Heart Disease Prediction using Azure ML

Ms. Banashree G. Bisalahalli
Department of CSE
K.L.E.Institute of Technology
Hubballi, India

Mrs. Shanta Kallur
Department of CSE
K.L.E.Institute of Technology
Hubballi, India

Puneeth N.Thotad
Department of MCA
K.L.E.Institute of Technology
Hubballi, India

***Abstract*— over a period of ten years, Heart disease is a prime reason of death in the world. Even in India, 1.7 millions of people killed due to severe heart disease. According to the Global Burden of disease 2016 report which has been released on 15 September 2017, there has been an exponential rapid growth of heart disease over 10 years. So, Azure Machine Learning studio is used to detect heart disease ratio of individual patients through machine learning techniques. It makes use of an existent dataset of heart disease patients from the Cleveland database of UCI repository. It is used to predict and examine the performance of heart disease patients where a current condition is valued from no fettle to likely fettle.**

***Keywords*— Heart disease, Microsoft Azure Machine Learning, Two-Class support Vector Machine, Two-Class Decision Jungle, Multi-Class Decision Jungle**

I. INTRODUCTION

In the world, past 10 years Heart disease is the serious cause of death. Even in India, 1.7 millions of people killed due to severe heart disease. According to the Global Burden of disease 2016 report which has been released on 15 September 2017, there has been an exponential rapid growth of heart disease over 10 years. Several distinct symptoms are in line with heart disease, which makes it cumbersome to diagnose it rapidly. A heart is the organ of all livings beings which plays an important role in blood pumping to the rest of the organs by means of blood vessels of the circulatory system. The term heart disease is close to all the various diseases affecting the heart. Heart disease is a term allocated to a large number of medical conditions related to the heart. Functioning in heart disease patient databases can be similar to the real-life application. Doctor knowledge is to allocate the weight to each attribute. More weight is allocated to the attribute is having a high tendency on disease prediction. Therefore it looks reasonable to employ the knowledge and experience of various specialists collected in databases towards assisting the diagnosis process. It also stipulates healthcare professionals an additional wellspring of knowledge for making choices. The healthcare sector field accumulates a huge quantity of health care data and that need to proceed out to cover unapparent information for operative decision making. The massive volume of data generated for prediction of heart disease is cumbersome and baggy to process out. In many cases diagnosis is based on patient present report and doctor experience. Thus the diagnosis is a complicated task that requires high knowledge and much better experience. To diagnose heart disease patient details there are some of the attributes are taken into considerations such as heart_disease_diag, thal, chest_pain_type,number_of_major_vessel,st_depression_induced_by_exercise,exercise_induced_angina,max_heart_rate, slope_of_peak_exercise and age. Guided by the worldwide increasing mortality of heart disease patients per year and the accessibility of a large number of patient data can be used to extract serviceable knowledge. Treatment of heart disease is quite expensive and most of the patient can't afford it. So, by providing effective treatment it also helps to reduce the cost of treatment. Researchers have been using machine learning techniques to assist health care professionals to diagnose of heart disease.

A. Introduction to domain

Machine learning is a stream of computer science that provides computers the capacity to learn without being explicitly programmed. The name machine learning came out from the course of pattern recognition and computational learning theory in artificial intelligence. Machine learning examines the study and building of algorithms that can learn from and make predictions on data.

Within the area of data analytics, machine learning is a process used to design complex models and algorithms that allow themselves to prediction; in commercial use, this is known as predictive analytics. To predict heart disease it will be using Microsoft Azure Machine learning studio is a collaborative, web tool is used to develop, examine and implement predictive analytics solutions on the data. The azure fair place has 25+ machine learning APIs. The fair place is a comfortable platform for data scientists to develop custom web services issues APIs and accuse of its usage.

B. Objectives

- To provides analysis report for doctors which give the overall accuracy of heart disease patients.

- To predict the score value for the heart disease patients.
- To provide a real time predicted value of patient through patient details.
- To provide necessary measures and precautions to the respective heart disease affected patients.

C. *Aim of the Proposed System*

Prediction of heart disease from the patient details provided using Microsoft Azure Machine Learning Studio.

II. SYSTEM METHODOLOGY

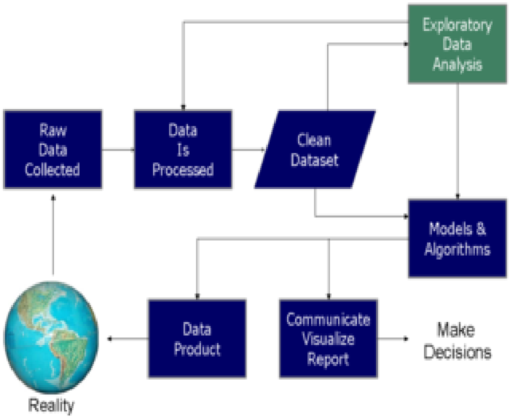


Fig 1: System Methodology

In the above fig 1, raw data is collected from the real world. Collected data is being processed and missing dataset is cleaned. Then the data model is trained and analyzed. After the analysis process, algorithms are applied to the trained model. After the training process, a report will be visualized and data product will be formed. Through the visualization, report decisions will be made and proper prediction will be analyzed though score values.

III. PROPOSED SYSTEM

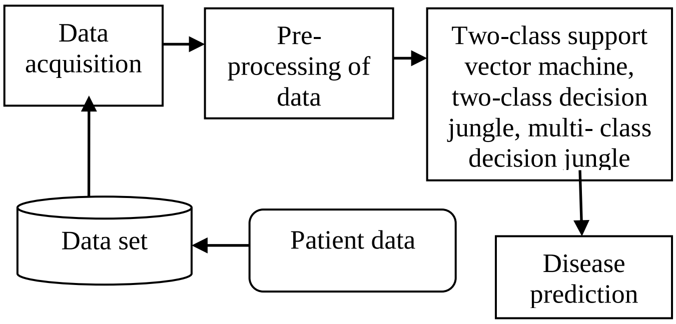

Data acquisition
Pre-processing of data
Two-class support vector machine, two-class decision jungle, multi- class decision jungle
Data set
Patient data
Disease prediction

Fig 2: Proposed system

In fig 2 represents a block diagram of a proposed system where patients detail is to import from web URL via HTTP. The import data refers to the data type of each column based on the values it contains and puts the data into Azure Machine Learning Studio workspace. The output of import data is a dataset that can be used in any experiment. Data acquisition in machine learning consists of two things: data and model. When collecting the data it must have enough features so that it can help to predict the disease and correctly train the learning model. In a pre-processing step, the data is to be cleaned and simplified. By pre-processing of the data, we can get more easily create meaningful features from data. After pre-processing, selected algorithms are applied on the machine learning model, used to measure the accuracy of the predictions. Scoring is the process of generating values or scores based on a trained machine learning model. The values or scores that are produced can represent disease predictions of future values.

A. System Architecture

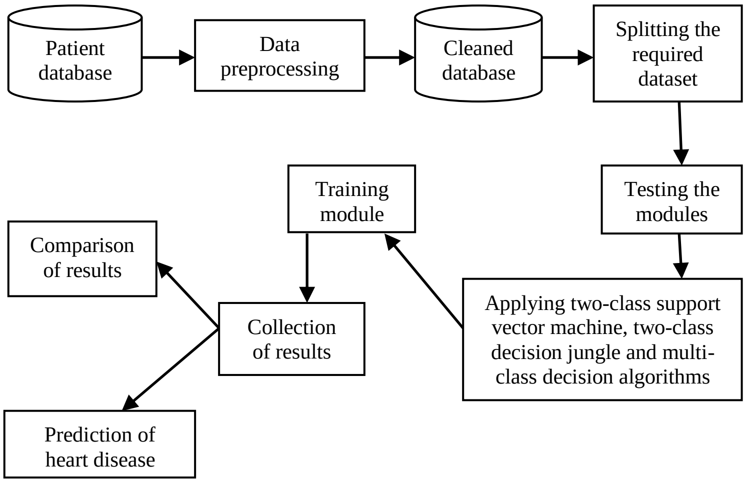

Patient database
Data preprocessing
Cleaned database
Splitting the required dataset
Testing the modules
Training module
Comparison of results
Collection of results
Applying two-class support vector machine, two-class decision jungle and multi-class decision algorithms
Prediction of heart disease

Fig 3: System Architecture

In the above figure, the patient database is collected from an existent dataset of heart disease patients from the Cleveland database of UCI repository. Later the data is being pre-processed where we can get more easily create meaningful features from data. Then the missing values, erroneous records or outliers are cleaned out. Then the selected attributes are used and they are split so that used for testing the modules. To test the modules we make use of Two-Class Support Vector machine, Two-Class Decision Jungle and Multi-Class Decision Jungle algorithms are applied on the training module. The scoring model evaluates the patient data to predict the heart diseases from the scale (value 1, 2, 3 and 4) and absence (value 0). The results are collected, compared, to analyze and diagnose heart disease in patient database and doctor will take necessary measures and precautions to cure heart disease.

B. Algorithms Used For Implementation

1. Two class Support Vector Machine (SVM)

SVM is one of the most used machines learning algorithms and its modes have been used in many applications from information fetching to text and image classifications. SVM can be used for classification and regression tasks. SVM model is a supervised learning model that needs labeled data.

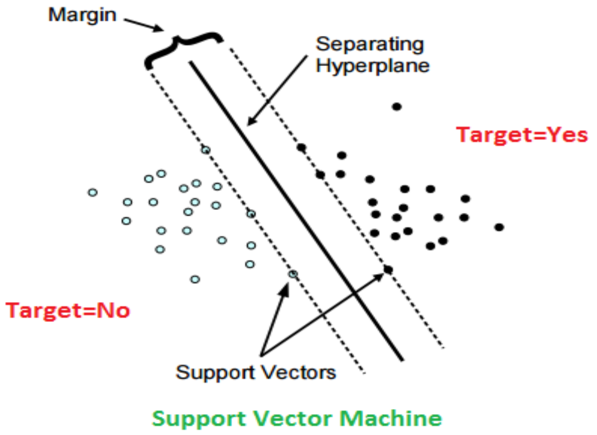
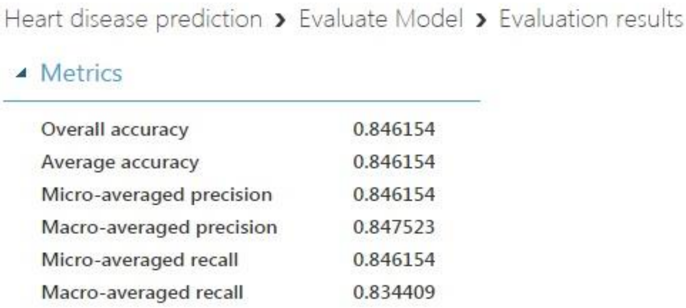


Fig 4: Two-Class Support Vector Machine

In fig 3 represents support Vector Machine. In the training process, the algorithm resolves input data and recognizes a
A pattern in multidimensional feature space refers a hyper plane. All input examples exhibit as points in this space and are mapped to output categories such that categories are divided into the wide and clear gap as possible.


Heart disease prediction › Evaluate Model › Evaluation results

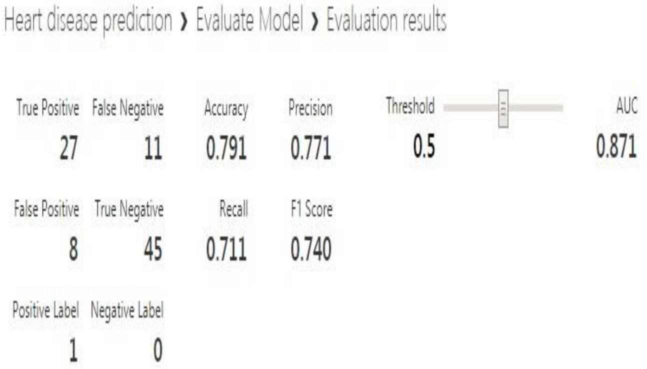
Metrics

Overall accuracy	0.846154
Average accuracy	0.846154
Micro-averaged precision	0.846154
Macro-averaged precision	0.847523
Micro-averaged recall	0.846154
Macro-averaged recall	0.834409

Fig 5: Evaluation results for Two-Class Support Vector Machine

2. Two-Class Decision Jungle

Two-Class Decision Jungle is used in Azure Machine Learning Studio to develop a machine learning model that is based on a supervised ensemble learning algorithm called decision jungle. This algorithm turned back to an untrained classier. Then a particular model is trained by using a Train model or Tune Model Hyper Parameters on a labeled training dataset. Later the trained model is used to make out predictions.

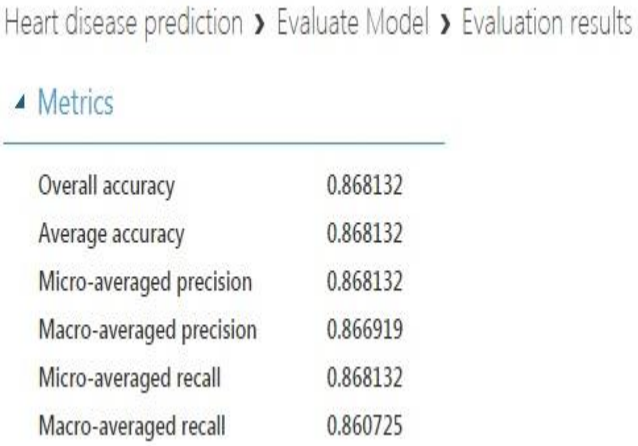

Heart disease prediction › Evaluate Model › Evaluation results

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
27	11	0.791	0.771	0.5	0.871
False Positive	**True Negative**	**Recall**	**F1 Score**		
8	45	0.711	0.740		
Positive Label	**Negative Label**				
1	0				

Fig 6: Evaluation results for Two-Class Decision Jungle

3. MULTI-CLASS DECISION JUNGLE

Multi-Class Decision Jungle algorithm is used in Azure Machine Learning Studio to develop a machine learning model that depends on supervised learning algorithm called decision jungle. After defining a model and its parameters, later to connect to a labeled trained dataset is used to train the model using one of the training modules. Then the trained model is used to predict and analyze a target that has a multiple values.


Heart disease prediction › Evaluate Model › Evaluation results

Metrics

Overall accuracy	0.868132
Average accuracy	0.868132
Micro-averaged precision	0.868132
Macro-averaged precision	0.866919
Micro-averaged recall	0.868132
Macro-averaged recall	0.860725

Fig 7: Evaluation results for Multi-Class Decision Jungle

C. Steps to configure algorithm/model in Microsoft Machine Learning

1. Add the selected algorithm in Microsoft Machine Learning Studio workspace.
2. Double click on the module to open the properties.
3. In the Resampling method, choose either bagging or replicate option.
4. To train the model we can choose single parameter or parameter range option by setting Create trainer mode.
5. Choose a number of decision DAGs it specifies the maximum number of graphs that can be created in the ensemble.
6. Select Maximum depth of the decision DAGs.
7. Then select a Maximum width of the decision DAGs.
8. Enter Number of optimization steps per decision DAG layer.
9. Select Allow unknown values for categorical features to build a group for unknown values in testing or validation data.
10. Set Create trainer mode to Single Parameter and connect to a labeled dataset.
11. Then run the experiment.
12. After the training process, the results are visualized on right-click on train model where a number of iterations have been created

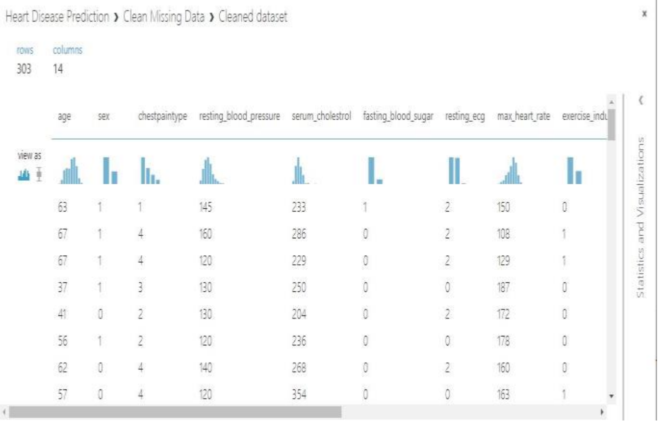
***C.** Applications*

- More efficiently deliver care, and predict prediction for better care with less complexity.

- Massive amounts of data in any given health care system and mining useful information to treat patients.
- It also provides healthcare professionals an extra source of intelligence for making decisions.
- It is used for better health policy-making and prevention of hospital errors, early detection, prevention of heart diseases and preventable hospital deaths.
- It can able to predict and diagnose with heart disease more probabilistically.

IV. RESULTS AND DISCUSSION

After the evaluation of the model, the overall accuracy of algorithms can be compared. The Two-Class Support Vector Machine has 84.6154%, Two-Class Decision Jungle has 79.1% and Multi-Class Decision Jungle has 86.8132%. So based on the better accuracy of the algorithm, the score value will be predicted and a doctor will take necessary precautions for the patient.


Heart Disease Prediction › Clean Missing Data › Cleaned dataset
view as
Statistics and Visualizations

rows	columns
303	14

age	sex	chestpaintype	resting_blood_pressure	serum_cholestrol	fasting_blood_sugar	resting_ecg	max_heart_rate	exercise_indu
63	1	1	145	233	1	2	150	0
67	1	4	160	286	0	2	108	1
67	1	4	120	229	0	2	129	1
37	1	3	130	250	0	0	187	0
41	0	2	130	204	0	2	172	0
56	1	2	120	236	0	0	178	0
62	0	4	140	268	0	2	160	0
57	0	4	120	354	0	0	163	1

Fig 8: Cleaned dataset

In the above figure, shows the cleaned dataset where the dataset contains missing values, erroneous records or outliers are cleaned out.

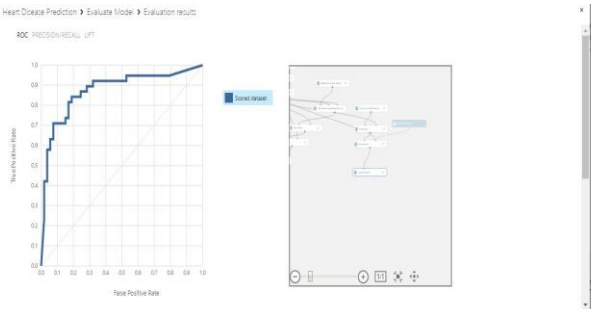


Fig 9: Evaluation results for Heart Disease Prediction

In the above figure, represent evaluation results for Heart Disease Prediction. The receiver operating characteristic (ROC) curve is an effective method for evaluating the performance of heart disease diagnosis. The ROC curve illustrates a scored dataset between a true positive rate and false positive rate.

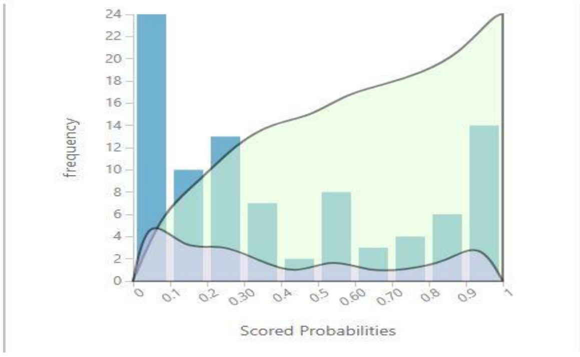
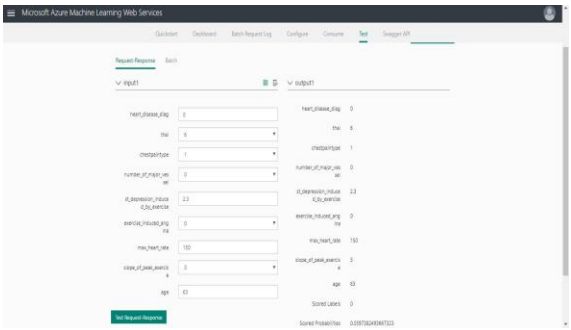


Fig 10: evaluation of scored probability

In the above figure, a graph shows a ratio of scored probabilities with respect to frequency.



input1

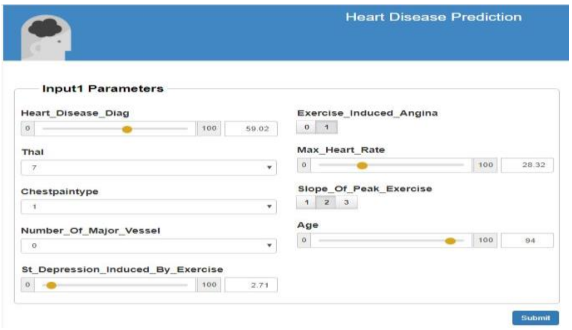
heart_disease_diag: 0
thal: 6
chestpaintype: 1
number_of_major_vessel: 0
st_depression_induced_by_exercise: 2.3
exercise_induced_angina: 0
max_heart_rate: 150
slope_of_peak_exercise: 3
age: 63

output1

heart_disease_diag: 0
thal: 6
chestpaintype: 1
number_of_major_vessel: 0
st_depression_induced_by_exercise: 2.3
exercise_induced_angina: 0
max_heart_rate: 150
slope_of_peak_exercise: 3
age: 63
Scored Labels: 0
Scored Probabilities: 0.0359738248366732

Fig 11: Test Request-Response

In the above figure have to input the patient details, it will test and analyze the report and generates scored labels and scored probabilities.

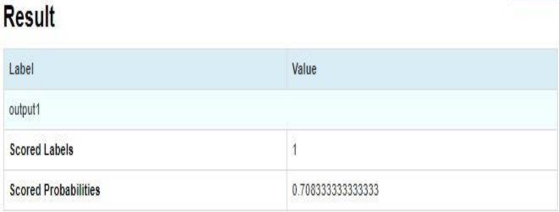


Input1 Parameters

Heart_Disease_Diag: 0 — 100 — 59.02
Thal: 7
Chestpaintype: 1
Number_Of_Major_Vessel: 0
St_Depression_Induced_By_Exercise: 0 — 100 — 2.71
Exercise_Induced_Angina: 0 1
Max_Heart_Rate: 0 — 100 — 28.32
Slope_Of_Peak_Exercise: 1 2 3
Age: 0 — 100 — 94

Fig 12: Import input parameters

In the above figure, a doctor will import the input parameters with respect attributes and then click on the submit button.



Result

Label	Value
output1	
Scored Labels	1
Scored Probabilities	0.708333333333333

Fig 13: Result of Heart Disease Prediction

In the above figure shows that result of the Heart Disease Prediction. It contains the label and value. The label includes the scored labels and scored probabilities.

V. CONCLUSION

By analyzing the experimental results, it is concluded Two-class Support Vector Machine, Two- Class Decision Jungle and Multi-Class Decision Jungle classification algorithms techniques turned out to be the best classifier for heart disease prediction in Microsoft Azure Machine Learning because it contains overall 83.509% accuracy. After the trained data set it evaluates the patient data to predict the heart diseases from the scale (value 1, 2, 3 and 4) and absence (value 0). By knowing the prediction value, a doctor will take necessary precautions for the patient.

REFERENCES

[1] Jyoti Soni, Ujma Ansari and Sunita Soni, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction,” International Journal of Computer Applications, Volume 17– No.8, March 2011.

[2] Vikas Chaurasia, “Early Prediction of Heart Diseases Using Data Mining Techniques,” Vol.1,208-217.

[3] A. Dhanasekar and Dr. R. Mala, “Analysis of Association Rule for Heart Disease Prediction from Large Datasets,” Vol. 5, Issue 10, October 2016.

[4] Rovina Dbritto, Anuradha Srinivasaraghavan and Vincy Joseph, “Comparative Analysis of Accuracy on Heart Disease Prediction using Classification Methods,” Volume 11 – No. 2, July 2016.

[5] P. V. AnkurMakwana, “Identify the patients at high risk of re-admissionin hospital in the next year,” International Journal of Science andResearch, vol. 4, pp. 2431–2434, 2015.

[6] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, “Computationalintelligence for heart disease diagnosis: A medical Knowledge driven approach,” Expert Systems with Applications, vol. 40, no. 1, pp. 96–104,2013.

[7] Y. Xing, J. Wang, Z. Zhao, and Y. Gao, “Combination data miningmethods with new medical data to predicting outcome of coronary heartdisease,” pp. 868–872, 2007.

[8] Asha Rajkumar, G.Sophia Reena, Diagnosis Of Heart Disease Using Datamining Algorithm, Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver. 1.0 September 2010.

[9] Sunita Soni, O.P.Vyas, Using Associative Classifiers for Predictive Analysis in Health Care Data Mining, International Journal of Computer Application (IJCA, 0975 – 8887) Volume 4– No.5, July 2010, pages 33-34.

[10] K.Srinivas, B.Kavihta Rani , A.Govrdhan , Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks, (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255.

[11] Ankita Dewan, Meghna Sharma,” Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification”, 2nd International Conference on Computing for Sustainable Global Development IEEE 2015 pp 704-706. [12] R. Alizadehsani, J. Habibi, B. Bahadorian, H. Mashayekhi, A. Ghandeharioun, R. Boghrati, et al., "Diagnosis of coronary arteries stenosis using data mining," J Med Signals Sens, vol. 2, pp. 153-9, Jul 2012.

[12] Carlos Ordonez, Edward Omincenski and Levien de Braal ,“Mining Constraint Association Rules to Predict Heart Disease”, Proceeding of 2001, IEEE International Conference of Data Mining, IEEE Computer Society, ISBN-0-7695-1119-8, 2001, pp: 433-440.

[13] Deepika. N, “Association Rule for Classification of Heart Attack patients ”, IJAEST, Vol 11(2), pp 253-257, 2011.

[14] Usha. K Dr, “Analysis of Heart Disease Dataset using neural network approach”, IJDKP, Vol 1(5), Sep 2011.