ANALYSIS USING MACHINE LEARNING FOR HOAX URLs

AN INDUSTRY ORIENTED MINI PROJECT REPORT
Submitted to

Jawaharlal Nehru Technological University Hyderabad

In partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted by

P PAVAN KALYAN	**(18E11A0578)**
O VENKATESH	**(18E11A0577)**
SYED MUQTHAR ALI	**(18E11A0584)**
B UDAY KIRAN	**(18E11A0551)**
G SURYA PRAKASH	**(19E15A0508)**

Under the Supervision of

Dr. K. Thirupal Reddy

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
BHARAT INSTITUTE OF ENGINEERING AND TECHNOLOGY
Accredited by NAAC, Accredited by NBA, (UG Programmes: CSE, ECE, EEE & Mechanical)
Approved by AICTE, Affiliated to JNTUH Hyderabad
Ibrahimpatnam-501510, Hyderabad, Telangana

JANUARY 2022

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BHARAT INSTITUTE OF ENGINEERING AND TECHNOLOGY

Accredited by NAAC, Accredited by NBA (UG Programs: CSE, ECE, EEE & Mechanical)
Approved by AICTE, Affiliated to JNTUH Hyderabad
Ibrahimpatnam -501 510, Hyderabad, Telangana

Certificate

*This is to certify that the Industry Oriented Mini Project work entitled "***ANALYSIS USING MACHINE LEARNING FOR HOAX URLs"** *is the* bonafide *work done*

By

P PAVAN KALYAN	**(18E11A0578)**
O VENKATESH	**(18E11A0577)**
SYED MUQTHAR ALI	**(18E11A0584)**
B UDAY KIRAN	**(18E11A0551)**
G SURYA PRAKASH	**(19E15A0508)**

In the Department of **Computer Science and Engineering,** ***BHARAT INSTITUTE OF ENGINEERING AND TECHNOLOGY,*** Ibrahimpatnam is submitted to **Jawaharlal Nehru Technological University, Hyderabad** in partial fulfillment of the requirements for the award of **Bachelor of Technology** degree in **Computer Science and Engineering** during **2018-2022.**

Guide:

Dr. K. Thirupal Reddy
Assistant Professor
Dept of CSE,
Bharat Institute of Engineering and Technology,
Ibrahimpatnam – 501 510, Hyderabad.

Head of the Department:

Mr. Suyash Agarwal
Assistant professor
Dept of CSE
Bharat Institute of Engineering and Technology,
Ibrahimpatnam – 501 510, Hyderabad.

Viva-Voce held on……

Internal Examiner **External Examiner**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BHARAT INSTITUTE OF ENGINEERING AND TECHNOLOGY

Accredited by NAAC, Accredited by NBA (UG Programs: CSE, ECE, EEE & Mechanical)
Approved by AICTE, Affiliated to JNTUH Hyderabad
Ibrahimpatnam -501 510,Hyderabad, Telangana

Vision of the Institution

To achieve the autonomous & university status and spread universal education by inculcating discipline, character and knowledge into the young minds and would them into enlightened citizens.

Mission of the Institution

Our mission is to impart education, in a conductive ambience, as comprehensive as possible, with the support of all the modern technologies and make the students acquire the ability and passion to work wisely, creatively and effectively for the betterment of our society.

Vision of CSE department

Serving the high-quality educational needs of local and rural students within the core areas of Computer Science and Engineering and Information Technology through a rigorous curriculum of theory, research and collaboration with other disciplines that is distinguished by its impact on academia, industry and society.

Mission of CSE department

The Mission of the department of Computer Science and Engineering is to work closely with industry and research organizations to provide high quality computer education in both the theoretical and applications of Computer Science and Engineering. Then department encourages original thinking, fosters research and development, evolve innovative applications of technology.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
BHARAT INSTITUTE OF ENGINEERING AND TECHNOLOGY
Accredited by NAAC, Accredited by NBA (UG Programs: CSE, ECE, EEE & Mechanical)
Approved by AICTE, Affiliated to JNTUH Hyderabad
Ibrahimpatnam -501 510,Hyderabad, Telangana

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

The Computer Science and Engineering program provides students with an in depth education in theconceptual foundations of computer science and in complex hardware and software systems. It allows them to explore the connections between computer science and a variety of other disciplines in engineering and outside. Combined with a strong education in mathematics, science, and the liberal arts it prepares students to be leaders in computer science practice, applications to other disciplines and research.

Program Educational Objective 1: (PEO1)

The graduates of Computer Science and Engineering will have successful career in technology or managerial functions.

Program Educational Objective 2: (PEO2)

The graduates of the program will have solid technical and professional foundation to continue higher studies.

Program Educational Objective 3: (PEO3)

The graduates of the program will have skills to develop products, offer services and create new knowledge.

Program Educational Objective 4: (PEO4)

The graduates of the program will have fundamental awareness of Industry processes, tools and technologies.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BHARAT INSTITUTE OF ENGINEERING AND TECHNOLOGY

Accredited by NAAC, Accredited by NBA (UG Programs: CSE, ECE, EEE & Mechanical)
Approved by AICTE, Affiliated to JNTUH Hyderabad
Ibrahimpatnam -501 510,Hyderabad, Telangana

PROGRAM OUTCOMES (POs)

PO1:	**Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
PO2:	**Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
PO3:	**Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
PO4:	**Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
PO5:	**Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
PO6:	**The engineer and society**: Apply reasoning informed by the contextualknowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
PO7:	**Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
PO8:	**Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
PO9:	**Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
PO10:	**Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
PO11:	**Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

PO12:	**Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BHARAT INSTITUTE OF ENGINEERING AND TECHNOLOGY

Accredited by NAAC, Accredited by NBA (UG Programs: CSE, ECE, EEE & Mechanical)
Approved by AICTE, Affiliated to JNTUH Hyderabad
Ibrahimpatnam -501 510,Hyderabad, Telangana

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1:	**Foundation of mathematical concepts:** To use mathematical methodologies to crack problem using suitable mathematical analysis, data structure and suitable algorithm.
PSO2:	**Foundation of Computer System:** The ability to interpret the fundamental concepts and methodology of computer systems. Students can understand the functionality of hardware and software aspects of computer systems.
PSO3:	**Foundations of Software development:** The ability to grasp the software development lifecycle and methodologies of software systems. Possess competent skills and knowledge of software design process. Familiarity and practical proficiency with a broad area of programming concepts and provide new ideas and innovations towards research.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
BHARAT INSTITUTE OF ENGINEERING ANDTECHNOLOGY
Accredited by NAAC, Accredited by NBA (UG Programs: CSE, ECE, EEE & Mechanical)
Approved by AICTE, Affiliated to JNTUH Hyderabad
Ibrahimpatnam -501 510,Hyderabad, Telangana

QUALITY OF THE PROJECT

I. Consideration to Factors

Factors *(Environment, Safety, Ethics, Cost)*	Type of Project *(Application, Product, Research, Review, etc.)*	Standards
This project has a high Impact of Safety and is feasible.	This project is a research based project and is applicable in improving safety on social media .	The type of the project is based on application and its standards. The standard of this project is midrange.

II. POs and PSOs addressed through the project with justification

S. No.	POs and PSOs addressed	Justification
1.	PO1	Engineering Knowledge: we have applied the Data Mining Techniques and Machine learning techniques to solve the problem.
2.	PO2	Problem analysis: We have analyzed the problem of compromising the safety of users in social media by the spammers.
3.	PO3	Design/Development of solutions: we designed a solution which meets the specified needs of societal safety.
4.	PO5	Modern tool usage: We selected and applied the appropriate techniques,modern engineering and IT tools.
5.	PO6	The Engineer and Society: This Project context is to improve the safety of society and reduce the legal issues.
6.	PO7	Individual and teamwork: We worked best as individual's and as a team in multi-disciplinary settings.
7.	PO11	Project Management and Finance: Understanding the engineering principles we applied those to the work in terms of finance and managed the project .
8.	PSO1	Foundation of mathematical concepts: we used methodologies like calculating the accuracy and percentages to acquire results.

DECLARATION

We hereby declare that this Mini Project work is titled **"ANALYSIS USING MACHINE LEARNING FOR HOAX URLs"** is a genuine project carried out by us, in **B. Tech (Computer Science and Engineering)** degree course of **Jawaharlal Nehru Technology University Hyderabad, Hyderabad** and has not been submitted to any other course or university for the award of my degree by us.

Candidate Name(s)	Roll Number	Signature
1. P PAVAN KALYAN	18E11A0578	
2. O VENKATESH	18E11A0577	
3. SYED MUQTHAR ALI	18E11A0584	
4. B UDAY KIRAN	18E11A0551	
5. G SURYA PRAKASH	19E15A0508	

DATE:

ABSTRACT

The internet is full of websites that are either fake, fraudulent or a scam. It's a sad fact of life. The evolution of the internet has brought with it a number of extremely convenient advances in the way we shop, bank, and interact with the world around us. At the same time, that evolution has also given way to new risks. A large number of people buy things online and pay for them using various online payment platforms. Several websites require users to enter sensitive data in order to authenticate.

Currently, the risk of network information insecurity is increasing rapidly in number and level of danger. The methods mostly used by hackers today is to attack end-to-end technology and exploit human vulnerabilities. These techniques include social engineering, phishing, pharming, etc. One of the steps in conducting these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As a results, malicious URL detection is of great interest now a days.

Some phishing websites make use of this information for various purposes. Automated detection systems have arisen as a means of countering misleading websites, although the most of them are rather basic in terms of fraud and detecting methods.

Here, in this work a website is created for the users with the algorithm for detecting the website and categorizing them into levels based on the risk factor. This will help to understand the nature and purpose of the URL by further detailed study of the threating URLs.

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of the task would be put incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success.

We avail this opportunity to express our deep sense of gratitude and hearty thanks to

Sri CH. Venugopal Reddy, Chairman & Secretary of BIET, for providing congenial atmosphere and encouragement.

We would like to thank **Prof. G. Kumaraswamy Rao**, Former Director & O.S. of DLRL Ministry of Defense, Sr. Director R&D, BIET, for having provided all the facilities and support.

We would like to thank our Department Admin Incharge **Mr. Suyash Agarwal**, Department Academic Incharge **Mrs. Shilpa Sudheendran** for their expert guidance and encouragement at various levels of our project.

We are thankful to our Project Coordinator **Dr. N. Srihari Rao**, Assistant Professor, Department of Computer Science and Engineering for their support and cooperation throughout the process of this project.

We are thankful to our guide **Dr. K. Thirupal Reddy**, Assistant Professor in the Department of Computer Science and Engineering for her sustained inspiring Guidance and cooperation throughout the process of this project. Her wise counsel and suggestions were invaluable.

We express our deep sense of gratitude and thanks to all the Teaching and Non-Teaching Staff of our college who stood with us during the project and helped us to make it a successful venture.

We place highest regards to our Parent, our Friends and Well-wishers who helped a lot in making the report of this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVATIONS

SYMBOL	DESCRIPTION
URL	Uniform resource locator
SRS	Software requirement specification
SVM	Support Vector Machine
RFC	Random Forest Classifier
LG	Logistic Regression
SGD	Stochastic Gradient Descent
API	Application programming interface
CGI	Common gateway interface
HTML	Hypertext Markup Language
DFD	Data flow diagram
UML	Unified modelling language

1. INTRODUCTION

The URL is usually the first and cheapest information we have about a website. Therefore, it is a natural choice to develop techniques that would recognize a malicious URL from benign. Moreover, accessing and downloading the content of the website can be time consuming and brings risks associated with downloading potentially harmful content.

The idea is to develop a web-based application for users will become aware of fraudulent events which are done by the attackers. Using this concept one can know the difference between original websites and abnormal websites. Further measures are taken to keep secure.

Online services have become an irreplaceable part of today's businesses, schools, banking, or personal lives. With their increasing popularity, the number of malicious websites is growing. A malicious website contains some unsolicited content with a purpose to gather sensitive data or install malware onto a user's machine. Usually, some interaction from the user part is needed, but in the case of a drive by download, malware is installed automatically without asking for permission.

The URL is usually the first and cheapest information we have about a website. Therefore, it is a natural choice to develop techniques that would recognize a malicious URL from benign. Moreover, accessing and downloading the content of the website can be time consuming and brings risks associated with downloading potentially harmful content.

2. LITERATURE REVIEW

Reference 1:

Title: "Detection of Phishing Websites from URLs by using Classification Techniques on WEKA".

Author: Buket Geyik and Emre kocyigit

Description: Many of the applications are developed and released to the world daily. Everyone is using applications and making easier life. However, with the development of the applications cyber-attacks have increased gradually and identity thefts have emerged. It is a type of fraud committed by intruders by using fake web pages to access people's private information such as userid, password, credit card number and bank account numbers, etc. Machine learning technology has been used to detect and prevent this type of intrusions. The anti-phishing method has been developed by detecting the attacks made with the technologies used.

Reference 2:

Title: "URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis".

Author: Mohammed Abutaha

Description: Phishing URLs mainly target individuals and/or organizations through social engineering attacks by exploiting the humans' weaknesses in information security awareness. These URLs lure online users to access fake websites, and harvest their confidential information, such as debit/credit card numbers and other sensitive information. This approach can be incorporated within add-on/middleware features in Internet browsers for alerting online users whenever they try to access a phishing website using only its URL.

Reference 3:

Title: "Phishing Web Page Detection Methods: URL and HTML Features Detection".

Author: Humam Faris.

Description: Phishing is a type of fraud on the Internet in the form of fake web pages that mimic the original web pages to trick users into sending sensitive information to phisher. The statistics presented by APWG and Phistank show that the number of phishing websites from 2015 to 2020 tends to increase continuously. To overcome this problem, several studies have been carried out including detecting phishing web pages using various features of web pages with various methods. A security detection device should require effectiveness, good performance, and deployable.

Reference 4:

Title: “Wide scope and fast websites phishing detection using URLs lexical features”.

Author: Ammar Yahya Daeef.

Description: Phishing is a considerable problem differs from the other security threats such as intrusions and Malware which are based on the technical security holes of the network systems. The weakness point of any network system is its Users. Phishing attacks are targeting these users depending on the trikes of social engineering. Despite there are several ways to carry out these attacks, unfortunately the current phishing detection techniques cover some attack vectors like email and fake websites. Therefore, building a specific limited scope detection system will not provide complete protection from the wide phishing attack vectors. This paper develops detection system with a wide protection scope using URL features only which is relying on the fact that users directly deal with URLs to surf the internet and provides a good approach to detect malicious URLs as proved by previous studies.

3. CONCLUSION OF DESCRIPTION

The goal of the project is to help the people by providing a method to stay secure while using online websites or any kind of web applications. We use machine learning algorithms to develop this web application. It contains the malicious URL predictor and later on it will be analyzed for further details. So that every user will not be affected by certain attacks. They will identify the nature of the URL and it is classified as Hoax or legitimate category. This reduces several kinds of attacks. The attackers will not be able to access the public data. The data will be kept secure.

4. OBJECTIVE

The main objective of this work is to predict the malicious URL by using machine learning algorithm. We use python to develop this web application. It contains the user interactive web page. An input URL is taken from the user and later on it is extracted to analyze the nature of the URL. After that prediction is done with the help of the trained data. The URL nature is classified finally.

4.1 FEASIBILITY STUDY

After requirement gathering, the team comes up with a rough plan of software process. At this step the team analyzes if a software can be made to fulfill all requirements of the user and if there is any possibility of software being no more useful. It is found out, if the project is financially, practically and technologically feasible for the organization to take up. There are many algorithms available, which help the developers to conclude the feasibility of a software project.

4.2 Types of Feasibility Study:

The feasibility study mainly concentrates on below five mentioned areas. Among this Economic Feasibility Study is most important part of the feasibility analysis and Legal Feasibility Study is less considered feasibility analysis.

4.2.1 Technical Feasibility:

In Technical Feasibility current resources both hardware software along with required technology are analyzed/assessed to develop project. This technical feasibility study gives report whether there exists correct required resources and technologies which will be used for project development. Along with this, feasibility study also analyzes technical skills and capabilities of technical team, existing technology can be used or not, maintenance and up-gradation is easy or not for chosen technology etc.

4.2.2 Operational Feasibility:

In Operational Feasibility degree of providing service to requirements is analyzed along with how much easy product will be to operate and maintenance after deployment. Along with this other operational scopes are determining usability of product, determining suggested solution by software development team is acceptable or not etc.

4.2.3 Economic Feasibility:

In Economic Feasibility study cost and benefit of the project is analyzed. Means under this feasibility study a detail analysis is carried out what will be cost of the project for development which includes all required cost for final development like hardware and software resource required, design and development cost and operational cost and so on. After that it is analyzed whether project will be beneficial in terms of finance for organization or not.

4.2.4 Legal Feasibility:

In Legal Feasibility study project is analyzed in legality point of view. This includes analyzing barriers of legal implementation of project, data protection acts or social media laws, project certificate, license, copyright etc. Overall, it can be said that Legal Feasibility Study is study to know if proposed project conforms legal and ethical requirement.

4.2.5 Schedule Feasibility:

In Schedule Feasibility Study mainly timelines/deadlines is analyzed for proposed project which includes how many times teams will take to complete final project which has a great impact on the organization as purpose of project may fail if it can’t be completed on time.

4.3 Software Development Life Cycle:

The Systems Development Life Cycle (SDLC), or Software Development Life Cycle in systems engineering, information systems and software engineering, is the process of creating or altering systems, and the models and methodologies use to develop these systems.

Fig 4.1 : software development life cycle

4.4 Requirement Analysis and Design

Analysis gathers the requirements for the system. This stage includes a detailed study of the business needs of the organization. Options for changing the business process may be considered. Design focuses on high level design like, what programs are needed and how are they going to interact, low-level design (how the individual programs are going to work), interface design (what are the interfaces going to look like) and data design

(what data will be required). During these phases, the software's overall structure is defined. Analysis and Design are very crucial in the whole development cycle. Any glitch in the design phase could be very expensive to solve in the later stage of the software development. Much care is taken during this phase. The logical system of the product is developed in this phase.

4.5 Implementation

In this phase the designs are translated into code. Computer programs are written using a conventional programming language or an application generator. Programming tools like Compilers, Interpreters, and Debuggers are used to generate the code. Different high level programming languages like PYTHON 3.6, Anaconda Cloud are used for coding. With respect to the type of application, the right programming language is chosen.

4.6 Testing

In this phase the system is tested. Normally programs are written as a series of individual modules, this subject to separate and detailed test. The system is then tested as a whole. The separate modules are brought together and tested as a complete system. The system is tested to ensure that interfaces between modules work (integration testing), the system works on the intended platform and with the expected volume of data (volume testing) and that the system does what the user requires (acceptance/beta testing).

4.7 Maintenance

Inevitably the system will need maintenance. Software will definitely undergo change once it is delivered to the customer. There are many reasons for the change. Change could happen because of some unexpected input values into the system. In addition, the changes in the system could directly affect the software operations. The software should be developed to accommodate changes that could happen during the post implementation period.

4.8 SDLC METHDOLOGIES

The software development paradigm helps developer to select a strategy to develop the software. A software development paradigm has its own set of tools, methods and procedures, which are expressed clearly and defines software development life cycle.

4.8.1 Spiral Model

Spiral model was defined by Barry Boehm in his 1988 article, “A spiral Model of Software Development and Enhancement. This model was not the first model to discuss iterative development, but it was the first model to explain why the iteration models.

As originally envisioned, the iterations were typically 6 months to 2 years long. Each phase starts with a design goal and ends with a client reviewing the progress thus far. Analysis and engineering efforts are applied at each phase of the project, with an eye toward the end goal of the project.

The following diagram shows how a spiral model acts like:

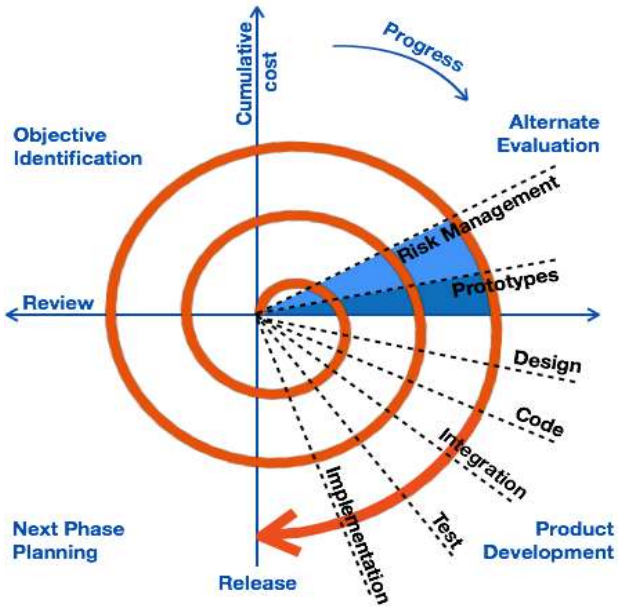


Fig 4.2: Spiral model

The steps for Spiral Model can be generalized as follows:

- The new system requirements are defined in as much details as possible.
- This usually involves interviewing a number of users representing all the external or internal usersand other aspects of the existing system.
- A preliminary design is created for the new system.
- A first prototype of the new system is constructed from the preliminary design. This is usually ascaled-down system, and represents an approximation of the characteristics of the final product.

5. PROBLEM STATEMENT

5.1 Existing System:

Current system is working with URL detection where a URL is analyzed by using Logistic Regression (LR) and Stochastic Gradient Descent (SGD). The prediction is not so effective. The accuracy is predicted by using these algorithms and that result percentage is low. Using existing system can analyze existing URL data sets. There are a lot of URLs are generating day by day which are different from one to other.

5.2 Proposed System:

Proposed system is a web application which there is a user interface, a user can give input URL to the input module. Then that URL is divided into different parts. That is tested by the previous trained datasets. Then prediction is done by using Support Vector Machine and Random Forest to show the comparison between them.

6. INPUT DESIGN AND OUTPUT DESIGN

6.1 INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

6.2 OBJECTIVES:

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.
2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also providesrecord viewing facilities.
3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

6.3 OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information

3. Create document, report, or other formats that contain information produced by the system. The output form of an information system should accomplish one or more of the following objectives

Convey information about past activities, current status or projections of the

- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

A second prototype is evolved by a fourfold procedure:

- Evaluating the first prototype in terms of its strengths, weakness, and risks.
- Defining the requirements of the second prototype.
- Planning an designing the second prototype.
- Constructing and testing the second prototype.
- At the customer option, the entire project can be aborted if the risk is deemed too great. Risk factors might involve development cost overruns, operating-cost miscalculation, or any other factor that could, in the customer's judgment, result in a less-than-satisfactory final product.
- The existing prototype is evaluated in the same manner as was the previous prototype, and if necessary, another prototype is developed from it according to the fourfold procedure outlined above.
- The preceding steps are iterated until the customer is satisfied that the refined prototype represents the final product desired.
- The final system is constructed, based on the refined prompt maintenance is carried on a continuing basis to prevent large scale failures and to minimize down time.

7. Software and Hardware Requirements:

7.1 Hardware Requirements

- Processor : I3 and above
- Ram : 4GB.
- Hard Disk : 256 GB.

7.2 Software Requirements:

- Server : local host
- Technology : Python
- Client-Side Technologies : Html, CSS
- IDE : Anaconda
- UML Design/E-R Modeling Tools : Star UML

8. DESIGN AND IMPLEMENTATION

8.1 STUDY OF THE SYSTEM

To provide flexibility to the users, the interfaces have been developed that are accessible through a browser. The GUI'S at the top level have been categorized as

8.2 Analysis:

Although the scale of this work is relatively small, to produce a professional solution is it imperative that the current problem is understood accurately. However, this task has been made doubly difficult by the lack of support from the company. Thankfully, the Application manager has been kind enough to spare me some of his own time to discuss the problem with me further. Therefore, this chapter is concerning with analyzing the current situation and expectations of the user for this system

8.2.1 Requirements:

The minimum requirements of the project are listed below:

- Examine the tools and methodologies required to gain an overview of the system requirements forthe proposed database.
- Evaluate appropriate website authoring and web graphic creation tools that can be used to develop web based forms for the proposed database
- Produce and apply suitable criteria for evaluating the solution

8.3 Requirement Analysis:

Taking into account the comparative analysis stated in the previous section we could start specifying the requirements that our website should achieve. It is very critical process that enables the success of a system or software project to be assessed. Requirements are generally split into two types: Functional and Non-functional requirements.

8.3.1 Functional requirements

Functional requirement should include function performed by a specific screen outline work-5flowsperformed by the system and other business or compliance requirement the system must meet.

Functional requirements specify which output file should be produced from the given file they describe the relationship between the input and output of the system, for

each functional requirement a detailed description of all data inputs and their source and the range of valid inputs must be specified. The functional specification describes what the system must do, how the system does it is described in the design specification.

If a user requirement specification was written, all requirements outlined in the user requirements specifications should be addressed in the functional requirements.

These are the requirements that the end user specifically demands as basic facilities that the system should offer. All these functionalities need to be necessarily incorporated into the system as a part of the contract. These are represented or stated in the form of input to be given to the system, the operation performed and the output expected. They are basically the requirements stated by the user which one can see directly in the final product, unlike the non-functional requirements.

- Input
- Enter URL
- Submit
- Extraction
- Comparison
- Analysis
- Legitimate URL
- Phishing URL

8.3.2 Non-functional requirement

These are basically the quality constraints that the system must satisfy according to the project contract. The priority or extent to which these factors are implemented varies from one project to other. They are also called non-behavioral requirements.

They basically deal with issues like:

- Portability
- Security
- Maintainability
- Reliability
- Scalability
- Performance
- Reusability
- Flexibility

8.3.3 UI Requirements

1. **Administrative user interface**

The administrative user interface' concentrates on the consistent information that is practically, part of the organizational activities and which needs proper authentication for the data collection. These interfaces help the administrators with all the transactional states like Data insertion, Data deletion and Date updating along with the extensive data search capabilities.

2. **The operational or generic user interface**

The 'operational or generic user interface' helps the end users of the system in transactions through the existing data and required services. The operational user interface also helps the ordinary users in managing their own information in a customized manner as per the included Flexibilities.

8.4 SOFTWARE REQUIREMENT SPECIFICATION

8.4.1 What is SRS?

Software Requirements Specification (SRS) is the starting point of the software developing activity. As system grew more complex it became evident that the goal of the entire system cannot be easily comprehended. Hence the need for the requirement phase arose. The software project is initiated by the client needs. The SRS is the means of translating the ideas of the minds of clients (the input) into a formal document (the output of the requirement phase).

The SRS phase consists of two basic activities:

Problem/Requirement Analysis:

The process is order and more nebulous of the two, deals with understand the problem, the goal and constraints.

Requirement Specification:

Here, the focus is on specifying what has been found giving analysis such as representation, Specification languages and tools, and checking the specifications are addressed during this activity.

The requirement phase terminates with the production of the validate SRS document. Producing the SRS document is the basic of this phase.

8.4.2 Role of SRS:

The purpose of the SRS is to reduce the communication gap between the clients and the developers. SRS is the medium though which the client and user needsare accurately specified. It forms the basis of software development. A good SRS should satisfy allthe parties involved in the system.

Purpose: The purpose of this document is to describe all external requirements for the E-learning System. It also describes the interfaces for the system.

Scope: This document is the only one that describes the requirements of the system. It is meant for the use by the developers, and will also by the basis for validating the final deliver system. Any changes made to the requirements in the future will have to go through a formal change approval process. The developer is responsible for asking for clarifications, where necessary, and will not make any alternations without the permission of the client.

- **Overview:** The SRS begins the translation process that converts the software Requirements into the language the developers will use. The SRS draws on the Use Cases from the user Requirement Document and analyses the situations from a number of perspectives to discover and eliminate inconsistencies, ambiguities and omissions before development progresses significantly under mistaken assumptions.

9. Python and its libraries

9.1 What is Python

Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Guido van Rossum began working on Python in the late 1980s, as a successor to the ABC programming language, and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000 and introduced new features, such as list comprehensions and a cycle-detecting garbage collection system (in addition to reference counting). Python 3.0 was released in 2008 and was a major revision of the language that is not completely backward-compatible. Python 2 was discontinued with version 2.7.18 in 2020.

Application

Python is a popular and in-demand skill to learn. Python is used in many real-world applications. Some of them are as follows:

1. Web Development.
2. Game Development.
3. Scientific and Numeric applications.
4. Artificial Intelligence and Machine Learning.
5. Desktop GUI.
6. Software Development.
7. Enterprise-level/Business Applications.
8. Image Processing and Graphic Design Applications.
9. Web Scraping Applications.
10. Operating Systems.

9.2 Types of Libraries available in Python

A Python library is a reusable chunk of code that you may want to include in your programs/ projects.

The Python Standard Library is a collection of exact syntax, token, and semantics of Python. It comes bundled with core Python distribution. We mentioned this when we began with an introduction.

9.2.1. Pandas

It provides fast, expressive, and flexible data structures to easily (and intuitively) work with structured (tabular, multidimensional, potentially heterogeneous) and time-series data.

9.2.2. Requests

Requests is a Python Library that lets you send HTTP/1.1 requests, add headers, form data, multipart files, and parameters with simple Python dictionaries.

9.2.3. NumPy

It has advanced math functions and a rudimentary scientific computing package.

9.2.4. BeautifulSoup

It may be a bit slow, BeautifulSoup has an excellent XML- and HTML- parsing library for beginners.

9.2.5. SciPy

Next up is SciPy, one of the libraries we have been talking so much about. It has a number of user-friendly and efficient numerical routines.

9.2.6. Flask

A web framework, Flask is built with a small core and many extensions.

9.3 Anaconda Environment

- Anaconda Individual Edition contains conda and Anaconda Navigator, as well as Python and hundreds of scientific packages. When you installed Anaconda, you installed all these too.
- Conda works on your command line interface such as Anaconda Prompt on Windows and terminal on macOS and Linux.
- Navigator is a desktop graphical user interface that allows you to launch applications and easily manage conda packages, environments, and channels without using command-line commands.
- You can try both conda and Navigator to see which is right for you to manage your packages and environments. You can even switch between them, and the work you do with one can be viewed in the other.

  To open Navigator

  From the Start menu, click the Anaconda Navigator desktop app.

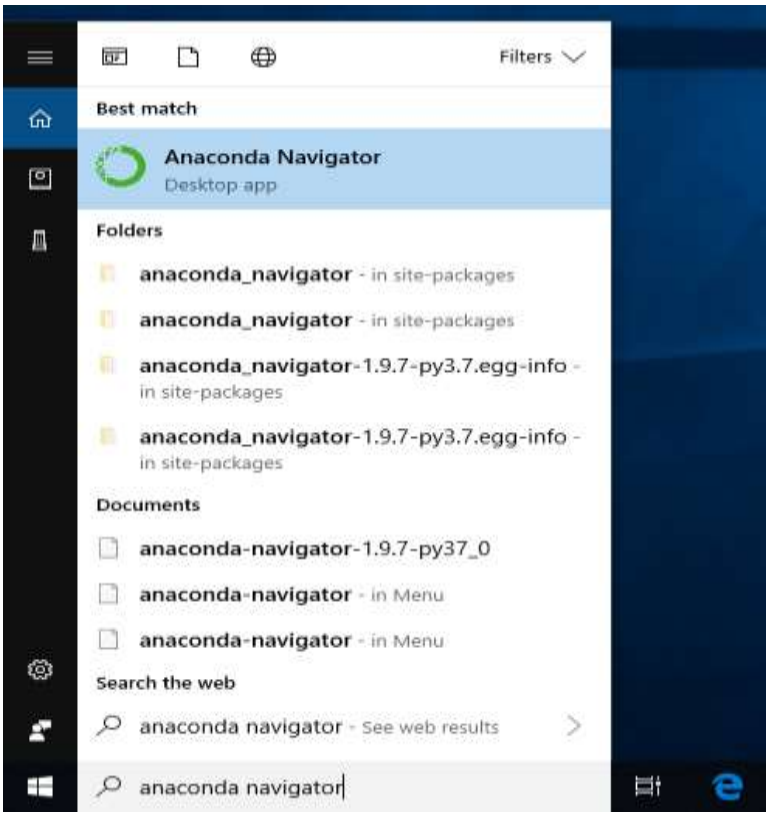


Fig 9.1: Anaconda Navigator

Run Python in a Jupyter Notebook

On Navigator's Home tab, in the Applications pane on the right, scroll to the Jupyter Notebook tile and click the Install button to install Jupyter Notebook.

Launch Jupyter Notebook by clicking Jupyter Notebook's Launch button.

This will launch a new browser window (or a new tab) showing the Notebook Dashboard.

IMPLEMENTATION

Data Collection: The dataset has been taken from an online open source. That is being analyzed and used it as training model for this work.

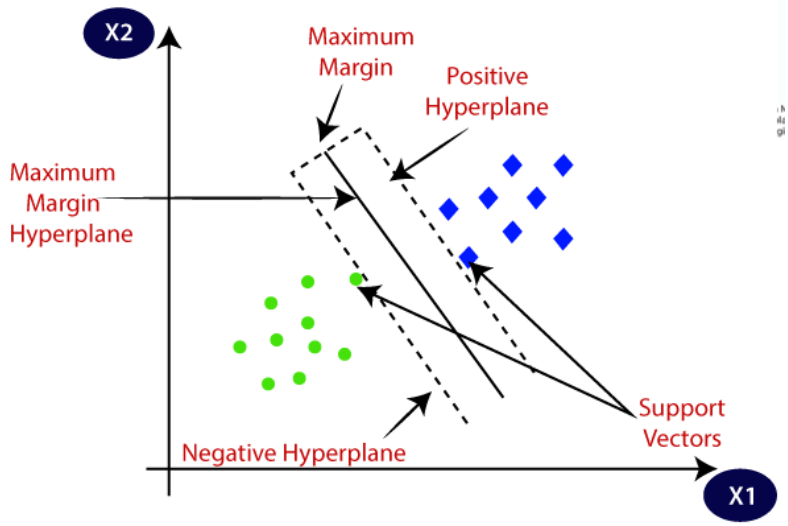
10. Algorithms

10.1 Support Vector Machine Algorithm

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



10.2 Types of SVM

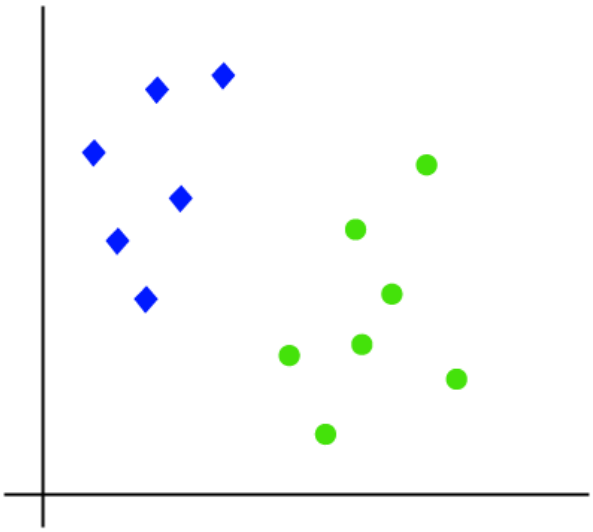
SVM can be of two types:

Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

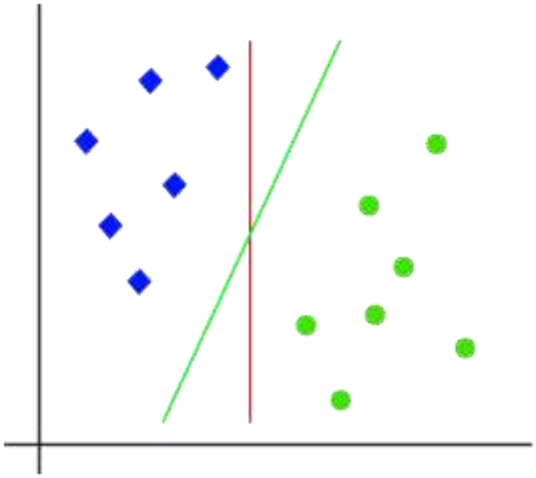
Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

10.2.1 Linear SVM:

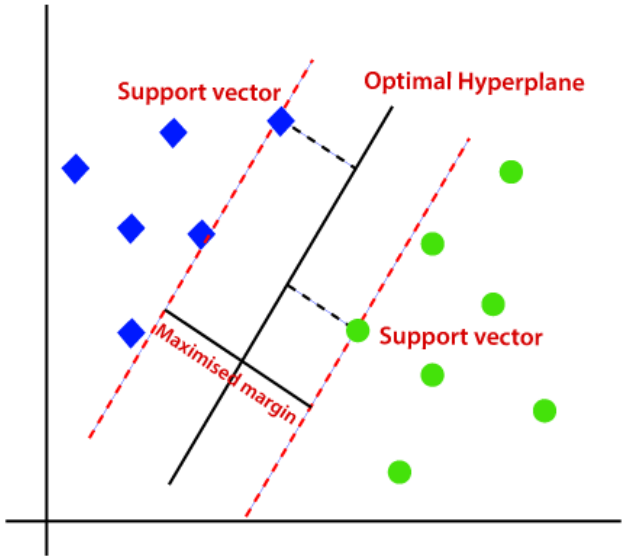
The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and x2. We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue. Consider the below image:



So, as it is 2-d space so by just using a straight line, we can easily separate these two classes. But there can be multiple lines that can separate these classes. Consider the below image:

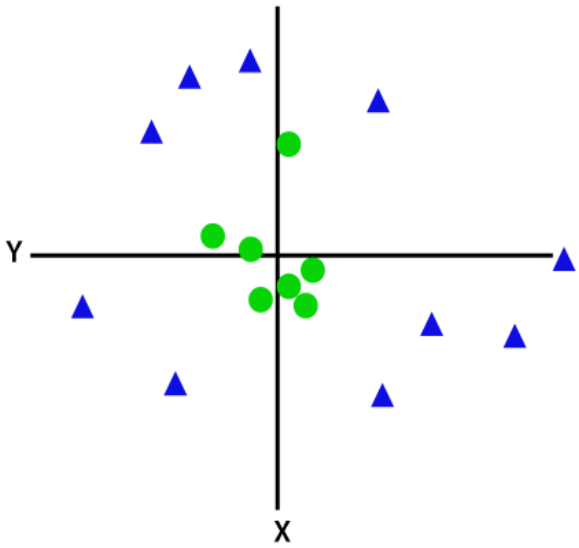


Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a hyperplane. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as margin. And the goal of SVM is to maximize this margin. The hyperplane with maximum margin is called the optimal hyperplane.



10.2.3 Non-Linear SVM:

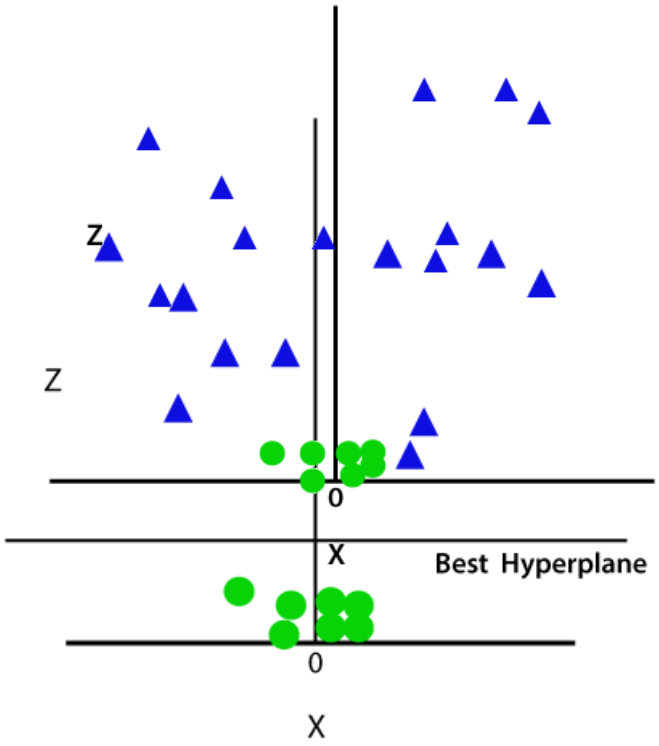
If data is linearly arranged, then we can separate it by using a straight line, but for non-linear data, we cannot draw a single straight line. Consider the below image:



So to separate these data points, we need to add one more dimension. For linear data, we have used two dimensions x and y, so for non-linear data, we will add a third dimension z. It can be calculated as:

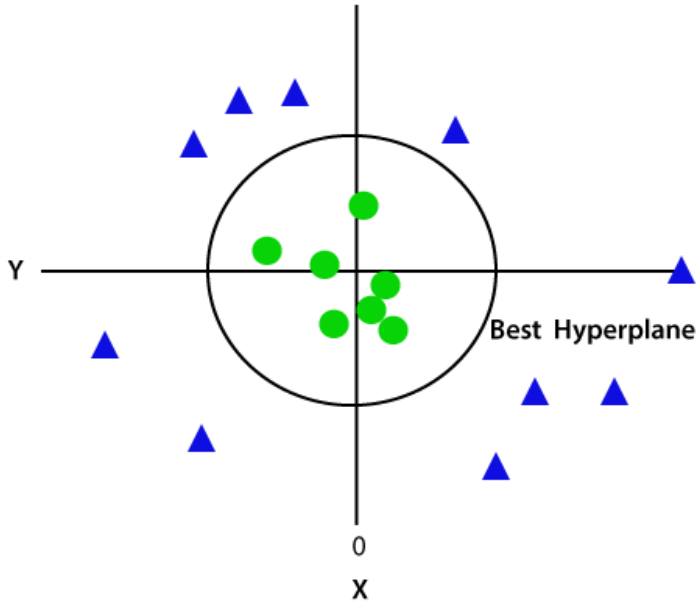
$z=x^2 +y^2$

By adding the third dimension, the sample space will become as below image:



So now, SVM will divide the datasets into classes in the following way. Consider the below image:

Since we are in 3-d Space, hence it is looking like a plane parallel to the x-axis. If we convert it in 2d space with z=1, then it will become as:



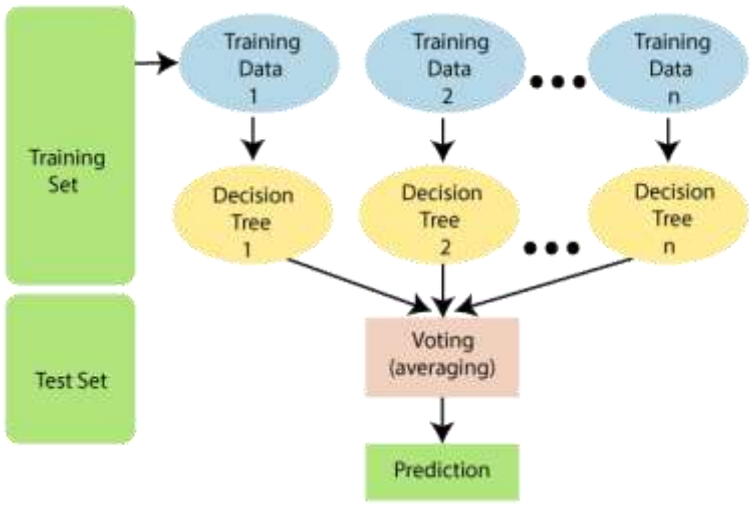
10.3 Random Forest Algorithm:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:



Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

11.URL

11.1 URL

A Uniform Resource Locator is a reference to a web resource that specifies its location on a computer network and a mechanism for accessing it.

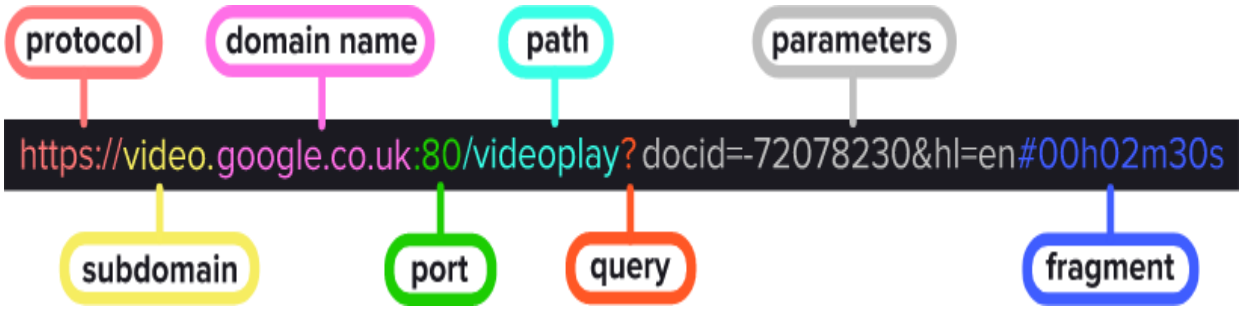


Fig 11.1: URL Structure

11.1.1 Protocol

The protocol declares how your web browser should communicate with a web server when sending or fetching a web page or document. The most common protocol is http which stands for Hypertext Transfer Protocol.

Another common protocol is https which stands for Hypertext Transfer Protocol Secure. You'll see this on secure pages, like shopping sites and log in pages. If you're visiting a site where you need to enter sensitive information, like bank details and passwords, make sure the protocol is declared as https. This means your web browser encrypts any information you provide so it can't be understood by any phishers who try to intercept the page during transfer.

11.1.2 Subdomain

A subdomain is a subdivision of the main domain name. For example, mail.doepud.com and calendar.doepud.com are subdomains of the domain name doepud.com.

11.1.3 Domain Name

A domain name is a unique reference that identifies a web site on the internet, for example doepud.co.uk. A domain name always includes the top-level domain (TLD), which in Doepud's case is UK. The co part is shorthand for commercial and combined.co.uk is called a second-level domain (SLD).

11.1.4 Port

The port number is rarely visible in URLs but always required. When declared in a URL it comes right after the TLD, separated by a colon. When it's *not* declared and in most cases where the protocol is http, port 80 is used. For https (secure) requests port 443 is used.

11.1.5 Path

The path typically refers to a file or directory on the web server, e.g. /directory/file.php. Sometimes the file name won't be specified, e.g. https://doepud.co.uk/blog/ so a web browser will automatically look inside the /blog/ folder for a file called index or default. If neither can be found, a *404 Not Found* error will usually be returned by the server.

11.1.6 Query

A query is commonly found in the URL of dynamic pages (ones which are generated from database or user-generated content) and is represented by a question mark followed by one or more parameters. The query directly follows the domain name, path or port number.

11.1.7 Parameters

Parameters are snippets of information found in the query string of a URL.

11.1.8 Fragment

A fragment is an internal page reference, sometimes called a named *anchor*. It usually appears at the end of a URL and begins with a hash (#) character followed by an identifier. It refers to a section *within* a web page.

11.2 What is CGI ?

CGI is actually an external application which is written by using any of the programming languages like C or C++ and this is responsible for processing client requests and generating dynamic content. In CGI application, when a client makes a request to access dynamic Web pages, the Web serverperforms the following operations :

- It first locates the requested web page i.e the required CGI application using URL.
- It then creates a new process to service the client's request.
- Invokes the CGI application within the process and passes the request information to the server.
- Collects the response from CGI application.
- Destroys the process, prepares the HTTP response and sends it to the client.

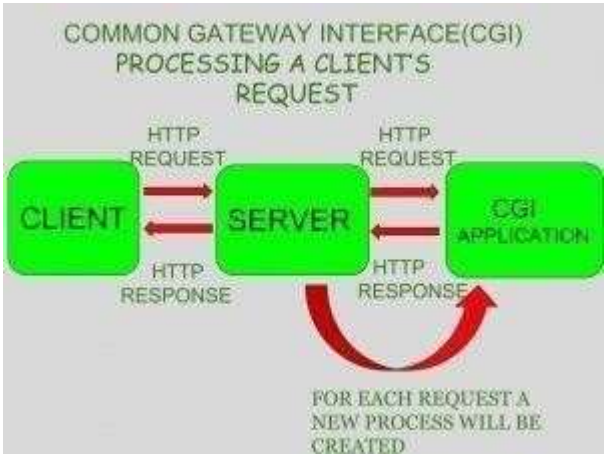


Fig 11.2: CGI

So, in CGI server has to create and destroy the process for every request. It's easy to understand that this approach is applicable for handling few clients but as the number of clients increases, the workload on the server increases and so the time taken to process requests increases.

12.UML

12.1 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object-oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of softwaresystem, as well as for business modeling and other non-software systems. The UML represents a collection of best engineering practices thathave proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects-oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects

Do we really need UML?

- Complex applications need collaboration and planning from multiple teams and hence require a clear and concise way to communicate amongst them.
- Businessmen do not understand code. So UML becomes essential to communicate with non programmer's essential requirements, functionalities and processes of the system.
- A lot of time is saved down the line when teams are able to visualize processes, user interactions and static structure of the system.
- UML is linked with object oriented design and analysis. UML makes the use of elements and forms associations between them to form diagrams. Diagrams in UML can be broadly classified as:

12.2 GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices

12.3 Types of UML Diagrams:

Structural Diagrams:

Capture static aspects or structure of a system. Structural Diagrams include: Component Diagrams, Object Diagrams, Class Diagrams and Deployment Diagrams.

Behavior Diagrams:

Capture dynamic aspects or behavior of the system. Behavior diagrams include: Use Case Diagrams, State Diagrams, Activity Diagrams and Interaction Diagrams.

12.3.1 USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor.

Roles of the actors in the system can be depicted

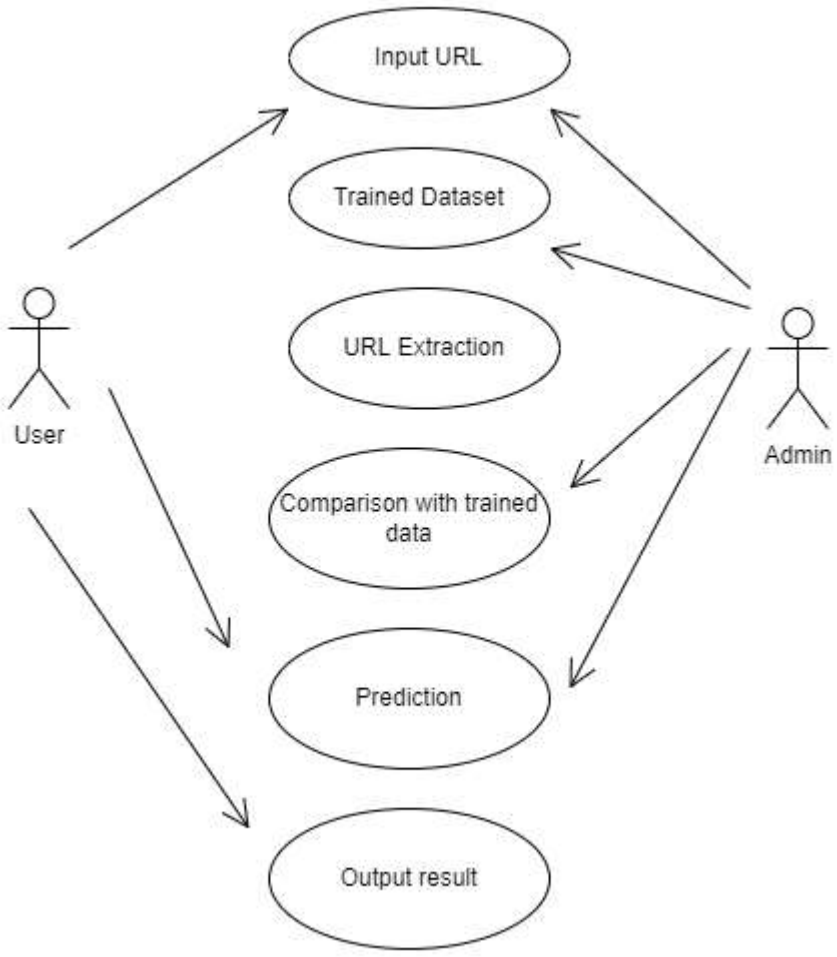


Fig 12.1: Use case Diagram

12.3.2 CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

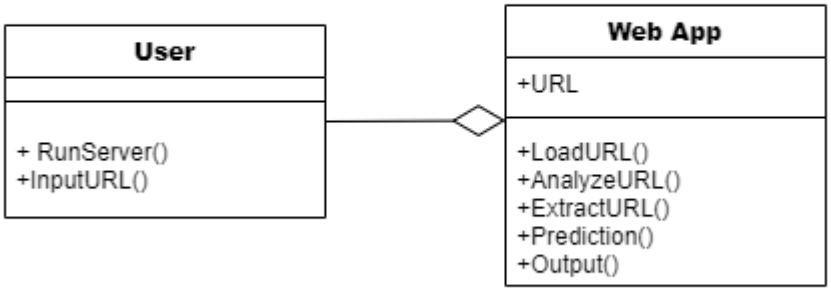


Fig 12.2: Class diagram

12.3.3 SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

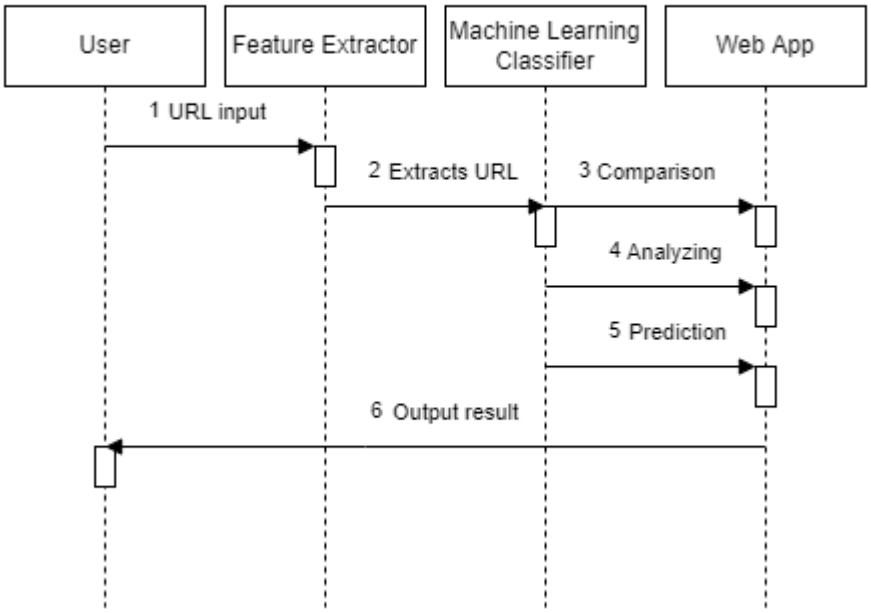


Fig 12.3: Sequence diagram

12.3.4 ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

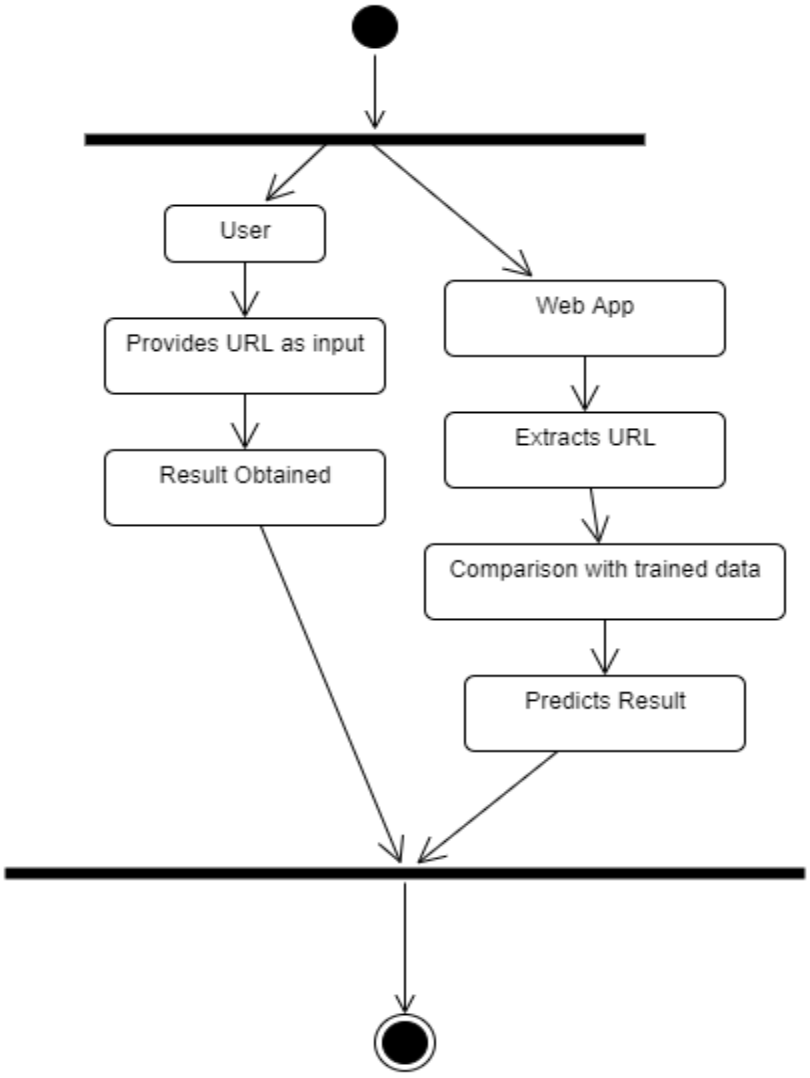


Fig 12.4: Activity Diagram

12.3.5 Experimental Studies:

TYPES OF TESTS

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

1) All field entries must work properly.

2) Pages must be activated from the identified link, The entry screen, messages and responses mustnot be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

13.Testing

13.1 Testing:

Software testing is a critical element of software quality assurance and represents the ultimate review of specification, design and code generation.

13.2 TESTING OBJECTIVES

- To ensure that during operation the system will perform as per specification.
- To make sure that system meets the user requirements during operation
- To make sure that during the operation, incorrect input, processing and output will be detected
- To see that when correct inputs are fed to the system the outputs are correct
- To verify that the controls incorporated in the same system as intended
- Testing is a process of executing a program with the intent of finding an error
- A good test case is one that has a high probability of finding an as yet undiscovered error.

The software developed has been tested successfully using the following testing strategies and any errors that are encountered are corrected and again the part of the program or the procedure or function is put to testing until all the errors are removed. A successful test is one that uncovers an as yet undiscovered error.

Note that the result of the system testing will prove that the system is working correctly. It will give confidence to system designer, users of the system, prevent frustration during implementation process etc.,

13.3 Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application

.it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

13.4 Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

13.5 Functional testing

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:

Valid Input	: identified classes of valid input must be accepted.
Invalid Input	: identified classes of invalid input must be rejected.
Functions	: identified functions must be exercised.
Output	: identified classes of application outputs must be exercised
Systems/Procedures	: interfacing systems or procedures must be invoke

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current testsis determined.

13.6 System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

13.7 White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It isused to test areas that cannot be reached from a black box level.

13.8 Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works

13.9 Test Cases:

Test case Id	Test case Description	Test Data	Expected Result	Actual Result	Pass/Fail
1	Admin Access	Admin Access To Server	Access Should be Successful	Access Was Successful	Pass
2	Input URL	Entering URL	URL input taken Successfully	Input Successful	Pass
3	View Output	Viewing predicted data	Output data is obtained Successful	Output Successful	Pass

Table 1: Test cases

14. RESULTS

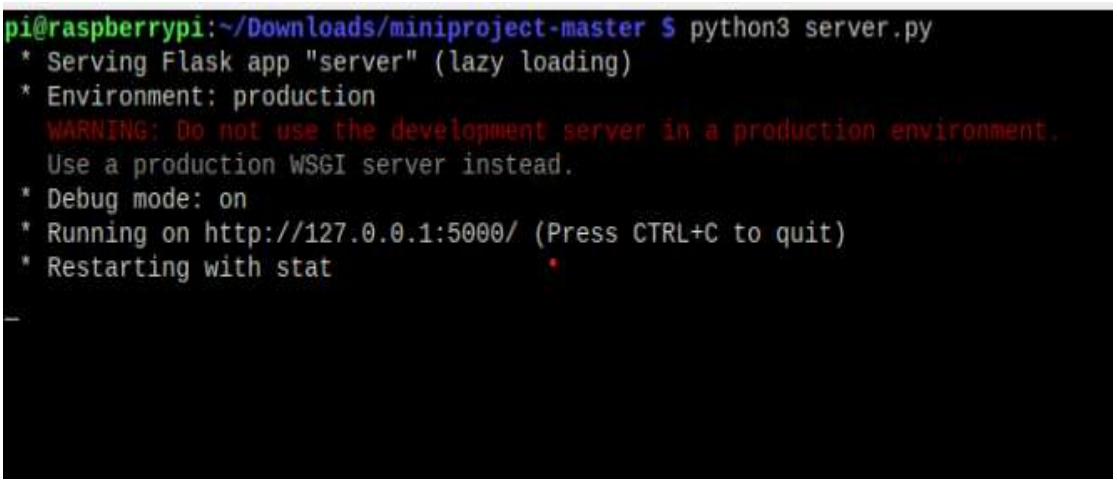


Fig 14.1: Screenshot of terminal server

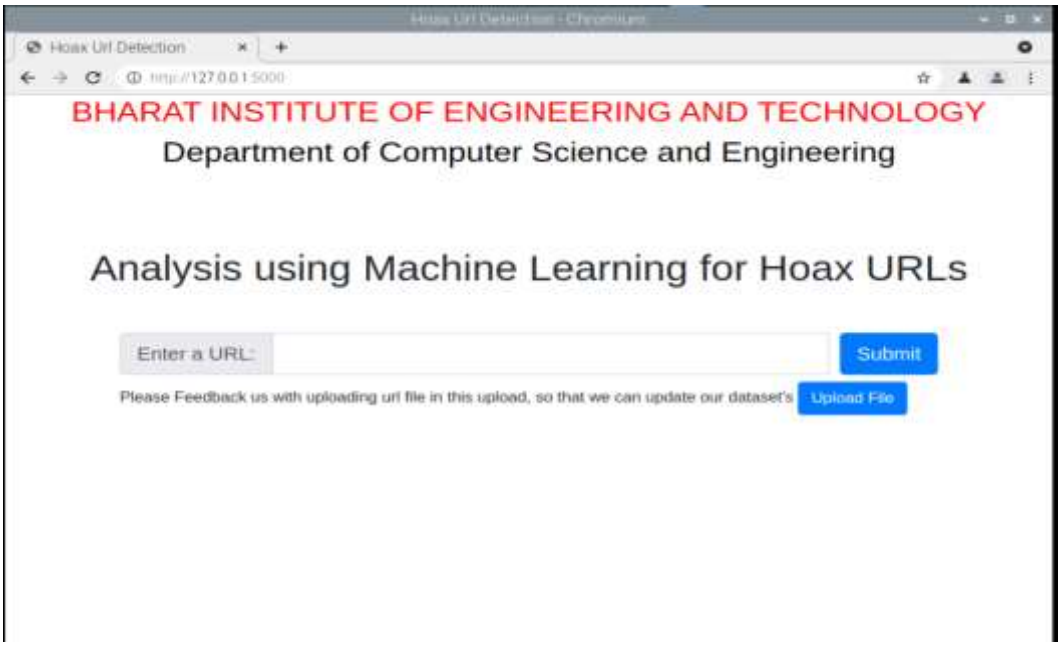


Fig 14.2: Screenshot of initial web page

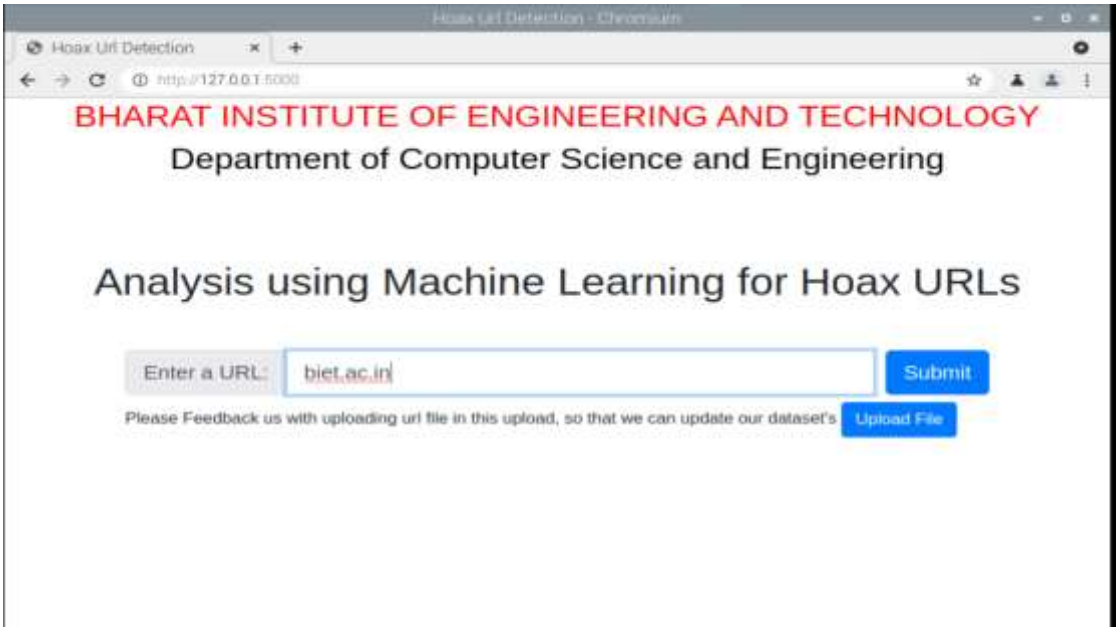


Fig 14.3: Screenshot of URL input page

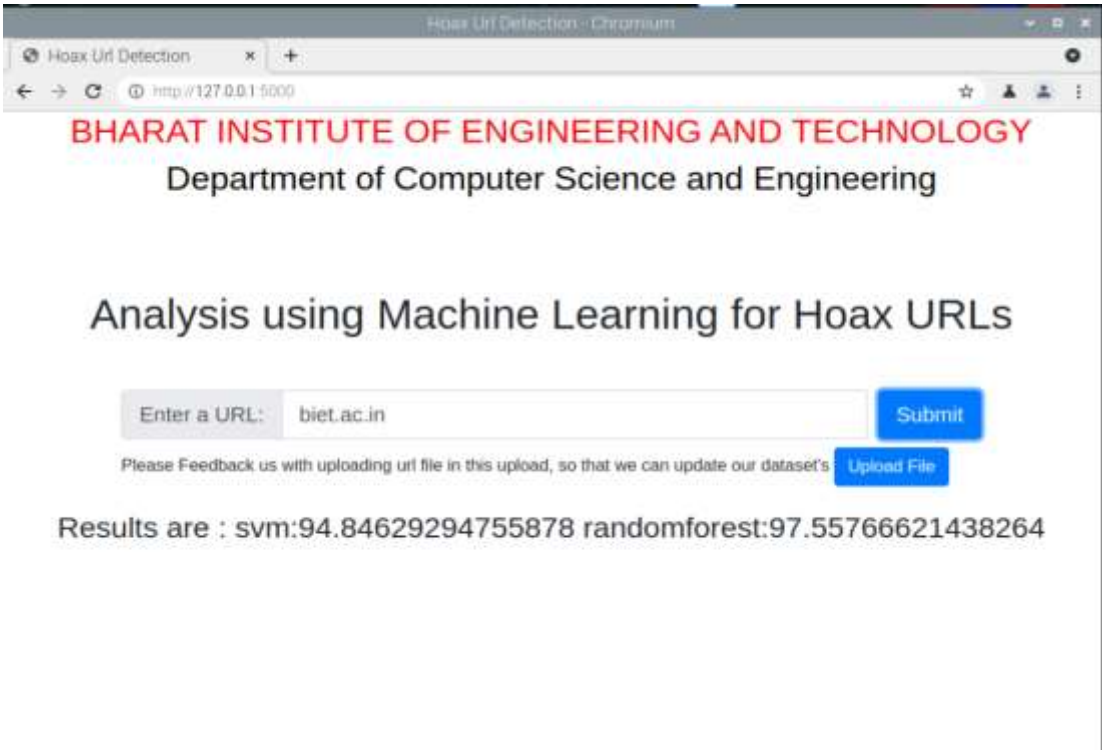


Figure 14.4: Screenshot of output page

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

15.CONCLUSION AND FUTURE SCOPE

15.1 Conclusion:

This work focus on the problem of detecting malicious URLs based on the information obtained from URLs string without the need to download page content. We presented and described a variety of features that can be extracted to represent a URL effectively. Then we explained the problem of appropriate representation of features and their possible necessary normalization. The user by giving input to the application will obtain the results. By this the threats will be decreased up to some extent. Day by day the data is generating which results indirectly in the creation of fraud things. An efficient result is obtained after training with machine learning algorithms.

This Project Work addressed the Program Outcomes (POs): PO1, PO2, PO3, PO5, PO6, PO7, PO11 and Program Specific Outcomes (PSOs): PSO1 & POS3 . These Program Outcomes (POs) and Program Specific Outcomes (PSOs) are attained by demonstrating the working model of the project.

15.2 Future Scope:

There will be improvement in the prediction rate. By using ensembling technique we can use variety of algorithms to increase the accuracy rate. And also processing rate will be enhanced. Database will be linked to the application so that the user can update the database frequently. This feature makes up to date function. A dynamic method is used to extract in a simple way to provide better results.

16.References

[1] Shantanu, B Janet, R Joshua Arul Kumar, Malicious URL Detection: A Comparative Study https://ieeexplore.ieee.org/document/9396014

[2] Dhanalakshmi Ranganayakulu and C. Chellappan, Detecting Malicious URLs in E-mail – An Implementation, AASRI Procedia, vol. 4, pp. 125-131, 2013, https://doi.org/10.1016/j.aasri.2013.10.020.

[3] Fuqiang Yu, Malicious URL Detection Algorithm based on BM Pattern Matching, International Journal of Security and Its Applications, vol. 9, pp. 33-44. https://www.researchgate.net/publication/283849287_Malicious_URL_Detection_Algorithm_based_on_BM_Pattern_Matching

[4] Frank Vanhoenshoven, Gonzalo Napoles, Rafael Falcon, Koen Vanhoof and Mario Koppen, Detecting Malicious URLs using Machine Learning Techniques, vol. 97. https://www.researchgate.net/publication/311583202_Detecting_Malicious_URLs_Using_Machine_Learning_Techniques

[5] Doyen Sahoo, Chenghao lua and Steven C. H. Hoi, Malicious URL Detection using Machine Learning: A Survey, arXiv:1701.07179v3, Aug 2019. Doyen Sahoo, Chenghao lua and Steven C. H. Hoi, "Malicious URL Detection using Machine Learning: A Survey", arXiv:1701.07179v3, Aug 2019.

[6] F. Vanhoenshoven, G. Nápoles, R. Falcon, K. Vanhoof and M. Köppen, "Detecting malicious URLs using machine learning techniques", 2016 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1-8, 2016. https://www.semanticscholar.org/paper/Detecting-malicious-URLs-using-machine-learning-Vanhoenshoven-N%C3%A1poles/5acc6b0e128ae8c1026fa85b79f328d74e18aaba

[7] Mohammed Abutaha, Mohammad Ababneh, Khaled Mahmoud, URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis, https://ieeexplore.ieee.org/document/9464539/authors#authors

[8] G. J. W. Kathrine, P. M. Praise, A. A. Rose and E. C. Kalaivani, "Variants of phishing attacks and their detection techniques", 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) Tirunelveli India 2019, pp. 255-259, 2019. https://ieeexplore.ieee.org/document/9464539/references#references

[9] A. Desai, J. Jatakia, R. Naik and N. Raul, "Malicious web content detection using machine leaning", 2017 2nd IEEE International Conference on Recent Trends in Electronics Information & Communication Technology (RTEICT), pp. 14321436, 2017.

[10] S. Parekh, D. Parikh, S. Kotak and S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection", 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 949952, 2018. https://www.semanticscholar.org/paper/A-New-Method-for-Detection-of-Phishing-Websites%3A-Parekh-Parikh/b8b2cb52c8600fecb5ec661eaab712f86b453b0b

[10] A.K. Jain and B.B. Gupta, "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning" in Cyber Security. Advances in Intelligent Systems and Computing, Singapore:Springer, vol. 729, 2018. https://www.springerprofessional.de/en/phish-safe-url-features-based-phishing-detection-system-using-ma/15726312

[11] Y. Sönmez, T. Tuncer, H. Gökal and E. Avc, "Phishing web sites features classification based on extreme learning machine", 2018 6th International Symposium on Digital Forensic and Security (ISDFS), pp. 1-5, 2018. https://www.researchgate.net/publication/325002990_Phishing_web_sites_features_classification_based_on_extreme_learning_machine

[12] Jair Cervantes et al., "A comprehensive survey on support vector machine classification: Applications challenges and trends", Neurocomputing, vol. 408, pp. 189-215, 2020.

[13] Happy Chapla, Riddhi Kotak, Mittal Joiser, A Machine Learning Approach for URL Based Web Phishing Using Fuzzy Logic as Classifier. https://ieeexplore.ieee.org/document/9002145/