

# **DETECTION OF FAKE SOCIAL MEDIA ACCOUNTS USING MACHINE LEARNING**

A PROJECT REPORT

*Submitted by,*

<b>Shravani. M</b>	<b>- 20191CSE0561</b>
<b>Narmada Gogineni</b>	<b>- 20191CSE0373</b>
<b>Sarojini T Habbli</b>	<b>- 20191CSE0534</b>
<b>Pagidela Venkata Mokshith Reddy</b>	<b>- 20191CSE0406</b>
<b>Pachipulusu Akash Kumar</b>	<b>- 20191CSE0405</b>

**Under the guidance of,  
Mr. Aarif Ahamed S  
Assistant Professor**

**Department of computer science and engineering  
in partial fulfilment for the award of the degree of  
Bachelors of Technology**

**In**

**COMPUTER SCIENCE AND ENGINEERING,**

**At**



**SCHOOL OF COMPUTER SCIENCE & ENGINEERING  
PRESIDENCY UNIVERSITY  
BENGALURU**

**JUNE 2023**

# **PRESIDENCY UNIVERSITY**

## **SCHOOL OF COMPUTER SCIENCE & ENGINEERING**

### **CERTIFICATE**

This is to certify that the Project report “**DETECTION OF FAKE SOCIAL MEDIA ACCOUNTS USING MACHINE LEARNING**” being submitted by “**SHRAVANI M - 20191CSE0561, NARMADA GOGINENI - 20191CSE0373, SAROJINI T HABBALI - 20191CSE0534, PAGIDELA VENKATA MOKSHITH REDDY - 20191CSE0406, PACHIPULUSU AKASH KUMAR - 20191CSE0405**”, in partial fulfilment of requirement for the award of degree of **Bachelor of Technology in Computer Science and Engineering** is a bonafide work carried out under my supervision.

**Mr. AARIF AHAMED S**

Assistant Professor-CSE

School of CSE&IS

Presidency University

**Dr. PALLAVI R**

Associate Professor & HOD

School of CSE&IS

Presidency University

**Dr. C. KALAIARASAN**

Associate Dean

School of CSE&IS

Presidency University

**Dr. MD. SAMEERUDDIN KHAN**

Dean

School of CSE&IS

Presidency University

# **PRESIDENCY UNIVERSITY**

## **SCHOOL OF COMPUTER SCIENCE & ENGINEERING**

### **DECLARATION**

We hereby declare that the work, which is being presented in the project report entitled **DETECTION OF FAKE SOCIAL MEDIA ACCOUNTS USING MACHINE LEARNING** in partial fulfilment for the award of Degree of **Bachelor of Technology in Computer Science and Engineering**, is a record of our own investigations carried under the guidance of **Mr. Aarif Ahamed S, Assistant Professor, Department of School of Computer Science & Engineering, Presidency University, Bengaluru.**

We have not submitted the matter presented in this report anywhere for the award of any other Degree.

<b>ID</b>	<b>Name</b>	<b>Signature</b>
<b>20191CSE0561</b>	<b>Shravani.M</b>	
<b>20191CSE0373</b>	<b>Narmada Gogineni</b>	
<b>20191CSE0534</b>	<b>Sarojini T Habbli</b>	
<b>20191CSE0406</b>	<b>Pagidela Venkata Mokshith Reddy</b>	
<b>20191CSE0405</b>	<b>Pachipulusu Akash Kumar</b>	

## **ABSTRACT**

**Internet social networks (ISNs) are increasingly popular and are more closely linked to people's social activities than ever before. They use its ISN to communicate with each other, share news, plan activities, and even run their own businesses online. Attackers and scammers are lured into its ISN to steal personal information, spread malicious activities, and spread misinformation. ISN is growing explosively and collects a large amount of personal data from users. Meanwhile, scientists are beginning to discover effective ways to detect suspicious activity and fraudulent accounts using already existing account classification algorithms and features. However, for some features work for the account is negatively affected or does not affect the results. Furthermore, the use of independent classification algorithms does not always give satisfactory results. To build the decision tree of this article, three techniques of feature selection and size reduction were used to effectively detect fake Instagram accounts. Three machine learning classification algorithms were used to determine if the interest account was real or fake: Decision trees, Random forests and Logistic regression.**

**keyword: Decision trees, random forests, logistic regression.**

## **ABBREVIATIONS**

<b>RAM</b>	-Random Access Memory
<b>HTML</b>	-Hyper-text Mark-up Language
<b>CSS</b>	-Cascading Style Sheets
<b>PIP</b>	-Package Installer for Python
<b>PEP</b>	-Python Enhancement Proposals
<b>REPL</b>	-Read-Eval-Print Loop
<b>BDFL</b>	-Benevolent Dictator For Life
<b>LEGB</b>	-Local, Enclosing, Global, and Built-in

## ACKNOWLEDGEMENT

First of all, we indebted to the GOD ALMIGHTY for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Dean, School of Computer Science & Engineering, Presidency University for getting us permission to undergo the project.

We record our heartfelt gratitude to our beloved Associate Dean **Dr. C. Kalaiarasan**, Professor **Dr. T K Thivakaran**, University Project-II In-charge, School of Computer Science & Engineering, Presidency University for rendering timely help for the successful completion of this project.

We would like to convey our gratitude and heartfelt thanks to the University Project-II Co-Ordinators **Mr. Mrutyunjaya MS**, **Mr. Sanjeev P Kaulgud**, **Mr. Rama Krishna K** and **Dr. Madhusudhan MV**.

We are greatly indebted to our guide **Mr. Aarif Ahamed**, Assistant Professor, School of Computer Science & Engineering, Presidency University for his inspirational guidance, valuable suggestions and providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

<b>Shravani. M</b>	<b>- 20191CSE0561</b>
<b>Narmada Gogineni</b>	<b>- 20191CSE0373</b>
<b>Sarojini T Habbli</b>	<b>- 20191CSE0534</b>
<b>Pagidela Venkata Mokshith Reddy</b>	<b>- 20191CSE0406</b>
<b>Pachipulusu Akash Kumar</b>	<b>- 20191CSE0405</b>

## Table of Contents

CERTIFICATE .....	i
DECLARATION .....	ii
ABSTRACT .....	iii
ABBREVIATIONS.....	iv
ACKNOWLEDGEMENT .....	v
LIST OF FIGURES.....	vii
CHAPTER-1 .....	1
INTRODUCTION .....	1
CHAPTER – 2 .....	4
REQUIREMENT ANALYSIS .....	4
2.1 HARDWARE REQUIREMENTS:.....	4
2.2 SOFTWARE REQUIREMENTS: .....	4
CHAPTER-3 .....	8
LITERATURE SURVEY .....	8
CHAPTER – 4 .....	11
EXISTING WORK.....	11
CHAPTER – 5 .....	12
PROPOSED WORK .....	12
CHAPTER – 6 .....	17
SYSTEM DESIGN .....	17
CHAPTER – 7 .....	18
IMPLEMENTATION.....	18
CHAPTER-8.....	41
TIMELINE FOR EXECUTION OF PROJECT .....	41
(GHANTT CHART).....	41
CHAPTER-9 .....	43
TESTING .....	43
CHAPTER-10.....	49
CONCLUSIONS AND FUTURE SCOPE .....	49
CHAPTER - 11 .....	50
REFERENCES .....	50

## LIST OF FIGURES

Figure 1 Picture showing the official python Webpage .....	5
Figure 2 Picture showing the PyCharm Download version .....	6
Figure 3 Welcome to PyCharm page .....	7
Figure 4 Installing the python packages in cmd .....	7
Figure 5 Random Forest Classifier .....	14
Figure 6 Example of Random forest classifier .....	15
Figure 7 Figure shows the System Design .....	17
Figure 8 Picture shows the USECASE Diagram .....	21
Figure 9 Picture shows the Class Diagram .....	22
Figure 10 Picture shows the Sequence Diagram .....	23
Figure 11 Picture shows the Collaboration Diagram .....	24
Figure 12 Picture shows the Activity Diagram .....	25
Figure 13 Picture shows the Deployment Diagram .....	26
Figure 14 Picture shows the Component Diagram .....	26
Figure 16 Picture DFD Diagram-1 .....	28
Figure 17 Picture DFD Diagram-2 .....	29
Figure 18 TIMELINE FOR EXECUTION OF PROJECT .....	41
Figure 19 Home Page .....	46
Figure 20 About Page .....	46
Figure 21 Data Page .....	47
Figure 22 Model Selection-1 Page .....	47
Figure 23 Model Selection-2 Page .....	48
Figure 24 Prediction Page .....	48



## **CHAPTER-1**

### **INTRODUCTION**

Internet social networks (ISNs) such as Facebook, Twitter, LinkedIn, and Google+ have become increasingly popular in recent years. People can use ISN to keep in touch with each other, share news, organize events, and even run their own e-business. Between 2014 and 2018, nonprofits that sponsored political ads on Facebook spent about \$2.53 million. ISN is vulnerable to Sybil attacks because they are open and subscribers store large amounts of personal information. In 2012, Facebook became aware of deceptive practices on its platform such as fake news posts, hate speech, sensationalism and polarization. But Internet social networks (ISNs) have also caught the attention of factfinders by quarrying and surveying vast amounts of statistics, examining and investigating client behavior, and detecting anomalous activities. Researchers conducted a study to predict, analyze, Explaining customer loyalty in social media-based online brand communities involves identifying the cognitive traits that effectively predict customer attitudes. The Facebook community has continued to experience growth, surpassing 2.2 billion monthly active users and 1.4 billion daily users, reflecting an 11% year-over-year increase. In the second quarter of 2018, Facebook reported a total revenue of \$13.2 billion, with \$13 billion coming solely from advertising. Similarly, Twitter reported nearly 1 billion Twitter members and 336 million monthly active clients during the same quarter in 2018. While Twitter witnessed steady revenue growth of \$2.44 billion in 2017, its profits declined by \$108 million compared to the previous year. In 2015, Facebook estimated that approximately 14 million of its monthly active users were unwanted, representing malicious fake accounts created in violation of the platform's terms of service. During the first quarter of 2018, Facebook released a report outlining its internal policies for enforcing community standards from October 2017 to March 2018. This report disclosed the removal of various types of unwanted content from Facebook, including graphic violence, adult nudity and sexual activity, terrorist propaganda, hate speech, spam, and fake accounts. In total, 837 million spam posts were removed, approximately 583 million fake accounts were deactivated, and around 81 million pieces of other infringing content were taken down. Despite these efforts to combat fake accounts, it is still estimated that.

About 88.5 million accounts are still fake accounts. For example, the presence of fake accounts on OSN has led advertisers, developers, and inventors to distrust the identity of reported users, and recently, US banks and financial institutions have been sued for his OSN. decided to sell it. I

opened a Twitter account. began exploiting the Twitter and Facebook accounts of loan applicants. before the loan is actually paid. The attackers believe the ISN user, whose account is "the key to the walled garden," unknowingly stole or artificially created photos and profiles of real people. increase. I think it could be stolen. use. Used to impersonate another person. Spreading fake news or stealing personal information. These fake accounts are commonly known as scammers. In both cases, such fake accounts harm users, but the motivation is usually offensive, such as sending spam her messages or stealing personal information. They aim to lure single naive users into fake relationships with phishing scams, leading to a bit of other fraud, smuggling and even political spheres. Data show that 42% of elderly and 18% of teens in the U.S. are very concerned about using fake social media accounts or bots to sell or influence products. I'm here. I am here. Another example: Romney's Twitter account suddenly exploded during the 2012 US presidential election. Most of them were later classified as fake followers. A real person designed to mimic American social media users. Similarly, during the run-up to the February 2013 Italian general election, online blogs and newspapers published statistics on the number of times key candidates were mistreated. Detecting these threatening accounts within ISN is essential to prevent a variety of malicious activities, ensure the security of users his accounts, and protect their private information. Researchers are trying to develop automated detection tools to identify fake accounts, but doing it manually is cumbersome and expensive. The researchers' work will enable ISN operators to efficiently and effectively detect fake accounts, potentially improving the user experience by preventing unwanted spam messages and other misleading content. I have. ISN operators may also increase the reliability of user identification numbers and allow third parties to consider user accounts. Information security and privacy are among the top requirements of social network users, and adherence and adherence to these requirements is increasing

About 88.5 million accounts are still fake accounts. For example, the presence of fake accounts on ISN has led advertisers, developers, and inventors to distrust the identity of reported users, and recently, US banks and financial institutions have been sued for his ISN. decided to sell it. I opened a Twitter account. The loan applicant responded on Twitter. They started abusing his Facebook account, which could negatively affect his income. before the loan is actually paid. Attackers claim an ISN user whose account is "the key to the walled garden" by stealing or artificially creating photos and profiles of real people without his knowledge. I believed. Was I here? Was I here? Used to impersonate others. Researchers are trying to develop automated detection tools to identify

fake accounts. but doing it manually is cumbersome and expensive. The researchers' work will enable ISN operators to efficiently and effectively detect fake accounts, potentially improving the user experience by preventing unwanted spam messages and other fraudulent content. In addition, ISN operators can increase the reliability of user identification numbers and allow third parties to consider user accounts. Information security and privacy are among the core requirements of social network users, and compliance and adherence to these requirements is on the rise. Researchers focus on Extract features to analyze user-level activity current clients to identify fake accounts. Number of posts, number of followers, profile. They apply trained machine learning techniques to classify real and fake accounts. Another approach is to use a diagram-level structure. In this case, the ISN is a modeled diagram, represented as a collection of nodes and edges. Each node represents an entity (such as an account) and each edge represents a relationship (such as a friendship). Therefore, automatic Sybil detection is not necessarily robust against enemy pounce and does not provide the desired accuracies. The random forest classifying algorithms was performed based on decisions obtained from a support vector machine (SVM). We also checked the detection capabilities of the classifier using his two additional sets of real and fake accounts unrelated to the original training dataset, as shown in . It outlines research conducted on the Twitter network as well as previous research on detecting fake profiles. It shows how the data was preprocessed and how the results were used to differentiate account into true and false accounts. Eventually, accuracy was examined and evaluated relative to all other techniques applied.

## **CHAPTER – 2**

### **REQUIREMENT ANALYSIS**

#### 2.1 Hardware Components

#### 2.2 Software Components

##### **2.1 HARD-WARE REQUIREMENTS:**

- |              |                               |
|--------------|-------------------------------|
| 1. Processor | - I3/Intel Processor          |
| 2. RAM       | - 4GB (minmum)                |
| 3. Hard Disk | - 160GB                       |
| 4. Key Board | - Standard Windows Keyboards  |
| 5. Mouse     | - Two or Three Buttons Mouses |
| 6. Monitor   | - SVGA                        |

##### **2.2 SOFTWARE REQUIREMENTS:**

- |                        |             |
|------------------------|-------------|
| 1. Operating System    | : Window 10 |
| 2. Servers side Script | : Python    |
| 3. IDE                 | : PyCharm   |

## Install software for ML project

### Installing Python

1. To download and deploy Python, go to the website <https://www.python.org/downloads/> and select a version.

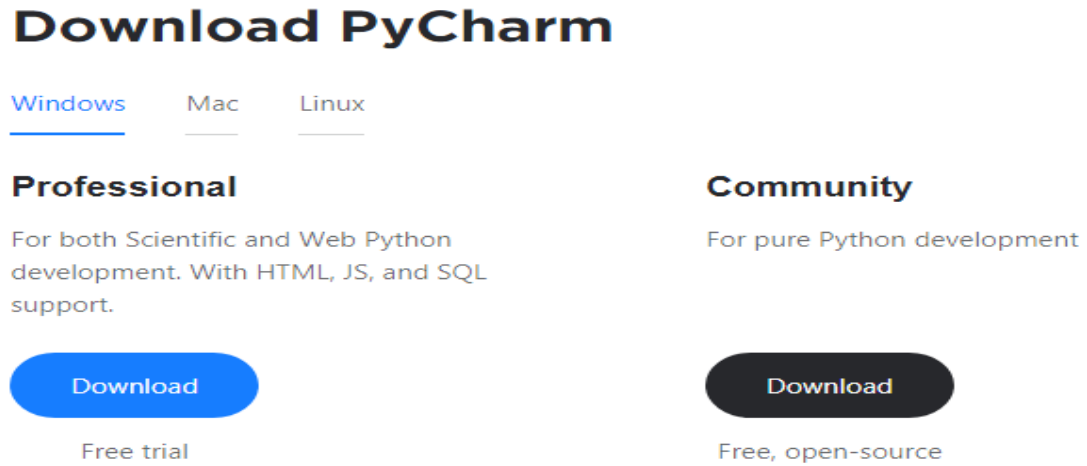


**Figure 1** Picture showing the official python Webpage

2. After the download is complete, run the exe file to install Python. Click Install Now.
3. At this point you can see that Python is installed.
4. When the process is complete, you will see a screen stating Indicates that the set-up was successful. Click Close-button.

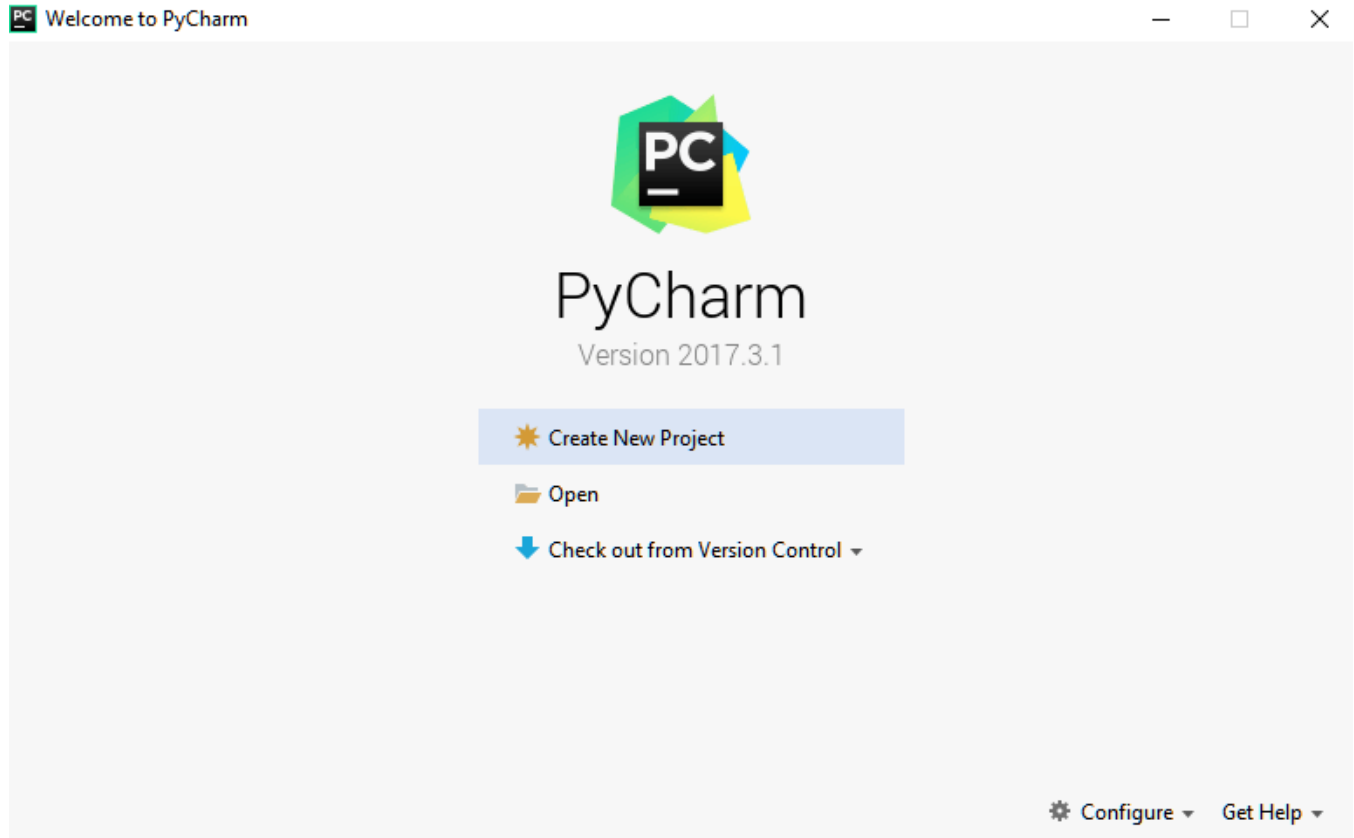
## Installing PyCharm

1. To download PyCharm, go to the website <https://www.jetbrains.com/idea/>. Go to [//www.pycharm.com/download/](https://www.pycharm.com/download/) and click the "Download-button" link in the display area.



**Figure 2 Picture showing the PyCharm Download version**

2. After the downloading is complete, run the exe file to install PyCharm. The set up wizard should have started. Click Next.



**Figure 3 Welcome to PyCharm page**

9. To run the project properly, you need to install some packages.
10. Open a cmd Anaconda as administrator.
11. A cmd opens at the specified path. Enter the "Pip Install Package Name" to install (e.g. Numpy, Pandas, Seaborn, Scikit-Learn, Matplotlib-pyplot).

```
C:\WINDOWS\system32>pip install numpy==1.18.5
Collecting numpy==1.18.5
  Downloading numpy-1.18.5-cp36-cp36m-win_amd64.whl (12.7 MB)
    |████████████████████████████████████████| 12.7 MB 939 kB/s
ERROR: tensorboard 2.0.2 has requirement setuptools>=41.0.0, b
Installing collected packages: numpy
Successfully installed numpy-1.18.5
```

**Figure 4 Installing the python packages in cmd**

### CHAPTER-3 LITERATURE SURVEY

Sl. No	Paper Title	Method	Advantages	Limitations
1	Fake Profile Detection Using Machine Learning Techniques	XG boost A Random forest Multi-layered neural network Googles Free GPU Utilization NVIDIA Tesla K80 GPU	Protecting users from scams and fraud: Fake accounts can be used to spread scams and phishing attempts, which can result in financial losses for users. By detecting and removing fake accounts, social media platforms can help prevent these types of attacks.	The primary drawback of this project is its limited scope, as it solely operates on accessible data and lacks real-time applicability.
2	Real Time Profile Analysis and Fake Detection Model for Improved Profile Security in Online Social Networks	PrLTA, PLTA , PHATA	Preserving the integrity of social media: Fake accounts can be used to manipulate social media engagement metrics such as likes, comments, and shares, which can manipulated content. By detecting and removing fake accounts, social media platforms can preserve the integrity of their platforms and maintain user trust.	The profile trust analysis is done by measuring the similarity among different profiles and only on higher valued PSM is selected and based on that the fake profile has been detected



Sl. No	Paper Title	Method	Advantages	Limitations
3	Auto-matic Detection of Fake Profile Using ML on Instagram	Artificial Neural networks (ANN), Random forest Algorithm, SVC Algorithm.	Ensuring the credibility of social media platforms and safeguarding users from potential harm necessitates the detection of fake accounts, which can be achieved with high accuracy, low complexity, exceptional efficiency, and without requiring highly skilled personnel.	High complexity, Requires skilled person
4	Detecting Fake Accounts in Online Social Networks at the Time of Registrations	Ianus, Louvain's variants of Ianus- Sync and Ianus Anomaly Ianus - FS Ianus- CD Ianus -FS-CD	Reducing the spread of misinformation: Fake accounts can be used to spread false information and propaganda, which can have harmful effects on individuals and society as a whole. By detecting and removing fake accounts, social media platforms can help reduce the spread of misinformation and promote accurate information.	Low Accuracy, And High ineffective

Sl. No	Paper Title	Method	Advantages	Limitations
5	Fake Identities in Online Social Media - A Comprehensive Survey	Feature Set: SVM, NLP, Naïve bayes	Improving advertising targeting: Fake accounts can be used to artificially inflate the number of followers and engagement metrics for certain accounts, which can skew advertising targeting and make it less effective. By detecting and removing fake accounts, social media platforms can improve the accuracy of their advertising targeting and make it more effective for advertisers.	Combining multiple models and creating one that works in real time can lead to better results while creating a complex and time consuming process. .

## **CHAPTER – 4**

### **EXISTING WORK**

With the growth of machine learning, computing technology can be segregated as conventional method and contraption study methods. However, it takes a lot of memory and the results are not accurate.

.

#### **DISADVANTAGES**

- Low Accuracy
- Requires skilled persons

## **CHAPTER – 5**

### **PROPOSED WORK**

This proposed application can be viewed as a useful system as it helps alleviate the limitations of conventional and other existing methods. We used powerful algorithms in a Python-based environment to design this systems.

#### **ADVANTAGES**

- Accuracy is good.
- No need of skilled persons

### **5.1 ALGORITHMS**

#### **5.1.1 LOGISTIC REGRESSION:**

Logistic regression is a widely used statistical model that predicts binary outcomes by estimating the likelihood of developing heart disease based on explanatory factors such as age, gender, blood pressure, and other relevant variables. By analyzing these factors, logistic regression provides insight into the status and likelihood of developing heart disease. Logistic regression is a commonly used statistical technique for binary data. A classification problem focused on predicting the probability that an event or outcome he will be classified into one of two classes. The name comes from the logistic function known as the sigmoid function, used to model the relationship between independent and dependent variables. In logistic regression, the probability  $P(y=1|x)$  represents the dependent probability, classifying the variable as 1 based on the values of the independent variables. The logistic function uses the natural logarithm base and the linear function denoted by  $e$ . For example, a combination of independent variables is expressed as: Logistic regression can predict the class membership of a dependent variable. Apply a threshold (usually 0.5) to the predicted probabilities. This threshold is a transformation that facilitates the assignment of class labels based on estimated probabilities, use a logistic regression model.

**Uses of logistic regression:**

Logistic regression is especially popular in online advertising, allowing marketers to predict the likelihood that users of a particular website will click on a particular ad based on the percentage of votes for or against it.

- Logistic Regression can also be used in the following cases.
- Health care identifies disease risk factors and prevention plans.
- A weather application that predicts snowfall and weather conditions.
- Vote requests to determine if a voter voted for a particular candidate. Insurance that predicts the likelihood of death of a policyholder before the end of the policy term based on certain criteria such as sex, age and physical examination. A banking service used to predict a credit seeker's default or likelihood of default based on annual income, past defaults, and past liabilities.

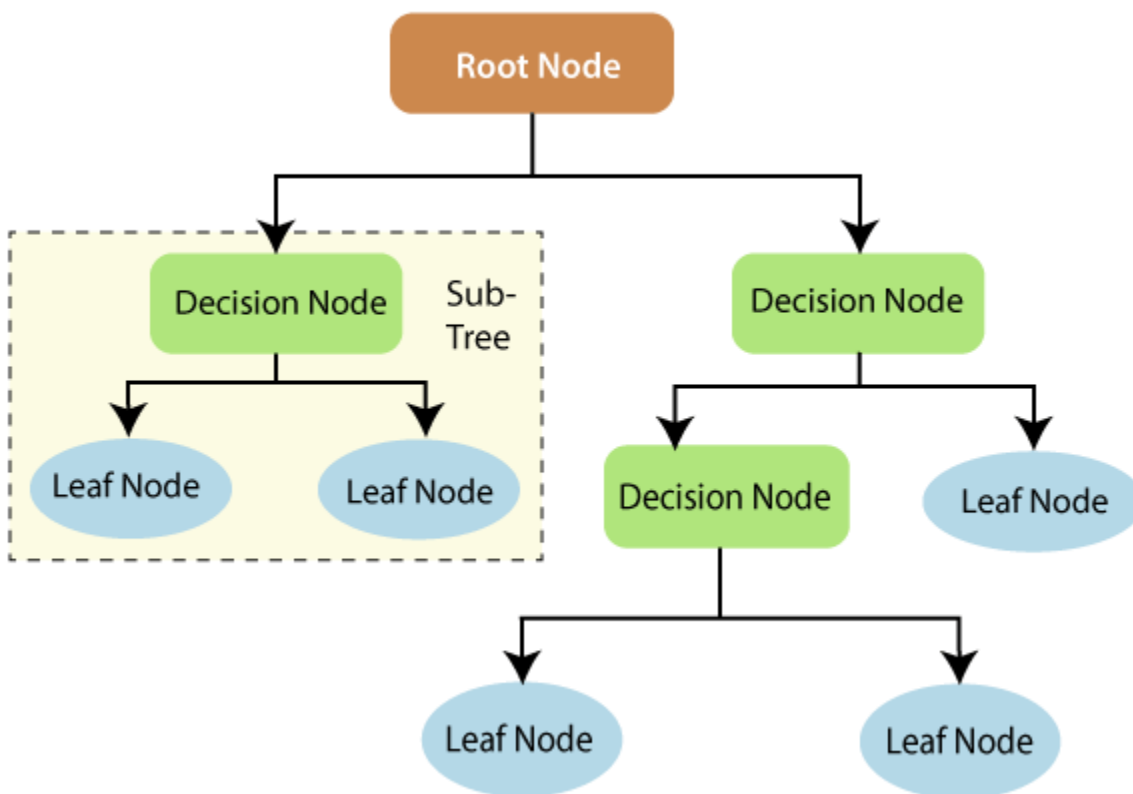
**5.1.2 DECISION TREE CLASSIFIER:**

The Decision Tree Classifier is a powerful algorithm that uses a tree-like structure to classify people with or without heart disease. By recursively partitioning the dataset based on specific medical history attributes, this algorithm is able to accurately classify patients and facilitate predictions about their cardiac status. The process of building a decision tree involves choosing the most appropriate feature to split the data at each node. Various algorithms such as ID3, C4.5 and CART use different measures such as information gain, Gini exponent, mean squared error to assess the quality of the splits. In the training phase, the decision tree algorithm aims to identify the best splits by assessing the ability of features to effectively separate the data and improve the purity or homogeneity of the resulting subsets. The goal is to minimize impurities or uncertainties within each subset, thereby creating homogeneous subsets that are more likely to belong to the same class. This process allows the decision tree to make accurate predictions based on learned patterns in the data.

In summary, the decision tree classifier is a widely used and intuitive algorithm for classification tasks. It constructs a tree-like structure by iteratively partitioning the data according to feature conditions, enabling it to make predictions based on the path traversed in the tree. This approach has gained popularity due to its effectiveness in providing accurate classification results in a straightforward and interpretable manner.

### 5.1.3 RANDOM FOREST CLASSIFIER:

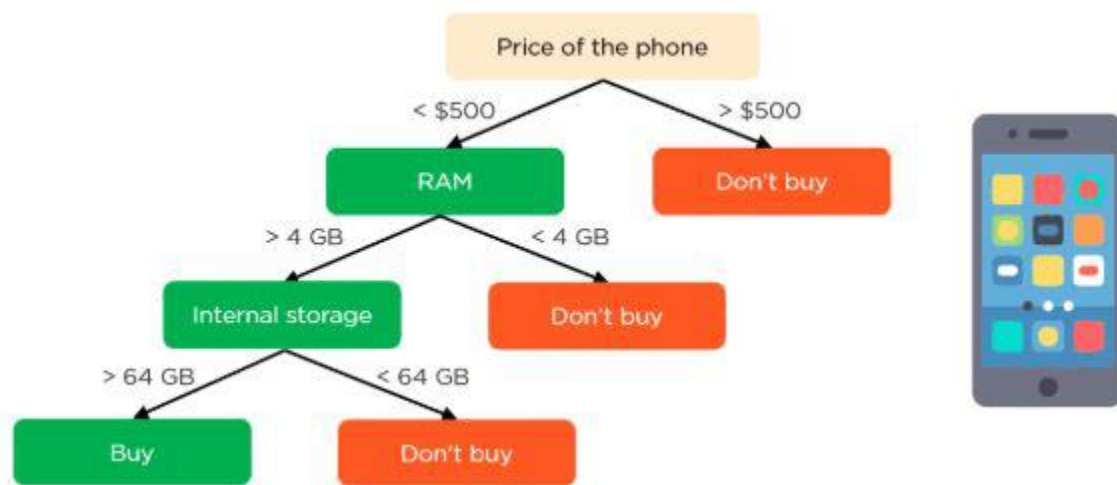
The random forest classifier is an ensemble learning algorithm that combines multiple decision trees to enhance accuracy and robustness. By aggregating predictions from various decision trees, this algorithm can effectively handle complex relationships and feature interactions, leading to improved heart disease predictions. Training a random forest can be computationally demanding, particularly when dealing with large datasets and a high number of trees. The prediction process of random forests can also be relatively slower compared to simpler models. After training all the trees, predictions are made by combining the results from each individual tree. In classification tasks, the random forest aggregates the predictions using majority voting, where the class with the highest number of votes across the trees is selected as the final prediction. For regression tasks, the predictions are averaged across the trees. Random forest classifiers are widely used in diverse.



**Figure 5 Random Forest Classifier**

Information theory provides detailed information about how decision trees work. Entropy and information acquisition are the building blocks of decision trees. An overview of these basic concepts will give you a better understanding of how decision trees are constructed. Entropy is a metric used to calculate uncertainty. Information gain is a measure of how the uncertainty of the

target variable is reduced compared to a set of independent variables. The concept of information retrieval involves gathering information about target variables (classes) using independent variables (characteristics). The randomness of the target variable (Y) and the unconditionality of Y (given X) are used to estimate the information gain. In this case, conditional anarchy is subtracted from Y's entropy. The information obtained is used to train the decision tree. This helps reduce the uncertainty in these trees. A high information gain means that a high degree of uncertainty (information entropy) is removed. Anarchy and information acquisition are important in branch splitting, a key activity in building decision trees. Let's look at a simple example to show how decision trees work. Suppose you want to predict whether a customer will buy a mobile phone. The basis for his decision is the characteristics of the phone.



**Figure 6 Example of Random Forest classifier**

Applying Decision Trees to Random Trees The main difference between decision trees and random forest algorithms is that the root node creation and node separation are random in the latter. Random Forest uses a bagging technique to generate desired predictions. Bagging uses multiple data samples (training data) instead of just one sample. The training dataset contains observations and features used to make predictions. Decision trees produce different results depending on the training data fed to the random forest algorithm. These scores are scored and the highest score is selected as the final score. The first example can still be used to illustrate how Random Forest works. Instead of a single decision tree, the random tree contains many decision trees. Suppose there are only four decision trees. In this case, the training data containing phone observations and

features is split into four root nodes. A root node can represent four characteristics (price, internal storage, camera, RAM) that influence a customer's choice. Random forests randomly select features to split nodes. A final prediction is made based on the results of the four trees. The outcome chosen in most decision trees is the final decision. If 3 trees predict a purchase and 1 tree predicts no purchase, the final prediction is a purchase. In this case, the customer is expected to purchase the phone.



## CHAPTER – 6 SYSTEM DESIGN

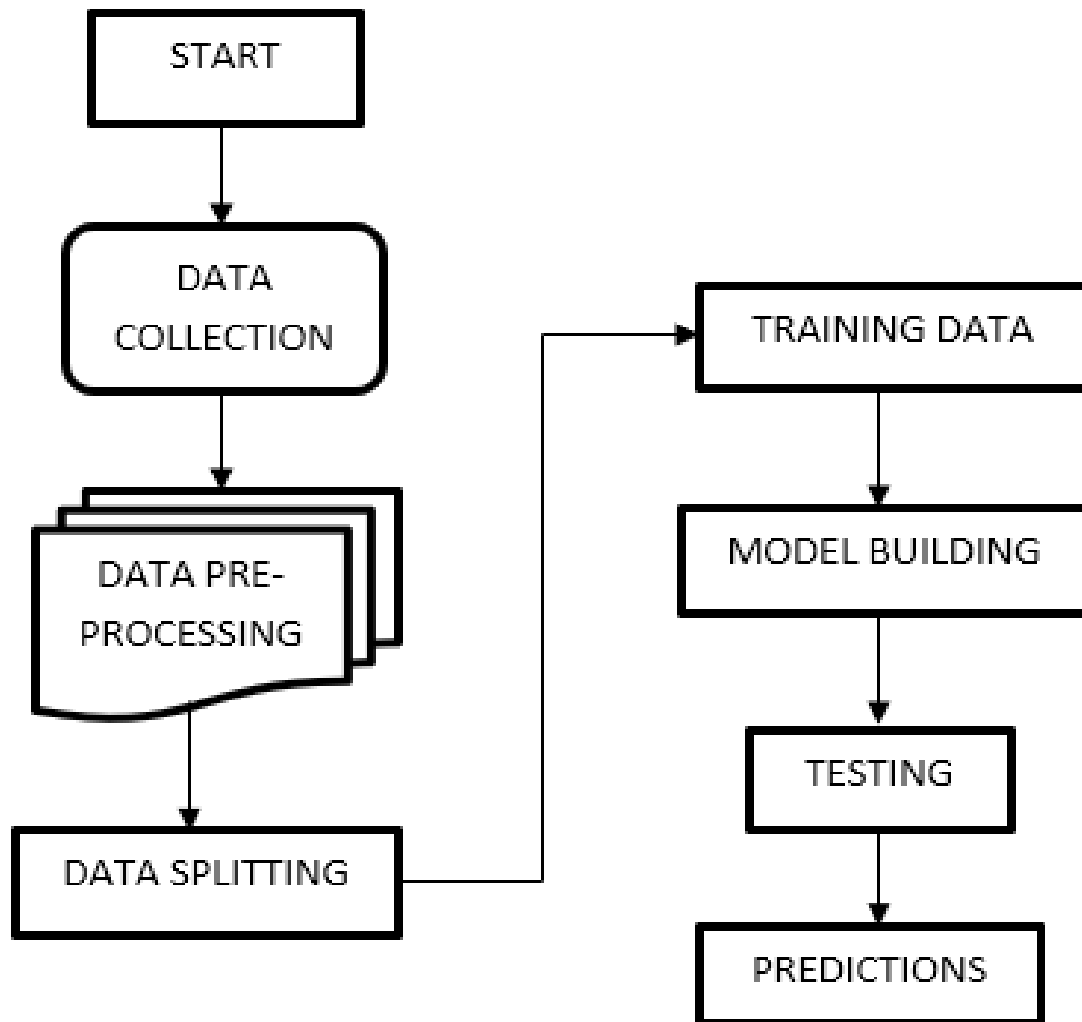


Figure 7 Figure shows the System Design

## **CHAPTER – 7 IMPLEMENTATION**

### **7.1 System:**

#### **7.1.1 Store Dataset:**

The system saves the records specified by the user.

#### **7.1.2 Model Training:**

The system receives the user's data and transfers it to the selected model.

#### **7.1.3 Graphs Generation:**

The system takes a user-specified dataset, selects a model, and generates an accuracy corresponding to the selected model.

### **7.2. User:**

#### **7.2.1 Load Dataset:**

The User can load the record they want to work with

#### **7.2.2 View Dataset:**

The Users can view records.

#### **7.2.3 Select model:**

Users can apply models to datasets to ensure accuracy.

#### **7.2.4 Prediction:**

Passing parameters to predict the output

### **7.3 STEPS -FOR EXECUTION OF THE PROJECT**

1. Installation of all the require packages
2. Defining the problem association.
3. Create a Django based User Interface.
4. Upload CSV file
5. View data
6. Model Selection
7. Predict output.

## **UML DIAGRAMS:**

Unified Modeling Language (UML) is a standardized general-purpose modeling language in the field of object-oriented software development. This is managed and produced by the Object Management Group. The aim is for UML to become a general language for creating models of object-oriented computer software. In its current form, UML consists of two main components. Supermodels and symbols. Other methods and processes may be added in the future. Unified Modelling Language is a standard language for specifying, visualizing, creating, and documenting the artifacts of software systems, business models, and other non-software systems. UML represents a set of technical best practices that have proven successful in modeling large and complex systems. UML is a very important part of object-oriented software development and the software development process. UML basically employs graphical documentation to speak to the plan of computer program projects.

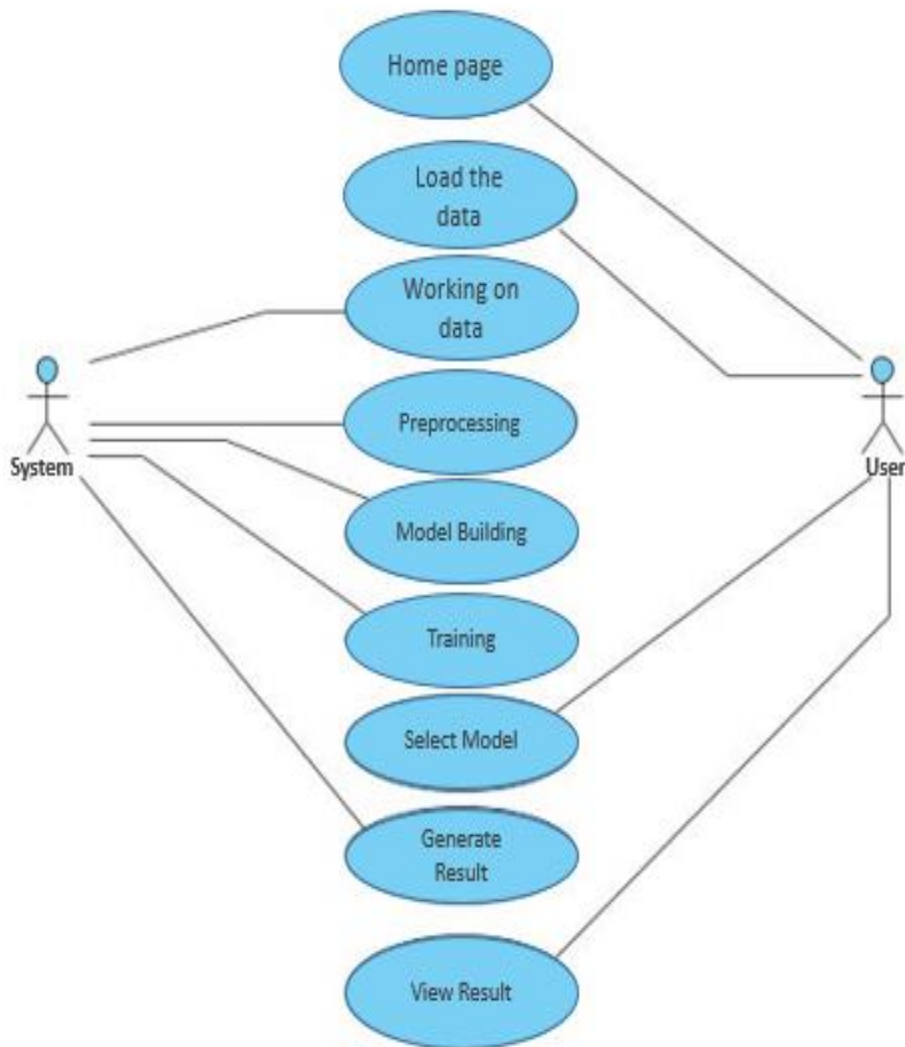
## **GOALS:**

The Essential objectives within the plan of the UML are as follows:

1. Give clients a get-set-go, visual displaying Dialect by which they can create and trade significant prototype.
2. Give extendible and specialized mechanisms to amplify the center concepts.
3. Be free of specific programming dialects and improvement handle.
4. Give a formulated premise to understand the displaying dialect.
5. Empower the development of Object-Oriented devices showcase.
6. Bolster highest advancement in collaboration's, systems, designs and components.
7. Coordinated with good bones.

## USECASE DIAGRAM:

In utilize cases graph within the Unified Modelling Languages (UML) may be a sort of behavioural chart characterized by and made from a Use-case examination. Its reason is to display a graphical outline of the usefulness given by a framework in terms of on-screen characters, their objectives (represents a utilize casee), and in any conditions among those utilized case. Most reasoned utilize case chart appears what framework capacities are designed.



**Figure 8** Picture shows the USECASE Diagram

## CLASS DIAGRAM:

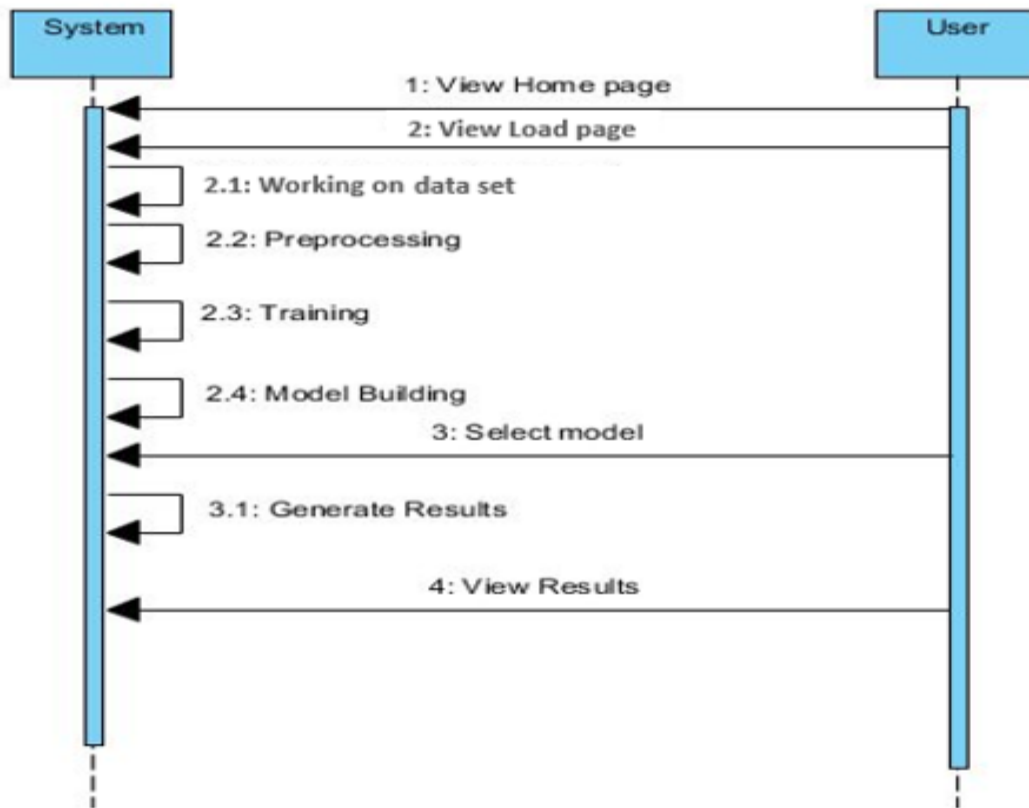
In program design, the Unified Modeling Language (UML) class diagram becomes a type of passive structural diagram that represents the structure of the framework by showing the classes of the system, their properties, operations (or strategies), etc. and their connections. You may. between classes. Describes which lessons contain data about computer program projects.



**Figure 9 Picture shows the Class Diagram**

## SEQUENCE DIAGRAM:

A



**Figure 10 Picture shows the Sequence Diagram**

Arrangement chart) could be a unique inter-action of graph that appears how forms work with 1 an-other and in what arrange. It may be a develop of a Message Grouping Charrrt. Grouping graphs are now and then called occasion graphs, occasion scenarios, and timming charts

## COLLABORATION DIAGRAM:

In collaboration chart the strategy call grouping is demonstrated by a few numbering strategy as appeared underneath. The number shows how the strategies are known one after-another. We have taken the same arrangement administration framework to portray the collaboration chart. The strategy calls are comparative to that of an arrangement chart. But the distinction is that the arrangement chart does not depict the question organization while the collaboration chart appears the process organization.

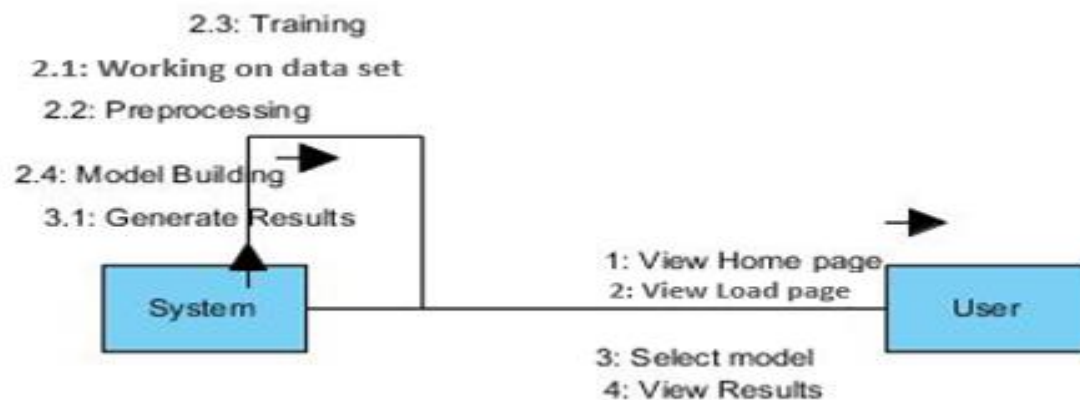
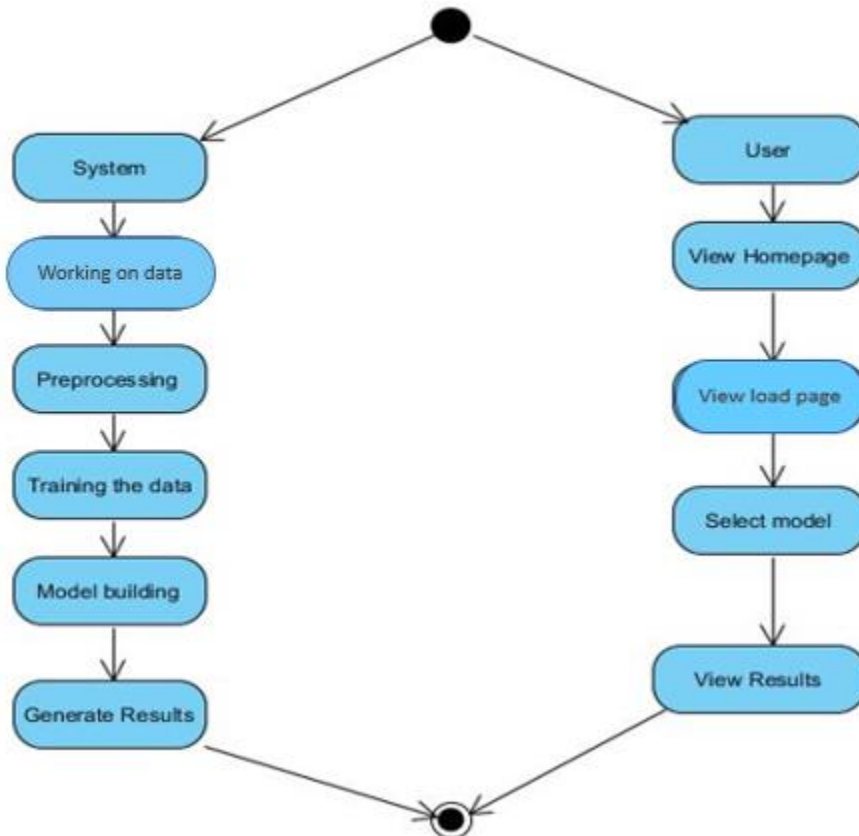


Figure 11 Picture shows the Collaboration Diagram



## ACTIVITY DIAGRAM:

Advancement charts are the graphic representation of work-flows that works out with bolster for choices, emphasis and concurrencies. Within the Bound together Displaying Dialect, gesture graph can be made use to portray the commerce and operating step-by-step in the frameworks.



**Figure 12** Picture shows the Activity Diagram

## DEPLOYMENT DIAGRAM

The dispatch graph talks about the sending view of a frame. It is associated with the component graph. Since components are sent with a dispatch tag. A sending card includes the hub. Hubs are nothing but hard physical products used to submit applications. .



**Figure 13 Picture shows the Deployment Diagram**

## COMPONENT DIAGRAM

Component diagrams are often created to illustrate the execution of subtle elements and ensure that all aspects of the desired functionality of the system are underpinned by planned improvements.

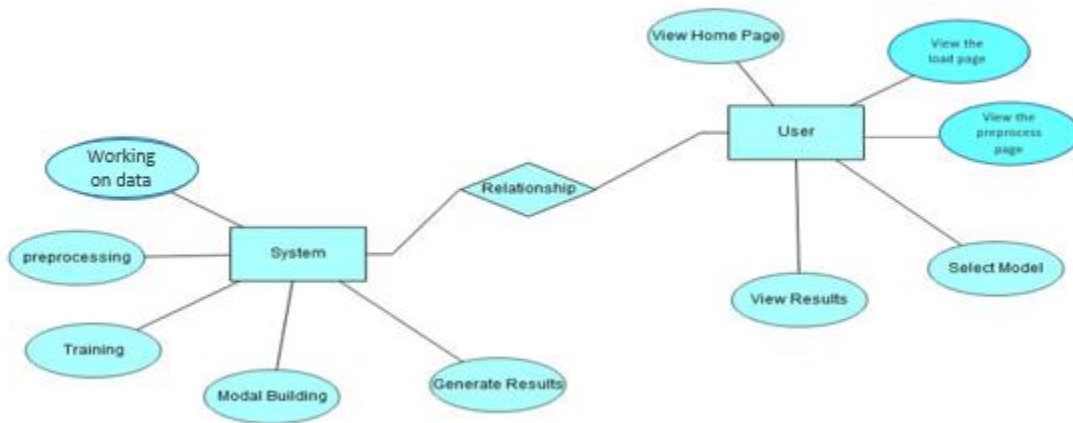
.



**Figure 14 Picture shows the Component Diagram**

## ER DIAGRAM

The entity-relationship (ER) model describes the structure of the database using diagrams called entity-relationship (ER) diagrams. An ER model is a database design or blueprint that can be deployed as a database. The main components of the E-R model are Entity Sets and Relationship Sets. An ER diagram shows the relationships between a set of entities. A material set is a collection of comparable materials, and these materials can have properties. In DBMS terminology, an entity is a table or attribute of a table in the database. Thus, by showing the relationships between tables and their attributes, an ER diagram shows a logical structure.



**Figure 15** Picture shows the ER Diagram

## DFD DIAGRAM:

Information stream charts (DFDs) have become the traditional way of visualizing data streams within frameworks. A complete and well-defined DFD can represent a large amount of framework requirements graphically. It can be manual, mechanized, or a combination of both. It shows how data enters and exits frames, what modifies the data, and where the data is stored. The purpose of the DFD is to indicate the scope and limits of the overall framework. It can be used as a communication device between framework researchers and anyone playing a role within the framework, and serves as a starting point for updating the framework.

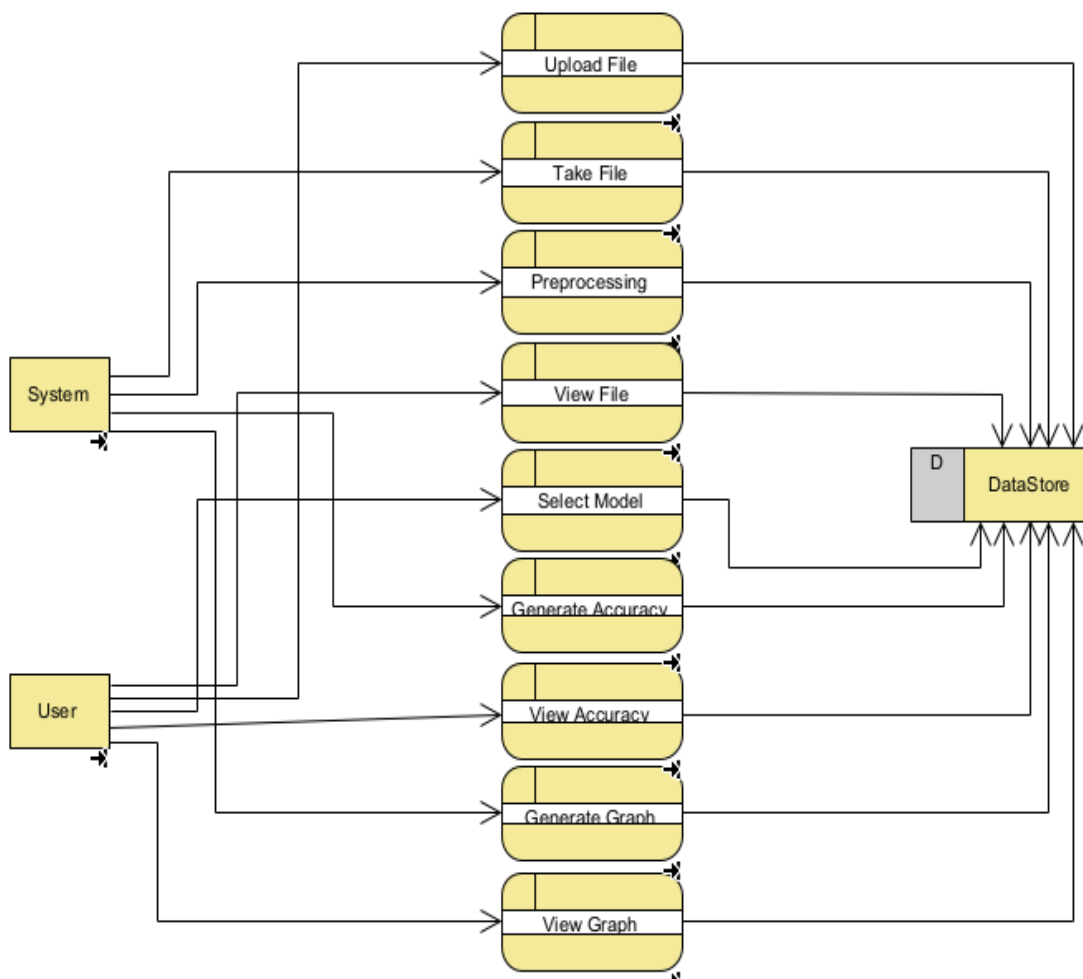
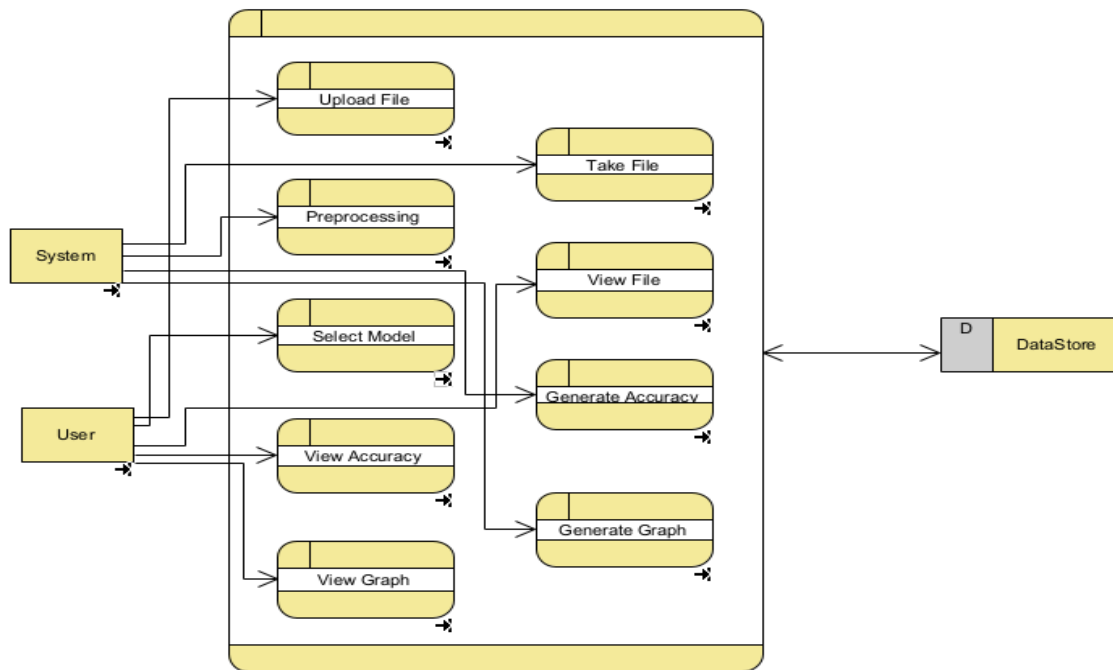


Figure 16 Picture DFD Diagram-1



**Figure 17 Picture DFD Diagram-2**

## INTRODUCTION TO PYTHON

### **What is handScriptit?**

So far, we've focused on Python's interactive programming capabilities. This is a very useful feature that allows you to type a program and run it immediately in interactive mode.

### **Scripts are reusable.**

A scrip is essentially a content data set containing the instructions that make up a Python program. Once you have created your script, you can run it as many times as you want without having to retype it each time. anyone can change scriptss.

Perhaps, and more importantly, you can use the content editor to create different forms of script by changing articulations from one record to the next. Any customizations can be made at this point. This makes it easy to create your own programs with minimal writing.

Almost any content editor can be used to create Python script records. If you prefer, you can use Microsoft Scratch Pad, Microsoft WordPad, Microsoft Word, or any other word processing program.

### **Differ-ence between a script and a program**

#### **Script:**

Scriptts are unmistakable from the center coode of the appli-cation, which is ordinarily composed in a distinctive dialect, and are frequently made or at slightest altered by the enduser.

#### **NumPy**

Numpy's fundamental question is the homojeneous multi-dimensional cluster. It could be a table of components (ordinarily numbers), all of the same sort, ordered by a tuple of positive integrability. In unfeelingly measurements are known tomahawks. The number of tomahawks is rank.

- Quick cluster functions
- 2D clusters, mullti-D clusters, straight polynomial math etc.

## **Inheritance**

Inheritance is a fundamental concept in Object-Oriented Programming (OOP) that allows new classes (called "child classes" or "subclasses") to be defined based on existing classes (called "parent classes" or "superclasses") will do so. . A child class inherits the attributes and methods of the parent class and can use, modify, or extend them directly.

ChildClass is the new class being created and ParentClass is the existing class that the child class inherits from. A child class can have additional attributes and methods of its own, in addition to those inherited from the parent class.

When an object is created from a subclass, it inherits the attributes and methods of both the subclass and the superclass. If both a parent class and a child class define a method, the child class method overrides the parent class method. This is called method over-riding.

In Python, a class can inherit from multiple parent classes (called multiple inheritance). A class definition can specify multiple parent classes separated by commas.

In addition to inheriting attributes and methods, child classes can also add new attributes, override parent class methods, and call parent class methods using the `super()` function. increase.

## **Exceptions**

I've talked approximately exemptions some time recently but presently I will conversation around them in profundity. Basically, special cases are occasions that alter program's stream, either intentioned or due to blunders.

## Model Building

**Model Development:** Construct the predictive model using the selected machine learning algorithm. This involves training the model on a portion of the data set, typically using techniques like supervised learning, and evaluating its performance on the remaining portion. Performance metrics such as accuracy, precision, recall, and F1- score can be used to assess the model's effectiveness in predicting heart disease.

Here is a general outline of the model development process:

- **Define the Problem:** Clearly understand the problem at hand, defining objectives, task type (classification, regression, etc.), target variable, and available input features.
- **Gather and Prepare Data:** Collect relevant data for the problem and preprocess it as necessary. Perform data cleaning, handle missing values, address outliers, and transform data into a suitable format for modeling.
- **Split Data:** Divide the dataset into training, validation, and testing sets. The training set is used for model training, the validation set for hyperparameter tuning and model selection, and the testing set for final evaluation.
- **Select Model:** Choose an appropriate machine learning or statistical model based on problem type, data characteristics, and specific requirements. Consider both traditional algorithms (e.g., Random Forest).
- **Train the Model:** Use the training set to train the chosen model. Feed input features and target variables, adjusting model parameters to minimize error or maximize the objective function.
- **Validate and Optimize:** Evaluate the model's performance on the validation set. Adjust hyperparameters (e.g., learning rate, regularization) using techniques like grid search or random search. Continuously assess performance and make necessary adjustments to improve accuracy



and generalization.

- **Evaluate Model Performance:** Assess the model's performance on the testing set using appropriate evaluation metrics (e.g., accuracy, precision, recall, mean squared error) to gauge its generalization to unseen data.
- **Refine and Iterate:** Analyze results and insights gained, refining the model if needed by iterating through steps 4 to 7. This may involve revisiting data preparation, feature engineering, or exploring different model architectures.
- **Deploy the Model:** Deploy the model in a production environment or integrate it into your application. Ensure scalability, efficiency, and appropriate handling of new data.
- **Monitor and Maintain:** Continuously monitor the model's performance in real- world scenarios. Update the model as needed to adapt to changing data patterns or evolving requirements. Maintain documentation of the model development process for future reference.

**Model Optimization:** Fine-tune the model's parameters and hyperparameters to improve its performance. This optimization process aims to find the best combination of parameter values that maximizes the model's accuracy. Techniques such as grid search or Bayesian optimization can be employed to systematically explore the parameter space and find the optimal configuration.

Here are some common techniques and considerations for model optimization:

- **Hyperparameter Tuning:** Adjusting the hyperparameters of the model can have a significant impact on its performance. Techniques like grid search, random search, or Bayesian optimization can be used to find the optimal combination of hyperparameters that maximizes performance.

- **Feature Engineering:** Enhancing the model's predictive power through feature engineering involves creating new features or transforming existing ones. Techniques like scaling, normalization, log transformations, polynomial features, interaction terms, or feature selection can be employed to extract relevant information from the data and improve the model's ability to capture patterns
- **Model Architecture Selection:** Choosing the appropriate architecture is crucial, especially for deep learning models. Experimenting with different architectures, layer types, activation functions, or network depths can help identify the one that best suits the problem. Transfer learning, where pre-trained models are fine-tuned for specific tasks, can also be beneficial.
- **Regularization Techniques:** Regularization methods prevent overfitting by reducing model complexity and improving generalization. Techniques like L1 or L2 regularization, dropout, or early stopping can be applied to achieve better performance.
- **Ensemble Methods:** Ensemble methods combine multiple models to make predictions, reducing variance and capturing complex relationships in the data. Techniques like bagging (e.g., random forests) or boosting (e.g., AdaBoost, gradient boosting) can improve performance. Experimenting with different ensemble techniques and sizes can lead to optimal results.
- **Batch Normalization:** Batch normalization is a technique commonly used in deep learning models to improve training efficiency and generalization. It normalizes the activations of each layer within a mini-batch during training, reducing internal covariate shift and stabilizing the learning process.
- **Early Stopping:** Early stopping is a regularization technique that halts the training process based on a validation metric. It prevents overfitting by monitoring validation performance and stopping training when performance deteriorate.

- **Model Compression:** Model compression techniques aim to reduce model size, making it more efficient and faster to deploy. Techniques like pruning, quantization, or knowledge distillation can reduce parameters or precision while maintaining performance.

- **Parallel Processing:** Utilizing parallel processing techniques, such as distributed computing or GPU acceleration, speeds up training and inference. This improves scalability and reduces training time for larger datasets or complex models.

- **Cross-Validation:** Cross-validation assesses model performance and generalization by iteratively training and evaluating on different subsets of data. It provides a robust estimate of performance and guides optimization decisions.

**Model Validation:** Validate the performance and generalizability of the developed model using a new and independent data set. This process ensures that the model is not overfitting to the training data and can accurately predict heart disease in real-world scenarios. Techniques like data set splitting or cross-validation can be used to validate the model's performance and assess its robustness.

Here are some key aspects and techniques involved in model validation:

- **Validation Set:** Prior to training the model, a portion of the available data is typically reserved as a validation set. This dataset is used to assess the model's performance during training and make informed decisions regarding model adjustments or hyperparameter tuning. It is important for the validation set to accurately represent the overall data distribution.

- **Evaluation Metrics:** Select appropriate evaluation metrics based on the problem type and objectives. Common metrics for classification tasks include accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (ROC AUC). Regression tasks often

employ metrics like mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), or R-squared. Choose metrics that align with the specific problem and provide meaningful insights into the model's performance.

- **Cross-Validation:** Utilize cross-validation techniques to estimate the model's performance. Cross-validation involves partitioning the data into multiple subsets or folds and iteratively training and evaluating the model on different combinations of these subsets. Popular methods include k-fold cross-validation, stratified k-fold cross-validation, or leave-one-out cross-validation. Cross-validation provides a robust estimation of the model's performance and aids in assessing its stability and generalization ability.

- **Overfitting and Underfitting Detection:** Monitor the model's performance on both the training and validation sets to identify overfitting and underfitting. Overfitting occurs when the model excessively fits the training data but struggles to generalize to unseen data, while underfitting arises when the model is too simple to capture the underlying patterns. Adjustments such as regularization, model complexity, or hyperparameter tuning can help mitigate these issues.

- **Model Comparison:** Compare the performance of different models or variations to determine the best-performing one. This may involve evaluating various algorithms, architectures, hyperparameter settings, or feature selections. Consider both statistical significance and practical significance when comparing models.

- **Validation on Unseen Data:** Once the model selection and optimization process is complete, evaluate the model's performance on a separate testing set that has not been used during training or validation. This ensures an unbiased assessment of the model's generalization ability to new, unseen data. The testing set should accurately represent the overall data distribution and be reserved solely for final.

- **Model Maintenance and Monitoring:** Continuously monitor the model's performance in real-world scenarios or production environments. Track key performance metrics, identify potential issues or drifts, and update the model as necessary. Regularly re-evaluate and validate the model using new data to ensure its ongoing reliability and accuracy.

**Model Interpretation:** Interpret the results of the trained model to gain insights into the key features associated with heart disease risk. This involves analyzing feature importance measures, such as the weights in logistic regression or the importance scores in decision trees. Techniques like partial dependence plots or SHAP (Shapely Additive Explanations) values can provide a deeper understanding of the relationships between the features and the predicted outcomes.

**Feature Importance:** Evaluate the significance of input features in the decision-making process of the model. Techniques such as permutation importance, feature importance from tree-based models, or coefficients from linear models offer insights into the impact of features on predictions. Visualize or rank feature importance to understand the relative importance of different features.

- **Partial Dependence Plots:** Analyze how the model's predictions vary with changes in specific input features while holding other features constant. Partial dependence plots reveal the relationship between individual features and the target variable. They help identify non-linear relationships, feature interactions, and potential outliers.

- **Instance Explanations:** Explain individual predictions to provide specific insights. Techniques like LIME (Local Interpretable Model-Agnostic Explanations) or SHAP (SHapley Additive exPlanations) generate local explanations for individual predictions. These methods highlight the contribution of each feature to the prediction, enabling interpretability at the instance level.

- **Rule Extraction:** Extract interpretable rules or decision trees that mimic the behavior of complex models. Decision trees or rule-based models offer transparent and easily understandable representations for non-experts. Rule extraction techniques simplify complex models while preserving accuracy, facilitating interpretation.

- **Model Visualization:** Visualize the model's structure, decision boundaries, or internal representations to gain insights into its functioning. Examples include visualizing decision trees, activation maps in deep learning models, or t-SNE plots to understand information processing and data clustering.

- **Global Explanations:** Obtain an overall understanding of the model's behavior through global explanations. Techniques like SHAP values, feature importance ranking, or aggregated partial dependence plots summarize the model's behavior across the entire dataset, highlighting consistent patterns or trends.

- **Model Ablation:** Assess the impact of removing or modifying specific features on the model's predictions. Ablation studies help identify crucial features or test the model's sensitivity to variables. Systematically removing or altering features and observing changes in predictions provides insights into feature importance and influence.

- **Bias and Fairness Analysis:** Evaluate the model for potential biases or unfairness. Investigate if the model's predictions disproportionately favor or discriminate against certain groups or demographics. Analyze predictions across different subgroups or protected attributes to identify and mitigate biases.

- **Documentation and Communication:** Document the interpretation process and the insights gained. Clearly communicate the model's behavior, limitations, and key findings to stakeholders, users, or regulators. Use visualizations, summaries, or interactive tools to make the interpretation accessible and understandable to a wide audience.

**Evaluation:** Continuously evaluate the deployed model's performance in real-world

- **Evaluation Metrics:** The selection of suitable evaluation metrics depends on the problem type and objectives. Classification tasks commonly use metrics such as accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (ROC AUC). Regression tasks employ metrics like mean squared error (MSE), root mean squared error (RMSE), mean absolute

error (MAE), or R-squared. It is important to choose metrics that align with the problem and offer meaningful insights into the model's performance.

- **Training and Testing Sets:** Splitting the available data into training and testing sets is essential. The training set is used to train the model, while the testing set evaluates its performance on unseen data. The testing set should be representative of the overall data distribution and not utilized during model development or optimization to avoid biased evaluation. Consider class imbalance or stratification if applicable.

- **Cross-Validation:** Cross-validation estimates a model's performance by partitioning the data into multiple subsets or folds. Iteratively, the model is trained and evaluated on different combinations of these subsets. Common methods include k-fold cross-validation, stratified k-fold cross-validation, or leave-one-out cross-validation. Cross-validation provides a robust estimate of the model's performance, assessing its stability and generalization ability.

- **Confusion Matrix:** Classification tasks benefit from utilizing a confusion matrix to visualize the model's performance in predicting different classes. It provides insights into true positives, true negatives, false positives, and false negatives, enabling calculation of various evaluation metrics like accuracy, precision, recall, and F1 score.

## **SYSTEM STUDY**

### **FEASIBILITY STUDY**

The possibility of the venture is dissected in this stage and trade proposition is put forward with an awfully common arrange for the extend and a few taken a toll gauge. Amid framework investigation the achievability ponder of the propped frame-work is to be carried out. This is often to guarante that the propose frame-work isn't a burden to the company. Reachability checking is fundamentally important to have a constant understanding of the framework's most important assumptions.

Three key contemplations included within the possibility examination are

- Conservative Possibility
- Specialized Achievability
- SOCIAL Possibility

#### **Conservative Possibility**

This review is done to consider the financial impact of the framework on the organization. There is a limit to how much money companies can invest in researching and improving frameworks. Consumption must be encouraged. This has therefore been achieved as the framework has also been created within the budget and most of the advancements used are fully available. In other words, we had to procure a custom-made product.

#### **Technical Accessibility**

This review is performed to check the technicality, ie the technical requirements of the framework. Each framework created should not impose large demands on the special resources available. This puts a higher demand on available special resources. This imposes high demands on the client. There must be some necessity in the framework as it is created, because running this framework requires some negligible or invalid changes, so to speak.

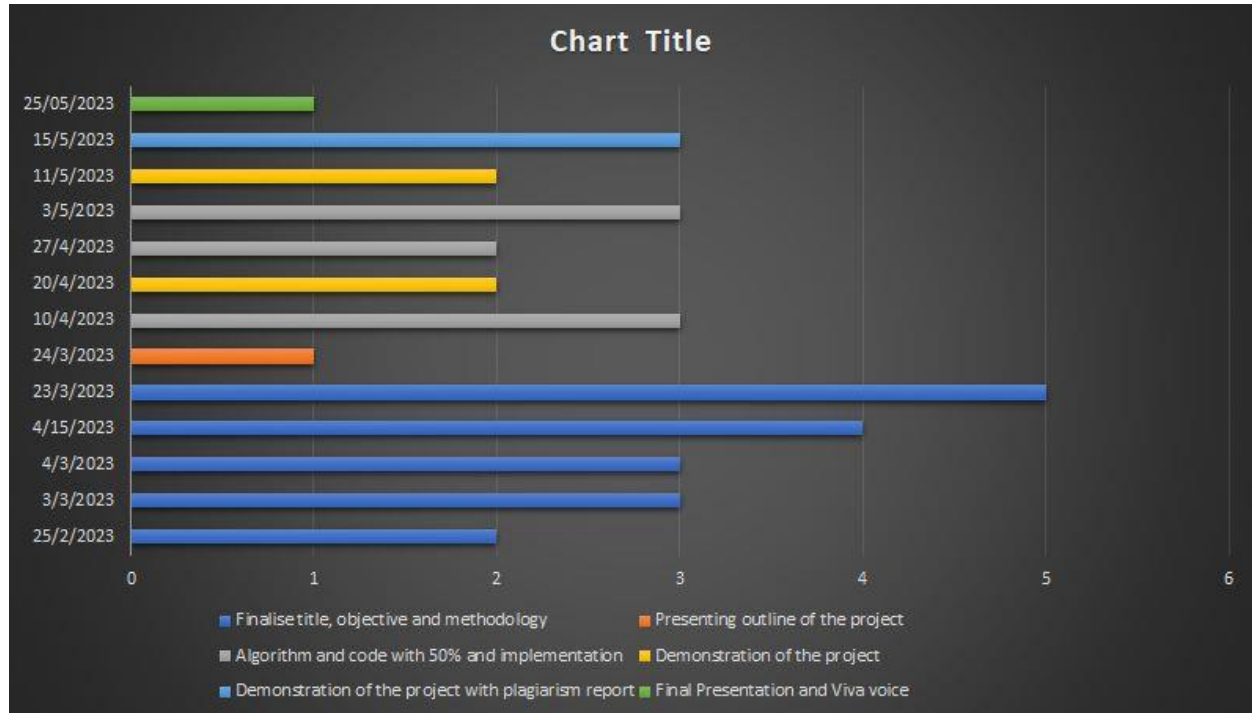
#### **Social Opportunities**

A point of consideration is to verify the extent of customer verification of the system. This is combined with a strategy of planning for the wise use of the system by the customer. The client should not feel undermined by the system and should perceive it as a requirement as a step.



## CHAPTER-8

### TIMELINE FOR EXECUTION OF PROJECT (GHANTT CHART)



**Figure 18 TIMELINE FOR EXECUTION OF PROJECT**

The project's timeline is displayed on the GANNT chart. It is an effective tool for illustrating a project timetable. It aids in the identification of various project tasks and offers a summary of the project's progress. According to our project's timeline, the project's preparation phase, which deals with determining the project's timeline, budget, and resources required, began in the month of February. A project management plan that specifies the project's goals, objectives, timetables, and deliverables is the result of the planning phase. Identifying potential obstacles, dangers, and challenges that could have an impact on the project's success as well as creating backup plans to deal with them may also be part of the preparation phase. Second, software analysis is a key component of specification. and hardware components required for proceeding with the project. The development phase of a project is where the planning and design begin to take shape. It is a period focused on creating the means to achieve the desired outcome and involves various stages of testing and refinement. This is also where the core team members begin working on their

individual tasks, bringing them together to build the full project. During this phase, it is important to closely monitor the progress of each task to ensure all teams are meeting their respective deadlines.

To start, planning offers an overview of precise project planning that is broken up into several pieces and evaluated or designed in accordance with the needs of the developer. The entire technical aspect of a project is dealt with in website design, where the project developer uses software or hardware components and technical skills to create a website in accordance with project specifications. In order to make the project technically sound and well-balanced from the perspective of the customer, development is also a technical aspect of the project in which the designer goes in-depth to gather packages, frameworks, etc. Before being introduced to the market or used in a demo, the project is tested for alterations following the successful completion of development. After the project is over, documentation is produced in which details about the complete project are recorded from beginning to end, specifically defining the work of the project.

## **CHAPTER-9**

### **TESTING**

The purpose of testing is to identify errors or flaws. Testing involves systematically searching for any possible faults or weaknesses in a work item, such as components, sub-assemblies, assemblies, or the final product. It is the process of executing a computer program to ensure that the software meets its requirements and fulfills user expectations without failing in an undesirable manner. Various types of tests exist, each catering to specific testing needs.

#### **TYPES OF TESTS**

##### **Unit Testing:**

Unit testing involves designing test cases that validate the correctness of the internal program logic and ensure that program inputs produce valid outputs. It covers all decision branches and internal code flow to ensure validation. Unit testing is performed on individual program units within the application. It occurs after the completion of a single unit and before integration. Typically, it is a non-intrusive and basic testing approach based on knowledge of the unit's development. Unit tests execute fundamental tests at the component level, focusing on specific business processes, applications, or system configurations. These tests ensure that each unique path of a business process adheres precisely to the documented specifications, with clearly defined inputs and expected outcomes.

##### **Integration Testing:**

Integration tests are designed to verify the proper functioning of integrated program components as a unified program. Testing is event-driven and focuses more on the overall outcome of screens or sections. Integration tests demonstrate that although the components were individually validated through successful unit testing, their integration is correct and consistent. Integration testing aims to uncover issues arising from the combination of components.

##### **Functional Testing:**

Functional tests provide systematic evidence that the tested functions are accessible according to business and technical requirements, system documentation, and user manuals. Functional testing focuses on the following aspects:

Valid Input: Acceptance of identified classes of valid input.

Invalid Input: Rejection of identified classes of invalid input.

Functions: Execution of identified functions.

Output: Generation of identified classes of application outputs.

Systems/Procedures: Activation of interrelated systems or procedures.

The organization and arrangement of functional tests are focused on requirements, key functionalities, or specific test cases. Additionally, comprehensive coverage related to identifying business process flows, data domains, predefined processes, and progressive forms should be considered for testing. After functional testing is completed, additional tests are identified, and the effectiveness of existing tests is determined.

### **System Testing**

System testing ensures that the entire integrated law meets the conditions and produces known and predictable issues. An illustration of system testing is a script- grounded system integration test. System testing is based on detailed process descriptions and flows, with a focus on predefined process connections and integration points.

### **White Box Testing**

White Box Testing is a type of testing where the software tester possesses an understanding of the internal mechanisms, structure, and programming language of the software, or at least its intended functionality. It's used to test areas that can not be penetrated at a black box position.

### **Black Box Testing**

Black Box Testing is the testing of a program without any knowledge of its internal workings, structure, or language. Black box tests, like utmost other types of tests, must be designed grounded on a definitive source document, similar as a specification or conditions document. It treats the program under test as a black box, where you can not" see" into it. The test provides inputs and observes labors without considering how the program operates.

### **Unit Testing**

Unit testing is generally conducted as part of a combined law and unit test phase in the software development lifecycle, although it isn't uncommon for rendering and unit testing to be conducted as separate phases.

**Test Strategy and Approach:**

For field testing, physical testing will be conducted, and detailed functional tests will be designed.

**Test Objectives:**

Ensure proper functioning of all field sections.

Activate pages from identified links.

Ensure timely display of entry screens, messages, and responses.

**Features to be Tested:**

Validate that entries are in the correct format.

Prevent duplicate entries.

Verify that all links direct the user to the appropriate page.

**Integration Testing:**

Program integration testing involves incrementally integrating two or more software components on a single platform to detect failures caused by interface defects. The goal of integration testing is to ensure error-free connections between components or software applications, such as components within a code or, at a higher level, program applications at the company level.

**Test Results:**

All mentioned test cases passed successfully without any defects encountered.

**Acceptance Testing:**

Client Acceptance Testing is a crucial stage in any project and requires significant involvement from the end user.

## OUT COMES

### Home Page

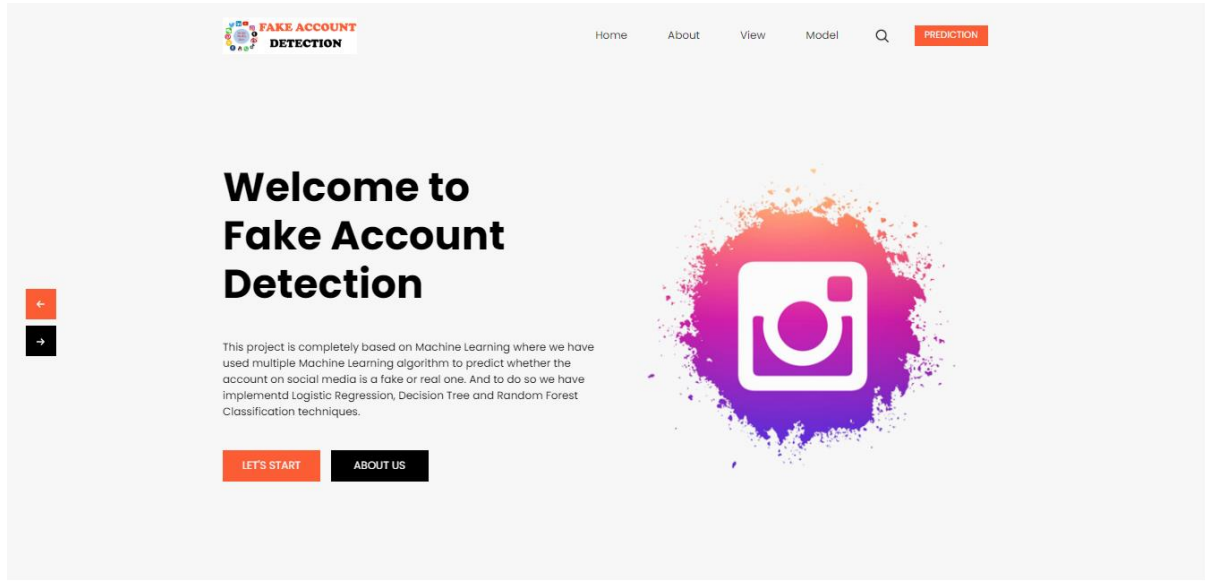


Figure 19 Home Page

### About Page

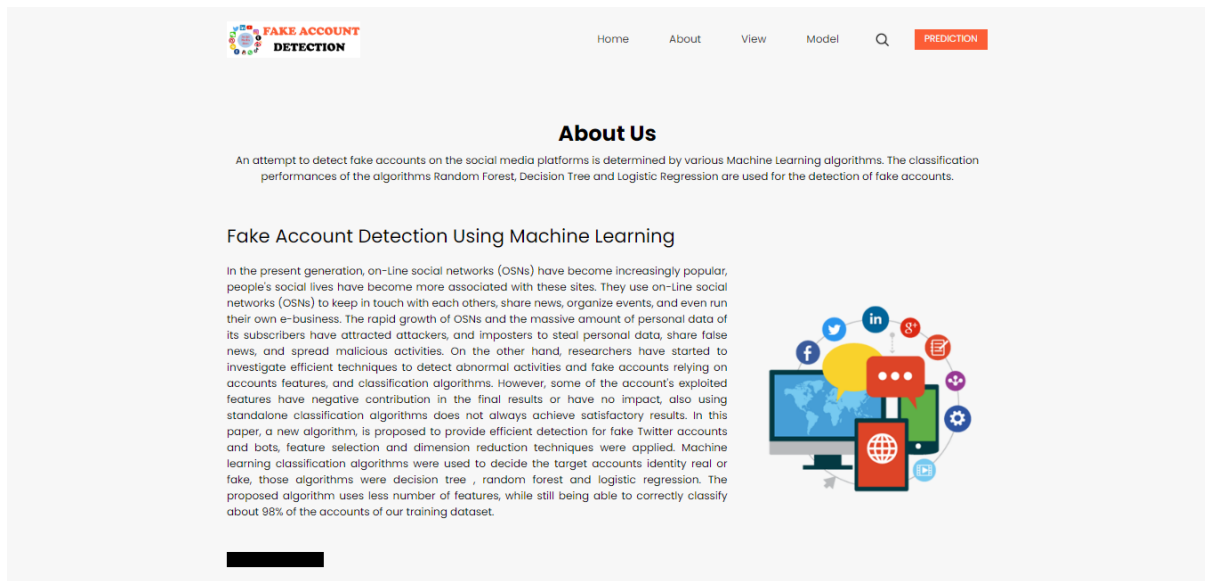
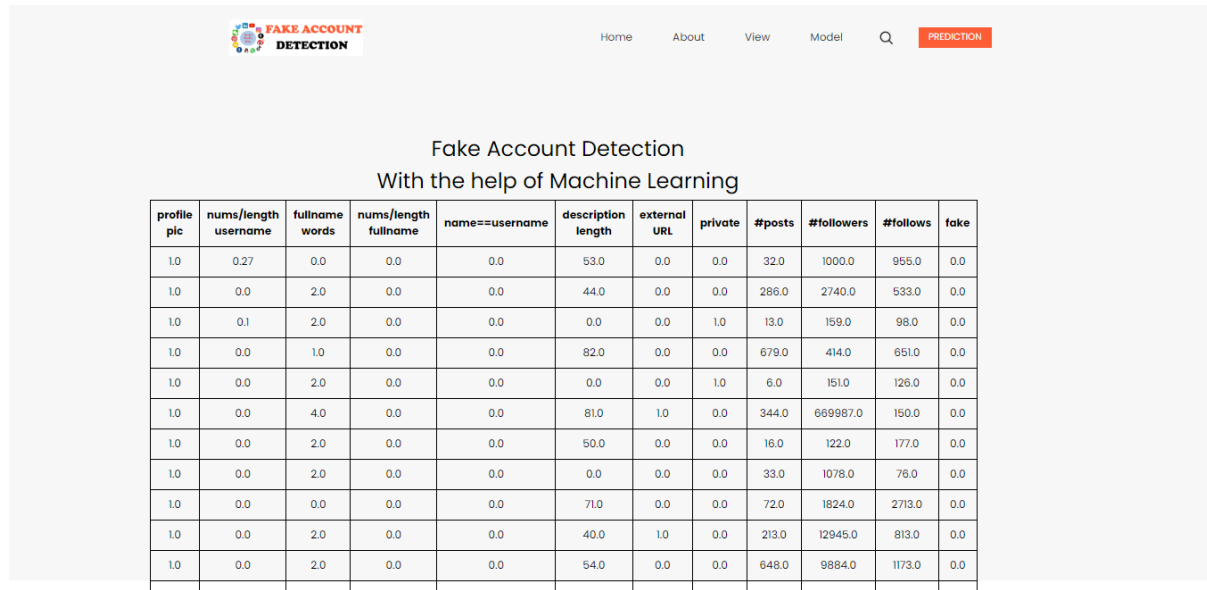


Figure 20 About Page

## Data

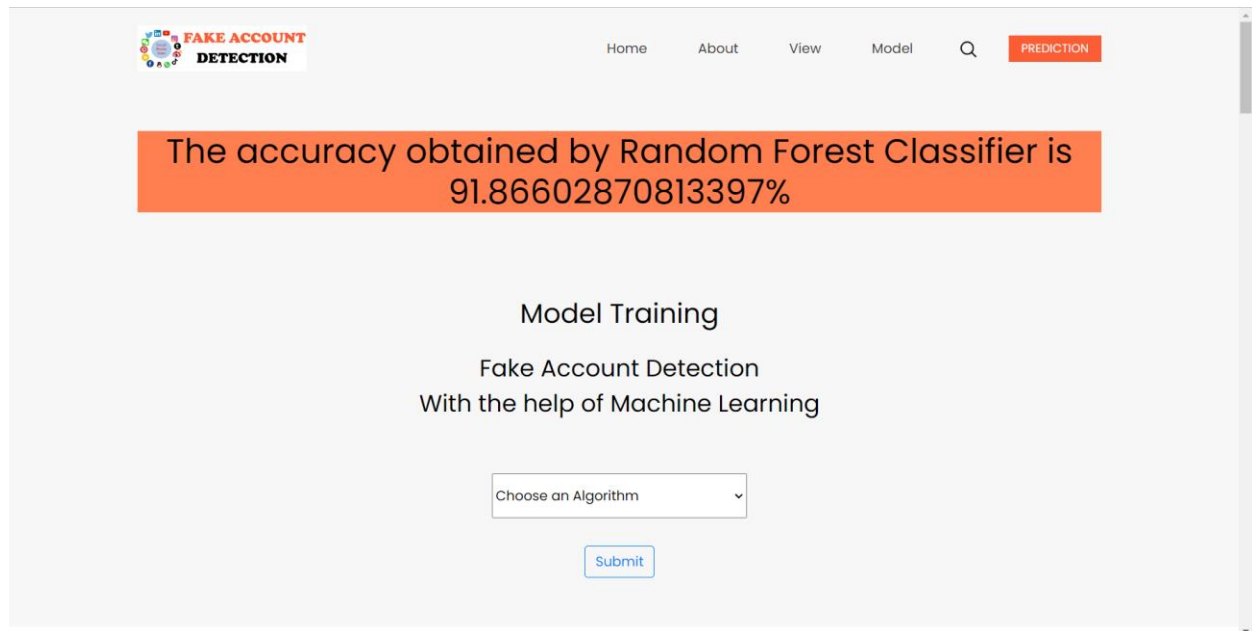


The screenshot shows the 'Data' page of the 'FAKE ACCOUNT DETECTION' web application. The page has a navigation bar with links for Home, About, View, Model, and a search icon, along with a 'PREDICTION' button. The main content area displays the title 'Fake Account Detection With the help of Machine Learning' above a table of 12 columns and 13 rows of data. The columns represent various user profile attributes and a final 'fake' classification column.

profile pic	nums/length username	fullname words	nums/length fullname	name==username	description length	external URL	private	#posts	#followers	#follows	fake
1.0	0.27	0.0	0.0	0.0	53.0	0.0	0.0	32.0	1000.0	955.0	0.0
1.0	0.0	2.0	0.0	0.0	44.0	0.0	0.0	286.0	2740.0	533.0	0.0
1.0	0.1	2.0	0.0	0.0	0.0	0.0	1.0	13.0	159.0	98.0	0.0
1.0	0.0	1.0	0.0	0.0	82.0	0.0	0.0	679.0	414.0	651.0	0.0
1.0	0.0	2.0	0.0	0.0	0.0	0.0	1.0	6.0	151.0	126.0	0.0
1.0	0.0	4.0	0.0	0.0	81.0	1.0	0.0	344.0	669987.0	150.0	0.0
1.0	0.0	2.0	0.0	0.0	50.0	0.0	0.0	16.0	122.0	177.0	0.0
1.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	33.0	1078.0	76.0	0.0
1.0	0.0	0.0	0.0	0.0	71.0	0.0	0.0	72.0	1824.0	2713.0	0.0
1.0	0.0	2.0	0.0	0.0	40.0	1.0	0.0	213.0	12945.0	813.0	0.0
1.0	0.0	2.0	0.0	0.0	54.0	0.0	0.0	648.0	9884.0	1173.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...

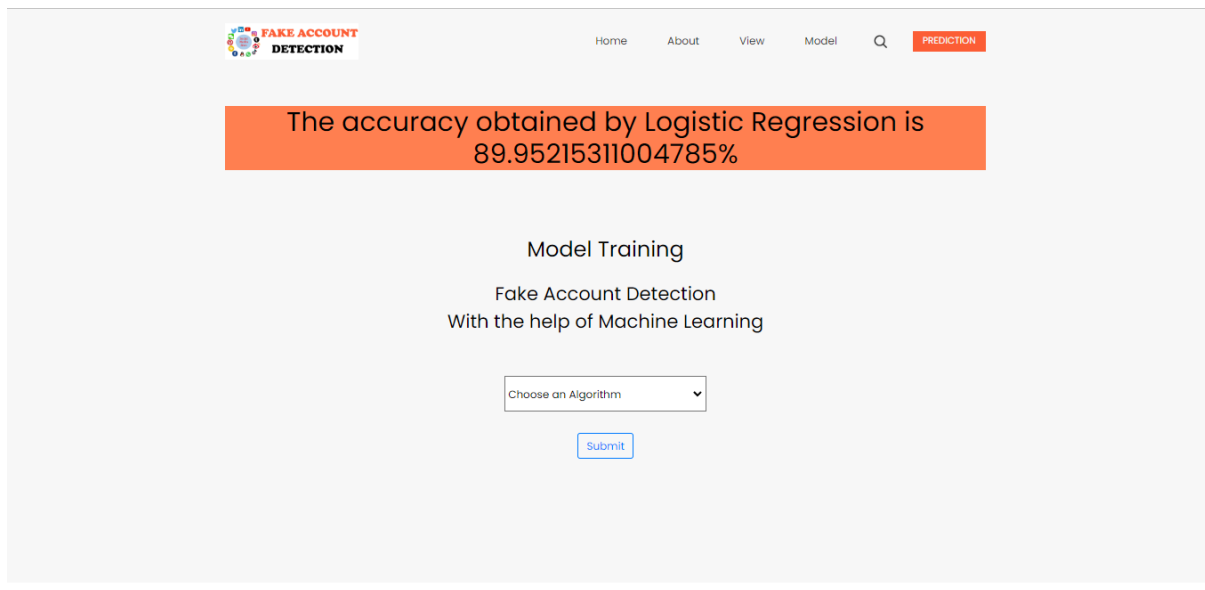
Figure 21 Data Page

## Model Selection



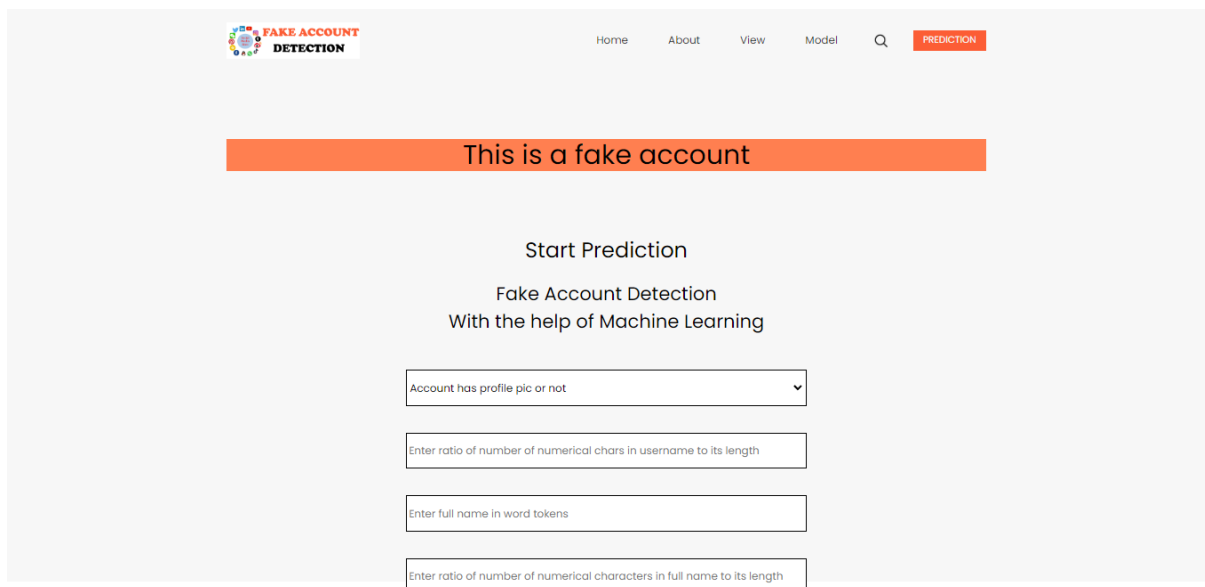
The screenshot shows the 'Model Selection' page of the 'FAKE ACCOUNT DETECTION' web application. The page features a navigation bar similar to the previous one. A large orange banner at the top displays the text: 'The accuracy obtained by Random Forest Classifier is 91.86602870813397%'. Below this, the title 'Model Training Fake Account Detection With the help of Machine Learning' is centered. A dropdown menu labeled 'Choose an Algorithm' is present, with a 'Submit' button below it.

Figure 22 Model Selection-1 Page



**Figure 23 Model Selection-2 Page**

## Prediction



**Figure 24 Prediction Page**



## **CHAPTER-10**

### **CONCLUSIONS AND FUTURE SCOPE**

The conclusions of the survey revealed that clients highly valued the webapp's modified features, real-time data, and user-friendly structure. According to clients, the website was easy to use and provided them with rapid access to check the account whether it is fake or genuine. The study also discovered a number of areas that needed to be improved, such as introducing more features, upgrading customization algorithms, along with providing better customer service.

The study demonstrated the potential for the Detecting fake account website to increase user satisfaction and engagement. The study's conclusions can aid in the development of future social media services web and mobile applications and improve the standard of services offered and overall user experience.

We may also incorporate a chatbot in this web application to improve the user experience.

After appropriate investigation and comparison, it was found that Arbitrary Woodland detailed the most extreme exactness. The exactness's for forecast can be encourage progressed

- Measure of dataset changes in future (directly an imperative).
- Lesson Dispersion of the target variable gets adjusted.

## **CHAPTER - 11**

### **REFERENCES**

1. E. Anupriya, N. Kumaresan, V. Suresh, S. Dhanasekaran, K. Ramprathap and P. Chinnasamy, "Fraud Account Detection on Social Network using Machine Learning Techniques," 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), Bhubaneswar, India, 2022, pp. 1-4, doi: 10.1109/ASSIC55218.2022.10088336.
2. E. Dubasova, A. Berdashkevich, G. Kopanitsa, P. Kashlikov and O. Metsker, "Social Network Users Profiling Using Machine Learning for Information Security Tasks," 2022 32nd Conference of Open Innovations Association (FRUCT), Tampere, Finland, 2022, pp. 87-92, doi: 10.23919/FRUCT56874.2022.9953858.
3. P. Harris, J. Gojal, R. Chitra and S. Anithra, "Fake Instagram Profile Identification and Classification using Machine Learning," 2021 2nd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2021, pp. 1-5, doi: 10.1109/GCAT52182.2021.9587858.
4. M. J. Ekosputra, A. Susanto, F. Haryanto and D. Suhartono, "Supervised Machine Learning Algorithms to Detect Instagram Fake Accounts," 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2021, pp.396-400,doi: 10.1109/ISRITI54043.2021.9702833.
5. K. Anklesaria, Z. Desai, V. Kulkarni and H. Balasubramaniam, "A Survey on Machine Learning Algorithms for Detecting Fake Instagram Accounts," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 141-144, doi: 10.1109/ICAC3N53548.2021.9725724.
6. A. Bhattacharya, R. Bathla, A. Rana and G. Arora, "Application of Machine Learning Techniques in Detecting Fake Profiles on Social Media," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-8, doi: 10.1109/ICRITO51393.2021.9596373.

7. G. Sansonetti, F. Gasparetti, G. D'aniello and A. Micarelli, "Unreliable Users Detection in Social Media: Deep Learning Techniques for Automatic Detection," in IEEE Access, vol. 8, pp. 213154-213167, 2020, doi: 0.1109/ACCESS.2020.3040604.
8. S. D. Muñoz and E. Paul Guillén Pinto, "A dataset for the detection of fake profiles on social networking services," 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2020, pp. 230-237, doi: 10.1109/CSCI51800.2020.00046.
9. N. Singh, T. Sharma, A. Thakral and T. Choudhury, "Detection of Fake Profile in Online Social Networks Using Machine Learning," 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), Paris, France, 2018, pp. 231-234, doi: 10.1109/ICACCE.2018.8441713.
10. S. Gheewala and R. Patel, "Machine Learning Based Twitter Spam Account Detection: A Review," 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2018, pp. 79-84, doi: 10.1109/ICCMC.2018.8487992.

# DETECTION OF FAKE SOCIAL MEDIA ACCOUNTS USING MACHINE LEARNING

1<sup>st</sup> Mr. Aarif Ahamed S

Department of Computer Science and  
Engineering  
Presidency University  
Bangalore, India.  
[aarif.ahamed@presidencyuniversity.in](mailto:aarif.ahamed@presidencyuniversity.in)

2<sup>nd</sup> Shravani. M

20191CSE0561  
Department of Computer Science  
and Engineering  
Presidency University  
Bangalore, India.  
[201910100323@presidencyuniversity.in](mailto:201910100323@presidencyuniversity.in)

3<sup>rd</sup> Narmada Gogineni

20191CSE0373  
Department of Computer Science  
and Engineering  
Presidency University  
Bangalore, India.  
[201910100227@presidencyuniversity.in](mailto:201910100227@presidencyuniversity.in)

4<sup>th</sup> Pagidela Venkata  
Mokshith Reddy

20191CSE0406  
Department of Computer Science  
and Engineering  
Presidency University  
Bangalore, India.  
[201910100960@presidencyuniversity.in](mailto:201910100960@presidencyuniversity.in)

5<sup>th</sup> Pachipulusu Akash  
Kumar

20191CSE0405  
Department of Computer Science  
and Engineering  
Presidency University  
Bangalore, India.  
[201910101141@presidencyuniversity.in](mailto:201910101141@presidencyuniversity.in)

6<sup>th</sup> Sarojini T Habbli

20191CSE0534  
Department of Computer Science  
and Engineering  
Presidency University  
Bangalore, India.  
[201910102074@presidencyuniversity.in](mailto:201910102074@presidencyuniversity.in)

**Abstract**— Online social networks (OSNs) have grown in popularity and are now more closely associated with people's social activities than ever before. They use OSNs to communicate with one another, exchange news, plan activities, and even operate their own online businesses. In order to steal personal information, spread malicious activities, and share false information, attackers and imposters have been drawn to OSNs because of their explosive growth and the vast quantity of personal data they collect from their users. On the other hand, academics have begun to look into effective methods for spotting suspicious activity and bogus accounts using account features and classification algorithms. However, some of the characteristics of the account that are exploited have an adverse effect on the results or have no effect at all. Additionally, using independent classification algorithms does not always produce satisfactory results. Three feature selection and dimension reduction techniques were used to create the decision tree in this paper, which is suggested to provide effective detection for fake Instagram accounts. To determine whether the target account was genuine or fake, Three

machine learning classification algorithms— Decision Tree, Random Forest, Logistic Regression were used.

**Keywords**—Decision Tree, Random Forest, Logistic Regression.

## Introduction

Online social network's (OSNs), such as Facebook, Twitter, LinkedIn, Google+ have become increasingly popular over last few years. People use OSNs to keep in touch with each other, share news, organize events, and even run their own e-business. For the period between 2014 and 2018 around 2.53 million U.S. dollars have been spent on sponsoring political ads on Facebook by non-profits. The open nature of OSNs and the massive amount of personal data for its subscribers have made them vulnerable to Sybil attacks. In 2012, Facebook noticed an abuse on their platform including publishing false news, hate speech, sensational and polarizing, and some others. However, online Social Networks (OSNs) have also attracted the interest of researchers for mining and analysing their massive amount of data, exploring and studying users behaviours as well as detecting their abnormal activities. In researchers have made a study to predict, analyse and explain

## CERTIFICATES



# PRESIDENCY UNIVERSITY

Presidency University Act, 2013 of the Karnataka Act No. 41 of 2013 | Established under Section 2(f) of UGC Act, 1956  
Approved by AICTE, New Delhi | Approved By BCI  
Bengaluru

**45 YEARS**  
OF ACADEMIC  
WISDOM

### Certificate of Presentation

\*\*\*\*\*

This is to certify that Mr./Ms. Shravani.M Under the Supervision of Dr./Mr./Ms. Mr. Aarif Ahmed from from PRESIDENCY UNIVERSITY, BANGALORE has successfully PRESENTED the paper at the National Conference on Recent Advancements and Challenges in Information Technology [NCRACIT-23] bearing the paper title Detection of fake social media accounts using machine learning and paper ID 275 held during 28th April 2023 - 29th April 2023.

 Dr. Gopal K Shyam Conference Chair, Prof. and Head, Dept. Of CSE	 Dr. Manujakshi B C Conference Chair, Associate Prof., Dept. Of CSE	 Dr. C Kalaarasan General Co-Chair, Associate Dean – SOCS&IS	 Dr. Md. Sameeruddin Khan General Chair, Dean – SOCS&IS
--	--	--	---



# PRESIDENCY UNIVERSITY


Presidency University Act, 2013 of the Karnataka Act No. 41 of 2013 | Established under Section 2(f) of UGC Act, 1956  
Approved by AICTE, New Delhi | Approved By BCI  
Bengaluru

**45 YEARS**  
OF ACADEMIC  
WISDOM

### Certificate of Presentation

\*\*\*\*\*

This is to certify that Mr./Ms. Narmada Gogineni Under the Supervision of Dr./Mr./Ms. Mr. Aarif Ahmed from from PRESIDENCY UNIVERSITY, BANGALORE has successfully PRESENTED the paper at the National Conference on Recent Advancements and Challenges in Information Technology [NCRACIT-23] bearing the paper title Detection of fake social media accounts using machine learning and paper ID 275 held during 28th April 2023 - 29th April 2023.

 Dr. Gopal K Shyam Conference Chair, Prof. and Head, Dept. Of CSE	 Dr. Manujakshi B C Conference Chair, Associate Prof., Dept. Of CSE	 Dr. C Kalaarasan General Co-Chair, Associate Dean – SOCS&IS	 Dr. Md. Sameeruddin Khan General Chair, Dean – SOCS&IS
--	--	--	---





# PRESIDENCY UNIVERSITY

Presidency University Act, 2013 of the Karnataka Act No. 41 of 2013 | Established under Section 2(f) of UGC Act, 1956  
Approved by AICTE, New Delhi | Approved By BCI  
Bengaluru



## Certificate of Presentation

\*\*\*\*\*

This is to certify that Mr./Ms. Pagidela Venkata Mokshith Reddy Under the Supervision of Dr./Mr./Ms. Mr. Aarif Ahmed from from PRESIDENCY UNIVERSITY, BANGALORE has successfully PRESENTED the paper at the National Conference on Recent Advancements and Challenges in Information Technology [NCRACIT-23] bearing the paper title Detection of fake social media accounts using machine learning and paper ID 275 held during 28th April 2023 - 29th April 2023.

Dr. Gopal K Shyam  
Conference Chair,  
Prof. and Head,  
Dept. Of CSE

Dr. Manujakshi B C  
Conference Chair,  
Associate Prof.,  
Dept. Of CSE

Dr. C Kalaiarasan  
General Co-Chair,  
Associate Dean – SOCS&IS

Dr. Md. Sameeruddin Khan  
General Chair,  
Dean – SOCS&IS



# PRESIDENCY UNIVERSITY

Presidency University Act, 2013 of the Karnataka Act No. 41 of 2013 | Established under Section 2(f) of UGC Act, 1956  
Approved by AICTE, New Delhi | Approved By BCI  
Bengaluru



## Certificate of Presentation

\*\*\*\*\*

This is to certify that Mr./Ms. Pachipulusu Akash Kumar Under the Supervision of Dr./Mr./Ms. Mr. Aarif Ahmed from from PRESIDENCY UNIVERSITY, BANGALORE has successfully PRESENTED the paper at the National Conference on Recent Advancements and Challenges in Information Technology [NCRACIT-23] bearing the paper title Detection of fake social media accounts using machine learning and paper ID 275 held during 28th April 2023 - 29th April 2023.

Dr. Gopal K Shyam  
Conference Chair,  
Prof. and Head,  
Dept. Of CSE

Dr. Manujakshi B C  
Conference Chair,  
Associate Prof.,  
Dept. Of CSE

Dr. C Kalaiarasan  
General Co-Chair,  
Associate Dean – SOCS&IS

Dr. Md. Sameeruddin Khan  
General Chair,  
Dean – SOCS&IS



# PRESIDENCY UNIVERSITY

Presidency University Act, 2013 of the Karnataka Act No. 41 of 2013 | Established under Section 2(f) of UGC Act, 1956  
Approved by AICTE, New Delhi | Approved By BCI  
Bengaluru



## Certificate of Presentation

\*\*\*\*\*

This is to certify that Mr./Ms. Sarojini T Habbli Under the Supervision of Dr./Mr./Ms. Mr. Aarif Ahmed from from PRESIDENCY UNIVERSITY, BANGALORE has successfully PRESENTED the paper at the National Conference on Recent Advancements and Challenges in Information Technology [NCRACIT-23] bearing the paper title Detection of fake social media accounts using machine learning and paper ID 275 held during 28th April 2023 - 29th April 2023.

Dr. Gopal K Shyam

Conference Chair,  
Prof. and Head,  
Dept. Of CSE

Dr. Manujakshi B C

Conference Chair,  
Associate Prof.,  
Dept. Of CSE

Dr. C Kalaiarasan

General Co-Chair,  
Associate Dean - SOCS&IS

Dr. Md. Sameeruddin Khan

General Chair,  
Dean - SOCS&IS

# PLAGIARISM REPORT

## DETECTION OF FAKE SOCIAL MEDIA ACCOUNTS USING MACHINE LEARNING

### ORIGINALITY REPORT

17%

SIMILARITY INDEX

10%

INTERNET SOURCES

4%

PUBLICATIONS

13%

STUDENT PAPERS

### PRIMARY SOURCES

1

[archive.interconf.center](http://archive.interconf.center)

Internet Source

1%

2

Submitted to Sreenidhi International School

Student Paper

1%

3

[www.ijert.org](http://www.ijert.org)

Internet Source

1%

4

Submitted to CSU, San Jose State University

Student Paper

1%

5

Sarah Khaled, Neamat El-Tazi, Hoda M. O. Mokhtar. "Detecting Fake Accounts on Social Media", 2018 IEEE International Conference on Big Data (Big Data), 2018

Publication

1%

6

[digi.landesbibliothek.at](http://digi.landesbibliothek.at)

Internet Source

1%

7

Submitted to Middle East College of Information Technology

Student Paper

1%