International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 06 | June -2017 www.irjet.net p-ISSN: 2395-0072

Comparing Data Mining Techniques Used For Heart Disease Prediction

Prof. Mamta Sharma [1], Farheen Khan[2] , Vishnupriya Ravichandran[2]

[1]Assistant Professor, [2]MCA SEM VI
[1,2]SIES College of Management Studies, Shri Chandrasekarendra Saraswathi Vidyapuram,

---

Abstract – Data mining is one of the richest areas of research that is more popular in health organizations. Data mining plays an effective role for uncovering new trends in healthcare organization which is helpful for all the parties associated with this field. Heart disease is the leading cause of death in the world over the past 10 years. Heart disease is a term that assigns to a large number of medical conditions related to heart. These medical conditions describe the irregular health condition that directly affects the heart and all its parts. The healthcare industry gathers enormous amount of heart disease data which are not "mined" to discover hidden information for effective decision making. Data mining techniques are useful for analyzing the data from many different dimensions and for identifying relationships. This paper explores the utility of various decision tree and neural network algorithms to classify and predict the disease.

Keywords – Heart disease, Data Mining, Soft computing, Decision Tree Techniques, Neural Networks.

I. INTRODUCTION

Due to heart disease, normal functionality of heart is affected. There are various factors which increases the risk of heart disease. Some of them are high blood pressure, cholesterol, family history of heart disease, obesity, hypertension, smoking, etc. In today's modern world, cardiovascular diseases are the highest flying diseases and in every year 12 million deaths occur over the world due to heart problem. In India casualties are also caused due to cardiovascular diseases and its diagnosis is very difficult process. Normally, these diseases can be analyzed using intuition of the medical specialist and it would be highly beneficial if the techniques used for analysis shall be improved with the medical information system. At reduced cost, if a decision support or computer based information system is developed then it will be helpful for accurate diagnosis.

Data mining combines statistical analysis, machine learning algorithms and database technology to extract hidden patterns and relationships from large databases [4]. Nowadays, many hospitals keep their present data in electronic form through some hospital database management system. These systems generate large volume of data on daily basis. This data may be in form of free text, structured as in databases or in form of images.[4] This data may be used to extract meaning information which may be used for decision making. Data mining which is one of the KDD (Knowledge discovery in database) concentrates on finding meaningful patterns from large datasets. These patterns can be further analyzed and the result can be used for further valuable decision making and analysis. It has helped to confirm the best prediction technique in terms of its accuracy and error rate on the specific dataset. [4] Researchers have been applying different data mining techniques to help health care professionals with improved accuracy in the diagnosis of heart disease. Neural network, Naïve Bayes, Genetic algorithm, Decision Tree, classification via clustering, and direct kernel self-organizing map are some techniques used.

By Using some data mining techniques heart disease prediction can be made simple by using various characteristic to find out whether the person suffers from heart attack or not, and it also takes less time to for the prediction and improve the medical diagnosis of diseases With good accuracy and minimizes the occurrence of heart attack [2]. Data mining along with soft computing techniques helps to unravel hidden relationships and diagnose diseases efficiently even with uncertainties and inaccuracies.

II. LITERATURE REVIEW

Numerous work has been done related to heart prediction system by using various data mining techniques and algorithms by many authors. The aim of all is to achieve better accuracy and to make the system more efficient so that it can predict the chances of heart attack. This paper aims at analyzing the various data mining techniques introduced in recent years for heart disease prediction. Different data mining techniques have been used in the diagnosis of CVD over different Heart disease datasets.

Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease. Statistical analysis has identified the risk factors associated with heart disease to be age, blood pressure, smoking habit, total cholesterol, diabetes, hypertension, family history of heart disease, obesity, and lack of physical activity.

Chaitrali S. Dangare and Sulabha S. Apte [3] showed that Artificial Neural Network outperforms other data mining techniques such as Decision Tree and Naïve Bayes. In this research work, Heart disease prediction system was developed using 15 attributes [3]. The research work included two extra attributes obesity and smoking for efficient diagnosis of heart disease in developing effective heart disease prediction system.

Aditya Methaila et.al. [11] In their research work focused on using different algorithms and combinations of several target attributes for effective heart attack prediction using data mining. Decision Tree has outperformed with 99.62% accuracy by using 15 attributes. Also, the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.

B. Venkatalakshmi and M.V Shivsankar in year 2014 [7] performed an analysis on heart disease diagnosis using data mining techniques Naïve Bayes and Decision Tree techniques. Different sessions of experiments were conducted with the same datasets in WEKA 3.6.0 tool. Data set of 294 records with 13 attributes was used and the results revealed that the Naïve Bayes outperformed the Decision tree techniques.

Due to the higher accuracy and learning rate the artificial neural network k(ANN) algorithms can also be used in the prediction of heart disease [9]. **Kumaravel et al.** have proposed automatic diagnosis system for heart diseases using neural network. In this system ECG data of the patients is used to extract features and 38 input parameters are used to classify 5 major types of heart diseases with accuracy of 63.6 - 82.9% [12].

In 2015, **Hnin Wint Khaing** presented an efficient approach for the prediction of heart attack risk levels from the heart disease database. Firstly, the heart disease database is clustered using the K-means clustering algorithm, which will extract the data relevant to heart attack from the database. The machine learning algorithm is trained with the selected significant patterns for the effective prediction of heart attack. They have employed the ID3 algorithm as the training algorithm to show level of heart attack.

III. TECHNIQUES USED FOR PREDICTION

A. NEURAL NETWORKS

An artificial neural network (ANN) is a computational model that attempts to account for the parallel nature of the human brain. An (ANN) is a network of highly interconnecting processing elements (neurons) operating in parallel. These elements are inspired by biological nervous systems.

Artificial neural networks work as a leading tool that helps doctors to evaluate, model and get sensible results from complex data. Most applications of artificial neural networks in medicine are diagnostic systems, biomedical analysis, image analysis, drug development. Feed-forward neural networks are widely and successfully used models for classification, forecasting and problem solving. A typical feed-forward back propagation neural network is proposed to diagnosis diseases. It consists of three layers: the input layer, a hidden layer, and the output layer [4].

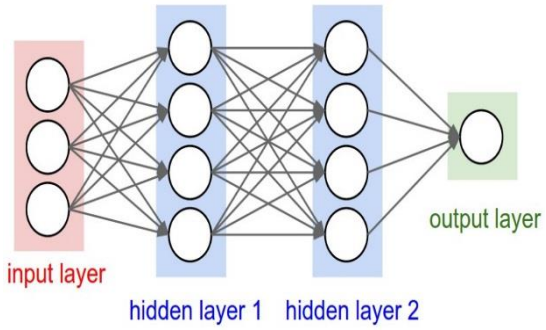

Figure 1: Artificial Neural Network

In a feed-forward neural network information always moves one direction; it never goes backwards. It allows signals to travel one-way only; from source to destination; there is no feedback. The hidden neurons are able to learn the pattern in data during the training phase and mapping the relationship between input and output pairs. Each neuron in the hidden layer uses a transfer function to process data it receives from input layer and then transfers the processed information to the output neurons for further processing using a transfer function in each neuron [5].

In other words, it is an emulation of biological neural system. It maps a set of input data onto a set of appropriate output data. It consists of 3 layers input layer, hidden layer & output layer. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input xi into neurons in hidden layer. Neuron of hidden layer adds input signal x_i with weights w_{ji} of respective connections from input layer. The output Y_j is function of $Y_j = f(\Sigma w_{ji} x_i)$ Where f is a simple threshold function such as sigmoid.

B) DECISION TREES

Applying Decision Tree techniques has shown useful accuracy in the diagnosis of heart disease. But assisting health care professionals in the diagnosis of the world's biggest killer demands higher accuracy. research seeks to improve diagnosis accuracy to improve health outcomes. Decision Trees (DTs) are a non-parametric supervised learning method used for classification. The main aim is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The structure of decision tree is in the form of a tree. Decision trees classify instances by starting at the root of the tree and moving through it until a leaf node. Decision trees are commonly used in operations research, mainly in decision analysis

Decision tree is similar to the flowchart in which every non-leaf nodes denotes a test on a particular attribute and every branch denotes an outcome of that test and every leaf node have a class label. . The node at the top most labels in the tree is called root node. There are many popular decision tree algorithms ID_3, C4.5, CART, and J_{48}

ID_3: - ID_3 stands for Iterative Dichotomiser 3. ID_3 adopt a greedy (i.e. nonbacktracking) approach in which decision trees are constructed in a top-down recursive divide and conquer manner. The resulting tree is used to classify the future samples

C4.5:- It builds decision tree from a set of training data in the same way as ID_3 using the concept of information entropy.C4.5 is often called as Statistical Classifier. understanding data. It can be used as a training tool to train nurses and medical students to diagnose patients with heart disease.

The Naïve Bayes Classifier technique is mainly applicable when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model recognizes the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state.

CART: - Classification and regression trees (CART) are a non-parametric decision tree learning technique that produces either classification or regression trees, epending on whether the dependent variable is categorical or numeric, respectively.

J_{48}: - J_{48} decision tree is the implementation of ID_3 algorithm, developed by WEKA project team. J_{48} is a simple C4.5 decision tree for classification. With this technique, a tree is constructed to model the classification process. Once the tree is build, it is applied to each tuple in the database and the result in the classification for that tuple.

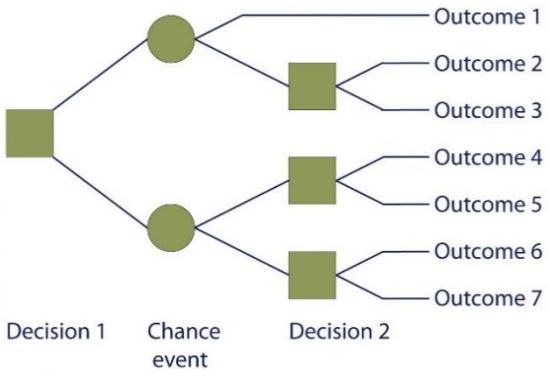

Figure 2 : Decision Trees

C) NAIVES BAYES THEOREM

Naïve Bayes or Bayes Rule is the basis for many machine learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and

Naive Bayes algorithm is preferred in the following cases.

- When the dimensionality of data is high
- When the attributes are independent of each other. Otherwise, attributes are assumed to be independent in order to simplify the computations involved and, in this sense, is considered “naïve”.
- When we expect more efficient output, as compared to other methods output.
- Exhibits high accuracy and speed when applied to large databases.

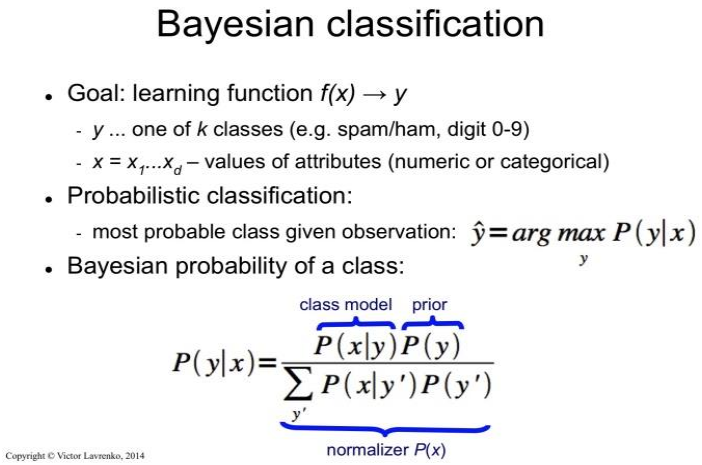

Figure 3 : Naïve Bayes algorithm

IV. ANALYSIS OF DATA

The dataset consist of 3 types of attributes. Input, Key and Prediction attributes. Commonly used attributes such as Age, Gender, Blood pressure, Pulse rate and Cholesterol are considered as input attributes of which Age and gender and non-modifiable attributes . Age is continuous and dynamic in nature where gender is static and constant. The other parameters have a continuous and Random Values. In order to get more appropriate results additional attributes such as Smoking and history of heart diseases also where included in the study. Smoking and Heart disease were the Modifiable attributes. Constant values were given to the smoking and heart disease to predict from the risk rate of heart disease. Patient id is considered as a key attribute which is unique for each and every user. Using this key attribute the patient and doctor retrieve record. Authenticity of the user is taken care by the application. The Prediction Attribute found the Risk level of the disease. The risk level was classified into three levels namely low risk, high risk, normal risk which indicates lesser than 50%, greater than 50% and 0 respectively[9].

V. CONCLUSION

Heart Disease is a fatal disease by its nature. This disease makes a life threatening complexities such as heart attack and death. The importance of Data Mining in the Medical Domain is realized and steps are taken to apply relevant techniques in the Disease Prediction. The various research works with some effective techniques done by different people were studied.

Study reveals that the Neural Networks with 15 attributes has outperformed over all other data mining techniques. Decision Tree has shown good accuracy with C4.5, ID_3, CART and J_{48}. Decision Tree has shown good accuracy with the help of genetic algorithm and feature subset selection. Naïve Bayes algorithm gives an average prediction with 90% accuracy. The following table shows the interpretation of various research papers we have studied:

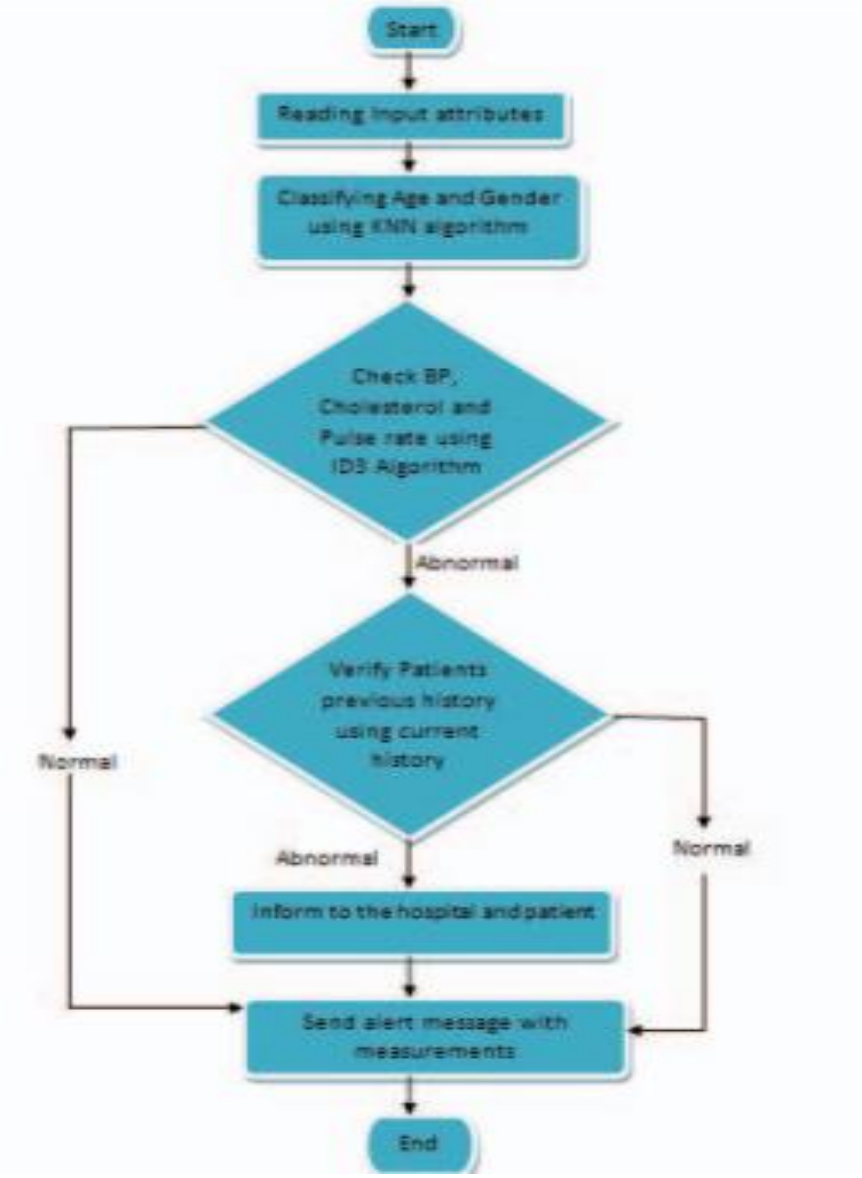

Figure 2: Flow chart of risk level Prediction system

Classification Techniques	Accuracy with (%)	
	13 attributes	15 attributes
Naïve Bayes	94.44	90.74
Decision Tree	96.66	99.62
Neural Networks	99.25	100

VI. REFERENCES

[1] G. E. Sakr, I. Elhajj, and H. Huijer, "Support vector machines to define and detect agitation transition", IEEE Transactions on Affective Computing, vol. 1, pp. 98–108, December 2015.
[2] Neha Chauhan and Nisha Gautam "An Overview of heart disease prediction using data mining techniques"
[3] 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 2014.
[4] Sujata Joshi and Mydhili K.Nair,"Prediction of Heart Disease Using Classification Based Data Mining Techniques", Springer India 2015, volume 2.
[5] Beant Kaur and Williamjeet Singh.," Review on Heart Disease Prediction System using Data Mining Techniques", IJRITCC ,October 2016.
[6] Frank Lemke and Johann-Adolf Mueller, "Medical data analysis using self-organizing data mining technologies," Systems Analysis Modeling Simulation , Vol. 43, Issue No. 10, 2003, pp. 1399-1408
[7] B.Venkatalakshmi, M.V Shivsankar, Heart Disease Diagnosis Using Predictive Data mining; International Journal of Innovative Research in Science,
Engineering and Technology Volume 3, Special Issue 3, March 2016.
[8] Beant Kaur h, Williamjeet Singh, "Review on Heart Disease Prediction System using Data Mining
Techniques", International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 10, pp.3003-08, October 2015.
[9] Nidhi Bhatlet and Kiran Jyoti, "An Analysis of Heart disease prediction system using different data mining
techniques", International Journal of Engineering Research and Technology,ISSN,volume 1,October-2016
[10] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network", In proceedings of IEEE International Conference on Computer and Communication Technology (ICCCT), pp. 741–745, November 2015.
[11] Aditya Methaila, Early Heart Disease Prediction Using Data Mining Techniques; CCSEIT, DMDB, ICBB, MoWiN, AIAP pp. 53–59, 2016.
[12] N. Kumaravel, K. Sridhar, and N. Nithiyanandam, "Automatic diagnoses of heart diseases using neural network", In Proceedings of the Fifteenth Biomedical Engineering Conference, pp. 319–322, March 2016