Prediction of Heart Disease Using Decision Tree

Mrs. Mehdi Khundmir Iliyas  
Faculty, MCA Department  
Allana Institute of Management Sciences Pune.  
mehdiaims@gmail.com

Mr. Imran Sadekh Shaikh  
MCA Student  
Allana Institute of Management Sciences Pune  
imraninfo.1996@gmail.com

Abstract –

**Purpose**— Prediction of heart disease at the early stage may reduce death ratio to some extent. This software helps prediction of heart disease at an early stage. Now a days healthcare organizations generates huge data but that are highly unorganized. If this data is organized in a proper way using data mining technique it can be easily use for the prediction of heart diseases.

**Objective**— To develop a heart disease prediction system using Decision Tree using J-48 algorithm with two method i.e Cross fold validation and Percentage Split for prediction and implementation.

**Design/Methodology/Approach**— In this paper we have taken Cleveland data from UCI repository. It consist of 303 records. A visualization of Heart disease is shown Using Power BI Dashboard. Where percentagewise male, female, age group, cholesterol level is shown for Heart disease. And developed a heart disease prediction system using Decision Tree using J-48 algorithm with different method for prediction and implementation.

**Findings**— The cause of heart attacks and strokes are usually heart disease level, chest pain, Restecg, Oldpeak etc. it is shown by the decision tree. As well as the cause of heart attacks and strokes are usually due to following risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol, hypertension, diabetes and hyperlipidemia.

**KEYWORDS**

Data mining, Heart Disease, CVDs, Decision Tree (J-48), Heart Disease Prediction System, Health Care Organization etc.

**INTRODUCTION**

Cardiovascular diseases are the main source of death all inclusive, a bigger number of individuals pass on every year from CVDs than from some other reason. An expected 17.9 million individuals died from CVDs in 2016, speaking to 31% of every worldwide death, out of these deaths, 85% are due to heart attack and stroke [1].

More than seventy five percent of CVD passings occur in low and center salary nations. Over three quarters of CVD deaths take place in low- and middle-income countries because they do not have the benefit of health care programs for early detection and treatment as compared to people in high-income countries as a result people die younger from CVDs and from other non-communicable diseases in their most productive years.

Peoples in low and middle income nations regularly don't have the advantage of social insurance programs for early discovery and treatment when contrasted with individuals in high-pay nations. Accordingly, numerous individuals in low-and center pay nations are recognized late over the span of the sickness and die younger from CVDs and other illnesses, frequently in their most gainful years.

Most cardiovascular infections can be avoided by tending to social hazard factors, for example,
Early recognition is important to recognize people in danger, accordingly individuals with above age 30 need early identification using counselling and medicines to advance way of life changes before disease progression happens.

This paper intends to make attention to all CVDs risk factors, way of life habits (diet, exercise, and smoking), blood pressure, glucose, and weight list for risk reduction and early identification of CVDs using medical help or our created systems.

The types of Cardiovascular diseases (CVDs) are:
1. Coronary heart disease – disease of the blood vessels supplying the heart muscle.
2. Cerebrovascular disease – disease of the blood vessels supplying the brain.
3. Peripheral arterial disease – disease of blood vessels supplying the arms and legs.
4. Rheumatic heart disease – damage to the heart muscle and heart valves from rheumatic fever, caused by streptococcal bacteria.
6. Deep vein thrombosis and pulmonary embolism – blood clots in the leg veins, which can dislodge and move to the heart and lungs.

Data Mining
Data mining is a ground-breaking method for preparing enormous data sets. It is utilized for computational and finding designs in enormous data sets. It is significant system for getting obscure examples and crucial data from huge data set. The principle objective of the data mining process is to remove concealed data from an enormous data set and change it into significant data for further use.

In data mining commonly four types of relationships are Classification, clustering, Associations, Sequential patterns. Also Different levels of analysis are available such as artificial neural networks, Genetic algorithms, Decision trees, Nave Bayes K-Means, Rule induction, Data visualization.[9]

Decision Tree:
A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision. Decision trees can be drawn by hand or created with a graphics program or specialized software. Informally, decision trees are useful for focusing discussion when a group must make a decision. Programmatically, they can be used to assign monetary/time or other values to possible outcomes so that decisions can be automated [18].

J48 Algorithm:
J48 is an algorithm used to generate a decision tree it’s also known as C4.5. Developed by Ross Quinlan mentioned earlier, C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier [18].

II.OBJECTIVE: Objective of this paper is to develop a system that predicts heart disease.
Prediction of heart disease at the early stage may reduce death ratio to some extent. This software helps prediction of heart disease at an early stage. This system accepts persons data, such as name, Age, Gender, Chest pain, Cholesterol etc. 13 parameters will be given to the system. The system will predict whether the person have heart disease or not. As well as the objective of this paper is to show how many peoples are suffering from heart disease, agewise genderwise we have shown a visualization using Power BI Dashboard that shows percentgewise male, female, age group and cholesterol level for Heart disease. For visualization we have taken data from UCI repository (i.e. The University of California, Irvine). Cleveland data.

III METHODOLOGY: This system has taken Cleveland data for training from UCI repository (i.e. The University of California, Irvine). Cleveland database consist of 303 records. Dashboard is created using Power BI. We uses decision tree for the prediction of heart disease. Cleveland data is provided to decision tree model, then it creates decision tree with different method such as 10 fold Cross Validation and Percentage Split. But the output are same for both method so we use percentage split with cross fold validation 20 and confidence level is 0.25%. and we are going to develop a system using decision tree, this system takes persons data, and predict whether the person have heart disease or not.

IV. IMPLEMENTED SYSTEM
The implemented system is a heart disease prediction system. This system has taken Cleveland data for training from UCI repository (i.e. The University of California, Irvine). Cleveland database consist of 303 records and 76 attributes out of which 14 attributes are considered which are important as follows.
Some Important Terminology:

Thal / Thalach:
Thalassemia is an inherited blood disorder characterized by less hemoglobin and fewer red blood cells in your body than normal.

Ca (Coronary artery):
Coronary arteries are the blood vessels of coronary circulation, which transports oxygenated blood to the actual heart muscle.

Fbs (Fasting blood sugar):
Fasting blood sugar level tests are used to diagnose diabetes.

SYSTEM DESCRIPTION USING DECISION TREE:

This system predicts the heart disease. It uses decision tree for the prediction of heart disease. Decision tree helps to take decision, it provides good decision making technique. It divides the instances in multiple trees. It contains root node, branch node and leaf node.

Cleveland data is provided to decision tree model, then it creates decision tree with different methods such as 10 fold Cross Validation and Percentage Split. But the output are same for both methods so we use percentage split with cross fold validation 20 and confidence level is 0.25%.

Cross Fold Validation

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

Percentage Split

Percentage split: Splits the data and separates x% of the data for learning and the rest of it for testing. Following are the method we used to create a tree.

In both the table correctly classified instance is 67 to 68 Percent, so we choose 68.1319% instance who’s tree size and tree leaves are small and it is easy to implement. The file is loaded in weka and perform this operation the weka give the output as follows in Fig.2. In this the Ca (Coronary artery) node is chosen after the heart disease node because its entropy and information gain is more than other and likewise the other branches are chosen.

Fig.2: Weka decision Tree rule

Here decision tree used for the prediction of heart disease. It contains root node, branch node and leaf node.

Heart disease use as main class variable, on the basis of that the decision tree is made, it contains following attributes, heart disease, Ca, Oldpeak, Trestbps, Chol, Fbs, Resteeg, chest pain, age & slope level.

First node contain heart disease in this data is categorized in two category Yes or No. If the tree says Yes it open another instances on the basis of Ca (Coronary artery), it will take further decision and draw the tree. In this we used test mode as cross fold validation 20 with confidence level is 25%.
After generating the tree it gives 70, 30 instances. 70 as correct classified and 30 is incorrect classified. It means that the module is good.

Fig.3: Decision Tree

**Entropy:**
Entropy is the measures of impurity disorder or uncertainty in a bunch of example. Entropy controls how a Decision Tree decides to split the data. It actually effects how a Decision Tree draws its boundaries.

\[ \text{Entropy} = - \sum p(X) \log p(X) \]

**V. Dashboard using Power BI:**

Using Cleveland database of 303 records Dashboard is created in Power BI. The main attribute for dashboard is: Heart disease, gender, age group, heart disease level, chest pain, Restecg, Oldpeak.

Out of 303 records, 45.21% people have a heart disease and 54.79 have no heart disease. In this the level of their disease is as follows- Critical 4.29%, High level 11.55%, Medium 11.56%, Low 17.82%. The low level and medium level are on the top level.

Here 8.25% are female and 36.96% are male, it seems that the majority of males are more than females.

The scatter plot diagram show the cholesterol wise Oldpeak. On ‘X’ axis the Oldpeak level is putted and on ‘Y’ axis the cholesterol level are putted.

In the last visualization the 100% stacked bar chart is used in gives the row wise and column wise data facility in both view we can visualize our data in this we put chest pain and the Restecg the chest pain such as asymptomatic and typical angina are more than other. Likewise the Restecg level definite and normal are highest value than other.

Fig.5 it showing the relation of age group and heart disease. After the age 40 the heart disease ratio is more. In fact, the risk of heart disease rises steadily and sharply with age. After 40 it might be a chance to high blood pressure, overweight, high cholesterol and diabetes related problem occur during this age or after this age.

After the age group 51 and above the chances are more to occur heart disease because of less attention towards their life and less focus towards their health. The men get highest heart attack at minimum age for men is 45 and for women is 55 at this age the platelets in the blood are especially "sticky" and prone to form clots.

**VI. RESULT DISCUSSION:**

1. No – person have No heart disease.
2. Low - Person have low level of heart disease.
3. Medium - person have middle level of heart disease.
4. High - Person have high level of heart disease.
5. Critical - Person have critical level of heart disease.
VII. CONCLUSION

This system successfully Predict the heart disease of a person. And the system is successfully developed using Decision tree using J48 Algorithm the develop system pull out the knowledge from historical database made by medical practitioner. In first step training is given to the system then prediction take place. The accuracy of this system is 68% true, and 32% false with the size of tree is 25 and leaves are 15 it’s the smallest one as compare to other method. As this approach gives 68% accuracy only. In future we try to increase the accuracy level of the system.

VIII. REFERENCES

[1] https://www.who.int/health-topics/cardiovascular-diseases/
[12] https://sourceforge.net/projects/weka/ 