IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.6, Special Issue.11, Dec 2018 E-ISSN: 2347-2693

A Novel Approach for Heart Disease Classification using Feature Selection

R.UmaDevi[1*], Raynuka Azhakarsamy[2] , J.G.R.Sathiaseelan[3]

[1]Department of Computer Science, National College, Tiruchirappalli, Tamil Nadu, India
[2,3]Department of Computer Science, Bishop Heber college, Trichy, India

Corresponding Author: umadevir95@gmail.com

Available online at: www.ijcseonline.org

Abstract— Heart disease is predicted by classification technique. The data mining tool WEKA has been utilized for implementing J48 classifier. Proposed work is framed with a specific end goal to enhance the execution of models. For improving the classification accuracy J48 is combined with Bagging and Feature Selection. Trial results demonstrated a critical change over in the current J48 classifier. This approach enhances the classification accuracy and reduces computational time.

Keywords— *Data mining, Heart diseases, WEKA, classification, J48, Bagging*

I. INTRODUCTION

Data mining is one of the greatest part substantial and energizing territories of research with the target of choice valuable direction from immense informational indexes. Heart disease is result of alteration of functionality and structure of heart. This leads to inadequate pumping due to which organs and tissues receives insufficient amount of oxygen for their metabolic needs Tissues could become necrotized due to lack of oxygen and may cause death. Heart problem is one of the major reason of death in Australia, USA, Canada, UK. Threateningly all doctors do not possess competence in every field and moreover there is shortage of medical expertise at certain places. Heart disease is said to be imperative reason of fatality in the number of different countries. In US said disease shatters 1 individual in every 34 second. Diagnosis is intricate process that should be precisely and intensively executed. On basis of doctors knowledge and experience, analysis is often made. This may bring about undesirable results along with excessive medical costs of treatment given to patients. This work focuses on predicting heart disease on basis of features. There are variety of methods available for classification of patient data in two classes either positive or negative [1]. Today's, Data mining is an appealing renowned in medicinal services organization as there is expect of vigorous symptomatic system for identifying mysterious and important data in wellbeing information. Malady expectation assumes the prominent critical part in the information mining. Since the mindfulness is extricated by the general population, it is additionally eluded as information expulsion or Knowledge identification from Data (KDD).

Heart Disease:

Cardiovascular illnesses are a noticeable amongst the most widely recognized maladies of the cutting edge world. There are sure things that expansion a man's odds of getting cardiovascular infection. Cardiovascular illness (CVD) alludes to any condition that influences the heart. Numerous CVD patients have manifestations, for example, chest torment (angina) and weakness, which happen when the heart isn't accepting sufficient oxygen. According to a study about 50 per cent of patients, be that as it may, have no manifestations until the point that a heart assault happens. Various variables have been appeared to build the danger of creating CVD. A portion of these are family history of cardiovascular infection, High levels of cholesterol, Low level of cholesterol, Hypertension, High fat eating routine, Lack of consistent exercise and Obesity. With such huge numbers of components to investigate for an analysis of cardiovascular ailment, doctors by and large make a conclusion by assessing a patient's present test outcomes. Past conclusions made on different patients with similar outcomes are likewise analysed by doctors. These perplexing methods are difficult. In this way, a doctor must be experienced and exceptionally talented to analyse cardiovascular ailment in a patient. Data mining has been vigorously utilized as a part of the therapeutic field, to incorporate patient conclusion records to help distinguish best practices. The troubles postured by expectation issues have brought about an assortment of critical thinking procedures. For instance, information mining strategies include Artificial Neural Networks and Clustering Techniques (K-Means Clustering). It is troublesome, notwithstanding, to think about the precision of the systems and decide the best one in light of the fact that their execution is information subordinate. A couple of studies have contrasted information mining and factual methodologies with take care of expectation issues. The correlation thinks about have essentially considered a particular informational collection. The Fig1 [2] shows that the normal artery and the diseased artery.

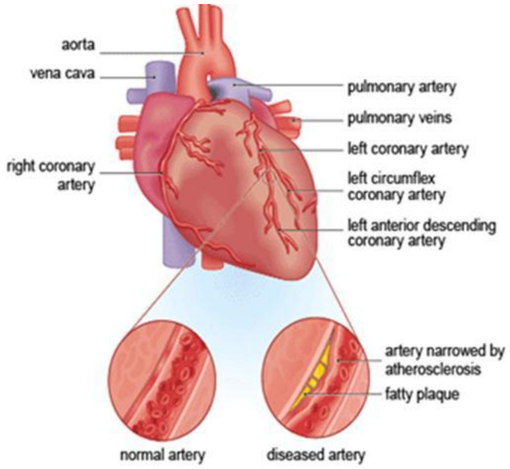


Fig 1: Heart Arteries

There are a many of elements that causes the heart disease can be increased:

- Smoking,
- High blood pressure,
- High cholesterol,
- Lack of physical activities,
- Alcohol intake,
- Obesity,
- Family history,
- Hyper tension etc..

II. LITERATURE SURVEY

Jay Gholap [3] tuned J48 algorithm for predicting soil fertility. A lot of information that are practically reaped in these days alongside the products must be broke down and ought to be utilized to their full extent. Different decision tree calculations can be utilized for the expectation of soil richness. This examination is demonstrated that J48 gives 91.90 % precision, consequently it can be utilized as a base leaner. With the assistance of other meta-calculations like attribute selection and boosting, J48 gives precision of 96.73% which makes a decent prescient model.

ShahedAnzarusSabab et al. [4] designed a paper as for cardio vascular disease using the classification technique and feature selection model. In this paper we endeavored to center around the significance of highlight choice in cardiovascular disease prognosis treatment utilizing diverse information mining calculation. Using proper attribute determination strategy, any classification algorithm can be enhanced fundamentally. Attribute with less commitment in dataset, frequently miss lead the order model and results in poor expectation exactness. In our work, we found that Naïve Bayes gave best outcome before property determination. Be that as it may, subsequent to playing out a controlled and cautious element choice, SVM ended up being the best classifier. Zone under ROC bend investigation indicated results to support us where each of the three classifiers demonstrated much better upgrades after component choice strategy. Notwithstanding this work, we will attempt to assess some more current calculations with better element determination system.

III. DATA SOURCES

The input data has been collected from the UCI Repository. The data size is 4545 records and 11 attributes.

Table 1 shows that brief description of the data set have been considered.

Table 1: Dataset Description

Data set	No of attributes	No of Instances
UCI Repository	11	4545

The attribute description as shown in the table2:

Table 2: Attribute data set for heart disease

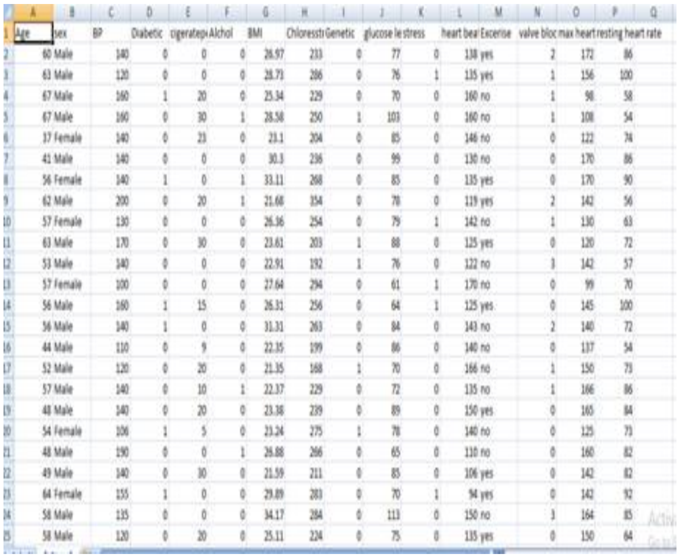
S.No	Name	Description
1	Age	Age in years
2	Sex	Male or Female
3	BP	Blood pressure
4	Diabetes	If Person having diabetes or not Yes=1,No=0
5	Alcohol	If Alcohol consume or not yes=1, no=0.
6	Genetic	If Genetically affected or not yes=1,no=0.
7	Stress	If Stress level if the person having stress or not yes=1 no=0.
8	Exercise	Person doing exercise or not if yes=1,no=0.
9	Valve block	Blocks in heart
10	Max heart rate	Maximum heart in 1min
11	Rest heart rate	Rest heart rate in 1min

3.1 Pre-processing

Data pre-processing is an information mining approach that requires changing a unique information into a reasonable organization. Genuine information is frequently deficient, changeable and insufficient in specific practices and it comprises of more blunders. Data pre-processing is a show strategy for performing such issues. Information pre-processing gets ready basic information for additionally preparing. There are many numbers of information pre-processing procedures. They are

- Data cleaning
- Data Integration
- Data reduction
- Data Transformation

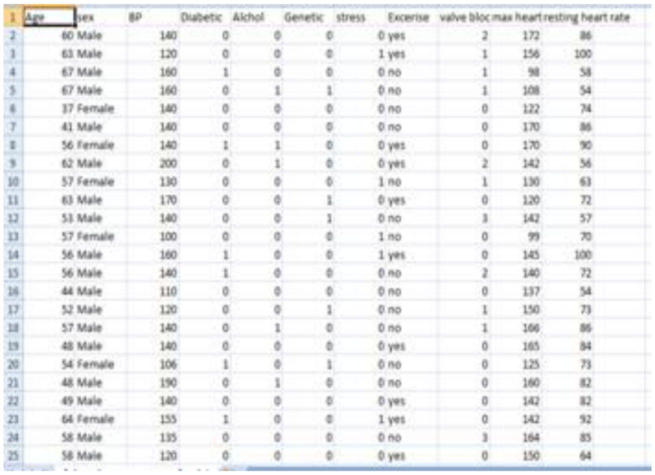
The proposed arrangement of Pre-processing utilized normalization. Normalization may build the exactness and execution of mining calculations including separation measures. The below figure 2 represent the raw data of the data set.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Age	sex	BP	Diabetic	cigeratep	Alchol	BMI	Chloresstr	Genetic	glucose le	stress	heart bea	Excerise	valve bloc	max heart	resting heart rate	
2	60	Male	140	0	0	0	26.97	233	0	77	0	138	yes	2	172	86	
3	63	Male	120	0	0	0	28.71	286	0	76	1	135	yes	1	156	100	
4	67	Male	160	1	20	0	25.34	229	0	70	0	160	no	1	98	58	
5	67	Male	160	0	30	1	28.58	250	1	103	0	160	no	1	108	54	
6	37	Female	140	0	23	0	23.1	204	0	85	0	146	no	0	122	74	
7	41	Male	140	0	0	0	30.3	236	0	99	0	130	no	0	170	86	
8	56	Female	140	1	0	1	33.11	268	0	85	0	135	yes	0	170	90	
9	62	Male	200	0	20	1	21.68	354	0	78	0	119	yes	2	142	56	
10	57	Female	130	0	0	0	26.36	254	0	79	1	142	no	1	130	63	
11	63	Male	170	0	30	0	23.61	203	1	88	0	125	yes	0	120	72	
12	53	Male	140	0	0	0	22.91	192	1	76	0	122	no	3	142	57	
13	57	Female	100	0	0	0	27.64	294	0	61	1	170	no	0	99	70	
14	56	Male	160	1	15	0	26.31	256	0	64	1	125	yes	0	145	100	
15	56	Male	140	1	0	0	31.31	263	0	84	0	143	no	2	140	72	
16	44	Male	110	0	9	0	22.35	199	0	86	0	140	no	0	137	54	
17	52	Male	120	0	20	0	21.35	168	1	70	0	166	no	1	150	73	
18	57	Male	140	0	10	1	22.37	229	0	72	0	135	no	1	166	86	
19	48	Male	140	0	20	0	23.38	239	0	89	0	150	yes	0	165	84	
20	54	Female	106	1	5	0	23.24	275	1	78	0	140	no	0	125	73	
21	48	Male	190	0	0	1	26.88	266	0	65	0	110	no	0	160	82	
22	49	Male	140	0	30	0	21.59	211	0	85	0	106	yes	0	142	82	
23	64	Female	155	1	0	0	29.89	283	0	70	1	94	yes	0	142	92	
24	58	Male	135	0	0	0	34.17	284	0	113	0	150	no	3	164	85	
25	58	Male	120	0	20	0	25.11	224	0	75	0	135	yes	0	150	64	

Fig 2: Raw Data

The figure 3 represent the pre processed data after the pre processed in using the different methods.



	Age	sex	BP	Diabetic	Alchol	Genetic	stress	Excerise	valve bloc	max heart	resting heart rate
2	60	Male	140	0	0	0	0	yes	2	172	86
3	63	Male	120	0	0	0	1	yes	1	156	100
4	67	Male	160	1	0	0	0	no	1	98	58
5	67	Male	160	0	1	1	0	no	1	108	54
6	37	Female	140	0	0	0	0	no	0	122	74
7	41	Male	140	0	0	0	0	no	0	170	86
8	56	Female	140	1	1	0	0	yes	0	170	90
9	62	Male	200	0	1	0	0	yes	2	142	56
10	57	Female	130	0	0	0	1	no	1	130	63
11	63	Male	170	0	0	1	0	yes	0	120	72
12	53	Male	140	0	0	1	0	no	3	142	57
13	57	Female	100	0	0	0	1	no	0	99	70
14	56	Male	160	1	0	0	1	yes	0	145	100
15	56	Male	140	1	0	0	0	no	2	140	72
16	44	Male	110	0	0	0	0	no	0	137	54
17	52	Male	120	0	0	1	0	no	1	150	73
18	57	Male	140	0	1	0	0	no	1	166	86
19	48	Male	140	0	0	0	0	yes	0	165	84
20	54	Female	106	1	0	1	0	no	0	125	73
21	48	Male	190	0	1	0	0	no	0	160	82
22	49	Male	140	0	0	0	0	yes	0	142	82
23	64	Female	155	1	0	0	1	yes	0	142	92
24	58	Male	135	0	0	0	0	no	3	164	85
25	58	Male	120	0	0	0	0	yes	0	150	64

Fig 3: Pre-processed data

3.2 Methodology:

Today's, data mining plays out a important part in the country. In this paper, data mining has been utilized to improve the execution utilizing highlight choice approach.

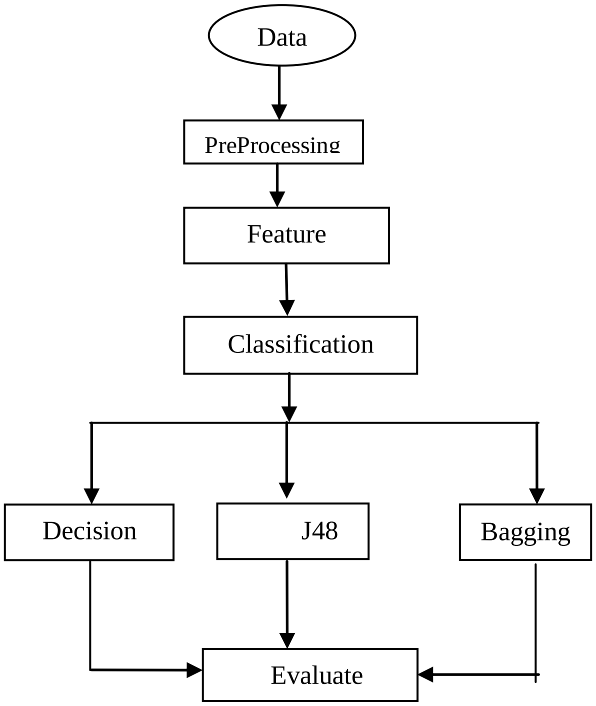


Fig 4: Frame work for JBAS

3.3 Feature selection:

Feature selection is likewise eluded as variable selection or attributes selection. For feature selection choice, various techniques are accessible, for example, chi-squared, correlation analysis, random forest, relief and so on. In feature choice approach, choosing the applicable properties and disposing of the unimportant qualities, different component determination strategies are connected to the pre processed dataset.

3.3.1 Classification:

Classification is a correct data mining strategy in view of machine learning. Fundamentally classification is utilized to arrange everything in an arrangement of data into one of predefined set of classes or gatherings. Characterization strategy makes utilization of numerical systems, for example, decision trees, linear programming, neural network and statistics.

3.3.2. Decision tree:

The introduction of the Decision Tree strategy in the treatment of coronary illness have been examined by the scientists with critical achievement. Decision tree is a tree-like structure, which comprises of interior hubs, branches and leaf nodes, in which each branch means a trait esteem, each inside hub signified a test on a characteristic which is utilized for and a leaf hub speaks to the anticipated classes or class disseminations. The characterization begins from the root hub, at that point crosses the tree in view of the prescient characteristic esteem. The technique includes information apportioning, information order, choice tree class choice, and the demand of lessening of blame trimming to make trimmed choice trees. Arrangement strategies are sorted as administered and unsupervised methodologies. The directed characterization strategies contain chi union and entropy while the unsupervised techniques incorporate indistinguishable width and identical repetition.

3.3.3. J48:

Classification is the way toward building a model of classes from an arrangement of records that contain class marks. Decision Tree Algorithm is to discover the manner in which the traits vector acts for various occurrences. Likewise on the bases of the preparation occurrences the classes for the recently produced examples are being discovered. This calculation produces the standards for the expectation of the objective variable. With the assistance of tree arrangement calculation the basic circulation of the information is effectively understandable.J48 is an extension of ID3. The extra highlights of J48 are representing missing qualities, decision tree pruning, continuous attribute esteem ranges, determination of principles, and so on. In the WEKA information mining instrument, J48 is an open source Java usage of the C4.5 calculation. The WEKA tool furnishes various alternatives related with tree pruning. If there should be an occurrence of potential over fitting pruning can be utilized as a device for précising. In different calculations the arrangement is performed recursively till each and every leaf is pure, that is the classification of the data performed to be as immaculate as would be possible. This calculation it produces the tenets from which specific personality of that information is created. The goal is logically generalization of a decision tree until the point when it picks up balance of flexibility and precision.

3.3.4. Bagging:

Bootstrap Aggregation famously knows as bagging, is a powerful and simple ensemble method.

An ensemble method is a technique that combines the predictions from many machine learning algorithms together to make more reliable and accurate predictions than any individual model. It means that we can say that prediction of bagging is very strong.

The main purpose of using the bagging technique is to improve Classification Accuracy.

- Accuracy Estimation
- Sampling with replacement
- Some may not be used, other may be used more than once.[4]

IV. EXPERIMENTAL RESULT

4.1 WEKA:

Weka is a workbench that contains a gathering of representation instruments and calculations for information investigation and prescient demonstrating, together with graphical UIs for simple access to these functions. Weka boost a few standard data mining tasks, all the more particularly, information pre-processing, classification, clustering, association, representation, and highlight determination. WEKA is a mainstream data mining apparatus. It is utilized to break down the most huge elements causing barrenness. It is likewise used to perform measurable investigation of every individual characteristic.

- In experimental analysis we used classification techniques namely J48, Bagging and Feature Selection to predict the Heart diseases using classification techniques to find accuracy and time consuming of heart disease dataset.
- In Experiment 1 we executed the J48 algorithm with the test mode is ten fold cross validation and data set has 11 attribute it has the accuracy of 64.63 and the time taken is 0.28 seconds.
- In Experiment 2 we executed the J48 with Bagging algorithm and the test mode is ten fold cross validation with the data set 11 attribute it has the accuracy of 65.54 and the time taken 0.25sec.
- In Experiment 3 we executed the J48, Bagging, and Attribute selection(JBAS) we selected 6 attributes uses of string attribute removal applying Percentage split with 97% it has accuracy of 72.05 and the time taken 0.22 sec.

4.2 Performance analysis:

- Table 3 represent the value of correlation analysis of different algorithms. They are TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC Area, PRC area and figure 4 represent the graphical representation of correlation analysis.
- Table 4 represent the accuracy rate, error rate and time taken of the different algorithms. Figure 5 represent the graphical representation of accuracy rate, error rate, error rate and time taken.

Table 3: Correlation Analysis of Algorithms

Algorithms	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**
J48	0.646	0.645	0.552	0.646	0.531	0.004	0.490	0.541
J48 & Bagging	0.655	0.644	0.614	0.655	0.531	0.047	0.492	0.543
JBAS	0.721	0.695	0.799	0.721	0.611	0.136	0.579	0.653

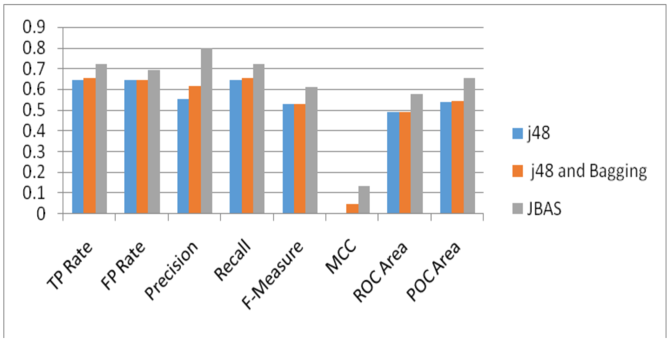


Fig 4: Graphical Representation of correlation analysis using algorithms

Table 4: Represent the time consuming time and accuracy check for different algorithms

Algorithms	Correctly Classified Instances	Incorrectly classified Instances	Time Taken	Mean Absolute Error	Relative Absolute Error
J48	64.83	35.36	0.28	45.25	100.05
J48 & Bagging	65.54	34.46	0.25	45.11	99.74
JBAS	72.05	27.94	0.22	42.78	98.35

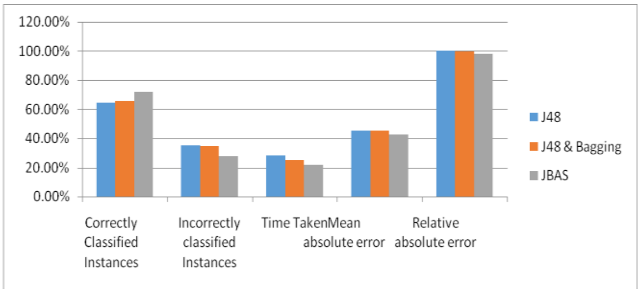


Fig 5: Represent the graphical representation of accuracy and time taken.

V. CONCLUSION

In this paper the proposed a new algorithm for predicting the heart diseases from medical records of Patient.We use Various Classification techniques they are J48,Bagging and Feature selection. We proposed algorithm J48, with other Meta algorithms like Bagging and attribute selection(JBAS) with the test mode has the percentage split 97% it increase the accuracy and time for predicting heart diseases.

REFERENCES

[1] KanikaPahwa, Ravinder Kumar," Prediction of Heart Disease Using Hybrid Technique For Selecting Features", 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON) GLA University, Mathura, Oct 26-28, 2017.

[2] Salha M. Alzahani, AfnanAlthopity, AshwagAlghamdi, BoushraAlshehri, and SuheerAljuaid "An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction", Lecture Notes on Information Theory Vol. 2, No. 4, December 2014.

[3] Jay Gholap" Performance Tuning Of J48 Algorithm For Prediction Of Soil Fertility",.

[4] ShahedAnzarusSabab, Ahmed IqbalPritom, Md. AhadurRahmanMunshi, Shihabuzzaman" Cardiovascular Disease Prognosis Using Effective Classification and Feature Selection Technique",

[5] Monika Gandhi,Dr.Shailendra Narayan Singh, "Predictions in Heart Disease Using Techniques of Data Mining", 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015).

[6]R.umadevi,M.umamaheswari,Dr.J.G.R.Sathiaseelan,"Cardiac Disease Prediction using Data mining techniques:Asurvey",International journal of Advance Research Trends in engineering and technology(IJARTET)vol 5,special issue12,April 2018,ISSN 2394-3777.

[7] SarathBabu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M," Heart Disease Diagnosis Using Data Mining Technique", International Conference on Electronics, Communication and Aerospace Technology ICECA 2017.

[8] Abhishek TANEJA," Heart disease Prediction System Using data Mining Techniques", Oriental Journal Of Computer Science & Technology ISSN: 0974-6471 December 2013, Vol. 6, No. (4): Pgs. 457-466 An International Open Free Access, Peer Reviewed Research Journal Published By: Oriental Scientific Publishing Co., India.

[9] AdityaMethaila, Prince Kansal, HimanshuArya, PankajKumar,"Early Heart Disease Prediction Using Data Mining Techniques", Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP - 2014 pp. 53–59, 2014. © CS&IT-CSCP2014 DOI : 10.5121/csit.2014.4807