

International Journal of Computer Science and Mobile Computing



A Monthly Journal of Computer Science and Information Technology

ISSN 2320-088X

IMPACT FACTOR: 7.056

IJCSMC, Vol. 9, Issue. 7, July 2020, pg.136 – 143

Cluster Based Mining for Prediction of Heart Disease

Suresh Kumar Pandey*; **Dr. Bharat Mishra***; **Dr. S.S.Gautam***

suresh.kumarpandey@rediffmail.com

*Department of Physical Sciences, Mahatma Gandhi Chitrakoot Gramodaya Vishwavidyalaya, Satna (M.P.)- India

Abstract: DM techniques are being extensively used in the area of healthcare and particularly for prediction of heart disease. Clinical dataset of heart patients has been created, which consists of 2139 instances and 8 attributes. After pre-processing of dataset, we have applied 4 clustering algorithms on this dataset in Weka. The clustering outputs are presented.

1. Introduction

Recent technological developments in the area of computational and storage technology have enabled organizations to store massive amounts of data. This data can be extracted and mined for getting extremely useful information. DM techniques are used to find novel and useful unknown patterns and predicting the outcome of future observation. Presently DM are used in various application areas. Some important application areas in which DM techniques are used includes Banking, Fraud detection, Bio-informatics, Marketing, Insurance, Telecommunication, e-Commerce, Education and Healthcare. Various DM techniques are classification, association, clustering, pattern analysis and prediction.

DM techniques are extremely useful for diagnosis, prognosis and treatment planning by health experts. Heart related diseases are getting significant attention of researchers since it is the prime cause of deaths. In the present work we have used clustering algorithms for heart dataset.

2. Related Work

MA. Jabbar et al. (2011), in their paper discussed about cluster based association rule mining for prediction of heart attack. They have suggested association rule mining based approach. In this approach the sequence number and clustering has been used for heart attack prediction. For their experimental evaluation they have used dataset of UCI machine learning repository- Cleveland heart diseases dataset of 10 patients. There are total 14 attributes selected for discovery association rules. The proposed algorithm has been implemented in C language. They have found that in their algorithm execution time for association rule mining is less and execution time changes as support increases. They have divided database into equal size partitions called clusters. For each cluster they have calculated frequent item sets which recording to them reduces main memory requirement because of consideration of only small cluster at a time. Due to this it becomes scalable and efficient.

Mamta Sharma et al.(2017), have compared different DM techniques for heart diseases prediction. They have described ANN, Naives Bayes theorem and decision trees (ID3, C4.5, CART, J48). The objective of their paper was to explore various decision tree and NN classification algorithms for disease prediction.

Nelofar Rehman (2017), have described data mining processed, DM techniques, DM algorithms and DM tools.

AH Chen et al. (2011), have developed heart disease prediction system. Their system consists multiple features. Different sections of the system are input clinical data, ROC curve display and prediction performance display. The classification accuracy achieved is 70%. The system has been developed in C and C# environment.

Elham Naikookar and Ebrahim Naderi (2018), have investigated hybrid ensemble model for detection and prediction of heart disease. The used dataset have 278 samples from SPECT database. The sensitivity, specificity and classification accuracy reported are 80%, 93% and 96% respectively.

Lovepreet Kaur (2014), has used Fuzzy C-Mean clustering for symptom prediction of heart disease. For their research work they have used 58 records and 14 attributes from Cleveland dataset. Fuzzy c-means clustering and genetic algorithms are used by the author and has achieved 86.6% accuracy.

Various other works in which clustering techniques have been used for heart disease prediction are - **Bala Sundar** et al. (2012), **Nallakaruppan M.K** et al. (2017), **Mirpouya Mirmozaffari** (2017), **Sarangam Kodati and Dr. R. Vivekanandam** (2018) and **Reetu Singh and E. Rajesh** (2019).

3. Cluster Analysis

Cluster analysis is used for data summarization, data compression and finding efficiently nearest neighbors. Different types of clusterings are hierarchical, partitional, exclusive, overlapping, fuzzy, partial and complete. Clustering algorithms are classified as prototype based clustering, density based clustering, scalable clustering algorithms and graph based clustering. Various important factors required to be considered for the selection of appropriate clustering algorithm depends upon type of clustering, characteristics of clusters, characteristics of datasets and attributes, number of data objects, noise & outliers, cluster description and domain specific issues (**Pang-Ning Tan** et al., 2008).

4. Clustering on Heart Dataset

We have created our own heart dataset which has 2139 instances and 8 attributes have been taken for experimental evaluation. Attributes considered are A2 (sex), A3 (age), A4 (chest pain), A5 (blood pressure), A6 (blood sugar), A7 (ecg), A8 (heart rate) and DE (disease). Attribute age(A3) is classified in 3 groups (1-young, 2- Middle age and 3- old age), attribute A4 is categorized in 4 groups(1-Asymptomatic, 2- Angina Pectoris, 3- non-angian, 4-typ-angian) , attribute A5 is classified in 5 groups (1-Normal, 2-elevated, 3-HBP Stage1, 4-HBP Stage2, 5-Hypertension Crisis), attribute A6 is classified in 3 groups (1-Normal, 2-pre diabetic, 3-diabetic), attribute A7 is classified in 3 groups (1- left ventricular hypertrophy, 2-normal, 3-st-t-wave-abnormality), attribute A8 is classified in 3 groups(1-Good, 2-average, 3-poor) and attribute DE is classified in 2 groups (1 negative and 2 positive).

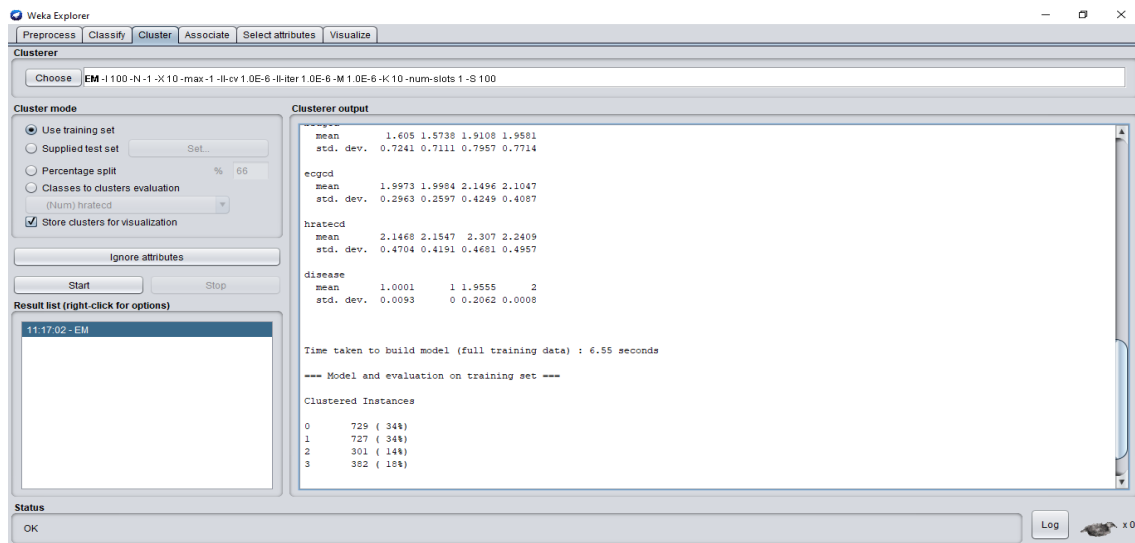


Figure 1: Clustering - EM (Expectation Maximisation)

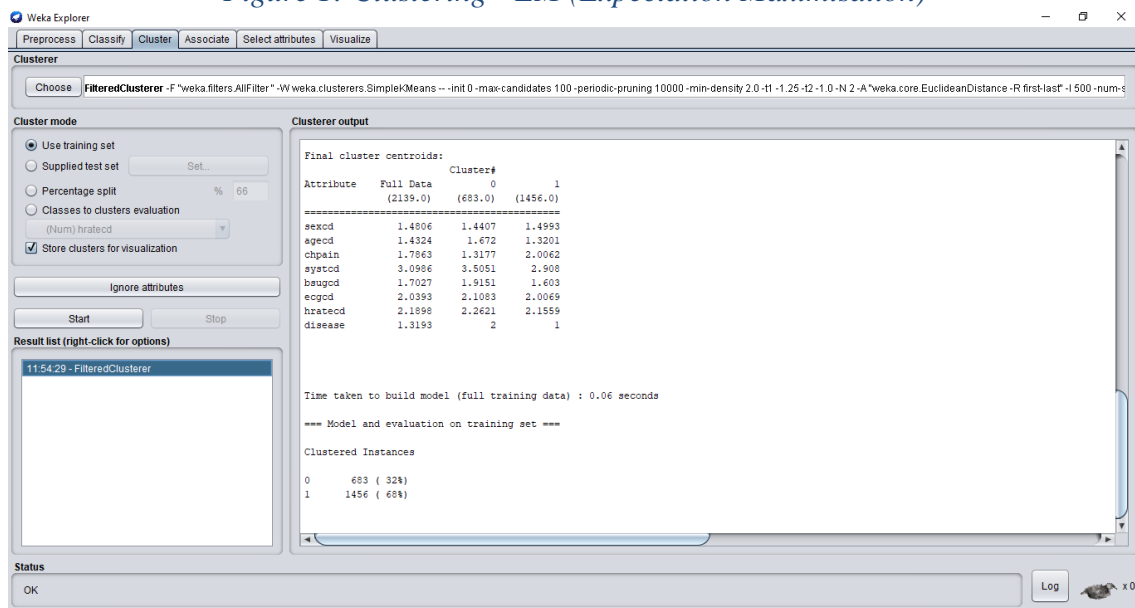


Figure 2: Clustering – Filtered Cluster

For our heart dataset we have used WEKA tool (**Waikato**) and applied EM (Expectation Maximisation), Filtered Cluster, Density Based Cluster and SimpleKMeans. Weka outputs using these methods are shown in Figure1 to Figure 4 respectively.

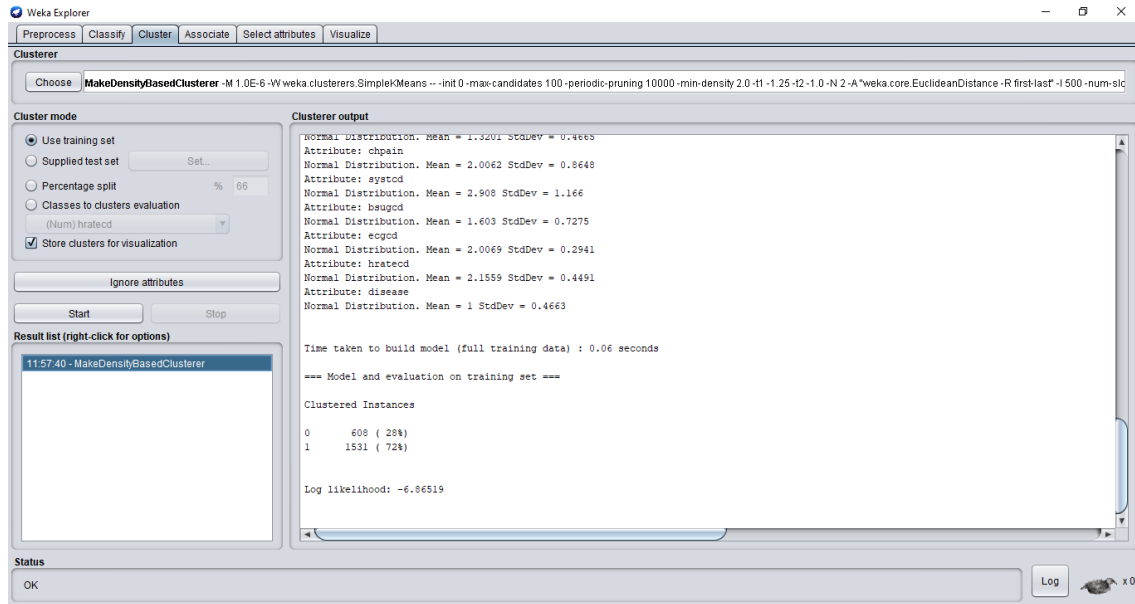


Figure 3: Clustering –Density Based Cluster

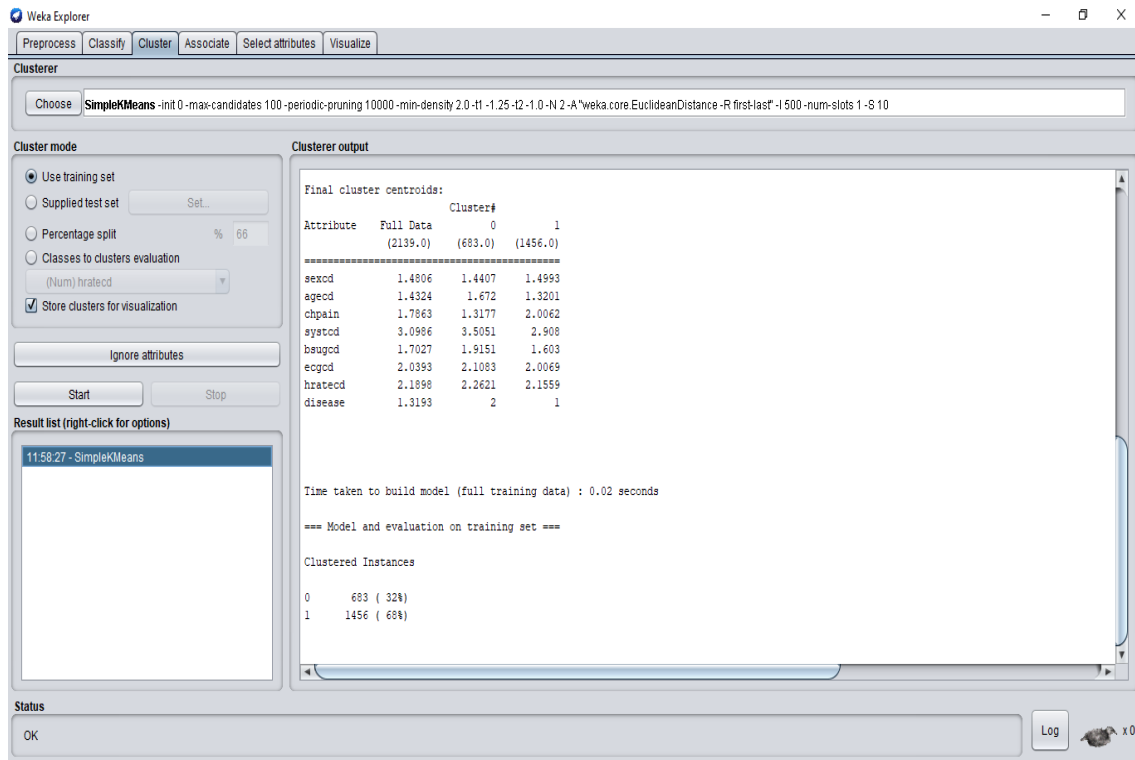


Figure 4: Clustering – SimpleKmeans

Table 1 shows the summary of Clustering Models – EM, filtered, Density Based and simple k-means for our Heart dataset.

Algorithm	No. Of Clusters	Cluster Instance	Time in Seconds
EM (Expectation Maximisation)	04	729(34%) 727(34%) 301(14%) 382(18%)	6.55
Filtered Cluster	2	683(32%) 1456(68%)	0.06
Density Based Cluster	2	608(28%) 1531(72%)	0.06
SimpleKMeans	2	683(32%) 1456(68%)	0.02

Table 1 : Summary of Clustering Models for Heart dataset

Figure 5 to Figure 8 shows cluster visualization with heart disease (blue – negative and orange-positive) for attributes blood pressure, chest pain, blood sugar and blood pressure & blood sugar respectively.

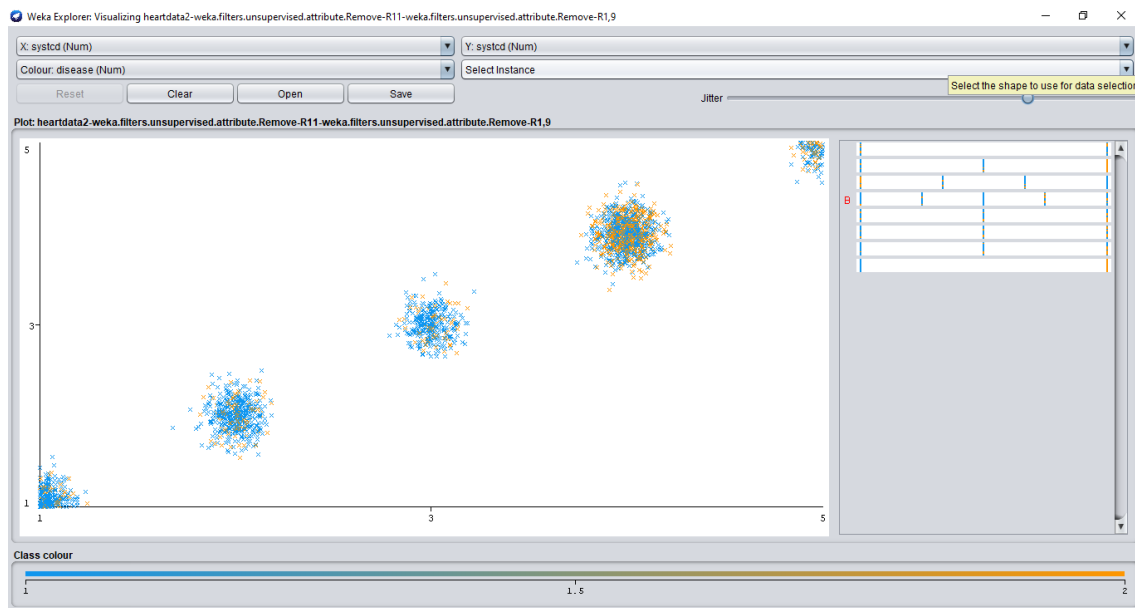


Figure 5: Cluster visualization- Attribute blood pressure

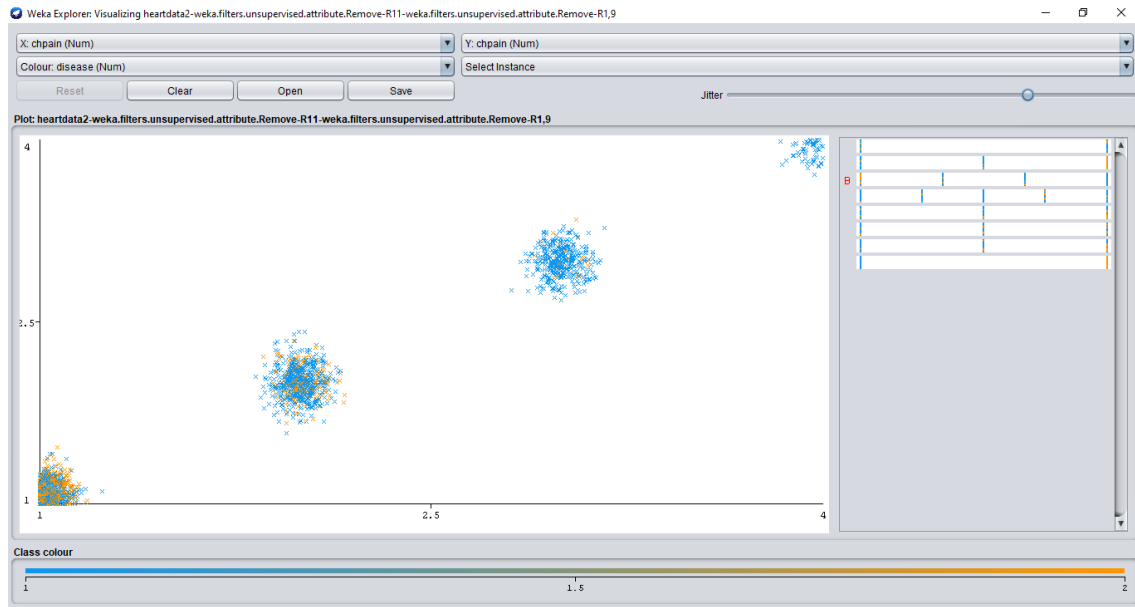


Figure 6: Cluster visualization- Attribute chest pain



Figure 7: Cluster visualization- Attribute blood sugar

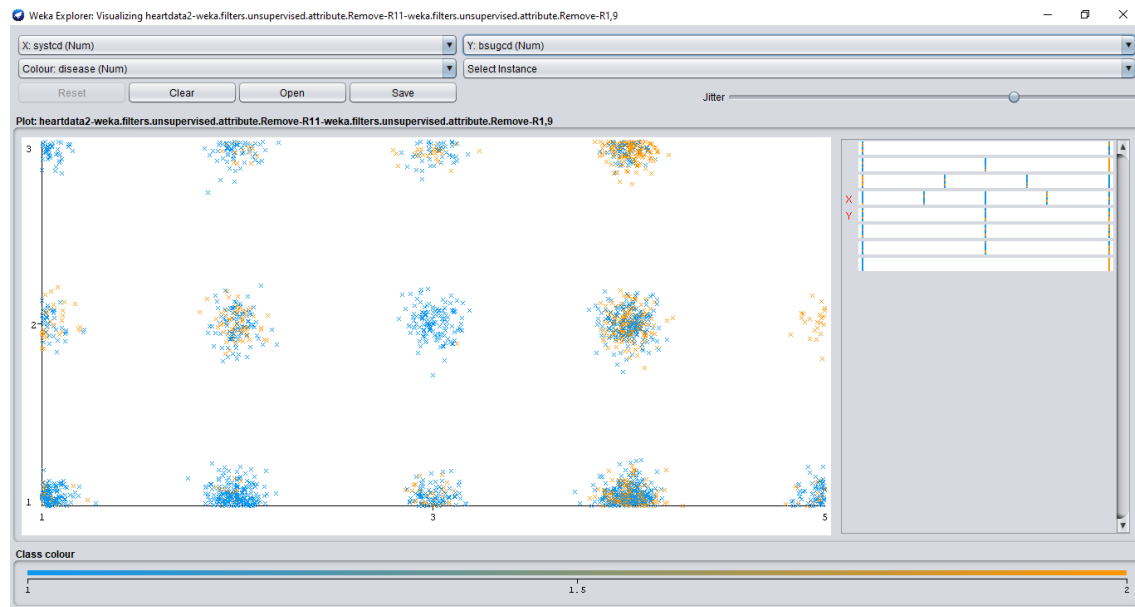


Figure 8: Cluster visualization- Attribute blood pressure & blood sugar

5. Conclusion

In this paper we have used 4 clustering algorithms to mine heart dataset created by us. The Weka output of used clustering algorithms are given. We have presented cluster visualization for few attributes of our dataset. In future we can apply other DM techniques and compare our dataset with other datasets for performance evaluation.

REFERENCES

1. **AH Chen, SY Huang, PS Hong, CH Cheng and EJ Line**, "HDPS : Heart Disease Prediction System", Computing in Cardiology, ISSN 0276-6574, 2011.
2. **Bala Sundar V, T Devi and N Saravanan**, "Development of a Data Clustering Algorithm for Predicting Heart", Volume 48 No.7, 2012.
3. **Elham Naikookar and Ebrahim Naderi**, "Hybrid Ensemble Framework for Heart Disease Detection and Prediction", (IJACSA), Vol. 9, No. 5, pp243-248, 2018.
4. **Lovepreet Kaur**, "Predicting Heart Disease Symptoms using Fuzzy C-means Clustering" IJAR CET, Volume 3 Issue 12, 2014.
5. **MA. Jabbar, Dr. Priti Chandra and B.L. Deekshatulu** "Cluster Based Association Rule Mining For Heart Attack Prediction", Vol. 32 No. 2, E-ISSN: 1817-3195, 2011.
6. **Mamta Sharma, Farheen Khan and Vishnupriya Ravichandran**, "Comparing Data Mining Techniques Used For Heart Disease Prediction", (IRJET), e-ISSN: 2395-0056, p-ISSN : 2395-0072, vol. 04 Issue: 06, 2017.
7. **Mirpouya Mirmozaffari, Alireza Alinezhad and Azadeh Gilanpour**, "Heart Disease Prediction with Data Mining Clustering Algorithms" , IJCCIE, Vol. 4, Issue 1), ISSN 2349-1469 EISSN 2349-1477, 2017.
8. **Nallakaruppan M.K , N. Deepa , Pranava Kumar R , Moin Ahmed , Tanuja**

- and Ajith Kumar**, “Heart Disease Prediction using Data Mining Techniques”, Volume 7 Issue No.12, 2017.
9. **Nelofar Rehman**, “Data Mining Techniques Methods Algorithms and Tools”, IJCSMC, Vol. 6, Issue 7, pp227-231, ISSN 2320-088X, 2017.
 10. **Pang-Ning Tan, Michael Steinbach and Vipin Kumar.**, “Introduction to Data Mining”, Pearson Education, Inc., ISBN 978-81-317-1472-0, 2008.
 11. **Reetu Singh and E. Rajesh**, “Prediction of Heart Disease by Clustering and Classification Techniques”, Vol.-7, Issue-5,E-ISSN: 2347-2693, , 2019.
 12. **Sarangam Kodati and Dr. R. Vivekanandam**, “Analysis of Heart Disease using in Data Mining Tools Orange and Weka”, Volume 18 Issue 1 Version 1.0, ISSN: 0975-4350, 2018.
 13. **Waikato Environment for Knowledge Analysis (WEKA)**, developed at the University of Waikato, New Zealand, www.cs.waikato.ac.nz/~ml/weka/