**WEB PHISHING DETECTION**

**IBM PROJECT REPORT**

**Submitted by**

**Keerthana Rajeswari R - 2019105022**

**Saheer Anas - 2019105047**

**Shruthilaya A - 2019105052**

**Vaishnavi S – 2019105604**

Logo

Description automatically generated

**INTRODUCTION**

1.1 PROJECT OBJECTIVES:

BY THE END OF THE PROJECT:

• We'll be able to understand the problem to classify if it is a regression or a classification kind of problem.

• We will be able to know how to pre-process/clean the data using different data preprocessing techniques.

• Applying different algorithms according to the dataset

• We will be able to know how to find the accuracy of the model.

• We will be able to build web applications using the Flask framework.

1.2 PURPOSE:

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 :

1. Web phishing aims to steal private information, such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity.

2. It will lead to information disclosure and property damage.

3. Large organizations may get trapped in different kinds of scams.

4. 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.

**LITERATURE SURVEY**

2.1 EXISTING PROBLEM

Phishing is the most popular attack vector for criminals and has grown 65% in the last year. Malicious links will lead to a website that often steals login credentials or financial information like credit card numbers. Attachments from phishing emails can contain malware that once opened can leave the door open to the attacker to perform malicious behavior from the user’s computer.

Due to their low bar of skill required to launch, phishing is a popular choice for cyber criminals. Many of them use phishing kits, which include all the technical materials needed to launch a phishing campaign.

More advanced phishing methods like spoofing (pretending to send emails from a legitimate source), spear phishing (personalizing emails to target specific people), and whaling (targeting high-level executives) remain popular and are even harder to detect by eye alone.

2.2 References

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Reference Paper | Algorithm | Advantage | Disadvantage |
| 1. | M. Chandrasekaran, et al., ”Phishing email detection based on structural properties”, in New York State Cyber Security Conference (NYS) , Albany, NY ,” 2006 | Support Machine  Vectors(SVM) Classifiers | With every data set an accuracy of 96% is reached and accuracy of 99% is with SVM. | Small data set-only 200 emails, therefore very time consuming. |
| 2. | P. R. a. D. L. Ganger, ”Gone phishing: Evaluating anti-phishing tools for windows. Technical report, ,” September 2006 | Bayesian statistics  100,000 email algorithms | Uses the feedback data from the Users of Microsoft | Uses fixed number of features and low recall measurement. |
| 3. | . M. Bazarganigilani, ”Phishing E-Mail Detection Using Ontology Concept and Nave Bayes Algorithm,” International Journal of Research and Reviews in Computer Science, vol. 2,no.2, 2011 | Semantic ontology  Concept by method  Information Gain and naïve Bayes algorithm classifies | Working of model is very simple, in five steps to be exact. | Accuracy is low in comparison with other methods. |
| 4. | I. Fette, et al., ”Learning to detect phishing emails,” in Proc. 16th International World Wide Web Conference (WWW 2007), ACM Press, New York, NY, USA, May 2007, pp. 649-656 | Random forest and support vector machines (SVM) as a classifier | Has a lot of different features included WHOIS query | Sizeable number of phishing and ham emails was not well classified leading to decrease in accuracy. |
| 5. | M. Chandrasekaran, et al., ”Phoney: Mimicking user response to detect phishing attacks,” in In: Symposium on World of Wireless, Mobile and Multimedia Networks, IEEE Computer Society, 2006, pp. 668-672 | PHONEY-Mimics user response | Is implemented between a user’s MTA and MUA | Collected data is so small and is time consuming. |
| 6. | A. Bergholz, et al., ”Improved phishing detection using model-based features,” in Proc. Conference on Email and Anti-Spam (CEAS). Mountain View Conf, CA, aug 2008 | Dynamic Markov chain and a class Topic Models | Has a large number of features which increases accuracy. | Due to the large number of features , there is a high memory requirement. |
| 7. | L. Ma, et al.,”Detecting phishing emails using hybrid features,”IEEE Conf, 2009, pp. 493-497 | Information gain algorithms, Decision tree algorithm, C4.5 | Has seven hybrid features and appears in five stages. | Used a non-standard data set. |

2.3 Problem Statement definition

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 use. The blacklist-based method is unable to keep up with the current phishing attacks as registering new domains has become easier. Moreover, comprehensive blacklist can ensure a perfect up-to-date database. Various other techniques such as page content inspection 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 the accuracy while implementing different error detection methods.

Problem statements:

1.How to process dataset for phishing detection?

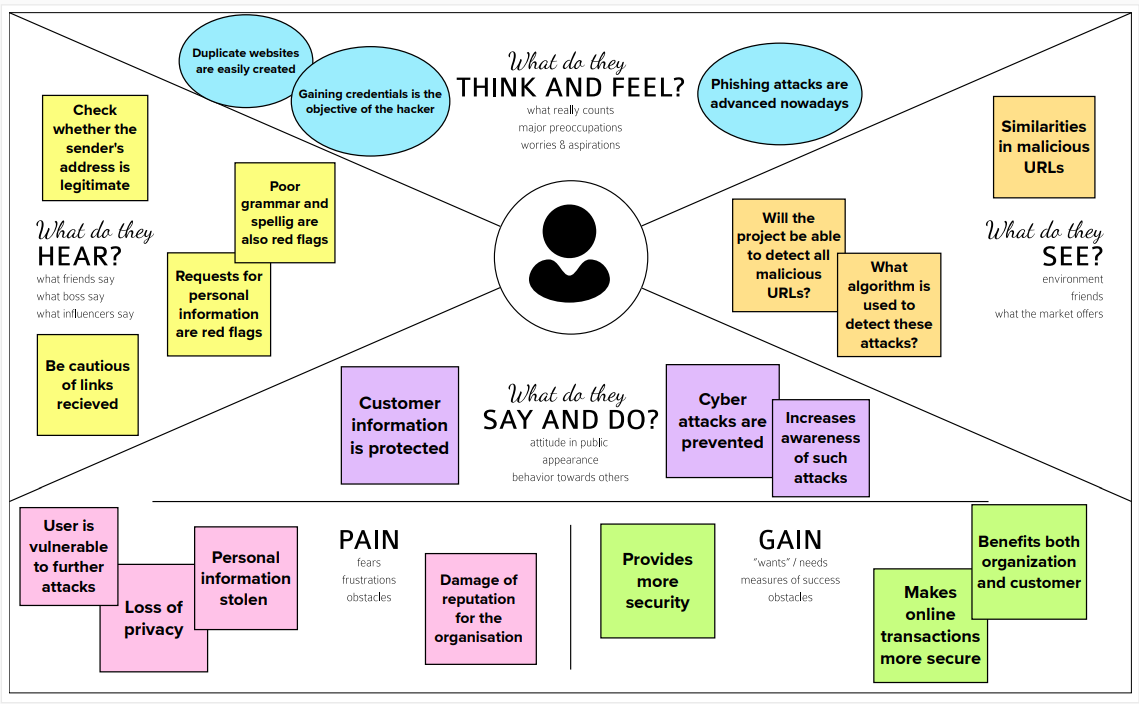
2.How to increase accuracy of [phishing websites](https://www.sciencedirect.com/topics/computer-science/phishing-website) algorithms?

3.How to reduce false negative rate in phishing websites algorithm?

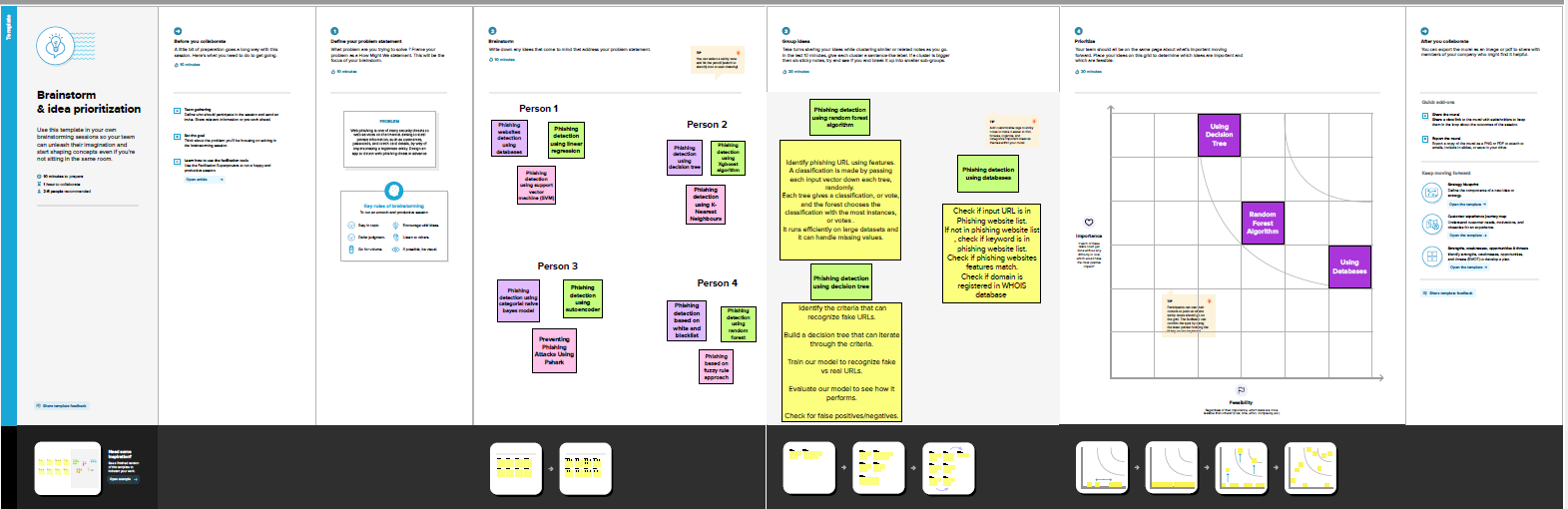
4.What are the best combinations of [classifiers](https://www.sciencedirect.com/topics/computer-science/classification-machine-learning) that can efficiently detect phishing attacks?

**3. IDEATION & PROPOSED SOLUTION**

3.1 EMPATHY MAP CANVAS

****

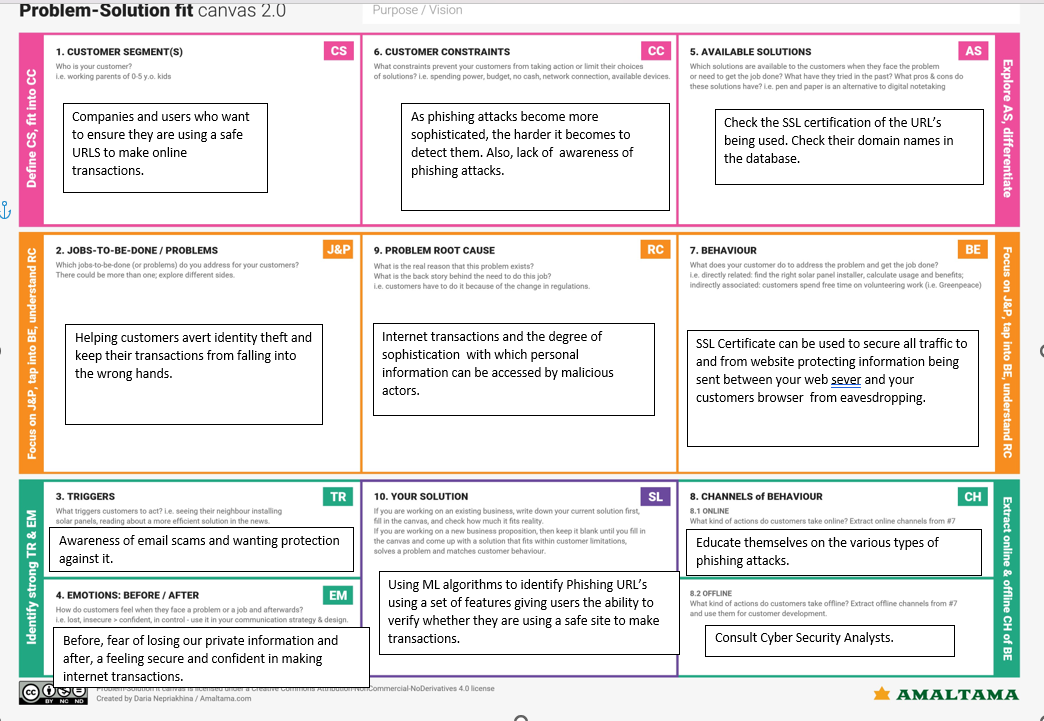
3.2 IDEATION AND BRAINSTROMING



3.3 PROPOSED SOLUTION

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Parameter** | **Description** |
|  | Problem Statement (Problem to be solved) | To detect the phishing activity in websites |
|  | Idea / Solution description | Identify phishing sites using features and machine learning algorithm. |
|  | Novelty / Uniqueness | Features chosen to help detect the phishing URLs for instance the number of dots used, phishing sites tend to use more than five dots. |
|  | Social Impact / Customer Satisfaction | * User security * Reduction in cyber crime * Social Awareness of malicious websites |
|  | Business Model (Revenue Model) | Sell it to companies to train their employees to prevent data theft. |
|  | Scalability of the Solution | Training users/employees of organisations to identify fraudulent URLs trying to phish confidential organizational data. |

3.4 PROPOSED SOLUTION FIT



**4.REQUIREMENT ANALYSIS**

4.4 Functional Requirements:

Following are the functional requirements of the proposed solution.

|  |  |  |
| --- | --- | --- |
| **FR No.** | **Functional Requirement (Epic)** | **Sub Requirement (Story / Sub-Task)** |
| FR-1 | User Input | User types in URL to be checked. |
| FR-2 | Feature Extraction | The model must extract appropriate features from the URL given by the user. |
| FR-3 | Prediction | Model must use an ML algorithm such as KNN, Logistic regression etc to make predictions. |
| FR-4 | Classification | This data is then provided to the classifier to make the final decision. |
| FR-5 | Displaying the result | Whether the given link is legitimate or not is displayed. |
| FR-6 | Clearing Queries | FAQ’s must be provided. |

4.5 NON-FUNCTIONAL REQUIREMENTS:

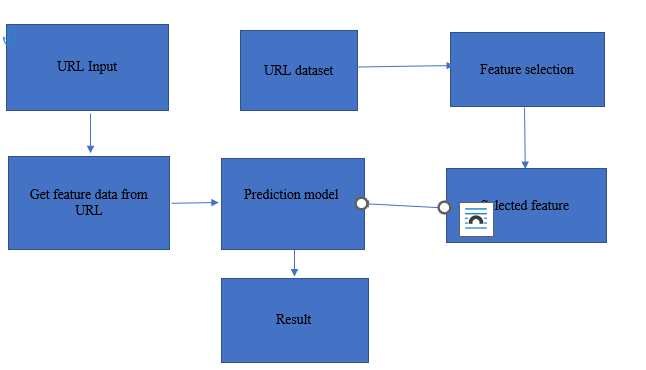
Following are the non-functional requirements of the proposed solution.

|  |  |  |
| --- | --- | --- |
| **NFR No.** | **Non-Functional Requirement** | **Description** |
| NFR-1 | Usability | It pertains to effectiveness ,efficiency and overall satisfaction of the user. |
| NFR-2 | Security | Users need to be protected from malicious attacks when using the site. |
| NFR-3 | Reliability | The model should be able to accurately predict whether a URL is malicious or not/ |
| NFR-4 | Performance | It is the ability of the application to always run acceptably. In time-critical scenarios, even the smallest delay in processing data can be unacceptable. |
| NFR-5 | Availability | The system is accessible to a  user at any given point in time. |
| NFR-6 | Scalability | The system should be able to perform efficiently to meet changing requirement. |

**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.

****

5.2 SOLUTION AND TECHNICAL ARCHITECTURE

TECHINCAL ARCHITECTURE:

Diagram, schematic

Description automatically generated

5.3 USER STORIES

Table

Description automatically generated

**6.PROJECT PLANNING AND SCHEDULING**

6.1 SPRINT PLANNING AND ESTIMATION

| **Sprint** | **Functional Requirement (Epic)** | **User Story Number** | **User Story / Task** | **Story Points** | **Priority** | **Team Members** |
| --- | --- | --- | --- | --- | --- | --- |
| Sprint-3 | User input | USN-1 | User can input the URL. | 2 | High | Saheer Anas |
| Sprint-3 |  | USN-2 | As a user i can input the URL in the required field and waiting for validation. | 2 | High | Saheer Anas |
| Sprint-1 | Feature extraction | USN-3 | The model must extract appropriate features from the URL given by the user. | 2 | High | Shruthilaya A |
| Sprint-2 | Prediction | USN-4 | Model must use an ML algorithm such as KNN, Logistic regression etc to make predictions. | 2 | High | Keerthana Rajeswari |
| Sprint-1 | Classifier | USN-5 | This data is then provided to the classifier to make the final decision | 2 | High | Vaishnavi |

* 1. SPRINT DELIVERY SCHEDULE

| **Sprint** | **Total Story Points** | **Duration** | **Sprint Start Date** | **Sprint End Date (Planned)** | **Story Points Completed (as on Planned End Date)** | **Sprint Release Date (Actual)** |
| --- | --- | --- | --- | --- | --- | --- |
| Sprint-1 | 20 | 6 Days | 29 Oct 2022 | 4 Nov 2022 | 20 | 4 Nov 2022 |
| Sprint-2 | 20 | 6 Days | 4 Nov 2022 | 10 Nov 2022 | 20 | 10 Nov 2022 |
| Sprint-3 | 20 | 6 Days | 11 Nov 2022 | 17 Nov 2022 | 20 | 17 Nov 2022 |
| Sprint-4 | 20 | 6 Days | 18 Nov 2022 | 24 Nov 2022 | 20 | 24 Nov 2022 |

6.3 REPORTS FROM JIRA

A picture containing application

Description automatically generated

**7.CODING AND SOLUTIONING**

7.1 FEATURE CODE 1

Feature sets are divided into four main categories:

**Address Bar-Based Features** – these are features extracted from the URL itself like,

* URL length : Function returns 1 if URL length is less than 54, 0 if length is between 54 and 75 , else returns -1 which indicates phishing.
* Whether it contains an IP address: Functions returns -1 if IP address is detected otherwise 1 is returned.
* Uses an URL shortening service like Tiny URL or Bitly:- A number of shortening services are stored in a separate variable, and re.search is used to see if any of there are in the URL given and if so, -1 is returned indicating it is a phishing URL.
* Employs redirection:- Checks for //, if it is in a position greater than seven in the URL, -1 is returned indicating it is phishing.

**Abnormal Features**

* + Loading images loaded in the body from a different URL :- Favicon module is used and if domain of image and URL match it returns 1 else -1 is returned indicating phishing.
  + Minimal use of meta tags: If more than one link is associated with page, it returns -1.
  + The use of a Server Form Handler (SFH):- If SFH refers to a domain then it returns -1.
  + Submitting information to email:-Using mail() or mailto() returns -1.
  + An abnormal URL :-If host name is not in URL, then it is classified as an abnormal URL.

7.2 Feature Code 2

**HTML and JavaScript-Based Features**– these can include things like:

* + Disabling the ability to right-clicks:- If right click is disabled , function returns -1 else it returns 1.
  + On mouseover:-If status bar changes , function returns -1.
  + Using pop-up windows:-If popup window contains textfield , then it is considered phishing and returns -1.
  + iFrame redirection:-If iframe is used it returns -1, else it returns 1.

**Domain-Based Features** – these can include:

* + Unusually young domains:-If age of domain is greater then 6 months it returns 1 and it considered legitimate.
  + Suspicious DNS record:-If there is no DNS record, then it is considered a phishing URL.
  + Low volume of website traffic:-If website rank it lesser than 100, it returns 1.
  + PageRank:- Most phishing websites do not have a page rank, the function checks for prank and if not, available it is considered a phishing URL.
  + Whether the site has been indexed by Google:- Most phishing websites are not indexed on Google, if webpage is indexed , function returns 1 and is considered legitimate.

**8.TESTING**

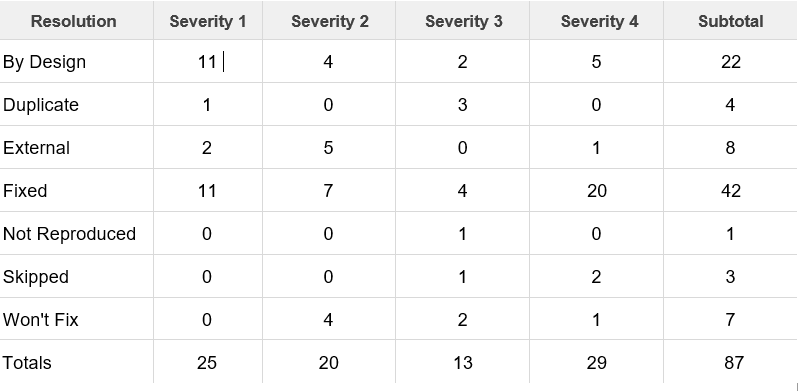
**8.1 Testcases Report**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | Date | 19-Nov-22 |  | |  |  |  |  |  |  |
| Team ID | PNT2022TMID35404 |
| Project Name | Project - Web Phishing Detection |
| Maximum Marks | 4 marks |
| Test case ID | Feature Type | Componen  t | Test Scenario | Pre-Requisite | Steps To Execute | Test Data | Expected Result | Actual  Result | Status | Comments | TC for  Automation(Y/N) | BUG ID | Executed By |
| 1 | Functional | Home Page | Verify user is able to see the  Landing Page 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://www.google.com | Should Display the Webpage | Working as expected | Pass |  | N |  | Vaishanavi.s |
| 2 | UI | Home Page | Verify the UI elements is Responsive |  | 1.Enter URL and click go   1. Type or copy paste the URL 2. Check whether the button is   responsive or not   1. Reload and Test Simultaneously | https://www.google.com | Should Wait for Response and then gets Acknowledge | Working as expected | Pass |  | N |  | Keerthana Rajeswari.R |
| 3 | 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://www.google.com | User should observe whether the website is legitimate or not. | Working as expected | Pass |  | N |  | Saheer Anas |
| 4 | Functional | Home Page | Verify user 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://www.google.com |  | Working as expected | Pass |  | N |  | Shruthilaya. A |
| 5 | Functional | Home Page | Testing the website with multiple URLs |  | 1.Enter URL ( https://phishingshield.herokuapp.com/) and click go   1. Type or copy paste the URL to   test   1. Check the website is legitimate   or not   1. Continue if the website is   secure or be cautious if it is not  secure | https://127.0.0.1 | User can identify whether the websites is secure or not | Working as expected | Pass |  | N |  | Shruthilaya. A |

8.3 User Acceptance Testing

**Defect Analysis**

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



**Test Case Analysis**

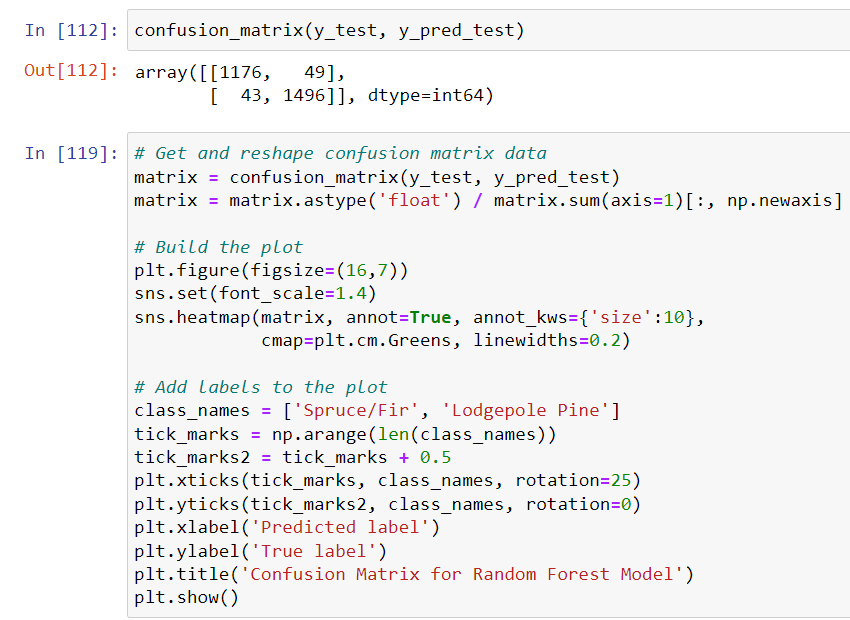
This report shows the number of test cases that have passed, failed, and untested

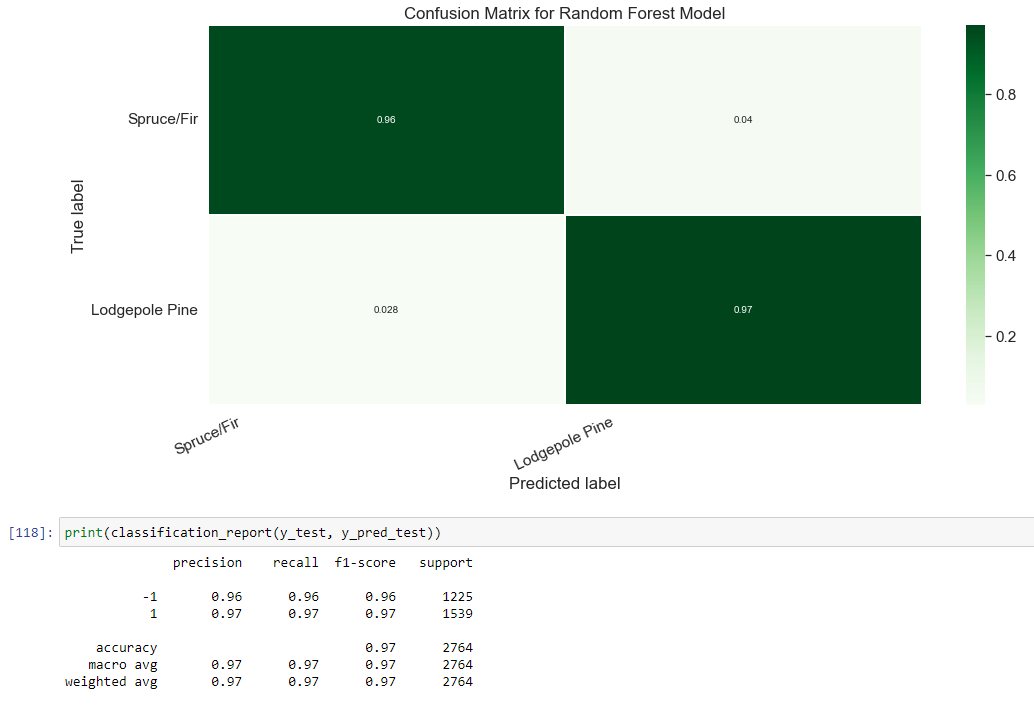
Table

Description automatically generated

**9.RESULTS**







**RANDOM FOREST MODEL**

**ACCURACY:0.966**

**CONFUSION MATRIX:** ([1176,49],[43,1496])

**10.ADVANTAGES AND DISADVANTAGES**

ADVANTAGES OF WEB PHISHING DETECTION:

* The algorithm provides a clear idea about the effective level of each classifier on phishing email detection
* It has a very high level of accuracy
* It can be evolved with time according to the classification of features
* Majority of the work is online
* It provides a secure connection between agent and user
* It can Mitigate zero-hour attacks.
* Requiring low resources on host machine
* We can construct our own ML classification models

DISADVANTAGES OF WEB PHISHING DETECTION:

* It is a time-consuming process
* The process needs feed continuously
* High computational cost involved in certain cases
* It involves a huge number of rules
* Can result in excessive queries with heavily loaded servers.

**11.CONCLUSION**

The most important way to protect the user from phishing attack is the education awareness. Internet users must be aware of all security tips which are given by experts. Every user should be trained not to blindly follow the links to websites where they must enter their sensitive information. It is essential to check the URL before entering the website. In future, system can upgrade to automatic detection of the web page and the compatibility of the application with the web browser. Additional work also can be done by adding some other characteristics to distinguishing the fake web pages from the legitimate web pages. These can be upgraded in the Darkphish web page.

There are many features that can be improved in the work, for various other issues. The heuristics can be further developed to detect phishing attacks in the presence of embedded objects like flash. Identity extraction is an important operation and it was improved with the Optical Character Recognition (OCR) system to extract the text and images. More effective inferring rules for identifying a given suspicious web page, and strategies for discovering if it is a phishing target, should be designed in order to further improve the overall performance of this system. Moreover, it is an open challenge to develop a robust malware detection method, retaining accuracy for future phishing emails. In addition, the dynamic and static features complement each other, and therefore both are considered important in achieving high accuracy.

**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**

PROJECT DEMO VIDEO:

<https://drive.google.com/file/d/13oaS4-PsSDDBMFkMHbUylAQPwlSsH4YC/view?usp=share_link>

GIT REPOSITORY:

https://github.com/IBM-EPBL/IBM-Project-1823-1658416078