Heart disease classification environment and building the predictive analytics by Machine Learning Algorithms

1Gowthami T S,2Dr.Lokesh.A,3Dr.Krishna Kumar P.R

1M.Tech Final Year Student, 2Professor, 3HoD,Dept.of CSE

1 Dept.of Computer Science & Engineering

1,2,3S.E.A College of Engineering and Technology, K.R.Puram, Bangalore-560049

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Abstract :* The real-time forecasting tools are becoming increasingly crucial in the healthcare industry today. Through this initiative, we will be able to anticipate heart disease in real-time by utilizing patient data from the model that employs the Logistic Regression Algorithm, thus enabling precise heart disease predictions through machine learning. From our results, we achieved an accuracy of 85% with both logistic regression and the random forest classifier. We conclude that logistic regression is preferable since the random forest classifier's performance is not as strong and it tends to be more time-consuming. Additionally, while it can be applied to unstructured data, our focus here is on structured data, making logistic regression a more straightforward option for implementation.

*IndexTerms* - K Nearest Neighbor Algorithm (KNN), Logistic Regression, Decision Trees (DT), Genetic algorithm (GA), and Naïve Bayes *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

# **Introduction**

Cardiovascular disease (CVD) includes various disorders impacting the circulatory system, such as angina pectoris, myocardial infarction, coronary artery disease, heart failure, arrhythmias, and others. From 1990 to 2013, the worldwide mortality rate due to cardiovascular diseases increased by 41%, rising from 12. 3 million to 17. 3 million deaths. Additionally, roughly half of the deaths in the United States and other developed countries are linked to these conditions. As a result, early detection of heart ailments is vital to reduce health complications. In modern healthcare, machine learning has gained popularity for diagnosing and predicting disease occurrence through data models. Recognizing heart disease is difficult owing to various contributing risk factors, including diabetes, hypertension, high cholesterol, irregular pulse rates, and several other factors. Various methods in data mining and neural networks have been employed to evaluate the severity of heart disease in patients. The classification of disease severity is carried out using different techniques such as the K-Nearest Neighbor Algorithm (KNN), Logistic Regression, Decision Trees (DT), Genetic Algorithms (GA), and Naïve Bayes (NB). Given the complex nature of heart disease, it should be approached carefully, since overlooking it can result in heart damage or untimely death. Both medical science and data mining viewpoints are utilized to reveal various types of metabolic syndromes. Data mining, especially classification, is integral in predicting heart disease and performing data analysis. Machine Learning is utilized across numerous domains globally, and the healthcare industry is included. It can considerably aid in forecasting the presence or absence of locomotor disorders, heart diseases, and more. Reliable predictions can provide significant insights to physicians, allowing them to customize their diagnoses and treatments for individual patients. Logistic regression is one of the frequently applied machine learning algorithms for investigations aimed at evaluating the risk of complex diseases. Therefore, this study aspires to determine the most vital predictors of cardiovascular diseases and assess the overall risk using logistic regression.

**1.1 Background**

Among all life-threatening illnesses, cardiovascular diseases, particularly heart attacks, are regarded as the most widespread. Healthcare professionals conduct various studies on heart-related ailments, collecting data on heart patients, their symptoms, and the evolution of their conditions. An increasing number of reports highlight patients suffering from common illnesses who exhibit standard symptoms. In today’s fast-paced society, individuals aspire to lead luxurious lives, often working tirelessly to earn substantial income and enjoy comfort. In this pursuit, they frequently neglect their health; their dietary habits shift, and their overall lifestyle transforms. This kind of living results in heightened stress levels, leading to issues such as hypertension and diabetes at an unusually young age. They often overlook the importance of rest, consuming whatever food is available without considering its quality. Instead of seeking professional medical advice when feeling unwell, they tend to self-medicate. Such small oversights can culminate in significant health risks, particularly heart disease. The term ‘heart disease’ encompasses a variety of conditions that impact the heart. The prevalence of heart disease is increasing (health topics, 2010). Reports from the World Health Organization indicate a staggering number of fatalities attributed to heart disease globally each year. It is also highlighted as one of the leading causes of death in Africa. Data mining has found applications in numerous fields such as marketing, customer relationship management, engineering, and medical analysis, including expert forecasting, web mining, and mobile computing. Recently, data mining has been effectively utilized in uncovering healthcare fraud and identifying cases of misuse. Data analysis is pivotal in the medical sector, providing a solid foundation for critical decision-making. It aids in the formulation of comprehensive research proposals. One of the most significant advantages of data analysis is its ability to minimize human bias in medical conclusions through proper statistical methods. By employing data mining for exploratory analysis, we can extract meaningful insights from large datasets. The healthcare sector generates vast amounts of data that hold concealed information, valuable for making informed decisions that lead to favorable outcomes. A heart prediction system will leverage data mining techniques to unveil user-friendly insights and identify new and hidden patterns within the data. The knowledge derived can be utilized by healthcare professionals to enhance service quality and mitigate the risks of adverse medication effects.

**1.2 Aims and Objectives of research work**

* The provided data is restructured, and the percentage of patients with heart disease as well as those without is calculated and presented.
* The percentage of heart disease cases is identified and displayed for both female and male patients.
* Age-based classification is performed, and a pie chart is created.
* The results of the physical examination are extracted, and the analysis of chest pain is stored in an array, followed by the creation of a bar graph.
* A pairwise correlation analysis of all columns is conducted, and the results are presented. ¬ Summary statistics, including count, mean, standard deviation, minimum, and maximum values for numeric variables, are compiled..

**2.0 PROBLEM STATEMENT**

Heart disease can be effectively managed through a blend of lifestyle modifications, medication, and sometimes surgery. With appropriate treatment, the symptoms of heart disease can be alleviated and the heart's performance enhanced. The anticipated outcomes can be utilized to prevent and therefore lower expenses related to surgical interventions and other costly procedures. The overarching aim of my work will be to accurately predict the existence of heart disease with minimal tests and attributes. The attributes evaluated form the foundational basis for tests and yield results that are more or less precise. Many additional input attributes can be considered, but our objective is to forecast the risk of heart disease with fewer attributes and greater speed. Decisions are frequently made based on the intuition and experience of doctors rather than the knowledge-rich data concealed within the dataset and databases. This approach results in undesired biases, mistakes, and increased medical expenses that hinder the quality of care offered to patients. Data mining presents significant opportunities for the healthcare sector to facilitate health systems in systematically leveraging data and analytics to pinpoint inefficiencies and best practices that enhance care and decrease costs. According to (Wurz and Takala, 2006) the chances to simultaneously improve care and cut costs could pertain to as much as 30% of total healthcare expenditure. The fruitful implementation of data mining in highly visible domains such as e-business, marketing, and retail has sparked its adoption across various industries and sectors. Among these sectors, healthcare is just beginning to uncover its potential. The healthcare landscape is still "information rich" yet "knowledge poor. " There exists a vast amount of data within healthcare systems. Nevertheless, there is a deficiency of effective analytical tools to unveil hidden connections and trends in the data for African contexts.

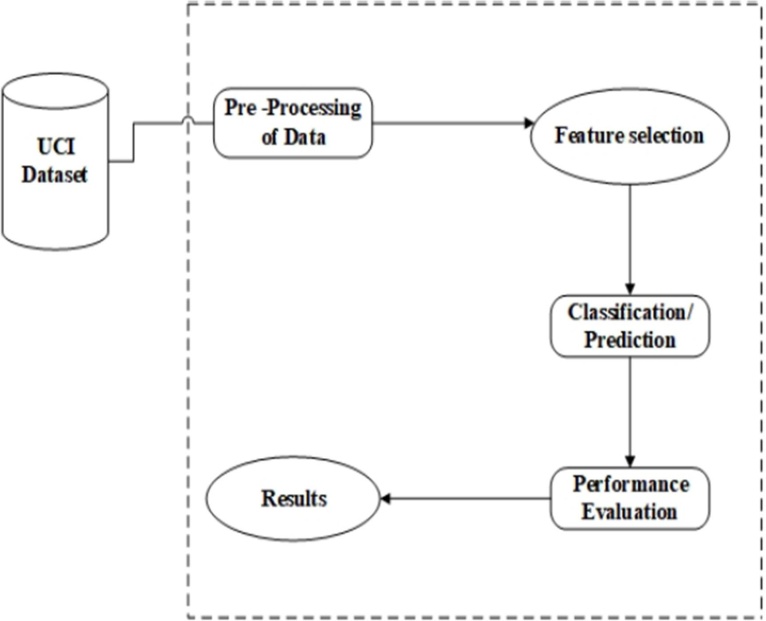
**2.1 Existing system**

In this system, data is collected from the patient. Then, by using user inputs and applying ML techniques, heart disease is evaluated. Next, the results produced are compared with the results of contemporary models in the same area and discovered to demonstrate enhancement. The data from heart disease patients obtained from the UCI laboratory is utilized to recognize patterns using NN, DT, Support Vector machines (SVM), and Naïve Bayes. The efficacy and precision of these algorithms are assessed by comparing their results. The proposed hybrid approach achieves an F-measure result of 87%, competing with other current methods

**2.2 Proposed system**

After assessing the outcomes from the current methodologies, we have employed Python and Pandas operations to execute heart disease classification for the data acquired from the UCI repository. It offers a user-friendly visual representation of the dataset, development environment, and predictive analytics construction. The ML process begins with a data pre-processing phase, succeeded by feature selection founded on data cleaning and evaluation of classification modeling performance. The random forest technique is utilized to enhance the accuracy of the results.

**3.0 Implementation**

****Fig1.1UMLdesign

* + 1. **Data Acquisition:** A dataset is a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer. The data set used in this case is the Cleveland Heart Disease dataset taken from the UCI repository. The data set consists of a data of 303 individuals.
    2. **Data pre processing:** In order to build an accurate ML model, data pre processing is required. Data pre process is the process of cleaning the data. This includes identification of missing data, noisy data and inconsistent data.

The dataset contains a total of 303 patient records, where 6 records are with some missing values. Those 6 records are removed from the dataset and the remaining 297 patient records are used for pre processing. Multiclass and binary classification is introduced to represent the attributes of the given dataset.

The dataset consists of a data of 303 individuals. There are 14 columns in the dataset described below:

* + - * Age : Displays the age of the individual.
      * Sex: Displays the gender of the individual using the convention:

1 =male

2 =female

* Chest pain type: Displays the chest pain experienced by the individual using the convention:
  + - * Resting Blood Pressure: Displays there sting vital sign value of a private in mmHg.
      * Serum Cholesterol: Displays these rum cholesterol Olin mg/dl.
      * Fasting Blood Sugar: Compares the fasting blood glucose value of with a standard value of 120mg/dl.

If fasting blood glucose>120mg/dlthen:1(true)

else:0(false)

Resting ECG: Displays resting electro car diographic results0= normal

1=having ST-T wave abnormality

2=left ventricular hypertrophy

* + - * Maxpulse achieved: Displays the maxpulse achieved.
      * Exercise induced angina:

1=yes

0=no

* + - * Displays the worth which is an integer or float. ST depression

Induced by exercise relative to rest.

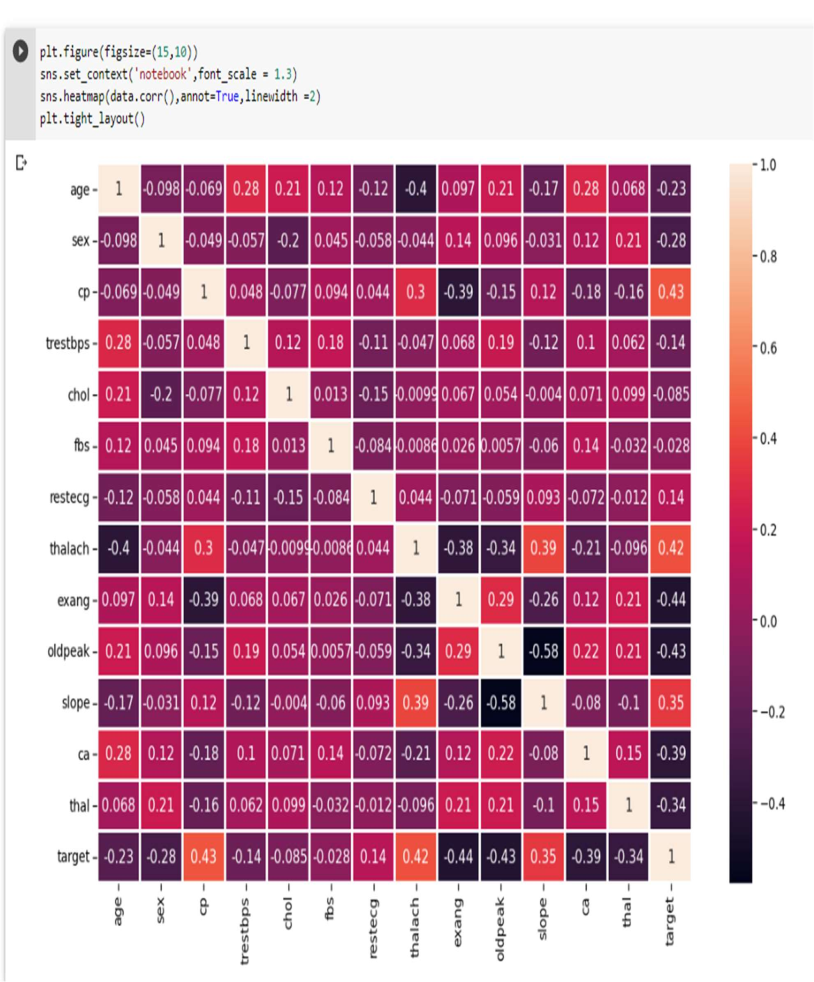
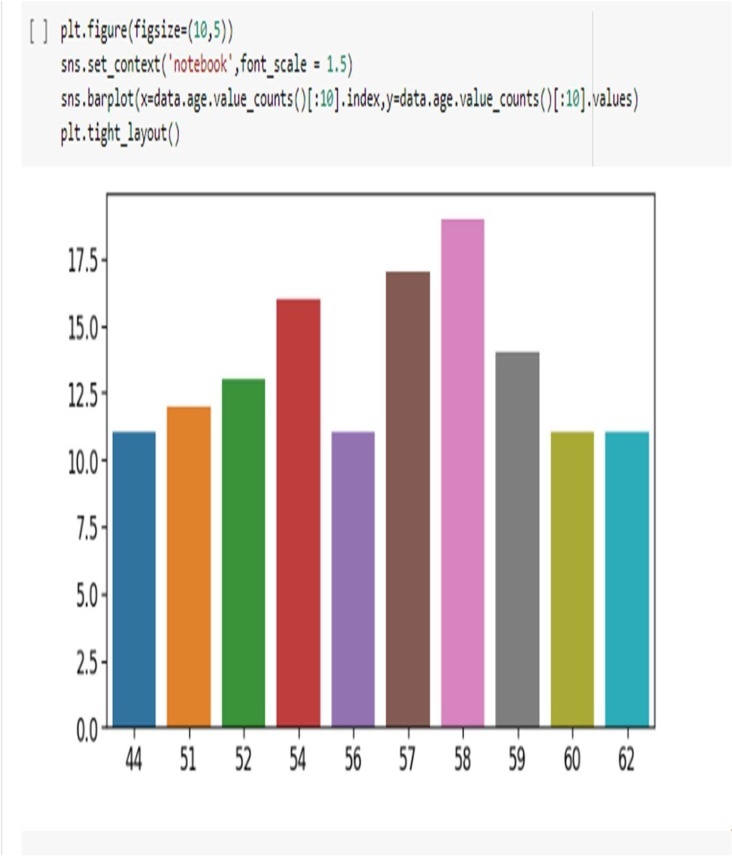
* + - * Peak exercise ST segment:

1=upsloping2=flat

3=down sloping

* + - * Number of major vessels(0–3) colored by fluoro scopy: displays the worth as integer or float.

**4.0 RESULTS**

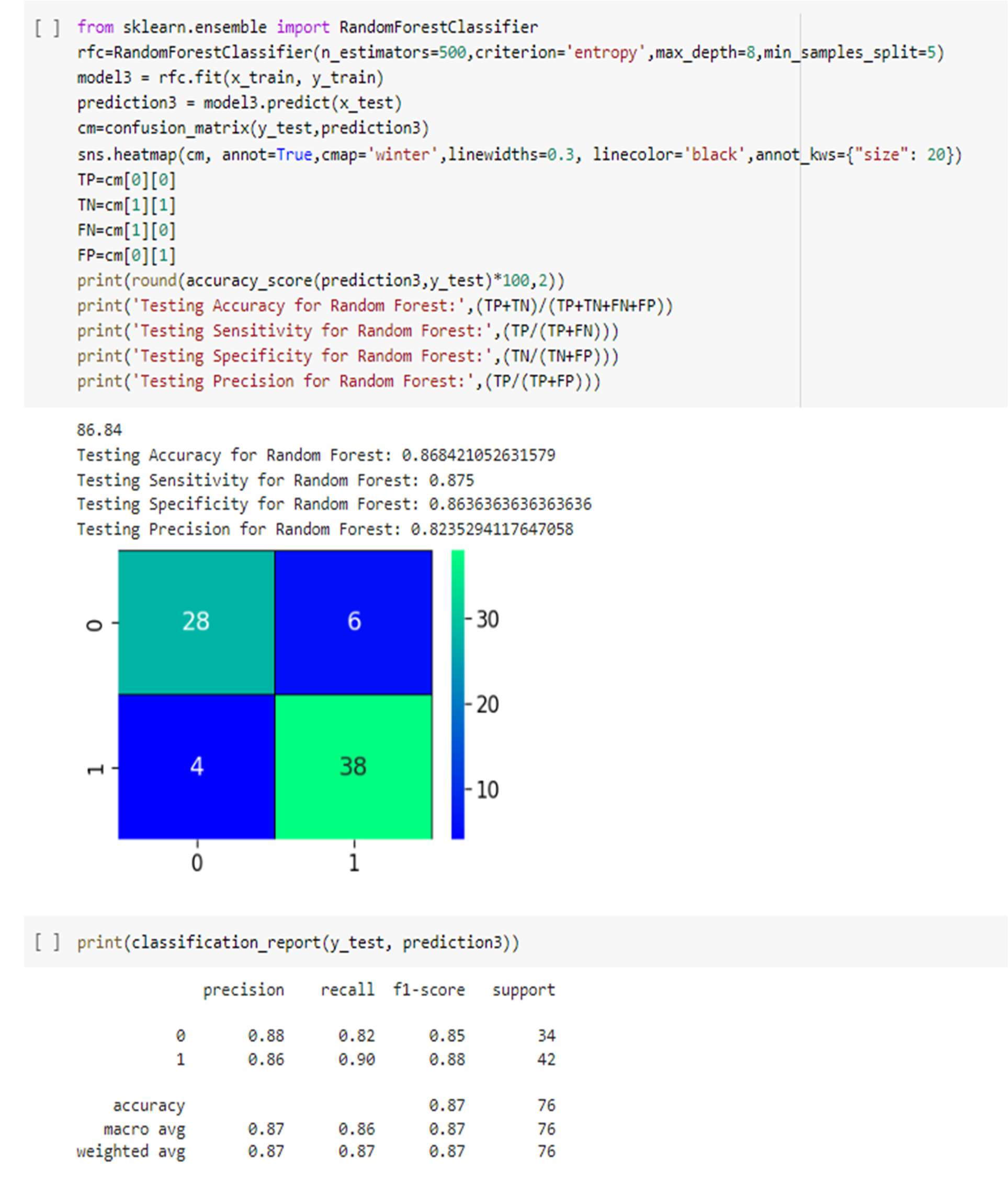
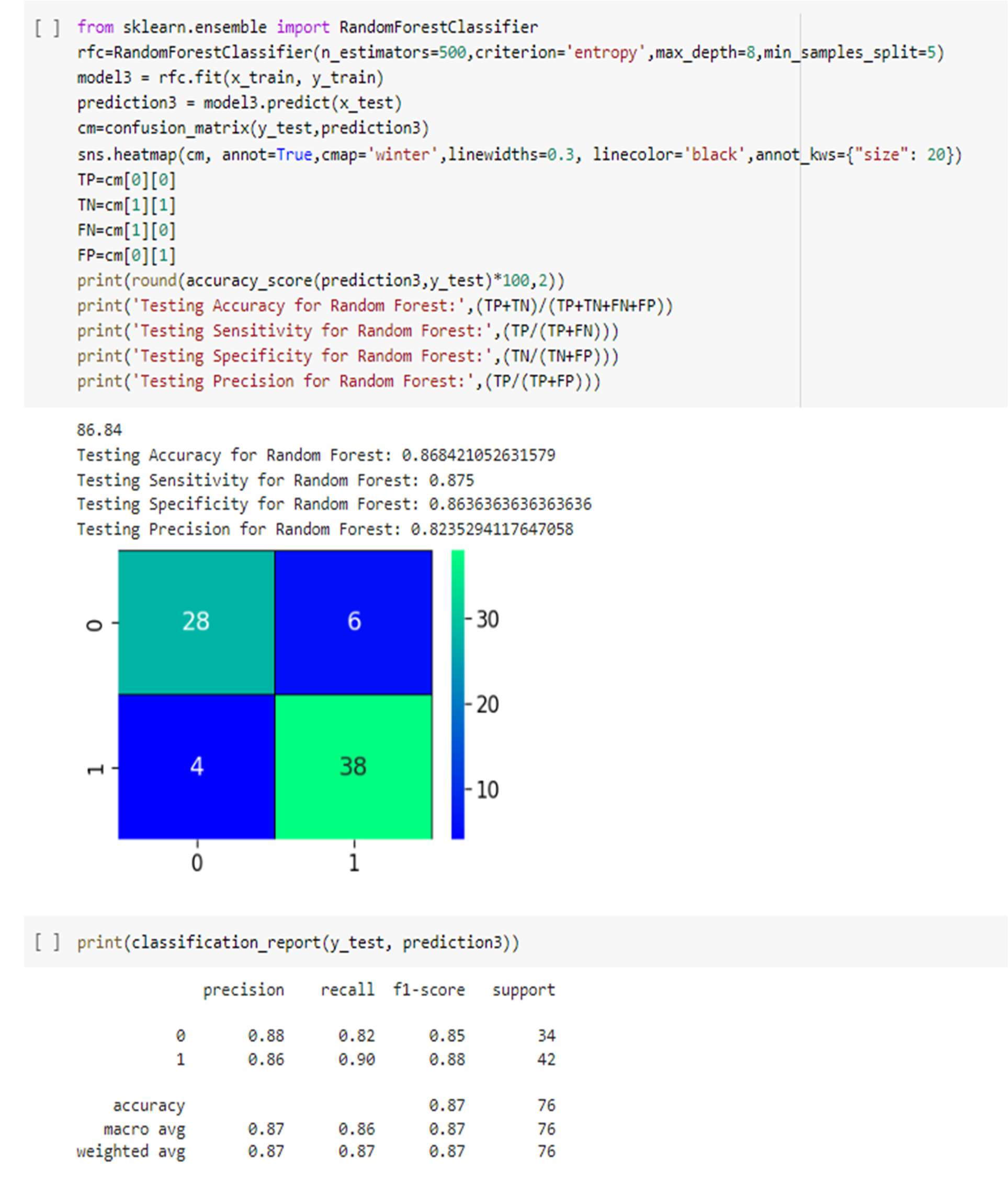
****

#### Fig 2.1 **Correlation Matrix Fig 3.1** Age(“age”) Analysis

# 

Fig.4.1 Classification of age pie chart.

#### The following results derived by Random Forest Classifier



**FUTURE WORK AND SCOPE**

The amount of Heart diseases can exceed the current scenario to reach the maximum point. Heart disease are complicated and each and every year lots of people are dying with this disease.It is difficult to manually determine the odds of getting heart disease based on risk factors previously shown. By using this system one of the major drawbacks of this work is that it’s main focus is aimed only to the application of classifying techniques and algorithms for heart disease prediction, by studying various data cleaning and mining techniques that prepare and build a dataset appropriate for data mining so that we 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. It is still an open domain waiting to get implemented in heart disease predication and increase the accuracy. Today’s, world most of the data is computerized and everything is in the cloud which can be accessed although it is not utilized 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 predicting the heart disease, we can use logistic regression algorithm, 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.

**REFERENCES**

[1] AvinashGolande, Pavan Kumar T. Heart disease prediction using effective machine learning techniques.

[2] The Lancet Global Health. The changing patterns of cardiovascular diseases and their risk factors in the states of India: The global burden of disease study 1990-2016.

[3] Himanshu Sharma, M A Rizvi. Prediction of heart disease using machine learning algorithms: A survey. [4] World health ranking.

[5] Himanshu Sharma, M A Rizvi. Prediction of heart disease using machine learning algorithms: A survey. [6] Sana Bharti,2015. Analytical study of heart disease prediction compared with different algorithms; International conference on computing, communication, and automation (ICCA2015).

[7] Monika Gandhi,2015. Prediction in heart disease using techniques of data mining, international conference on futuristic trend in computational analysis and knowledge management (ABLAZE2015)

[8] SarathBabu, 2017.Heart disease diagnosis using data mining technique, international conference on electronics, communication and aerospace technology (ICECA2017)

[9] A H Chen, 2011. HDPS: heart disease prediction system; 2011 computing in cardiology

[10] Reddy Prasad,Pidaparthi Anjali, S.Adil, N.Deepa(Feb 2019) Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning