JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2250 - 2259
https://publishoa.com
ISSN: 1309-3452

Heart Disease Prediction using Data Mining Techniques

Isha Salve[1#], Dilipkumar A. Borikar[2]

[1,2]Shri Ramdeobaba College of Engineering and Management, Nagpur, INDIA

[1]salveiy@rknec.edu

[2]borikarda@rknec.edu

ABSTRACT

Studies conducted by World Health Organization and Centres for Disease Control and Prevention have shown that heart diseases have appeared as the primary cause of deaths. Heart disease is responsible for deaths in all age groups and is frequent among all genders. A good answer to the issue of heart diseases is to be able to predict what a patient's health situation will be in the near future so the doctors can start treatment much sooner which will present good results. Data mining techniques are very efficient to less the diagnostic errors to better the patient's good health. By utilizing data mining techniques, it takes shorter time for the prediction of the disease with more accuracy. The purpose of this paper is to process a heart disease dataset and draw fine distinctions to predict the disease in the future using data mining techniques. We will be able to decide efficient algorithm from this paper which could be used to predict the disease the application of classification techniques like Random Forest, Naïve Bayes, KNN and Decision tree for the detection of heart disease has been used. Classification tree uses many factors including age, blood sugar and blood pressure; it can discern the probability of patients fallen in CD by using few diagnostic tests which save money and time.

Keywords- Big Data Analytics, Data Mining, Machine Learning, Component Analysis, Accuracy.

I. INTRODUCTION

The heart is the rigid working muscle in the body. The average heart beats 100,000 times day and night, to provide oxygen and nutrients throughout the whole body. Blood pumped by the heart also commutes unwanted products like carbon dioxide to the lungs so it can be removed from the body. Right heart functioning is important to protect life. Coronary artery disease (CAD), frequently known as heart disease is a state in which cholesterol, calcium, and many other fats assemble in the arteries that supply blood to the heart. This material thickens forming a plaque that blocks blood flow to the heart. When a coronary artery shrinks due to plaque build on or some other reasons, the heart muscle needs for oxygen and a person experiences a chest pain. Table I brings out the description of different types of heart diseases and the symptoms categorizing each of these, whereas Figure 1 gives general description of heart diseases.

TABLE I TYPES OF CVD, SYMPTOMS AND RISK FACTORS

Types of CVD	Description	Symptoms	Risk Factor
Coronary Heart Disease	Ischemic heart disease (IHD); most common type.	Heart attack, Angina at chronic condition.	High BP, high BC, unhealthy diet, diabetes, physical inactivity.
Stroke	Common forms of CVD and three categories: Ischemic stroke, Transient ischemic attack, Haemorrhagic stroke.	Brain damage, leading to a weakness often on one side of the body.	High BC, unhealthy diet, physical inactivity, diabetes, advancing in age.
Congenital heart disease	Malformations of heart pr central blood vessel at birth or during gestation.	Breathlessness or failure to attain normal growth and development.	Maternal alcohol and medicines use, maternal infection, poor maternal nutrition,
Other cardiovascular diseases	Tumours of the heart, vascular tumour of the brain, disorder of heart muscle, heart valve disease.		

There are two types of heart disease category for threat. They are as follows

1) *Controllable factors.*
2) *Uncontrollable factors.*

The Controllable features are smoking, drinking, weight, blood pressure and cholesterol which can be controlled by the people to lessen the heart diseases. The Uncontrollable elements are sex, age, history of the family which cannot be controlled by people to lessen the heart diseases. There are many types of the heart diseases in the world.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2250 - 2259
https://publishoa.com
ISSN: 1309-3452

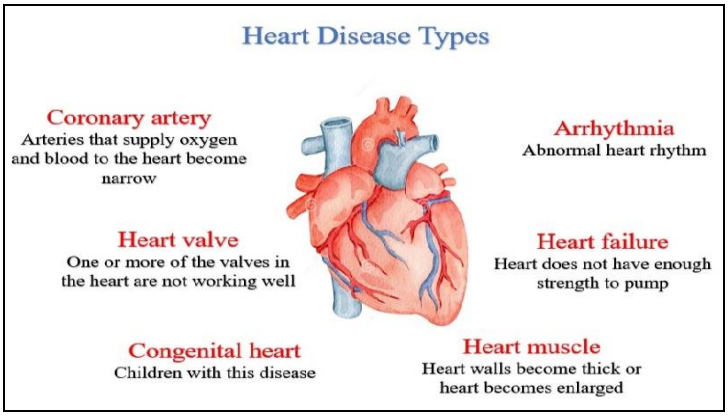

Fig.1. Types of Heart Diseases

A. Congenital heart disease

The Congenital heart disease is a kind of heart disease that has been arising in the heart since birth of people. Some of the instances of congenital heart disease are:

1) *Septal Defects:* The septal defects have the hole between the two chambers of the heart of people.
2) *Obstruction Defects*: In the obstruction defects there is partial or complete blockage of the flow of blood among the chambers of the heart.
3) *Cyanotic Heart Disease:* The cyanotic heart disease has the scarcity or lesser number of the oxygen around the human body.

B. Arrhythmia

Arrhythmia takes place because of the changes in the normal heartbeat of the humans. When the electrical impulses in the human body heart fails to synchronize the heartbeat rate of the heart then the arrhythmia comes up. The electrical impulses balance the heart to continue the heartbeat rate as fix in any condition. Changes in the heartbeat rate are so ordinary, and most of the people encounter it.

C. Coronary Artery Disease

The main function of the coronary arteries is to provide the nutrients to the muscle of the heart and circulation of oxygen through the blood. The Coronary arteries can be destroyed or diseased because of the cholesterol. The cholesterol causes the coronary arteries to provide the fewer amounts of oxygen and nutrients to the body.

D. Heart failure

The Heart failure is also known as the congestive heart failure, the main purpose for heart failure is there is no proper circulation of blood throughout the human body effectively.

E. Heart Muscle Disease (Cardiomyopathy)

The Heart Muscle disease is also called as the Cardiomyopathy. The Heart muscle disease arise when the walls of the human heart are become thicker or enlargement of the heart. This disease is the primary purpose to lessen supply of the blood to the whole human body and therefore it results into the failure of the heart.

F. Heart Valve Disease

There are four valves for the Human heart. These four valves are accountable for pumping of the blood to the whole human body and to supply that the heart keeps the forward flow of the blood in human heart.The symptoms of heart disease considerably depend upon which of the malaise felt by an individual. Some of the symptoms are not usually recognized by ordinary people. Nevertheless, some common symptoms consist of chest pain, breathlessness, and heart palpitations.

The chest pain is mostly common to many types of heart disease known as angina or angina pectoris, which takes place when a part of the heart does not get enough oxygen. Angina may be activated by stressful circumstances or exertion and generally lasts for 10 minutes. Heart attacks can also take place as an outcome of different types of heart disease. The signs of a heart attack are related to angina besides that they can happen during rest and can incline to be more serious. The symptoms of a cardiac disease occasionally be similar to indigestion, heartburn and a stomach ache can happen, as well as a heavy feeling in the chest. Other symptoms of a cardiac disease consist pain that proceed through the body, for instance from the chest to the arms, neck, back, abdomen, or jaw, dizzy sensations, nausea and vomiting [24].

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2250 - 2259
https://publishoa.com
ISSN: 1309-3452

For prediction of this disease, data mining plays a pivotal part in healthcare zone. Data Mining is a task of extracting the important decision making of information from an accumulative of past records for future examination or prediction. The data may be concealed and is not recognizable without the use of data mining. The classification is one data mining technique through which the future results or predictions can be made based on the historical data. The medical data mining made a feasible solution to combine the classification techniques and supply computerised training on the dataset that farther guides to inspecting the hidden patterns in the medical data sets which is used for the prediction of the patient's future health. Therefore, by using data mining it is likely to supply perception on a patient's history and is able to provide clinical help through the analysis. For clinical examination of the patients, these patterns are very much important. In other words, data mining algorithms is one of the important parts for recognizing the occurrence of cardiac disease before the occurrence. The algorithms will be trained and tested for predictions that decide the person's health of being affected by heart disease.

With the assist of data mining techniques, medical practitioners will be able to make future predictions. The data mining techniques and prediction models are in charge for making error free predictions if a patient is likely to get a heart disease in the future. The prediction models make prediction on patients based on earlier patterns in the past. If we can identify patients who are endangered to a heart disease in the future, then the doctor can take suitable action to help the patient. This can remarkably reduce the number of deaths of patients if not eradicate it completely. This big data prediction analysis will certainly come in easy to use to researchers and if the models are made better upon, it might even be used in real life.

In this research work, the data mining techniques are made use for making the predictions. A comparative analysis of the four data mining classification algorithms namely Naïve Bayes, K Nearest Neighbour, Decision Tree and Random Forest are used for making predictions. The survey is done at several levels of cross validation and several percentages of percentage split evaluation methods respectively. The predictions are made using the classification algorithms where the heart disease dataset is used for training.

II. LITERATURE SURVEY

Prediction of heart disease using data mining techniques has been an undergoing task for the past decades. Most of the research papers have also worked and implemented ample of data mining techniques for diagnosis of heart disease such as Decision Tree, Naive Bayes, Neural network, naturally defined groups, algorithm and support vector machine showing different stages of accuracies (M Kumari 2004; N Kausar et al 2010; Das, Kawaljeet Bhatla et al. 2007; K. Kaurand 2005; Chaitrali S. Dangare et al. 2012; M. Nikhil Kumar et al. 2019; Srinivas Rani et al. 2010; Dr P Alli et al. 2010; Sonali S. Jagtap 2017;) on different datasets of patients from around the world. One of the foundations on which the papers vary are the choosing of features on which the methods have been used. Many other authors have pinpointed different parameters and datasets for testing the accuracies. In certain, researchers have been investigating and enquiring the application of the Decision Tree technique in the diagnosis of heart disease with expected success.

M Kumari predicted using Decision Tree which is one of the data mining classifiers. The demonstration of this algorithm is analysed based on many factors like accuracy rate, sensitivity and error rate. Author tells the accuracy of the Decision Tree which was used is 79% [1]. Then by using biomedical mining algorithms heart disease is predicted, the author used supervised machine learning procedures in classification model. Moreover, they used decision tree with an error rate of 0.2775 and accuracy of 79.05% [3]. Decision tree is used in various application area like production, medicine manufacturing and monetary analysis [4]. N. Kausar et al. illustrated comparison between decision tree and Naive Bayes algorithm and used UCI data set for risk prediction and culminated that decision tree has a higher accuracy of 96.4% than Naïve Bayes [5]. Kawaljeet Bhatla et al. worked on three classifiers namely Decision Tree, Naive Bayes and Neural network for likelihood of Cardio-vascular disease. It shows that neural network has more accurateness and after neural network, Decision Tree mastered other data mining algorithms [7].

K. Kaurand Lalit performed research on many experiments with KNN, Naïve Bayes and Decision Tree. Decision tree is the most precise. Subsequently pre-processing the data of Naïve Bayes and Decision Tree have been augmented, they use Tanagra tool to classify their data [8]. The study was carried out via C4.5, Decision Tree for identification of heart stroke. The percentages of this algorithm are 75%, 65% and 75% for Myocardial Infarction disease [9]. Chaitrali S. Dangare et al. worked and studied on methods like Decision Tree and Naïve Bayesian which have been surpass by artificial neural networks. In this analysis, 15 attributes were used for the CVD analysis system. Obesity and smoking were added as two extra attributes for the treatment of CVD in enhancing an effective prediction system [14]. M. Nikhil Kumar et al. has used 8 algorithms which incorporates Decision Tree, J48 algorithm, Logistic model tree algorithm, Random Forest algorithm, Naïve Bayes, KNN, Support Vector Machine, and Nearest Neighbour to predict heart diseases. The accuracy attained for the prediction is at peak when using more features [28].

Researches have led to the fact that data mining technique plays a vital role in healthcare sector so that the environment is much healthier and all the objectives are contented on time [2]. Dr. P Alli et al. introduced a new system which depends on data mining techniques to diagnose cardiovascular disease. They gathered various patterns to evaluate the blockage in heart and related diseases. They also terminated that decision tree is secured to use since it has higher accuracy rate than any other techniques [6]. Sonali S. Jagtap worked on WEKA tool. The author apprises us about the

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2250 - 2259
https://publishoa.com
ISSN: 1309-3452

design and development of futuristic prediction techniques of the health system for heart diseases for these methods DT, Naïve Bayesian and k-mean [10]. Detailed points about different classification techniques are used for knowledge extraction by using data mining. Using this information mining strategies specifically Naïve Bayesian, Neural Network and DT were examined on therapeutic data utilizing these classifiers [11].

J. Vijayshree et al. introduced and proved DSS for the prognosis and analysis of heart illness, described by data mining and a hybrid intelligent methodology (genetic algorithm and fuzzy logic) [12]. Hlaudi Daniel Masethe et al. proposed a likelihood model for CVD illness utilizing mining methods Bayesian Net, J48, Naïve Bayesian, SIMPLE CART, and REP Tree algorithms, using medicinal dataset of patients. By applying these algorithms, it strongly suggested that information mining strategies are capable of predicting a resultant label for heart illness. Assessment of perplexity (confusion) matrix concludes that, a prognostic model of 89 cases with different causes of heart strokes were positive [15]. Amandeep Kaur et al. affirmed that Data mining is a vital phase of the KDD process that can be acclimated for disease management, prediction and diagnosis in healthcare sectors. This paper reviews different approaches in data mining that have been used for the prediction heart disease [26].

Research describes the research gaps that occurred in the betterment of heart illness analysis. It recommends a model that was systematically close to those gaps which can be used to discover a reliable performance as that attained in identifying CVD illness [13]. Aditi Gavhane et al. contemplated heart attack for prior diagnosis to lessen the number of deaths. Machine Learning plays a pivotal role in this paper. This prediction helps people to come out from the danger zone of their life. KNN algorithm and Random Forest algorithm was used to predict the heart attack beforehand [25]. Himanshu Sharma et al. proved that machine learning algorithms and deep learning approaches creates a new door of possibilities for accurate prediction of a heart attack. This paper delivers a lot of data about the latest methods in Machine learning and deep learning. A methodical comparison has been given to help new researchers working in this area. [27]

III. PROPOSED WORK

The architecture of the proposed system for implementation is shown in Figure 2. A medical expert or the user of the prediction system begins by feeding a dataset containing various factors contributing towards a heart disease.. The dataset is then pre-processed where it is cleaned against null and duplicate rows. Now the most updated dataset with no null values, if required, is selected to be performed algorithms on. The dataset is split into training and testing chunks since it's a global approach of any data analysis programming language like Python and R. Generally, the ratio is 1/4 for testing/training (which is 20% testing and 80% training).

As we have a label attached for every row which is essentially its column name, we don't really need to perform clustering. We move ahead by classification algorithms one by one. These algorithms are Random Forest, KNN, Naive Bayes, and Decision Tree. After we perform prediction using these algorithms, we select the best algorithm based on its accuracy, f1 score, recall and precision which is technically called as confusing matrix. For different datasets, a different algorithm may be more accurate than others. After an algorithm is selected, a prediction is being performed using it.

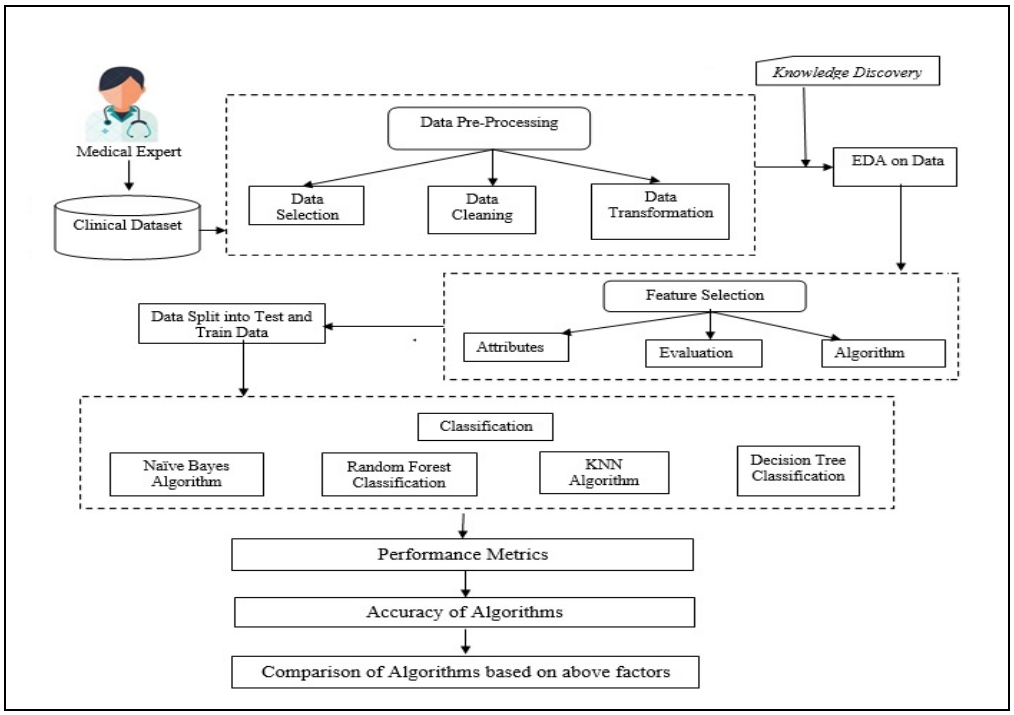

Fig. 2. Architecture of the system

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2250 - 2259
https://publishoa.com
ISSN: 1309-3452

A. Data Pre-Processing

This data set dates contains of four databases: Cleveland, Hungary, Switzerland, and Long Beach. It consists of 14 attributes. The fields refer to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease. The attributes composition of the heart disease dataset is shown in Table II. .

TABLE II DATASET WITH ATTRIBUTES AND SAMPLES

Attribute	**Descripton**	**Sample No**				
Age	The person's age in years.	63	37	41	56	57
Sex	1 = male; 0 = female	1	1	0	1	0
Cp	Chest pain type	3	2	1	1	0
Trestbps	Resting blood pressure.	145	130	130	120	120
Chol	Cholesterol measurement in mg/dl.	233	250	204	236	354
Fbs	Fasting blood sugar.	1	0	0	0	0
Restecg	Resting electrocardiographic results	0	1	0	1	1
Thalach	Person's maximum heart rate achieved	150	187	172	178	163
Exang	Exercise induced angina	0	0	0	0	1
Oldpeak	ST depression induced by exercise relative to rest.	2.3	3.5	1.4	0.8	0.6
Slope	slope of the peak exercise ST segment	0	0	2	2	2
Ca	The number of major vessels	0	0	2	2	2
Thal	A blood disorder called thalassemia,	1	2	2	2	2
Target	Presence or absence of heart disease, values (0,1)	1	1	1	1	1

The data mining task begins from the pre-processing step, in accordance with feature engineering, to choose different combination of features and classification modelling, to make models for prediction using data mining techniques. The feature selection and modelling were replicated for all the combination of the attributes. The loop iterates as a subset containing at least 3 attributes that are taken from the 13 attributes and the model is pertained to it. The performance of each model generated, formed on the chosen attributes and data mining technique during each iteration, is taped and the product of the results is shown after the complete process is completed.

The dataset was pre-processed after collection of data. There were no missing values in Cleveland dataset. The data pre-processing task starts from examining the dataset of 303 patients where 165 (54%) have a heart disease (target=1). The resulting dataset thus contains only 0 and 1 as the diagnosis value, where 0 being the absence and 1 being the presence of a heart disease.

B. Exploratory Data Analysis

In EDA, initially we need to allot tar variable to target columns. Then we avail unique function which will prints all the unique values in the given variable of the dataset.

target_temp = hearts.target.value_counts()

The above code computes the complete values in the target variable. Here we denoted "1" is the number of people suffering from heart disease and "0" is the number of people who are not suffering from heart disease. Thus, the number of people suffering from heart disease is "165" and the number of people not suffering from heart disease is "138". Plainly, from this, we can presume that this is a classification issue with target variables having values "0" and "1".

Here we discover the percentage of people that are suffering and the people who are not suffering from heart disease and they are 45.54% and 54.45% respectively. Using unique function prints all the unique values for the specific variable of the dataset. Here we have two features, where "1" is denotes for the number of males and "0" is for the number of females.

The Table III represents the meaning of every attribute in the dataset. The dataset contains several attributes which are to be evaluated to predict the heart disease result. The table shows each attribute, along with its features. The features are nothing but values associated with that attribute. For example, features of sex could be M or F. The other column is description, describing the features. So, M will be male and F will be female. The next columns are a bar plot showing the count of each feature for each attribute. And the last column is a comment for the whole row.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2250 - 2259
https://publishoa.com
ISSN: 1309-3452

TABLE III UNIQUE FEATURE OF ATTRIBUTES

Attributes	Features	Description	Comment
Cp	Type 0, 1, 2, 3	type "0" is for typical anginal type "1" is for atypical anginal type "2" is for non-anginal type "3" is for asymptotic	Comparing vulnerability of heart disease from different types of chest pains
Fbs	1, 0	1 - >120 mg/dl 0 - <120 mg/dl	fbs does not have much effect on heart problem.
Restecg	Type 0, 1, 2	0-definite left ventricular hypertrophy 1-normal 2-having ST-T wave abnormality	people having type "2" are much less likely to have heart problems on comparison to type "0" and type "1".
Exang	0, 1	"0" if for people not having exang "1" is for people having exang.	that people having type "1" are much less likely to have heart problems as compared to type "0".
Slope	0, 1, 2	0: down-sloping 1: flat 2: up-sloping	people having slope "2" have much more heart problems as compared to slope "0" and slope "1".
Ca	0, 1, 2, 3, 4	Number of major vessels	ca=4 have a very high number of heart problems as compared to the rest of the people.
Thal	0, 1, 2, 3	0-null, 1-fixed defect, 2-normal blood flow, 3-reversible defect.	type "0" has a high chance of having a heart problem.

C. Feature Selection

Among the 13 features used in heart disease prediction, only 'age' and 'sex' features are personal data of each patient. The remaining 11 attributes are all clinical features which are collected from different medical analysis. In this experiment, a collection of features was chosen to be used with 4 data mining classification techniques: Random Forest, Naïve Bayes, K Nearest Neighbor and Decision Tree. In this analysis, primarily, all possible collective features from the 13 attributes were taken and each combination was tested by implying the 4 data mining techniques. Next, the experiment was replicated to select all possible clusters of 4 features from the 14 attributes.

D. Spliting Data into Training and Testing

The data is split into training and testing. The approach is fit on the given data. 20% of the data is taken for testing while 80% is being used for training the model.

In this approach, the complete dataset is divided into 5 subsets and then processed 5-times. 4 subsets are used as training sets and the remaining 1 subset is used as a testing set. Let's say subsets range fromN1 to N5. In the first iteration, we take one of the subsets N as a testing dataframe and rest as training. The iterations follow until all of the subsets(N1-N5) are selected for testing. Lastly, the results are displayed by averaging each 5 iterations.

E. Classification

After determing the features, the models were made with the 4 classification techniques in data mining: Random Forest, Naïve Bayes, K Nearest Neighbor and Decision Tree. Cross validation technique was accustomed to proof the performance of the models.

1) Random Forest

Random forests (RF) [29] are amalgamation of tree predictors using decision tree such that each of the tree rely on the values of a random vector illustrated independently and with the similar classification for all trees in the forest. The general error of a forest of tree classifiers depends on the robustness of the single trees in the forest and the correlation between them. They are stronger with respect to noise. It is a supervised classification algorithm mostly used for the prediction and it is contemplated as superior because of its huge amount of trees in the forest which gives better accuracy than decision trees. Generally, the trees are trained individually and the predictions of the trees are merged by averaging. Random forest algorithm can utilize both for classification and the regression based on the problem domain.

The random forest algorithm is given below:

1) *Arbitrarily select k attributes from entire m attributes, where k << m.*
2) *Neighboured by the k attributes, calculate the node "d" using the best split point.*
3) *Divide the node into children nodes using the best split.*

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2250 - 2259
https://publishoa.com
ISSN: 1309-3452

4) *Keep repeating 1 to 3 steps upto l number of nodes has been accomplished.*
5) *Construct a forest by following steps 1 to 4 for n number times to make n number of trees*.

Intially, the k attributes are selected out of total m attributes. In the next level, in each individual tree arbitrarily choose k attributes in respect to get the root node by using the best split approach. Thereafter next level includes determining the children nodes using the same best split approach for the heart disease dataset. Likewise, the tree is created from the root node and till all the leaf nodes are created from the features. This arbitrarily formed tree also forms the random forest that is utilized for making heart disease prediction in patients.

2) Decision Tree

Decision Tree (DT) [30] is a uncomplicated and also easy to execute classifier. The chunk through attribute to approach in depth patient's health profile is only accquired in Decision Trees. Decision tree makes classification or regression models in the shape of a tree making it basic to rectify and handle. Decision trees can manage both sequential and numerical data. The algorithm implements by searching the information gain of the attributes and taking the attributes out for splitting the branches in threes. The information gain for the tree is recognized using the given Equation (1).

$$E(S) = -P(P)\log_2 P(P) - P(N)\log_2 P(N) \quad \text{Eqn. (1)}$$

The decision tree algorithm is given as follows:

1) *Recognize the information gain for the features in the dataset.*
2) *Classify the information gain for the heart disease dataset in decreasimg order.*
3) *After the classification of the information gain allot the good feature of the dataset at the root of the tree.*
4) *Then comtemplate the information gain using the same above formula.*
5) *Divide the nodes based on the elevated information gain value.*
6) *Repeat the process till all features are put as leaf nodes in all the branches of the tree*.

3) Naïve Bayes

Naïve Bayes (NB) is an analytical classifier which presumes no enthrall between features. Naive Bayes [31] is formed on Bayes rule and it presumes that features are not dependent on each other. The working of Naïve Bayes classifier is given below:

1) Training Step: By expecting predictors to be tentatively independent given for a class, the method the constaints of a probability distribution which is known as the prior probability from the training data.
2) Prediction Step: For undetermined test data, the method contemplates the posterior probability of the dataset which is belonging to each class. The method lastly distributes the test data formed upon the highest posterior probability from the set.

4) KNNK-Nearest Neighbour

K-Nearest Neighbour (KNN) is a basic algorithm, which stocks all cases and distributes new cases formed on similarity measure. Right from training point to sample point distance is calculated, and the point with lowest distance is called nearest neighbor. Structure based NN techniques are formed on shapes of data like orthogonal structure tree (OST), ball tree, k-d tree, axis tree, nearest future line and central line. Nearest neighbor classification is utilized primarily when all the features are continuous.

1) *Observe the K training instances which are nearest to unknown instance.*
2) *Take out the most frquently exisiting classification for these K instances.*

F. Performance Matrix

The metrics utilized for the research work is described in this section.

1) Precision

Precision is the portion of notable instances between the retrieved instances.

$$Precision = TP / (TP+FP) \quad \text{Eqn. (2)}$$

2) Recall

Recall is the minute segment of suitable instances that have been recovered over the total quantity of germane instances.

$$Recall = TP / (TP + FN) \quad \text{Eqn. (3)}$$

3) F-Score

The f-score (or f-measure) is contemplated based on the two times the precision times recall classified by the sum of precision and recall.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2250 - 2259
https://publishoa.com
ISSN: 1309-3452

The performance of the data mining classification models was computed using three performance measures: accuracy, precision and recall. Review of data mining techniques pivot on the best performing models that can generate great accuracy in heart disease prediction because accuracy and precision recall f-score are the most instinctive evaluation metrics on performance. For each classifier, performances have been scaled differently.

IV. RESULTS

This section brings out the performance of heart disease prediction system. The analysis of the results concludes that the proposed method gives the accurate diagnosis of heart disease. The accuracy of different algorithms applied is depicted in Figure 3. The comparisons for other performance measures are detailed in Table IV.

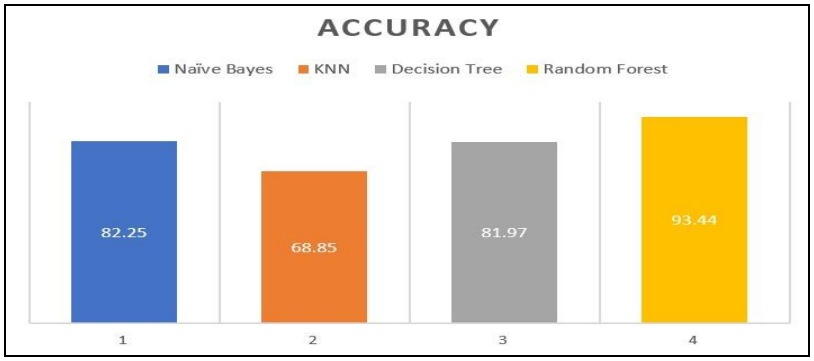

Fig. 3. Accuracy of Algorithms

TABLE IV COMPARISON OF PERFORMANCES OF ALGORITHMS

Parameter	Naïve Bayes	KNN	Decision Tree	Random Forest
Accuracy (%)	85.25	68.85	81.97	93.44
Precision	0.837	0.741	0.87	0.8611
Recall	0.911	0.676	0.794	0.822
F-Score	0.8857	0.707	0.83	0.855
False Negative Rate	8.823	32.352	20.588	11.764

Amidst all classification techniques [Table III], accuracy of Random Forest was best with 93.44% and Naïve Bayes approach with an accuracy 85.25% was second best. Here KNN is the worst algorithm with accuracy of 68.85%. Since KNN is a distance-based algorithm, the cost of calculating distance between a new point and each existing point is very high which in turns degrades the performance of the algorithm. Hence it impacts the performance. KNN is also sensitive to outliers and missing values.

V. CONCLUSION

This cardiac disease prediction model with an accuracy of 93.44% will be a help to many people mainly to medical professionals to scale different context. They will have a better understanding of a person's health and they can simply understand age related health factors and thus they can alert a patient in advance. Patients on the contrary can also seek a doctor in advance and go through health examination and therefore can avert the occurrence of any cardiac disease. Hence, this model helps to build belief and grows a sense of safety among people.

REFERENCES

[1] National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents (2004). The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents; Paediatrics 114:555-76.

[2] American Heart Association; Cardiovascular Disease (CVDs) fact sheet, reviewed January 2018.

[3] Dinarević S, Mulaosmanović V (2005) Primary prevention of Hypertension in Sarajevo Children: Role of Adiposity. 29th UMEMPS Congress Union of Middle Eastern and Mediterranean Paediatric Societies, pp. 154-156.

[4] Dinarević S, Mesihović H, Simeunović S, Zulić I (1994) Dyslipoproteinaemia in Children with Heart Disease. Intercontinental Cardiol 3:126-9.

[5] Kapur G, Ahmed M, Pan C, Mitsnefes M, Chiang M, et al. (2010) Secondary hypertension in overweight and stage 1 hypertensive children: A Midwest Paediatric Nephrology Consortium report. J Clin Hypertens12:34-39.

[6] Dinarević S, Hasanbegović S (2010) Problem of obesity in children and youth in Canton Sarajevo. Pediatr Res 68:1091.

[7] The challenge of obesity in the WHO European Region and the strategies for response. Copenhagen: WHO Regional Office for Europe.

[8] Daniels SR, Arnett DK, Eckel RH,Gidding SS, Hayman LL, et al. (2005) Overweight in children and adolescents: Pathophysiology, consequences, prevention and treatment. Circulation 111:1999-2012.

[9] Berenson GS, Wittingly WA, Tracy RE,Newman WP, Srinivasan SR, et al. (1992) Atherosclerosis of the aorta and coronary arteries and cardiovascular risk factors in persons aged 6 to 30 years and studied at necropsy (The Bogalusa Heart Study). Am J Cardiol 70:851-858.

[10] McNiece KL, Gupta-Malhotra M, Samuels J, Bell C, Garcia K, et al. (2007) National High Blood Pressure Education Program Working Group: Left ventricular hypertrophy in hypertensive adolescents: Analysis of risk by 2004 National High Blood Pressure Education Program Working Group staging criteria. Hypertension 50:392-5.

[11] Sonali S. Jagtap, "Prediction and Analysis of Heart Disease", International Journal of Innovative Research in Computer and Communication engineering, February 2017.

[12] Monika Gandhi and Dr. Shailendra Narayan Singh, "Prediction in Heart Disease Using Techniques of Data Mining", International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015).

[13] J. Vijayshree and N.Ch.Sriman Narayana lyengar, "Heart Disease Prediction System using data mining and Hybrid Intelligent Techniques: A Review", IJBSBT-2016.

[14] Mai Shouman, et al., "Using Data mining techniques in heart disease diagnosis and treatment", IEEE, 2012.

[15] Chaitrali S. Dangare, Sulabha S. Apte, "Improved study of Heart Disease Prediction System using Data Mining Classification Techniques"; International Journal of Computer Applications (0975-888) Volume 47- No.10, June 2012.

[16] Hlaudi Daniel Masethe, Mosima Anna Masethe, M. Akhil B. L. Deekshatulu and P. Chandra "Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science, 2014.

[17] M.Akhil jabbar, B.L Deekshatulua Priti Chandra b, "Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm", Procedia Technology. vol. 10 pp.85-94 Conference Paper, IEEE, 2013.

[18] Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Diseases Prediction ,4th International Conference on Computing Communication And Automation(ICCCA), IEEE, 2018

[19] K. Hafeeza, R Mohanraj, "Classification of Multi Disease Diagnosing and Treatment Analysis Based on Hybrid Mining Technique", March 2014, Volume 06, Issue No. 03, Pages 108-116.

[20] Nidhi Bhatla, Kiran Jyoti, "An Analysis of heart disease prediction using different data mining techniques", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, Vol. 1 Issue 8, October - 2012

[21] Manlik Kwong, Heather L. Gardner, Neil, Virginia, "Optimization of Electronic Medical Records for Data Mining Using a Common Data Model", Research Article, Volume 37, Science Direct, December 2019

[22] K. Gomathi Kamaraj, D. Shanmuga Priyaa, "Multi Disease Prediction using Data Mining Techniques", International Conference in Trends in Information, Management, Engineering and Science [ICTIMES-2018]

[23] Robert Nisbet, Gary Miner, Jhon Elder, "Handbook of Statistical Analysis and Data Mining Applications 1st Edition", May 2009, Page Count 864.

[24] Carlos Ordonez, "Improving Heart Disease Prediction using Constrained Association Rules", University of Tokyo, 2004.

[25] Aditi Gavhane, Gouthami Kokkula, Isha Panday and Kailash Devadkar,"Prediction of Heart Disease using Machine Learning", Proceedings of the 2nd International conference on Electronics Communication and Aerospace Technology(ICECA), IEEE, 2018.

[26]]"Heart Diseases Prediction using Data Mining Techniques: A survey" Amandeep Kaur and Jyoti Arora International Journal of Advanced Research in Computer Science IJARCS, IEEE, 2019.

[27] Himanshu Sharma and M A Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms: A Survey" International Journal on Recent and Innovation Trends in Computing and Communication vol. 5 no. 8, IEEE, 2019.

[28] M. Nikhil Kumar K. V. S. Koushik and K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools", International Journal of Scientific Research in Computer Science Engineering and Information Technology IJSRCSEIT, IEEE, 2019

[29] A. Liaw and Matthew Wiener, “Classification and Regression by Random Forest”, R News, Vol. 2, No. 3, pp. 18-22, 2002

[30] J. Ross Quinlan, “Induction of Decision Trees”, Machine Learning, Vol. 1, No. 1, pp. 81-106, 1986.

[31] K. Ming Leung, “Naive Bayesian Classifier”, Master Thesis, Department of Computer Science and Engineering, Polytechnic University, 2007.