***Project Report on***

*A Comparative Study of Classification Algorithm for Spam*

*Data Analysis*

*Submitted in the partial fulfilment of the requirements for*

*the award of Degree of B. Tech*

***By***

***Muskaan Singh (1806810200)***

***Ritika Verma (1806810262)***

***Aarti Goswami (1806810004)***

*Under the Supervision of: -*

*Dr. Vimal Kumar*

*(Professor, Department of CSE)*



*Department of Computer Science & Engineering*

*Meerut Institute of Engineering and Technology,*

*Meerut – 250 005*

*Dr. A.P.J. Abdul Kalam Technical University, U.P., Lucknow*

*[2018-2022]*

# DECLARATION

*I hereby declare that this submission is my own work and that, to the best of my knowledge and belief,*

*it contains no material previously published or written by another person nor material which to a*

*substantial extent has been accepted for the award of any other degree or diploma of the university or*

*other institute of higher learning, except where due acknowledgment has been made in the text.*

*Signature:*

*Name*: *Muskaan Singh*

*Roll No.: 1806810200*

*Date* *:*

*Signature:*

*Name: Ritika Verma*

*Roll No.: 1806810262*

*Date* *:*

*Signature:*

*Name: Aarti Goswami*

*Roll No. :1806810004*

*Date* *:*

**CERTIFICATE**

*This is to certify that Project Report entitled –a comparative study of classification algorithm for spam*

*data analysis which is submitted by Muskaan Singh (1806810200), Ritika Verma (1806810262),*

*Aarti Goswami (1806810004) in partial fulfilment of the requirement for the award of degree B. Tech*

*. in Department of CSE, Of Dr. A.P.J. Abdul Kalam Technical University, U.P., Lucknow., is a record*

*of the candidate own work carried out by him/her under my/our supervision. The matter embodied*

*in this Project report is original and has not been submitted for the award of any other degree.*

**Date:** **Supervisor**

**Dr. Vimal Kumar**

**(Professor, CSE Department)**

**MIET, Meerut**

## ACKNOWLEDGEMENT

*It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during*

*B. Tech Final Year. We owe special debt of gratitude to our guide Prof. (Dr.) Vimal Kumar,*

*Department of CSE, Meerut Institute of Engineering and Technology, Meerut for his constant support*

*and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have*

*been a constant source of inspiration for us. It is only his cognizant efforts that our endeavours have*

*seen light of the day*.

*We also take the opportunity to acknowledge the contribution of Mrs. Priyanka Bhardwaj, Department*

*of CSE, Meerut Institute of Engineering and Technology, Meerut for his full support and assistance*

*during the development of the project.*

*We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of*

*the department for their kind assistance and cooperation during the development of our project. Last*

*but not the least, we acknowledge our friends for their contribution in the completion of the project.*

###### ABSTRACT

Phishing is described as gaining private information from a user via hacking into an affiliate’s website. To combat phishing, a variety of strategies have been offered. This menace, however, cannot be eliminated by a single miraculous bullet. Data mining is an effective method for detecting phishing assaults. An intelligent approach to identifying phishing attempts is shown in this article. We employ a variety of data mining techniques to classify websites as real or fraudulent. To construct an accurate intelligent phishing analysis system, many categories are employed.

The performance of data mining approaches was assessed using classification accuracy, ROC (area under receiver) operating characteristic (AUC) curves and F-size. The results demonstrate that Random Forest performs the best among the categorization algorithms, with a 97.36 percent accuracy rate. The random forest algorithm is very fast and can handle various phishing analysis sites.

###### LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **Figure No** | **Figure Description** | **Page No** |
| Figure 5.1 | E-R Diagram | 29 |
| Figure 5.2 | Data Flow Diagram | 30 |

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Contents** | **Page No.** |
|  | GTU Completion Certificate |  |
|  | Acknowledgement | 4 |
|  | Abstract | 5 |
|  | List Of Figure | 6 |
|  | Index | 7 |
| **1.** | **Introduction** |  |
|  | 1.1 Objectives | 11 |
|  | 1.2 Scope | 11 |
|  | 1.3 Limitations of existing system | 11 |
| **2.** | **Project Planning** |  |
|  | 2.1 Software Process Model | 13 |
|  | 2.2 Tools and Technology | 15 |
|  | 2.2.1 Hardware Requirement | 15 |
|  | 2.2.2 Software Requirement | 15 |
|  | 2.3 Team Organization | 16 |
|  | 2.4 Time Line Chart | 17 |
| **3** | **Literature Survey** |  |
|  | 3.1 Technology and Literature Review | 19 |
|  | 3.1.1 Technology Review (Machine Learning) | 19 |
|  | 3.2 Technology Review (Python) | 20 |
|  | 3.3 Technology Review (Voice Assistant) | 21 |

|  |  |  |
| --- | --- | --- |
| **4** | **System Analysis** |  |
|  | 4.1 Feasibility Study | 23 |
|  | 4.1.1 Operational Feasibility | 24 |
|  | 4.1.2 Technical Feasibility | 24 |
|  | 4.1.3 Time Schedule Feasibility | 24 |
|  | 4.1.4 Economical Feasibility | 24 |
|  | 4.1.5 Implementation Feasibility | 25 |
|  | 4.2 Requirement Specification | 25 |
|  | 4.2.1 Functional Requirements | 25 |
|  | 4.2.2 Non-Functional Requirements | 25 |
| **5** | **System Design** |  |
|  | 5.1 Design Methodology | 27 |
|  | 5.2 E-R Diagram | 29 |
| **6** | **Screen Layout** |  |
|  | * Home Screen | 30 |
|  | * Entertaining Modules | 37 |
|  | * Media Features | 38 |
|  | * Gaming Features | 39 |
| **7** | **Conclusion** |  |
|  | * Conclusion | 41 |
|  | * Scope for further enhancement | 41 |
| **8** | **Reference** |  |
|  | * Reference | 43 |

**CHAPTER-1 INTRODUCTION**

###### 1.1 INTRODUCTION

The purpose of this document is to present a detailed description of the of the URL Phishing Detection. It will explain the purpose and features of the system, the interfaces of the system, what the system will do, the constraints under which it must operate and how the system will react to external stimuli. This document is intended for both the stakeholders and the developers of the system and will be proposed to the Meerut institute of Engineering and Technology for its approval.

A phishing URL is created to get the personal data of the user, such as usernames and passwords, or to attack or send some malicious data to the user’s system. Ideally, the attacker manipulates the user to click the links and get sensitive information. Phishers can clone the legitimate link data to trick the user into filling in the sensitive information. Phishing links can be used to get a user’s confidential details also. This is a very difficult condition for users. The phisher can misuse it for personal gain. According to a survey, phishing attacks are increasing day by day. Therefore, a lot of effort has been made in this area to minimize these phishing attacks. By viewing the content of a website or web page, or using URL metadata, we can determine if the site is a phishing site or not. In our project, we deal with website URL metadata, whether it is a phishing site or not.

By using metadata in the URL, we no longer need to attempt phishing websites or download any of their content, making it much more secure access. We can look into certain parts of the URL, such as the number of slashes, keywords in part of the URL path, etc. After getting the necessary information, we only need information data about a series of URLs to be classified using some algorithms. In our project, we used the Support Vector Machine (SVM) and Random Forest algorithms.

###### Key Concepts:

* 1. **Objectives:**

Aim of the project is to develop a system that can prevent user from being phished from hacker while surfing websites. And along with prevention also aware user by providing the information about phishing via chatbot.

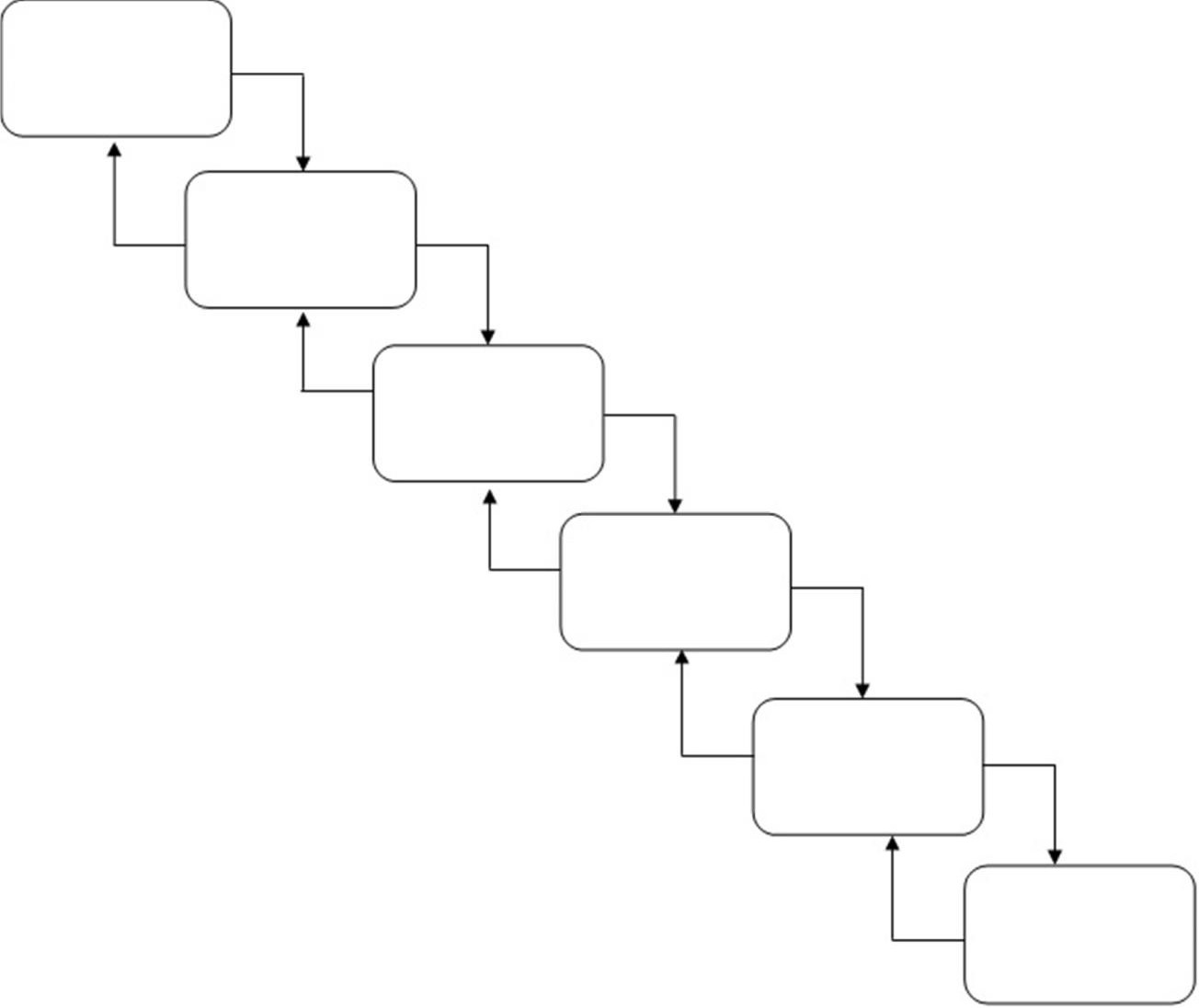
###### Scope:

This is describing the cyber security system which is called phishing website detection which is based on machine learning. It provides security through phishing detection. This system aims to provide a security system which holds previous information and characteristics of websites and to increase the accuracy of the extension. A hacker implements their phishing abilities in such a way that it looks as similar as to the legitimate because of which the user can’t find any difference and fill credentials to it and fooled them easily. So by using different methods we detect phishing URLs, we use extensions so that if the site is legitimate the pop-up message appears.

###### Why is the particular topic chosen?:

Nowadays phishing attacks are widely spread through social media sites, fake applications, SMS and emails. These attacks contain URL phishing, email spoofing, website spoofing, voice phishing, mass target, image phishing, clone phishing. A hacker implements their phishing abilities in such a way that it looks as similar as to the legitimate because of which the user can’t find any difference and fill credentials to it and fooled them easily. So by using different methods we detect phishing URL’s, we use extensions so that if the site is legitimate the pop-up message appears.

### CHAPTER-2 PROJECT PLANNING



###### Key Concepts:

* 1. Software Process Model
  2. Tools and Technology
     1. Hardware Requirement
     2. Software Requirement
  3. Team Organization
  4. Time Line Chart

###### SOFTWARE PROCESS MODEL:

Our system uses Incremental model for software development. Following figure shows the figure of our system’s process model:

Requirement Gathering

System Analysis

System Design

Code Implementation

System Testing

System Maintenance

“The process that deals with technical and management issues of software development is called software process.”

* In a Project model the first phase is requirement gathering. In this phase you gathering all the related requirements of system what you are going to developed.
* Second phase is System analysis. The system analysis is ”What is done, and How it is done” This is the most important and critical state of project or system. In this phase the basic requirements of the software must be understood by an Analyst.
  + All the requirements are then well documented and discussed further with the client or end-user, for reviewing.
  + After the analysis of system you will get the actual or real time problem of the system.
* Third phase is system design. This phase is between the analysis and implementation stages. In design phase you will design the project. It has some attributes such as:
  + Data Structure
  + Software architecture
  + Interface representation
  + Algorithm details
  + The all requirements are translated in some easy to represents from using which coding can be done effecting and efficiently. The design needs to be documented for further use.
* Forth phase is implementation. It is a step in which design is translated into machine readable form. If design is done in sufficient detail then coding can be done effectively. Programs are created in this phase. The coding is done with the help of some programming language. In our system we are using Python as implementation language.
* Fifth phase is testing. It begins when coding is done. While performing testing the major focus is on logical internals of the software. It also ensures execution of all the paths, functional behaviors. If any error in your system occurs, it will detect that error and the fix the error and meet the customer requirements. There are many criteria for testing, like white box testing, black box testing, etc.
* Sixth and last phase is maintenance. This is huge phase of this model. It maintains the system after deploying it on the particular space like any computer system.

###### TOOLS AND TECHNOLOGY:

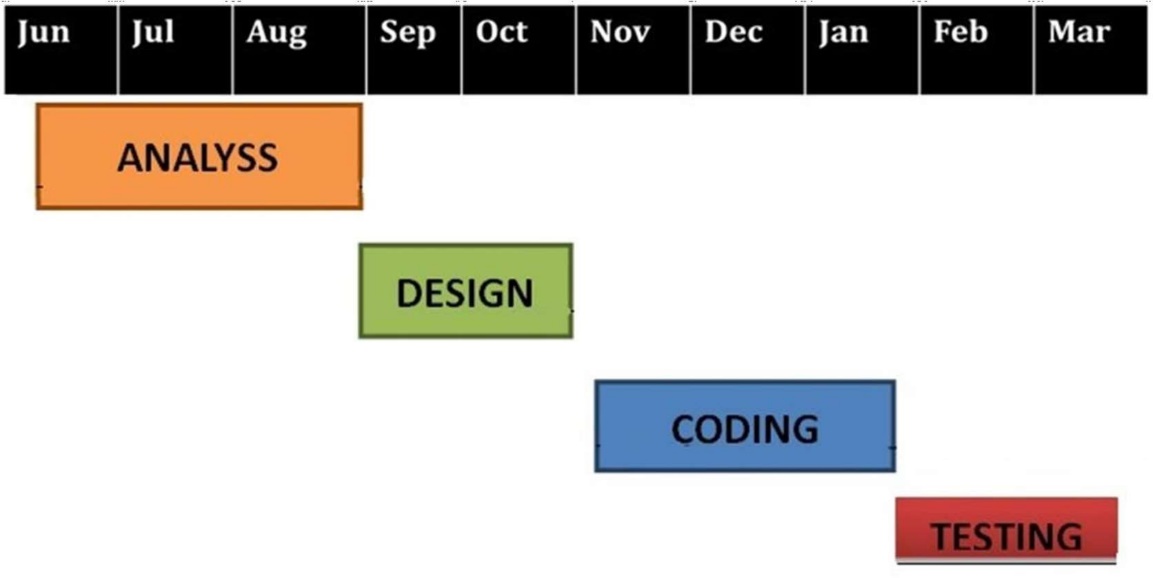
* + 1. **Hardware Requirements:**

All the hardware needed here are generally the basic configuration of a typical computer. A list of the hardware requirement used in the system given below:

* **CPU :** Intel core i3-6006U 2.0GHz
* **RAM SIZE : >**4 GB
* CACHE MEMORY : >1 MB
* **OS :** Windows
* **SENSORS :** Camera, microphone
  + 1. **Software Requirements:**
* Console (CMD )
* IDE(Visual Studio Code, Spyder, Jupyter Notebook, pycharm, IDLE)
* Python
* Tensor flow API
* Dataset(movies, news)

###### TEAM ORGANIZATION:

|  |  |  |  |
| --- | --- | --- | --- |
| **RESPONSIBILITIES** | **ROLES** | **ROLES** | **ROLES** |
| Analysis | Muskaan | Ritika | Aarti |
| Design | Muskaan | Ritika | Aarti |
| Design Review | Muskaan | Ritika | Aarti |
| Coding | Muskaan | Ritika | Aarti |
| Font-site Design | Muskaan | Ritika | Aarti |
| Design Review | Muskaan | Ritika | Aarti |
| Testing | Muskaan | Ritika | Aarti |
| Documentation | Muskaan | Ritika | Aarti |



* 1. **TIME LINE CHART:**

**ANALYSIS**

### CHAPTER-3 LITERATURE SURVEY

###### TECHNOLOGY AND LITERATURE REVIEW:

* + 1. Technology Review (Machine Learning)

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or Machine learning algorithms are often categorized as supervised or unsupervised. Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset,the learning algorithm produces an inferred function to make predictions about the output values.

The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn’t figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it learn from it. Otherwise, acquiring unlabeled data generally doesn’t require additional resources.

###### Technology Review (Python)

Python is an interpreter, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object- oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

Python was conceived in the late 1980s as a successor to the ABC language. Python 2.0, released 2000, introduced features like list comprehensions and a garbage collection system capable of collecting reference cycles. Python 3.0, released 2008, was a major revision of the language that is not completely backward-compatible, and much Python 2 code does not run unmodified on Python 3. Due to concern about the amount of code written for Python 2, support for Python 2.7 (the last release in the 2.x series) was extended to 2020. Language developer Guido van Rossum shouldered sole responsibility for the project until July 2018 but now shares his leadership as a member of a five-person steering council. Python interpreters are available for many operating systems. A global community of programmers develops and maintains Python, an open source reference implementation. A non-profit organization, the Python Software Foundation, manages and directs resources for Python development.

###### Technology Review (Selenium)

Selenium is an open-source tool that automates web browsers. It provides a single interface that lets you write test scripts in programming languages like Ruby, Java, NodeJS, PHP, Perl, Python, and C#, among others.

A browser-driver then executes these scripts on a browser-instance on your device (more on this in a moment).

### Selenium WebDriver

Also known as Selenium 2.0, WebDriver executes test scripts through browser-specific drivers. It consists of:

##### API

Application Programming Interface. Ports test scripts you write in Ruby, Java, Python, or C# to Selenese (Selenium’s own scripting language), through bindings.

##### Library

Houses the API and language-specific bindings. Although plenty of third-party bindings exist to support different programming languages, the core client-side bindings supported by the main project are: Selenium Java (as selenium jar files), Selenium Ruby, Selenium dotnet (or Selenium C#, available as .dll files), Selenium Python, and Selenium JavaScript (Node).

##### Driver

Executable module that opens up a browser instance and runs the test script. Browser-specific—for instance, Google develops and maintains Chromedriver for Selenium to support automation on Chromium/Chrome.

##### Framework

Support libraries for integration with natural or programming language test frameworks, like Selenium with Cucumber or Selenium with TestNG.

### CHAPTER-4 SYSTEM ANALYSIS

###### Key Concepts:

* 1. Feasibility Study
     1. Operational Feasibility
     2. Technical Feasibility
     3. Time Schedule Feasibility
     4. Economical Feasibility
  2. Requirement Specification
     1. Functional Requirements
     2. Non-Functional Requirements

###### Feasibility Study:

Feasibility is a practical extent to which a project can be performed successfully. To evaluate feasibility, a feasibility study is performed, which determines whether the solution considered to accomplish the requirements is practical and workable in the software or not. Such information as resource availability, cost estimate for software development, benefits of the software to organization, and cost to be incurred on its maintenance are considered. The objective of the feasibility study is to establish the reasons for developing software that is acceptable to users and adaptable to change.

A feasibility study is a short, focused study, which aims to answer a number of questions:

* + - Does the system contribute to the overall objectives of the organization?
    - Can the system be implemented using current technology and within given cost and schedule constraints?
    - Can the system be integrated with systems which are already in place?
    - Operational Feasibility.
    - Technical Feasibility.
    - Financial & Economic Feasibility.
    - Time Schedule Feasibility.

###### Operational Feasibility:

The main purpose of checking Operational Feasibility is to find out whether the system will be functional after its development and installation or not. The outcomes of the operational feasibility are as follows:

* In this application user can create a group and manage the members of the group.
* User can track the position of another member.
* It eliminates the limitation ofan existing system

###### Technical Feasibility:

The main purpose of checking Technical Feasibility is to examine whether the current technology is sufficient for the development of the system. The outcomes of the technical feasibility are as follows:

###### Time Schedule Feasibility:

A project will fail if it takes too long to be completed before it is useful. Typically this means estimating how long the system will take to develop, and if it can be completed in a given time period using some methods like payback period. Given our technical expertise, are the project deadlines reasonable? Some projects are initiated with specific deadlines. You need to determine whether the deadlines are mandatory or desirable.

###### Economical Feasibility:

In economic analysis the procedure is to determine the benefits and savings that are expected from a candidate system and compare them with costs. If the expected benefits equal or exceed costs, the system can be judged to be economically feasible. Economic analysis is used for evaluating the effectiveness of the proposed system. The economic feasibility will review the expected costs to see if they are in-line with the projected budget or if the project has an acceptable return on investment. At this point, the projected costs will only be a rough estimate. A rough estimate of the project schedule is required to determine if it would be feasible to complete the systems project within a required timeframe. The required timeframe would need to be set by the organization.

* + 1. **IMPLEMENTATION FEASIBILITY:**

The main purpose of checking Operational Feasibility is to find out whether the system will be functional after its development and installation or not. The outcomes of the operational feasibility are as follows:

* + - * Service Provider website helps the user for getting information about new available
      * Service product quickly by presenting the data in proper format online. Graphical representation of data helps the user in taking proper decision in time.
      * So, it is supposed to improve the working efficiency of user. So, this application is operationally feasible

###### Requirement Specification:

* + 1. **Functional Requirements:**

###### Software Requirements:

* Console (CMD)
* IDE(Visual Studio Code, Spyder, Jupyter Notebook)
* Python
* Tensor flow API
* Dataset

###### Hardware Requirements:

* Camera
* Windows Operating System
* Screen
* Minimum 1 GB Storage Space
* Efficient RAM( around 1 GB )
* Sensor

###### Non-Functional Requirements:

* Stable internet connectivity
* Medium configured computer

### CHAPTER-5 SYSTEM DESIGN

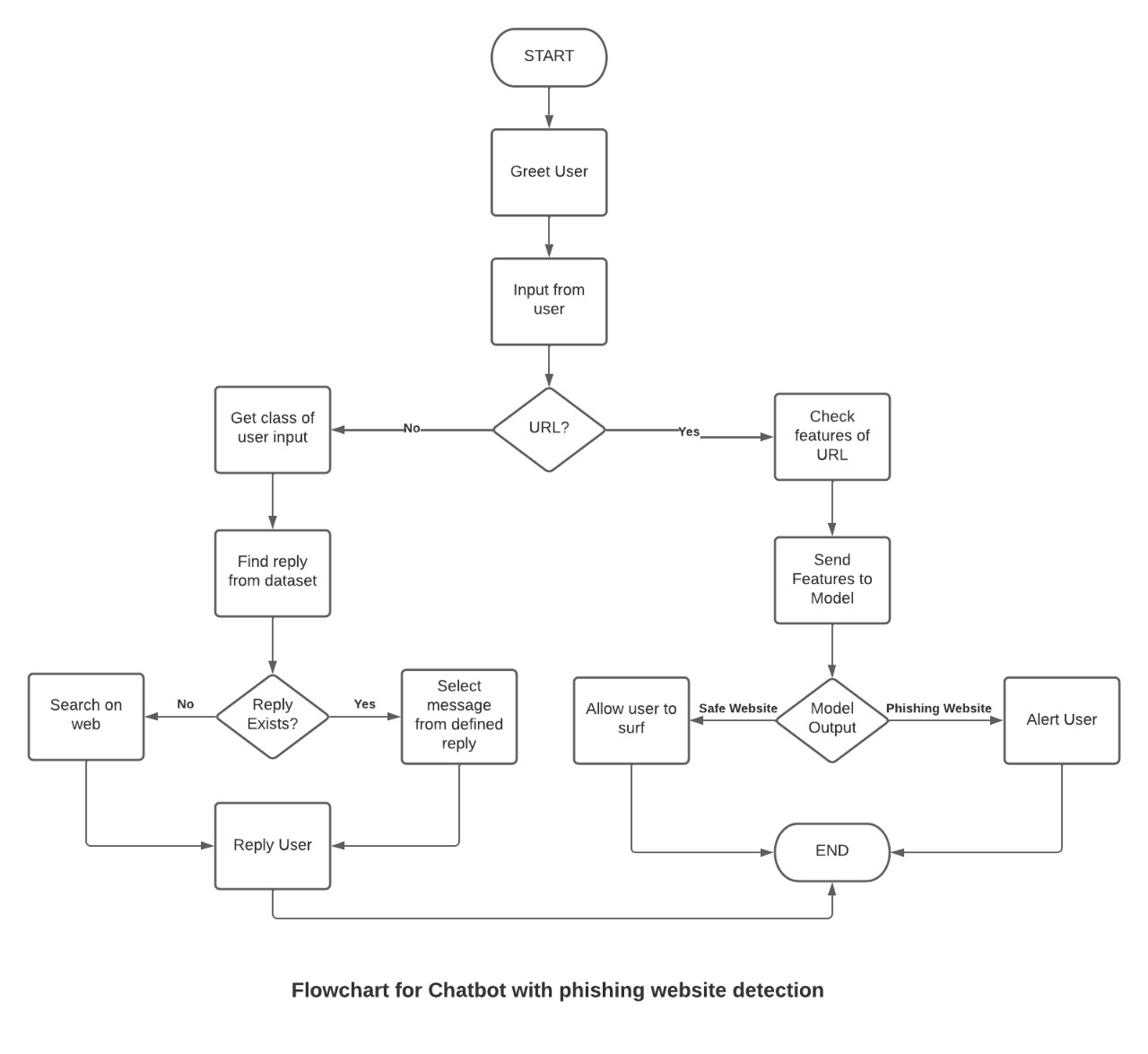


Fig:5.1

###### 

**CHAPTER-6 SCREEN LAYOUT**

**Screenshots:**

**Home page**

****

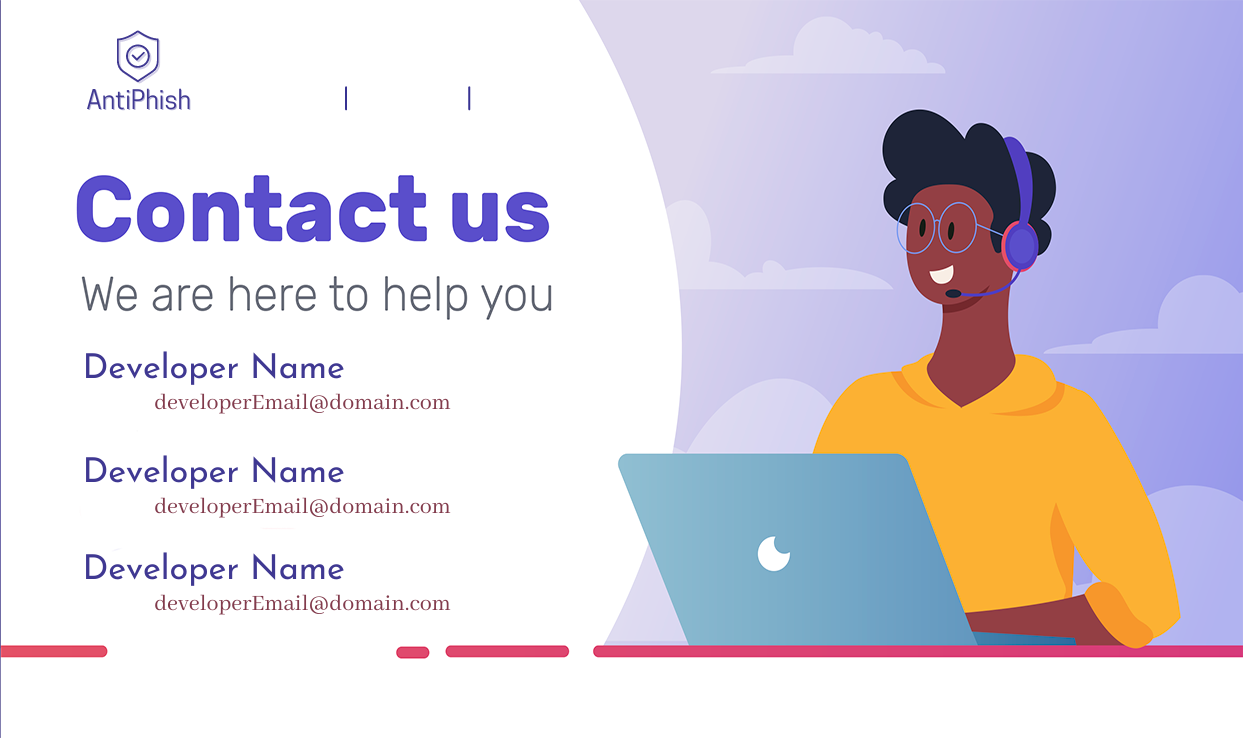
**About us:**

****

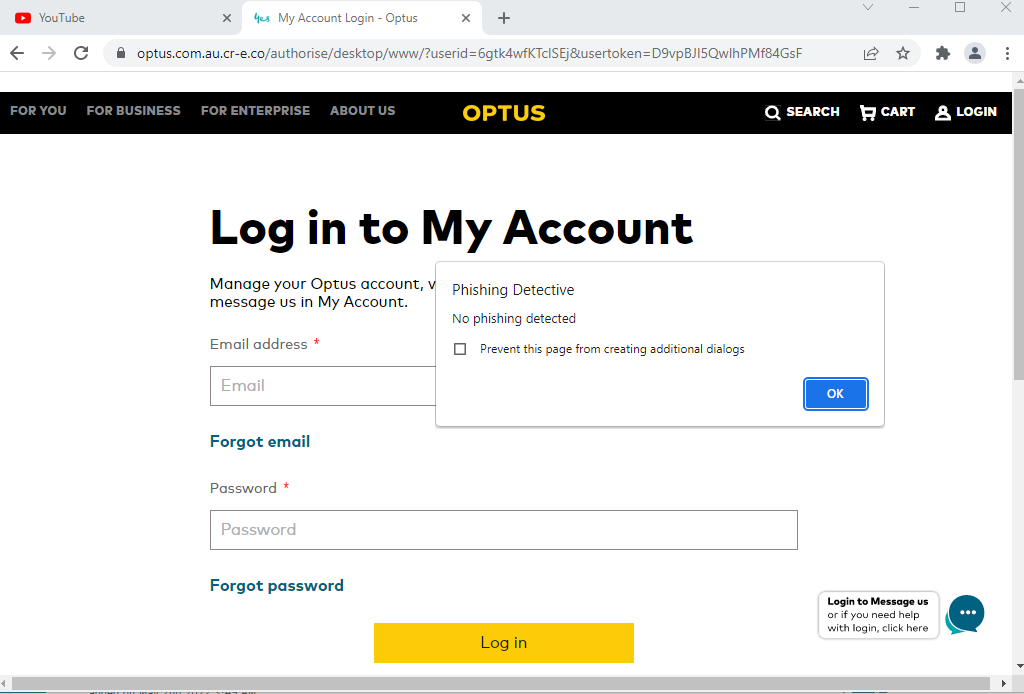


**What is Phishing**

**Contact us:**

****

**Example**



## CHAPTER-7 CONCLUSION

###### 7.1 Conclusion:

Thus concluding the research we have seen how the phishing attack is a very important concern in cyber security. We also reviewed some old methods for phishing detection and their drawbacks. We trained and tested four different supervised ml algorithms using a dataset from 'The UCI machine learning repository and showed their results. We selected the Decision tree algorithm as best based on its speed and accuracy. Our proposed tool can easily be deployed and used in realtime for the detection of phishing URLs and preventing cyber attacks.

###### Scope for further enhancement:

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

## CHAPTER-8 REFERENCE

###### Reference:

The following documents were referenced in developing this Synopsis:

[1] K. Jain and B. B. Gupta, PHISH-SAFE: URL features-based phishing detection system using machine learning, vol. 729. Springer Singapore, 2018.

[2] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, “Machine learning based phishing detection from URLs,” Expert Syst. Appl., vol. 117, pp. 345–357, 2019.

37

##### 38

##### 39

\

## 