A

Mini Project Report

on

HEART DISEASE PREDICTION

Submitted in partial fulfillment of the
Requirements for the award of degree of

Bachelor of Technology

in

Computer Science and Engineering – Data Science

by

DEVARASETTI VARUN (20eg110104)

EEGA TEJASWINI (20eg110106)

VUPPALANCHI SRIVATSA (20eg110130)

BIRELLA SAI SUHAS REDDY (21eg510101)

Under the Guidance of
Dr. Rakesh Roshan
Assistant Professor

Department of Data Science

ANURAG UNIVERSITY

Venkatapur (V), Ghatkesar (M), Medchal(D)., T.S-500088

(2020-2024)

DEPARTMENT OF DATA SCIENCE

CERTIFICATE

This is to certify that the project entitled **"HEART DISEASE PREDICTION"** being submitted by **DEVARASETTI VARUN, EEGA TEJASWINI, VUPPALANCHI SRIVATSA, BIRELLA SAI SUHAS REDDY** bearing the Hall Ticket number **20eg110104, 20eg110106, 20eg110130, 21eg510101** respectively in partial fulfillment of the requirements for the award of the degree of the **Bachelor of Technology** in **Computer Science and Engineering – Data Science** to **Anurag University** is a record of bonafide work carried out by them under my guidance and supervision from June 2023 to September 2023.

The results presented in this project have been verified and found to be satisfactory. The results embodied in this project report have not been submitted to any other University for the award of any other degree or diploma.

Internal Guide

Dr. Rakesh Roshan
Assistant Professor

External Examiner

Head of the Department
Department of Data Science

ACKNOWLEDGEMENT

It is our privilege and pleasure to express profound sense of respect, gratitude and indebtedness to our guide **Dr. Rakesh Roshan**, Department of Data Science, Anurag University, for his/her indefatigable inspiration, guidance, cogent discussion, constructive criticisms and encouragement throughout this dissertation work.

We express our sincere gratitude to **Dr. M.Sridevi**, Head,Department of Data Science Anurag University, for her suggestions, motivations, and co-operation for the successful completion of the work.

We extend our sincere thanks to **Dr. V. Vijaya Kumar**, **Professor & Dean,** School of Engineering, Anurag University, for his encouragement and constant help.

Devarasetti Varun
(20EG1 10104)

EegaTejaswini
(20EG110106)

Vuppalanchi Srivatsa
(20EG110130)

Sai Suhas Reddy
(21eg510101)

DECLARATION

We hereby declare that the project work entitled "**HEART DISEASE PREDICTION**" submitted to the **Anurag University** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in Computer Science and Engineering – Data Science is a record of an original work done by us/me under the guidance of **Dr.Rakesh Roshan, Assistant professor** and this project work have not been submitted to any other university for the award of any other degree or diploma.

Devarasetti Varun
(20EG110104)

EegaTejaswini
(20EG110106)

Vuppalanchi Srivatsa
(20EG110130)

Sai Suhas Reddy
(21eg510101)

ABSTRACT

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications.

Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analyzing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

Our Heart disease prediction System is intended to assist patients in recognizing their heart state early and receiving treatment at an earlier stage, allowing them to avoid any serious conditions by using Machine Learning Algorithms.

KEY WORDS: Cardiovascular disease, Mortality rate, Expertise, Machine learning, Prediction system

INDEX Page.no

LIST OF FIGURES

1.INTRODUCTION

Heart disease is a leading cause of morbidity and mortality worldwide, making it a critical public health concern. Timely and accurate diagnosis of heart disease is crucial for effective treatment and prevention. Machine Learning (ML) offers a powerful tool to enhance the prediction and early detection of heart disease by analyzing complex patterns and relationships in patient data.

This introductory guide explores the application of ML in predicting heart disease, providing an overview of the key concepts and steps involved in developing a predictive model. By leveraging ML techniques, healthcare professionals can make more informed decisions, potentially saving lives and improving patient outcomes.

Importance of Heart Disease Prediction:

Understanding why early detection and prediction of heart disease are vital for improving patient care and reducing healthcare costs. The importance of heart disease prediction lies in its capacity to save lives, reduce healthcare costs, and enhance the quality of life for individuals. Early prediction enables timely interventions and preventative measures, mitigating the risk of heart disease and its associated complications. By allocating healthcare resources more efficiently, it optimizes healthcare delivery, making it more accessible and cost-effective. Moreover, heart disease prediction informs public health strategies, allowing for targeted interventions and policy development to combat this leading cause of mortality.

What is Machine Learning ?

Machine Learning (ML) is a subset of artificial intelligence (AI) that focuses on the development of algorithms and models that enable computer systems to learn and make predictions or decisions based on data without explicit programming. In other words, ML allows machines to improve their performance on a specific task through experience and exposure to data.

Why Machine Learning ?

Machine learning is employed in heart disease prediction due to its ability to analyze complex patient data, extract valuable insights, and offer accurate risk assessments. This technology is crucial in handling the ever-growing volume of healthcare data, automating risk assessments, and continuously improving prediction accuracy. It aids healthcare providers in making informed decisions, optimizing resource allocation, and contributing to ongoing research in cardiology, ultimately leading to better patient care and outcomes.

1.1 MOTIVATION

The motivation for this heart disease prediction project lies in utilizing predictive analytics to identify individuals at high risk, empower them with personalized interventions, and make a substantial positive impact on public health by reducing the burden of heart disease. The motivation for building a heart disease prediction project is rooted in the urgent need to address one of the leading causes of death worldwide. Heart disease is responsible for a significant loss of lives, and its prevalence continues to rise. By developing an accurate and effective prediction model, we aim to empower individuals with knowledge about their heart health, enabling early intervention and lifestyle changes that can save lives. This project not only has the potential to improve individual well-being but also contributes to healthcare system efficiency, reducing the economic burden associated with heart disease treatment. Furthermore, it aligns with the broader goal of advancing medical research by harnessing the power of data-driven approaches, ultimately leading to better understanding, prevention, and management of heart disease. Ultimately, building this project is a critical step toward a healthier future for individuals and communities worldwide.

1.2 PROBLEM DEFINITION

Heart disease is difficult to recognize due to a variety of risk factors such as high blood pressure, cholesterol, and abnormal pulse rate. Because of the disease's complexity, it must be handled with care. Otherwise, the effects of heart or death may occur. Accurate decision-making and optimal treatment are required to address cardiac risk in less Time.

The problem at hand is to develop a machine learning model that accurately predicts the risk of heart disease in individuals based on a set of clinical and demographic features. The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either it are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications

1.3. OBJECTIVE OF THE PROJECT

Prediction of this heart disease is regarded as one of the most important subject.

So our primary objective in this project is to develop a machine learning-based predictive model that can accurately assess the risk of heart disease in individuals. Explore the heart disease datasets using exploratory data analysis (EDA) and to design the system using the Machine Learning model.To predict the future possibility of heart disease by implementing the Logistic Regression algorithm(Which is highly accurate). This model will utilize a diverse set of patient attributes and historical health data to make reliable predictions.

2. LITERATURE SURVEY

[1] Purushottam ,et ,al proposed a paper “Efficient Heart Disease Prediction System” using hill climbing and decision tree algorithms .They used Cleveland dataset and preprocessing of data is performed before using classification algorithms. The Knowledge Extraction is done based on Evolutionary Learning (KEEL), an open source data mining tool that fills the missing values in the data set.A decision tree follows top-down order. For each actual node selected by hill-climbing algorithm a node is selected. by a test at each level. The parameters and their values used are confidence. Its minimum confidence value is 0.25. The accuracy of the system is about
65.7%.

[2] Sonam Nikhar et al proposed paper "Prediction of Heart Disease Using Machine Learning Algorithms" their research gives point to point explanation of Naïve Bayes and decision tree classifier that are used especially in the prediction of Heart Disease. Some analysis has been led to think about the execution of prescient data mining strategy on the same dataset, and the result decided that Decision Tree has the highest accuracy than Bayesian classifier.

[3] Lakshmana Rao et al, proposed “Machine Learning Techniques for Heart Disease Prediction” in which the contributing elements for heart disease are more. So, it is difficult to distinguish heart disease. To find the seriousness of the heart disease among people different neural systems and data mining techniques are used.

[4] Anjan N. Repaka et al, proposed a model stated the performance of prediction for two classification models, which s analyzed and compared to previous work. The experimental results show that accuracy is improved in finding the percentage of risk prediction of our proposed method in comparison with other models.

3. ANALYSIS

This chapter in the report gives a detailed analysis and explanation of methodologies adopted in the proposed system along with its advantages. The software requirements specification section in the report discusses both hardware and software prerequisites needed for the project. It is further followed by the overall description of the flow of activities performed.

3.1 EXISTING SYSTEM

Heart disease is often referred to as a silent killer, as it can lead to a person's death without obvious symptoms. The nature of the disease has caused growing anxiety about its consequences, leading to continued efforts to predict the possibility of this deadly condition in advance. Various tools and techniques are regularly being experimented with to meet the present-day health needs, and Machine Learning techniques can play a crucial role in this regard.

While heart disease can manifest in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting data from various sources, classifying it under suitable headings, and finally analyzing it to extract the desired information, we can draw meaningful conclusions. This technique can be effectively adapted to predict the likelihood of heart disease. As the well-known quote goes, "Prevention is better than cure," early prediction and its control can be instrumental in preventing and reducing the death rates associated with heart disease.

3.2 PROPOSED SYSTEM

Here, creating a system for heart disease prediction aims to involves outlining the components, technologies, and processes that will be used to develop and deploy the predictive model enhancing the accuracy, personalization, and efficiency by incorporating advanced technologies and comprehensive data analysis.

This system will employ Logistic Regression Algorithm to analyze a wide range of patient data including Features such as Age, sex, chest pain, Fasting blood pressure,ECG,Max_heart rate,oldpeak,slope etc.
Logistic Regression Algorithm provides highly accurate prediction system with 85% More Accuracy in Less Time.

By implementing this proposed system for heart disease prediction, the goal is to provide healthcare professionals with a powerful tool for early detection and personalized risk assessment, ultimately improving patient care and outcomes in the context of heart disease.

3.3 SOFTWARE REQUIREMENT SPECIFICATION

The software requirements are always subject to change when it comes to the extent of the accuracy the user desires or the flexibility in which the deployment is needed. This project can be made by using the following software and hardware which are enough for students' purposes in the industry or market.

- Python 3.0 and more
- Numpy
- Pandas
- Matplotlib and Seaborn
- Visual Studio Code
- HTML/CSS.
- Streamlit.

3.3.1 PURPOSE

The primary purpose of a heart disease prediction project is to develop a predictive model that can accurately assess an individual's risk of developing heart disease. Early detection and risk assessment are critical for effective prevention, intervention, and treatment. Therefore, the purpose of the project is as follows:

A. **Early Detection:**
   To identify individuals at risk of heart disease before clinical symptoms manifest, enabling early intervention and preventive measures.
B. **Personalized Healthcare:**
   To provide personalized risk assessments, allowing healthcare professionals to tailor treatment and lifestyle recommendations based on an individual's specific risk factors.
C. **Advancement in Healthcare:**
   To contribute to the advancement of healthcare through the application of machine learning and data-driven approaches in improving heart disease risk assessment and management.

3.3.2 SCOPE

The scope of a heart disease prediction project using ML defines the project's boundaries and objectives, specifying what the project will encompass and what it won't. Here's the scope of such a project:
Data collection, data preprocessing, model development, model evaluation, deployment, documentation and reporting, future enhancements.
Defining the scope of the project is crucial to ensure that the project's objectives are clear, attainable, and aligned with the needs of healthcare and patient care. It also helps manage expectations and resources effectively.

3.3.3 OVERALL DESCRIPTION

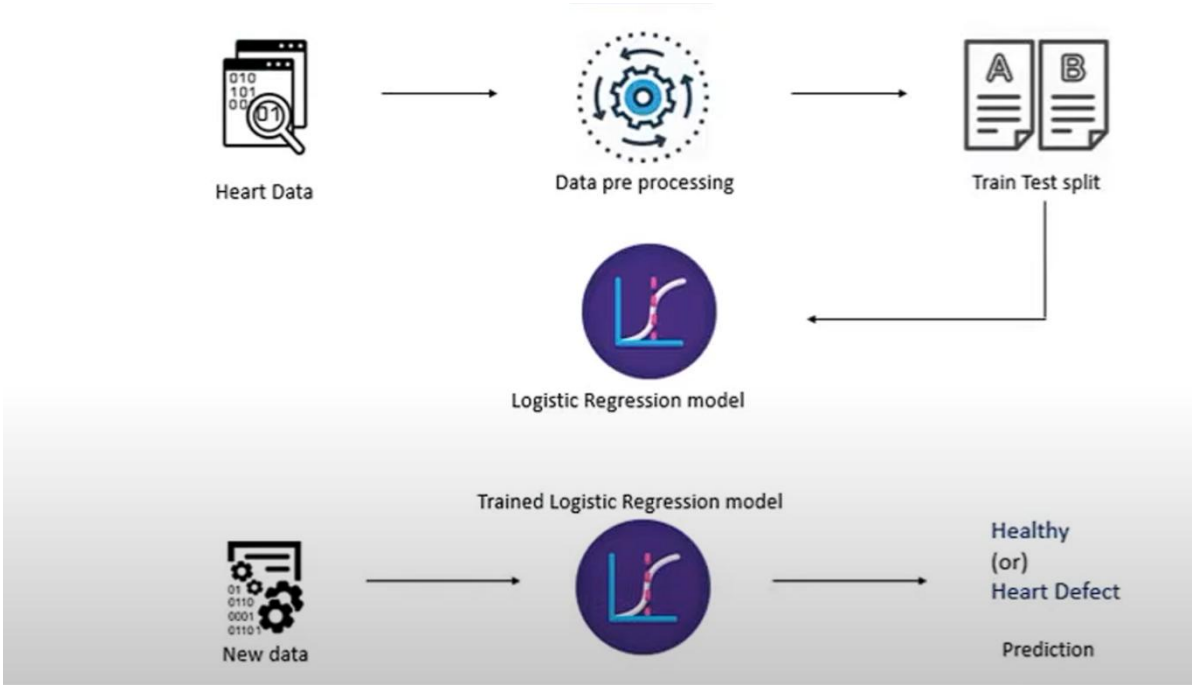


Figure 3.1 Overall Description

The first step involves collecting relevant data from various sources. This data can include medical records, patient history, lifestyle data (e.g., diet, exercise), and possibly genetic information.Once collected, the data needs to be preprocessed.
The dataset is typically split into training and validation sets. The model is trained on the training set and then validated on the validation set to assess its performance. Techniques like cross-validation may be used to ensure robustness. Once a satisfactory model is developed and validated, it is deployed in a real-world setting.

4. DESIGN

4.1. UML DIAGRAMS

We have represented the design of the system with the help of UML diagrams. The following UML diagrams visually represent the system along with its main actors, roles, actions, and classes. They help in easily understanding the overall system.

4.1.1. USE CASE DIAGRAM

Use cases share different kinds of relationships. Defining the relationship between two use cases is the decision of the software analysts of the use case diagram. A relationship between two use cases is basically modeling the dependency between the two use cases. The reuse of an existing use case by using different types of relationships reduces the overall effort required in developing a system.

Figure 4.1 Use Case Diagram

4.1.2. CLASS DIAGRAM

A class diagram in the Unified Modeling Language (UML) is a type of static Structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects.

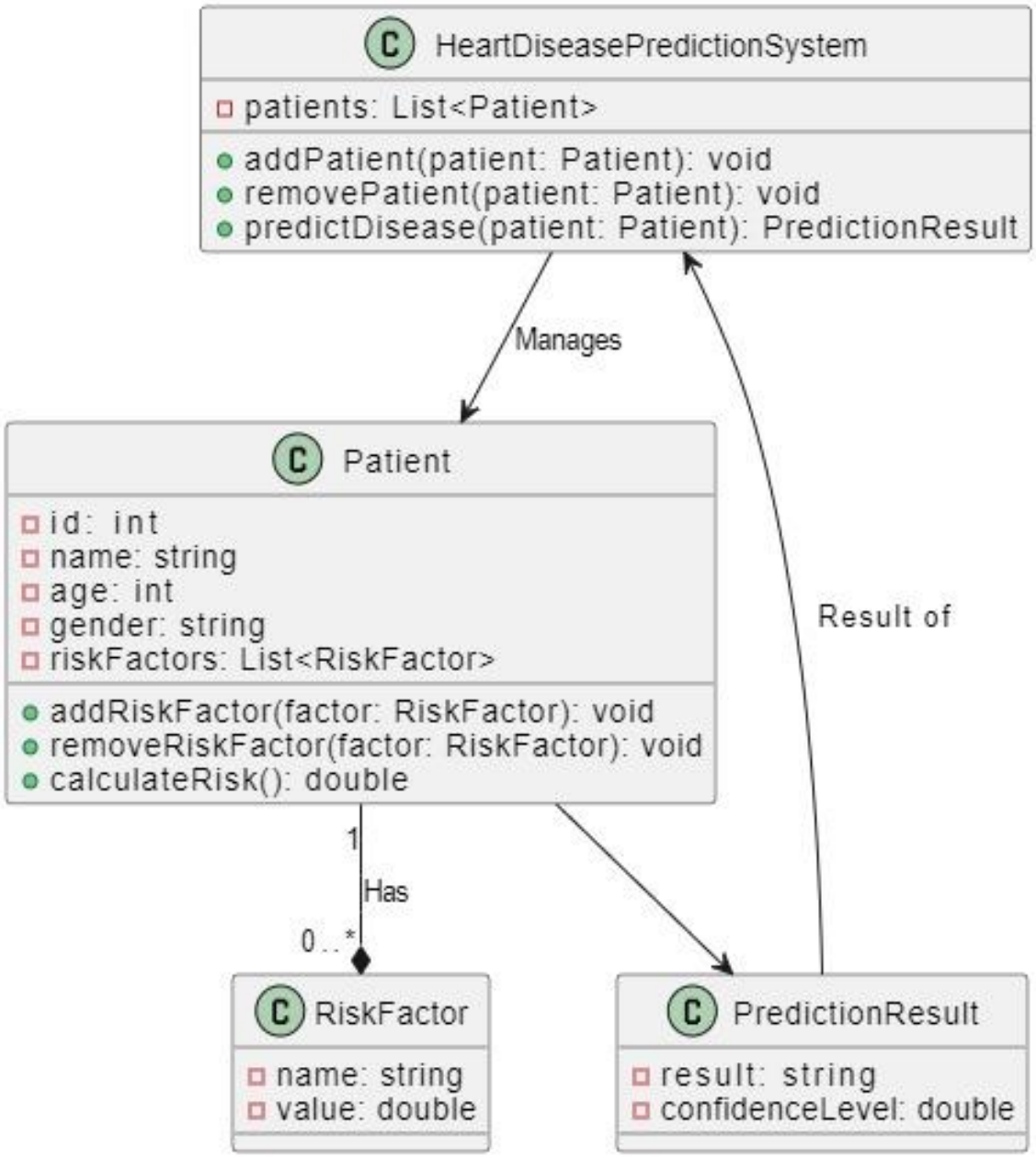


Figure 4.2 Class Diagram

4.1.3. ACTIVITY DIAGRAM

Activity diagram is another important diagram in UML to describe the dynamic aspects of a system. Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. The control flow is drawn from one operation to another. This flow can be sequential, branched, or concurrent. Activity diagrams deal with flow control by using different elements such as fork, join.

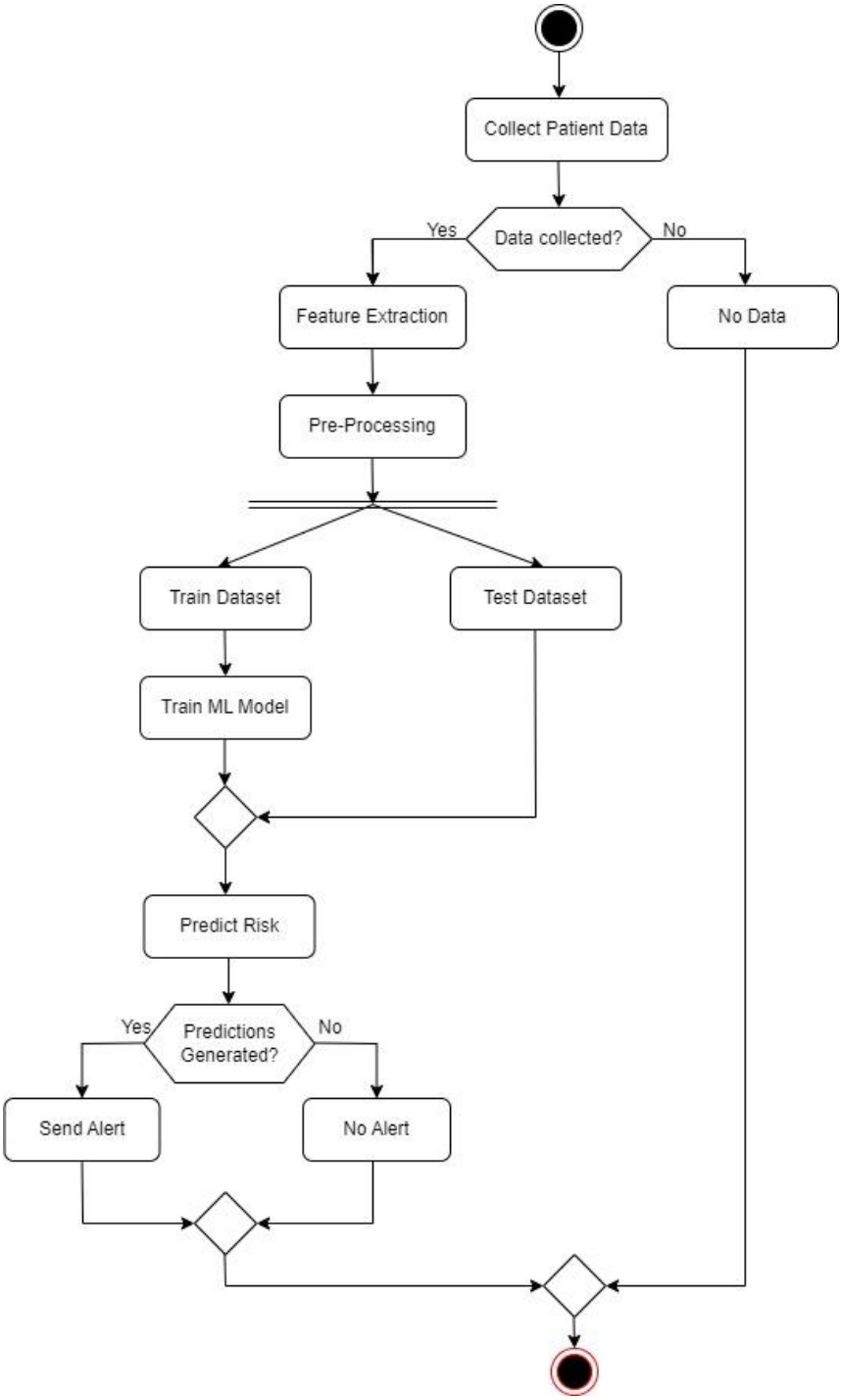


Figure 4.3 Activity Diagram

4.1.4. SEQUENCE DIAGRAM

A sequence diagram or system sequence diagram (SSD) shows process interactions arranged in time sequence in the field of software engineering. It depicts the processes involved and the sequence of messages exchanged between the processes needed to carry out the functionality.

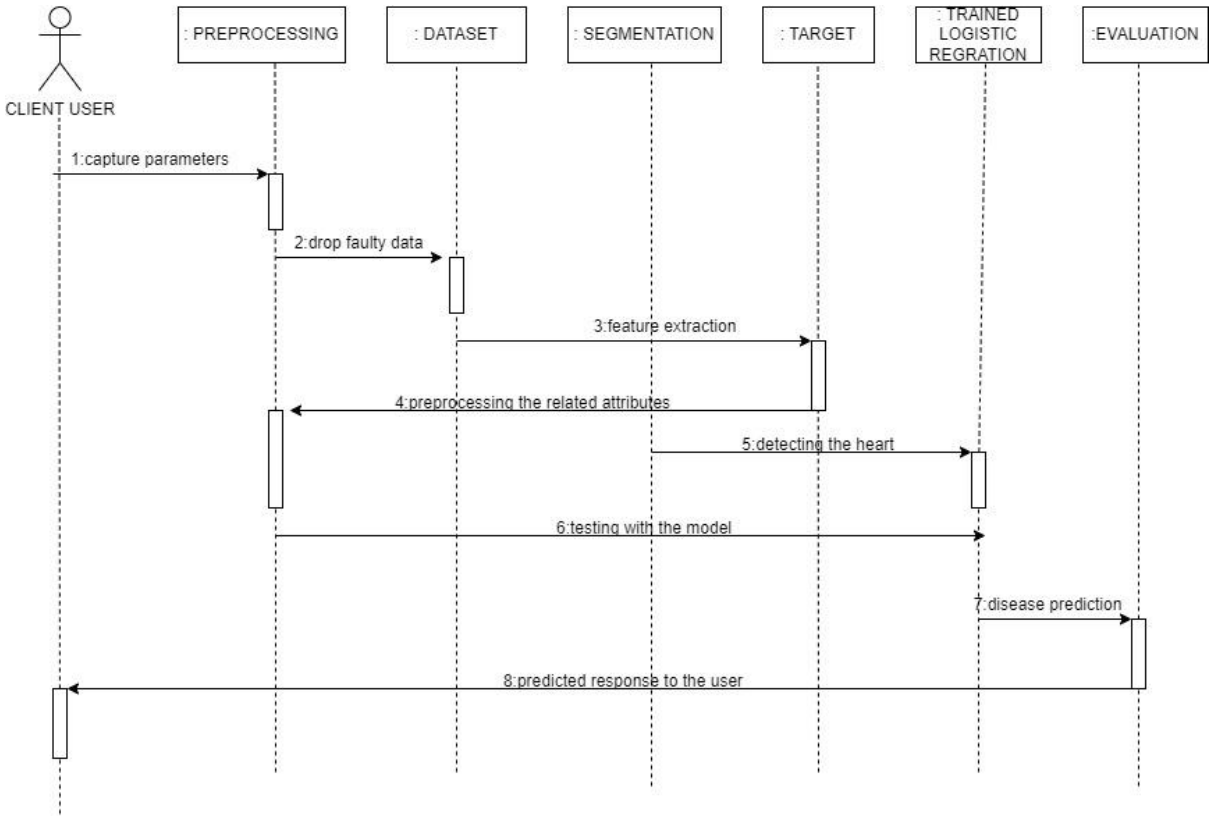


Figure 4.4 Sequence Diagram

5. IMPLEMENTATION

The implementation of the project has been carried out in a step-by-step manner. A detailed description of each module is given below and it is followed by an introduction to the technologies used in implementing the project.

5.1. MODULE DESCRIPTION

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

1.) Collection of Dataset

2.) Selection of attributes

3.) Data Pre-Processing

4.) Balancing of Data

5.) Disease Prediction

Collection of dataset :

Initially, we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 70% of training data is used and 30% of data is used for testing. The dataset used for this project is Heart Disease UCI. The dataset consists of 76 attributes; out of which, 14 attributes are used for the system.

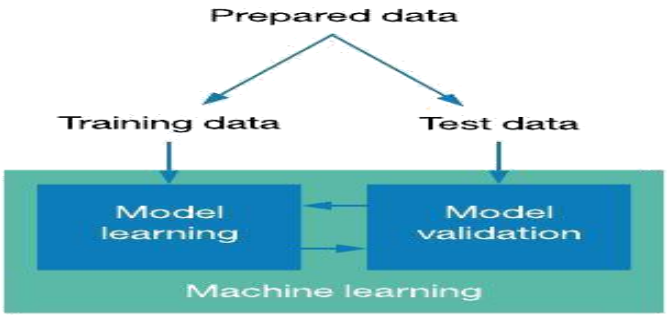


Figure 5.1 Machine Learning Model

Selection of attributes:

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction. The Correlation matrix is used for attribute selection for this model.

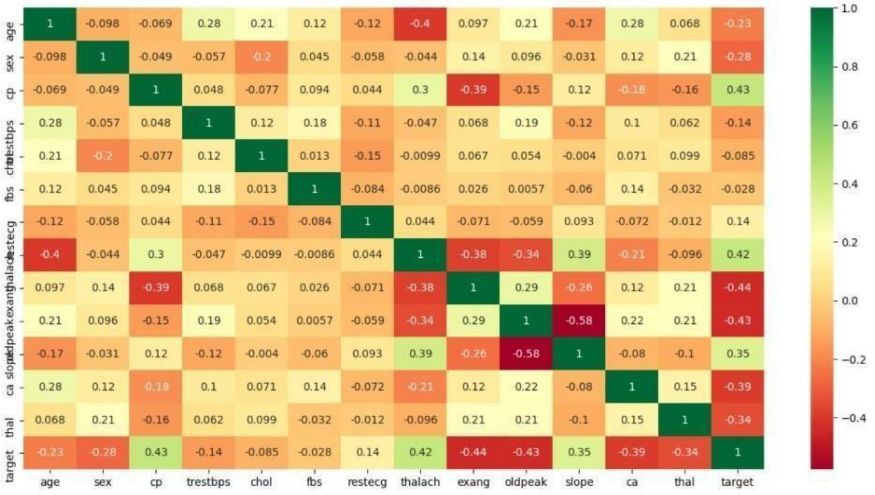
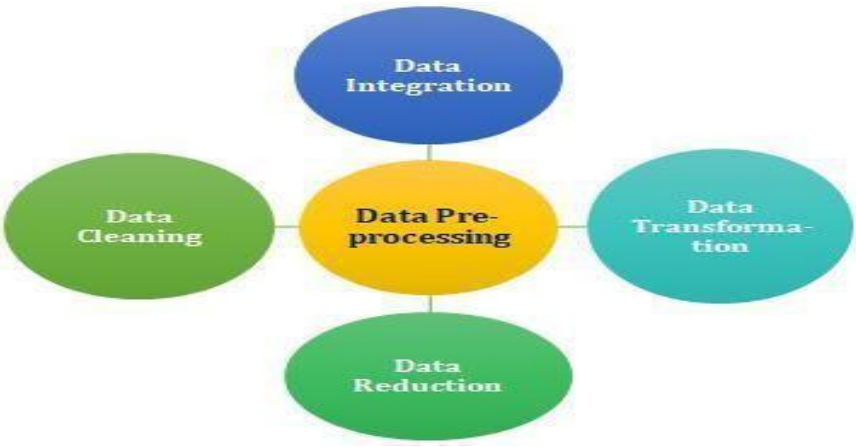


Figure 5.2 Correlation matrix

Pre-processing of Data :

Data pre processing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Pre processing of data is required for improving the accuracy of the model.

Figure 5.3 Pre processing



Balancing of Data:

Imbalanced datasets can be balanced in two ways. They are Under Sampling and Over Sampling

(a) Under Sampling:

In Under Sampling, dataset balance is done by the reduction of the size of the ample class. This process is considered when the amount of data is adequate. (b) Over Sampling:

In Over Sampling, dataset balance is done by increasing the size of the scarce samples.This process is considered when the amount of data is inadequate.

Prediction of Disease:

Machine learning algorithm like Logistic Regression, is used for classification. This algorithm that gives the highest accuracy is used for heart disease prediction.

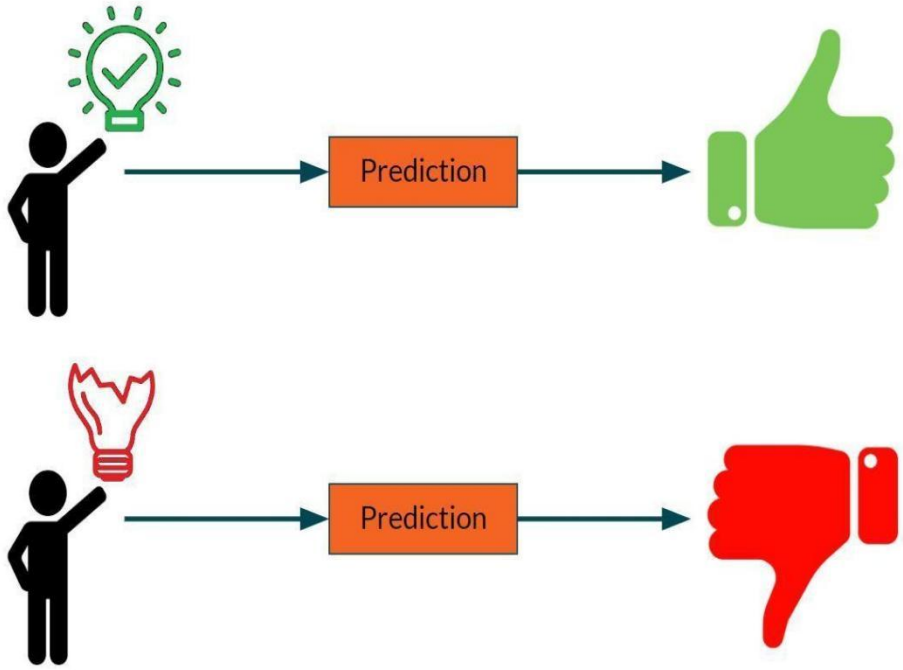


Figure 5.4 Prediction

5.2. INTRODUCTION TO TECHNOLOGIES USED

Python

Python is a high-level, versatile, and dynamically-typed programming language known for its simplicity and readability. It is widely used in various domains, including web development, data analysis, machine learning, artificial intelligence, scientific computing, and more.

Numpy

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

Pandas

Pandas is a popular open-source Python library used for data manipulation and analysis. It provides easy-to-use data structures and functions for working with structured data, such as tabular data in spreadsheets or databases. Pandas is a fundamental tool in data science, machine learning, and data analysis workflows.

Matplotlib

Matplotlib is a widely used open-source Python library for creating static, animated, and interactive visualizations in various formats. It is a fundamental tool in data visualization and scientific computing, providing a flexible and extensive set of functionalities to create high-quality plots and charts.

Seaborn

Seaborn is a Python data visualization library based on Matplotlib. It is designed to make creating informative and attractive statistical graphics easier. Seaborn provides a high-level interface for creating a variety of statistical plots, including complex visualizations, with just a few lines of code.

sklearn

Scikit-Learn, often abbreviated as sklearn, is a popular and versatile open-source Python machine learning library. It provides a wide range of tools and algorithms for various machine learning tasks, including classification, regression, clustering, dimensionality reduction, and model evaluation.

Google colab

Google Colab, short for Google Colaboratory, is a free cloud-based platform provided by Google that allows you to write, run, and share Jupyter Notebook documents with Python code. It's designed for data science, machine learning, and deep learning tasks and has gained popularity in the data science and research communities.

5.3 SOURCE CODE

```
# Import necessary libraries
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, classification_report, confusion_matrix
import matplotlib.pyplot as plt

# Load the dataset (replace 'heart_disease.csv' with your dataset path)
data = pd.read_csv('heart_disease.csv')
# Replace any missing values if necessary (e.g., data = data.fillna(method='ffill'))
# Split data into features (X) and target (y)
X = data.drop('target', axis=1)
y = data['target']
# Split the dataset into training and testing sets
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
# Create and train the Logistic Regression model
clf = LogisticRegression(random_state=42)
clf.fit(X_train, y_train)
# Make predictions on the test set
y_pred = clf.predict(X_test)
# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred)
print(Accuracy: {accuracy:.2f}')
print('Confusion Matrix:\n', conf_matrix)
print('Classification Report:\n', class_report)
plt.barh(range(len(feature_names)), coefficients[sorted_idx])
plt.yticks(range(len(feature_names)), [feature_names[i] for i in sorted_idx])
plt.xlabel('Coefficient Value')
plt.show()
```

RESULTS AND DISCUSSIONS

After performing the machine learning approach for training and testing we find that accuracy of the Logistic Regression is better compared to other algorithms. Accuracy is calculated with the support of the confusion matrix of the algorithm, here the number count of TP, TN, FP, FN is given and using the equation of accuracy, value has been calculated and it is concluded that extreme gradient boosting is best with 85% accuracy

Screen Shots

Copy of final code

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

```
[5] import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
    from sklearn.model_selection import train_test_split
    from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import accuracy_score
    from sklearn.metrics import confusion_matrix
    from sklearn import metrics
```

```
heart_data = pd.read_csv('/content/heart (1).csv')
```

```
[7]
heart_data.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

```
[8] heart_data.tail()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

Figure 6.1 Execution Image

```
[9] heart_data.shape

(303, 14)
```

```
heart_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

```
[11] heart_data.isnull().sum()

age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

Figure 6.2 Execution Image

```
[12] heart_data.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Figure 6.3 Execution Image

```
plt.figure(figsize=(8, 6))
sns.countplot(x=heart_data["target"])
plt.xlabel('target')
plt.ylabel('Frequency')
plt.title('Distribution of Target')
plt.show()
```

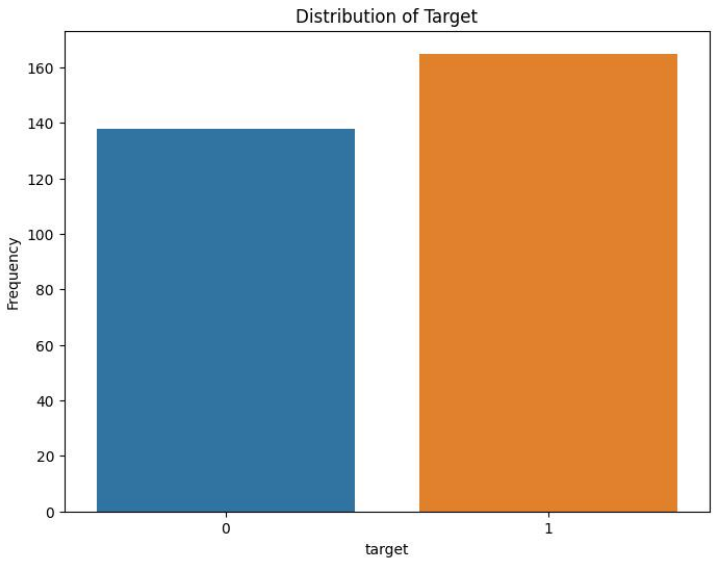


Figure 6.4 Execution Image

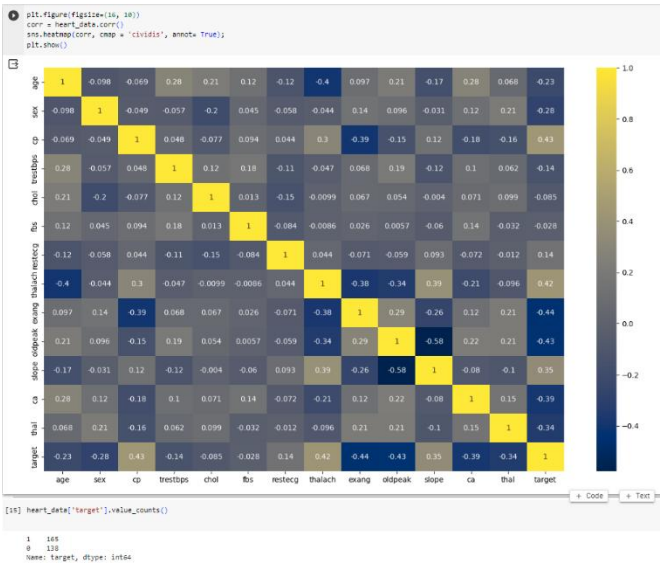


Figure 6.5 Execution Image

```
print(X.shape, X_train.shape, X_test.shape)

(303, 13) (242, 13) (61, 13)
```

```
[21] model = LogisticRegression()
```

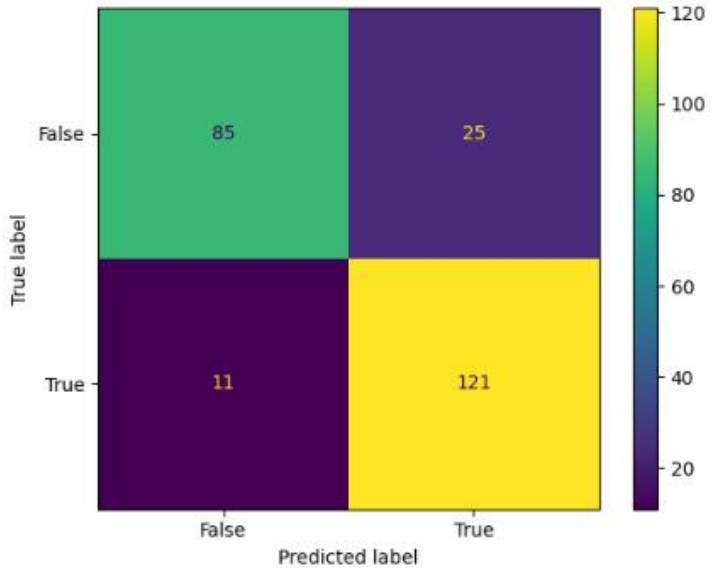
```
[22] model.fit(X_train, Y_train)

/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
▾ LogisticRegression
LogisticRegression()
```

```
[23] X_train_prediction = model.predict(X_train)
     training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

```
[24] print('Accuracy on Training data : ', training_data_accuracy)

     Accuracy on Training data :  0.8512396694214877
```

Figure 6.6 Execution Image

```
[25] confusion_mat = confusion_matrix(Y_train, X_train_prediction)

     cm_display = metrics.ConfusionMatrixDisplay(confusion_matrix = confusion_mat, display_labels = [False, True])

     cm_display.plot()
     plt.show()
```




```
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

```
[27] print('Accuracy on Test data : ', test_data_accuracy)

     Accuracy on Test data :  0.819672131147541
```

Figure 6.7 Execution Image

```
#(age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ac,thal)
input_data = [29,1,1,130,204,0,0,202,0,0,2,0,2]

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

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

input_data = pd.DataFrame(input_data_reshaped, columns=X_train.columns)
prediction = model.predict(input_data)
print(prediction)

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

```
[1]
The Person has Heart Disease
```

Figure 6.8 Execution Image

DEPLOYMENT:

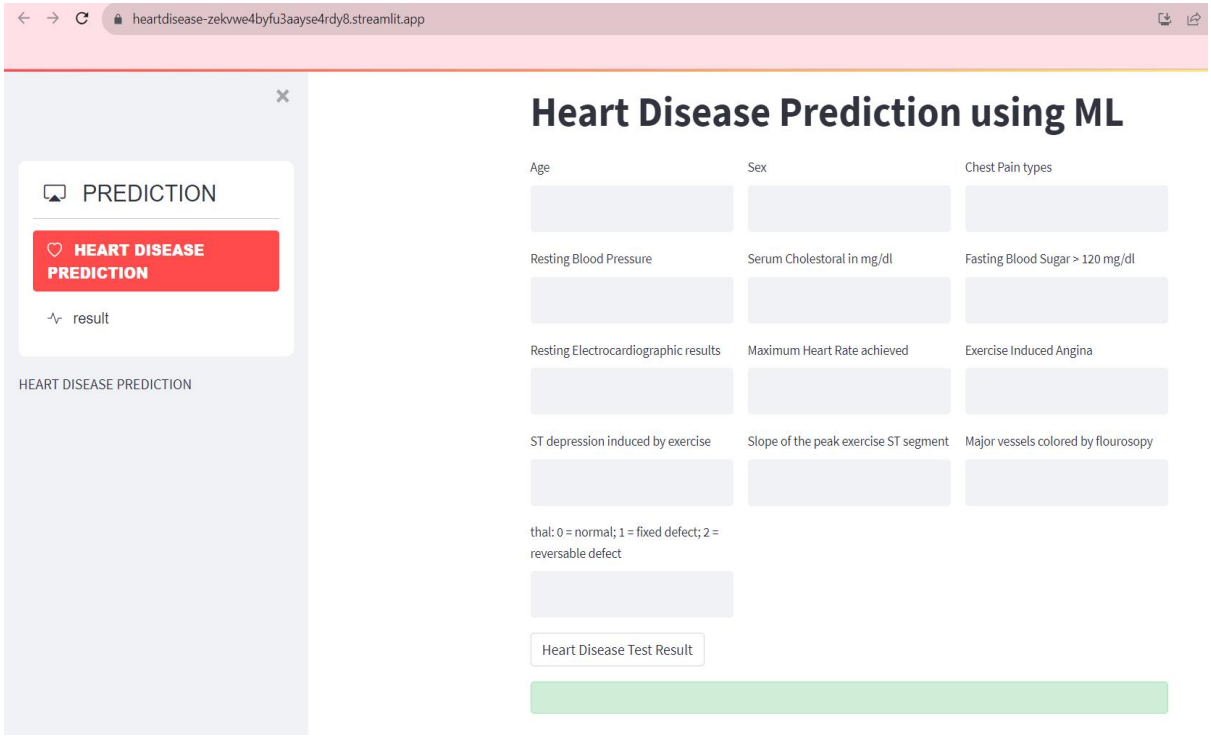


Figure 7.1 Execution Image

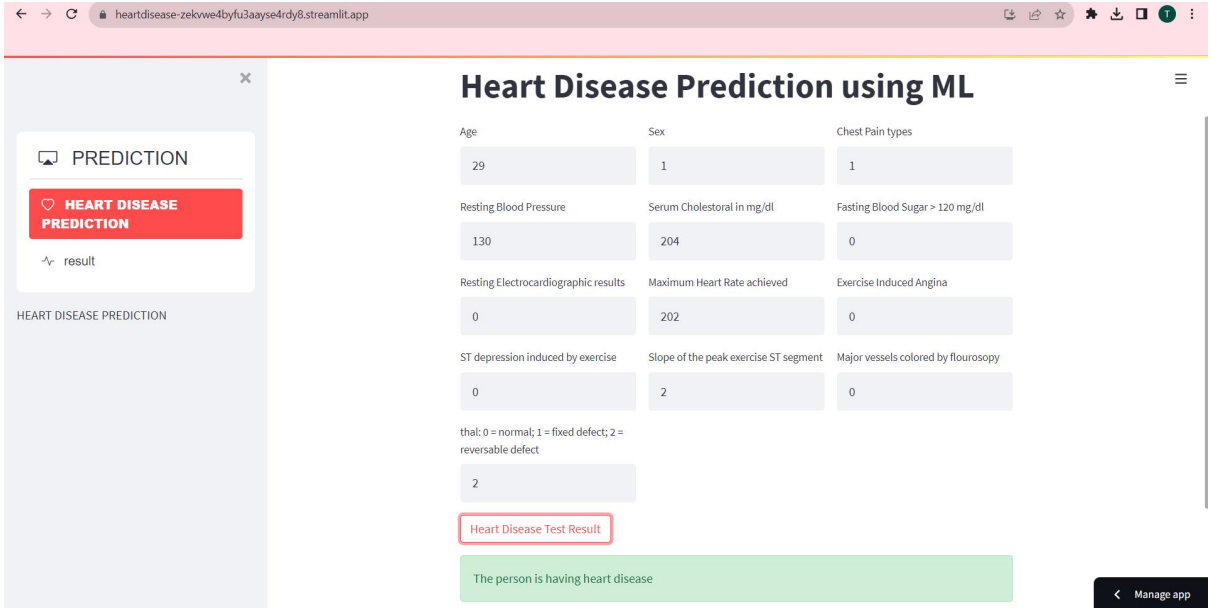


Figure 7.2 Execution Image

THE PERSON IS HAVING HEART DISEASE

CONCLUSION AND FUTURE ENHANCEMENT

Heart diseases are a major killer in India and throughout the world, application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. The early prognosis of heart disease can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. The number of people facing heart diseases is on a raise each year. This prompts for its early diagnosis and treatment. The utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients.

After performing the machine learning approach for training and testing we find that accuracy of the Logistic Regression is better compared to other algorithms. Accuracy is calculated with the support of the confusion matrix of each algorithm, here the number count of TP, TN, FP, FN is given and using the equation of accuracy, value has been calculated and it is concluded that logistic Regression is best with 85% accuracy.

Finally comparing the results We get that accuracy of logistic Regression is about 85% . Heart diseases are one of the major concerns of society and the number of people affected by these diseases is increasing day by day and it is important to find a solution so with the help of data analytics and machine learning models ,we can determine these diseases and have a better chance of treating it.

BIBLIOGRAPHY

[1] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8

[2] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.

[3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

[4] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.

[5] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259-64). IEEE. ICCRDA 2020 IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757-899X/1022/1/012072 9

[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.

[7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

[8] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International Mutli-Conference on Automation, Computing,Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE.