WEB PHISHING DETECTION USING MACHINE LEARNING

APPLIED DATA SCIENCE DOMAIN

TEAM ID: PNT2022TMID07498

A PROJECT REPORT

Submitted by

SUBAVARSHINI. S

NAVANEETHA KRISHNAVENI. A

MYTHILI. R

VARSHINI.P

COMPUTER SCIENCE AND ENGINEERING

P. A. COLLEGE OF ENGINEERING AND TECHNOLOGY

(Autonomous) Pollachi, Coimbatore Dt. - 642 002

NOVEMBER 2022

P. A. COLLEGE OF ENGINEERING AND TECHNOLOGY

BONAFIDE CERTIFICATE

Certified that this project report **"WEB PHISHING DETECTION USING MACHINE LEARNING"** is the work of **"SUBAVARSHINI.S (721719104081), NAVANEETHA KRISHNAVENI.A (721719104056), MYTHILI.R (721719104053), VARSHINI.P (721719104089)"** who carried out the project work under our supervision.

SIGNATURE
Dr. D. CHITRA
Professor
HEAD OF THE DEPARTMENT
Computer Science and Engineering
P. A. College of Engineering and Technology

SIGNATURE
FACULTY MENTOR
Dr. M. UMASELVI
Associate Professor
Computer Science and Engineering
P. A. College of Engineering and Technology

SIGNATURE
FACULTY EVALUATOR
Mrs. P. SANGEETHA
Assistant Professor
Computer Science and Engineering
P. A. College of Engineering and Technology

Submitted to the Viva- Voce Examination held on --------------------------------

INTERNAL EXAMINER **EXTERNAL EXAMINER**

PROJECT REPORT FORMAT

WEB PHISHING DETECTION

1.INTRODUCTION

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques.

The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as “blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including: fast-flux, in which proxies are automatically generated to host the web-page; algorithmic generation of new URLs; etc. Major drawback of this method is that, it cannot detect zero-hour phishing attack.

Heuristic based detection which includes characteristics that are found to exist in phishing attacks in reality and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high.

To overcome the drawbacks of blacklist and heuristics based method, many security researchers now focused on machine learning techniques. Machine learning technology consists of a many algorithms which requires past data to make a decision or prediction on future data. Using this technique, algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites including zero- hour phishing websites.

1.1 Project Overview

In emerging technology industry which deeply influence today’s security problems has given a non-ease of mind to some employer and home users. Occurrences that exploit human vulnerabilities have been on the upsurge in recent years.In the dimension of new era there are many security systems being developed to ensure security is given the utmost priority and prevention to be taken from being hacked by those who are involved in cyber-criminal and essential prevention is also taken as high consideration in organization to ensure network security is not being breached. Cyber security employee are currently searching for trustworthy and steady detection techniques for phishing websites detection. Due to wide usage of internet to perform various activities such as online bill payment, banking transaction, online shopping, and, etc. Customer face numerous security threats like cybercrime. There are many cybercrime that are extensively executed for example spam, fraud, cyber terrorisms and phishing. Among this phishing is known as the popular cybercrime today. Phishing has become one amongst the highest 3 most current forms of law-breaking in line with recent reports, and both frequency of events and user susceptible ness has enlarged in recent years, more combination the danger of economic damage.

Phishing is a type of practice done on the Internet where individual data are obtained by illegal approaches.It supply of obtaining sensitive information, as an example, usernames, passwords, and positive identification points of interest, often for malignant reasons, by taking up the looks of an electronic correspondence. Phishing attack will be enforced in varied kind like Email phishing, web site phishing, spear phishing, Whaling, Tab off his guard, Evil twin phishing etc. Phishing is known as webpage violence. Phishing is often done by email spoofing or texting, and it typically guides user to enter points of interest at a fake web site which look and feel the same. It tries to handle the increasing range of phishing got to be met by clients in awareness and alternative efforts to ascertain protection numerous anti-phishing tools. A number of sites have currently created optional instruments for applications, like maps for redirection but clients ought to not utilize similar passwords anywhere on the net. [8] The primary key feature is to allow user to inquire whether visited websites is original or fake. This paper proposes a security tool called as Detecting Phishing Website Using Machine Learning.

1.2 Purpose

Phishing attack is used to steal confidential information of a user. Fraud websites appears like genuine websites with the logo and graphics of genuine website. This project aims to detect fraud or phishing website using machine learning techniques. Phishing website is one of the internet security problems that target the human vulnerabilities rather than software vulnerabilities. It can be described as the process of attracting online users to obtain their sensitive information such as usernames and passwords. The objective of this project is to train machine learning models and deep neural networks on the dataset created to predict phishing websites. Both phishing and legitimate URLs of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measured and compared.

2 LITERATURE SURVEY

1.LONGFETWU ET AL, "EFFECTIVE DEFENSE SCHEMES FOR PHISHING ATTACKS ON MOBILE COMPUTING PLATFORMS," IEEE 2016, PP.6678- 6691.

In This Paper, Author did a Comprehensive Study on the Security Vulnerabilities Caused by Mobile Phishing Attacks, Including the Web Page Phishing Attacks. Author Propose MobiFish, a novel Automated Lightweight AntiPhishing Scheme for Mobile Platforms. MobiFish Verifies the Validity of Web Pages, Applications, and Persistent Accounts by Comparing the Actual identity to the Claimed identity. Author Propose MobiFish, a novel Automated Lightweight AntiPhishing Scheme for Mobile Platforms. MobiFish Verifies the Validity of Web Pages, Applications, and Persistent Accounts by Comparing the Actual identity to the Claimed identity.

2. SURBHI GUPTA ET AL..," A LITERATURE SURVEY ON SOCIAL NETWORKING ATTACKS: PHISHING ATTACK," IN INTERNATIONAL

CONFERENCE ON COMPUTING, COMMUNICATION AND AUTOMATION. (ICCCA2016).

To fool an online user into elicit personal information. The prime objective of this review is to do literature survey on Social Engineering Attack: Phishing Attack and Techniques to Detect Attack. The paper discusses various types of Phishing Attacks such as Tabnapping, spoofing emails, Trojan horse, hacking and how to prevent them. Every organization has security issues that have been of great concern to users, site developers, and specialists, in order to defend the confidential data from this type of social engineering attack.

3.“A PRACTICAL GUIDE TO ANOMALY DETECTION IMPLICATIONS OF MEETING NEW FFIEC MINIMUM EXPECTATIONS FOR LAYERED SECURITY”[ACCESSED: 08 JAN 2015]

Commercial and retail account holders at financial instructions of all sizes are under attack by sophisticated, organized, wellfunded cyber criminals. Anomaly detection solutions are readily available, are deployed quickly and immediately and automatically protect all account holders against all types of fraud attack with minimal Disruption to legitimate online banking activity. Implementing anomaly detection will not only meet FFIEC Expectations, it will decrease the total cost of fraud and will increase customer loyalty and trust

4.SANS INSTITUTE,” PHISHING: AN ANALYSIS OF A GROWING PROBLEM”. 2007. 1417 [ACCESSED: 23 MAY 2017]

This paper gives an in-depth Analysis of Phishing: what it is, the technologies and Security Weaknesses it takes advantage of, the dangers it Poses to end users.In this Analysis Author Explain the Concepts and technology behind Phishing, Show how the threats much more than just a nuisance or passing trend, and discuss how gangs of Criminals are using these Scams to make a great deal of Money.Unfortunately, a growing number of cyber-thieves are using these same systems to manipulate us and Steal our Private information

2.1 EXISTING PROBLEM

The problem with phishing is that attackers constantly look for new and creative ways to fool users into believing their actions involve a legitimate website or email. Phishers have become more skilled at forging websites to appear identical to the expected location, even including logos and graphics in the phishing emails. to make them more convincing. There are dangerous new advanced phishing methods that utilise personal information that is easily available to the public in order to produce plausible and believable attacks that directly target victims. Methods such as social phishing and context aware phishing are perfect examples of attacks utilising the massive amount of public information to increase the effectiveness of their scams. One study shows that victims are 4.5 times more likely to fall for a phishing attempt if it is from a personal contact or personally relates to them. “Phishers have also started to develop a psychology behind their emails that plays off urgency, greed or trust.

Combined with the legitimate look and feel of the spoofed websites, even more cautious and aware users can fall victim to their attacks” These methods all fall within the classification of spear-phishing, where the attacks directly target specific victims with something in common that they can exploit. Spear-phishing requires some information about the victims – their bank, where they work, what sites they've ordered from recently – to produce a targeted attack, and much of this data can easily be found by combing profiles, blogs and other websites.

Some phishing attacks even incorporate malware such as worms or trojans into the emails they send, which then directly compromise the security of the victim's computer and create another tool from which they can select victims and send out attacks. Phishers have also started to develop a psychology behind their emails that plays off urgency, greed or trust. Combined with the legitimate look and feel of the spoofed websites, even more cautious and aware users can fall victim to their attacks.

2.2 REFERENCES

[1] Gunter ollmnn,"the phishing Guide Understanding & Preventing Phishing Attacks", IBMInternet Security Systems, 2007. https://resources.infosecinstitute.com/category/enterprise /phishing/the-phishing-landscape/phishing-data-attackstatistics/#gref

[2] Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013

[3] Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available:

[4] https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016

[5] http://dataaspirant.com/2017/01/30/how-decision-treealgorithm-works/

[6] http://dataaspirant.com/2017/05/22/random-forestalgorithm-machine-learing/

[7] https://www.kdnuggets.com/2016/07/support-vectormachines-simple-explanation.html

[8] www.alexa.com

[9] www.phishtank.com

2.3 PROBLEM STATEMENT DEFINITION

Phishing detection techniques do suffer low detection accuracy and high false alarm especially when novel phishing approaches are introduced. Besides, the most common technique used, blacklist-based method is inefficient in responding to emanating phishing attacks since registering new domain has become easier, no comprehensive blacklist can ensure a perfect up-to-date database.

Furthermore, page content inspection has been used by some strategies to overcome the false negative problems and complement the vulnerabilities of the stale lists. Moreover, page content inspection algorithms each have different approach to phishing website detection with varying degrees of accuracy. Therefore, ensemble can be seen to be a better solution as it can combine the similarity in accuracy and different error-detection rate properties in selected algorithms. Therefore, this study will address a couple of research:

- How to process raw dataset for phishing detection?
- How to increase detection rate in phishing websites algorithms?
- How to reduce false negative rate in phishing websites algorithm?
- What are the best compositions of classifiers that can give a good detection rate of phishing website?

3. IDEATION & PROPOSED SOLUTION

Ideation is the process where you generate ideas and solutions through sessions such as Sketching, Prototyping, Brainstorming, Brainwriting, Worst Possible Idea, and a wealth of other ideation techniques. Ideation is also the third stage in the Design Thinking process.

Steps for ideation process:

Regardless of length, a thorough ideation process is built on the following steps:

- Find your problem.
- Charter to define your problem and objectives.
- Research to find stimulus.
- Utilize ideation methods.
- Objectively screen and score your ideas

3.1 EMPATHY MAP CANVAS

An empathy map is *a* **collaborative tool teams can use to gain a deeper insight into their customers.** Much like a user persona, an empathy map can represent a group of users, such as a customer segment. The empathy map was originally created by Dave Gray and has gained much popularity within the agile community.

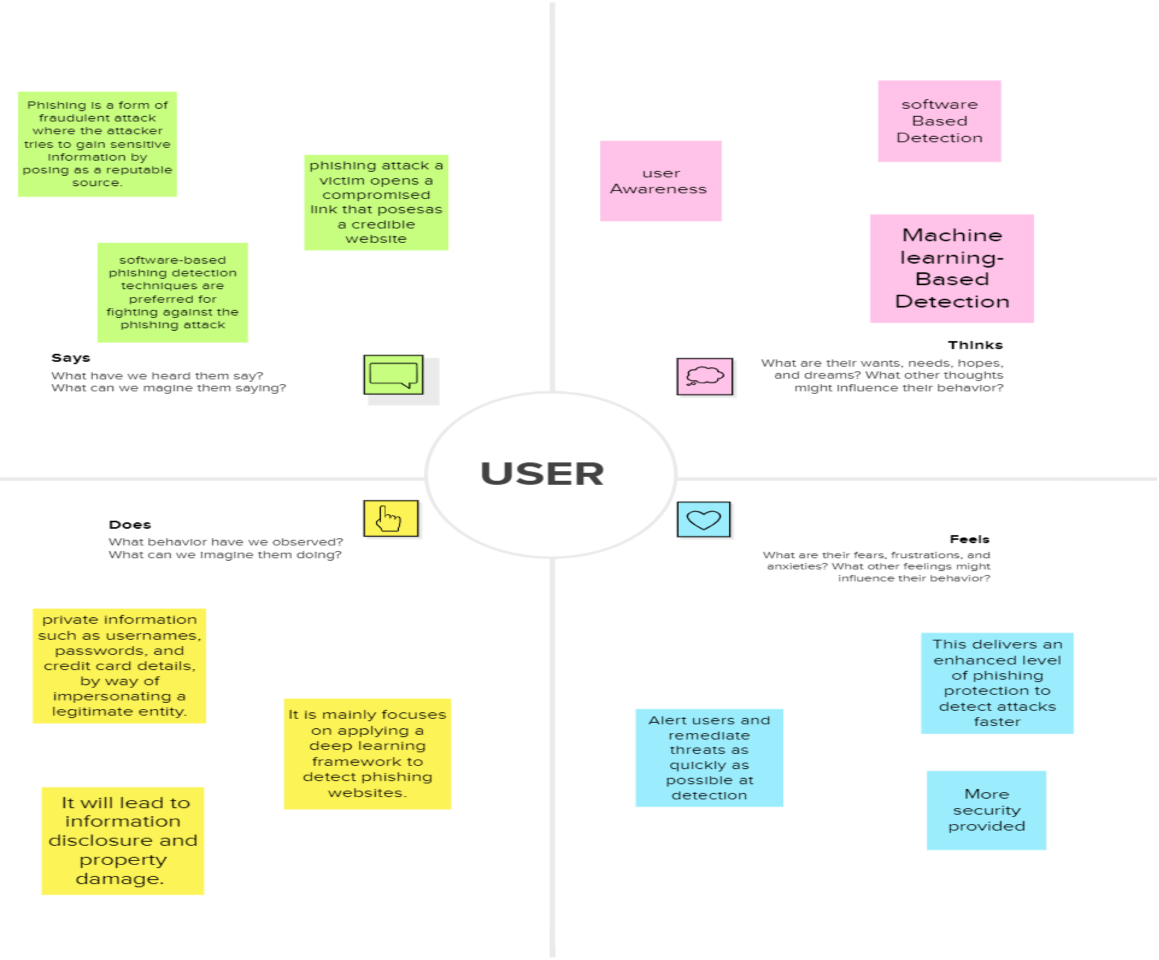

Phishing is a form of fraudulent attack where the attacker tries to gain sensitive information by posing as a reputable source.
phishing attack a victim opens a compromised link that posesas a credible website
software-based phishing detection techniques are preferred for fighting against the phishing attack
Says
What have we heard them say?
What can we magine them saying?
software Based Detection
user Awareness
Machine learning-Based Detection
Thinks
What are their wants, needs, hopes, and dreams? What other thoughts might influence their behavior?
USER
Does
What behavior have we observed?
What can we imagine them doing?
private information such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity.
It is mainly focuses on applying a deep learning framework to detect phishing websites.
It will lead to information disclosure and property damage.
Feels
What are their fears, frustrations, and anxieties? What other feelings might influence their behavior?
This delivers an enhanced level of phishing protection to detect attacks faster
Alert users and remediate threats as quickly as possible at detection
More security provided

Fig 3.1 empathy map canvas

The empathy map **represents a principal user and helps teams better understand their motivations, concerns, and user experience**. Empathy mapping is a simple yet effective workshop that can be conducted with a variety of different users in mind, anywhere from stakeholders, individual use cases, or entire teams of people.

An empathy map canvas **helps brands provide a better experience for users by helping teams understand the perspectives and mindset of their customers**. Using a template to create an empathy map canvas reduces the preparation time and standardizes the process so you create empathy map canvases of similar quality.Empathy is important because it helps us understand how others are feeling so we can respond appropriately to the situation. It is typically associated with social behaviour and there is lots of research showing that greater empathy leads to more helping behaviour.

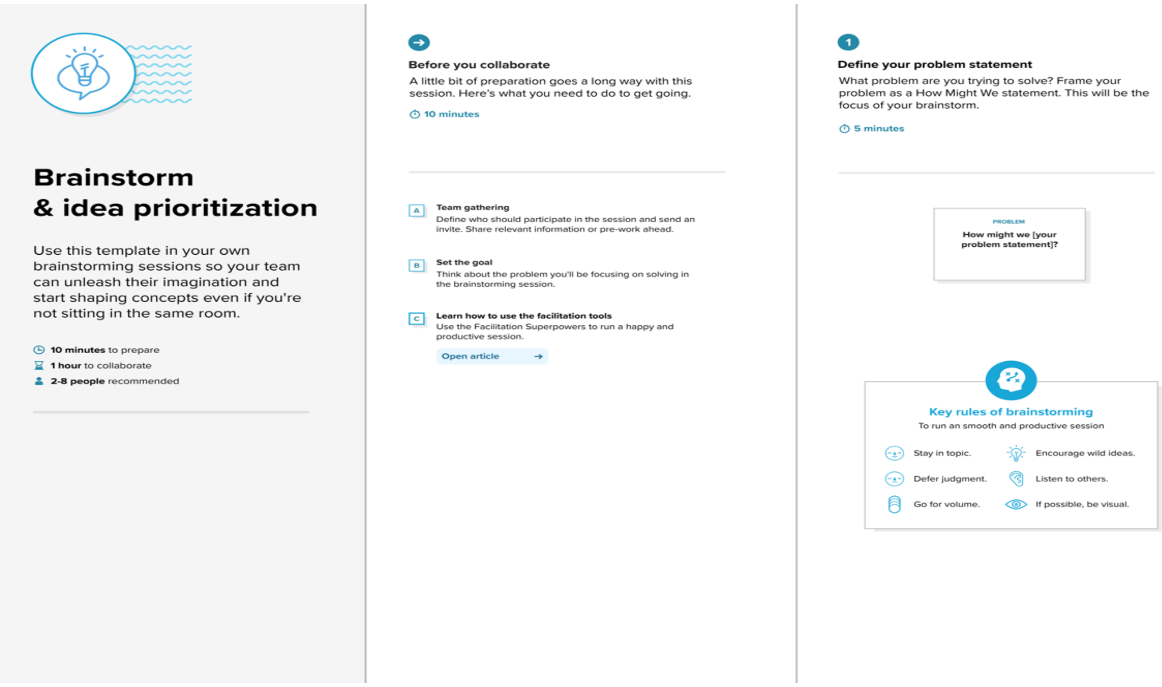
3.2 IDEATION & BRAIN STORMING

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.

Reference: https://www.mural.co/templates/empathy-map-canvas

Step-1: Team Gathering, Collaboration and Select the Problem Statement



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

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

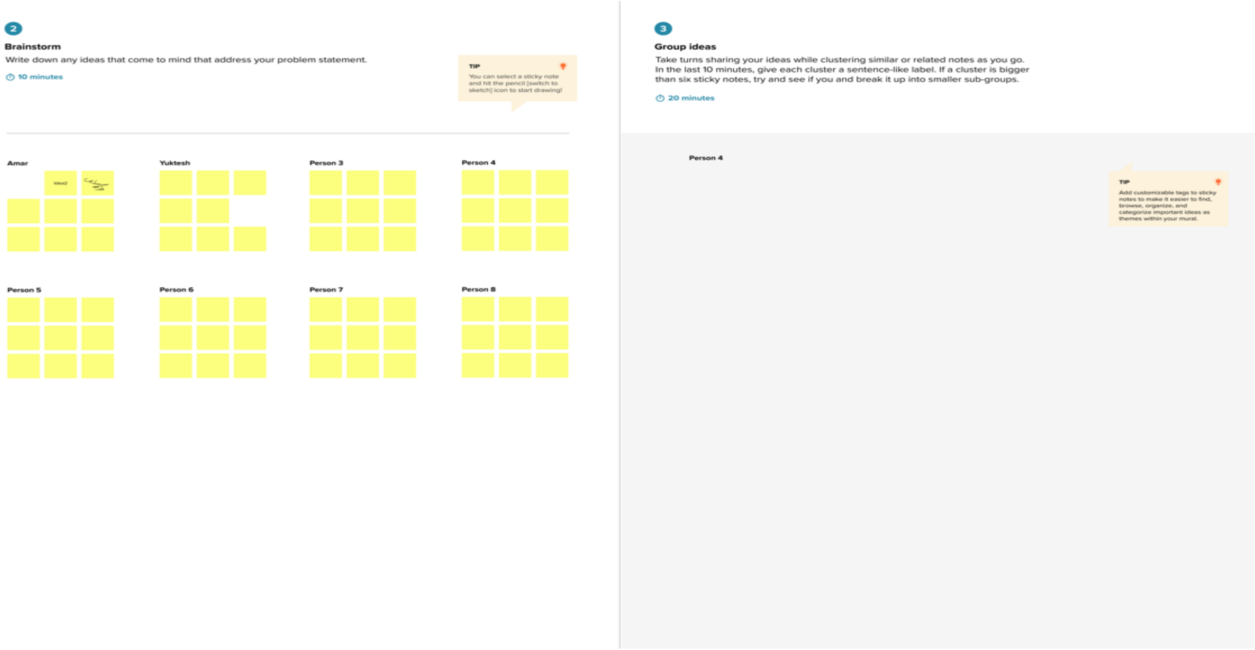
PROBLEM
How might we [your problem statement]?

Key rules of brainstorming
To run an smooth and productive session

Stay in topic.
Encourage wild ideas.
Defer judgment.
Listen to others.
Go for volume.
If possible, be visual.

Fig 3.2.1 ideation & brain storming

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] icon to start drawing!

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 and break it up into smaller sub-groups.

20 minutes

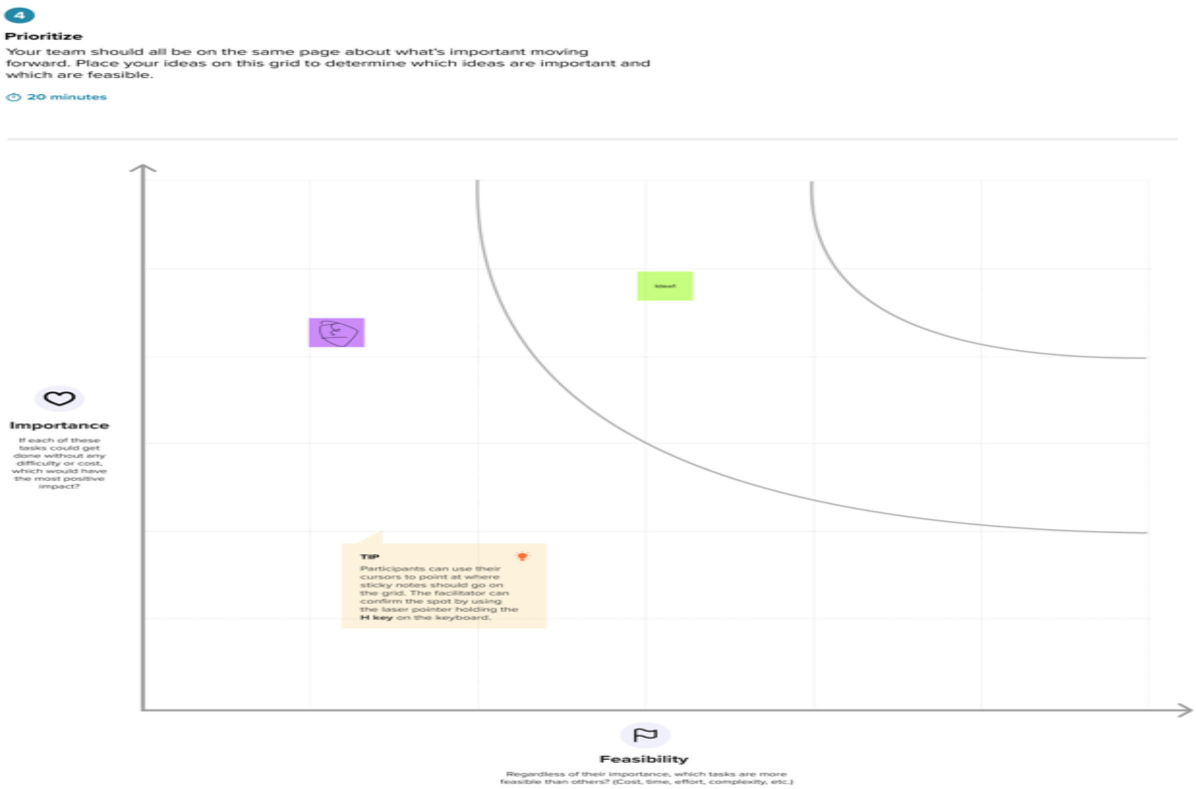
TIP

Add customizable tags to sticky notes to make it easier to find, browse, organize, and categorize important ideas as themes within your mural.

fig 3.2.2

brainstorm, idea listing and grouping

Step-3: Idea Prioritization



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.

If each of these tasks could get done without any difficulty or cost, which would have the most positive impact?

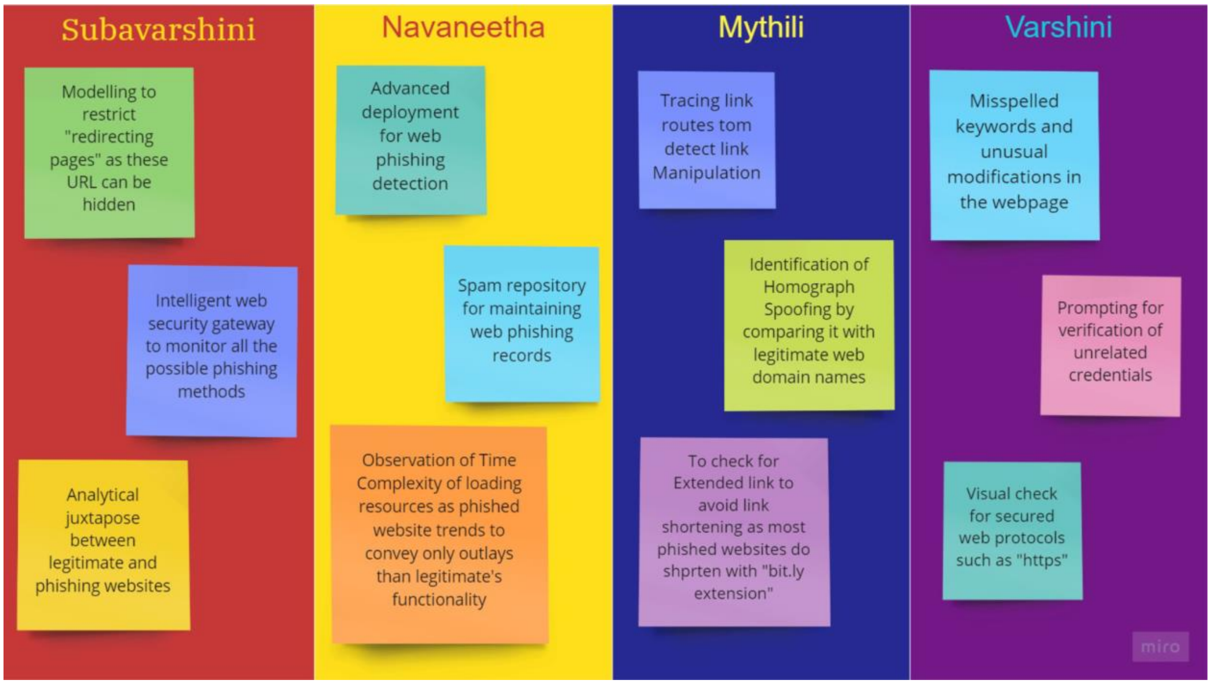
TIP

Participants can use their cursors to point at where sticky notes should go on the grid. The facilitator can confirm the spot by using the laser pointer holding the **H key** on the keyboard.

Regardless of their importance, which tasks are more feasible than others? (Cost, time, effort, complexity, etc.)

Fig 3.2.3 Idea Prioritization

BRAIN STORMING



Subavarshini
Modelling to restrict "redirecting pages" as these URL can be hidden
Intelligent web security gateway to monitor all the possible phishing methods
Analytical juxtapose between legitimate and phishing websites
Navaneetha
Advanced deployment for web phishing detection
Spam repository for maintaining web phishing records
Observation of Time Complexity of loading resources as phished website trends to convey only outlays than legitimate's functionality
Mythili
Tracing link routes tom detect link Manipulation
Identification of Homograph Spoofing by comparing it with legitimate web domain names
To check for Extended link to avoid link shortening as most phished websites do shprten with "bit.ly extension"
Varshini
Misspelled keywords and unusual modifications in the webpage
Prompting for verification of unrelated credentials
Visual check for secured web protocols such as "https"
miro

fig 3.2.4 brain storming

3.3 Proposed solution

Problem statement (problem to be solved)

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.

Idea / solution description

our aim is to detect and predict e-banking phishing websites, we proposed an intelligent, flexible and effective system that is based on using classification algorithms. We implemented classification algorithms and techniques to extract the phishing datasets 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 detection rate. Once a user makes a transaction online when he makes payment through an e-banking website our system will use a data mining algorithm to detect whether the e-banking website is a phishing website or not.

Novelty / uniqueness

Phishing is a form of fraudulent attack where the attacker tries to gain sensitive information by posing as a reputable source. In a typical phishing attack, a victim opens a compromised link that poses as a credible website. The uniqueness of wind energy:

- Detect attacks faster
- Alert users and remediate threats as quickly as possible.

Social impact/customer satisfaction Phishing is one of the top cyber-crimes that impact consumers and businesses all around the world. It is the most common scams on the Internet. Phishing is known as the process in which someone attempts to obtain sensitive information such as usernames, passwords, social security number or financial information and personal information such as birthdates, name and addresses by masking themselves as a trustworthy or familiar entity. With social networking on the rise, people are sharing their personal information everywhere, and have no idea if a website is truly what it seems to be.

Business model (Revenue model)

It includes the following in its Business model

- Anti-Phishing
- Web Scraping
- Spam Filter
- Detecting Fake Websites
- Second Authorization Verification

6 Scalability of the solution

Machine Learning Algorithm: this solution work on prediction, based on known properties learned from the training data set. Spam Filters: this solution is more effective than all the solution we have seen in our study by far because it works on the context of the e-mail and also observes the URL.

Anti-Phishing Plug-in (Browser Extension): In this solution browser capability is extended. Now browses keep the track of users information and generates warning if found something is go wrong.

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.

Achieving problem-solution fit is essential to the success of any new business. Because without it, you're essentially just guessing that your idea is going to work. And if you want to be successful, you need more than just a guess.

Problem-solution fit is a term used to describe the point validating that the base problem resulting in a business idea really exists and the proposed solution actually solves that problem.

Validate that the problem exists:

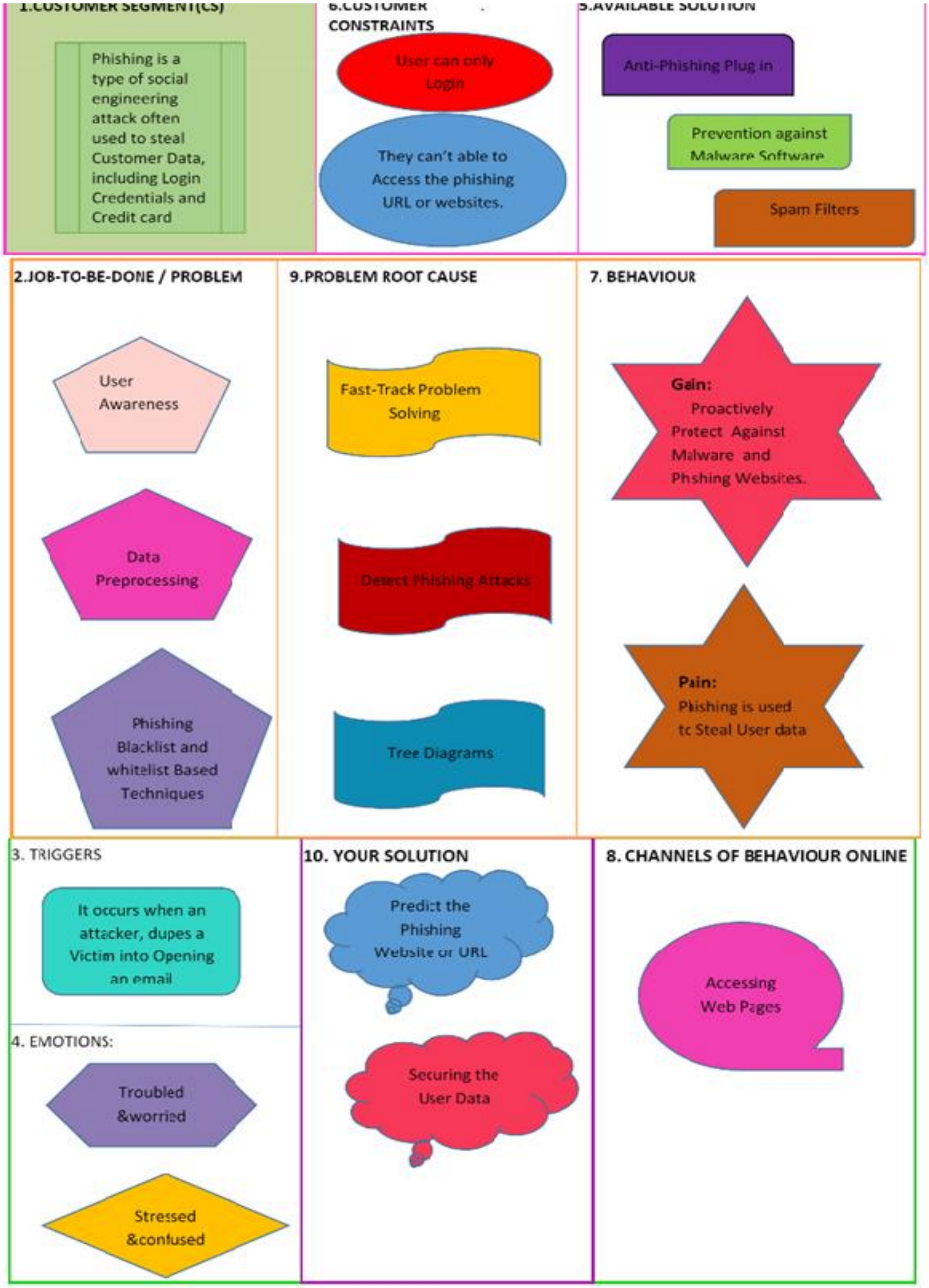
When you validate your problem hypothesis using real-world data and feedback. That is, you gather information from real users to determine whether or not they care about the pain point you’re trying to solve.

Validate that your solution solves the problem:

When you validate that the target audience appreciates the value your solution delivers to them.The problem-solution fit precedes the product development and forms the foundation upon which a company is built. It helps you answer the basics startup-related questions before you even start your startup.

- Do people actually have the problem that you think they have?
- How do they solve the problem now?

Does your proposed solution make a meaningful difference?



1.CUSTOMER SEGMENT(CS)

Phishing is a type of social engineering attack often used to steal Customer Data, including Login Credentials and Credit card

6.CUSTOMER CONSTRAINTS

User can only Login

They can't able to Access the phishing URL or websites.

5.AVAILABLE SOLUTION

Anti-Phishing Plug in

Prevention against Malware Software

Spam Filters

2.JOB-TO-BE-DONE / PROBLEM

User Awareness

Data Preprocessing

Phishing Blacklist and whitelist Based Techniques

9.PROBLEM ROOT CAUSE

Fast-Track Problem Solving

Detect Phishing Attacks

Tree Diagrams

7. BEHAVIOUR

Gain: Proactively Protect Against Malware and Phshing Websites.

Pain: Phishing is used to Steal User data

3. TRIGGERS

It occurs when an attacker, dupes a Victim into Opening an email

4. EMOTIONS:

Troubled &worried

Stressed &confused

10. YOUR SOLUTION

Predict the Phishing Website or URL

Securing the User Data

8. CHANNELS OF BEHAVIOUR ONLINE

Accessing Web Pages

Fig 3.3.1 problem solution fit

4.REQUIREMENT ANALYSIS

Functional requirements are the desired operations of a program, or system as defined in software development and systems engineering. The systems in systems engineering can be either software electronic hardware or combination software-driven electronics.

Two types of Functional Requirements

- functional requirement
- non-functional requirements

4.1 FUNCTIONAL REQUIREMENTS

A function of software system is defined in functional requirement and the behavior of the system is evaluated when presented with specific inputs or conditions which may include calculations, data manipulation and processing and other specific functionality.

- Our system should be able to load air quality data and preprocess data.
- It should be able to analyze the air quality data.
- It should be able to group data based on hidden patterns.
- It should be able to assign a label based on its data groups.
- It should be able to split data into trainset and testset.

FR No.	**Functional Requirement (Epic)**	**Sub Requirement (Story / Sub-Task)**
FR-1	User Registration	Registration through Form Registration through Gmail Registration through Linked IN
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	Change Password	Users have the option to change their password to something more familiar to them.
FR-4	Invite Users	Users must have the ability to invite other people to join the service via email invitation.
FR-5	Block User	Users must have the functionality to actively block any other member of the web application, disabling them from viewing to avoid the phishing.
FR-6	Edit trusted user friend list	Users must be allowed to modify their trusted users, giving them the power to decide who they can share photos with – sending or receiving.

4.2 NON-FUNCTIONAL REQUIREMENTS

Non-functional requirements describe how a system must behave and establish constraints of its functionality. This type of requirements is also known as the system's quality attributes. Attributes such as performance, security, usability, compatibility are not the feature of the system, they are a required characteristic. They are "developing" properties that emerge from the whole arrangement and hence we can't compose a particular line of code to execute them.

NFR No.	Non-Functional Requirement	Description
NFR-2	**Security**	Appropriate user authentication should be provided. Security questions will need to be setup when a user registers as a user. Security questions will need to be answered when a user forgets their password.Separate account types for admin and members so that no member can access the database and only admin can update the database
NFR-3	**Reliability**	There are two main types of users of this system. The member, which is a registered user that has logged in to the system and the admin. The members are assumed to have basic knowledge of the computers and internet browsing. Admin can provide help to the users.
NFR-4	**Performance**	The system should be fast and accurate. System will handle expected and non-expected errors in a manner that will prevent information loss and long downtime period. System should be able to handle large amounts of data. System should accommodate high number of photos and users without any fault
NFR-5	**Availability**	This service will be available on laptops, tablets and mobile devices.
NFR-6	**Scalability**	Admins can create changes to the system, but members or other users cannot make any changes. The quality of the website and database is maintained in such a way that it can be very user friendly to all users. The user should be able to download and install the system very easily

5 PROJECT DESIGN

Project design is an early phase of a project where the project's key features, structure, criteria for success, and major deliverables are planned out. The aim is to develop one or more designs that can be used to achieve the desired project goals. importance of project designs are They help your team understand how to move through a project in the correct way. They help you avoid omitting important steps or items. They help you look more professional. They put the 'know how' in the business, instead of in employees.

5.1 DATA FLOW DIAGRAM

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.

Example: DFD Level 0 (phishing detection)

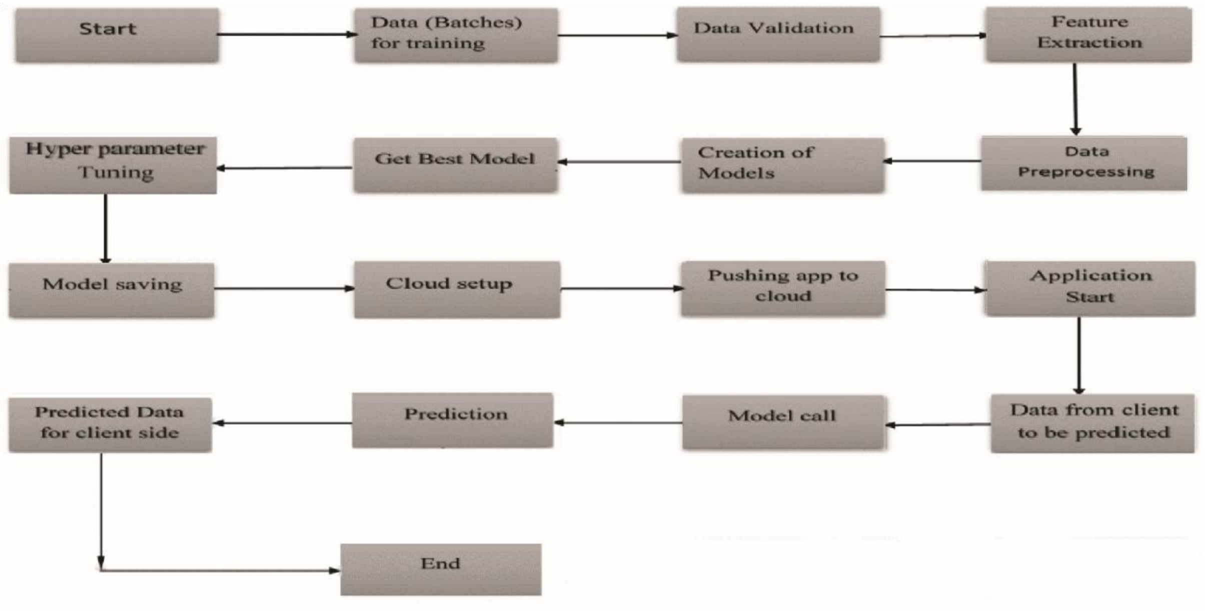


fig:5.1.1 data flow diagram

Example: (Simplified)

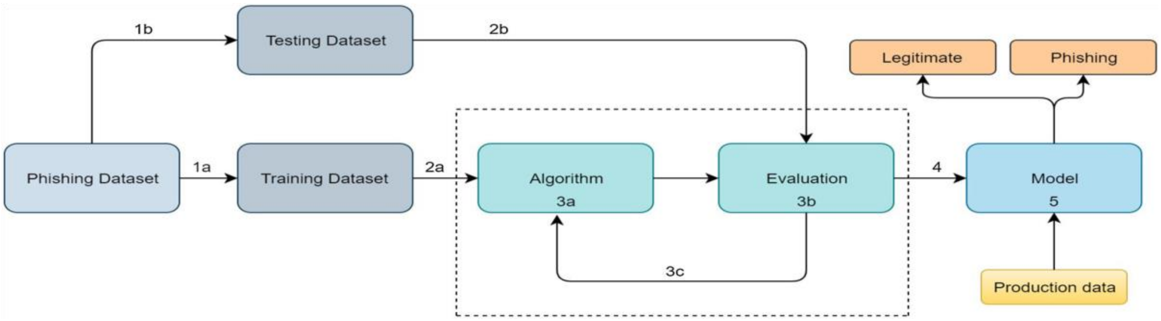


Fig 5.1.2 simplified DTD

- ML approaches are popular for phishing websites detection and it becomes a simple classification problem.
- To train a machine learning model for a learning-based detection system, the data at hand must-have features that are related to phishing and legitimate website classes.
- A batch of input data is given as input for training to the machine learning model to predict the legitimate traffic.
- By reducing features, dataset visualization becomes more efficient and understandable.
- The most significant classifiers that were used in various studies and are found to give good phishing attack detection accuracy are C4.5, k-NN, and SVM.

5.2 SOLUTION & TECHNICAL ARCHITECTURE

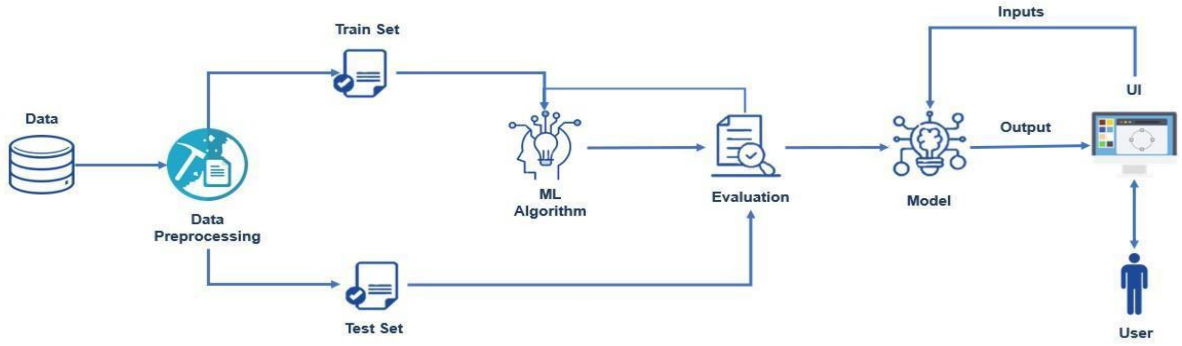


Fig:5.2.1 Technical Architecture

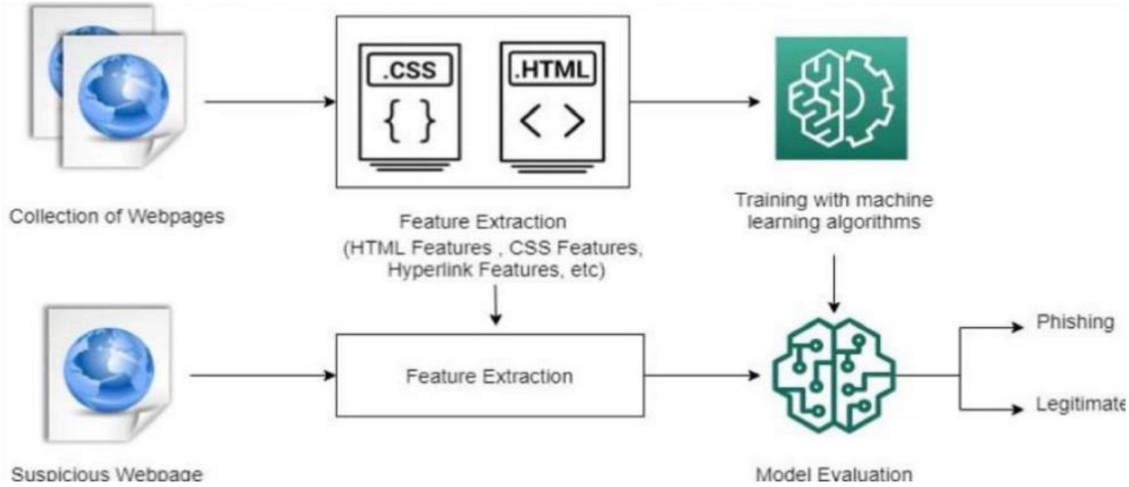


Fig:5.2.2 Solution Architecture

5.3 USER STORIES

Use the below template to list all the user stories for the product.

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	User Registration	USN-1	Registration through Form Registration through Gmail Registration through Linked IN	I can access my account / dashboard	High	Sprint-1
	User Confirmation	USN-2	Confirmation via Email Confirmation via OTP	As a conditions that a software product must meet to be accepted by a user, a customer, or other systems.	High	Sprint-2
	Change Password	USN-3	Users have the option to change their password to something more familiar to them.	I can register & access the dashboard user cannot submit a form without completing all the mandatory fields	Medium	Sprint-2
Admin (Web User)	Invite Users	USN-4	Users must have the ability to invite other people to join the service via email invitation.	As a end-user testing, is a phase of software development in which the software is tested in the real world	High	Sprint-3

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
				by its intended audience.		
	Block User	USN-5	Users must have the functionality to actively block any other member of the web application, disabling them from viewing to avoid the phishing.	1. Tap Block & Report . If you use a Google Workspace account and the person is in your organization, this option is Block.To confirm, tap Block.	High	Sprint-1
	Edit trusted user friend list	USN-6	Users must be allowed to modify their trusted users, giving them the power to decide who they can share photos with – sending or receiving.	To make sure that when *user* sends friend requests and another *user accept* it the *friend list* will change for both *users*.	High	Sprint-3

6.PROJECT PLANNING & SCHEDULING

In Scrum Projects, Estimation is done by the entire team during Sprint Planning Meeting. The objective of the Estimation would be to consider the User Stories for the Sprint by Priority and by the Ability of the team to deliver during the Time Box of the Sprint. During sprint planning, we break the stories down into tasks, estimate those tasks, and compare the task estimates against our capacity. It's that, not points, that keep us from overcommitting in this sprint. No need to change the estimate.

How to run a sprint planning

- Examine team availability.
- Establish velocity for your team.
- Plan your sprint planning meeting.
- Start with the big picture.
- Present new updates, feedback, and issue.
- Confirm team velocity and capacity.
- Go over backlog items.
- Determine task ownership.

6.1 SPRINT PLANNING & ESTIMATION

Product Backlog, Sprint Schedule, and Estimation

Sprint	Functional Requirement (EPIC)	User Story Number	User Story/ Task	Story Points	Priority	Team Members
Sprint 1	Registration	USN-1	As a user ,I can register for the application by entering email, password and conforming my password.	5	High	JEEVITHA V SANGEETHA A NARMATHA T RAJASRI K
Sprint 1		USN-2	As a user, I will receive confirmation email once I have registered for the application	4	High	JEEVITHA V SANGEETHA A NARMATHA T RAJASRI K
Sprint 1		USN-3	As a user ,I can register for application through phone number	4	High	JEEVITHA V SANGEETHA A NARMATHA T RAJASRI K
Sprint 1		USN-4	As a user, I can register for the application through Gmail	3	Medim	JEEVITHA V SANGEETHA A NARMATHA T RAJASRI K
Sprint 1	Login(User)	USN-5	As a user, I can log into application by entering email& password	5	High	JEEVITHA V SANGEETHA A NARMATHA T RAJASRI K
Sprint 2	Dashboard	USN-6	Once I have logged in , I can see my dashboard.	6	Medim	JEEVITHA V SANGEETHA A NARMATHA T RAJASRI K
Sprint 2	Web access	USN-7	As a customer I can access the website to predict phishing Websites.	7	High	JEEVITHA V SANGEETHA A NARMATHA T RAJASRI K
Sprint 2	Prediction	USN-8	As a customer when I enter the website details the website should predict	7	High	JEEVITHA V SANGEETHA A NARMATHA T RAJASRI K

			the the website is phishing or not.			
Sprint 3	Analysis	USN-9	As a customer, I wish to store my prediction and make analysis.	10	Medium	JEEVITHA V SANGEETHA A NARMATHA T RAJASRI K
Sprint 3	Security	USN-10	As a customer I expect my data to be secured.	10	Medim	JEEVITHA V SANGEETHA A NARMATHA T RAJASRI K
Sprint 4	Database Access	USN-11	An administrator I should maintain the website. And update the website regularly.	20	Low	JEEVITHA V SANGEETHA A NARMATHA T RAJASRI K

6.2 SPRINT DELIVERY SCHEDULE

The outcome of the sprint is a deliverable, albeit with some increments. The scrum is used for projects like Web Technology or development of a product for the new market, i.e. the product with many requirements or fast-changing requirement. When does Sprint Planning take place? Sprint planning occurs on the first day of a new sprint. The event should occur after the sprint review and retrospective from the previous sprint so that any output from those discussions can be considered when planning for the new sprint. The deliverables of a sprint aren't as predictable as they are for other projects. Sprint participants have produced sketches and drawings, writing, photographs, comic strips, videos and fully coded working prototypes. The answer is whatever's right to answer the problem.

Sprint	Total story points	Duration	Sprint start date	Sprint end date (Planned)	Story points completed (as on planned End date)	Sprint Release date(Actual)
Sprint 1	20	6 days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint 2	20	6 days	31 Oct 2022	5 Nov 2022	20	5 Nov 2022
Sprint 3	20	6 days	7 Nov 2022	12 Nov 2022	20	12 Nov 2022

Sprint 4	20	6 days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

7. CODING & SOLUTIONING

Coding Solutions is a highly competitive job accelerator and talent refinement program that recruits and transitions college graduates with past programming experience or technical degrees into professional careers with Alabama companies and organizations at no cost to the graduates.

7.1 FEATURE 1

```
from flask import Flask, request, render_template
import numpy as np
import rfc
import pandas as pd
from sklearn import metrics
import requests
import json
import warnings
import pickle
warnings.filterwarnings('ignore')
from feature import FeatureExtraction
API_KEY = "ITkxxANDG6N3roYajqDgCD4qZOdsELITcTYseAecmJKu"
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}
file = open("pickle/model.pkl","rb")
rfc = 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 =rfc.predict(x)[0]
 #1 is safe
 #-1 is unsafe
 y_pro_phishing = rfc.predict_proba(x)[0,0]
 y_pro_non_phishing = rfc.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)
 payload_scoring = {"input_data": [{"field": ["PrefixSuffix-",
                              "SubDomains",
                              "HTTPS",
                              "AnchorURL",
                              "WebsiteTraffic"], "values": [1, -1]}]}
 response_scoring = requests.post('https://us-south.ml.cloud.ibm.com/ml/v4/deployments/f44a67e9-
 ebef-4ab8-abf0-91b5e1307e64/predictions?version=2022-11-13', json=payload_scoring,
  headers={'Authorization': 'Bearer ' + mltoken})
 print("Scoring response")
 print(response_scoring.json())


 if __name__ == "__main__":
    app.run(debug=True)
```

7.2 FEATURE 2

```
import ipaddress
import re
import socket
import time
import urllib.request
from datetime import date, datetime
from urllib.parse import urlparse
import domain
import requests
import response
import whois
from bs4 import BeautifulSoup
from dateutil.parser import parse as date_parse
from googlesearch import search
class FeatureExtraction:
features = []
def __init__(self,url):
self.features = []
self.url = url
self.domain = ""
self.whois_response = ""
self.urlparse = ""
self.response = ""
self.soup = ""
url=""
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:
return 1

# 9.DomainRegLen
def DomainRegLen(self):
try:
expiration_date = self.whois_response.expiration_date
creation_date = self.whois_response.creation_date
try:
if(len(expiration_date)):
expiration_date = expiration_date[0]
except:
pass
try:
if(len(creation_date)):
creation_date = creation_date[0]
except:
pass

age = (expiration_date.year-creation_date.year)*12+ (expiration_date.month-creation_date.month)
if age >=12:
return 1
return -1
except:
return -1

# 10. Favicon
def Favicon(self):
try:
for head in self.soup.find_all('head'):
```

```
for head.link in self.soup.find_all('link', href=True):
dots = [x.start(0) for x in re.finditer('\.', head.link['href'])]
if self.url in head.link['href'] or len(dots) == 1 or domain in head.link['href']:
return 1
return -1
except:
return -1

# 11. NonStdPort
def NonStdPort(self):
try:
port = self.domain.split(":")
if len(port)>1:
return -1
return 1
except:
return -1

# 12. HTTPSDomainURL
def HTTPSDomainURL(self):
try:
if 'https' in self.domain:
return -1
return 1
except:
return -1

# 13. RequestURL
def RequestURL(self):
try:
for img in self.soup.find_all('img', src=True):
dots = [x.start(0) for x in re.finditer('\.', img['src'])]
if self.url in img['src'] or self.domain in img['src'] or len(dots) == 1:
success = success + 1
i = i+1

for audio in self.soup.find_all('audio', src=True):
dots = [x.start(0) for x in re.finditer('\.', audio['src'])]
if self.url in audio['src'] or self.domain in audio['src'] or len(dots) == 1:
```

```
success = success + 1
i = i+1

for embed in self.soup.find_all('embed', src=True):
dots = [x.start(0) for x in re.finditer('\.', embed['src'])]
if self.url in embed['src'] or self.domain in embed['src'] or len(dots) == 1:
success = success + 1
i = i+1

for iframe in self.soup.find_all('iframe', src=True):
dots = [x.start(0) for x in re.finditer('\.', iframe['src'])]
if self.url in iframe['src'] or self.domain in iframe['src'] or len(dots) == 1:
success = success + 1
i = i+1

try:
percentage = success/float(i) * 100
if percentage < 22.0:
return 1
elif((percentage >= 22.0) and (percentage < 61.0)):
return 0
else:
return -1
except:
return 0
except:
return -1

# 14. AnchorURL
def AnchorURL(self):
try:
i,unsafe = 0,0
for a in self.soup.find_all('a', href=True):
if "#" in a['href'] or "javascript" in a['href'].lower() or "mailto" in a['href'].lower() or not (
urllib.request.urlopen in a['href'] or self.domain in a['href']):
unsafe = unsafe + 1
i = i + 1

try:
```

```
percentage = unsafe / float(i) * 100
if percentage < 31.0:
return 1
elif ((percentage >= 31.0) and (percentage < 67.0)):
return 0
else:
return -1
except:
return -1

except:
return -1

# 15. LinksInScriptTags
def LinksInScriptTags(self):
try:
i,success = 0,0

for link in self.soup.find_all('link', href=True):
dots = [x.start(0) for x in re.finditer('\.', link['href'])]
if self.url in link['href'] or self.domain in link['href'] or len(dots) == 1:
success = success + 1
i = i+1

for script in self.soup.find_all('script', src=True):
dots = [x.start(0) for x in re.finditer('\.', script['src'])]
if self.url in script['src'] or self.domain in script['src'] or len(dots) == 1:
success = success + 1
i = i+1
try:
percentage = success / float(i) * 100
if percentage < 17.0:
return 1
elif((percentage >= 17.0) and (percentage < 81.0)):
return 0
else:
return -1
except:
return 0
```

```
except:
return -1

# 16. ServerFormHandler
def ServerFormHandler(self):
try:
if len(self.soup.find_all('form', action=True))==0:
return 1
else :
for form in self.soup.find_all('form', action=True):
if form['action'] == "" or form['action'] == "about:blank":
return -1
elif self.url not in form['action'] and self.domain not in form['action']:
return 0
else:
return 1
except:
return -1

# 17. InfoEmail
def InfoEmail(self):
try:
if re.findall(r"[mail\(\)|mailto:?]", self.soap):
return -1
else:
return 1
except:
return -1

# 18. AbnormalURL
def AbnormalURL(self):
try:
if self.response.text == self.whois_response:
return 1
else:
return -1
except:
return -1
```

```
# 19. WebsiteForwarding
def WebsiteForwarding(self):
try:
if len(self.response.history) <= 1:
return 1
elif len(self.response.history) <= 4:
return 0
else:
return -1
except:
return -1

# 20. StatusBarCust
def StatusBarCust(self):
try:
if re.findall("<script>.+onmouseover.+</script>", self.response.text):
return 1
else:
return -1
except:
return -1

# 21. DisableRightClick
def DisableRightClick(self):
try:
if re.findall(r"event.button ?== ?2", self.response.text):
return 1
else:
return -1
except:
return -1

# 22. UsingPopupWindow
def UsingPopupWindow(self):
try:
if re.findall(r"alert\(", self.response.text):
return 1
else:
return -1
```

```
except:
return -1

# 23. IframeRedirection
def IframeRedirection(self):
try:
if re.findall(r"[<iframe>|<frameBorder>]", self.response.text):
return 1
else:
return -1
except:
return -1

# 24. AgeofDomain
def AgeofDomain(self):
try:
creation_date = self.whois_response.creation_date
try:
if(len(creation_date)):
creation_date = creation_date[0]
except:
pass
today  = date.today()
age = (today.year-creation_date.year)*12+(today.month-creation_date.month)
if age >=6:
return 1
return -1
except:
return -1

# 25. DNSRecording
def DNSRecording(self):
try:
creation_date = self.whois_response.creation_date
try:
if(len(creation_date)):
creation_date = creation_date[0]
except:
pass
```

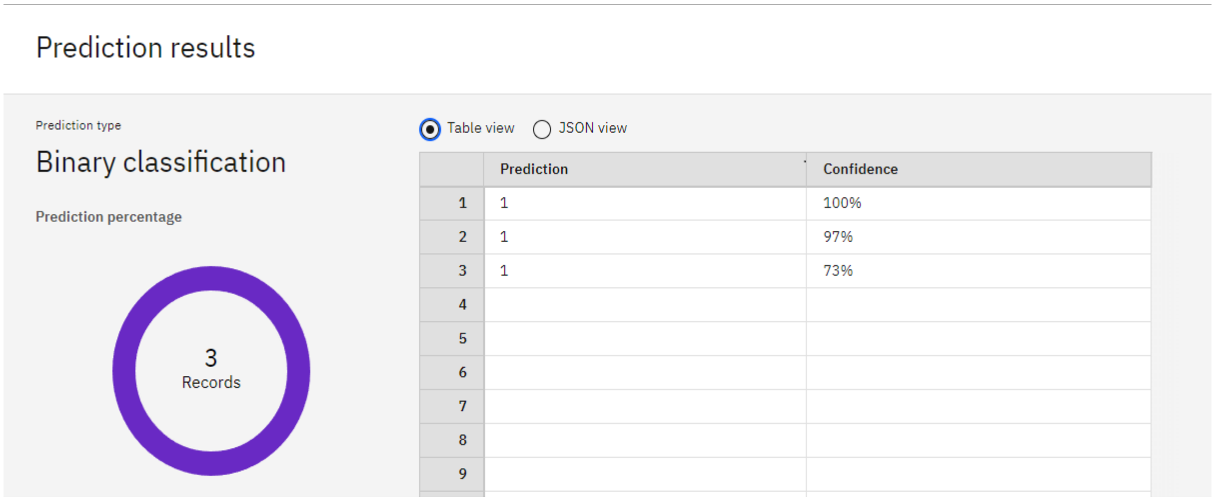
```
today  = date.today()
age = (today.year-creation_date.year)*12+(today.month-creation_date.month)
if age >=6:
return 1
return -1
except:
return -1

# 26. WebsiteTraffic
def WebsiteTraffic(self):
try:
rank = BeautifulSoup(urllib.request.urlopen("http://data.alexa.com/data?cli=10&dat=s&url=" + urllib.request.urlopen).read(), "xml").find("REACH")['RANK']
if (int(rank) < 100000):
return 1
return 0
except :
return -1

# 27. PageRank
def PageRank(self):
try:
prank_checker_response = requests.post("https://www.checkpagerank.net/index.php", {"name": self.domain})
global_rank = int(re.findall(r"Global Rank: ([0-9]+)", prank_checker_response.text)[0])
if global_rank > 0 and global_rank < 100000:
return 1
return -1
except:
return -1

# 28. GoogleIndex
def GoogleIndex(self):
try:
site = search(self.url, 5)
if site:
return 1
else:
```

```
return -1
except:
return 1

# 29. LinksPointingToPage
def LinksPointingToPage(self):
try:
number_of_links = len(re.findall(r"<a href=", self.response.text))
if number_of_links == 0:
return 1
elif number_of_links <= 2:
return 0
else:
return -1
except:
return -1

# 30. StatsReport
def StatsReport(self):
try:
url_match = re.search(
'at\.ua|usa\.cc|baltazarpresentes\.com\.br|pe\.hu|esy\.es|hol\.es|sweddy\.com|myjino\.ru|96\.lt|ow\.ly', urllib.request.urlopen)
ip_address = socket.gethostbyname(self.domain)
ip_match=re.search('146\.112\.61\.108|213\.174\.157\.151|121\.50\.168\.88|192\.185\.217\.116|78\.46\.211\.158|181\.174\.165\.13|46\.242\.145\.103|121\.50\.168\.40|83\.125\.22\.219|46\.242\.145\.98|'
        '107\.151\.148\.44|107\.151\.148\.107|64\.70\.19\.203|199\.184\.144\.27|107\.151\.148\.108|107\.151\.148\.109|119\.28\.52\.61|54\.83\.43\.69|52\.69\.166\.231|216\.58\.192\.225|'
        '118\.184\.25\.86|67\.208\.74\.71|23\.253\.126\.58|104\.239\.157\.210|175\.126\.123\.219|141\.8\.224\.221|10\.10\.10\.10|43\.229\.108\.32|103\.232\.215\.140|69\.172\.201\.153|'
        '216\.218\.185\.162|54\.225\.104\.146|103\.243\.24\.98|199\.59\.243\.120|31\.170\.160\.61|213\.19\.128\.77|62\.113\.226\.131|208\.100\.26\.234|195\.16\.127\.102|195\.16\.127\.157|'
        '34\.196\.13\.28|103\.224\.212\.222|172\.217\.4\.225|54\.72\.9\.51|192\.64\.147\.141|198\.200\.56\.183|23\.253\.164\.103|52\.48\.191\.26|52\.214\.197\.72|87\.98\.255\.18|209\.99\.17\.27|'
        '216\.38\.62\.18|104\.130\.124\.96|47\.89\.58\.141|78\.46\.211\.158|54\.86\.225\.156|54\.82\.156\.19|37\.157\.192\.102|204\.11\.56\.48|110\.34\.231\.42', ip_address)
if url_match:
return -1
elif ip_match:
return -1
return 1
except:
return 1
def getFeaturesList(self):
return self.features
```

8.TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionalityof components, sub assemblies, assemblies and/or a finished product it is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

8.1 TEST CASES



Prediction results

Prediction type

Binary classification

Prediction percentage

3
Records

● Table view ○ JSON view

	Prediction	Confidence
1	1	100%
2	1	97%
3	1	73%
4		
5		
6		
7		
8		
9		

fig 8.1.1 binary classification for table view

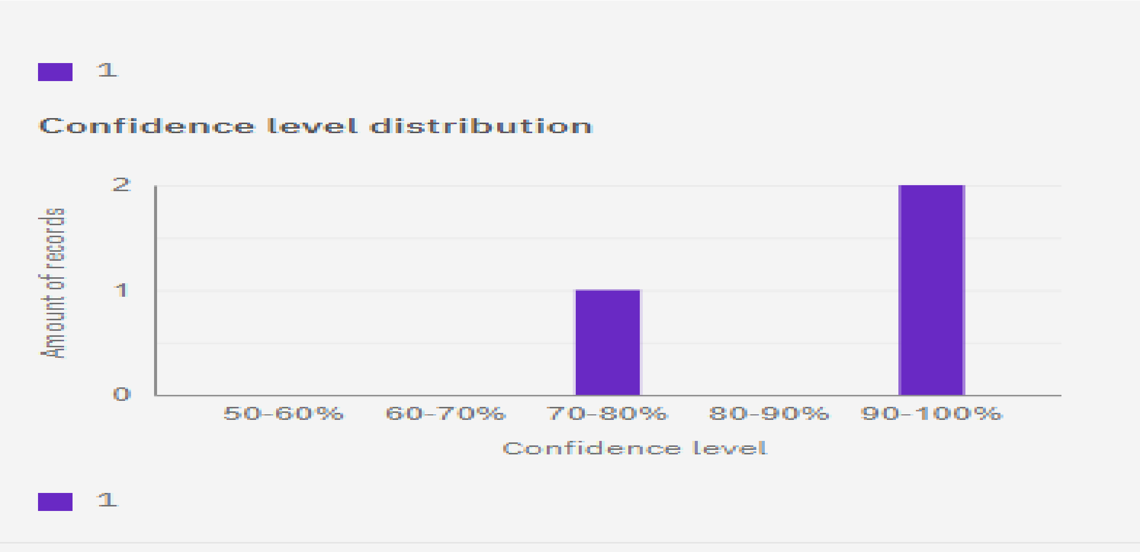
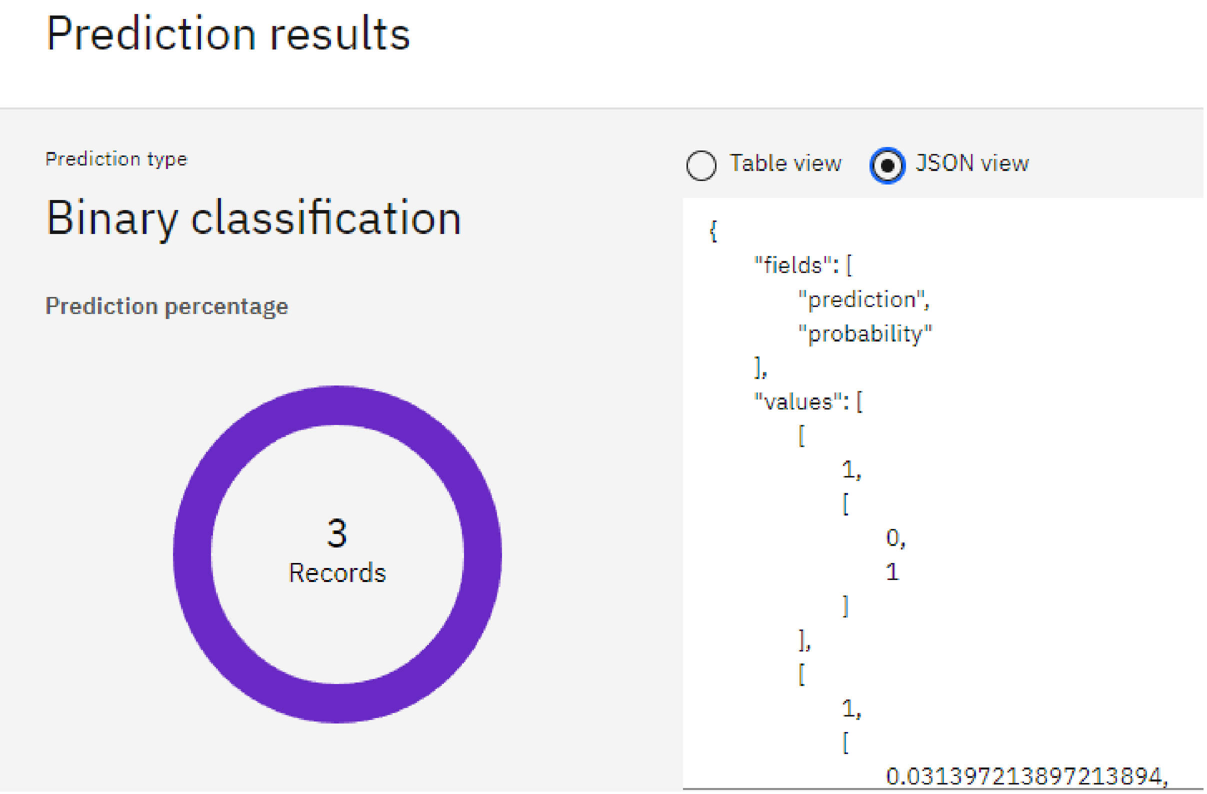


fig 8.1.2 confidence level distribution for table view



Prediction results

Prediction type

Binary classification

Prediction percentage

○ Table view ● JSON view

```
{
    "fields": [
        "prediction",
        "probability"
    ],
    "values": [
        [
            1,
            [
                0,
                1
            ]
        ],
        [
            1,
            [
                0.031397213897213894,
```

fig 8.1.3 binary classification for JSON view

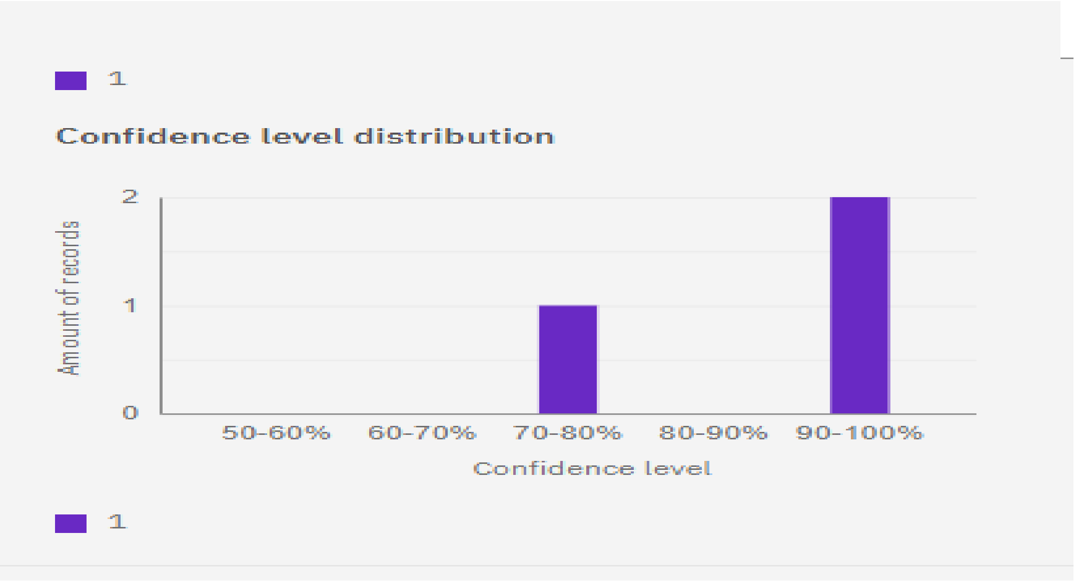


fig 8.1.4 confidence level distribution for JSON view

8.2 USER ACCEPTANCE TESTING

8.2.1 UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application,and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputsand expected results.

8.2.2 INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

8.2.3 VALIDATION TESTING

An engineering validation test (EVT) is performed on first engineering prototypes, to ensure that the basic unit performs to design goals and specifications. It is important in

Identifying design problems, and solving them as early in the design cycle as possible, is the key to keeping projects on time and within budget. Too often, product design and performance problems are not detected until late in the product development cycle when the product is ready to be shipped. The old adage holds true: It costs a penny to make a change in engineering, a dime in production and a dollar after a product is in the field.Verification is a Quality control process that is used to evaluate whether or not a product,service, or system complies with regulations, specifications, or conditions imposed at the start of a development phase. Verification can be in development, scale-up, or production.This is often an internal process.Validation is a Quality assurance process of establishing evidence that provides a high degree of assurance that a product, service, or system accomplishes its intended requirements.

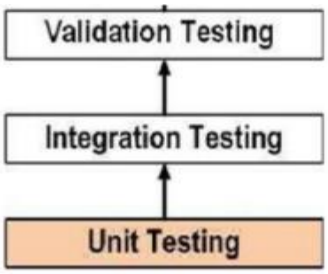


fig 8.2.3.1 validation testing

8.2.4 SYSTEM TESTING

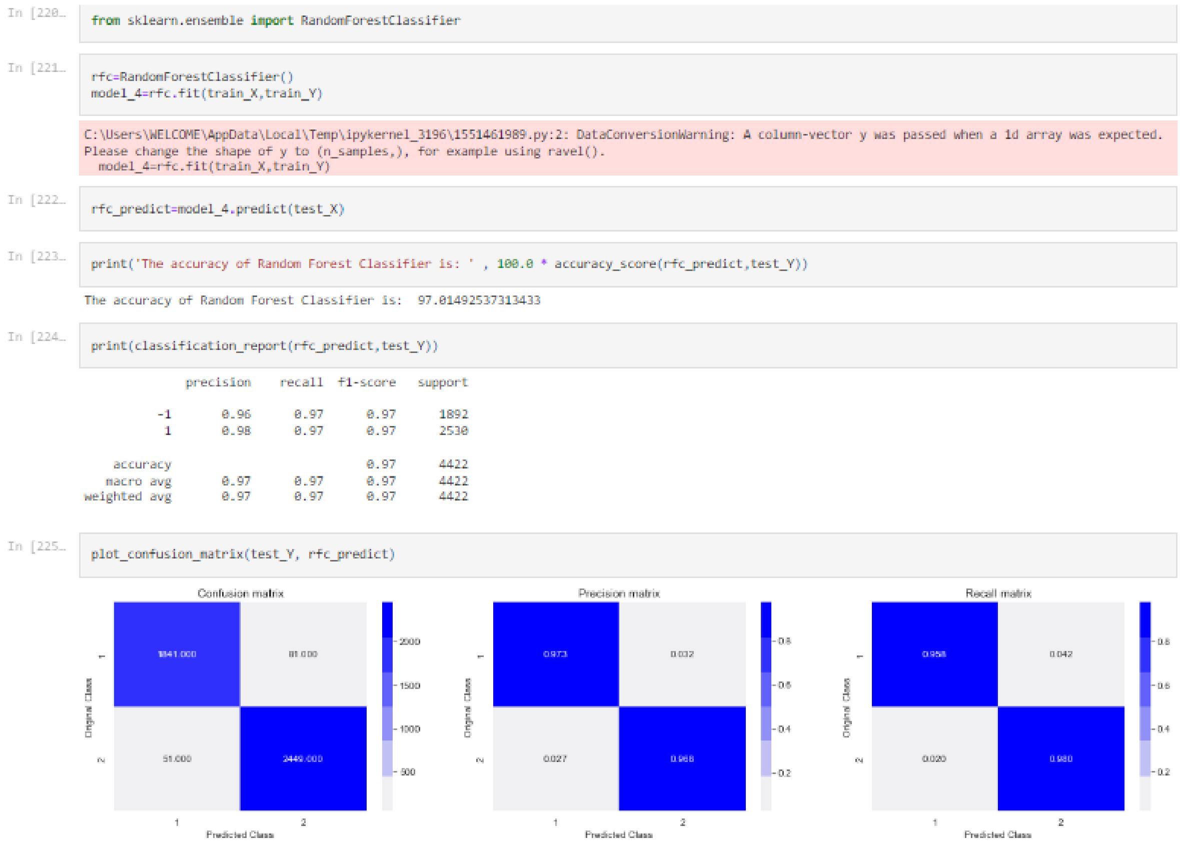
System testing of software or hardware is testing conducted on a complete, integrated system to evaluate the system's compliance with its specified requirements. System testing falls within the scope of black box testing, and as such, should require no knowledge of the inner design of the code or logic.As a rule, system testing takes, as its input, all of the "integrated" software components that have successfully passed integration testing and also the software system itself integrated with any applicable hardware system(s).System testing is a more limited type of testing; it seeks to detect defects both within the "inter-assemblages" and also within the system as a whole.System testing is performed on the entire system in the context of a Functional Requirement Specification(s) (FRS) and/or a System Requirement Specification (SRS).System testing tests not only the design, but also the behavior and even the believed expectations of the customer. It is also intended to test up to and beyond the bounds defined in the software/hardware requirements specification(s).

9.RESULTS

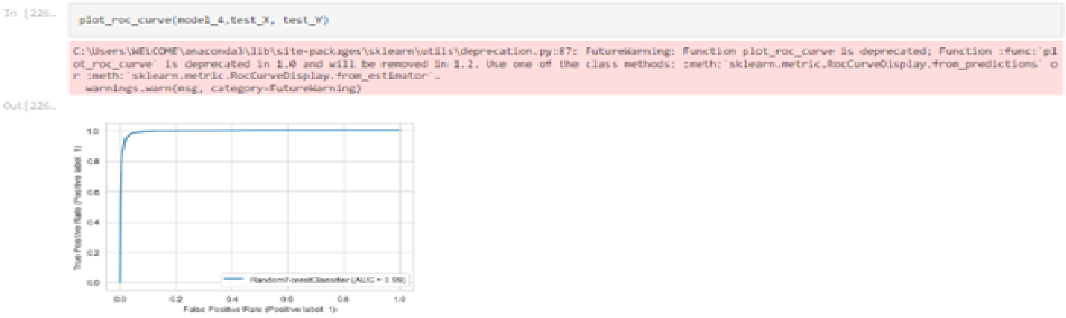
9.1 PERFORMANCE METRICS

- accuracy_score
- classification_report
- plot_confusion_matrix
- plot_roc_curve

fig 9.1.1



```
In [220… from sklearn.ensemble import RandomForestClassifier

In [221… rfc=RandomForestClassifier()
         model_4=rfc.fit(train_X,train_Y)

C:\Users\WELCOME\AppData\Local\Temp\ipykernel_3196\1551461989.py:2: DataConversionWarning: A column-vector y was passed when a 1d array was expected.
Please change the shape of y to (n_samples,), for example using ravel().
  model_4=rfc.fit(train_X,train_Y)

In [222… rfc_predict=model_4.predict(test_X)

In [223… print('The accuracy of Random Forest Classifier is: ' , 100.0 * accuracy_score(rfc_predict,test_Y))

The accuracy of Random Forest Classifier is:  97.01492537313433

In [224… print(classification_report(rfc_predict,test_Y))

              precision    recall  f1-score   support

          -1       0.96      0.97      0.97      1892
           1       0.98      0.97      0.97      2530

    accuracy                           0.97      4422
   macro avg       0.97      0.97      0.97      4422
weighted avg       0.97      0.97      0.97      4422

In [225… plot_confusion_matrix(test_Y, rfc_predict)
```



```
In [226… plot_roc_curve(model_4,test_X, test_Y)

C:\Users\WELCOME\anaconda3\lib\site-packages\sklearn\utils\deprecation.py:87: FutureWarning: Function plot_roc_curve is deprecated; Function :func:`pl
ot_roc_curve` is deprecated in 1.0 and will be removed in 1.2. Use one of the class methods: :meth:`sklearn.metric.RocCurveDisplay.from_predictions` o
r :meth:`sklearn.metric.RocCurveDisplay.from_estimator`.
  warnings.warn(msg, category=FutureWarning)

Out[226…
```

1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
RandomForestClassifier (AUC = 0.99)

performance metrics

10. ADVANTAGES & DISADVANTAGES

ADVANTAGES

- This system can be used by many E-commerce or other 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.
- A mailbox-level anti-phishing solution offers an additional layer of protection by analyzing account information and understanding users' communication habits.
- This delivers an enhanced level of phishing protection to detect attacks faster, alert users and remediate threats as quickly as possible.
- Alert users and remediate threats as quickly as possible at detection
- private information such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity.
- It will lead to information disclosure and property damage.

DISADVANTAGES

- If Internet connection fails, this system won’t work.
- All websites related data will be stored in one place.

The problem with phishing is that attackers constantly look for new and creative ways to fool users into believing their actions involve a legitimate website or email. Phishers have become more skilled at forging websites to appear identical to the expected location, even including logos and graphics in the p

11.CONCLUSION

This paper aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 97.14% detection accuracy using random forest algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data. In future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used. we get very good performance in classifiers namely, Random Forest computation duration and accuracy. The main idea behind ensemble algorithms is to combine several weak learners into a stronger one, this is perhaps the primary reason why machine learning is used in practice for most of the classification problems

12.FUTURE SCOPE

As we have already a real time implementation in our project the scope for future work for our project would be creating a GUI or web extension which would help our user if he accesses any phishing websites by any chance. The efficient of our product can be increased drastically provided your given access to the current fishing website data collection. As cyber- crime is a very prominent in our generation. The scope of future work for this project is perennial.

13.APPENDIX

13.1 Source code

Style.css

```
*,
*::after,
*::before {
  margin: 0;
  padding: 0;
  box-sizing: inherit;
  font-size: 62,5%;
}
body {
  padding: 10% 5%;
  background: rgb(25, 169, 185);
  justify-content: center;
  align-items: center;
  height: 100vh;
  color: rgb(81, 12, 12);
}
form__label {
  font-family: 'Roboto', sans-serif;
  font-size: 1.2rem;
  margin-left: 2rem;
  margin-top: 0.7rem;
  display: block;
  transition: all 0.3s;
  transform: translateY(0rem);
}
.form__input {
  top: -24px;
  font-family: 'Roboto', sans-serif;
  color: #333;
  font-size: 1.2rem;
  padding: 1.5rem 2rem;
  border-radius: 0.2rem;
  background-color: rgb(123, 140, 164);
  border: none;
  width: 75%;
  display: block;
  border-bottom: 0.3rem solid transparent;
  transition: all 0.3s;
}

.form__input:placeholder-shown + .form__label {
  opacity: 0;
  visibility: hidden;
  -webkit-transform: translateY(+4rem);
  transform: translateY(+4rem);
}



.button {
  appearance: button;
  background-color: transparent;
  background-image: linear-gradient(to bottom, rgb(255, 255, 255), #e4dbf8);
  border: 0 solid #e6eaeb;
  border-radius: .5rem;
  box-sizing: border-box;
  color: #482307;
  column-gap: 1rem;
  cursor: pointer;
  display: flex;
  font-family: ui-sans-serif,system-ui,-apple-system,system-ui,"Segoe UI",Roboto,"Helvetica Neue",Arial,"Noto Sans",sans-serif,"Apple Color Emoji","Segoe UI Emoji","Segoe UI Symbol","Noto Color Emoji";
  font-size: 100%;
  font-weight: 700;
  line-height: 24px;
  margin: 0;
  outline: 2px solid transparent;
  padding: 1rem 1.5rem;
  text-align: center;
  text-transform: none;
  transition: all .1s cubic-bezier(.4, 0, .2, 1);
  user-select: none;
  -webkit-user-select: none;
  touch-action: manipulation;
  box-shadow: -6px 8px 10px rgba(20, 170, 25, 0.1),0px 2px 2px rgba(9, 187, 33, 0.2);
```

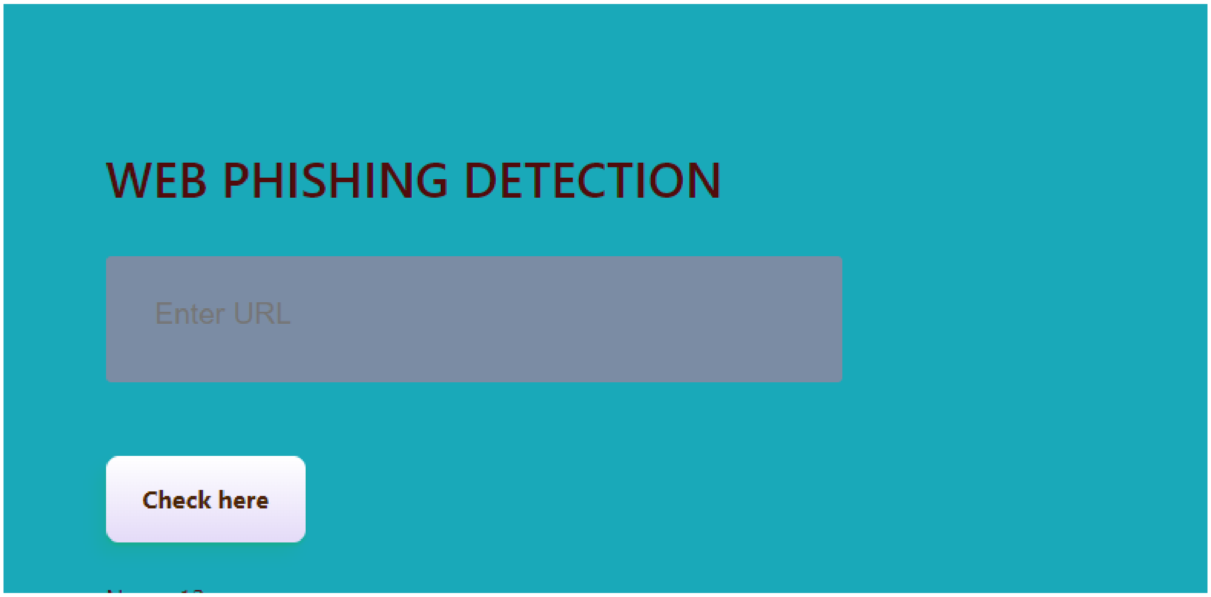
```
}

.button:active {
  background-color: #27aca9;
  box-shadow: -1px 2px 5px rgba(81,41,10,0.15),0px 1px 1px rgba(81,41,10,0.15);
  transform: translateY(0.125rem);
}

.button:focus {
  box-shadow: rgba(72, 35, 7, .46) 0 0 0 4px, -6px 8px 10px rgba(81,41,10,0.1), 0px 2px 2px
rgba(81,41,10,0.2);
}



.main-body{
  display: flex;
  flex-direction: row;
  width: 75%;
  justify-content:space-around;
}

.button1{
  appearance: button;
  background-color: transparent;
  background-image: linear-gradient(to bottom, rgb(160, 245, 174), #37ee65);
  border: 0 solid #e5e7eb;
  border-radius: .5rem;
  box-sizing: border-box;
  color: #482307;
  column-gap: 1rem;
  cursor: pointer;
  display: flex;
  font-family: ui-sans-serif,system-ui,-apple-system,system-ui,"Segoe UI",Roboto,"Helvetica
Neue",Arial,"Noto Sans",sans-serif,"Apple Color Emoji","Segoe UI Emoji","Segoe UI
Symbol","Noto Color Emoji";
  font-size: 100%;
  font-weight: 700;
  line-height: 24px;
  margin: 0;
  outline: 2px solid transparent;
  padding: 1rem 1.5rem;
  text-align: center;
  text-transform: none;
  transition: all .1s cubic-bezier(.4, 0, .2, 1);
  user-select: none;
  -webkit-user-select: none;
  touch-action: manipulation;
  box-shadow: -6px 8px 10px rgba(81,41,10,0.1),0px 2px 2px rgba(81,41,10,0.2);
  display: none;
}

.button2{
  appearance: button;
  background-color: transparent;
  background-image: linear-gradient(to bottom, rgb(252, 162, 162), #f51707);
  border: 0 solid #e5e7eb;
  border-radius: .5rem;
  box-sizing: border-box;
  color: #482307;
  column-gap: 1rem;
  cursor: pointer;
  display: flex;
  font-family: ui-sans-serif,system-ui,-apple-system,system-ui,"Segoe UI",Roboto,"Helvetica
Neue",Arial,"Noto Sans",sans-serif,"Apple Color Emoji","Segoe UI Emoji","Segoe UI
Symbol","Noto Color Emoji";
  font-size: 100%;
  font-weight: 700;
  line-height: 24px;
  margin: 0;
  outline: 2px solid transparent;
  padding: 1rem 1.5rem;
  text-align: center;
  text-transform: none;
  transition: all .1s cubic-bezier(.4, 0, .2, 1);
  user-select: none;
  -webkit-user-select: none;
  touch-action: manipulation;
  box-shadow: -6px 8px 10px rgba(81,41,10,0.1),0px 2px 2px rgba(81,41,10,0.2);
  display: none;
}

.right {
  right: 0px;
  width: 300px;
}

@media (max-width: 576px) {
  .form {
    width: 100%;
  }
 }
.abc{
  width: 50%;
}
```

INDEX.HTML

```
<!DOCTYPE html>
<html lang="en">
<head>
    <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,classifier,python">
    <meta name="author" content="Narma12">


    <!-- BootStrap -->
    <link rel="stylesheet"
href="https://stackpath.bootstrapcdn.com/bootstrap/4.5.0/css/bootstrap.min.css"
        integrity="sha384-
9aIt2nRpC12Uk9gS9baDl411NQApFmC26EwAOH8WgZl5MYYxFfc+NcPb1dKGj7Sk"
crossorigin="anonymous">
    <link href="static/style.css" rel="stylesheet">
    <title>Web Phishing detection</title>
</head>
<body>
<div class=" container">
    <div class="row">
        <div class="form col-md" id="form1">
            <h2>WEB PHISHING 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>
<p>Narma12</p>
</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-
Q6E9RHvbIyZFJoft+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>
```

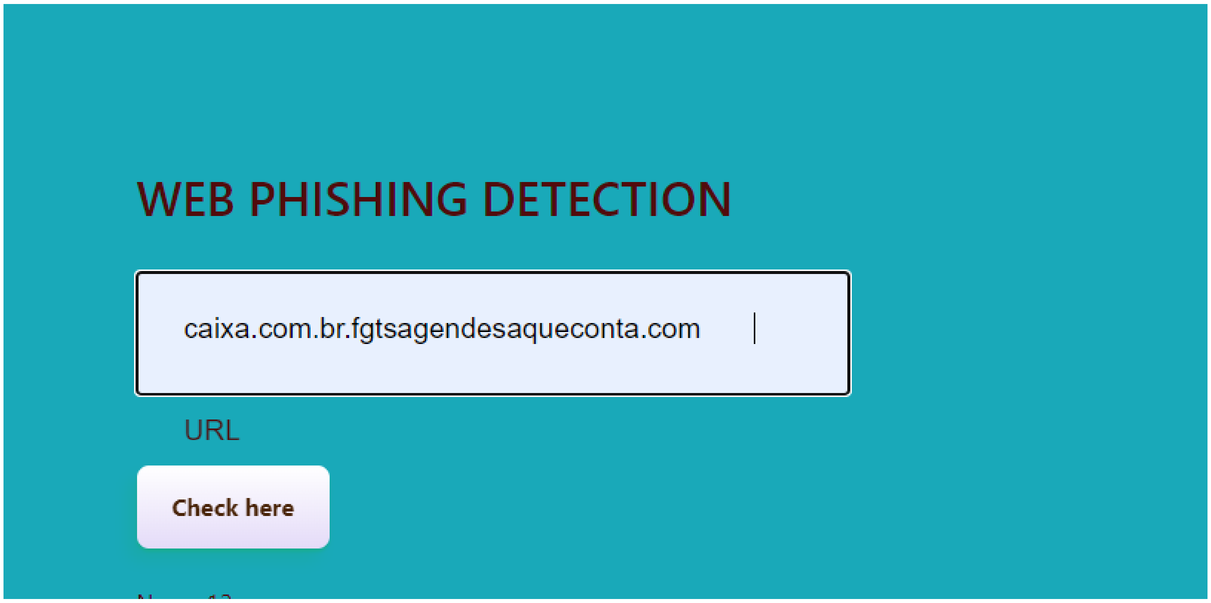
13.2 SCREEN SHOTS

Step1: Home Page of web Phishing Detection



WEB PHISHING DETECTION

Step 2: Enter the Website URL for Detecting the Website is Phishing or not

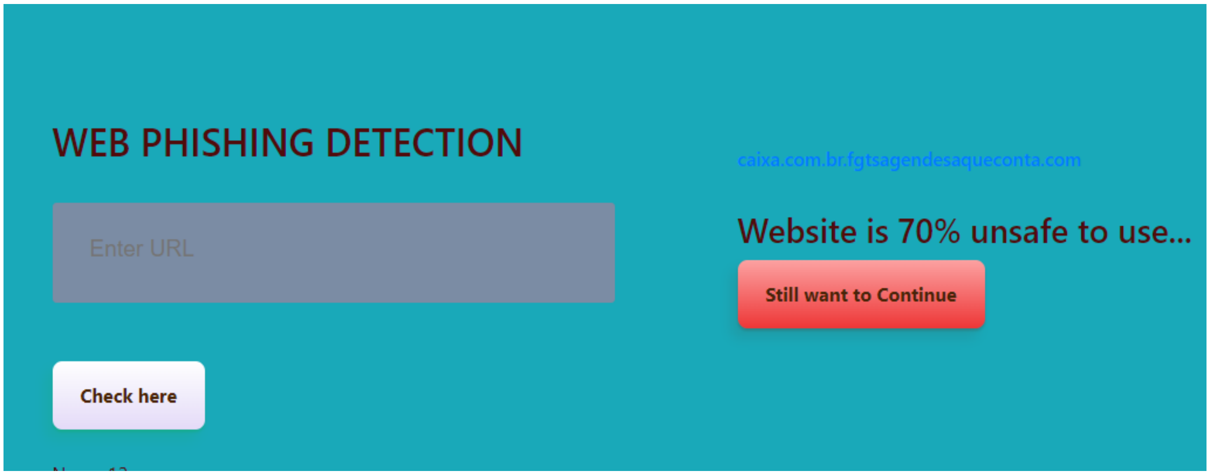


WEB PHISHING DETECTION

caixa.com.br.fgtsagendesaqueconta.com

URL

Step 3: The Website is phishing it shows the warning Message like Website is unsafe to use



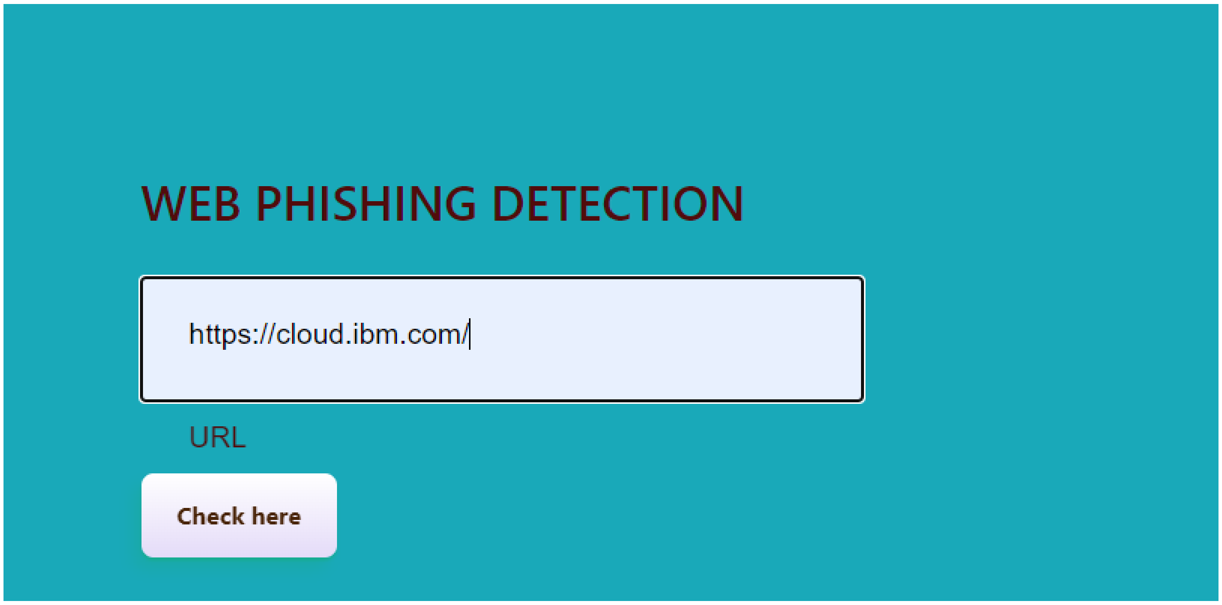
WEB PHISHING DETECTION

Enter URL

caixa.com.br.fgtsagendesaqueconta.com

Website is 70% unsafe to use...

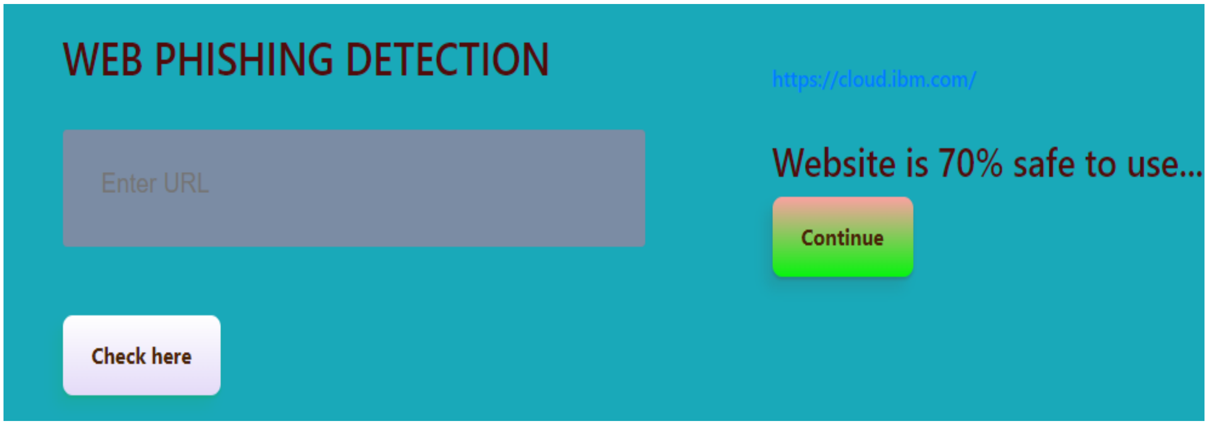
Step 4: Enter the Website URL for Detecting the Website is Phishing or not



WEB PHISHING DETECTION

URL: https://cloud.ibm.com/

Step 5: The Website is Legitimate it shows the Message safe to use



WEB PHISHING DETECTION

Enter URL

https://cloud.ibm.com/

Website is 70% safe to use...

13.2 GITHUB & PROJECT DEMO LINK

https://youtu.be/NNPUkToXHuE