WEB PHISHING DETECTION

NALAIYA THIRAN PROJECT BASED LEARNING ON PROFESSIONAL READLINESS FOR INNOVATION, EMPLOYMENT AND ENTERPRENEURSHIP

A PROJECT REPORT

TEAM ID: PNT2022TMID33847

- JEYANTHI LAKSHMI G 950819106022
- BALAVIKA K 950819106010
- LAVANYA G 950819106034
- INDUMATHI P 950819106308

BACHELOR OF ENGINEERING IN ELECTRONICS AND COMMUNICATION

GOVERNMENT COLLEGE OF ENGINEERING, TIRUNELVELI.

PROJECT REPORT

Project Name	WEB PHISHING DETECTION
Team ID	PNT2022TMID33847

Team Members:

- *Jeyanthi Lakshmi G*
- *Balavika K*
- *Lavanya G*
- *Indumathi P*

1. INTRODUCTION:

Phishing is described as **a fraudulent activity that is done to steal confidential user information such as credit card numbers, login credentials, and passwords**. It is usually done by using email or other forms of electronic communication by pretending to be from a reliable business entity.

There are a number of users who purchase products online and make payments through e-banking. There are e-banking websites that ask users to provide sensitive data such as username, password & credit card details, etc., often for malicious reasons. This type of **e-banking website is known as a phishing website.** Web service is one of the key communications software services for the Internet. Web phishing is one of many security threats to web services on the Internet.

Common threats of web phishing:

- Web phishing aims to **steal private information**, such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity.
- It will lead to information disclosure and property damage.
- Large organizations may get trapped in different kinds of scams.

Our project mainly focuses on **applying a machine-learning algorithm to detect Phishing websites.**

PROJECT OVERVIEW:

The project aims at creating a website for detecting phishing websites.

The project involves the following steps:

- Collection of dataset consisting of URL features.
- Preprocess or clean the data.
- Analyze the pre-processed data.
- Train the machine with preprocessed data using an appropriate machine learning algorithm.
- Save the model and its dependencies.
- Build a Web application using a flask that integrates with the model built.

PURPOSE:

- The purpose of web phishing detection is to prevent stealing of sensitive user information like passwords.
- The main purpose of this project is to help large organizations and individuals to perform safe and secure online transactions.
- This prevents money loss and property damage.

2. LITERATURE SURVEY:

A survey and classification of web phishing detection schemes

Authors: Gaurav Varshney, Manoj Misra, Pradeep K.Atrey
Published on: 26 October 2016.

Abstract:

Phishing is a fraudulent technique that is used over the Internet to deceive users with the goal of extracting their personal information such as username, passwords, credit card, and bank account information. The key to phishing is deception. Phishing uses email spoofing as its initial medium for deceptive communication followed by spoofed websites to obtain the needed information from the victims. Phishing was discovered in 1996, and today, it is one of the most severe cybercrimes faced by the Internet users. Researchers are working on the prevention, detection, and education of phishing attacks, but to date, there is no complete and accurate solution for thwarting them. This paper studies, analyzes, and classifies the most significant and novel strategies proposed in the area of phished website detection, and outlines their advantages and drawbacks. Furthermore, a detailed analysis of the latest schemes proposed by researchers in various subcategories is provided. The paper identifies advantages, drawbacks, and research gaps in the area of phishing website detection that can be worked upon in future research and developments. The analysis given in this paper will help academia and industries to identify the best anti-phishing technique.

Web Phishing Detection Using a Deep Learning Framework

Authors: Ping Yi, Yuxiang Guan, Futai Zou, Yao Yao, Wei Wang, Ting Zhu.
Published on: 26 Sept 2018

Abstract:

Web service is one of the key communications software services for the Internet. Web phishing is one of many security threats to web services on the Internet. Web phishing aims to steal private information, such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity. It will lead to information disclosure and property damage. This paper mainly focuses on applying a deep learning framework to detect phishing websites. This paper first designs two types of features for web phishing: original features and interaction features. A detection model based on Deep Belief Networks (DBN) is then presented. The test using real IP flows from ISP (Internet Service Provider) shows that the detecting model based on DBN can achieve an approximately 90% true positive rate and 0.6% false positive rate.

Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text

Authors: M.A. Adebowale, K.T. Lwin, E. Sánchez, M.A. Hossain
Published on: 25 April 2018

Abstract:

A phishing attack is one of the most significant problems faced by online users because of its enormous effect on the online activities performed. In recent years, phishing attacks continue to escalate in frequency, severity and impact. Several solutions, using various methodologies, have been proposed in the literature to counter the web-phishing threats. Notwithstanding, the existing technology cannot detect the new phishing attacks accurately due to the insufficient integration of features of the text, image and frame in the evaluation process. The use of related features of images, frames and text of legitimate and non legitimate websites and associated artificial intelligence algorithms to develop an integrated method to address these together. This paper presents an Adaptive Neuro-Fuzzy Inference System (ANFIS) based robust scheme using the integrated features of the text, images and frames for web-phishing detection and protection.

Systematization of Knowledge (SoK): A Systematic Review of Software-Based Web Phishing Detection

Authors: Zuochao Dou; Issa Khalil; Abdallah Khreishah; Ala Al-Fuqaha; Mohsen Guizani
Published on: 13 September 2017

Abstract:

Phishing is a form of cyber attack that leverages social engineering approaches and other sophisticated techniques to harvest personal information from users of websites. The average annual growth rate of the number of unique phishing websites detected by the Anti Phishing Working Group is 36.29% for the past six years and 97.36% for the past two years. In the wake of this rise, alleviating phishing attacks has received a growing interest from the cyber security community. Extensive research and development have been conducted to detect phishing attempts based on their unique content, network, and URL characteristics. Existing approaches differ significantly in terms of intuitions, data analysis methods, as well as evaluation methodologies. This warrants a careful systematization so that the advantages and limitations of each approach, as well as the applicability in different contexts, could be analyzed and contrasted in a rigorous and principled way. This paper presents a systematic study of phishing detection schemes, especially software based ones. Starting from the phishing detection taxonomy, we study evaluation datasets, detection features, detection techniques, and evaluation metrics. Finally, we provide insights that we believe will help guide the development of more effective and efficient phishing detection schemes.

Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions

Authors: M. Vijayalakshmi,S. Mercy Shalinie,Ming Hour Yang,Raja Meenakshi U.
First published: 23 September 2020

Abstraction:

Internet dragged more than half of the world's population into the cyber world. Unfortunately, with the increase in internet transactions, cybercrimes also increase rapidly. With the anonymous structure of the internet, attackers attempt to deceive the end-users through different forms namely phishing, malware, SQL injection, man-in-the-middle, domain name system tunnelling, ransomware, web trojan, and so on. Amongst them, phishing is the most deceiving attack, which exploits the vulnerabilities in the end-users. Phishing is often done through emails and malicious websites to lure the user by posing themselves as a trusted entity. Security experts have been proposing many anti-phishing techniques. Till today there is no single solution that is capable of mitigating all the vulnerabilities. A systematic review of current trends in web phishing detection techniques is carried out and a taxonomy of automated web phishing detection is presented. The objective of this study is to acknowledge the status of current research in automated web phishing detection and evaluate their performance. This study also discusses the research avenues for future investigation.

Using Domain Top-page Similarity Feature in Machine Learning-Based Web Phishing Detection

Authors: Nuttapong Sanglerdsinlapachai; Arnon Rungsawang

Published on: 2010 Third International Conference on Knowledge Discovery and Data Mining.

Abstract:

This paper presents a study on using a concept feature to detect web phishing problem. Following the features introduced in Carnegie Mellon Anti-phishing and Network Analysis Tool (CANTINA), we applied additional domain top-page similarity feature to a machine learning based phishing detection system. We preliminarily experimented with a small set of 200 web data, consisting of 100 phishing webs and another 100 non-phishing webs. The evaluation result in terms of f-measure was up to 0.9250, with 7.50% of error rate.

Web Phishing Detection Based on Page Spatial Layout Similarity

Authors: Weifeng, Hua Lu, Baowen Xu, Hongji Yang
Published On: July 8, 2012

Abstract:

Web phishing is becoming an increasingly severe security threat in the web domain. Effective and efficient phishing detection is very important for protecting web users from loss of sensitive private information and even personal properties. One of the keys of phishing detection is to efficiently search the legitimate web page library and to find those page that are the most similar to a suspicious phishing page. Most existing phishing detection methods are focused on text and/or image features and have paid very limited attention to spatial layout characteristics of web pages. In this paper, we propose a novel phishing detection method that makes use of the informative spatial layout characteristics of web pages. In particular, we develop two different options to extract the spatial layout features as rectangle blocks from a given web page. Given two web pages, with their respective spatial layout features, we propose a page similarity definition that takes into account their spatial layout characteristics. Furthermore, we build an R-tree to index all the spatial layout features of a legitimate page library. As a result, phishing detection based on the spatial layout feature similarity is facilitated by relevant spatial queries via the R-tree. A series of simulation experiments are conducted to evaluate our proposals. The results demonstrate that the proposed novel phishing detection method is effective and efficient.

Machine learning based phishing detection from URLs

Authors: Ozgur KoraySahingozaEbubekirBuberbOnderDemirbBanuDiric .
Published on: 7 May 2018

Abstract:

Due to the rapid growth of the Internet, users change their preference from traditional shopping to the electronic commerce. Instead of bank/shop robbery, nowadays, criminals try to find their victims in the cyberspace with some specific tricks. By using the anonymous structure of the Internet, attackers set out new techniques, such as phishing, to deceive victims with the use of false websites to collect their sensitive information such as account IDs, usernames, passwords, etc. Understanding whether a web page is legitimate or phishing is a very challenging problem, due to its semantics-based attack structure, which mainly exploits the computer users' vulnerabilities. Although software companies launch new anti phishing products, which use blacklists, heuristics, visual and machine learning-based approaches, these products cannot prevent all of the phishing attacks. In this paper, a real time anti-phishing system, which uses seven different classification algorithms and natural language processing (NLP) based features, is proposed. The system has the following distinguishing properties from other studies in the literature: language independence, use of a huge size of phishing and legitimate data, real-time execution, detection of new websites, independence from third-party services and use of feature-rich classifiers. For measuring the performance of the system, a new dataset is constructed, and the experimental results are tested on it. According to the experimental and comparative results from the implemented classification algorithms, Random Forest algorithm with only NLP based features gives the best performance with the 97.98% accuracy rate for detection of phishing URLs.

Phishing Detection Based on Machine Learning and Feature Selection Methods

Authors: Mohammad Almseidin, AlMaha Abu Zuraiq, Mouhammd Al-kasassbeh, Nidal Alnidami

Abstract: With increasing technology developments, the Internet has become everywhere and accessible by everyone. There are a considerable number of web-pages with different benefits. Despite this enormous number, not all of these sites are legitimate. There are so called phishing sites that deceive users into serving their interests. This paper dealt with this problem using machine learning algorithms in addition to employing a novel dataset that related to phishing detection, which contains 5000 legitimate web-pages and 5000 phishing ones. In order to obtain the best results, various machine learning algorithms were tested. Then J48, Random forest, and Multilayer perceptron were chosen. Different feature selection tools were employed to the dataset in order to improve the efficiency of the models. The best result of the experiment achieved by utilizing 20 features out of 48 features and applying it to Random forest algorithm. The accuracy was 98.11%

PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning

Authors: Ankit Kumar Jain, B. B. Gupta

Abstract:

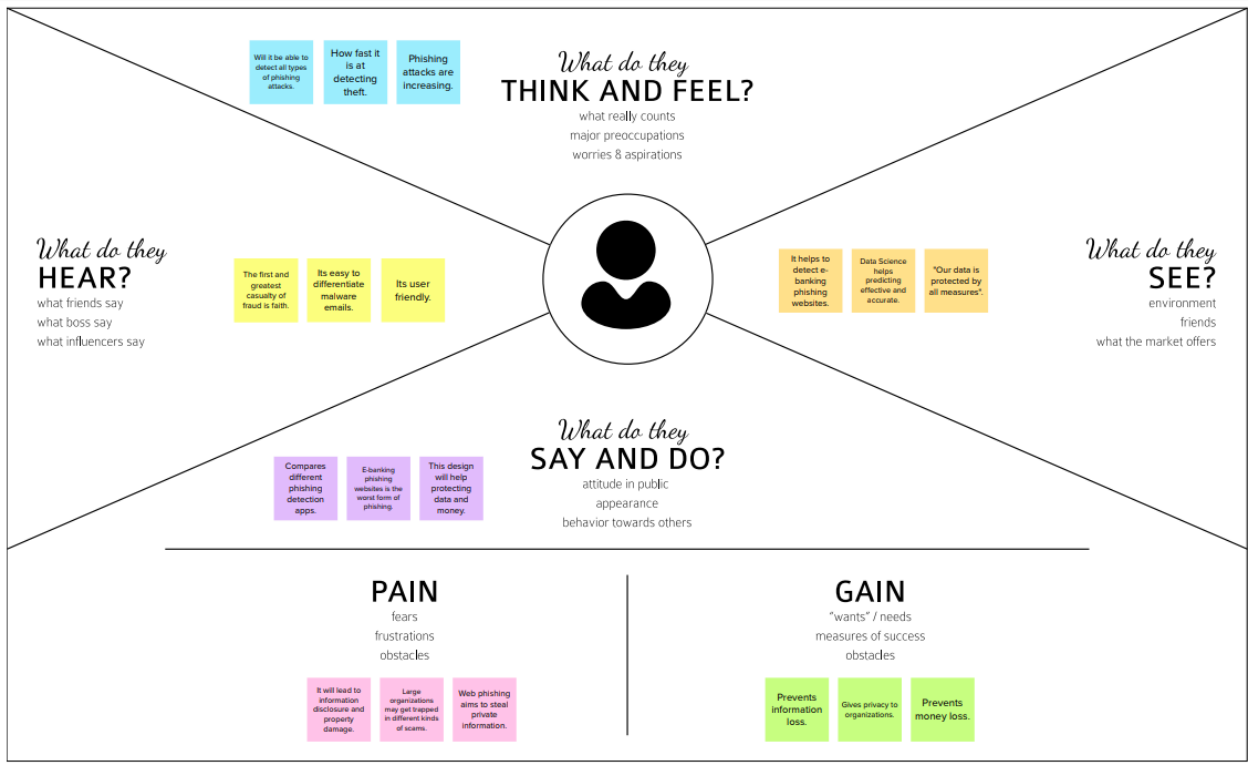
Today, phishing is one of the most serious cyber-security threat in which attackers steal sensitive information such as personal identification number (PIN), credit card details, login, password, etc., from Internet users. In this paper, we proposed a machine learning based anti phishing system (i.e., named as PHISH-SAFE) based on Uniform Resource Locator (URL) features. To evaluate the performance of our proposed system, we have taken 14 features from URL to detect a website as a phishing or non-phishing. The proposed system is trained using more than 33,000 phishing and legitimate URLs with SVM and Naïve Bayes classifiers. Our experiment results show more than 90% accuracy in detecting phishing websites using SVM classifier.

PROBLEM STATEMENT DEFINITION:

The main purpose of the project is to detect the fake or phishing websites who are trying to get access to the sensitive data or by creating the fake websites and trying to get access of the user personal credentials.

3. IDEATION AND PROPOSED SOLUTION:

EMPATHY MAP CANVAS:



WEB PHISHING DETECTION

IDEATION AND BRAINSTORMING:

https://app.mural.co/invitation/mural/ml6371/1664244835456?sender=ud1a3ab4dde794ed4f2e89750&key=8b1b8f1d-ca1a-4d72-893a-0d47866e5766

PROPOSED SOLUTION:

S.NO	PARAMETERS	DESCRIPTION
1.	**Problem Statement (Problem to be solved)**	Phishing is a major problem, which uses both social engineering and technical deception to get users' important information such as financial data, emails, and other private information.
2.	**Idea / Solution description**	Use anti-phishing protection and anti-spam software to protect yourself when malicious message slip through to your computer.
3.	**Novelty / Uniqueness**	Proposed web technology features improve phishing detection accuracy.
4.	**Social Impact / Customer Satisfaction**	Organizations and individuals can protect their data and maintain privacy. It provides safe and secure money transactions. So customers are well satisfied and design has positive social impacts.
5.	**Business Model (Revenue Model)**	Our model targets customers and organizations that use online money transaction. This will help gain profit.
6.	**Scalability of the Solution**	The design will be suitable and performs with full efficiency according to rising demands. The performance of model does not change with existing external situations.

PROBLEM SOLUTION FIT:

4.REQUIREMENT ANALYSIS:

Functional Requirement:

Following are the functional requirements of the proposed solution.

FR No.	**Functional Requirement (Epic)**	**Sub Requirement (Story/ Sub-Task)**
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIN
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	Objective	Todetect phishing websites.
FR-4	Area of focus	The main area of focus is organizations where ecommerce plays a major roleand in areas where confidentiality of data is required

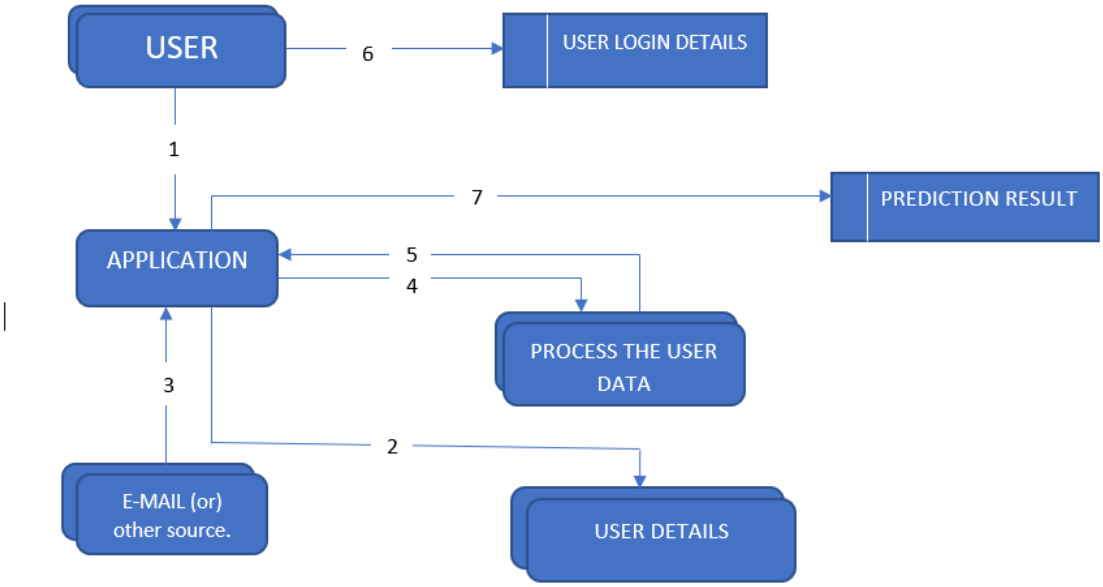
NON FUNCTIONAL REQUIREMENTS:

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

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	A set of specifications that describe the system's operationcapabilities and constraints andattempt to improve its functionality.
NFR-2	**Security**	Assuring all data inside the system or its part will be protected against malware attacks or unauthorized access.
NFR-3	**Reliability**	Our model is based on machine learning. The approach showed an accuracy of 98.3% which is sofar the best integrated solution of web phishing detection.
NFR-4	**Performance**	Performance basically gives system parameters to reach ourgoal. Parameters for the proposed system are accurate predicted value which is compared to the existing system.
NFR-5	**Availability**	The system is accessible to a user at any givenpoint in time.
NFR-6	**Scalability**	The design will be suitable andperforms with full efficiency according to rising demands. The performance of model doesnot changewith existing external situations.

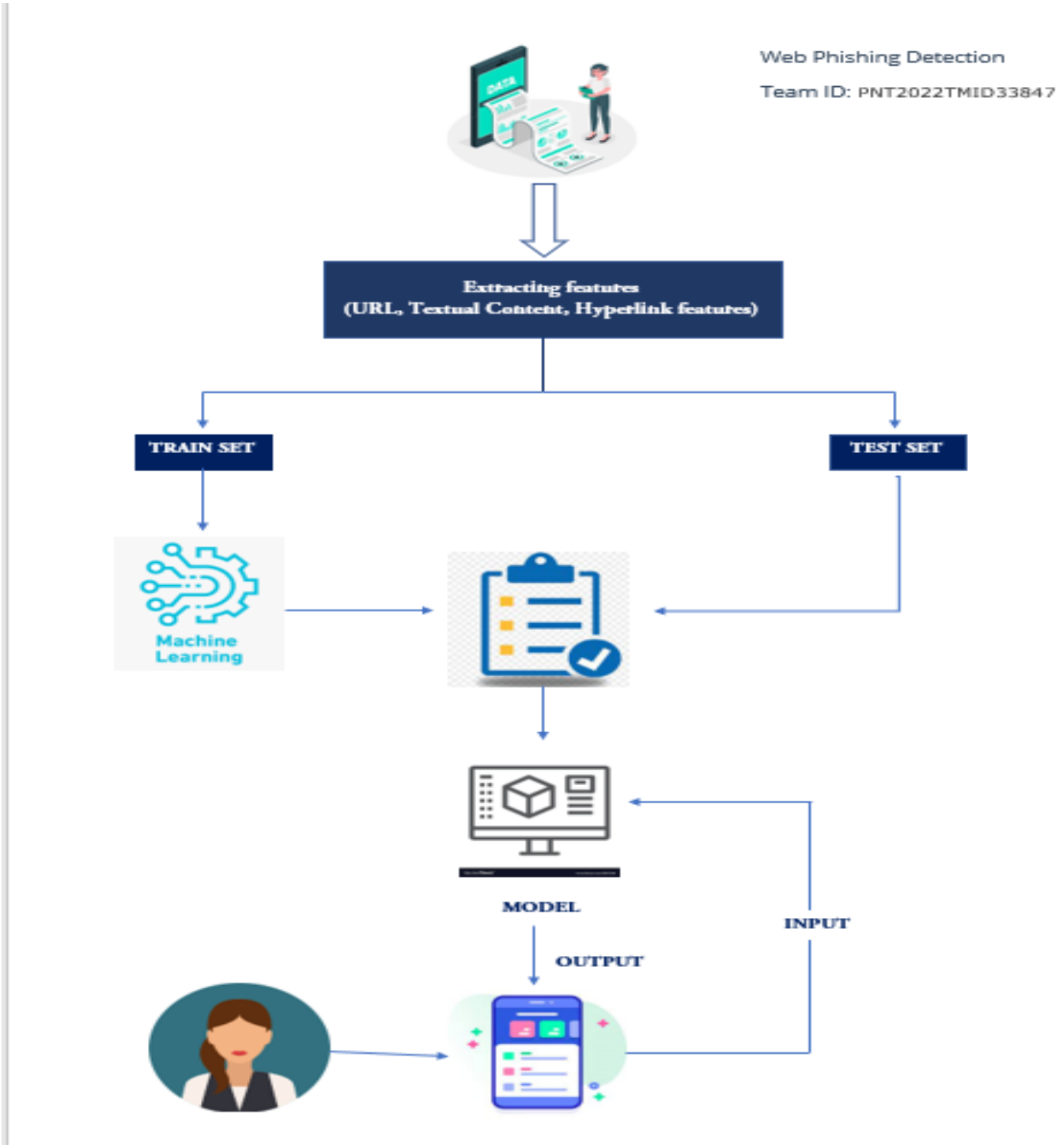
5. PROJECT DESIGN:

DATA FLOW DIAGRAMS:

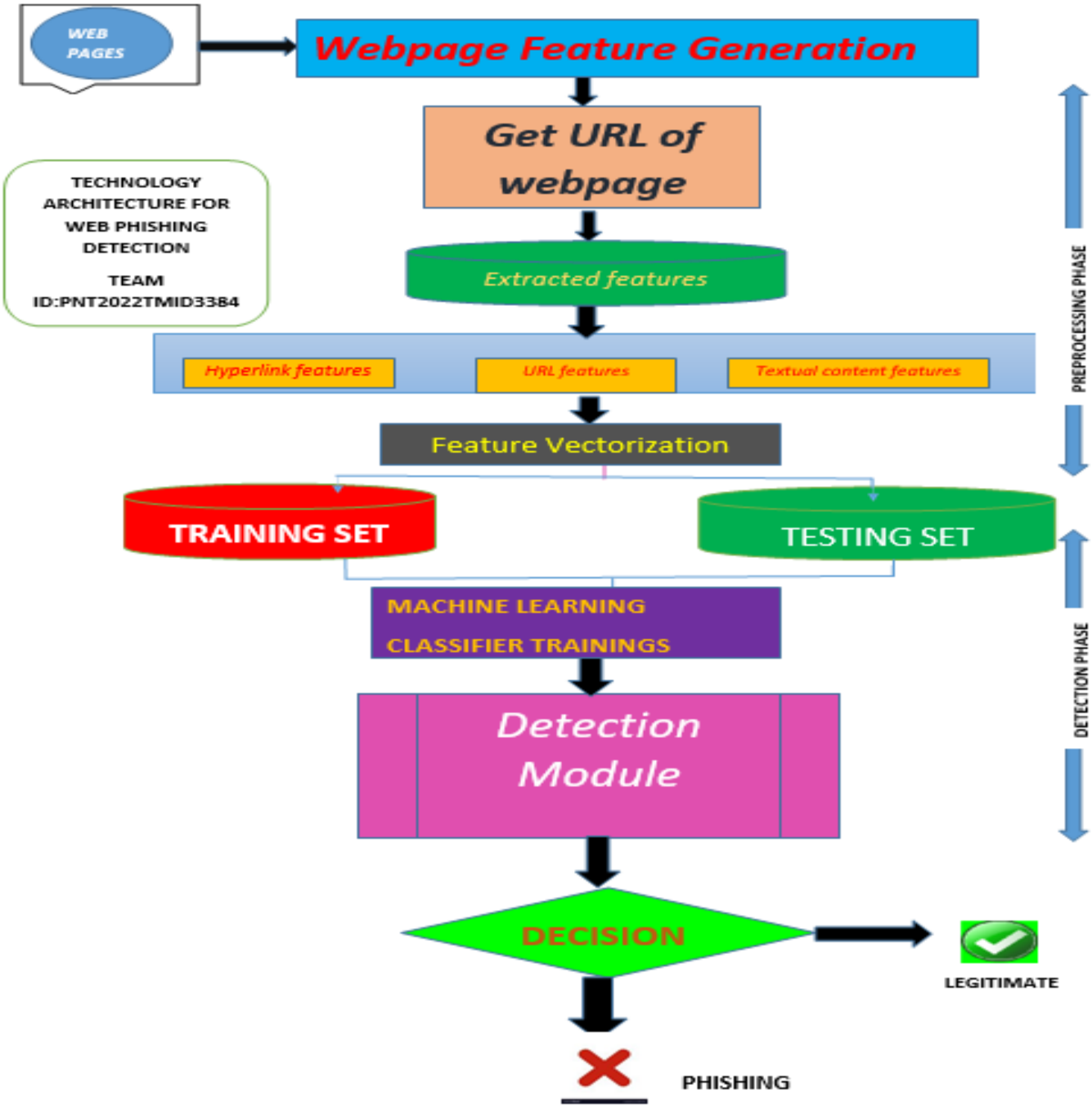


1. User interacts with web phishing detection app with his valid login credentials.
2. App checks if the user is a valid user.
3. The user pastes the unknown url in the app.
4. The app processes the user data.
5. The app displays the resulting prediction.
6. The user data is saved for further use.
7. The prediction results are saved for further use.

SOLUTION ARCHITECTURE:



TECHNICAL ARCHITECTURE:



USER STORIES:

User Type	Functional Requirement	User Story Number	User Story /Task	Acceptance Criteria	Priority	Release
Customer (Mobile User)	Registration	US1	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
		US2	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
		US3	As a user, I can register for the application through Facebook.	I can register & access the dashboard with Facebook Login.	Low	Sprint-2
		US4	As a user, I can register for the application through Gmail.	I can register & access the dashboard with Gmail Login.	Medium	Sprint-1

	Login	US5	As a user, I can log into the application by entering email & password.	I can login to the app using the same email and password and access the resources.	High	Sprint-1
	Dashboard	US6	As a user, I can easily navigate through dashboard and I can use the dashboard to get details about app and instruction to use the app.	I can login to the app using the same email and password and access the resources.	High	Sprint-1
Customer (Web user)	Login and Dashboard	US7	As a web app user, I can login into application by using my email and password and I can access all resources same as mobile users.	I can login to the app using the same email and password and access the resources.	High	Sprint-1
Customer Care Executive	Login	CCE1	As a CCE I can login to app using my id and password and I can interact with	I can login using my mail and password.	High	Sprint-1

			user.			
	Dashboard	CCE2	As a CCE I can access dashboard using id and password and I can see all user queries, explain app usage and attend their queries.	I can login using my mail and password.	High	Sprint-1
Administrator	Login and Dashboard	A1	As an administrator, I can login and access dashboard and manage and direct activites.	I can login using my company id and password.	High	Sprint-1

6.PROJECT PLANNING AND SCHEDULING:

SPRINT PLANNING AND ESTIMATION:

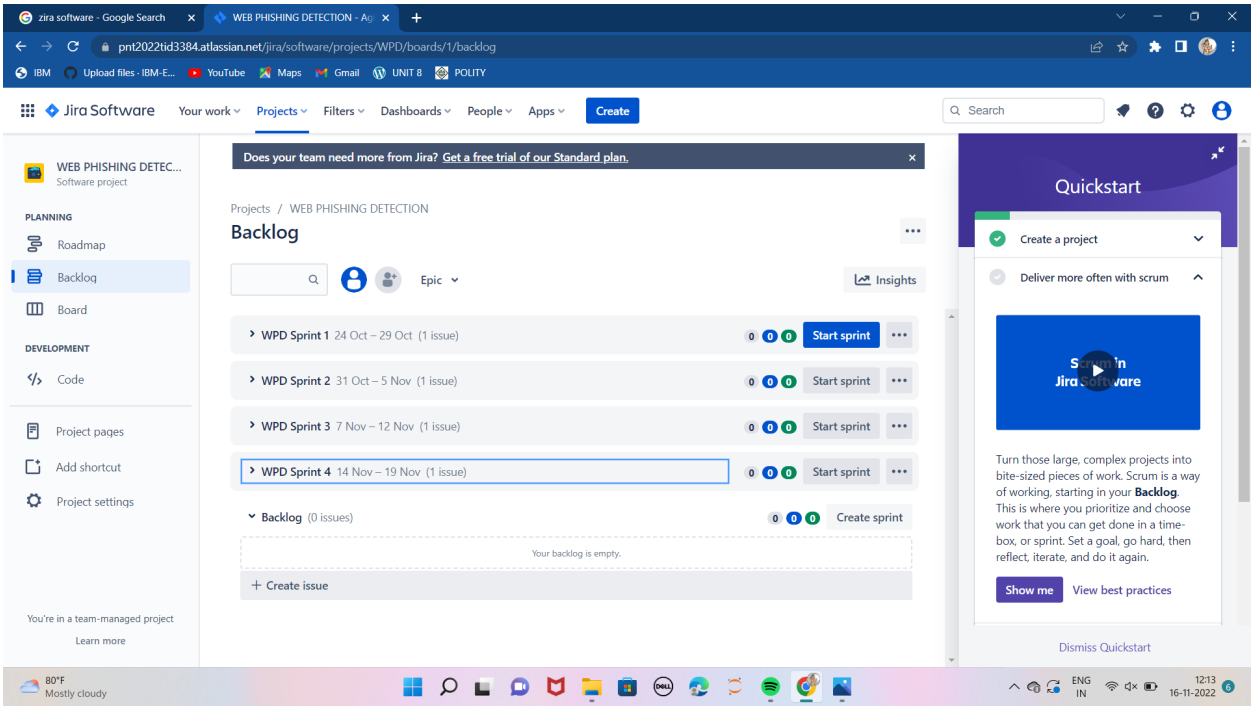
Sprint	Functional Requirement	User Story Number	User Story /Task	Story Points	Priority	Team Members
Sprint-1	Registration	US1	As a user, I can register for the application by entering my email, password, and confirming my password.	2	High	Jeyanthi Lakshmi G
Sprint-1		US2	As a user, I will receive confirmation email once I have registered for the application.	2	High	Balavika K
Sprint-2		US3	As a user, I can register for the application through Facebook.	2	Low	Indumathy P
Sprint-1		US4	As a user, I can register for the application through Gmail.	2	Medium	Lavanya G
Sprint-1	Login	US5	As a user, I can log into the application by entering email & password.	2	High	Jeyanthi Lakshmi G
Sprint-1	Dashboard	US6	As a user, I can easily navigate through dashboard and I can use the dashboard to get details about app and instruction to use the app.	2	High	Balavika K
Sprint-1	Login and Dashboard	US7	As a web app user, I can login into application by using my email and password and I can access all resources	2	High	Indumathy P Lavanya G

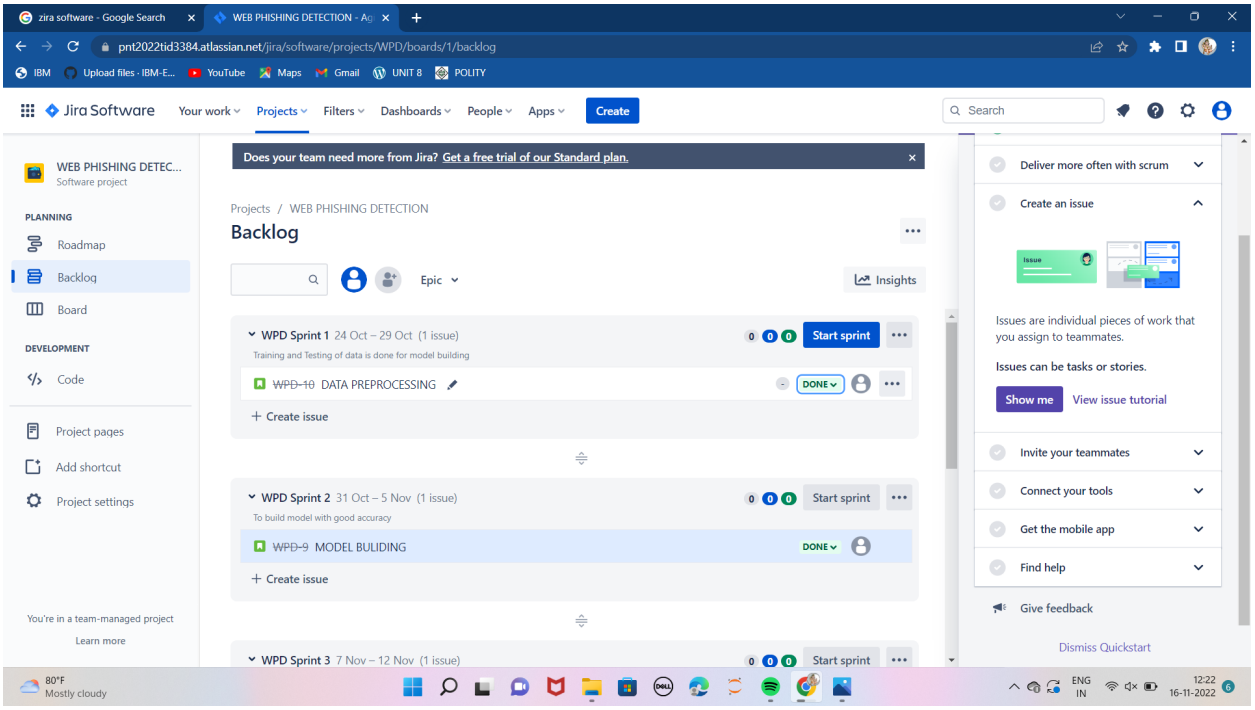
			same as mobile users.			
Sprint-1	Login	CCE1	As a CCE I can login to app using my id and password and I can interact with user.	2	High	Jeyanthi Lakshmi G
Sprint-1	Dashboard	CCE2	As a CCE I can access dashboard using id and password and I can see all user queries, explain app usage and attend their queries.	2	High	Balavika K
Sprint-1	Login and Dashboard	A1	As an administrator, I can login and access dashboard and manage and direct activities.	2	High	Indumathy P Lavanya G

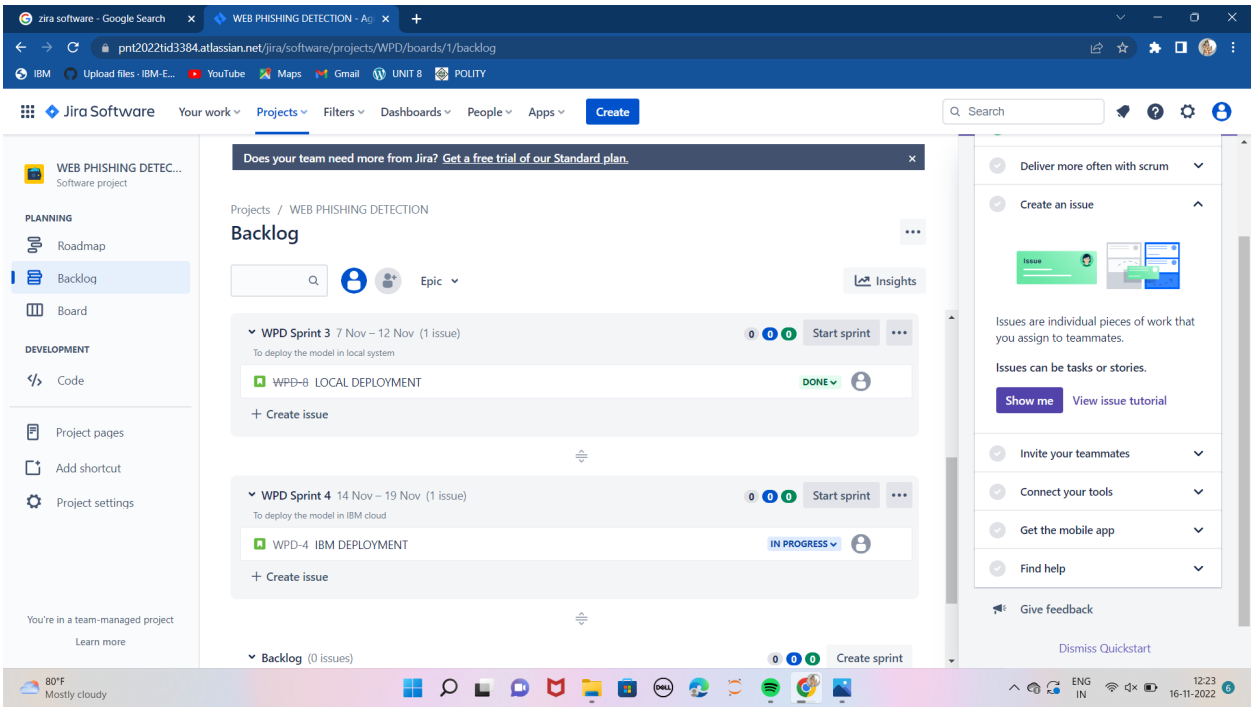
SPRINT DELIVERY SCHEDULE:

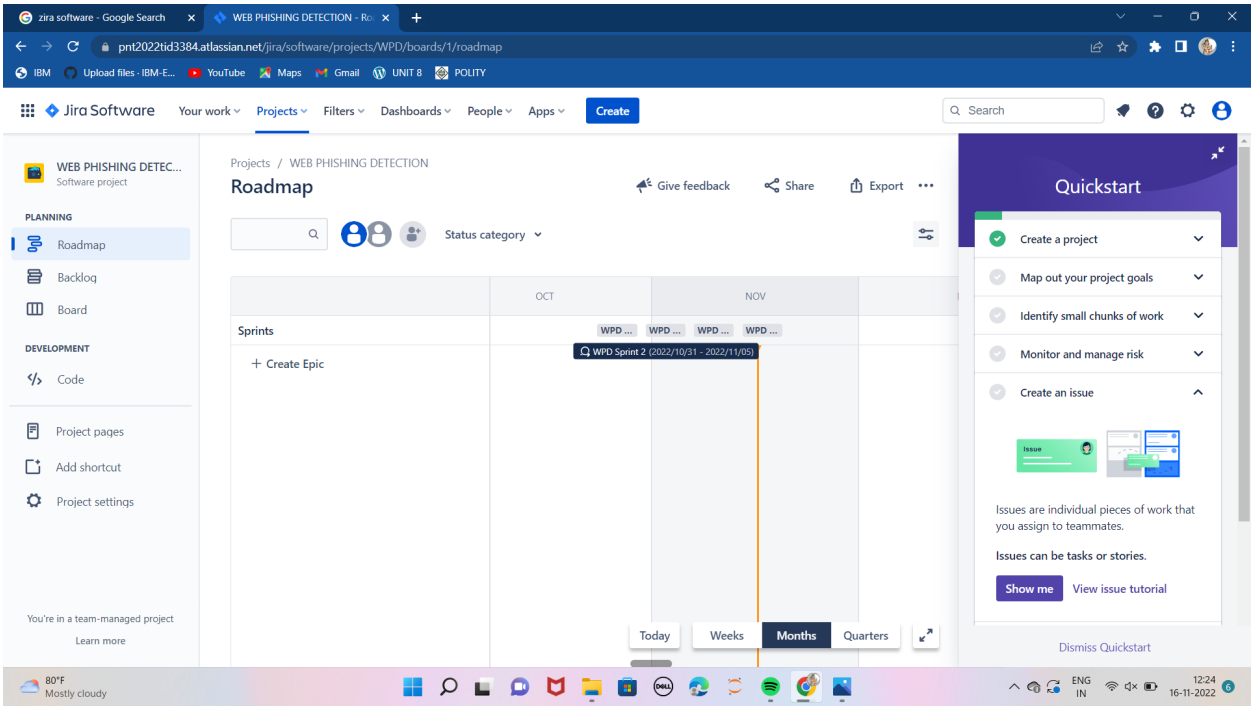
Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

REPORTS FROM JIRA:






zira software - Google Search
WEB PHISHING DETECTION - Ag
pnt2022tid3384.atlassian.net/jira/software/projects/WPD/boards/1/backlog
Jira Software
Your work
Projects
Filters
Dashboards
People
Apps
Create
Search
WEB PHISHING DETEC...
Software project
PLANNING
Roadmap
Backlog
Board
DEVELOPMENT
Code
Project pages
Add shortcut
Project settings
You're in a team-managed project
Learn more
Does your team need more from Jira? Get a free trial of our Standard plan.
Projects / WEB PHISHING DETECTION
Backlog
Epic
Insights
WPD Sprint 3 7 Nov – 12 Nov (1 issue)
To deploy the model in local system
Start sprint
WPD-8 LOCAL DEPLOYMENT
DONE
Create issue
WPD Sprint 4 14 Nov – 19 Nov (1 issue)
To deploy the model in IBM cloud
WPD-4 IBM DEPLOYMENT
IN PROGRESS
Backlog (0 issues)
Create sprint
Deliver more often with scrum
Create an issue
Issues are individual pieces of work that you assign to teammates.
Issues can be tasks or stories.
Show me
View issue tutorial
Invite your teammates
Connect your tools
Get the mobile app
Find help
Give feedback
Dismiss Quickstart
80°F
Mostly cloudy
ENG
IN
12:23
16-11-2022


zira software - Google Search
WEB PHISHING DETECTION - Ro
pnt2022tid3384.atlassian.net/jira/software/projects/WPD/boards/1/roadmap
Jira Software
Your work
Projects
Filters
Dashboards
People
Apps
Create
Search
WEB PHISHING DETEC...
Software project
PLANNING
Roadmap
Backlog
Board
DEVELOPMENT
Code
Project pages
Add shortcut
Project settings
You're in a team-managed project
Learn more
Projects / WEB PHISHING DETECTION
Roadmap
Give feedback
Share
Export
Status category
OCT
NOV
Sprints
WPD ...
WPD Sprint 2 (2022/10/31 - 2022/11/05)
Create Epic
Today
Weeks
Months
Quarters
Quickstart
Create a project
Map out your project goals
Identify small chunks of work
Monitor and manage risk
Create an issue
Issues are individual pieces of work that you assign to teammates.
Issues can be tasks or stories.
Show me
View issue tutorial
Dismiss Quickstart
80°F
Mostly cloudy
ENG
IN
12:24
16-11-2022

7.CODING & SOLUTIONING:

FEATURES:

RANDOM FOREST CLASSIFIER ALGORITHM:

Random Forest Classifier Algorithm

```
In [32]: # model building
         from sklearn.ensemble import RandomForestClassifier
         rf=RandomForestClassifier(n_estimators=15,max_depth=3)

In [33]: rf.fit(X_train,y_train)

Out[33]: RandomForestClassifier(max_depth=3, n_estimators=15)

In [34]: test_pred = rf.predict(X_test)

In [35]: train_pred = rf.predict(X_train)
```

A random forest classifier. A random forest is **a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting**.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. **It performs better results for classification problems.**

Below are some points that explain why we should use the Random Forest algorithm:

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

8.USER ACCEPTANCE TESTING:

1. Purpose of Document:

The purpose of this document is to briefly explain the test coverage and open issues of the Web Phishing Detection project at the time of the release to User Acceptance Testing (UAT).

2. Defect Analysis:

This report shows the number of resolved or closed bugs at each severity level, and how they were resolved.

Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	10	4	2	3	20
Duplicate	1	0	3	0	4
External	2	3	0	1	6
Fixed	11	2	4	20	37
Not Reproduced	0	0	1	0	1
Skipped	0	0	1	1	2
Won't Fix	0	5	2	1	8
Totals	24	14	13	26	77

3. Test Case Analysis:

This report shows the number of test cases that have passed, failed, and untested.

Section	Total Cases	Not Tested	Fail	Pass
Print Engine	7	0	0	7
Client Application	51	0	0	51
Security	2	0	0	2
Outsource Shipping	3	0	0	3
Exception Reporting	9	0	0	9
Final Report Output	4	0	0	4
Version Control	2	0	0	2

9.RESULTS:

PERFORMANCE METRICS:

- ➤ Performance metrics is used to check whether the model is best for the given problem.
- ➤ Evaluation metrics used for classification machine learning algorithm are ***Accuracy score, Confusion metrics, Classification report: precision, recall, f1 score.***

EVALUATION METRICS

In [36]:
```
# evaluating the model
from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
```

In [37]:
```
print('Testing accuracy: ' , accuracy_score(y_test,test_pred))
print('Training accuracy: ',accuracy_score(y_train,train_pred))
```

```
Testing accuracy:  0.9158878504672897
Training accuracy:  0.9167743602998191
```

In [38]:
```
pd.crosstab(y_test,test_pred)
```

Out[38]:

col_0	-1	1
Result		
-1	1332	166
1	113	1706

In [39]:
```
print(classification_report(y_test,test_pred))
```

```
              precision    recall  f1-score   support

          -1       0.92      0.89      0.91      1498
           1       0.91      0.94      0.92      1819

    accuracy                           0.92      3317
   macro avg       0.92      0.91      0.91      3317
weighted avg       0.92      0.92      0.92      3317
```

10.ADVANTAGES:

- High Level of Accuracy.
- Mitigate zero hour attacks.
- Constuct own classification models.

DISADVANTAGES:

- Time consuming.
- Costly.
- Need large mail server and high memory requirement.

11.CONCLUSION:

We achieved 97.14% detection accuracy using random forest algorithm with lowest false positive rate.

In future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used.

12.FUTURE SCOPE:

For future enhancements, we intend to build the phishing detection system as a scalable web service which will incorporate online learning so that new phishing attack patterns can easily be learned and improve the accuracy of our models with better feature extraction.

13.APPENDIX:

GITHUB REPOSITORY:

https://github.com/IBM-EPBL/IBM-Project-30253-1660142900

PROJECT DEMO LINK:

https://drive.google.com/drive/folders/1Y3CSf_fMXjvSuTL6cWh4tJ8mkukF8wu5?usp=sharing