.

WEB PHISHING DETECTION

Submitted by

K.MOHAMED AMEER KHAN (Team Leader) (950619106012)

S.JOSEPH SELVIN (Team Member) (950619106004)

A.PRAVEEN KUMAR (Team Member) (950619106017)

B.KALEESWAR@PRAVIN (Team Member) (950619106302)

TEAM ID PNT2022TMID49921

For the project

HX8001 PROFESSIONAL READYINESS FOR INNOVATION

EMPLOYABLITY AND ENTREPRENEURSHIP

in the department of

ELECTRONICS AND COMMUNICATION ENGINEERING

EINSTEIN COLLEGE OF ENGINEEING, TIRUNELVELI-627 012

ANNA UNIVERSITY : CHENNAI 600 025

NOVEMBER:2022

BONAFIED CERTIFICATE

Certified this Report "WEB PHISHING DETECTION", for the project, is the bonafied work of "**K.MOHAMED AMEERKHAN(950619106012), A.PRAVEEN KUMAR (950619106017),S. JOSEPH SELVIN (950619106004) B.KALEESWAR@PRAVIN (950619106302)** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported here in does not form part of any other thesis or dissertation on the basis of which a degree or award was co-offered on the earlier occasion on this or any other candidate.

SIGNATURE
Dr. P. REVATHY
HOD/ECE

SIGNATURE
Dr. P.REVATHY
MENTOR

EVALUATOR
Mrs. T. Viji
AP/ECE

SPOC
Mr.S.ARUNSINGH
AP/ECE

ACKNOWLEDGEMENT

We have successfully completed the project with blessings showered onus by god, the almighty, A project of this nature needs co-operation and support from many for successful completion.

We express our heartfelt thanks to **Mr.A.MATHIVANAN, M.sc.,(Agri),** Managing Trustee of Einstein college of Engineering, Tirunelveli, for his mortal support and device.

Our thanks to **Prof.A.AMUTHAVANAN, BE., M.S (USA).B.L,**Chairman of our college for making necessary arrangements to do this project.

Our heartly thanks to **Prof.A.EZHILVANAN, MBA.,** Secretary of our college for making necessary arrangements to do this project.

We wish to express our gratitude to **Dr.VELAYUTHAM,M.E,Ph.D.,FIE.** Principal for the support he provided us to carry out this project successfully.

We are very much thankful to **Dr.P.REVATHY** , Head of the Department , Electronics and communication Engineering who is always a constant of inspiring us.

We are extend our sincere thanks to our project evaluator **Mrs.T.VIJI M.E, SPOC Mr.S.ARUNSINGH M.Tech** friends for their help in completing this project.

ABSTRACT

With raising in-depth amalgamation of the Internet and social life, the Internet is looking differently at how people are learning and working, meanwhile opening us to growing serious security attacks. The ways to recognize various network threats, specifically attacks not seen before, is a primary issue that needs to be looked into immediately. The aim of phishing site URLs is to collect the private information like user's identity, passwords and online money related exchanges. Phishers use the sites which are visibly and semantically like those of authentic websites. Since the majority of the clients go online to get to the administrations given by the government and money related organizations, there has been a vital increment in phishing threats and attacks since some years. As technology is growing, phishing methods have started to progress briskly and this should be avoided by making use of anti-phishing techniques to detect phishing. Machine learning is a authoritative tool that can be used to aim against phishing assaults. There are several methods or approaches to identify phishing websites. The machine learning approaches to detect phishing websites have been proposed earlier and have been implemented. The central aim of this project is to implement the system with high efficiency, accuracy and cost effectively. That is been achieved. The project is implemented using 4 machine learning supervised classification models. The four classification models are K-Nearest Neighbor, Kernel Support vector machine, decision tree and random forest classifier. It was established that the Random forest classifier provides best accuracy for the selected dataset and gives an accuracy score of 96.82%.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

As a new type of cyber security threat, phishing websites appear frequently in recent years, which have led to great harm in online financial services and data security . It has been projected that the vulnerability of most web servers have led to the evolution of most phishing websites such that the weakness in the web server is exploited by phishers to host counterfeiting website without the knowledge of the owner. It is also possible that the phisher hosts a new web server independent of any legitimate web server for phishing activities.the most researchers claimed that the method used in carrying out phishing can be different across regions. Furthermore, he also deduced that the phishers in America and China region have different approaches that he categorized into two on the basis of region:

Most researchers have worked on increasing the accuracy of website phishing detection through multiple techniques. Several classifiers such as Linear Regression, K-Nearest Neighbor, C5.0, Naïve Bayes, Support Vector Machine (SVM), and Artificial Neural Network among others have been used to train datasets in identifying phishing websites. These classifiers can be classified into two techniques: either probabilistic or machine learning. Based on these algorithms, several problems regarding phishing website detection have been solved by different researchers. Some of these algorithms were evaluated using four metrics, precision, recall, F1-Score, and accuracy.

1.1 Project Overview

Title of the study	Author	Briefly description of study	Experimental results	Study limitations
Intelligent flushing website detection system using fuzzy techniques	(Aburrouse et al., 2008)	Tie proposed model is based on FL operators which is used to characterize me website flushing factors and indicators as fuzzy variables and produces six measures and criteria's of website phishing attack dimensions with a layer structure.	The experimental results showed me significance and importance of the phishing website criteria (URL Domain Identity) represented by layer one and the variety influence of the phishing characteristic layers on the final phishing website rate.	The approach does not look for deviations from stored patterns of normal phishing behavior and for previously described patterns of behavior that is likely to indicate phishing.
An anti-Phishing ppproach that uses training intervention for flushing websites detection is likely to indicate phishing.	(Ainajim and Munro, 2009)	This paper proposes and evaluates a novel anti-phishing approach that uses training intervention for Phishing websites detection (APTIPWD) in comparison to an existing approach (sending anti-phishing tips by emails) and control group.	There is a significant positive effect of using the APTIFWD in comparison on with the existing approach and control group in helping users properly judging legitimate and phishing websites.	Nil
Identifying vulnerable websites by analysis of common strums in phishing URLs	(Wardmaner et al., 2009)	The propos ed method involves applying a – longest common substring algorithm to known phishing URLs, and investigating me results of that string to identify common vulnerabilities, exploits, and attack tools which may be prevalent among those who lack servers for phishing.	The result demonstrated mat these application paths may be used as a basis for further investigation to expose and document the primary exploits and tools used by hackers to compromise web servers, which could lead to the revelation of the aliases or identities of me criminals.	Nil
Associative classification techniques for predicting e-banking phishing websites	(Aburrous et al., 2010)	The research proposed an intelligent, resilient aid effective model that is based on using association and classification data mining algorithms. They used a number of different listing data mining association and classification techniques	The experimental results demonstrated me feasibility of using associative classification techniques in real applications and its better performance as compared to other traditional classifications algorithms.	

1.2 PURPOSE

There are number of users who purchase products online and make payment through various websites.There are multiple websites who ask user to provide sensitive data such as username, password or credit card details etc. often for malicious reasons. This type of websites is known as phishing website. In order to detect and predict phishing website, we proposed an intelligent, flexible and effective system that is based on using classification Data mining algorithm. We implemented classification algorithm and techniques to extract the phishing data sets criteria to classify their legitimacy. The phishing website can be detected based on some important characteristics like URL and Domain Identity, and security and encryption criteria in the final phishing detection rate. Once user makes transaction through online when he makes payment through the website our system will use data mining algorithm to detect whether the website is phishing website or not. This application can be used by many E-commerce enterprises in order to make the whole transaction process secure. 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. Admin can add phishing website url or fake website url into system where system could access and scan the phishing website and by using algorithm, it will add new suspicious keywords to database.System uses machine learning technique to add new keywords into database.

CHAPTER 2

LITERATURE SURVEY

In this Review, Many Papers have studied to know the details about web phishing detection. Machine learning, classification algorithm also the other technique other machine learning techniques can be involved here,explainidentificationtechniquesofeachpaper.

SURVEYPAPERS

- Rao et al proposed a novel classification approach that use heuristic based feature extraction approach. In this, they have classified extracted features into three categories such as URL Obfuscation features, Third-Party-based features,Hyperlink-based features. Moreover, proposed technique gives 99.55% accuracy.Drawback of this is that as this model uses third-party features, classification of website dependent on speed of third-party services .Also this model is purely depends on the quality and quantity of the training set and Broken links feature extraction has limitation of more execution time for the websites with more number of links
- Chunlinetal proposed approach that primarily focus on character frequency features In this they have combined statistical analysis of URL with machine learning technic to get results that is more accurate for classifcation of malicious URL. Also they have comparedsix machine learning algorithm to verify the effictiveness of proposed algorithm which gives 99.7% precision.
- Sudhans huetal .used association data mining approach.They have proposed rule based classification technique for phishing website detection .They have concluded that association classification algorithm is better than any other algorithms be cause of their simple rule transformation.They achieved 92.67% accuracy by extracting 16 features but this is not upto mark so proposed algorithm can be enhanced forefficient detectionrate…
- .M.Amaadetal.presented a hybrid model for classification of phishing website . In this paper, proposed model carried out into phase.In phase1, they individually perform classification techniques, and select the best three models based on high accuracy an do the performance criteria.While in phase2, they further combined

individual model with best three model and makes hybrid model that gives better accuracy than individual model.They achieved 97.75% accuracy on testing dataset.There is limitation of this model that it requires more time to build hybridmodel.

- Hosseinetal .developed an open-source framework knownas“Fresh-Phish”.Forphishing websites, machine-learning data can be created using this framework.In this, they have used reduced features set andusing python for building query .They build a large labelled dataset and analyse several machine-learning classifiers against this dataset.Analysis of this gives very good accuracy using machine-learning classifiers. These analyses how long time it takes to train the model.
- Guptaetal .proposed a novel anti phishing approach that extracts features from client-side only. Proposedapproachisfastandreliableasitisnotdependent on third party but it extracts features only from URL and source code. In this paper, they have achieved99.09%of overall detection accuracy for phishing website.This paper have concluded that his a sit can web page written in HTML.Non- HTML web page cannot detect by this approach

2.1 EXISTING PROBLEM

Phishing is a typical type of social designing assault intended to gather client data, for example, login certifications and Visa data. At the point when a casualty opens an email, text, or instant message subsequent to being hoodwinked into doing as such by a culprit acting like a dependable source, it happens. The beneficiary is in this manner fooled into clicking a hazardous connection, which might introduce malware, lock the framework as a feature of a ransomware assault, or uncover private data.

Phishing is additionally consistently used to get sufficiently close to corporate or administrative organizations as a component of bigger assaults like high level determined danger (APT) occurrences. In the last situation, workforce is compromised to evade safety efforts, engender malware inside a protected setting, or get to private data.

As well as experiencing huge monetary misfortunes, an organization that is the casualty of such an assault habitually has its piece of the pie, notoriety, and client certainty decline. A security emergency from which an association will experience difficulty recuperating could result from a phishing endeavor, contingent upon its expansiveness.

2.2 REFERENCES

S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs,"Who falls for phish?: a demographic analysis of phishing suscepti-bility and effectiveness interventions," in Proceedings of the 28th international conference on Human factorsin computing systems, ser.CHI '10. New York, NY, USA: ACM, 2010, pp. 373–382.

B. Krebs, HBGary Federal hacked by Anonymous, accessed December 2011…

B. Schneier, "Lockheed Martin hack linked to RSA's SecurID breach,"http://www. schneier.com/blog/archives/2011/05/lockheed martin.html,2011, accessed December 2011

C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classifi-cation of phishing pages," in NDSS '10, 2010.

X. Dong, J. Clark, and J. Jacob, "Modelling user-phishing interaction,"in Human System Interactions, 2008 Conference on, may 2008, pp. 627–632.

W. D. Yu, S. Nargundkar, and N. Tiruthani, "A phishing vulnerability analysis of web based systems," in Proceedings of the 13th IEEE Sym-posium on Computers and Communications (ISCC 2008). Marrakech,Morocco: IEEE, July 2008, pp. 326–331.

Anti-Phishing Working Group (APWG), "Phishing activity trends report second half 2010," http://apwg.org/reports/apwg report h2 2010.pdf, 2010, accessed December 2011.

Anti-Phishing Working Group (APWG), "Phishing activity trends re-port — first half 2011," http://apwg.org/reports/apwg trends report h1 2011.pdf, 2011, accessed December 2011.

***2.3* PROBLEM SOLUTION DEFINITION**

Phishing is a type of social engineering attack often used to steal user data. Phishing attacks are becoming more and more sophisticated, and our algorithms are suffering to keep up with this level of sophistication. They have low detection rate and high false alarm especially when novel phishing approaches are used. The blacklist-based method is unable to keep up with the current phishing attacks as registering new domains has become easier. Moreover, a blacklist can ensure a

perfect up-to-date database.Variousothertechniquessuchaspagecontentinspection algorithms have been used to combat the false negatives but as each algorithm uses a different approach, their accuracy varies. Therefore, a combination of the two can increase theaccuracywhileimplementingdifferenterrordetectionmethods.

Since scams are widespread and nobody wants to fall for web phishing, regular users whosearchtheinternetforinformationneedawaytomakesurethelinks they click are safe.

Example:

ProblemStatement (PS)	I am (Customer)	I'm trying to	But	Because	Which makes me feel
PS-1	Internet user	Browse the internet	Identify a scam	An attacker makes an reputable entity	Unsafe about my information that is shared over the network
PS-2	Enterprise user	Open emails in the cloud server	I detect malicios protocol	They are not cryptographically signed	Emails are un verified and third-partyintrusion

CHAPTER 3

IDEATION & PROPOSED SOLUTION

In this activity we are prepared the empathy map canvas to capture the user Pains & Gains,brainstorming, problem solution fit and proposed solution based on the feasibility & importance.

2.1 EMPATHYMAP

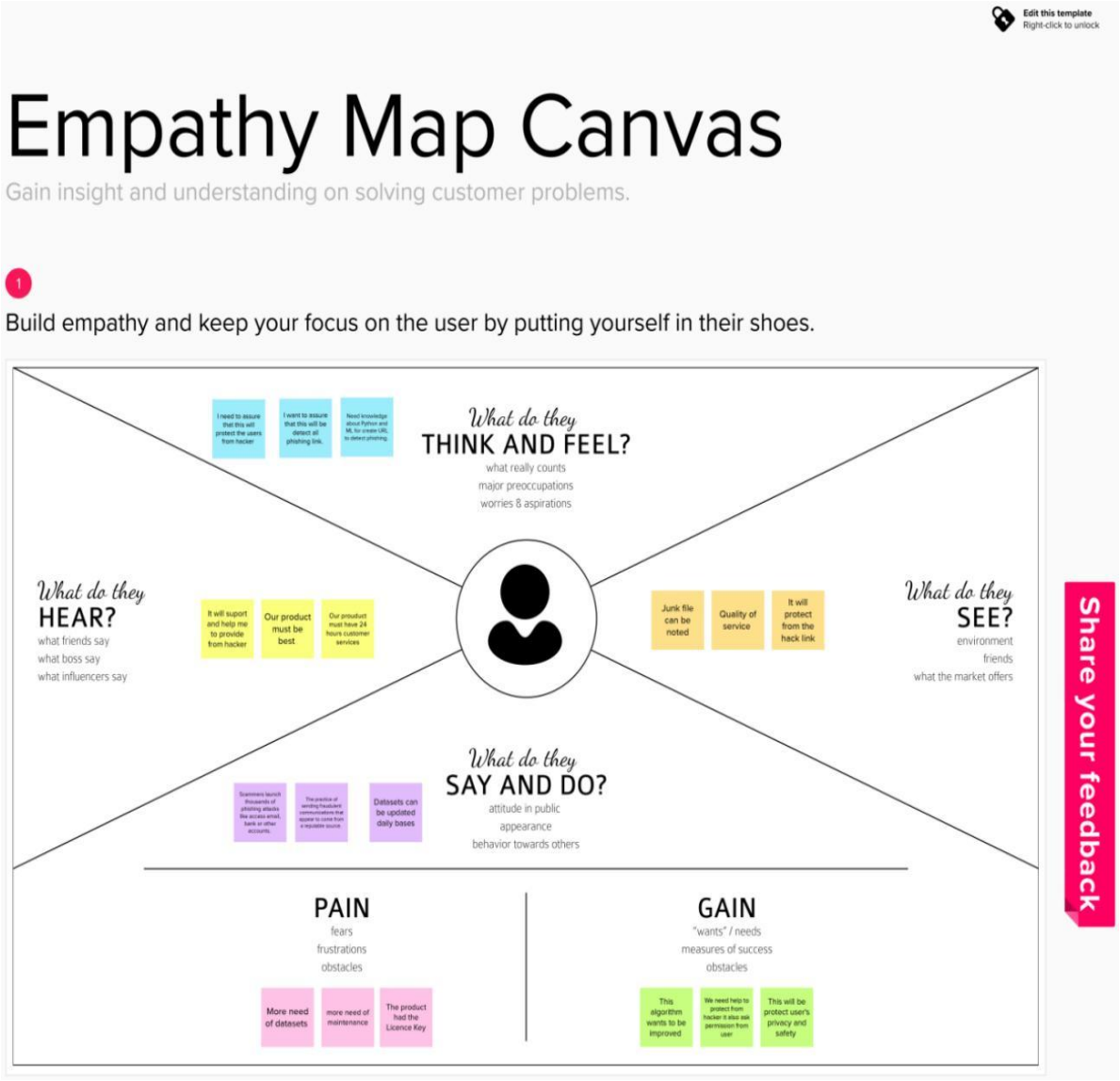


Fig 3.1.1 Empathy Map Canvas

2.2 IDEATION & BRAINSTROMING

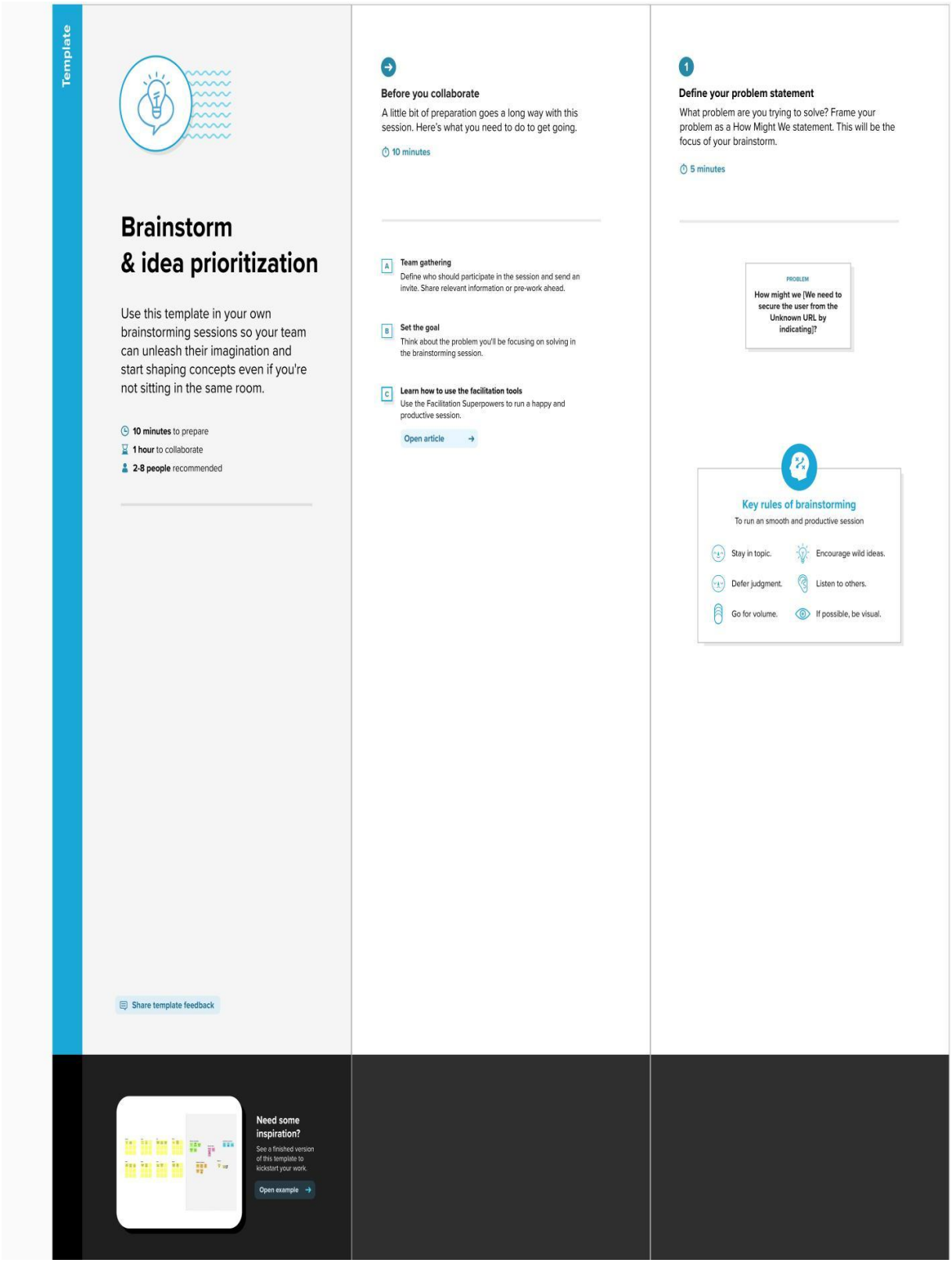


Fig 3.2.1 shows Brainstroming

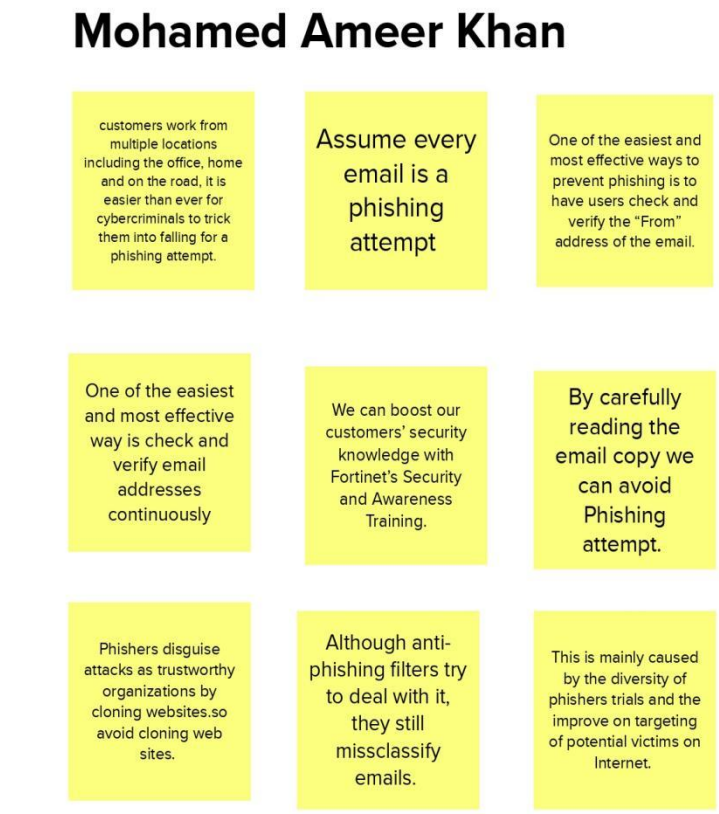


Fig 3.2.2 shows the ideas

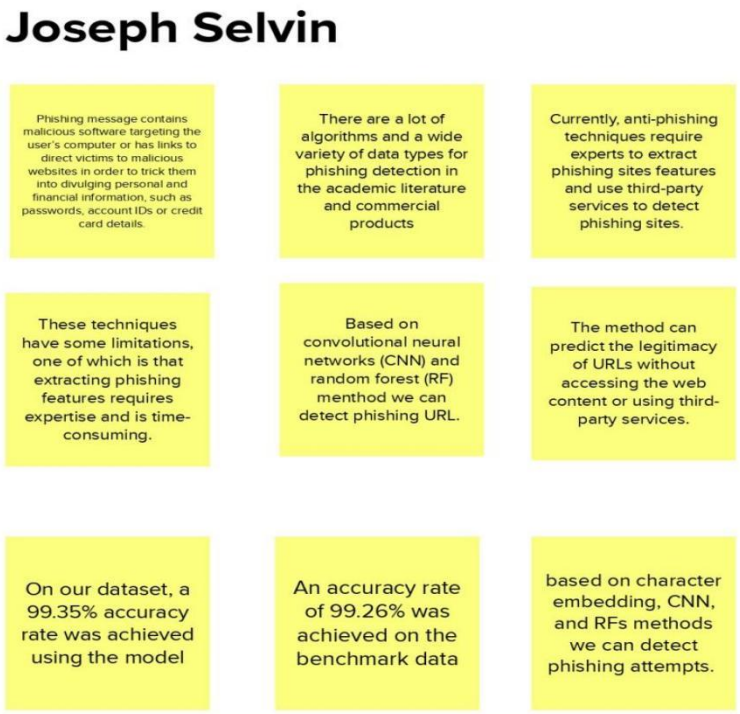


Fig 3.2.3 shows the ideas

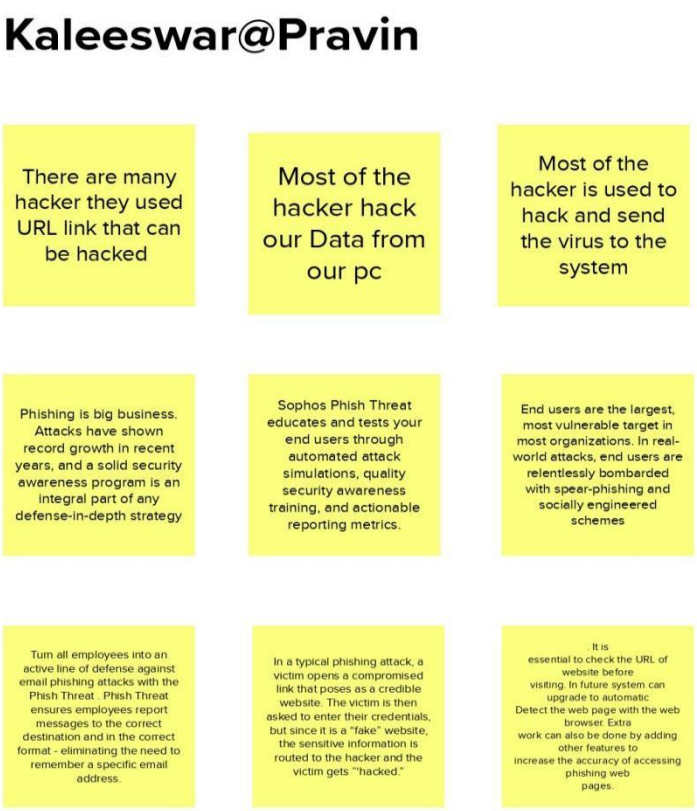


Fig 3.2.4 shows the ideas

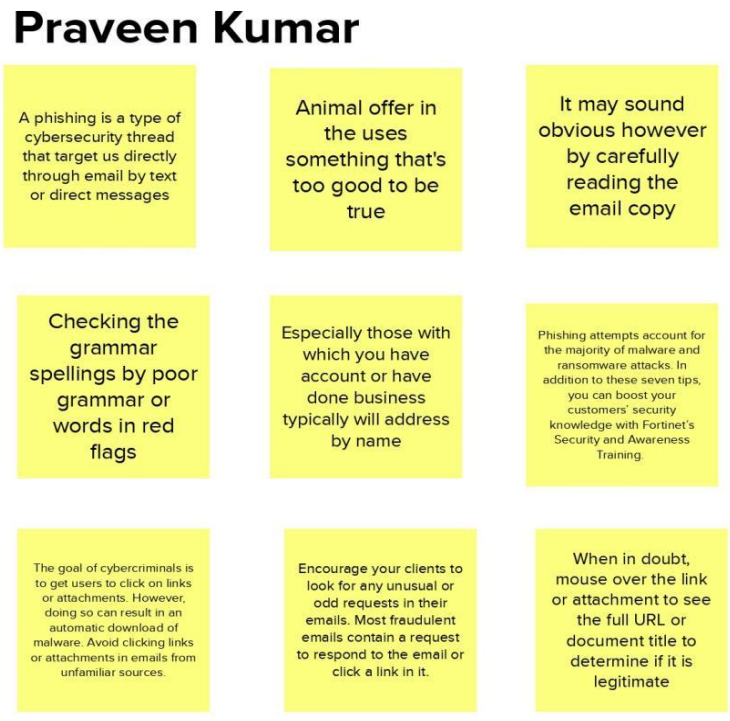


Fig 3.2.5 shows the ideas

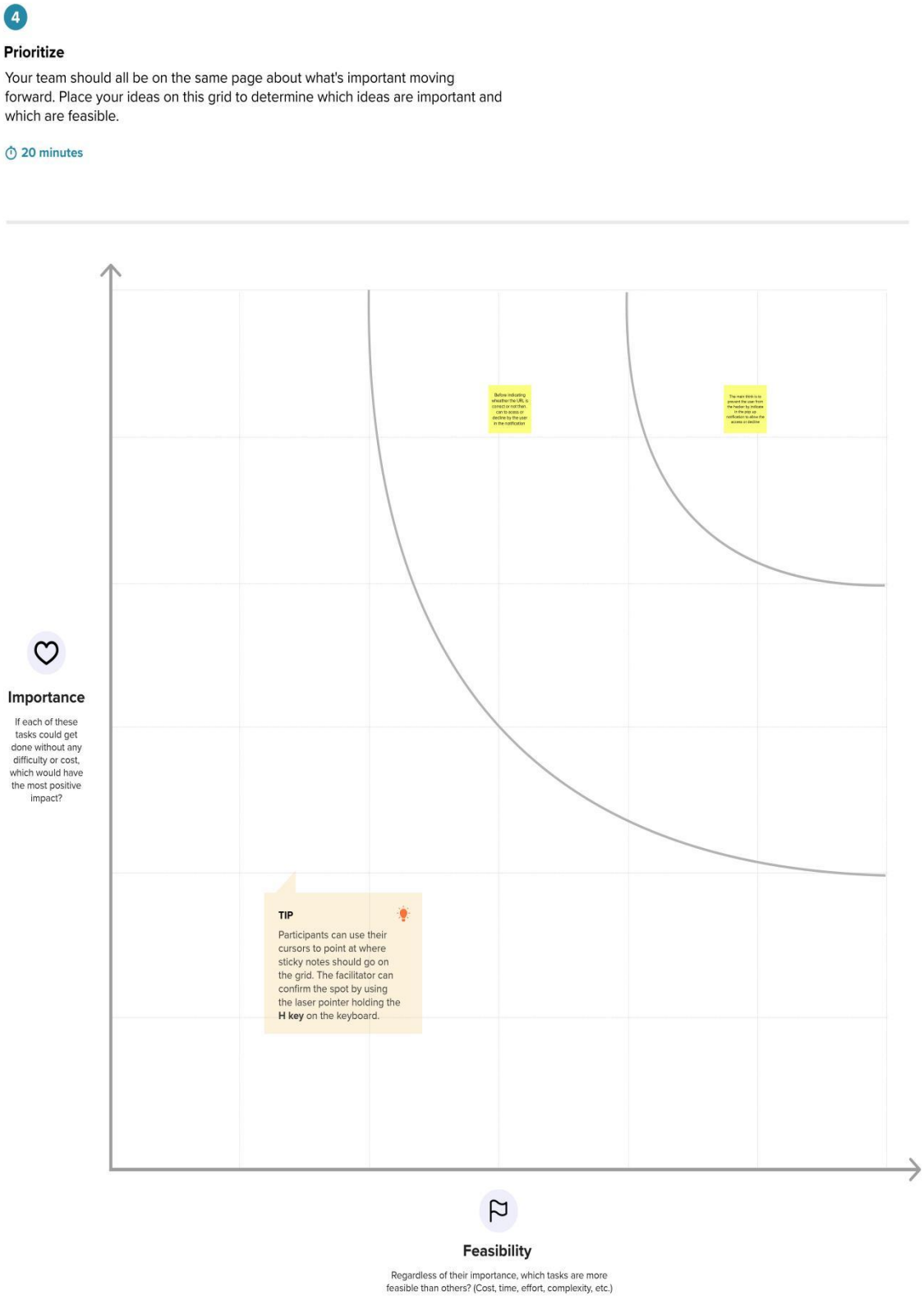


Fig 3.2.6 shows the priority of information

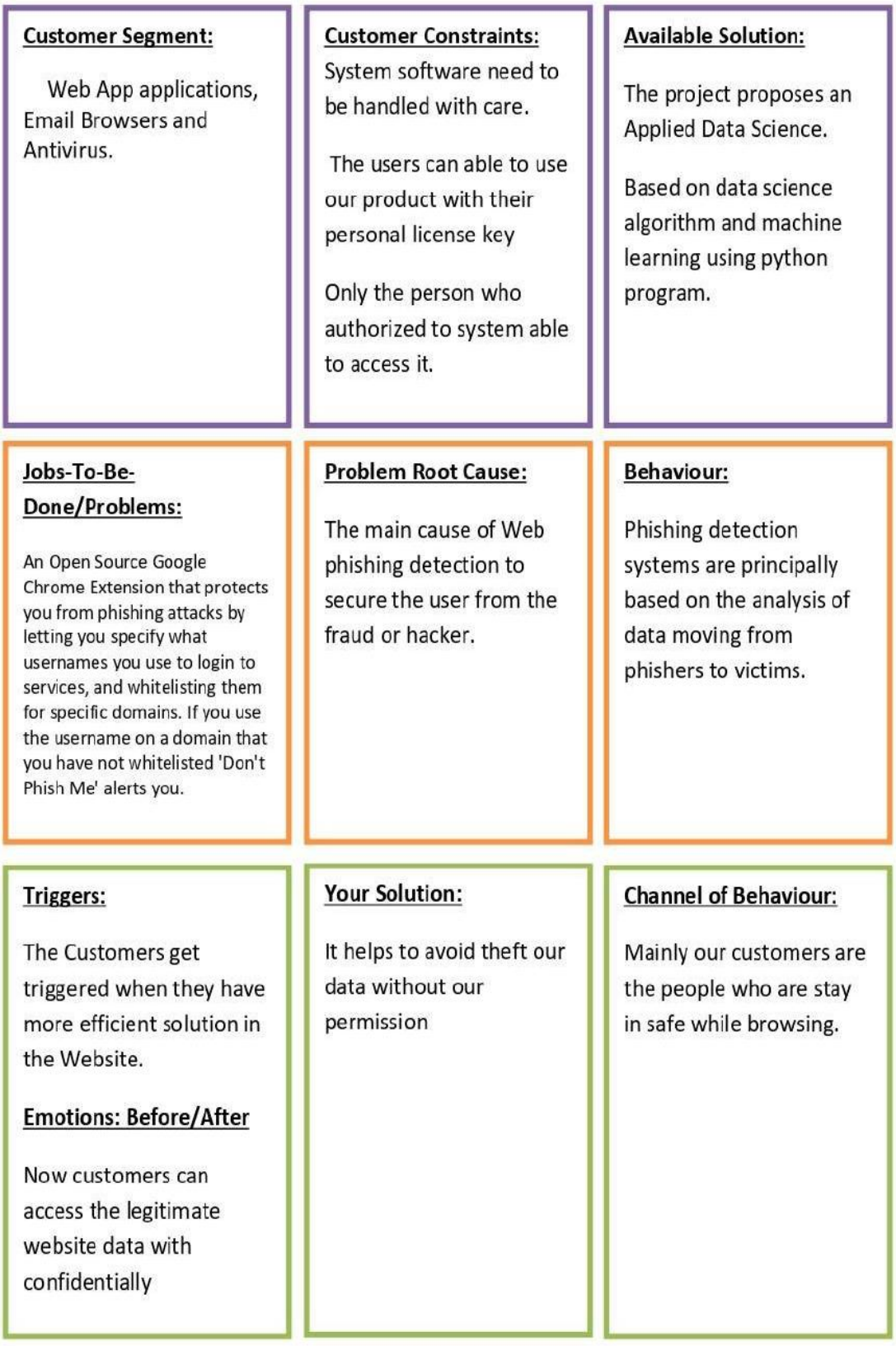
3.2 PROPOSED SOLUTION

Projectteamshallfillthefollowinginformationinproposedsolutiontemplate.

Below(Table: Proposed solution 3.3.1)

S.No.	Parameter	Description
1.	Problem Statement (Problemtobesolved)	Hacker can access our data without our permission weneed tosolveit
2.	Idea/Solution description	When the user to access the data by himself only
3.	Novelty/Uniqueness	The URL send by some one it must be checked by my product and it will not harm then it can be access by my product
4.	Social Impact/ Customer Satisfaction	It will give the awareness from the hacker they can theft our data without our permission
5.	Business Model(Revenue Model)	It must be licensed product by the machine learn coding
6.	Scalability of the Solution	The product must berenewal within 6months to the updated version to Avoid the hacker.

2.3 PROBLEM SOLUTION FIT



CHAPTER 4

REQUIREMENT ANALYSIS

One most popular directory-based approach is Phish Tank . What Phish Tank offers is a very collaborative data house regarding all the phishing websites on the World Wide Web with information's regarding the website to indicate how severe it is to have users aware. Another feature Phish Tank offers, is their open API availability for allowing researchers, developers to integrate it into their phishing tools without any cost behind it. Based on that availability, Phish Tank is considered as a directory that contains all phishing websites reported by community members around the world which aids developers when they use their API for phishing detection purposes. Another API that exists which helps in developing tools for phishing detection is from Google called Safe Google Browsing API , but it also follows the same directory based approach as Phish Tank. The downside of this approach is that there is always a constant influx of new phishing websites in the web and this cannot always be updated in a real time moment and it will also require huge rate of contribution from community members to always update their directories with new updated phishing websites that exist. Detection of phishing methods is based on server base side and client base side. One of the existing google chrome extension plugin that follows a rule-based concept is called Phish Detector, which allows eligibility to detect phishing websites without the usage of an external web service. While implementing such plugin is much easier from a client side, it cannot be compared to how accurately it will be whilst being compared to machine learning approach techniques.

Another tool that works with rule-based concepts is PhishNet, where it uses a predictive method for blacklisting. The rules that are being adapted are being matched with a term called Top Level Domain (TLD) directory structure, IP addresses, and headers of HTTP responses. Stanford developed tool called SpoofGuard works just like PhishNet mentioned above, but it considers rule-based approach using DNS, URLs, clickable links and images presented on the web-page. Author from the research paper "Feature extraction and classifying websites that are malicious based on their URL" used a technique where he extracts the features to make a feature matrix that was substantially used to classify URLs. In their development, they extracted roughly 133 features and they only use sub-part of it which they concluded

as feasible for their project. It was also not understood why they didn't specify their reasoning for choosing specific parameters to declare websites as malicious or not otherwise. Parameters that were set and used together with related respective algorithms in their projects were different to what we do plan to use for our development and project implementation. For our master's project, we have decided to only use one single algorithm that suites best and identifying what features that were given from our dataset would help us respectively. A research article "Comparing machine learning techniques for detection" a comparative study was done between six various classifiers to understand which classifier will fit and work best in distinguishing a phishing URL and a legitimate URL. The authors concluded that Random Forest classifiers fit and worked best due to having lowest error rate among the five remaining classifier that was used for the comparative study. In an article regarding how using machine learning can aid in detecting phished URL , the author raises questions and awareness how phishing in general on the internet has rose significantly over the recent years and also discusses technique implementation on feature extraction and understanding what ideal machine learning algorithm will best fit the classification. While we don't particularly follow their exact measures of extracting features like details on traffic, page rank detail features, this paper provides an ideal footstep for understanding what features can be extracted based on our requirements for our project. While in the paper, the author hasn't represented any indication of the best algorithm that fits these projects, we in our master thesis project will give a statistical analysis on why Random Forest classifiers is the best fit together with its accuracy for the chosen algorithm. Netcraft phishing protection basically works as a big neighborhood watch scheme. Once someone reports a potential phishing website into the community, it will then be investigated and if it's proven as a phishing website, the targeted URL will be blocked for their community members. Phish Detector has a 100% success rate on detecting phishing attacks on online banking websites. To obtain positive accurate results, only use this tool on banking websites as this extension does not work on other website domains.

Another existing phishing detector that exists in today's market is called cascaded phishing detector. It basically functions as a client side and as well as server-side tool where the client side is developed as a chrome extension. This is then followed up by injecting certain scripts to the respected websites to extract the relevant and related corresponding HTML DOMs. This extension compared to the existing extensions that

exist in the market only gives priority to the HTML DOMs to identify the prospect of being phished while disregarding the other parameters.

4.1Functional System Requirement

Extension plugin should provide a warning pop-up when they visit a website that is phished; therefore it should strictly follow the following:

a. Extension plugin ability to present the pop-up to the users screen should be quick enough to the point, users will be aware before entering any confidential or sensitive details into a phishing website.

b. Extension plugin should not need the facilities and services from an 3rd party service or APIs, due the reason that those services will always the potential to leak users browsing data and pattern when it gets compromised by hackers

c. Extension plugin will have the capability to also detect latest and new phishing websites

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 Gmal
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	User Profile	Confirm by their license Confirm by their password
FR-4	User Login	Via Gmail Via Mobile Number Via License Key

(Table Functional requirements 4.1.1)

Non-Functional System Requirement

Graphical User Interface design Interface developed should be done with the understanding that it must meet the simplicity of what users would like to see when they need an extension for detecting things, and also it needs to adhere to non IT literate users as well. It must also provide the exact information on what the user wants like identifying a phishing website quickly without needing to click on many options. The process of identifying phishing website should be taken directly from the web-page user wants to view through their URL and the result from it should be

easily understood by the users. Most importantly, the extension plugin should have a popup that will notify the user regarding the website status of being phished.

Nonfunctional requirements describe how a system must behave and establish constraints of its functionality. . Any attributes required by the customer are described by the specification. We must include only those requirements that are appropriate for our project. Some Non-Functional Requirements are as follows:

- Reliability
- Maintainability
- Performance
- Portability
- Scalability
- Flexibility

Software requirements:

a. PyCharm software
b. Python language
c. Google chrome browser
d. Scikit-learn
e. NumPy
f. Liac-arff for dataset

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

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	One result of our research is also a classification of anti-phishing toolbar applications.
NFR-2	**Security**	Our product provides all type of security to the customer datas.
NFR-3	**Reliability**	comparing their URLs against a blacklist of known fake URLs.
NFR-4	**Performance**	Application can be provides the phishing less environment against phishing attacks
NFR-5	**Availability**	Available in all browsers.
NFR-6	**Scalability**	Detect and isolate both phishing e-mail senders and phishing web servers.

(Table Functional requirements 4.2.1)

CHAPTER 5

PROJECT DESIGN

5.1 Data Flow Diagram

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.

Architecture:

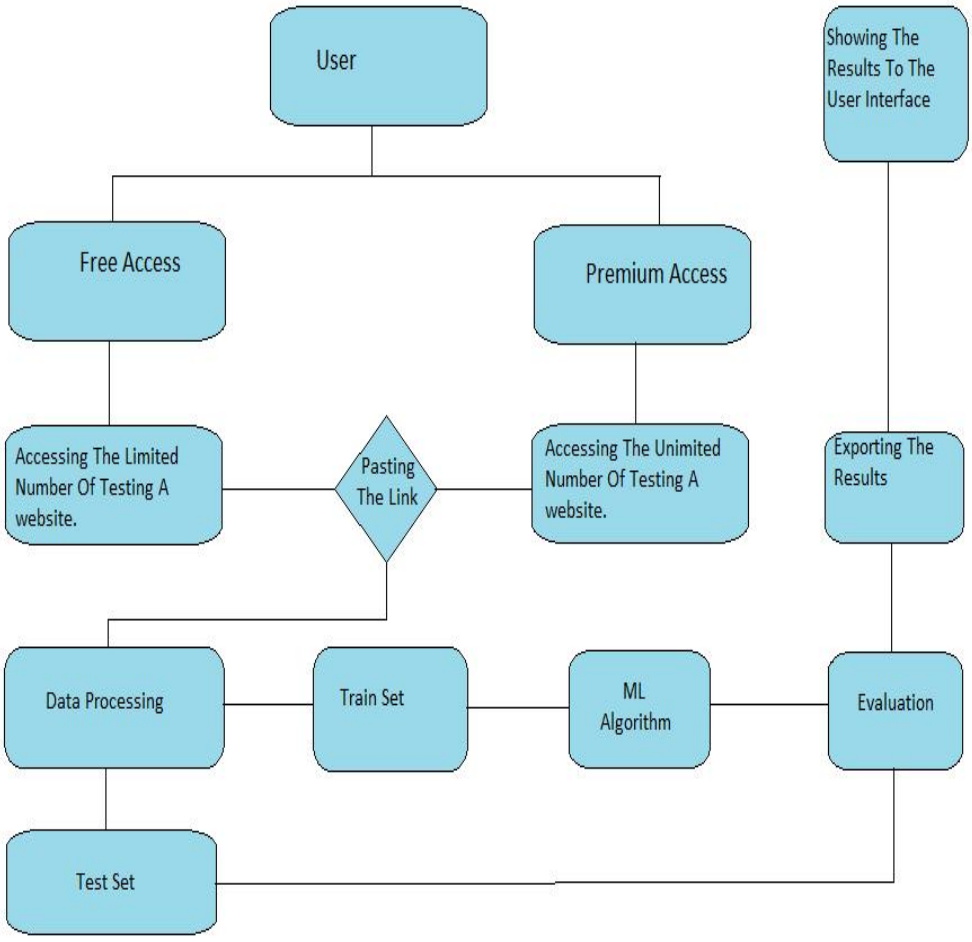


Fig 5.1.1 Data Flow Diagrams

5.2 SOLUTION & TECHNICAL ARCHITECTURE

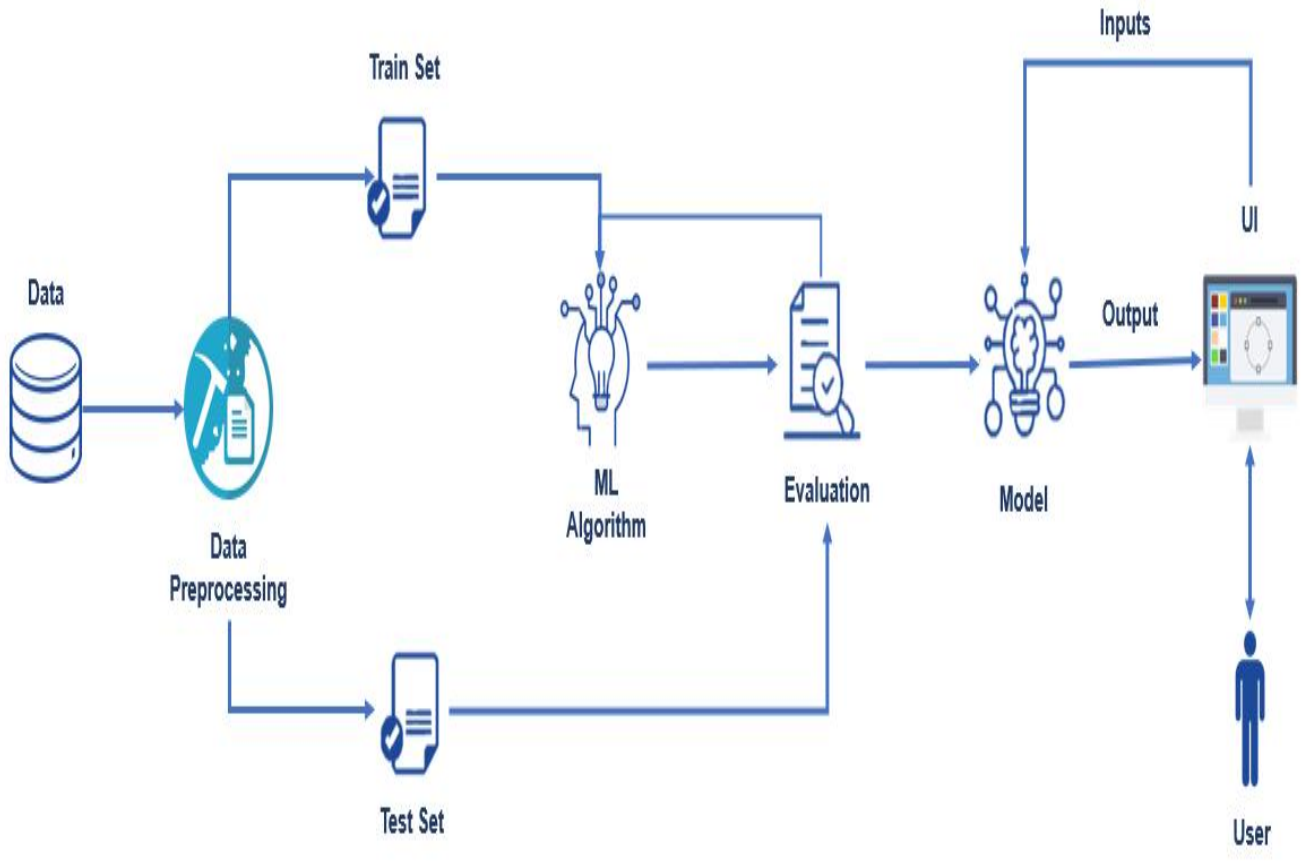


Fig Technical Architecture 5.2.1

5.3 User Stories

A user story is a lightweight method for quickly capturing the "who", "what" and "why" of a product requirement. In simple terms, user stories are stated ideas of requirements that express what users need. User stories are brief, with each element often containing fewer than 10 or 15 words each. User stories are "to-do" lists that help you determine the steps along the project's path. They help ensure that your process, as well as the resulting product, will meet your requirements.Use the below template to list all the user stories for the product.

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and 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		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)	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	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
Customer Care Executive	Dashboard					
Administrator						

CHAPTER 6

PROJECT PLANNING & SCHEDULING

S.NO	ACTIVITY TITLE	ACTIVITY DESCRIPTION	DURATION
1	Projectpreparation	Assign teammembers,Creater epository in the Github,download rocket-chat essential sand join respective projectchannel.	1WEEK
2	Attend class	Attend sessions on IBM ,team leader assign task to each member of theproject,attend quiz ,submit assignment.	1WEEK
3	Working on different phases of project	Ideation phase-literature survey, Project design phase I-proposed solution , solution architecture ,project design phase II-customer journey ,data flow,technical architecture,planning phase-milestones,tasks,sprint schedule.	4WEEK
4	Developing project	Develop the code,test and push itto GitHub,clarifyqueries.	2WEEK
5	Budget and scope of project	Analyze and making the project budget and discuss with team for budget prediction.	1WEEK

Table Sprint planning and estimation 6.1.1

6.2 SPRINT DELIVERY SCHEDULE

Product Backlog, Sprint Schedule , and Estimation

Use the below template to create product backlog and sprint schedule

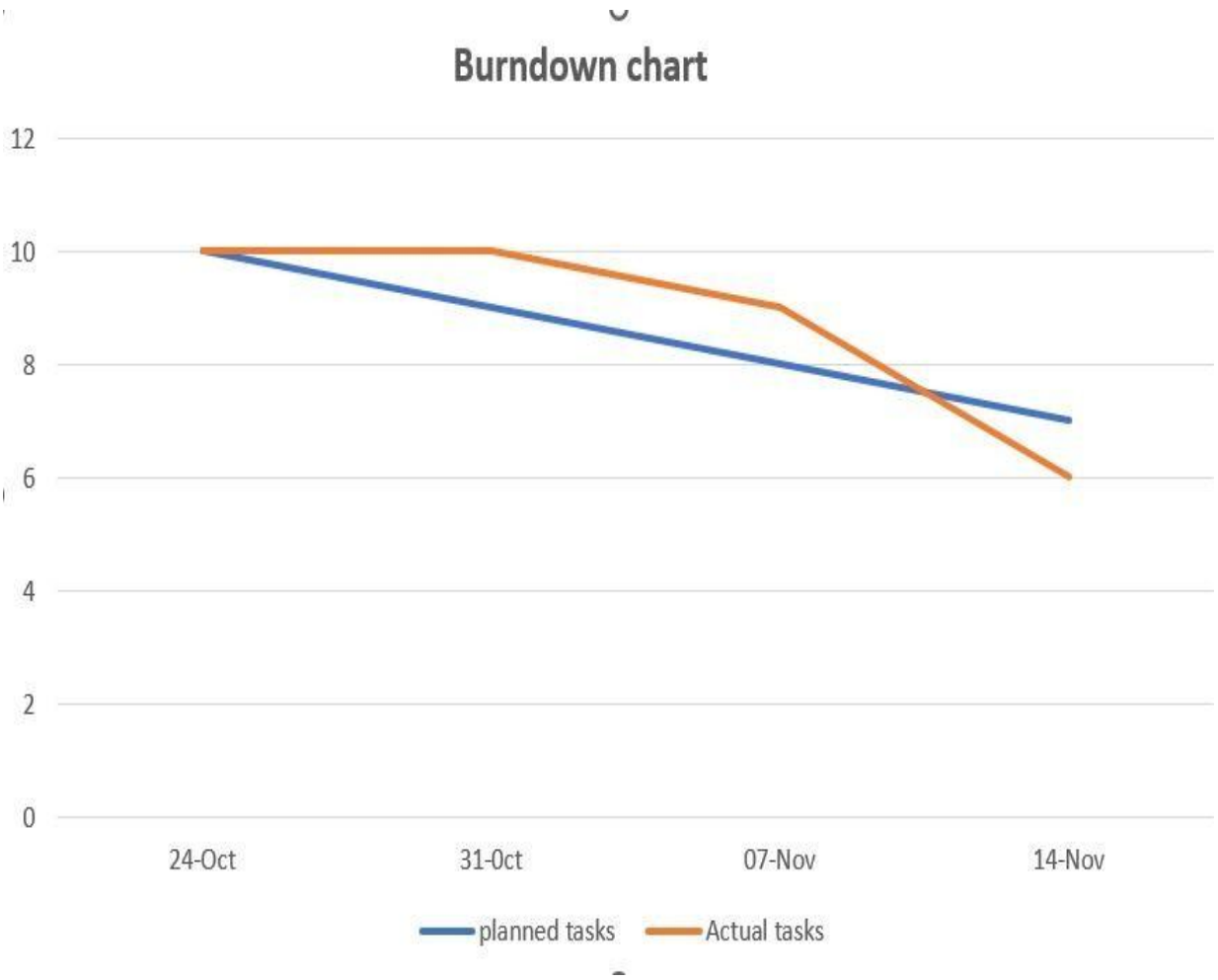
Table Sprint Delivery schedule 6.2.1

Sprint	FunctionalRequirement(Epic)	User StoryNumber	UserStory/Task	StoryPoints	Priority	TeamMembers
Sprint-1	Registration	USN-1	As a user, I can register for the application byentering my email, password, and confirmingmypassword.	2	High	A.Praveenkumar
Sprint-1		USN-2	As a user, I will receive confirmation email oncel have registered forthe application	1	High	B.Kaleeswar@pravin
Sprint-2		USN-3	Asauser,Icanregisterforthe applicationthrough Facebook	2	Low	K.Mohamed Ameerkhan
Sprint-1		USN-4	Asauser,Icanregisterforthe applicationthrough Gmail	2	Medium	S.JosephSelvin
Sprint-1	Login	USN-5	As a user, I can log into the application byEntering email &password	1	High	K.MohamedAmeerkhan

6.3 Burndown Chart

Project Tracker ,Velocity & Burn down Chart:

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date(Planned)	Story PointsCompleted(ason PlannedEndDate)	Sprint Release Date (Actual)
Sprint-1	20	6Days	24Oct2022	29Oct2022	20	29Oct2022
Sprint-2	20	6Days	31Oct2022	05Nov2022	20	06Nov2022
Sprint-3	20	6Days	07Nov2022	12Nov2022	20	14Nov2022
Sprint-4	20	6Days	14Nov2022	19Nov2022	20	19Nov2022



Velocity:

AV= sprint delivery

Velocity

= 20/6
= 3.33

CHAPTER 7

CODING & SOLUTIONING
(Explain the features added in the project along with code)

7.1 FEATURE 1 Software Requirement Specification (SRS)

PANDAS

Pandas is an open-source, BSD-authorized Python library giving elite, simple to-utilize information structures and information examination instruments for the Python programming language. Python with Pandas is utilized in a wide scope of fields including scholastic and business areas including money, financial matters, Statistics, examination, and so on. In this instructional exercise, we will get familiar with the different highlights of Python Pandas and how to utilize them practically speaking. This instructional exercise has been set up for the individuals who try to become familiar with the essentials and different elements of Pandas. It will be explicitly valuable for individuals working with information purging and examination. In the wake of finishing this instructional exercise, you will wind up at a moderate dimension of ability from where you can take yourself to more elevated amounts of skill. You ought to have a fundamental comprehension of Computer Programming phrasings. A fundamental comprehension of any of the programming dialects is an or more. Pandas library utilizes the vast majority of the functionalities of NumPy. It is recommended that you experience our instructional exercise on NumPy before continuing with this instructional exercise. 2.4.5 ANACONDA Anaconda constrictor is bundle director. Jupyter is an introduction layer.Boa constrictor endeavors to explain the reliance damnation in python—where distinctive tasks have diverse reliance variants—in order to not influence distinctive venture conditions to require diverse adaptations, which may meddle with one another. Jupyter endeavors to fathom the issue of reproducibility in investigation by empowering an iterative and hands-on way to deal with clarifying and imagining code; by utilizing rich content documentations joined with visual portrayals, in a solitary arrangement. Boa constrictor is like pyenv, venv and minconda; it's intended to accomplish a python situation that is 100% reproducible on another condition, autonomous of whatever different forms of a task's conditions are accessible. It's somewhat like Docker, however limited to the Python biological system. Jupyter is an astounding introduction device for expository work; where you can display code in

"squares," joins with rich content depictions among squares, and the consideration of organized yield from the squares, and charts created in an all around planned issue by method for another square's code. Jupyter is extraordinarily great in expository work to guarantee reproducibility in somebody's exploration, so anybody can return numerous months after the fact and outwardly comprehend what somebody attempted to clarify, and see precisely which code drove which representation and end. Regularly in diagnostic work you will finish up with huge amounts of half-completed note pads clarifying Proof-of-Concept thoughts, of which most won't lead anyplace at first. A portion of these introductions may months after the fact—or even years after the fact— present an establishment to work from for another issue.

PYTHON

Python is a translated, object-arranged, abnormal state programming language with dynamic semantics. Its abnormal state worked in information structures, joined with dynamic composing and dynamic authoritative, make it appealing for Rapid Application Development, just as for use as a scripting or paste language to interface existing segments together. Python's basic, simple to learn language structure underlines intelligibility and hence decreases the expense of program support. Python underpins modules and bundles, which empowers program seclusion and code reuse. The Python translator and the broad standard library are accessible in source or parallel structure without charge for every single significant stage, and can be openly appropriated. Frequently, software engineers begin to look all starry eyed at Python on account of the expanded efficiency it gives. Since there is no aggregation step, the alter test-troubleshoot cycle is staggeringly quick.

Troubleshooting Python programs is simple: a bug or awful information will never cause a division blame. Rather, when the mediator finds a blunder, it raises a special case. At the point when the program doesn't get the special case, the translator prints a stack follow. A source level debugger permits assessment of nearby and worldwide factors, assessment of discretionary articulations, setting breakpoints, venturing through the code a line at any given moment, etc. The debugger is written in Python itself, vouching for Python's contemplative power. Then again, frequently the speediest method to troubleshoot a program is to add a couple of print proclamations

to the source: the quick alter test investigate cycle makes this straightforward methodology successful. Python is an item situated, abnormal state programming language with incorporated unique semantics essentially for web and application improvement. It is amazingly alluring in the field of Rapid Application Development since it offers dynamic composing and dynamic restricting alternatives. Python is generally basic, so it's anything but difficult to learn since it requires a one of a kind language structure that centers around coherence. Designers can peruse and interpret Python code a lot simpler than different dialects. this decreases the expense of program upkeep and improvement since it enables groups to work cooperatively without huge language and experience obstructions. Moreover, Python underpins the utilization of modules and bundles, which implies that projects can be planned in a secluded style and code can be reused over an assortment of tasks. When you've built up a module or bundle you need, it very well may be scaled for use in different tasks, and it's anything but difficult to import or fare these modules. A standout among the most encouraging advantages of Python is that both the standard library and the mediator are accessible for nothing out of pocket, in both parallel and source structure. There is no restrictiveness either, as Python and all the important instruments are accessible on every single real stage. In this way, it is a tempting alternative for designers who would prefer not to stress over paying high improvement costs.

7.2 FEATURE

HTML Home Page

```
<!doctype html>
<html lang="en">
<head>
<link rel="stylesheet" type="text/css"
href="{{url_for('static',filename='css/style.css')}}">
<meta charset="utf-8">
<meta name="viewport" content="width=device-width, initial-scale=1">
<title>Home</title>
<link
href="https://cdn.jsdelivr.net/npm/bootstrap@5.2.2/dist/css/bootstrap.min.css"
rel="stylesheet" integrity="sha384-
Zenh87qX5JnK2Jl0vWa8Ck2rdkQ2Bzep5IDxbcnCeuOxjzrPF/et3URy9Bv1WT
Ri" crossorigin="anonymous">
</head>
<body class="bg-co">
<div class="bg-nav text-light d-flex flex-column flex-md-row align-items-center
pb-3 mb-4 border-bottom">
```

```
<h5 class="my-0 mr-md-auto font-weight-bold mt-3" style="font-size:20px;opacity: 0.5; font-family: Georgia, serif; font-weight: bold; padding-left: 50px;">URL Prediction</h5>
<nav class="d-inline-flex mt-2 mt-md-0 ms-md-auto ">
<a class="me-3 py-2 text-light text-decoration-none mt-3" style="font-family: Georgia, serif;font-weight: bold;margin-right:20px; " href="/predicturl">Predict URL</a>
<a class="me-3 py-2 text-light text-decoration-none mt-3" style="font-family: Georgia, serif;font-weight: bold;margin-right:20px; " href="/addurl">Add url</a>
<a class="me-3 py-2 text-light text-decoration-none mt-3" style="font-family: Georgia, serif;font-weight: bold; margin-right:20px;" href="/project_details">Project Details</a>
<a class="py-2 text-light text-decoration-none mt-3" style="font-family: Georgia serif;font-weight: bold; margin-right:20px;" href="/about">About</a>
</nav>
</div>
<div class="container bg-co">
<div class="row">
<div class="col-md-6 ">
<div  style="margin-top: 120px;margin-left: 190px;">
<form class="form"  action="/predicturl">
<center><img  src="{{url_for('static',  filename='Code.jpg')}}"  style="height: 40%; width: 40%; opacity: 0.5; " class="img img-responsive img-circle  mx-auto d-block" /><br>
<input type="submit" style="background-color: rgba(0, 0, 0, 0.801);color: white font-weight: bolder;"class="btn  btn-lg btn-block mx-auto "  value="PREDICT YOUR URL">
</center>
</form>
</div>
</div>
</div>
</div>
</div>
<script src="https://cdn.jsdelivr.net/npm/@popperjs/core@2.11.6/dist/umd/popper.min.js" integrity="sha384-oBqDVmMz9ATKxIep9tiCxS/Z9fNfEXiDAYTujMAeBAsjFuCZSmKbSSUnQlmh/jp3" crossorigin="anonymous"></script>
<script src="https://cdn.jsdelivr.net/npm/bootstrap@5.2.2/dist/js/bootstrap.min.js" integrity="sha384-IDwe1+LCz02ROU9k972gdyvl+AESN10+x7tBKgc9I5HFtuNz0wWnPclzo6p9vxnk" crossorigin="anonymous"></script>
<script src="https://cdn.jsdelivr.net/npm/bootstrap@5.2.2/dist/js/bootstrap.bundle.min.js" integrity="sha384-OERcA2EqjJCMA+/3y+gxIOqMEjwtxJY7qPCqsdltbNJuaOe923+mo//f6V8Qbsw3" crossorigin="anonymous"></script></body></html>
```

Numpy

- NumPy is an open-source library in Python that provides support in mathematical, scientific, engineering, and data science programming.
- To perform large mathematical operations and statistical operations Numpy is an incredible library.
- Numpy is basically a simple programming language that works superbly well for the multi-dimensional arrays and matrices multiplication.
- In 2005, Numpy was created by Travis Oliphant and as it is open-source so anyone can access it freely.
- Numpy is a great tool for any scientific project and it also contains a powerful n-dimensional array object.
- NumPy Library is written partially in Python and the parts of NumPy that require fast computation are written in C or C++.

Where is Numpy used?

- Below we have some usecases where NumPy is effective to use:
- Numpy is very useful in performing operations that are related to linear algebra and for its handling of random numbers.
- NumPy can efficiently implement multi-dimensional array objects (that are in the form of rows and columns).
- Numpy works efficiently with reshaping of matrices, random numbers, and Fourier transforms, etc.
- Numpy was designed for scientific computation.
- One thing is important to note here that TensorFlow and Scikit learn also uses NumPy array to compute the matrix multiplication in their back end

Why to use Numpy in Python?

- Because, in Python, Lists are used in order to serve the purpose of the array but lists are very slow to process. Hence we use Numpy in Python because it provides an array object that is up to 50x faster than traditional Python lists. And Python has other modules too, which makes data analysis and presentation very easy. So Numpy library is used with Python along with other Python libraries like Matplotlib, Scikit Learn, etc for AI/ML and Data analysis purposes.
- In NumPy, the array object is commonly known as ndarray. Numpy provides a lot of supporting functions for performing operations on its array object and with these functions, working with ndarray becomes very easy.

- Also, the NumPy arrays are more compact than Python Lists in terms of the size.

- NumPy uses much less memory in order to store data and it provides an easy mechanism of specifying the data types. Thus code can be optimized easily.

- Now you must be thinking, that how NumPy works faster than lists. Don't worry, we have an answer for your question.

- NumPy arrays are mainly stored at one continuous place in memory contrary to lists. Thus you can access and manipulate them very efficiently and this behavior is commonly known as locality of reference. Due to this reason, Numpy is faster than lists. Numpy is optimized to work with latest CPU architecture.

- Like we mentioned above, NumPy is also used along with packages like SciPy (Scientific Python) and Matplotlib (plotting library in python).

This combination is mainly a replacement for MatLab(which is a popular platform for technical computing). Also, Python is an alternative to MatLab and is now seen as a modern and complete programming language.

CHAPTER 8

TESTING

8.1 Test Cases

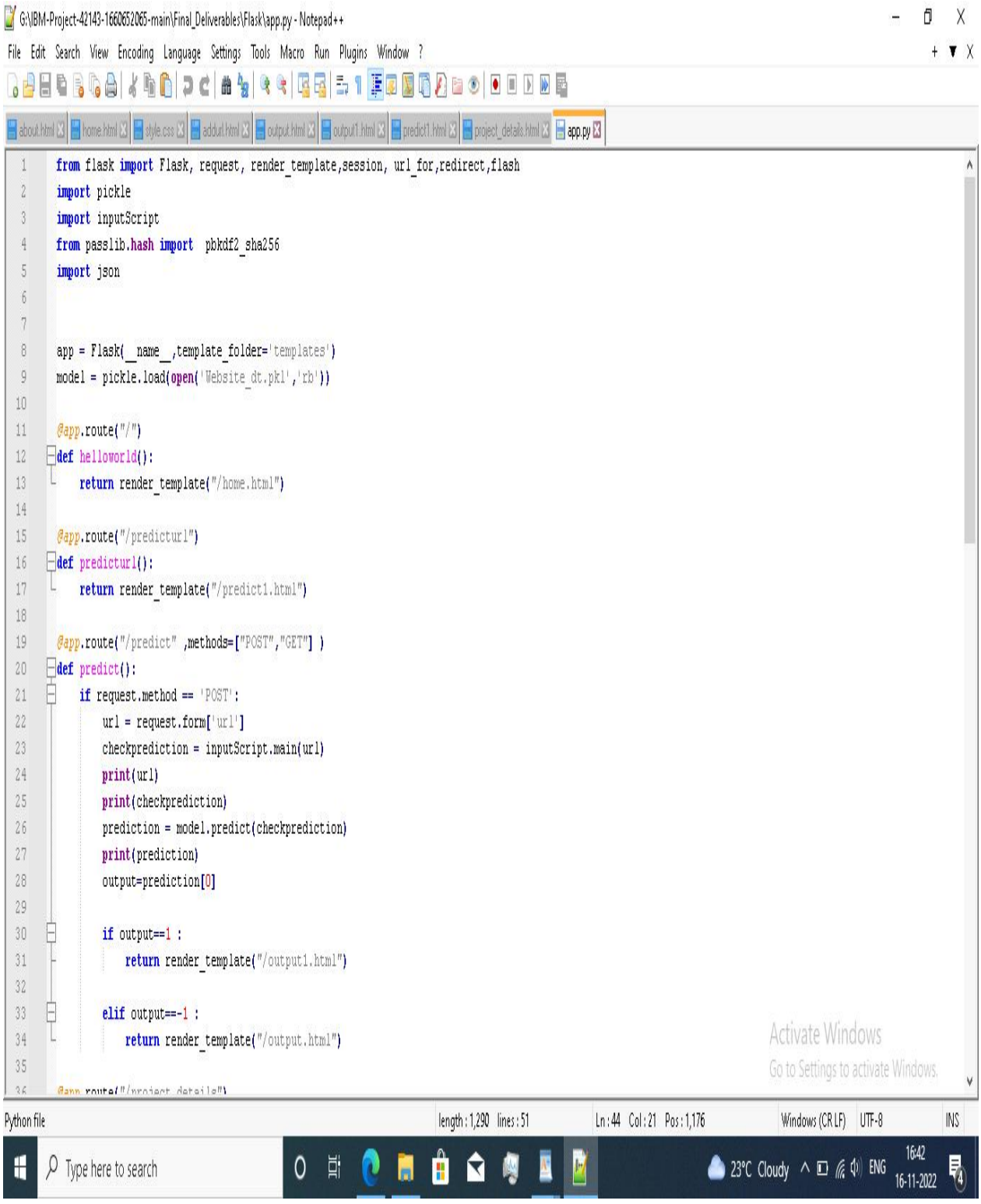
Testing is most important in the Web Phishing Detection, In the real time there are many new types of phishing attacks are arises and small scale and large scale industries and peoples who are suffering by this attack. so , the dataset need to update frequently

And also , its not a one way communication to the user while using our product he can able to communicate with our team by the "Help Page" where he can able to mention their problems and also send a complaint or a problem with our team.

Software testing is a critical element of software quality assurance and represents the ultimate review of specification, design and coding. In fact, testing is the one step in the software engineering process that could be viewed as destructive rather than constructive.

A strategy for software testing integrates software test case design methods into a well-planned series of steps that result in the successful construction of software. Testing is the set of activities that can be planned in advance and conducted systematically. The underlying motivation of program testing is to affirm software quality with methods that can economically and effectively apply to both strategic to both large and small-scale systems

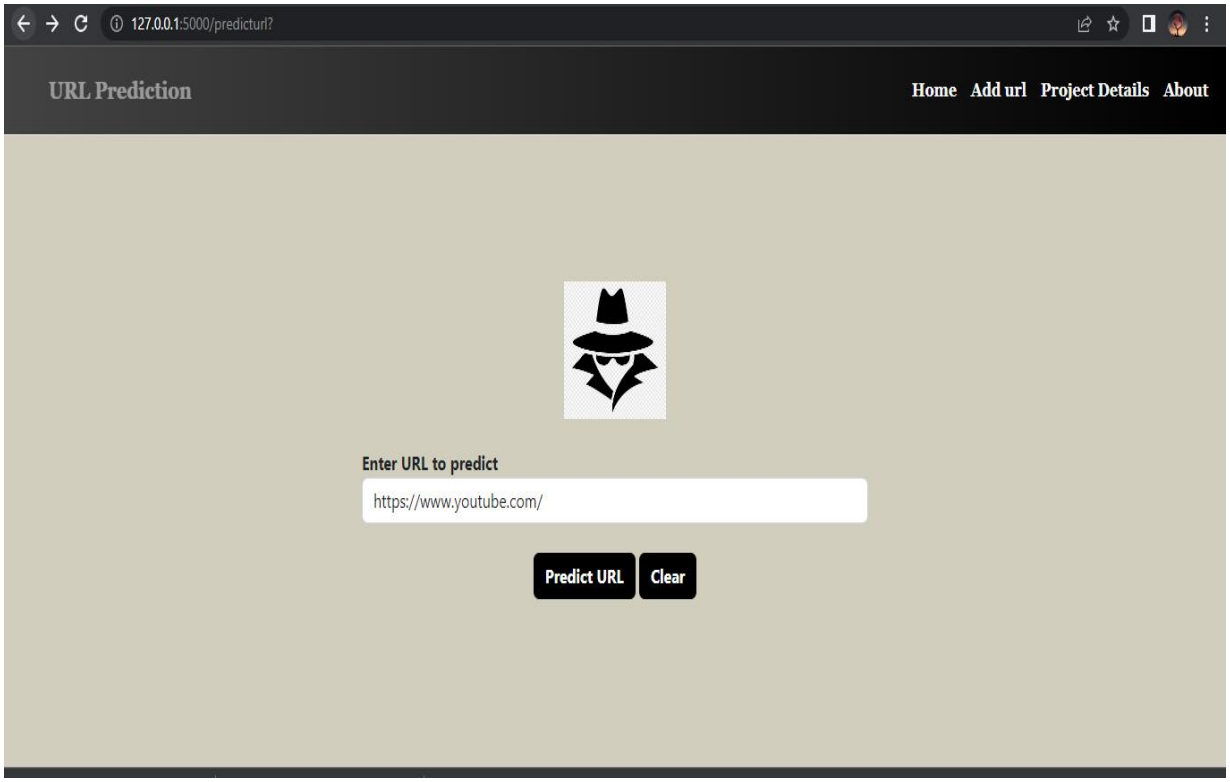
8.2 User Acceptance Testing



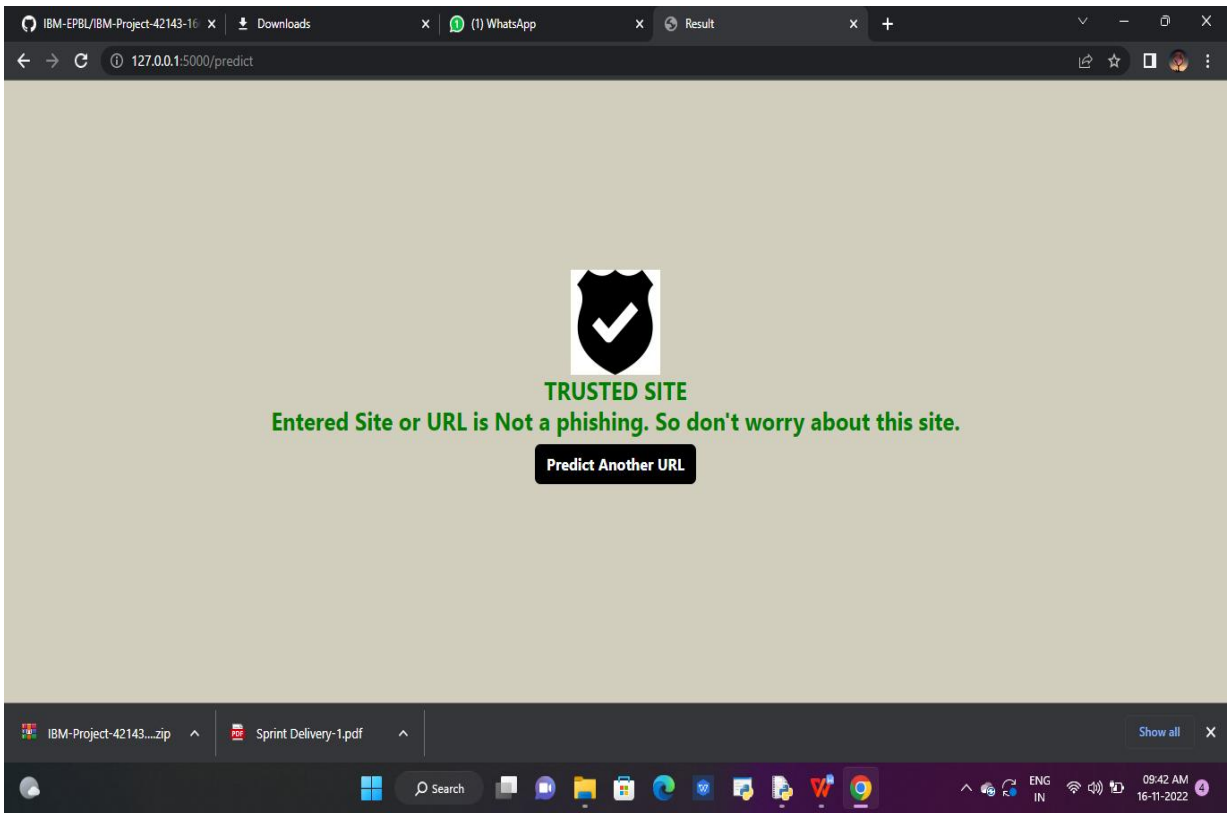
CHAPTER 9

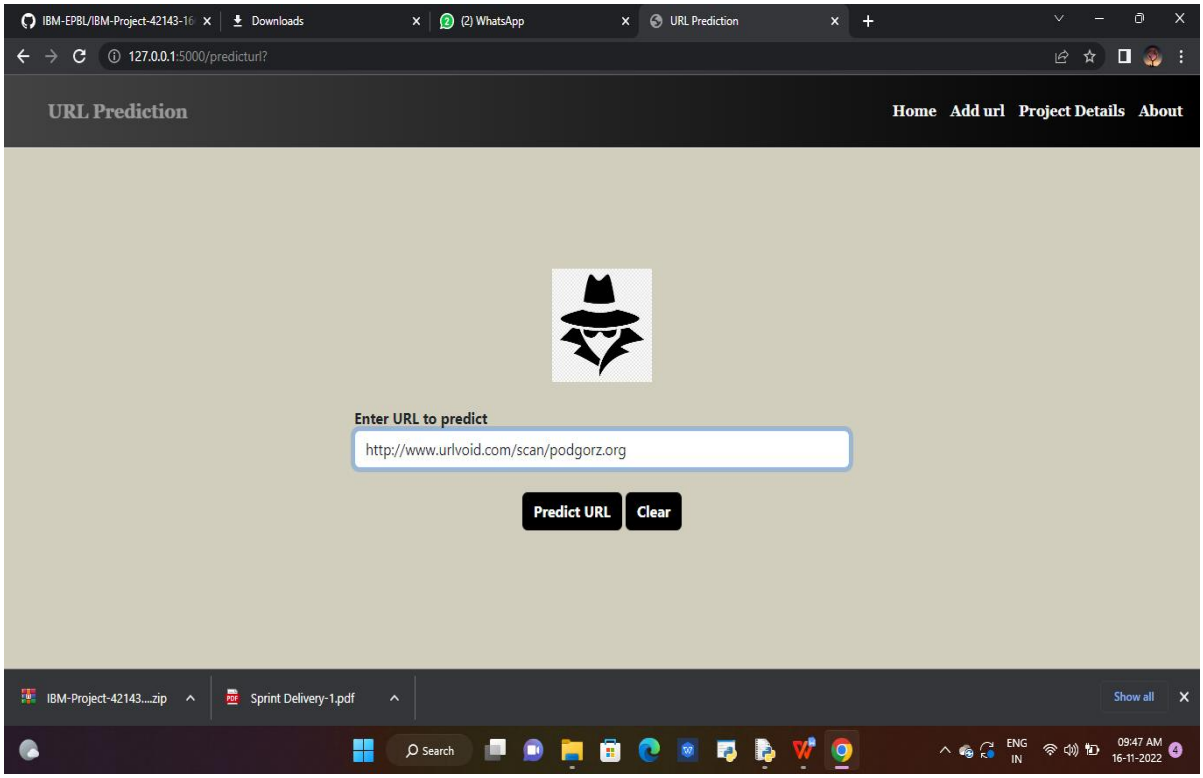
RESULT

9.1 Performance Metrics

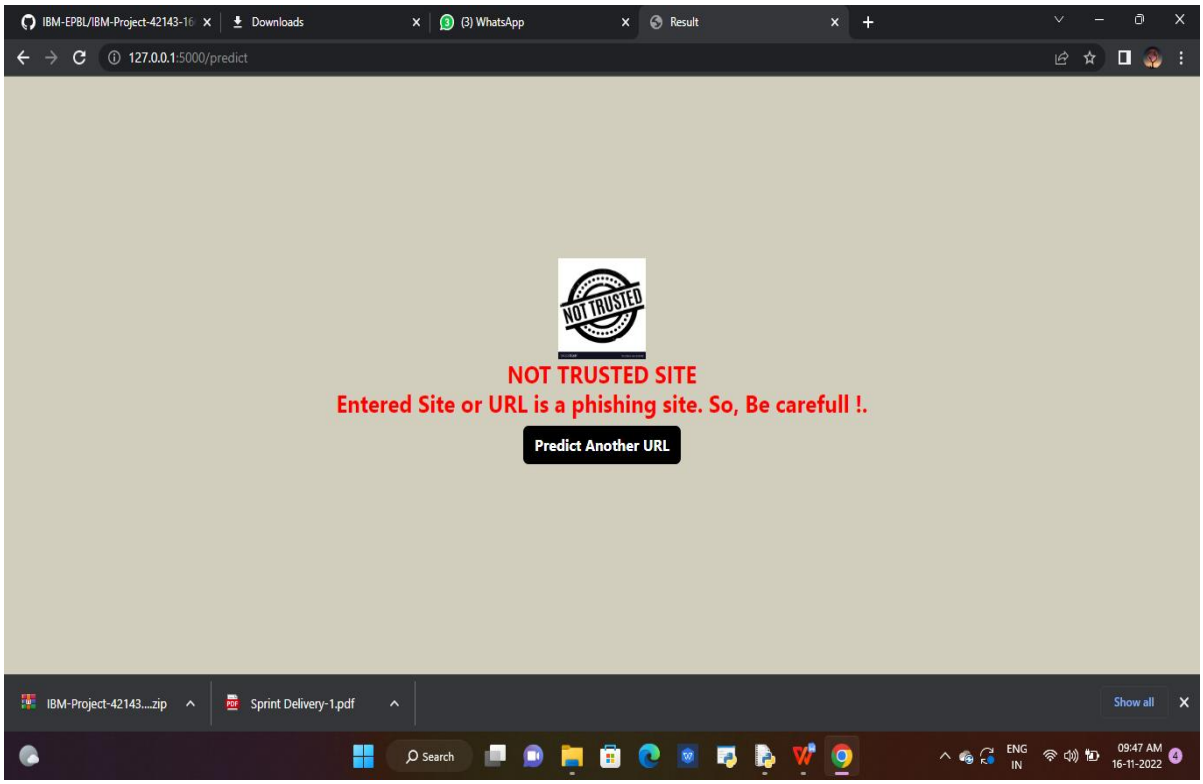


In the above fig we had entered a site like eg: www.youtube.com





For example took some scenario that some user try to enter some websites that before he went to our product and check this is a legitimate URL or not.



CHAPTER 10

ADVANTAGES AND DISADVANTAGES

9.2 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.
- The system improves the security of EFT's.
- Use of OTP provides second level of security.
- The system is cost effective and cheaper than usual EFT systems.
- New card requires less time for creation.
- It is safe communication technique.
- Main Advantage of this System is that other users will not know what communication is held by the 2 user.
- DES algorithm can hardly be cracked.

9. 3 Disadvantages

- If Internet connection fails, this system won't work.
- All websites related data will be stored in one place.
- If card is lost there is no way to interact with system.
- Needs a security guard at EFT center.
- OTP may take time to be received on mobile.
- It will need much more memory to safe encrypted message, and decrypt receiving messages.
- Process takes longer time than normal communication.
- It requires active internet connection else error may occur

CHAPTER 11

11. CONCLUSION

The demonstration of phishing is turning into an advanced danger to this quickly developing universe of innovation. Today, every nation is focusing on cashless exchanges, business online, tickets that are paperless and so on to update with the growing world. Yet phishing is turning into an impediment to this advancement. Individuals are not feeling web is dependable now. It is conceivable to utilize AI to get information and assemble extraordinary information items. A lay person, completely unconscious of how to recognize a security danger shall never invite the danger of making money related exchanges on the web. Phishers are focusing on installment industry and cloud benefits the most. The project means to investigate this region by indicating an utilization instance of reorganizing phishing sites utilizing ML. It aimed to build a phishing detection mechanism using machine learning tools and techniques which is efficient, accurate and cost effective. The project was carried out in Anaconda IDE and was written in Python. The proposed method used four machine learning classifiers to achieve this and a comparative study of the four algorithms was made. A good accuracy score was also achieved. The four algorithms used are K-Nearest neighbor, Kernel Support Vector Machine, De51 Conclusion & future works Detection of phishing websites using machine learning techniques decision Tree and Random Forest Classifier. All the four classifiers gave promising results with the best being Random Forest Classifier with an accuracy score of 96.82%. The accuracy score might vary while using other datasets and other algorithms might provide better accuracy than random forest classifier. Random forest classifier is an ensemble classifier and hence the high accuracy. This model can be deployed in real time to detect the URLs as phishing or legitimate.Ensemble methods is a ML technique that combines many base models to generate an optimal predictive model. Further reaching future work would be combining multiple classifiers, trained on different aspects of the same training set, into a single classifier that may provide a more robust prediction than any of the single classifiers on their own. to complete the system. Looking even further out, the methodology needs to be evaluated on how it might handle collection growth. The collections will ideally grow incrementally over time so there will need to be a way to apply a classifier incrementally to the new data, but also potentially have this classifier receive feedback that might modify it over time.

CHAPTER 12

10. FUTURE SCOPE

Further work can be done to enhance the model by using ensembling models to get greater accuracy score. Ensemble methods is a ML technique that combines many base models to generate an optimal predictive model. Further reaching future work would be combining multiple classifiers, trained on different aspects of the same training set, into a single classifier that may provide a more robust prediction than any of the single classifiers on their own. The project can also include other variants of phishing like smishing, vishing, etc. to complete the system. Looking even further out, the methodology needs to be evaluated on how it might handle collection growth. The collections will ideally grow incrementally over time so there will need to be a way to apply a classifier incrementally to the new data, but also potentially have this classifier receive feedback that might modify it over time.

Although the use of URL lexical features alone has been shown to result in high accuracy (97%), phishers have learned how to make predicting a URL destination difficult by carefully manipulating the URL to evade detection. Therefore, combining these features with others, such as host, is the most effective approach .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.

CHAPTER 13

APPENDIX

Source code

Github & Project Demo link

13.1 SOURCE CODE

13.1.1 Python Flask Coding

```
from flask import Flask, request, render_template,session, url_for,redirect,flash
import pickle
import inputScript
from passlib.hash import pbkdf2_sha256
import json
app = Flask(__name__,template_folder='templates')
model = pickle.load(open('Website_dt.pkl','rb'))
@app.route("/")
def helloworld():
    return render_template("/home.html")
@app.route("/predicturl")
def predicturl():
    return render_template("/predict1.html")
@app.route("/predict" ,methods=["POST","GET"] )
def predict():
    if request.method == 'POST':
        url = request.form['url']
        checkprediction = inputScript.main(url)
        print(url)
        print(checkprediction)
        prediction = model.predict(checkprediction)
        print(prediction)
        output=prediction[0]
        if output==1 :
            return render_template("/output1.html")
        elif output==-1 :
            return render_template("/output.html")
@app.route("/project_details")
def support():
    return render_template("/project_details.html")
@app.route("/addurl")
def addurl():
    return render_template("/addurl.html")
@app.route("/about")
def about():
    return render_template("/about.html")
if __name__ =="__main__":
    app.run(debug=True)
```

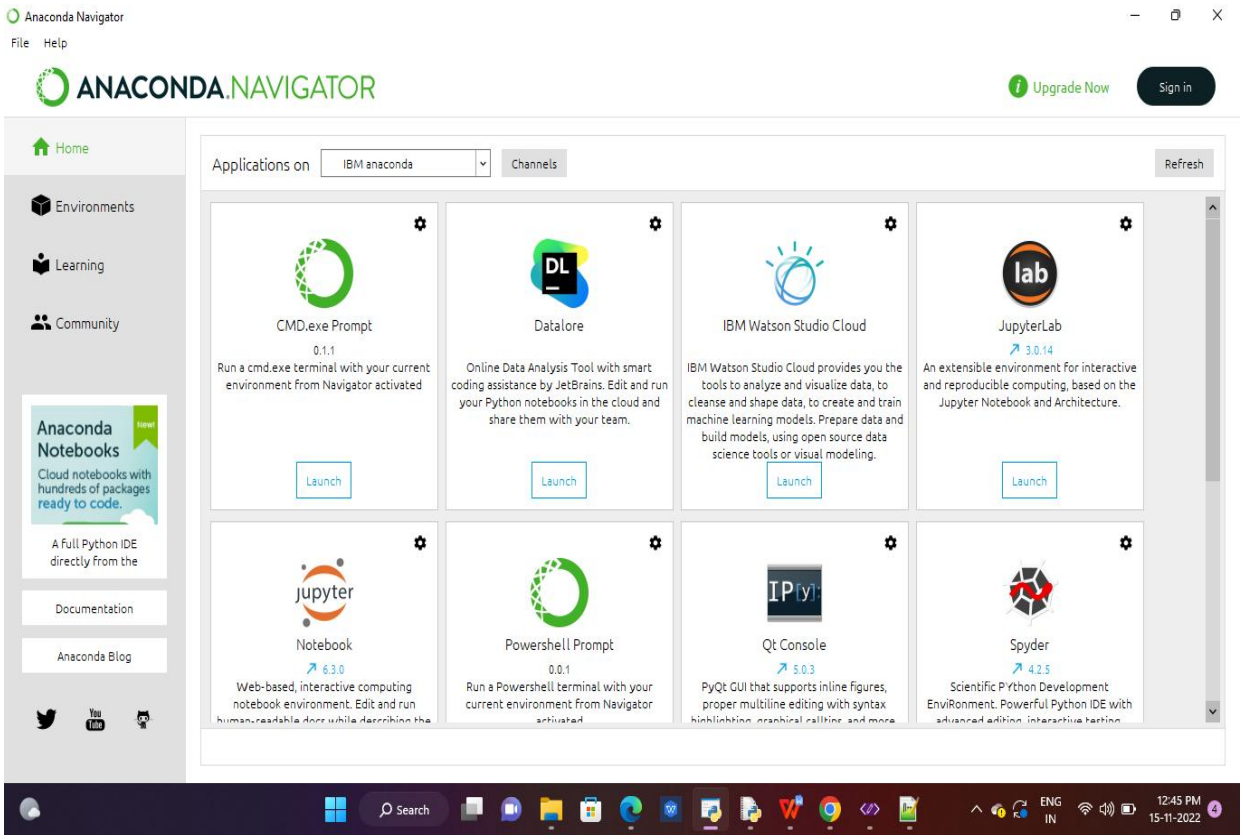


Fig 13.1.1.1 show the Running of jupyter notebook

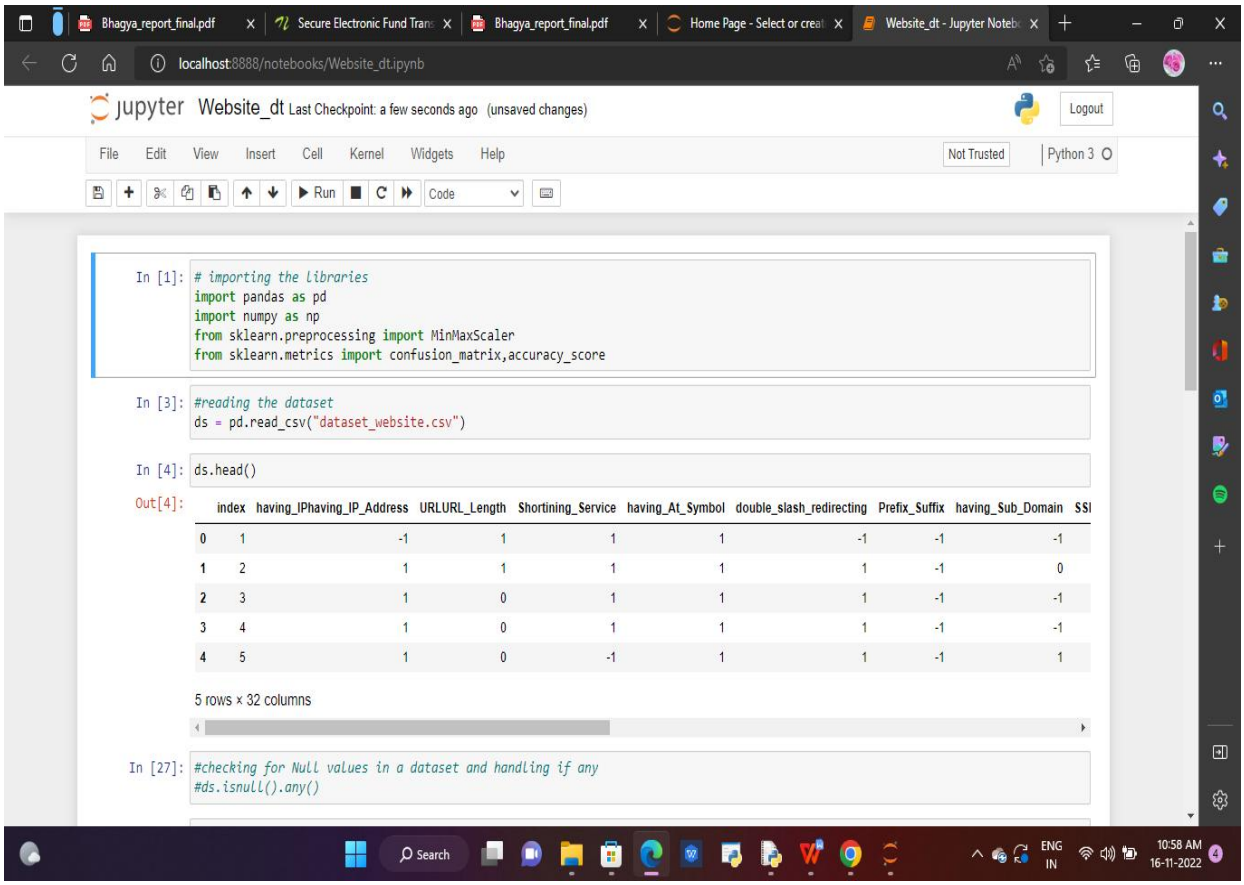


Fig 13.1.1.2 show the read of data set

GITHUB & PROJECT DEMO LINK

- https://github.com/IBM-EPBL/IBM-Project-43589-1660718318

- https://drive.google.com/file/d/1ilNIQNZp1eN9G-MtFeCAN8jLCjFyI-Bh/view?usp=drivesdk