1.INTRODUCTION

1.1 Project Overview

Phishing costs Internet users billions of dollars per year. It refers to luring techniques used by identity thieves to fish for personal information in a pond of unsuspecting Internet users. Phishers use spoofed e-mail, phishing software to steal personal information and financial account details such as usernames and passwords. This paper deals with methods for detecting phishing Web sites by analyzing various features of benign and phishing URLS by Machine learning techniques. We discuss the methods used for detection of phishing Web sites based on lexical features, host properties and page importance properties. We consider various machine learning algorithms for evaluation of the features in order to get a better understanding of the structure of URLS that spread phishing. The fine-tuned parameters are useful in selecting the apt machine learning algorithm for separating the phishing sites from benign sites.

1.2 Purpose

Phishing is one of the techniques which are used by the intruders to get access to the user credentials or to gain access to the sensitive data. This type of accessing the is done by creating the replica of the websites which looks same as the original websites which we use on our daily basis but when a user click on the link he will see the website and think its original and try to provide his credentials.

To overcome this problem we are using some of the machine learning algorithms in which it will help us to identify the phishing websites based on the features present in the algorithm. By using these algorithm we cam be able to keep the user personal credentials or the sensitive data safe from the intruders.

2. LITERATURE SURVEY

2.1 Existing Problem

The Main Probelm is to attempts to steal your money on your identity, by getting you to restal personal information-anch as credit card numbers, bank information, or passwords on websites that pretend to be legitimate.

2.2 References

1.https://www.researchgate.net/publication/328541785_Phishing_Website_Detection_using_Machine_Learning_Algorithms

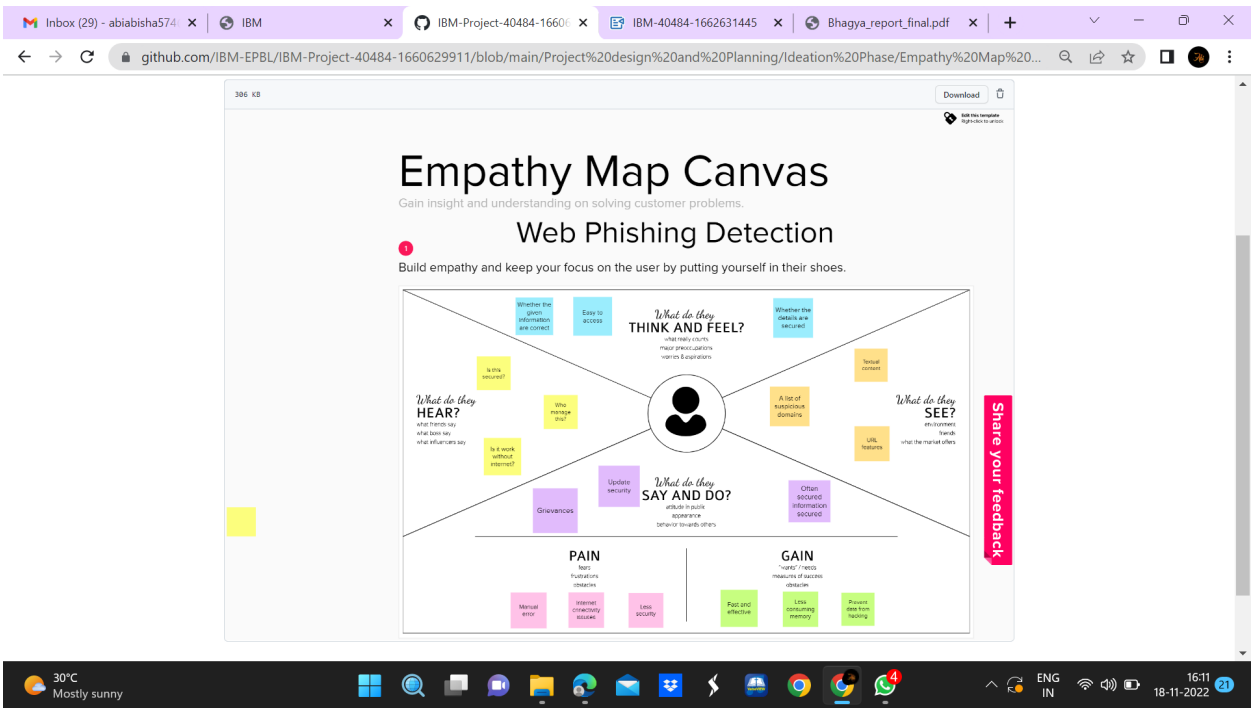
2.3 Problem Statement Definition

Internet has dominated the world by dragging half of the world's population exponentially into the cyber world. With the booming of internet transactions, cybercrimes rapidly increased and with anonymity presented by the internet, Hackers attempt to trap the end-users through various forms such as phishing, SQL injection, malware, man-in-the-middle, domain name system tunnelling, ransomware, web trojan, and so on. Among all these attacks, phishing reports to be the most deceiving

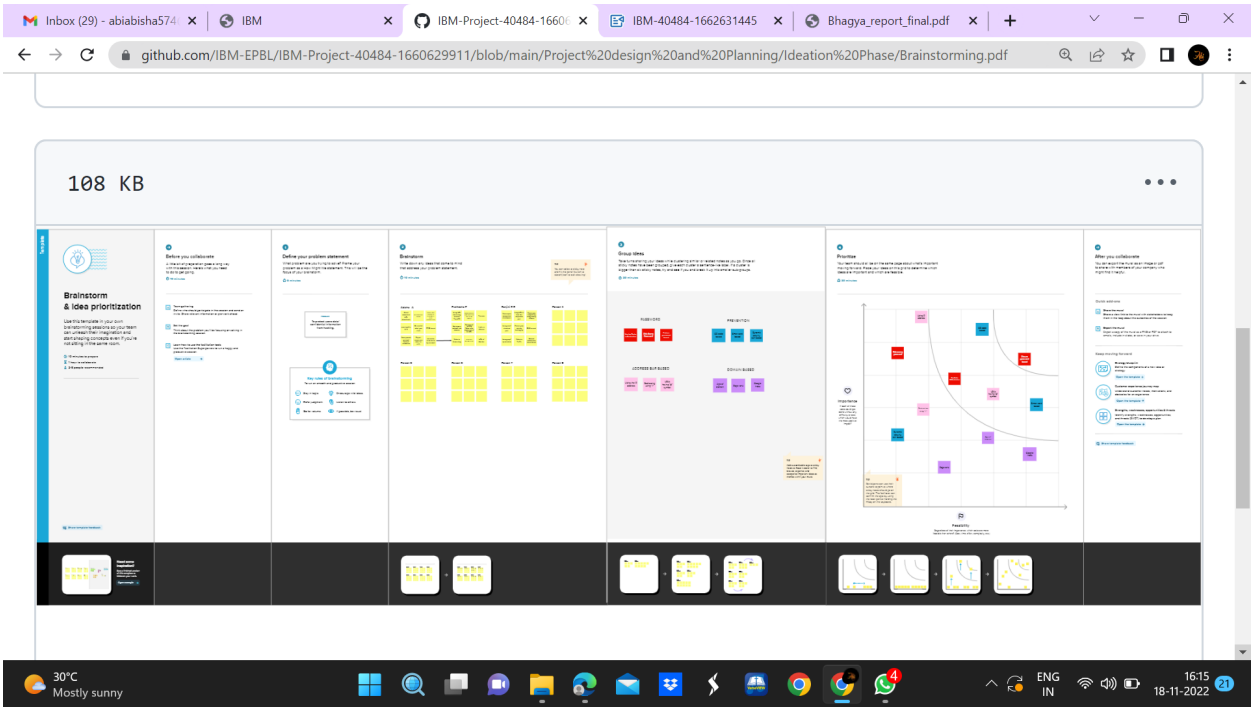
attack. Our main aim of this paper is classification of a phishing website with the aid of various machine learning techniques to achieve maximum accuracy and concise model.

3.IDEATION & PROPOSED SOLUTION

3.1 EMPATHY MAP CANVAS



3.2 IDEATION AND BRAINSTROING



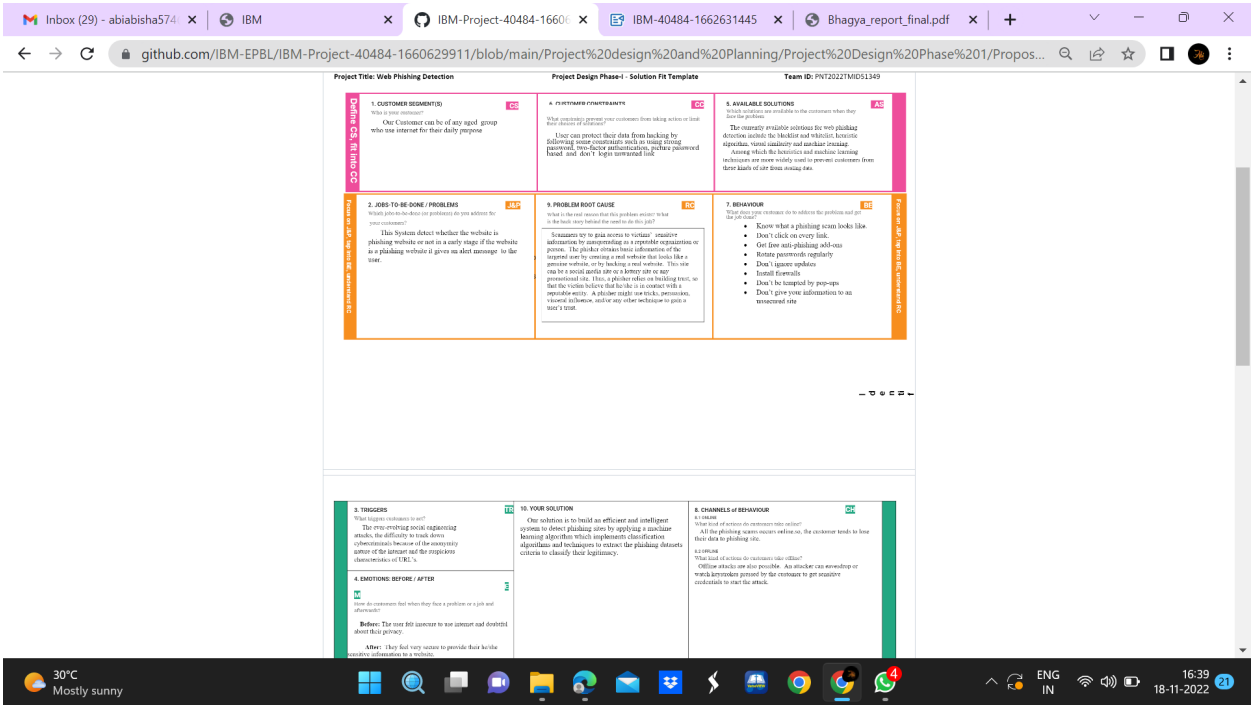
3.3 Proposed Solution

Proposed Solution Template:

Project team shall fill the following information in proposed solution template.

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	To detect whether the e-banking website is phishing website or not.
2.	Idea / Solution description	By using Machine Learning to design an efficient and adaptive phishing detection algorithm that has the ability to adapt the newer data or phishing attack vectors discovered in the wild and it's detection based on some important characteristics such as URL, domain identity and security.
3.	Novelty / Uniqueness	Using classification algorithm and techniques for automated web phishing detection approach by this approach it check the web URL . If the website is not secured it give an alert message to the user.
4.	Social Impact / Customer Satisfaction	This web phishing detection project attains the customer satisfaction by discarding various kinds of malicious websites to prevent their privacy. This project is not only capable of using by a single individual, a large social community and an organization can use this web phishing detection to protect their privacy. This project helps to block various malicious websites simultaneously.
5.	Business Model (Revenue Model)	This developed model can be used as an enterprise applications by organisations which handles sensitive information and also can be prevent the loss of potential important data.
6.	Scalability of the Solution	This project's performance rate will be high and it also provide many capabilities to the user without reducing its efficiency to detect the malicious websites. Thus, scalability of this project will be high.

3.4 Problem Solution Fit



Project Title: Web Phishing Detection

Project Design Phase-I - Solution Fit Template

Team ID: PNT2022TMID51349

Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS

Who is your customer?

Our Customer can be of any aged group who use internet for their daily purpose

4. CUSTOMER CONSTRAINTS CC

What constraints prevent your customers from taking action or limit their choices of solutions?

User can protect their data from hacking by following some measures such as using strong password, two-factor authentication, perform password based and don't login unwanted link

5. AVAILABLE SOLUTIONS AS

Which solutions are available to the customers when they face the problem

The currently available solutions for web phishing detection include the blacklist and whitelist, heuristic algorithm, visual similarity and machine learning.

Among which the heuristics and machine learning techniques are more widely used to prevent customers from these kinds of site from stealing data.

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

2. JOBS-TO-BE-DONE / PROBLEMS J&P

Which jobs-to-be-done (or problems) do you address for your customers?

This System detect whether the website is phishing website or not in a early stage if the website is a phishing website it gives an alert message to the user.

9. PROBLEM ROOT CAUSE RC

What is the real reason that this problem exists? What is the basic story behind the need to do this job?

Scammers try to gain access to victims' sensitive information by masquerading as a reputable organization or person. The phisher obtains basic information of the targeted user by creating a real website that looks like a genuine website, or by hacking a real website. This site can be a social media site or a lottery site or any promotional site. Thus, a phisher relies on building trust, so that the victim believes that he/she is in contact with a reputable entity. A phisher might use tricks, persuasion, visceral influence, and/or any other technique to gain a user's trust.

7. BEHAVIOUR BE

What does your customer do to address the problem and get the job done?

- Know what a phishing scam looks like.
- Don't click on every link.
- Get free anti-phishing add-ons
- Rotate passwords regularly
- Don't ignore updates
- Install firewalls
- Don't be tempted by pop-ups
- Don't give your information to an unsecured site

3. TRIGGERS TR

What triggers customers to act?

The ever-evolving social engineering attacks, the difficulty to track down cybercriminals because of the anonymity nature of the internet and the suspicious characteristics of URL's.

4. EMOTIONS: BEFORE / AFTER EM

How do customers feel when they face a problem or a job and afterwards?

Before: The user felt insecure to use internet and doubtful about their privacy.

After: They feel very secure to provide their he/she sensitive information to a website.

10. YOUR SOLUTION

Our solution is to build an efficient and intelligent system to detect phishing sites by applying a machine learning algorithm which implements classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy.

8. CHANNELS of BEHAVIOUR CH

8.1 ONLINE

What kind of actions do customers take online?

All the phishing scams occurs online so, the customer tends to lose their data to phishing site.

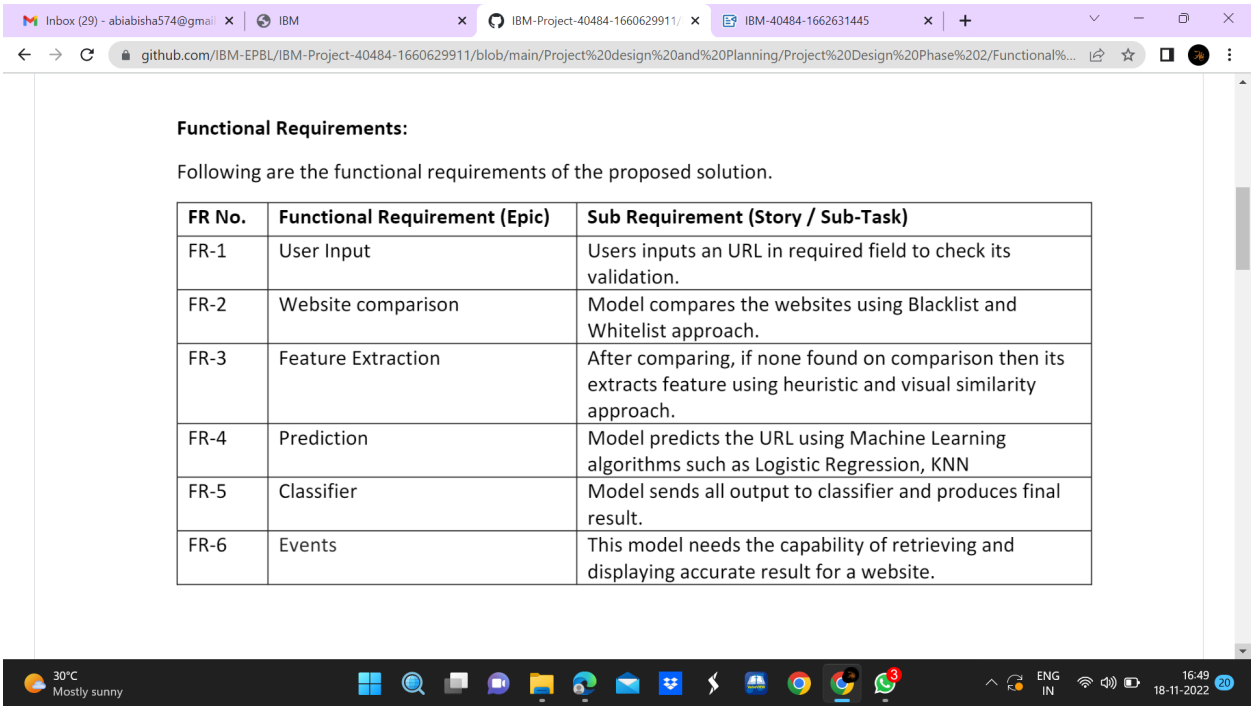
8.2 OFFLINE

What kind of actions do customers take offline?

Offline attacks are also possible. An attacker can eavesdrop or watch keystrokes pressed by the customer to get sensitive credentials to start the attack.

4.REQUIREMENT ANALYISIS

4.1 Functional Requirements

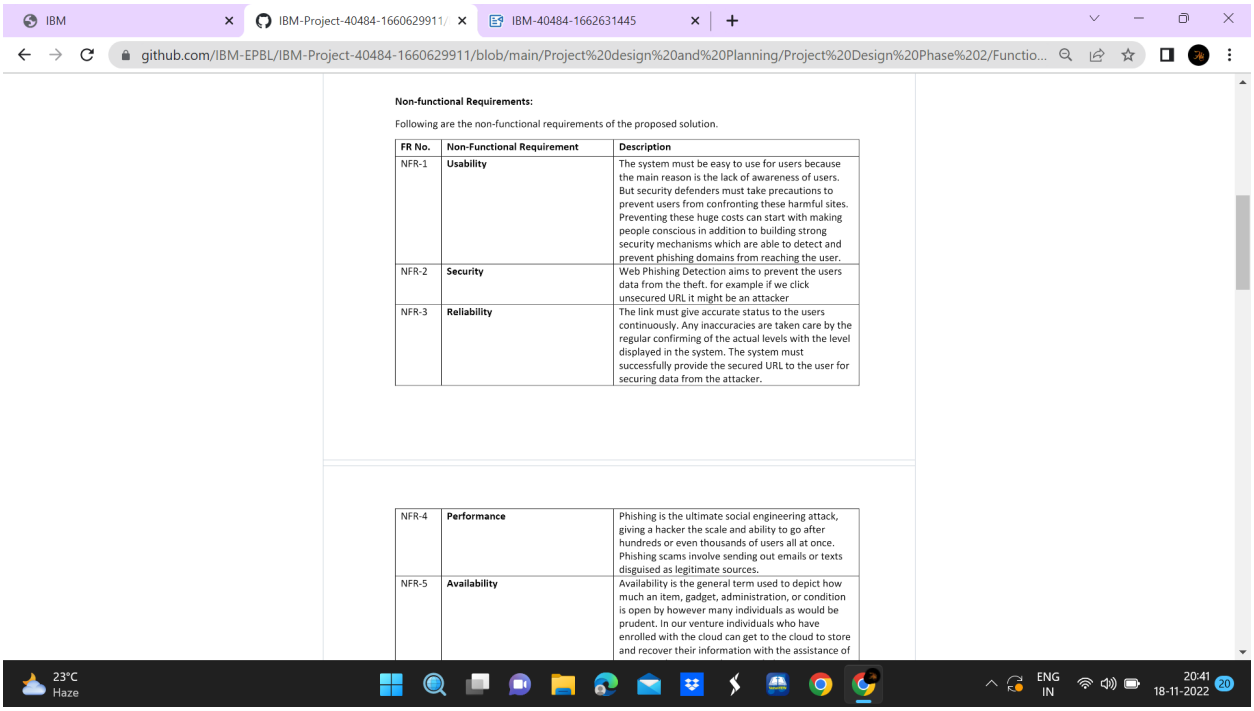


Functional Requirements:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Input	Users inputs an URL in required field to check its validation.
FR-2	Website comparison	Model compares the websites using Blacklist and Whitelist approach.
FR-3	Feature Extraction	After comparing, if none found on comparison then its extracts feature using heuristic and visual similarity approach.
FR-4	Prediction	Model predicts the URL using Machine Learning algorithms such as Logistic Regression, KNN
FR-5	Classifier	Model sends all output to classifier and produces final result.
FR-6	Events	This model needs the capability of retrieving and displaying accurate result for a website.

4.2 Non-Functional Requirements



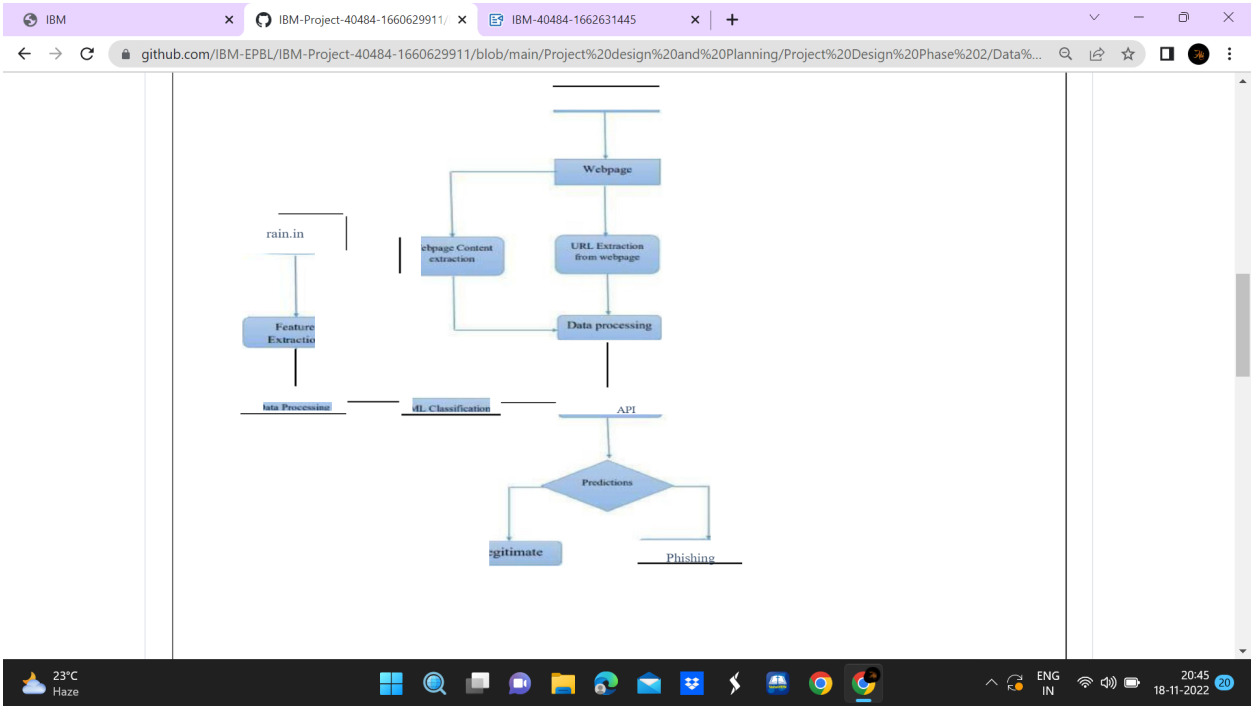
Non-functional Requirements:

Following are the non-functional requirements of the proposed solution.

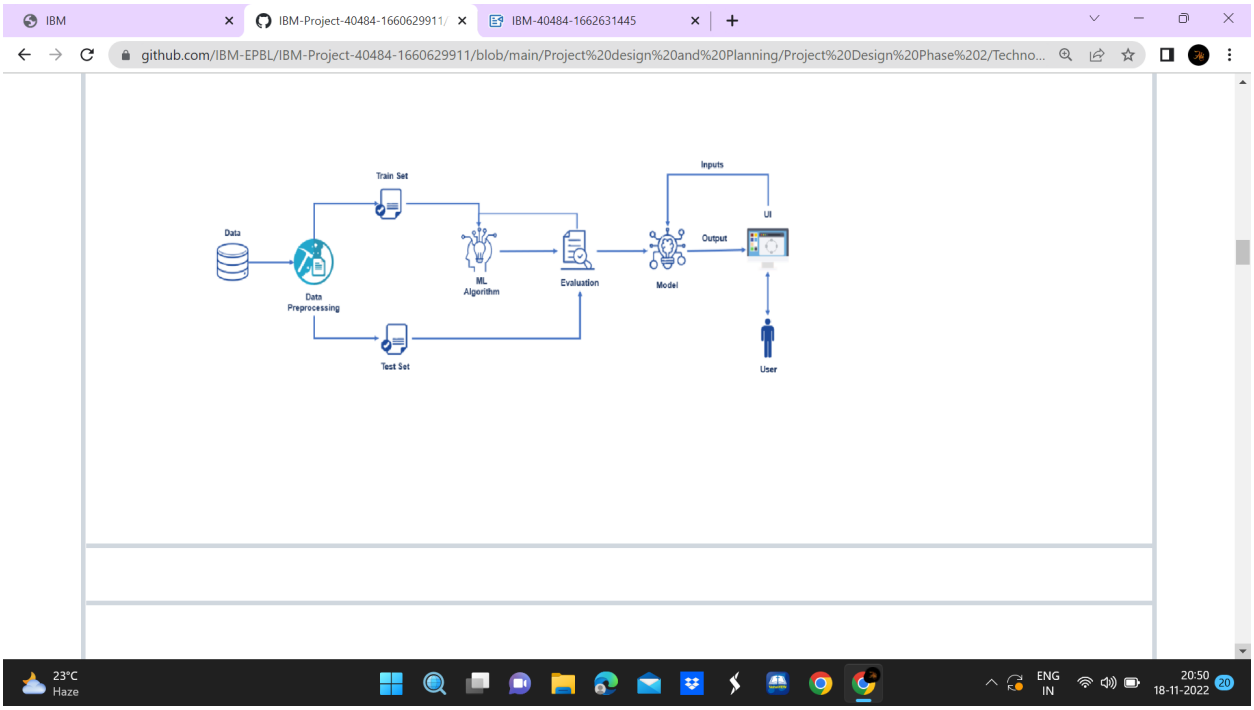
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The system must be easy to use for users because the main reason is the lack of awareness of users. But security defenders must take precautions to prevent users from confronting these harmful sites. Preventing these huge costs can start with making people conscious in addition to building strong security mechanisms which are able to detect and prevent phishing domains from reaching the user.
NFR-2	**Security**	Web Phishing Detection aims to prevent the users data from the theft. for example if we click unsecured URL it might be an attacker
NFR-3	**Reliability**	The link must give accurate status to the users continuously. Any inaccuracies are taken care by the regular confirming of the actual levels with the level displayed in the system. The system must successfully provide the secured URL to the user for securing data from the attacker.
NFR-4	**Performance**	Phishing is the ultimate social engineering attack, giving a hacker the scale and ability to go after hundreds or even thousands of users all at once. Phishing scams involve sending out emails or texts disguised as legitimate sources.
NFR-5	**Availability**	Availability is the general term used to depict how much an item, gadget, administration, or condition is open by however many individuals as would be prudent. In our venture individuals who have enrolled with the cloud can get to the cloud to store and recover their information with the assistance of

5.PROJECT DESIGN

5.1 Data Flow Diagram



5.2 Solution & Technical Architecture



5.3 User Stories

IBM | IBM-Project-40484-1660629911/ | IBM-40484-1662631445

github.com/IBM-EPBL/IBM-Project-40484-1660629911/blob/main/Project%20design%20and%20Planning/Project%20Design%20Phase%202/Data%...

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)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirmingmy password.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Facebook	I can register & access the dashboard with Facebook Login	Low	Sprint-2
		USN-4	As a user, I can register for the application through Gmail		Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password		High	Sprint-1
	Dashboard					
Customer (Web user)	User input	USN-1	As a user i can input the particular URL in the required field and waiting for validation.	I can go access the website without any problem	High	Sprint-1
Customer Care Executive	Feature extraction	USN-1	After i compare in case if none found on comparison then we can extract feature using heuristic and visual similarity approach.	As a User i can have comparison between websites for security.	High	Sprint-1
Administrator	Prediction	USN-1	Here the Model will predict the URL websites using Machine Learning algorithms such as Logistic Regression, KNN	In this i can have correct prediction on the particular algorithms	High	Sprint-1
	Classifier	USN-2	Here i will send all the model output to classifier in order to produce final result.	I this i will find the correct classifier for producing the result	Medium	Sprint-2

23°C Haze | ENG IN | 20:45 18-11-2022

6.PROJECT PLANNING AND SCHEDULING

6.2 Sprint Delivery Schedule

Project Tracker, Velocity & Burndown Chart: (4 Marks)

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	05 Nov 2022	20	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

Velocity:
Imagine we have a 10-day sprint duration, and the velocity of the team is 20 (points per sprint). Let's calculate the team's average velocity (AV) per iteration unit (story points per day)

$$AV = \frac{sprint\ duration}{velocity} = \frac{20}{10} = 2$$

7. CODING AND SOLUTIONING

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pickle
# %matplotlib inline

# Filter the uneccesary warnings
import warnings
warnings.filterwarnings("ignore")

df1 = pd.read_csv('Phishing.csv')
df = pd.DataFrame()
df['SSLfinal_State']=df1['SSLfinal_State']
df['URL_of_Anchor']=df1['URL_of_Anchor']
df['Prefix_Suffix']=df1['Prefix_Suffix']
df['web_traffic']=df1['web_traffic']
df['Domain_registeration_length']=df1['Domain_registeration_length']
df['Result']=df1['Result']

df['Result'] = df['Result'].map({-1:0, 1:1})
df['Result'].unique()
#to check null values in the dataframe
df.isnull()

from sklearn.model_selection import train_test_split
X=df.drop("Result",axis=1).values
y=df["Result"].values
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=101)

```

```
from sklearn.ensemble import RandomForestClassifier
error= []
# Will take some time
for i in range(550,600):
    rfc = RandomForestClassifier(n_estimators=i)
    rfc.fit(X_train,y_train)
    pred_i = rfc.predict(X_test)
    error.append(np.mean(pred_i != y_test))

rfc = RandomForestClassifier(n_estimators=571)
rfc.fit(X_train,y_train)


pickle.dump(rfc,open('model.pkl','wb'))
model=pickle.load(open('model.pkl','rb'))
```

8.TESTING

8.1 Test Cases

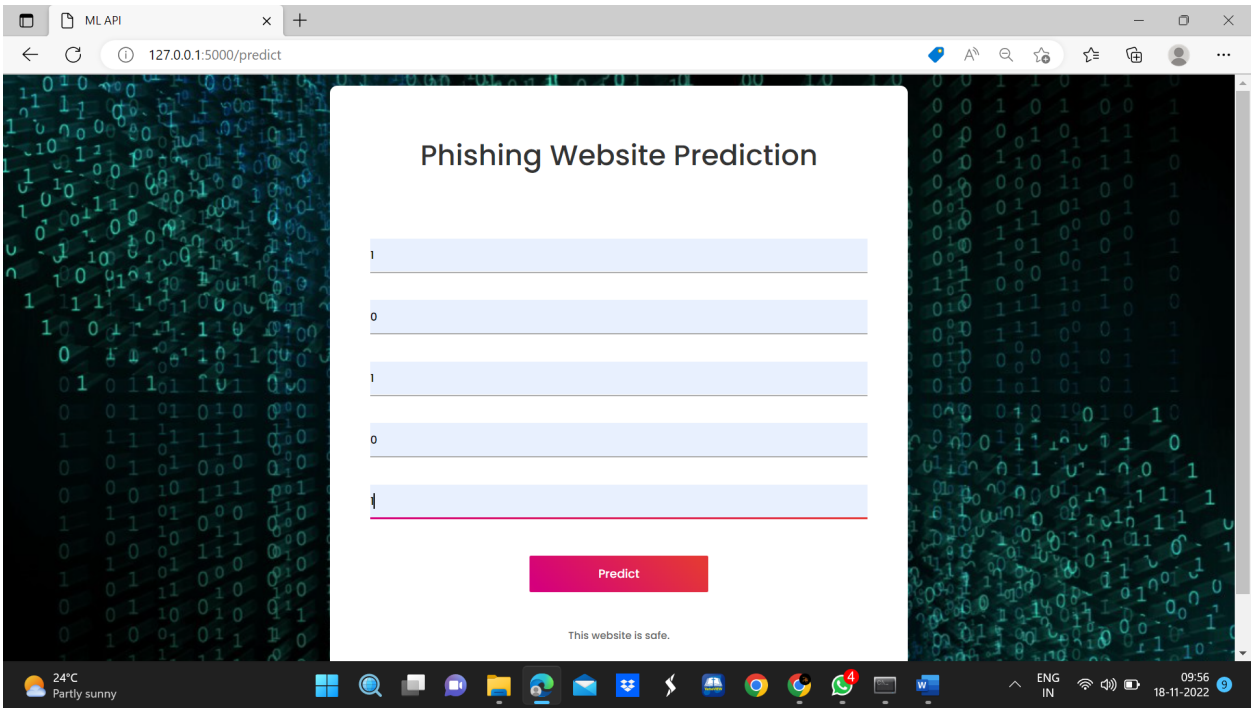
A test case is a document, which has a set of test data, preconditions, expected results and postconditions, developed for a particular test scenario in order to verify compliance against a specific requirement.

Test Case acts as the starting point for the test execution, and after applying a set of input values, the application has a definitive outcome and leaves the system at some end point or also known as execution postcondition.

8.2 User Acceptance Testing

User Acceptance Testing (UAT) is a type of testing performed by the end user or the client to verify/accept the software system before moving the software application to the production environment. UAT is done in the final phase of testing after functional, integration and system testing is done.

9.RESULTS


ML API
127.0.0.1:5000/predict
Phishing Website Prediction
1
0
1
0
1
Predict
This website is safe.
24°C
Partly sunny
ENG
IN
09:56
18-11-2022

10. ADVANTAGES AND 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.

Disadvantages:

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

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.

12. FUTURE SCOPE

Further work can be done to enhance the model by using ensembling models to get greater accuracy score. Ensemble methods is a ML technique that combines many base models to generate an optimal predictive model. Further reaching future work would be combining multiple classifiers, trained on different aspects of the same training set, into a single classifier that may provide a more robust prediction than any of the single classifiers on their own.

The project can also include other variants of phishing like smishing, vishing, etc. to complete the system. Looking even further out, the methodology needs to be evaluated on how it might handle collection growth. The collections will ideally grow incrementally over time so there will need to be a way to apply a classifier incrementally to the new data, but also potentially have this classifier receive feedback that might modify it over time.

Github link: https://github.com/IBM-EPBL/IBM-Project-40484-1660629911

Project Demo Link:https://drive.google.com/file/d/1zmmy7GBDyfQ-yNdyMV3tK9PSbXZ1i_7T/view?usp=sharing