IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.6, Special Issue.11, Dec 2018 E-ISSN: 2347-2693

Heart Disease Prediction System Using Data Mining Classification Techniques

D.Bharathi[1*], P.Sundari [2]

[1,2]Department of Computer Science, National College, Trichy, India

[*]*Corresponding Author: bharathi2010malli@gmail.com*

Available online at: www.ijcseonline.org

***Abstract*—** The Healthcare industry is generally "information rich", but unfortunately not all the data are mined which is required for discovering hidden patterns & effective decision making. Data mining techniques are used to notice knowledge in database and for medical research, mainly in Heart disease prediction. This paper has analyzed prediction system for Heart disease using more number of input attributes. The system uses medical terms such as sex, blood pressure, cholesterol Family history, Smoking , Poor diet , High blood pressure , High blood cholesterol , Obesity , Physical inactivity , Hyper tension etc like 13 attributes to predict the likelihood of patient getting a Heart disease. This research thesis added two more attributes i.e. obesity and smoking. The data mining classification techniques, namely Decision Trees, Naive Bayes, and Support vector machine are analyzed on Heart disease database. The show of these techniques is compared, based on accuracy. As per our results accuracy of Support Vector machine, Decision Trees, and Naive Bayes are 98%,85.5%, and 65.74% respectively. Our analysis shows that out of these three classification models support vector machine predicts Heart disease with highest accuracy.

***Keywords*-** Data Mining, Naïve Bayes, Support Vector Machine, Decision Tree, Weka Tool.

I. INTRODUCTION

Data mining is a process of extracting interesting knowledge or patterns from large databases. There are several techniques that have been used to discover such kind of knowledge, most of them resulting from machine learning and statistics. The greater part of these approaches focus on the discovery of accurate knowledge. Though this knowledge may be useless if it does not offer some kind of surprisingness to the end user. The tasks performed in the data mining depend on what sort of knowledge someone needs to mine. Data mining technique are the result of a time-consuming process of study and product development.

The heart is important organ of human body part. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Our Life is dependent on healthy working of the heart. The term Heart disease refers to blood vessel system & disease of heart within it.

A number of factors have been shown that increases the risk of Heart disease: Family history, Smoking, Poor diet, High blood pressure, High blood cholesterol, Obesity, Physical inactivity, Hyper tension, Factors like these are used to analyze the Heart disease. In many cases, diagnosis is generally based on patient's current test results & doctor's experience. Thus the diagnosis is a complex task that requires much experience & high skill.

II. LITERATURE SURVEY

In [1] discuss the show analysis of classification data mining technique for heart disease prediction. The three algorithms used in this work are naïve bayes, WAC and apriori. The performance valuation is based on classification matrix, it display the frequency of correct and incorrect prediction. The analyze model is view in Lift charts, Bar charts and Pie charts. This struggle can be further enhanced and expanded by using other data mining techniques like Time series, Clustering and Association rules. Instead of categorical data, the continuous data can be used.

In [4] the fuzzy specialist system is planned for heart disease diagnosis with reduced number of attribute. The author find that how genetic algorithm and fuzzy logic combine together for efficient and cost effectual diagnosis of heart disease. The GA and two models of fuzzy system Mamdani and Takagi-sugeno were used to find the cost. The dataset were taken in Cleveland clinic establishment dataset. The input field is a set of all the certain kind and the output of the system is to get a value '1' or '0' that indicates the presence or absence of disease. It is additional improved by using art classifiers like

Decision tree, Naïve bayes, Classification via clustering and SVM classifier.

In [5] data mining classification technique namely RIPPER classifier, decision tree, Artificial Neural Network and Support vector machine are used for predicting cardiovascular heart disease. The arrangement factor used for compare these technique are sensitivity, accuracy, specificity, error rate, true positive rate and false positive rate. To gauge the unbiased estimate of prediction models the author used 10 fold cross validation method. This model was industrial by using data mining classification tool weka version 3.6. It contains 14 attributes and 303 instances. Error rates for RIPPER, Artificial Neural Networks, Support vector machine and Decision tree are 02.756, 0.2248, 0.1588 and 0.2755 respectively. The accuracy of RIPPER, Artificial Neural Networks, Support Vector Machine and Decision tree are 81.10%, 80.07%, 84.13% and 79.15% respectively. Four classification models, the Support Vector Machine has given least error rate and highest accuracy. The writer concluded that the Support Vector Machine is the best technique for predicting the cardiovascular disease. In future in order to improve the efficiency of the classification techniques by creating metamodels.

In [6] heart attack symptom is predicted using biomedical data mining technique. The creator used data classification which is based on supervised machine learning algorithms. For data classification the Tanagra tool is used. Using entropy based cross validations and partitioned techniques, the data is evaluated and the results are compared. The algorithms used in these techniques are Knearest neighbours, K-means and Mean Clustering Algorithm (EMC) is the extension of the K-mean algorithm for clustering process which reduces the number of iterations. In this paper result the author analyzed that the mean clustering algorithm perform well when compare to other algorithm. To run the data the time taken is very fast and it gives the result of accuracy about 83.25%. An enhanced by applying unsupervised machine learning algorithm.

In [8] association rule mining method is used for predicting heart attack. In this paper creator future a novel method CBARBSN for association rule mining based on sequence numbers and clustering the transactional database for predict heart attack. The two important step of this process first the medical data is transformed into binary and the planned method is applied to the binary transactional data. The data is collected from Cleveland database. The medical data contains 14 attributes. From the results author concluded that the proposed algorithm performs better than the existing ARNBSN ( ) algorithm. The performance of the algorithms is compared based on the execution time.

In [9] the coronary artery disease was efficiently diagnosed by rotation forest algorithm in order to support clinical decision-making process. It utilize the Artificial Neural Networks with Levenberg-Marquardt back propagation algorithm of rotation forest ensemble method as base classifier. The algorithm is implemented in matlab. From an experiment, the author diagnosed the disease by comparing the performance of base classifiers in terms of sensitivity, accuracy, AUC and specificity on two things i) without rotation forest classifier, the greatest performance of classifiers and ii) with rotation forest algorithm it actually improve the performance of classifier. Result it is experiential that Levenberg-Marquardt was the best classifier with or without random forest. The accuracy is improved to 94.2% of original classification accuracy which is an improvement of 8%. In prospect the proposed work may be used to develop efficient expert systems for the diagnosis of heart disease.

III. METHODOLOGY

A. Navie Bayes

Naïve Bayes is a supervised probability machine learning classifier method that assumes terms occur independently. This can be used to in classifying textual documents in simplest method. The Naïve Bayes model computes the posterior probability of a class, based on the allocation of words in the document this illustration works with the BOWs feature extraction which ignores the situation of the word in the document .Bayesian classification represent a supervised learning method as well as statistical method for classification. It is easy probabilistic classifier based on Bayesian theorem with strong independence assumption. It is for the most part suited when the dimensionality of input is high. They can predict the probability that a given tuple belongs to a particular class. This classification is named after Thomas Bayes who proposed the bayes theorem. Bayesian formula can be written as

P(H | E) = [P(E | H) * P(H)] / P(E)

The basic idea of Bayes's rule is that the outcome of a hypothesis or an event (H) can be predicted based on some evidences (E) that can be observed from the Bayes's rule.

B. Support Vector Machine

SVM perform classification by finding the hyper plane that maximizes the margin between the two classes. The vectors (cases) that define the hyper plane are the support vectors. In machine learning, support vector machines SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Classification SVM Type 1

For this type of SVM, training involves the minimization of the error function:

$$\frac{1}{2} w^T w + C\sum_{i=1}^{N} \xi_i$$

subject to the constraints:

$$y_i\left(w^T \phi(x_i) + b\right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, i = 1,\ldots, N$$

where C is the capacity constant, w is the vector of coefficients, b is a constant, and ξ_i represents parameters for handling no separable data (inputs). The index i labels the N training cases. Note that $y \in \pm 1$ represents the class labels and xi represents the independent variables. The kernel ϕ is used to transform data from the input (independent) to the feature space. It should be noted that the larger the C, the more the error is penalized. Thus, C should be chosen with care to avoid over fitting.

Decision Trees

The decision tree approach is more powerful for classification problems. There are two steps in this techniques building a tree & applying the tree to the dataset. There are many popular decision tree algorithms C4.5. From these C4.5 algorithm is used for this system. C4.5 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data, which leads to poor accuracy in predications. The C4.5 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

C. WEKA TOOL

Weka is a landmark system in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time. The Weka or woodhen (Gallirallus australis) is an endemic bird of New Zealand. It provides many different algorithms for data mining and machine learning. Weka is open source and freely available. It is also platform-independent.

Advantages

- As weka is fully implemented in java programming languages, it is platform independent & portable.
- It is freely available under GNU General Public License.
- Weka s/w contain very graphical user interface, so the system is very easy to access.
- There is very large collection of different data mining algorithms.

Disadvantages

- Lack of possibilities to interface with other software
- Performance is often sacrificed in favour of portability, design transparency, etc.
- Memory limitation, because the data has to be loaded into main memory completely

IV. EXPERIMENT AND RESULTS

In this study mainly classification algorithms such as Naive Bayes, Decision Trees and SVM algorithm is used for predicting the Heart Disease from the given data set instances and the proposed algorithms are applied on type Heart Disease dataset in the WEKA tool and the performance is measured. The heart is main organ of human body part. It is nothing more than a pump, which pump blood through the body. If circulation of blood in body is incompetent the organ like brain suffers and if heart stops working altogether, death occurs within minutes. The term Heart disease refers to blood vessel system & disease of heart within it. A number of factors have been shown that increases the risk of Heart disease: Family history, Smoking , Poor diet , High blood pressure , High blood cholesterol , Obesity , Physical inactivity , Hyper tension etc.,Factors like these are used to analyze the Heart disease. In many cases, diagnosis is generally based on patient's current test results & doctor's experience. Thus the diagnosis is a complex task that requires much experience & high skill.

A. DATA SOURCE

The publicly available heart disease database is used. The Cleveland Heart Disease database consists of 1000 records & Statlog Heart Disease database consists of 970 records [12]. The data set contains 3 types of attributes: Key, Input & Predictable attribute which are listed below.

Attribute Information:

Only 14 attributes used:

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (the predicted attribute)

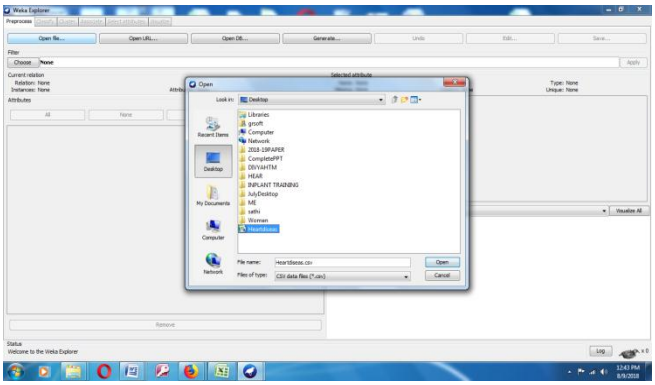


Figure 4.1.1 Upload Heart Disease Dataset

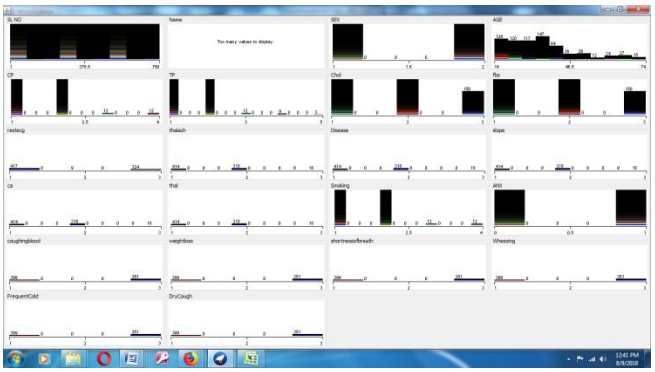


Figure 4.1.2 View Heart Disease Chart

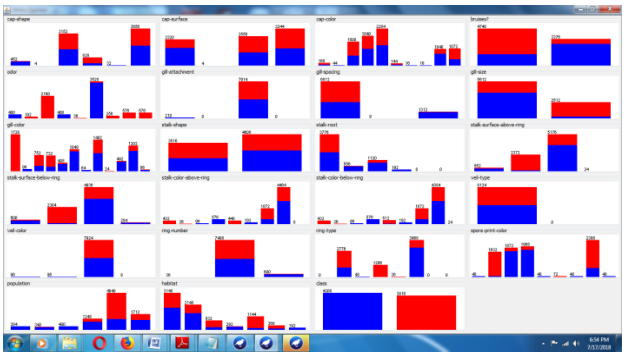
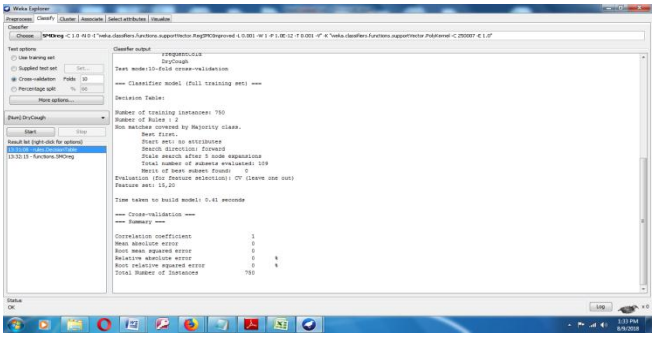
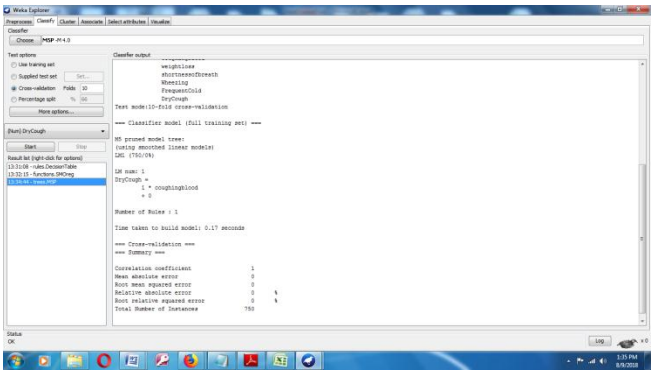


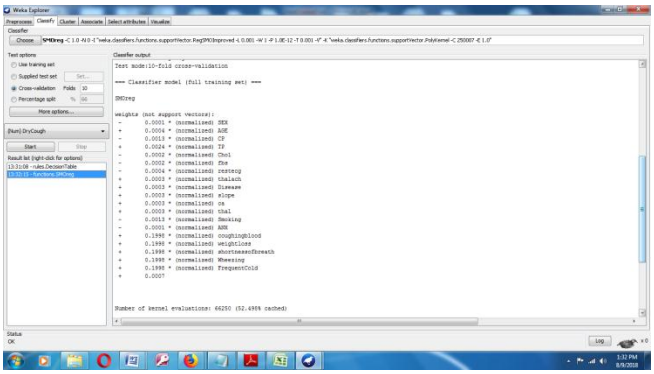
Figure 4.1.3 Over All Chart



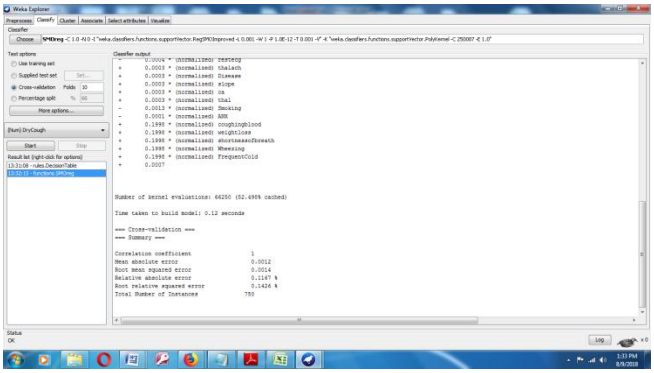
4.1.4 Decision Tree Heart Disease Dataset Execution Second



4.1.5 Navie Bayes Heart Disease Dataset Execution Second



4.1.6 SVM Heart Disease Dataset Execution Second



4.1.7 SVM Heart Disease Dataset Classification techniques

B. Performance Calculated Using Correlation Coefficient

Experiments are performed on the heart disease datasets. Machine with configuration of windows 7 system and 2-GB of RAM is used. The results were compared to experiments with WEKA implementations of Navie Bayes, Decision Tree and Support Vector machine the techniques run to ensure that the results are comparable for Correlation coefficient.

Table 4.2.1: The table shows Correlation coefficient

ALGORITHM	CORRELATION

	COEFFICIENT
Navie Bayes	0.9485
Decision Tree	0.9961
SVM	1

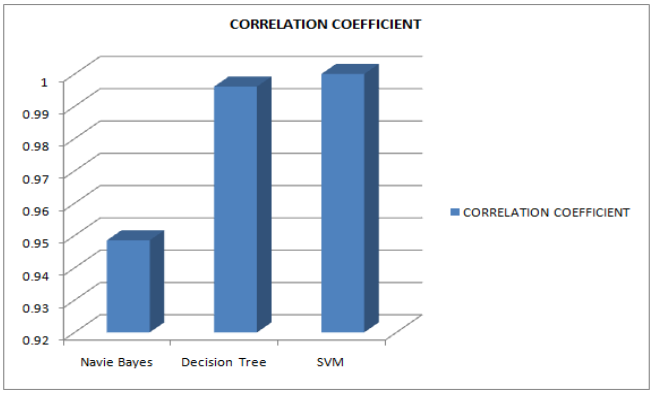


Fig 4.2.1 Evaluation measures of Correlation coefficient

Relative Absolute Error

Experiments are performed on the heart disease datasets. Machine with configuration of windows 7 system and 2-GB of RAM is used. The results were compared to experiments with WEKA implementations of Navie Bayes, Decision Tree and Support Vector machine the techniques run to ensure that the results are comparable for Relative absolute error.

Table: 4.2.2 The table shows relative absolute error

ALGORITHM	RAE
Navie Bayes	0.9632
Decision Tree	0.9971
SVM	0.3707

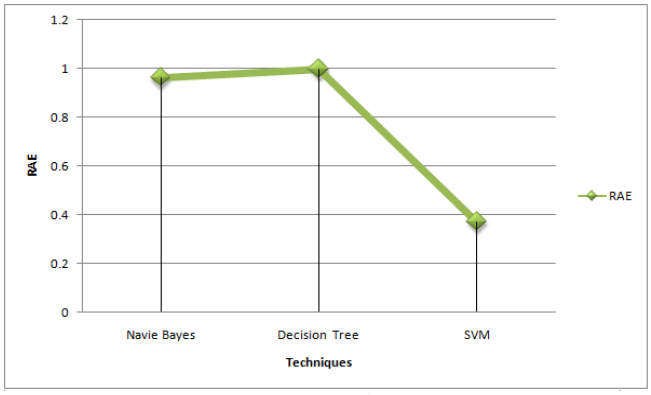


Fig: 4.2.2 Evaluation measures of Relative absolute error

Execution Time Second

Experiments are performed on the heart disease. Machine with configuration of windows 7 system and 2-GB of RAM is used. The results were compared to experiments with WEKA implementations of Navie Bayes, Decision Tree and Support Vector machine the techniques run to ensure that the results are comparable for Relative absolute error.

Table 4.2.3 Execution Time Second

ALGORITHMS	TIME TAKEN (IN MIL SECS)
Navie Bayes	0.03
Decision Tree	0.13
SVM	0.02

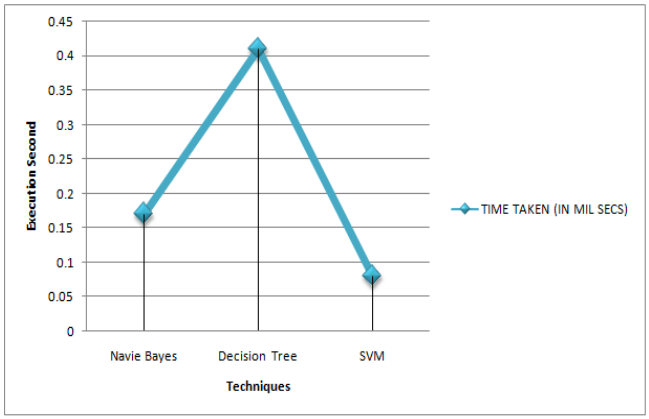


Figure: 4.2.3 Representing Dataset measured in Execution Second(ms)

V. CONCLUSION

Around 18 million people 7% of the Indians are affected by heart disease. Heart disease is mostly affected 65 above age group. This paper mainly focuses on three different categories of heart diseases namely coronary artery disease, cardiovascular disease and cardiomyopathy. The overall objective of our work is to predict more accurately the presence of heart disease. In this thesis more input attributes obesity and smoking are used to get more accurate results. Three data mining classification techniques were applied namely Decision trees, Naive Bayes & Support vector machine. From results it has been seen that SVM provides accurate results as compare to Decision trees & Naive Bayes.

REFERENCES

[1] Frawley and G. Piatetsky -Shapiro, Knowledge Discovery in Databases: An Overview. Published by the AAAI Press/ The MIT Press, Menlo Park, C.A 1996.

[2] Yanwei, X.; Wang, J.; Zhao, Z.; Gao, Y., "Combination data mining models with new medical data to predict outcome of coronary heart disease". Proceedings International Conference on Convergence Information Technology 2007, pp. 868 – 872.

[3] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008

[4] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1 (January - June 2007).

[5] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007.

[6] Shantakumar B.Patil, Y.S.Kumaraswamy "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network". ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656.

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

[8] 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.

[9] Franck Le Duff, Cristian Munteanb, Marc Cuggiaa, Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in health technology and informatics, Vol. 107, No. Pt 2, pp. 1256-9, 2004.

[10] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol. 3, No. 3, 2008.

[11] Dr. Yashpal Singh, Alok Singh chauhan "Neural Networks in data mining" Journal of Theoretical and Applied Information Technology , 2005 - 2009 JATIT.

[12] N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra, "Performance Analysis of Classification Data Mining Techniques over Heart Disease Data base" [IJESAT] international journal of engineering science & advanced technology ISSN: 2250–3676, Volume-2, Issue-3, 470 – 478

[13] Blake, C.L., Mertz, and C.J.: "UCI Machine LearningDatabases", Cleveland heart disease dataset"

[14] E.P. Ephzibah, "A Hybrid Genetic-Fuzzy Expert System for Effective Heart Disease Diagnosis" D.C. Wyld et al. (Eds.): ACITY 2011, CCIS 198, pp. 115–121, 2011. © Springer-Verlag Berlin Heidelberg 2011

[15] Esra Mahsereci Karabulut & Turgay İbrikçi "Effective Diagnosis of Coronary Artery Disease Using The Rotation Forest Ensemble Method" June 2011 / Accepted: 30 August 2011 / Published online: 13 September 2011 # Springer Science+Business Media, LLC 2011

[16] V.V.Jaya Rama krishniah, D.V.Chandra Sekar, Dr.K.Ramchand H Rao, "Predicting the Heart Attack Symptoms using Biomedical Data Mining Techniques" Volume 1, No. 3, May 2012 ISSN – 2278-1080 The International Journal of Computer Science & Applications (TIJCSA)

[17] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006

[18] S.Sarumathi, N.S.Nithya, "Effective Heart Disease Prediction System Using Frequent Feature Selection Method" International ournal of Communications and Engineering Volume 01– No.1, Issue: 01 March2012

[19] Mrs.G.Subbalakshmi, Mr. K. Ramesh, Mr. M. Chinna Rao, "Decision Support in Heart Disease Prediction System using Naïve Bayes" Indian Journal of omputer Science and Engineering (IJCSE), ISSN: 0976-5166 Vol. 2 No. 2 Apr-May 2011

[20] R. Sumathi, E. Kirubakaran "Enhanced Weighted K-Means Clustering Based Risk Level Prediction for Coronary Heart Disease" European Journal of Scientific Research ISSN 1450- 216X Vol.71 No.4 (2012), pp. 490-500 © Euro Journals Publishing, Inc. 2012

[21] Dr. K. Usha Rani "Analysis of Heart Diseases Dataset Using Neural Network Approach" (IJDKP) Vol.1, No.5, September 2011

