Human Heart Disease Prediction System Using Data Mining Techniques

Sidra Javed[1], Hamza Javed[2], Ayesha Saddique[3], Beenish Rafiq[4]

(Received October 03, 2018; Revised November 30, 2018; Accepted December 13, 2018)

DOI: 10.33317/SSURJ.V8.I2.74

***Abstract*— Prediction of heart disease is a big concern now a days because everyone is busy and due to heavy load of work people do not give attention to their health. To diagnose a disease is a big challenge. The issue is to extract data that have some meaningful knowledge. For this purpose, data mining techniques are used to extract meaningful data. Decision Tree and ID3 are used to predict heart diseases. Many researchers and practitioners are familiar with prediction of heart diseases and wide range of techniquesis available to predict disease. To address this problem, DecisionTree is used to predict the heart disease. In this study the collected data is pre-processed, Decision Tree algorithm and ID3 were then applied to predict the heart disease.**

***Index Terms*— Decision Tree, ID3 Algorithm, Data Mining, Decision Support System (DSS), knowledge Discovery from Databases (KDD).**

I. INTRODUCTION

A. *Basic Knowledge of Human Heart*

The heart is an imperative part or an organ of the human body. Life cannot exist without proper functioning of the heart also it will impact the other body parts of humans, for instance, mind, kidney, and so on. The heart is a pump, which circulates the blood within the body. If blood circulation in the body is affected, then numerous organs like cerebrum will suffer, likewise if heart stops working, death occurs within minutes. Life is absolutely subject to compelling working of the heart. The term ‘Heart Disease’, insinuates sickness of heart and vessel structure inside it.

There are number of components which fabricate the threat of heart disease:

- Family history of coronary disease.
- Smoking.
- Poor eating philosophy.
- High heartbeat.
- Cholesterol.
- High blood cholesterol.
- Obesity.
- Physical idleness.

There are some indications of a heart ambush which can include:

- Discomfort, weight, vastness or desolation in the midriff, arm or underneath the breastbone.
- Discomfort radiating to the back, jaw, throat or arm.
- Fullness, acid reflux or smothering feeling (may feel like heartburn).
- Sweating, nausea, hurling or shakiness.
- Extreme deficiency, anxiety or shortness of breath.
- Rapid or irregular heart beats.

B. *Overview of Data Mining*

Because of immense amount of information accessibility and the need to change over that information into valuable learning, Data Mining Systems can be helpful. As of late, data mining has discovered its importance in relatively every field including medicinal services. The plenitude of information and the need of effective investigation instrument for that information, portrayed as "information rich yet data poor" circumstance, as shown in Fig. 1.

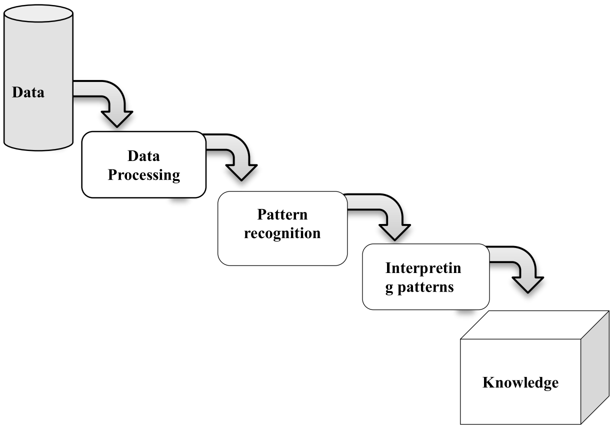


Fig. 1: Knowledge Flow Diagram

The extensive dataset, move toward becoming information tombs. Information mining has some different names, e.g, learning mining from databases, design investigation, information extraction, information digging and information prehistoric studies. Mining process is more than the information investigation it incorporates characterization, grouping and affiliation lead revelation, etc.

The sympathetic or parasympathetic is a critical situation which occurs when a nerve of nervous system fails. It can effect on different parts of body like blood pressure, breathing problem and some others. Doctors can check the regulation of blood pressure and heart beat by taking four simple, basic and effective tests which consist of:

[1] Student of MSCS, Department of Computer Science, University of Agriculture, Faisalabad, Pakistan. sidra.javed1@yahoo.com
[2] Student of BSCS, Department of Computer Science, University of Agriculture, Faisalabad, Pakistan. hamza_javaid327@yahoo.com
[3] Student of MSCS, Department of Computer Science, University of Agriculture, Faisalabad, Pakistan. ashiayesha11287@gmail.com
[4] Student of MSCS, Department of Computer Science, University of Agriculture, Faisalabad, Pakistan. beenishrafiq45@yahoo.com

- Firstly, deep breathing test; in which a patient breathes as fast as he or she can breathe for two minutes.
- Secondly, hand claps test; where patients are requested to keep up hand clap for 3 minutes.
- Third one, mind tension test; in which some mind calculations are performed by patient.
- Last and fourth test is the position or posture of upright standing observation, in which some time fainting, decrease in blood pressure and increase in pulse rate occur of a patient.

Data mining provides a set of tools and methods for withdrawal of useful and important information for classification purposes like Rule Based Classifier, Decision Tree Induction Nearest Neighbor Classifier, Bayesian Classifier and SupportVector Machine. Classifica tion and association are among the main tasks of data mining. The purpose of classification techniques is to process a large amount of data and predict categorical class labels. In this a medical record is carried out to make a decision based system of diagnosis and treatment of a patient [1].

In present day, computers and information technology play an important role in our daily life. Information technology with advanced computer programming is making some new frameworks. There are many challenges and tasks that are faced by IT specialists. One of them is to make an algorithm or a scenario according to health care to predict heart disease. For this purpose the patient's data are collected and stored and then different data mining techniques are applied to attain or predict heart disease by these data. There is a big issue of data quality. To overcome this issue, the stored data is used to predict heart problem and according to these predictions some clinical care is given to patient. Due to computer based data there is less chance of errors, increase safety level of patients, decreasing undesirable variety and increase patient health as a result. This is all due to data mining techniques and machine learning that are used to predict data. Data mining is an important step of Knowledge Discovery from Databases (KDD), which consist of an iterative sequence of data cleaning, data integration, data choice, data mining and pattern recognition and data presentation [2].

II. LITERATURE REVIEW

Data mining predicts the future by comparing past and future results and plays an important role in analyzing the data. For this purpose many big and well developed companies stored and maintain data for many years and then compare these data for the knowledge of business. To analyze this information, data mining techniques are used like Naïve Bayes, Decision Tree, KNN, ANN, SVM, WAC and MLP [3].

The heart disease is the largest cause of death. To predict heart disease a better Decision Support System is developed by using Naïve Bayes and Jelinek Mercer that are data mining classifiers. An intelligent hybrid algorithm is developed for better performance to predict heart disease. To predict heart disease different data mining methods are used with a limited number of characters [6]. Different data mining methods like J48, Naive Bayes, Bayes Net, REPTree and SimpleCart, performed excellent, at the chosen number of characters. But, KStar, J48, SMO, Bayes Net and MLP that are data mining methods are contrast.

There are numerous reasons that cause coronary illness like weight, age and cholesterol and so on. To coronary illness numerous information mining techniques are utilized like Decision tree, Naïve Bayes and KNN. Fluffy manage based framework is likewise used to anticipate coronary illness. In this framework, malady is anticipated by the patient's record. Neural System give better outcome on fifteen characters and Decision Tree likewise give amazing outcome with hereditary calculation and highlight subset choice. A prior information mining strategy is utilized to demonstrate information graphically. That can be extremely useful, if all things are considered [5].

In UK, there are some issues of white and black people. So, keeping in mind of that a research was done to predict heart disease. Some characters are identified that effect on heart and cause of heart disease by using digital device like mobile. The research was completed to know how much mobile app is sufficient and effective to predict heart disease and this way diabetes is also checked out [6], [7].

III. MATERIALS AND METHODS

This research is based on prediction of heart disorder. To expect heart disorder Decision Tree and ID3 are used. ID3 is a precursor of C4.5 algorithm that is used in Decision Tree. The Decision Tree is a choice aid device that makes use of a tree like graph or model of choices and their feasible outcomes, together with danger event outcomes, resource fees, and applications. It's far one manner to display a set of rules that best incorporates conditional manipulate statements, as seen in Fig. 2.

First of all, the records are accumulated from hospitals. The Decision Tree is carried out on those facts to expect heart ailment. Many specific attributes are created in step with heart disease that have an effect on working of coronary heart like a chest ache, blood strain, age, gender, cholesterol level , family records and respiration. On the foundation of these one-of-a-kind attributes, heart ailment is expected.

The usual manner of predicting the error rate of a mastering approach given a set pattern of information is through the use of stratified tenfold cross-validation. Sizable assessments on numerous one-of-a-kind datasets, with one of a kind learning technique, have proven that ten is about the right variety of folds to get the first-class estimate of errors. To be able to degree the steadiness of the proposed model, the information is divided into education and checking out information with 10-fold pass validation to evaluate the accuracy of our gaining knowledge of version. In this case, we will divide the dataset into 10. The algorithm requires submission of data in a specified format. The conversion of raw data into machine understandable format is called, 'Preprocessing'. The data preparation phase covers all activities to construct the final dataset from the initial raw data. These raw data can be stored in several formats including text, excel or other database types of files. Then the raw data is changed into data sets with a few appropriate characteristics. Values are absent in most datasets with many causes contributing towards it. The raw data usually have a great deal of noise, which is a random error or variance in a measured variable. It cannot be used directly for processing, with the machine learning algorithms. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Its routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Cleaning and filtering of the data have to be necessarily carried out with respect to the data in data mining algorithm to avoid the creation of deceptive or inappropriate rules or patterns. To make the data appropriate for the mining process, it needs to be transformed. Data integration merges data from multiple sources into a coherent data store, like a Data Warehouse or a Data Cube, see Fig. 3.

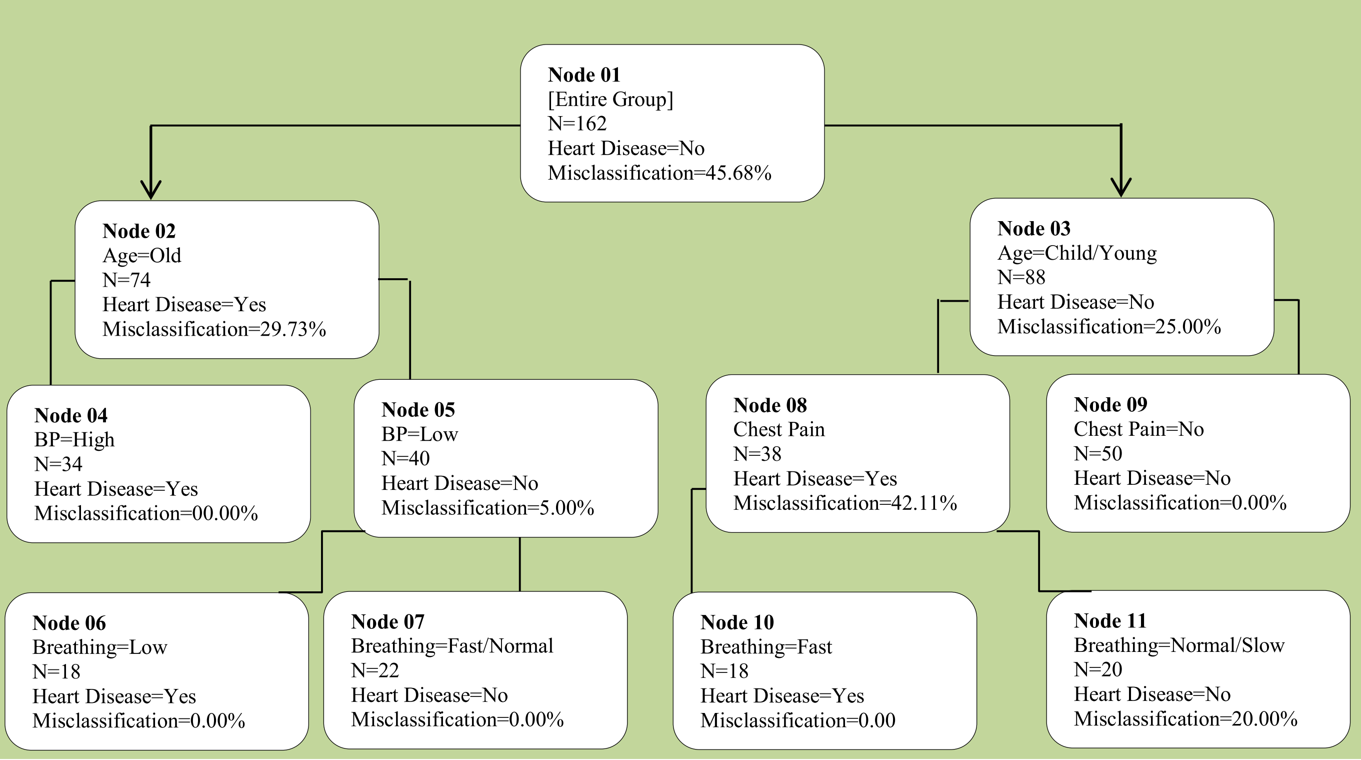
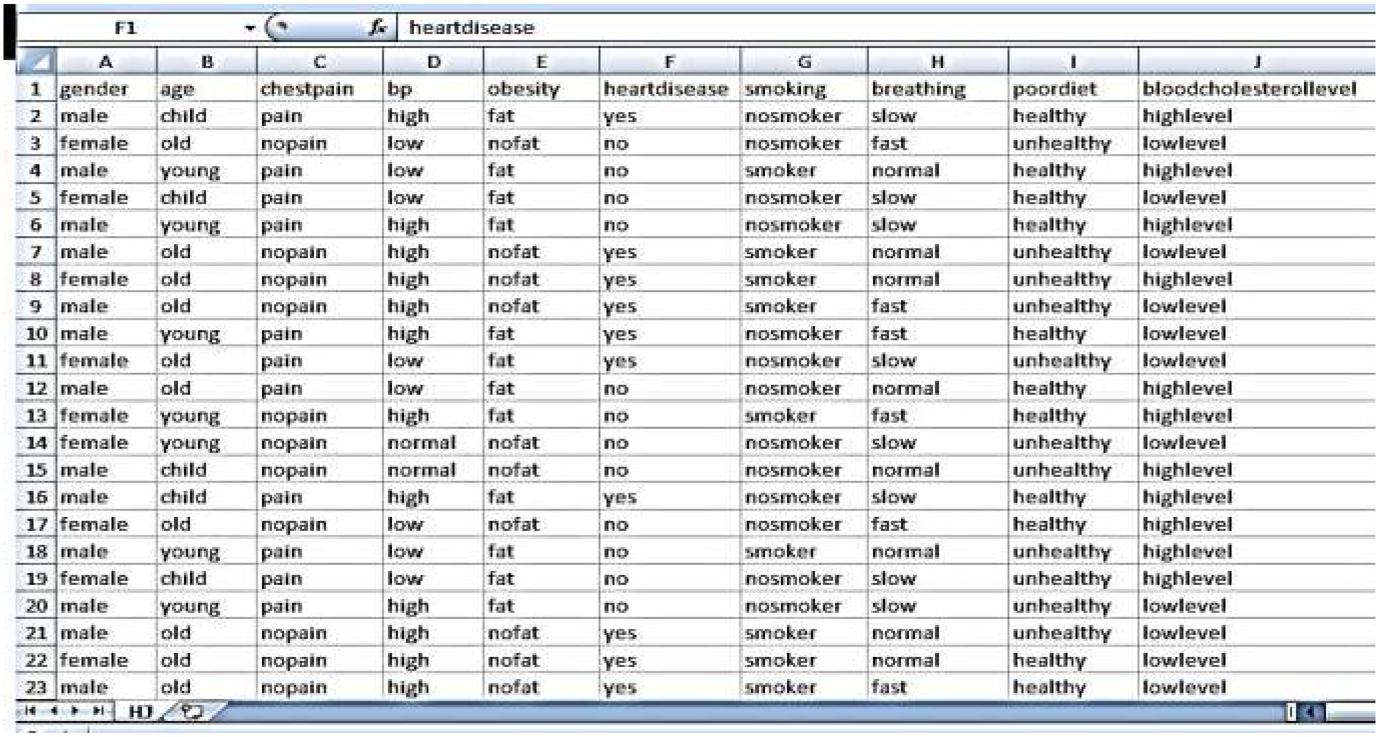


Fig. 2: Overview of Decision Tree



	A	B	C	D	E	F	G	H	I	J
1	gender	age	chestpain	bp	obesity	heartdisease	smoking	breathing	poordiet	bloodcholesterollevel
2	male	child	pain	high	fat	yes	nosmoker	slow	healthy	highlevel
3	female	old	nopain	low	nofat	no	nosmoker	fast	unhealthy	lowlevel
4	male	young	pain	low	fat	no	smoker	normal	healthy	highlevel
5	female	child	pain	low	fat	no	nosmoker	slow	healthy	lowlevel
6	male	young	pain	high	fat	no	nosmoker	slow	healthy	highlevel
7	male	old	nopain	high	nofat	yes	smoker	normal	unhealthy	lowlevel
8	female	old	nopain	high	nofat	yes	smoker	normal	unhealthy	highlevel
9	male	old	nopain	high	nofat	yes	smoker	fast	unhealthy	lowlevel
10	male	young	pain	high	fat	yes	nosmoker	fast	healthy	lowlevel
11	female	old	pain	low	fat	yes	nosmoker	slow	unhealthy	lowlevel
12	male	old	pain	low	fat	no	nosmoker	normal	healthy	highlevel
13	female	young	nopain	high	fat	no	smoker	fast	healthy	highlevel
14	female	young	nopain	normal	nofat	no	nosmoker	slow	unhealthy	lowlevel
15	male	child	nopain	normal	nofat	no	nosmoker	normal	unhealthy	highlevel
16	male	child	pain	high	fat	yes	nosmoker	slow	healthy	highlevel
17	female	old	nopain	low	nofat	no	nosmoker	fast	healthy	highlevel
18	male	young	pain	low	fat	no	smoker	normal	unhealthy	highlevel
19	female	child	pain	low	fat	no	nosmoker	slow	unhealthy	highlevel
20	male	young	pain	high	fat	no	nosmoker	slow	unhealthy	lowlevel
21	male	old	nopain	high	nofat	yes	smoker	normal	unhealthy	lowlevel
22	female	old	nopain	high	nofat	yes	smoker	normal	healthy	lowlevel
23	male	old	nopain	high	nofat	yes	smoker	fast	healthy	lowlevel

Fig. 3: Sample Data in Excel Sheet

Data reduction can reduce the data size by aggregating and eliminating redundant features. Using the data mining techniques, the focus is on specific fields that allow exploration of the data, by selecting and filtering some fields as input, output fields and predictive fields. Out of the 535 records, 400 related to patients highly vulnerable to heart diseases, and the remaining 135 patients found less vulnerable to heart disease.

The Fig. 4 shows the ROC (Receiver Operating Characteristic) graph line of true positive rate and false positive rate. The upper area of the graph line shows the value of ROC of heart disease that is yes.

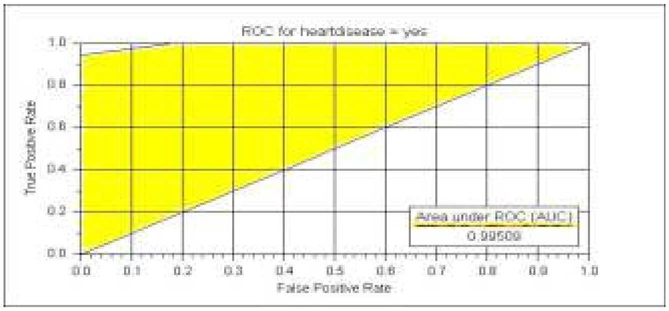


Fig. 4: ROC of Heart Disease

The Fig. 5 show s the graph line of number terminal node and misclassification cost of variables.

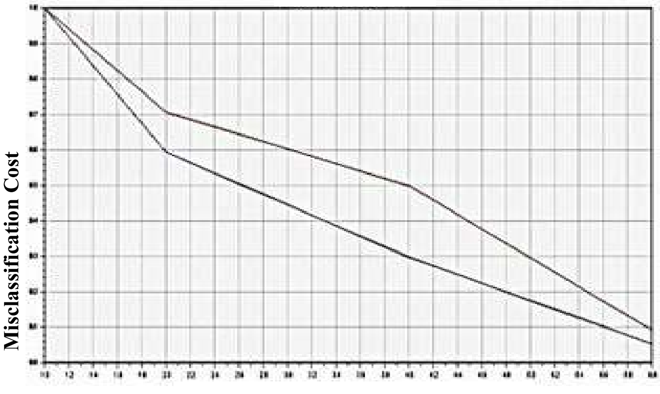


Fig. 5: Model Size and Error Rate

IV. RESULTS AND DISCUSSIONS

The model using the chosen data mining methods were created by categorizing the dataset based on certain attribute value pairs. The results of the experiments over the model are summarized. The Decision Tree outperformed the other methods providing the highest classification accuracy. This confirms research question, that the best method is the Decision Tree with the highest classificaiton accuracy. Not only in overall accuracy, even in terms of precision and recall of the classes, Decision Tree, exhibited a very consistent performance in the final results. Different classifierscan be chosen according to the requirements and desires, but in this research, ID3 algorithm is applied to get desired results. ID3 is applied with different tests like cross validation, use training set, suspected data set and percentage split test sets to get desired results.

To evaluate the performance of the algorithms different performance metrics were considered thatAare ccuracy, Precision, F-measure, ROC curve value, TP rate and FP rate. Four experiments were conducted in two different scenarios, in first scenario all attributes were used and in second scenario selected attributes were used and data set was in, Attribute-Relation File Format (ARFF) that is supported by Weka. This study shows that the data mining can be used to predict about heart disease efficiently and effectively. The results or the outcomes of this research may be used as assistant tool to help in making more consistent diagnosis of heart diseases. Decision Tree works best as the numbers of attributes are going to decrease and its accuracy increases with decreasing number of attributes. Also it takes less time to predict the disease. In the future different severe diseases can be predicted like diabetes, blood cancer and brain tumor etc.

V. CONCLUSION

The primary emphasis in this research is on the use of various data mining calculations; in particular, Decision Tree and ID3 on heart dataset to foresee the danger of heart sicknesses in the light of their prescient precision. Subsequently, an examination of the results of the different characterization procedures has been made and a higher level of precision of the decisionrte is found. For future research, stacking procedures can be utilized to build the precision of Decision Trees and diminish the quantity of leaf hubs. The principle center is around the significance of information mining in restorative frameworks.

REFERENCES

[1] Patel, S. et al. (2017). Heart Disease Prediction Using Data Mining. *International Research Journal of Engineering and Technology (IRJET), 4*(1), 1705-1707.

[2] Babu, S. et al., (2017).*Heart Disease Diagnosis Using Data Mining Technique.* Internat ional Conference on Electronics, communication and Aerospace Technology Electronics, communication and Aerospace Technology ICECA, (750 753), Coimbatore : IEEE.

[3] Keerthana, T. K. (2017). Heart Disease Prediction System using Data Mining Method. *Interna tional Journal of Engineering Trends and Technology (IJETT), 47*(6), 361-363.

[4] Idri A. & Kadi, I. (2017). *A. Data Mining- Based Approach For Cardiovascular Dysautonomias Diagnosis and Treatment.* IEEE CIT- 17th. IEEE Inter national Conference on Computer and Information Technology, Helsinki : Finland.

[5] Gandhi, M. & Singh, S. N. (2015). *Predictions in Heart Disease Using Techniques of Data Mining.* International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), (520–525), Greater Noida : IEEE.

[6] Raihan, M. et al., (2016). *Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction Using Clinical Data and Data Mining Approaches, a Prototypes Design*. 19th, International Conference on Computer and Information Technology (ICCIT) , (299–303), Dhaka : IEEE.

[7] Sultana, M., Haider, A. & Uddin, M. S. (2016). *Analysis of data mining techniques for heart disease prediction*. 3rd International Conference on Electrical Engineering and Information and Communication Technology (ICEEICT) , (1-5). IEEE.