K.L.N College of Information Technology, Pottapalayam

Department of (B.TECH-IT)

Sub.Code & Sub.Name: HX 8001 & Professional Readiness for Innovation, Employability and Entrepreneurship

"Project Report"

"WEB PHISHING DETECTION"

Team ID: PNT2022MID52526

Guided by,
Ms.B.K.Hemalatha,
HOD(I/c)/ECE,
KLNCIT

Submitted by,

1.A.Pooja(910719205014)

-Team Leader

2. A.P.Abirami(910719205002)

-Team Member

3. S.Chelsea Evelyn Christina

(910719205005) – Team Member

4. P.Shiny Jaculine Mary

(910719205022) – Team Member

TABLE OF CONTENT

1.INTRODUCTION

Phishing is a social engineering attack that aims at exploiting the weakness found in system processes as caused by system users. For example, a system can be technically secure enough against password theft, however unaware end users may leak their passwords if an attacker asked them to update their passwords via a given Hypertext Transfer Protocol (HTTP) link, which ultimately threatens the overall security of
the system, technical vulnerabilities (e.g. Domain Name System
(DNS) cache poisoning) can be used by attackers to construct far more persuading socially-engineered messages (i.e. use of legitimate, but spoofed, domain names can be far more persuading than using different domain names).

1.1 PROJECT OVERVIEW

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.

This Guided Project mainly focuses on applying a machine-learning algorithm to detect Phishing websites.

In order to detect and predict e-banking phishing websites, we proposed an intelligent, flexible and effective system that is based on using classification algorithms. We implemented classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy. The e-banking 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 a user makes a transaction online when he makes payment through an e-banking website our system will use a data mining algorithm to detect whether the e-banking website is a phishing website or not.

1.2 PURPOSE

Various fraudulent websites have been built on the World Wide Web in the previous decade to resemble reputable websites and steal financial assets from users and organizations. This type of online scam is known as phishing, and it has cost the internet community and other stakeholders hundreds of millions of dollars. As a result, robust countermeasures that can identify phishing are required.
These are the challenges to be addressed in this project: a. Reduce the rate of financial theft from users and organizations online. b. Educate Internet Users on the deception of phishers. c. Educate Internet users on the countermeasures of a phishing attack. arious fraudulent websites have been built on the World Wide Web in the previous decade to resemble reputable websites and steal financial assets from users and organizations. This type of online scam is known as phishing, and it has cost the internet community and other stakeholders hundreds of millions of dollars.
As a result, robust countermeasures that can identify phishing are required. These are the challenges to be addressed in this project: a. Reduce the rate of financial theft from users and organizations online. b. Educate Internet Users on the deception of phishers. c. Educate Internet users on the countermeasures of a phishing attack. To accomplish the project's purpose, the following particular objectives have been established: i. dataset collection and pre-processing;
ii. machine-learning model selection and development ;
iii. development of a web-based application for detection;
iv. Integration of the developed model to a web application.

2.LITERATURE SURVEY:

2.1 Existing Problem

An extensive review was done on existing works of literature and machine learning models on detecting phishing websites to best decide the classification models to solve the problem of detecting phishing websites. Hence, Series of these machine learning classification models such as Decision Tree, Support Vector Machine, XGBooster, Multilayer perceptions, Auto encoder Neural Network and Random Forest was deployed on the dataset to distinguish between phishing and legitimate URLs. The best model with high training accuracy out of all the deployed models was selected then integrated into a developed web application. Thus, a user can enter a URL link on the web application to predict if it is phishing or legitimate.

2.2 References

Abdelhamid, N., Thabtah F., & Abdel-Jaber, H. Phishing detection: A recent intelligent machine learning comparison based on models' content and features," 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, 2017, pp. 72-77, DOI: 10.1109/ISI.2017.8004877.

Anjum N. S., Antesar M. S., & Hossain M.A. (2016). A Literature Review on Phishing Crime, Prevention Review and Investigation of Gaps. Proceedings of the 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), Chengdu, China, 2016, pp. 9-15, DOI: 10.1109/SKIMA.2016.7916190.

Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). A survey of phishing email filtering techniques, Proceedings of IEEE Communications Surveys and Tutorials, vol. 15, no. 4, pp. 2070–2090.

Ashritha, J. R., Chaithra, K., Mangala, K., & Deekshitha, S. (2019). A Review Paper on Detection of Phishing Websites using Machine Learning.Proceedings of International Journal of Engineering Research & Technology (IJERT), 7, 2. Retrieved from www.ijert.org.

Ayush, P. (2019). Workflow of a Machine Learning project. Retrieved from https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94

Camp W. (2001). Formulating and evaluating theoretical frameworks for career and technical education research. Journal of Vocational Education Research, 26(1), 4-25.

DeepAI (n.d.). About clinical psychology. Retrieved from https://deepai.org/machine-learning-glossary-and-terms/feature-extraction

Engine K., & Christopher K. (2005). Protecting Users Against Phishing Attacks. Proceedings of the Oxford University Press on behalf of The British Computer Society, Oxford University, 0, 2005, Retrieved from: https://sites.cs.ucsb.edu/~chris/research/doc/cj06_phish.pdf

Gandhi, V. (2017). A Theoretical Study on Different ways to identify the Phishing URL and Its Prevention Approaches: presented at International Conference on Cyber Criminology, Digital Forensics and Information Security at DRBCCC Hindu College, Chennai. Retrieved from https://www.researchgate.net/publication/319006943_A_Theoretical_Study_on_Different_ways_to_Identify_the_Phishing_URL_and_Its_Prevention_Approaches

Gupta, B. B., Tewari, A., Jain, A. K., & Agrawal, D. P. (2016). Fighting againstphishing attacks: state of the art and future challenges, Neural Computing and Applications.https://www.imperva.com/learn/application-security/phishing-attack-scam/

Noel, B. (2016). Support Vector Machines: A Simple Explanation. Retrieved from https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html

Osanloo, A., & Grant, C. (2016). Understanding, selecting, and integrating a theoretical framework in dissertation research: creating the blueprint for your “house”. Administrative issues journal: connecting education, practice and research 4(2), 7.
Peng, T., Harris, I., & Sawa, I. (2018). Detecting Phishing Attacks Using Natural Language Processing and Machine Learning. Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301.

Pamela (2021). Phishing attacks. Retrieved from https://www.khanacademy.org/computing/computers-andinternet/xcae6f4a7ff015e7d:online-data-security/xcae6f4a7ff015e7d:cyber-attacks/a/phishing-attacks

Rami, M. M., Fadi, T., & Lee, M. (2015). Phishing Websites Features. Retrieved from https://eprints.hud.ac.uk/id/eprint/24330/6/MohammadPhishing14July2015.pdf

Rishikesh, M., & Irfan, S. (2018a). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications, 23, 45. doi:10.5120/ijca2018918026

Rishikesh, M., & Irfan, S. (2018b). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications, 23, 45-46. doi:10.5120/ijca2018918026

Rahul, S. (2017). How the decision tree algorithm works. Retrieved from https://dataaspirant.com/how-decision-tree-algorithm-works/

Rishikesh, M., & Irfan, S. (2018c). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications, 23, 46-47. doi:10.5120/ijca2018918026

Saimadhu, P. (2017). How the random forest algorithm works in machine learning. Retrieved from https://dataaspirant.com/random-forest-algorithm-machine-learing/

2.3 Problem Statement Definition

Internet has dominated the world by dragging half of the world's population exponentially into the cyber world. With the booming of internet transactions, cybercrimes rapidly increased and with anonymity presented by the internet.Hackers attempt to trap the end-users through various forms such as phishing, SQL injection, malware, man-in-the-middle,domain name system tunnelling, ransomware, web trojan, and so on.Among all these attacks, phishing reports to be the most deceiving attack.

3.IDEATION& PROPOSED SOLUTION

3.1 Empathy Map

To avoid installationof harmful malwares

WHAT DO THEY THINK AND FEEL

Concerned about their privacy and safety.

When the data is obtained,they start to steal their information.

WHAT DO THEY SAY AND DO

Making the victim's believe they are in real problem to steal money or data.

Constant fear &frustration of getting their money stolen.

PAIN & GAIN

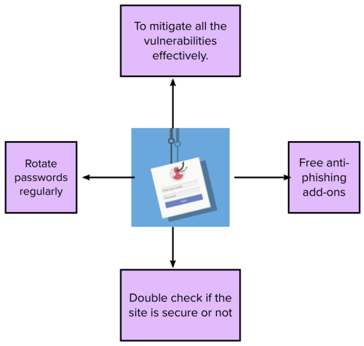
Secure personal data,payment process and web browsing.

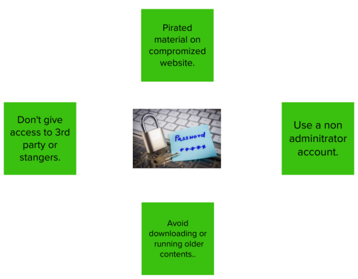
They are mearly aimed at stealing personal data.

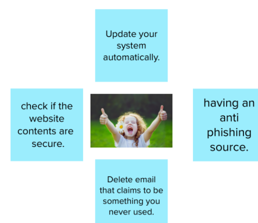
HEAR & SEE

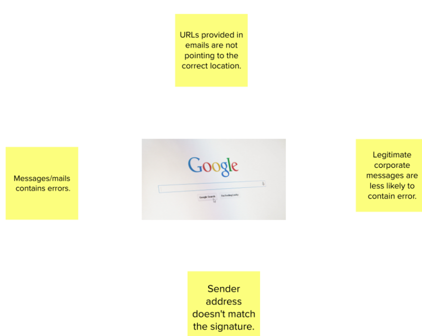
Innocent People who lost their money to Phishing means.

3.2 Ideation& Brainstorming









BRAINSTROMING

P.Shiny Jaculine Mary

- Low detection accuracy
- High false alarm
- No comprehensive blacklist
- Suspicious activity in the page
- Inefficient in responding
- User Vulnerability

A.P.Abirami

- Scamming people for money
- Harmful ransomware
- Stealing sensitive details
- Affects the performance of the affected system
- Spread malicious code onto recipients' computers.
- Convince you to willingly send money or valuables

A.Pooja

- Leading to information leakage and blackmail.
- Protecting password by using combination of special characters, alphanumeric.
- Always filter email.
- Do not trust the mail that says spam.
- Check before entering passwords in 3rd party sites.
- Check if the URL contains any suspicious characters,it coul be a phishhing site.

S.Chelsea Evelyn Christina

- Attacker sends a fraudulent message.
- It tricks a person into revealing sensitive information.
- As of 2020 phishing is by far the most common attack.
- Phishing attacks have become increasingly sophisticated.
- Prevention of phishing can be done by using user training and public awareness.
- Phishing allows to view a persons details while navigating each site without their knowledge.

3.3 PROPOSED SOLUTION

S.No.	Parameter	Description
1.	Idea / Solution description	This study explores data science and machine learning models that use datasets obtained from open-source platforms to analyze website links and distinguish between phishing and legitimate URL links..
2.	Novelty / Uniqueness	The model will be integrated into a web application, allowing a user to predict if a URL link is legitimate or phishing. This online application is compatible with a variety of browsers enhance better results in the identification and prevention of phishing attacks.
3.	Social Impact / Customer Satisfaction	By using our phishing detection, both the organisation and their customers can be safe and can avoid identity theft, data stealing etc..
4.	Business Model (Revenue Model)	Phishing could often gain a foothold in corporate or governmental networks as a part of larger attacks, such Threats lead to severe financial losses in addition to declining market share,reputation and consumer trust.
5.	Scalability of the Solution	The proposed model focuses on identifying the phishing attack based on checking phishing websites features, Blacklist and WHOIS database. A few selected features can be used to differentiate between legitimate and spoofed web pages. These selected features are many such as URLs, domain identity, security & encryption, source code, page style and contents, web address bar and social human factor.This paper presents a proposal for scalable detection and isolation of phishing and deployment of the machine learning algorithms.

3.4 Problem Solution fit

Customer Segment	Anyone who uses web browser,surfs the internet, ● Organisation ● Individuals.
Problems	1. Breach of privacy 2. Loss of data,reputation 3. Identity theft 4. Victim to malware,ransomeware
Triggers	1. site is blocked ,phishing site 2. tigger warning displayed
Emotions	BEFORE: Constant fear of losing their data and insecure of privacy breach AFTER: Feeling Protected and safe.
Available Solution	1. Blacklist 2. Anti-spam software 3. Firewall
Customer Constraints	No adequate knowledge,constrain at implementing of resources and need of internet access
Behaviour	What to do and not to do
Channels of Behaviour	Online- Tend to lose their data online phishing site. Offline- By learning via books and other resources.
Problem root cause	IT's difficult for someon to determine if the site is legit or not.
Our Solution	To implement Machine learning (Decision tree algorithm).

4.REQUIREMENT ANALYSIS

4.1 Functional requirement

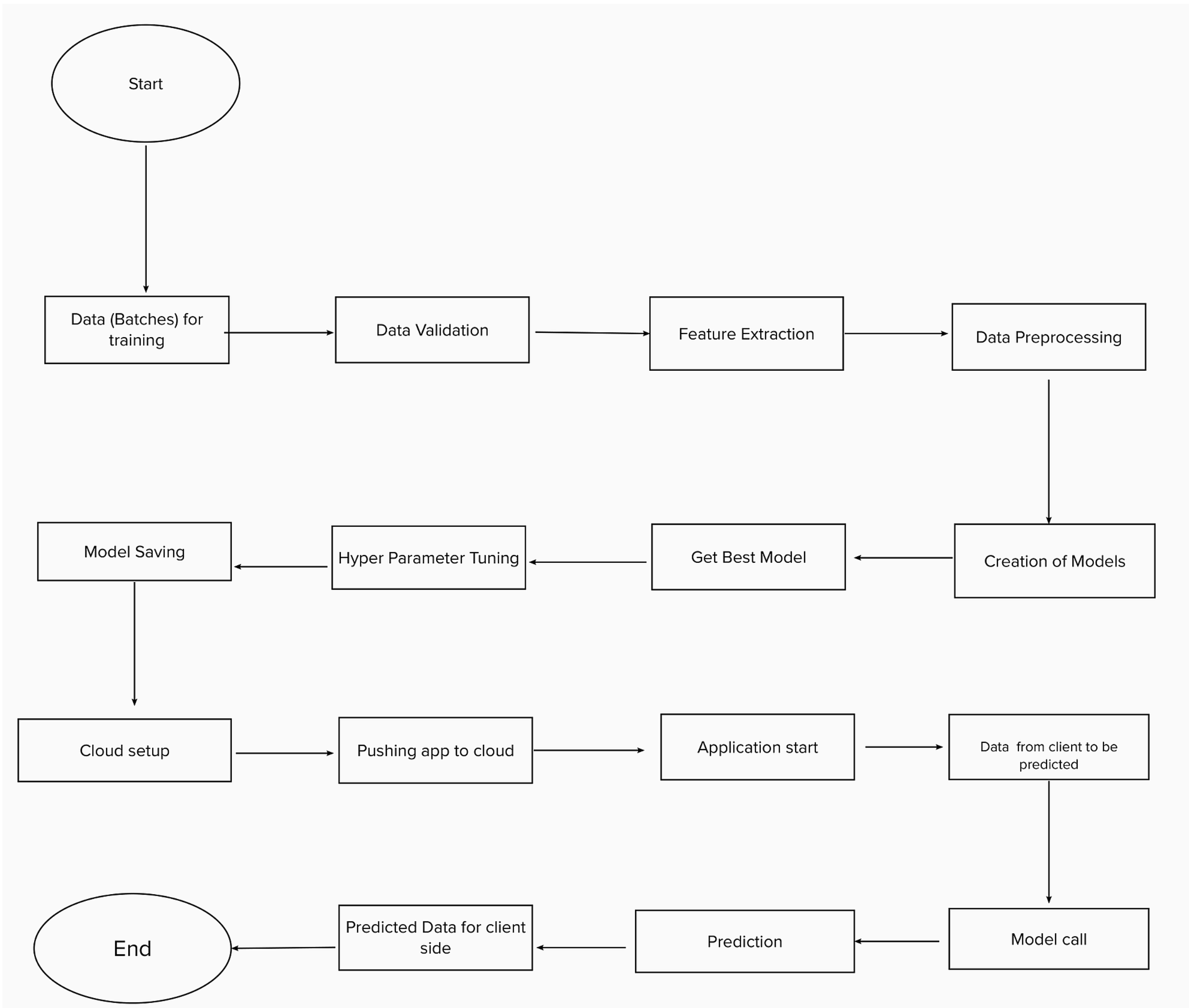
FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIN
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	Website Analyze&Preprocessing	Our system should be able to load air quality data and preprocess data. It should be able to analyze the air quality data
FR-4	Prediction	It should be able to group data based on hidden patterns. It should be able to assign a label based on its data groups.
FR-5	Classification	It should be able to split data into trainset and testset. • It should be able to train model using trainset. It must validate trained model using testset.
FR-6	Result	It should be able to display the trained model accuracy. It should be able to accurately predict the air quality on unseen data.

4.2 Non-Functional requirements

FR No.	Non-Functional Requirement	Description
NFR-1	Usability	Since the writing computer programs is extremely straightforward, it is simpler to discover and address the imperfections and to roll out the improvements in the undertaking
NFR-2	Security	High level of security is ensured.
NFR-3	Reliability	It enlists the different permutations and combinations a system can be reused in many other applications which gives better prediction, as well as gives a new approach to prediction techniques.
NFR-4	Performance	The user interface allows the user to interact with the system at a very comfortable level with no hassles.
NFR-5	Availability	To depict how much an item, gadget, administration, or condition is open by however many individuals as would be prudent.
NFR-6	Scalability	low data transfer capacity and substantial number of clients

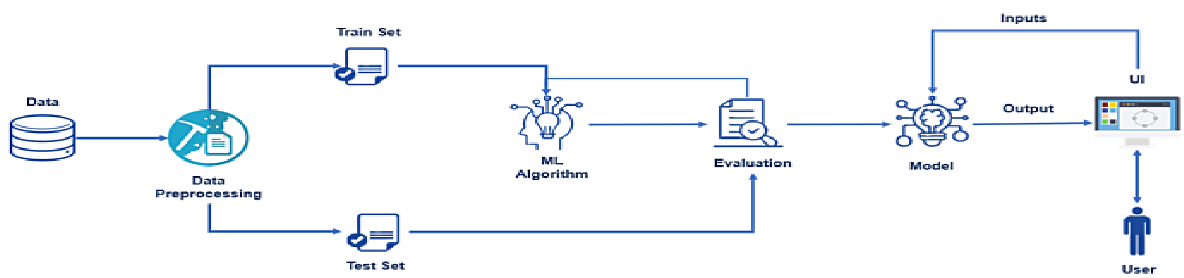
5.PROJECT DESIGN

5.1 Data Flow Diagrams

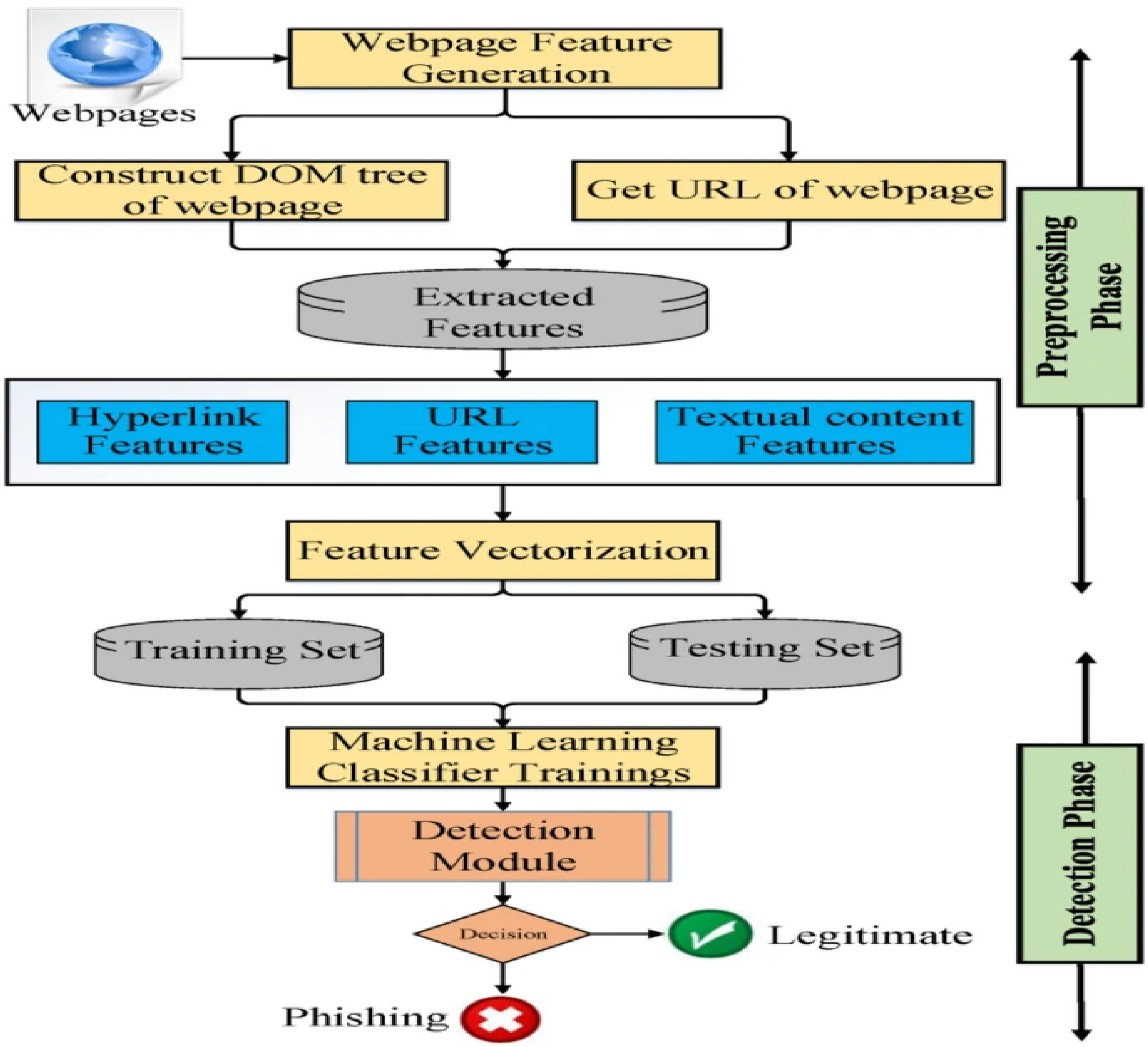


5.2 Solution & Technical Architecture

Technical Architecture



SOLUTION ARCHITECTURE:



5.3 User Stories

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	Access my information any time from the cloud.	Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password	Login anywhere with my information.	High	Sprint-1
	Dashboard					
Customer (Web user)	User Input	USN-1	As a user i can input the particular URL in the required field .	I can access the website without any problem.	High	Sprint-1
Customer Care Executive	Extraction	USN-1	After i compare in case if none found on comparison then we can extract feature using heuristic and visual similarity approach.	As a user i can have comparison between websites for security.	Medium	Sprint-1
Administrat or	Processing & Prediction	USN-1	Here the Model will predict the URL websites using Machine Learning algorithms such as Logistic Regression,sv m.	In this i can have correct prediction on the particular algorithms.	Medium	Sprint-1

	Classification	USN-2	Here I will send all the model output to classifier in order to produce final result.	In this i will find the correct classifier for producing the result	Medium	Sprint-1
	Detection	USN-3	After the extraction purpose the model will be able to categorize it from other safe website through data mining classification technique through ML.	I can determine whether the website is from secure website or not.	High	Sprint-1
	End result	USN-4	I can access, verify my results.	I can view the final output given to me by the administrator.	High	Sprint-1

6.PROJECT PLANNING&SCHEDULING

6.1 Sprint planning & Estimation

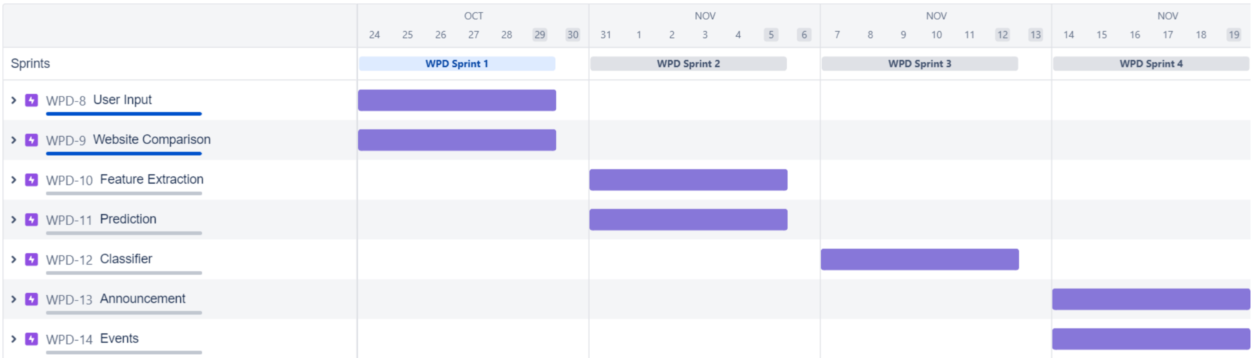
Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Homepage	USN-1	As a user, I can enter by just entering the site's URL or clicking	2	Medium	A.P.Abirami A.Pooja

			the site’s link..			
Sprint-1		USN-2	As a user, I will receive information and pieces of Phishing scams and prevention.	1	Low	A.P.Abirami
Sprint-4	Result	USN-3	As a user, I will know the site's legitimacy.	2	Low	A.Pooja Shiny jaculine
Sprint-2	Prediction	USN-4	As a user, I can just sit and watch the site predicting the URl	2	Medium	Chelsea
Sprint-3	Training The Model on IBM	USN-5	TASK- To make access and prediction	1	High	Shiny jaculine mary
Sprint-3	Deploying Model in IBM cloud	USN-6	TASK- Deploying the model on cloud and running it to predict the site’s.	2	High	A.P.Abirami A.Pooja

6.2 Sprint Delivery Schedule

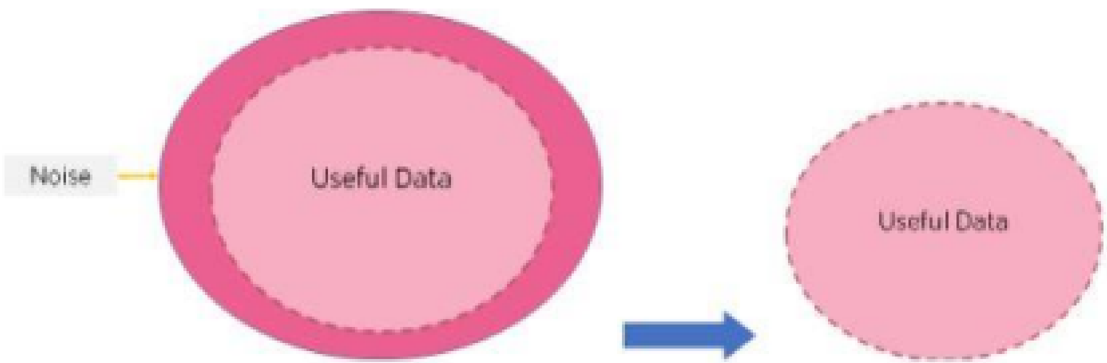
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

6.3 Reports from JIRA



7.CODING&SOLUTIONING

7.1 Feature 1



i. Supervised Models:

Supervised feature selection refers to the method which uses the output label class for feature selection. They use the target variables to identify the variables which can increase the efficiency of the model.

ii. Unsupervised Models:

Unsupervised Feature selection refers to the method which does not need the output label class for feature selection. We use them for unlabeled data. shows the flow of the feature selection model.

7.2 Feature 2

i. Using the IP Address

If an IP address is used as an alternative to the domain name in the URL, such as "http://125.98.3.123/fake.html", users can be sure that someone is trying to steal their personal information. Sometimes, the IP address is even transformed into hexadecimal code as shown in the following link "http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".

Rule: IF{If The Domain Part has an IP Address → Phishing
Otherwise → Legitimate

(2) Long URL to Hide the Suspicious Part

i.Phishers can use a long URL to hide the doubtful part in the address bar. For example:

http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e8110

04d58f5b74f8dc1e7c2 e8dd4105e8@phishing.website.html

To ensure the accuracy of our study, we calculated the length of URLs in the dataset and produced an average URL length. The results showed that if the length of the URL is greater than or equal to 54 characters then the URL is classified as phishing. By reviewing our dataset, we were able to find 1220 URL lengths equals 54 or more which constitute 48.8% of the total dataset size.

ii. Presence of @ symbol in URL: If @ symbol is present in URL then the feature is set to 1 else set to 0. Phishers add special symbol @ in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol .

iii. The number of dots in Hostname: Phishing URLs have many dots in URL. For example, http://shop.fun.amazon.phishing.com, in this URL phishing.com is an actual domain name, whereas the use of the "amazon" word is to trick users to click on it. The average number of dots in benign URLs is 3. If the number of dots in URLs is more than 3 then the feature is set to 1 else to 0.

8.TESTING

8.1 Test Cases

Testcase 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
LoginPage_TC_OO1	Functional	Home Page	Verify user is able to see the LandingPage when user can type the URL in the box		1.Enter URL and click go 2.Type the URL 3.Verify whether it is processing or not.	https://phishing-shield.herokuapp.com/	Should Display the Webpage	Working as expected	Pass		N		A.P.Abirami
LoginPage_TC_OO2	UI	Home Page	Verify the UI elements is Responsive		1.Enter URL and click go 2. Type or copy paste the URL 3. Check whether the button is responsive or not 4. Reload and Test Simultaneously	https://phishing-shield.herokuapp.com/	Should Wait for Response and then gets Acknowledge	Working as expected	Pass		N		A.Pooja
LoginPage_TC_OO3	Functional	Home page	Verify whether the link is legitimate or not		1.Enter URL and click go 2. Type or copy paste the URL 3. Check the website is legitimate or not 4. Observe the results	https://phishing-shield.herokuapp.com/	User should observe whether the website is legitimate ornot.	Working as expected	Pass		N		A.P.Abirami
LoginPage_TC_OO4	Functional	Home Page	Verifyuser is able to access the legitimate website or not		1.Enter URL and click go 2. Type or copy paste the URL 3. Check the website is legitimate or not 4. Continue if the website is legitimate or be cautious if it is not legitimate.	https://phishing-shield.herokuapp.com/	Application should show that Safe Webpage or Unsafe.	Working as expected	Pass		N		A.P.Abirami
LoginPage_TC_OO5	Functional	Home Page	Testing the website with multiple URLs		1.Enter URL (https://phishing-shield.herokuapp.com/) and click go 2. Type or copy paste the URL to test 3. Check the website is legitimate or not 4. Continue if the website is secure or be cautious if it is not secure	https://www.google.com/	User can able to identify the websites whether it is secure or not	Working as expected	Pass		N		A.P.Abirami

8.2 User Acceptance Testing

1. Defect Analysis

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	10	2	4	20	36
Not Reproduced	0	0	1	0	1
Skipped	0	0	0	0	0
Won't Fix	0	0	2	1	3
Totals	23	9	12	25	60

2. 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	10	0	0	10
Client Application	50	0	0	50
Security	5	0	0	4
Outsource Shipping	3	0	0	3
Exception Reporting	10	0	0	9
Final Report Output	10	0	0	10
Version Control	4	0	0	4

9.RESULTS

9.1 Performance Metrics

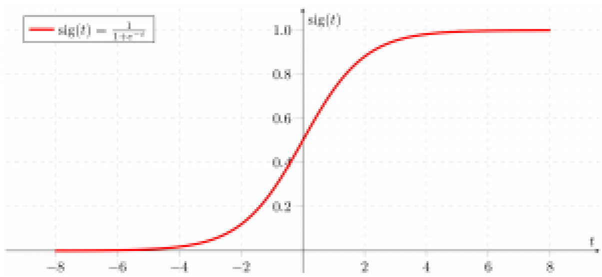
S.No.	Parameter	Values	Screenshot
1.	Metrics	Regression Model: Logistic Regression MAE – 0.26142017186793304 MSE - 0.5228403437358661 RMSE - 0.7230769971004928 R2 score - - 2.888673182487615 Classification Model: Decision Tree Classifier Confusion Matrix - array([[61, 249], [26, 1875]]) Accuracy Score- 0.8756218905472637 Classification Report – refer screenshot	Attached Below
2.	Tune the Model	Hyperparameter Tuning - Validation Method -	Attached Below

1.METRICS:

REGRESSION MODEL: LOGISTIC REGRESSION

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X.

Contrary to popular belief, logistic regression is a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.



EVALUATION METRICS:

Here are some evaluation metrics used for regression they are,

- R2 Score:

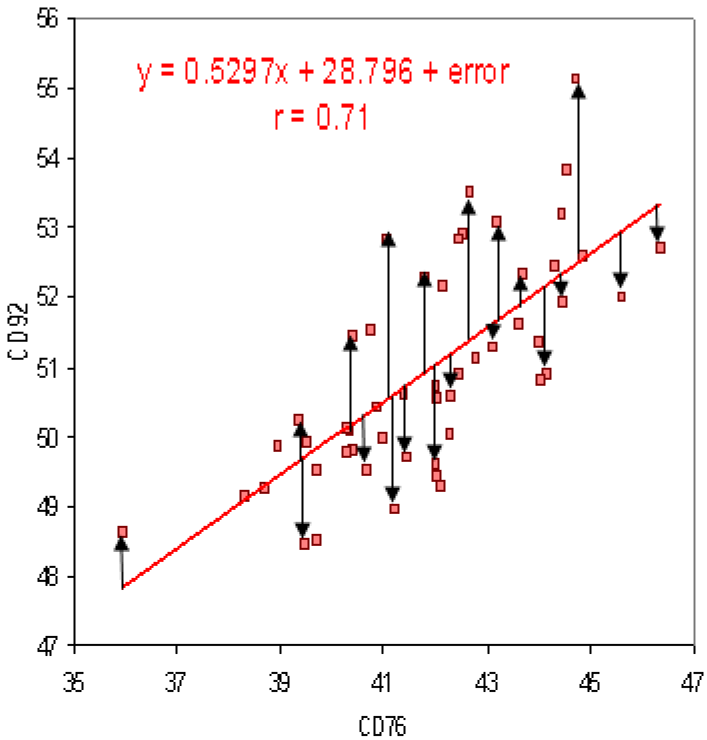
A low value would show a low level of correlation, meaning a regression model that is not valid, but not in all cases.The r2 score varies between 0 and 100%. It is closely related to the MSE , but not the same.

- Mean Square Error(MSE)

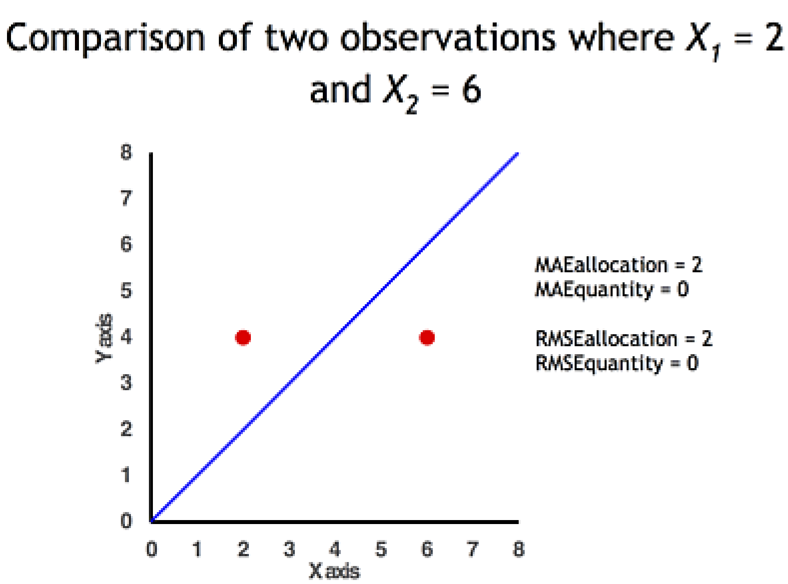
Mean square error (MSE) is the average of the square of the errors. The larger the number the larger the error. Error in this case means the difference between the observed values y1, y2, y3, … and the predicted ones pred(y1), pred(y2), pred(y3), … We square each difference (pred(yn) – yn)) ** 2 so that negative and positive values do not cancel each other out.

- Root Mean Square Error (RMSE)

RMSE is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.



Mean Absolute Error(MAE)



- *x axis= true value ; y axis= prediction*

- *Mean Absolute Error* is a model evaluation metric used with regression models. The mean absolute error of a model with respect to a test setis the mean of the absolute values of the individual prediction errors on over all instances in test set. Each prediction error is the difference between the true value and the predicted value for the instance.

mae=∑ni=1abs(yi−λ(xi))n

CLASSIFICATION MODEL: DECISION TREE CLASSIFIER

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.

building the Decision Tree Classifier model

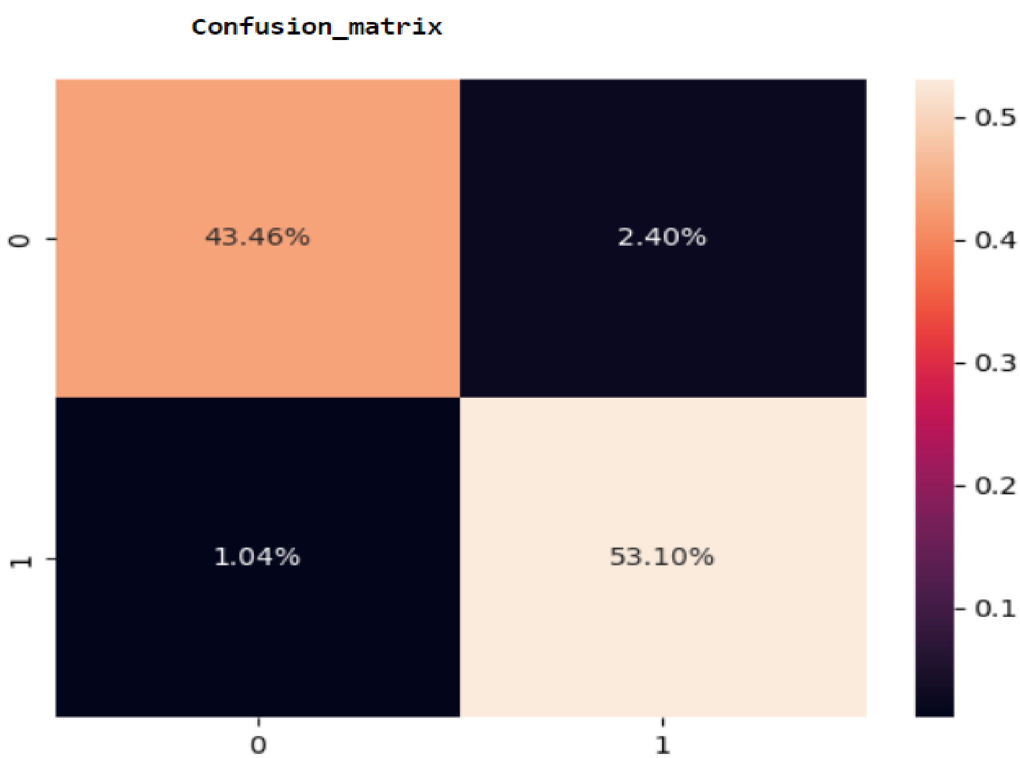
```
[44] # Decision Tree model
     from sklearn.tree import DecisionTreeClassifier

     # instantiate the model
     tree = DecisionTreeClassifier(max_depth = 5)
     # fit the model
     tree.fit(x_train, y_train)

     DecisionTreeClassifier(max_depth=5)

[45] #prediction on test data
     pred2=tree.predict(x_test)
     pred2

     array([1, 1, 1, ..., 1, 1, 1])
```

EVALUATION METRICS:

Some of the evaluation metrics is as follows

- Confusion matrix

Confusion Matrix is a performance measurement for machine learning classification.Accuracy score



- Classification report

Precision: It is calculated with respect to the predicted values.
Recall: It is calculated with respect to the actual values in dataset.
F1-score: It is the harmonic mean of precision and recall.
Support: It is the total entries of each class in the actual dataset.

evaluation metrics

```
[63] from sklearn import metrics
```

```
[47] metrics.confusion_matrix(y_test,pred2)

     array([[  61,  249],
            [  26, 1875]])
```

```
[53] print('DT model Accuracy Score:',metrics.accuracy_score(y_test,pred2))

     DT model Accuracy Score: 0.8756218905472637
```

```
[54] acc=metrics.accuracy_score(y_test,pred2)
     acc

     0.8756218905472637
```

```
[55] #error
     1-acc

     0.12437810945273631
```

```
[65] from sklearn.metrics import classification_report
```

```
report = classification_report(y_test,pred2)
print("Classification report:")
print(report)

Classification report:
              precision    recall  f1-score   support

          -1       0.70      0.20      0.31       310
           1       0.88      0.99      0.93      1901

    accuracy                           0.88      2211
   macro avg       0.79      0.59      0.62      2211
weighted avg       0.86      0.88      0.84      2211
```

2.TUNE THE MODEL: DECISION TREE CLASSIFIER

HYPERPARAMETER TUNING:

tuning the model

hyperparameter tuning

```
[80] from sklearn.tree import DecisionTreeClassifier
```

```
[81] tree = DecisionTreeClassifier(max_depth = 5,random_state=42)
     tree.fit(x_train, y_train)
     tree.score(x_train, y_train)

     0.885119855269109
```

```
[88] tree = DecisionTreeClassifier(max_depth = 5,random_state=42)
     tree.fit(x_train, y_train)
     print('The Training Accuracy for max_depth 5 is:',format(5),tree.score(x_train, y_train))
     print('The Validation Accuracy for max_depth 5 is:',format(5),tree.score(x_train, y_train))

     The Training Accuracy for max_depth 5 is: 5 0.885119855269109
     The Validation Accuracy for max_depth 5 is: 5 0.885119855269109
```

10.ADVANTAGES &DISADVANTAGES

ADVANTAGES:

i. Will be able to differentiate between phishing(0) and legitimate(1) URLs .

ii. It Will help reduce phishing data breaches for an organization

iii. It Will be helpful to individuals and organizations iv. It is easy to use.

SAFETY: No data loss occurs in this system.

QUALITY: The project is developed with the help of Anaconda Navigator software which meets the requirement of the user, the project is checked whether the phases individually have a served its purpose.

DISADVANTAGES:

- Need of internet to search
- Need feed continously
- only applicable for detecting URLs.

11.CONCLUSION

Phishing has becoming a serious network security problem, causing financial loss of billions of dollars to both consumers and e-commerce companies. Phishing attacks can be detected through a combination of customer reportage, bounce monitoring, image use monitoring, honey pots and other techniques. Email authentication technologies such as Sender-ID and cryptographic signing, when widely deployed, have the potential to prevent phishing emails from reaching users. Personally identifiable information should be included in all email communications. Systems allowing the user to enter or select customized text and imagery are particularly promising. Anti-phishing toolbars are promising tools for identifying phishing sites and heightening security when a potential phishing site is detected. By IPDCM it includes the detection of phishing websites through ensemble classifiers and categorizing the phishing websites according to the various streams as online payments, Banking etc.

12.FUTURE SCOPE

In future if we get structured dataset of phishing we can perform phishing detection much more faster than any other technique.In future we can use a combination of any other two or more classifier to get maximum accuracy. We also plan to explore various phishing techniques that uses Lexical features, Network based features,Content based features, Webpage based features and HTML and JavaScript features of web pages which can improve the performance of the system. In particular, we extract features from URLs and pass it through the various classifiers.

13.APPENDIX

SOURCE CODE

phishing_notebook

In [1]:

```
#importing libs
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import confusion_matrix,accuracy_score
```

In [2]:

```
#import dataset
ds=pd.read_csv("dataset_website.csv")
ds.head()
```

Out[2]:

ut[2]:

	index	having_IPhaving_IP_Address	URLURL_Length	Shortining_Service	having_At_Symbol	double_slash_redirecting	Prefix_Suffix	having_Sub_Domain	SSI
0	1	-1	1	1	1	-1	-1	-1	
1	2	1	1	1	1	1	-1	0	
2	3	1	0	1	1	1	-1	-1	
3	4	1	0	1	1	1	-1	-1	
4	5	1	0	-1	1	1	-1	1	

5 rows × 32 columns

In [3]:

```
#null values
ds.info()
ds.isnull().any()#no null values
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 11055 entries, 0 to 11054
Data columns (total 32 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   index                        11055 non-null  int64
 1   having_IPhaving_IP_Address   11055 non-null  int64
 2   URLURL_Length                11055 non-null  int64
 3   Shortining_Service           11055 non-null  int64
 4   having_At_Symbol             11055 non-null  int64
 5   double_slash_redirecting     11055 non-null  int64
 6   Prefix_Suffix                11055 non-null  int64
 7   having_Sub_Domain            11055 non-null  int64
 8   SSLfinal_State               11055 non-null  int64
 9   Domain_registeration_length  11055 non-null  int64
 10  Favicon                      11055 non-null  int64
 11  port                         11055 non-null  int64
 12  HTTPS_token                  11055 non-null  int64
 13  Request_URL                  11055 non-null  int64
 14  URL_of_Anchor                11055 non-null  int64
 15  Links_in_tags                11055 non-null  int64
 16  SFH                          11055 non-null  int64
 17  Submitting_to_email          11055 non-null  int64
 18  Abnormal_URL                 11055 non-null  int64
 19  Redirect                     11055 non-null  int64
 20  on_mouseover                 11055 non-null  int64
 21  RightClick                   11055 non-null  int64
 22  popUpWidnow                  11055 non-null  int64
 23  Iframe                       11055 non-null  int64
 24  age_of_domain                11055 non-null  int64
 25  DNSRecord                    11055 non-null  int64
 26  web_traffic                  11055 non-null  int64
 27  Page_Rank                    11055 non-null  int64
 28  Google_Index                 11055 non-null  int64
 29  Links_pointing_to_page       11055 non-null  int64
 30  Statistical_report           11055 non-null  int64
 31  Result                       11055 non-null  int64
dtypes: int64(32)
```

memory usage: 2.7 MB

Out[3]:

```
index                    False
having_IPhaving_IP_Address    False
URLURL_Length                False
Shortining_Service           False
having_At_Symbol             False
double_slash_redirecting     False
Prefix_Suffix                False
having_Sub_Domain            False
SSLfinal_State               False
Domain_registeration_length  False
Favicon                      False
port                         False
HTTPS_token                  False
Request_URL                  False
URL_of_Anchor                False
Links_in_tags                False
SFH                          False
Submitting_to_email          False
Abnormal_URL                 False
Redirect                     False
on_mouseover                 False
RightClick                   False
popUpWidnow                  False
Iframe                       False
age_of_domain                False
DNSRecord                    False
web_traffic                  False
Page_Rank                    False
Google_Index                 False
Links_pointing_to_page       False
Statistical_report           False
Result                       False
dtype: bool
```

In [4]:

```
#split data independent and dependent
```

```
#remove index coln in independent dataset
x=ds.iloc[:,1:31].values
y=ds.iloc[:,-1].values
print(x,y)
[[-1  1  1 ...  1  1 -1]
 [ 1  1  1 ...  1  1  1]
 [ 1  0  1 ...  1  0 -1]
 ...
 [ 1 -1  1 ...  1  0  1]
 [-1 -1  1 ...  1  1  1]
 [-1 -1  1 ... -1  1 -1]] [-1 -1 -1 ... -1 -1 -1]
```

In [5]:

```
#splitting data into train and test
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0)
```

In [6]:

```
x
```

Out[6]:

```
array([[-1,  1,  1, ...,  1,  1, -1],
       [ 1,  1,  1, ...,  1,  1,  1],
       [ 1,  0,  1, ...,  1,  0, -1],
       ...,
       [ 1, -1,  1, ...,  1,  0,  1],
       [-1, -1,  1, ...,  1,  1,  1],
       [-1, -1,  1, ..., -1,  1, -1]], dtype=int64)
```

In [7]:

```
y
```

Out[7]:

```
array([-1, -1, -1, ..., -1, -1, -1], dtype=int64)
```

In [8]:

```
# Creating a Decision Tree model
from sklearn.tree import DecisionTreeClassifier
DecisionT=DecisionTreeClassifier()
DecisionT.fit(x_train,y_train)
```

Out[8]:

```
DecisionTreeClassifier()
```

In [9]:

```
y_pred5=DecisionT.predict(x_test)
```

from sklearn.metrics import accuracy_score
dec_tree=accuracy_score(y_test,y_pred5)
print(dec_tree)
0.9647218453188603

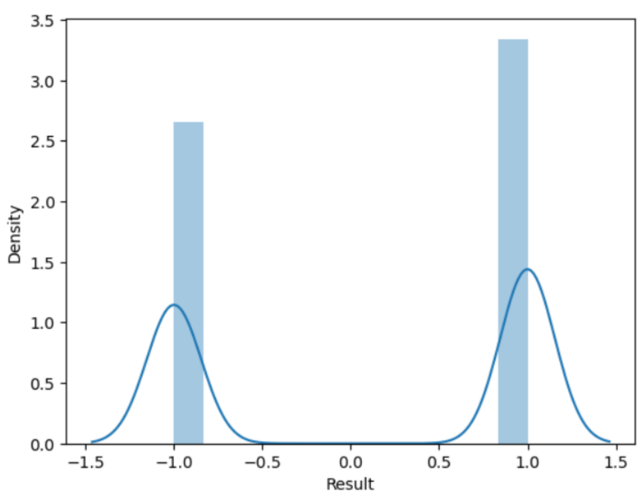
In [10]:

import pickle
pickle.dump(,open('PhisingWebsite.pkl','wb'))

```
In [6]: sns.distplot(data['Result'])
```

```
C:\Users\jm\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated function and w
ill be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexib
ility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

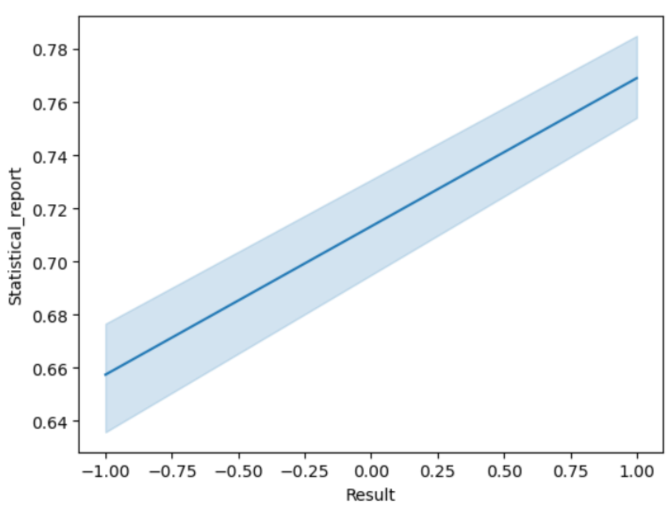
```
Out[6]: <AxesSubplot:xlabel='Result', ylabel='Density'>
```




```
In [7]: sns.lineplot(data['Result'], data['Statistical_report'])
```

```
C:\Users\jm\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as keyword args:
x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit key
word will result in an error or misinterpretation.
  warnings.warn(
```

```
Out[7]: <AxesSubplot:xlabel='Result', ylabel='Statistical_report'>
```



ibm_app.py

```python
import flask
from flask import request, render_template
from flask_cors import CORS
import requests

# NOTE: you must manually set API_KEY below using information retrieved from your IBM Cloud account.
API_KEY = "2ev1GR8SAtWLwWssYOE18Lsh2PnIrqX2baPc6kSY84cf"
token_response = requests.post('https://iam.cloud.ibm.com/identity/token', data={"apikey":
 API_KEY, "grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})
mltoken = token_response.json()["access_token"]

header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' + mltoken}

app=Flask(__name__)

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

@app.route('/')
@app.route('/About.html')

def About():
    return render_template("About.html")


```

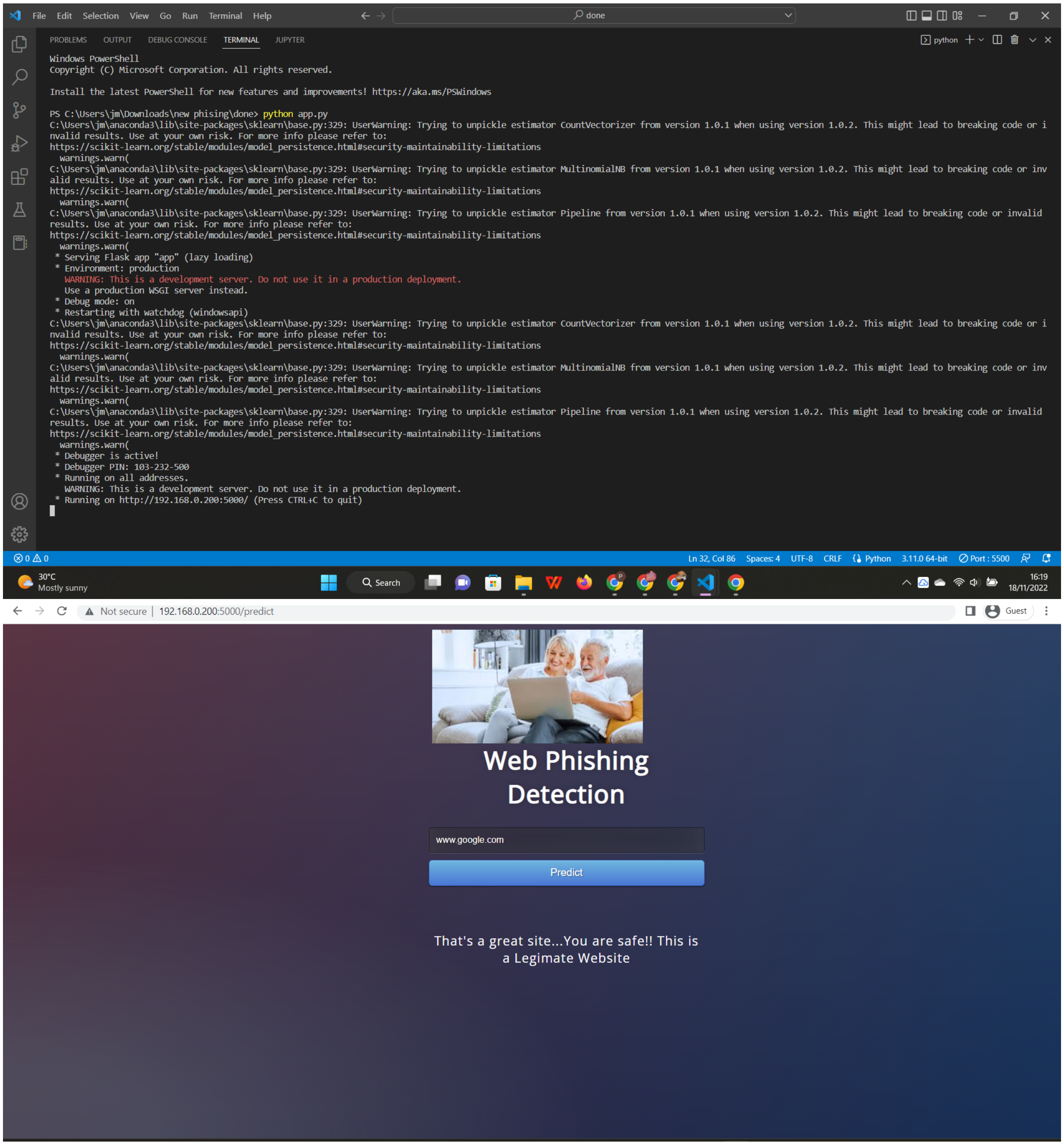
```
    # NOTE: manually define and pass the array(s) of values to be scored in the next line
payload_scoring = {"input_data": [{"fields": [array_of_input_fields], "values": [array_of_values_to_be_scored, another_array_of_values_to_be_scored]}]}
response_scoring = requests.post('https://us-south.ml.cloud.ibm.com/ml/v4/deployments/2de01c97-1fd7-44b5-aec3-f15bb3d28d2e/predictions?version=2022-11-10', json=payload_scoring,
 headers={'Authorization': 'Bearer ' + mltoken})
print("Scoring response")
print(response_scoring.json())
  # showing the prediction results in a UI# showing the prediction results in a UI
pred=print(predictions['predictions'][0]['values'][0][0])
if(pred != 1):
    print("The Website is secure. you are safe....")
else:
    print("The Website is not Legitimate !!BEWARE!!")
if __name__ == "__main__":
    app.run(debug=True,port=5500)
```

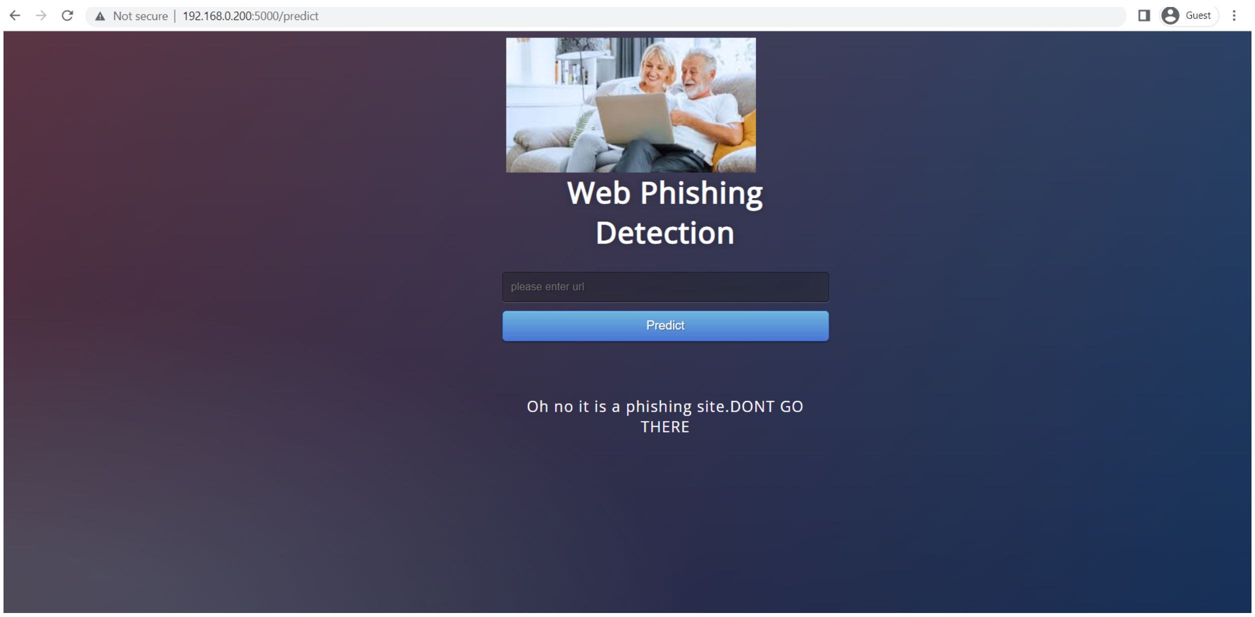
app.py

```
import pickle
import warnings
import numpy as np
import pandas as pd
from flask import Flask, render_template, request
from sklearn import metrics
warnings.filterwarnings('ignore')
from feature import FeatureExtraction
app = Flask(__name__)
phishing = pickle.load(open('Phishing_Website.pkl','rb'))
@app.route('/')
```

```python
@app.route('/web.html')
def Home():
      return render_template("web.html")



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

       url = request.form["url"]
       obj = FeatureExtraction(url)
       x = np.array(obj.getFeaturesList()).reshape(1,30)

       y_pred =phishing.predict(x)[0]
       #1 is safe
       #-1 is unsafe
       y_pro_phishing = phishing.predict_proba(x)[0,0]
       y_pro_non_phishing = phishing.predict_proba(x)[0,1]
       # if(y_pred ==1 ):
       # pred = "It is {0:.2f} % safe to go ".format(y_pro_phishing*100)
       return render_template('predict.html',xx =["It is {0:.2f} % Safe to go ".format(y_pro_non_phishing*100), "It is {0:.2f} % Unsafe to go ".format(y_pro_phishing*100)],url=url)
       # else:
       #     return render_template("predict.html", xx ="Your are on the wrong site. Be cautious!")


if __name__ == "__main__":
   app.run(debug=True,port=5000)
```

EXECUTION & OUTPUT:


File Edit Selection View Go Run Terminal Help
done
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL JUPYTER
python
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.
Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows
PS C:\Users\jm\Downloads\new phising\done> python app.py
C:\Users\jm\anaconda3\lib\site-packages\sklearn\base.py:329: UserWarning: Trying to unpickle estimator CountVectorizer from version 1.0.1 when using version 1.0.2. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/modules/model_persistence.html#security-maintainability-limitations
warnings.warn(
C:\Users\jm\anaconda3\lib\site-packages\sklearn\base.py:329: UserWarning: Trying to unpickle estimator MultinomialNB from version 1.0.1 when using version 1.0.2. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/modules/model_persistence.html#security-maintainability-limitations
warnings.warn(
C:\Users\jm\anaconda3\lib\site-packages\sklearn\base.py:329: UserWarning: Trying to unpickle estimator Pipeline from version 1.0.1 when using version 1.0.2. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/modules/model_persistence.html#security-maintainability-limitations
warnings.warn(
* Serving Flask app "app" (lazy loading)
* Environment: production
WARNING: This is a development server. Do not use it in a production deployment.
Use a production WSGI server instead.
* Debug mode: on
* Restarting with watchdog (windowsapi)
C:\Users\jm\anaconda3\lib\site-packages\sklearn\base.py:329: UserWarning: Trying to unpickle estimator CountVectorizer from version 1.0.1 when using version 1.0.2. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/modules/model_persistence.html#security-maintainability-limitations
warnings.warn(
C:\Users\jm\anaconda3\lib\site-packages\sklearn\base.py:329: UserWarning: Trying to unpickle estimator MultinomialNB from version 1.0.1 when using version 1.0.2. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/modules/model_persistence.html#security-maintainability-limitations
warnings.warn(
C:\Users\jm\anaconda3\lib\site-packages\sklearn\base.py:329: UserWarning: Trying to unpickle estimator Pipeline from version 1.0.1 when using version 1.0.2. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/modules/model_persistence.html#security-maintainability-limitations
warnings.warn(
* Debugger is active!
* Debugger PIN: 103-232-500
* Running on all addresses.
WARNING: This is a development server. Do not use it in a production deployment.
* Running on http://192.168.0.200:5000/ (Press CTRL+C to quit)
Ln 32, Col 86 Spaces: 4 UTF-8 CRLF Python 3.11.0 64-bit Port : 5500
30°C
Mostly sunny
Search
16:19
18/11/2022
Not secure | 192.168.0.200:5000/predict
Guest
Web Phishing Detection
www.google.com
Predict
That's a great site...You are safe!! This is a Legimate Website



GITHUB&PROJECT DEMO

LINK- IBM-12430-1659451175

https://github.com/IBM-EPBL/IBM-Project-12430-1659451175

PROJECT DEMO LINK

DRIVE-demo link

https://drive.google.com/drive/folders/1YVZtgdDfvjzjTzEBsNo-sea6tqWMExIl