ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Volume 2, Special Issue 19, October 2016

AN IMPROVED MODEL OF CARDIOVASCULAR DISEASE PREDICTION SYSTEM USING DATA MINING MODELS

Sivasankari .S[1], SankaraSubramanian.R[2]

M.Phil. Research Scholar, Department of Computer Science, Erode Arts and Science College, Erode-638009, Tamilnadu, India.[1]

Associate Professor, Department of Computer Science, Erode Arts and Science College, Erode-638009, Tamilnadu, India.[2]

sivamirtha@gmail.com[1], rsankarprofessor@gmail.com[2]

***ABSTRACT*-----People who lack prediction in healthcare will undergo severe treatment. Early prediction of the health related issues will reduce psychological stress and gives enormous time to identify the specialist in the respective field to acquire pre-determined treatment. In this scenario, predictive analysis plays a vital role. Predictive analysis uses techniques and statistical measures to get insight into vast patient's information and analyzing the same data to predict the possible causes for the health issues and its impact on individual patients. Today quality of the healthcare and treatment outcome relies heavily on Data Mining field to exchange information for accurate detection of the life threatening causes. The quality of health care system can be improved by employing an intelligent system via the Data Mining Techniques. Data Mining Techniques are used to reveal the hidden patterns from the vast collection of patient's data. The aim of this research is to develop an intelligent system to predict the Cardiovascular Disease which is the most dreadful disease and the main cause of premature death of the human mankind. This system involves efficient techniques for sharing the health care information to make quick and accurate decision. This paper has analyzed prediction of Cardio vascular disease with 14 attributes as its measurement. Predictive and Descriptive are the two different Data Mining models which include Classification and Clustering techniques respectively are used for this research. When classification is used in conjunction with clustering, it produces considerable improvement in learning the accuracy particularly in detecting the Outliers. Weka 3.8 tool is utilized for implementing the Data Mining Models.**

***KEYWORDS*---- Cardiovascular Disease, Classification, Clustering, Data Mining, Random Forest, K-Median, Weka 3.8.**

I. INTRODUCTION

Data mining [2] is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so that they may be used in an enterprise's decision making.

A. Data Mining Models:

There are two forms of data analysis that can be used for extracting models.

- Predictive Model
- Descriptive model

Predictive model uses Classification, Regression technique and Descriptive model uses Clustering technique[3],Association rules. Predictive model gives result from the already collection of data while Descriptive model elaborate the data.

B. Prediction versus Classification:

Prediction is a statement about what will happen or might happen in the future. So prediction is an important event to avoid undesirable defects especially in health care If the user trying to classify existing data, it would be called as a classification. When we use a classification model to predict the treatment outcome for a new patient, it would be prediction. Now a days Data Mining techniques are more useful in generating useful patterns with 100% accuracy in almost every field. The major challenge in health care organization is providing best accuracy. As it is related to life, the accuracy must be 100%. So these systems handle efficient algorithm and techniques to predict the disease. Automated medical prediction system is more useful for early prediction of the disease. This system can help in predicting cardiovascular disease with less medical test, cost and effort.

II. CARDIOVASCULAR DISEASE

Cardiovascular diseases [4] are the leading cause of death globally. Cardiovascular disease (CVD) is a class of diseases that involve the heart and blood vessels. It includes Coronary Artery diseases (CAD) (commonly known as a heart attack)such as Angina and Myocardial infraction CVDs are stroke, hypertensive heart disease, Cardiomyopathy heart arrhythmia

congenital heart disease, Rheumatic heart diseases. And Valvular heart disease which includes Cere- brovascular disease, Peripheral and thrombosis diseases. It is estimated that 90% of cardiovascular disease (CVD) is preventable.

The factors used to predict Cardiovascular Diseases are: Age, Gender, Blood Pressure, Cholesterol, Blood Sugar ,High Tri – glycerides, E.C.G, Angina, Family history, Obesity, Smoking, Drugs, Pre-Menopause, Rheumatic fever & Swelling.

The Risk factors of Coronary Artery disease include High Blood Pressure, High Triglycerides, Cholesterol, Angina.

The Risk factors of Rheumatic heart disease include Rheumatic Fever with Blood Pressure.

The Risk factors of Cerebrovascular Disease include High Blood pressure, Angina, Cholesterol except High Triglycerides.

III. LITERATURE REVIEW

Mr.S. Dhayanand [6] et.al analyzed SVM and Naïve Bayes Supervised machine learning algorithm to predict kidney disease. His research work mainly focus on finding best classification algorithm to predict Kidney disease. Finally based on the accuracy and execution time SVM is chooses as best algorithm.

Kiyong Noh et al [9] used Associative Classification method to predict the Heart Disease Which is a sub Category of Cardiovascular Disease.

Carlos Ordonez [10] applies Constrained Associative Classification technique for the prediction of the Heart Disease to get highest accuracy.

G.Purusothaman and P.Krisnakumari [11] uses MLP and Genetic algorithm to predict heart disease and its accuracy is compared and found Genetic algorithm has the highest accuracy by using WEKA Tool.

R.Tamilarasi [12] compared Naive Bayes, K nearest Neighbor, Decision tree, Artificial Neural Network and find K nearest Neighbour gives accurate result for predicting Heart Disease.

Durairaj,M, Revathi.V [13] analyze MLP Back Propagation algorithm to predict Heart Disease and its accuracy is listed.

Chaitrali S.Dangare [15] developed the prediction of Heart Disease system using three data mining techniques namely Decision Tree, Naive Bayes and Neural Network and it is compared based on its accuracy. From results it has been seen that Neural Networks provide accurate results.

Gunsai Pooja Dineshgar [16] developed An Intelligent Heart system Prediction system (IHDPS) using K-Mediod algorithm.

Abhishek Taneja [17] developed Heart Disease Prediction system using Naïve Bayes, J48, MLP and find J48 gives best result than other algorithm.

B.K. Bharadwaj [24] use classification technique to analyze and predict the student's academic performance for Training and placement.

K. Lakshmi et.al [25] has used three types of Data Mining techniques Artificial Neural Network, Decision Tree, Logical Regression for the prediction of Kidney Dialysis Survivability, based on the accuracy result ANN is suggested for the Kidney dialysis to get better results.

IV.PROPOSED PREDICTION SYSTEM

The main objective of this proposed system is to build High Intelligent Cardiovascular Disease Prediction System that predicts the disease accurately. To develop this system 14 input attributes are used. Data mining classification technique is used to classify the patients as Low Risk and High Risk and Clustering technique is used to cluster the patients for Coronary Artery Disease[18],Cerebrovascular Disease[19] and Rheumatic Heart Disease [20].The Data Mining technique Random Forest algorithm is used for classification. In this paper we analyses 14 attributes to predict the cardiovascular disease.

A. *Data Base*

A database is a collection of information or data that is stored in a database for managing ,accessing, updating and retrieving. Example of database application include computer -ized library systems, reservation systems etc. The database used in this research paper is collected from a medical practitioner. Microsoft Excel is used to create the Database. Microsoft Excel is a spreadsheet developed by Microsoft for Windows, Mac OS and Android. The advantages of using Microsoft Excel are that users can maximize the value of their data, creating charts, using conditional formatting, identifying trends, bringing data together and utilizing online access. It can also be easily connected to any types of tools and software.

The Data Base used in this research paper contains 1000 records, 14 measurable attributes, a predictive variable and a class variable. Predictive Variable is the summation of all the measurable attributes. Class Variable is used to classify the patient as High Risk and Low Risk.

IF PREDICTIVE VARIBLE ≥ 4
HIGH RISK;
ELSE
LOW RISK;

B. *Input Attributes*

Table . I List of Attributes

S.no	Attribute	Description	VALUES
1	Age	Age in Years	Numeric
2	Gender	Male /Female	1=Female,0=Male
3	BP	Blood Pressure	1=High,0=Low High=Above 140; Low=Below 90;
4	Cho	Cholesterol	1=High,0=Low High=Above200 ;
5	Sugar	Diabetes	1=High,0=Low High=Above125;
6	HT	High Tri Glycerides	1=High,0=Low High=Above 150 ;
7	Angina	Chest pain Type	1=Unstable 0=Stable
8	ECG	Electro Cardio Gram	1=Abnormal 0=Normal
9	Smoke	Smoking	1=Yes,0=No
10	Obese	Obesity	1=Yes,0=No
11	Drug	Using drugs	1=Yes,0=No
12	Family History	Heredity Information	1=Yes,0=No
13	Pre-M	Pre-Menopause	1=Yes,0=No
14	FFT	Rheumatic Fever &swelling	1=Yes,0=No

Table I. shows 14 attributes with its description and values.

C. Techniques used

- Classification
- Clustering

Classification is the separation or grouping of objects into classes using the class label. If classes are created non-empirically, the classification is called Apriori Classification; if they are created empirically the classification is called Posteriori Classification.

Clustering is a technique that partition objects of multi-dimensional data into meaningful disjoint subgroups.

1) Classification Technique

Random forest: The algorithm was developed by Leo Bremen and Adele Cutler. Random forests [21] or Random Decision Forests are an ensemble learning method for Classification Regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes(classification) or mean prediction (regression) of the individual trees. The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging.

Algorithmic steps:

- Let N be the number of test cases, M is the number of variables in the classifier.
- M is the number of input variables to be used to determine the decision in a given node; m is a variable that must be much smaller than M
- Choose a training set for this tree and use the rest of the test cases to estimate the error.
- For each tree node randomly choose m variables on which to base the decision. Calculate the best part from the m variables of the training set.

While classifying the data it is shown that patients are classified as low risk and high risk using Random Forest algorithm with 100% accuracy result.

2) *Clustering Technique*

K-Median: It is a variation of k-mean clustering where instead of calculating the mean value for each cluster to determine its centroid, it calculates the median value. Median value is the center value of the given data set. This has the effect of minimizing error over all clusters. It uses Manhattan distance is used to minimizing error over all clusters with respect to the 1-norm distance metric.

This Manhattan distance metric is also known as L1 distance or L1 norm city block distance, Manhattan length, Rectilinear distance, Murkowski's L1 distance, taxi-cab metric, or city block distance.

L1-norm is also known as least absolute deviations (LAD), least absolute errors (LAE). It is basically minimizing the sum of the absolute differences **(S)** between the target value (Y_i) and the estimated values ($f(x_i)$).

Algorithmic steps:

A median value is the middle value of a set of values arranged in order. Two steps are repeated until the assignments have no longer change.

Assignment step: Assigns each observation to the cluster with the closest center.

Update step: Calculates the new median of each feature of each

Cluster to be the new center of that cluster.

The distance between two points measured along axes at right angles. In a plane with p_1 at (x_1, y_1) and p_2 at (x_2, y_2), it is

$$|x_1 - x_2| + |y_1 - y_2|.$$

This proposed system form 4 clusters [7] to cluster the patients.

Cluster 0: Low Risk Patients

Cluster 1: Coronary Artery Disease Patients

Cluster 2: Rheumatic Heart Disease Patients

Cluster 3: Cerebrovascular Disease Patients

V. FLOW CHART

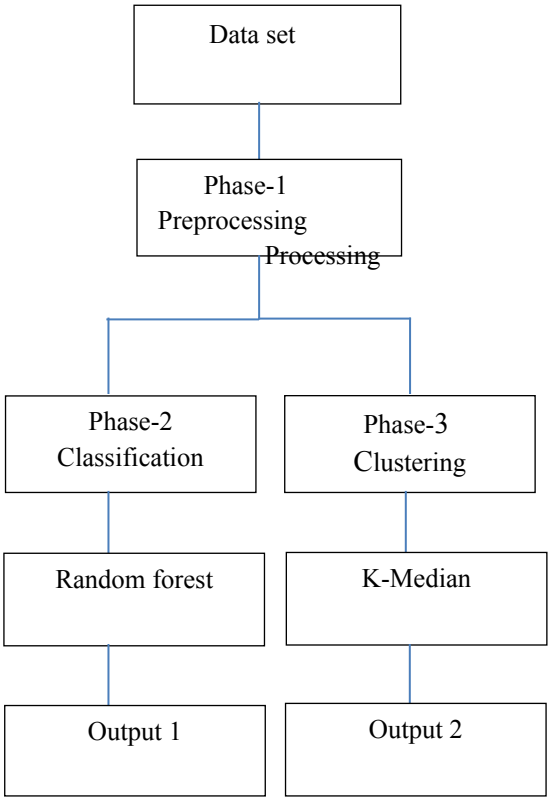

Fig.1 Flow Chart

From Fig.1,

This research paper works in 3 Phase: Phase 1-PreProcessing
Phase 2-Classification
Phase 3-Clustering

For Preprocessing ADD ID method is used. Random Forest algorithm is used for Classification and K-Median algorithm is used for Clustering.

VI.WEKA TOOL

Waikato Environment for Knowledge Analysis (Weka) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License. It is an open source environment in which we can apply various Data Mining Techniques such as Classification, Clustering using various in-built algorithms. We can also view classification errors and clustering assignments in graphical format.

Advantages of Weka:

- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data processing and modeling techniques.
- Ease of use due to its graphical user interfaces.
- Availability of measures and confusion matrix.
- Ease of use due to its graphical user interfaces.
- Availability of Graph, Tree Visualization.

Weka supports several standard Data mining task more specifically Data Pre-Processing, clustering, and classification, Regression, Visualization and Feature Selection. Weka support only ARFF (Attribute Relation File Format) files. The file can be easily converted to ARFF format if it is CSV (Comma Separated Value) file.

VII. RESULTS

The dataset consists of 1000 records and 14 attributes. .The Data Mining tool Weka 3.8. is used. This system works in three phases.

PHASE 1: Pre Processing

It is a method to Clean, Integrate, Transform the data before processing.

Fig .2 Before ADD ID

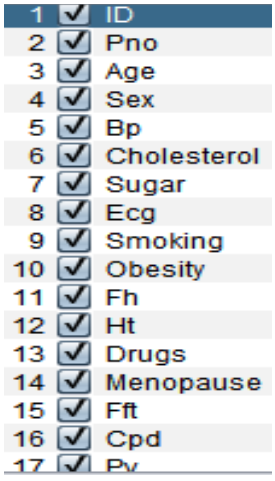

Fig .3 After ADD ID

Method **Method**

There are so many options in Weka for Pre-processing like Replacing the attribute, Rename the attribute, Add instance, Remove the attribute and so on. This research paper use Add ID. It is one of the Pre-processing methods, as this research paper handles patient list the ID is created before processing the data.

PHASE 2: Classification

In second Phase the classification is done to classify how many patients are in Low Risk and High Risk. Random forest algorithm is applied for classification and the accuracy is calculated by using confusion matrix and measures.

A. Confusion Matrix

Table II. Confusion Matrix

	A	B
A	TP	FN
B	FP	TN

From Table II, It contains True Positive, True Negative, False Positive, False Negative. As this research paper has 2 classes Low Risk and High Risk 2x2 matrix is used.

B. Classification Results

Table III. Confusion Matrix for Random Forest

	A	B
A	427	0
B	0	573

From the Table III: Wrongly Classified Data= {0 ,0}
Correctly Classified Data= {427-High Risk, 573-Low Risk}

C. Results based on measures

Table IV. Measures

Measures	**Random Forest**
Correctly Classified Instances	100%
Time taken to Classify	0.14 seconds
TP Rate	1.000
Fp Rate	0.000
Precision	1.000
Recall	1.000
F-Measure	1.000
MCC	1.000
ROC Area	1.000
PRC Area	1.000

From Table IV, Random forest algorithm classify the dataset with 100% accuracy with minimum time.

D. Chart for Random Forest

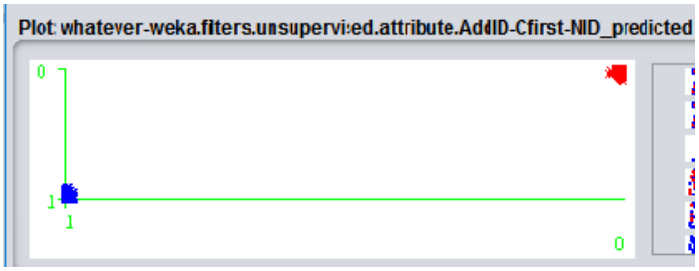

Figure 4 . Classification of Random Forest

From the Figure.1: X ,Y Axis-Shows the Classification of Data;
1-HighRisk Patients, 0-LowRisk Patients;
{1, 1}-TRUE POSITIVE
{1, 0}-FALSE POSITIVE
{0, 1}-FALSE NEGATIVE
{0, 0}-TRUE NEGATIVE

There are no values in {1,0}, {0,1}.So it shows the data are correctly classified. Out of 1000 patients 427 patients are classified as High Risk people and 573 patients are classified as Low Risk people.

PHASE 3:

In third Phase K-median clustering is done to cluster the patient as No Risk patients, Coronary artery disease patient, cerebrovascular disease patients and Rheumatic Heart disease patients. Totally four clusters are formed.

A. Clustering Results:

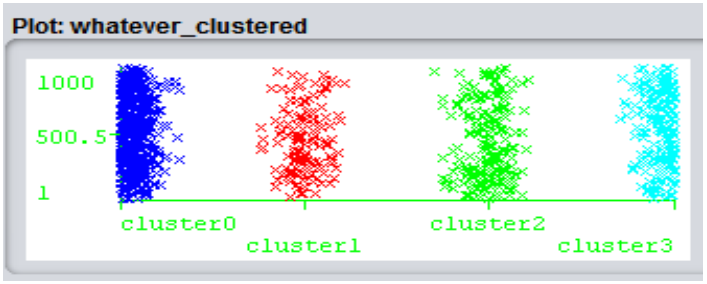

Figure 5. K-Median Clustering

From the Figure 2, X Axis - Number of clusters;
Y Axis - Number of patients;
X shows the 1000 data in the dataset which forms 4 clusters.
There are four clusters of patients.
They are clustered as: Low Risk patients,
Coronary Artery Disease patients,
Rheumatic Heart Disease patients,

Cerebrovascular patients.

Cluster Instances:

Table V. Clustered Data

Clusters	No. of Patients	Category
Cluster 0	454 (45%)	Low Risk patients
Cluster 1	140 (14%)	Coronary Artery Disease patients
Cluster 2	210 (21%)	Rheumatic Heart Disease patients
Cluster 3	196 (20%)	Cerebrovascular Disease patients

From Table V, It shows the patients categorized into 4 clusters.

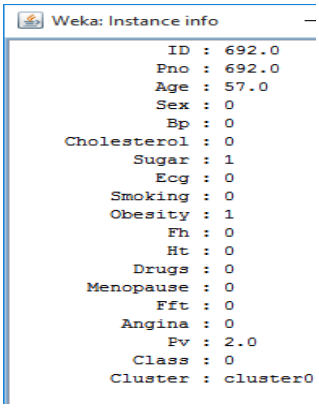

Fig.6 Cluster 0

From Fig .6, it shows the patients are in Low Risk.

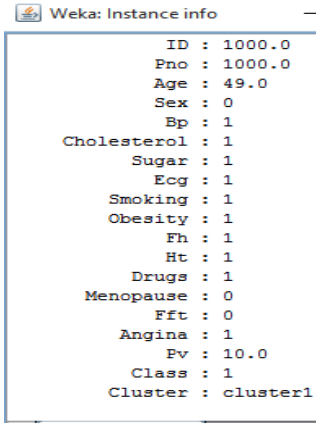

Fig.7 Cluster 1

From Fig .7, it shows the patients are in Coronary Artery Disease, as Blood Pressure, Cholesterol, Sugar, ECG, Smoking, Obesity, High Triglycerides, Drugs, Angina, are 1.

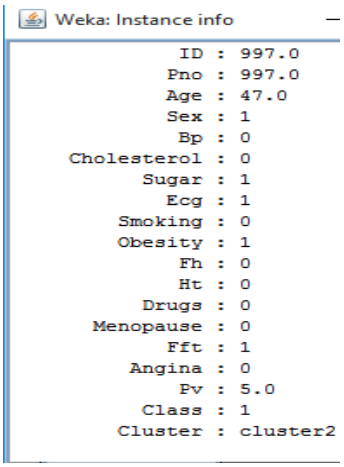

Fig.8 Cluster 2

From Fig .8, it shows the patients are in Rheumatic heart disease, as FFt (Rheumatic Fever) is 1.

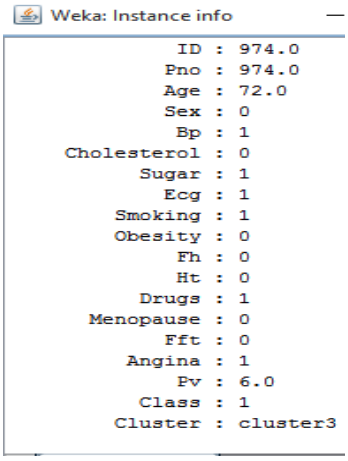

Fig.9 Cluster 2

From Fig .8, it shows the patients are in Cerebrovascular disease, as Blood Pressure, Sugar, ECG, Smoking, Drugs, Angina are 1.

VIII.CONCLUSION

There will be many benefits in the quality of life as the use of predictive analysis increase in the patients. The objective of this system is to predict the Cardio vascular Disease accurately and to cluster the patients according to the disease. This system use 14 attributes as measurements to predict the disease. To predict the disease more accurately data mining classification technique Random Forest is used and its accuracy is listed. To cluster the patient data mining clustering technique K-Median is used. Four Clusters are formed to classify the patients as Low Risk patients, Coronary artery disease patient, cerebrovascular disease patients and Rheumatic Heart disease patients.

IX.FUTURE WORK

This system can be further expanded by using various Data Mining classification & clustering algorithms. And also by cluster the patients for other types of Cardio Vascular Disease.

REFERENCES

[1] "*Health information technology — HIT*" [online] Health IT. Gov. Retrieved 5 August 2014.

[2] Jiawei Han and Michelin Kamber, "*Data Mining: Concepts and Techniques*," Morgan Kaufmann Publishers, ISBN 1-55860-489-8. August 2000

[3] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[4] *Global Atlas on Cardiovascular Disease Prevention and Control* (PDF). World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization. pp. 3–18. ISBN 978-92-4-156437-3,Sep 20,2011.

[5] Eugene F. Krause (1987). *Taxicab Geometry*. Dover [online] ISBN 0-486-25202-7.k-median)

[6] Dr.S.Vijarani, Mr.S. Dhayanand, Data Mining Classification Algorithms for Kidney disease, Prediction, IJCI,Vol ,4.No.4,August 2015.

[7] Kaufman L.,Rousseau P.J. *Finding groups in data. An introduction to cluster analysis*. – John Wiley & Son 2005.

[8] Christopher Whelan Greg Harrell, and Jin Wang Valdosta *'Understanding the K-Medians Problem'* State University, Valdosta, Georgia 31698, USA,2015.

[9] Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee,and Keun Ho Ryu,,*Associative Classification Approach for Diagnosing Cardiovascular Disease,* Springer,vol:345,pp:721-727,2006.

[10] Carlos Ordonez, *Improved Heart Disease Prediction Using Constrained Association Rules*" Seminar Presentation at University of Tokyo,2004.

[11] G.Purusothaman and P. Krisnakumari, *A Survey of Data Mining Techniques on Risk Prediction: Heart* Disease", Vol 8(12), 58385,June 2015.

[12] R. Tamilarasi, Dr.R. Porkodi, *A study and Analysis of Disease Prediction Techniques in Data Mining for Health care*, March 2015.

[13] Durairaj M, Revathi V, *Prediction of Heart Disease Using Back Propagation MLP Algorithm*", ISSUE 08,AUGUST 2015.

[14] S.Vijayarani, S.Sudha,*A Study of Heart Disease Prediction in Data Mining*",Vol.2,No.5,October2012.

[15] Dr.Sulabha S. Apte, PhD,Chaitrali S.Dangare *Improved study of Heart Disease Prediction System using Data Mining Classification Techniques,*vol47-No.10.June 2012.

[16] Mrs.Lolita singh,Gunsai Pooja Dineshgar``*A Review on Data Mining for Heart Disease Prediction*''Volume 5,Issue 2, February 2016.

[17] Abhishek Taneja``*Heart Disease Prediction System using Data MiiningTechniques*"ISSN.0974-6471 Dec Vol. 6 , No(4). 2013

[18] "*Coronary heart disease - causes, symptoms, prevention"* . Southern cross healthcare group Retrieved 15 September 2013.

[19] *"Cerebrovascular disease - NHS Choices - Risks and prevention"* .[online] www.nhs.uk. Retrieved 01-09-2015.

[20] http://www.world-heart-federation.org/press/fact-sheets/rheumatic-heart-disease/[online]

[21] Geurts, P.; Ernst, D.; Wehenkel, L.. *"Extremely randomized trees"(PDF). Machine Learning*. **63**:342 .Doi:10.1007/s10994-006-6226-1(2006)

[22] Ashfaq Ahmed .K, Sultan Aljahdali and Syed Naimatullah Hussain,*Comparative Prediction performance with support Vector Machine and Random Forest Classification Techniques*, International Journal of Computer Applications Volume 69-No.11,pp no 12-16. 2013

[23] Madhuri V.Joseph Data Mining:A Comparaative studyon various techniques and Methods,Volume 3,Issue 2,Feb 2013.

[24] B.K.Bharadwaj and S.Pal.,*Data Mining :A prediction for performance improvement using classification* , International Journal of Computer Science and Information Security (IJCSIS) Vol.9,No.4,pp.136-140,2011.

[25] Lakshmi. K.R,Nagesh .Y,Veerakrishna.M *Performance Comparison of three Data Mining techniques for Predicting Kidney Dialysis Survivability,* International Journal of Advances in Engineering & Technology,Mar,Vol.7,Issue 1(2014)