A

FIELD BASED PROJECT REPORT

on

CYBERSECURITY AWARNESS IN ONLINE EDUCATION A CASE STUDY ANALYSIS

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

Submitted by

(BATCH : 05)

Y.Sri charitha 227Y1A12B7

U.Anjana 227Y1A1270

Under the Guidance

of

Mr.Y.APPA RAO

Associate Professor

DEPARTMENT OF INFORMATION TECHNOLOGY

MARRI LAXMAN REDDY INSTITUTE OF TECHNOLOGY AND MANAGEMENT

(AUTONOMOUS)

JUNE 2024

EMPOWERING THROUGH INNOVATIONS
MLRS

MARRI LAXMAN REDDY
INSTITUTE OF TECHNOLOGY AND MANAGEMENT
(AN AUTONOMOUS INSTITUTION)
(Approved by AICTE, New Delhi & Affiliated to JNTUH, Hyderabad)
Accredited by NBA and NAAC with 'A' Grade & Recognized Under Section2(f) & 12(B)of the UGC act,1956

CERTIFICATE

This is to certify that the field based project report titled **"Cyber security awarness in online education a case study analysis"** is being submitted by **Y.Sricharitha(227Y1A12B7), U.Anjana(227Y1A1270)** in **II B.Tech II Semeste**r **Information Technology** is a record bonafide work carried out by him. The results embodied in this report have not been submitted to any other University for the award of any degree.

Internal Guide **HOD**
(Mr .Y.APPARAO) (Dr.M.NAGA LAXMI)

Principal
(Dr.R.MURALI PRASAD)

MARRI LAXMAN REDDY
INSTITUTE OF TECHNOLOGY AND MANAGEMENT
(AN AUTONOMOUS INSTITUTION)
(Approved by AICTE, New Delhi & Affiliated to JNTUH, Hyderabad)
Accredited by NBA and NAAC with 'A' Grade & Recognized Under Section2(f) & 12(B)of the UGC act,1956

DECLARATION

I here by declare that the Field Based Project Report entitled, **"Cyber security awarness in online education a case study analysis"** submitted for the B.Tech degree is entirely my work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree.

Date: 29-06-2024

Y.Sricharitha
(207Y1A12B7)
U.Anjana
(227Y1A1270)

MARRI LAXMAN REDDY
INSTITUTE OF TECHNOLOGY AND MANAGEMENT
(AN AUTONOMOUS INSTITUTION)
(Approved by AICTE, New Delhi & Affiliated to JNTUH, Hyderabad)
Accredited by NBA and NAAC with 'A' Grade & Recognized Under Section2(f) & 12(B)of the UGC act,1956

ACKNOWLEDGEMENT

I am happy to express my deep sense of gratitude to the principal of the college

Dr.R.Murali Prasad, Professor, Marri Laxman Reddy Institute of Technology & Management, for having provided me with adequate facilities to pursue my project.

I would like to thank **Dr.M.Nagalakshmi ,** Professor and Head, Department of Information Technology , Marri Laxman Reddy Institute of Technology & Management, for having provided the freedom to use all the facilities available in the department, especially the laboratories and the library.

I am very grateful to my project guide,**Mr.Y.AppaRao** Associate Prof., Department of Information Technology , Marri Laxman Reddy Institute of Technology & Management, for his extensive patience and guidance throughout my project work.

I sincerely thank my seniors and all the teaching and non-teaching staff of the Department of Information Technology for their timely suggestions, healthy criticism and motivation during the course of this work.

I would also like to thank my classmates for always being there whenever I needed help or moral support.With great respect and obedience, I thank my parents and brother who were the backbone behind my deeds.

Finally, I express my immense gratitude with pleasure to the other individuals who have either directly or indirectly contributed to my need at right time for the development and success of this

ABSTRACT

This study examines how well-versed Kyrgyz-Turkish Manas School students are in cybersecurity through online learning. An first-year, second-year, and PhD sample of 517 students from all of the university's faculties participated in the study. Despite the fact that there are a great deal of cyberattacks happening all over the world, our research reveals that the pupils had no knowledge of cybersecurity or the overall effects of assaults. By focusing on issues pertaining to harmful software, safe passwords, and social media security, an analysis of understanding of cybersecurity was conducted. It has been discovered that students' knowledge of cybersecurity is lacking, despite the fact that we live in a technologically advanced age where every aspect of our lives is connected to the net through distance learning. More study has led to the conclusion that security instruction should be provided to students in order to shield them from hacks and improve their use of the web.

TABLE OF CONTENTS

LIST OF FIGURES

1.INTRODUCTION

With the spread of technology and the penetration of the internet into every aspect of daily life, cyber security has begun to be of great importance for both individuals and states alike [1]. Although these innovations have made our lives easier, the increase in cyber attacks has made it necessary to take measures in this area [2]_[4]. In addition, one of the most basic points is that the types of cyber attack, in other words the malicious use of cyberspace, have changed in the last 20 years. This has led to the use of new ``cyber'' concepts and risks in the literature.

A cyber attack is defined by Hathaway *et al.* as follows: ``A cyber-attack consists of any action taken to undermine the functions of a computer network for a political or national security purpose'' [6]. The most basic question to ask is `Does this definition define cyber attacks today?' Today, saying that cyberattacks are carried out only for political purposes is insufficient when it comes to trying to understand the nature of cyber attacks. This is because new cyber concepts have emerged that have changed the nature of cyber attacks. What remains similar is the use of computers in attacks. In this context, cybercrimes are defined as crimes committed through computers [7]. The Department of Justice of the USA defines a cybercrime as ``any violations of criminal law that involve knowledge of computer technology for their perpetration, investigation or prosecution'' [8]. On the one hand, it is important to explain what cyber security is. Although the concept does not have any common definition, the International Telecommunication Union (ITU) defines cyber security as ``the collection of tools, policies, security concepts, security safeguards, guidelines, risk management approaches, actions, training, best practices, assurance and technologies that can be used to protect the cyber environment and organization and user's assets. Cyber security strives to ensure the attainment and maintenance of the security properties of the organization and user's assets against relevant security risks in the cyber environment'' [9]. Although there are now more complex structures in cyber attacks and cyber security compared to the past, the ability to perform cyber attacks has developed. The capacity to learn through websites that almost every computer user can access has increased. This is especially so the new generation, called the Z generation. They are often completely involved with computer technologies and can easily perform any activity they want by using it [10]_[13].On the other hand, this situation has also led to the emergence of new situations regarding computer technologies, or cyber security awareness as it is called in the literature. Although the Z generation has grown up with the internet and with computer technologies, sometimes they do not know what kind of problems they may encounter or they do not know how to deal with the problems arising from the continual development of internet technologies [14], [15].

Cyber security awareness, or information security awareness, has become an important issue today. The number of studies on this subject, which affects every aspect of daily life, is increasing. First of all, defining cyber security awareness is important to better gain a full understanding of the subject. Shaw *et al.* defined the concept as; ``the degree of understanding of users about the importance of information security and their responsibilities and acts to exercise sufficient levels of information security control to protect the organization's data and networks'' [16]. As can be understood from the definition, the important points are evaluated in two ways. Firstly, it emphasizes the importance and responsibilities to do with information security. Secondly, it is aimed at knowing and applying information security control practices at an adequate level to protect the information.

Hwang *et al.* defined information security awareness as a phenomenon that aims to enable users to recognize the security vulnerabilities or problems that may arise and to respond in an appropriate way. Naturally, it also intends to keep the security phenomenon on the internet at the forefront of the user's minds [17]. Khan *et al.* made similar points to Hwang. Khan *et al.* defined information security awareness as the fact that users have information about security and act within the framework of the known rules [18]. Zilka, on the other hand, defines cyber security awareness as a phenomenon that aims to increase the level of knowledge about the online applications that users use so then they can stay safe in response to online risks [19]. Within the framework of these definitions, it can be clearly seen that security awareness training should be provided to improve cyber security awareness [20]. Within the framework of this information, it will also be questioned what kind of information the students have about cyber security awareness during the online education period and whether they want to receive training in this direction. The second aim of this study is to obtain data for use by further studies on how students can increase their cyber security awareness based on the theoretical framework findings.

1.1.EXISTING SYSTEM

Several studies have been conducted to measure the level of cybersecurity awareness among students and academics. For example, Ismailova and Muhametjanova studied the cybercrime risk awareness in the Kyrgyz Republic with 172 participants. The results show that the students were not familiar with cybercrime.Another survey was done in New Zealand in 2016 to measure cybersecurity awareness among individuals between the ages of 8-21. This was conducted by Trimula, Sarrafzadeh, and Pang. According to the authors, most of the students were not aware of the presence of cyber threats and they did not know the term cybersecurity Ahmed *et al.* examined the cybersecurity awareness of the people of Bangladesh . Their research states that the sample did not have enough information about cybersecurity. The authors made a recommendation that a guide should be prepared so then people can become consciously aware of cybersecurity [26]. The Department of Computer Science at Yobe State University conducted a survey that showed that although the students were aware of cybersecurity, they did not know how to protect the data that they have.

Today, social media accounts are very popular among students. Sometimes people can be defrauded and their information stolen through their social media accounts. Kirwan *et al.* conducted a study on this subject involving Malaysian students. They investigated whether the sample of students knew about this subject and whether they had been the victim of this type of fraud [28]. The results of their survey showed that more than 30% of students had been a victim of a social networking site scam .

Senthilkumar and Sathiskumar surveyed cybersecurity awareness among college students in Tamil Nadu. They found that the students were able to protect themselves from cyber threats. Zwilling *et al.* conducted a survey among undergraduate and graduate students. The survey was conducted on students from various countries . The results revealed that internet users are aware of cyber risks and simple precautions are taken by them. The authors claimed that there is a link between cyber awareness and cyber knowledge .

Disadvantages :

- The system is not implemented SECURITY VULNERABILITIES AND CYBER THREATS.
- The system is not implemented AWARENESS OF CYBERCRIMES AND LAWS.

Advantages :

- Before measuring the level of awareness of an ordinary computer user about the risks of cyberattacks, it is important to determine whether they have basic security knowledge. For this reason, while creating the framework of this survey study, an attempt was made to understand whether the basis of possible unawareness in relation to the field of cybersecurity is a lack of knowledge.
- Since it is predicted that most of the participants are a population that uses passwords, has social media accounts, and installs various software on their computers, the questions were chosen in this direction

1.2.PROPOSED SYSTEM

Cybersecurity awareness, or information security awareness, has become an important issue today. The number of studies on this subject, which affects every aspect of daily life, is increasing. First of all, defining cybersecurity awareness is important to better gain a full understanding of the subject. Shaw *et al.* defined the concept as; ``the degree of understanding of users about the importance of information security and their responsibilities and acts to exercise sufficient levels of information security control to protect the organization's data and networks'' [16]. As can be understood from the definition, the important points are evaluated in two ways. Firstly, it emphasizes the importance and responsibilities to do with information security. Secondly, it is aimed at knowing and applying information security control practices at an adequate level to protect the information.

Hwang *et al.* defined information security awareness as a phenomenon that aims to enable users to recognize the security vulnerabilities or problems that may arise and to respond in an appropriate way. Naturally, it also intends to keep the security phenomenon on the internet at the forefront of the user's minds . Khan *et al.* made similar points to Hwang. Khan *et al.* defined information security awareness as the fact that users have information about security and act within the framework of the known rules . Zilka, on the other hand, defines cybersecurity awareness as a phenomenon that aims to increase the level of knowledge about the online applications that users use so then they can stay safe in response to online risks Within the framework of these definitions, it can be clearly seen that security awareness training should be provided to improve cybersecurity awareness .

1.3.Literature Survey

This study examines how well-versed Kyrgyz-Turkish Manas University students are in cybersecurity through online learning. An college, graduate, and PhD sample of 517 people from all of the university's faculties participated in the study. Despite the fact that there are a great deal of cyberattacks happening all over the world, our research reveals that the pupils had no interest in cybersecurity or the overall effects of cyberattacks. By focusing on issues pertaining to harmful software, safe passwords, and social media security, an analysis of understanding of cybersecurity was conducted. It has been discovered that students' knowledge of cybersecurity is lacking, despite the fact that we live in a technologically advanced age where every aspect of our lives is connected to the internet through distance learning. This study measures the level of cybersecurity knowledge attained by Kyrgyz-Turkish Manas college pupils through distance learning. A total of 517 college, master's, and PhD students from all of the university's faculties made up the sample for the study. Our study's findings demonstrate that students' lack of awareness about cybersecurity and the overall effects of cyberattacks is despite the fact that there are numerous cyberattacks taking place all over the world. By focusing on hazardous software, secure passwords, and social media safety questions about cybersecurity knowledge were put together. As COVID-19 took hold and groups and classrooms all over the world had to close their doors, online learning started to become the standard. As of April 2020, 186 countries and more than 1.2 billion students were affected by educational institution closures. The school system there has also changed as a result of this pandemic. Online learning is becoming more popular both academics and students because to free and distance learning platforms. The cognitive computing and human intelligence revolution is known as "education 5.0." The main obstacle that one will encounter as the world moves in this direction is how to deal with the risks involved with safety online in the digital age. In order to protect against hacker attacks online, information security awareness might be crucial. The main objective of this paper is to analyse the level of information security awareness, associated risks, and overall effect on the institutions among professors, researchers, freshmen, and employees in schools in the Middle East. The findings show that those polled lack the necessary knowledge and comprehension of the significance of information security concepts and how they are used in their daily job. The Web is becoming a more integral part of many people's, organisations', and countries' daily lives. It has, in large part, positively impacted how people speak. Additionally, it has set up novel business possibilities and given countries the chance to conduct their governments online. Cyber comes with a lot of risks even if it provides an unending array of services and opportunities.

SYSTEM STUDY

FEASIBILITY STUDY:

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are :

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY:

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY:

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY:

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

2. REQUIREMENTS

2.1.System Requirements:

<u>Hardware System Configuration</u>:-

➢ Processor	Pentium –IV
➢ RAM	4 GB (min)
➢ Hard Disk	20 GB
➢ Key Board	Standard Windows Keyboard
➢ Mouse	Two or Three Button Mouse
➢ Monitor	SVGA

2.2.SOFTWARE REQUIREMENTS:

➢ **Operating system**	Windows 7 Ultimate.
➢ **Coding Language**	Python.
➢ **Front-End**	Python.
➢ **Back-End**	Django-ORM
➢ **Designing**	Html, css, javascript.
➢ **Data Base**	MySQL (WAMP Server).

3.MODULES

Service Provider:

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Browse and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of Cyber security Awareness, View Cyber security Awareness Type Ratio, Download Trained Data Sets, View Cyber security Awareness Type Ratio Results, View All Remote Users.

View and Authorize Users:

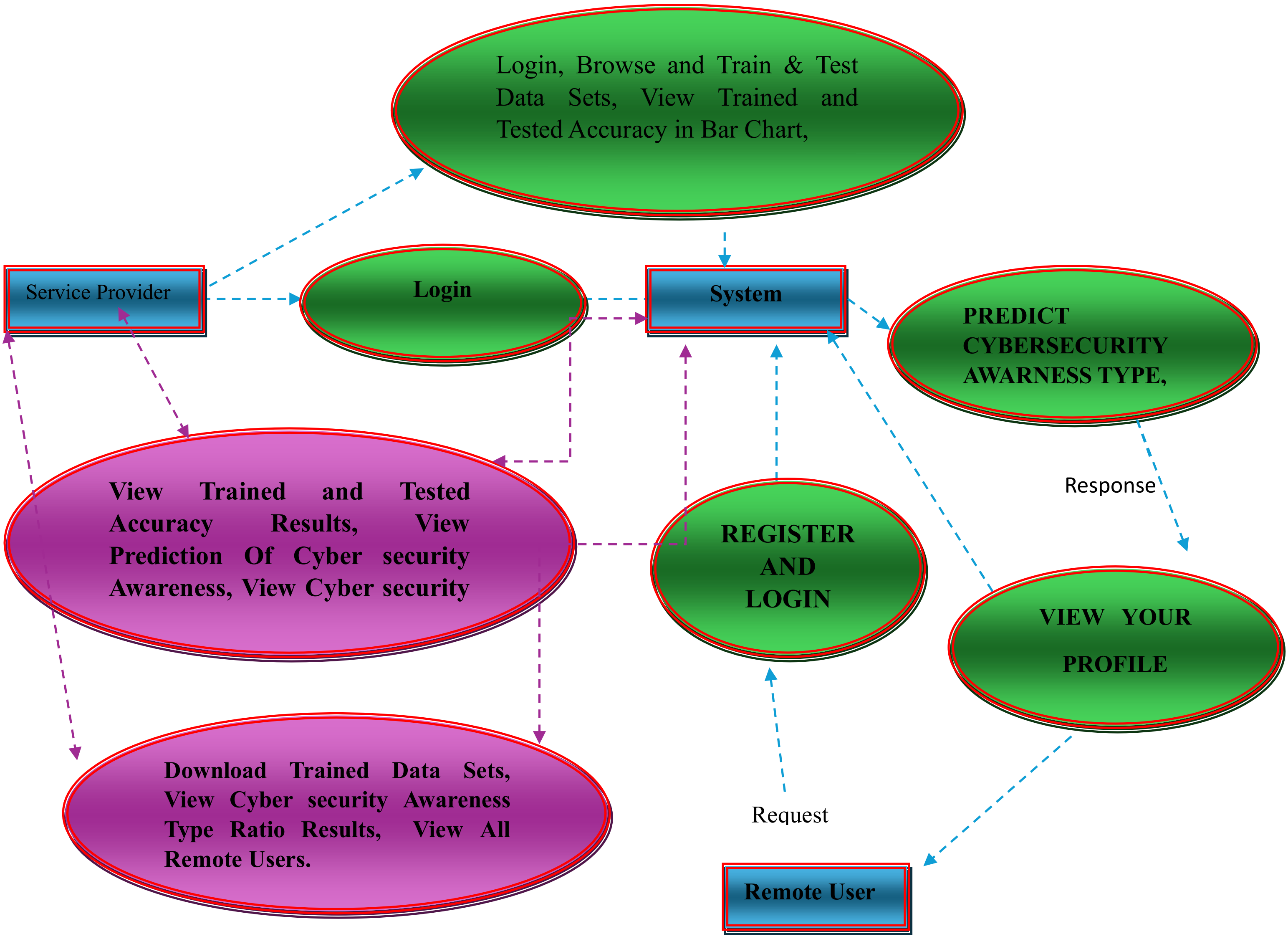
In this module, the admin can view the list of users who all registered. In this, the admin can view the user’s details such as, user name, email, address and admin authorizes the users.

Remote User:

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT CYBERSECURITY AWARNESS TYPE, VIEW YOUR PROFILE.

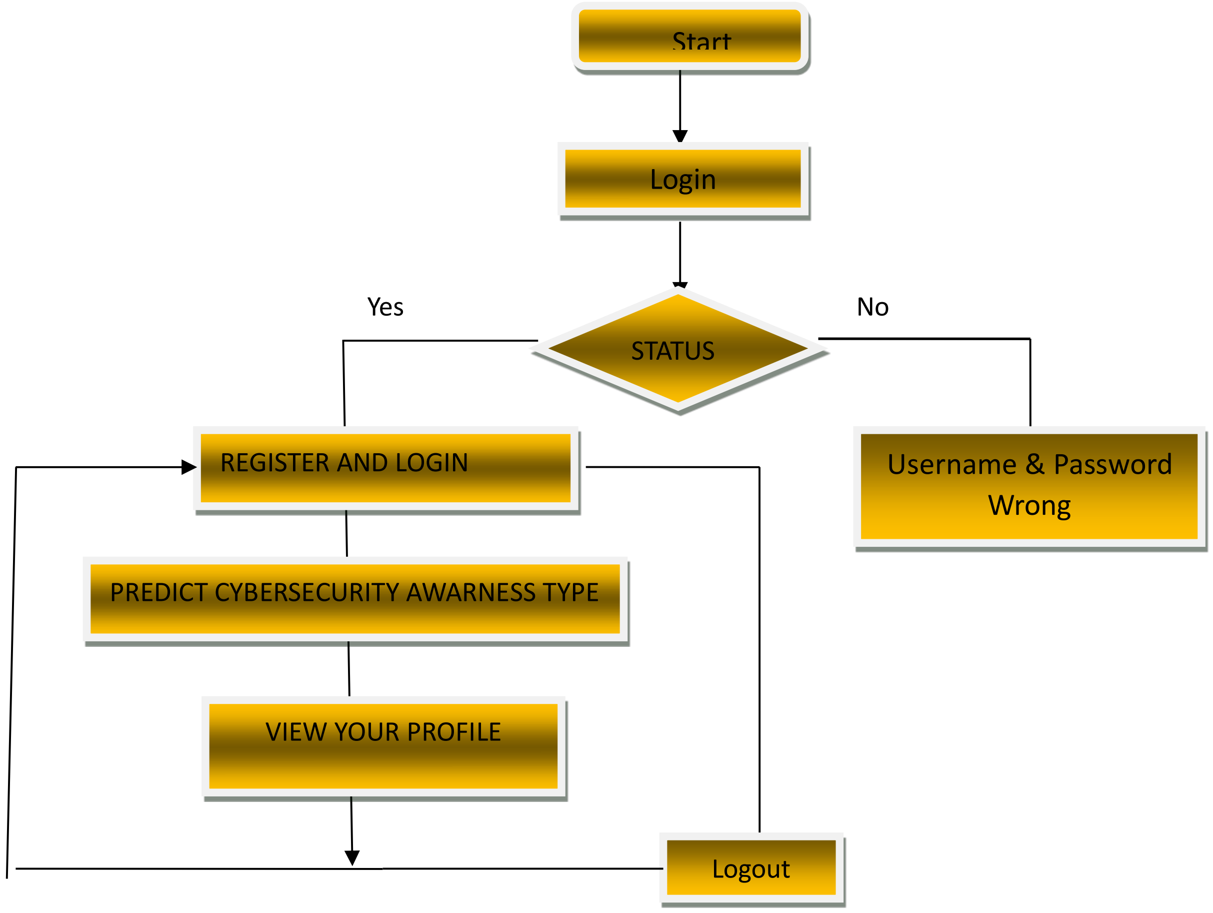
4.DESIGN

Data Flow Diagram:


Login, Browse and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart,
Service Provider
Login
System
PREDICT CYBERSECURITY AWARNESS TYPE,
Response
View Trained and Tested Accuracy Results, View Prediction Of Cyber security Awareness, View Cyber security
REGISTER AND LOGIN
VIEW YOUR PROFILE
Download Trained Data Sets, View Cyber security Awareness Type Ratio Results, View All Remote Users.
Request
Remote User

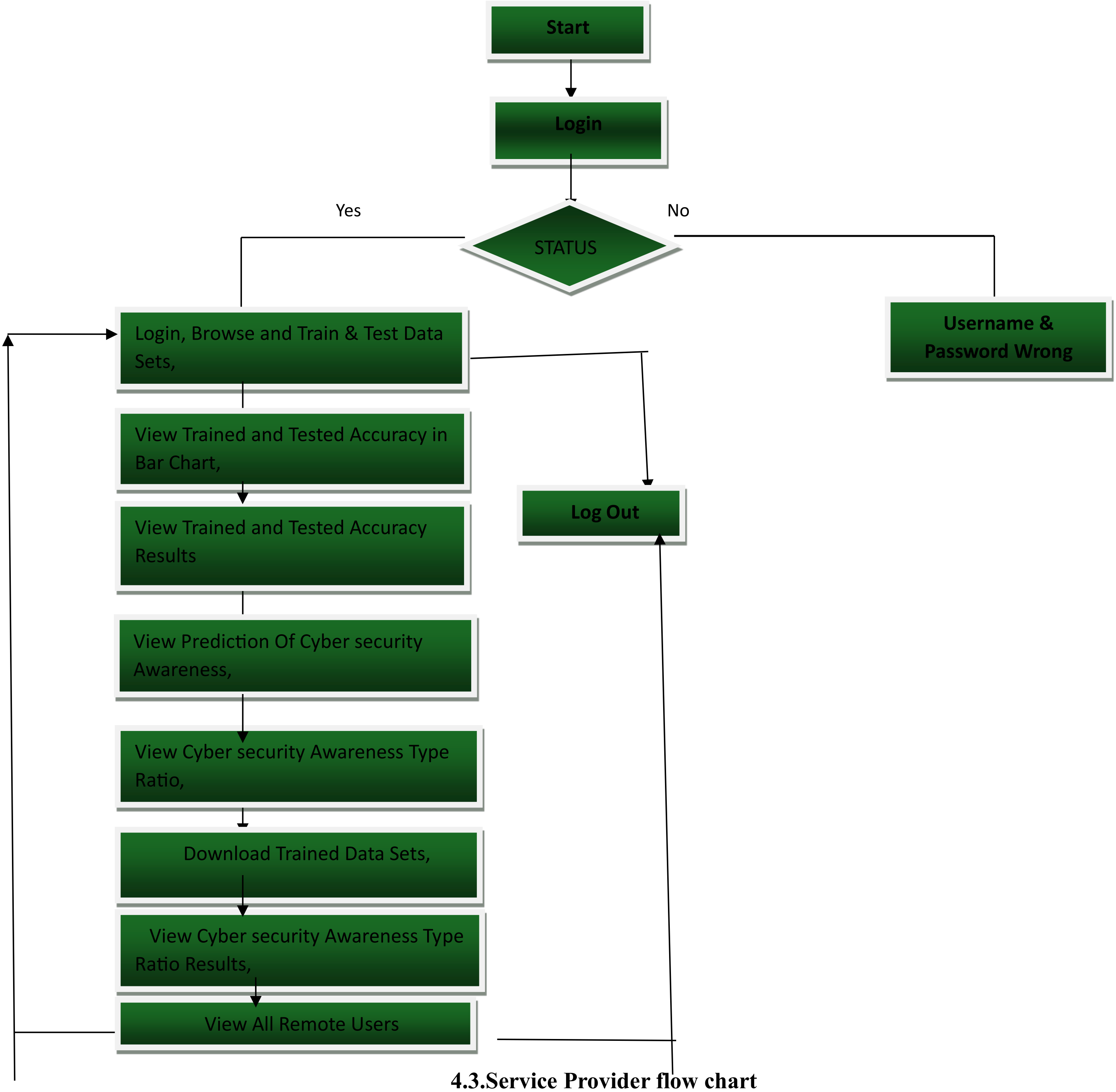
4.1.Data flow diagram

Flow Chart : Remote User



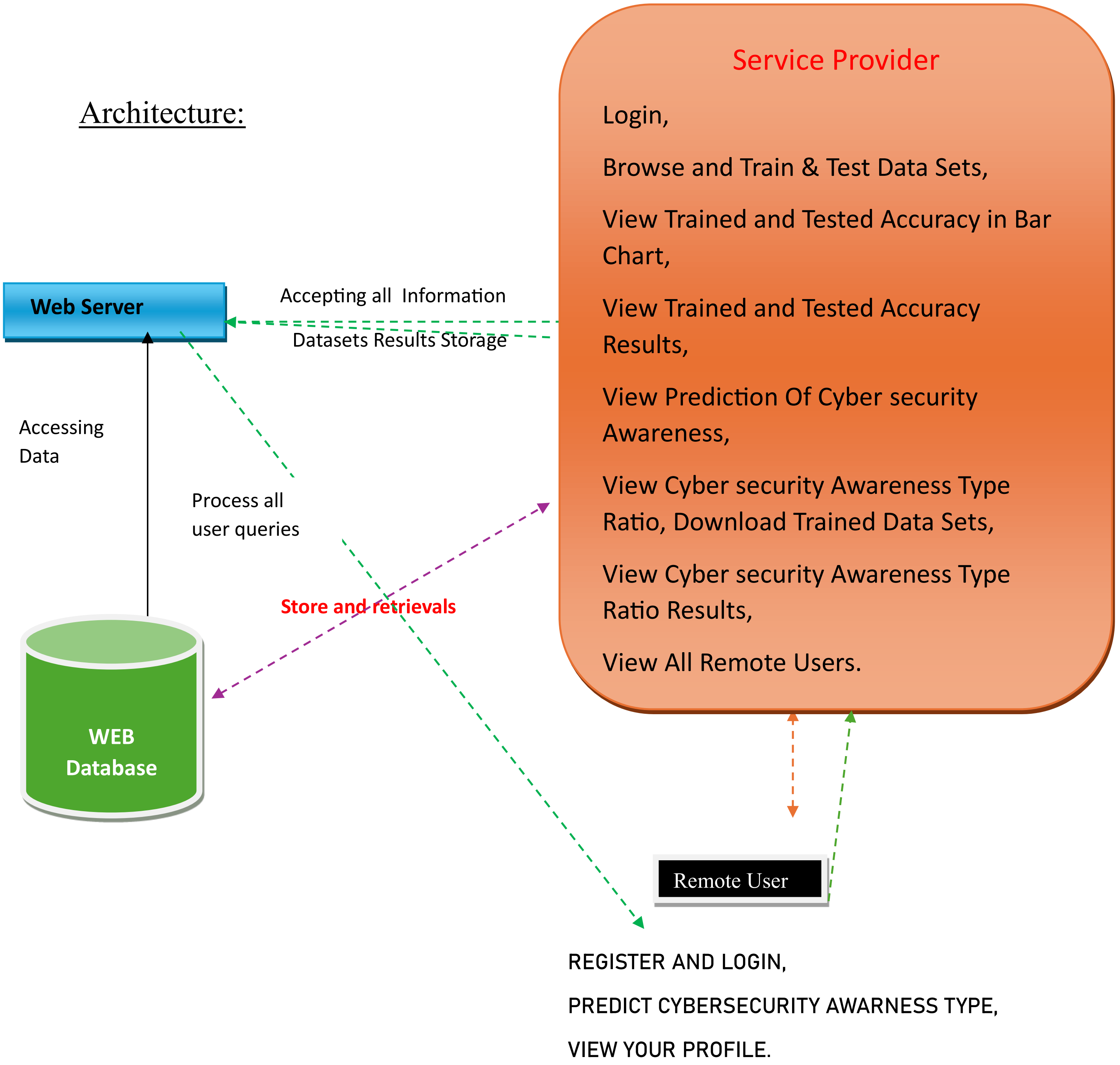
4.2.Remote user flowchart

Flow Chart : Service Provider


Start
Login
Yes
No
STATUS
Login, Browse and Train & Test Data Sets,
Username & Password Wrong
View Trained and Tested Accuracy in Bar Chart,
Log Out
View Trained and Tested Accuracy Results
View Prediction Of Cyber security Awareness,
View Cyber security Awareness Type Ratio,
Download Trained Data Sets,
View Cyber security Awareness Type Ratio Results,
View All Remote Users

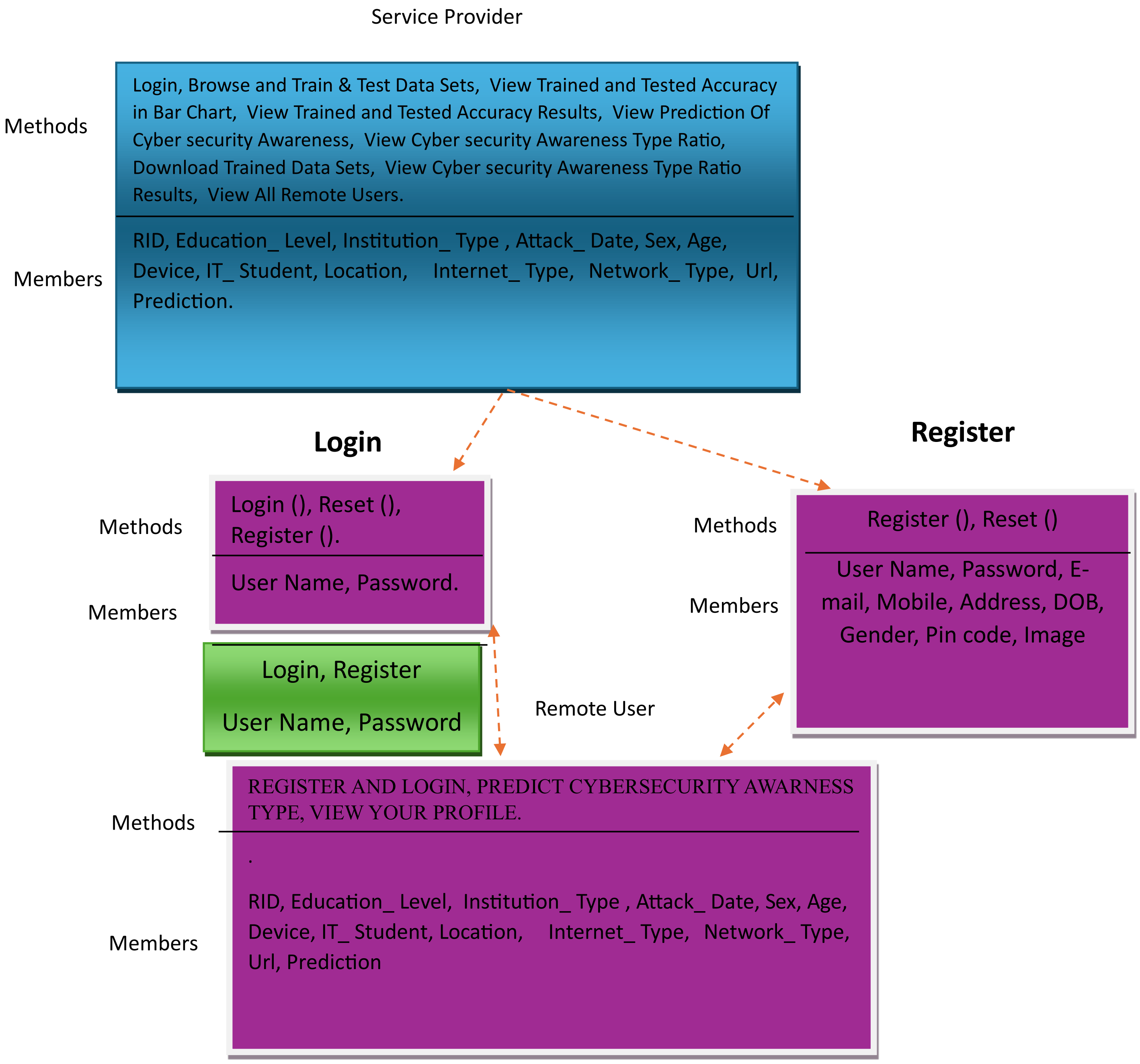
4.3.Service Provider flow chart

Architecture:


Service Provider
Login,
Browse and Train & Test Data Sets,
View Trained and Tested Accuracy in Bar Chart,
View Trained and Tested Accuracy Results,
View Prediction Of Cyber security Awareness,
View Cyber security Awareness Type Ratio, Download Trained Data Sets,
View Cyber security Awareness Type Ratio Results,
View All Remote Users.
Web Server
Accepting all Information
Datasets Results Storage
Accessing Data
Process all user queries
Store and retrievals
WEB Database
Remote User
REGISTER AND LOGIN,
PREDICT CYBERSECURITY AWARNESS TYPE,
VIEW YOUR PROFILE.

4.4.Architecture

Class Diagram :



Service Provider

Methods

Login, Browse and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of Cyber security Awareness, View Cyber security Awareness Type Ratio, Download Trained Data Sets, View Cyber security Awareness Type Ratio Results, View All Remote Users.

Members

RID, Education_ Level, Institution_ Type , Attack_ Date, Sex, Age, Device, IT_ Student, Location, Internet_ Type, Network_ Type, Url, Prediction.

Login

Methods

Login (), Reset (), Register ().

Members

User Name, Password.

Login, Register

User Name, Password

Register

Methods

Register (), Reset ()

Members

User Name, Password, E-mail, Mobile, Address, DOB, Gender, Pin code, Image

Remote User

Methods

REGISTER AND LOGIN, PREDICT CYBERSECURITY AWARNESS TYPE, VIEW YOUR PROFILE.

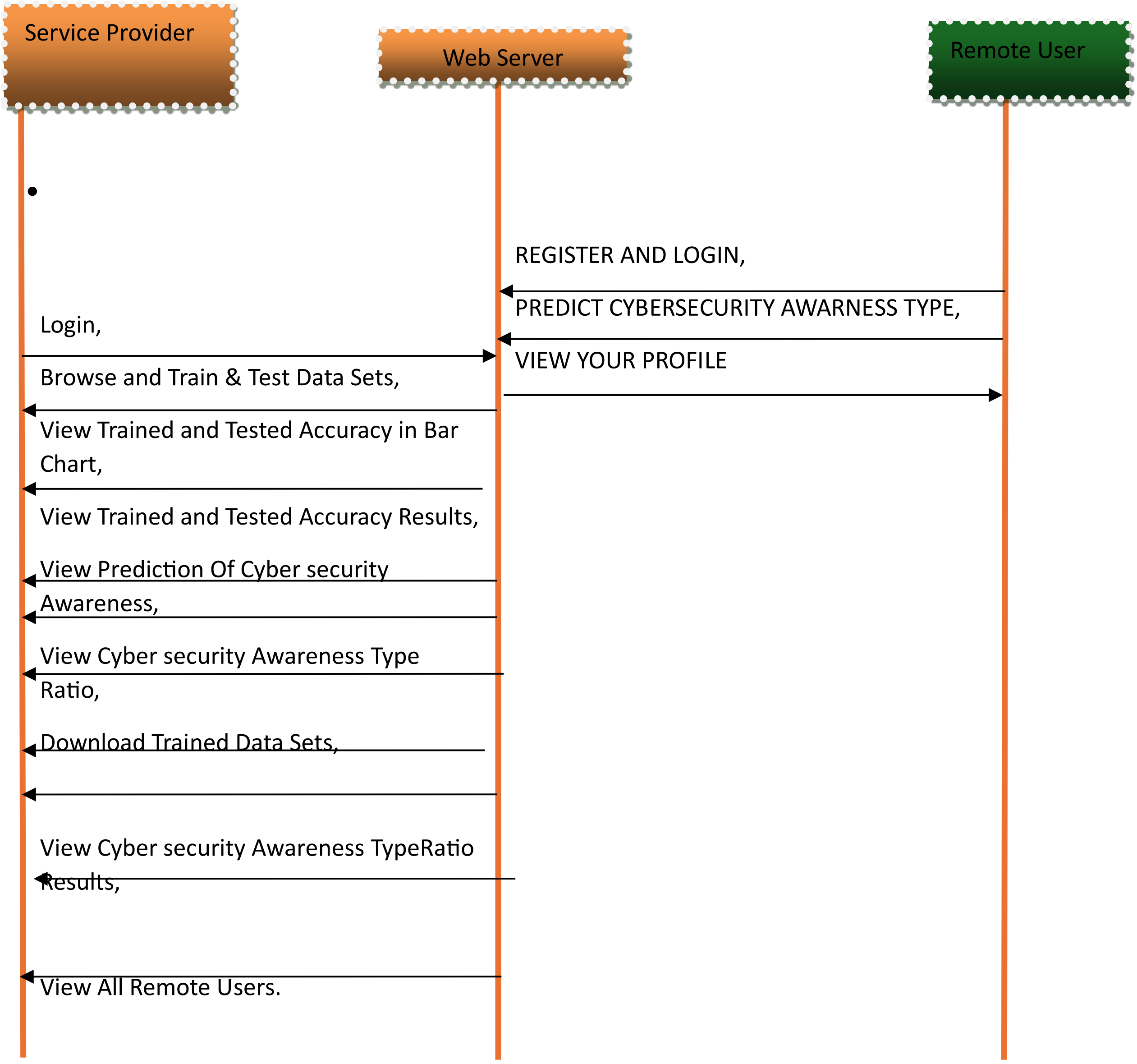
.

Members

RID, Education_ Level, Institution_ Type , Attack_ Date, Sex, Age, Device, IT_ Student, Location, Internet_ Type, Network_ Type, Url, Prediction

4.5.Class diagram

Sequence Diagram :


Service Provider
Web Server
Remote User
REGISTER AND LOGIN,
PREDICT CYBERSECURITY AWARNESS TYPE,
VIEW YOUR PROFILE
Login,
Browse and Train & Test Data Sets,
View Trained and Tested Accuracy in Bar Chart,
View Trained and Tested Accuracy Results,
View Prediction Of Cyber security Awareness,
View Cyber security Awareness Type Ratio,
Download Trained Data Sets,
View Cyber security Awareness TypeRatio Results,
View All Remote Users.

4.6.Sequence diagram

5.SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS:

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.

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.

Functional test:

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 invoked.

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 tests is determined.

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.

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 is used to test areas that cannot be reached from a black box level.

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.

Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach:

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

Test objectives:

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not 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.

Integration Testing:

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

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

Acceptance Testing:

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

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

SYSTEM TESTING

TESTING METHODOLOGIES

The following are the Testing Methodologies:

- **Unit Testing.**
- **Integration Testing.**
- **User Acceptance Testing.**
- **Output Testing.**
- **Validation Testing.**

Unit Testing:

Unit testing focuses verification effort on the smallest unit of Software design that is the module. Unit testing exercises specific paths in a module's control structure to ensure complete coverage and maximum error detection. This test focuses on each module individually, ensuring that it functions properly as a unit. Hence, the naming is Unit Testing.

During this testing, each module is tested individually and the module interfaces are verified for the consistency with design specification. All important processing path are tested for the expected results. All error handling paths are also tested.

Integration Testing:

Integration testing addresses the issues associated with the dual problems of verification and program construction. After the software has been integrated a set of high order tests are conducted. The main objective in this testing process is to take unit tested modules and builds a program structure that has been dictated by design.

The following are the types of Integration Testing:

1.Top Down Integration

This method is an incremental approach to the construction of program structure. Modules are integrated by moving downward through the control hierarchy, beginning with the main program module. The module subordinates to the main program module are incorporated into the structure in either a depth first or breadth first manner. In this method, the software is tested from main module and individual stubs are replaced when the test proceeds downwards.

2.Bottom-up Integration

This method begins the construction and testing with the modules at the lowest level in the program structure. Since the modules are integrated from the bottom up, processing required for modules subordinate to a given level is always available and the need for stubs is eliminated. The bottom up integration strategy may be implemented with the following steps:

- The low-level modules are combined into clusters into clusters that perform a specific Software sub-function.
- A driver (i.e.) the control program for testing is written to coordinate test case input and output.
- The cluster is tested.
- Drivers are removed and clusters are combined moving upward in the program structure.

The bottom up approaches tests each module individually and then each module is module is integrated with a main module and tested for functionality.

One method involves removing shallow characteristics from URLs. The other two, on the other hand, construct accurate feature representations of URLs and use shallow features to evaluate URL legitimacy. The final output of our system is calculated by adding the outputs of all divisions. Extensive testing on a dataset collected from the Internet indicate that our system can compete with other cutting-edge detection methods. while consuming a fair amount of time. For phishing detection, Vahid Shahrivari et al. used machine learning approaches. They used the logistic regression classification method, KNN, Adaboost algorithm, SVM, ANN and random forest. They found random forest algorithm provided good accuracy. Dr.G. Ravi Kumar used a variety of machine learning methods to detect phishing assaults. For improved results, they used NLP tools. They were able to achieve high accuracy using a Support Vector Machine and data that had been pre-processed using NLP approaches. Amani Alswailem et al. tried different machine learning model for phishing detection but was able to achieve more accuracy in random forest. Hossein et al. created the “Fresh-Phish” open-source framework. This system can be used to build machine-learning data for phishing websites. They used a smaller feature set and built the query in Python. They create a big, labelled dataset and test several machine-learning classifiers on it. Using machine-learning classifiers, this analysis yields very high accuracy. These studies look at how long it takes to train a model. X. Zhang suggested a phishing detection model based on mining the semantic characteristics of word embedding, semantic feature, and multi-scale statistical features in Chinese web pages to detect phishing performance successfully. To obtain statistical aspects of web pages, eleven features were retrieved and divided into five classes. To obtain statistical aspects of web pages, eleven features were retrieved and divided into five classes. To learn and evaluate the model, AdaBoost, Bagging, Random Forest, and SMO are utilized. The legitimate URLs dataset came from DirectIndustry online guides, and the phishing data came from China's Anti-Phishing Alliance. With novel methodologies, M. Aydin approaches a framework for extracting characteristics that is versatile and straightforward. Phish Tank provides data, and Google provides authentic URLs. C# programming and R programming were utilized to btain the text attributes. The dataset and third-party service providers yielded a total of 133 features. The feature selection approaches of CFS subset based and Consistency subset-based feature selection were employed and examined with the WEKA tool. The performance of the Nave Bayes and Sequential Minimal Optimization (SMO) algorithms was evaluated, and the author prefers SMO to NB for phishing detection.

6. METHODOLOGY

A phishing website is a social engineering technique that imitates legitimate webpages and uniform resource locators (URLs). The Uniform Resource Locator (URL) is the most common way for phishing assaults to occur. Phisher has complete control over the URL's sub-domains. The phisher can alter the URL because it contains file components and directories.This research used the linear-sequential model, often known as the waterfall model. Although the waterfall approach is considered conventional, it works best in instances where there are few requirements. The application was divided into smaller components that were built using frameworks and hand-written code.

6.1 RESEARCH FRAMEWORK

The steps of this research in which some selected publications were read to determine the research gap and, as a result, the research challenge was defined. Feature selection, classification and phishing website detection were all given significant consideration. It's worth noting that most phishing detection researchers rely on datasets they've created. However, because the datasets utilized were not available online for those who use and check their results, it is difficult to assess and compare the performance of a model with other models. As a result, such results cannot be generalized. For the preparation of this dissertation, I used Python as the primary language. Python is a language that is heavily focused on machine learning. It includes several machine learning libraries that may be utilized straight from an import. Python is commonly used by developers all around the world to deal with machine learning because of its extensive library of machine learning libraries. Python has a strong community, and as a result, new features are added with each release.

Data Collection:

The phishing URLs were gathered using the open source tool Phish Tank. This site provides a set of phishing URLs in a variety of forms, including csv, json, and others, which are updated hourly. This dataset is used to train machine learning models with 5000 random phishing URLs.

Data Cleaning

Fill in missing numbers, smooth out creaking data, detect and delete outliers, and repair anomalies to clean up the data.

Data Pre-processing

Data preprocessing is a cleaning operation that converts unstructured raw data into a neat, well-structured dataset that may be used for further research. Data preprocessing is a cleaning operation that transforms unstructured raw data into well-structured and neat dataset which can be used for further research.

The data is split into 8000 training samples and 2000 testing samples, before the ML model is trained. It is evident from the dataset that this is a supervised machine learning problem. Classification and regression are the two main types of supervised machine learning issues. Because the input URL is classed as legitimate or phishing, this data set has a classification problem. The following supervised machine learning models were examined for this project's dataset training: Decision Tree , Multilayer Perceptron, Random Forest,Autoencoder Neural Network , XGBoost and Support Vector Machines.

6.2 ADDRESS BASED CHECKING

Below are the categories been extracted from address based

1. Domain of the URL :Where domain which is present in the URL been extracted

2. IP Address in the URL :The presence of an IP address in the URL is checked. Instead of a domain name, URLs may contain an IP address. If an IP address is used instead of a domain name in a URL, we can be certain that the URL is being used to collect sensitive information.

3. "@" Symbol in URL :The presence of the'@' symbol in the URL is checked. When the “@” symbol is used in a URL, the browser ignores anything before the “@” symbol, and the genuine address is commonly found after the “@” symbol.

4. Length of URL :Calculates the URL's length. Phishers can disguise the suspicious element of a URL in the address bar by using a lengthy URL. If the length of the URL is larger than or equal to 54 characters, the URL is classed as phishing in this project.

5. Depth of URL :Calculates the URL's depth. Based on the'/', this feature determines the number of subpages in the given address.

6. Redirection "//" in URL :The existence of"//" in the URL is checked. The presence of the character"//" in the URL route indicates that the user will be redirected to another website. The position of the"//" in the URL is calculated. We discovered that if the URL begins with “HTTP,” the “//” should be placed in the sixth position. If the URL uses “HTTPS,” however, the “//” should occur in the seventh place.

7. Http/Https in Domain name :The existence of "http/https" in the domain part of the URL is checked. To deceive users, phishers may append the “HTTPS” token to the domain section of a URL.

8. Using URL Shortening Services :URL shortening is a means of reducing the length of a URL while still directing to the desired webpage on the "World Wide Web." This is performed by using a “HTTP Redirect” on a short domain name that points to a webpage with a long URL.

9. Prefix or Suffix "-" in Domain :Checking for the presence of a '-' in the URL's domain part. In genuine URLs, the dash symbol is rarely used. Phishers frequently append prefixes or suffixes to domain names, separated by (-), to give the impression that they are dealing with a legitimate website.

6.3 DOMAIN BASED CHECKING

This category contains a lot of features that can be extracted. This category contains a lot of features that can be extracted. The following were considered for this project out of all of them.

DNS Record :In the case of phishing websites, the WHOIS database either does not recognize the stated identity or there are no records for the host name .

Web Traffic :This function determines the number of visitors and the number of pages they visit to determine the popularity of the website. In the worst-case circumstances, legitimate websites placed among the top100,000, according to our data. Furthermore, it is categorized as "Phishing" if the domain has no traffic or is not recognized by the Alexa database.

- Age of Domain :This information can be retrieved from the WHOIS database. Most phishing websites are only active for a short time. For this project, the minimum age of a legal domain is deemed to be 12 months. Age is simply the difference between the time of creation and the time of expiry.
- End Period of Domain :This information can be gleaned from the WHOIS database. The remaining domain time is calculated for this feature by determining the difference between the expiry time and the current time. For this project, the valid domain's end time is regarded to be 6 months or fewer.

6.4 HTML AND JAVASCRIPT BASED CHECKING

Many elements that fall within this group can be extracted. The following were considered for this project out of all of them.

IFrame Redirection : IFrame is an HTML tag that allows you to insert another webpage into the one you're now viewing. The “iframe” tag can be used by phishers to make the frame invisible, i.e., without frame borders. Phishers employ the “frame border” attribute in this case, which causes the browser to create a visual boundary.

Status Bar Customization : Phishers may utilize JavaScript to trick visitors into seeing a false URL in the status bar. To get this feature, we'll need to delve into the webpage source code, specifically the "on Mouseover" event, and see if it alters the status bar.

Disabling Right Click : Phishers disable the right-click function with JavaScript, preventing users from viewing and saving the webpage source code. This functionality is handled in the same way as "Hiding the Link with on Mouseover." Nonetheless,we'll look for the ``event"event. button==2" in the webpage source code and see if the right click is disabled for this functionality.

Website Forwarding : The number of times a website has been redirected is a narrow line that separates phishing websites from authentic ones. We discovered that authentic websites were only routed once in our sample. Phishing websites with this functionality, on the other hand, have been redirected at least four times.

Implementation : We'll examine the implementation component of our artefact in this area of the report, with a focus on the description of the developed solution. This is a task that requires supervised machine learning.

List-Based Approaches:Jain and Gupta proposed an auto-updated, whitelist-based approach to protect against phishing attacks on the client side in 2016. The experimental results demonstrate that it achieved 86.02% accuracy and less than a 1.48% false-positive rate, which indicates a false warning for phishing attacks. The other benefit of this approach is fast access time, which guarantees a real-time environment and products.

Heuristic Strategies: Tan et al. introduced a phishing detection approach named PhishWHO, which consists of three phases. First, it obtains identity keywords by a weighted URL token system and ensembles the N-gram model from the page's HTML. Secondly, it puts the keywords into mainstream search engines to find the legitimate website and the legal domain. Next, it compares the legal domain and the target website's domain to determine if the target website is a phishing website or not. Chiew et al. used a logo image from the website to distinguish if the website was legal. In this paper, the authors extracted a logo from web page images by some machine learning algorithms and then queried the domain via the Google search engine with a logo as a keyword. Therefore, some researchers also called this category search engine-based approach.

Machine Learning-Based Methods: Machine learning-based countermeasures are proposed to address dynamic phishing attacks with higher accuracy performance and lower false positive rates than other methods. Consequently, the machine learning approach consists of six components: data collection, feature extraction, model training, model testing, and predicting.The flowchart of each part. Existing machine learning-based phishing website detection solutions are based on this flowchart to optimize one or more parts to obtain better performance.

Tiny URL Detection: Since tiny URLs do not present the real domain, resource direction, or search parameters, rule-based feature selection techniques might be useless for tiny URLs. Due to tiny URLs generated by different services, it is hard to convert them to original URLs. Furthermore, tiny URLs are short strings that are unfriendly for natural language processing to extract character-level features. If tiny URLs are not specially processed during data cleansing and preprocessing, they are likely to cause false or missed alarms. Internet products are also essential in terms of user experience, and users are also sensitive to false alarms of Internet security products. 6.4. Response Time for Real-Time Systems Rule-based models depend on rule parsing and third-party services from a URL string. Therefore, they demand a relatively long response time in a real-time prediction system that accepts a single URL string as an input in each request from a client. Phishing attacks spread to various communication media and target devices, such as personal computers and other smart devices. It is a big challenge for developers to cover all devices with one solution. Language independence and running environment independence should be taken into consideration to reduce system development complexity and late maintenance costs.

6.5 DATASET

We collected the datasets from the open-source platform called Phishing tank. The dataset that was collected was in csv format. There are 18 columns in the dataset, and we transformed the dataset by applying data pre-processing technique. To see the features in the data we used few of the data frame methods for familiarizing. For visualization, and to see how the data is distributed and how features are related to one another, a few plots and graphs are given. The Domain column has no bearing on the training of a machine learning model. We now have 16 features and a target column. The recovered features of the legitimate and phishing URL datasets are simply concatenated in the feature extraction file, with no shuffling. We need to shuffle the data to balance out the distribution while breaking it into training and testing sets. This also eliminates the possibility of over fitting during model training.

6.6 MACHINE LEARNING MODELS

Decision Tree Classifier :For classification and regression applications, decision trees are commonly used models. They basically learn a hierarchy of if/else questions that leads to a choice. Learning a decision tree is memorizing the sequence of if/else questions that leads to the correct answer in the shortest amount of time. The method runs through all potential tests to discover the one that is most informative about the target variable to build a tree.**Random Forest Classifier** :Random forests are one of the most extensively used machine learning approaches for regression and classification. A random forest is just a collection of decision trees, each somewhat different from the others. The notion behind random forests is that while each tree may do a decent job of predicting, it will almost certainly overfit on some data. They are incredibly powerful, frequently operate effectively without a lot of parameters adjusting, and don't require data scalability.

6.7 LIBRARIES USED

Pandas:It's a Python-based machine learning library. Pandas is a free and open-source programming language. Pandas is a programming language that is commonly used for dataset loading and data analytics. Pandas is used for machine learning in a variety of domains, including economics, finance, and others. It is extremely user-friendly and can display datasets in a tabular style for easier comprehension.

Sklearn: Sklearn is one of the most essential Python libraries for machine learning. Sklearn includes several tools for statistical classification, modelling, regression, dimensionality reduction and clustering.

Numpy: Numpy is a Python-based machine learning package. In Python, Numpy is used to deal with arrays. NumPy is used for all calculations using 1-d or 2-d arrays. Numpy also has routines for working with linear algebra and the Fourier transform.

MAPTlotlib: MAPTlotlib is a library for data visualization. It's a Python open-source module for plotting graphs from model results. These diagrams can aid in comprehending the circumstance of the outcomes. For easier comprehension, several components of the results can be graphically formatted.

6.8 EVALUATION

In this section, we use different models of machine learning for evaluating the accuracy. It has been explained about the different models in below sections. Where in this project the models are examined, with accuracy as the primary metric. In final stage we have compared the model accuracy. In all circumstances the testing and training datasets are splinted into 20:80 ratio.

Experiment 1/ Feature Distribution :Here in below figure shows how the data is distributed and how features are related to one another, a few plots and graphs are given.

Experiment 2/ Decision Tree Classifier :The method runs through all potential tests to discover the one that is most informative about the target variable to build a tree.

Where we are predicting the accuracy of the model on the samples collected on both trained and test samples. On this we found accuracy of test and training datasets are 82.6% and 81%. Below is the execution of Decision tree classifier algorithm. To generate model various parameters are set and the model is fitted in the tree. The samples are divided into X and Y train, X and Y test to heck the accuracy of the model.

Experiment 3/ Random Forest Classifier :We can limit the amount of over fitting by averaging the outcomes of numerous trees that all operate well and over fit in diverse ways. To construct a random forest model, you must first determine the number of trees to construct. They are incredibly powerful, frequently operate effectively without a lot of parameters adjusting, and don't require data scalability. Where we are predicting the accuracy of the model on the samples collected on both trained and test samples. On this we found accuracy of test and training datasets are 83.4% and 81.4%.

Experiment 4/ MLP :MLPs can be thought of as generalized linear models that go through numerous phases of processing before deciding. Below is the execution of the MLP algorithm. To generate a model various parameters are set and the model is fitted in the tree. The samples are divided into X and Y train, X and Y test to check the accuracy of the model. Where we are predicting the accuracy of the model on the samples collected on both trained and test samples. On this we found accuracy of test and training datasets are 86.3% and 85.9%.

Experiment 5/ XGBoost :Below is the execution of XGBoost algorithm. To generate a model various parameters are set and the model is fitted in the tree. The samples are divided into X and Y trains, X and Y tests to check the accuracy of the model. Where we are predicting the accuracy of the model on the samples collected on both trained and test samples. On this we found accuracy of test and training datasets are 86.4% and 86.6%.

Experiment 6/ Auto encoder :The auto-encoder must learn to encode the input to the hidden neurons with fewer neurons. In an auto encoder, the predictors (x) and output (y) are identical. To generate model various parameters are set and the model is fitted in the tree. The samples are divided into X and Y train, X and Y test to check the accuracy of the model. Where we are predicting the accuracy of the model on the samples collected on both trained and test samples. On this we found accuracy of test and training datasets are 81.8% and 81.9%.

Experiment 7/ SVM :An SVM training algorithm creates a model that assigns new examples to one of two categories, making it a non-probabilistic binary linear classifier, given a series of training examples that are individually designated as belonging to one of two categories. To generate model various parameters are set and the model is fitted in the tree. The samples are divided into X and Y train, X and Y test to check the accuracy of the model. Where we are predicting the accuracy of the model on the samples collected on both trained and test samples. On this we found accuracy of test and training datasets are 81.8% and 79.8%.

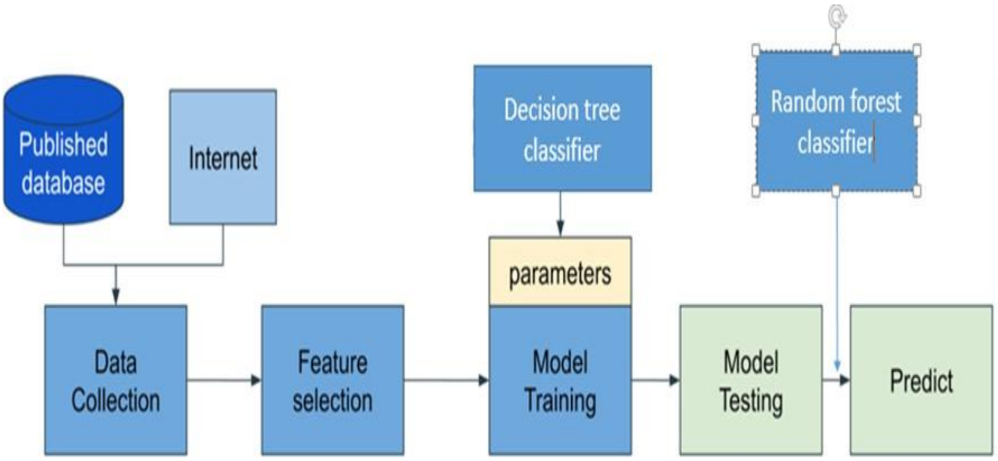
Performance Evaluation:

The evaluation of performance was carried out during the testing process. The original dataset would be divided into training data and test data, usually 80% and 20%, respectively. When evaluating the classifier's behavior on the testing dataset, there were four statistical numbers: the number of correctly identified positive data points (TP), the number of correctly identified negative data points (TN), the number of negative data points labeled by the classifier as positive (FP), and the number of positive data points labeled by the model as negative (FN).

There are several broadly used metrics to evaluate performance. The classification accuracy is the ratio of correct predictions to total predictions: accuracy = TP + TN /TP + TN + FN + FP.In binary classification cases, it is known that random selection has 50% accuracy. In unbalanced datasets, sometimes high accuracy does not mean that the model is excellent. For instance, among the 10,000 data, 9000 were legitimate websites, and 1000 were phishing websites, so when the prediction model did nothing, it could reach 90%. Accuracy is misleading when the class sizes are substantially different. Precision is the percentage of correctly identified positive data points among those predicted as positive by the model. The number of false-positive cases (FP) reflects the false warning rate. In real-time phishing detection systems, this directly affects the user experience and trustworthiness: Precision = $TP/TP + FP$.The recall is the portion of positive data points labeled as such by the model among all truly positive data points. The number of false-negative cases (FN) represents the number of phishing URLs that has not been detected. Leak alarms mean that users are likely to receive an attack that could result in the theft of sensitive information. Misleading users can do more harm to users than not detecting them:

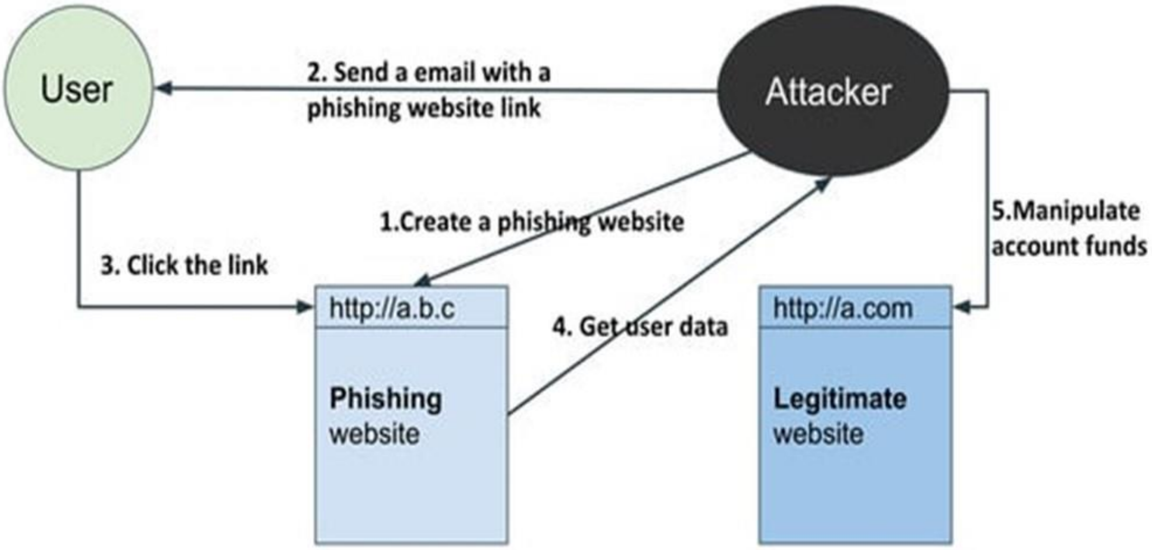
Recall = $TP /TP + FN$.The F-measure or F-score is the combination of precision and recall. Generally, it is formulated as shown below: $F\beta = (\beta 2 + 1)\times$ Precision × Recall/ $\beta 2$ × Precision + Recall $\beta \in (0, \infty)$.Here, β quantifies the relative importance of the precision and recall such that $\beta = 1$ stands for the precision and recall being equally important, which is also called F1. The F-score does the best job of any single statistic, but all four work together to describe the performance of a classifier: F1 = 2 × Precision × recall /Precision + recall = $TP /TP + 1/2 (FP + FN)$.In addition, many researchers use the N-fold cross-validation technique to measure performance for phishing detection [4,30,31]. The N-fold cross-validation technique is widely used on small datasets for evaluating machine learning models' performance. It is a resampling procedure that divides the original data samples into N pieces after shuffling the dataset randomly. One of the pieces is used in the testing process, and others are applied to the training process. Commonly, N is set as 10 or 5.

6.9 ARCHITECTURE DIAGRAM



6.9. Architecture diagram

6.10 FLOW DIAGRAM



6.10. Flow diagram

6.11 DECISION TREE ALGORITHM

There’s not much mathematics involved here. Since it is very easy to use and interpret it is one of the most widely used and practical methods used in Machine Learning.It is a tool that has applications spanning several different areas. Decision trees can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.Root Nodes – It is the node present at the beginning of a decision tree from this node the population starts dividing according to various features.Decision Nodes – the nodes we get after splitting the root nodes are called Decision Node.Leaf Nodes – the nodes where further splitting is not possible are called leaf nodes or terminal nodes. Sub-tree – just like a small portion of a graph is called sub-graph similarly a subsection of this decision tree is called sub-tree.Pruning – is nothing but cutting down some nodes to stop overfitting.

Entropy:Entropy is nothing but the uncertainty in our dataset or measure of disorder. Let me try to explain this with the help of an example.Suppose you have a group of friends who decides which movie they can watch together on Sunday. There are 2 choices for movies, one is “Lucy” and the second is “Titanic” and now everyone has to tell their choice. After everyone gives their answer we see that “Lucy” gets 4 votes and “Titanic” gets 5 votes. Which movie do we watch now? Isn’t it hard to choose 1 movie now because the votes for both the movies are somewhat equal.This is exactly what we call disorderness, there is an equal number of votes for both the movies, and we can’t really decide which movie we should watch. It would have been much easier if the votes for “Lucy” were 8 and for “Titanic” it was 2. Here we could easily say that the majority of votes are for “Lucy” hence everyone will be watching this movie.

Information Gain:Information gain measures the reduction of uncertainty given some feature and it is also a deciding factor for which attribute should be selected as a decision node or root node.It is just entropy of the full dataset – entropy of the dataset given some feature. To understand this better let’s consider an example:Suppose our entire population has a total of 30 instances. The dataset is to predict whether the person will go to the gym or not. Let’s say 16 people go to the gym and 14 people don’tNow we have two features to predict whether he/she will go to the gym or not. Feature 1 is “Energy” which takes two values “high” and “low”Feature 2 is “Motivation” which takes 3 values “No motivation”, “Neutral” and “Highly motivated”.Let’s see how our decision tree will be made using these 2 features. We’ll use information gain to decide which feature should be the root node and which feature should be placed after the split.Let’s calculate the entropy:To see the weighted average of entropy of each node we will do as follows:Now we have the value of E(Parent) and E(Parent|Energy), information gain will be:Our parent entropy was near 0.99 and after looking at this value of information gain, we can say that the entropy of the dataset will decrease by 0.37 if we make “Energy” as our root node.Similarly, we will do this with the other feature “Motivation” and calculate its information gain.In this example “Energy” will be our root node and we’ll do the same for sub-nodes. Here we can see that when the energy is “high” the entropy is low and hence we can say a person will definitely go to the gym if he has high energy, but what if the energy is low? We will again split the node based on the new feature which is “Motivation”.

You must be asking this question to yourself that when do we stop growing our tree? Usually, real-world datasets have a large number of features, which will result in a large number of splits, which in turn gives a huge tree. Such trees take time to build and can lead to overfitting. That means the tree will give very good accuracy on the training dataset but will give bad accuracy in test data.There are many ways to tackle this problem through hyperparameter tuning. We can set the maximum depth of our decision tree using the max_depth parameter. The more the value of max_depth, the more complex your tree will be. The training error will off-course decrease if we increase the max_depth value but when our test data comes into the picture, we will get a very bad accuracy. Hence you need a value that will not overfit as well as underfit our data and for this, you can use GridSearchCV.

Another way is to set the minimum number of samples for each spilt. It is denoted by min_samples_split. Here we specify the minimum number of samples required to do a split. For example, we can use a minimum of 10 samples to reach a decision. That means if a node has less than 10 samples then using this parameter, we can stop the further splitting of this node and make it a leaf node.

Pruning:It is another method that can help us avoid overfitting. It helps in improving the performance of the tree by cutting the nodes or sub-nodes which are not significant. It removes the branches which have very low importance. There are mainly 2 ways for pruning:Pre-pruning – we can stop growing the tree earlier, which means we can prune/remove/cut a node if it has low importance while growing the tree.Post-pruning – once our tree is built to its depth, we can start pruning the nodes based on their significance.

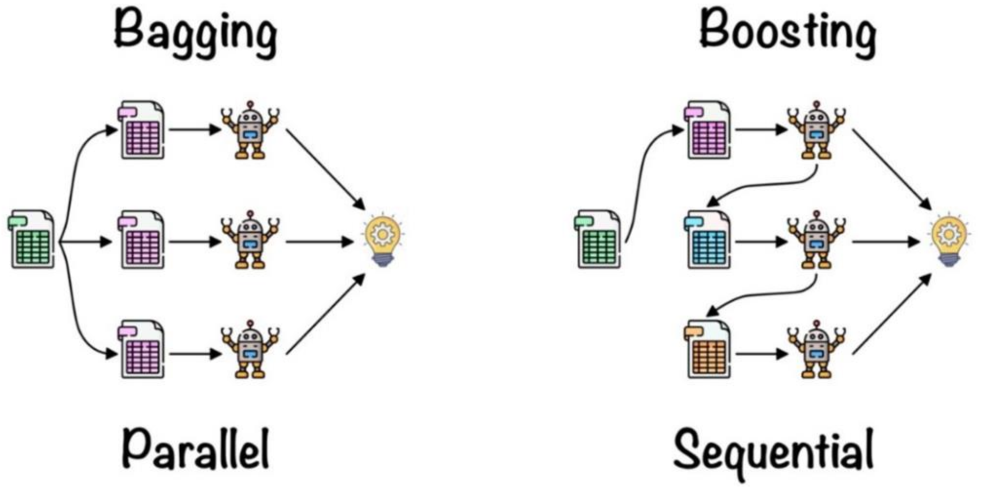
6.12 RANDOM FOREST

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.Let's dive into a real-life analogy to understand this concept further. A student named X wants to choose a course after his 10+2, and he is confused about the choice of course based on his skill set. So he decides to consult various people like his cousins, teachers, parents, degree students, and working people. He asks them varied questions like why he should choose, job opportunities with that course, course fee, etc. Finally, after consulting various people about the course he decides to take the course suggested by most of the people.

Ensemble uses two types of methods:

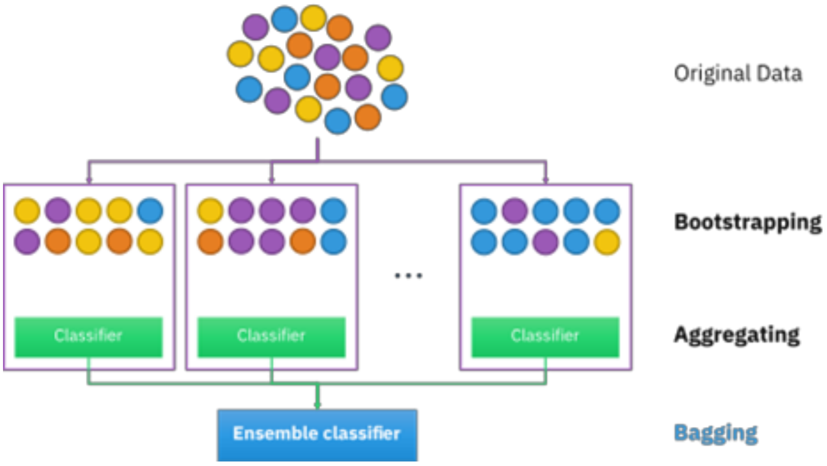
1.Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random
Forest.

2.Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.



6.12.1. Bagging & Boosting

Bagging:Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.



Steps involved in random forest algorithm:

Step 1: In Random forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for

Classification and regression respectively.

For example: consider the fruit basket as the data as shown in the figure below. Now a number of samples are taken from the fruit basket and an individual decision tree is constructed for each sample. Each decision tree will generate an output as shown in the figure. The final output is considered based on majority voting. In the below figure you can see that the majority decision tree gives output as an apple when compared to a banana, so the final output is taken as an apple.

Important Features of Random Forest

Diversity- Not all attributes/variables/features are considered while making an individual tree, each tree is different..Immune to the curse of dimensionality- Since each tree does not consider all the features, the feature space is reduced.Parallelization-Each tree is created independently out of different data and attributes.This means that we can make full use of the CPU to build random forests.Train-Test split- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.Stability- Stability arises because the result is based on majority voting/ averaging.

7. CONCLUSION

When the results of the survey conducted involving Kyrgyz Turkish Manas University students were examined, it could be seen that the majority of the students did not have sufficient knowledge about internet use and cyber threats. At the same time, they were found to lack technical knowledge of many issues including whether the websites they visit have security certificates or whether their information can be stolen by a hacker through deception. Since cyber threats affect people from all educational backgrounds, it would not be appropriate to provide this information only in the departments that provide technical education. The results of this study also show that the students who received cyber security education were more competent in terms of computer use and basic network security subjects. Almost all of the students who did not receive the education were eager for the same education. The study revealed that taking this education would be beneficial to the students to help them use the internet more securely. Cyber security awareness training can not only teach the students to be prepared for possible cyber threats but also inform them about the legal dimension of cybercrime.

The awareness levels can be re-measured after basic cyber security training is given to the students as a pilot application in future studies. Cyber skills can be tested through hands-on activities where the effects of the training can be explored. The same study can also be repeated with different demographics, for example, with students from a different country. In this way, it can be understood whether the lack of cyber security awareness is a regional or local problem. Apart from this, future studies may offer possible solutions by measuring the proficiency of the students or a different demographics in specific areas such as social media, password security, and malware.

This study, in its current form, has some limitations as it only measures the cyber security awareness of the students from a certain university based on a questionnaire. This study can be re-evaluated by adding other methods such as interviews and assessment/evaluation exams. More qualitative and quantitative results will be useful to increase the reliability of the study. After adding new methods, the framework of the study can also be visualized to increase its readability and coherence. *Ethical Statement*: This study was approved by the Faculty of Economics and Management of Kyrgyz Turkish Manas University document number R.30.2021/IBF-1745. (03/02/2021). *Conflict of Interest*: The authors declared there to be no conflict of interest.

8. REFERENCES

[1] M. Zwilling, G. Klien, D. Lesjak, .. Wiechetek, F. Cetin, and H. N. Basim, ``Cyber security awareness, knowledge and behavior: A comparative study," *J. Comput. Inf. Syst.*, vol. 62, no. 1, pp. 82_97, Jan. 2022.

[2] A. A. Karim, P. M. Shah, F. Khalid, M. Ahmad, and R. Din, ``The role of personal learning orientations and goals in Students' application of information skills in Malaysia," *Creative Educ.*, vol. 6, no. 18, pp. 2002_2012, 2015.

[3] N. Kaloudi and J. Li, ``The AI-based cyber threat landscape: A survey," *ACM Comput. Surv.*, vol. 53, no. 1, pp. 1_34, Jan. 2021.

[4] J. Jang-Jaccard and S. Nepal, ``A survey of emerging threats in cybersecurity," *J. Comput. Syst. Sci.*, vol. 80, no. 5, pp. 973_993, Aug. 2014.

[5] G. Pogrebna and M. Skilton, *Navigating New Cyber Risks: How Businesses Can Plan, Build and Manage Safe Spaces in the Digital Age*, London, U.K.: Palgrave Macmillan, 25, Jun. 2019.

[6] O. Hathaway, R. Crootof, P. Levitz, H. Nix, A. Nowlan, W. Perdue, and J. Spiegel, ``The law of cyber-attack," *California Law Rev.*, vol. 100, no. 4, 817-885, 2012.

[7] F. Forester and P. Morrison, *Computer Ethics*. Cambridge, MA, USA: MIT Press, 2001.

[8] D. Parker. (1989). *Computer Crime: Criminal Justice Resource Manual*. Accessed: Jan. 2, 2022. [Online]. Available: https://www.ncjrs.gov/pdf_les1/Digitization/118214NCJRS.pdf

[9] *De_nition of Cybersecurity*. Accessed: Mar. 2, 2022. [Online]. Available: https://www.itu.int/en/ITU-T/studygroups/com17/Pages/cyber security.aspx

[10] G. Pons-Salvador, X. Zubieta-Méndez, and D. Frias-Navarro, ``Internet use by children aged six to nine: Parents' beliefs and knowledge about risk prevention," *Child Indicators Res.*, vol. 11, no. 6, pp. 1983_2000, Dec. 2018.

[11] T. Correa, J. D. Straubhaar, W. Chen, and J. Spence, ``Brokering new technologies: The role of children in their parents' usage of the internet," *New Media Soc.*, vol. 17, no. 4, pp. 483_500, Apr. 2015.

[12] M. Micheli, ``What is new in the digital divide? Understanding internet use by teenagers from different social backgrounds," in *Communication and Information Technologies Annual*. Bingley, U.K.: Emerald Group Publishing, 2015, pp. 55_87.

[13] N. H. Abd Rahim, S. Hamid, and M. L. Mat Kiah, ``Enhancement of cybersecurity awareness program on personal data protection among youngsters in Malaysia: An assessment," *Malaysian J. Comput. Sci.*, vol. 32, no. 3, pp. 221_245, Jul. 2019.

[14] F. Quayyum, D. S. Cruzes, and L. Jaccheri, ``Cybersecurity awareness for children: A systematic literature review," *Int. J. Child-Computer Interact.*, vol. 30, Dec. 2021, Art. no. 100343.

[15] J. P. Hourcade, *Child-Computer Interaction*. Scotts Valley, CA, USA:

CreateSpace Independent Publishing Platform, 2015.

[16] R. S. Shaw, C. C. Chen, A. L. Harris, and H.-J. Huang, ``The impact of information richness on information security awareness training effectiveness,'' *Comput. Educ.*, vol. 52, no. 1, pp. 92_100, Jan. 2009.

[17] I. Hwang, R. Wake_eld, S. Kim, and T. Kim, ``Security awareness: The _rst step in information security compliance behavior,'' *J. Comput. Inf. Syst.*, vol. 61, no. 4, pp. 345_356, Jul. 2021.

[18] B. Khan, ``Effectiveness of information security awareness methods based on psychological theories,'' *Afr. J. Bus. Manage.*, vol. 5, no. 26, pp. 10862_10868, Oct. 2011.

[19] G. Cohen Zilka, ``Awareness of eSafety and potential online dangers among children and teenagers,'' *J. Inf. Technol. Education: Res.*, vol. 16, pp. 319_338, 2017.

[20] M. Adams and M. Makramalla, ``Cybersecurity skills training: An attacker-centric gami_ed approach,'' *Technol. Innov. Manage. Rev.*, vol. 5, no. 1, pp. 5_14, Jan. 2015.

[21] *Statista*. Accessed: Feb. 5, 2022. [Online]. Available: https://www.statista.com/topics/1145/internet-usage-worldwide/

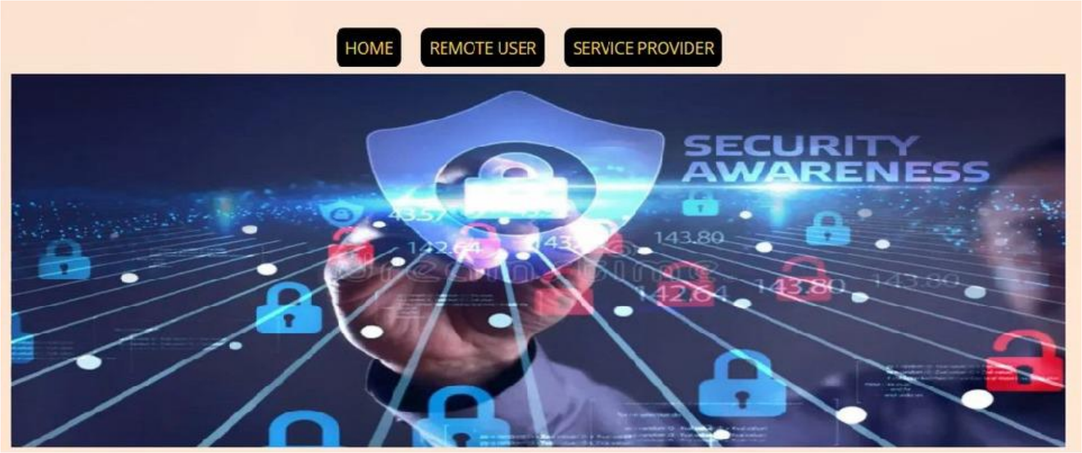
[22] A. Cuthbertson. (Jan. 5, 2022). *Ransomware Attacks Rise 250 Percent in 2017, Hitting U.S. Hardest*. Newsweek. [Online]. Available: https://www.newsweek.com/ransomware-attacks-rise-250-2017-us-wannacry-614034

[23] R. Ismailova and G. Muhametjanova, ``Cyber crime risk awareness in Kyrgyz republic,'' *Inf. Secur. J., A Global Perspective*, vol. 25, nos. 1_3, pp. 32_38, Apr. 2016.

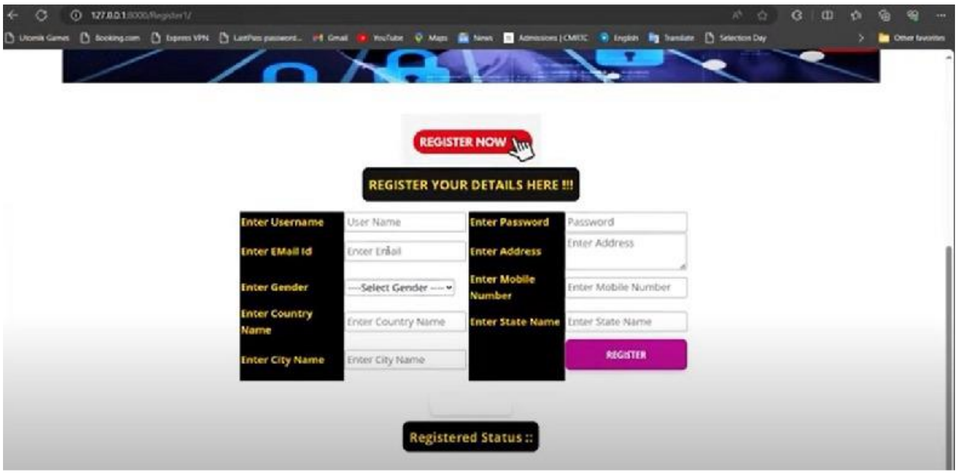
[24] A. Moallem, *Cybersecurity Awareness Among Students and Faculty*. Boca Raton, FL, USA: CRC Press, 2018.

[25] S. S. Tirumala, A. Sarrafzadeh, and P. Pang, ``A survey on internet usage and cybersecurity awareness in students,'' in *Proc. 14th Annu. Conf. Privacy, Secur. Trust (PST)*, Auckland, NewZealand, Dec. 2016, pp. 223_228.

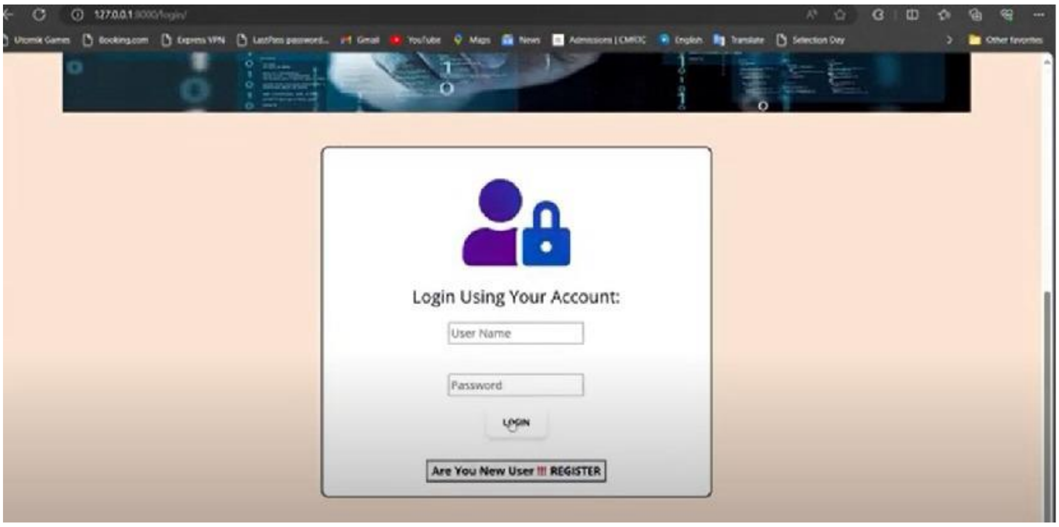
9. OUTPUT SCREENSHOTS



9.1.Home page



9.2.Registration page



Login Using Your Account:

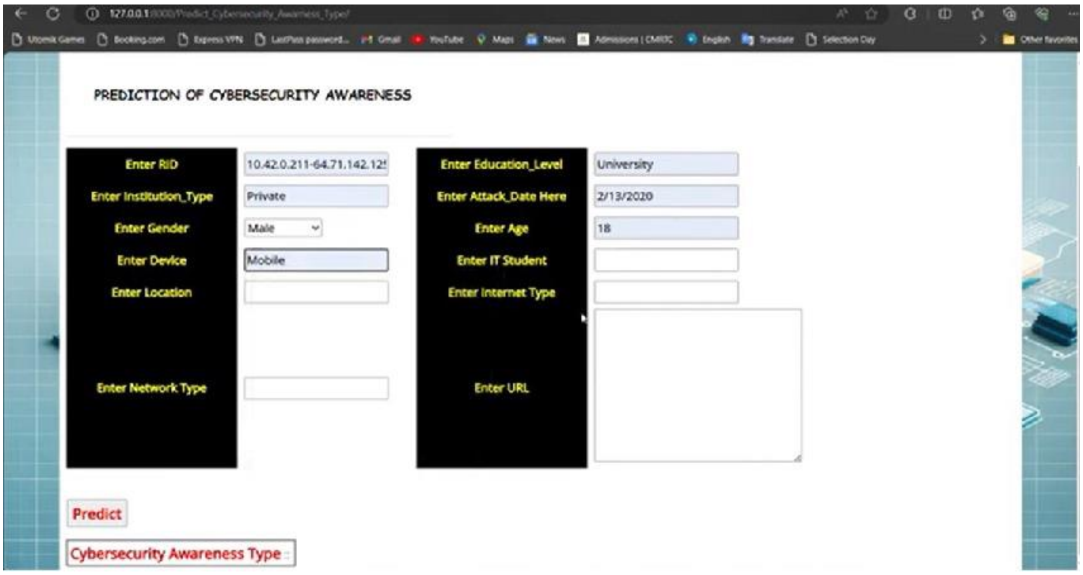
User Name

Password

LOGIN

Are You New User !!! REGISTER

9.3.Login page



PREDICTION OF CYBERSECURITY AWARENESS

Enter RID	10.42.0.211-64.71.142.12!	Enter Education_Level	University
Enter Institution_Type	Private	Enter Attack_Date Here	2/13/2020
Enter Gender	Male	Enter Age	18
Enter Device	Mobile	Enter IT Student	
Enter Location		Enter Internet Type	
Enter Network Type		Enter URL	

Predict

Cybersecurity Awareness Type :

9.4.Prediction page