Usage of Data Mining Techniques in Predicting the Heart Diseases Decision Tree & Random Forest Algorithm

G S Mallikarjuna Rao, K Anitha

Abstract: ***Nowadays, heart disease is the main cause of several deaths among all other diseases. Due to the lack of resources in the medical field, the prediction of heart diseases becomes a major problem. For early diagnosis and treatment, some classification algorithms such as Decision Tree and Random Forest Algorithm are used. The data mining techniques compare the accuracy of the algorithm and predict heart diseases. The main aim of this paper is to predict heart disease based on the dataset values. In this paper we are comparing the accuracy of above two algorithms. To implement these methods the following steps are used. In first phase, a dataset of 13 attributes is collected and it was applied on classification techniques using the Decision tree and Random Forest Algorithms. Finally, the accuracy is collected for both the algorithms. In this paper we observed that random forest is generating better results than decision tree in prediction of heart diseases.***

Index Terms: Classification, Decision Tree, Heart disease, Random Forest.

I. INTRODUCTION

Many new techniques and algorithms of data mining have been used for predicting heart diseases. Data mining proved its efficiency in most of the areas to achieve improved accuracy and performance mostly in the medical field. Classification is termed as one of the data mining techniques which are used to predict group membership for data instances.Machine learning techniques are one of the existing techniques which have a transparent diagnostic knowledge which gives more accurate results. Generally, Machine learning is classified into two types, one is Supervised learning and the other is Unsupervised learning. In Supervised learning, the output or an outcome for the given input has known itself and the machine must be able to map or assign the given input to the output.

In Unsupervised learning, the outcome or output for the given inputs is unknown. Here input data are given and the model is run on it. Decision Tree Algorithm is a supervised algorithm. This will solve the problems by using a Tree like representation. Each internal node of the tree represents an attribute and each leaf node represents a class label.

A Random Forest algorithm is a supervised algorithm. As the name suggests, this algorithm creates the forest with a number of trees.

Heart diseases are seen in all the classes of people in recent times. Cardiovascular disease is the leading cause of death. Hence, continued efforts are being done to predict the possibility of getting heart diseases. Cardiovascular heart diseases are of two types which are most common in people with diabetes. They are,

- Coronary Artery diseases
- Cerebral Vascular disease.

Some symptoms of heart diseases are a pain in chest, shoulders, arms, jaws, breath shortness,

Giddiness and nausea. The major problem of Coronary heart diseases is high blood pressure and also diabetes which may weaken the heart.

Heart Disease: Heart Disease is a term used to refer any disorder that affects the heart. Heart diseases are of several forms.

A. Coronary Heart Disease: Coronary heart disease also known as Coronary Artery disease (CAD) or Ischemic heart disease (IHD). These diseases come under cardio vascular disease which involves the valves of the heart.

Symptoms include:

Heart burn

Shortness of breath

Chest pain, discomfort which radiates to neck or back

B. Heart Attack: Heart attack is the most common term used for myocardial infarction (MI). It occurs due to interruption of blood supply to a part of heart leading to damage to the heart muscle. The most common symptoms are pain on shoulder, back or jaw and also pain on the left side of the chest.

C. Heart Failure: Inability to pump the blood for body functioning. Heart attack, Coronary artery disease may lead to Heart failure. Symptoms may include leg swelling, fatigue and breathing. Heart disease is considered as a silent killer which leads to death without any obvious symptoms occurred in all classes of people. As a well know quote says “Prevention is better than Cure”, which means early diagnosing the problem can be helpful to decrease the death rates further.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author

GS Mallikarjuna Rao*, Department of Computer Applications,Gayatri Vidya Parishad College of Engineering(Autonomous),Madhurawada, Visakhapatnam - 530 048,Andhra Pradesh, India.

K Anitha, Department of Computer Applications, Gayatri Vidya Parishad College of Engineering (Autonomous), Madhurawada, Visakhapatnam - 530 048, Andhra Pradesh, India.

This paper discusses about Decision tree and Random Forest Algorithms. The Random Forest algorithm is robust and much more efficient due to the following features:

Algorithm performs both classification and regression tasks.

Algorithm Doesn’t overfit the model.

Handles missing and noisy data so no need of preprocessing the dataset as it maintains the accuracy for missing data.

Handles Large dataset with high dimensionality.

Several techniques are used to detect heart diseases based on a large number of attributes. In order to reduce the attributes from a large dataset, a KNN classification approach is used to reduce the attributes list but takes more amount of time to perform classification and then reduces the attributes which may not be accurate in prediction of heart disease for a particular dataset. Naïve Bayes is one of the most popular classification algorithms used in data mining. By using the Naïve Bayes algorithm diagnosing a heart disease is possible based on the list of attributes and provides the calculation of yes/no probability. But the disadvantage is loss of accuracy and also strong feature independence.

II. RELATED STUDY

For diagnosing of heart diseases most of the papers have implemented several data mining techniques such as Naïve Bayes, Neural network, Kernel density, automatically defined groups, bagging algorithm and support vector machine showing different levels of accuracy on multiple databases of patients at different levels. For testing the accuracy of algorithms, many authors have specified different attributes and different datasets. In particular, researchers have been investigating the application of decision tree algorithm in the diagnosis of heart disease with considerable success. In [1] the authors have made a study and developed a system for diagnosing the congenital heart diseases using decision support system. The system also uses the Back propagation Neural Networks based on the data and signs of the heart disease and symptoms are obtained from the patient. The system results the 90% accuracy. In [2] The authors designed a system by using a artificial neural network which is frequently used in the medical fields for prediction. This paper shows both merits and demerits of machine learning techniques like SVM, Naïve Bayes, Neural Network.C.D. Katsis in et al [3] proposed a way using Correlation Feature Selection (CFS) procedure & an Artificial Immune Recognition System (AIRS) classifier for breast cancer prediction. By adopting the SVM technique the accuracy results as 76.33% on data collected for 52 patients among 4726 cases. The paper by Pushpa in et al [4] proposed a system for predicting heart diseases using data mining techniques: Decision Tree and Naïve Bayes and showed the Decision tree provide more accuracy than naïve bayes for a dataset related to heart diseases collected from UCI laboratory. The study done by B.J Jaidhan & Shanmuk Srinivas et al[5]. A developed a system for detecting the fraudulent credit card transactions by using a machine learning technique – Random Forest algorithm and showed that the efficiency increased by 0.267% compared to the traditional models. The study done by Sellappanpalaniappan, Rafiah Awang developed a prototype Intelligent Heart Disease Prediction System(IHDPS) using data mining techniques. The heart diseases are predicted based on the patterns, relationships between medical factors related to heart disease.In et al[7] proposed a system for prediction of heart diseases using data mining techniques such as Naïve bayes, Decision tree algorithm, KNN, Neural networks. This paper proves that more number of attributes results high level of accuracy. Results shows that each technique has its infrequent strength in realizing the objectives of the defined mining goals.The study done by Kalaiselvi proposed a system for prediction of heart diseases using KNN algorithm. The author suggested that Average KNN is used to reduce the attributes based on binary classification from a huge dataset and then applying the clustering for division of records.

Decision Tree Algorithm

After calculating the entropy and information gain for a particular given data set the root node is calculated and the tree is generated as follows.

Fig 2 depicts the decision tree generation for heart disease prediction which consists of root node and child node attributes. Root node contains the maximum entropy among all other attributes that is CP.

Random Forest Algorithm

Random Forest Algorithm increases the predictive power of the algorithm and also helps to prevent over-fitting. The random forest algorithm is an ensemble of a randomized decision tree. Each decision tree gives a vote for the prediction of the target variable. Random forest algorithm chooses the prediction that gets the most vote. In the Random forest, the system uses multiple random decision trees for better accuracy. The subset of trees is taken from the decision tree and prediction can be done by considering the majority of the votes given by each subtree. Thereby the majority vote is considered as a target variable/final decision

III. PROPOSED METHODOLOGY

In the Proposed system, we are using Decision tree and Random forest algorithms to predict heart diseases. As the Naïve Bayes classifier requires a small dataset to predict there will be a loss of accuracy which is the disadvantage of the naïve Bayes classifier. Where in the decision tree and random forest algorithm requires less effort for prediction as it is in the tree-like structure[8] [9]. Using a decision tree and random forest interpretation of a complex decision tree model can be simplified by its appearance. Even a naïve person can understand the logic easily.The main aim of this paper is to predict the presence or absence of heart disease in one particular heart patient record. Here classification techniques like Random Forest and Decision tree algorithms were used [10]. This paper even compared the accuracy of both the algorithms. Figure 1 depicts the architecture of the proposed model for the prediction of heart disease. Where it collects the data and applies the classification algorithms such as the Decision Tree algorithm and the Random Forest Algorithm. After the classification the result will be predicted and also the accuracy will also be calculated.

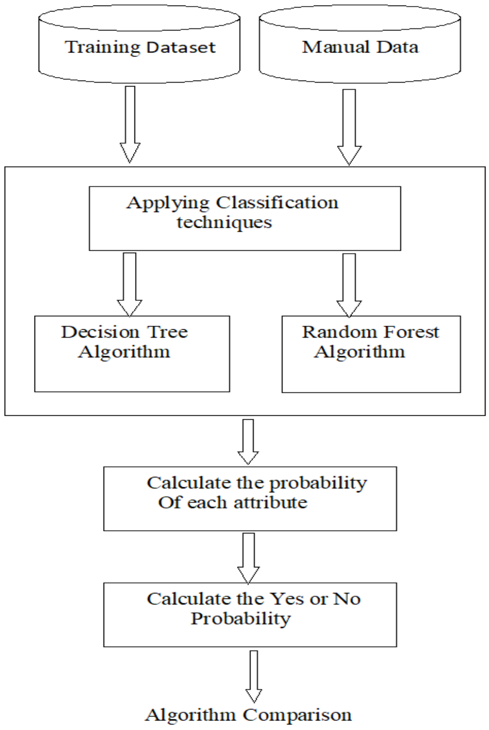
Training Dataset
Manual Data
Applying Classification techniques
Decision Tree Algorithm
Random Forest Algorithm
Calculate the probability Of each attribute
Calculate the Yes or No Probability
Algorithm Comparison

Figure 1. Architecture of the proposed model

Algorithms

Input: Dataset Individual patients' data
Output: Prediction result or (Graph).

1. Begin
2. Decisions will be depicted from the root node to terminal/leaf node by using ID3
3. To generate a tree, root node is the most important aspect for decision tree. Among 13 attributes, root node can be identified by calculating entropy and Information gain by ID3.
4. Entropy H(S) is used to measure the amount of uncertainty in the given data set.
   $H(S) = \sum_{c \in C} -p(c) \log_2 p(c)$
5. Information gain is used to measure how much uncertainty in dataset was reduced after splitting the dataset S on attribute A.
   $IG(A,S) = H(S) - \sum_{t \in T} p(t) H(t)$
6. End

A dataset containing 13 attributes related to heart diseases is collected. The attributes are age, gender, cp(Chest pain), fbs(sugar level), chol(Cholesterol), thal(heart rate achieved), slope, rest ECG, etc. Collection of data can be done in both static and dynamic way. Where in a static way 13 attribute dataset is collected form UCI laboratory and applied the classification algorithms to that dataset. Where in a dynamic way those attributes are collected for one particular patient dynamically and perform the prediction using both classification techniques. For the dataset, no preprocessing is required as the dataset is not containing any missing fields and duplications are avoided. Classification is a process where we categorize data into given number of classes. Classification can be performed on both structured and unstructured data. Classification technique is used to identify the target variable value. Some classification algorithms are Logistic Regression, Naïve Bayes, K – Nearest Neighbors, Decision Tree, Random Forest and Support vector machine. Classification model classifies unordered and discrete values or data. In this prediction process classification techniques used is Decision Tree and Random Forest Algorithm.

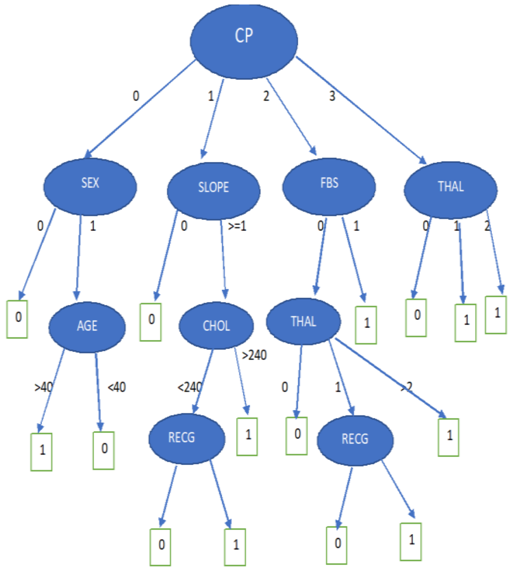

Figure 2: Decision tree structure for prediction of a disease

.

IV.EXPERIMENT AND RESULT

Here we have considered a sample data set containing heart disease related data. The table consists of 303 records and 15 columns. Our dataset contains 14 attributes and one target variable (0 and 1). Where 0 represent No and 1 represents Yes.For the above dataset the output will be predicted for both algorithms is shown in the Table 2. By following the above generated decision tree, the data set results are displayed. The output directly shows the possibility of getting a heart disease for each row of the dataset. To know the accurate result of a particular person Random Forest provides its best result than decision tree in the form of a graph which displays both Yes probability and No probability.

The proposed system also provides the accurate result for each record dynamically in the form of a graph. For the first instance of dataset results a patient with heart disease. This can be computed by using the above generated decision tree and majority of the votes given by the randomly generated trees. The probability of both the algorithms is shown in the below graph fig 3.

The Random forest algorithm gives the Yes probability as 0.96 and No probability as 0.4. Whereas the decision tree gives yes/no probability directly. Hence it shows 1.0 as yes. Finally, both algorithms say the patient is having a heart disease. In Fig 4. Depicts the comparison between the decision tree and Random forest algorithm for the 5th instance of dataset. Where Random Forest algorithm provides Yes probability as 0.37 and No probability as 0.63. Whereas the Decision tree algorithm gives 1.0 as No probability. Therefore, The above graph results patient with No Heart Disease.

Table 1: Sample heart disease prediction dataset.

Patient Name	Age	Sex	CP	Tresbps	Chol	fbs	Restecg	Thalach	Exang	oldpeak	Slope	Ca	Thal	Target
LILY	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
KOEL	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
JACK	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
GEORGE	68	1	2	180	274	1	0	150	1	1.6	1	0	3	0
MIA	62	0	0	160	164	0	0	145	0	6.2	0	3	3	0
JACOB	56	1	3	120	193	0	0	162	0	1.9	1	0	3	1
HARRY	65	0	0	150	225	0	0	114	0	1	1	3	3	0
SOPHIA	41	1	1	120	157	0	1	182	0	0	2	0	2	1
NOAH	38	1	2	138	175	0	1	173	0	0	2	4	2	1
CHARLIE	49	1	2	120	188	0	1	139	0	2	1	3	3	0
ISABELLA	59	1	0	140	177	0	1	162	1	0	2	1	3	0
BOB	57	1	2	128	229	0	0	150	0	0.4	1	1	3	0
OLIVER	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
FREDDIE	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
STEPPIE	59	1	0	164	176	1	0	90	0	1	1	2	1	0

Table 2. Prediction result of Entire Dataset

Patient Name	Age	Random Forest	Decision tree
LILY	41	1	1
KOEL	56	1	1
JACK	57	1	1
GEORGE	68	0	0
MIA	62	0	0
JACOB	56	1	1
HARRY	65	0	0
SOPHIA	41	1	1
NOAH	38	1	1
CHARLIE	49	0	0
ISABELLA	59	0	0
BOB	57	0	0
OLIVER	54	1	1
FREDDIE	48	1	1
STEPPIE	59	0	0

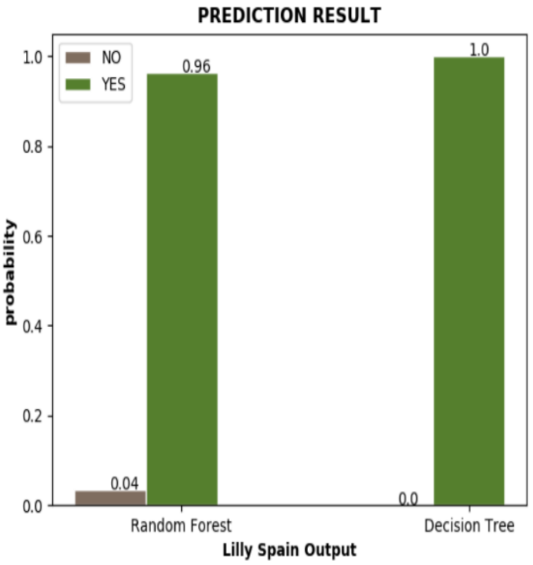

Figure 3. Comparison of Decision Tree and Random Forest Algorithm (Test case 1)

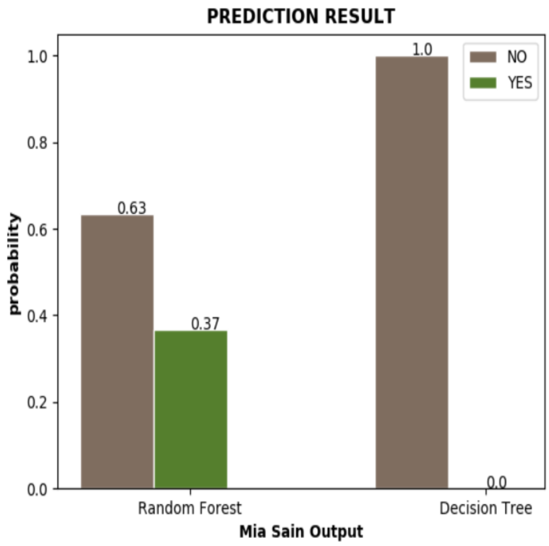

Figure 4. Comparison of Decision Tree and Random Forest Algorithm (Test case 2)

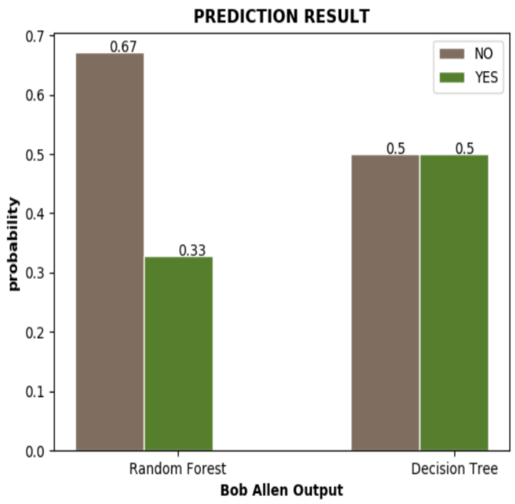

Figure 5. Comparison pf Decision Tree and Random Forest Algorithm (Test case 3)

In Fig 5. Depicts the comparison between the decision tree and Random forest algorithm for the 10th instance of dataset. Where Random Forest algorithm provides Yes probability as 0.33 and No probability as 0.67. Here the Decision tree algorithm gives 0.5 probabilities for both yes and no cases. As the random Forest gives the accurate result so by considering the Random Forest output the above graph results patient with No Heart Disease. Similarly, a new patient with 13 attributes are also predicted by using these data mining techniques that is by considering already generated decision tree. The results show that Random Forest Algorithm provides more accuracy than compared to the Decision Tree. The accuracy for Decision Tree algorithm is 98.1% whereas the accuracy for Random Forest algorithm is 98.6%. Which results Random Forest algorithm is so far better than the Decision tree algorithm.

V. CONCLUSION

Prediction of Heart Diseases Using a Data Mining Approaches (Decision Tree and Random Forest Algorithm) gives the higher efficiency and reduces complexity based on the attribute reduction. Initial attributes used. Age,Sex,CP,Tresbps,Chol,fbs,Restecg,Thalach,Exang,oldpeakSlope,Ca,Thal Target. The Random forest algorithm performs well and classifies the dataset of the Heart Disease into two classes well when compared to traditional methods.

Finally, proposed work reduces the cost for different medical tests and helps the patients to take precautionary measures well in advance. In the future, the same method can also be applied in predicting and diagnosing other disease types.

REFERENCES

1. VanishreeK,JyothiSingaraju,"Decision Support System for congential heart disease diagnosis based on signs and symptoms using neutral network" International Journal of computer applications,April 2011 Vol 19 no.6.
2. S.Kharya, D. Dubey, and S. Soni - Predictive Machine Learning Techniques for Breast Cancer Detection, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013, 1023-1028.
3. C. D. Katsis, I. Gkogkou, C.A. Papadopulos, Y.Goletsis, P. V. Boufounou, G. Stylios "Using artificial immune recognition systems in order to detect early breast cancer." International Journal of Intelligent Systems and Applications 5.2 (2013): 34.
4. Priyanka N, Dr. Pushpa Ravi Kumar, "Usage of Data mining techniques in predicting the heart diseases – Naïve Bayes and Decision Tree." International Conference on circuits Power and Computing Technologies 2017.
5. B.J.Jaidhan, B.Divya Madhuri, K.Pushpa, B.V.S. Lakshmi Devi, Shanmukh Srinivas A, "Application of Big Data Analytics and Pattern Recognition Aggregated with Random Forest for Detecting Fraudulent Credit Card Transactions(CCFD – BPRRF)" International Journal of Recent Technology and Engineering(IJRTE), Vol. 7 -March 2019.
6. SellappanPalaniappan, Rafiah Awang, "Intelligent heart disease prediction system using data mining prediction" ACS International Conference on Computer Systems and Applications - 2008.
7. J. Thomas, R. Theresa Princy, "Human heart disease prediction system using data mining techniques" 2016 International Conference on Circuit, Power and Computing Technologies(ICCPCT).
8. Amiripalli, S. S., & Bobba, V. (2019). An Optimal TGO Topology Method for a Scalable and Survivable Network in IOT Communication Technology. Wireless Personal Communications, 1-22.
9. Shanmuk Srinivas Amiripalli, VeeramalluBobba (2019), Impact of trimet graph optimization topology on scalable networks, Journal of Intelligent & Fuzzy Systems 36 (3), 2431-2442. DOI: 10.3233/JIFS-169954
10. C.Kalaiselvi, "Diagnosing of Heart diseases using average k – nearest neighbor algorithm of data mining" published in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom).
11. Amiripalli, S. S., & Bobba, V. (2018). Research on network design and analysis of TGO topology. International Journal of Networking and Virtual Organisations, 19(1), 72-86.

AUTHORS PROFILE

G.S.Mallikarjuna Rao is working as Associate Professor & HOD, Department of Computer Applications, Gayatri Vidya Parishad College of Engineering(A), Visakhapatnam(AP). He holds M.Tech(CSE), ME(I.E) and B.Tech(ECE) all from Andhra University College of Engineering, Visakhapatnam. He is a life member of Computer Society of India and Indian Society for Technical Education.

K.Anitha is working as Software Developer at Savvy Software Solutions, Visakhapatnam.She was awarded Master of Computer Applications in the year 2019 from Gayatri Vidya Parishad College of Engineering(A) affiliated to JNTU College of Engineering Kakinada.