Machine Learning based Heart Disease Prediction System

[1] *Sophia S*, [2]*Priyadharshini R*,[3] *Princy S*, [4]*Sophia B*, [5]*Vidhusha M*
Department of CSE, Dr. T. Thimmaiah Institute of Technology

Abstract: Heart attack disease is one of the leading causes of the death worldwide. In today's common modern life, deaths due to the heart disease had become one of major issues, that roughly one person lost his or her life per minute due to heart illness. Predicting the occurrence of disease at early stages is a major challenge nowadays. Machine learning when implemented in health care is capable of early and accurate detection of disease. In this work, the arising situations of heart disease illness are calculated. Datasets used have attributes of medical parameters. The datasets are been processed in python using ML Algorithm i.e., Decision tree classifier Algorithm. This technique uses the past old patient records for getting prediction of new one at early stages preventing the loss of lives. In this work, reliable heart disease prediction system is implemented using strong Machine Learning algorithm which is the Random Forest algorithm. Which read patient record data set in the form of CSV file. After accessing dataset, the operation is performed and effective heart attack level is produced. Advantages of proposed system are High performance and accuracy rate and it is very flexible and high rates of success are achieved.

Keywords- ML: Machine Learning, Vector Quantization, Questionnaire, CSV: Comma- Separated Values, Random Forest algorithm, Decision Trees.

I. INTRODUCTION

Heart disease effects the functioning of the heart. World Health Organization had made a survey and made a conclusion that 10 million people are affected with heart disease and lost their lives. The problem that the Healthcare industry faces in today's life is early prediction of disease after a person is affected. Records or data of medical history is very large and the data in real world might be incomplete and inconsistent. In past predicting the disease effectively and treatment to patients might not be possible for every patient at early stages under these circumstances [2].

The datasets are been processed in python using ML Algorithm i.e., Random Forest Algorithm. This technique uses the past old patient records for getting prediction of new one at early stages preventing the loss of lives.

Many scientists tried to build a model which is capable of predicting the heart disease in the early stage, but they are not able to build a perfect model. Every proposed system has disadvantages in its own way. In the existing system, Shen et al. had initially, proposed a system which is based on self-applied questionnaire. In this system the user needs to enter all the symptoms which he is suffering from, based on that the result is predicted. This study is based on the analysis data collected in SAQ.

Chen et al. came up with an idea to predict heart disease. He used the technique of Vector Quantization which is one of the artificial intelligence techniques for classification and prediction purpose. Training of neural networks is performed using back propagation to evaluate the prediction system. In the testing phase approximately 80% accuracy is achieved on testing set. Practical use of data collected from previous records is time consuming. Low accuracy rate.

So to overcome this we are implementing Random forest algorithm in order to achieve accurate results in less time. Machine learning is given a major priority in modern life in many applications and in healthcare sector. Prediction is one of area where machine learning plays a vital role, our topic is to predict heart disease by processing patient"s dataset and a data of patients i.e., user of whom we need to predict the chances of occurrence of a heart disease.

II. RELATED STUDY

In the existing systems, heart disease prediction system is developed using various algorithms but has some disadvantages.
Using Machine learning techniques [5]
Every machine learning algorithm will follow these techniques,

A. Preprocessing The database or dataset which we use contains NaN values i.e., „not a number" values. The program which we use cannot process these non- numerical values, so it is mandatory to covert these values to numerical values. The approach followed is, the NaN values are replaced by the mean of the column [1].

B. Splitting The data in the database is divided into 2 types i.e., training and testing sets. Training set is of 80% data and testing set is of 20% data.

C. Classification The selected data from the database under training set is trained with different algorithms like KNN, Adaptive boost, Decision Tree and K-mean, all these are machine learning algorithms [1].

1. Decision Tree There are various kinds if decision trees. The main difference among them is, they use to first-rate the class feature. The characteristic that reduces entropy in an entropy system and makes use of information gain is known as tree root.

Firstly, information gain of all attributes in the dataset is estimated to select a tree root. Then the attribute that makes use of information gain will be specified [6].

2. KNN This method is one of the simplest and efficient methods of classification. At the time of quality check, some reliable constant controls of probability densities are difficult to understand because the user is not aware of them. So this KNN classification method is implemented to calculate such type of calculations. With the help of training datasets the location of K- nearest neighbor is predicted. Euclidean distance is used to find how close the training dataset is from target. Find the k-nearest neighbors and assign them to group of rows which is examined. Repeat the step for the rows outstanding in the target set. In this application the highest value of K can be selected, after that the software application automatically builds a similar parallel model on the values of K up to the maximum value defined. KNN algorithm with support of WEKA tool concludes that training dataset, input and output variables must derive in. The best value of K is used to build parallel models on all the values of K up to max known value.

3. K-mean clustering It is an unsupervised learning algorithm, in which the dataset contains unlabeled data and also class labels are not known. The algorithm main aim is to make a group of present data. The algorithm recursively allocates K groups. These groups are formed on their similarities between them. Each group consists of centroid K. There as K groups Benn formed, when the new value is given, k-means algorithm allocates it to some specific group based on its similarities. As centroid is key to the group, using centroid the new variable is assigned to a specific group [6].

4. Adaptive boost (Adaboost)It is one of the efficient techniques. It is used in binary classification problem which increases the execution of decision trees. Since it is mainly used for classification than regression it is also been discussed to as discrete adaptive boost. Using adaptive boost, it is possible to increase the presentation of machine learning algorithms. The models slightly increase the accuracy rate on a given classification problem. The algorithm that is commonly used with adaptive boost is decision tree algorithm but with only one level. The decision trees are small and contain single decision for the classification, are named as decision stumps. [11] Classification algorithms (Naive bayes, neural networks, support-vector machines model)

- It takes input as training and testing data sets to the application.
- It uses data mining environment in order to prediction of heart disease.
- It takes Cardio vascular data (CVD) sets from various other Sources in prediction of heart disease.
- Evaluating the datasets 5 classic fiction Algorithm are been used.
- The CVD datasets have two defining attributes True/False for the prediction of heart disease.
- These all CVD data sets are been compressed as ARFF file with their data labels and utilized as WEKA data mining tool for prediction.
- When the user gives input to the application, it will be taken and maintained by SQL server at the background. As the input been generated as EXCEL file is converted to ARFF file using WEKA tool as the application accepts only ARFF file.
- When the new instance been given to prediction system, it classifies the new instance and generates its class label [7].

III. ENHANCED PREDICTION METHOD OF HEART DISEASE

Our aim is to build an application of heart disease prediction system using robust Machine Learning algorithm which is Random Forest algorithm. A CSV file is given as input. After the successful completion of operation, the result is predicted and displayed.

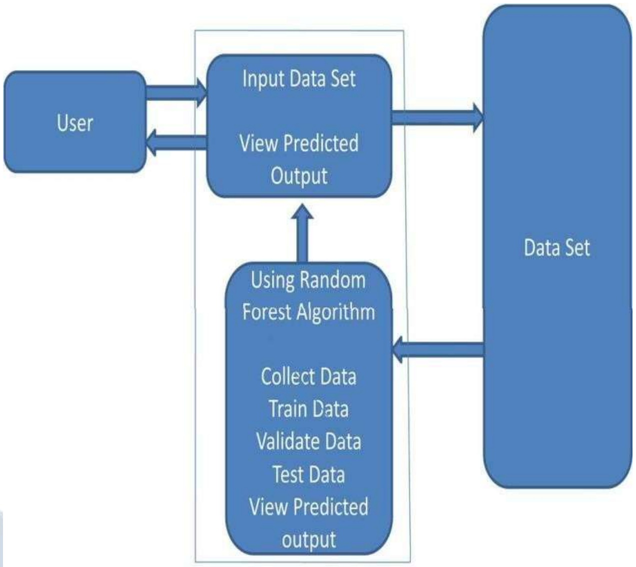

Fig-1: Architecture diagram.

The working principle of the system is shown in fig-1. The user enters the input which is compared with the data present in the existing data set by using the Random Forest Algorithm. [10] Random Forest algorithm is an efficient ML algorithm that comes under supervised learning technique. It is be used for both Regression and Classification problems. To solve a complex problem, it uses a process of combining multiple classifiers, to increase the accuracy and performance of the model. "Random Forest is known as classifier that contains more number of decision trees on different subsets of the given dataset and considers the average to improve the predictive accuracy of that dataset." Instead of depending on single decision tree, the RFA algorithm takes the result from each decision tree and it predicts the final output as shown in fig-2.

The accuracy of the result depends on the number of trees, more the trees higher is the accuracy rate. And also avoids the problem of over fitting. The Working process of the algorithm can be explained in the following steps:

Step-1: First step is to choose the K data points from the selected training set.
Step-2: Build as many as decision trees associated with the selected data points as in fig-3
Step-3: Select the number of decision trees you wish to build i.e., N Fig-2
Step-4: Repeat steps 1 and 2.
Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Training phase is of 70% and testing phase is of 30%. The advantages of proposed model are High performance and accuracy rate. It is very flexible and high rates of success are achieved. The data i.e., attributes in the data set are categorized in the following way while building the decision tree:

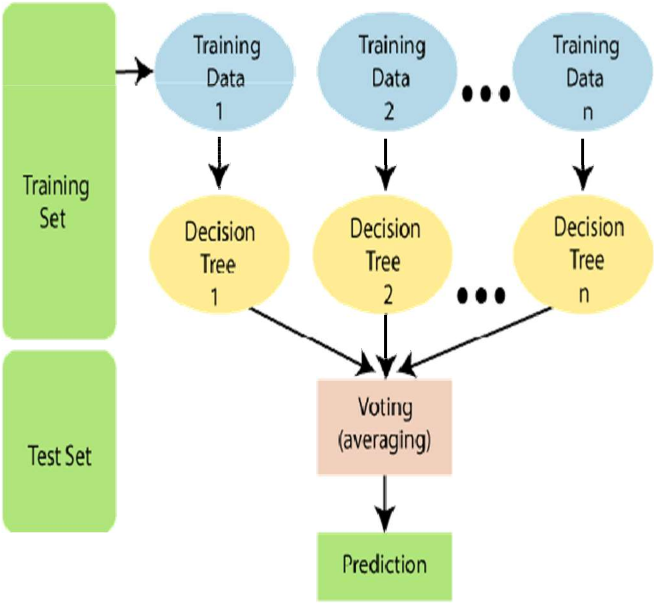

Fig-2: Procedure of random forest algorithm.

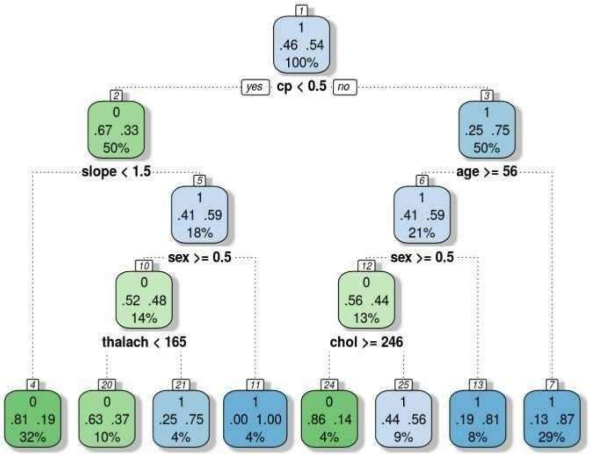

Fig-3 Building decision tree.

IV. RESULT ANALYSIS

The main aim of our project is to know whether a person is having any heart disease or not. And give suggestions how to proceed further. Using Random Forest algorithm, it is possible to achieve high accuracy rate. The data set which we have used is as follows (sample)

Table-1 Sample data set

Age	63	37	41	56
Cp	3	2	1	1
Trestbps	145	130	130	120
Chol	233	250	204	236
Fbs	1	0	0	0
Thalach	150	187	172	178
Exang	0	0	0	0
Old Peak	2.3	3.5	1.4	0.8
Thal	1	2	2	2
Target	1	1	1	1

The above mentioned attributes in table-1 are enough to predict if a person is affected with heart disease or not. Each attribute in the data set result of functionality of heart. For example, Cp- The type of chest pain categorized into 4 values. (1. Typical angina 2.Atypical angina 3.Non-angical pain 4. Asymptomatic) The data set represents attributes as in fig-4 are shown below and also listed in table-1.

- Trestbps- Level of blood pressure at resting mode.
- Chol-Serum cholesterol in mg/dl.
- Fbs- Blood sugar levels on fasting (if>120mg/dl represented as 1 otherwise 0)
- Restingecg- Results of electrocardiogram while at rest.
- Exang- Angina induced by exercise (0-No, 1-Yes)
- Old peak- Exercise induced ST depression in comparison with state of rest.

Table-2: Sample data set with results.

AGE	63	37	17	56
CP	1	1	0	1
TRESTBPS	3	2	0	0
CHOL	233	250	170	200
FBS	1	0	1	1
RESTECG	0	1	0	1
THALCH	150	187	77	79
EXANG	0	0	1	1
OLDPEAK	2.3	3.5	0	1
SLOPE	0	0	2	2
THAL	1	2	3	3
TARGET	1	1	1	1
HEART DISEASE	YES	YES	NO	NO

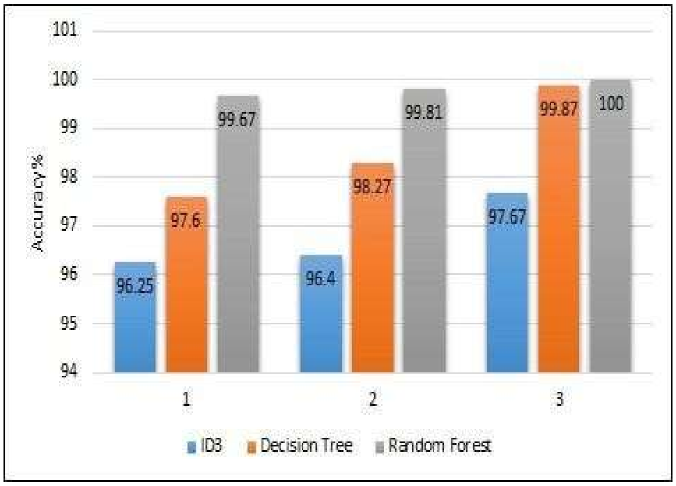

Fig-5: Graph comparing accuracy of various algorithms

From fig-5 we can say that the application when implemented using random forest algorithm has more accuracy rate when compared to other algorithms.

V. CONCLUSION

Random Forest algorithm is an efficient algorithm which is an ensemble learning method for regression and classification techniques. The algorithm constructs N of Decision trees and outputs the class that is the average of all decision trees output. So accuracy of prediction at early stages is achieved effectively. Processing of healthcare data i.e., data related to heart will help in early detection of heart disease or abnormal condition of heart which results in saving of long term deaths. Heart disease prediction is a major challenge in the present modern life. With this application if the patient/user is away from reach of doctor, he/she can make use of the application in prediction of disease just by entering the report values. And can proceed further whether to consult a doctor or not.

VI. FUTURE SCOPE

In future this application can extended by updating some features like, if the user is effected with heart disease all his family members will be notified with a message in early passed to the nearest hospital. Another feature is there should be online doctor consultation with the nearest doctor available.

REFERENCES

[1] *Kaan Uyar and Ahmet İlhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks" in B.V ICTASC,Elsevier,pp*

[2] *Ashish Chhabbi, Lakhan Ahuja, Sahil Ahir and Y.K. Sharma, "Heart Disease Prediction Using Data Mining Techniques", © IJRAT Special Issue National Conference "NCPC-2016", pp. 104-106, 19 March 2016.*

[3] *Berry JD, Lloyd-Jones DM, Garside DB, et al. Framingham risk score and prediction of coronary heart disease death in young men. Am Heart J. 2007;154(1):80–6.*

[4] *Theresa Princy and R, J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques", © IEEE ICCPCT, 2016.*

[5] *Kaur h Beant and Williamjeet Singh, "Review on Heart passed to the nearest hospital. Another feature is there should be online doctor* Disease Prediction System using Data Mining Techniques", © *IJRITCC*, vol. 2, no. 10, pp. 3003-08, 2014.

[6]Kirmani, M.M., Ansarullah, S.I.: Prediction of heart disease using decision tree a data mining technique. IJCSN Int. J.Comput. Sci. Netw. **5**(6), 885–892 (2016)

[7] *Salam Ismaeel, Ali Miri et al., "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis", IEEE Canada International Humanitarian Technology*

[8] *Tahira Mahboob, Rida Irfan and Bazelah Ghaffar et al."Evaluating ensemble prediction of coronary heart disease using receiver operating characteristics" ©2017 IEEE*

[9] *Ammar Asjad Raja, Irfan-ul-Haq , Madiha Guftar Tamim Ahmed Khan "Intelligence syncope Disease Prediction Framework using DM- techniques" FTC 2016 –Future Technologies Conference 2016.*

[10] *M.A. Jabbar, B.L.Deekshatulu, and Priti Chandra, " Intelligent heart disease prediction system using random forest and evolutionary approach", Journal of Network and Innovative Computing, Vol. 4, pp.174-184, 2016.*

[11] *N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Machine Learning,(1997).*

[12] *Ayon Dey, Jyoti Singh, N. Singh "Analysis of supervised machine learning algorithms for heart disease prediction".*