WEB PHISHING DETECTION
PROJECT REPORT

Submitted By

HARINI B	210819205011
JAYASRI S	210819205017
HARINI D	210819205012
SAFANA PARVEEN J	210819205041

In partial fulfilment for the award of the degree

Of

BACHELOR OF TECHNOLOGY

In

INFORMATION TECHNOLOGY

KINGS ENGINEERING COLLEGE,IRUNGATTUKOTAI

ANNA UNIVERSITY:CHENNAI 600025

BONAFIDE CERTIFICATE

Certified that project report"Web Phishing Detection"is bonafide work of"HARINI B,JAYASRI S,HARINI D,SAFANA PARVEEN J"who carried out thisproject work under my supervision.

SIGNATURE

Dr.G.MANIKANDAN

HEAD OF THE DEPARTMENT

Professor

Dept of Information Technology,
Kings Engineering College,
Irungattukottai,
Chennai-602 117

SIGNATURE

Dr.J Briso Becky Bell

SUPERVISOR

Assistant Professor

Dept of Information Technology
Kings Engineering College,
Irungattukottai,
Chennai-602 117.

ACKNOWLEDGEMENT

We thank God for his blessings and also for giving as good knowledge and strength in enabling us to finish our project.Our deep gratitude goes to our founder late **Dr.D.SELVARAJ,M.A.,M.Phil.,**for his patronage in the completion of our project.We like to take this opportunity to thank our honourable chairperson **Dr.S.NALINI SELVARAJ,M.COM.,MPhil.,Ph.D.**and honourable director,**MR.S.AMIRTHARAJ,M.Tech.,M.B.A** for their support given to us to finish our project successfully.We wish to express our sincere thanks to our beloved principal.**Dr.T.JOHN ORAL BASKAR.,M.E.,Ph.D** for his kind encouragement and his interest towards us.

We are extremely grateful and thanks to our professor **Dr.G.MANIKANDAN,**head of Information Technology,Kings Engineering College,for his valuable suggestion,guidance and encouragement.We wish to express our sense of gratitude to our project supervisor **Dr.J Briso Becky Bell,**Assistant Professor of Information Technology Department,Kings Engineering College and project evaluator **Dr.M Robinson Joel**,Associate Professor of Information Technology Department,Kings Engineering College whose idea and direction made our project a grand success.We express our sincere thanks to our parents,friends and staff members who have helped and encouraged us during the entire course of completing this project work successfully.

TABLE OF THE CONTENT

CHAPTER 1

INTRODUCTION:

13.1 :PROJECT OVERVIEW

The terms"heart disease"and"cardiovascular disease"are frequently used interchangeably.Heart disease is a general term that covers a wide range of heart related medical conditions.The irregular health state that directly affects the heart and all of its components is characterized by these medical conditions.In order to forecast cardiac disease,this study discusses various data mining,big data,and machine learning techniques.Building an important model for the medical system to forecast heart disease or cardiovascular illness requires the use of data mining and machine learning.Our application helps the user in finding out if they have heart disease or not.They can find out by entering details such as their heart rate,cholesterol,blood pressure etc.A dashboard is also attached along with the results for better understanding where they can compare their blood pressure and similar metrics with other users.This project focuses on Random Forest Classifier.The accuracy of our project is 87%for which is better than most other systems in terms of achieving accuracy quickly.

13.2 :PURPOSE

This project's goal is to determine,depending on the patient's medical characteristics—such as gender,age,chest pain,fasting blood sugar level,etc.—whether they are likely to be diagnosed with any cardiovascular heart illnesses.The leading cause of death in the developed world is heart disease.Heart disease cases are rising quickly every day,thus it's crucial and worrisome to predict any potential illnesses in advance.This diagnosis is a challenging task that requires accuracy and efficiency.Therefore,there needs to be work done to help prevent the risks of having a heart attack or stroke.It is the main factor in adult deaths.By using a person's medical history,our initiative those who are most likely to be diagnosed with a cardiac condition.It can assist in identifying disease with less medical tests and effective therapies,so that patients can be treated appropriately.It can identify anyone who is experiencing any heart disease symptoms,such as chest pain or high blood pressure.Around the world,machine learning is applied in many different fields.There is no exception in the healthcare sector.Machine learning may be crucial in determining whether locomotor disorders,heart illnesses,and other conditions are present or absent.If foreseen well in advance,such information can offer valuable insights to doctors,who can then customise their diagnosis and course of care for each patient.

The EHDPS predicts the likelihood of patients getting heart disease.It enables significant knowledge,eg,relationships between medical factors related to heart disease and patterns,to be established.We have employed the multilayer perceptron neural network with backpropagation as the training algorithm

.

CHAPTER 2

LITERATURE SURVEY

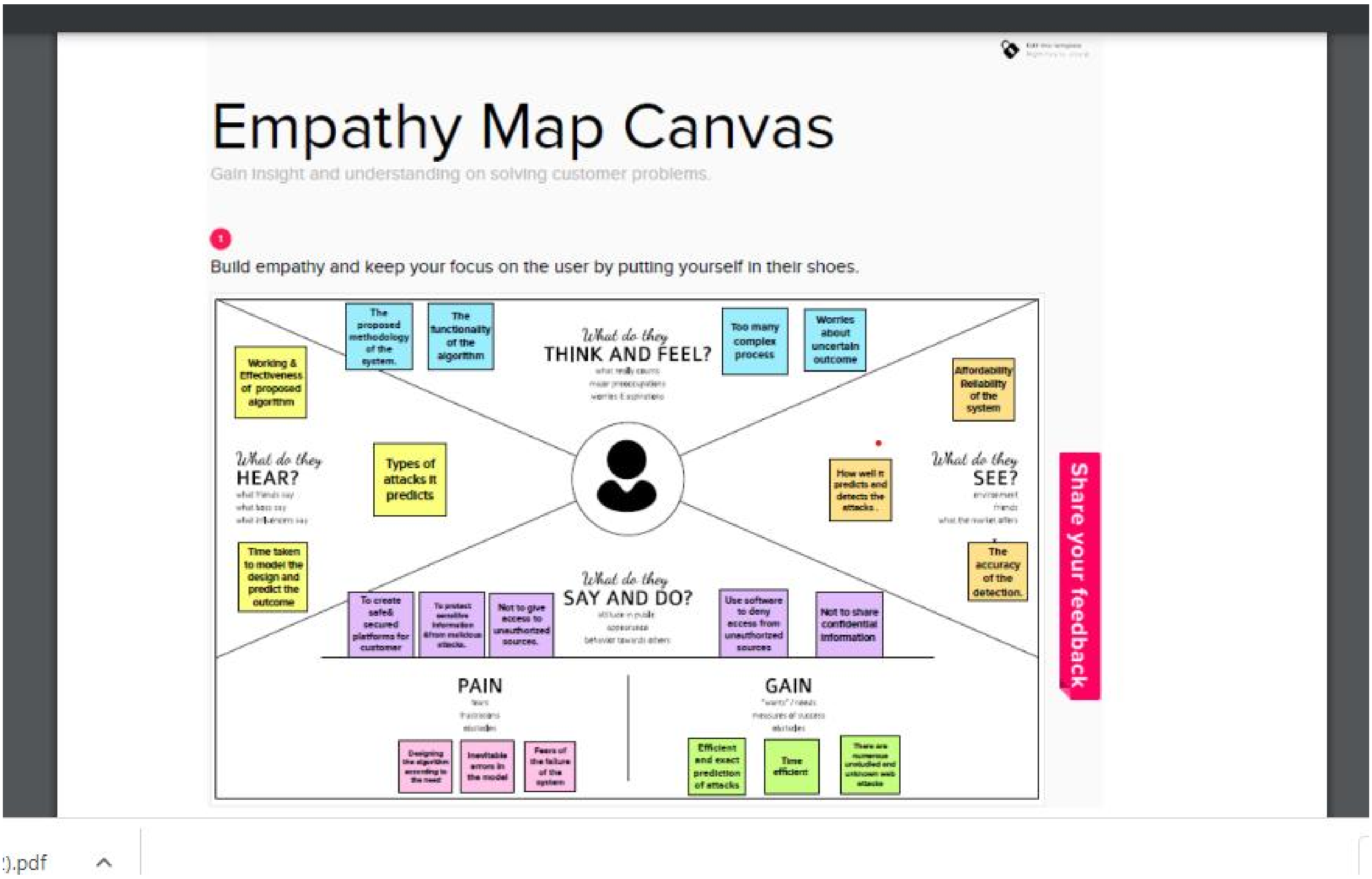
2.1 EXISTING PROBLEM

Phishing is a widespread method of tricking unsuspecting people into disclosing personal information by using fake websites.Phishing website URLs are designed to steal personal information such as user names,passwords,and online banking activities.Phishers employ webpages that are visually and semantically identical to legitimate websites.As technology advances,phishing strategies have become more sophisticated,necessitating the use of anti-phishing measures to identify phishing.Machine learning is an effective method for combating phishing assaults.This study examines the features utilised in detection as well as machine learning-based detection approaches.Phishing is popular among attackers because it is easier to persuade someone to click on a malicious link that appears to be legitimate than it is to break through a computer\'s protection measures.The malicious links in the message body are made to look like they go to the faked organisation by utilising the spoofed organization\'s logos and other valid material.We\'ll go through the characteristics of phishing domains(also known as fraudulent domains),the qualities that distinguish them from real domains,why it\'s crucial to detect them,and how they can be discovered using machine learning and natural language processing technique

CHAPTER 3

IDEATION&PROPOSED SOLUTION

3.1 EMPATHY MAP CANVAS



Empathy Map Canvas

Gain insight and understanding on solving customer problems.

1

Build empathy and keep your focus on the user by putting yourself in their shoes.

What do they THINK AND FEEL?

what really counts
major preoccupations
worries & aspirations

- The proposed methodology of the system.
- The functionality of the algorithm
- Too many complex process
- Worries about uncertain outcome

What do they HEAR?

what friends say
what boss say
what influencers say

- Working & Effectiveness of proposed algorithm
- Types of attacks it predicts
- Time taken to model the design and predict the outcome

What do they SEE?

environment
friends
what the market offers

- Affordability Reliability of the system
- How well it predicts and detects the attacks .
- The accuracy of the detection.

What do they SAY AND DO?

attitude in public
appearance
behavior towards others

- To create safe& secured platforms for customer
- To protect sensitive information &from malicious attacks.
- Not to give access to unauthorized sources.
- Use software to deny access from unauthorized sources
- Not to share confidential information

PAIN

fears
frustrations
obstacles

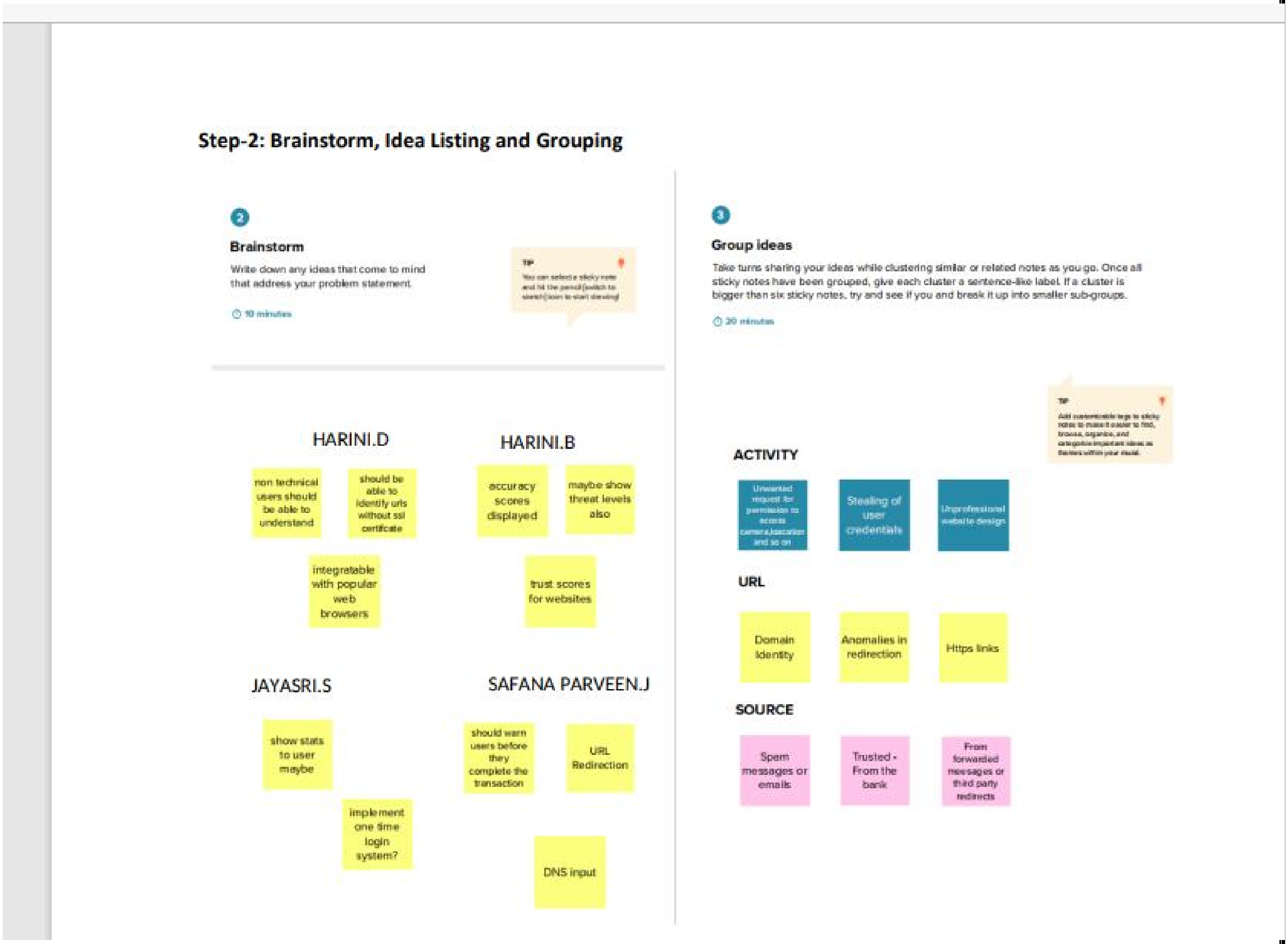
- Designing the algorithm according to the need
- Inevitable errors in the model
- Fears of the failure of the system

GAIN

"wants" / needs
measures of success
obstacles

- Efficient and exact prediction of attacks
- Time efficient
- There are numerous unstudied and unknown web attacks

3.2 IDEATION&BRAINSTORMING



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!

HARINI.D

non technical users should be able to understand

should be able to identify urls without ssl certificate

integratable with popular web browsers

HARINI.B

accuracy scores displayed

maybe show threat levels also

trust scores for websites

JAYASRI.S

show stats to user maybe

implement one time login system?

SAFANA PARVEEN.J

should warn users before they complete the transaction

URL Redirection

DNS input

3

Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. Once all sticky notes have been grouped, 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.

ACTIVITY

Unwanted request for permission to access camera,location and so on

Stealing of user credentials

Unprofessional website design

URL

Domain Identity

Anomalies in redirection

Https links

SOURCE

Spam messages or emails

Trusted - From the bank

From forwarded messages or third party redirects

PROPOSED SOLUTION

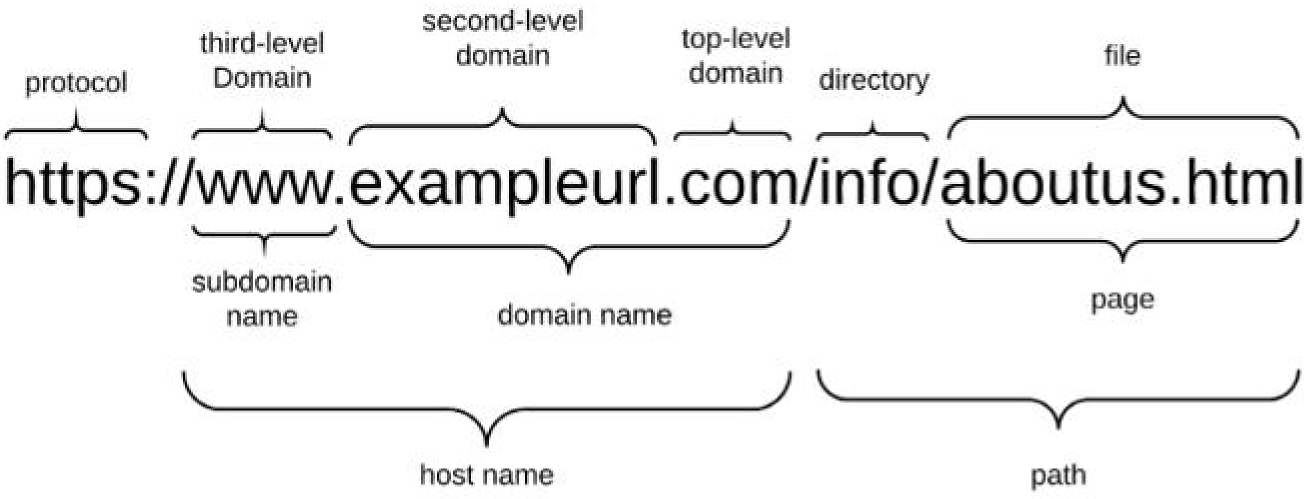
Today's growing phishing websites pose significant threats due to their extremely undetectable risk.They anticipate internet users to mistake them as genuine ones in order to reveal user information and privacy,such as login ids,pass-words,credit card numbers,etc.without notice.This paper proposes a new approach to solve the anti-phishing problem.The new features of this approach can be represented by URL character sequence without phishing prior knowledge,various hyperlink information,and textual content of the webpage,which are combined and fed to train the XGBoost classifier.One of the major contributions of this paper is the selection of different new features,which are capable enough to detect 0-h attacks,and these features do not depend on any third-party services.In particular,we extract character level Term Frequency-Inverse Document Frequency(TF-IDF)features from noisy parts of HTML and plaintext of the given webpage.Moreover,our proposed hyperlink features determine the relationship between the content and the URL of a webpage.Due to the absence of publicly available large phishing data sets,we needed to create our own data set with 60,252 webpages to validate the proposed solution.This data contains 32,972 benign webpages and 27,280 phishing webpages.For evaluations,the performance of each category of the proposed feature set is evaluated,and various classification algorithms are employed.From the empirical results,it was observed that the proposed individual features are valuable for phishing detection.8 However,the integration of all the features improves the detection of phishing sites with significant accuracy.The proposed approach achieved an accuracy of--96.76%with only 1.39%false-positive rate on our dataset,and an accuracy of 98.48%with 2.09%false-positive rate on benchmark dataset,which outperforms the existing baseline approaches.

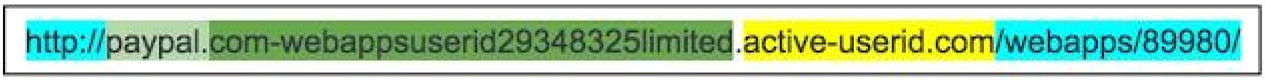
Phishing is a form of fraud in which the attacker tries to learn sensitive information such as login credentials or account information by sending as a reputable entity or person in email or other communication channels.

Typically a victim receives a message that appears to have been sent by a known contact or organization.The message contains malicious software targeting the user’s computer or has links to direct victims to malicious websites in order to trick them into divulging personal and financial information,such as passwords,account IDs or credit card details.

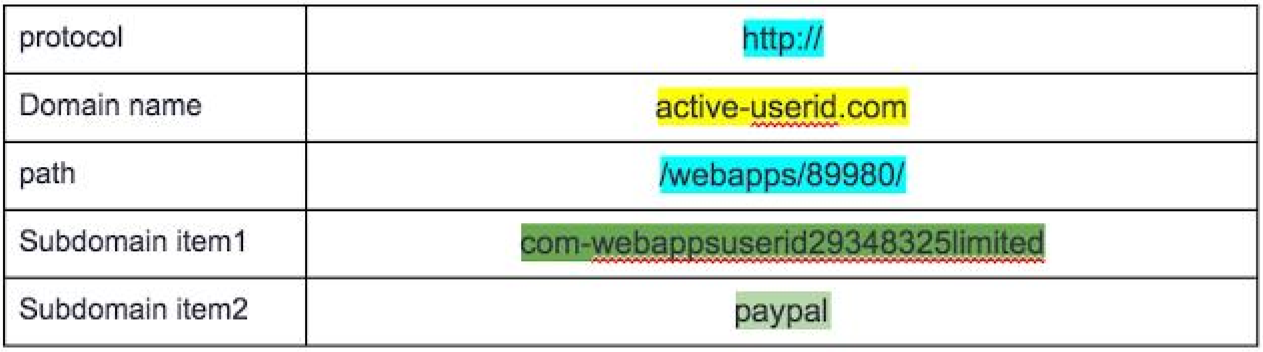
Phishing is popular among attackers,since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer’s defense systems.The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization’s logos and other legitimate contents.

Phishing domain(or Fraudulent Domain)characteristics,the features that distinguish them from legitimate domains,why it is important to detect these domains,and how they can be detected using machine learning and natural language processing techniques.





http://paypal.com-webappsuserid29348325limited.active-userid.com/webapps/89980/



protocol	http://
Domain name	active-userid.com
path	/webapps/89980/
Subdomain item1	com-webappsuserid29348325limited
Subdomain item2	paypal

Although the real domain name is active-userid.com,the attacker tried to make the domain look like paypal.com by adding FreeURL.When users see paypal.com at the beginning of the URL,they can trust the site and connect it,then can share their sensitive information to the this fraudulent site.This is a frequently used method by attackers.

Other methods that are often used by attackers are Cybersquatting and Typosquatting.

Cybersquatting(also known as domain squatting),is registering,trafficking in,or using a domain name with bad faith intent to profit from the goodwill of a trademark belonging to someone else.The cybersquatter may offer selling the domain to a person or company who owns a trademark contained within the name at an inflated price or may use it for fraudulent purposes such as phishing.For example,the name of your company is“abcompany”and you register as abcompany.com.Then phishers can register abcompany.net,abcompany.org,abcompany.biz and they can use it for fraudulent purpose.

Typosquatting,also called URL hijacking,is a form of cybersquatting which relies on mistakes such as typographical errors made by Internet users when inputting a website address into a web browser or based on typographical errors that are hard to notice while quick reading.URLs which are created with

Typosquatting looks like a trusted domain.A user may accidentally enter an incorrect website address or click a link which looks like a trusted domain,and in this way,they may visit an alternative website owned by a phisher.

There are a lot of algorithms and a wide variety of data types for phishing detection in the academic literature and commercial products.A phishing URL and the corresponding page have several features which can be differentiated from a malicious URL.For example;an attacker can register long and confusing domain to hide the actual domain name(Cybersquatting,Typosquatting).In some cases attackers can use direct IP addresses instead of using the domain name.This type of event is out of our scope,but it can be used for the same purpose.Attackers can also use short domain names which are irrelevant to legitimate brand names and don’t have any Free Url addition.But these type of web sites are also out of our scope,because they are more relevant to fraudulent domains instead of phishing domains.

The purpose of Phishing Domain Detection is detecting phishing domain names.Therefore,passive queries related to the domain name,which we want to classify as phishing or not,provide useful information to us.Some useful Domain-Based Features are given below.

- Its domain name or its IP address in blacklists of well-known reputation services?
- How many days passed since the domain was registered?
- Is the registrant name hidden?

Page-Based Features are using information about pages which are calculated reputation ranking services.Some of these features give information about how much reliable a web site is.Some of Page-Based Features are given below.

- Global Pagerank
- Country Pagerank
- Position at the Alexa Top 1 Million Site

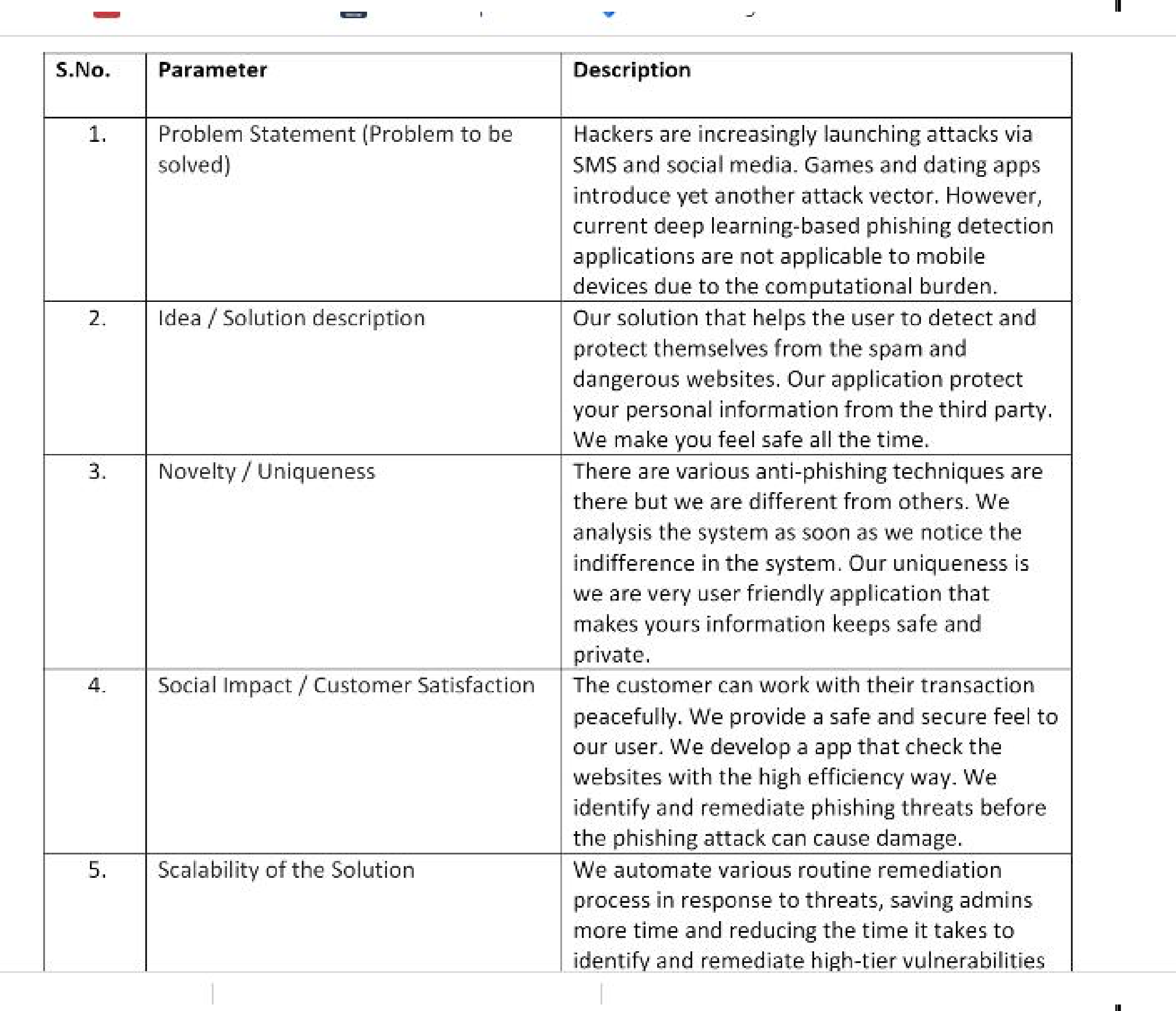
Page-Based Features give us information about user activity on target site.

Some of these features are givenSome below.

- Estimated Number of Visits for the domain on a daily,weekly,or monthly basis
- Average Pageviews per visit
- Average Visit Duration
- Web traffic share per country
- Count of reference from Social Networks to the given domain
- Category of the domain
- Similar websites etc.

3.3 PROBLEM SOLUTION FIT

The Problem-Solution Fit simply means that we have found a problem with our customer and that the solution we have realized for it actually solves the customer’s problem.It helps entrepreneurs,marketers and corporate innovators identify behavioural patterns and recognize what would work and why.The purpose is to solve complex problems in a way that fits the state of your customers and succeed faster and increase your solution adoption by tapping into existing mediums and channels of behaviour



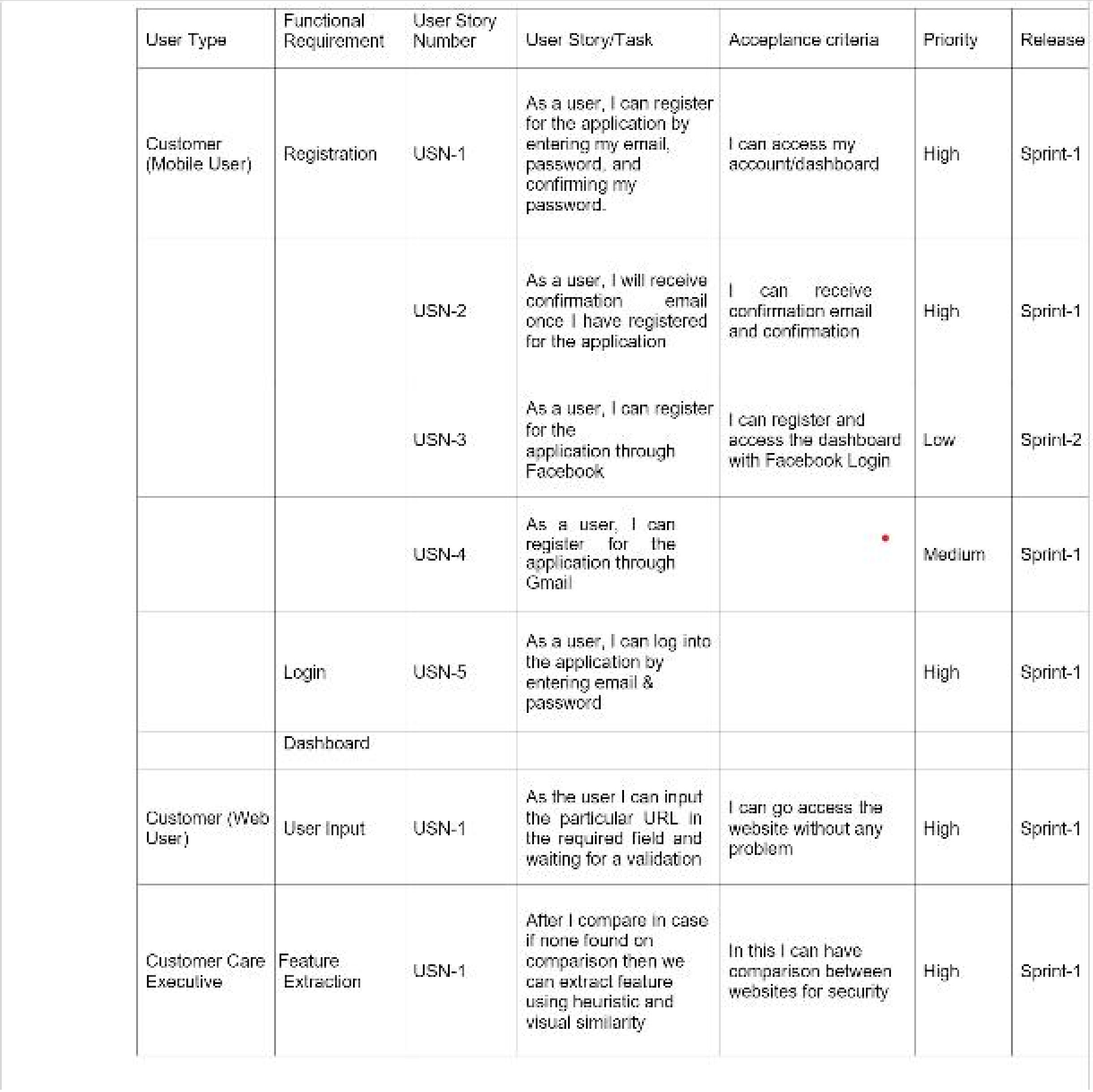
S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Hackers are increasingly launching attacks via SMS and social media. Games and dating apps introduce yet another attack vector. However, current deep learning-based phishing detection applications are not applicable to mobile devices due to the computational burden.
2.	Idea / Solution description	Our solution that helps the user to detect and protect themselves from the spam and dangerous websites. Our application protect your personal information from the third party. We make you feel safe all the time.
3.	Novelty / Uniqueness	There are various anti-phishing techniques are there but we are different from others. We analysis the system as soon as we notice the indifference in the system. Our uniqueness is we are very user friendly application that makes yours information keeps safe and private.
4.	Social Impact / Customer Satisfaction	The customer can work with their transaction peacefully. We provide a safe and secure feel to our user. We develop a app that check the websites with the high efficiency way. We identify and remediate phishing threats before the phishing attack can cause damage.
5.	Scalability of the Solution	We automate various routine remediation process in response to threats, saving admins more time and reducing the time it takes to identify and remediate high-tier vulnerabilities

CHAPTER 4

REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENTS

- Using the IP address
- Long URL to hide the Suspicious Part
- Using URL shortening services TinyURL
- URLs having@symbol
- Redirecting using//
- Adding Prefix or Suffix Separated by(-)to the Domain.



User Type	Functional Requirement	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 confirming my 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 and confirmation	High	Sprint-1
		USN-3	As a user, I can register for the application through Facebook	I can register and 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 the user I can input the particular URL in the required field and waiting for a 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	In this I can have comparison between websites for security	High	Sprint-1

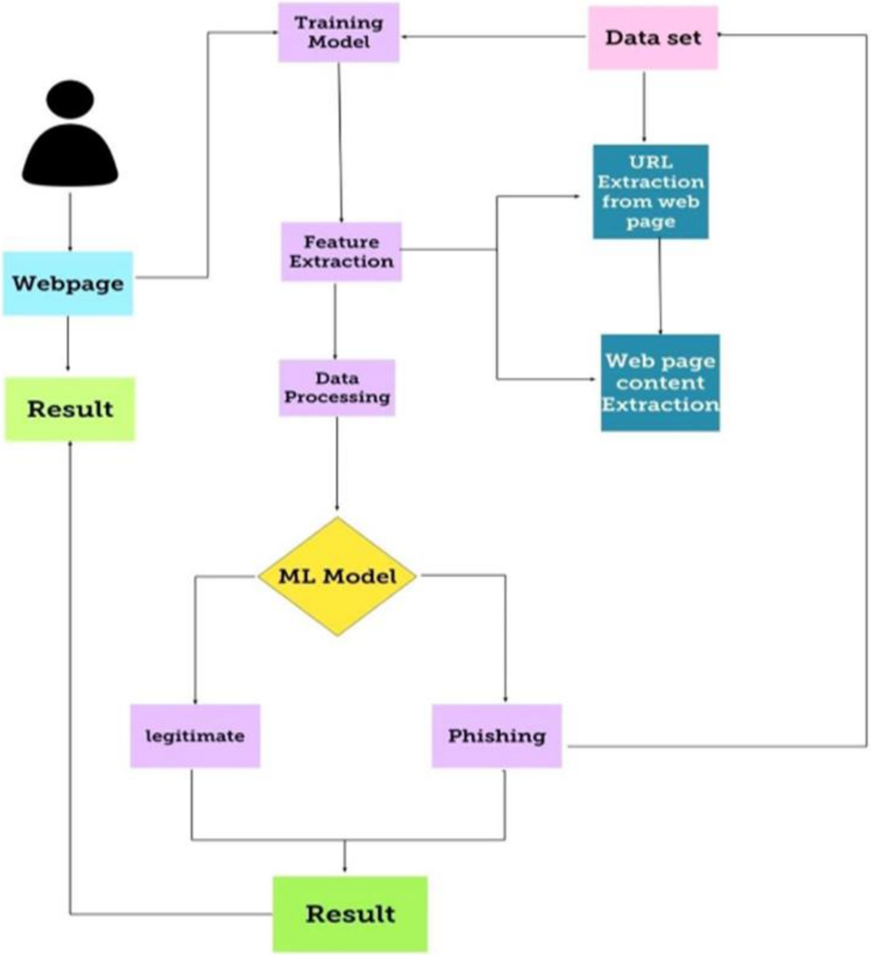
4.2 NON-FUNCTIONAL REQUIREMENTS

- Performance/Response time requirement This IDS photo gallery that we are developing will be used as the main performance system which interacts with members and administrators.
- Therefore,it is expected that the system would perform all the requirements that are specified.
- 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 have error testing to identify invalid username or password and email System should be able to handle large amounts of data System should accommodate high number of photos and users without any fault.
- Safety Requirement It has to be taken into account that at any given time the database may crash due to a virus or operating system failure.
- Therefore,it is required to take the database backup to prevent losing it.Appropriate UPS facility should be installed in case of power supply failure.

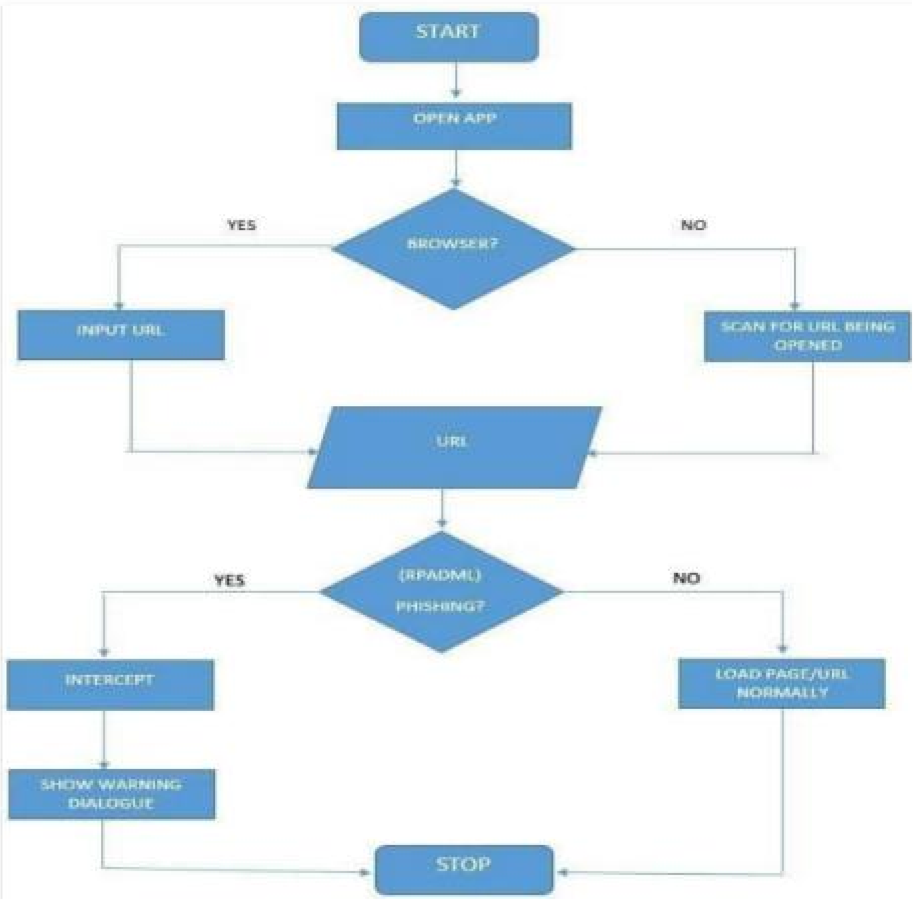
CHAPTER 5

PROJECT DESIGN

5.1 DATA FLOW DIAGRAMS



5.2 SOLUTION&TECHNICAL ARCHITECTURE



5.3 UseStories:

S.No	Component	Description	Technology
1.	Classification of phishing atttacks	Phishing websites are challenging to an organization.	Machine learning
2.	Phishing detection appproaches	Phishing detection schemes which detect phishing on the server side are better than phishing prevention.	Machine learning
3.	Normal dataset	It obtains the browsing habits of users from different sources	Machine Learning
4.	Phishing dataset	Phishtank is a familiar phishing website benchmark dataset	Machine Learning
5.	Testing data	Test data is data which has been specifically identified for use in tests,typically of a computer program.	Machine Learning

6.	Machine learning model	A machine learning model is a file that has been trained to recognize certain types of patterns.You train a model over a set of data,providing it an algorithm that it can use to reason over and learn from those data	Machine Learning.
7.	Improve model performance	Accuracy is one metric for evaluating classification models.Informally,accuracy is the fraction of predictions our model got right.	Machine Learning
8.	Checking accuracy	A data accuracy check,sometimes called a data sanity check,is a set of quality validations that take place before using data.	Machine Learning

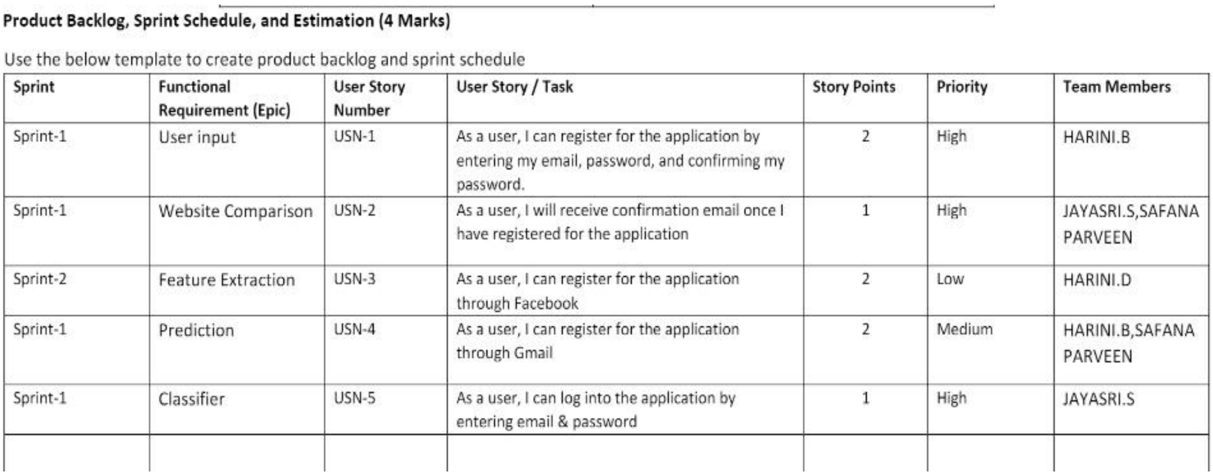
5.4:Application Characteristics:

S.No	Characteristics	Description	Technology
1.	Collection of data	Data collection is the process of gathering,measuring,and analyzing accurate data from a variety of relevant sources to find answers to research problems,answer questions,evaluate outcomes,and forecasttrends and probabilities	Machine Learning
2.	EDA Analysis	Exploratory Data Analysis(EDA)is an approach to analyze the data using visual techniques.It is used to discover trends,patterns,or to check assumptions with thehelp of statistical summary and graphical representations	Technology used
3.	Train&Test split of data	The train-test split is used to estimate the performance of machine learning algorithms that are applicable for prediction-based Algorithms/Applications.This method is a fast and easy procedure to perform such that we can compare our own machine learning model results to machine results.	Technology used

CHAPTER 6

PROJECT PLANNING&SCHEDULING

6.1 SPRINT PLANNING&ESTIMATION



Product Backlog, Sprint Schedule, and Estimation (4 Marks)

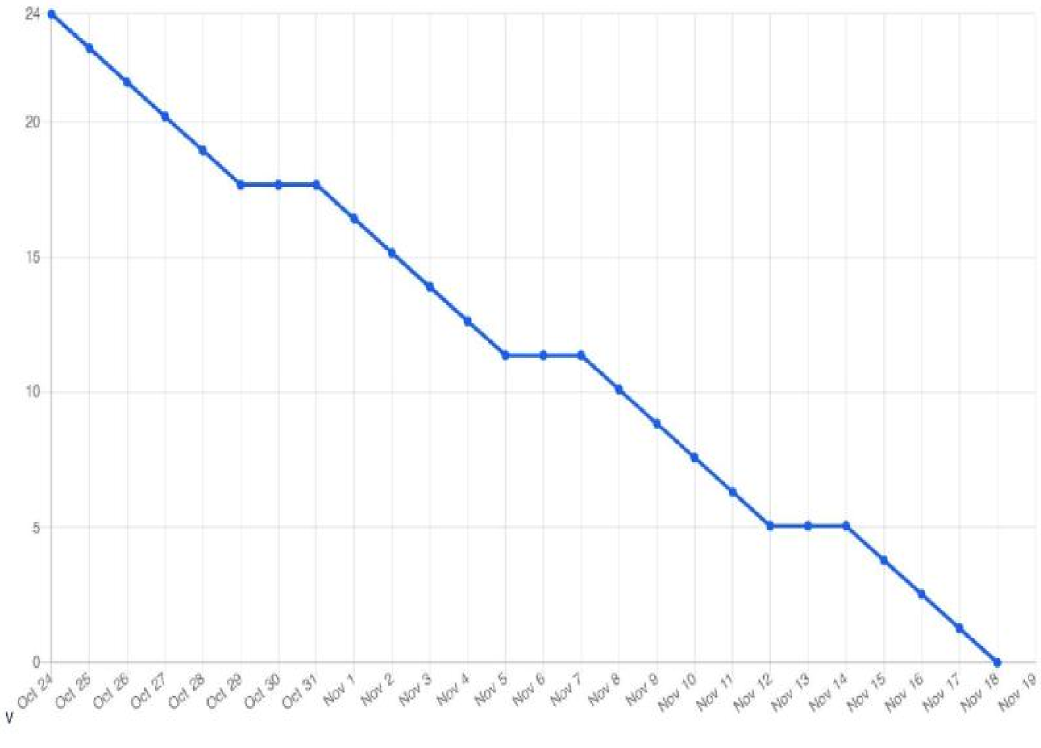
Use the below template to create product backlog and sprint schedule

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	User input	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	2	High	HARINI.B
Sprint-1	Website Comparison	USN-2	As a user, I will receive confirmation email once I have registered for the application	1	High	JAYASRI.S,SAFANA PARVEEN
Sprint-2	Feature Extraction	USN-3	As a user, I can register for the application through Facebook	2	Low	HARINI.D
Sprint-1	Prediction	USN-4	As a user, I can register for the application through Gmail	2	Medium	HARINI.B,SAFANA PARVEEN
Sprint-1	Classifier	USN-5	As a user, I can log into the application by entering email & password	1	High	JAYASRI.S

6.2 SPRINT DELIVERY SCHEDULE

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	1	3 Days	24 Oct 2022	26 Oct 2022	1	26 Oct 2022
Sprint-2	1	3 Days	31 Oct 2022	02 Nov 2022	1	02 Nov 2022
Sprint-3	1	3 Days	07 Nov 2022	09 Nov 2022	1	09 Nov 2022
Sprint-4	1	3 Days	14 Nov 2022	16 Nov 2022	1	16 Nov 2022

6.3 REPORTS FROM JIRA



CHAPTER 7

CODING&SOLUTIONING

7.1 FEATURE 1

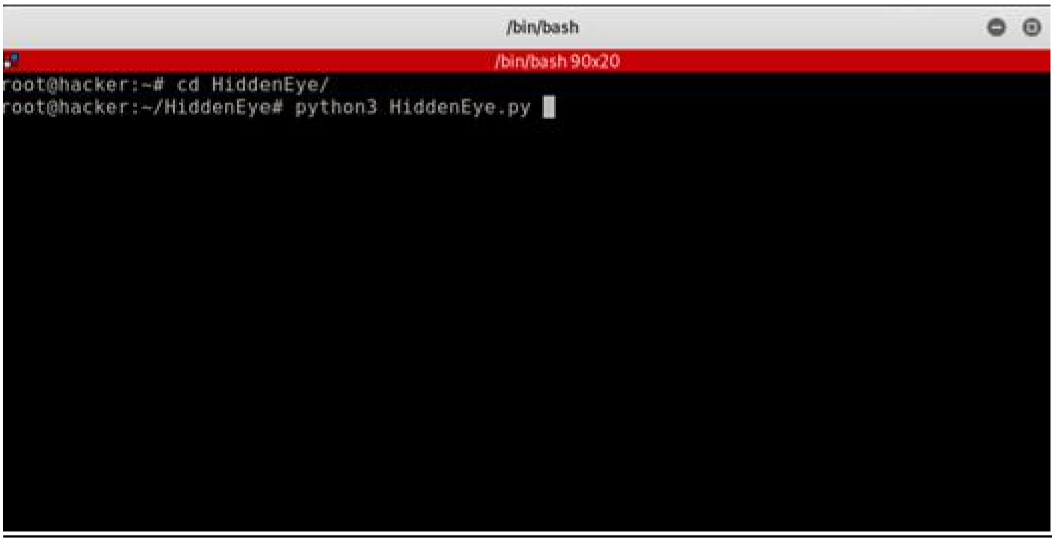
Prediction Model:When applied to a nonlinear data set,the random forest technique performs better than the decision tree.The collection of decision trees known as a random forest was produced by several root nodes.The random forest algorithm can achieve more accuracy quickly and produce expected results.

Phishing is a type of social engineering attack of tricking an individual to enter the sensitive information like usernames,passwords and credit card details.It can be done by any individual with a mere basic requirement of Kali Linux(or any other Linux Distribution).

Steps to create a phishing page:

- Open Kali Linux terminal and paste the following code:

git clonehttps://github.com/DarkSecDevelopers/HiddenEye.git



```
root@hacker:~# cd HiddenEye/
root@hacker:~/HiddenEye# python3 HiddenEye.py
```

7.2 FEATURE 2

1. **Spear Phishing–**

This attack is used to target any specific organization or an individual for unauthorized access.These types of attacks are not initiated by any random hacker,but these attacks are initiated by someone who seeks information related to financial gain or some important information.Just like the phishing attack spear-phishing also comes from a trusted source.This type of attack is much successful.It is considered to be one of the most successful methods as both of the attacks(that is phishing and spear-phishing)is an online attack on users.

2. **Clone Phishing–**

This attack is actually based on copying the email messages that were sent from a trusted source.Now the hackers alter the information by adding a link that redirects the user to a malicious or fake website.Now,this is sent to a large number of users and the person who initiated it watches who clicks on the attachment that was sent as a mail.This spreads through the contacts of the user who has clicked on the attachment.

3. **Catphishing–**

It is a type of social engineering attack that plays with the emotions of a person and exploits them to gain money and information.They target them through dating sites.It is a type of engineering threat.

4. **Voice Phishing–**

Some attacks require to direct the user through fake websites,but some attacks do not require a fake website.This type of attack is sometimes referred to as vishing.Someone who is using the method of vishing,use modern caller id spoofing to convince the victim that the call is from a trusted source.They also use IVR to make it difficult for the legal authorities to trace,block,monitor.It is used to steal credit card numbers or some confidential data of the user.This type of phishing can cause more harm.

5. **SMS phishing–**

These attacks are used to make the user revealing account information.This attack is also similar to the phishing attack used by cybercriminals to steal credit card details or sensitive information,by making it look like it came from a trusted organization.Cybercriminals use text messages to get personal information by trying to redirect them to a fake website.This fake website looks like that it is an original website.

As android phones or smartphones are mostly used by the user,cybercriminals use this opportunity to perform this type of attack.Because they don’t have to go through the trouble of breaking firewalls and then accessing the system of the user to steal data.

Symptoms of the phishing:

- It may request the user to share personal details like the login credentials related to the bank and more.
- It redirects to a website if the user clicks on the link that was sent in the email.
- If they are redirected to a website it may want some information related to the credit card or banking details of the user.

Preventive measures of phishing:

- Do not try to open any suspicious email attachments.
- Do not try to open any link which may seem suspicious.
- Do not try to provide any sensitive information like personal information or banking information via email,text,or messages.
- Always the user should have an antivirus to make sure the system is affected by the system or not.

7.3 FEATURE 3

Login Algorithm:

1. Input the credentials(email and password).
2. If already logged in user is taken to home page
3. Else,check for validity of credentials
4. If wrong credentials entered,notification is displayed to user and user stays in login page.
5. On correct credentials,user is taken to home page**.**

Code**:**

7.3.1 FEATURE 4:

```
#To perform operations on dataset
import pandas as pd
import numpy as np


#Machine learning model
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier


#Visualization
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
```

Next we read and split the dataset:

```
df=pd.read_csv('.../dataset.csv')

dot_file='.../tree.dot'

confusion_matrix_file='.../confusion_matrix.png'
```

And then print the results:

```
print(df.head())

-1 1 1.1 1.2-1.1-1.2-1.3-1.4-1.5 1.3 1.4-1.6 1.5-1.7 1.6...-1.9-1.10 0 1.7
1.8 1.9 1.10-1.11-1.12-1.13-1.14 1.11 1.12-1.15-1.16

0 1 1 1 1 1-1 0 1-1 1 1-1 1 0-1...1 1 0 1 1 1 1-1-1 0-1 1 1 1-1

1 1 0 1 1 1-1-1-1-1 1 1-1 1 0-1...-1-1 0 1 1 1 1 1-1 1-1 1 0-1-1

2 1 0 1 1 1-1-1-1 1 1 1-1-1 0 0...1 1 0 1 1 1 1-1-1 1-1 1-1 1-1

3 1 0-1 1 1-1 1 1-1 1 1 1 1 0 0...1 1 0-1 1-1 1-1-1 0-1 1 1 1 1

4-1 0-1 1-1-1 1 1-1 1 1-1 1 0 0...-1-1 0 1 1 1 1 1 1 1-1 1-1-1 1
```

This dataset contains 5 rows and 31 columns,where each column contains a value for each of the attributes we discussed in the above section.

3—Train The Model

As always,the first step in training a machine learning model is to split the dataset into testing and training data:

```
X=df.iloc[:,:-1]

y=df.iloc[:,-1]



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

Since the dataset contains boolean data,it’s always best to use a Decision Tree,RandomForest Classifier or Logistic Regression algorithm since these models work best for classification.In this case,I chose to work with a Decision Tree,because it’s straightforward and generally gives the best results when trying to classify data.

```
model=DecisionTreeClassifier()

model.fit(Xtrain,ytrain)
```

4—Evaluate The Model

Now that the model is trained,let’s see how well it does on the test data:

```
ypred=model.predict(Xtest)

print(metrics.classification_report(ypred,ytest))

print("\n\nAccuracy
Score:",metrics.accuracy_score(ytest,ypred).round(2)*100,"%")
```

We used the model to predict *Xtest* data.Now let’s compare the results to *ytest* and see how well we did:

```
precision recall f1-score support

-1 0.95 0.95 0.95 1176

1 0.96 0.96 0.96 1588



micro avg 0.96 0.96 0.96 2764

macro avg 0.96 0.96 0.96 2764

weighted avg 0.96 0.96 0.96 2764



Accuracy Score:96.0%
```

Not bad!We made literally no modifications to the data and achieved an accuracy score of 96%.From here,you can dive deeper into the data and see if there’s any transformation that can be done to further improve the accuracy of prediction.

5—Identify False Positives&False Negatives

The results of any decision tree evaluation are likely to contain both false positives(URLs that are actually valid,but that our model indicates are not),as well as false negatives(URLs that are actually bad,but our model indicates are fine).To help resolve these instances,let’s draw out a confusion matrix(a table with 4 different combinations of predicted and actual values)for our results.The matrix will help us identify:

- True Positives
- True Negatives
- False Positives(Type 1 Error)
- False Negatives(Type 2 Error)

```
mat=confusion_matrix(ytest,ypred)

sns.heatmap(mat.T,square=True,annot=True,fmt='d',cbar=False)

plt.xlabel('true label')

plt.ylabel('predicted label');

plt.savefig(confusion_matrix_file)
```

As you can see,the number of false positives and false negatives are pretty low compared to our true positives and negatives,so we can be pretty sure of our results.

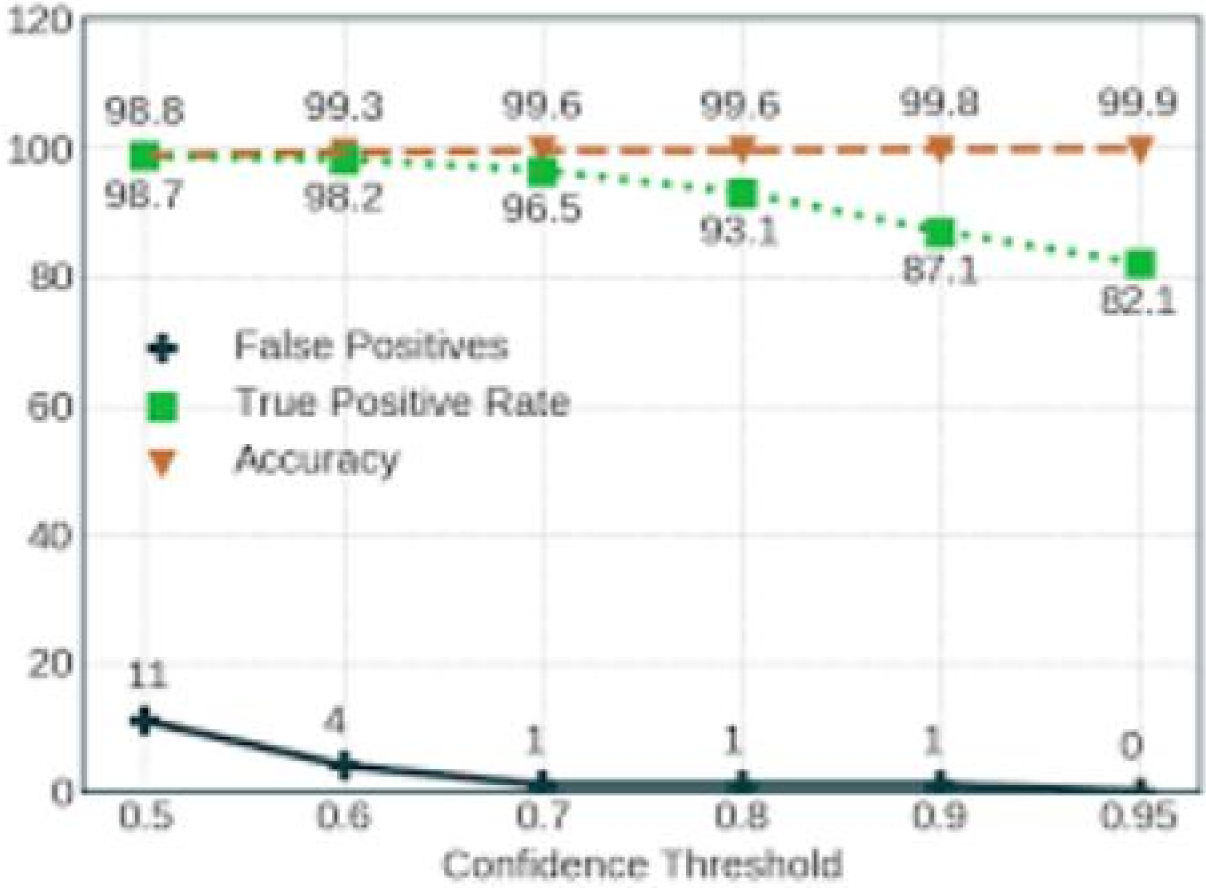
To see how the decision tree panned out in making these decisions,we can visualize it with sklearn,matplotlib and sns.

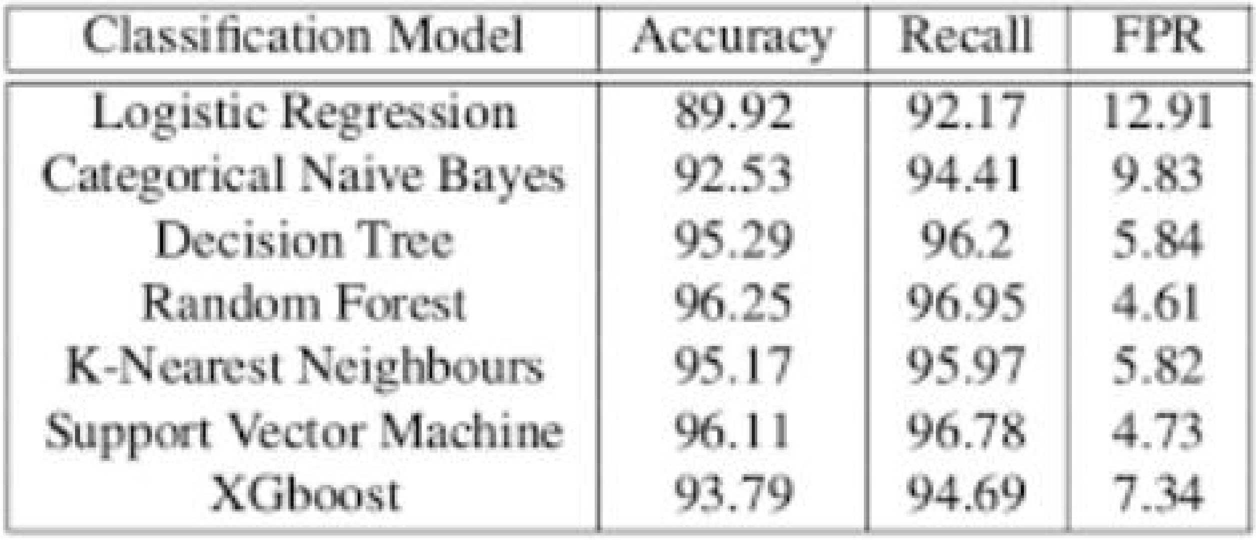
```
export_graphviz(model,out_file=dot_file,feature_names=X.columns.values)>>dot-Tpng
tree.dot-o tree.png
```

We use *export_graphviz* to create a dot file of the decision tree,which is a text file that lets us visualize the actual bifurcations in decisions.Then,using the command line tool *dot* we convert the text file to a PNG image which shows our final“tree”of decisions(open it in a new tab to view the details

7.3.DATABASE SCHEMA

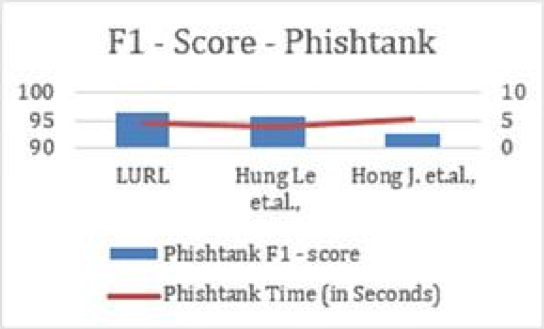
NoSQL databases like MongoDB offer high performance,high availability,and easy scalability.MongoDB is a documentoriented database which stores data in JSON-like documents with dynamic schema.It means you can store your records without worrying aboutthe data structure such as the number of fields or types of fields to store values.MongoDB documents are similar to JSON objects.Details like name,email,password of the registered user are stored so that when the user tries to login,authentication takes place and the user is logged in.

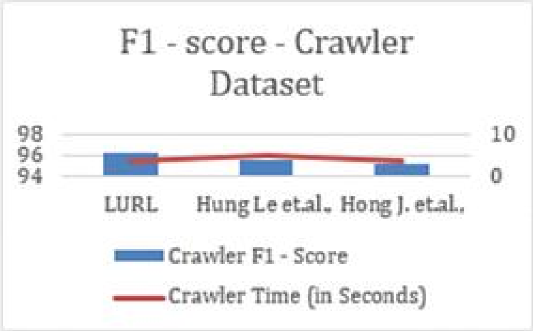


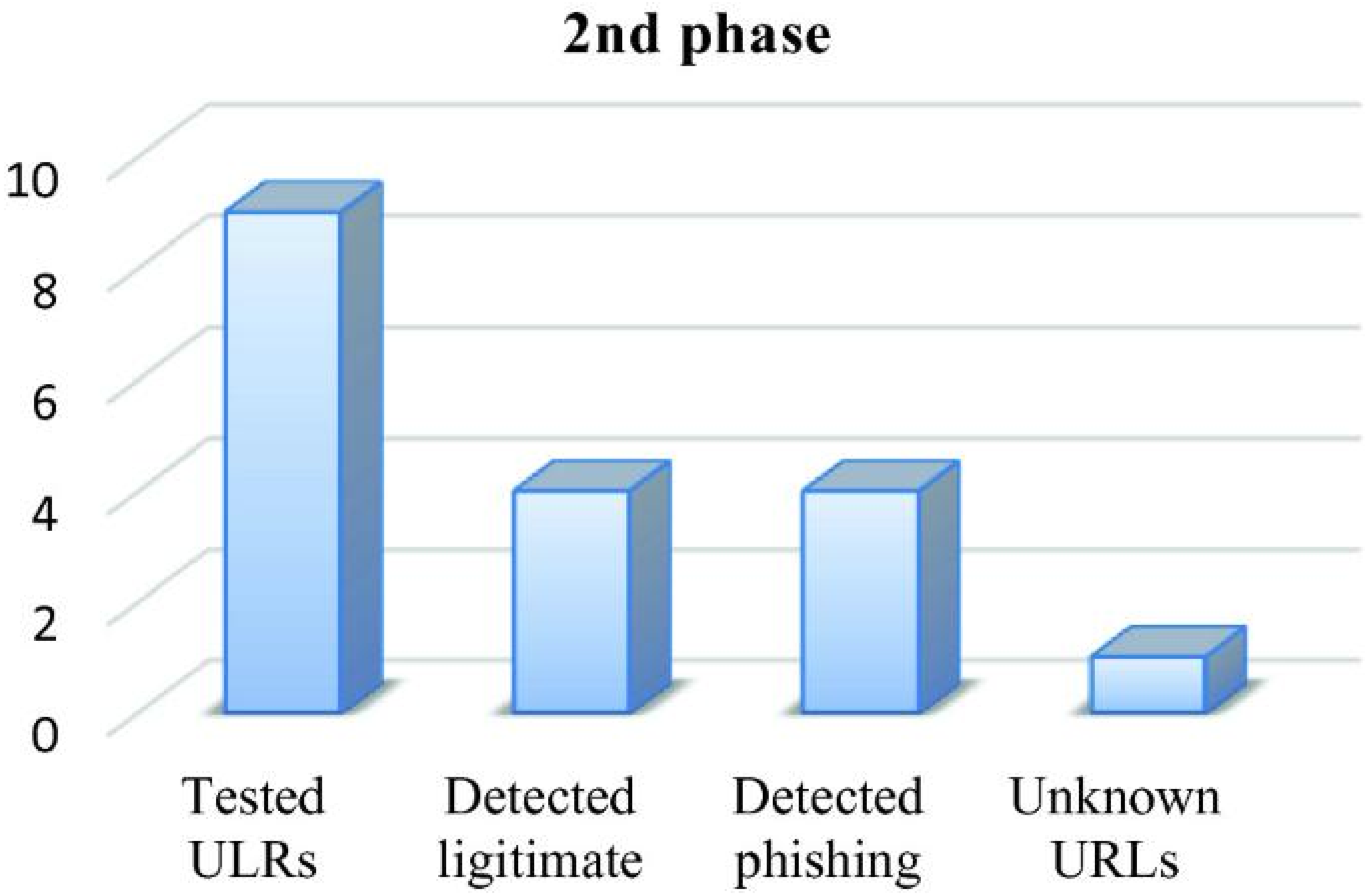


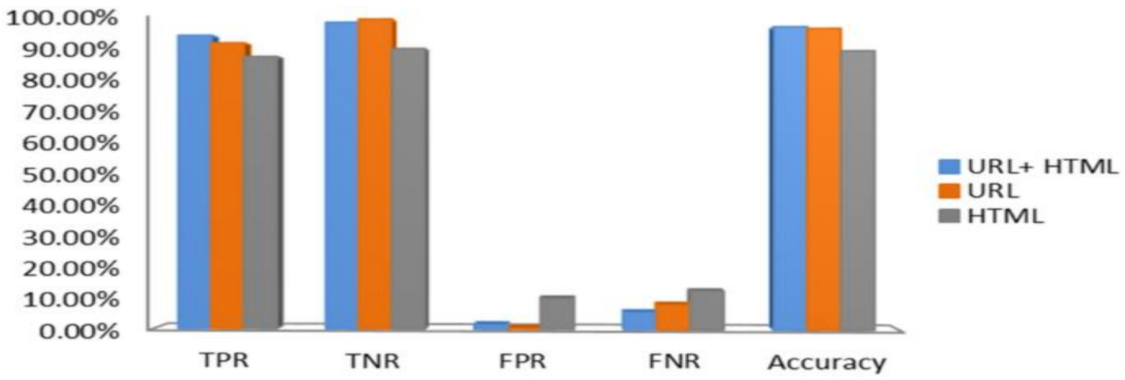
Classification Model	Accuracy	Recall	FPR
Logistic Regression	89.92	92.17	12.91
Categorical Naive Bayes	92.53	94.41	9.83
Decision Tree	95.29	96.2	5.84
Random Forest	96.25	96.95	4.61
K-Nearest Neighbours	95.17	95.97	5.82
Support Vector Machine	96.11	96.78	4.73
XGboost	93.79	94.69	7.34

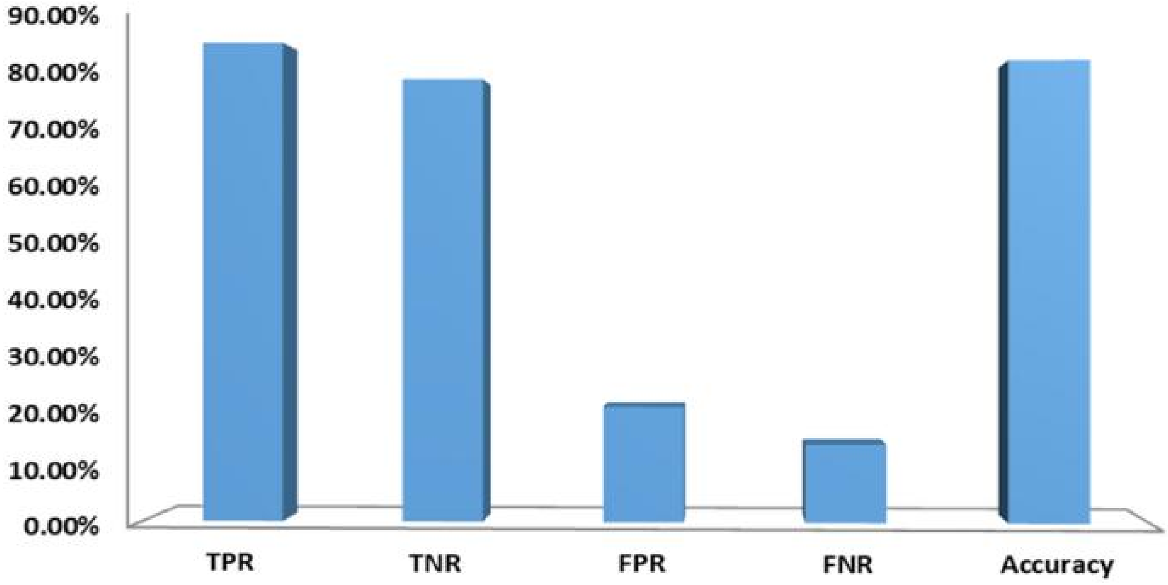
Table 1. Classification Models Results (in percentage)

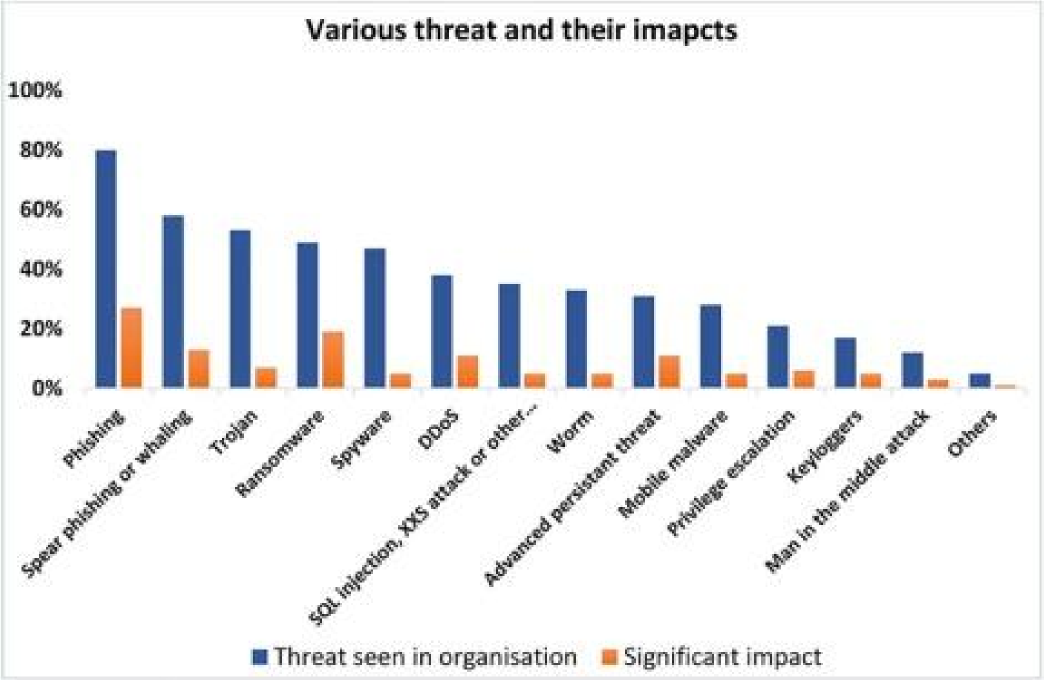

F1 - Score - Phishtank
100
95
90
10
5
0
LURL
Hung Le et.al.,
Hong J. et.al.,
Phishtank F1 - score
Phishtank Time (in Seconds)

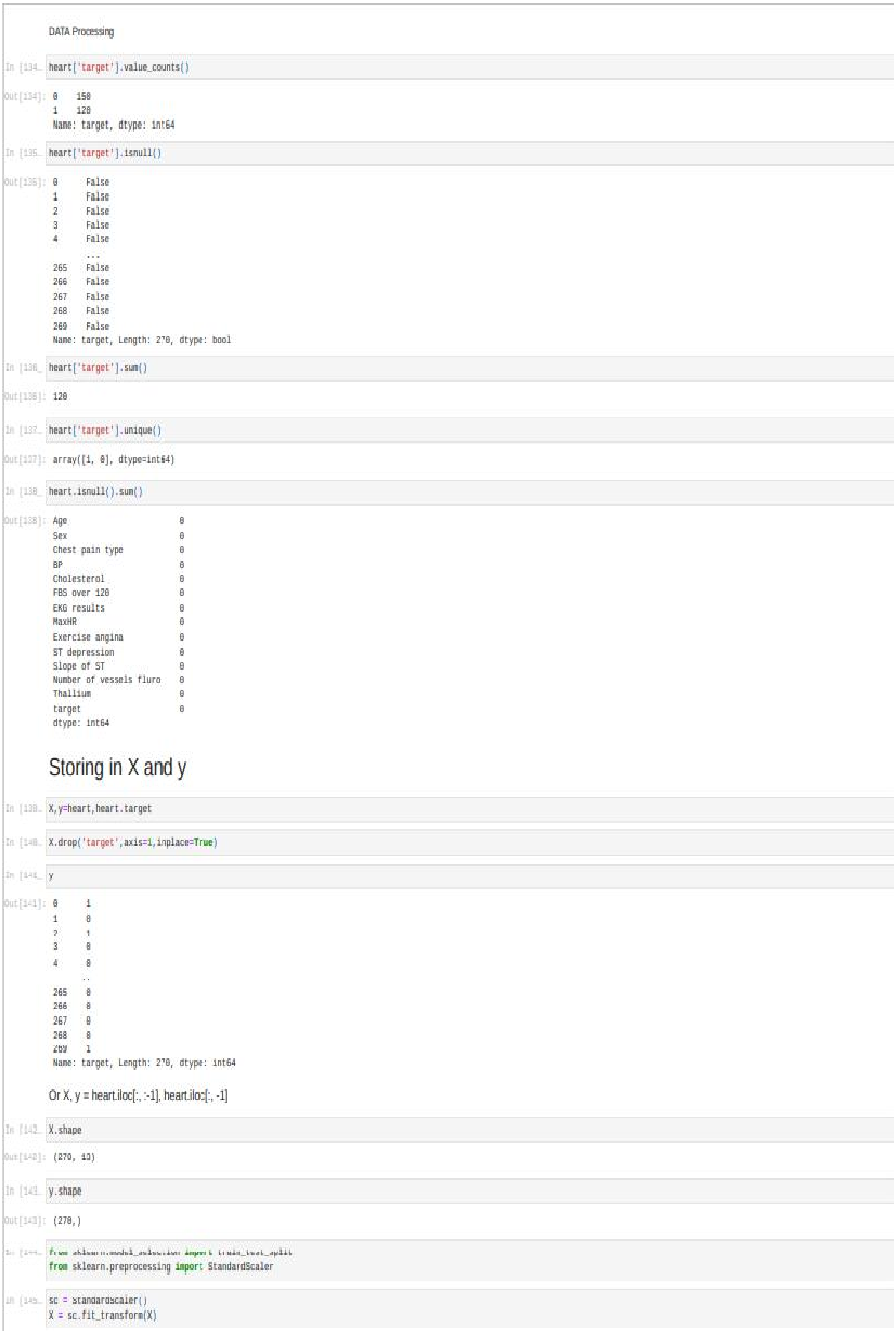

F1 - score - Crawler Dataset
98
96
94
10
0
LURL
Hung Le et.al.,
Hong J. et.al.,
Crawler F1 - Score
Crawler Time (in Seconds)


2nd phase
10
8
6
4
2
0
Tested ULRs
Detected ligitimate
Detected phishing
Unknown URLs


100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
TPR
TNR
FPR
FNR
Accuracy
URL+ HTML
URL
HTML


90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
TPR
TNR
FPR
FNR
Accuracy


Various threat and their imapcts
100%
80%
60%
40%
20%
0%
Phishing
Spear phishing or whaling
Trojan
Ransomware
Spyware
DDoS
SQL injection, XXS attack or other...
Worm
Advanced persistant threat
Mobile malware
Privilege escalation
Keyloggers
Man in the middle attack
Others
■ Threat seen in organisation
■ Significant impact



DATA Processing

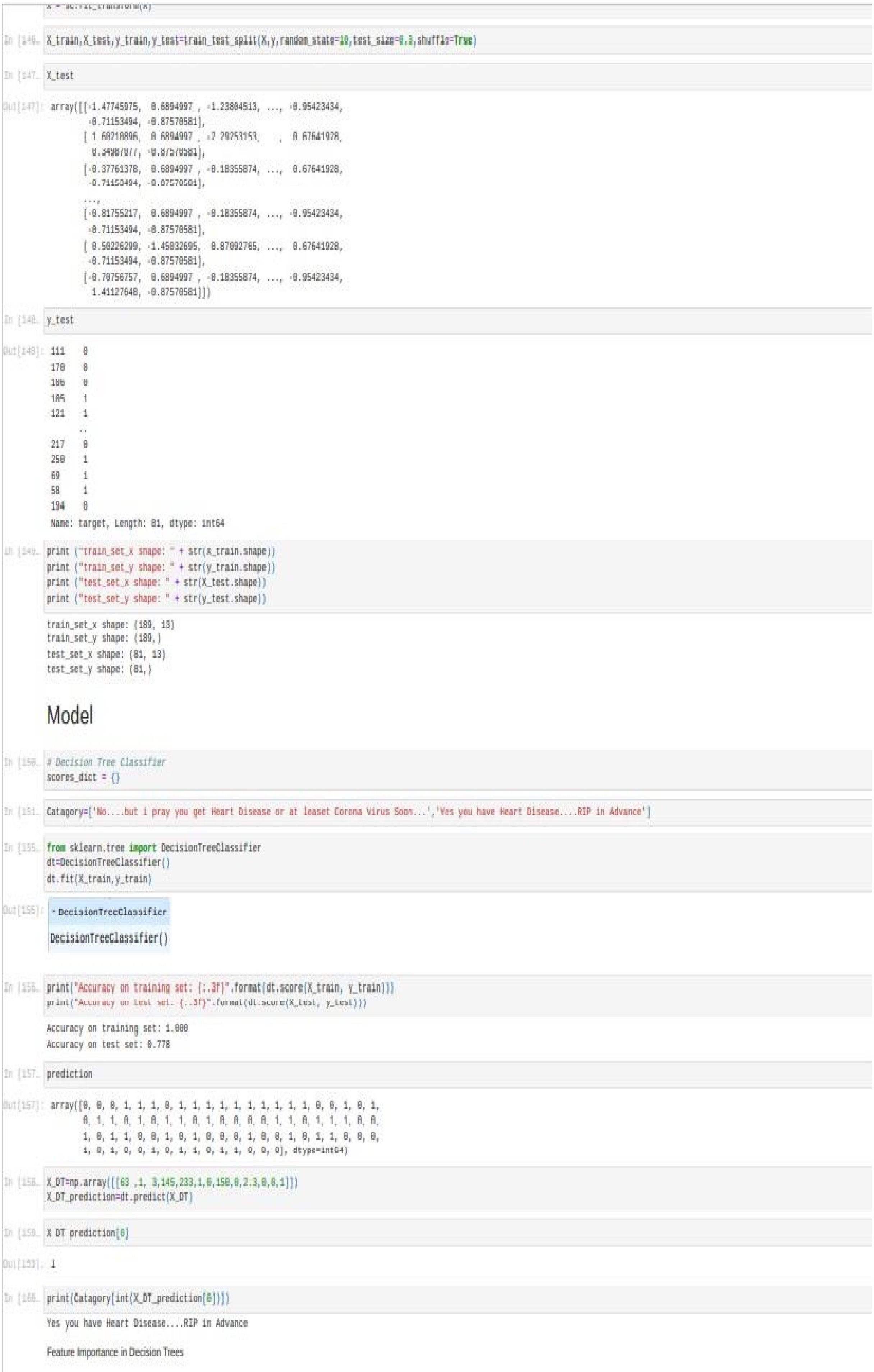
```
In [134…  heart['target'].value_counts()

Out[134]: 0    150
          1    120
          Name: target, dtype: int64

In [135…  heart['target'].isnull()

Out[135]: 0      False
          1      False
          2      False
          3      False
          4      False
                 ...
          265    False
          266    False
          267    False
          268    False
          269    False
          Name: target, Length: 270, dtype: bool

In [136…  heart['target'].sum()

Out[136]: 120

In [137…  heart['target'].unique()

Out[137]: array([1, 0], dtype=int64)

In [138…  heart.isnull().sum()

Out[138]: Age                        0
          Sex                        0
          Chest pain type            0
          BP                         0
          Cholesterol                0
          FBS over 120               0
          EKG results                0
          MaxHR                      0
          Exercise angina            0
          ST depression              0
          Slope of ST                0
          Number of vessels fluro    0
          Thallium                   0
          target                     0
          dtype: int64
```

Storing in X and y

```
In [139…  X,y=heart,heart.target

In [140…  X.drop('target',axis=1,inplace=True)

In [141…  y

Out[141]: 0      1
          1      0
          2      1
          3      0
          4      0
                ..
          265    0
          266    0
          267    0
          268    0
          269    1
          Name: target, Length: 270, dtype: int64
```

Or X, y = heart.iloc[:, :-1], heart.iloc[:, -1]

```
In [142…  X.shape

Out[142]: (270, 13)

In [143…  y.shape

Out[143]: (270,)

In [144…  from sklearn.model_selection import train_test_split
          from sklearn.preprocessing import StandardScaler

In [145…  sc = StandardScaler()
          X = sc.fit_transform(X)
```



```
In [146… X_train,X_test,y_train,y_test=train_test_split(X,y,random_state=10,test_size=0.3,shuffle=True)
```

```
In [147… X_test
```

```
Out[147]: array([[-1.47745075,  0.6894997 , -1.23884513, ..., -0.95423434,
                  -0.71153494, -0.87570581],
                 [ 1.60210896,  0.6894997 , -2.29253153, ...,  0.67641928,
                   0.34987977, -0.87570581],
                 [-0.37761378,  0.6894997 , -0.18355874, ...,  0.67641928,
                  -0.71153494, -0.87570581],
                 ...,
                 [-0.81755217,  0.6894997 , -0.18355874, ..., -0.95423434,
                  -0.71153494, -0.87570581],
                 [ 0.50226299, -1.45032695,  0.87092765, ...,  0.67641928,
                  -0.71153494, -0.87570581],
                 [-0.70756757,  0.6894997 , -0.18355874, ..., -0.95423434,
                   1.41127648, -0.87570581]])
```

```
In [148… y_test
```

```
Out[148]: 111    0
          170    0
          186    0
          105    1
          121    1
                ..
          217    0
          250    1
          69     1
          58     1
          194    0
          Name: target, Length: 81, dtype: int64
```

```
In [149… print ("train_set_x shape: " + str(X_train.shape))
          print ("train_set_y shape: " + str(y_train.shape))
          print ("test_set_x shape: " + str(X_test.shape))
          print ("test_set_y shape: " + str(y_test.shape))
```

```
train_set_x shape: (189, 13)
train_set_y shape: (189,)
test_set_x shape: (81, 13)
test_set_y shape: (81,)
```

Model

```
In [150… # Decision Tree Classifier
          scores_dict = {}
```

```
In [151… Catagory=['No....but i pray you get Heart Disease or at leaset Corona Virus Soon...','Yes you have Heart Disease....RIP in Advance']
```

```
In [155… from sklearn.tree import DecisionTreeClassifier
          dt=DecisionTreeClassifier()
          dt.fit(X_train,y_train)
```

```
Out[155]: ▸ DecisionTreeClassifier
          DecisionTreeClassifier()
```

```
In [156… print("Accuracy on training set: {:.3f}".format(dt.score(X_train, y_train)))
          print("Accuracy on test set: {:.3f}".format(dt.score(X_test, y_test)))
```

```
Accuracy on training set: 1.000
Accuracy on test set: 0.778
```

```
In [157… prediction
```

```
Out[157]: array([0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1,
                 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0,
                 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0,
                 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0], dtype=int64)
```

```
In [158… X_DT=np.array([[63 ,1, 3,145,233,1,0,150,0,2.3,0,0,1]])
          X_DT_prediction=dt.predict(X_DT)
```

```
In [159… X_DT_prediction[0]
```

```
Out[159]: 1
```

```
In [160… print(Catagory[int(X_DT_prediction[0])])
```

```
Yes you have Heart Disease....RIP in Advance
```

Feature Importance in Decision Trees

CHAPTER 8

TESTING

8.1 TEST CASES

Testcase 1: Logging in with registered login details.

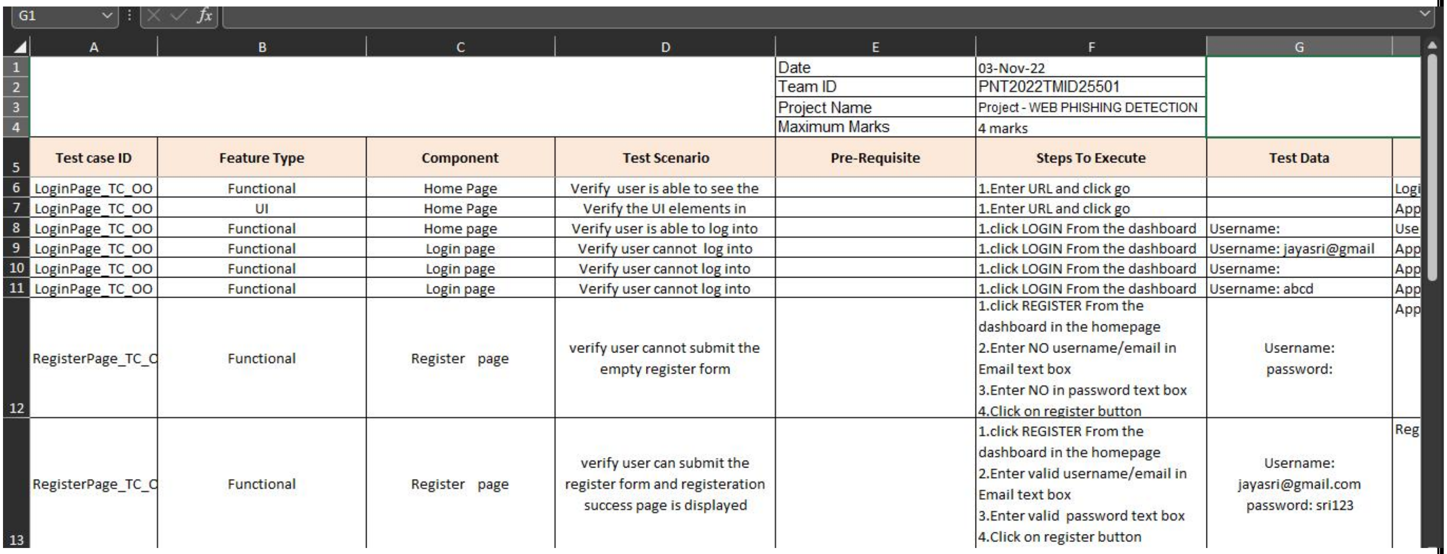
Testcase 2: Logging in with invalid login details.

Testcase 3: Registering with existing user’s details.

Testcase 4: Entering wrong values while filling medical related details.

Testcase 5: Producing visualisations for given input.

8.2 USER ACCEPTANCE TESTING



Date	03-Nov-22
Team ID	PNT2022TMID25501
Project Name	Project - WEB PHISHING DETECTION
Maximum Marks	4 marks

Test case ID	Feature Type	Component	Test Scenario	Pre-Requisite	Steps To Execute	Test Data	
LoginPage_TC_OO	Functional	Home Page	Verify user is able to see the		1.Enter URL and click go		Logi
LoginPage_TC_OO	UI	Home Page	Verify the UI elements in		1.Enter URL and click go		App
LoginPage_TC_OO	Functional	Home page	Verify user is able to log into		1.click LOGIN From the dashboard	Username:	Use
LoginPage_TC_OO	Functional	Login page	Verify user cannot log into		1.click LOGIN From the dashboard	Username: jayasri@gmail	App
LoginPage_TC_OO	Functional	Login page	Verify user cannot log into		1.click LOGIN From the dashboard	Username:	App
LoginPage_TC_OO	Functional	Login page	Verify user cannot log into		1.click LOGIN From the dashboard	Username: abcd	App
RegisterPage_TC_C	Functional	Register page	verify user cannot submit the empty register form		1.click REGISTER From the dashboard in the homepage 2.Enter NO username/email in Email text box 3.Enter NO in password text box 4.Click on register button	Username: password:	App
RegisterPage_TC_C	Functional	Register page	verify user can submit the register form and registeration success page is displayed		1.click REGISTER From the dashboard in the homepage 2.Enter valid username/email in Email text box 3.Enter valid password text box 4.Click on register button	Username: jayasri@gmail.com password: sri123	Reg

CHAPTER 9

RESULTS

9.1 PERFORMANCE METRICS

1. Hours worked:50 hours

2. Stick to Timelines:100%

3. Stay within budget:100%

4. Consistency of the product:85%

5. Efficiency of the product:85%

6. Quality of the product:80%

CHAPTER 10

ADVANTAGES&DISADVANTAGES

ADVANTAGES:

- Smooth User Interface
- Accuracy is achieved quickly

DISADVANTAGES:

- Random forest can be used for both classification and regression tasks,but it is not more suitable for Regression tasks.

CHAPTER 11

CONCLUSION

This paper aims to enhance the detection method to detect phishing 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.

CHAPTER 12

FUTURE SCOPE

Phishing is a considerable problem differs from the other security threats such as intrusions and Malware which are based on the technical security holes of the network systems.The weakness point of any network system is its Users.Phishing attacks are targeting these users depending on the trikes of social engineering.Despite there are several ways to carry out these attacks,unfortunately the current phishing detection techniques cover some attack vectors like email and fake websites.Therefore,building a specific limited scope detection system will not provide complete protection from the wide phishing attack vectors.

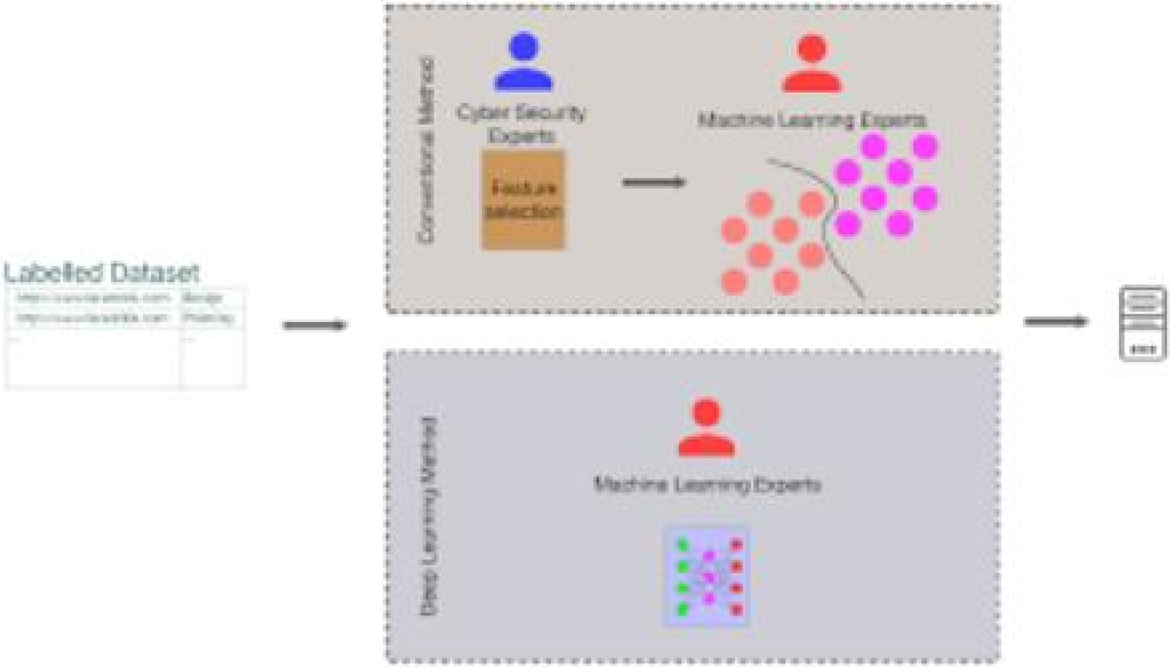
CHAPTER 13

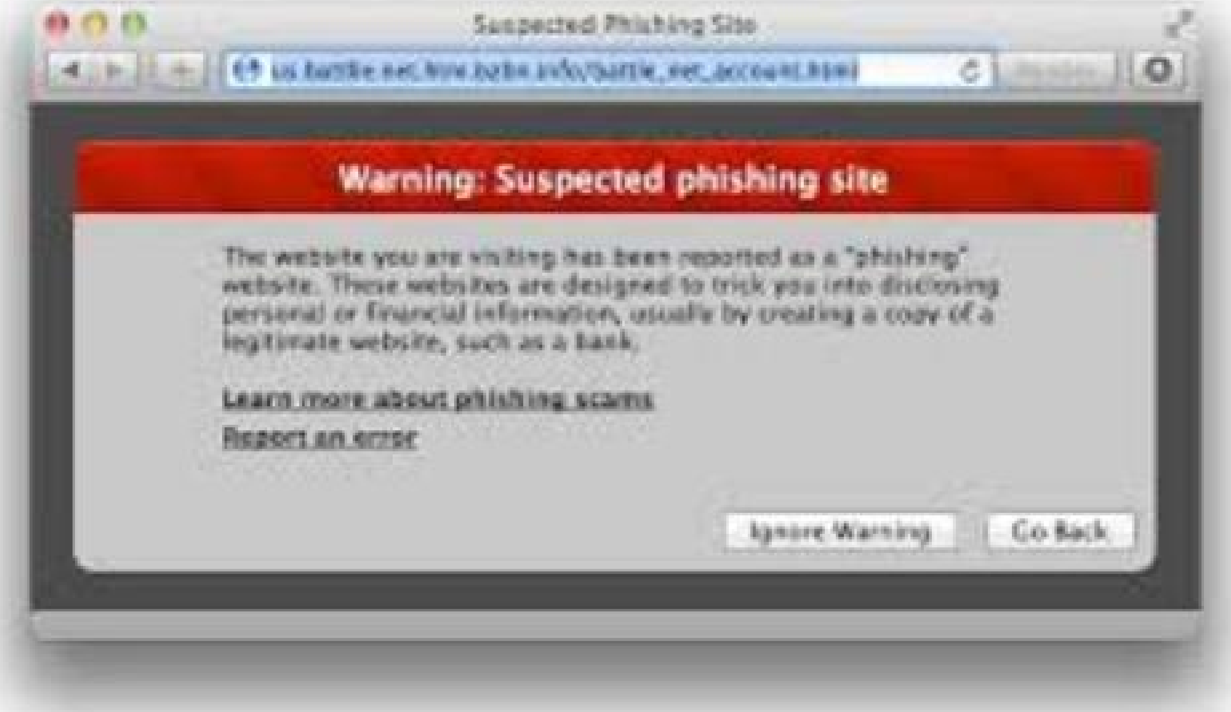
APPENDIX

PROJECT DEMONSTRATION LINK:https://youtu.be/mnNkPE5JoMY

GITHUB LINK: https://github.com/IBM-EPBL/IBM-Project-26570-1660029758

APPENDIX A1:SCREENSHOTS





Warning: Suspected phishing site

The website you are visiting has been reported as a "phishing" website. These websites are designed to trick you into disclosing personal or financial information, usually by creating a copy of a legitimate website, such as a bank.

Learn more about phishing scams

Report an error