FORECASTING THE HEART DISEASE BY MACHINE LEARNING

Minor project report submitted
in partial fulfillment of the requirement for award of the degree of

Bachelor of Technology
in
Computer Science & Engineering

By

B.VAMSHIKRISHNA	(20UECS0112)	**(VTU 15583)**
M.KRISHNA SANDEEP	(20UECS0626)	**(VTU 15554)**
G.SRINUVASULU	(20UECS0363)	**(VTU 15986)**

Under the guidance of
DR.A.SURESH,M.E,Ph.D.,
ASSOCIATE PROFESSOR

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
SCHOOL OF COMPUTING

VEL TECH RANGARAJAN DR. SAGUNTHALA R&D INSTITUTE OF SCIENCE & TECHNOLOGY

(Deemed to be University Estd u/s 3 of UGC Act, 1956)
Accredited by NAAC with A++ Grade
CHENNAI 600 062, TAMILNADU, INDIA

April, 2023

FORECASTING THE HEART DISEASE BY MACHINE LEARNING

Minor project report submitted
in partial fulfillment of the requirement for award of the degree of

Bachelor of Technology
in
Computer Science & Engineering

By

B.VAMSHIKRISHNA	(20UECS0112)	**(VTU 15583)**
M.KRISHNA SANDEEP	(20UECS0626)	**(VTU 15554)**
G.SRINUVASULU	(20UECS0363)	**(VTU 15986)**

Under the guidance of
DR.A.SURESH,M.E,Ph.D.,
ASSOCIATE PROFESSOR

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
SCHOOL OF COMPUTING

VEL TECH RANGARAJAN DR. SAGUNTHALA R&D INSTITUTE OF SCIENCE & TECHNOLOGY

(Deemed to be University Estd u/s 3 of UGC Act, 1956)
Accredited by NAAC with A++ Grade
CHENNAI 600 062, TAMILNADU, INDIA

April, 2023

CERTIFICATE

It is certified that the work contained in the project report titled "FORECASTING THE HEART DISEASE BY MACHINE LEARNING" by "B.VAMSHIKRISHNA (20UECS0112), M.KRISHNA SANDEEP (20UECS0626), G.SRINUVASULU (20UECS0363)" has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

Signature of Supervisor
Dr.A.Suresh
Associate Professor
Computer Science & Engineering
School of Computing
Vel Tech Rangarajan Dr. Sagunthala R&D
Institute of Science & Technology
April, 2023

Signature of Head of the Department
Dr.M.S.Muralidhar
Computer Science & Engineering
School of Computing
Vel Tech Rangarajan Dr. Sagunthala R&Dy
Institute of Science & Technology
April, 2023

Signature of the Dean
Dr. V. Srinivasa Rao
Professor & Dean
Computer Science & Engineering
School of Computing
Vel Tech Rangarajan Dr. Sagunthala R&D
Institute of Science & Technology
April, 2023

DECLARATION

We declare that this written submission represents our ideas in own words and where others ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

B.VAMSHIKRISHNA

Date: / /

M.KRISHNA SANDEEP

Date: / /

G.SRINUVASULU

Date: / /

APPROVAL SHEET

This project report entitled " FORECASTING THE HEART DISEASE BY MACHINE LEARNING" by (B.VAMSHIKRISHNA (20UECS0112), (M.KRISHNA SANDEEP (20UECS0626), (G.SRINUVASULU (20UECS0363) is approved for the degree of B.Tech in Computer Science & Engineering.

Examiners **Supervisor**

Dr.A.Suresh, M.E,Ph.D.,

Date: / /

Place:

ACKNOWLEDGEMENT

We express our deepest gratitude to our respected **Founder Chancellor and President Col. Prof. Dr. R. RANGARAJAN B.E. (EEE), B.E. (MECH), M.S (AUTO),D.Sc., Foundress President Dr. R. SAGUNTHALA RANGARAJAN M.B.B.S.** Chairperson Managing Trustee and Vice President.

We are very much grateful to our beloved **Vice Chancellor Prof. S. SALIVAHANAN,** for provid- ing us with an environment to complete our project successfully.

We record indebtedness to our **Professor & Dean, Department of Computer Science & Engineering, School of Computing, Dr. V. SRINIVASA RAO, M.Tech., Ph.D.,** for immense care and encouragement towards us throughout the course of this project.

We are thankful to our **Head, Department of Computer Science & Engineering, Dr.M.S. MURALI DHAR, M.E., Ph.D.,** for providing immense support in all our endeavors.

We also take this opportunity to express a deep sense of gratitude to our Internal Supervisor **Dr.A.SURESH,M.E,Ph.D** for his cordial support, valuable information and guidance, he helped us in completing this project through various stages.

A special thanks to our **Project Coordinators Mr. V.ASHOK KUMAR, M.Tech., Ms.C.SHYAMALA KUMARI,M.E.,** for their valuable guidance and support throughout the course of the project.

We thank our department faculty, supporting staff and friends for their help and guidance to complete this project.

B.VAMSHIKRISHNA (20UECS0112)
M.KRISHNA SANDEEP (20UECS0626)
G.SRINUVASULU (20UECS0363)

ABSTRACT

Healthcare expenditures are overwhelming national and corporate budgets due to asymptomatic diseases including cardiovascular diseases. Therefore, there is an urgent need for early detection and treatment of such diseases Machine learning is one of the trending technologies which used in many spheres around the world including healthcare industry for predicting diseases. The aim of this study is to identify the most significant predictors of heart disease and predicting the overall risks by using logistic regression. Thus, binary logistic model which is one of the classification algorithm in machine learning is used in this study to identify the predictors. The early prognosis of cardiovascular diseases can aid in making decisions to lifestyle changes in high risk patients and in turn reduce their complications. According to recent survey by WHO organisation 17.5 million people dead each year. It will in- crease to 75 million in the year 2030. Medical professionals working in the field of heart disease have their own limitation, they can predict chance of heart attack upto 67% accuracy, with the current epidemic scenario doctors need a support system for more accurate prediction of heart disease. Machine learning algorithm and opens new door opportunities for precise predication of heart attack.

Keywords: Cardiovascular Diseases, Heart Disease, Logistic Regression, Machine Learning.

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligenc
ANN	Artificial Neural Network
BP	Blood Pressure
DNN	Deep Neural Network
KNN	K-Nearest Neighbors
LR	Logistic Regression
LVQ	Learning Vector Quantization
ML	Machine Learning
SVM	Support Vector Machine
UCI	University of california Irvine

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 Introduction

Heart disease is one of the leading causes of death worldwide, early detection and prevention can significantly improve the outcome. Machine learning, as a subfield of artificial intelligence, has shown great promise in the medical field in recent years, including in the prediction of heart disease. The goal of a heart disease prediction system using machine learning is to develop an algorithm that can accurately identify individuals who are at risk of developing heart disease, based on a set of well-defined risk factors. The system can then be used to provide personalized health advice, early screening and treatment, and ultimately reduce the number of heart disease-related deaths. Heart disease prediction systems using machine learning algorithms have become increasingly popular in recent years. These systems use various techniques to analyze large amounts of data related to heart disease risk factors and make predictions about an individual's likelihood of developing heart disease.

The system typically involves the collection of various medical data, such as demographic information, blood pressure, cholesterol levels, family history, and lifestyle factors like smoking and exercise habits. These data are then fed into ML algorithms that can analyze the data and identify patterns or relationships between different variables. The algorithm use this information to predict the likelihood of an individual developing heart disease.

The advantages of using machine learning algorithms for heart disease prediction are that they can quickly process large amounts of data, identify complex patterns that might not be easily detected by humans, and can be trained to improve their accuracy over time. By predicting heart disease risk early, these systems can help doctors and patients take preventive measures to reduce the risk of heart disease and improve health outcomes.

Overall, heart disease prediction systems using machine learning are a promising tool for improving cardiovascular health and reducing the burden of heart dis- ease. In this system, various machine learning techniques like decision trees, ran- dom forests, and artificial neural networks can be used to analyze and predict heart disease based on a large dataset of patient information and medical records.

The scope of the system includes the development of an accurate, reliable, and efficient prediction model, which can be used in real-world healthcare settings to improve the quality of life for millions of people worldwide.

1.2 Aim of the project

The aim of the project is the prediction of heart disease with high rate of accuracy. It identifies the most significant predictors of heart disease and predicting the risks by using logistic regression algorithm in machine learning. And also any non-medical employee can use this software and predict the heart disease and reduce the time complexity of the doctors. The aim of this study is to identify the most sig- nificant predictors of heart disease and predicting the overall risks by using logistic regression. Thus, binary logistic model which is one of the classification algorithmin machine learning is used in this study to identify the predictors. The early prog- nosis of cardiovascular diseases can aid in making decisions to lifestyle changes in high risk patients and in turn reduce their complications.

1.3 Project Domain

Machine Learning(ML) is a type of Artificial Intelligence(AI) that allows soft- ware applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine Learning algorithms use historical data as input to predict new output values. Machine Learning is important because it gives enterprises a view of trends in customer behaviour and business operational patterns, as well as supports the development of new products. Many of today's leading com- panies, such as Facebook, Google and Uber, make machine learning a central part of their operations. Machine Learning has become a significant competitive differen- tiator for many companies.

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. Machine Learning is one of the most exciting technologies that one would have ever come across. As it is evident fromthe name, it gives the computer that makes it more similar to humans. The abilityto learn. Machine Learning is actively being today, perhaps in many more places than one would expect. Machine Learning is an essential skill for any aspiring data analyst and data scientist, and also for those who wish to transform a massive amountof raw data into trends and predictions.

1.4 Scope of the Project

The scope of the project is to develop an efficient to predict the presence of the heart disease. The steps are the UCI dataset is used to predict the disease, the fea- tures are selected based on high positive correlation values with the target and used random order of data, the performance of the model is evaluated by five different training and testing ratio of dataset, To check the behaviour of the model with low to high training and testing data.

The scope of this project is restricted to discovering associations in data us- ing logistic regression, in order to improve the prediction rate. As a result, this project conducts a comparative analysis of the results of logistic regression in ma- chine learning algorithms. The trial results verify that Logistic Regression algorithm has achieved the highest accuracy of 89 percent compared to other machine learn-ing algoithms. To predicts the chances of Heart disease and classifies patient's risk level by implementing different data mining techniques. The scope of the projectis to effectively predict if the patient suffers from heart disease. The doctors enters the input values from the patient's health report. The data is divided into tested and trained data set and apply Machine learning Algorithm.

Chapter 2

LITERATURE REVIEW

Senthil kumar Mohan et al. [2020] [1], implemented the competent technique using hybrid machine learning methodology. The hybrid approach is combinationof random forest and linear method. The data set and subsets of attributes were collected for prediction.The subset of some attributes were chosen from the pre pro- cessed knowledge(data) set of cardiovascular disease. After pre processing, the hy- brid techniques were applied and diagnosis the cardiovascular disease.

Mamatha Alex P et al. [2019] [2], implemented KNN, Random Forest,Support Vector Machine. Comparing with the mentioned classification techniques in data mining to predict the higher accuracy for diagnosing the heart disease in Artificial Neural Network.

Aakash Chauhan et al. [2019] [3], described manual task that additionally helps in extracting the information directly from the electronic records.To generate strong association rules, they have applied frequent pattern growth association on patient's data set.This will facilitate in decreasing the amount of services and have shown the tangible majority of the rules helps with the best prediction of coronary diseases .

Bo Jin et al. [2020] [4], proposed a electronic health record (EHR) data from real-world datasets related to congestive heart disease to perform the experiment and predict the heart disease before itself. They tend to used one-hot encryption and word vectors to model the diagnosing events and coronary failure events victimization the essential principles of an extended memory network model. By viewing results, teey tend to reveal the importance of respecting the sequential nature of clinical records.

R.Sethukkarasi et al. [2019] [5], implemented a new fuzzy technique to diagnose the facts of the disease in patient reports. The generalized database is configured for decision-making from a reduced set of attributes, which is the output of the genetic algorithm. Four layered fuzzy neural networks are used for efficient modeling and

inference with time dependence under uncertainty.They also implemented a system for predicting heart disease and has applied three mining classification techniques: Decision trees, Naive Bayes Neural Networks. The results show neural networks are superior to decision trees and Naive Bayes.

Shadab Adam Pattekari et al. [2019] [6], implemented the Naive Bayesian method to develop decision support in the heart disease prediction system. The sys- tem finds hidden knowledge from past heart disease databases. This is the most effective model for predicting heart disease. The model can respond to complex queries and has its own strength in terms of ease of model interpretation,access to more information and accuracy

Sudha et al. [2018] [7], described algorithms such as Naive Bayes, Decision tree, and Neural Network to detect stroke disease. Classification techniques as decision trees, Bayesian classifiers, and back propagation neural networks were used in this study. Records with irrelevant data were deleted from the data warehouse before the mining process occurred.

D. P. Shukla et al. [2018] [8], implemented a task of designing a system to realize the probability of coronary heart disease. They divided all the parameters into two levels based on the importance of the parameters and gave each step extra weight. Finally, the final decision is made considering both levels. They proposed a neuro fuzzy integration approach at two levels. This error rate is very low and work effi- ciency is high. They concluded that this method could be used to perform an analysis of other diseases.

Chapter 3

PROJECT DESCRIPTION

3.1 Existing System

The existing system used for predicting heart disease using machine leaning algorithm consists of Support Vector Machine(SVM). In this system input is dataset and it is divided into training data and testing data in the ratio of 70:30. The initial step is preprocessing the dataset because dataset from real world consists of redun- dant data, incomplete data, irrelevant data or it may contain errors. By using Data Preprocessing, we can able to remove all these errors. Later Preprocessed dataset will be sent to SVM algorithm as input. The output is fuzzy model is used to analyze whether a person has heart disease or not.

3.2 Proposed System

To predict the heart disease in our proposal system we will use dataset as input. For the training and testing, the machine learning model need appropriate dataset with necessary features. The prediction accuracy has been increased by selecting appropriate and concerned features from the dataset. Standard dataset from UCI Machine Learning repository has been used in our proposed system. Dataset contains of several features they are age, gender, Blood Pressure (BP), collastrol, chest pain, heart beat level, rest, sugar level, resting ECG. Data pre-processing is the first and important step in the proposed model. It is a technique that helps to transform really world data into understandable format. Dataset is divided into training data and testing data Training data ensures that model learns only from the training data and tests its performance with the testing data.The training data contains 75 % of total dataset and testing data contains 25 % of total dataset. Preprocessed dataset sent as input to LR model where it will undergo several computations and output will beused to predict whether a person has heart disease or not.

3.3 Feasibility Study

3.3.1 Economic Feasibility

The economical feasibility in our project Heart Disease using Machine Learning system is cost and time saving. By using computer based system we can reduce the cost and the at the same time we can save the time. The implementation of this project will take up no cost as the previous systems. People no need to spend much money on treatment.

3.3.2 Technical Feasibility

Technical feasibility in our project is to collect the patient details regarding age, gender, Blood Pressure (BP), collastrol, chest pain, heart beat level, rest, sugar level, resting ECG to check whether a person has heart disease or not. They no need to go to hospital every time. This system also helpful to doctors they no need to spend their time in labs, they can detect whether a person has heart disease or not from their computer itself.

3.3.3 Social Feasibility

Social Feasibility of this project can help patients to predict heart disease at an early stage. Early detection and symptomatic treatment helps to ensure the healthy life and well-being of pre-heart disease patients. The death ratio of heart disease can be controlled if the disease is diagnosed and preventative measures are taken in early-stage. Early detection of diabetes can prevent serious health complications. Whena problem with blood sugar is found, doctors and patients can take steps to prevent permanent damage to the heart, kidneys, eyes, nerves, blood vessels, and other vital organs. Lifestyle changes like a healthy diet, daily exercise, and weight management can improve our health.

3.4 System Specification

3.4.1 Hardware Specification

- Intel Core:i3
- Processor:Pentium IV or higher

• Processor speed:1.6GHz
• RAM:4GB RAM minimum, 8GB RAM-recommended
• Disk Space:2GB of available disk space minimum, 4GB-Recommended

3.4.2 Software Specification

• Operating System:Windows7 or higher
• Platform:Google Colaboratory
• Language:Python 3.10.2

3.4.3 Standards and Policies

IEEE P7002™ - Standard for Data Privacy Process.

This standard specifies how to manage privacy issues for systems or software that collect personal data. It will do so by defining requirements that cover corporate data collection policies and quality assurance. It also includes a use case and data model for organizations developing applications involving personal information. This helps designers by providing ways to measure privacy controls in their systems utilizing privacy impact assessments.

IEEE P3333.1.3™ - Standard for the Deep Learning Based Assessment of Visual Experience Based on Human Factors.

This standard defines deep learning-based metrics of content analysis and quality of experience (QoE) assessment for visual contents, which is an extension of Stan- dard for the Quality of Experience (QoE) and Visual-Comfort Assessments of Three-Dimensional (3D) Contents Based on Psychophysical Studies (IEEE 3333.1.1™) and Ultra High Definition (UHD) Contents(IEEE 3333.1.2™).

IEEE P2841™ - Framework and Process for Deep Learning Evaluation.

This document defines best practices for developing and implementing deep learning algorithms and defines a framework and criteria for evaluating algorithm reliability and quality of the resulting software systems.

Chapter 4

METHODOLOGY

4.1 General Architecture

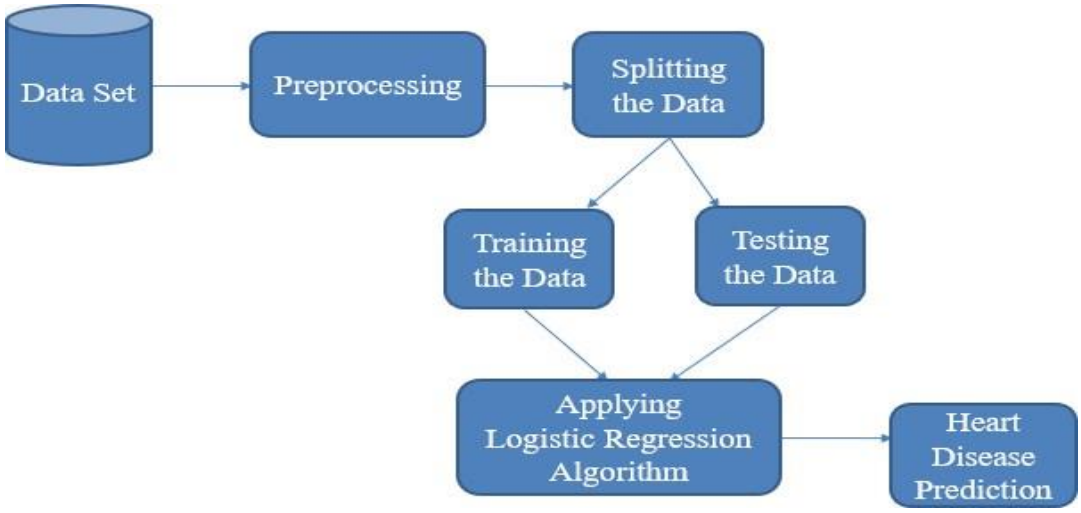


Figure 4.1: **Architecture Diagram of Heart Disease prediction**

The figure 4.1 shows architecture diagram depicts how the overall components involved and the flow of input into a compliments in each module to produce output. Architecture diagram tells us about process of the proposed system. In this initially we collect dataset from UCI, then we will pre-process the dataset. Then the pre-processed dataset will be divided into training and testing data. Training data ensures that model learns only from the training data and tests its performance with the testing data. The training data contains 75% of total dataset and testing contains 25% of total dataset.Pre-processed dataset sent as input to LR model where it will undergo several computations and output will be used to predict whether a personhas heart disease or not.

4.2 Design Phase

4.2.1 Data Flow Diagram

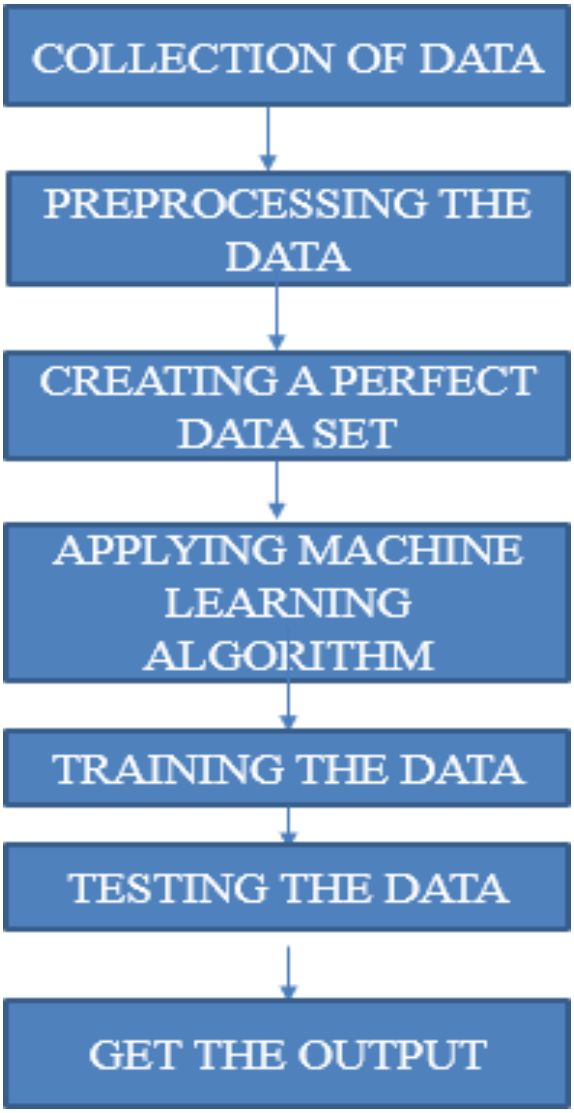


Figure 4.2: **Data Flow Diagram of Heart Disease prediction**

The figure 4.2 shows data flow diagram about how the data is getting transmitted and check for the given condition to produce the expected output. Dataflow diagram depicts flow of data from input to output. LR is feed forward neural network that is here data flows only in forward direction from left to right. In LR a node is visited only once and it cannot be revisited.

4.2.2 Use Case Diagram

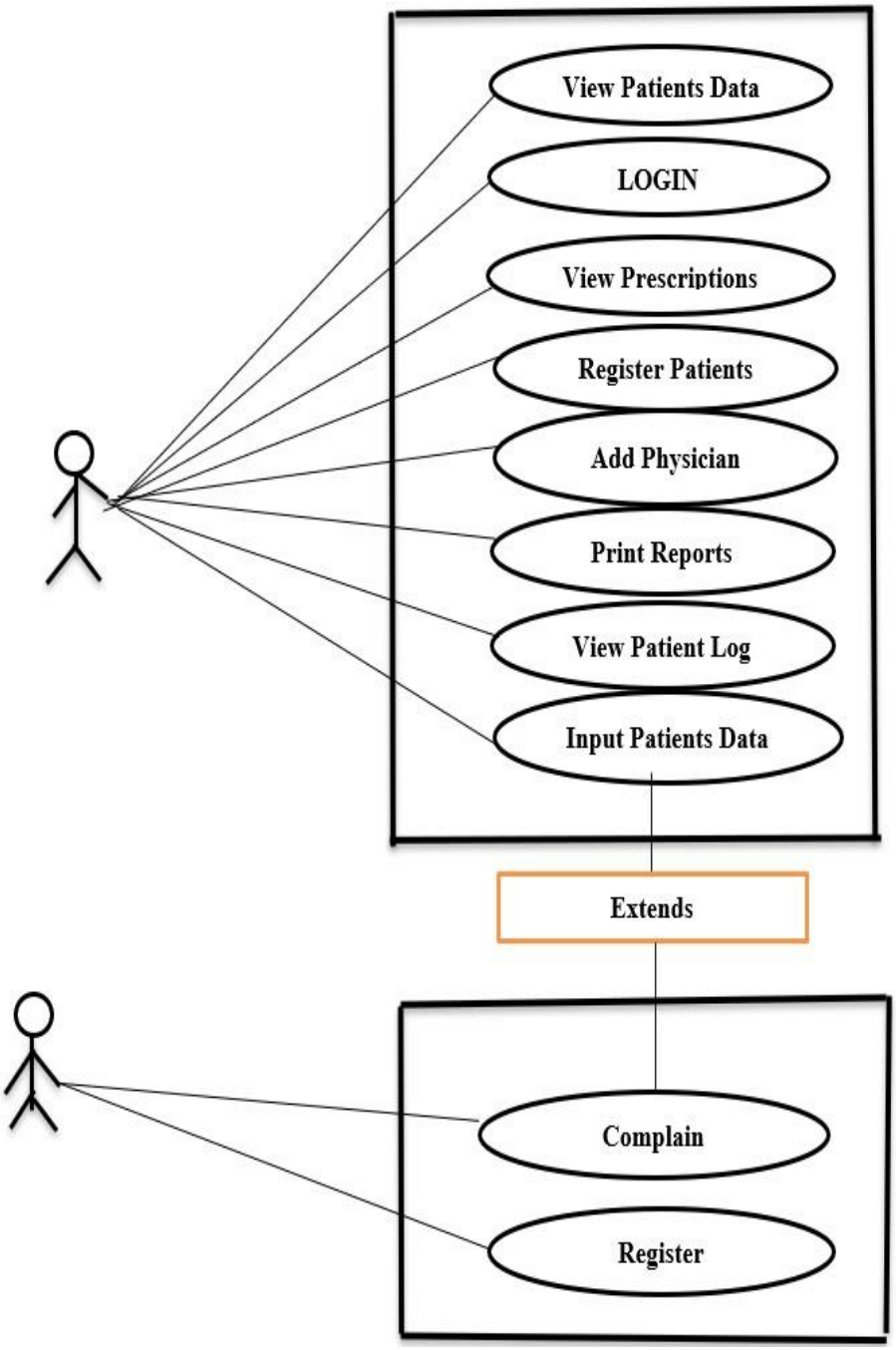


Figure 4.3: **Use Case Diagram of Heart Disease prediction**

The figure 4.3 shows use case diagram depicts the rules and responsibilities of each entity in the project is maintained by mapping the entity to its corresponding responsibility. In our proposed system admin controls the system and admin is responsible for importing the dataset, pre-processing the data into training and testing. Then admin train the LR algorithm using training data. User sends input to the trained LR model, which gives the output. The output is heart disease prediction.

4.2.3 Class Diagram

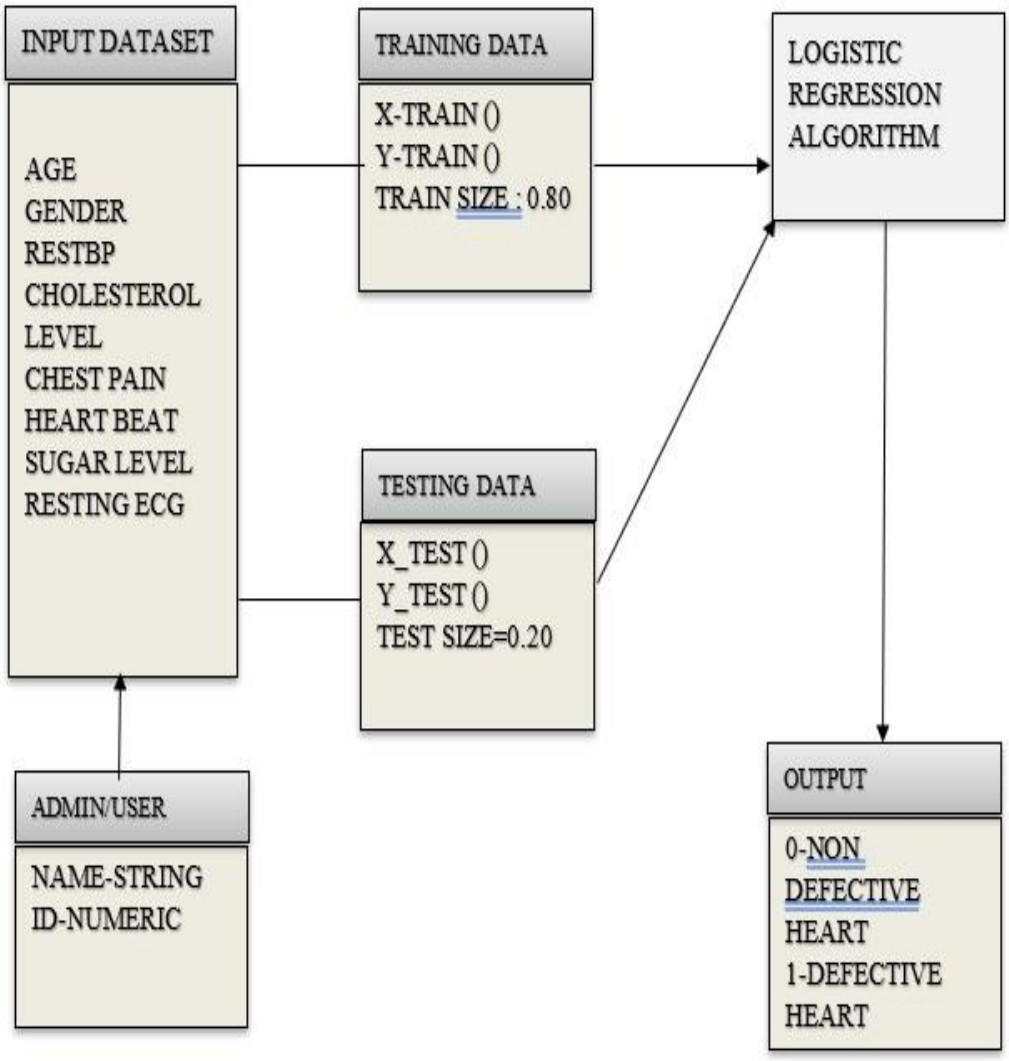


Figure 4.4: **Class Diagram of Heart Disease prediction**

The figure 4.4 shows class diagram includes the attributes and variables related with each entity or module of our project to make the programmer to code easilyand better optimistic way. Input is dataset and it has several parameters such as age, gender, Blood Pressure (BP), collastrol, chest pain, heart beat level, rest, sugar level, resting ECG,outcome. All the input parameters must be numeric.Then we divide input data into training and testing data. Size of Training data is 0.75 and testingdata is 0.25. Then we will train our model using LR algorithm. Then it gives an output either 0 or 1.0 implies NonHeart Disease Person, 1 implies Heart Disease Person.

4.2.4 Sequence Diagram

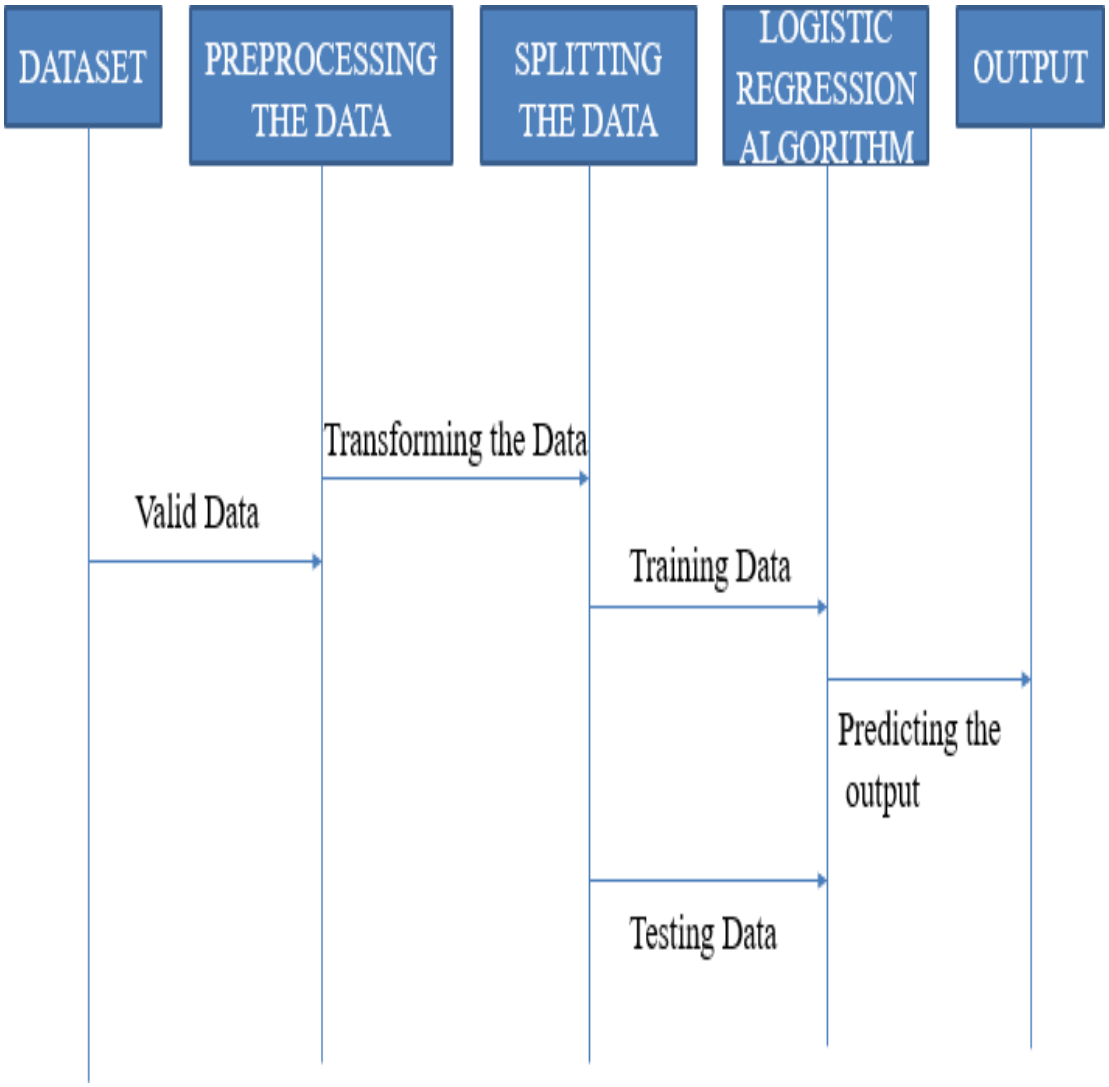


Figure 4.5: **Sequence Diagram of Heart Disease prediction**

The figure 4.5 shows sequence diagram depicts the sequence of data from input to output. A sequence diagram shows the sequence of messages passed between objects. Sequence diagrams can also show the control structures between objects. When the dataset is imported it checks for errors if there are ni errors, it sends a message valid data to next object which is pre-processing the data. It transforms the data into understandable format and sends a message to next object. Then the datais divided into training and testing data. LR algorithm is used to train the algorithm, and it passes the message about output which contains information of heart disease prediction.

4.2.5 Collaboration diagram of Heart Disease prediction

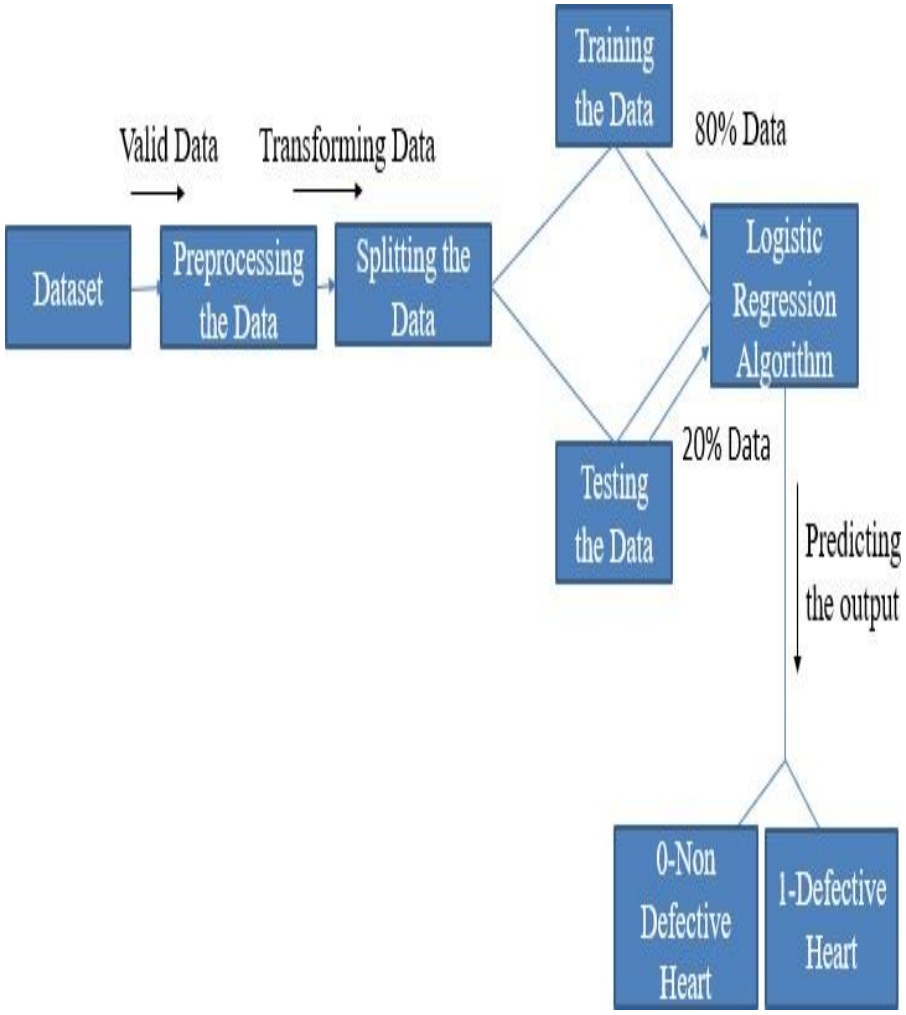


Figure 4.6: **Collaboration diagram of Heart Disease prediction**

The figure 4.6 shows collaboration diagram depicts about the relationship be- tween the objects in the system. Both the sequence and the collaborative diagram represent the same information but differently. Collaboration diagrams show how objects associate with each other. Collaborative diagrams show objects, their links, and their messages. They can also contain simple class instances and class utility instances. Each collaboration diagram provides a view of the interactions or struc- tural relationships that occur between objects and object-like entities in the current model. It shows how proposed is system interacts and relationship between objects, how data is transforming from input to output.

4.2.6 Activity Diagram

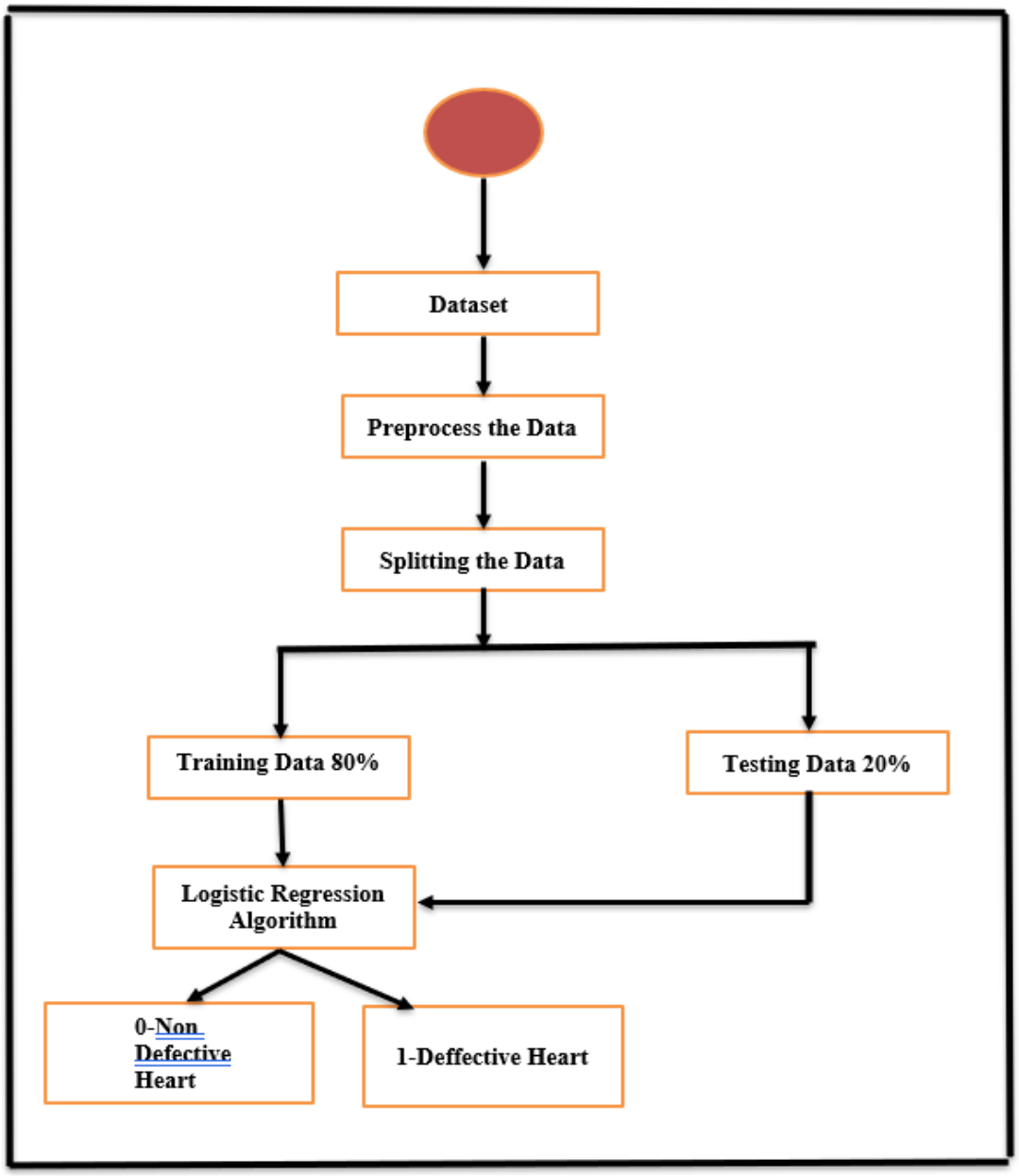


Figure 4.7: **Activity Diagram of Heart Disease prediction**

The figure 4.7 shows activity diagram depicts the control flow from input to output through various decision paths. Activity diagram is basically a flowchart to represent the flow from one activity to another activity. In the proposed system the data flows only in forward direction since we are using LR feed forward neural network. In the system ourselves trained the LR algorithm using training data by sending both input and output, then we will send testing data to trained algorithm for the prediction of heart disease.

4.3 Algorithm

4.3.1 Algorithm

Step-1:Start
Step-2:Import the dataset
Step-3:Pre-process the data
Step-4:Split the data set into training and testing data.
Step-5:Apply the LR algorithm to training data.
Step-6:Predicting the output(0-Non-Heart Disease,1-Heart Disease)
Step-7:Stop

4.4 Module Description

4.4.1 Collection of Dataset

To predict the heart disease in our proposed system ourselves used dataset as input. For the training and testing, the machine learning model need appropriate dataset with necessary features.The prediction accuracy has been increased by se- lecting appropriate and concerned features from the dataset. Standard dataset from (University of Calinfomia Irvine)UCI Machine Learning repository has been used in our proposed system. Dataset consists of several features they are age,gender, Blood Pressure (BP), collastrol, chest pain, heart beat level, rest, sugar level, resting ECG.

4.4.2 Pre-Processing the Dataset

Data Pre-processing is the first and important step in the proposed model. It is a technique that helps to transform real world data into understandable format. When collected data from real world it consists of redundant data, incomplete data, irrele- vant data, incomplete data or it may contain errors. By using Data Pre-processing, we can able to remove all these issues. We will be using Data Cleaning as Pre- processing method in our proposed system.

4.4.3 Splitting Dataset into Training and Testing data

In the proposed system the data is divided into training and testing data. Training data ensures that model learns only from the training data and tests its performance with the testing data. The training data contains 75% of total dataset and testing data contains 25% of dataset.

4.4.4 Acquiring the output using LR algorithm

Pre-Processed dataset will send to machine learning algorithm which is LR. Input Layer is the first layer in the DNN network. To this layer dataset will be send as input. This layer does not consist of any computation. The only use of this layer is to send the features to the next layer which is hidden layer. The hidden layer is the layer which presents in between input layer and output layer. In Deep Neural Network more than one hidden layer are present. Before the information passes to the output layer computations will be performed. The output layer is the last layer of the Deep Neural Network. This layer is responsible to generate the output variables. The input layers represent the actual input ports in networks. The regular densely connected layers with their output dimensions use SoftMax activation and linear activation with the load initialization function. To get an output the activation layer will apply an activation function.

4.5 Steps to execute/run/implement the project

4.5.1 Step 1: Download Google Colaboratory

Platform to execute for proposed system is Google Colaboratory. First we have to install Google Colaboratory using command prompt. By using python 3.10.2in Google Colaboratory to execute the project. After installing this we just haveto type Google Colaboratory in command prompt then Google Colaboratory will automatically opens in our browser

4.5.2 Step 2: Uploading Data Set

Input of the project is dataset and downloading it from kaggle. Then we have to save the dataset in csv format. Then will upload the dataset file in Google Colaboratory.

4.5.3 Step 3: Create New File

Then the team will create a new file in Google Colaboratory. In that file we write code of our project and it will be executed.

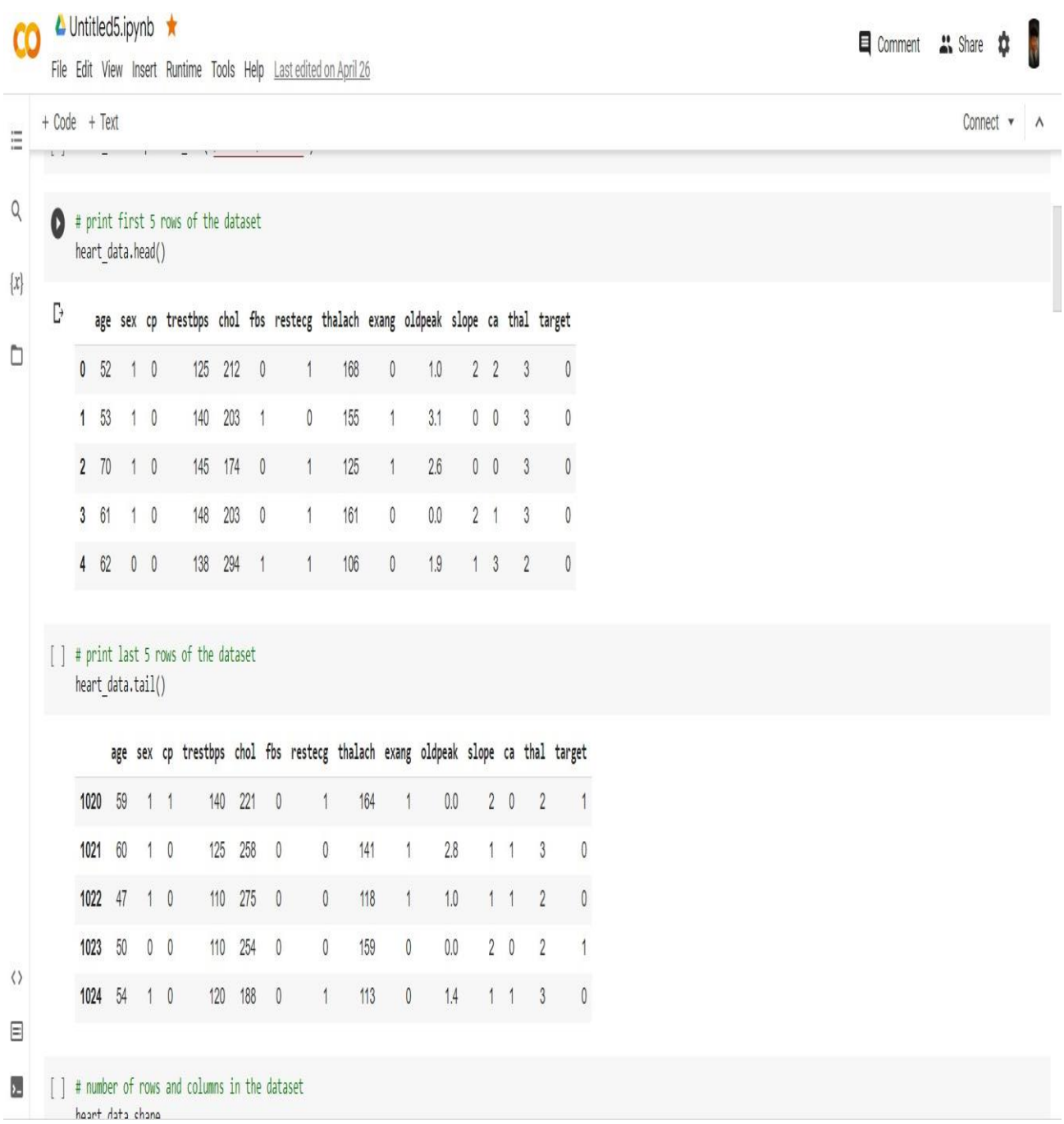
Chapter 5

IMPLEMENTATION AND TESTING

5.1 Input and Output

Input of the proposed system is Dataset. Standard dataset from (University of Calinfomia Irvine) UCI Machine Learning repository has been used for this work. Dataset consists of several features such as Age, Gender, Rest BP,Cholestrol level, Chest pain, Heart beat level, Sugar level, and Resting ECG. Then the dataset is pre-processed to transform the data into understandable format. Then the pre-processed dataset is divided into training data and testing data. Initially data we will send training data which consists of both input and output to LR algorithm to train the model. LR has neurons which we call as nodes are connected with layers. Laterwill send testing data to trained LR model to predict heart disease. Output of the proposed model is either 0 or 1.0 implies Non-Defective heart. 1 implies Defective heart. The proposed system generates precision matrix which consists of accuracy, precision, recall and f1-score. Interpret the results of the model to gain insights into the underlying factors that contribute to heart disease. Visualize the data, model, and predictions using techniques such as scatter plots, histograms, heatmaps, decision trees, or feature importance plots. Communicate the findings and limitations of the model to stakeholders, including healthcare professionals, patients, and policymak- ers. Identify the most relevant features for predicting heart disease and engineer new features if necessary. Feature selection can be done using techniques such as correla- tion analysis, feature importance ranking, or domain expertise. Feature engineering involves creating new features based on the existing ones, such as calculating ratios, adding interactions, or deriving statistical measures.

5.1.1 Input Design of Dataset



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

Figure 5.1: **Input Design of Dataset**

5.1.2 Output Design Of Dataset

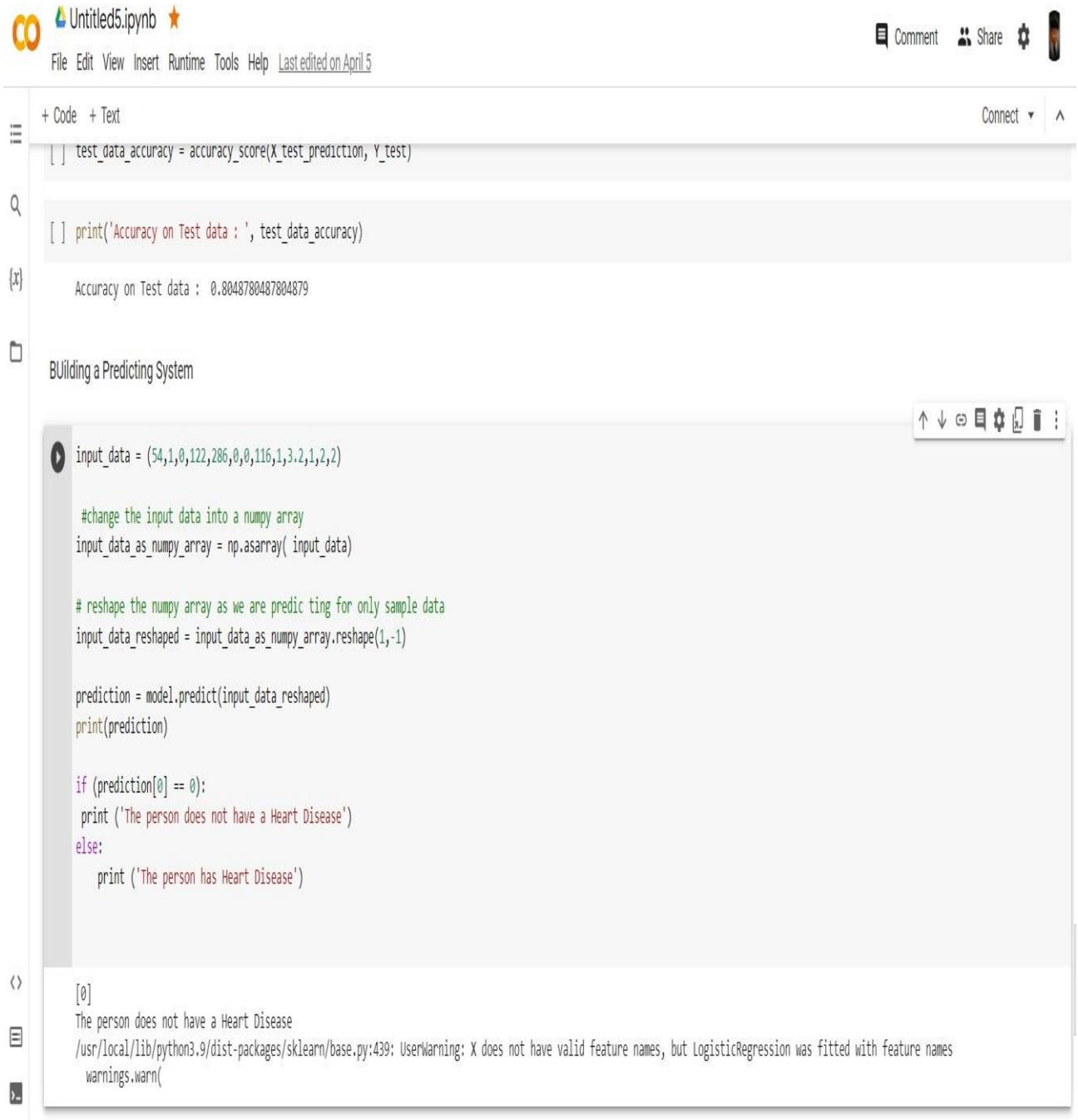


Figure 5.2: **Output Design Of Dataset**

5.2 Testing

The purpose of testing is to find the errors. Testing is the process of trying to find every conceivable fault or weeakness in a work product. It provides a way to check the performance of components,sub-assemblies, assemblies and/or a finished product. It is the process of working software with the intent of ensuring that the Software system meets its necessary and user expectations and does not fail in an unacceptable manner. There are many types of testing. Each test type address a specific testing requirements

5.3 Types of Testing

5.3.1 Unit testing

Unit testing is intended to check tiny items of pratially instead of the system as entire. Unit testing will be performed from very cheap up, beginning with smallest and lowest level modules and continuing one at a time. For every module inn bottom-up testing a brief program is employed to execute the module and provides the required information. So the module is asked to perform the approach it'll once embedded with within the larger system. The code and it will check the connection of the modules and collects the data and checks the code

Input

```
1
2 input_data = (71,0,0,112,149,0,1,125,0,1.6,1,0,2)
3
4  #change the input data into a numpy array
5 input_data_as_numpy_array = np.asarray( input_data)
6
7 # reshape the numpy array as we are predic ting for only sample data
8 input_data_reshaped = input data as numpy array.reshape(1,-1)
9
10 prediction = model.predict(input_data_reshaped)
11 print(prediction)
```

Test result

It will display if there are any duplicates in the dataset, if there are none it will display 0. Output:0

5.3.2 Integration testing

Integration testing is any type of software testing that seeks to verify the inter- faces between components against a software design. Its primary purpose is to ex- pose the defects associated with the interfacing of modules. This testing is done to determine/check whether the communication interfacing between the nodes is done properly or not which in turn effects the function of the modules as it's interlinked one to another.

Input

```
1
2 # training the LogisticRegression model with Training data
3 model.fit(X train, Y train)
```

Test result

Increase the number of iterations (maxiter)orscalethedataasshownin : https : //scikit learn.org/stable/modules/preprocessinghtml Please also refer to the documentation for alternative solver options: https://scikit-learn.org/stable/modules/linearmodel.htmllogisticregressionextrawarningmsg =L OGIST ICSOLV ERCONV ERGENCEMSG, LogisticRegression()

5.3.3 System testing

System testing is a software testing method in which the functionalities of soft- ware applications are tested without having knowledge of internal code structure, implementation details and internal paths. This testing is done to check by giving input and checking whether expected or appropriate output is producing or not.

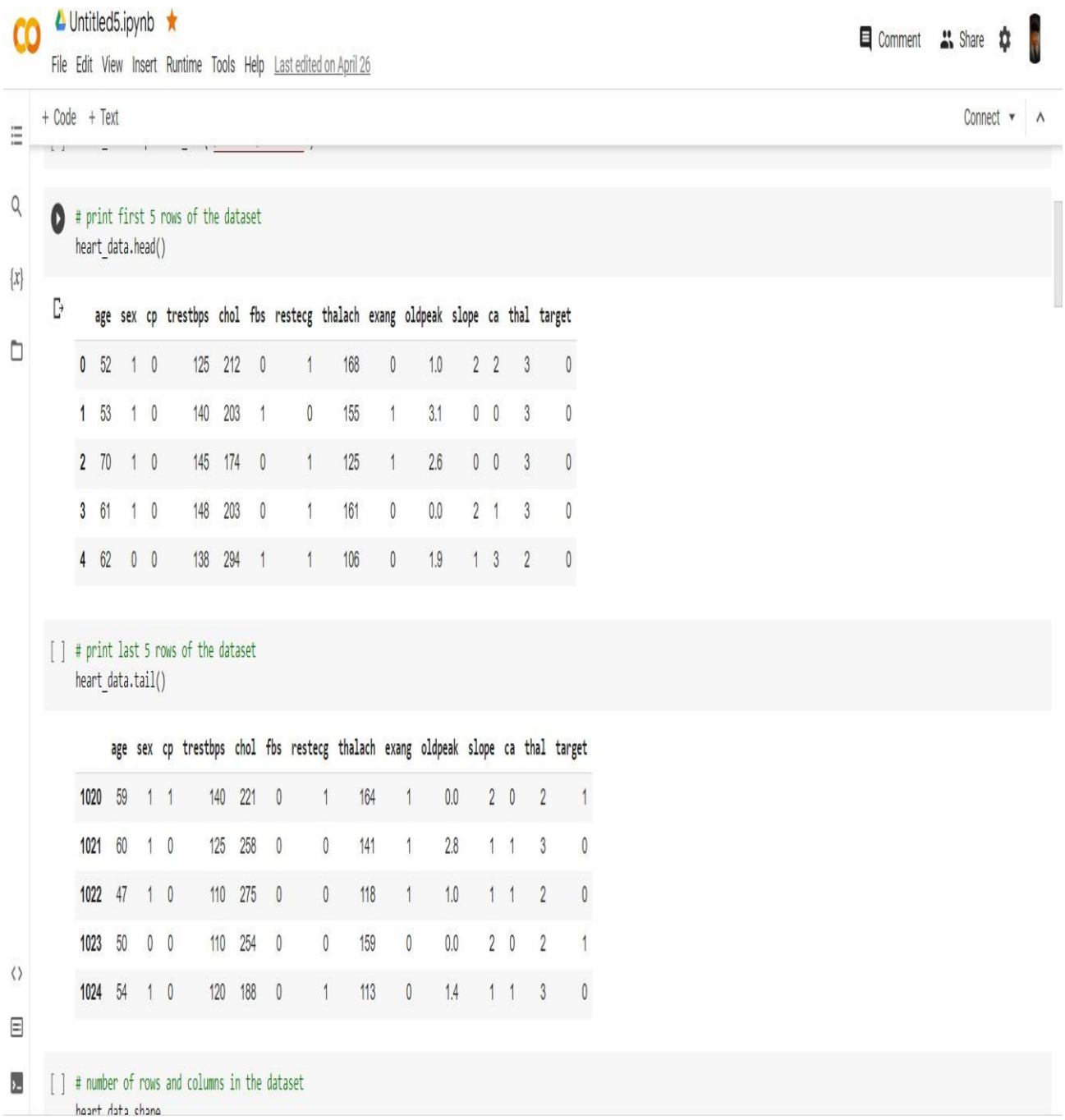
Input

```
input_data = (71,0,0,112,149,0,1,125,0,1.6,1,0,2)

 #change the input data into a numpy array
input_data_as_numpy_array = np.asarray( input_data)

# reshape the numpy array as we are predic ting for only sample data
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0] == 0):
 print ('The person does not have a Heart Disease')
else:
    print ('The person has Heart Disease')
```

Test Result

[0] The person does not have a heart disease
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names
”X does not have valid feature names”

5.3.4 Test Result



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

Figure 5.3: **Test Image of Heart Disease prediction**

Chapter 6

RESULTS AND DISCUSSIONS

6.1 Efficiency of the Proposed System

Proposed system uses LR algorithm to predict the heart disease. LR is supervised machine learning algorithm, so we can train the system by providing both input and output. If we send new data system can easily provide output without because itis already trained. LR produce better results for large amount of data better than other algorithms. Proposed system has high accuracy, sensitivity, specificity. In the proposed system initially pre-processing the dataset so that we can errors. Then we split data into training and testing data. Training data ensures that model learns only from the training data and tests its performance with the testing data. So that if we send new data to proposed system it can able to produce better without errors.

6.2 Comparison of Existing and Proposed System

Existing system:

Existing system uses two algorithms SVM and ANN to predict whether a person is diabetic or not. Whereas in proposed we are using DNN algorithm to predict heart disease. In the existing system we will send pre-processed dataset to both SVM and ANN algorithms as input and output of these two two will send as input to fuzzy model and it produces an output and gives information about whether a person have heart disease or not.

Proposed system:

Cardio Vascular Diseases (CVDs) are the most common cause of death in the modern society constituting to about 35total deaths. People underestimate the value of health in the modern society. This proposal aims in the production of a success-

fully built web application that would allow people to obtain a reliable source to predict whether they are at any risk to any CVDs during their lifetime. The browser provides the user with predictions from a total of 5 different machine learning modelswhich are logistic regression classifier, na¨ıve bayes classifier, decision tree classi- fier, SVC classifier and random forest classifier.

6.3 Sample Code

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
# loading the csv data to a Pandas DataFrame
heart_data = pd.read_csv('/content/data.csv')
# print first 5 rows of the dataset
heart_data.head()
# print last 5 rows of the dataset
heart_data.tail()
# number of rows and columns in the dataset
heart_data.shape
# getting some info about the data
heart_data.info()
# checking for missing values
heart_data.isnull().sum()
# statistical measures about the data
heart_data.describe()
# checking the distribution of Target Variable
heart_data['target'].value_counts()
X = heart_data.drop(columns='target', axis=1)
Y = heart_data['target']
print(X)
print(Y)
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify=Y, random_state=2)
print(X.shape, X_train.shape, X_test.shape)
model = LogisticRegression()
# training the LogisticRegression model with Training data
model.fit(X_train, Y_train)
# accuracy on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
print('Accuracy on Training data : ', training_data_accuracy)
# accuracy on test data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
print('Accuracy on Test data : ', test_data_accuracy)
```

```
input_data = (71,0,0,112,149,0,1,125,0,1.6,1,0,2)

 #change the input data into a numpy array
input_data_as_numpy_array = np.asarray( input_data)

# reshape the numpy array as we are predic ting for only sample data
input_data_reshaped=input data as numpy array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0] == 0):
 print ('The person does not have a Heart Disease')
else:
    print ('The person has Heart Disease')
```

Output

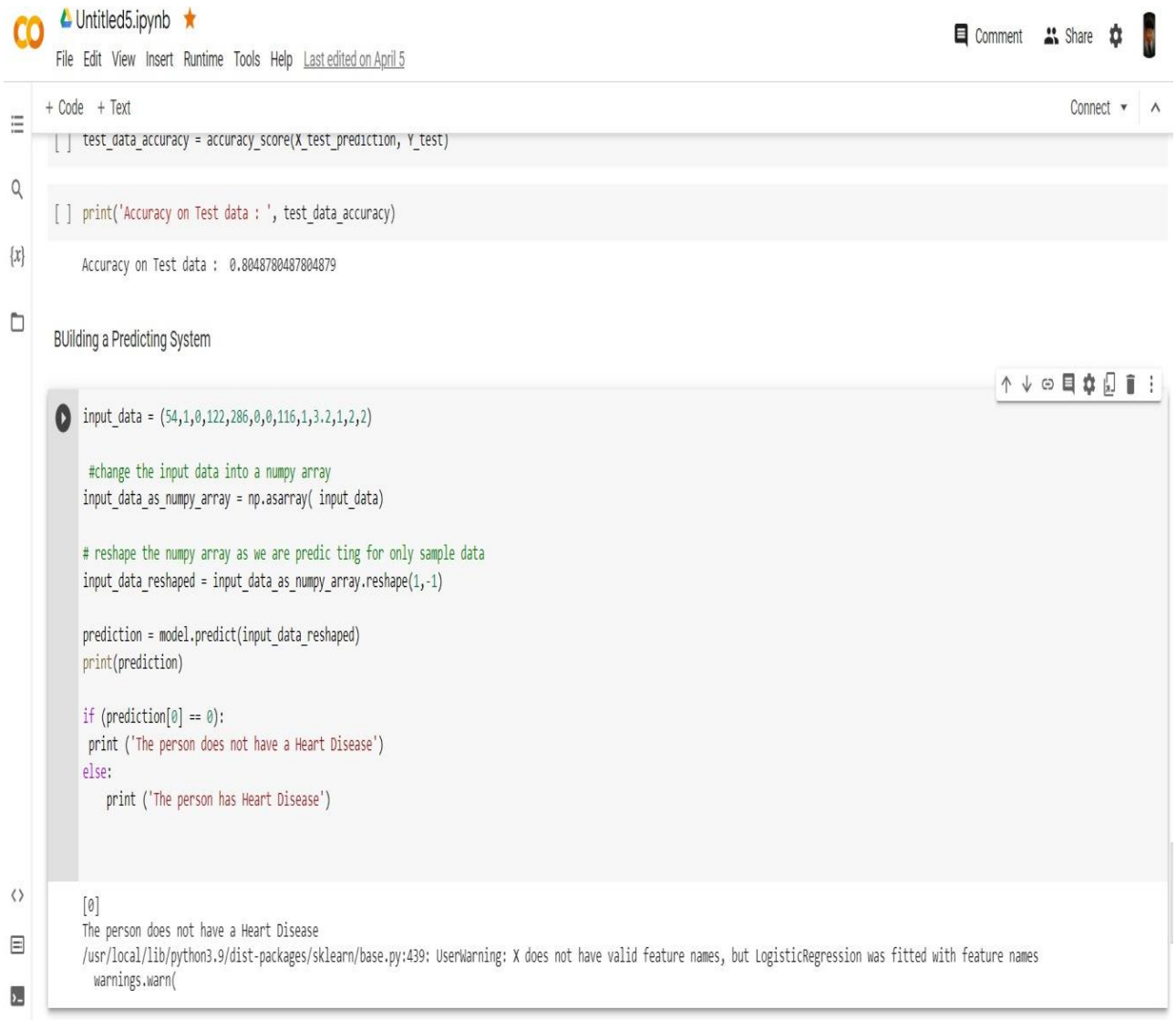


Figure 6.1: **Output of Heart Disease prediction**

Chapter 7

CONCLUSION AND FUTURE ENHANCEMENTS

7.1 Conclusion

One of the important areas in industry of medical is prediction of cardiovascular disease, with the available data of the patient to predict the absence and presence of cardiac disease. There are several techniques and methods are present for prediction of cardiovascular disease. In this project, Logistic Regression in Machine Learning algorithm used to classify the heart disease. As the behaviour of Logistic Regression is as training increases the accuracy of prediction also increased. The Logistic Regression classifier achieved 87.10results outperformed compared to previous research work. The limitation is only UCI dataset is used in the study and future work try to implement on multiple datasets.

The amount of heart disease can exceed the control line and reach to maximum point. Heart diseases are complicated and each and every year lots of people are dying with this disease by using this all systems one of the major drawbacks of these works is mainly focus only to the application of classify techniques and algorithmsfor heart disease prediction. So that I can use this Machine Learning in that Logistic Regression algorithms by predicting if patient has heart disease or not. Any non- medical employee can use this software and predict the heart disease and reduce the time complexity of the doctors.

7.2 Future Enhancements

Today's,world most of the data is computerized, the data is distributed and it is not utilizing properly. By analyzing the available data we can also use for unknown patterns. The primary motive of this research is the prediction of heart diseases with

high rate of accuracy. For prediction the heart diseases we can use logistic regression algorithm, navie bayes, sklearn in machine learning. The future scope of the paper is the prediction of heart diseases by using advanced techniques and algorithms in less time complexity.

For similar studies, it is better to use two or more algorithms. More algorithms create more results, and the best result can be chosen as the decision. In addition, Logistic Regression itself is an algorithm that is quite effective in providing predictions related to cardiovascular disease. In analysing using logistic regression, binary data is needed to simplify the development of an algorithm model. Besides, in implementing the logistic regression algorithm, future validation is also required. We recommend that you validate the data and the results of using the algorithm, more than twice. It is suggested that a better level of accuracy is obtained so that the anal- ysis and evaluation of the effects of implementing the logistic regression model can be.

Chapter 8

PLAGIARISM REPORT

REPOSTSEO

PLAGIARISM SCAN REPORT

Date	May 03, 2023
Exclude URL:	NO

Unique Content	91%	Word Count	4492
Plagiarized Content	9%	Records Found	1
Paraphrased Plagiarism	0		

CONTENT CHECKED FOR PLAGIARISM:

Healthcare expenditures are overwhelming national and corporate budgets due to asymptomatic diseases including cardiovascular diseases. Therefore, there is an urgent need for early detection and treatment of such diseases Machine learning is one of the trending technologies which used in many spheres around the world including healthcare industry for predicting diseases. The aim of this study is to identify the most significant predictors of heart disease and predicting the overall risks by using logistic regression. Thus, binary logistic model which is one of the classification algorithm in machine learning is used in this study to identify the predictors. The early prognosis of cardiovascular diseases can aid in making decisions to lifestyle changes in high risk patients and in turn reduce their complications. According to recent survey by WHO organisation 17.5 million people dead each year. It will increase to 75 million in the year 2030. Medical professionals working in the field of heart disease have their own limitation, they can predict chance of heart attack up to 67curacy, with the current epidemic scenario doctors need a support system for more accurate prediction of heart disease. Machine learning algorithm and opens

Chapter 9

SOURCE CODE & POSTER PRESENTATION

9.1 Source Code

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
# loading the csv data to a Pandas DataFrame
heart_data = pd.read_csv('/content/data.csv')
# print first 5 rows of the dataset
heart_data.head()
# print last 5 rows of the dataset
heart_data.tail()
# number of rows and columns in the dataset
heart_data.shape
# getting some info about the data
heart_data.info()
# checking for missing values
heart_data.isnull().sum()
# statistical measures about the data
heart_data.describe()
# checking the distribution of Target Variable
heart_data['target'].value_counts()
X = heart_data.drop(columns='target', axis=1)
Y = heart_data['target']
print(X)
print(Y)
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify=Y, random_state=2)
print(X.shape, X_train.shape, X_test.shape)
model = LogisticRegression()
# training the LogisticRegression model with Training data
model.fit(X_train, Y_train)
# accuracy on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
print('Accuracy on Training data:', training_data_accuracy)
# accuracy on test data
```

```
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
print('Accuracy on Test data :', testdataaccuracy)
input_data = (71,0,0,112,149,0,1,125,0,1.6,1,0,2)

 #change the input data into a numpy array
input_data_as_numpy_array = np.asarray( input_data)

# reshape the numpy array as we are predic ting for only sample data
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0] == 0):
 print ('The person does not have a Heart Disease')
else:
    print ('The person has Heart Disease')
input_data = (60,0,0,150,258,0,0,157,0,2.6,1,2,3)

 #change the input data into a numpy array
input_data_as_numpy_array = np.asarray( input_data)

# reshape the numpy array as we are predic ting for only sample data
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0] == 0):
 print ('The person does not have a Heart Disease')
else:
    print ('The person has Heart Disease')
```

9.2 Poster Presentation

FORECASTING THE HEART DISEASE BY MACHINE LEARNING

Department of Computer Science & Engineering
School of Computing
1156CS601 – MINOR PROJECT
WINTER SEMESTER 21-22

ABSTRACT

Healthcare expenditures are overwhelming national and corporate budgets due to asymptomatic diseases including cardiovascular diseases. Therefore, there is an urgent need for early detection and treatment of such diseases.

Machine learning is one of the trending technologies which used in many spheres around the world including healthcare industry for predicting diseases. The aim of this study is to identify the most significant predicators of heart diseases and predicting the overall risks by using logistic regression.

•Thus, binary logistic model which is one of the classification algorithms in machine learning is used in this study to identify the predicators.Further, data analysis is carried out in Python using JupyterLab in order to validate the logistic regression.

TEAM DETAILSMEMBER

<vtu15583/B.VAMSHIKRISHNA>
<vtu15554/M.KRISHNA SANDEEP>
< vtu15986/G.SRINUVASULU>
<1.Phone no 9381387876>
<2. Phone no 8688762434>
<3.Phone no 9347618450>
<Student 1. vtu15583@veltech.edu.in >
<Student 2. vtu15554@veltech.edu.in>
<Student 3. vtu15986@veltech.edu.in>

INTRODUCTION

The number of deaths due to cardiovascular diseases increased by 41% between 1990 and 2013, climbing from 12.3 million deaths to 17.3 million deaths in globally. Logistic regression is one such relatively used machine learning algorithms for studies involving risk assessment of complex diseases.

To initiate with the work we can use different types of techniques and algorithms. In this paper, machine learning techniques are used to increase the accuracy rate. In machine learning technique we can use the logistic regression algorithm. It recognizes who all are having any symptoms of heart disease such as chest pain or high blood pressure and can help in diagnosing disease with less medical tests and effective treatments, so that they can be cured accordingly.

METHODOLOGIES

MODULE 1: Collection of data

To predict diabetes in our proposed system we will use the dataset as input. For the training and testing, the machine learning model needs an appropriate dataset with the necessary features. The prediction accuracy has been increased by selecting appropriate and concerned features from the dataset. Dataset consists of several features they are Age, Sex, Rest BP, Cholesterol level, Chest Pain, Heart Beat, Sugar Level, and Resting ECG.

MODULE 2: Preprocessing the Dataset

To predict diabetes in our proposed system we will use the dataset as input. For the training and testing, the machine learning model needs an appropriate dataset with the necessary features. The prediction accuracy has been increased by selecting appropriate and concerned features from the dataset. Dataset consists of several features they are Age, Sex, Rest BP, Cholesterol level, Chest Pain, Heart Beat, Sugar Level, and Resting ECG.

Module 3: Splitting data set in Training and Testing Data

In our proposed system the data is divided into training data and testing data. Training data ensures that the model learns only from the training data and its performance with the testing data. The training data contains 80% of total dataset and testing data contains 20% of the dataset.

Module 4: Acquiring the output using Logistic Regression algorithm

The pre-Processed dataset will send to a machine learning algorithm which is Logistic Regression. The input layer is the first layer in the logistic regression algorithm. To this layer, the dataset will be sent as input. This layer does not consist of any computation. The only use of this layer is to send the features to the next layer which is the hidden layer. The hidden layer is the layer that presents in between the input layer and output layer. In Logistic Regression more than one hidden layer are present. Before the information passes to the output layer computations will be performed

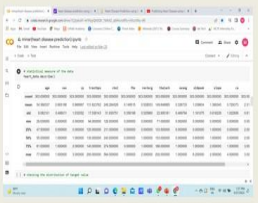


Figure 1. Input

Figure 2. Output

CONCLUSIONS

This Heart Disease detection system assists a patient based on his/her clinical information of them been diagnosed with a previous heart disease. There are a number of medical databases that we can work on as these Machine learning techniques are better and they can predict better than a human being which helps the patient as well as the doctors.

Therefore, in conclusion this project helps us predict the patients who are diagnosed with heart diseases by cleaning the dataset and applying logistic regression to get a high accuracy.

ACKNOWLEDGEMENT

1. Dr. A. SURESH/Associate professor
2. 9843849880
3. drtkamaleshwar@veltech.edu.in

References

[1] Avinash Golande, Pavan Kumar T, ”Heart Disease Prediction Using Effective Machine Learning Techniques”,International Journal of Recent Technology and Engineering,vol8,pp.944-950(2021)

[2] Basavaraj S Pol, Chethan C, Luimath, Nagaraj M , ”Prediction of Heart Disease using Machine Learning”,International journal of Recent Technology and Engineering,8,(2S10),pp 474-477(2020))

[3] Berina Ali, “Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases” 6th Mediterranean Conference on Embedded Computing (MECO), IEEE, pp: 1-4, (2021)

[4] E.Andres, G.Badr, E.Hassani, C.B.Rjeily, Medical Data Mining for Heart Diseases and Future of Sequential Mining in Medical Field,in Machine Learning Paradigms,vol8,pp.71-99(2021)

[5] Fahd Saleh Alotaibi, ”Implementation of Machine Learning Model to predict Heart Failure Disease”,(IJACSA)International Journal of Advanced Computer Science and Applications,vol.10,No.6(2022)

[6] G.Parthiban and S.K.Srivatsa, ”Applying machine learning methods in diagnosing heart disease”, International journal of Applied Information Systems,vol3,pp.25-30(2020)

[7] H. Sharma, “Prediction of Heart Disease using Machine Learning Algorithms: A Survey,” Int. J. Recent Innov. Trends Comput. Commun., vol. 5, no. 8, pp. 99–104(2022)

[8] Prerana T H M1, Shivaprakash N C2 , Swetha N3, ”Prediction of Heart Disease Using Machine Learning Algorithms- Naive Bayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS” International Journal of Science and Engineering Volume 3,pp: 90-99(2021)

[9] S. Hafeez, H. Krishna, T. Maneesha, R. Nichenametla, “Prediction of HeartDisease Using Machine Learning Algorithms,” Int. J. Eng. Technol., vol. 7, pp. 363–366 (2018)