

# **WEB PHISHING DETECTION USING MACHINE LEARNING A MINI PROJECT REPORT**

**Submitted by**

**SHREE BHARGAV RK (AC19UCS154)**

**SANTHOSH PRIYAN S (AC19UCS098)**

**SHASHANK S (AC19UCS104)**

**YOGA PRIYAN M (AC19UCS142)**

*in partial fulfillment for the award of the degree  
of*

**BACHELOR OF ENGINEERING  
in  
COMPUTER SCIENCE AND ENGINEERING**

.....

**ADHIYAMAAN COLLEGE OF ENGINEERING**

**DR. M.G.R NAGAR, HOSUR-635130**

**ANNA UNIVERSITY: CHENNAI 600 025**

**NOVEMBER 2022**

# CONTENTS

## CHAPTER 1 – INTRODUCTION

1.1	Project Overview .....	4
1.2	Purpose .....	4

## CHAPTER 2 – LITERATURE SURVEY

2.1	Existing problem .....	5
2.2	References .....	6
2.3	Problem Statement Definition .....	6

## CHAPTER 3 – IDEATION & PROPOSED SOLUTION

3.1	Empathy Map Canvas.....	7
3.2	Ideation & Brainstorming.....	8
3.3	Proposed solution .....	11
3.4	Problem Solution Fit .....	12

## CHAPTER 4 – REQUIREMENT ANALYSIS

4.1	Functional Requirements.....	13
4.2	Non - Functional Requirements .....	14

## CHAPTER 5 - PROJECT DESIGN

5.1	Data Flow Diagrams.....	16
5.2	Solution & Technical Architecture .....	17
5.3	User Stories .....	18

**CHAPTER 6 - PROJECT PLANNING & SCHEDULING**

6.1 Sprint Planning & Estimation..... 20

6.2 Sprint Delivery Schedule..... 22

6.3 Reports From JIRA..... 23

**CHAPTER 7 - CODING & SOLUTIONING**

7.1 Feature 1 .....24

7.2 Feature 2 .....27

**CHAPTER 8 - TESTING**

8.1 Test Cases .....30

8.2 User Acceptance Testing.....30

**CHAPTER 9 - RESULTS**

9.1 Performance Metrics ..... 31

**CHAPTER 10 - ADVANTAGES & DISADVANTAGES .....33**

**CHAPTER 11 - CONCLUSION.....34**

**CHAPTER 12 - FUTURE SCOPE ..... 35**

**CHAPTER 13 - APPENDIX**

Source code ..... 36

GitHub & Project Demo Link ..... 41

# **CHAPTER 1**

## **INTRODUCTION**

### **1.1 INTRODUCTION**

#### **1.1. Project Overview**

There are a number of users who purchase products online and make payments through e-banking. There are e-banking websites that ask users to provide sensitive data such as username, password & credit card details, etc., often for malicious reasons. This type of e-banking website is known as a phishing website. Web service is one of the key communications software services for the Internet. Web phishing is one of many security threats to web services on the Internet.

#### **1.2. Purpose**

To preserve confidentiality, To protect the user from phishing websites, To develop a user-friendly environment, To prevent or mitigate harm or destruction of computer networks, applications, devices, and data.

## **CHAPTER 2**

### **LITERATURE SURVEY**

In this survey paper they have mentioned how phishing attacks appear, how the phishers use email or message, as evidence to target the individual or business by sending the link to victim people and deceive them with a large no of phishing emails or messages every day, so many of the corporations or individual are not able to recognize them all. so, here they have mentioned various types of phishing attacks like Learning Model Algorithm, CAT boost classifier, Decision tree, support vector machine and random forest.

#### **2.1 Existing problem**

Phishing Website Detection model is trained using an existing dataset which contains URLs, each with unique features, and is applied to three different machine learning classifiers - support vector machine, logistic regression and Naïve Bayes. After training and testing the algorithms, it is observed that Naïve Bayes classifier recorded the highest accuracy

## 2.2 References

<https://drive.google.com/file/d/1PdQhfWsZv9YO10tEOo9hiYr5fv11ArA-/view?usp=drivesdk>

<https://drive.google.com/file/d/1PAn-nVBSvA3Ivx8te1XbbEAYBZHTz9RX/view?usp=drivesdk>

<https://drive.google.com/file/d/1PlOjJnAFsD6IYAYqCH702ALfT2ua2I1D/view?usp=drivesdk>

## 2.3 Problem Statement Definition

Scam emails have become a noticeable problem in society today. The purpose of this paper is to research the effect age and education level has on a person's ability to identify if an email is credible or not. Research has indicated that scammers often target the elderly. Exploring why this is the case and if factors such as education also have an effect on judgment regarding emails is therefore worth investigating. This study aims to investigate the following. How does the effectiveness of phishing mail vary in recipients of different ages and educational levels?

# CHAPTER 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 behaviors and attitudes.

It is a useful tool to help teams 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.




## 3.2 Ideation & Brainstorming

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.

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.

### Step-1: 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.

- 🕒 10 minutes to prepare
- 🕒 1 hour to collaborate
- 👤 2-8 people recommended

Share template feedback

➔

**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.

🕒 10 minutes

A

**Team gathering**

Define who should participate in the session and send an invite. Share relevant information or pre-work ahead.

B

**Set the goal**

Think about the problem you'll be focusing on solving in the brainstorming session.

C

**Learn how to use the facilitation tools**

Use the Facilitation Superpowers to run a happy and productive session.

Open article ➔

1

**Define your problem statement**

What problem are you trying to solve? Frame your problem as a How Might We statement. This will be the focus of your brainstorm.

🕒 5 minutes

PROBLEM

The main aim of Web Phishing is to identify the phishing links and avoid phishing done by attackers which will be useful for an organization or any individual

Key rules of brainstorming

To run a smooth and productive session

🗨️

Defer judgment.

🗨️

Listen to others.

🗨️

Go for volume.

💡

Encourage wild ideas.

👁️

If possible, be visual.

Stay in topic.

8



## Step-2: Brainstorm, Idea Listing and Grouping

2

### Brainstorm

Write down any ideas that come to mind that address your problem statement.

10 minutes

#### TIP

You can select a sticky note and hit the pencil (switch to sketch) (can't select drawing)

#### Shree Bhargav RK

- Pharming has to be avoided in an organization
- While dealing with clients the messages which are to be sent and received must be in communication media that is encrypted
- Should maintain the customer data as secured. Otherwise there will be reputational damage.
- Operating system and the softwares should be up to date
- Online Industries
- Clients also should be aware of suspicious links due to the malicious activities done by attackers
- Only open the link that begins with https and avoid using the links starts with http.
- Update the security measures around the sensitive data of the organization

#### Santhosh Priyan S

- Delete the email without opening it.
- Anti-phishing technology is designed to identify and block phishing mails
- Use safe browsing in chrome
- Avoid clicking embedded URLs
- EMAIL
- Purchase an extra line of security
- Manually block the sender
- Convert HTML email into text only email messages or disable HTML email messages

#### Shashank S

- Choose a secure ecommerce platform
- Sharing the usernames and passwords should be avoided as it lead to threatening to privacy
- Use a secure connection for online
- Employ an address and card verification system
- Ecommerce Site
- Set up system alerts for suspicious activity
- Don't store sensitive data
- Require strong passwords

#### Yogapriyan M

- Avoid mails and text messages that fraudsters send which is pretend to be from a bank
- If we came to know that our bank account is hacked then we should contact the bank authorities
- Bank password should be changed frequently
- Don't give credit card informations in unknown sites
- E-Banking
- "Remember password" in google should not be enabled for bank transactions
- Don't do the transactions in the sites which is not trustful
- Contact the company if we get suspicious mail from the bank name

3

### Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. In the last 10 minutes, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you can break it up into smaller sub-groups.

20 minutes

#### Detection

The links that is send should be checked with the database and compare whether any attacks done by that link

The links must be classify as blacklist and whitelist and pass the information about the link to the user

A Combined blacklist-based ,heuristic and web based approach using algorithms can detect the fraud sites

Anti phishing tools can be used to reduce the phishing attacks

#### Protection

Protect the data by taking backup of it.

Close immediately the popup which is opened without your permission

Verify SSL Certification that is to check whether the link begins with https.

Never click and download software or files from an unknown source as some programs like trojan can be installed affect our data in the system

#### Prevention

They use fake DNS names that are similar but not identical with the target website, it must be also aware of it.

Visual link and actual link may vary, that must be aware of it.

Firewalls and antivirus must be installed in the system

Have a Data Security platform to spot signs of an attack

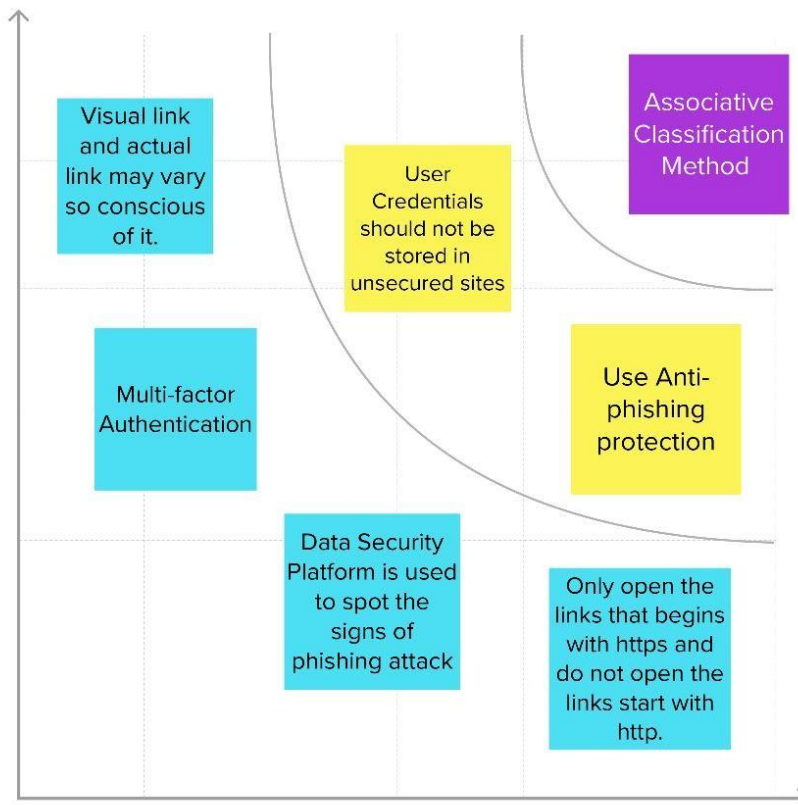
## Step-3: Idea Prioritization

4

### Prioritize

Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.

20 minutes



→

### After you collaborate

You can export the mural as an image or pdf to share with members of your company who might find it helpful.

### Quick add-ons

A

#### Share the mural

Share a view link to the mural with stakeholders to keep them in the loop about the outcomes of the session.

B

#### Export the mural

Export a copy of the mural as a PNG or PDF to attach to emails, include in slides, or save in your drive.

### Keep moving forward



#### Strategy blueprint

Define the components of a new idea or strategy.

[Open the template →](#)



#### Customer experience journey map

Understand customer needs, motivations, and obstacles for an experience.

[Open the template →](#)



#### Strengths, weaknesses, opportunities & threats

Identify strengths, weaknesses, opportunities, and threats (SWOT) to develop a plan.

[Open the template →](#)

[Share template feedback](#)

### 3.3 PROPOSED SOLUTION

- Problem Statement (Problem to be solved) - Phishing links causes major theft of data and privacy which has damage for an organization or any individual
- Idea / Solution description - Identifying the phishing links and avoiding phishing done by attackers which will be useful for an organization or any individual.
- Novelty / Uniqueness - We are using TF-IDF algorithm and various algorithms to make an accurate prediction of phishing.
- Social Impact / Customer Satisfaction - It provides an enhanced level of phishing protection to detect attacks faster, alert users and remediate threats as quickly as possible and provides cyber security awareness in the society
- Business Model (Revenue Model) - With using: Can stop the attack done by attackers and prevent the loss of data. Without using: Cause financial loss for victims. Put their personal information at risk.
- Scalability of the Solution - This can help the organizations all over the world to prevent their intellectual property and the reputation

## 3.4 PROBLEM 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 behavioral patterns and recognize what would work and why

Define CS, fit into CC	<b>1. CUSTOMER SEGMENT(S)</b> <span>CS</span> <ul style="list-style-type: none"> <li>● Business Organization</li> <li>● Online Banking Sector</li> <li>● Those who use Websites and URL 's for surfing through internet</li> </ul>	<b>6. CUSTOMER CONSTRAINTS</b> <span>CC</span> <p>Provides full access to scan the transaction process of the user and no breakdown of server connections</p>	<b>5. AVAILABLE SOLUTIONS</b> <span>AS</span> <p>This is applied to three different machine learning classifier - support vector machine, logistic regression and Naive Bayes. After training and testing the algorithms, it is observed that Naive Bayes classifier recorded the highest accuracy</p>	Explore AS, differentiate
	<b>2. JOBS-TO-BE-DONE / PROBLEMS</b> <span>J&amp;P</span> <p>To identify the phishing sites and to protect users Credentials from hackers</p>	<b>9. PROBLEM ROOT CAUSE</b> <span>RC</span> <p>Having the data without any protection using anti phishing technologies, So that attacker creates fake website and steal the data.</p>	<b>7. BEHAVIOUR</b> <span>BE</span> <p>Customer finds the web phishing detection websites or applications and also the customer should provide all the transaction details of whole process</p>	
Focus on J&P, tap into BE, understand RC	<b>3. TRIGGERS</b> <span>TR</span> <p>Customer will get triggered because of data get stolen, theft of money and loss of privacy.</p>	<b>10. YOUR SOLUTION</b> <span>SL</span> <p>The links that gets checked for identifying phishing ,and we will be using various algorithm for making accurate prediction. Especially we are using Ada Boost Algorithm to make high accuracy prediction.</p>	<b>8.CHANNELS of BEHAVIOR</b> <span>CH</span> <p><b>8.1 ONLINE</b></p> <p>Pass the URL as input and identify whether it is a phishing site or not.</p> <p><b>8.2 OFFLINE</b></p> <p>Using the phishing detection application to predict the phishing sites in offline mode(offload the app).</p>	Identify strong TR & EM
	<b>4. EMOTIONS: BEFORE / AFTER</b> <span>EM</span> <p><b>BEFORE :</b> Believing that the data is protected and secured in the Organization.</p> <p><b>AFTER :</b> Feeling depressed as the data and money have been stolen.</p>			

## **CHAPTER - 4**

### **REQUIREMENT ANALYSIS**

#### **4.1 FUNCTIONAL REQUIREMENT**

1. User Registration
  - a. Registration through Form
  - b. Registration through Gmail
  - c. Registration through Linked-IN
2. User Confirmation
  - a. Confirmation via Email
  - b. Confirmation via OTP
3. User Authentication
  - a. Authentication can be done with password or 2 step-verification
4. User input
  - a. Users need to download the extension and then input anURL(Uniform Resource Locator) in the necessary field to check its validation.
5. Website Evaluation
  - a. Model evaluates the website using Blacklist and Whitelist approach

## 6. Extraction and Prediction

- a. It retrieves features based on heuristics and visual similarities. The URL is predicted by the model using Machine Learning methods such as Logistic Regression and KNN

## 7. Real Time monitoring

- a. The use of Extension plugin should provide a warning pop-up when they visit a website that is phished. Extension plugin will have the capability to also detect latest and new phishing websites

## 8. Authentication

- a. Authentication assures secure sites, secure processes and enterprise information security.

# **4.2 NON - FUNCTIONAL REQUIREMENT**

## 1. Usability

- a. You will be able to secure your personal data andalso protect the data from phishing websites.

## 2. Security

- a. It is a secured website which protects these sensitive data to the user and prevents malicious attack.

### 3. Reliability

- a. It specifies the likelihood that the system or its component will operate without failure for a specified amount of time under prescribed conditions.

### 4. Performance

- a. The performance of web phishing detection is high and it is very efficient as it is very easy to understand and has a high security and scale-able.

### 5. Availability

- a. Users can access the website via any browser from anywhere at any time, because it's act as an extension.

### 6. Scalability

- a. This application can be accessed online without paying. It can detect any websites with a higher accuracy.

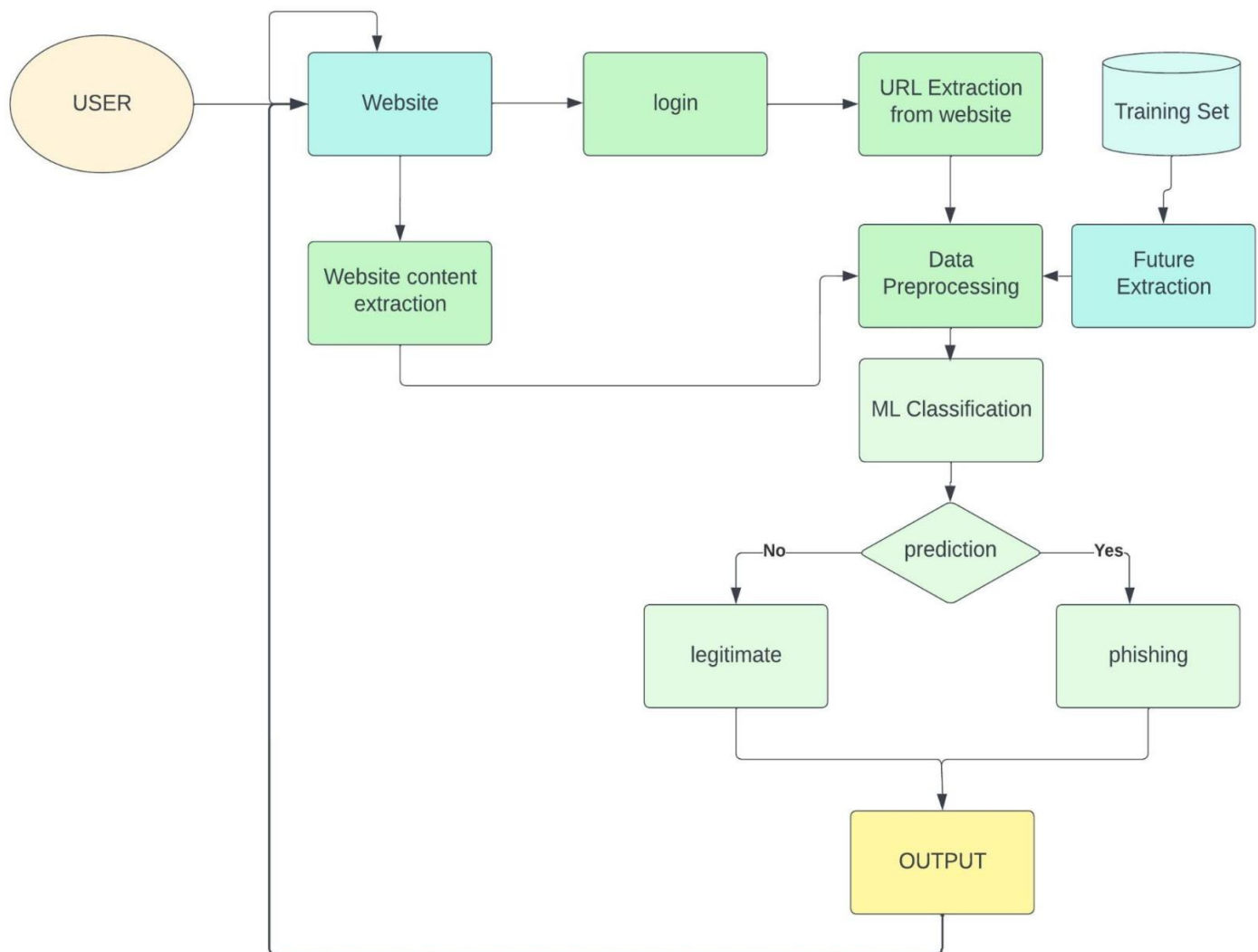
## CHAPTER 5

### PROJECT DESIGN

#### 5.1 DATA FLOW DIAGRAM

##### 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.



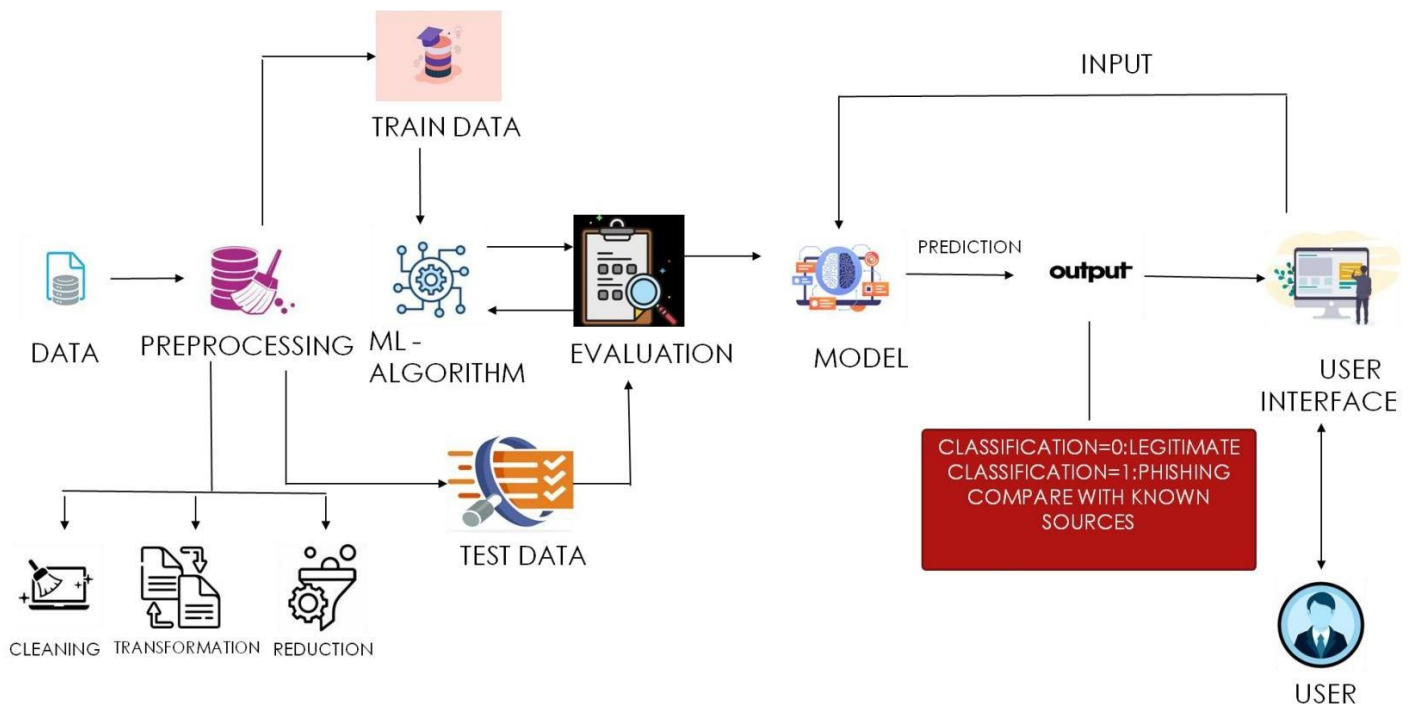


## 5.2 SOLUTION & TECHNICAL ARCHITECTURE

Solution architecture is a complex process – with many sub-processes – that bridges the gap between business

problems and technology solutions. Its goals are to:

- Find the best tech solution to solve existing business problems.
- Describe the structure, characteristics, behavior, and other aspects of the software to project stakeholders.
- Define features, development phases, and solution requirements.
- Provide specifications according to which the solution is defined, managed, and delivered.



## 5.3 USER STORIES

### USER TYPE: Customer (Mobile user)

- **FUNCTIONAL REQUIREMENTS (Epic): Registration**

- **USN 1** - As a user, I can register for the application by entering my email, password, and confirming password.
- **USN 2** - As a user, I will receive a confirmation email once I have registered for the application.
- **USN 3** - As a user, I can register for the application via Facebook.
- **USN 4** - As a user, I can register for the application via Gmail.

- **FUNCTIONAL REQUIREMENTS (Epic): Login**

- **USN 5** - As a user, I can log into the application by entering email & password.

- **FUNCTIONAL REQUIREMENTS (Epic): Dashboard**

### USER TYPE: Customer (Web user)

- **FUNCTIONAL REQUIREMENTS (Epic): User input**

- **USN 1** - As a user I can input the particular URL in the required field and wait for validation.

## **USER TYPE: Customer Care Executive**

- **FUNCTIONAL REQUIREMENTS (Epic) : Feature extraction**

- **USN 1** - After i compare in case if none is found on comparison then we can extract features using heuristic and visual similarity approach.

## **USER TYPE: Administrator**

- **FUNCTIONAL REQUIREMENTS (Epic): Prediction**

- **USN 1** - Here the Model will predict the URL websites using Machine Learning algorithms such as Logistic Regression, KNN

- **FUNCTIONAL REQUIREMENTS (Epic): Classifier**

- **USN 2** - Here I will send all the model output to the classifier in order to produce the final result.

## **CHAPTER 6**

### **PROJECT PLANNING & SCHEDULING**

#### **6.1 Sprint Planning & Estimation**

##### **Sprint -1:**

**USN-1** - As a user I can register for the application by entering my email, password and confirming my password.

##### **Team members:**

- SANTHOSH PRIYAN

**USN-2** - As a user, I can receive confirmation email once I have register for the application.

##### **Team members:**

- YOGA PRIYAN M

**USN-3** – As a user, I can login into application by entering email and password.

##### **Team member:**

- SHREE BHARGAV RK

## **Sprint-2:**

**USN-4** - As a user, I can easily navigate through dashboard and I can use the dashboard to get details about app and instruction to use the app

### **Team Members:**

- SHREE BHARGAV RK

**USN-5** - As a user, UI of second page will be displayed

### **Team member:**

- SHASHANK

## **Sprint – 3:**

**USN-6** - As a user, I can able to paste the URL to check whether it is phishing or not.

### **Team members:**

- SANTHOSH PRIYAN

## **Sprint – 4:**

**USN-7** - If the model Predict the URL as Phishing site or not with accuracy rate above 95%.

### **Team members:**

- YOGA PRIYAN M

**USN-8** - As a user, I will get the result that the given URL is a phishing or not.

### **Team member:**

- SHASHANK

## 6.2 Sprint Delivery Schedule

### Sprint - 1

- **Total Points** - 20
- **Duration** - 6 Days
- **Sprint Start date** - 24 Oct 2022
- **Sprint End date (Planned)** - 29 Oct 2022
- **Sprint Points Completed (as on Planned End Date)** - 20
- **Sprints Release Date (Actual)** - 29 Oct 2022

### Sprint - 2

- **Total Points** - 20
- **Duration** - 6 Days
- **Sprint Start date** - 31 Oct 2022
- **Sprint End date (Planned)** - 05 Nov 2022
- **Sprint Points Completed (as on Planned End Date)** - 20
- **Sprints Release Date (Actual)** - 05 Nov 2022

### Sprint - 3

- **Total Points** - 20
- **Duration** - 6 Days
- **Sprint Start date** - 07 Nov 2022
- **Sprint End date (Planned)** - 12 Nov 2022
- **Sprint Points Completed (as on Planned End Date)** - 20
- **Sprints Release Date (Actual)** - 12 Nov 2022

## Sprint - 4

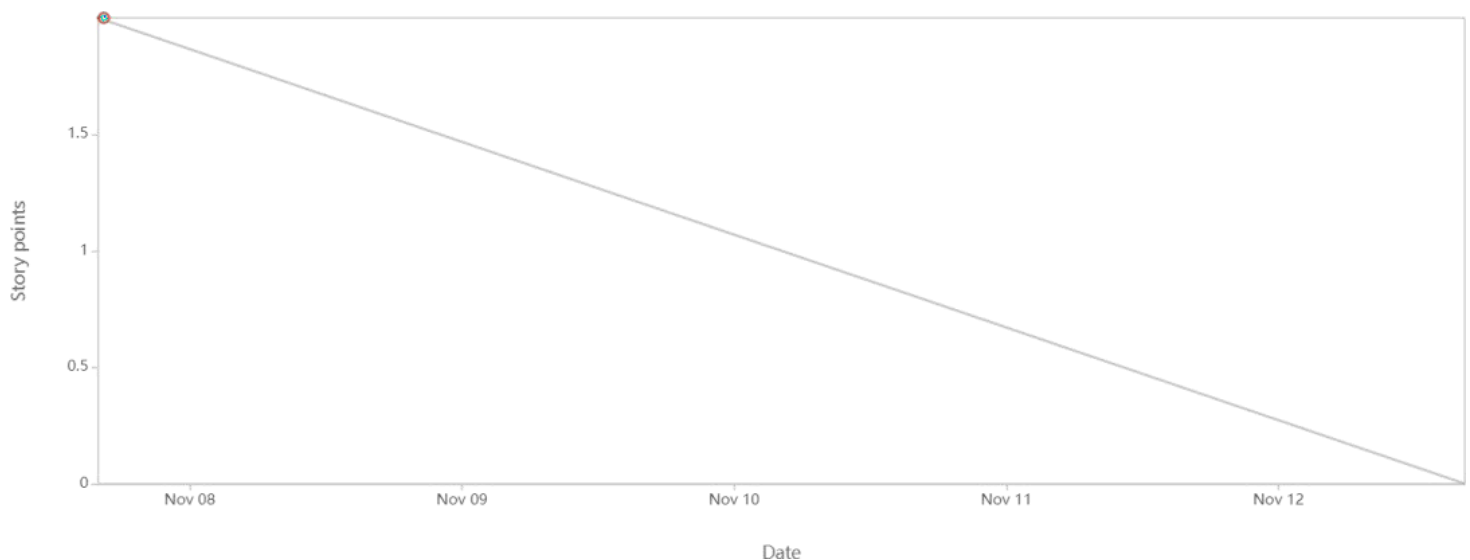
- **Total Points** - 20
- **Duration** - 6 Days
- **Sprint Start date** - 14 Nov 2022
- **Sprint End date (Planned)** - 19 Nov 2022
- **Sprint Points Completed (as on Planned End Date)** - 20
- **Sprints Release Date (Actual)** - 19 Nov 2022

## 6.3 Reports from JIRA

### Burndown Chart:

A burndown chart is a graphical representation of work left to do versus time.

It is often used in agile software development methodologies such as Scrum. However, burn down charts can be applied to any project containing measurable progress over time.



# CHAPTER 7

## CODING & SOLUTIONING

### 7.1 Feature 1

#### User interface feature:

Here the user gets an interface where the user can enter into a detailed form to predict the links that are secured or not, they can copy paste any links to check whether the link is secured or not. It is done through HTML and CSS languages.

#### Code:

```
<!DOCTYPE html>
<html lang="en">
<head>
  <center> <h1> WEB PHISHING DETECTION</h1> </center>
  <meta charset="UTF-8">
  <meta http-equiv="X-UA-Compatible" content="IE=edge">
  <meta name="viewport" content="width=device-width, initial-scale=1.0">
  <meta name="description" content="This website is develop for identify the safety of url.">
  <meta name="keywords" content="phishing url,phishing,cyber security,machine learning,classifer,python">
  <meta name="author" content="Balajee A V">

  <!-- Bootstrap -->
  <link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.5.0/css/bootstrap.min.css"
    integrity="sha384-9aIt2nRpC12Uk9gS9baDI411NQApFmC26EwAOH8WgZl5MYYxFfc+NcPb1dKGj7Sk"
    crossorigin="anonymous">

  <link href="static/style.css" rel="stylesheet">
  <title>URL detection</title>
</head>

<body>
  <center>  </center>

  <div class="container">
    <div class="row">
```



```

<div class="form col-md" id="form1">
  <h2>PHISHING URL DETECTION</h2>

  <br>
  <form action="/" method="post">
    <input type="text" class="form_input" name='url' id="url" placeholder="Enter URL" required="" />
    <label for="url" class="form_label">URL</label>
    <button class="button" role="button">Check here</button>
  </form>

</div>

<div class="col-md" id="form2">

  <br>
  <h6 class="right"><a href= "{{ url }}" target="_blank">{{ url }}</a></h6>

  <br>
  <h3 id="prediction"></h3>
  <button class="button2" id="button2" role="button" onclick="window.open('{{ url }}') target="_blank">Still
want to Continue</button>
  <button class="button1" id="button1" role="button" onclick="window.open('{{ url }}')
target="_blank">Continue</button>
</div>
</div>
<br>
</div>

<!-- JavaScript -->
<script src="https://code.jquery.com/jquery-3.5.1.slim.min.js"
  integrity="sha384-DfXdz2htPH0lsSSs5nCTpuj/zy4C+OGpamoFVy38MVBnE+IbbVYUew+OrCXaRkfj"
  crossorigin="anonymous"></script>
<script src="https://cdn.jsdelivr.net/npm/popper.js@1.16.0/dist/umd/popper.min.js"
  integrity="sha384-Q6E9RHvblyZFJoft+2mJbHaEWldlvI9IOYy5n3zV9zzTtmI3UksdQRVvoxMfooAo"
  crossorigin="anonymous"></script>
<script src="https://stackpath.bootstrapcdn.com/bootstrap/4.5.0/js/bootstrap.min.js"
  integrity="sha384-OgVRvuATP1z7JjHLkuOU7Xw704+h835Lr+6QL9UvYjZE3Ipu6Tp75j7Bh/kR0JKI"
  crossorigin="anonymous"></script>

<script>

  let x = '{{xx}}';
  let num = x*100;
  if (0<=x && x<0.50){
    num = 100-num;
  }

```

```

let txtx = num.toString();
if(x<=1 && x>=0.50){
    var label = "Website is "+txtx +"% safe to use...";
    document.getElementById("prediction").innerHTML = label;
    document.getElementById("button1").style.display="block";
}
else if (0<=x && x<0.50){
    var label = "Website is "+txtx +"% unsafe to use..."
    document.getElementById("prediction").innerHTML = label ;
    document.getElementById("button2").style.display="block";
}

```

```

</script>

```

```

</body>

```

```

</html>

```

## Feature Screenshot:

## User Interface



## 7.2 Feature 2

### Details Acquisition feature

Here the user can enter their details regarding whether the link(url,http) is secured or not and they can get into a conclusion with the help of backend code.

#### Code:

```
import pickle

import warnings

import numpy as np

import pandas as pd

from flask import Flask, render_template, request

from sklearn import metrics

warnings.filterwarnings('ignore')

from feature import FeatureExtraction

file = open("model.pkl", "rb")

gbc = pickle.load(file)

file.close()

app = Flask(__name__)
```

```

@app.route("/", methods=["GET", "POST"])

def index():

    if request.method == "POST":

        url = request.form["url"]

        obj = FeatureExtraction(url)

        x = np.array(obj.getFeaturesList()).reshape(1,30)

        y_pred =gbc.predict(x)[0]

        #1 is safe

        #-1 is unsafe

        y_pro_phishing = gbc.predict_proba(x)[0,0]

        y_pro_non_phishing = gbc.predict_proba(x)[0,1]

        # if(y_pred ==1 ):

        pred = "It is {0:.2f} % safe to go ".format(y_pro_phishing*100)

        return render_template('index.html',xx =round(y_pro_non_phishing,2),url=url )

    return render_template("index.html", xx =-1)


if __name__ == "__main__":

    app.run(debug=True,port=2002)

```

## Feature Screenshot:



### PHISHING URL DETECTION

`https://www.bing.com/`

URL

Check here



### PHISHING URL DETECTION

`http://www.mybonk.com`

URL

Check here

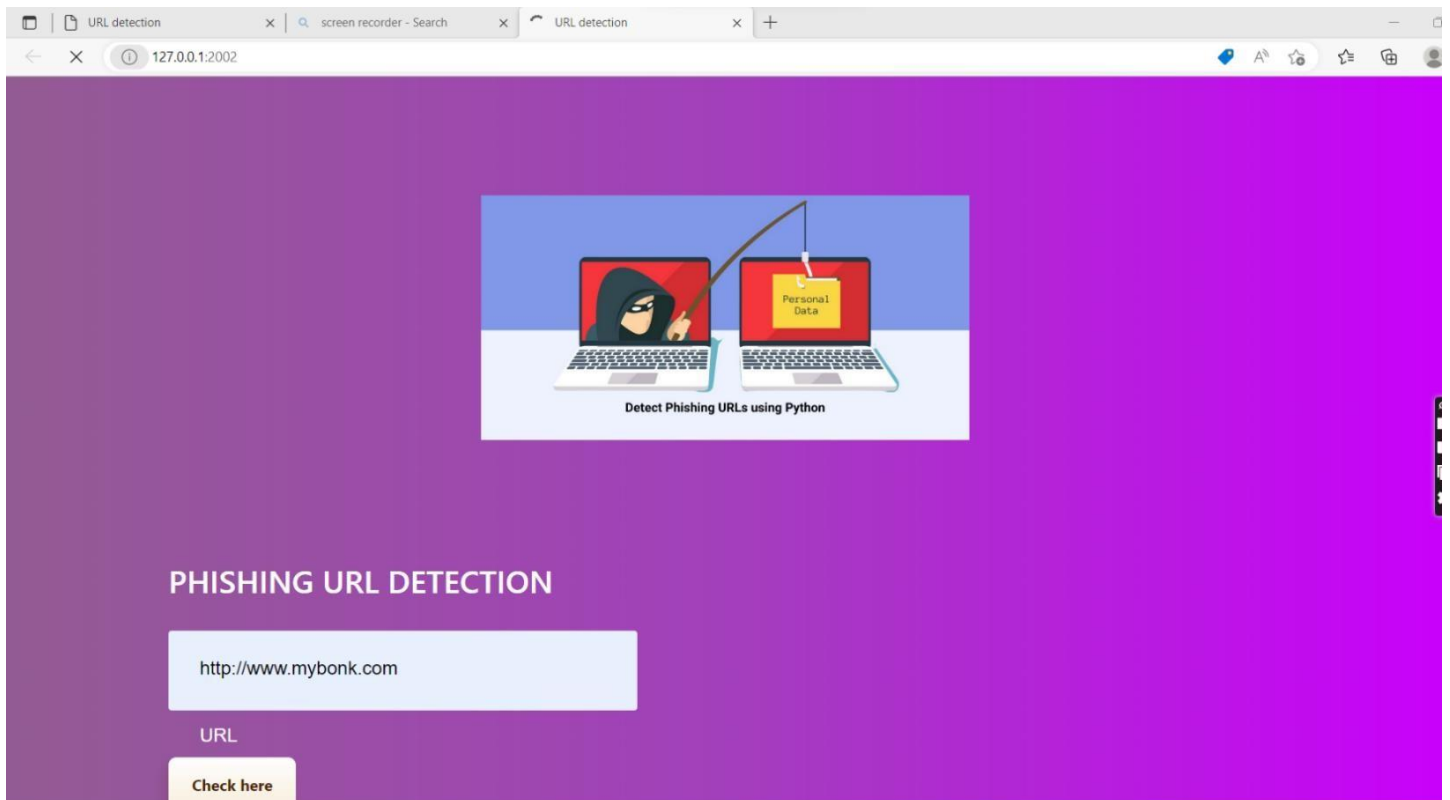
# CHAPTER 8

## TESTING

### 8.1 Test Cases

- We want to test the link or sites whether it is a phishing site or not
- The result is to be predicted by the machine learning algorithm (Cat Boost Classifier gives highest accuracy)
- We want to display the output that the site entered in the checkbox is safe or unsafe.

### 8.2 User Acceptance Testing



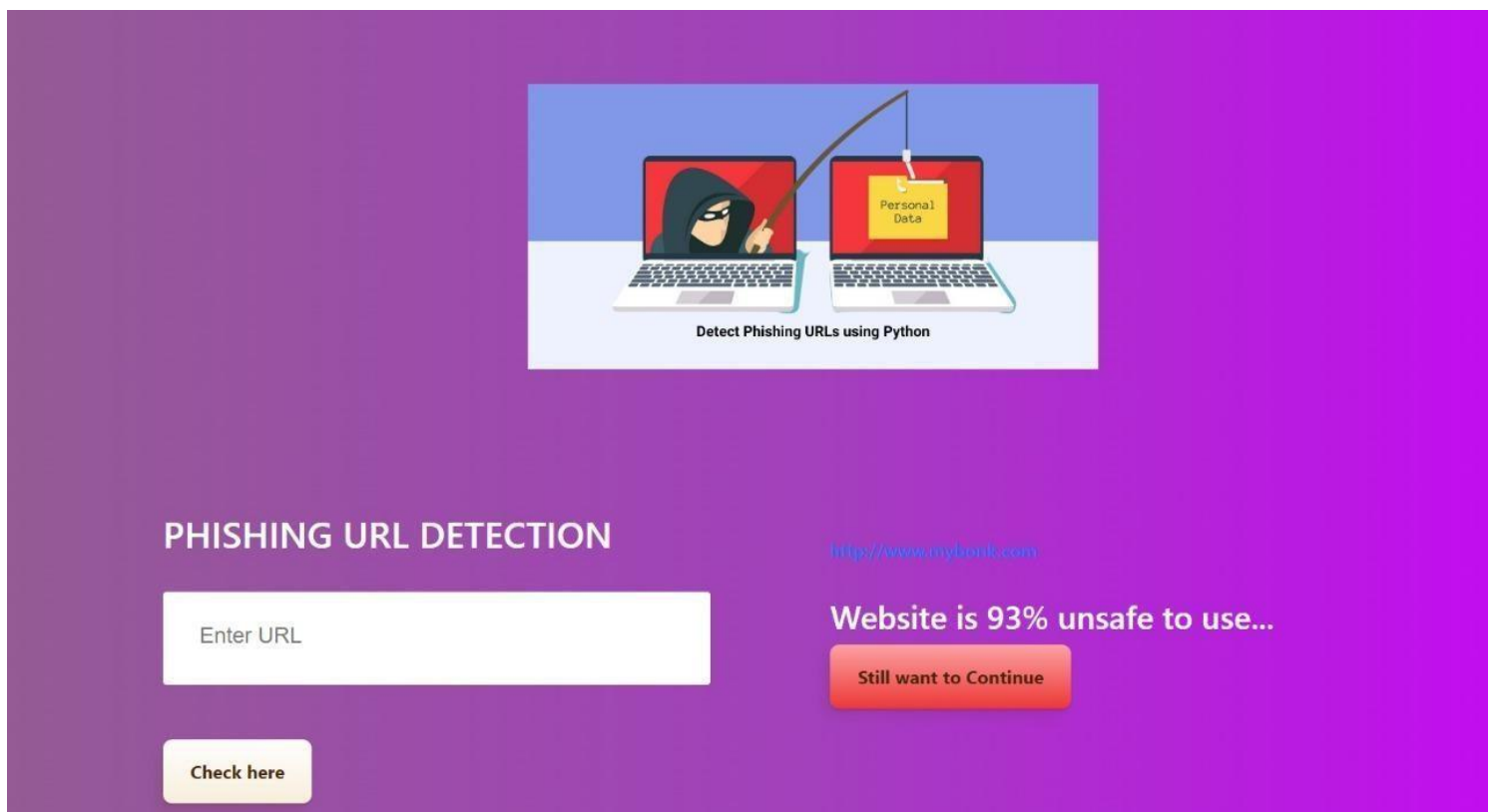
The User can test any website URL to check the whether the website is secured or not

## CHAPTER 9 RESULTS

### 9.1 Performance Metrics

With this user interface users can easily predict the links and website whether it is a fake link or not, based on the process of Machine Learning. Users can also Check the percentage of the link and also it tells that the link is safe or not.

Some sample images of the output are provided below:



The above image tells that the website is unsafe to use.



## PHISHING URL DETECTION

Enter URL

Check here

<https://www.bing.com/>

Website is 99% safe to use...

Continue

The above image tells that the website is safe to use.



## **CHAPTER 10**

### **ADVANTAGES & DISADVANTAGES**

#### **ADVANTAGES:**

- Build a Secure connection between the user's mail transfer agent.
- Build a Secure connection between the mail user agent.
- Provide a clear idea about the effective level of each classifier on phishing detection.
- High level of accuracy.
- Creates a new type of features.
- Fast in Classification.
- It reduces the time consumption.
- Easy To use.

#### **DISADVANTAGES:**

- The Algorithm we used will performs better than other algorithms only when we have categorical data.
- It can perform bad if the variables are not properly tuned.
- The current model only works for predefined data set.

## **CHAPTER 11**

### **CONCLUSION**

Thus, the user is able to easily identify whether the URL is a Phishing site or not in this Web Application, by using the algorithms of Machine learning. So that the user can be aware of those links and their data will be protected.

## **CHAPTER 12**

### **FUTURE SCOPE**

- We can develop this project by using the Adaboost and xgboost Algorithm to increase the accuracy rate, by comparing the other algorithms.
- We can increase the prediction of the datasets
- We can increase the accuracy of the model
- User login and storing the user credentials will be introduced in upcoming modules.
- Real time data can also be predicted based certain API's and packages that can be handled using advanced machine algorithms
- It can also be extended to mobile apps where users can access from remote.

## CHAPTER 13

### APPENDIX

#### Source code:

#### Source code for Flask Integration:

```
import pickle

import warnings

import numpy as np

import pandas as pd

from flask import Flask, render_template, request

from sklearn import metrics

warnings.filterwarnings('ignore')

from feature import FeatureExtraction


file = open("model.pkl", "rb")

gbc = pickle.load(file)

file.close()


app = Flask(__name__)


@app.route("/", methods=["GET", "POST"])

def index():

    if request.method == "POST":
```

```

url = request.form["url"]

obj = FeatureExtraction(url)

x = np.array(obj.getFeaturesList()).reshape(1,30)


y_pred =gbc.predict(x)[0]

#1 is safe

#-1 is unsafe

y_pro_phishing = gbc.predict_proba(x)[0,0]

y_pro_non_phishing = gbc.predict_proba(x)[0,1]

# if(y_pred ==1 ):

pred = "It is {0:.2f} % safe to go ".format(y_pro_phishing*100)

return render_template('index.html',xx =round(y_pro_non_phishing,2),url=url )

return render_template("index.html", xx =-1)


if __name__ == "__main__":

    app.run(debug=True,port=2002)

```

**This Below link is used for re-directing to the web page**

```

PS C:\Users\shashank\Desktop\sandy dupi> & C:/Users/shashank/AppData/Local/Programs/Python/Python310/python.exe "c:/Users/shashank/Desktop/sandy dup
i/app.py"
* Serving Flask app 'app'
* Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on http://127.0.0.1:2002
Press CTRL+C to quit
* Restarting with watchdog (windowsapi)
* Debugger is active!
* Debugger PIN: 822-776-768

```

## Source code for IBM model deployment:

```
from flask import Flask, render_template, request

from feature import FeatureExtraction

import numpy as np

# import joblib

# import json

import requests

app = Flask(__name__)


API_KEY = "ihLwr2uokhHtIZYKQLCE_ADiHWqfvPivU52qNuMMarGA"

token_response = requests.post('https://iam.cloud.ibm.com/identity/token', data={"apikey": API_KEY,
"grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})

mltoken = token_response.json()["access_token"]


header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' + mltoken}


@app.route('/', methods=['GET'])

def hello():

    return render_template('index.html')

@app.route("/", methods=["POST"])

def index():

    url = request.form['url']

    obj = FeatureExtraction(url)

    x = obj.getFeaturesList()
```

```
#print(x)
```

```
#x1 = [obj.getFieldList()]
```

```
payload_scoring={  
    "input_data": [  
        {  
            "fields": [  
                "having_IPhaving_IP_Address",  
                "URLURL_Length",  
                "Shortining_Service",  
                "having_At_Symbol",  
                "double_slash_redirecting",  
                "Prefix_Suffix",  
                "having_Sub_Domain",  
                "SSLfinal_State",  
                "Domain_registration_length",  
                "Favicon",  
                "port",  
                "HTTPS_token",  
                "Request_URL",  
                "URL_of_Anchor",  
                "Links_in_tags",  
                "SFH",  
                "Submitting_to_email",  
                "Abnormal_URL",
```

```

        "Redirect",
        "on_mouseover",
        "RightClick",
        "popUpWidnow",
        "Iframe",
        "age_of_domain",
        "DNSRecord",
        "web_traffic",
        "Page_Rank",
        "Google_Index",
        "Links_pointing_to_page",
        "Statistical_report"
    ],
    "values": [x]
}

]

}

response_scoring = requests.post('https://us-south.ml.cloud.ibm.com/ml/v4/deployments/a3ad9075-6686-4181-8a33-ddf1191ecd8d/predictions?version=2022-11-08', json=payload_scoring, headers={'Authorization': 'Bearer ' + mltoken})

predictions = response_scoring.json()

predict = predictions['predictions'][0]['values'][0][0]

phishing_prob = predictions['predictions'][0]['values'][0][1][0]

if(predict==1):

    url = "https://" + url

```



```
pred = "It is {0:.2f} % safe to go ".format(100-(predictions['predictions'][0]['values'][0][1][0]*100))

x=round(predictions['predictions'][0]['values'][0][1][1],2)

print(x)

return render_template('index.html',xx =x,url=url )

if __name__ == '__main__':

    app.run(port=5500,debug=True)
```

## The generated Api key:

```
API_KEY = "ihLwr2uokhHtIZYKQLCE_ADiHWqfvPivU52qNuMMarGA"
```

## GitHub & Project Demo Link

### GitHub project link:

<https://github.com/IBM-EPBL/IBM-Project-12226-1659442967>

### Project Demo Link:

[https://drive.google.com/drive/folders/1heE81u6VOwO7AaIRkxWWd-7dgIpwV\\_ei?usp=sharing](https://drive.google.com/drive/folders/1heE81u6VOwO7AaIRkxWWd-7dgIpwV_ei?usp=sharing)