Heart Disease Prediction System using Data Mining Clustering Techniques

Meenu Singla
Department of computer science
Punjabi University,
Patiala, India

Kawaljeet Singh
Department of computer science
Punjabi University,
Patiala, India

ABSTRACT
Medical errors are generally costly and harmful. They caused a large number of deaths worldwide annually. A clinical decision support system offers the opportunity to reduce medical errors and also to improve patient safety. Certainly one of the most crucial aspect in applying such a systems is the diagnosis and therapy for heart diseases. This is because statistics demonstrate that a heart disease is one of the premiere factor behind deaths throughout the world. Data mining techniques are quite effective in designing clinical support systems and having the ability to discover hidden patterns and relationships in medical data. Till now, Data mining classification techniques is implemented to analyze the different kinds of heart based problems. This paper is aimed at developing a heart disease prediction system using data mining clustering techniques.

Keywords
Heart Disease Prediction, Data Mining Clustering Techniques, WEKA Tool

1. INTRODUCTION
As the time moves, world is changing rapidly and people want to live a very luxurious life, therefore they work like a device in order to get lot of money and live a relaxed life. But, they forget to take care of themselves; their entire lifestyle is changing as their food habits are changing. In this kind of lifestyle, they become tensed and having blood pressure or sugar problem. It leads to a major threat called heart disease which is the most essential organ which also affects the other human body parts. A number of factors exist which are helpful in prediction of heart disease such as smoking, hyper tension, blood pressure poor diet, high blood cholesterol, obesity, physical inactivity, family history etc. In several cases, diagnosis is generally based on current test results of the patients and experience of the medical doctor. Hence the diagnosis becomes a difficult task that will require much experience and high skill [1].

The World Health Statistics 2012 report enlightens the proven fact that one in three adults worldwide have raised blood pressure level – a condition that causes around 1 / 2 of all deaths from stroke and heart diseases. A major challenge facing healthcare industry is quality of service. It means correctly diagnosing the disease and administering treatments that are effective. A huge amount of healthcare data is gathered by the healthcare industry which is, unfortunately, not "mined" to discover hidden information for effective decision making [2]. Machine learning techniques are utilized intensively in the field of medication for the prediction of diseases like cardiovascular disease, lung, cancer carcinoma of the breast etc [3]. Effective and efficient automated heart problem prediction systems could be beneficial in healthcare sector for coronary disease prediction. This automation will even reduce the amount of tests to be taken by the patient. Hence, it will eventually save besides cost and also the time of both, analysts and patients [4]. Our work attempts to present a detailed study about the different data mining clustering techniques for heart disease prediction which is often deployed in these automated systems.

The remainder of this paper is structured as follows: Section 2 presents the taxonomy of heart disease prediction system. To attain the objectives, a research model is described in section 3. Section 4 discusses the literature work done by various researchers in the field of heart disease prediction. Proposed algorithm is presented in section 5. At last section 6 provides the conclusions and future work.

2. TAXONOMY OF HEART DISEASE PREDICTION SYSTEM
Various heart disease prediction systems have been developed which are shown as below.

2.1 Perceptron based heart disease prediction system
A multilayer perceptron (MLP) is based on decision support system for the diagnosis of heart diseases. The MLP system has used back propagation as a learning algorithm. Three different evaluation methods cross validation, holdout and bootstrapping, have been adopted to assess the performance of the system. Analysis showed that system achieved high diagnosis accuracy (>90%) [5].

2.2 Coactive neuro-fuzzy inference system using genetic algorithm
A coactive neuro fuzzy inference system (CANFIS), combined the neural network and fuzzy logic approach. It is then integrated with genetic algorithm to diagnose the presence of the heart disease [6].

2.3 Clinical decision support system
A clinical decision support system (CDSS) is developed for diagnosis of cardiovascular heart disease by combining four classifiers i.e. support vector machine (SVM), artificial neural networks (ANN), decision trees (DT) and bayesian networks (BN) [7].

2.4 Intelligent heart disease prediction system
An Intelligent Heart Disease Prediction System (IHDPS) has been developed by using three data mining techniques, decision trees, naïve bayes and neural network techniques. Naïve Bayes achieves the highest accuracy [8].

2.5 Clinical decision support system using fuzzy rules

A weighted fuzzy rule-based clinical decision support system (CDSS) is proposed for the diagnosis of heart disease. The weighted fuzzy rules are applied on the rule base of the fuzzy system before carrying out prediction on the designed fuzzy-based CDSS [6].

3. RESEARCH MODEL

To attain the objective, step-by-step research methodology is used in this dissertation. The different phases used to achieve this work are shown in figure 1.

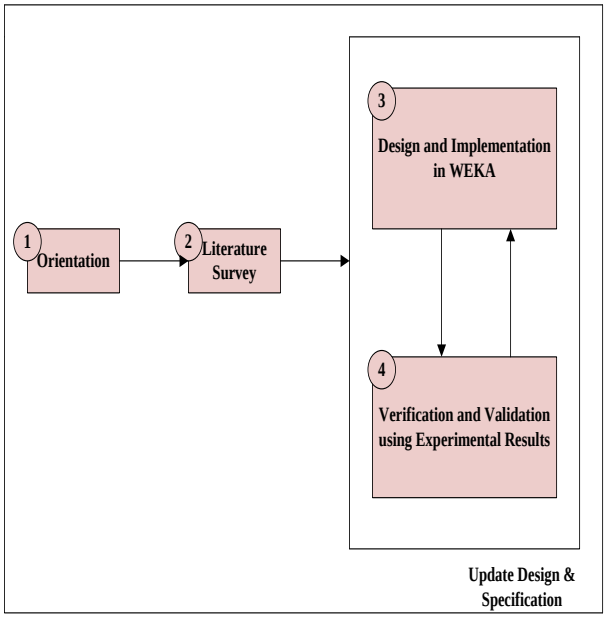


Fig 1: Research model

3.1 Orientation

This research work starts with the orientation in the field of data mining. By consulting websites, reading articles, participating in seminars and discussing data mining techniques and the research questions of this research are formulated. To achieve high quality information, this research employs a structured method known as literature review.

3.2 Literature survey

To explore the available knowledge on the area of data mining, role of data mining techniques in heart disease prediction, a literature review will be conducted using a systematic approach.

3.3 Simulation environment

An appropriate simulation environment will be made based upon which the proposed algorithm is developed and it will provide detail of investigational set-up and the outcomes of the model in various circumstances.

3.4 Experimental results

By applying proposed algorithm, experimental results of different algorithms will be analyzed.

3.5 Performance analysis

A comparison of different methods will be done to know about which methodology is the best. Comparisons table and diagrams will be made based upon the outcomes of the experimental results.

4. LITERATURE SURVEY

Banu et al. (2013) applied maximal frequent item set algorithm (MAFIA) for heart disease prediction. In this, data is estimated using entropy based cross validation and partition. C4.5 algorithm is used as the training algorithm to show rank of heart attack with the decision tree [3].

Amin et al. (2013) compared the performance of six clinical decision support system (CDSS) which use different data mining techniques for heart disease prediction and diagnosis [9].

Pattekari et al. (2012) developed an intelligent system using naive bayes for the prediction of heart disease. The proposed system answers difficult queries, each one having its own strength and have access to detailed information and accuracy [10].

Dangare et al. (2012) analyzed prediction systems for heart disease using more number of inputs attributes i.e. obesity and smoking by implementing decision trees, naive bayes and neural networks. Analysis showed that neural networks achieve highest accuracy [1].

Bhatla et al. (2012) implemented three classification techniques neural network, decision tree and genetic algorithm for the prediction of heart disease. The result shows that neural networks outperformed over all other data mining techniques [4].

Palaniappan et al. (2008) developed an intelligent heart disease prediction system (IHDPS) using naive bayes, neural network and decision trees. It was web-based, user friendly & expandable [8].

Lee et al. (2007) proposed a technique by implementing several classifiers i.e. support vector machine (svm), decision tree, classification based on multiple association rule (CMAR), bayesian classifier [2].

Guru et al. (2007) proposed the system that predicts sugar, heart disease and blood pressure by using neural networks. The neural network back propagation algorithm is implemented for training and testing of data [11].

Ordonez et al. (2004) identified the problem of constrained association rules for heart disease prediction. The resultant dataset contains records of heart disease patients [12].

5. PROPOSED ALGORITHM

A new algorithm is proposed which provides much better results and an improvement over existing methods. In this section, we elaborate the whole framework of our new approach as shown in figure 2.

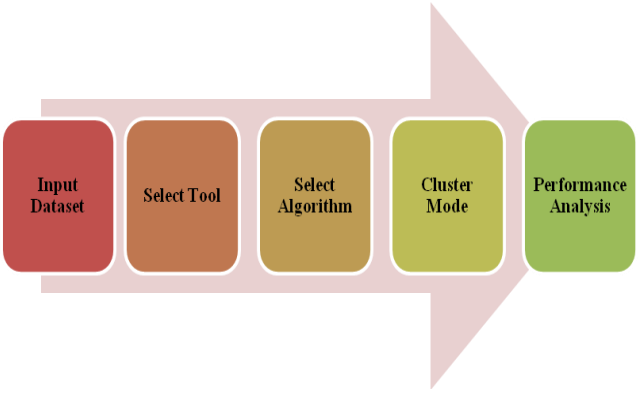


Fig 2: Block diagram of proposed technique

5.1 Input and read the dataset

Firstly, dataset selected is pima-diabetes dataset [13]. There are 768 classified instances and 9 attributes in this dataset, which are used to evaluate performance or to predict the heart disease.

5.2 Select tool

After selecting the dataset, it is loaded into the WEKA tool [14].

5.3 Select algorithm

When the dataset is loaded, then subsequent step is to apply clustering algorithm heart disease prediction and diagnosis.

5.4 Cluster mode

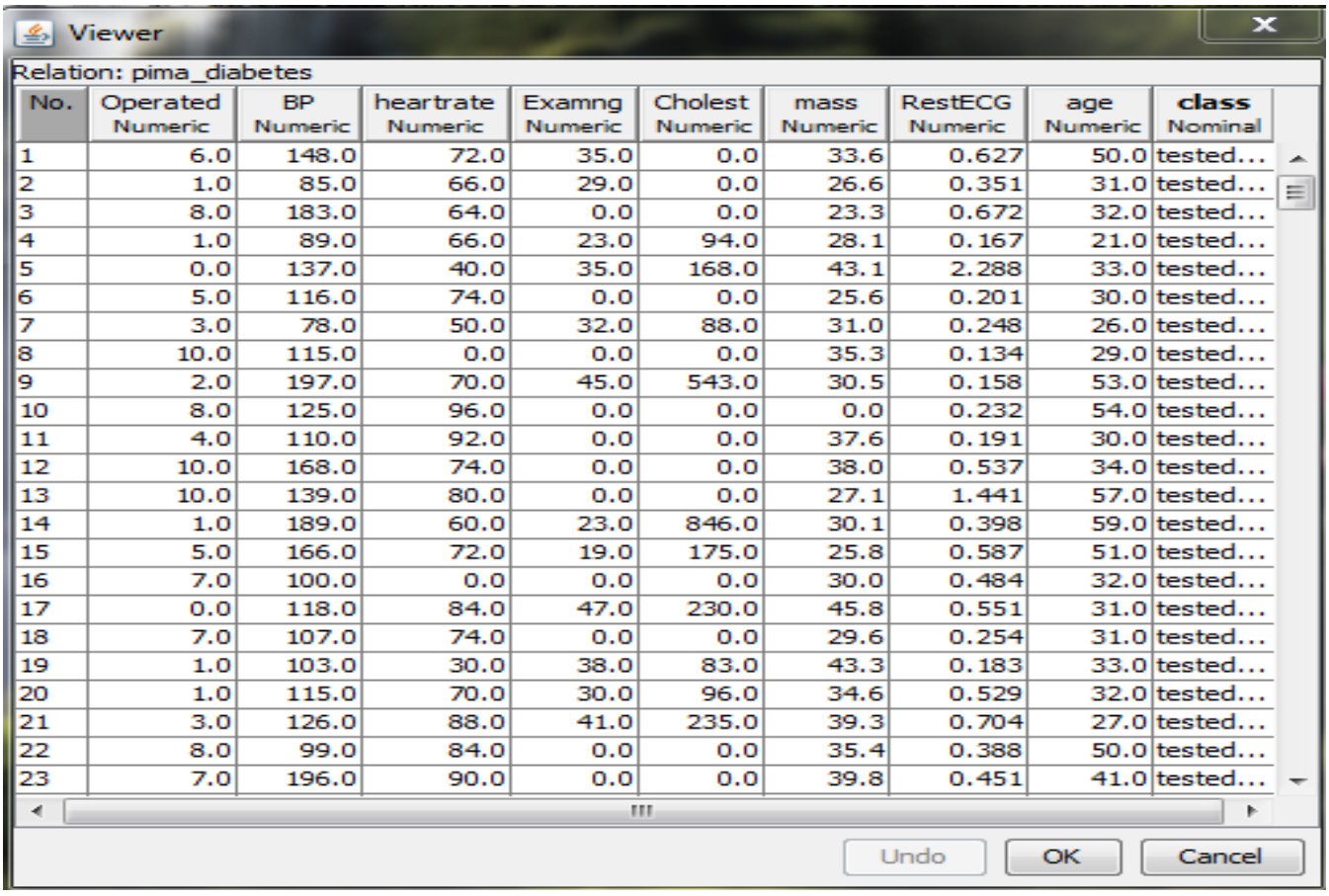
In order to have a good measure of the performance of the clustering algorithms, the cluster mode is repeatedly performed in the WEKA tool.

5.5 Performance analysis

The next step is to evaluate and compare the performance of different algorithms.

6. RESULTS AND DISCUSSION

In order to do performance analysis, different metrics are considered to evaluate the performance of proposed technique. The overall objective of this study is to predict the presence of heart disease accurately in minimum time. Java code has been used for the implementation of the clustering algorithm and then used WEKA [13] for simulation of the model. For experimental work, pima diabetes dataset [12] is implemented on WEKA [13] to evaluate the performance of proposed algorithm i.e. simple k-means, EM, farthest first. Pima diabetes dataset contains 768 instances. Each patient is characterized with 9 attributes from which 8 are numeric and the last is nominal having two values tested-positive and tested-negative as shown in figure 3. All the instances are classified according to the values of their feature vectors. Attributes histogram is defined in figure 4 which shows the distribution of all attributes for the classes tested negative and tested positive. Tested negative and tested positive shows the presence and absence of heart disease respectively.



No.	Operated Numeric	BP Numeric	heartrate Numeric	Examng Numeric	Cholest Numeric	mass Numeric	RestECG Numeric	age Numeric	class Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	tested...
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	tested...
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	tested...
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	tested...
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	tested...
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	tested...
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	tested...
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	tested...
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	tested...
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	tested...
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	tested...
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	tested...
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	tested...
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	tested...
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	tested...
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	tested...
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	tested...
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	tested...
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	tested...
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	tested...
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	tested...
22	8.0	99.0	84.0	0.0	0.0	35.4	0.388	50.0	tested...
23	7.0	196.0	90.0	0.0	0.0	39.8	0.451	41.0	tested...

Fig 3: Dataset description

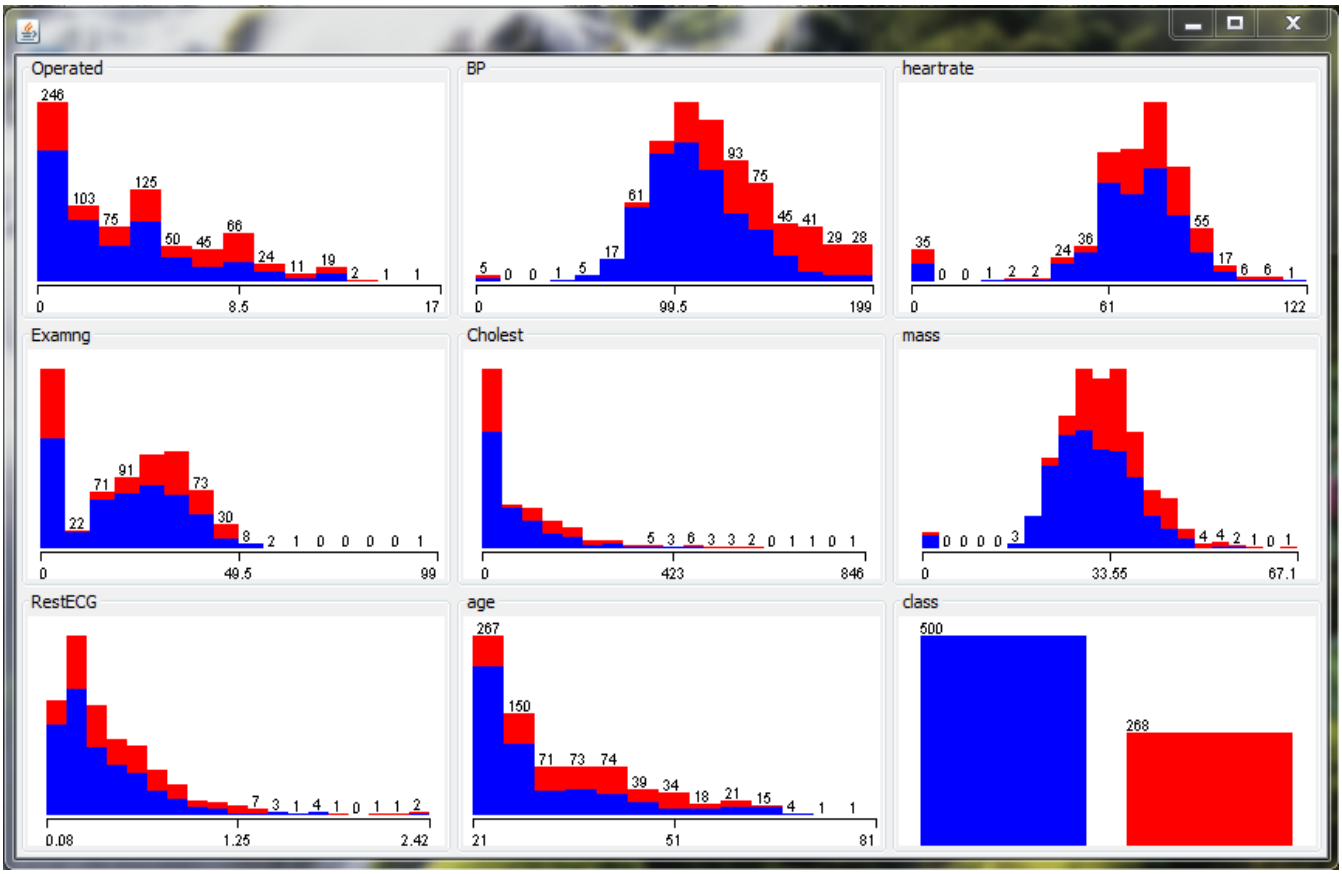


Fig 4: Attribute description

The statistical values of all attributes have been described as shown in table 1. The statistical values include minimum, maximum, mean deviation and standard deviation of all attributes. After getting the results of different clustering algorithms in WEKA [13] tool, the accuracy of different algorithms is compared by considering the three evaluation parameters. These parameters are no. of clusters, ratio of cluster and time taken. A comparative analysis of these algorithms is presented in table 2.

Table 1. Attributes statistical values

Attribute No.	Name	Min Value	Max Value	Mean Value	Standard Deviation
1.	Operated	0	17	3.845	3.37
2.	BP	0	199	120.895	31.973
3.	Heart Rate	0	122	69.105	19.356
4.	Examng	0	99	20.536	15.952
5.	Cholest	0	846	79.799	115.244
6.	Mass	0	67.1	31.993	7.884
7.	RestECG	0.078	2.42	0.472	0.331
8.	Age	21	81	33.241	11.76

Table 2. Performance analysis

Algorithm	No. of Clusters	Ratio to Each Cluster	Model Construction Time
K-Mean	2	65:35	0.03 Seconds
EM Technique	3	30:26:44	11.87 Seconds
Farthest First	2	80:20	0.02 Seconds

7. CONCLUSION AND FUTURE SCOPE

This research work presets the various clustering techniques, k-mean, EM and the farthest first algorithm for the prediction of heart disease. Result shows that farthest first clustering algorithm is the best algorithm as compared to other algorithms. Because the ratio of correctly classified instances to the cluster is maximum and the time taken to build the model is minimum. This system can be further expanded. It can use more number of input attributes and it can be further expanded by increasing the no. of the clusters. The same experiment can also be performed on other data mining tool such as R. Also the ensembling of classifiers can also be done to evaluate their performance with the individual classifiers. Above algorithms can be applied to other datasets in order to observe whether the same algorithm gives the highest accuracy or not.

8. REFERENCES

[1] Dangare, C.S., Apte, S.S., Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, International Journal of Computer Applications, 47(10), pp. 0975-888, (June 2012).

[2] Lee, H.G., Noh, K.Y., & Ryu, K.H., Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV, Emerging Technologies in Knowledge Discovery and Data Mining, pp. 218-228, (May 2007).

[3] Banu, M.A.N., Gomathy, B., Disease predicting system using data mining techniques, International Journal of Technical Research and Applications, 1(5), pp. 41-45, (2013).

[4] Bhatla, N., Jyoti, K., An Analysis of Heart Disease Prediction using Different Data Mining Techniques, International Journal of Engineering Research & Technology (IJERT), 1(8), pp. 1-4, (October 2012).

[5] Yana, H., Jiangb, Y., Zhenge, J., Pengc, C., & Lid, Q., A multilayer perceptron-based medical decision support system for heart disease diagnosis, Expert Systems with Applications, 30, pp. 272–281, (2006).

[6] Parthiban, L., Subramanian, R., Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm, International Journal of Biological and Medical Sciences, 3(3), pp. 157-160, (2008).

[7] Eom, J.H., Kim, S.C., & Zhang, B.T., AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction. Expert Systems with Applications, 34(4), pp.2465-2479, (2008).

[8] Palaniappan, S., Awang, R., Intelligent Heart Disease Prediction System Using Data Mining Techniques, International Journal of Computer Science and Network Security, 8(8), pp. 343-350, (August 2008).

[9] Amin, S.U., Agarwal, K., & Beg, R., Data Mining in Clinical Decision Support Systems for Diagnosis, Prediction and Treatment of Heart Disease, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2(1), pp. 218-223, (January 2013).

[10] Pattekari, S. A., Parveen, A., Prediction system for heart disease using naive bayes, International Journal of Advanced Computer and Mathematical Sciences, 3(3), pp. 290-294, (2012).

[11] Guru, N., Dahiya, A., & Rajpal, N., Decision Support System for Heart Disease Diagnosis Using Neural Network, Delhi Business Review, 8(1), pp. 99-101, (2007).

[12] Ordonez, C., Improving Heart Disease Prediction Using Constrained Association Rules, Seminar Presentation at University of Tokyo, (2004).

[13] Pima-diabetes Dataset [online]. Available: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes [Accessed 3 May 2015].

[14] Weka [online]. Available: http://www.cs.waikato.ac.nz/ml/weka/ [Accessed 1 January 2015].