International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 1, pp: 907-911 www.ijaem.net ISSN: 2395-5252

Heart Disease Prediction using Machine Learning

Mrs.S.Jansi rani, AP (Sr.Gr)/IT[1], Nithyasree C[2], Ramyasri G[3], Swethapriya D[4]
[1]*Dept. of Information Technology Sri Ramakrishna Engineering College Coimbatore, India*
[2]*Dept. of Information Technology Sri Ramakrishna Engineering College Coimbatore, India*
[3]*Dept. of Information Technology Sri Ramakrishna Engineering College Coimbatore, India*
[4]*Dept. of Information Technology Sri Ramakrishna Engineering College Coimbatore, India*

Date of Submission: 25-06-2020 Date of Acceptance: 13-07-2020

ABSTRACT: Heart diseases is one of the most important factors of mortality in the present world. Clinical data analysis is a very vast area and the prediction of cardiovascular disease becomes a critical challenge in it. The large quantity of data produced in the healthcare industry should be assisted by Machine learning in terms of making decisions and predictions.Machine Learning is also being used in Internet of Things (IoT) for several developments. There are many studies which provide only a glimpse in predicting heart disease using Machine Learning techniques. In this paper, we propose a method to predict cardiovascular disease by applying Machine Learning techniques which results in improving the accuracy. Using different combination of features and several classification techniques, the prediction model is designed. On experiment with the prediction model, it produces an enhanced performance level with an accuracy of 88.7%.

KEYWORDS: ***Machine Learning; heart disease prediction; feature selection; prediction model; classification algorithms; cardiovascular disease (CVD)***

I. INTRODUCTION

With several contributory risk factors such as diabetes, blood pressure, cholesterol, abnormal pulse rate - identification of heart disease becomes difficult. To find out the severity of heart disease among humans, various techniques in data mining and neural networks have been employed. Among various methods, K- Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic Algorithm (GA), and the Naïve Bayes (NB) have been used to find the severity of the heart disease. Since the nature of heart disease is complex, the complications must be handled carefully. If not doing so, it may affect the heart or it can also lead to premature death. In discovering various sorts of metabolic syndromes, the perspective of medical science and data mining is used. In predicting heart disease and its data investigation, data mining with classification plays a major role.

Decision trees are also being used in predicting the accuracy of heart disease. For knowledge abstraction in prediction of heart disease, various known methods of data mining are used. Numerous readings from different models have been carried out to generate a prediction model. Diagnosis of heart disease is done with genetic algorithm. Cleveland dataset from UCI machine learning repository has been used for experimental validation.

One of the powerful evolutionary algorithm, Particle Swarm Optimization (PSO) is introduced and some rules are generated for heart disease prediction.

The proposed system expects to have a better accuracy rate. Initially the data in the dataset is cleaned and pre- processed followed by feature selection and reduction. The feature selection plays a prominent role in the prediction of heart disease. In classification modelling, the clustering of datasets is done on the basis of the variables and criteria of Decision Tree (DT). The results of the classification method have proved a higher degree of accuracy and performance in the prediction of heart disease.

II. RELATEDWORK

Effective Heart disease prediction using Hybrid Machine Learning aims to enhance the performance level with an improved accuracy rate. The hybrid approaches used in previous projects has limited accuracy rate, and size indatasets.

BerinaAlic et al. [11] presented a study that was designed to perform a review of Artificial Neural Network and Bayesian network and their application in classification of diabetes and

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 1, pp: 907-911 www.ijaem.net ISSN: 2395-5252

cardiovascular diseases. The purpose of the study is to show the comparison of these machine learning techniques and to discover the best option for achieving the highest output accuracy of classification. In the comparison, different vales for the network accuracy have been achieved. In ANN, the accuracy of classification lies between 72.2 and 99%. On the other hand, in BN the accuracy of classification lies between 71 and 99.5%

Theresa Princy R. et al. [3] proposed a paper that gives us the survey about different classification techniques used for predicting the human heart disease and its risk level based on some attributes, for each person. Classification techniques such Naïve Bayes- used to predict heart disease through probability, KNN- to find values of the factors of heart disease , Decision Tree Algorithm- to provide the classified report for the heart disease, Neural Network- to provide the minimized error of the prediction of heart disease, etc., is used to find the risk level of patient. It is found that, by using more number of attributes- the accuracy of the risk level is high . The risk level of each class was identified with the help of ID3 algorithm.

Mahmood and Kuppa et al. [1] proposed a new pruning method with the aim of improving classification accuracy of heart diseases and reducing tree size. A combination of pre-pruning and post- pruning was used for pruning C4.5 decision tree classifier. With the dataset available in online, the new decision tree is compared with the benchmark algorithms. Resulting in the accuracy of 76.51%, the new prediction method has reduced the tree size.

Bashir et al. [4] proposed the use of majority vote from three different classifiers namely Naïve Bayes, Decision Tree and Support Vector Machine for heart disease diagnosis. The Cleveland database from the UCI machine learning repository was used for the experiment. Each classifier was individually trained using training set. Obtained decisions from three classifiers and combined them based on majority voting scheme. By the proposed framework with two output classes, accuracy of 81.82% is resulted.

Chaurasiaet al. [6] have used the commonest types of decision tree algorithms for the prediction of heart diseases. CARD, ID3 and DT were applied with the same dataset available at Cleveland database in UCI repository, and evaluated using 10-fold cross validation method. CARD decision tree has presented the highest the highest classification accuracy with 83.49%, followed by DT with 82.50% and finally 72.93% for ID3.

Shouman et al. [8] were focusing on the improvement of decision tree accuracy for diagnosis of heart disease. K-means clustering was integrated into the decision tree in order to enhance the diagnosis of heart disease. The Cleveland database in UCI repository has been utilized. The highest accuracy obtained was 83.9% by applying the inlier method with tow clusters.

III. HEART DISEASE PREDICTION ALGORITHMSUSING MACHINE LEARNING

The data is collected from UCI Machine Learning repository and Cleveland dataset is selected for the experiment. It is divided into training and testing data Training set is the one on which is trained and fits in our model basically to the parameters but test data is used only to assess the performance of model. Training data is used to fit in the model. Test data is used to validate the model. Initially the dataset undergoes preprocessing. The preprocessed data is divided into training and testing data. Then the machine learning models are applied on the training and testing data to predict the heart disease. Then the accuracy of each model is evaluated and the model which shows the highest accuracy among all is predicted. The project is implemented using Anaconda Spyder.
The system undergoes three main processes namely,1.Data Pre-Processing, 2. Feature selection and reduction and 3.Classification Modelling, 4. Performance measures.

A. DataPre-Processing

Data is pre-processed after collection of various records. The dataset contains a total of 303 patient records, where 6 records are with some missing values. Those 6 records have been removed from the dataset and the remaining 297 patient records are used in pre-processing. The multi- class variable and binary classification are introduced for the attributes of the given dataset and is used to check the presence or absence of disease

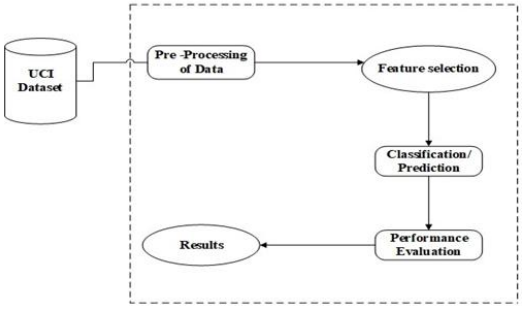

Figure 1. System Architecture for Heart Disease prediction

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 1, pp: 907-911 www.ijaem.net ISSN: 2395-5252

In the instance of the patient having a heart disease, the value is set to 1, else the value is set to 0 indicating the absence of heart disease in the patient. The preprocessing of data is carried out by converting medical records into diagnosis values. The results of data pre-processing for 297 patient records indicate the 137 records show the value of 1 establishing the presence of heart disease while the remaining 160 reflected the value of indicating the absence of heart disease.

B. Feature selection and reduction

Among the 13 attributes of the dataset, two attributes pertaining to age and sex are used to identify the personal information of the patient. Remaining eleven attributes are considered important as they contain vital clinical records. To learn the severity of heart disease, clinical records are used. In the experiment, several (ML) techniques are used namely, Naïve Bayes, SVM, Random Forest, Decision Tree, Neural Network (Keras), K-Nearest Neighbor algorithm. With 13 attributes, the experiment was repeated with all the ML techniques.

C. Classification Modelling

The algorithms used for building predictive models in the proposed system are:

1. Support Vector Machine(SVM)
2. DecisionTree
3. Random forest
4. Naïve bayes

1) Support Vector Machine(SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm used for problems subjected to both classification and regression. However, it is highly used in classification problems. With the value of each feature being the value of a particular coordinate, we plot each data item as a point in n-dimensional space . Then, the classification is performed by finding the hyper-plane that differentiate the two classes very well. Support Vector Machine is the frontier which best segregates the two classes.

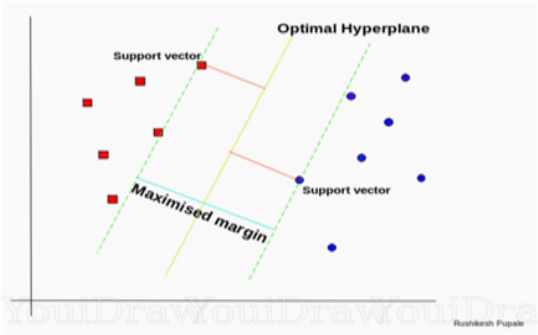

Figure 2. Overview of SVM

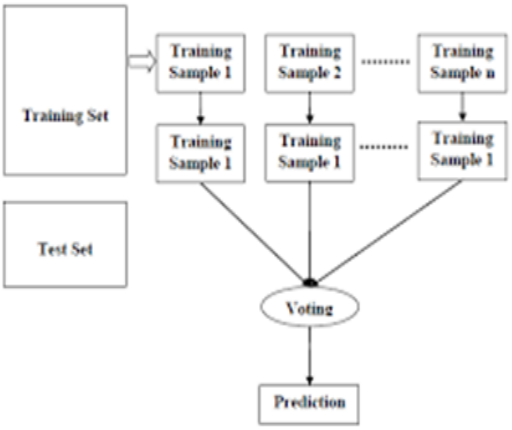

Figure 2. Overview of Random Forest

2) Decision Tree

Decision tree is used for classification as well as for regression problems. It is one of the most popular algorithms used in machine learning.When training a dataset to classify a variable, the idea of the decision is to divide the data into smaller datasets supported a particular feature value until the target variables all fall into one category. Based on the maximum information gain, the computer splits the dataset. By using decision tree, we can create a training model which is used to predict the class or value of target variables by learning decision rules inferred from prior data (training data). The decision tree algorithm tries to solve the problem, by using tree presentation.The internal node corresponds to attribute and leaf node to class label.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 1, pp: 907-911 www.ijaem.net ISSN: 2395-5252

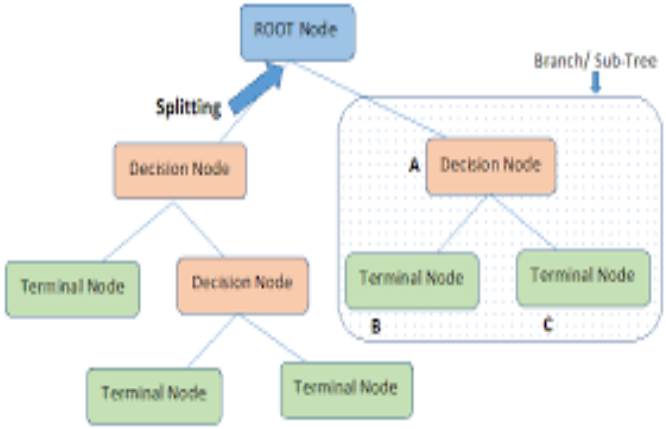

Figure 3. Overview of decision tree

3) Random Forest

Random forest is a treebased classification algorithm.

As the name indicates, the algorithm creates a forest with a large number of trees. It is an ensemble algorithm which combines multiple algorithms. It creates a set of decision trees from a randomsamples and makes a final decision based on majority voting. The Random forest algorithm is effective in handling missing values but it is prone to over – fitting

4) Naïve Bayes

The NB is a classification supervised learning algorithm.It is based on conditional probability theorem the class of a new feature vector. The NB uses the training dataset to seek out othe conditional probability value of vectors for a given class. After computing the probability conditional value of every vector, the new vectors class is computed supported its conditionality probability. NB is used for text-concerned problem classification.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Heart disease data was collected from the UCImachine learning repository. There are four databases (i.e. Cleveland, Hungary, Switzerland and VA Long Beach). The Cleveland database was selected for this experiment because it is a commonly used database for ML researchers with comprehensiveand complete records. The dataset contains 303 records. Although the Cleveland dataset has 76 attributes, the dataset provided in the repository furnishes information for the subset of only 14 attributes. The data source of Cleveland dataset is the Cleveland clinic foundation. There are 13 attributes that feature in the prediction of heart disease, where only one attribute serves as the output or the predicted attribute to the presence of heart disease in a patient. The Cleveland dataset contains an attribute named *num* to show the diagnosis of heart disease in patients on different scales, from 0 to 4. In this scenario, 0 represents the absence of heart disease and all values from 1 to 4 represents patients with heart disease, where scaling refers to the severity of the disease (4 being the highest)

The project is implemented using Anaconda. It is a free and open source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large- scale processing, predictive analytics, etc.), that actually aims to simplify package management and deployment. Spyder is a free integrated development environment (IDE) that is included with anaconda. It includes editing, interactive testing, debugging. The version of python used in the process is python 3.5

TABLEI. AVERAGEACCURACY

Algorithm	Average Accuracy
Support Vector Machine (SVM)	85.25%
Decision Tree	86.89%
Random Forest	90.16%
Naïve Bayes	86.89%
Neural Network	89.5%

On comparing the Decision tree, SVM, Random Forest, Naïve Bayes and KNN, it is clear that the Random Forest model gives highest accuracy rate for heart disease prediction. Results based on the accuracy, indicated Random Forest that was the best performing (90.16%), following by Neural Network (89.5%), Decision Tree (86.89%), Naïve Bayes (86.89%) and SVM (85.25%). SVM gave marginally lower accuracy.

V. CONCLUSION

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the very early stages and preventive measures are adopted as soon as possible. The average accuracy achieved by Random Forest is 90.16%. These results are better than those obtained using Decision tree, KNN, Neural network, Naïve ayes and SVM.

In Future, many studies can be conducted for those which results in restrictions of feature selection for algorithmic use.

REFERENCES

[1]. A.M.Mahmood and M.R. Kuppa, "Early detection of clinical parameters in heart disease by improved decision tree algorithm", 2010 second Vaagdevi International Conference on Information Technology for Real world Problems, Warangal, IEEE, 2010

[2]. S.Abdullah and R.R.Rajalaxmi, "A data mining model for predicting the coronary heart disease using random forest classifier," in Proc. Int. Conf.Recent Trends Comput. Methods, Communication controls, Apr.2012, pp.22_25.

[3]. Theresa Princy R.J. Thomas, "Human Heart Disease Prediction system using data mining techniques" International conference on circuit, Power and computing technologies [ICCPCT],2016.

[4]. S.Bashir , U.Qamar, and M.Y.Javed, "An ensemble based decision support framework for intelligent heart disease diagnosis",Inernational Conference o Information Society (i-Society 2014), London, IEEE, 2014

[5]. H.Alkeshuosh, M.Z.Moghadam, I.AlMansoori, and M.Abdar, "Using PSO algorithm for producing best rules in diagnosis of heart disease," in Proc.Int.Conf.Comput.Appl.(ICCA), Sep. 2017, pp.306_311.

[6]. V.Chaurasia and S.Pal, "Early prediction of heart diseases using data mining techniques" Caribbean Journal of Science and Technology, vol.1 208-217,2013.

[7]. N.Al-milli, "Backpropagation neural network for prediction of heart disease," J.Theor>appl.Inf>Technol., vol.56, no.1, pp.131_135, 2013.

[8]. M.Shouman, T.Turner and R.Stocker, "Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients", Proceedings of the International Conference on Data Mining, 2012.

[9]. Devi, S.P.Rajamohana, K.Umamaheshwari, R.Kiruba, K.Karunya, and R.Deepika, "Analysis of Neural networks based heart disease prediction system," in Proc.11thInt.Conf.hum.syst.Interact.(HSI), Gdansk, Poland, Jul.2018,pp.233_239

[10]. P.K.Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," J.King Saud Univ.Comput.Inf.Sci.Vol.24.no.1, pp. 27_40, Jan. 2012.

[11]. BerinaAlic, LejlaGurbeta, AlmirBAdnjevic, "Machine Learning Techniques for Classification of Diabetes and Cardiovascular disease" 2017 6th Mediterranean conference on Embedded cpmpting, 11-15 June 2017, Montenegro

International Journal of Advances in Engineering and Management
ISSN: 2395-5252

IJAEM

Volume: 02 Issue: 01 DOI: 10.35629/5252

www.ijaem.net
Email id: ijaem.paper@gmail.com