WEBPHISHINGDETECTION

Domain : Applied Data

ScienceA PROJECTREPORT

Submittedby

NAVEENKUMAR.S (920819104025)

AJAYKUMAR.M (920819104003)

SRIVATHSKARTHIC.G (920819104042)

GUNA SEAKAR.J (920819104012)

inpartialfulfillmentfortheawardofthedegree

of

BACHELOROFENGINEERING

in

COMPUTERSCIENCEANDENGINEERING

NPRCOLLEGEOFENGINEERING&TECHNOLOGYN

ATHAM,DINDIGUL.

ANNAUNIVERSITY::CHENNAI600025

TABLEOF CONTENTS

CHAPTER 1

INTRODUCTION

PROJECTOVERVIEW:

Phishingisaformoffraudinwhichtheattacksthistolearnsensitiveinformationsuchas logincredentials(or)accountinformationbysendingasareputable entity (or) person in email (or) other communication channels. Typically avictim receives a message that appears to have been sent by a known contact (or)organization. Phishing is popular among attackers , it is easier to tick someone intoclickingamaliciouslinkwhichseemslegitimatethantypingtobreakthroughacomputes defences, systems. So regarding this solution detecting phishing domains is aclassifications problem. So we need labeled data which has sample as phish. Domains& legitimate domains in the training plans. Collecting legitimate domains is anotherproblem. For this purpose, sites reputation services are among used. This servicesanalysis & Rank, available websites, 4 decision can be considered as an improvedwasted of else structures. Which segregates legitimates of phishing websites. So byimplementing this. We could develop a noble Machine learning tool for phishingdetection.

PURPOSE:

Our purpose is to develop a tool that is capable of detecting phising websites. This improvises the security status of the uses machine by using the legitimate website could provide a safe uses experiences. So by using our tool the uses data could be prevented from intruders.

We have included decisior the algorithm to regregatewheather website is phsihing of lighitimate along with this rain forest algorithm is used to improvise the accuracy. So as a combinature of such facilities we provide a issue internet experiences for the end consumers.

CHAPTER 2
LITERATURE SURVEY

- Detectionof phishing websites by using machinelearning basedURLAnalysisHelmetkmkmaz,Banypiri,Dzguskaraysahingoz.

This approach is URL Analysis k- nearest neighbourhood (kNN),supportvectormachine(VMM),DecisionTree(DT)withmeritscomprisingofeffective prediction of phishing domains based on logistic regression. Using hybridalgorithm the accuracy is enhanced. The main drawbacks are time consumption ofhighprocessingpower.

- Phishingwebsitedetectionbasedonmachinelearningalgorithmbyweighbhai.

This is a phishing detection based on machine learning & logisticregression. The benefits include high threshold value due to which accuracy increases&By using logistic regression. The speed can be improvised and regarding drawbacksdataprocessingis required.

2.2.REFERENCES:

TITLE	YEAR	AUTHORS	TECHNOLOGY ADOPTED	MERITS	DEMERITS
Real Time Detection of Phishing Websites	2016	ABDULGHANI ALI AHMED, NURUL AMIRAH ABDULLAH	Checking Uniform Resources Locators (URLs)	Increased accuracy by 32% than previous proposed systems. Identifies URL redirecting to other web pages by analyzing the URL type.	Accuracy depends on heuristic band and depends on discriminative features. Only checks validity of URLs.
Detection of Phishing Websites by Using Machine Learning-Based URL Analysis	2020	Mehmet Korkmaz, Ozgur Koray Sahingoz, Banu Diri	Machine Learning-Based URL Analysis (Logistic Regression (LR), K-Nearest Neighborhood (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), XGBoost, Random Forest (RF) and Artificial Neural Network (ANN))	Logistic Regression is the algorithm that generates effective predictions of phishing domain elements. Hybrid algorithms enhance the accuracy.	Takes time to identify the phishing website. Requires high processing power.
Phishing Website Detection Based on Machine Learning Algorithm	2020	Weiheng Bai	Machine Learning Algorithm (Logistic regression classifier)	With high threshold value, it has high accuracy rate. Logistic regression classifies being used improves the speed.	Not completely reliable because transmission of packets does not reflect the proximity of location . Data preprocessing is required.
Machine Learning Techniques for Detection of Website Phishing: A Review for Promises and Challenges	2021	Ammar Odeh, Ismail Keshta, Eman Abdelfattah	Machine learning, Deep learning (Heuristic and automated techniques)	Use of Machine learning improvised the URLs. Stacking model has improved the accuracy by detecting legitimate websites.	Large binds of datasets are difficult to handle. Low accuracy and hypertuning.

2.3 Problemstatement:

When a person asks the internet he/she might be concerned with the securitybeing provided. So when that person opens a link/website & enters their credentials ,he/shegetshacked(or)phished.

So in order to avoid this we have developed a tool backed up with machinelearning which uses certain criteria to evaluates this URL & declares a websitesasphisjhing(or)legitimateone.

IdeationPhase
DefinetheProblemStatements

Date	19September2022
TeamID	PNT2022TMID48632
ProjectName	WebPhishingDetection
MaximumMarks	2Marks

CustomerProblemStatementTemplate:

Create aproblem statementto understandyour customer's pointof view. The Customer Problem Statement template helps you focus on what matterstocreateexperiencespeoplewilllove.

A well-articulated customer problem statement allows you andyourteamtofindtheidealsolutionforthechallengesyourcustomersface.Throughout the process, you’ll also be able to empathize with your customers, whichhelpsyoubetterunderstandhowtheyperceiveyourproductorservice.


I am
Student
I'm trying to
Opens an compromised link and enters the credentials
But
Gets hacked
Because
It was a fake website
Which makes me feel
Insecure
I am
Shopkeeper
I'm trying to
Opens a link from bank and enters the account credentials
But
His bank account details and his credit card details gets hacked
Because
The link was fake and the hacker aims for the money from the account
Which makes me feel
Untrust worthy
I am
Homemaker
I'm trying to
make transactions for the purchased items from checkouts
But
Ended up as she purchases from a fake website
Because
The website has an advertisement for discounts
Which makes me feel
Disappoint
I am
Employer
I'm trying to
complete the tasks by the end of the day
But
found a file through mail and opens it and his company details get hacked
Because
The file seemed to be phishing and fake one
Which makes me feel
Annoyed and disturbed
miro

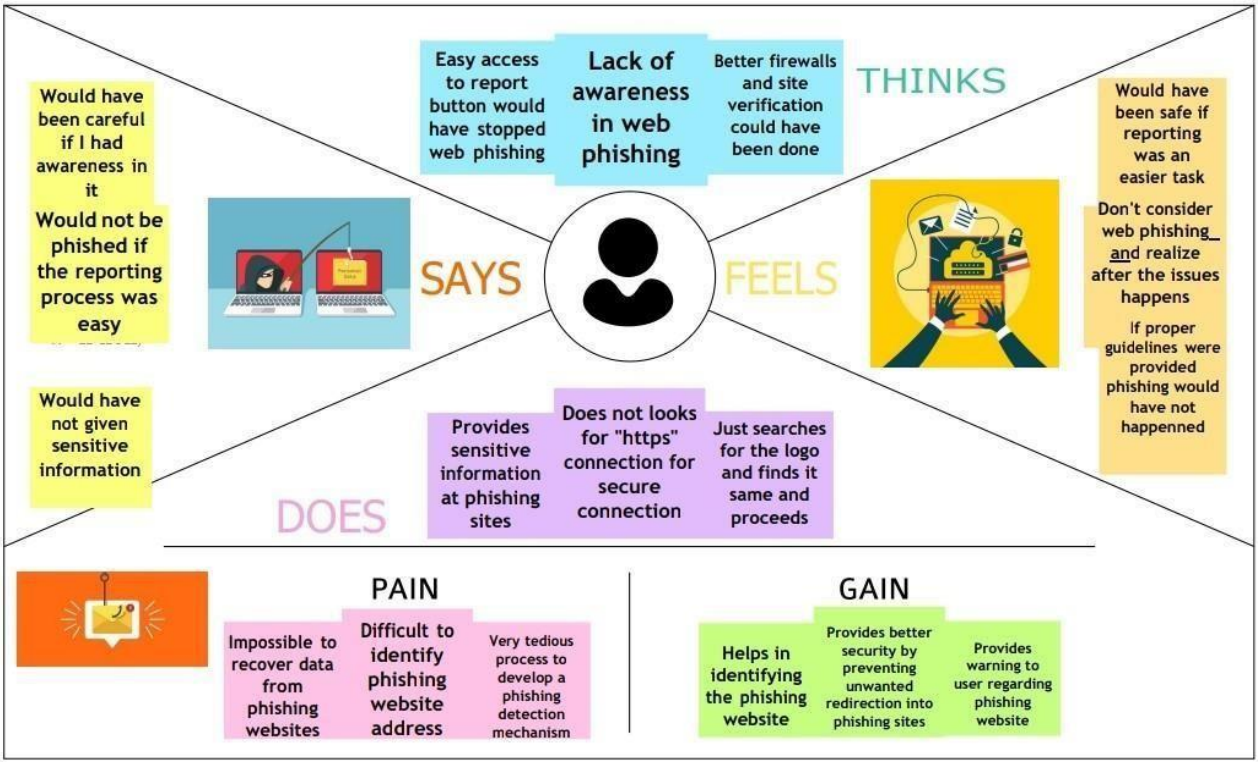
ProblemStatement(PS)	Iam(Customer)	I'mtryingto	But	Because	Whichmake smefeel
PS-1	Student	Opens acompromisedlinkand enters thecredentials	GetsHacked	Itwasafakewebsite	Insecure
PS-2	Shopkeeper	Opensalinkfromabank and enterstheaccountdetails	His bankaccountdetails andhis creditcard detailsgetshacked	The link wasfakeand thehacker aims forthemoneyfromtheaccount	Untrustworthy
PS-3	Homemaker	Maketransactions forthe purchaseditemsfromcheckouts	Ended up asshepurchasesfromanfakewebsite	The website hadanadvertisementfordiscounts	Disappointed
PS-4	Employer	Complete thetasksbyendoftheday	Found a filethroughmailandopensit,andhiscompanydetails getshacked	The fileseemedto be phishingandfakeone	Annoyedanddisturbed

CHAPTER 3
IDEATION PHASE

Empathymap:

An**empathymap**isacollaborativevisualizationusedtoarticulatewhatweknowabout aparticulartypeofuser.It externalizesknowledge about usersinorderto

1) Createasharedunderstandingofuserneeds, and
2) Aidindecisionmaking.



Explanation:

✓ Whatdotheythink&feel?

- Feelsafetouseinternet.
- Avoidstressduetodataloss.
- Eliminatesforgery.

✓ Whatdo theysee ?

- MachinelearningtoolthatdetectstheURL
- Popupsayingsafe/dangertouse.

✓ Whatdo theysee &do ?

- Safebrowsingenvironment.
- Stressfreesurfingontheinternet.

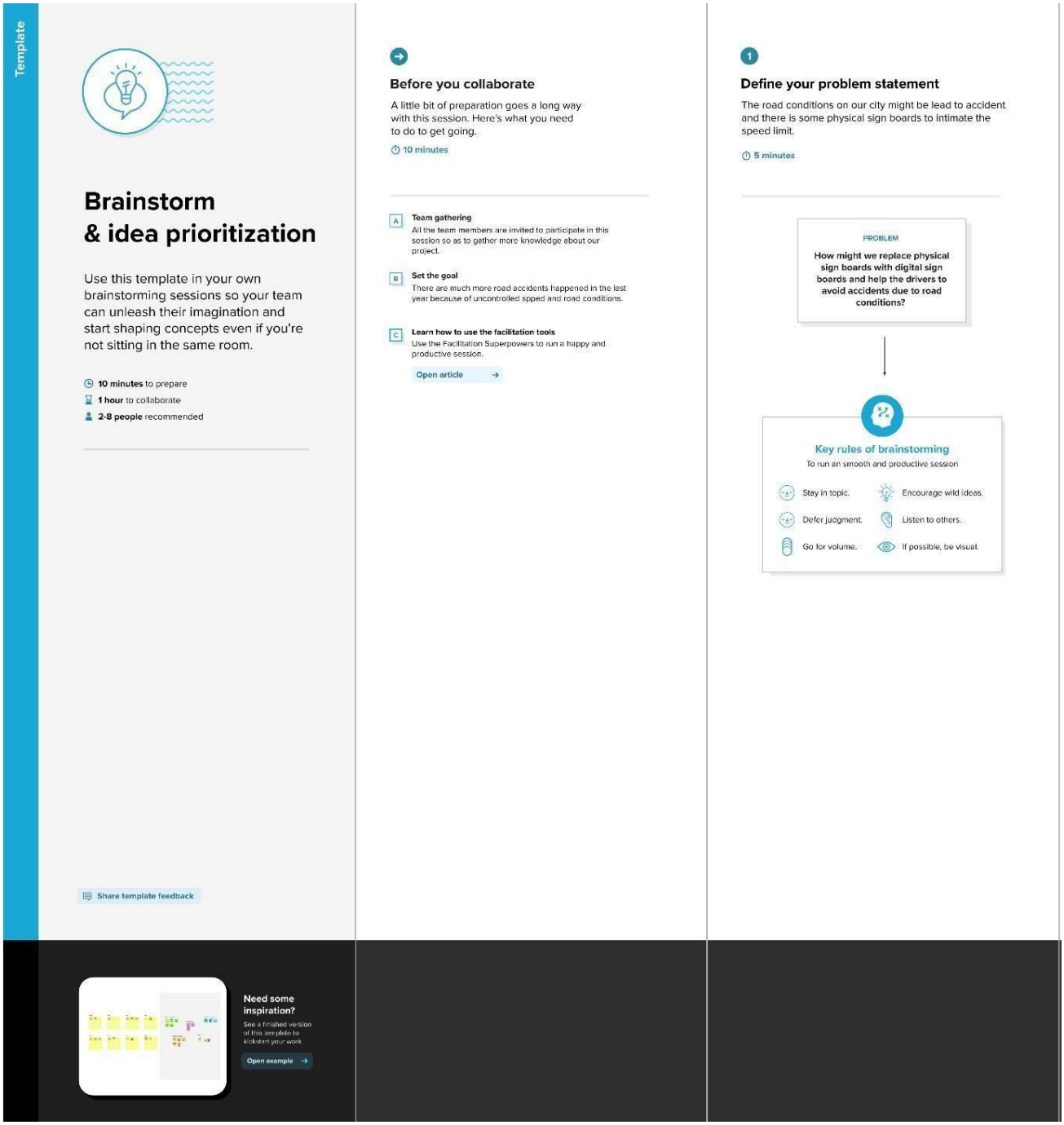
✓ Whatdotheyhear?

- Theymightbeabasicsoftwarepre-built onalldevices.

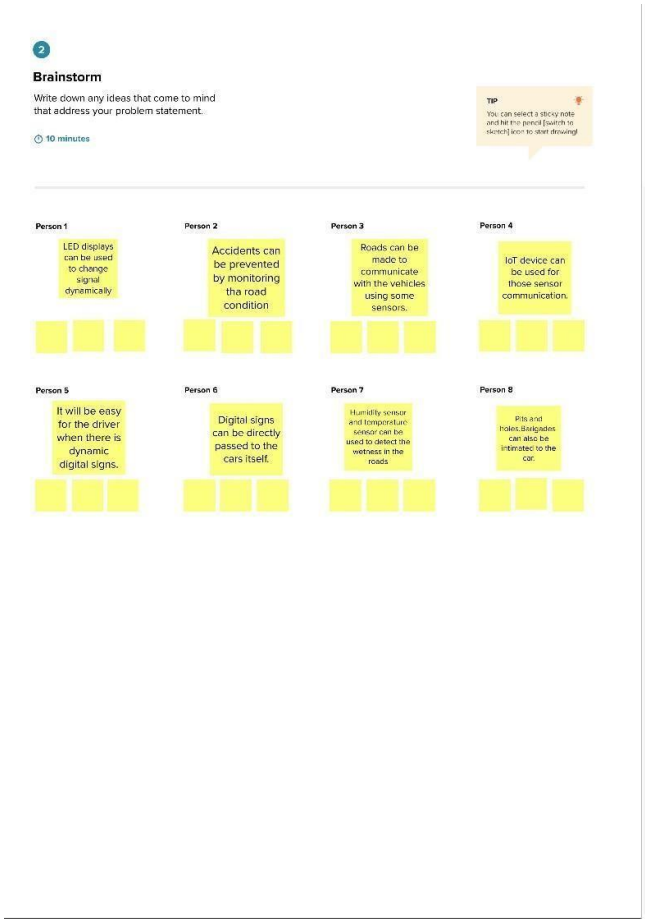
IdeationandBrainstroming:

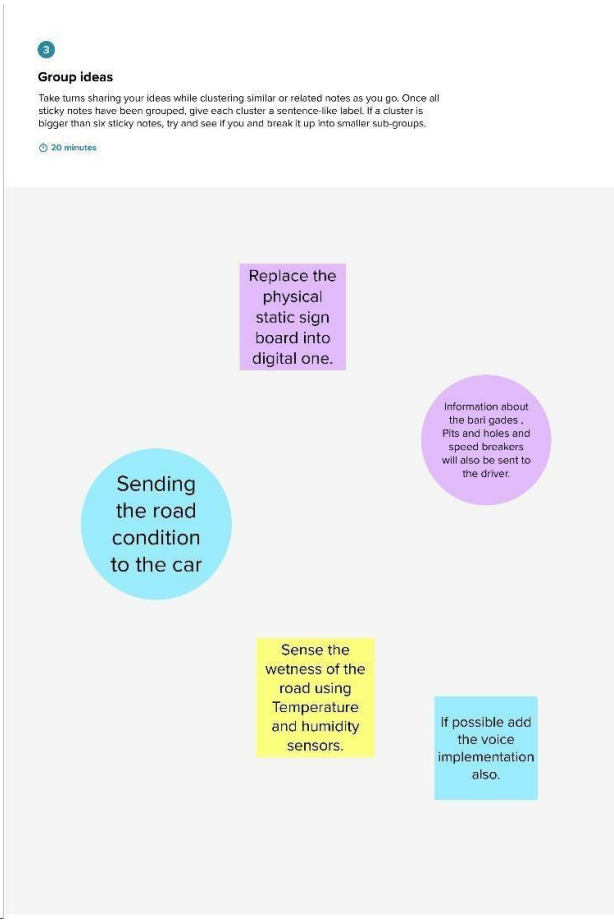
Brainstormingprovidesafreeandopenenvironmentthatencourageseveryone within a team to participate in the creative thinking process that leads toproblem solving. Prioritizing volume over value,out-of-the-box ideas are welcomeandbuiltupon,andallparticipantsareencouragedtocollaborate,helpingeach otherdeveloparichamountofcreativesolutions.

Step-1:TeamGathering,CollaborationandSelectthe ProblemStatement

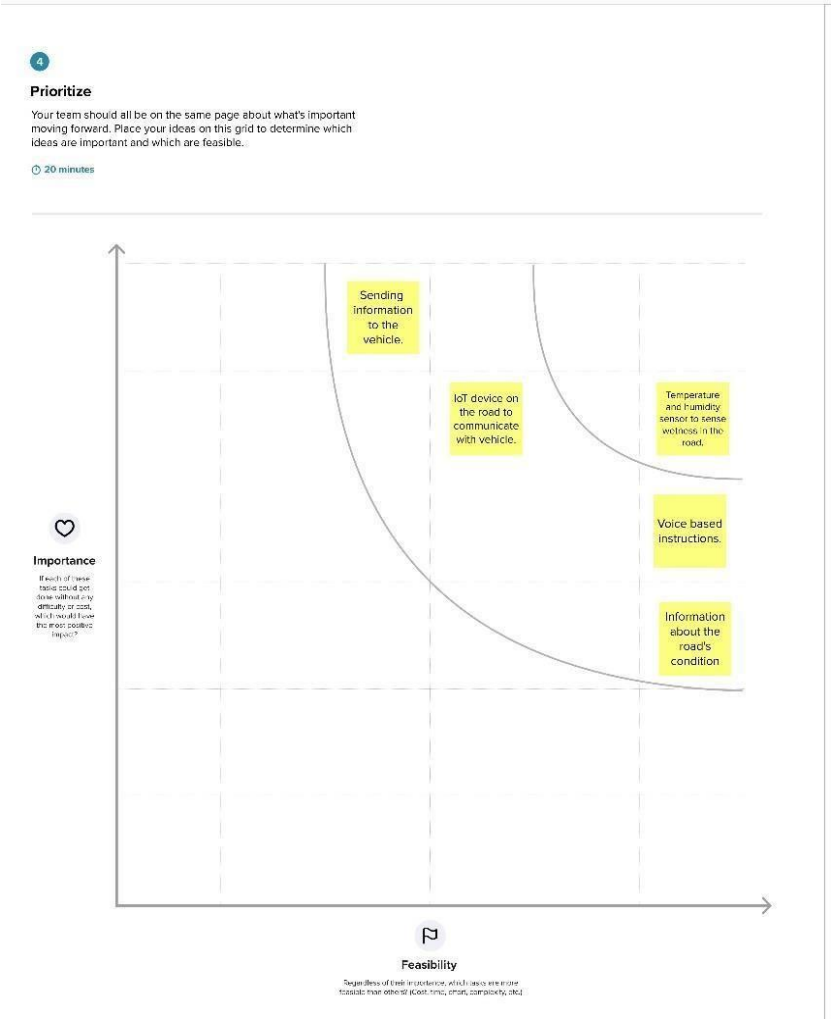


Step-2:Brainstorm,IdeaListingandGrouping:





Step-3:IdeaPrioritization



ProposedSolution:

S.No.	Parameter	Description
1.	ProblemStatement (Problemtobesolved)	There are a number of users who purchaseproducts online and make payments through e-banking. There are e-banking websites that askusers to provide sensitive data such asusername, password & credit card details, etc.,often for malicious reasons. Web phishing isoneofmanysecuritythreatstowebservicesonthe Internet. Commonthreatsofwebphishing: • Webphishingemphasizesonst ealprivateinformation • Itwillleadtoinformationdisclosure andpropertydamage. • Largeorganizations maygettrappedindifferentkindsofs cams.
2.	Idea/Solutiondescription	The proposed system handles the dataset andclassifies the dataset to be tested as genuine ornot by verifying certain criteria that arenecessary validate it . The datasets are nowevaluatedwherethenumberofcriteriathatha s been fulfilled plays a vital role. So basedonthe results that we get from the tools that weused (Decision tree algorithm), the website isdeclaredgenuineor not.
3.	Novelty/ Uniqueness	Implementation of Random Forestalgorithmalong with the decision treeimprovisestheaccuracyofthedetection.
4.	SocialImpact /CustomerSatisfaction	The proposed system prevents the user orcustomerfromfallingpreytothephishingandsc am websites by detecting thephishingwebsites.
5.	BusinessModel (RevenueModel)	By including premium subscription facility toprovide enhanced features to the customer wegenerateerevenuetothedevelopment team.
6.	ScalabilityoftheSolution	In this emerging world of technology, phishingurlswithnewscammingmethodscouldbei dentified usingoursystem.

ProblemSolutionFit:

Project Title: Web Phishing Detection | **Project Design Phase-I - Solution Fit Template** | **Team ID: PNT2022TMID48632**

Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS

All internet users

6.CUSTOMER CONSTRAINTS CC

- Difficult to detect exact replica of the actual site.
- Difficult to scan for malwares in files downloaded from that website.

5. AVAILABLE SOLUTIONS AS

- Alerts the user that the website redirects to another website.
- Alerts the user if the website automatically downloads files into the system.

Explore AS, differentiate

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

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

- Improper firewall protection leads to security threats.
- Difficult to retrieve data from phishing websites Tedious process to develop an phishing detection software.

9. PROBLEM ROOT CAUSE RC

Certain systems doesn't support the software as it's requirements are based on a baseline.

7. BEHAVIOUR BE

- User pretends to be precautious while surfing on the internet.

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

Identify strong TR & EM

3. TRIGGERS TR

- Seamless browsing facilitated with safety
- Protection from virus and malware

4. EMOTIONS: BEFORE / AFTER EM

- Before implementation of our projects people weren't aware whether they have been phished, but after which they started browsing without any stress

10. YOUR SOLUTION SL

- Linear regression technique and Decision tree algorithm are used to classify the website based on the dataset that was used to train the AI model
- Random forest algorithm is used to improvise the accuracy of the detection

8. CHANNELS of BEHAVIOUR CH

ONLINE

- The customer shares his/her feedback on the product through the online portals

OFFLINE

- The customer also shares his/her experience on the product to their peer group which seeks popularity for the product

Identify strong TR & EM

CHAPTER4

REQUIREMENT ANALYSIS

FUNCTIONALREQUIREMENTS

Functionalrequirementsmayinvolvecalculations,technicaldetails,datamanipulatio n and processing, and other specific functionality that define what a systemis supposed to accomplish. Behavioral requirements describe all the cases where thesystemusesthefunctionalrequirements,thesearecapturedinusecases.

Followingarethefunctional requirementsoftheproposedsolution.

FRNo.	FunctionalRequirement(Epic)	SubRequirement(Story/Sub-Task)
FR-1	HomePage	Theuserfindsthehomepageeasytonavigateandfeelscomfor tablewiththeuserinterface
FR-2	Sign up	The userwill get authenticationovertheiraccountbysecuritymea sures
FR-3	Login	The user can register using either their googleaccountortheirmobilenumber
FR-4	Dashboard	Theusercangothroughthefacilitiesprovidedbythe product
FR-5	Prediction	Userwouldbepromptedwithapopupindicat ingthetrustfulnessof thewebsite
FR-6	Resultspage	Theuserwouldbeabletoanalyzewebsitewhether it’sgenuineornot
FR-7	Reporting	Theusercanreportforanybugsoraskanyqueriesonthepro duct

NON FUNCTIONAL REQUIREMENTS:

Non-FunctionalRequirementsaretheconstraintsortherequirementsimposed onthe system. They specify the quality attribute of the software. Non-FunctionalRequirementsdealwithissueslikescalability,maintainability,performance,portability,security,reliability,andmanymore.

Followingarethenon-functionalrequirementsoftheproposedsolution.

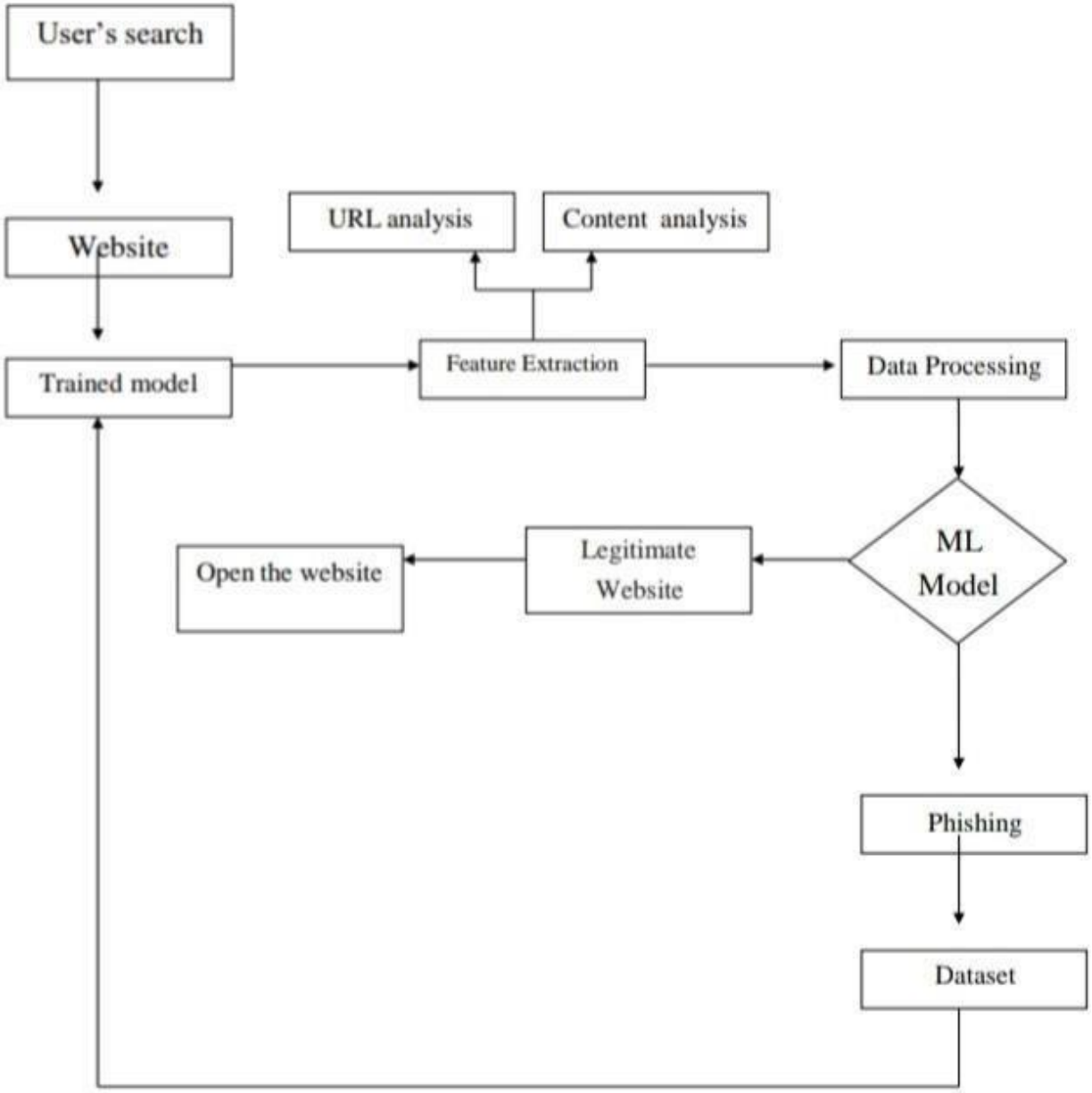
FRNo.	Non-FunctionalRequirement	Description
NFR-1	**Usability**	The proposed system is efficient and easy toconfigureindetectingthephishingwebsites.
NFR-2	**Security**	Thesystemissecuredasitpreventstheuserfromtheunauthorizedaccessandpreventsthemfrom thethreat.
NFR-3	**Reliability**	The system will perform the tasks it was supposedtodo.
NFR-4	**Performance**	Thesystemwillperformthetasksefficientlyandwitha goodaccuracyrate.
NFR-5	**Availability**	The proposed system is always available wheneveritisrequiredto beexecuted.
NFR-6	**Scalability**	Thesystemisscalableto handletheincreasinganddecreasingworkloads.

CHAPTER5

PROJECT DESIGN

DataFlowDiagrams:

ADataFlowDiagram(DFD)isatraditionalvisualrepresentationoftheinformation flow within a system. A neat and clear DFD can depict the right amount ofthe system requirement graphically. It shows how data enters and leaves the system,what charges the information, and where data is stored. A data flow diagram (DFD)maps out the flow of information for any process or system. It uses defined symbols likerectangles, circles and arrows, plus short text labels, to show data inputs,outputs,storage points and the routes between each destination. Data flowcharts can range fromsimple, even hand-drawn process overviews, to in-depth, multi-level DFDs that digprogressively deeper into how the data is handled. They can be used to analyze anexisting system or model a new one. Like all the best diagrams and charts, a DFD canoften visually “say” things that would be hard to explain in words, and they work for bothtechnicalandnontechnicalaudiences,fromdevelopertoCEO.


User's search
Website
URL analysis
Content analysis
Trained model
Feature Extraction
Data Processing
ML Model
Legitimate Website
Open the website
Phishing
Dataset

User Stories:

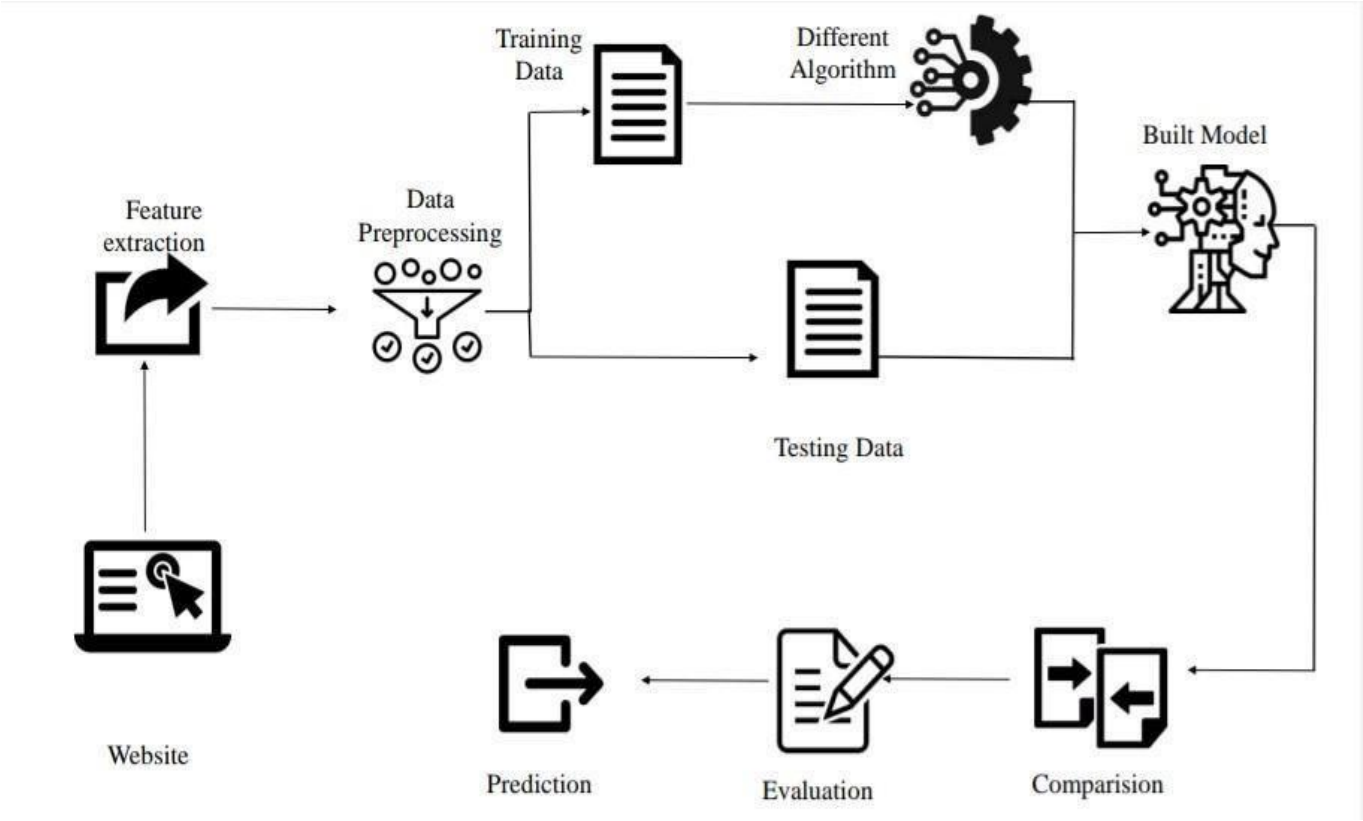
User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
General Users	Home Page	USN-1	The user finds the home page easy to navigateand feels comfortable with the user interface	The every user can access and visit the Homepage	High	Sprint1
Client user(Customer)	Sign up	USN-2	The user will get authentication over their account by security measures	The client customer would register for the application using their email and setting up their own passwords for accessing the application.	High	Sprint 2
		USN-3	The user will be able to authorise their accountonly if they remember their authentication key (biometrics if suitable hardware available,password)	The client user can access the application if their registration is confirmed by verifying with their email.	Medium	Sprint 2
	Login	USN-4	The user can register using either their google account or their mobile number	The client user can login after getting registered	Medium	Sprint 2
	Dashboard	USN-5	The user can go through the facilities providedby the product	The user can view their profile and status in their dashboard	Low	Sprint 3
Administrator	Prediction	USN-6	User would be prompted with a pop up indicating the trustfulness of the website	The user can accurately forecast about the algorithms used.	High	Sprint 3
	Results page	USN-7	The user would be able to analyse websitewhether it's genuine or not	The results of the website on whether the web pages is genuine or not.	Medium	Sprint 4
Customer Care Executive	Reporting	USN-8	The user can report for any bugs or ask any queries on the product	Bugs or queries can be enquired by the user.	High	Sprint 4

SOLUTIONANDTECHNICALARCHITECTURE:SO

LUTIONARCHITECTURE:

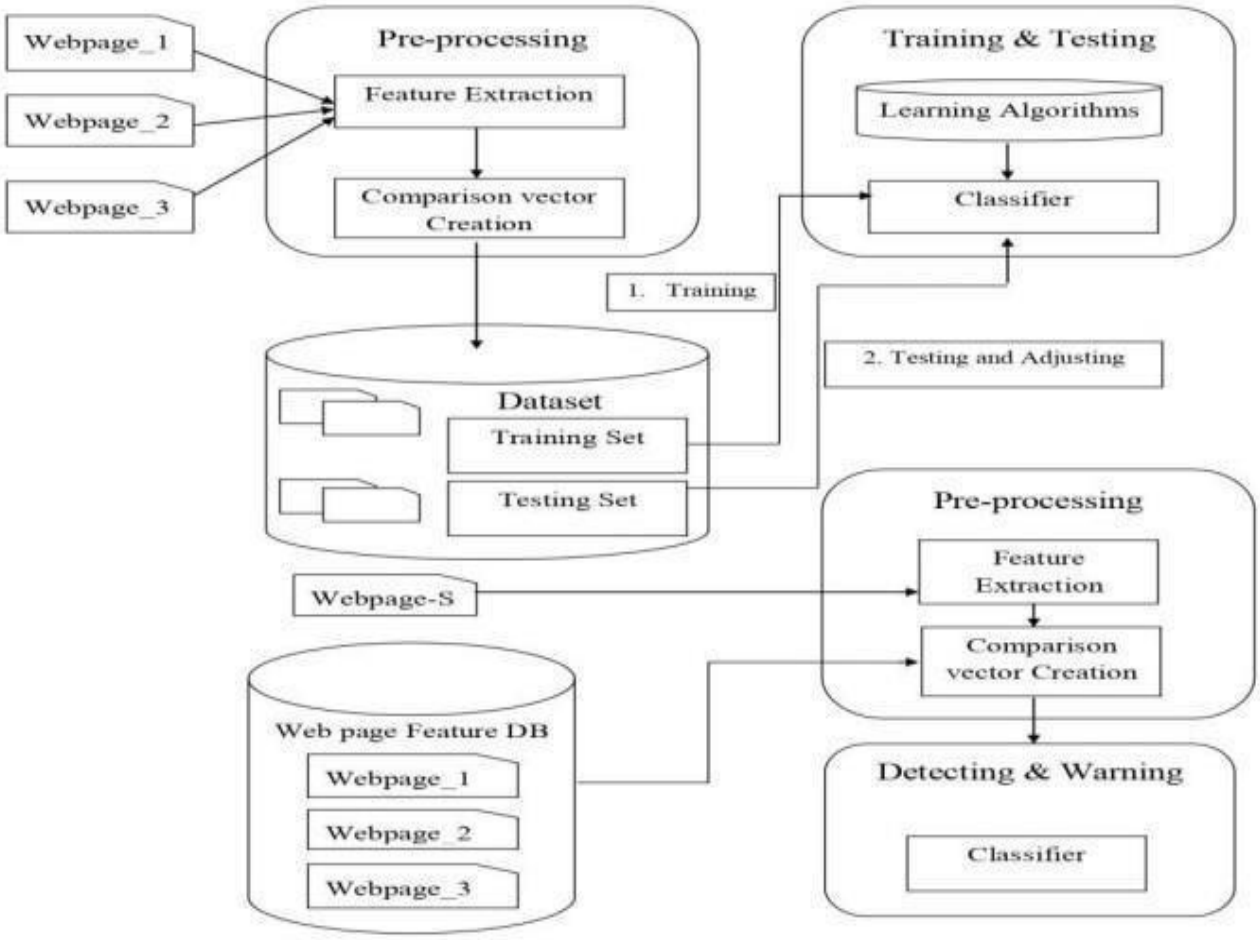
Solution architecture is a complex process – with many sub-processes – that bridges thegapbetweenbusinessproblems andtechnologysolutions.Itsgoalsareto:

- Findthebesttechsolutiontosolveexistingbusinessproblems.
- Describethestructure,characteristics,behavior,andotheraspectsofthesoftwaretoprojectstakeholders.
- Definefeatures,developmentphases,andsolutionrequirements.
- Providespecificationsaccordingtowhichthesolutionisdefined,managed,anddelivered.



TECHNICAL ARCHITECTURE:

Technology architecture deals with the deployment of application componentson technology components. A standard set of predefined technology components isprovidedinordertorepresentservers,network,workstations,andsoon.



USERSTORIES:

Table-1:Components&Technologies:

S.No	Component	Description	Technology
1.	User Interface	How user interacts with application e.g. Web UI, Mobile App, Chatbot etc.	HTML, CSS, JavaScript
2.	Application Logic	Logic for a process in the application	Flask (Python)
3.	Database	Data Type, Configurations etc.	MySQL
4.	Cloud Database	Database Service on Cloud	IBM Watson.
5.	File Storage	File storage requirements	IBM Block Storage ,MongoDB
6.	Machine Learning Model	Purpose of Machine Learning Model	Decision tree algorithm
7.	Infrastructure (Server / Cloud)	Application Deployment on Local System / Cloud Local Server Configuration: Cloud Server Configuration :	Local, IBM Cloud

Table-2:ApplicationCharacteristics:

S.No	Characteristics	Description	Technology
1.	Open-Source Frameworks	The package Sckit Learn in Python is used to handle Machine Learning Algorithms	Machine Learning
2.	Security Implementations	Typosquatting, Cybersquatting	Cyber security
3.	Scalable Architecture	The system will be able to detect maximum of the recently updated phishing websites and is highly scalable to use.	Technology used
4.	Availability	The system is always available whenever it is required to be executed by balancing the load traffic among the servers.	IBM Cloud Load Balancers
5.	Performance	The system would have efficiency and good accuracy rate in detecting the phishing websites.	Machine Learning algorithm(Decision tree algorithm)

CHAPTER 6

PROJECTPLANNINGANDSCHEDULING

6.1SPRINTPLANNINGANDESTIMATION:

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Home Page	USN-1	The user finds the home page easy to navigate and feels comfortable with the user interface	10	High	G. Srivaths Karthic
Sprint-2	Sign up	USN-2	The user will get authentication over their account by security measures	10	High	G. Srivaths Karthic S. Naveenkumar
Sprint-2		USN-3	The user will be able to authorise their account only if they remember their authentication key (biometrics if suitable hardware available, password)		Medium	S. Naveenkumar M. Ajaykumar J. Guna Seakar
Sprint-2	Login	USN-4	The user can register using either their google account or their mobile number	10	Medium	G. Srivaths Karthic
Sprint-3	Dashboard	USN-5	The user can go through the facilities provided by the product	5	Low	S. Naveenkumar
Sprint-3	Prediction	USN-6	User would be prompted with a pop up indicating the trustfulness of the website	15	High	M. Ajaykumar J. Guna Seaker
Sprint-4	Results page	USN-7	The user would be able to analyse website whether it's genuine or not	5	Medium	G. Srivaths Karthic S. Naveenkumar
Sprint-4	Reporting	USN-8	The user can report for any bugs or ask any queries on the product	15	High	S. Naveenkumar

ProjectTracker,Velocity&BurndownChart:

S.No	Characteristics	Description	Technology
1.	Open-Source Frameworks	The package Sckit Learn in Python is used to handle Machine Learning Algorithms	Machine Learning
2.	Security Implementations	Typosquatting, Cybersquatting	Cyber security
3.	Scalable Architecture	The system will be able to detect maximum of the recently updated phishing websites and is highly scalable to use.	Technology used
4.	Availability	The system is always available whenever it is required to be executed by balancing the load traffic among the servers.	IBM Cloud Load Balancers
5.	Performance	The system would have efficiency and good accuracy rate in detecting the phishing websites.	Machine Learning algorithm(Decision tree algorithm)

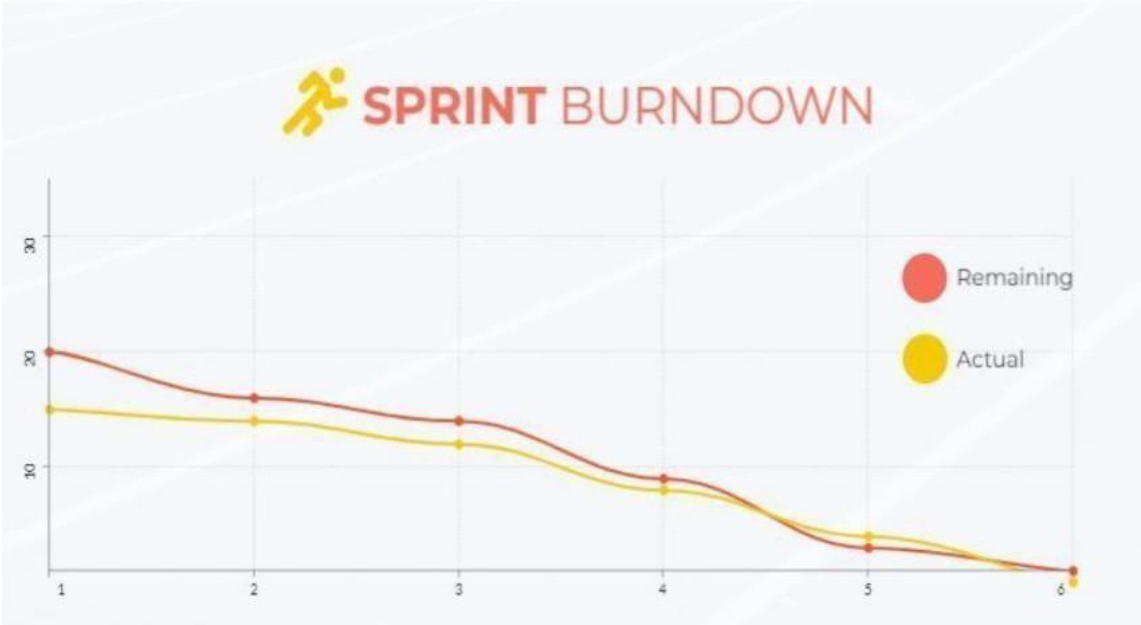
Velocity:

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

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

BurndownChart:

A burn down chart is a graphical representation of work left to doversus time. Itisoften used in agile software development methodologies suchasScrum. However, burn downchartscanbeappliedtoanyprojectcontainingmeasurableprogressover time.



CHAPTER 7

CONCLUSION

Internet is something that could not be eliminated from the daily routine. So having so muchimportance in it, we must also ensure the safety of ourselves while using it. So the model developedby us would certainly being a great change in this society. We also make use of large data set to trainthismodel.Sobydoingthismodelisrobustenoughtodetectthephishingwebsite.

So along with this reporting facility is provided to help users to address their issues whilebrowsing. So this tool is highly purposeful to be used by all internet users. This enhances the dataintegrityandbrowsingexperience.

CHAPTER 8

FUTURESCOPE

In near future we have also planned to update the tool with phishing techniques so that it facilitates the detection process. It also improvises the accuracy of detection.

SOURCECODE

```python
from flask import Flask;
from flask import render_template, request, redirect;
import pymysql
import pickle
import numpy as np
import streamlit as st

app = Flask(__name__)

model = pickle.load(open("ml.pkl","rb"))


def sql_connector():
    conn = pymysql.connect(user="root", password="root12345", db="webphishingdetection", host="localhost")
    c = conn.cursor()
    return conn, c


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

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

@app.route("/h1")
def h1():
    return render_template('h1.html')

@app.route("/predict",methods=["Post","GET"])
def predict():
```

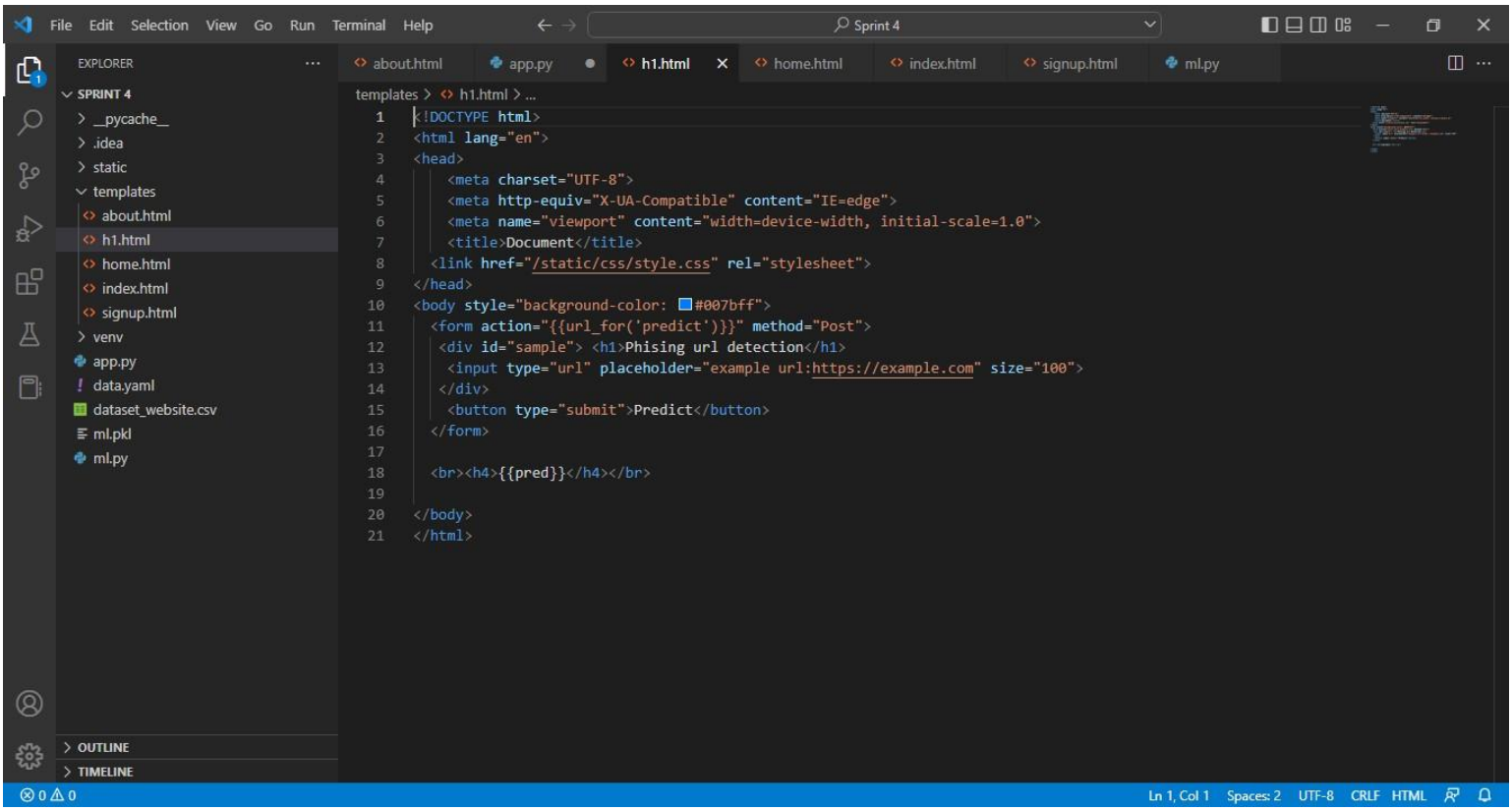
```html

<!doctype html>
<html lang="en">
  <head>
    <meta charset="utf-8">
    <meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">
    <meta name="description" content="">
    <meta name="author" content="Mark Otto, Jacob Thornton, and Bootstrap contributors">
    <meta name="generator" content="Jekyll v3.8.5">
    <title>WEB PHISHING DETECTION</title>



    <!-- Bootstrap core CSS -->
<link href="/static/css/bootstrap.min.css" rel="stylesheet" integrity="sha384-ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/j6cY/iJTQUOhcWr7x9JvoRxT2MZw1T" c


    <style>
      .bd-placeholder-img {
        font-size: 1.125rem;
        text-anchor: middle;
        -webkit-user-select: none;
        -moz-user-select: none;
        -ms-user-select: none;
        user-select: none;
      }

      @media (min-width: 768px) {
        .bd-placeholder-img-lg {
          font-size: 3.5rem;
        }
      }
```

```html
<!doctype html>
<html lang="en">
  <head>
    <meta charset="utf-8">
    <meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">
    <meta name="description" content="">
    <meta name="author" content="Mark Otto, Jacob Thornton, and Bootstrap contributors">
    <meta name="generator" content="Jekyll v3.8.5">
    <title>WEB PHISHING DETECTION</title>



    <!-- Bootstrap core CSS -->
<link href="/static/css/bootstrap.min.css" rel="stylesheet" integrity="sha384-ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/


    <style>
      .bd-placeholder-img {
        font-size: 1.125rem;
        text-anchor: middle;
        -webkit-user-select: none;
        -moz-user-select: none;
        -ms-user-select: none;
        user-select: none;
      }

      @media (min-width: 768px) {
        .bd-placeholder-img-lg {
          font-size: 3.5rem;
        }
      }
```

```html
<!doctype html>
<html lang="en">
  <head>
    <meta charset="utf-8">
    <meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">
    <meta name="description" content="">
    <meta name="author" content="Mark Otto, Jacob Thornton, and Bootstrap contributors">
    <meta name="generator" content="Jekyll v3.8.5">
    <title>Signin</title>

    <!-- Bootstrap core CSS -->
<link href="/static/css/bootstrap.min.css" rel="stylesheet" integrity="sha384-ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/


    <style>
      .bd-placeholder-img {
        font-size: 1.125rem;
        text-anchor: middle;
        -webkit-user-select: none;
        -moz-user-select: none;
        -ms-user-select: none;
        user-select: none;
      }

      @media (min-width: 768px) {
        .bd-placeholder-img-lg {
          font-size: 3.5rem;
        }
      }
    </style>
    <!-- Custom styles for this template -->
    <link href="/static/css/signin.css" rel="stylesheet">
```

```html
<!doctype html>
<html lang="en">
  <head>
    <meta charset="utf-8">
    <meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">
    <meta name="description" content="">
    <meta name="author" content="Mark Otto, Jacob Thornton, and Bootstrap contributors">
    <meta name="generator" content="Jekyll v3.8.5">
    <title>Signup</title>

    <!-- Bootstrap core CSS -->
<link href="/static/css/bootstrap.min.css" rel="stylesheet" integrity="sha384-ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/


    <style>
      .bd-placeholder-img {
        font-size: 1.125rem;
        text-anchor: middle;
        -webkit-user-select: none;
        -moz-user-select: none;
        -ms-user-select: none;
        user-select: none;
      }

      @media (min-width: 768px) {
        .bd-placeholder-img-lg {
          font-size: 3.5rem;
        }
      }
    </style>
    <!-- Custom styles for this template -->
    <link href="/static/css/signin.css" rel="stylesheet">
```

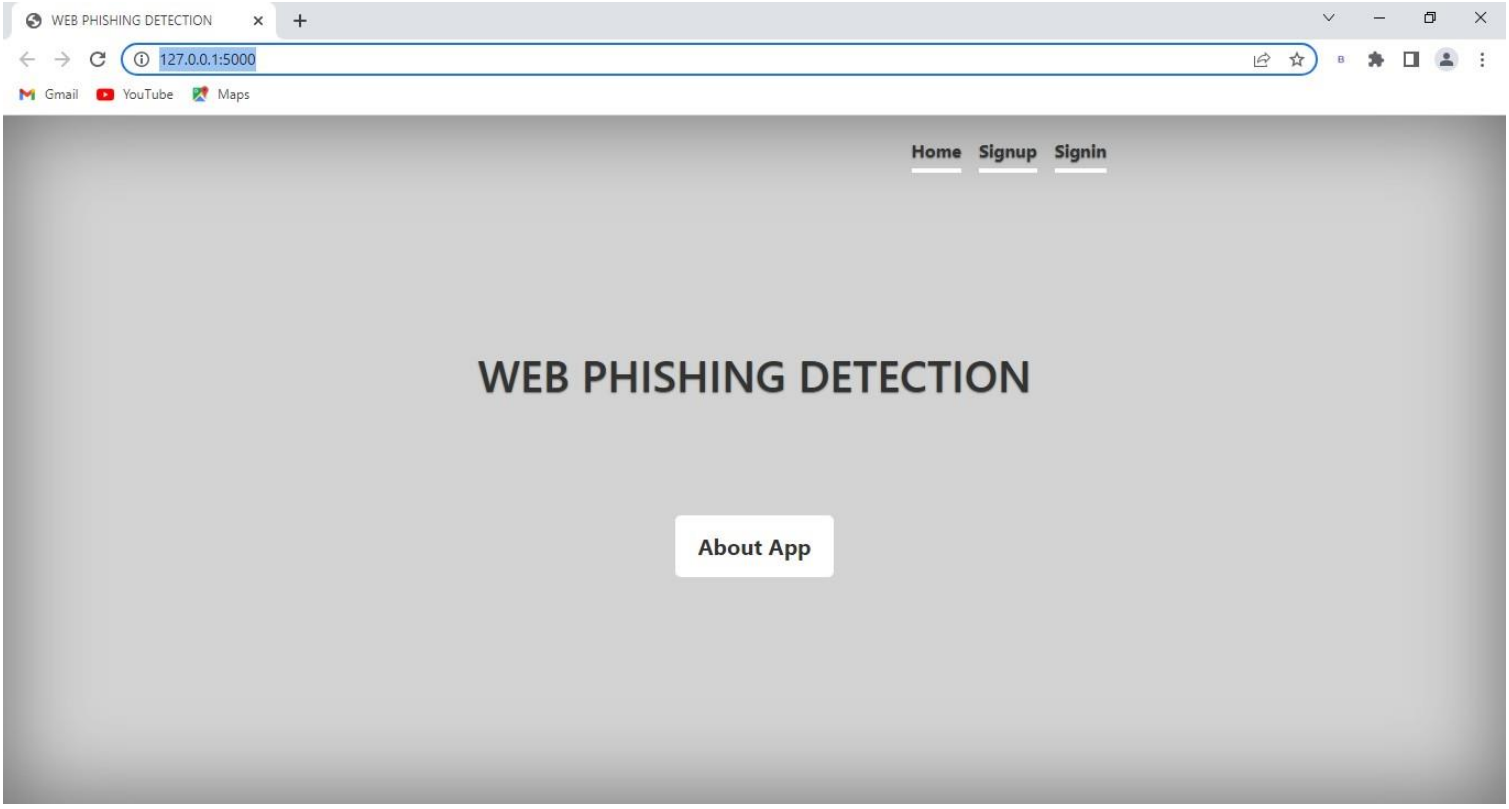
```python
import pickle
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics
from sklearn.metrics import confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.tree import export_graphviz

df = pd.read_csv("dataset_website.csv")
print(df.head())

df = np.array(df)

X = df[:, :-1]
y = df[:, -1]

y = y.astype('int')
X = X.astype('int')


Xtrain, Xtest, ytrain, ytest = train_test_split(X, y,test_size=2, random_state=0)

model = DecisionTreeClassifier()
model.fit(Xtrain, ytrain)
x = model.predict(Xtest)
y = model.predict(Xtrain)

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

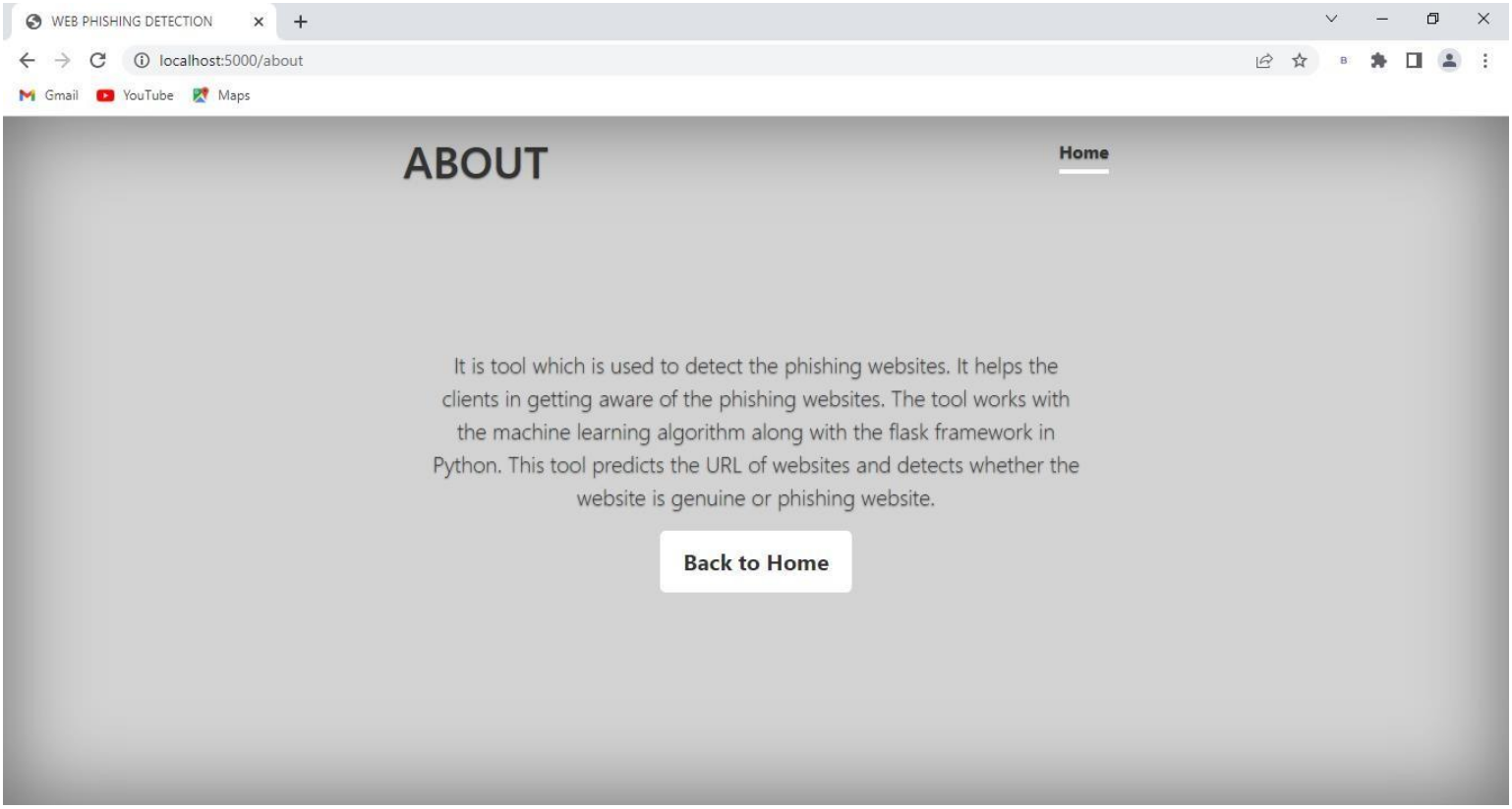

File Edit Selection View Go Run Terminal Help
Sprint 4
EXPLORER
SPRINT 4
__pycache__
.idea
static
templates
about.html
h1.html
home.html
index.html
signup.html
venv
app.py
data.yaml
dataset_website.csv
ml.pkl
ml.py
OUTLINE
TIMELINE
about.html
app.py
h1.html
home.html
index.html
signup.html
ml.py
templates > h1.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">
<title>Document</title>
<link href="/static/css/style.css" rel="stylesheet">
</head>
<body style="background-color: #007bff">
<form action="{{url_for('predict')}}" method="Post">
 <h1>Phising url detection</h1>
<input type="url" placeholder="example url:https://example.com" size="100">

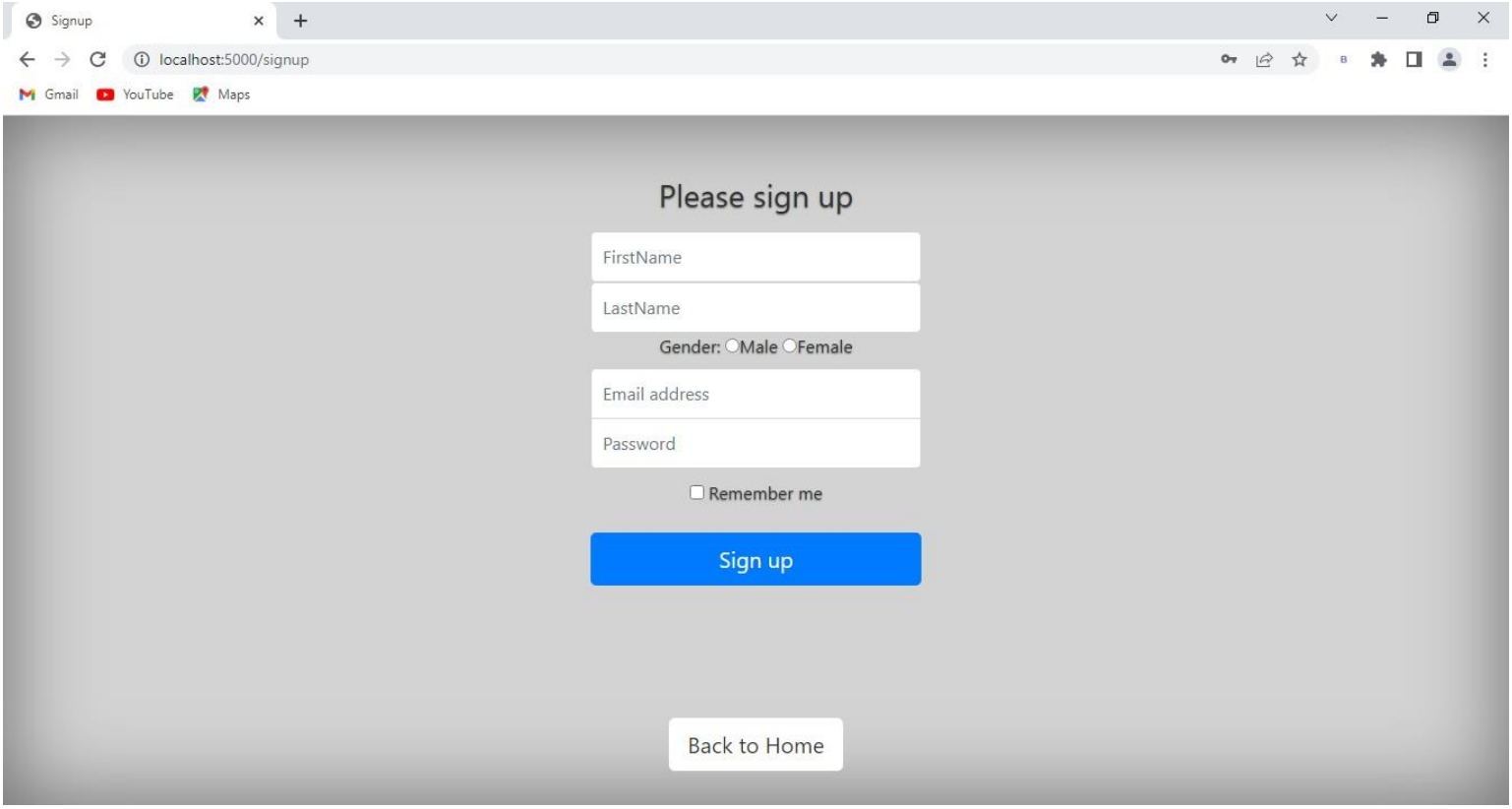
<button type="submit">Predict</button>
</form>

<h4>{{pred}}</h4></br>
</body>
</html>
Ln 1, Col 1 Spaces: 2 UTF-8 CRLF HTML

OUTPUT


WEB PHISHING DETECTION
127.0.0.1:5000
Gmail
YouTube
Maps
Home
Signup
Signin
WEB PHISHING DETECTION
About App





GITHUB LINK –https://github.com/IBM-EPBL/IBM-Project-42668-1660703736DEMOLINK – https://drive.google.com/file/d/1EynOE7Tfe1rGl8-QPzNKlUYN_GTRqm0p/view?usp=share_link