PREDICTION AND PERFORMANCE EVALUATION OF HEART DISEASE USING MACHINE LEARNING CLASSIFIERS

Dr. Dhananjay[1], Divya[2]

[1] Professor, Computer Science and Engineering Department, GNDEC Bidar, Karnataka, India
[2]Student, Computer Science and Engineering Department, GNDEC Bidar, Karnataka, India

ABSTRACT- *Coronary illness is one of the most prevalent causes around the globe. Traditionally statistical measures are employed to distinguish ailments in clinical diagnosis. The selection of the classical method of data analysis influences the accuracy of results. It is observed that predicting and identifying the coronary illness needs complex diagnostic data set. The majority of these clinical datasets are scattered, and not structured. Hence, information mining is challenging for extracting data from a broad database. This work target investigates the presentation of dissimilar information mining procedures. The proposed method employs Gaussian Naive Bayes, Logistic Regression (LR), Decision Tree (ID3), K-Nearest Neighbor, Random Forest taxonomy, on a dataset which contains dissimilar traits like sexual orientation, age, chest torment form, circulatory strain, glucose. The information mining strategy results in novel coronary illness prediction in early stages and improves the accuracy of results.*

Keywords: coronary illness prediction, data mining, machine learning, dataset, AI methods, classification, result analysis.

1. INTRODUCTION

According to the World Health Organization (WHO), 17.9 million individuals pass on consistently because of heart-related diseases. This figure is expected to grow further. With the increase in population and infection, it has become necessary to analyze the diagnostic data along with providing accurate treatment in the early stage of illness. The technological advancement and progress in clinical biomedical research are providing a helping hand in the treatment of disease. However, the information gathered is huge and, commonly, this information can be loud. As, these datasets which are excessively overpowering for human personalities to grasp, can be effortlessly examined using different AI methods. For diagnosis entire medical history and physical tests are used. These tests produce a large amount of data and hence machine learning can be used for finding important features from a large amount of data. Due to this specialty of machine learning, it can be utilized in combination with clinical science for the exact determination of coronary illness. As several machine learning techniques have been evolved and in order to achieve the best accuracy of a model ensembles are widely used. Thus, these calculations have gotten helpful, as of late, to predict the presence or absence of heart-related ailments precisely. In this paper we are using 5 calculations as indicated by heart-related illness at the starting period by using several highlights ultimately, we are waiting for the exactness about the result of looking at it.

1.1 Overview of Heart Disease

The human heart is an organ to siphons blood every through the body via way of the circulatory frame, provide oxygen as well as a supplement to the tissue an expelling carbon dioxide dissimilar squander. The term Heart affliction implies the disease of the heart plus vessel system within it. The heart means "cardio." Therefore, every heart disorder fits in with the class of cardiovascular infirmities.

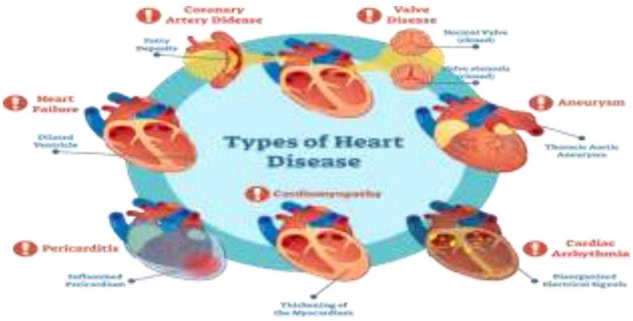


Figure 1 Types of Heart Disease

2. LITERATURE SURVEY

The author in [1] describes the Multivariate investigation strategy suggested a measurable assessment technique utilize in quantitative exploration to smash down numerous factors.[2] S. Palaniappan, R. Awang, describes disclosure of shrouded example plus links frequently go unexploited. Propelled information mining method canister assist cure this circumstance. [3]and [4] in this examination, an exhibition of arrangement strategy was contrasting through foreseeing the nearness of the patient getting a coronary illness [5] proposed that Cardiovascular Disease (CVD) is a significantly imperative supporter of the disappointment of worth just as the amount of life wherever during the globe. Moreover, said about the Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation Random Forest Classification Framework. The author in [6] considered presentations of KNN, Multilayer Perceptron, Radial Basis Function, Single Conjunctive Rule Learner, and Support Vector Machines. [7] This technique gives the most extraordinary accuracy in getting ready data. The general thought is to produce a decision tree that gives equality of flexibility and precision. [8] Considered different part assurance checks and evaluated the display of Naïve Bayes for the finish of coronary sickness patients. The author in [9] depicted that the standard Logistic Regression method gives supported execution over Classification and Regression Trees.

3. PROBLEM STATEMENT

3.1 Existing System

The measure of information in the medical industry is expanding day by day. It is a provoking task to deal with a ton of data and concentrate the helpful information for amazing decision making. Diagnosis of Heart Disease is a very long tiresome and difficult task. The diagnosis of coronary illness in the traditional way incorporates clinical assessment and the many blood tests. Diagnosis plus treatment of coronary illness is perplexing, especially in creating a nation, because of the absence of indicative gadget plus a lack of doctor plus dissimilar asset influencing appropriate expectation as well as healing of the cardiovascular patient. Consequently, information mining is an empowering field of AI and subsequently fit for dealing with this sort of issue very well. For clarifying various kinds of veritable issues, data mining is a novel field for finding hidden patterns and the significant data from a huge dataset. Since it is exceptionally exhausting to remove any significant information without mining a gigantic database. In brief, it is a basic technique for examining information from different points of view and assembling information. In the clinical field, Data mining plays a crucial role in the analysis of disease.

3.2 Proposed System

The proposed system uses the data from the dataset and creates the model which predicts if a patient has heart disease. These days, individuals can face any cardiovascular breakdown side effects at any phase of a lifetime. pc innovation and machine learning trial be utilizing to construct programs to assist specialists in choose coronary illness in the elementary phase. Beginning phase location of the infection as well as predicting the likelihood of a person to be at risk of coronary illness preserve lessen the demise pace. In healthcare services AI strategy techniques are mainly used for making a decision, sickness diagnosing, and giving better treatment to the patients at nearly low cost. Dimensionality decrease might be done as a future work with the goal that the number of blood tests for heart sickness will be reduced and furthermore time required diagnosing ailment. Along these lines, this forecast framework for coronary illness would encourage Cardiologists in taking faster choices so more patients can get medicines inside a shorter period, bringing about sparing a huge number of lives.

4. SYSTEM DESIGN

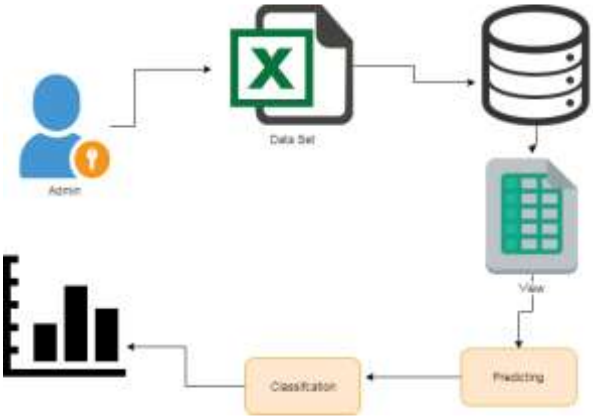


Figure 2: System Architecture

The above figure 2 shows the framework design as the reasonable model that characterizes the structure, conduct, and more perspectives on a framework.

A planning portrayal is a standard outline and representation of a framework, filtered through with the ultimate objective that supports thinking about the structures and practices of the structure. First of all, we collect different patient's data. load the dataset, analyze the features Out of many attributes we only select 13 attributes through feature selection. After extracting the features, we apply the classification algorithms Decision tree, LR, KNN, Naïve Bayes, and Random forest. Admin will train the data and can predict the best classification algorithm based on the accuracy of the result. Admin can plot the graph for the same depending on the accuracy.

5. IMPLEMENTATION

Algorithms used for Prediction and Classification of Heart Disease.

5.1 Random Forest

Random Forest [5] is a troupe sort technique to facilitate depends on the Decision Tree calculation. At the preparation phase, it delivers an immense numeral of plants as well as makes a wood of Decision Trees. This calculation takes a bit of the dataset plus afterward constructs a tree, rehash this progression pro making timberland via unification the produced trees. At the test phase, every tree predicts a class name for every test information, as well as the dominant part estimation of the class name, is doled out to the test information. Accordingly, it indicated sensible execution than the ordinary choice tree calculation of this information. every grouping calculation has its characteristics. pro assorted qualities, the yield of each categorization

5.2 K-Nearest neighbor (KNN)

KNN [6] classifies the test information utilizing the training set legitimately. To characterize any test information, it initially computes K esteem, which indicates the quantity of K-Nearest Neighbors. In this strategy, K-Nearest Neighbors indicated a horrible presentation in light of the fact to KNN arranges test information frankly from the dataset, no preparation is performed before the test.

5.3 Decision Tree (ID3)

Decision Tree (ID3) [7] is a greedy calculation to follows a recursively top-down avaricious method. A choice tree is like the flowchart wherein each non-leaf hub indicates a test on a precise superiority as well as every limb signifies an effect of to test and every leaf hub has a class mark. At the point when the test information plan contained qualities out of this agreed range, the classifier execution was influenced plus along these outlines predict an unsuitable class name.

5.4 Gaussian Naïve Bayes

Naive Bayes [8] classification model does the classification process based on probability. The probability of A with respect to B is given as follows.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

At the preparation phase, naive Bayes [8] determined the mean plus standard divergence of every superiority. When test information design contains those qualities esteems, it influences the classifier execution moreover in some cases gives an unsuitable yield name.

5.5 Logistic Regression

Logistic Regression [9] is a Machine Learning calculation that utilizes pro the group issue, it is a prescient analysis estimate, moreover reliant on the idea of likelihood. The replica can be prepared pro a fixed or as much as no ages via utilizing stochastic angle drop. Coefficients esteem is refreshed until the replica predicts the right class mark pro every preparation information.

5.6 EXPERIMENTAL SETUP

The data pre-processing

The information is gathered as of the UCI AI repository [10]. The informational collection is named Heart Disease train Dataset. which contains 13 attributes and 303 patient records as well as utilizes 10-overlay Cross-Validation to partition the information keen on two segments as train dataset as well as test datasets.

Dataset (Traindata.csv)

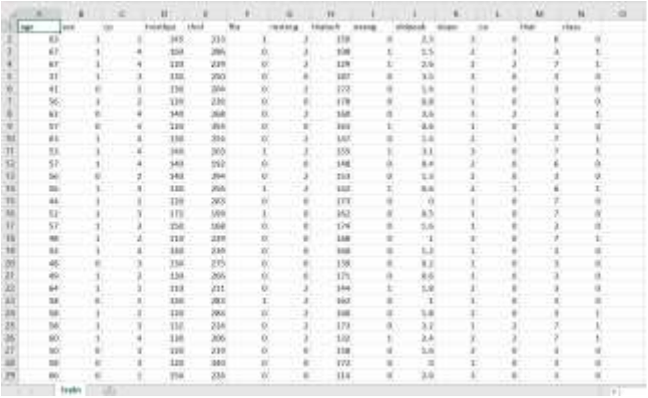
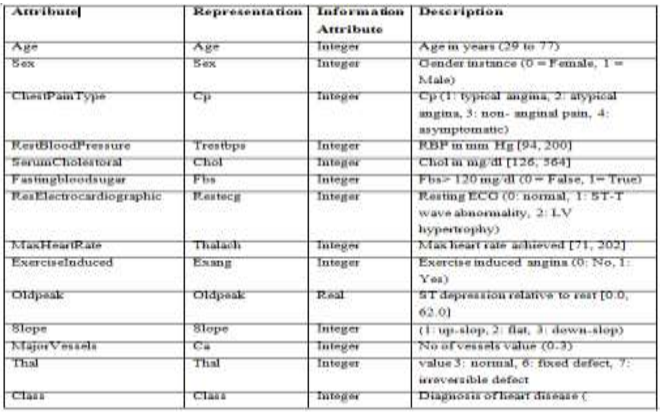


Figure 3: Dataset

Attributes description of the dataset



Attribute|	Representation	Information Attribute	Description
Age	Age	Integer	Age in years (29 to 77)
Sex	Sex	Integer	Gender instance (0 = Female, 1 = Male)
ChestPainType	Cp	Integer	Cp (1: typical angina, 2: atypical angina, 3: non- anginal pain, 4: asymptomatic)
RestBloodPressure	Trestbps	Integer	RBP in mm Hg [94, 200]
SerumCholestoral	Chol	Integer	Chol in mg/dl [126, 564]
Fastingbloodsugar	Fbs	Integer	Fbs> 120 mg/dl (0 = False, 1= True)
ResElectrocardiographic	Restecg	Integer	Resting ECG (0: normal, 1: ST-T wave abnormality, 2: LV hypertrophy)
MaxHeartRate	Thalach	Integer	Max heart rate achieved [71, 202]
ExerciseInduced	Exang	Integer	Exercise induced angina (0: No, 1: Yes)
Oldpeak	Oldpeak	Real	ST depression relative to rest [0.0, 62.0]
Slope	Slope	Integer	(1: up-slop, 2: flat, 3: down-slop)
MajorVessels	Ca	Integer	No of vessels value (0-3)
Thal	Thal	Integer	value 3: normal, 6: fixed defect, 7: irreversible defect
Class	Class	Integer	Diagnosis of heart disease (

Figure 4: Attributes description of the dataset

Analysis of data

It is defined as the process of cleaning, transforming, filling missing values, and modeling the data to give us helpful information for healthcare decision making. The purpose of this is to pre-process the data and to get useful information by data and taking the decisions based upon the data analysis**.**

Programming Environment

In this project, PyCharm is being used as an IDE for python programming language**.** Scikit-learn library contains a lot of efficient tools for ML and statistical modeling including classification, regression, clustering, etc. Pandas is a widely popular library used for data analysis and data manipulation. you need it to import the Pandas package to use it. Using NumPy, mathematical, and logical operations on arrays can be performed. SQLyog Community GUI tool for MySQL.

Admin:

Admin after login, train the dataset through the following five algorithms.

1. Decision Tree (ID3)
2. k-Nearest Neighbor (KNN)

3. Naïve Bayes (NB)
4. Logistic deterioration
5. Random Forest

A different function of Admin:
•After the preparation administrator spirit test the exactness of information as of the preparation document.

•Then the administrator will locate the best group calculation, called Random Forest.

•Admin additionally can plot the illustration of the precision of the five calculations.

User:

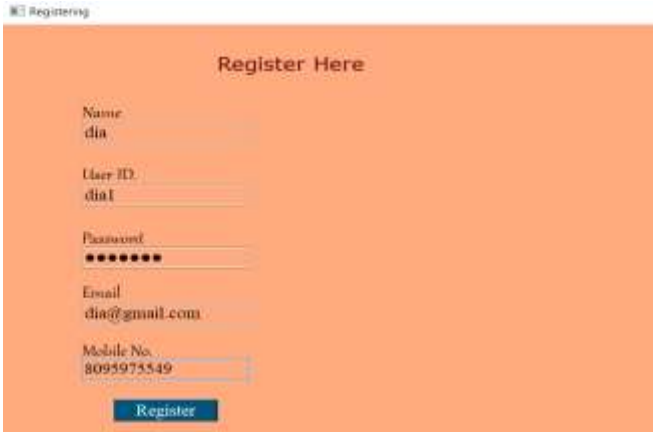
The consumer is the end consumer of the application; our application will assist the client via predicting heart infection by preparing the past patient's dataset the given calculation.

•User can enroll through their own subtleties as well as after login consumer transfer single patient record in the CSV document.

• User can see the outcome through the forecast of the calculation.

6. RESULT ANALYSIS

Below figure 5 shows the registration form for the user. Figure 6 identifies whether a person is having heart disease or not by displaying positive if the person is suffering from heart or else negative. The classification approach is shown in figure 7. The accuracy prediction for k-nearest neighbor is shown in figure 8 and for naïve Bayes is shown in figure 9, for logistic regression algorithm is shown in figure 10, for decision tree algorithm is shown in figure 11, and random forest algorithm is shown in figure 12.



Registering

Register Here

Name
dia

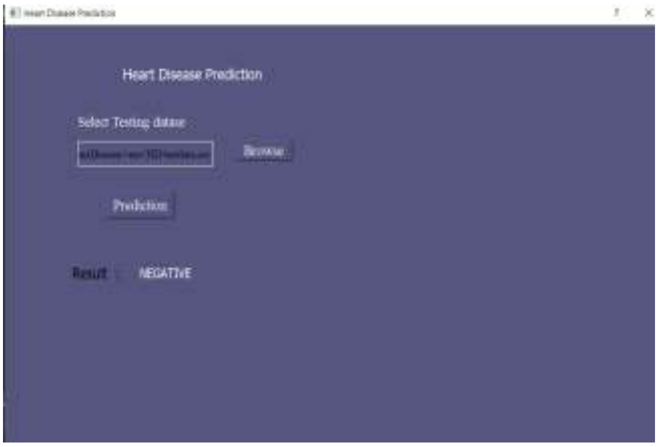
User ID
dia1

Password
●●●●●●●

Email
dia@gmail.com

Mobile No.
8095975549

Figure 5: **Registration form for the user**



Heart Disease Prediction

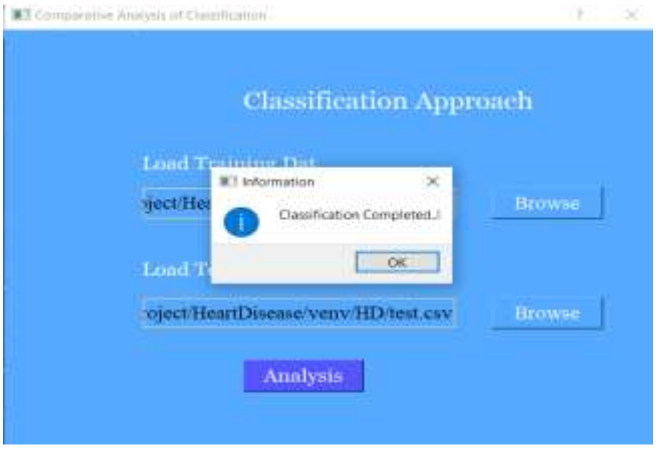
Heart Disease Prediction

Select Testing datase

Result NEGATIVE

Figure 6: **Showing the result**

(Positive or negative)


Comparative Analysis of Classification

Classification Approach

Load Training Dat

ject/He

Information

Classification Completed..!

Load T

oject/HeartDisease/venv/HD/test.csv

Figure 7: **Classification Approach**

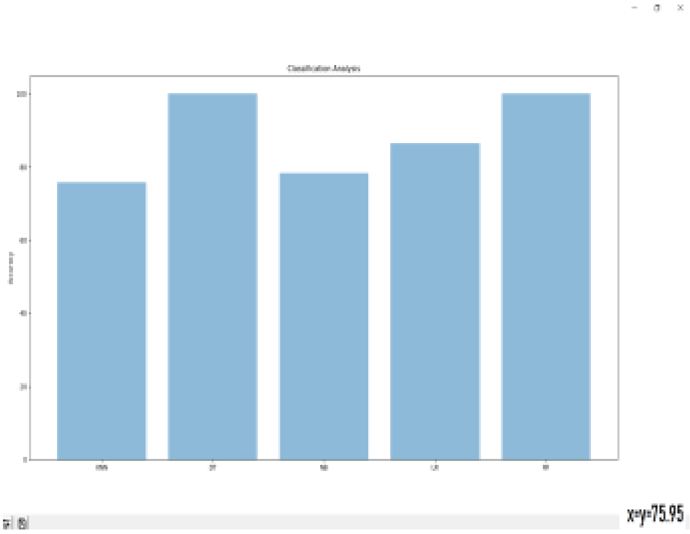


Figure 8: **Accuracy of the K nearest Neighbor Algorithm.**

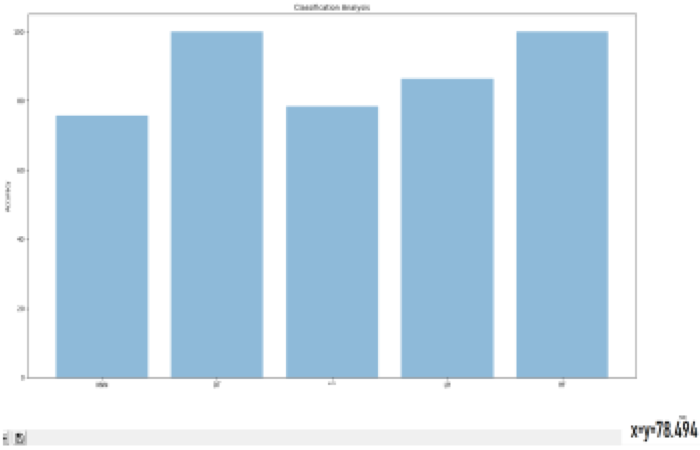


Figure 9: **Accuracy of Naïve Bayes Algorithm**

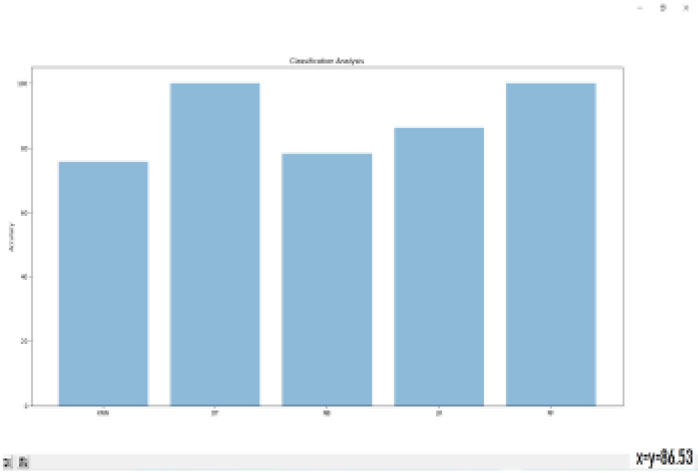


Figure 10: **Accuracy calculation of Logistic Regression.**

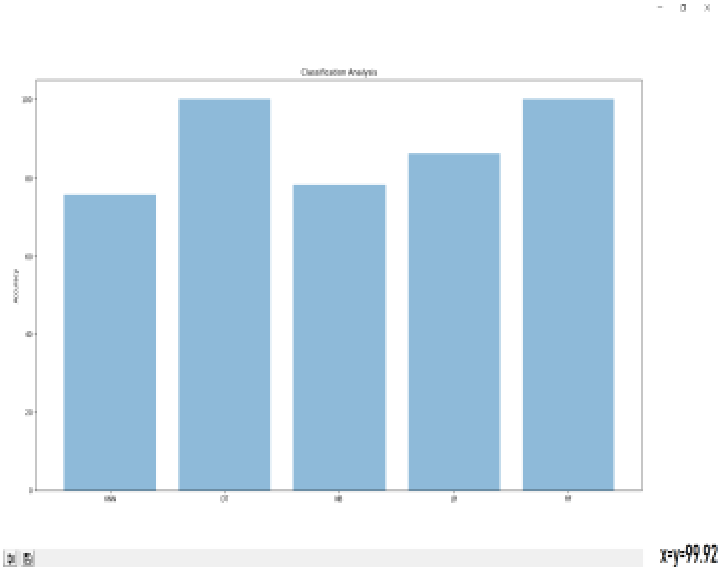


Figure 11: **Accuracy of Decision Tree Algorithm.**

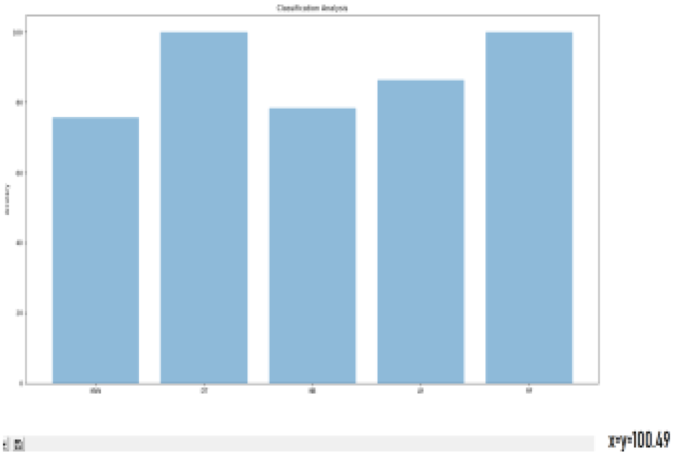


Figure 12: **Accuracy of the Random Forest Algorithm.**

7. CONCLUSION

In this paper, we have proposed a method to predict heart disease at an earlier stage using machine learning classifiers. The foremost goal of this venture is to think about various machine learning calculations plus foresee if someone in particular, given dissimilar individual qualities moreover indication, will get coronary illness otherwise not. The proposed technique helps to minimize the noisy data of a patient. The fundamental intention of this venture is to look at the precision of assorted machine learning calculation. Data mining Algorithms such as KNN, Naïve Bayes, Decision Tree, Logistic Regression, and Random Forest are considered for the study. The results of the five classification methods are based on the accuracy and performance of the model. The resulting classification of effective data helps to find the treatment to the heart disease patients with better cost and facilitate the management. For the given data set the accuracy using Naïve Bayes is 78.49, KNN is 75.95, LR is 86.53, a Decision tree is 99,92 and Random forest is 100.4.

REFERENCES

[1] M. Kamber and P. J. Han, Data Mining Concepts, and Techniques, 3rded., 2012.

[2] S. Palaniappan, R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques," IJCSNS International Journal of Computer Science and Network Security, vol. 8, no. 8, August 2008.

[3] A. Khemphila, V. Boonjing "Comparing Performances of Logistic Regression, Decision trees, and Neural Networks for Classifying heart disease Patients," 2010 IEEE International Conference on Computer Information Systems and Industrial Management Systems (CISIM), pp.193-199, 2010.

[4] M. Sultana, A. Haider and M. S. Uddin, "Analysis of Data Mining Techniques for Heart Disease Prediction," 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2016.

[5] S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan, and T. Zhu, "CardiovascularRisk Prediction Method Based on CFS Subset Evaluation Random Forest Classification Framework," 2017 IEEE 2nd International Conference on Big Data Analysis, 2017.

[6] S. Pouriyeh, S. Vahid, G. Sannino, G. D. Pietro and H. Arabnia, J. Gutierrez, "A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease," 22nd IEEE symposium on Computers and Communication (ISCC 2017): Workshops- ICTS4eHealth, 2017.

[7] M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi, "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining with Decision Trees," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, May 2010.

[8] S.Fathima and N. Hundewale, "Comparison of Classification Techniques- Support Vector Machines and Naive Bayes to predict the Arboviral Disease-Dengue," *IEEE International Conference on Bioinformatics and Biomedicine Workshops*, 2011.

[9] P. C. Austin, J. V. Tu, J. E. Ho, D. Obligation, D. S. Lee, "Using Methods from Data Mining and Machine Learning Literature for Disease Classification and Prediction: A Case Study Examining Classification of Heart Failure Subtypes," Journal of Clinical Epidemiology 66 (2013) pp. 398-407, 2013.

[10] UCI Machine learning repository (online).
Available:http://archive.ics.edu/ml/datasets/heart+disease