International Journal of Research in Engineering, Science and Management
Volume-3, Issue-6, June-2020
www.ijresm.com | ISSN (Online): 2581-5792

Application of Regression Tree in Detecting Heart Abnormalities

Aishwarya Shukla[1], Nikita Chaudhary[2*], Mahak Saxena[3]
[1,2,3]*Student, Department of Computer Science and Engineering, Galgotias College of Engineering & Technology, Greater Noida, India*
*Corresponding author: nktchaudhary29@gmail.com

***Abstract*: The world is changing rapidly. Twenty-five years ago, no one could have even imagined of the world in which we are living today. Indeed, this new world is amazing and easier to live but all these things have also paid their part in causing stress in human life by taking works or jobs done by a human and giving it to a machine. This led human towards diseases. Heart diseases are the most brutal one among all the diseases which call human body home sweet-home. So, this project aims to develop such a system which takes common human characteristics such as blood pressure, oxygen saturation, ECG, heart beat rate etc. and determines if any abnormal activity is going on in that heart. The research paper works to solve above discussed problem and it uses CART algorithm to do so.**

***Keywords*: Cardiovascular Diseases (CVD), Heart attack, Stroke, Arrhythmia, ECG, Regression tree.**

1. Introduction

In today's modern world one thing that people do not have is time. Everyone is living such a fast pace life that they barely can take out some time from their busy schedule to spend it with their loved ones or family. Even food and cuisines have adopted this culture by giving birth to a food category call fast-food or instant food. People hardly get any time exercise or medicate they just simply eat sand continue with their loves.

All this misbalance and toxic life style leads to disease and health issues. CVDs or heart diseases are the most brutal and dangerous one among all of them because they are hard to detect. Finding about them at last stage or past the early stage is not very helpful. Sometimes these diseases get confused with something completely different and patient ends up getting wrong diagnosis which wastes both money time and puts patient's life into danger. So, it's very important to find out about diseases at an early stage and be very clear about that problem is something related with the heart making sure that there is hope for the patient with real and correct causes in hand.

Traditional methods of finding about heart disease are fine but they do take time to process all the information and produce result. This research paper aims at developing such methods which takes most simple and common human characteristics and predict if patient's heart is going through some unexpected activities making it crystal clear that problem is related with heart. So that further tests can be made to find out exact cause and then its solution.

This is paper mostly uses CART algorithm to perform the task. This algorithm is taken and then modified according to the need of project and trained over hundreds of test data set. Some other algorithms like REPTREE, J48 this project also taken into consideration but they were not he most suitable one for this project.

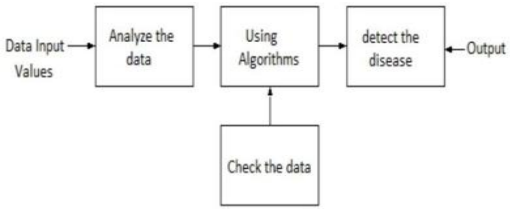


Fig. 1. Block diagram

Also the concept of machine learning is also used in this paper as we know that all these methods and models are used to train the machine and to make it performable without much interaction of the user. Data mining techniques are also used so that the best of the best recommendation can be given.

2. Literature Survey

There are various research works that have been explored to find computational methods to predict heart diseases and choose our method.

1. A. Sahaya Arthy et. al. (2018) Survey on Heart Disease Prediction used the techniques of data mining are Association, Classification, Prediction, Ontology. In this survey paper, analyzed various existing techniques used for predicting heart disease with the support of data mining and ontology methods. From this survey we gained the knowledge of how to apply data mining techniques to predict the heart disease. Previously existing system was designed with single algorithm which has not provided results with better accuracy.

2. Animesh Hazra et. al. (2017), discussed in detail the cardiovascular disease and different symptoms of heart attack. The different types of classification and clustering algorithms and tools were used.

3. Sharan Monica L et. al. (2016) surveyed current techniques of knowledge discovery in databases using data

mining techniques such as J48, NB Tree and simple CART to predict heart disease more accurately with reduced number of attributes in the WEKA tool. J48, is an open source Java implementation of the C4.5 which uses information gain to take decisions. Naive Bayes classifier creates models with predictive capabilities, preferably for continuous dataset. Classification and Regression Trees (CART) is used to display important data relationships very quickly. These three Decision tree algorithms were applied in WEKA. J48 was the quickest to be built (0.08 sec) whereas CART gave the highest accuracy of 92.2%.

4. S. Florence et. al. (2014) proposed a system which uses neural network and the Decision tree (ID3) for the prediction of heart attacks. The dataset used is provided by the UCI machine learning repository. CART, ID3, C4.5 Decision tree algorithms used Gini index to measure the impurity of a partition or set of training attributes. The dataset contains six attributes like age, sex, cardiac duration, signal, possibility of attack etc. The final outcome is the class label. Depending upon the attribute values present in the dataset, the corresponding class label is predicted. 75% of the data is used for training and 25% is used for testing the system.

The knowledge obtained from the classification is used to test the system. In the neural network, the input layer has 6 nodes, the hidden layer has 3 nodes and the output layer consists of 2 nodes. Finally, it shows 2 outputs, that is the possibility of heart attacks. The prediction is done using the tool called RapidMiner Studio. Results are generated by using Decision tree as well as neural networks. They have used this method to predict whether there is an attack or not.

3. Proposed Work

This project uses patient data to decide whether a patient will have any heart abnormalities in future or based on recent diagnosis reports by comparing the past diagnosed reports and patterns of the patient that were diagnosed with cardiovascular diseases or heart abnormalities

According to recent statistics and reports by World Health Organization eight out of ten people are diagnosed with heart diseases and in future it is estimated that the number will increase tremendously.

This project mainly focusses on Arrhythmias which is caused by irregularities in heart rhythm. Many factors can affect your heart's rhythm, such as having had a heart attack, smoking congenial heart defects, and stress. Some substances or medicines may also cause arrhythmias.

Symptoms of arrhythmias include

- Fast or slow heart beat
- Skipping beats
- Lightheadedness or dizziness
- Chest pain
- Shortness of breath
- Sweating

A. Algorithms

There are couple of algorithms there to build a decision tree, lets discuss about some of them.

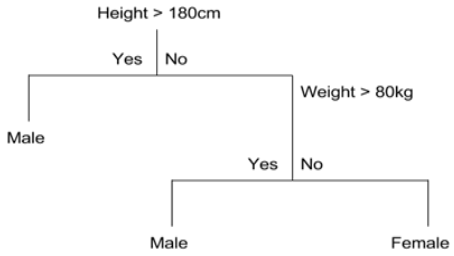
- Classification and Regression Trees (CART): Uses Gini Index (Classification) as metric.
- Iterative Dichotomiser3 (ID3): uses Entropy function and Information gain as metrics.
- C4.5 Algorithm

B. Classification and Regression Tree

The representation for the CART model is a binary tree. This is your binary tree from algorithms and data structures, nothing too fancy. Each root node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric).

The leaf nodes of the tree contain an output variable (y) which is used to make a prediction.

Given a dataset with two inputs (x) of height in centimeter and weight in kilograms the output of sex as male or female, below is a crude example of a binary decision tree (completely fictitious for demonstration purposes only).



C. Gini Index

For attribute selection we use Gini index method or Information gain method.

Gini index or Gini coefficient is a statistical measure of distribution developed by the Italian statistician Corrado Gini in 1912. It is often used as a gauge of inequality, measuring income distribution or, less commonly, wealth distribution among a data set.

In this out of different attributes we choose the best attribute on which we further built three. It measures the impurity(inequality).

Gini index is used as filter to reduce the no. of candidate item sets. It is used to select the best attribute. Those attributes with minimum Gini index are selected for rule generation.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Where Pi=Count of specific class level/total count of attribute in whole data set.

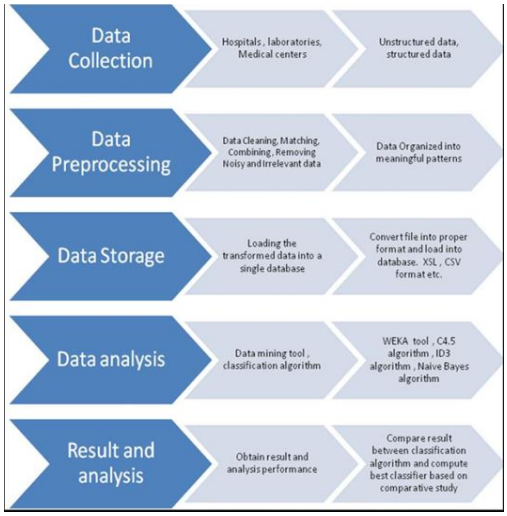
It performs binary splitting (i.e. yes /no).

D. Iterative Dichotomiser3

ID3 is an algorithm developed by Ross Quinlan used to

generate a decision tree from a dataset. To construct a decision tree, ID3 uses a top-down, greedy search through the given sets, where each attribute at every tree node is tested to select the attribute that is best for classification of a given set. Therefore, the attribute with the highest information gain can be selected as the test attribute of the current node. ID3 is based on Occam's razor. In this algorithm, small decision trees are preferred over the larger ones. However, it does not always construct the smallest tree and is, therefore, a heuristic algorithm. For building a decision tree model, ID3 only accepts categorical attributes. Accurate results are not given by ID3 when there is noise and when it is serially implemented. Therefore, data is preprocessed before constructing a decision tree. For constructing a decision tree information gain is calculated for each and every attribute and attribute with the highest information gain becomes the root node. The rest possible values are denoted by arcs. After that, all the outcome instances that are possible are examined whether they belong to the same class or not. For the instances of the same class, a single name class is used to denote otherwise the instances are classified on the basis of splitting attribute.

E. Data Set




```
1.Load the data
2.Initialize K to your chosen number of neighbors
3. For each example in the data
      3.1 Calculate the distance between the query example
          andthe current example from the data.
      3.2 Add the distance and the index of the
          example to an ordered collection
4. Sort the ordered collection of distances and indices
   from smallest to largest (in ascending order) by the
   distances
5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries
7. If regression, return the mean of the K labels
8. If classification, return the mode of the K labels
```

4. Implementation

The implementation consists of many sub sections which are a sequence of steps to be followed to solve any problem. These are:

1. Collection of Data
2. Preparation of Data
3. Data Storage
4. Data Analysis

A. Collection of Data

The data sets which are used to test and train our algorithms. As this project focus on major human traits which all have their part in defining the human heart condition, fall and rise in those trait's value can affect the human heart majorly either in a positive or negative way.

Out of all the human traits that can adversely affect human heart here we collect only those which have major contribution in heart related diseases. The data used in this project is taken from Kaggale Repository.

B. Preparation of Data

The Second step consists transforming and preprocessing of data. Data collected from healthcare organizations obtained into one form understandable by data mining tool.

Data comes from different resources each with its own different form. This different form of data needs transformation and preprocessing. This transformation and preprocessing consist following steps, Data Cleaning, Matching, Combining, Removing noisy and relevant Data.

C. Data Storage

The third step consists data storage process. In data storage, transformed data is stored into a one database with the same format.

D. Data Analysis

After data storage, next step is data analysis. Data analysis is most important phase in this proposed model. It encloses following procedure: First, it includes applying data mining methods or classification algorithms on patients' data.

5. Result

The results show that predictions made by applying this algorithm have a very high accuracy rate. This study will be also being helpful for those who further want to explore this topic. Our method is able to predict chances of a heart disease being present for given data. Below is the table which shows probability of result being true positive, false positive, true negative and false negative.

Accuracy: 76.92%

	True yes	True No	Class precision
Prediction Yes	81	24	77.14%
Prediction No	18	59	76.62%
Class recall	81.82%	71.08%	

6. Conclusion

Many medical techniques are out there to determine and help cure heart diseases. Some are time consuming and some are not.

With each technique are associated their advantages and disadvantages. In this research paper we investigate experiments conducted on heart diseases determination to find out which one works better.

The results show that predictions made by applying this algorithm have a very high accuracy rate. This study will be also be helpful for those who further want to explore this topic.

References

[1] Animesh Hazra, et al., "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques" Advances in Computational Sciences and Technology, Volume 10, Number 7 pp. 2137-2159, 2017.

[2] Sharan Monica.L et al., "Analysis of Cardio Vasular Disease Prediction using Data Mining Techniques", International Journal of Modern Computer Science, vol. 4, no. 1, pp. 55-58, 2016.

[3] S. Florence, N. G. Bhuvaneswari Amma, G. Annapoorani, and K. Malathi, "Predicting The Risk of Heart Attacks using Neural Network and Decision Tree", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 11, pp. 7025-7028, 2014.

[4] Lakshmishree et. al., "Prediction of Heart Disease Based on Decision Trees," International Journal for Research in Applied Science & Engineering Technology, 2017.

[5] Abdar, Moloud, et al., "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases," International Journal of Electrical and Computer Engineering, vol. 5, no. 6, pp. 1569-1576, 2015.

[6] Ashwini Shetty A, and Chandra Naik, "Different Data Mining Approaches for Predicting Heart Disease", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, Special Issue 9, pp. 277-281, May 2016.