Heart Disease Prediction System Using CRISP-ADM and Decision Trees

V.Krishnaiah[1] ,Dr.G.Narsimha[2], Dr.N.Subhash Chandra[3]
[1] Assistant Professor, Dept. of CSE, CVR College of Engineering, Ibrahimpatnam(M),R.R.Dist., AP, India.
Email: varkala.krish@gmail.com
[2] Associate Professor,Dept. of CSE, JNTUH College of Engineering, Kondagattu, Karinagar(D), AP, India.
Email: narsimha06@gmail.com
[3]Professor of CSE & Principal, Holy Mary inst. of Tech. & Sc., Keesara (M), R.R.Dist., AP, India.
Email: subhashchandra_n@yahoo.co.in

***Abstract*—This paper aim is to average the use of techniques of decision trees, in combination with the management model CRISP-ADM, to help in the prediction of heart diseases. It is widely based on decision trees, an important concept in the field of artificial intelligence. This paper focuse on discussing how these trees are able to assist in the result making process of identifying heart diseases by the analysis of information provided from the hospitals. This information is captured with the help of techniques and the CRISP-DM management model of data mining in large prepared databases logged from hospital day to day transactions.**

***Index Terms*—Heart disease, Data mining, prediction, Decision tree. CRISP-ADM, etc.**

I. INTRODUCTION

Accurate and error-free of diagnosis and treatment given to patients has been main issue decorated in medical service present days. These systems produce enormous amounts of data which take the form of numbers, text, charts and images. This data may consist a lot of hidden information which can be use in behind the clinical decision making. The main inspiration for this article is: "How we can turn the data into useful information to support decision making by healthcare practitioners?"

A good prediction system for heart disease can be proved as a better tool for improving the efficiency of a hospital and clinicians. It is very important for clinician as well as patients to know the future holds of heart disease patients for planning the better treatment. The taking of use of data mining approaches in present healthcare system is increasing quickly, because the success of these approaches to classification and prediction systems has enhanced, mainly in relation to helping medical practitioners in their decision making. This article can play an important role in civilizing patient outcomes, cost reduction of medicine, and further advance clinical studies.

II. METHODOLOGY

*Purpose***:** Heart disease prediction using decision trees uses the one of the prominent models of classification which is decision trees to predict the status of possibility of patient having disease or not using the training dataset.

Scope: Clinical diagnosis is regarded as an main yet difficult task that needs to be executed accurately and well. The mechanization of this system would be very useful. Efficient and accurate implementation of mechanical system wants a relative study of different techniques available. This paper intention is to analyze the different predictive/ descriptive data mining techniques proposed in present years for the diagnosis of heart disease.

Classification:

Classification is a data mining function that assigns items in a group to target categories or classes. The objective of classification is to accurately predict the target class for each case in the data.

The Decision tree growing algorithm is mentioned beneath.

Tree Growing (S, A, y):

Here:
S - Training Set
A - Input attributes Set
y - Target attribute
Create a new tree T with a single root node.

IF one of the Stop Criteria is fulfilled THEN
Mark the root node in T as a leaf with the mainly general value of y in S as a tag.
ELSE

Find a discrete function f (A) of the input attributes values such that splitting S according to f (A)'s outcome (v1... vn) gains the best splitting metric.

IF best splitting metric >threshold THEN
Label t by f (A)

FOR each outcome v_i of f (A): Set Sub tree *I* = Tree Growing ($\sigma f(A) = v_i$ *S*, A, y).
Connect the root node of *T* to Sub tree *I* with an edge that is labeled as v_i

END FOR
ELSE

Mark the root node in T as a leaf with the most general value of y in S as a tag.

END IF
END IF

RETURN

DOI:10.32377/cvrjst0502

Top-Down Algorithmic structure for Decision Trees Induction:

The possibility vector has a part of 1 (the variable *x* gets only one value), then the variable is defined as clean. On the other hand, if all components are equal, the level of infection reaches maximum. Given a training set *S*, the possibility vector of the target attribute *y* is defined as:

The good–of–split due to discrete attribute *ai* is defined as reduction in dirtiness of the target attribute after partitioning *S* according to the values vi,j a *dom* (*ai*):

$$\Delta\Phi(a_i,S)=\phi(P_y(S))-\sum_{j=1}^{|dom(a_i)|}\frac{|\sigma_{a_i=v_{i,j}}S|}{|S|}\cdot\phi(P_y(\sigma_{a_i=v_{i,j}}S))$$

Information Gain: Information gain is an impurity-based criterion that uses the entropy measure (origin from information theory) as the impurity measure

$$InformationGain(a_i,S)=$$
$$Entropy(y,S)-\sum_{v_{i,j}\in dom(a_i)}\frac{|\sigma_{a_i=v_{i,j}}S|}{|S|}\cdot Entropy(y,\sigma_{a_i=v_{i,j}}S)$$

Where:

$$Entropy(y,S)=\sum_{c_j\in dom(y)}-\frac{|\sigma_{y=c_j}S|}{|S|}\cdot\log_2\frac{|\sigma_{y=c_j}S|}{|S|}$$

The CRISP-ADM Methodology:

The CRISP-ADM methodology is describe in conditions of a hierarchical method representation, comprise four levels of abstraction (from general to specific): phases, generic tasks, specialized tasks, and procedure instances.

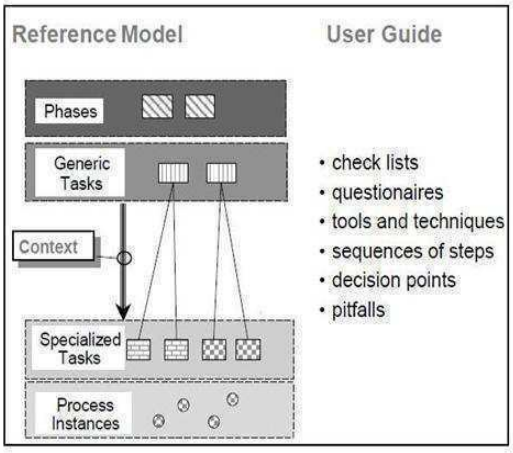

Figure 1. Four level breakdown of the crisp-dm methodology

The CRISP-ADM methodology distinguish between the Reference Model and the User Guide. while the Reference Model presents a fast outline of phases, tasks, and their outputs, and describes what to do in a data mining project, the User Guide gives more thorough instructions and hints for each phase and each task within a phase and depicts how to do a data mining project.

TABLE I.
DATASET ATTRIBUTE EXPLANATION
ATTRIBUTE INFORMATION 15 ARE USED.

1. Id: patient identification number

2. Age: age in years (that will be changed to nominal attribute with values ranging young, idle, old, very old.)

3. Sex: sex (1 = male; 0 = female)

4. Cp: chest pain type
   -- Value 1: typical angina
   -- Value 2: atypical angina
   -- Value 3: non-angina pain
   -- Value 4: asymptomatic

5. Trestbps: resting blood pressure (in mm Hg on entrance to the hospital)

6. Chol: serum cholesterol in mg/dl

7. Smoking: (1 = yes; 0 = no (is or is not a smoker)

8. Fbs: (fasting blood sugar > 120 mg/dl)
   (1 = true; 0 = false)

9. Restecg: resting electrocardiographic results
   --Value 0: normal
   --Value 1: having ST-T wave abnormality (T wave inversions and/ ST elevation or depression of > 0.05 mV)
   --Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

10. Thalach: greatest heart rate achieved

11. Exang: exercise induced angina
    (1 = yes; 0 =no)

12. Slope: the slope of the max out exercise ST segment
    -- Value 1: up sloping
    -- Value 2: flat
    -- Value 3: down sloping

13. Thal :(3 = normal; 6 = fixed defect;
    7 = reversible defect)

14. Overweight :(yes =1, no=0)

15. Alcohol intake: (Never, Past, Current)

The goal of our project is to establish a standardized process which can be reliably performed by marketing people with only little data mining skills and little time to experiment with different approaches.

DOI:10.32377/cvrjst0502

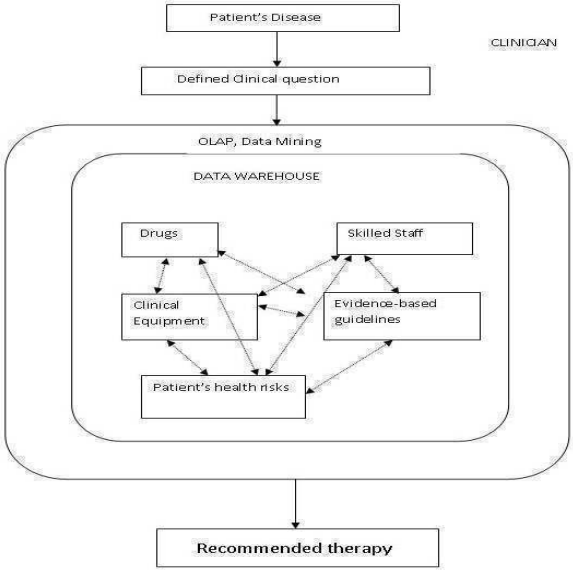

Figure 2: clinical task

III..DESIGN

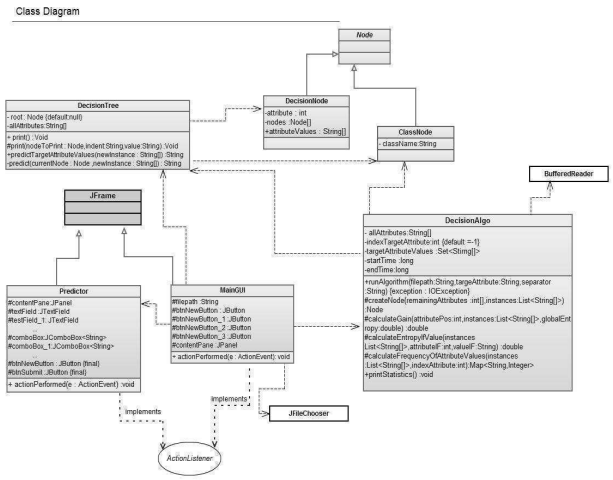

Figure 3: class diagram

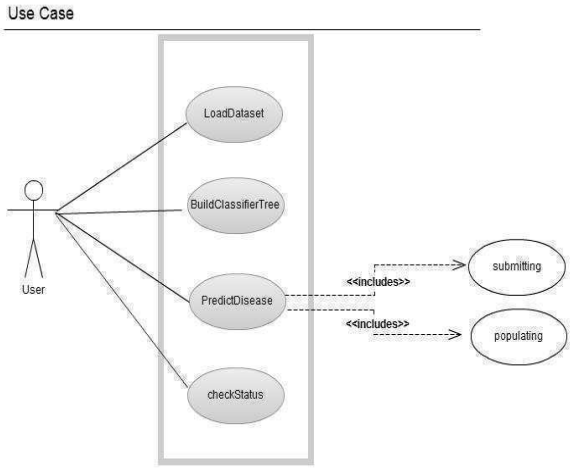

Figure 4: use case diagram

The above use case diagram has user as actor. The user first selects the document set on which the classification must be performed by clicking the "Load Dataset" button. The user can then go for a classification model build based on the loaded dataset. Once the dataset is built new patient details (symptoms) can be entered through the predictor frame. Once the predictor is appropriately populated he can then know the status of the heart disease.

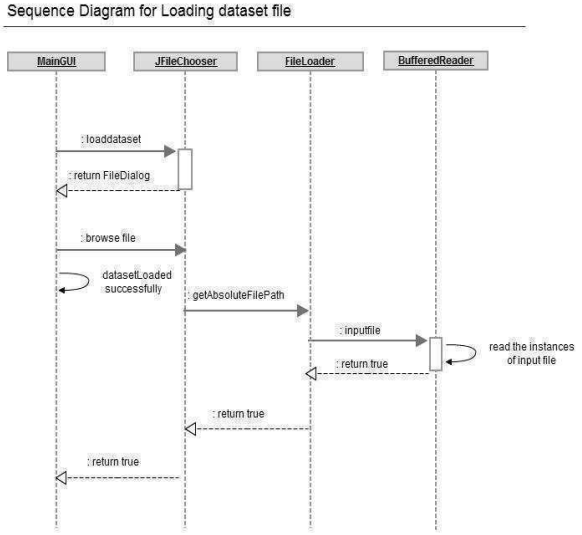

Figure 5:Sequence Diagram for loading dataset file

The above sequence diagram how the different objects come into existence while loading the dataset from local or remote system. Once the instance of Main GUI is running the user clicks the load dataset button and get a J File Chooser which is used to browse for the file we are interested in. Once we are done with choosing the file File Loader object will load the file into the memory. Using the Buffered Reader object we read the instances of dataset.

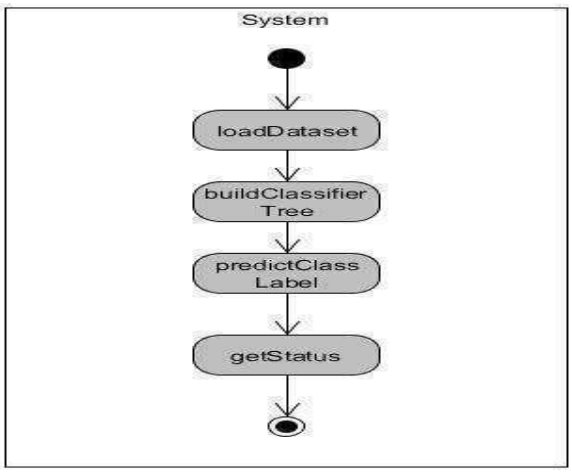

Figure 6:Activity Diagram to get the disease status of patient

DOI:10.32377/cvrjst0502

The above activity diagram has user as actor. The user first selects the document set on which the classification must be performed by clicking the "Load Dataset" button. The user can then go for a classification model build based on the loaded dataset. Once the dataset is built new patient details (symptoms) can be entered through the predictor frame. Once the predictor is appropriately populated he can then know the status of the heart disease.

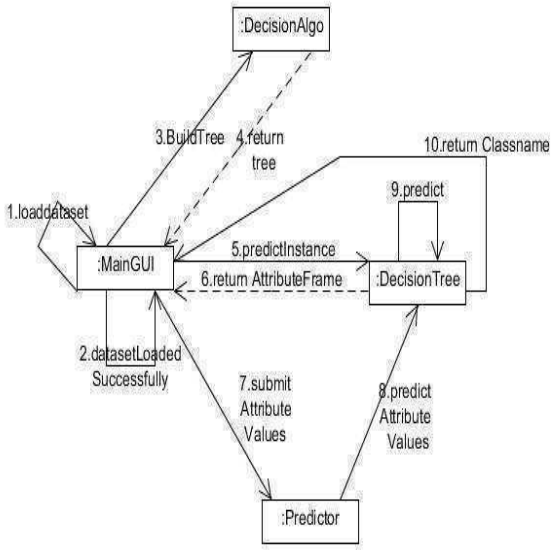

Figure 7: Collaborating Diagram for Predicting the class label of an instance.

The over collaboration diagram shows the different objects come into existence user is trying to predict the class label of a particular instance given by user has already entered the details of a new patient

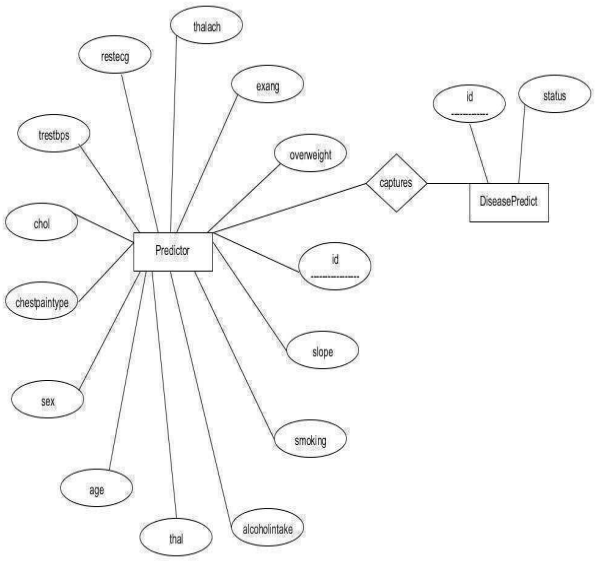

Figure 8: ER diagram

IV.IMPLIMENTATION

Test_Case_id: 01

Test_Case _name: Check whether the application main window is executed and displayed properly or not.
Assumptions: Everything is ok with the IDE and java runtime system is installed apriori.
Procedure: We need to run the main program which is name after "Main GUI. Java".

Expected Result: The Frame is displayed properly.
Actual Result: The Frame is displayed as intended.
Status: PASS.

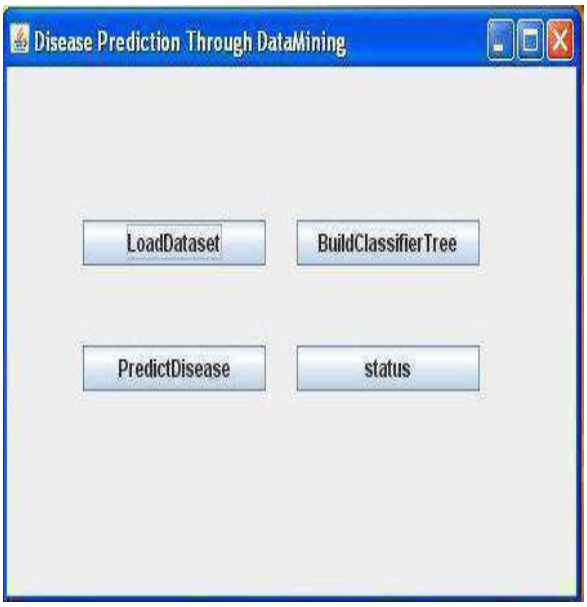

Figure9: Test case 1

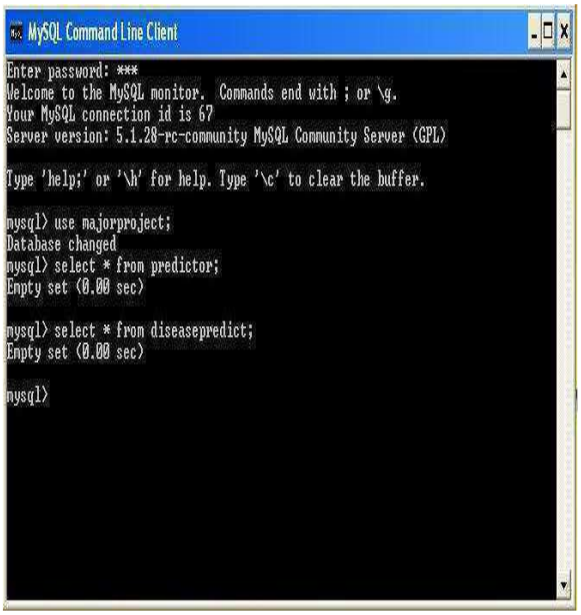

Figure 10: Test case 2

DOI:10.32377/cvrjst0502

Test_Case_id: 02

Test_Case _name: Check the status of database tables when no records are available.
Assumptions: The database software we are using MySql is properly configured and the database and tables are accordingly created with no default values initially.

Procedure: We need to click open"MySql" command prompt and check the status of tables with the query "select * from predictor".

Expected Result: At the start of the execution the query should show "Empty Set".
Actual Result: The query resulted in "Empty Set".
Status: PASS.

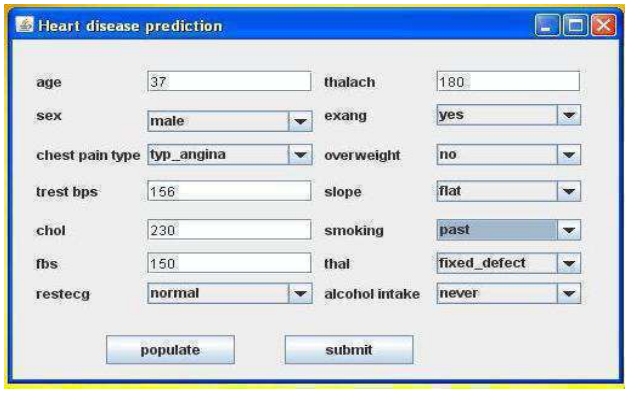

Figure 11: Test case 3

Test_Case_id: 03

Test_Case _name: Check whether the "Predict Disease" button is working properly or not.
Assumptions: Everything is ok with the IDE and java runtime system is installed apriori. And the main class of "Main GUI java" is successfully run and displays the frame. A dataset on heart disease is loaded successfully. And a Classifier is built.
Procedure: We need to click on the "Predict Disease" button.

Expected Result: The Predictor frame can open up where we can give different attributes of heart attack symptoms. And a Dialog box should open up to show the unique id of the patient.

Actual Result: The Predictor Frame opened up successfully where we can give different attributes of heart attack symptoms. And a Dialog box open to show the unique id of the patient.
Status: PASS.

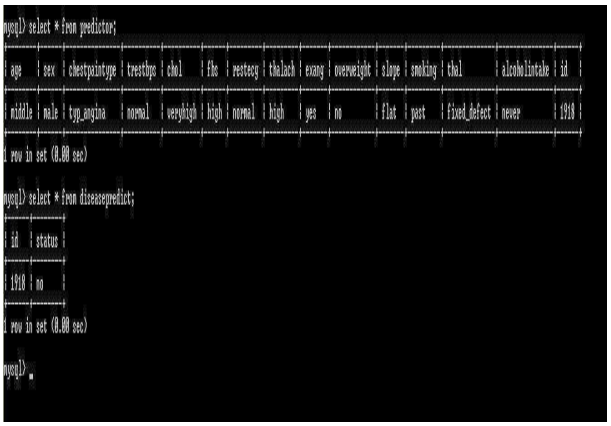

Figure 12: Test case 4

Test_Case_id: 04

Test_Case _name: Check the status of database tables "diseases predict "when the patient status is predicted.
Assumptions: The database software we are using MySql is properly configured and the database and tables are accordingly created with no default values initially. And by successfully enter the values using the predictor frame which is opened by clicking the "Predict Disease" button. And user wants to find the status by selecting the status button.

Procedure: We need to click open"MySql" command prompt and check the status of tables with the query "*select * from disease predict".

Expected Result: The query should show tuples list according to patients.

Actual Result: The query resulted in some tuples.
Status: PASS

V.CONCLUSION

At the present days, the decisions taken by experts and practitioners from many different branches of action must be rapid, accurate and with the possible lowest level problems caused by these decisions. Not with position this fact, due to the difficulty of factors and methods, specialists are level to making incorrect conclusions in their work. Based on the implementation of our proposed decision tree, and the test results on a sample actual database, we conclude that the decision trees with a criteria for data mining help in decision making, particularly in the managing of large data.

DOI:10.32377/cvrjst0502

ACKNOWLEDGMENT

The authors would like be grateful CVR College of Engineering, Hyderabad, for providing its amenities.

REFERENCES

[1] Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[2] CRISP-DM. Available at http://www.crisp-dm.org. Accessed in 1 July 2010.

[3] CUNHA, R. The CRISP-DM Process Model. Available in http://www.cin.ufpe.br/~compint/aulas-IAS/kdd 042/AulaCRISP-DM-OK.ppt.

[4] "Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes", Special report: 2001 – 2003, New Mexico.

[5] HianChyeKoh and Gerald Tan ,"Data Mining Applications in Healthcare", Journal of healthcare information management, Vol. 19, Issue 2, Pages 64-72, 2005.“Heart disease” from http://wikipedia.org

[6] Richard N. Fogoros, M.D, The 9 Factors that Predict Heart Attack90 of heart attacks are determined by these modifiable risk factors,About.com Guide.SellappanPalaniappan, RafiahAwang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968-5/08/$25.00 ©2008 IEEE.

[7] Hand, D. J. (1997): *Construction and Assessment of Classification Rules*. John Wiley & SonsLtd., Sussex, England.

[8] Heart attack dataset from http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

 DOI:10.32377/cvrjst0502