ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 7, July 2022

A Comparative Study on Decision Tree and Random Tree Approach in Predicting Heart Diseases

[1] Sudhanva S, [2] Tarun T N, [3] Sujay C V, [4] Vishal D A, [5] Yuvraaj, [6] Spandana S G

[1] [2] [3] [4] [5] Dayananda Sagar Academy Of Technology And Management, Bangalore, Karnataka, India.
[6] Faculty, Dayananda Sagar Academy Of Technology And Management, Bangalore, Karnataka, India.
Corresponding Author Email: [1] ssudhanva2k1@gmail.com, [2] taruntn291@gmail.com, [3] sujay.suraj@gmail.com,
[4] vdavishu@gmail.com, [5] yuviyuvraaj246@gmail.com

Abstract— ***Heart disease is one of the leading causes of death in the globe. All doctors cannot be equally proficient in every domain and well versed and skillfull doctors can't be available at all time. An automated medical diagnosis system would improve medical treatment while simultaneously lowering expenses. Predicting the course of illness is a difficult endeavor. Data mining is used to infer diagnostic principles automatically and assist professionals in making the diagnosis process more trustworthy. Researchers employ a variety of data mining approaches to assist health care practitioners in predicting cardiac disease. A classification model would be a fit model for predicting diseases accurately. We present you with a comparison of two classification models, Decision trees and Random Forest Models. We find that Random Forest method performs better than Decision Tree model.***

Keywords— ***Heart Disease, Random Forest, Decision Trees, Data Mining.***

I. INTRODUCTION

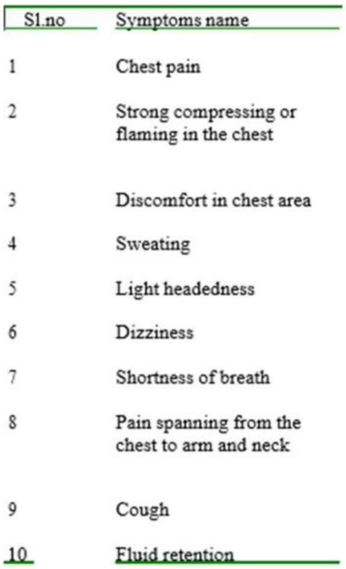
By and large, the determination of heart illness depends on a muddled blend of clinical and obsessive information; this intricacy prompts extravagant clinical expenses, which affect the nature of clinical consideration. As per WHO measurements, 33% of the total populace kicked the bucket from coronary illness in 2010, and coronary illness was the primary driver of death in emerging nations [1]. As per an American heart affiliation research, 33% of all grown-ups in the US have at least one sorts of coronary illness. Computational science is habitually utilized in the interpretation of organic information into clinical practice, as well as in the examination of natural peculiarities in view of clinical data. One of the main commitments of computational science is the finding of biomarkers in coronary illness. For medical purposes, this technique involves the production of a prescient model and the combination of different types of information and information. Information mining plays had a fundamental impact in coronary illness research in late many years. Patients' treatment depends on the arrangement of their coronary illness. Insights and machine learning are the two primary philosophies that have been utilized to figure the state of coronary illness in view of clinical information articulation. Decision tree is similar to a tree with every internal branch signifying a test on a prescient component and each branch indicating a property estimation, is one of the most common order models. anticipated classes or class circulations are addressed by a leaf hub. Beginning at the highest (root) node of the tree, an unlabeled element is arranged by crossing the tree contingent upon the upsides of the prescient traits of this article. The objective (subordinate) variable has discrete result esteems, each worth addressing a class mark, and choice tree methods expect that the information objects are depicted by a proper arrangement of characteristics, where each prescient quality takes few disjoint potential qualities and the objective (subordinate) variable has discrete result values. [2]

Data mining (DM) is a "nontrivial extraction of verifiable, new, and conceivably usable data from information", which is the essential phase of Knowledge Discovery in Databases (KDD). [3] It utilizes ML and measurable approaches to uncover beforehand obscure subject matters. The KDD technique ordinarily incorporates the means of Data Selection, Pre-processing of data, Data Transformation, DM (enlistment of significant patterns), and understanding the result. Several data mining techniques, such as decision tree, Naive Bayes, kernel density , neural network, support vector machine and bagging algorithm, are utilized in the detection of cardiac disease, with varying degrees of accuracy.[4] Some of the common decision tree algorithms are: ID3, C4.5, ART, See5, IFN.[5]

II. LITERATURE REVIEW

A. Heart Disease

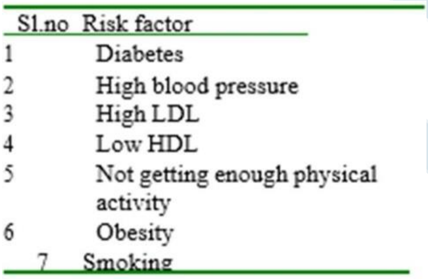
Any type of abnormal conditions related to blood vessels and muscles in the heart may lead to heart diseases, often known as coronary heart disease (CHD). Heart disease can begin as early as 18 years old, and people are only diagnosed with the condition when the blockage is greater than 70%. These obstructions grow over time, causing the membrane that covers the blockage to burst owing to increased pressure. Heart disease occurs when the chemicals produced by a damaged membrane mix with blood and cause a blood clot



Sl.no	Symptoms name
1	Chest pain
2	Strong compressing or flaming in the chest
3	Discomfort in chest area
4	Sweating
5	Light headedness
6	Dizziness
7	Shortness of breath
8	Pain spanning from the chest to arm and neck
9	Cough
10	Fluid retention

Table 1: Common symptoms of coronary diseases [6]

Risk factors are factors that raise the likelihood of a blockage. There are two types of risk factors: modifiable and non-modifiable risk factors. Age, gender, and heredity are non-modifiable risk factors. These risk factors are unchangeable.



Sl.no	Risk factor
1	Diabetes
2	High blood pressure
3	High LDL
4	Low HDL
5	Not getting enough physical activity
6	Obesity
7	Smoking

Table 2: Some of the risk factors related to cardiac diseases[8]

B. Decision Trees

A decision tree can be described with a flowchart-like structure where the internal nodes signify a "test" on an attribute(example : If a random person selected is Male or Female), a branch describes the test's conclusion and the leaf node signifies the class label.[14] The categorization rules are represented by the pathways from root to leaf. A decision tree serves as a visual and analytical decision support tool in decision analysis to calculate the expected values of competing alternatives.

There are three sorts of nodes in a decision tree:

Squares are commonly used to symbolise decision nodes. Chance nodes are usually depicted as circles.

End nodes - triangles are commonly used to depict them.

Decision trees are extensively made of use in operations research and management. If judgments must be made online with no recall and limited information, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. As a descriptive approach, decision trees can be employed to calculate conditional probabilities.

Undergraduate students in health economics, business and public health institutions are taught decision trees, impact diagrams, utility functions, and other decision analytic tools and procedures. These are some illustrations of operations research or management science methodologies.

C. Random Forest

One of the most effective ensemble classification approaches is the random forest algorithm. The RF method is extensively utilised in probability estimation and prediction. Many decision trees make up RF. Every decision tree provides a vote that indicates the object's class decision. Bell Labs' Tin Kam HO proposed the random forest item for the first time in 1995.

In a random forest, there are four significant tuning parameters: 1) The total aggregate of trees (n tree) 2) The minimal size of the node 3) The characteristics used to separate each node 4) The characteristics used to separate each node for each tree (m try).

The benefits of the random forest algorithm are described below:

1) The random forest algorithm is a reliable ensemble learning method.
2) Random forest method produces reliable results with huge datasets and therefore is a better choice if the dataset is adequately large
3) Random Forest approach works well with a large number of input variables.
4) The random forest method determines which factors are essential for categorization.
5) It can deal with data that is missing.
6) For class unbalanced data sets, Random Forest includes ways for balancing error.
7) This method's generated forests can be preserved for future use
8) Random forest solves the issue of overfitting.
9) RF is less susceptible to outliers in training data.
10) Parameters may be simply specified in RF, which avoids the requirement for tree pruning.

Randomization is used to choose the optimal node to split on when building individual trees in random forest.

III. METHODOLOGY

A. Decision Trees

The gain ratio decision tree is the type of decision tree employed in this study. The decision tree for the gain ratio is based on the entropy (information gain) strategy, which chooses the splitting attribute that reduces entropy and hence increases information gain [7]. The difference between both the information gain and the data redundancy is the information gain. Unique information substance, as well as the volume of data required. The information is used to rate the characteristics.

There are three sorts of nodes in a decision tree:

Squares are commonly used to symbolise decision nodes. Chance nodes are usually depicted as circles.

End nodes - triangles are commonly used to depict them.

Decision trees are extensively made of use in operations research and management. If judgments must be made online with no recall and limited information, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. As a descriptive approach, decision trees can be employed to calculate conditional probabilities.

Undergraduate students in health economics, business and public health institutions are taught decision trees, impact diagrams, utility functions, and other decision analytic tools and procedures. These[13] are some illustrations of operations research or management science methodologies.

B. Random Forest

One of the most effective ensemble classification approaches is the random forest algorithm. The RF method is extensively utilised in probability estimation and prediction. Many decision trees make up RF. Every decision tree provides a vote that indicates the object's class decision. Bell Labs' Tin Kam HO proposed the random forest item for the first time in 1995.

In a random forest, there are four significant tuning parameters: 1) The total aggregate of trees (n tree) 2) The minimal size of the node 3) The characteristics used to separate each node 4) The characteristics used to separate each node for each tree (m try).

The benefits of the random forest algorithm are described below:

1. The random forest algorithm is a reliable ensemble learning method.
2. Random forest method produces reliable results with huge datasets and therefore is a better choice if the dataset is adequately large
3. Random Forest approach works well with a large number of input variables.
4. The random forest method determines which factors are essential for categorization.
5. It can deal with data that is missing.
6. For class unbalanced data sets, Random Forest includes ways for balancing error.
7. This method's generated forests can be preserved for future use
8. Random forest solves the issue of overfitting.
9. RF is less susceptible to outliers in training data.
10. Parameters may be simply specified in RF, which avoids the requirement for tree pruning.

Randomization is used to choose the optimal node to split on when building individual trees in random forest

IV. METHODOLOGY

A. Decision Trees

The gain ratio decision tree is the type of decision tree employed in this study. The decision tree for the gain ratio is based on the entropy (information gain) strategy, which chooses the splitting attribute that reduces entropy and hence increases information gain [9]. The difference between both the information gain and the data redundancy is the information gain. Unique information substance, as well as the volume of data required. The information is used to rate the characteristics.

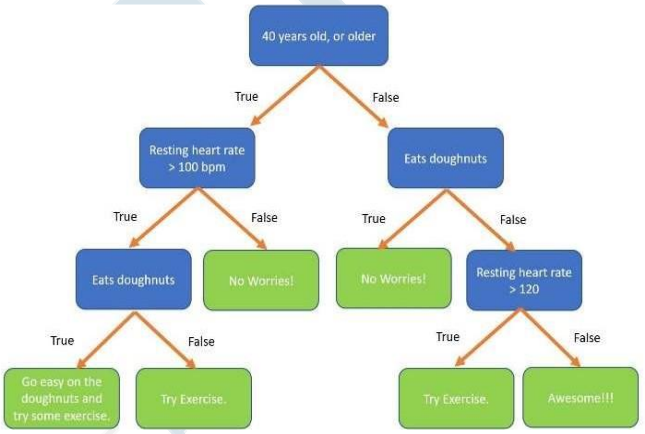


Fig 2: An example of a Decision Tree

The top-ranked features then are picked as prospective characteristics for inclusion in the classifier. to locate the In order to split the decision tree's attributes, one must first compute the information gain for each attribute and then choose one. The feature that increases the amount of information gained Each attribute's information gain is determined using the formula

Here j is the total number of target attribute classes. The number of instances of class I divided by the set of instances is called Pi (i.e. the probability of occurrence of I). To lessen the impact of prejudice The Australians devised a version called as performance gain as a result of the application of information gain. Ross Quinlan is a researcher [10]. The information gain metric favours tests with many outcomes. That is to say, It favours properties with a high number of possible values [11]. The information gain is adjusted using the Gain Ratio. Each attribute to ensure that the attribute values are diverse and consistent.

Information Ratio = Gain Ratio Knowledge Obtained / Split

B. Random Forest

In most cases, the diagnosis of heart-related disease is based on medical and pathological data; this difficulty results in much expensive medical costs, which in turn decides the quality of medical care. Data mining is used to predict automatic diagnostic principles and to assist professionals in making the diagnostic process more reliable. Researchers have used a variety of methods to extract data from health workers in predicting heart disease.To predict heart disease, scientists use four data mining algorithms: Nave go, random forest, line deceleration, and resolution tree.

The random forest, compared to other algorithms, has a high accuracy of 83.70 percent. It is a machine learning algorithm that builds a few decision trees. The final decision is made based on the majority of the decision tree. The decision tree suffers from low bias and high diversity. The random forest converts high variations to low variations. RF contains a lot of cutting trees. Each of multiple decision trees provides a poll that reflects the decision by category of the object.[12] There are three main planning elements in a random forest 1) Number of trees 2) Minimum node 3) Number of elements used to divide each node 3) Number of elements used to divide each area in each tree (m try).Randomization is used to select the node to be separated when building each tree in a random forest.

V. RESULT ANALYSIS

Hold-out & Cross-Validation were used in the experiments. We segregated samples into 2 unique data sets in the Hold out technique. 75% of the information the dataset is used to both train and construct the classifier. The other 25% of the data is employed for testing. In tenfold cross validation every instance of the data set is considered and classified into ten distinct groups, nine of which are utilised for training remainder is used for testing. The procedure repeats ten times and the average fold accuracy is calculated. We discovered that the Random Forest method worked better than the Decision Tree model after experimenting with the two algorithms, Decision Tree and Random Forest.

On the Cleveland Dataset, the Random Forest had an accuracy of 83.70, whereas Decision Trees had an accuracy of 82.43. When there is no heart illness, the probability of testing the result for heart disease is low . Positive predictive value(PPV) can be defined as the probability that any cardiac ailment is present for all patients whose diagnosis came positive.

According to the aforementioned experimental results, the Random Forest technique effectively reduces dimensionality and improves accuracy with dominant features. Overall, the Random Forest method beats other methods. This could be the result of random forest considering multiple decision trees and then achieving the result by polling each decision tree. This indirectly aids the patient's number of diagnostic tests for heart disease prognosis.

VI. CONCLUSION

For the prediction of heart disease, the random forest data mining technique proved as a better approach.

The Sensitivity value of 85.8 percent was derived through the experimental study. For prediction, the specificity is 82.3 and the accuracy is 83.70. Using the random forest method, we were able to predict heart disease with an accuracy of 83.70% in the planned study.

We compared Random Forest method to Decision Trees Algorithm

For successful categorization of cardiac disease, Random Forest technique beats standard classification algorithms.

This sort of study may be utilised to successfully forecast heart disease risk factors and to assist health care practitioners in heart disease prediction.

REFERENCES

[1] Wu R, Peters W, Morgan MW. The next generation clinical decision support: linking evidence to best practice. J Healthc Inf Manag, 2002; 16:50-5.

[2] M.A.Jabbar,B.L.Deekshatulu & Priti Chandra. Intelligent heart disease prediction system using random forest and evolutionary approach,

[3] Rajkumar A, Reena GS. Diagnosis of heart disease using datamining algorithm. Global Journal of Computer Science and Technology 2010; 10:38-43.

[4] Anbarasi M, Anupriya E, Iyengar NCHSN. Enhanced prediction of heart Disease with feature subset selection 2010.

[5] Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. International Journal of Computer Science and Network Security 2008; 8:343-50.

[6] J. Han, M. Kamber, Data Mining: Concepts and Techniques, 2nd Edition, Morgan Kaufmann, 2006.

[7] T. Mitchell, Machine Learning, McGraw Hill, 1997

[8] Saaol times, Monthly magazine" Modifiable risk factors of heart disease", pp 6-10, July (2015)

[9] Khan MG,"Heart disease diagnosis and therapy", a practical approach,2nd Edition Springer,pp544(2015)

[10] M.A.Jabbar,B L Deekshatulu,Priti Chandra ,"classification of heart disease using artificial neural network and feature subset selection",GJCST,Vol13, issue 3,2013

[11] Madhumita Pal and Smita Parija 2021 J. Phys.: Conf. Ser. 1817 012009

[12] Atul Kumar Pandey ,Prabhat Pandey ,K.L. Jaiswal ,Ashish Kumar Sen A Heart Disease Prediction Model using Decision Tree

[13] Sellappan Palaniappan Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques. August 2008.

[14] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). "Effective Heart Disease Prediction using Hybrid Machine Learning Techniques". IEEE Access, 1–1. doi:10.1109/access.2019. 2923707