REPORT ON
WEB PHISHING DETECTION
DOMAIN: APPLIED DATA SCIENCE

TEAM ID: PMT2022TMID29707

TEAM LEADER: SHANMUGAPRIYA D

TEAM MEMBERS:

1. MAHALAKSHMI M
2. SINDHU S
3. SWETHA D

TABLE OF CONTENTS

1. INTRODUCTION

1.1 PROJECT OVERVIEW:

Phishing attack is a simplest way to obtain sensitive information from innocent users. Large organizations may get trapped in different kinds of scams. Web phishing is one of many security threats to web services on Internet. There are a number of users who purchase products online and make payments through e-banking. There are e-banking websites that ask users to provide sensitive data such as username, password & credit card details, etc., often for malicious reasons. This type of e-banking website is known as a phishing website. Web phishing aims to steal private information, such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity.

1.2 PURPOSE:

The purpose of Phishing Website Detection is detecting phishing website names. Therefore, passive queries related to the domain name, which we want to classify as phishing or not, provide useful information to us.

2. LITERATURE SURVEY

A literature review is **a survey of scholarly sources (such as books, journal articles, and theses) related to a specific topic or research question**. It is often written as part of a thesis, dissertation, or research paper, in order to situate your work in relation to existing knowledge.

We collected the relevant information on our Web Phishing Detection project and we existed the solutions. We all gathered together and referred the following points through research publications.

2.1 REFERENCES:

1. Abdul ghani Ali Ahmed, Nurul Amirah Abdullah, International Conference on Computer Applications & Information Security(ICCAIS),2016

2. Amani Alswailen, Bashayr Alabdullah, Norah Alrumayh, Dr. Aram Alsedrani,International Conference and Workshop on Computing and Communication (IEMCON),2019

3. Mehmet Korkmaz, Ozgur Koray Sahingoz, Banu Diri, International Conference on Computing, Communication and Networking Technologies (ICCNT),2020.

4, Emre Kocyigit, Kubra Erensoy, Buket Geyik, International Conference on Innovative Computation Technologies[ICICT] IEEE,MAY 15,2021.

5. Ammar Odeh, Ismail Keshta, Eman Abdelfattah,International Conference on Innovative Computation Technologies [ICICT] IEEE, May 17,2021.

6. Rishikesh Mahajan, Irfan Siddarame,International Conference on Innovative Computation Technologies [ICICT] IEEE,2018.

7. Hochreiter and Schmidhuber, Information Technology, Networking,Electronics and Automation Control Conference,2020.

8. Happy Chapla,Riddhi Kotak, Mittal Joiser, International Conference on Communication and Electronics Systems,2019.

9. Bhagwat M. D,Dr. Patil P. H. , Dr. T. S. Vishwanath, International Conference on Intelligent Communication Technologies and Virtual Mobile Networks,2021

10. Chee-Yu Wu, Cheng-Chung Kuo,Chu-Sing Yang, Institute of Computer and Communication Engineering National Cheng Kung University Tainan, Taiwan,2019.11. Chon Lin Tan, Kang Leng Chiew, SaNah Sze, Faculty of Computer Science and Information Technology University Sarawak, Malaysia,JUNE 2014.

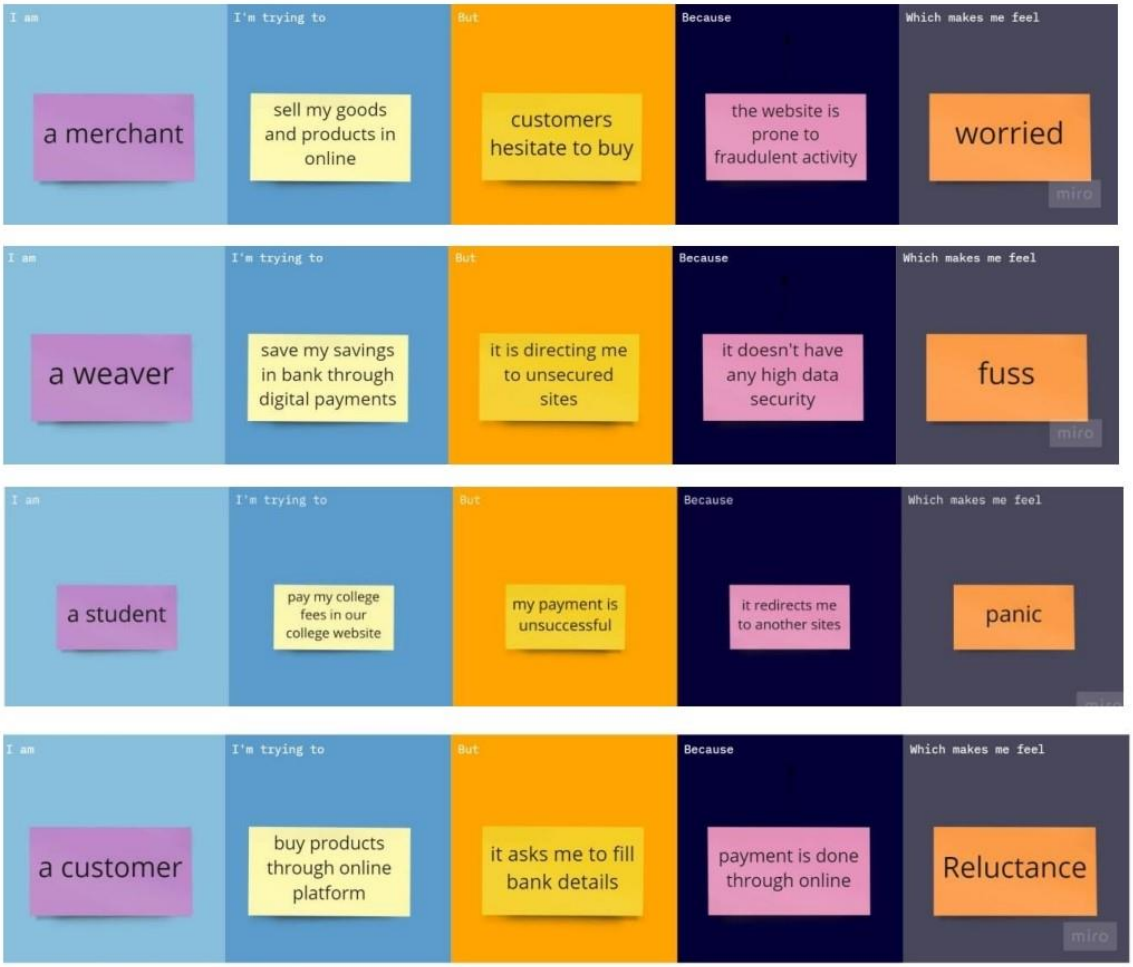
12. Andrew J. Park, Ruhi Naaz Quadari, Herbert H. Tsang, Department of Computing Science Thompson Rivers University Kamloo.

2.2 PROBLEM STATEMENT DEFINITION

Problem-solution essays consider the problems of a particular situation, and give solutions to those problems. They are in some ways similar to cause and effect essays, especially in terms of structure. Problem-solution essays are actually a sub-type of another type of essay, which has the following four components:

- Situation
- Problem
- Solution
- Evaluation

By this problem statement we found some problems faced by the people like customers, students, merchant and weaver etc.



3. IDEATION & PROPOSED SOLUTION

3.1 EMPATHY MAP CANVAS

Empathy maps provide a glance into who a user is as a whole and are **not chronological** or sequential.

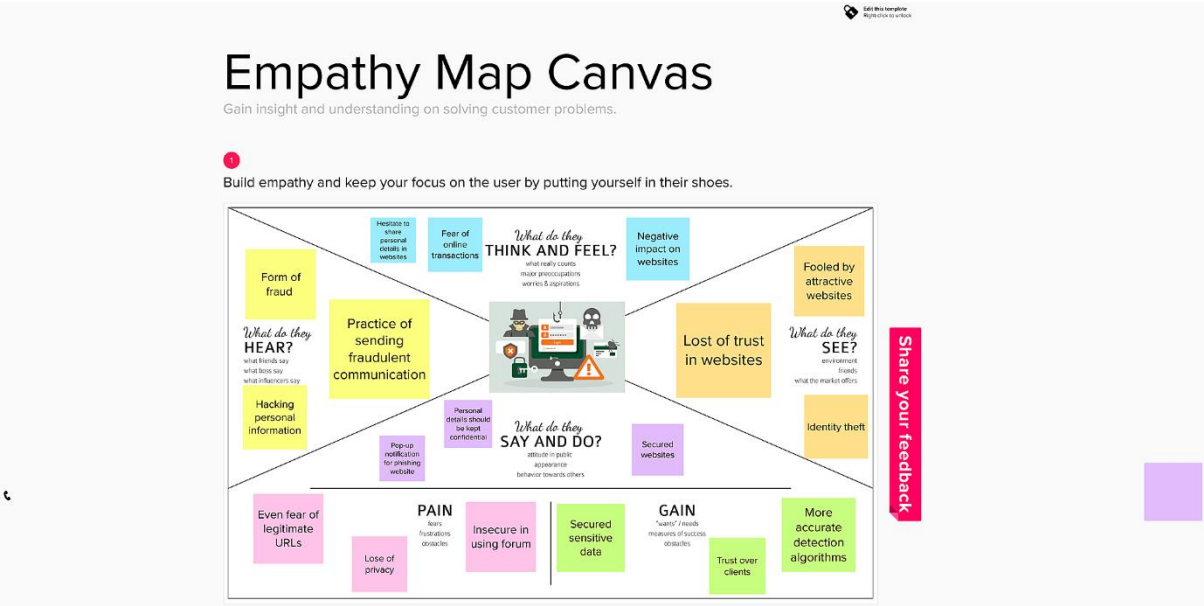
The ***Says*** quadrant contains what the user says out loud in an interview or some other usability study. Ideally, it contains verbatim and direct quotes from research.

The ***Thinks*** quadrant captures what the user is thinking throughout the experience. what occupies the user's thoughts? What matters to the user? It is possible to have the same content in both *Says* and *Thinks.*

The ***Does*** quadrant encloses the actions the user takes. From the research, what does the user physically do? How does the user go about doing it.

The ***Feels*** quadrant is the user's emotional state, often represented as an adjective plus a short sentence for context. Ask yourself: what worries the user? What does the user get excited about? How does the user feel about the experience?

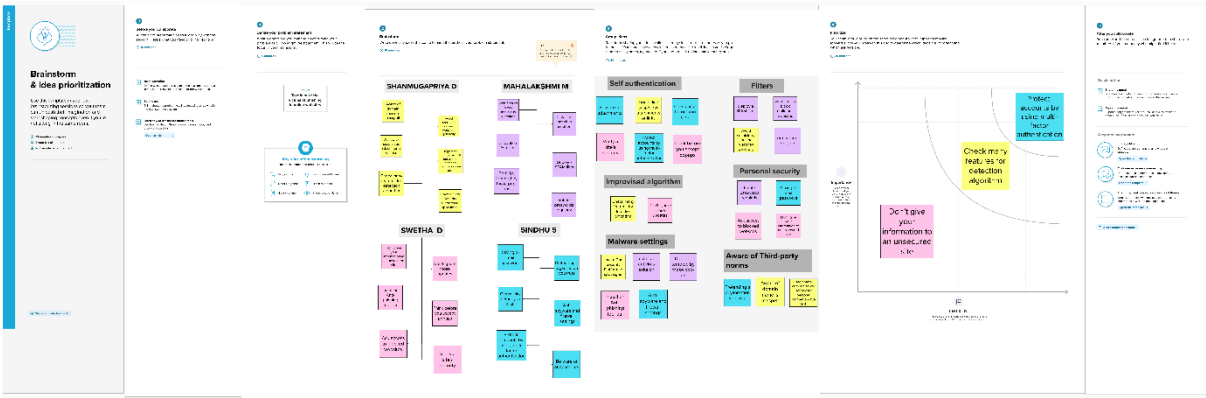
The empathy mapping is done using Mural platform.



3.2 BRAIN STORMING

Brainstorming is part design thinking. You use it in the ideation phase. It's extremely popular for design teams because they can expand in all directions.

Brainstorming is a method design teams use to generate ideas to solve clearly defined design problems.



3.3 PROPOSED SOLUTION

Your proposed solution should **relate the current situation to a desired result and describe the benefits that will accrue when the desired result is achieved**. So, begin your proposed solution by briefly describing this desired result.

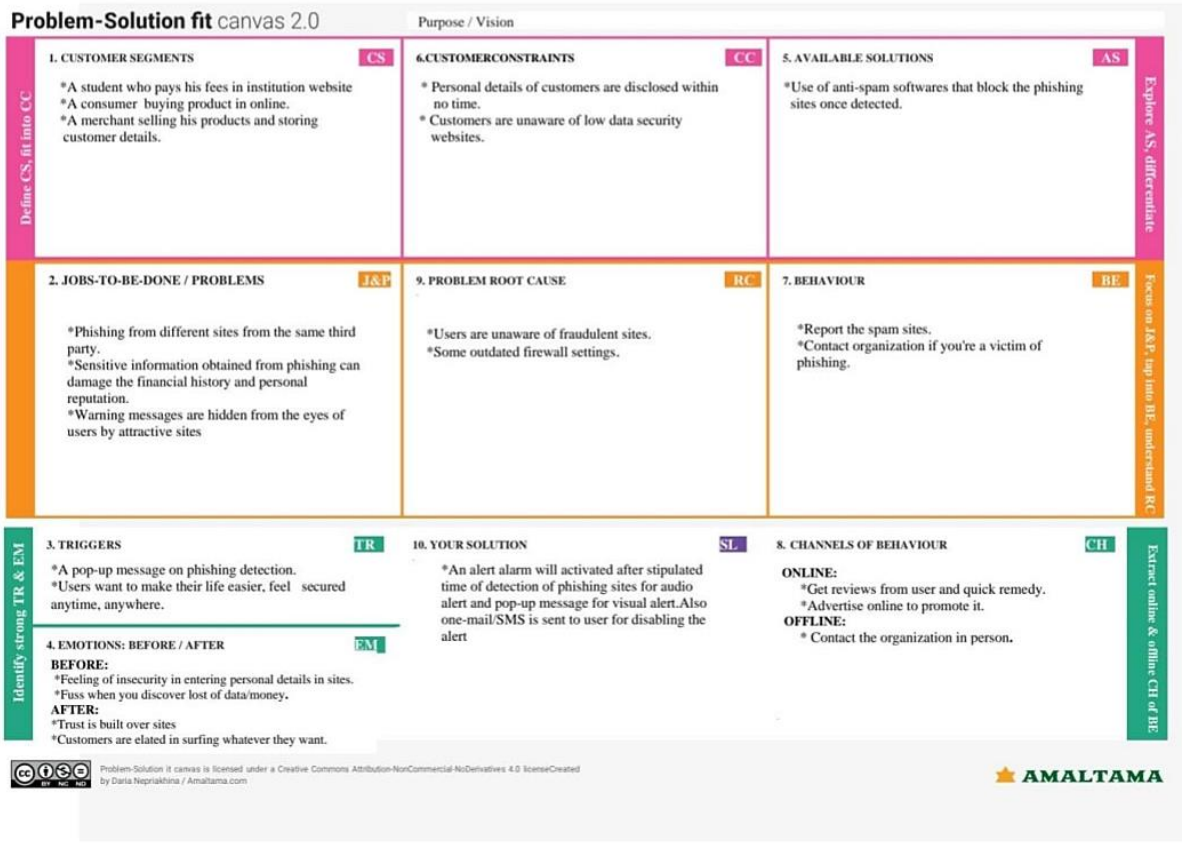
Proposed Solution Template:

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

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	User are carried out by attraction in phishing sites and it eventually leads to ignore the alert message. In rare cases false detection leads the users helpless.
2.	Idea / Solution description	An alert alarm will activated after stipulated time of detection of phishing sites for audio alert and pop-up message for visual alert. Also one-mail/SMS is sent to user for disabling the alert
3.	Novelty / Uniqueness	Besides visual alert and sound alert, an e-mail / SMS sent will allow to know the phishing site. Incase of false detection, there will be an option to disable the alert system and surf in that site under the own risk of the user through the mail received.
4.	Social Impact / Customer Satisfaction	By this, customer can get rid of phishing sites and feels secure on surfing through internet. Flexibility in accessing the false detection at user's own risk.
5.	Business Model (Revenue Model)	It yields expected revenue in training a model that best fits the algorithm to be used.
6.	Scalability of the Solution	It is scalable even the users are increased because the added alert system will decrease the victims.

3.4 PROBLEM SOLUTION FIT

This occurs when you have evidence that customers care about certain jobs, pains, and gains. At this stage you've proved the existence of a problem and have designed a value proposition that addresses your customers' jobs, pains and gains.



4. REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENTS

The functional requirements of the proposed solution are user registration,user confirmation, user cancellations, prediction of phishing website, security alert to the user and confirmation from the user.

4.2 NON-FUNCTIONAL REQUIREMENTS

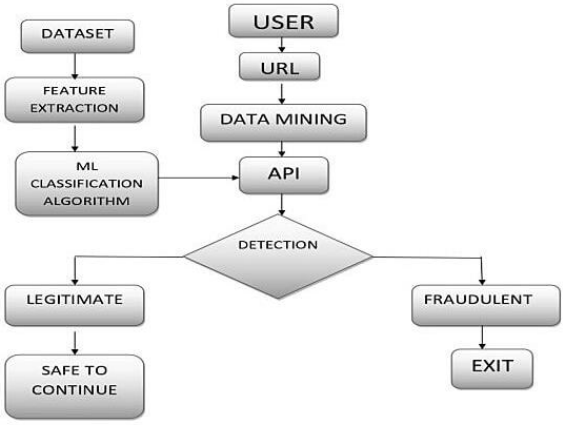
The non-functional requirements of the proposed solution are usability,security,reliability, performance, availability and scalability.

5. PROJECT DESIGN

5.1 DATA FLOW DIAGRAM:

A data-flow diagram is a way of representing a flow of data through a process or a system. The DFD also provides information about the outputs and inputs of each entity and the process itself. A data-flow diagram has no control flow — there are no decision rules and no loops.

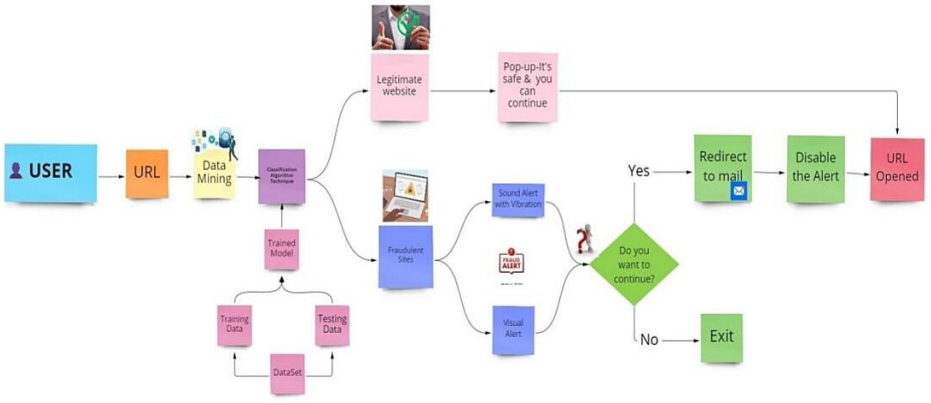
Data Flow Diagrams:



5.2 SOLUTION &TECHNICAL ARCHITECTURE

Solution architecture, term used in information technology with various definitions such as; "A description of a discrete and focused business operation or activity and how IS/IT supports that operation"

Solution Architecture Diagram:



5.3 USER STORIES

A user story is a short, simple description of a feature told from the perspective of the person who desires the new capability, usually a user or customer of the system

User Stories

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 & 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	I can receive confirmation email	Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password	I can receive confirmation email	High	Sprint-1
	Dashboard	USN-6	As a user, I can fill the application through dashboard.	I can access my dashboard	High	Sprint-1
Customer (Web user)	User input	USN-1	As a user I can input the particular URL in the required field and wait for validation.	I can access the website without any problem	High	Sprint-1
Customer Care Executive	Feature extraction	USN-1	As a user,I can extract feature using heuristic and visual similarity approach.	I can have comparison between websites for security.	High	Sprint-1
Administrator	Prediction	USN-1	I will predict the URL websites using Machine Learning algorithms such as Logistic Regression	I can correctly Predict on the particular algorithms	High	Sprint-1
	Classifier	USN-2	I will send all the output model to classifier in order to produce final result	I will find the correct classifier for producing the result	Medium	Sprint-2

6. PROJECT PLANNING &SCHEDULING

6.1 SPRINT PLANNING & ESTIMATION:

Sprint planning is an event in scrum that kicks off the sprint. The purpose of sprint planning is to define what can be delivered in the sprint and how that work will be achieved. Sprint planning is done in collaboration with the whole scrum team.

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

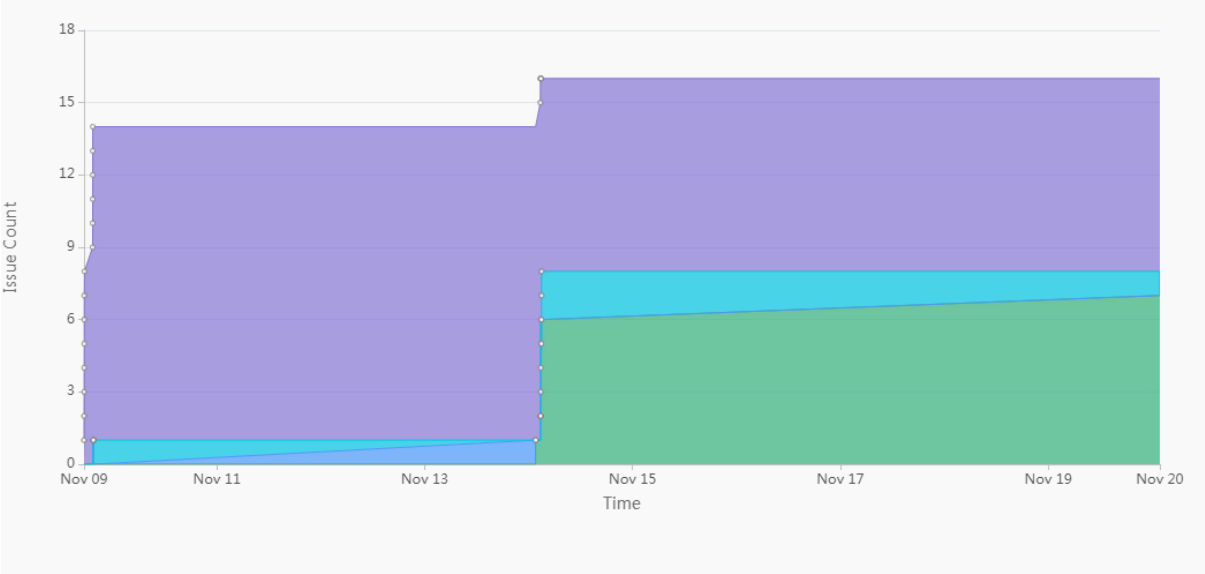
Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint ReleaseDate (Actual)
Sprint-1	13	6 Days	24 Oct 2022	29 Oct 2022	29 Oct 2022	
Sprint-2	13	6 Days	31 Oct 2022	05 Nov 2022	05 Nov 2022	
Sprint-3	13	6 Days	07 Nov 2022	12 Nov 2022	12 Nov 2022	
Sprint-4	13	6 Days	14 Nov 2022	19 Nov 2022	19 Nov 2022	

6.2 SPRINT DELIVERY SCHEDULE:

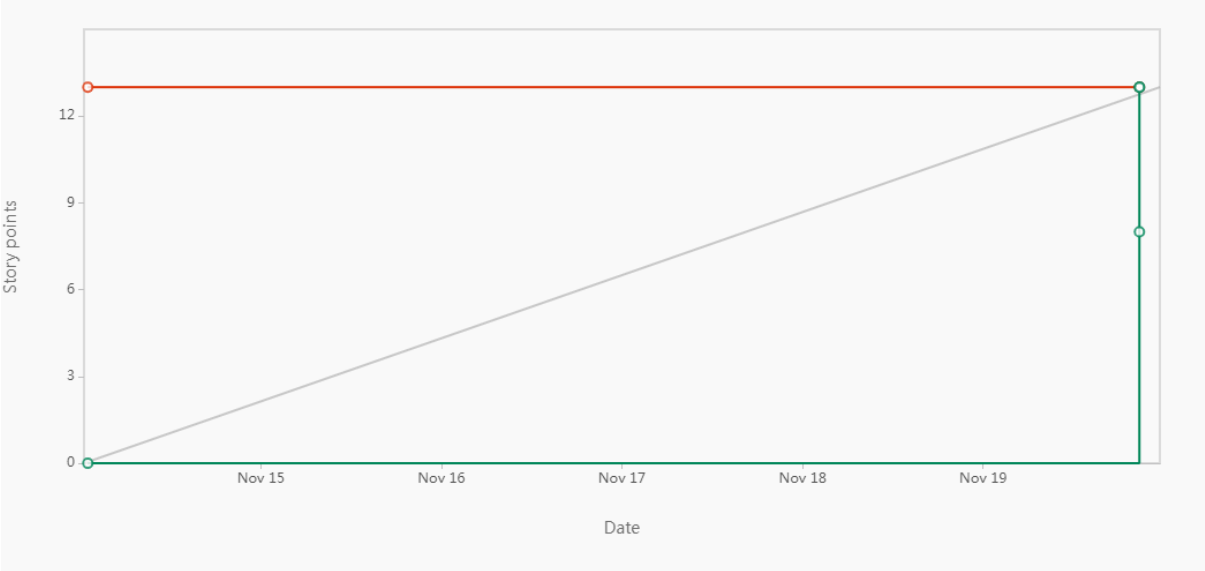
Since sprints take place over a fixed period of time, it's critical to avoid wasting time during planning and development. And this is precisely where sprint scheduling enters the equation.

6.3 REPORTS FROM JIRA :

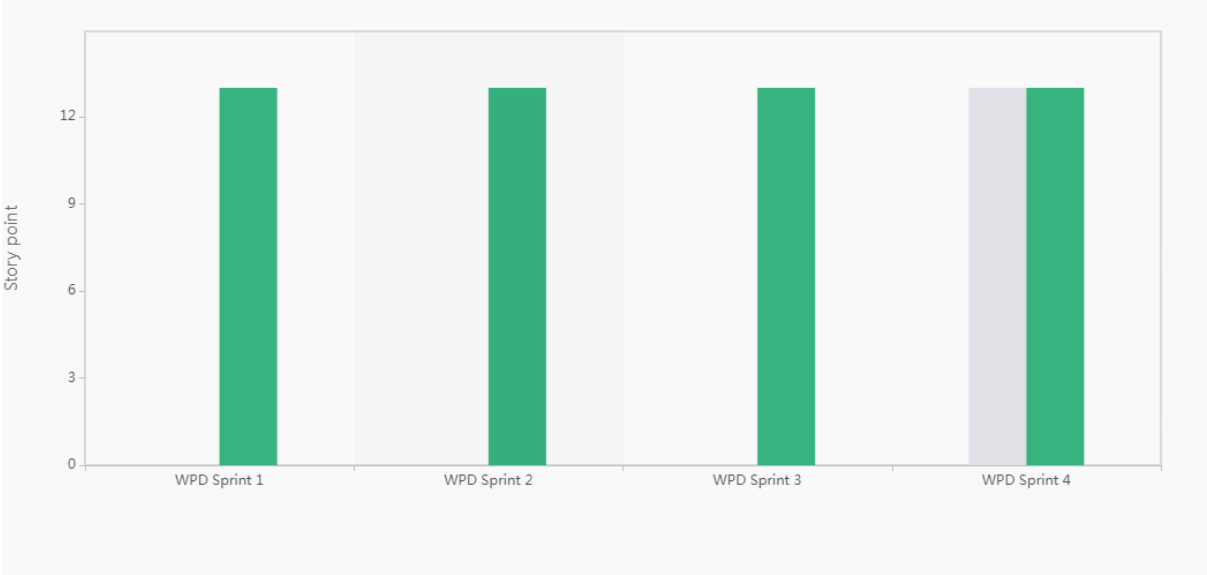
Cumulative diagram:



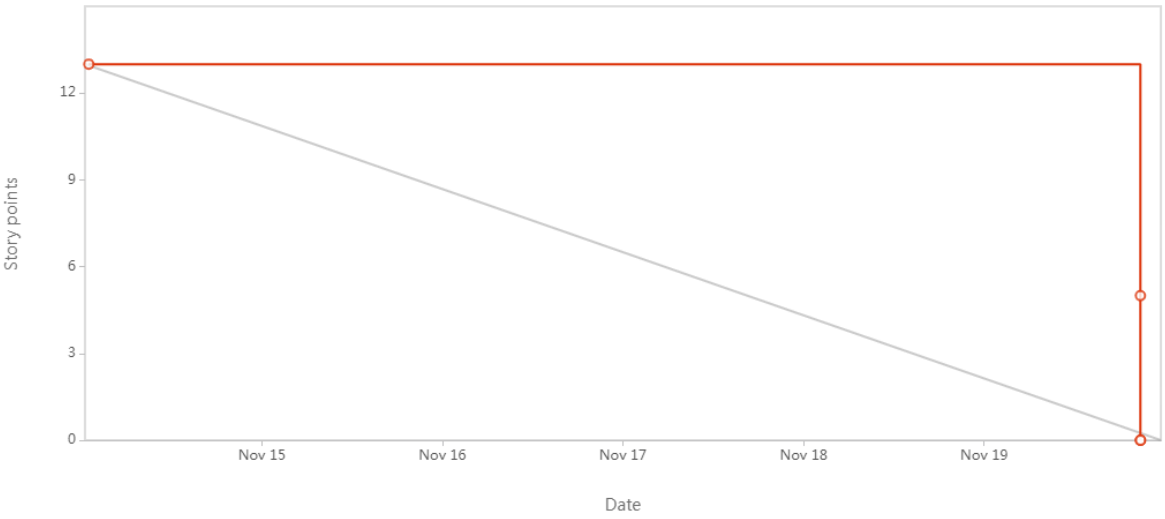
Burnup:



Velocity chart:



Burndown chart:



7.CODING & SOLUTIONING

app.py:

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <meta name="description" content="This website is develop for identify the safety of url.">
    <meta name="keywords" content="phishing url,phishing,cyber security,machine learning,classifier,pytho
    <meta name="author" content="VAIBHAV BICHAVE">

    <!-- BootStrap -->
    <link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.5.0/css/bootstrap.min.css
        integrity="sha384-9aIt2nRpC12Uk9gS9baDl411NQApFmC26EwAOH8WgZl5MYYxFfc+NcPb1dKGj7Sk" crossorigin="
    <link href="static/styles.css" rel="stylesheet">
    <title>Web Phishing Detection IBM</title>

</head>

<body>
     <h2> BE AWARE OF YOUR CONFIDENTIAL DATA! </h2>
<div class=" container">
    <div class="row">
        <div class="form col-md" id="form1">
            <h2>PHISHING URL DETECTION</h2>

            <br>
            <form action="/pred" method ="POST">
                <input type="text" class="form__input" name ='url' id="url" placeholder="Drop Your URL He
                <label for="url" class="form__label">URL</label>
                <button class="button" role="button" >HIT TO PREDICT</button>
```

login.html:

app.py | login.html | final1.css | index.html - KISHORE\...\templates | index.html - Downloads | styles.css | index.html - OneDrive\Desktop

```
import numpy as np
from flask import Flask, request, jsonify, render_template,redirect,url_for
import pickle
#importing the inputScript file used to analyze the URL

from flask import Flask, request, render_template
import numpy as np
import pandas as pd
from sklearn import metrics
import warnings
import pickle
warnings.filterwarnings('ignore')
from feature import FeatureExtraction

file = open("model.pkl","rb")
gbc = pickle.load(file)
file.close()


app = Flask(__name__)
@app.route('/', methods=['GET', 'POST'])
def login():
    if request.method == 'POST':

        return redirect(url_for('index'))

    return render_template('login.html')


@app.route("/pred", methods=["GET", "POST"])
def index():
    if request.method == "POST":
```

index.html:

```html
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <meta name="description" content="This website is develop for identify the safety of url.">
    <meta name="keywords" content="phishing url,phishing,cyber security,machine learning,classifier,pytho
    <meta name="author" content="VAIBHAV BICHAVE">

    <!-- BootStrap -->
    <link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.5.0/css/bootstrap.min.css
        integrity="sha384-9aIt2nRpC12Uk9gS9baDl411NQApFmC26EwAOH8WgZl5MYYxFfc+NcPb1dKGj7Sk" crossorigin="
    <link href="static/styles.css" rel="stylesheet">
    <title>Web Phishing Detection IBM</title>

</head>

<body>
     <h2> BE AWARE OF YOUR CONFIDENTIAL DATA! </h2>
<div class=" container">
    <div class="row">
        <div class="form col-md" id="form1">
            <h2>PHISHING URL DETECTION</h2>

            <br>
            <form action="/pred" method ="POST">
                <input type="text" class="form__input" name ='url' id="url" placeholder="Drop Your URL He
                <label for="url" class="form__label">URL</label>
                <button class="button" role="button" >HIT TO PREDICT</button>
```

IBM cloud development:

In [72]:

```
#deploy
deployment=wml_client.deployments.create(
    artifact_uid=model_id,
    meta_props=deployment_props
)
```

```
#######################################################################################

Synchronous deployment creation for uid: '00151be4-5731-4fac-b86c-186f7b783fc3' started

#######################################################################################


initializing
Note: online_url is deprecated and will be removed in a future release. Use serving_urls instead.

ready


------------------------------------------------------------------------------------------------
Successfully finished deployment creation, deployment_uid='a770214b-0598-4fa1-8932-5085ebc7e152'
------------------------------------------------------------------------------------------------
```

Model Building:

7.Comparison of Models

In [ ]:
```
result =pd.DataFrame({'ML Model':Model,'Test Accuracy':test})
result
```

Out[ ]:

	ML Model	Test Accuracy
0	LogisticRegression	91.678
1	Random Forest	96.970
2	Decision Tree	96.291
3	K-Nearest Neighbours	96.970
4	SVM	96.970

8.Sorting

In [ ]:
```
result.sort_values(by=['Test Accuracy'], ascending=False)
```

Out[ ]:

	ML Model	Test Accuracy
1	Random Forest	96.970
3	K-Nearest Neighbours	96.970
4	SVM	96.970
2	Decision Tree	96.291
0	LogisticRegression	91.678

8. TESTING

8.1 TEST CASES

3. TestCaseAnalysis

ThisreportshowsthenumberoftestcasesthathavepassedfailedanduntestedbyourWebPhishingmodel

Section	TotalCases	NotTested	Fail	Pass
Login	20	0	5	15
RedirectingtoDetectionPage	30	0	7	23
Userinput	68	0	10	58
SecurityPopup	10	0	0	10
URLValidation	68	0	10	58
DetectionRate	70	0	49	21
RedirectingtotheGivenURL	70	0	54	16
FinalModelOutput	52	0	29	23

8.2 USER ACCEPTANCE TESTING

				Date	5-Nov-22								
				Team ID	PNT2022TMID29707								
				Project Name	Real time communication powered by AI for specially abled								
				Maximum Marks	4 marks								
Test case ID	Feature Type	Component	Test Scenario	Pre-Requisite	Steps To Execute	Test Data	Expected Result	Actual Result	Status	Comments	TC for Automation(Y/N)	BUG ID	Executed By
Detection_01	User Interface webpage	Register Page	Verify user is able to see the Registration page	Active server and internet connection with frontend code for Detection	1.Enter Website URL and Search the URL 2.Display the Register Page to the user	http://127.0.0.1:5000	Register Page will be display with the Process of Front end	Worked as expected	Active	User can view the registration page	Yes	--	PNT2022TMID29707 Team
Detection_02	User interface webpage	Register page	Verify user is able to register with user credentials	Active server and internet connection with frontend code for Detection	1.Enter Website URL and Search the URL 2.Display the Register Page to the User 3.Able to register in registration page	http://127.0.0.1:5000	Successfully registered	Worked as expected	Active	User can register	Yes	--	PNT2022TMID29707 Team
Detection_03	User interface webpage	Login page	Verify user is able to see the login page	Active server and internet connection with frontend code for Detection HTML Search Tag with the valid URL	1.Enter Website URL and .Search the URL 2.Display the login Page to the user	http://127.0.0.1:5000	Login Page will be display with the Process of Front end	Worked as expected	Active	user can view the login page	yes	--	PNT2022TMID29707 Team
Detection_04	User interface webpage	Login page	Verify user is able to login using user credentials	Active server and internet connection with frontend code for Detection HTML Search Tag with the valid URL	1.Enter Website URL and Search the URL 2.Display the Login Page to the User 3.Able to login in login page	http://127.0.0.1:5000	Successful login	Worked as expected	Active	User can login into web app	Yes	--	PNT2022TMID29707 Team
Detection_05	prediction	DetectionPage	Check the url	Algorithm using html to process the prediction	1.Enter URL . 2.hit to predict	https://google.com	Safety of URL is displayed	Worked as expected	Active	The model will protect the url	yes	--	PNT2022TMID29707 Team
					1.Enter the url		The predicted result						

9. RESULTS

9.1 PERFORMANCE METRICS:

Validation Method

Validation Method

```
#Logistic Regression
pred3= log_reg.predict(x_test)
s3=accuracy_score(y_test,pred3)*100
results('LR',s3)
print("Accuracy score of LR :",accuracy_score(y_test,pred3)*100)
#Random Forest
pred2=R_model.predict(x_test)
s2=accuracy_score(y_test,pred2)*100
print("Accuracy score  of RF :",s2)
results('RF',s2)
#KNN
pred5=modellin.predict(x_test)
s5=accuracy_score(y_test,pred5)*100
print("Accuracy score of KNN :",s5)
results('SVM-LIN',s5)
#SVM Linear
pred5=modellin.predict(x_test)
s5=accuracy_score(y_test,pred5)*100
print("Accuracy score SVM :",s5)
results('SVM-LIN',s5)
#SVM Poly
pred6 = modelpoly.predict(x_test)
s6=accuracy_score(y_test,pred6)*100
print("Accuracy score of SVM Poly :",s6)
results('SVM poly', s6)
```

```
Accuracy score of LR : 92.67299864314789
Accuracy score  of RF : 97.33152419719585
Accuracy score of KNN : 92.67299864314789
Accuracy score SVM : 92.67299864314789
Accuracy score of SVM Poly : 95.74853007688829
```

10. ADVANTAGES AND DISADVANTAGES

10.1 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.

10.2 DISADVANTAGES

- If Internet connection fails, this system won't work.
- All websites related data will be stored in one place
- It will need much more memory to safe encrypted message, and decrypt receiving messages.
- Process takes longer time than normal communication.
- Sometimes there will be a delay in popup message.

11. CONCLUSION

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 defence 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. With the help of machine learning algorithms like, Random Forest, Decision Tree, Neural network and Linear model we can classify data into phishing, suspicious and legitimate. This can be done based on unique features of phishing websites and user does not need to check individual websites.

We've trained and evaluated supervised ML algorithms with our dataset. We tried training with four different algorithms and tested the accuracy score of each algorithm and choose the best algorithm for our model. We split the dataset into 7:3, 70% for training and 30% for testing. The results of each algorithm are mentioned below:

- Logistic Regression-0.9167
- Random forest-0.9687
- Decision Tree-0 9620
- KNN-(K nearest neighbour)-0.9696
- SVC (Support vector Machine)-0.9696

12. FUTURE SCOPE

The existing project can achieve more security and we can extend it to detect phishing with higher accuracy in the future, we can create an app filter that can scan app data before installation. We can check the app Data like the permission that the app is asking for and reviews of users to see whether the app is good or not, also according to search majority of the phishing links are shared via emails to the target user. So it is necessary to add a filter over emails also. Hence we can add filter phishing detection filter also and check all emails before clicking any links in it.

13. APPENDIX

GitHub Link: GitHub

Demo Video Link: video link