ISSN 2350-1049

International Journal of Recent Research in Interdisciplinary Sciences (IJRRIS)
Vol. 3, Issue 4, pp: (15-22), Month: October - December 2016, Available at: www.paperpublications.org

Heart Diseases Diagnosis Using Data Mining Techniques

[1]Dr. R. Muralidharan, [2]D. Vinodhini

[1]M.Sc. M.Phil. MCA., Ph.D. Vice Principal & HOD in CS Rathinam College of Arts & Science Coimbatore, India
[2]M.Sc., Department of Computer Science, Rathinam College of Arts & Science, Coimbatore, India

Abstract: **In health care domain, data mining plays a vital role for predicting diseases. For detecting a disease number of tests should be required from the patient but number of test should be reduced while using data mining techniques. The data mining technique analyze the test parameters and it concludes the associative relation between the parameters that reduces the tests and the reduced test plays key role in time and performance. In this study medical terms such as sex, blood pressure, and cholesterol like nineteen input attributes are used. In this paper association among various attributes which are the causative factors of heart diseases are analyzed. The patient's records are observed before prediction and the factors are grouped as per its severity level. In this system the level of causative factors are categorized using K-Means clustering technique and it distinguishes the risky and non-risky factors. Frequent risk factors are mined from the clinical heart database using Apriori algorithm. The risk factors are taken for this study to predict the risk level and find the co-ordination among the factors that helps the medical people to predict the disease with minimum tests and treatments.**

Keywords: **Associative Mining algorithm, Data mining, Heart disease, K-Means Clustering.**

I. INTRODUCTION

Data mining is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data Mining is a very crucial research domain in recent research world. The techniques are useful to elicit significant and utilizable knowledge which can be perceived by many individuals. Data mining programs consists of diverse methodologies which are predominantly produced and used by commercial enterprises and biomedical researchers. These techniques are well disposed towards their respective knowledge domain. The heart is very important part of human body, which pumps blood into the entire body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient working of the heart. The term Heart disease refers to disease of heart & blood vessel system within it. This study is closely related to heart diseases prediction and it provides an awareness to the patient to prevent them from unexpected heart problem. This paper analyzes the heart disease predictions using classification algorithms. Medicinal data mining has high potential for exploring the unknown patterns in the data sets of medical domain. These patterns can be used for medical analysis in raw medical data. Heart disease was the major cause of casualties in the world. Half of the deaths occur in the countries like India, United States are due to cardiovascular diseases. Medical data mining techniques like Association Rule Mining, Clustering, and Classification Algorithms are applied to analyze the different kinds of heart based problems.

Data mining is a process of extracting/equating the meaningful data from the large database by using different tools and techniques. Data mining (sometimes called information or knowledge discovery) is the process of evaluating data from different outlooks and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of diagnostic tools for analyzing data. It allows users to investigate data from many different dimensions or angles, categorize it, and summarize the relationships identified.

ISSN 2350-1049

International Journal of Recent Research in Interdisciplinary Sciences (IJRRIS)
Vol. 3, Issue 4, pp: (15-22), Month: October - December 2016, Available at: www.paperpublications.org

Technically, data mining is the process of finding correlations or forms among dozens of fields in large relational databases. Data mining or knowledge discovery in databases (KDD) is an interdisciplinary field where we integrate techniques from different fields including data base systems, statistics, mathematics, high performance computing, artificial intelligence and machine learning.

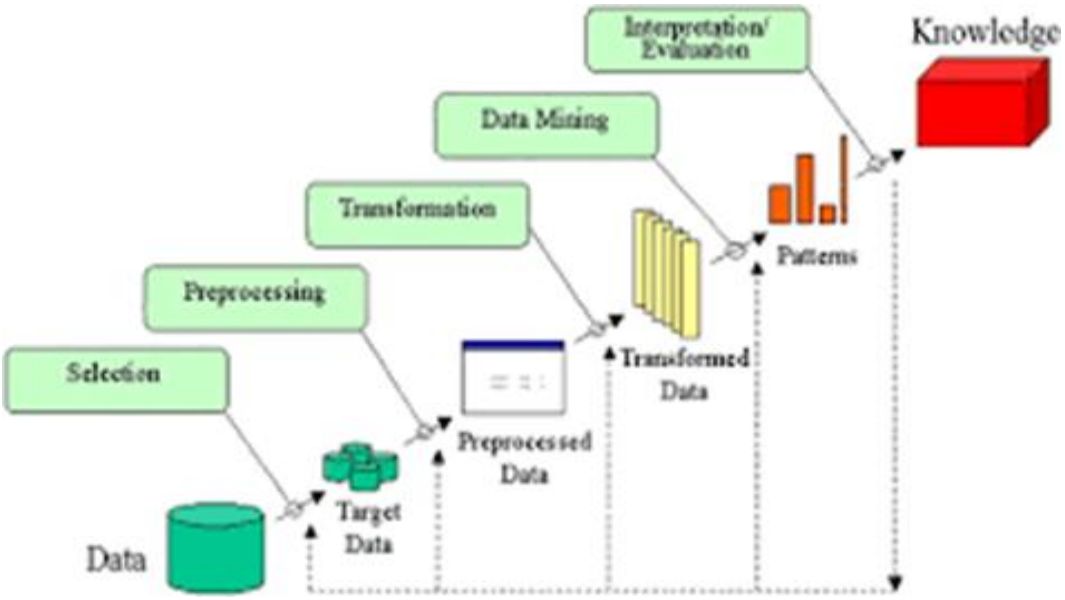

Fig. 1 Basic idea about KDD

This data mining techniques are used in health care domain for prediction of problem, disease detection, optimizing the solution and so on. Recent technologies are nowadays able to provide a lot of information on health-related activities, which can then be analyzed in order to find important information and to collect relevant information. This data mining techniques are used for disease detection, pattern recognition by using multiple application. Data mining is about to identify the similarities between searching the valuable business information from the large database systems such as finding linked products in gigabytes of store scanner data or the mining a mountain for a vein of valuable dataset. Both kind of processes required either shifting through an immense amount of material, or to perform the search intelligently so that exactly match will be performed. Data mining can be done on a database whose size and quality are sufficient. Data mining techniques are often used to study health factors.

The purpose of the study is to analyze the human health using classification techniques and predict the risk level of heart diseases based on the health factors. The study applied data mining methods like k-means and associative mining facilitate an improvement in the interpretation of the clinical data set.

II. REVIEW OF LITERATURE

The paper titled *"The Survey of Data Mining Applications and Feature Scope"* - focused a variety of techniques, approaches and different areas of the research which are helpful and marked as the important field of data mining Technologies. As they are aware that many MNC's and large organizations are operated in different places of the different countries**.** Each place of operation may generate large volumes of data. Corporate decision makers require access from all such sources and take strategic decisions. The data warehouse is used in the significant business value by improving the effectiveness of managerial decision-making. In an uncertain and highly competitive business environment, the value of strategic information systems such as these are easily recognized however in today's business environment, efficiency or speed is not the only key for competitiveness. These types of huge amount of data's are available in the form of tera- to peta-bytes which has drastically changed in the areas of science and engineering. To analyze, manage and make a decision of such type of huge amount of data we need techniques called the data mining which will transforming in many fields. They imparts more number of applications of data mining and also focuses scope of the data mining which will helpful in the further research.

"A Literature Review of Data Mining Techniques Used in Healthcare Databases" - presented an overview of the current research being carried out using the data mining techniques for the diagnosis and prognosis of various diseases. The goal of this study is to identify the well-performing data mining algorithms used on medical databases. The following algorithms have been identified: Decision Trees, Support Vector Machine, Artificial neural networks and their Multilayer Perceptron model, Naive Bayes, Apriori, Fuzzy Rules. Analyses show that it is very difficult to name a single data mining algorithm as the most suitable for the diagnosis and/or prognosis of diseases. At times some algorithms perform better than others, but there are cases when a combination of the best properties of some of the aforementioned algorithms together results more effective.

"An Empirical study on prediction of Heart disease using Classification Data mining techniques" - the use of pattern recognition and data mining techniques into risk prediction models in the clinical domain of cardiovascular medicine is proposed. The data is to be modelled and classified by using classification data mining technique. Some of the limitations of the conventional medical scoring systems are that there is a presence of intrinsic linear combinations of variables in the input set and hence they are not adept at modelling nonlinear complex interactions in medical domains. This limitation is handled in this research by use of classification models which can implicitly detect complex nonlinear relationships between dependent and independent variables as well as the ability to detect all possible interactions between predictor variables.

"Analysis of Attribute Association in Heart Disease Using Data Mining Techniques" - In data mining association rule mining represents a promising technique to find hidden patterns in large data bases. The main issue about mining association rules in a medical data is the large number of rules that are discovered, most of which are irrelevant. A rule-based decision support system (DSS) is presented for the diagnosis of coronary vascular disease (CVD). The dataset used for the DSS generation and evaluation consists of 1897 subjects, each one characterized by 21 features, including demographic and history data, as well as laboratory examinations. Such number of rules makes the search slow. However, not all of the generated rules are interesting, and some rules may be ignored. In medical terms, association rules relate disease data measures the patient risk factors and occurrence of the disease. Association rule medical significance is evaluated with the usual support and confidence metrics. Association rules are compared to predictive rules mined with decision trees, a well-known machine learning technique. In this paper [12] they proposed a new system to find the strength of association among the attributes of a given data set.

"A Fuzzification Approach for Prediction of Heart Disease" - Data Mining operations and approaches are the improvement over the statistical methods that enables a user to perform the future analysis based on current dataset. One of such analysis provided by data mining approaches is the predication based analysis. In their work [13], the heart disease prediction system is designed. The heart disease prediction is actually an expert system application which requires the authenticated dataset to process. A Fuzzy based soft computing approach is been implemented on multiple parameters to predict the heart disease. In this paper [13], the earlier work done in the area of medical disease prediction is studied as well a new fuzzy rule based approach is suggested to perform the heart disease prediction.

"Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" - The healthcare environment is still "information rich but knowledge poor". There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. This research paper [14] intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction. Number of experiment has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and some time Bayesian classification is having similar accuracy as of decision tree but other predictive methods like Neural Networks, Classification based on clustering are not performing well. The second conclusion is that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.

A Heart Disease Prediction Model using Decision Tree - In this paper [15], they developed a heart disease prediction model that can assist medical professionals in predicting heart disease status based on the clinical data of patients. Firstly, they select 14 important clinical features, i.e., age, sex, chest pain type, Blood Pressure, cholesterol, fasting blood sugar, resting ECG, max heart rate, exercise induced angina, old peak, slope, number of vessels colored and diagnosis of heart

disease. Secondly, they developed a prediction model using J48 decision tree for classifying heart disease based on these clinical features against un-pruned, pruned and pruned with reduced error pruning approach. Finally, the accuracy of Pruned J48 Decision Tree with Reduced Error Pruning Approach is more better then the simple Pruned and Un-pruned approach. Their result obtained that which shows that fasting blood sugar is the most important attribute which gives better classification against the other attributes but its gives not better accuracy.

III. RESEARCH METHODOLOGIES

In this work, an architecture data mining technique based heart disease prediction system combining the prediction system with mining technology was used. In this model we have used one of the classification algorithms called association mining.

Simulated heart disease dataset containing 1000 patient records are used for this research work. This dataset contains 19 symptoms as shown in Table 1. They are the symptoms of various heart diseases.

TABLE I. Heart Disease Attributes for the Study

S. No	Attributes	Values
1	Male and Female	Age < 30 , Age >30 to <50, Age>50 and Age <70, Age>70.
2	Smoking	Never , Past, Current
3	Overweight	Yes, No
4	Alcohol Intake	Never , Past, Current
5	High salt diet	Yes, No
6	High saturated fat diet	Yes, No
7	Exercise	Never, Regular
8	Sedentary Lifestyle/inactivity	Yes, No
9	Hereditary	Yes, No
10	Bad cholesterol	Very High >200, High 160 to 200, Normal <160
11	Blood Pressure	Normal (130/89), Low (< 119/79) High (>200/160)
12	Blood sugar	High(>120&<400), Normal(>90&<120), Low (<90)
13	Heart Rate	Low (< 60bpm), Normal (60 to 100) High (>100bpm)
14	Defect type	Normal, Fixed, Reversible defect
15	Chest pain type	typical type 1, typical type angina non-angina pain, asymptomatic
16	Resting electrographic results	Normal having ST_T wave abnormal left ventricular hypertrophy
17	number of major vessels colored by fluoroscopy	0-3 values
18	Dry or persistent Cough	Yes, No
19	Skin rashes or Unusual Spots	Yes, No

The data and item sets that occur frequently in the data base are known as frequent patterns. The frequent patterns that is most significantly related to specific heart disease types and are helpful in predicting the disease and its type is known as Significant frequent pattern. It has four levels of risk like low level, intermediate level, high level and very high level. Based on the predicted risk values the range of risk will be assigned. In this study every individual health factors which are taken for the study is analyzed and the causative factors are classified using data mining technique.

ISSN 2350-1049

International Journal of Recent Research in Interdisciplinary Sciences (IJRRIS)
Vol. 3, Issue 4, pp: (15-22), Month: October - December 2016, Available at: www.paperpublications.org

Association Rules:

Association rules are if / then statement that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An association rule has two parts an antecedent (if) and a consequent (then). An antecedent is an item found the data and consequence t is an item that is found in combination with the antecedents. Association rules are created y analyzing data for frequent if / then patterns and using the criteria support and confidence to identify the most important relationships. They are divided into separate categories in the data mining and used in the Weka to perform the operations.

Apriori Algorithm:

The Apriori algorithm is a popular and foundational member for correlation based 'Data Mining Kernels' used today. It is used to process the data into more useful forms, in particular connection between set of items.

Dataset:

The clinical data sets are gathered from various health care centres. From the collected records data preprocessing is applied to carry out the factor which are relevant for this study. Nearly 1000 health records are gathered and the data are pre-processed. The following table shows the sample pre-processed dataset.

TABLE II. Sample Data Table

Attributes	**Values**	**Values**	**Values**
Gender	Male	Male	Female
Age	63	68	45
Smoking	Yes	No	No
Weight (kgs.)	70	56	65
Alcohol Intake	Yes	No	No
High salt diet	Yes	No	Yes
High saturated fat diet	No	Yes	No
Exercise	No	Yes	No
Sedentary Lifestyle/inactivity	Yes	No	Yes
Hereditary	No	Yes	No
Bad cholesterol	170	160	180
Blood Pressure	130/89	160/100	140/90
Blood sugar	100	120	110
Heart Rate (bpm)	70	80	85
Defect type	Normal	Fixed	Normal
Chest pain type	typical type 1	No	typical type 1
Resting electrographic results	Normal	Normal	Abnormal
Number of major vessels colored by fluoroscopy	0	0	0
Dry or persistent Cough	Yes	No	No
Skin rashes or Unusual Spots	No	No	No

The above table shows sample dataset. First the data set is converted as binary data set and it is analyzed using Weka tool. This binary data set illustrates patients' health information in binary mode. Collections of Patient medical details used for transaction databases and set of associations can be represented as binary incidence matrices with columns corresponding to the factors and rows corresponding to the Patients. The matrix entries represent presence (1) or absence (0) of a risk factor in a particular patient.

ISSN 2350-1049

International Journal of Recent Research in Interdisciplinary Sciences (IJRRIS)
Vol. 3, Issue 4, pp: (15-22), Month: October - December 2016, Available at: www.paperpublications.org

TABLE III. Binary Data Set

Attributes	Values	Values	Values
Age	1	1	0
Smoking	1	0	0
Weight (kgs.)	1	0	1
Alcohol Intake	1	0	0
High salt diet	1	0	1
High saturated fat diet	0	1	0
Exercise	0	1	0
Sedentary Lifestyle/inactivity	1	0	1
Hereditary	0	1	0
Bad cholesterol	1	0	1
Blood Pressure	0	1	0
Blood sugar	0	0	0
Heart Rate (bpm)	0	0	0
Defect type	0	1	0
Chest pain type	1	0	1
Resting electrographic results	0	0	1
number of major vessels colored by fluoroscopy	0	0	0
Dry or persistent Cough	1	0	0
Skin rashes or Unusual Spots	0	0	0

Data Preprocess:

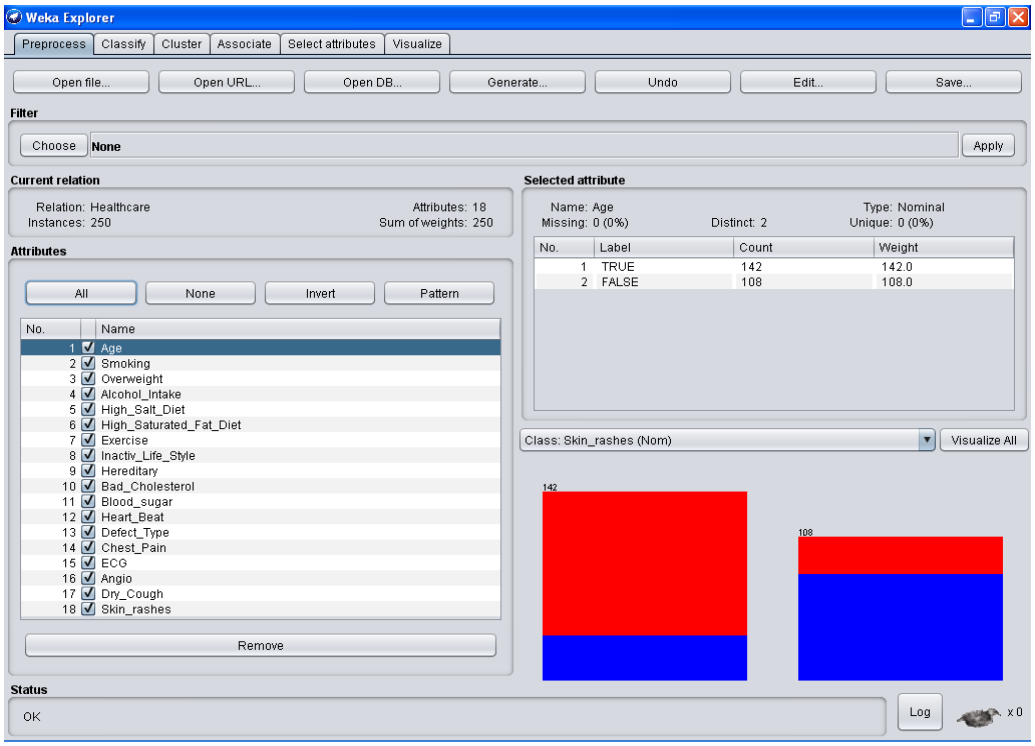

Fig.3.1 Data Preprocess

The Fig. 3.1 shows the attributes which are involving in the study. There are nineteen factors are taken for this analysis. The attributes are associated with one and another. This analysis find the support and confidence level of heart diseases according various factors.

ISSN 2350-1049

International Journal of Recent Research in Interdisciplinary Sciences (IJRRIS)
Vol. 3, Issue 4, pp: (15-22), Month: October - December 2016, Available at: www.paperpublications.org

IV. OBSERVATIONS AND RESULTS

When the apriori algorithm was implemented on the healthcare.arff file, it produced best association rules for that healthcare.arff or dataset. Now the observation took place, to observe the result of the same operation on the same dataset by the use of weka tool.

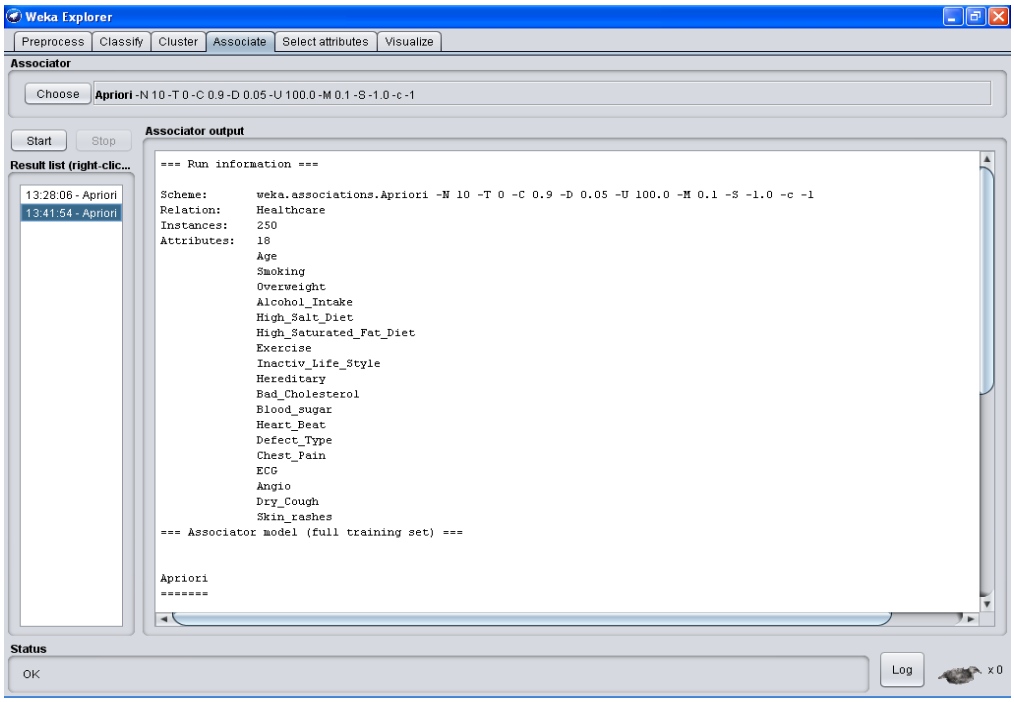

Fig.3.2 (a) Apriori Result

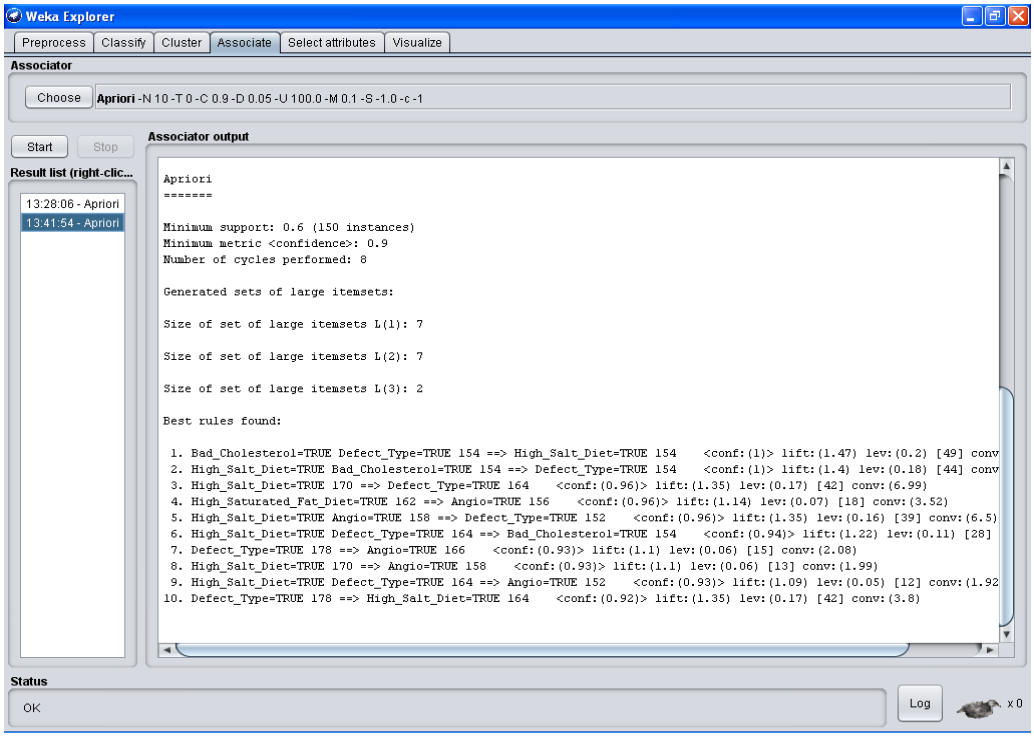

Fig.3.2 (b) Apriori Result

For a given data set, figure 3.2 (a & b) shows Apriori algorithm mined risk factors that have minimum support of all transactions and it provides 10 best association rules.

V. CONCLUSION AND FUTURE ENHANCEMENT

Healthcare data set includes the patients' health information. The information is classified as hereditary, habitual and medical test details. In which the seventeen causative factors which are gathered from domain experts are obtained for the study. Factors are analyzed and the co-occurrence of the factors is found by generating candidate item set and pruning. Finally the supportive measure of the combined factors is obtained through Apriori. For a given patient records risk and non-risk factors are grouped using K-means clustering. The ranges of the causative factors are analyzed to predict the risk level of the heart disease. It is no doubt that Apriori algorithm successfully finds the frequent elements from the database. In future the work can be expanded and enhanced for the automation of various types of disease prediction. It also extended to find various types of diseases with the use of these factors.

REFERENCES

[1] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi, "The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science, Engineering and Information Technology.

[2] Elma Kolçe, Neki Frasheri, "A Literature Review of Data Mining Techniques Used in Healthcare Databases", ICT Innovations 2012 Web Proceedings - Poster Session ISSN.

[3] T.John Peter., "An Empirical study on prediction of Heart disease using Classification Data mining techniques" IEEE-International Conference On Advances In Engineering, Science And Management.

[4] K.Srinivas., G.Raghavendra Rao and A.Govardhan., "Analysis of Attribute Association in Heart Disease Using Data Mining Techniques", International Journal of Engineering Research and Applications.

[5] Nitika, Madan Lal Yadav, "A Fuzzification Approach for Prediction of Heart Disease", International Journal of Engineering Trends and Technology (IJETT).

[6] Jyoti Soni Ujma Ansari Dipesh Sharma., "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (IJCA).

[7] Atul Kumar Pandey, Prabhat Pandey, K.L., Jaiswal, Ashish Kumar Sen, "A Heart Disease Prediction Model using Decision Tree", IOSR Journal of Computer Engineering.