# *Usage of Data mining techniques in predicting the Heart diseases – Naïve Bayes & Decision tree*

*Priyanka N B.E., (M.Tech)* [1]
*CS & E dept*
*AIT College,*
*Chikkamagaluru, Karnataka state,India .*
*priyanka.nagraj05@gmail.com*

*Dr.Pushpa RaviKumar B.E., M.Tech., Ph.D* [2]
*CS & E dept*
*AIT College, Chikkamagaluru, India*
*flowersunpr@yahoo.co.in*

***Abstract : Heart disease is the leading cause of death among all other diseases, even cancers. The number of men & women facing heart disease is on a raise each year. This prompts for its early diagnosis & treatment. Due to lack of resources in the medical field, the prediction of heart disease occasionally may be a problem. Utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity & patients. This issue can be resolved by adopting Data mining techniques. This paper intends to adopt Naïve Bayes & Decision tree – two data mining techniques for the effective prediction of Heart disease. It compares the efficiency & accuracy of the two techniques to decide among them the best.***

***Keywords: Heart Disease, Naïve Bayes, Decision tree, Classification.***

## I INTRODUCTION

Today’s era ia witnessing an alarmig ascent in the incidences of numerous life style disorders like Heart disease, Hypertension, Obesity, etc. Among these, Heart disease is on top of the list which is the leading cause of death in most of the countries. It includes diseases of heart muscles, valves, conduction system, heart attack & others. Myocardial infarction or heart attack is the major one among all other types of heart diseases. Heart diseases are seen in all the classes of people in recent times, incontrast to previous days when it was disease of rich class people.

Heart disease is even being highlighted as a silent killer which leads to death of a person without obvious symptoms. This nature of the disease is the cause of growing anxiety about the disease & its consequences. Hence continued efforts are being done to predict the posibility of this deadly disease in prior. So that various tools & techniques are regularly being experimented to suit the present day health needs. Data mining techniques can be a boon in this regard.

Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analyzing to extract the desired data we can arrive at a conclusion. This technique can be very well adapted to the do the prediction of heart disease. As the well known quote says “Prevention is better than cure”, early prediction & its control can be helpful to prevent & decrease the death rates due to heart disease.

### Heart disease

Heart disease is a term used for any type of disorder that affects the heart. Heart

978-1- 5090-4967- 7/17/$31.00 © 2017 IEEE

disease means the same as cardiac disease. Depending upon the pathology occurred, the heart disease are of various forms.

### A. Coronary heart disease

Coronary Artery disease (CAD) is also called by the name Ischemic heart disease (IHD). It comes under cardiovascular diseases which involve the valves of the heart. It comprises of a group of diseases, angina, myocardial infarction & sudden cardiac arrest.

The symptoms include,

- Chest pain / discomfort which radiates to shoulder, neck or back
- Heartburn
- Shortness of breath

### B. Heart attack

Heart attack is the common term used for Myocardial infarction (MI). It is due to the interruption of blood supply to a part of heart leading to damage to the heart muscle.

Chest pain is the common symptom which may pass on to shoulder, neck, back or jaw. It may present in the centre or left side of the chest.

### C. Arrhythmia

Cardiac Arrythmia is also known commonly as irregular heart beat is a group of conditions. Here, the heart beat will be too fast or slow or irregular. Sypmtoms may not be present many a times. Symptoms may include palpitations or feeling of gap between heart beats.

### D. Heart failure

Heart failure is the inability of the heart to pump the blood for the body funtioning. Coronary artery disease, heart attack, valve diease etc. can lead to heart failure. Sypmtoms are difficulty in breathing, fatigue & leg swelling. Chest pain is not present always in heart failure.

## II LITERATURE SURVEY

The paper by Vanishree. K developed a system for diagnosis of congenital heart disease using decision support system. It used Back propagation Neural Network with MLP. It was based on the data – signs of the heart disease, symptoms & result of evaluation obtained from the patient. It showed 90 % accurate results [3].

The study done by Kharya highlighted the fact that artificial neural network is the frequently used technique for prediction in the medical field. The paper also demonstrated the merits & de merits of the machine learning techniques like Decision tree, Naïve Bayes, Neural network & SVM [4].

C.D. Katsis et al study devised ways using Correlation Feature Selection (CFS) procedure & an Artificial Immune Recognition System (AIRS) classifier to diagnose breast cancer. Data for the study was collected from 53 patients among the 4726 cases. Biopsy was taken in all the patients & it was kept as standard parameter to validate the methodology. The features along with the biopsy result were used for the analysis in 53 patients. Adoption of SVM technique resulted in 70.00 + 6.33% accurate result [5].

## III METHODOLOGY

This paper aims to predict the Heart disease based on the data set values stored by using the MSQL database as backend server.

Using this training data set values it is possible to predict the presence or absence of Heart disease in that particular patient record. Here classification techniques like Naïve Bayes & Decision tree algorithm were utilized. It is a web based application. In the front end .NET framework which acts as client. This paper even compared the accuracy of results of prediction obtained between the two algorithms.

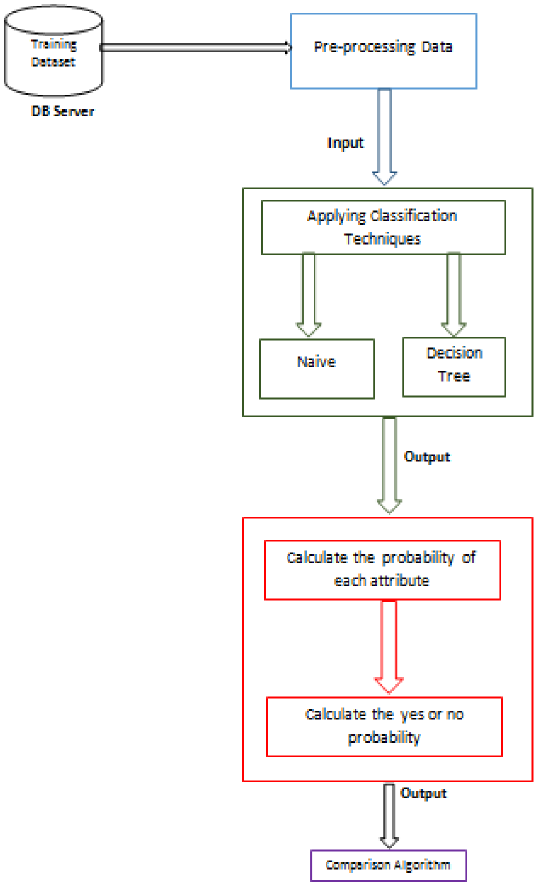

Training Dataset
DB Server
Pre-processing Data
Input
Applying Classification Techniques
Naive
Decision Tree
Output
Calculate the probability of each attribute
Calculate the yes or no probability
Output
Comparison Algorithm

**Fig1: Architecture of the proposed model**

The Fig 1 depicts the architecture of the proposed model used in the prediction of heart disease. It consists of 3 steps,

### A. Data base server

The data set was collected from UCI laboratory. It contains 13 attributes which include sex, serum cholesterol level, resting ECG etc. This data set was stored using MSQL database. It acts as a DB server (backend server). Connection was established between DB server & client (front end). Queries are posed on the server using this DB & each individual data is retrieved.

### B. Data preprocessing

In this step data is taken out from UCI repository in a recognized format. Missing fields are evacuated in this process & thus the data is transformed. Mean is entered in place of missing attributes.

### C. Classsification

In data mining mainly two methods are used, supervised & unsupervised learning. Supervised learning uses a training set inorder to learn model parameters. But no such training set like k – means custering is utilised in unsupervised learning. Data mining has got two most frequent modelling goals – classification & prediction. Classification model classifies discrete, unordered values or data.

In this prediction process, the classification techniques utilized are,

- Decision tree
- Naïve Bayes

### i. Decision tree

In this type of classification the knowledge is represented in a tree diagram. Schematic representation will be in the form of tree to depict the decisions. Rules are framed through these decisions to classify the data. The data classification is in the form of root nodes to begin with, ending with the terminal node of the attributes. Nodes are having attribute names, positive values are the edges & different classes represent the leaves of the tree. This paper has used the popular ID3 algorithm.

Fig 2 depicts the tree structure of prediction of heart disease which consists of root node & child node attributes of heart disease. Root node contains the maximum entropy value like old peak & the child node values like number of major vessels & chest real.

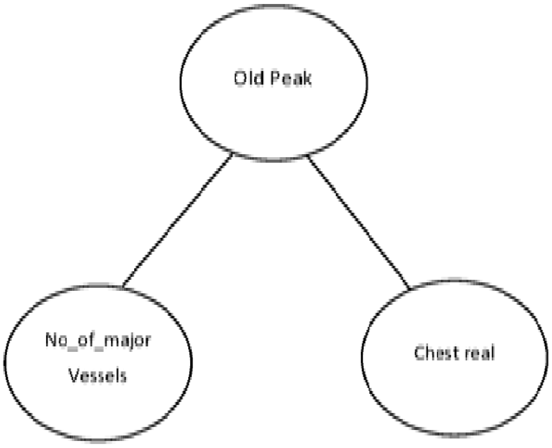
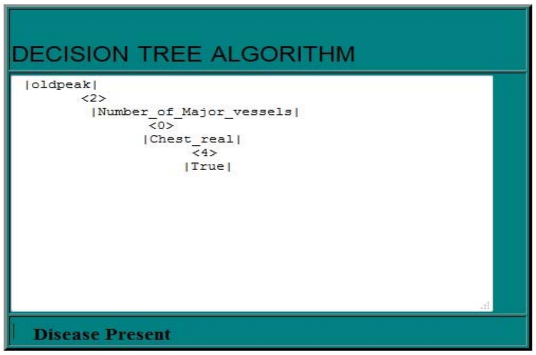


**Fig 2: Tree structure for prediction of disease attributes**

Fig 3 depicts the performance of the decision tree algorithm and it predicts the presence of heart disease based on the 3 attributes like oldpeak, Number_of_major_vessels and chest_real.



DECISION TREE ALGORITHM

```
|oldpeak|
        <2>
         |Number_of_Major_vessels|
                  <0>
                 |Chest_real|
                       <4>
                      |True|
```

Disease Present

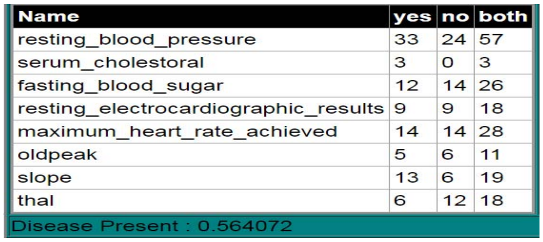
**Fig3: Performance result of decision tree**

### ii. Naïve Bayes

It is one of the popular classification techniques of algorithms used in data mining. It is a probability classifier. It links the attributes mutually & is dependent on the number of parameters. The principle here is that the variables provided are independent. It generates accurate results with appropriate calculation & provides fast results. It is based on Bayes theorem & the formula is,

$$p(z|xc) = \frac{p(xc|z)p(z)}{p(xc)}$$

Fig 4 shows the calculation of the yes/no probability using naïve. It calculates yes or no probability for each individual attributes and finally shows the highest probability attributes and their resultant values. It also predicts whether the disease is present or absent with accuracy value 56%.



| Name | yes | no | both |
|---|---|---|---|
| resting_blood_pressure | 33 | 24 | 57 |
| serum_cholestoral | 3 | 0 | 3 |
| fasting_blood_sugar | 12 | 14 | 26 |
| resting_electrocardiographic_results | 9 | 9 | 18 |
| maximum_heart_rate_achieved | 14 | 14 | 28 |
| oldpeak | 5 | 6 | 11 |
| slope | 13 | 6 | 19 |
| thal | 6 | 12 | 18 |

Disease Present : 0.564072

**Fig 4: Calculation of yes/no probability using Naïve bayes**

Fig 5 shows the graphical representation of the result. The graph is generated by using the lifting chart. X-axis is the number data set values and Y- axis represents the Probability of yes or no with the percentage values.

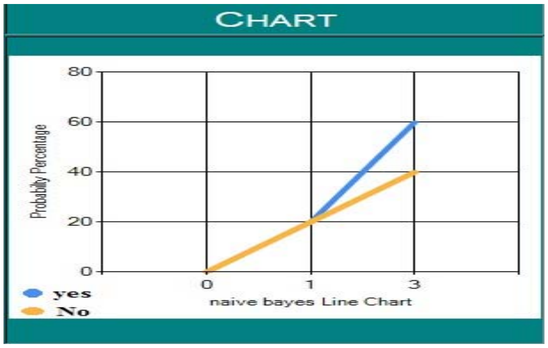
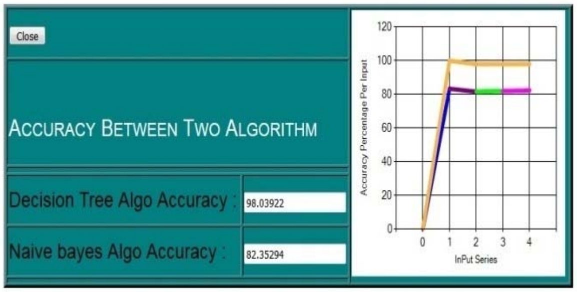


**Fig 5: Performance of naïve bayes**

## IV RESULT & DISCUSSION

In this system, training data set is tested in predicting the heart disease. 13 attributes were taken for the prediction system of the disease. Two algorithms were selected for comparison – Decision tree & Naïve Bayes. The most accurate & effective system was tested among the two. Study found that Decision tree was more precise in its prediction of heart disease. Various test cases included in the paper prove the above said fact in this regard.

Fig 6 depicts the comparison between the decision tree and naïve bayes algorithms with the accuracy vales like 98.03 and 82.35. Fig 7 shows the test cases of these two algorithms with different accuracy values like 98.21 and 82.14 and Fig 8 shows one more test case result like 90 and 70.

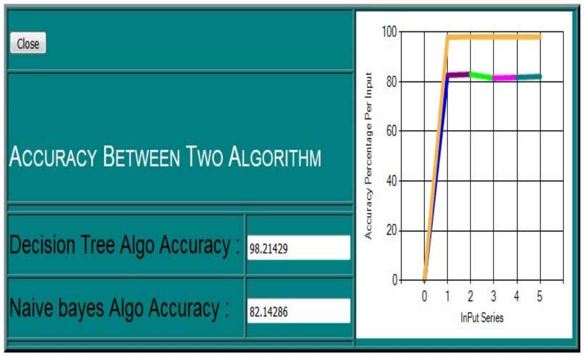


Accuracy Between Two Algorithm

Decision Tree Algo Accuracy : 98.03922

Naive bayes Algo Accuracy : 82.35294

**Fig 6: Comparison of Decision and naïve bayes algorithm (Test case 1)**

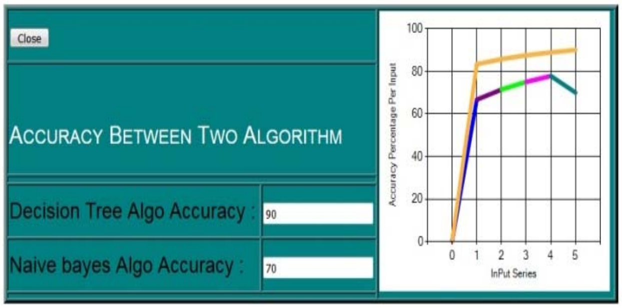


Accuracy Between Two Algorithm

Decision Tree Algo Accuracy : 98.21429

Naive bayes Algo Accuracy : 82.14286

**Fig 7: Comparison of Decision and naïve bayes algorithm (Test case 2)**



Accuracy Between Two Algorithm

Decision Tree Algo Accuracy : 90

Naive bayes Algo Accuracy : 70

**Fig 8: Comparison of Decision and naïve bayes algorithm (Test case 3)**

## V CONCLUSION

The study compared the accuracy of the results obtained through two algorithms – Naïve Bayes & Decision tree. Even though both of them were good enough in predicting the Heart disease using various parameters, Decision tree was found to be the best. It gave the most accurate result whether the patient had the possibility of the heart disease. This system can also be used in future projects to detect the specific type of heart disease in particular. Thereby the diagnosis & management of Heart disease can be made simpler.

## REFERENCES

[1] L. Burke, J. Ma, K. Azar, G. Bennett, E. Peterson,Y. Zheng, W. Riley,J. Stephens, S. Shah, B. Suffoletto, T. Turan, B. Spring, J. Steinbergerand C. Quinn, *"Current Science on Consumer Use of Mobile Healthfor Cardiovascular Disease Prevention", Circulation,* vol. 132, no. 12,pp. 1157-1213, 2015.

[2] M. Raihan1, Saikat Mondal2, Arun More3, Md. Omar Faruqe Sagor4, Gopal Sikder5, Mahbub Arab Majumder5, Mohammad Abdullah Al Manjur5 and Kushal Ghosh ," *Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction using Clinical Data and Data Mining Approaches, a Prototype Design*", International Conference on Computer and Information Technology, December 18-20, 2016.

[3]Vanishree K,Jyothi Singaraju*,"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.

[4]. 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.

[5] 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.

[6] Ms. Ishtake S.H , Prof. Sanap S.A., "*Intelligent Heart DiseasePrediction System Using Data Mining Techniques",*International J. of Healthcare & Biomedical Research,Volume: 1, Issue: 3, April 2013, Pages 94-101.

[7] G. M. Nasira , N. Hemageetha, "*Vegetable Price Prediction Using Data Mining Classification Technique",* Proceedings of the International Conference on Pattern Recognition,Informatics and Medical Engineering , March 21-23, 2012.

## BIOGRAPHIES

**Ms.Priyanka N** is a student of Computer Science from, Adhichunchanagiri Institute of Technology, Chikkamagaluru, Presently pursuing M.Tech (CS) from this college. She received B.E from Malnad college of Engineering, affiliated to VTU University; Hassan in the year 2015.She has published two papers in reputed International journals.

**Dr.Pushpa RaviKumar**B.E., M.Tech., Ph.D., LMISTE is working as professor and Head,Department of computer science & Engineering,AIT college,Chikkmagaluru.She had 16yr of teaching experience & 5yr of research experience.She had completed her Ph.D from VTU in 2014 as full time research scholar at R.V college of engineering Bangalore. She had completed her M.Tech from R.V. College of Engineering in the year 2007.She has presented and published many research papers in National & International conferences & journals. Her research interest in Data mining, Neural networks, Social network analysis & Computer networks.