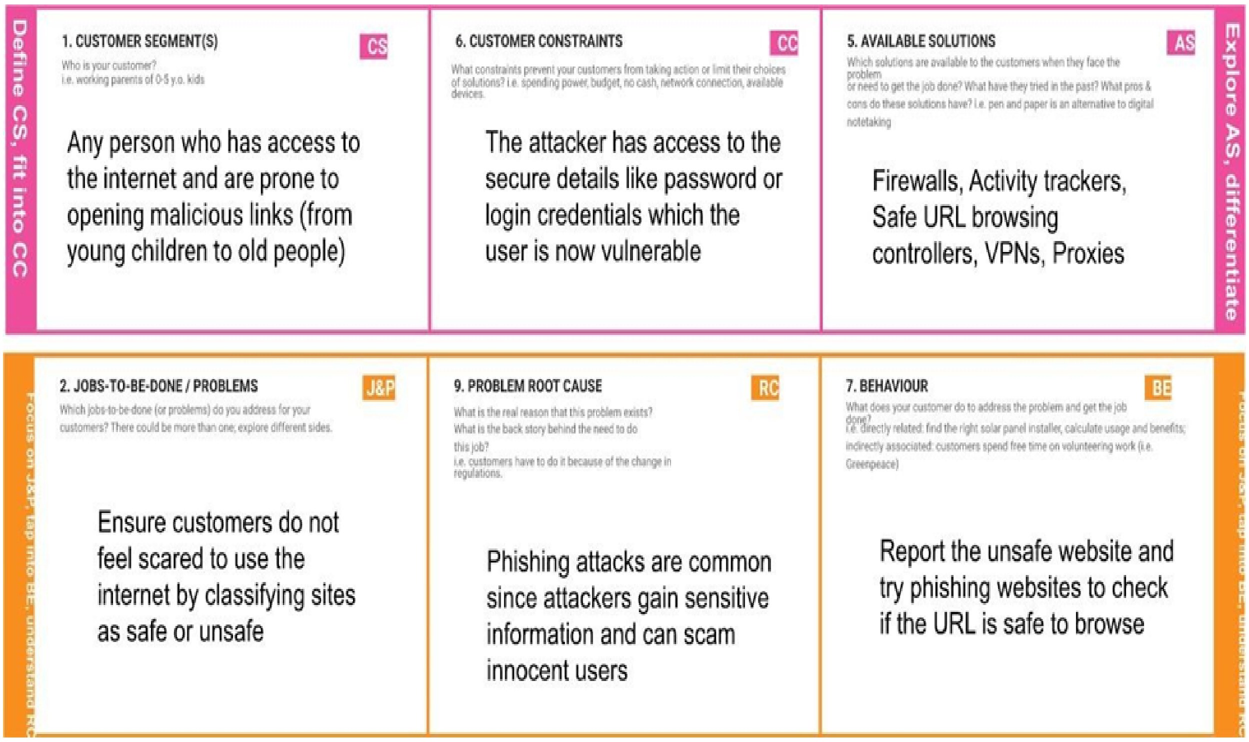
3.	Novelty / Uniqueness	1. The proposed classification algorithm helps to identify the phishing site in an effective manner and blocks the site while avoiding the property damage forthe users 2. Security alert 3. The proposed model helps users to avoid getting trapped in different kindsof scams. 4. Our model-will recognize fake vs real URLs
4.	Social Impact / Customer Satisfaction	1. It will save the users from fraudulent websites and reduced global economical losses caused by web phishing every year. 2. It gives a reliable way to detect web phishing and scamming sites. 3. It provides a secured and confidential environment for e-banking.
5.	Business Model (Revenue Model)	1. Our project can be used in e-commerce and online e-banking transactions.
6.	Scalability of theSolution	1. It will be useful for a wide range of users from individual users to corporate,banks and universities. 2. Helps in reducing economical loss caused by these web phishing incidents and also protects from confidential. 3. It identifies the suspicious phishing mails and enhances the security software.

3.4 Proposed Solution Fit

The Problem-Solution Fit simply means that you have found a problem with your customer and that the solution you have realized for it actually solves the customer’s problem. It helps entrepreneurs, marketers and corporate innovators identify behavioural patterns and recognize what would work andwhy.

Purpose:

Solve complex problems in a way that fits the state of your customers.Succeed faster and increase your solution adoption by tapping into existing mediums and channels of behaviour.Sharpen your communication and marketing strategy with the right triggers and messaging.Increase touch-points with your company by finding the right problem-behaviour fit and building trust by solving frequent annoyances, or urgent or costly problems.Understand the existing situation in order to improve it for your target group.

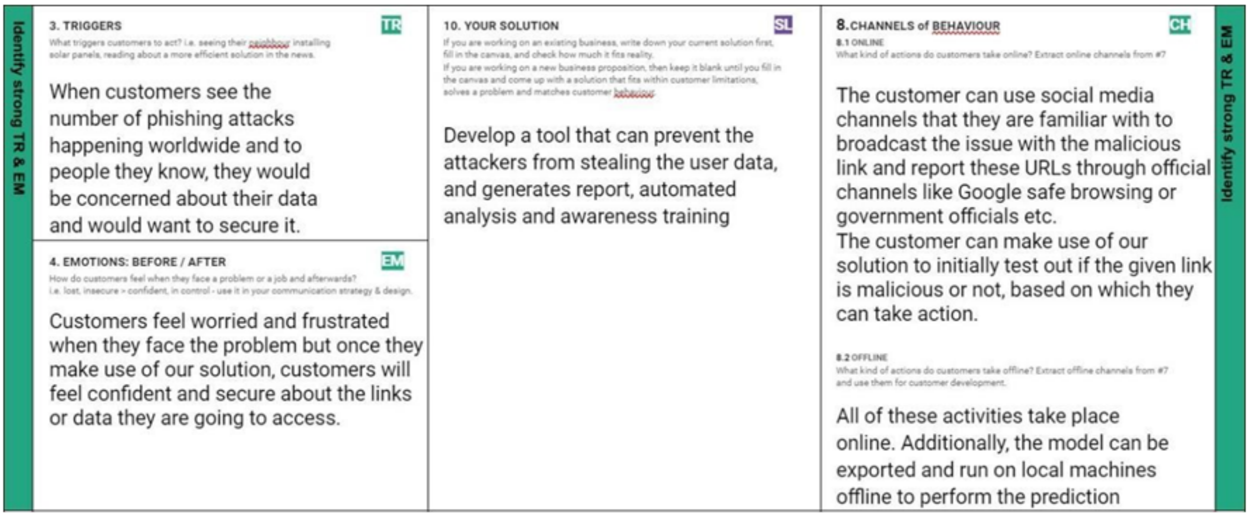


Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS Who is your customer? i.e. working parents of 0-5 y.o. kids Any person who has access to the internet and are prone to opening malicious links (from young children to old people)	6. CUSTOMER CONSTRAINTS CC What constraints prevent your customers from taking action or limit their choices of solutions? i.e. spending power, budget, no cash, network connection, available devices. The attacker has access to the secure details like password or login credentials which the user is now vulnerable	5. AVAILABLE SOLUTIONS AS Which solutions are available to the customers when they face the problem or need to get the job done? What have they tried in the past? What pros & cons do these solutions have? i.e. pen and paper is an alternative to digital notetaking Firewalls, Activity trackers, Safe URL browsing controllers, VPNs, Proxies
2. JOBS-TO-BE-DONE / PROBLEMS J&P Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides. Ensure customers do not feel scared to use the internet by classifying sites as safe or unsafe	9. PROBLEM ROOT CAUSE RC What is the real reason that this problem exists? What is the back story behind the need to do this job? i.e. customers have to do it because of the change in regulations. Phishing attacks are common since attackers gain sensitive information and can scam innocent users	7. BEHAVIOUR BE What does your customer do to address the problem and get the job done? i.e. directly related: find the right solar panel installer, calculate usage and benefits; indirectly associated: customers spend free time on volunteering work (i.e. Greenpeace) Report the unsafe website and try phishing websites to check if the URL is safe to browse

Explore AS, differentiate

Focus on J&P, tap into BE, understand RC



3. TRIGGERS

What triggers customers to act? i.e. seeing their neighbour installing solar panels, reading about a more efficient solution in the news.

When customers see the number of phishing attacks happening worldwide and to people they know, they would be concerned about their data and would want to secure it.

4. EMOTIONS: BEFORE / AFTER

How do customers feel when they face a problem or a job and afterwards?
i.e. lost, insecure > confident, in control - use it in your communication strategy & design.

Customers feel worried and frustrated when they face the problem but once they make use of our solution, customers will feel confident and secure about the links or data they are going to access.

10. YOUR SOLUTION

If you are working on an existing business, write down your current solution first, fill in the canvas, and check how much it fits reality.
If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behaviour.

Develop a tool that can prevent the attackers from stealing the user data, and generates report, automated analysis and awareness training

8.CHANNELS of BEHAVIOUR

8.1 ONLINE

What kind of actions do customers take online? Extract online channels from #7

The customer can use social media channels that they are familiar with to broadcast the issue with the malicious link and report these URLs through official channels like Google safe browsing or government officials etc.
The customer can make use of our solution to initially test out if the given link is malicious or not, based on which they can take action.

8.2 OFFLINE

What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development.

All of these activities take place online. Additionally, the model can be exported and run on local machines offline to perform the prediction

4.REQUIREMENT ANALYSIS

4.1 Functional Requirements

FR No.	Functional Requirement(Epic)	Sub Requirement (Story / Sub-Task)
FR-1	Evaluate the dataset	download dataset and analyse
FR-2	Test and train thedataset	Use variousmode ls to testand train thedataset
FR-3	Get the bestmode	Model with most accuracy is selected
FR-4	It is implemented on awebsite	It findswhether a websiteis a phishing siteor not
FR-5	Enter Details aboutthe Yes/No in all thefieldswebsite	
FR-6	Submit to get accuracy	After entering thedetails to get accuracy, Click on submit

4.2 Non-Functional Requirements

NFR No.	Non-Functional Requirement(Epic)	Description

NFR-1	Usability	It is a websitewhich can be used in any platformto check whethera website is a phishingsite or not.
NFR-2	Security	It is highly secure as the details entered are contained within the website and it cannotbe accessed by others.
NFR-3	Reliability	The accuracyof the model can be brought up to morethan 90 percent.
NFR-4	Performance	Only one model is used to detect whetheror not a site is phishing, so it gives the result instantaneously after thedetails are enteredand it is submitted.
NFR-5	Availability	
NFR-6	Scalability	It is scalable to applications and other anti-virus software.

5. Project Design

5.1 Data Flow Diagrams

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.

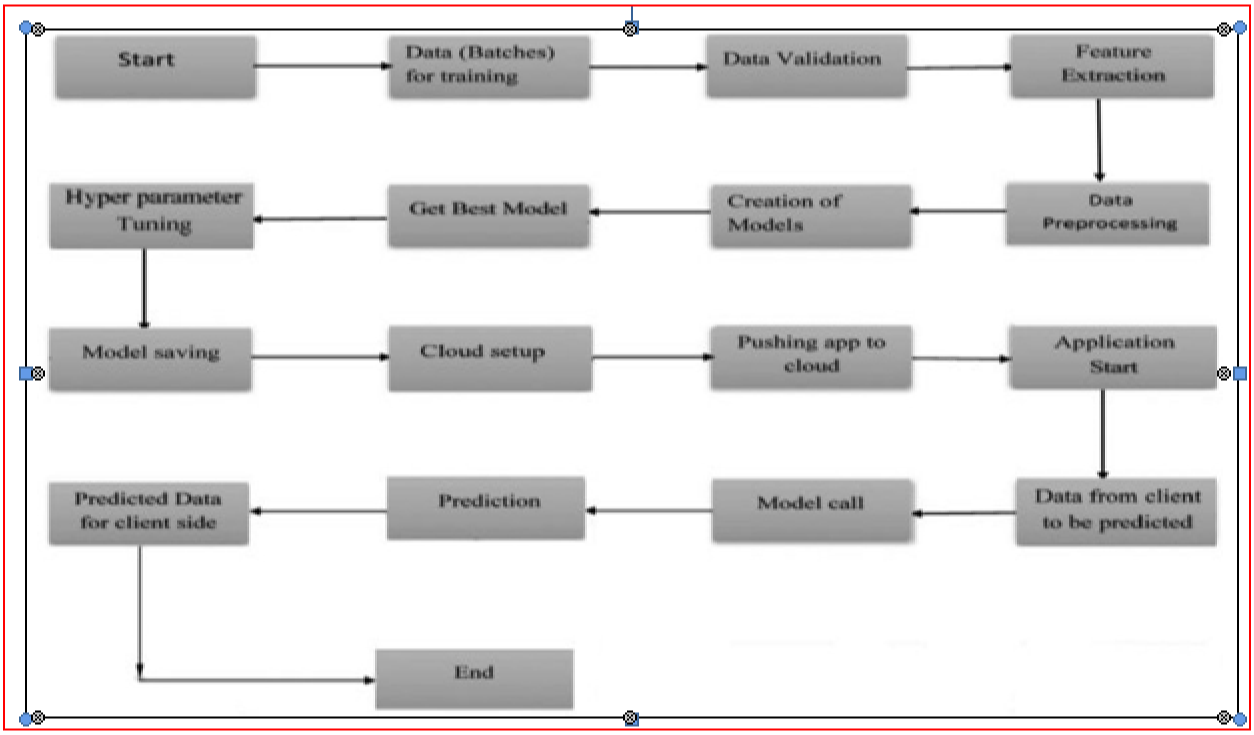


Fig)Data flow diagram

User Type	Functional Requirement (Epic)	User Story Number	User Story /Task	Acceptance criteria	Priority	Release
Customer	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my	I can access my account / dashboard	High	Sprint 1

			password.			
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint 1
		USN-3	As a user, I can register for the application through Facebook	I can register & access the dashboard with Facebook Login	Low	Sprint 2
		USN-4	As a user, I can register for the application through Gmail		Medium	Sprint 1
	Login	USN-5	Login As a user, I can log into the application by entering email &password		High	Sprint 1

5.2 Solution & Technical Architecture

Technical Architecture:

The Deliverable shall include the architectural diagram as below and the information as per the table1 & table 2.

Example :

Table 1 : Technical Characteristics

Sl/No	Component	Description	Technology
1	User Interface	How user interacts with application Web UI	HTML, CSS
2	Application Logic-1	Logic for a process in the application	Python
3	Machine Learning Model	Purpose of Machine Learning Model	ML Classifiers, etc. Table-2:

Table 2 : Application Characteristics:

Sl/No	Characteristics	Description	Technology
1	Open-Source Frameworks	It is a website which can be used on any platform to check whether a website is a phishing site or not.	HTML , CSS.
2	Performance	Only one model is used to detect whether or not a site is phishing, so it gives the result instantaneously after the details are entered and it is submitted	ML Classification Models

5.3 User Stories

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority

Sprint 1	Home page	USN-1	As a user, I can explore the resources of the homepage for the functioning	**5**	**Low**
Sprint 1	User Input	USN-2	As a user, I will inputs an URL in the required field to check its validation	**5**	**Low**
Sprint 1	Website comparison	USN-3	model checks for the feature extraction for prediction	**20**	**High**
Sprint-2	Feature Extraction	USN-4	After comparison if non found on comparison then it extract feature using heuristic and visual similarities.	**20**	**High**
Sprint-2	prediction	USN-5	Model predicts the URL using machine learning algorithms	**10**	**Medium**

			such as logistic Regression.		
Sprint-3	classifier	USN-6	Model sends all the output to the classifier and produces the final result.	**20**	**High**

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Sprint	**Functional Requirement**	**User Story Number**	**User Story /Task**	**Story Points**	**Priority**	**Team Members**
Sprint-1	Registration	US1	As a user, I can register forthe application by entering my email, password, and confirming my password.	2	High	Jeyanthi LakshmiG
Sprint-1		US2	As a user, I will receive confirmation email once I have registered for the application.	2	High	Balavika K

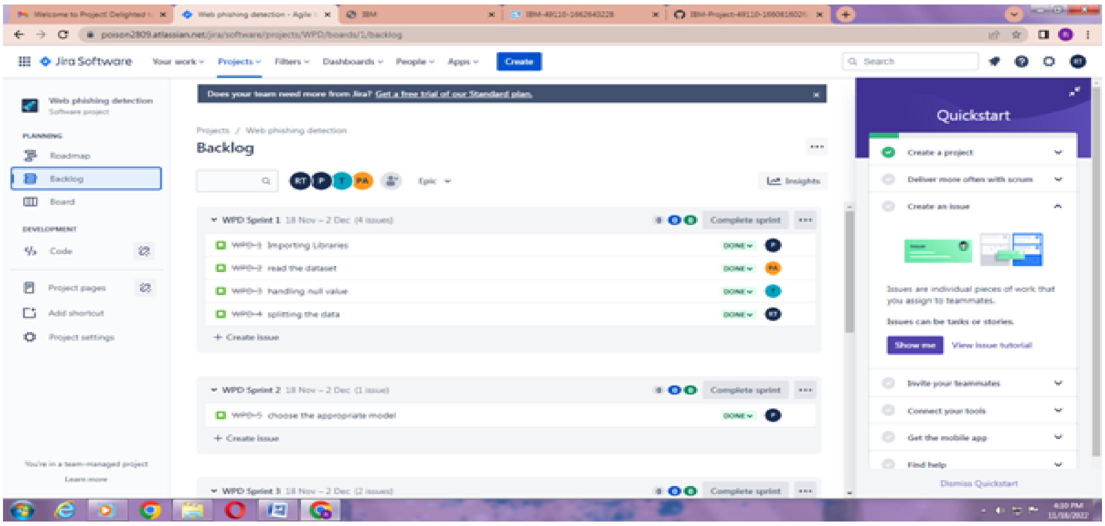
Sprint-2		US3	As a user, I can register forthe application through Facebook.	2	Low	Indumathy P
Sprint-1		US4	As a user, I can register forthe application through Gmail.	2	Medium	Lavanya G
Sprint-1	Login	US5	As a user, I can log into theapplication by entering email & password.	2	High	Jeyanthi LakshmiG
Sprint-1	Dashboard	US6	As a user, I can easily navigate through dashboardand I can use the dashboardto getdetails about app and instruction to use the app.	2	High	Balavika K
Sprint-1	Login and Dashboard	US7	As a web app user, I can login into application by using my email and password and I can access allresources same as mobile users.	2	High	Indumathy PLavanya G
Sprint-1	Login	CCE1	As a CCE I can login to appusing my idand password and I can interact with user.	2	High	Jeyanthi LakshmiG
Sprint-1	Dashboard	CCE2	As a CCE I can access dashboard using id and password and I can see alluser queries, explain app usage and attend their queries.	2	High	Balavika K

Sprint-1	Login and Dashboard	A1	As an administrator, I can login and access dashboard and manageand direct activities.	2	High	Indumathy PLavanya G

6.2 Sprint Delivery Schedule

Sprint	Total StoryPoints	Duration	SprintStart Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	SprintReleaseDate (Actual)
Sprint- 1	20	6 Days	24 Oct 2022	29 Oct 2022	20	7 Oct 2022
Sprint- 2	20	6 Days	31 Oct 2022	05 Nov 2022	20	10 Nov 2022
Sprint- 3	20	6 Days	07 Nov 2022	12 Nov 2022	20	17 Nov 2022
Sprint- 4	20	6 Days	14 Nov 2022	18 Nov 2022	20	17 Nov 2022

6.3 Reports from JIRA



Does your team need more from Jira? Get a free trial of our Standard plan.

Projects / Web phishing detection

Backlog

WPD Sprint 1 18 Nov – 2 Dec (4 issues)

- WPD-1 Importing Libraries DONE
- WPD-2 read the dataset DONE
- WPD-3 handling null value DONE
- WPD-4 splitting the data DONE

+ Create issue

WPD Sprint 2 18 Nov – 2 Dec (1 issue)

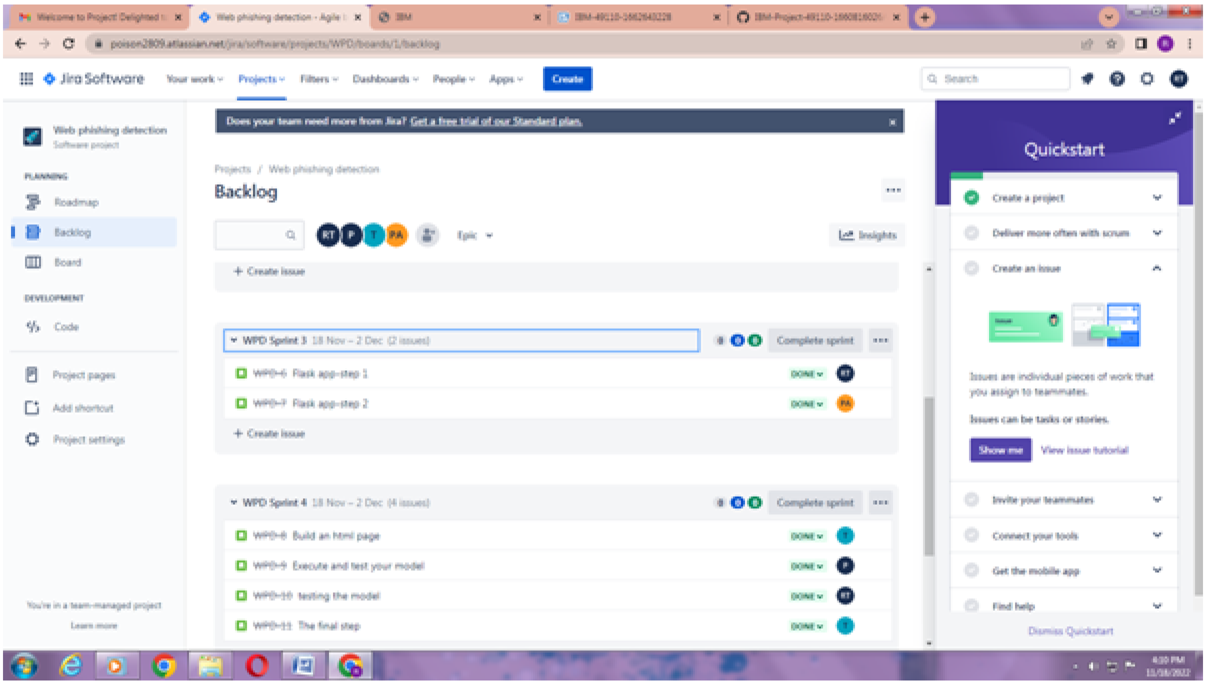
- WPD-5 choose the appropriate model DONE

+ Create issue

WPD Sprint 3 18 Nov – 2 Dec (2 issues)

Issues are individual pieces of work that you assign to teammates.

Issues can be tasks or stories.



poison2809.atlasian.net/jira/software/projects/WPD/boards/1/backlog

Does your team need more from Jira? Get a free trial of our Standard plan.

Web phishing detection
Software project

Projects / Web phishing detection

Backlog

WPD Sprint 3 18 Nov – 2 Dec (2 issues)

- ~~WPD-6~~ Flask app-step 1 DONE
- ~~WPD-7~~ Flask app-step 2 DONE

WPD Sprint 4 18 Nov – 2 Dec (4 issues)

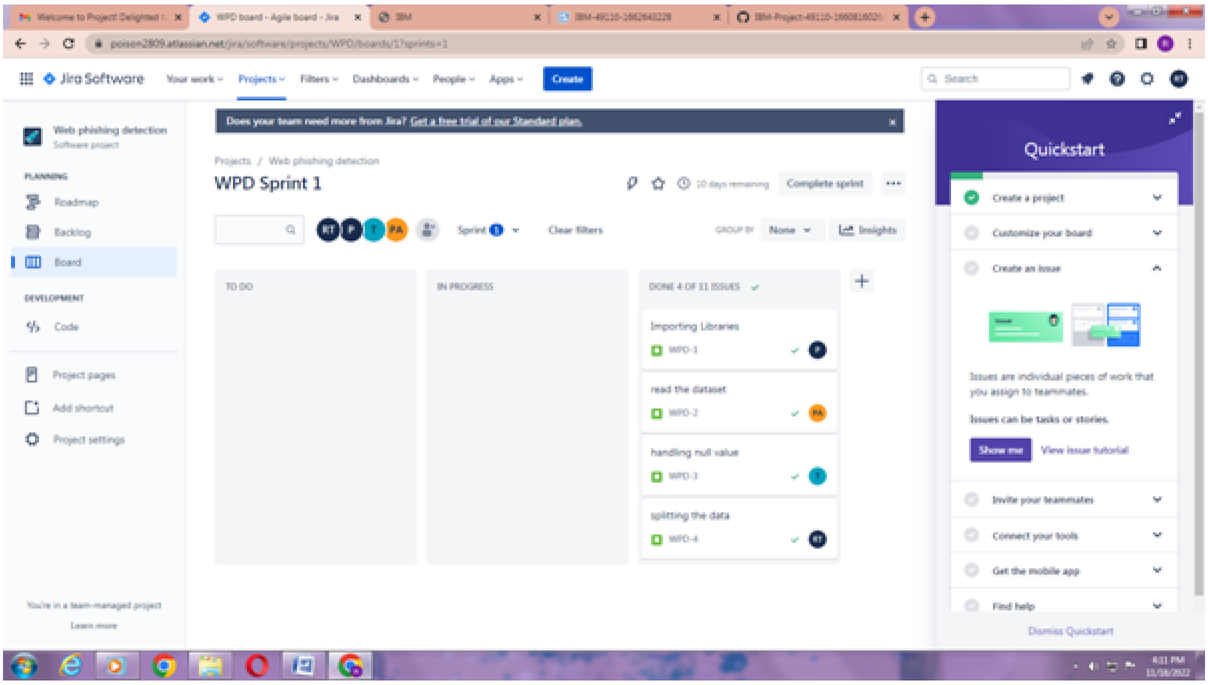
- ~~WPD-8~~ Build an html page DONE
- ~~WPD-9~~ Execute and test your model DONE
- ~~WPD-10~~ testing the model DONE
- ~~WPD-11~~ The final step DONE

Quickstart

Issues are individual pieces of work that you assign to teammates.

Issues can be tasks or stories.

You're in a team-managed project



poison2809.atlasian.net/jira/software/projects/WPD/boards/1?sprints=1

Does your team need more from Jira? Get a free trial of our Standard plan.

Web phishing detection
Software project

Projects / Web phishing detection

WPD Sprint 1

10 days remaining

GROUP BY None

TO DO	IN PROGRESS	DONE 4 OF 11 ISSUES ✓
		Importing Libraries WPD-1
		read the dataset WPD-2
		handling null value WPD-3
		splitting the data WPD-4

Quickstart

Issues are individual pieces of work that you assign to teammates.

Issues can be tasks or stories.

You're in a team-managed project

7. CODING & SOLUTIONING

7.1 Features

```html
<html>
<head>
    <title>Web Phishing detection</title>
    <link rel="stylesheet" href="/static/css/index.css"/>
</head>
<body>
    <div class="btn">
        <h1>Detect The Fake Websites</h1>
        <form class="from">
            <input type="button" id="btn"
onclick="window.location.href='http://localhost:5000/predict';" value="Get Started" />
          </form>
    </div>
</body>
</html>
```

7.2 Features

```python
from flask import Flask,request,jsonify,render_template
import pickle
import numpy as np
from inputScript import FeatureExtraction

app=Flask(__name__,template_folder='templates')

with open("D:\\Randam forest\\flask\\phishing_website.pkl","rb") as pickle_file:
    loaded_model=pickle.load(pickle_file)



@app.route('/')
def home():
    return render_template('index.html')



@app.route('/predict')
def predict():
    return render_template('final.html')


@app.route('/y_predict', methods=['POST'])
def y_predict():
    url = request.form['URL']
    url1 = FeatureExtraction(url)
    x = np.array(url1.getFeaturesList()).reshape(1, 30)

    prediction = loaded_model.predict(x)[0]

    print(prediction)
    if (prediction == 1):
        return "Your are safe!! This is a Legitimate website"
    else:
        return "You are on the wrong site Be cautious!"


@app.route('/predict_api', methods=['POST'])
def predict_api():
    data = request.ger_json(force=True)
    prediction = loaded_model.y_predict([np.array(list(data.values()))])
    output = prediction[0]
    return jsonify(output)


if __name__ == '__main__':
    app.run()
```

7.3 Database Schema

```
import ipaddress
import re
import urllib.request
from bs4 import BeautifulSoup
import socket
import requests
from googlesearch import search
import whois
from datetime import date, datetime
from urllib.parse import urlparse

from urllib3.util import response, url


class FeatureExtraction:
    features = []

    def __init__(self, url):
        self.features = []
        self.url = url
        self.domain = ""
        self.whois_response = ""
        self.urlparse = ""
        self.response = ""
        self.soup = ""

        try:
            self.response = requests.get(url)
            self.soup = BeautifulSoup(response.text, 'html.parser')
        except:
            pass

        try:
            self.urlparse = urlparse(url)
            self.domain = self.urlparse.netloc
        except:
            pass

        try:
            self.whois_response = whois.whois(self.domain)
        except:
            pass

        self.features.append(self.UsingIp())
        self.features.append(self.longUrl())
        self.features.append(self.shortUrl())
        self.features.append(self.symbol())
        self.features.append(self.redirecting())
        self.features.append(self.prefixSuffix())
        self.features.append(self.SubDomains())
        self.features.append(self.Hppts())
        self.features.append(self.DomainRegLen())
        self.features.append(self.Favicon())

        self.features.append(self.NonStdPort())
        self.features.append(self.HTTPSDomainURL())
        self.features.append(self.RequestURL())
        self.features.append(self.AnchorURL())
        self.features.append(self.LinksInScriptTags())
        self.features.append(self.ServerFormHandler())
        self.features.append(self.InfoEmail())
        self.features.append(self.AbnormalURL())
        self.features.append(self.WebsiteForwarding())
        self.features.append(self.StatusBarCust())

        self.features.append(self.DisableRightClick())
        self.features.append(self.UsingPopupWindow())
        self.features.append(self.IframeRedirection())
        self.features.append(self.AgeofDomain())
        self.features.append(self.DNSRecording())
        self.features.append(self.WebsiteTraffic())
        self.features.append(self.PageRank())
        self.features.append(self.GoogleIndex())
        self.features.append(self.LinksPointingToPage())
        self.features.append(self.StatsReport())

    # 1.UsingIp
    def UsingIp(self):
        try:
            ipaddress.ip_address(self.url)
            return -1
        except:
            return 1

    # 2.longUrl
    def longUrl(self):
        if len(self.url) < 54:
            return 1
        if len(self.url) >= 54 and len(self.url) <= 75:
                return 0
            return -1

        # 3.shortUrl
        def shortUrl(self):
            match =
    re.search('bit\.ly|goo\.gl|shorte\.st|go2l\.ink|x\.co|ow\.ly|t\.co|tinyurl|tr\.im|is\.gd|cli\.gs|'

    'yfrog\.com|migre\.me|ff\.im|tiny\.cc|url4\.eu|twit\.ac|su\.pr|twurl\.nl|snipurl\.com|'

    'short\.to|BudURL\.com|ping\.fm|post\.ly|Just\.as|bkite\.com|snipr\.com|fic\.kr|loopt\.us|'

    'doiop\.com|short\.ie|kl\.am|wp\.me|rubyurl\.com|om\.ly|to\.ly|bit\.do|t\.co|lnkd\.in|'

    'db\.tt|qr\.ae|adf\.ly|goo\.gl|bitly\.com|cur\.lv|tinyurl\.com|ow\.ly|bit\.ly|ity\.im|'

    'q\.gs|is\.gd|po\.st|bc\.vc|twitthis\.com|u\.to|j\.mp|buzurl\.com|cutt\.us|u\.bb|yourls\.org|'

    'x\.co|prettylinkpro\.com|scrnch\.me|filoops\.info|vzturl\.com|qr\.net|1url\.com|tweez\.me|v\.
    gd|tr\.im|link\.zip\.net',
                              self.url)
            if match:
                return -1
            return 1

        # 4.Symbol@
        def symbol(self):
            if re.findall("@", self.url):
                return -1
            return 1

        # 5.Redirecting//
        def redirecting(self):
            if self.url.rfind('//') > 6:
                return -1
            return 1

        # 6.prefixSuffix
        def prefixSuffix(self):
            try:
```

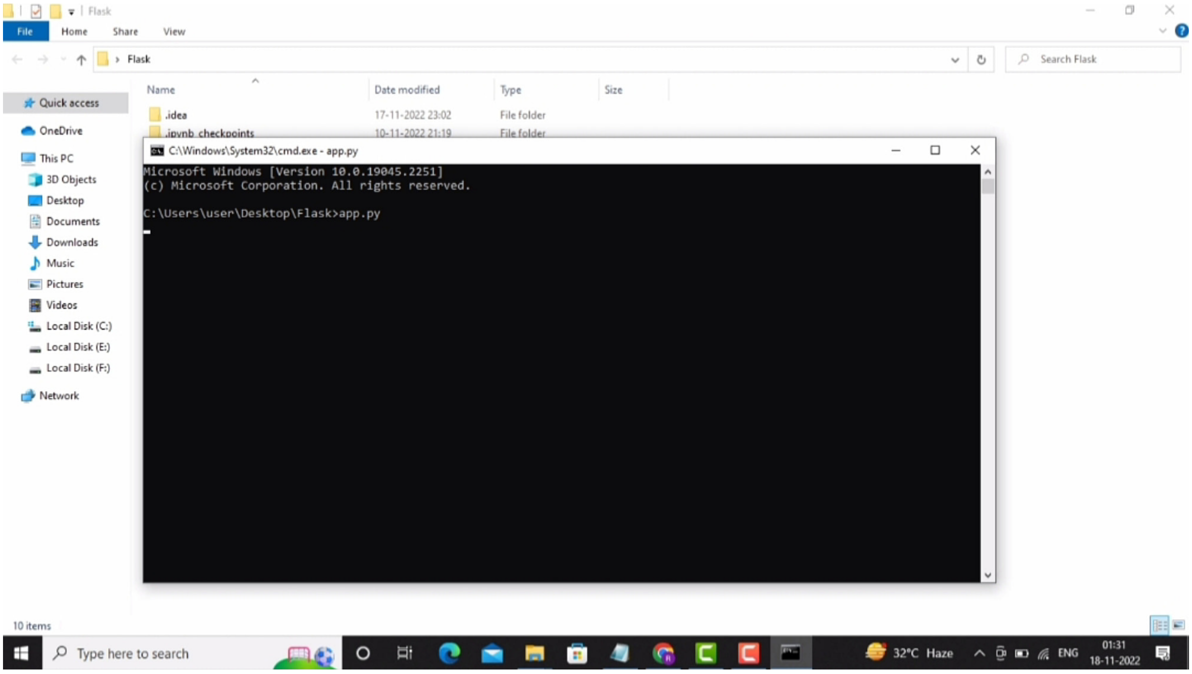
```
        match = re.findall('\-', self.domain)
        if match:
            return -1
        return 1
    except:
        return -1

# 7.SubDomains
def SubDomains(self):
    dot_count = len(re.findall("\.", self.url))
    if dot_count == 1:
        return 1
    elif dot_count == 2:
        return 0
    return -1

# 8.HTTPS
def Hppts(self):
    try:
        https = self.urlparse.scheme
        if 'https' in https:
           return 1
        return -1
    except:
```

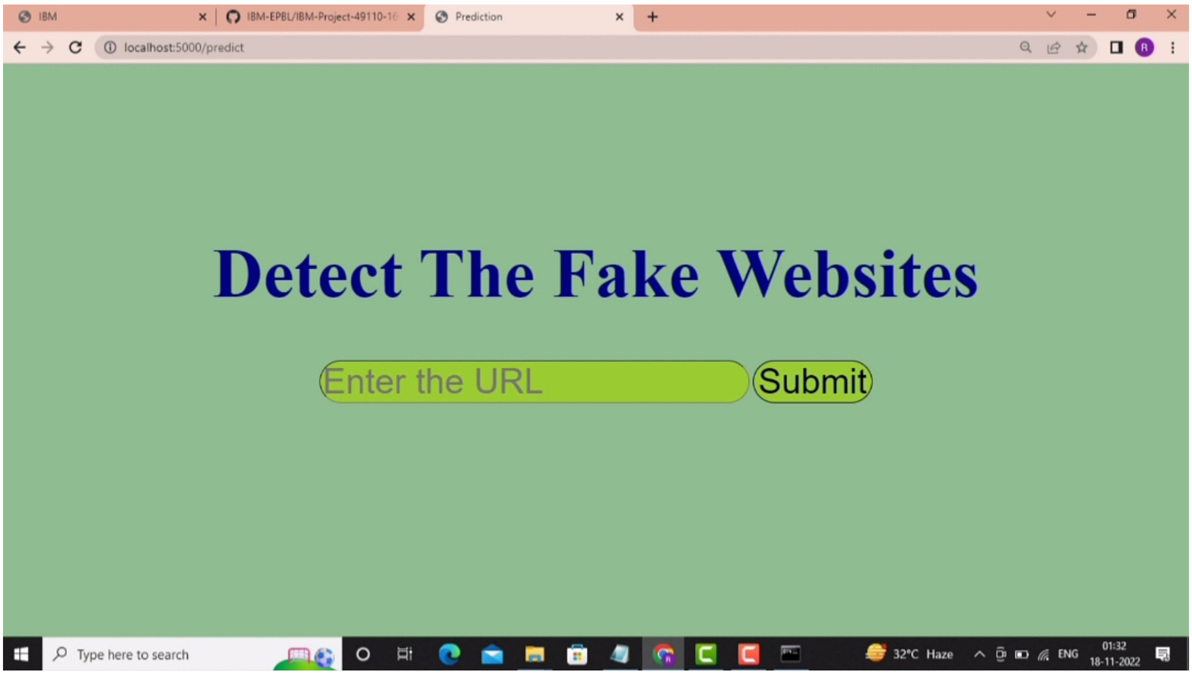
8.Testing

8.1 Test Cases

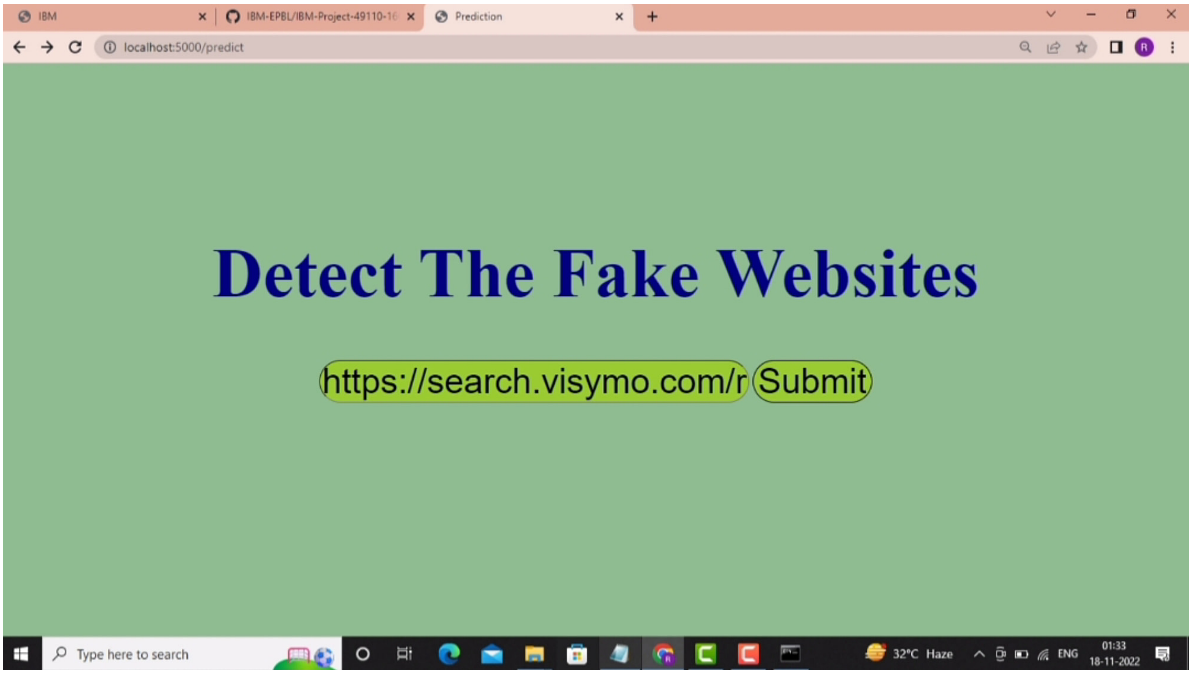


```
Microsoft Windows [Version 10.0.19045.2251]
(c) Microsoft Corporation. All rights reserved.

C:\Users\user\Desktop\Flask>app.py
```

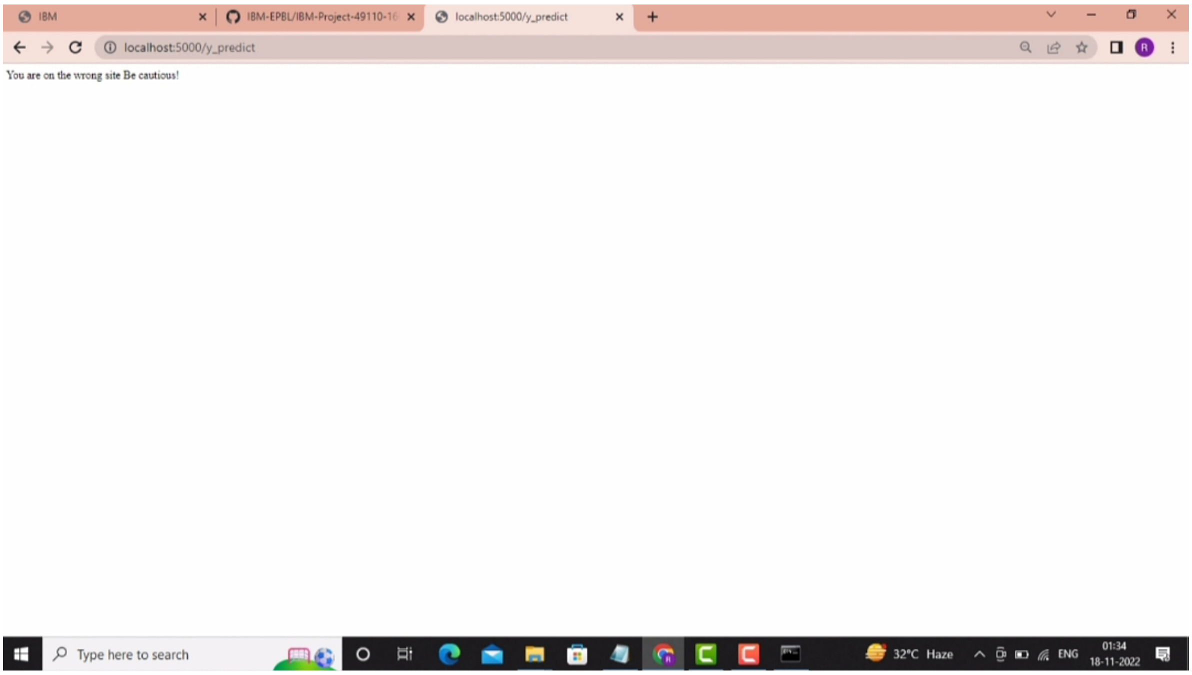


Detect The Fake Websites



Detect The Fake Websites

https://search.visymo.com/r



You are on the wrong site Be cautious!

RESULTS:

a. Performance Metrics

i. Accuracy

The accuracy metric is one of the simplest Classification metrics to implement, and it can be determined as the number of correct predictions to the total number of predictions.

ii. Confusion Matrix

A confusion matrix is a tabular representation of prediction outcomes of any binary classifier, which is used to describe the performance of the classification model on a set of test data when true values are known. The confusion matrix is simple to implement, but the terminologies used in this matrix might be confusing for beginners.

10.ADVANTAGES & DISADVANTAGES

Advantages:

- This system can be used by many E-commerce Websites in order to have good customer relationship.
- User can make online payment securely.
- Data mining algorithm used in this system provides better performance as compared to other traditional classifications algorithms.
- With the help of this system user can also purchase products online without any hesitation.

Disadvantages:

- If Internet connection fails, this system won't work.
- All e-banking websites related data will be stored in one place.
- System will match the review with those keywords which are in database rest of the words are not considered by the system.

11. CONCLUSION

The most important way to protect the user from phishing attack is the education awareness. Internet user must be aware of all the security tips which are given by experts. Every user must be trained to blindly follow the links to the websites where they have to send their sensitive information. It is essential to check the URL before entering the websites.

Here we have proposed a Random Forest Classification algorithm to predict the phishing
website based on their features. User can enter any URL to predict whether the website is phishing website or not.

Our proposed system has the accuracy of 93%. As we have implemented this algorithm by considering the URL and Domain Identity criteria, there are different criteria needs to work in future and to have an accuracy of 100%

12. FUTURE SCOPE

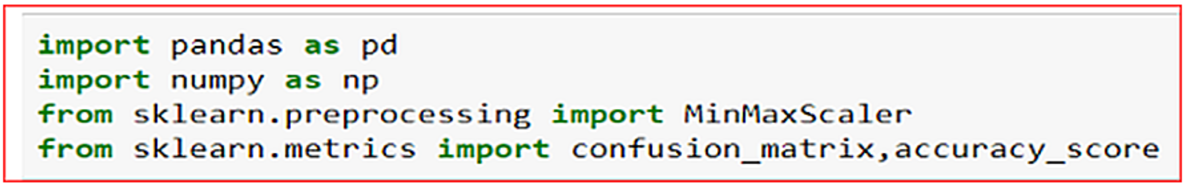
Today most of the banking happens while you are sipping coffee or taking an important call. ATMs are at your doorstep. Banking services are accessible 24x7. There are more plastic cards in your wallet than currency notes. A huge part of this change is due to advent of IT. Banks today operate in a highly globalized, liberalized, privatized and a competitive environment. In order to survive in this environment banks have to use IT. Indian banking industry has witnessed a tremendous developments due to sweeping changes that are taking place in the information technology. This work involves descriptive research design as my project is questionnaire based. Descriptive research includes survey and fact-finding enquiries kinds. The major purpose of descriptive research is description of the state of affairs, as it exists at present. For this study the sample size is 50 people of the area New Delhi, who were using the E-Banking services.

13. APPENDIX

APPENDIX A

Source Code

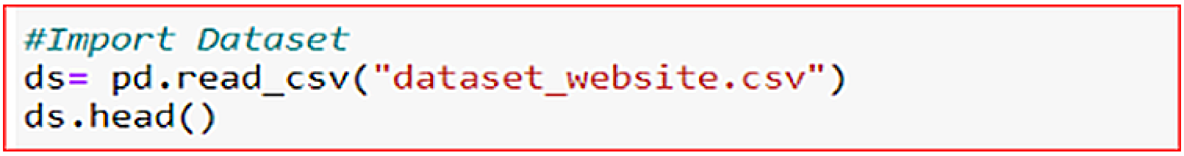
Importing the libraries



```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import confusion_matrix,accuracy_score
```

Figure 1: Snapshot importing the libraries

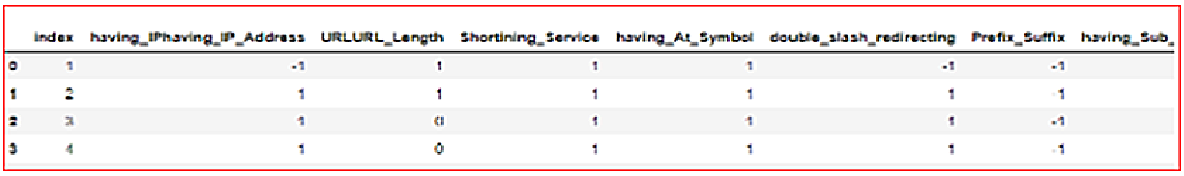
Reading the dataset



```
#Import Dataset
ds= pd.read_csv("dataset_website.csv")
ds.head()
```

Figure 2: Reading the dataset

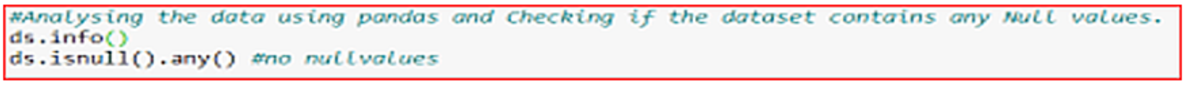
Sample output of the dataset rows :



	index	having_IPhaving_IP_Address	URLURL_Length	Shortining_Service	having_At_Symbol	double_slash_redirecting	Prefix_Suffix	having_Sub_
0	1	-1	1	1	1	-1	-1	
1	2	1	1	1	1	1	-1	
2	3	1	0	1	1	1	-1	
3	4	1	0	1	1	1	-1	

Figure 3: Sample output of the dataset

Handling Null Values



```
#Analysing the data using pandas and Checking if the dataset contains any Null values.
ds.info()
ds.isnull().any() #no nullvalues
```

Figure 4: Handling null values

Identifying Independent & dependent variables:

Figure5: Identifying Independent & dependent variables

Splitting the data:

Figure 6: Splitting the data

Logistic Regression

Figure7: Logistic Regression

RANDOM FOREST

Figure8: Random Forest

Sample Snapshots

Integrated development environment

Figure 9: Snapshot - Anaconda IDE