Available online at www.ijrp.org

IJRP.ORG
International Journal of Research Publications

International Journal of Research Publications

Prediction of Heart Disease Using Data Mining Techniques: A Case Study

Shamim Hasnanin Shadid, Ahmed Shafkat, Ms. Fauzia Yasmeen, Sabbir Ahmed Sibli, Md. Rumman Rafi

[1,2]B.Sc in CSE, Bangladesh Army University of Engineering and Technology, Bangladesh. . [3]Assistant Professor,[2,4] Teaching Assistant, [5] Lecturer. [2,3,4,5] Fareast International University, Banani, Dhaka, Bangladesh. . [1]hasnainbauet@gmail.com , [2]a.shafkat@gmail.com, [3]fauzia.cse@fiu.edu.bd, [4]sabbir.cse@fiu.edu.bd, [5]rumman.eee@fiu.edu.bd.

Abstract

Data mining is the process of rearranging through large datasets to identify patterns and establish relationships between them to solve problems through data analysis. Data mining tools allow enterprise to predict future trends. A pattern is useful, interesting and easily understood by human if it is valid for a given test and with some degree of certainty. Though the data amount generated in predicting heart disease is huge and complex advance data mining techniques can process the data. Heart disease is one the disease that causes the maximum causalities. This problem is identified long before but no proper actions been taken to combat this problem. This paper set out goal to finding which method would be best for predicting the diseases using data of four different dataset from four different places. Therefore, this article tries to finding which method would be best for predicting the diseases using data of four different datasets from four different places. This is a comparative study on the efficiency of different data mining techniques such as Decision Tree (DT), K-Nearest Neighbor (kNN), Naive Bayes, Logistic Regression in predicting heart diseases. The Data Mining techniques are analyzed and the accuracy of prediction is noted for each method used. The result showed that heart diseases can be predicted with accuracy of above 80%.

Keywords: Decision Tree, K-Nearest Neighbor, Naive Bayes, Logistic Regression.

1. Introduction

Nowadays, data mining is established as a novel field of extracting hidden patterns from extensively huge dataset. It can be used to take certain decisions, estimate and predict using different algorithms. Medical science

is a field where large data is generated from different patient's symptoms and clinic reports. Heart disease is a prevalent problem in modern world. It is also one of the lethal diseases in third world country. In today's world most of the data in medical sector is computerized but not utilized properly. It is stacked up in a database like old handwritten records and put to no use. This data can be harnessed to predict diseases such as Cancer, Cardiovascular Diseases and many more. Data mining techniques are used to predict different stages of cancer by using the different cancer cell photos for each stage. Similarly, heart disease can be predicted using different factors which include family history, cholesterol, diabetes, exercise, chest pain etc. The diagnosis of this disease is a long process and it should be diagnosis properly and correctly. Due to limitations of medical experts and lack of data they put their patient at risk. This paper will show the effectiveness of DT, Naive Bayes, Logistic Regression and K- Nearest Neighbors algorithms. It was noticed that a technique does not always work for a given scenario and differs due to the selected attributes or size of the data. This paper aims to extract hidden pattern and find out best methods for predicting heart disease rate using data mining techniques in certain conditions that will help to diagnosis. The article observed some of the techniques that work very efficiently with a certain scenario. This paper will provide the selective knowledge to build smart heart disease prediction system by comparing methods with each other.

2. Literature Review

A huge number of studies have been done on heart disease. Different study use different datamining techniques for predicting heart disease. This paper analyzes the different datamining techniques which are used in recent past years for predicting heart disease. Some papers use various datamining techniques and other use only one technique for the diagnosis of heart disease.

Andrea D'souza analyse the different datamining techniques to find useful one for medical analyst to diagnosis heart disease. This paper use aprori techniques for generation of frequent item set and K Means Clustering Algorithm, Artificial Neural Network. Finally a study was performed to find the most useful one.

Chaitrali S.Dangare et.al use there datamining techniques Neaural Networks, Decision Trees and Naïve Bayes. They built Intelligent Heart disease Prediction System for diagnosis heat disease using historical heart database. Medical terms like age, sex, blood pressure like 13 attributes are used to develop the system. For more accurate result they use two more attributes obesity and smoking and considered as important attributes.

V. Manikandan et.al this paper used association rule mining to extract the relations of item sets. Mafia algorithms are used to classify the data which resulted in better accuracy. Cross validation and partition techniques were used for data evaluation and result was compared. 19 attributes were used with MAFIA Algorithm.

K. Srinivas et.al they made an application in healthcare and heart disease prediction using datamining techniques. Naïve Bayes, Decision Tree and Artificial Neural Network were used for potential classification of large volume of heart disease data. Tanagra data mining tool was used 14 attributes and 3000 instance was in the training set. Different types of testing results represented by the instances to predict more accurately heart disease. Cross-Validations were used for comparing the results. In comparison result Naïve Bayes showed the best performance.

Jyoti Sonia,et.al To diagnose the presence of heart disease in patient they used three classifiers Naïve Bayes, Decision Tree and Classification via clustering.The process of grouping the similar elements is called clustering.

They used WEKA 3.6.0 tool for datamining and dataset contained 909 records and 13 different attributes. Genetic search was incorporated and all attributes were made categorical and inconsistency was resolved to enhance the prediction result. Decision showed the better result than two other techniques.

Nidhi Vatla et.al In this paper an intelligent Heart Disease Prediction System was propose. They used different data mining techniques like Artificial Neural Network, Decision Tree and Naïve Bayes.

3. Methodology

This work used two different platforms for analyzing data from various datasets. The one this Spyder (python 3.6), Jupyter (python 3.6) and another is Rapid Miner. The four different supervised machine learning algorithms, i.e. Naïve Bayes, K-Nearest Neighbor, Decision Tree are used to analyzed the datasets. Publicly available 4 datasets are used from UC Irvine School of Information and Computer Science (UCI) heart disease directory.
Creators:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Attributes:

Table 1: Description of 14 attributes

Serial No	Attributes	Description
1	Sex	Male or Female
2	Age	Age in years
3	Cp	Chest pain type
4	Thal	Segment
5	Chol	Serum cholesterol
6	Slope	Relative to rest
7	Thestbps	Resting blood pressure
8	Ca	Slope of the peak exercise ST
9	Exang	Exercise induce angina
10	Restecg	Resting electrographic result
11	oldpeak	ST depression induced by exercise
12	Fbs	Fasting blood sugar
13	Thalach	Maximum heart rate achieved
14	Num	The predicted attribute

Key attributes:

Patient ID: Patient’s Identification Number

Predictable attribute:

Diagnosis: Value 1 = < 50 % (no heart disease)

Value 0 = > 50 % (has heart disease)

Initially dataset contained some fields, in which some value in the records was missing. These were identified and replaced with most appropriate values using Replace Missing Values filter from Weka. The Replace Missing Values filter scans all records & replaces missing values with mean mode method. This process is known as Data Pre-Processing. Figure 3.1 shows an example of raw dataset and figure 3.2 is represent a dataset after cleaning procedure.

	Age	A	B	C	D	E	F	G	H	I	J	K	L	M
171	57	0	1	130	308	0	0	98	0	1.0	2	?	?	0
200	43	1	4	150	247	0	0	130	1	2.0	2	?	?	1
31	39	1	2	120	?	0	1	146	0	2.0	1	?	?	0

Figure 3.1: Dataset before cleaning

	Age	A	B	C	D	E	F	G	H	I	J	K	L	M
32	39	1	2	120	200	0	0	160	1	1.0	2	0	0	0
106	49	1	3	140	187	0	0	172	0	0.0	0	0	0	0
214	51	0	4	160	303	0	0	150	1	1.0	2	0	0	1

Figure 3.1: Dataset after cleaning

At first, this work applied Naïve Bayes algorithm to the datasets. Naïve Bayes classifier works on the basis of Bayes theorem. Naïve Bayes classifier assumes that an attribute value on a given class is independent of values of other attributes.

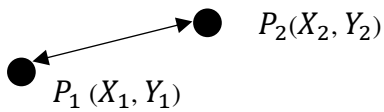
$$P(H|X) = P(X|H)\,P(H) / P(X). \qquad (I)$$

By applying the algorithm to Cleveland dataset, the result is shown in figure 3.3.

Name	Type	Size	Value
X	object	(303, 13)	array([[63, 1, 1, ..., 3, '0', '6'], [67, 1, 4, ..., 2, '3', '3'], ...
X_test	object	(61, 13)	array([[59, 1, 1, ..., 2, '0', '7'], [39, 0, 3, ..., 1, '0', '3'], ...
X_train	object	(242, 13)	array([[57, 1, 3, ..., 1, '1', '7'], [56, 1, 4, ..., 2, '1', '3'], ...
accuracy	float64	1	0.098360655737704916
df	DataFrame	(303, 14)	Column names: Age, A, B, C, D, E, F, G, H, I, J, K, L, M
y	int64	(303,)	array([0, 2, 1, ..., 3, 1, 0], dtype=int64)
y_test	int64	(61,)	array([1, 0, 0, ..., 0, 1, 0], dtype=int64)
y_train	int64	(242,)	array([0, 1, 1, ..., 3, 4, 0], dtype=int64)

Fig 3.3: Implementation of Naïve Bayes on Cleveland dataset.

After that, this work focuses on K-Nearest Neighbor algorithm for same conditions. kNN is a simple algorithm that stores all possible available cases and classifies new cases based on a similarity measure e.g. distance function. There are many methods are used for measuring distance such as Euclidian distance measure. K-Nearest Neighbor also used for statistical estimation and pattern recognition.



Euclidian Distance: $D=\sqrt{(X_2 - X_1)^2 - (Y_2 - Y_1)^2}$ *(II)*

After applying the kNN theorem to Cleveland dataset figure 3.4 has been found.

Name	Type	Size	Value
X	object	(303, 13)	array([[63, 1, 1, ..., 3, '0', '6'], [67, 1, 4, ..., 2, '3', '3'], ...
X_test	object	(61, 13)	array([[59, 1, 1, ..., 2, '0', '7'], [39, 0, 3, ..., 1, '0', '3'], ...
X_train	object	(242, 13)	array([[57, 1, 3, ..., 1, '1', '7'], [56, 1, 4, ..., 2, '1', '3'], ...
accuracy	float64	1	0.49180327868852458
df	DataFrame	(303, 14)	Column names: Age, A, B, C, D, E, F, G, H, I, J, K, L, M
y	int64	(303,)	array([0, 2, 1, ..., 3, 1, 0], dtype=int64)
y_test	int64	(61,)	array([1, 0, 0, ..., 0, 1, 0], dtype=int64)
y_train	int64	(242,)	array([0, 1, 1, ..., 3, 4, 0], dtype=int64)

Fig 3.4: Implementation of kNN on Cleveland dataset.

Then, this work used Logistic Regression on the dataset. Logistic Regression is a classifier used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous value, logistic regression transforms its output by using the logistic sigmoid function to returned a discrete probability value which can then be mapped to two or more discrete classes. The relationship between the categorical dependent variable and one or independent variables is measured in logistic regression by estimating probabilities using a logistic/sigmoid function. The implementation result of this algorithm is shown in figure 3.5.

$Y=1/1+e\ [-(c+x_1 * w_1+x_2 * w_2+\ldots\ldots\ldots\ldots..+x_e * w_e)]$
(III)

Here, the output is binary.

Name	Type	Size	Value
X	object	(303, 13)	array([[63, 1, 1, ..., 3, '0', '6'], [67, 1, 4, ..., 2, '3', '3'], ...
X_test	object	(61, 13)	array([[59, 1, 1, ..., 2, '0', '7'], [39, 0, 3, ..., 1, '0', '3'], ...
X_train	object	(242, 13)	array([[57, 1, 3, ..., 1, '1', '7'], [56, 1, 4, ..., 2, '1', '3'], ...
accuracy	float64	1	0.5901639344262295
dataset	DataFrame	(304, 12)	Column names: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13
df	DataFrame	(303, 14)	Column names: Age, A, B, C, D, E, F, G, H, I, J, K, L, M
y	int64	(303,)	array([0, 2, 1, ..., 3, 1, 0], dtype=int64)
y_pred	int64	(61,)	array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
y_test	int64	(61,)	array([1, 0, 0, ..., 0, 1, 0], dtype=int64)
y_train	int64	(242,)	array([0, 1, 1, ..., 3, 4, 0], dtype=int64)

Fig 3.5: Implementation of logistic regression on Cleveland dataset.

Finally, the decision tree approach is more considered and powerful for classification problems. There are two steps in these techniques first step is building a tree and then applying the tree to the dataset. CART, ID3, C4.5, CHAID, and J48 are popular decision tree algorithms. ID3 algorithm begins with the original set as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set and calculates the entropy (or information gain) of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set is then split or partitioned by the selected attribute to produce subsets of the data. The algorithm continues to recur on each subset, considering only attributes never selected before. By applying ID3 on the datasets, the figure of Heat map (figure: 3.6), Swarm plot (figure: 3.7), Histogram of Swarm plot (figure: 3.8) is shown below consequently.

Fig 3.6: Heat map of Hungarian Dataset

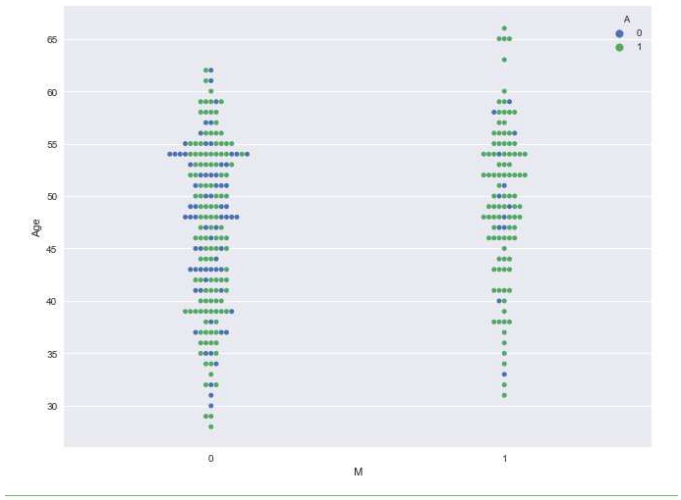


Fig 3.7: Swarm plot of Hungarian dataset

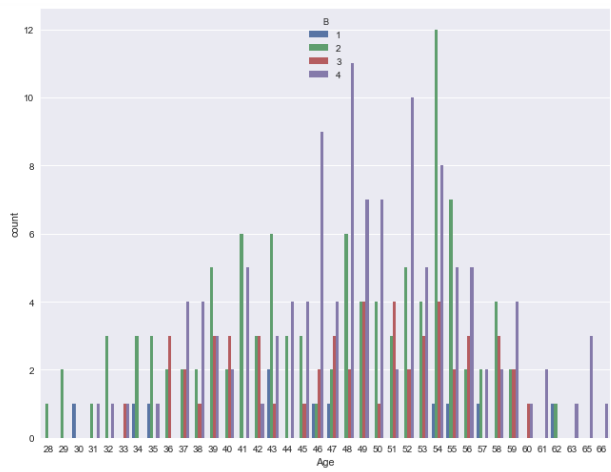


Fig 3.8: Histogram of swarm plot

After implementing each algorithm for different datasets like above procedures, this paper finds different algorithms shows better result for different cases. Next four figurers (Figure 3.9, 3.10, 3.11, 3.12) will shows the implementation result of different algorithms in this four datasets.

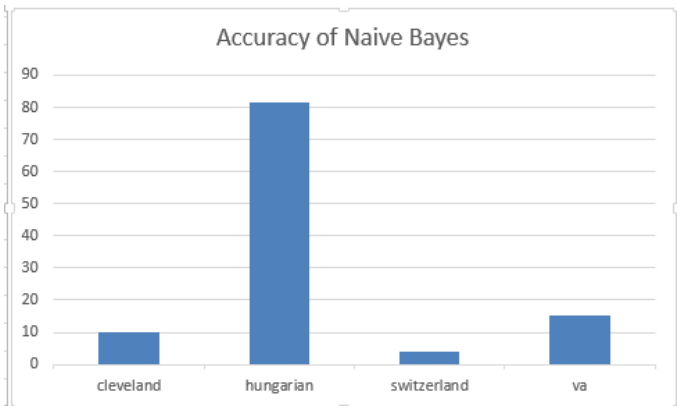


Fig 3.9: Accuracy curve of Naïve Bayes

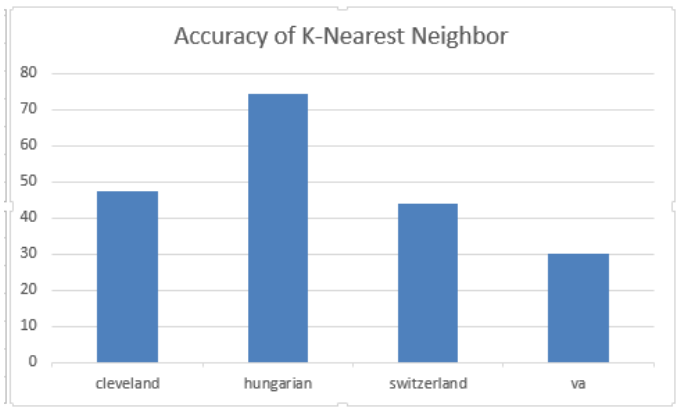


Fig 3.10: Accuracy Curve of K-Nearest Neighbor

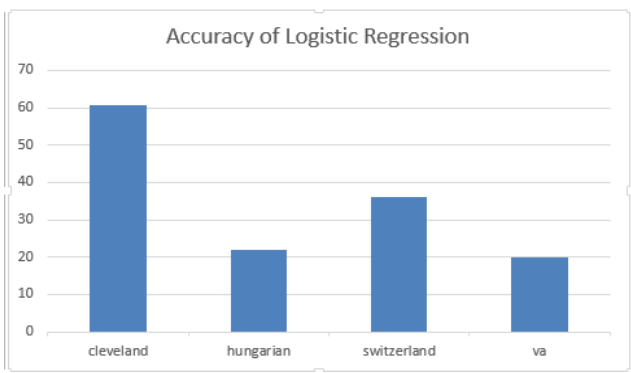


Fig 3.11: Accuracy Curve of Logistic Regression

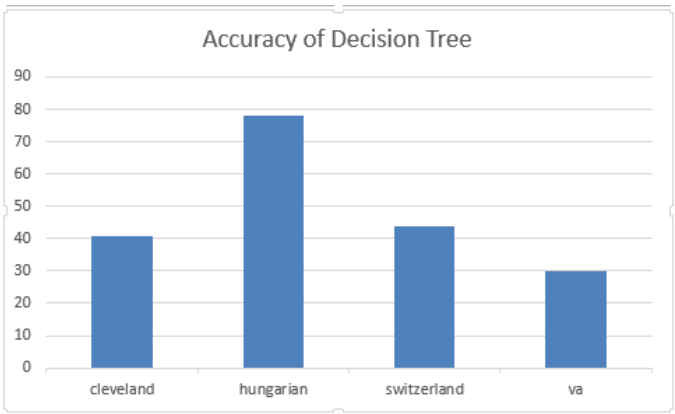


Fig 3.12: Accuracy Curve of Decision Tree

4. Result Analysis

Here, in this work we show the accuracy of different algorithms on various datasets form better understanding which algorithms works better on which situation.

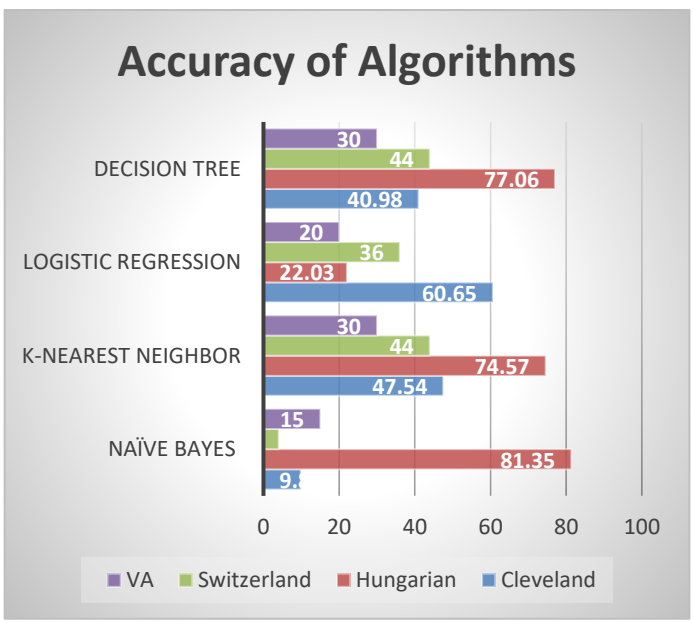


Figure 4.1: Accuracy of Algorithms

We also calculate the accuracy of different datasets by applying four supervised machine learning algorithms to show how all these publicly available UCI heart disease datasets works on those algorithms. This will give clear view to others when they will work on these datasets. This work proved single algorithm is not enough sufficient for all kinds of datasets. Different algorithm shows better performance for various kinds of dataset.

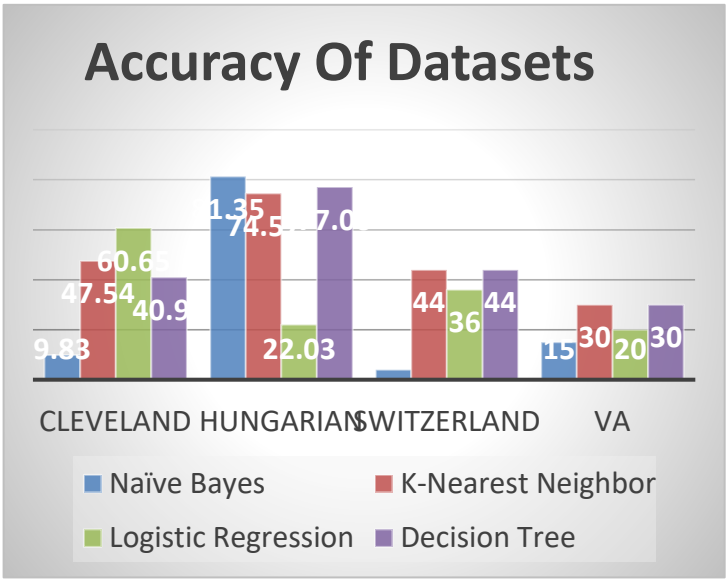


Figure 4.2: Accuracy of Dataset

5. Conclusion

Most of the data in today's world is computerized, these are usually scattered and not properly utilized. These data if analyzed for hidden patterns can be put to good use. Therefore, it has become a wide area for research with increasing importance and facilities. The main motive of this research was to create a basic data mining which can be used to predict heart diseases and also to find the efficiency of the data mining in this particular data set by the four chosen algorithm i.e. Naïve Bayes, K-Nearest Neighbour, Logistic Regression and Decision Tree. The performance of each algorithm depends on datasets. Some theories provide better results for a particular dataset. Data mining can be used to build a software which help predict heart diseases. This software could be used by any non-medical employee of the hospital making it easy for the patients and saving time for the doctors.

In future two or more techniques can be combined together for better performance and compare it with existing ones. Then it will show the differences between existing algorithm and hybrid algorithm so one could have clear idea which is going to be the best approach for his or her study.

References

1. Syed Immamul Ansarullah, Pradeep Kumar Sharma, Abdul Wahid, Mudasir M Kirmani "Heart Disease Prediction System using Data Mining Techniques: A study" International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 08 , Aug-2016.
2. Abhishek Taneja "Heart Disease Prediction System Using Data Mining Techniques"ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY ISSN: 0974-6471 Vol. 6, No. (4): December 2013.

3. Andrea D'Souza.*"*Heart Disease Prediction Using Data Mining Techniques" International Journal of Research in Engineering and Science (IJRES) ISSN (Online): 2320-9364, Volume 3 Issue 3 ǁ March 15.
4. Nidhi Bhatla, Kiran Jyoti, " An Analysis of Heart Disease Prediction using Different Data Mining Techniques" International Journal of Engineering and Technology Vol.1 issue 8 2012.
5. Beant Kaur and Williamjeet Singh.," Review on Heart Disease Prediction System using Data Mining Techniques", IJRITCC ,October 2014.
6. R. Chitra, Review Of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques; Ictact Journal On Soft Computing, July 2013,volume: 03, Issue: 04.
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. Clevelanddatabase:http://archive.ics.uci.edu/ml/dataets/Heart+Disease.