Biomedical (Cardiac) Data Mining: Extraction of SignificantPatterns for Predicting Heart Condition

By
Mamuna Fatima

2011-NUST-MSPHD- CSE (E)-016
MS-11 (CSE)

Submitted to the Department of Computer Engineering in fulfillment of the requirements for the degree of

MASTER OF SCIENCE
In
SOFTWARE ENGINEERING

Thesis Supervisor
Dr. Shoab Ahmed Khan

Thesis Co-Supervisor
Dr. UsmanQamar

College of Electrical & Mechanical Engineering
National University of Sciences & Technology
2013

DECLEARATION

I hereby declare that this thesis work has been built on my personal efforts under the supervision of Dr. Shoab Ahmed Khan. All the data sources have been referenced and there is no plagiarized data contained in this research. No data of this thesis has been shared as a part of any other research work to be presented in any other institute or university for the fulfillment of degree requirement.

Student Signature

ACKNOWLEDGEMENT

I am grateful to Almighty Allah whogranted me the strength and courage to accomplish this task in the best possible way, and also sincerelythankful to my loving parents for their ‘dua’ and look after.

I am thankfulto many learned people, for their help, guidance and sincere cooperation without which this achievement was not possible. At the most I am absolutely indebted and highly grateful to my supervisors Dr. Shoab Ahmed Khan and Dr. UsmanQamarfor their regular guidance and assistance which made it possible for me to accomplish this task. I also acknowledged the guidance of my Committee members. Their suggestions were very viable in completion of this research.

Besides, I would like to thank Mr. Ali RazaAnjum and IqraBasharatfor their constant help and support.

ABSTRACT

There is a huge amount of ‘knowledge-enriched data’ in hospitals, which needs to be processed in order to extract useful information from it. The knowledge-enriched data is very useful in making valuable medical decisions. However, there is a lack of effective analysis framework to discover hidden relationships in data. The objective of this research is to propose a framework for cardiac data mining that mines the historical unstructured data of heart patients and extract significant features and patterns which will not only enable doctors to predict heart attack but also provide in-depth insight to write better prescription in future. This work is based on a large amount of unstructured data in the form of patients medical reports collected from a renowned cardiac hospital in Pakistan. Firstly data preparation is done in which the unstructured (textual) data of heart patients is converted to structured (tabular) form and then pre-processed to make it suitable to apply different data mining techniques. After data preprocessing, data mining techniques are used in which clustering, correlationand association rule mining techniquesare applied onthe dataset. The output from this exercise takes the form of trends, patterns and rules which can then be used for heart condition prediction besides helping medical practitioners in making intelligent verdicts. Finally, performance evaluation of the selected k-Means algorithm is performed with other clustering algorithms on the basis ofsome internal evaluation indexes.Further the generated rules are evaluated using statistical measures such as support, confidence, lift, completeness, interestingness and comprehensibility to extract significant rules only.

Keywords:*Data Mining, Unsupervised learning,Clustering, Heart Disease,MiningUnstructured Data,K-Means*

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Illustration
CHD	Coronary heart disease
CAD	Coronary artery disease
ECG	Electrocardiogram
ETT	Exercise Tolerance Test
CT	Computerized tomography
2D Echo	Two-dimensional echocardiogram
DBSCAN	Density-based spatial clustering of applications with Noise
RF	Random Forest
DWH	Data Warehouse
IHDPS	Intelligent Heart Disease Prediction System
CANFIS	Coactive neuro-fuzzy inference system
GUI	Graphical user Interface
CRISP-DM	Cross Industry Standard Process for Data Mining
ANOVA	Analysis-of-variance
SVCAD	Single Vessel Coronary Artery disease
DVCAD	Double Vessel Coronary Artery disease
TVCAD	Triple Vessel Coronary Artery disease
BMI	Body Mass Index
LVEF	Left Ventricular Ejection Fraction
NIDDM	Non-Insulin-Dependent Diabetes Mellitus
MI	Myocardial Infarction
ETL	Extraction, Transformation and Loading
T	T-Wave

RBBB	Right bundle branch block
LBBB	Left bundle branch block
BP	Blood Pressure
HR	Heart rate

Chapter 1

INTRODUCTION

This chapter presents the detailed introduction of the research. The basic purpose is to introduce the main concepts, research motivation, problem definition and contribution of this research.

The diagnosis of heart disease is intricate and tedious task which depends upon various symptoms and factors. The health care industry contains huge amount of unstructured data that unfortunately is not mined properly for effective decision making. The preeminent option for this is to makeuse of this data by applying some data mining techniques along with expert knowledge of medical specialistsfor valuable knowledge discovery and diagnosis process which enhances the quality of clinical decisions[10].

1.1 MAIN CONCEPTS

1.1.1 CORONARY HEARTDISEASESIGNIFICANCE

Coronary heart disease (CHD) is the mainkiller in developed countries and the risk of CHD is increasing in developing countries like Pakistan[2].However, there remain distinct differences in how severe the burden is affecting the various populations [3].

CHD claims 17.3 million lives a year throughout the world while in Pakistan this disease claims about 200,000 lives per year that is 410/100,000 of the population and the condition is frightening as the number is constantlygrowing. To reduce burden of heart diseases in Pakistan, there is a need of creating awareness among public about this disease, its symptoms, risk factors and preventive measures to be taken to avoid heart-related problems [6].Pakistani population has one of the highest risks of CHD in the world. “In Pakistan, 30 to 40 percent of all deaths are due to cardiovascular diseases. The CHD is now becoming the leading cause of death in Pakistan.

Some complications ofCoronary Heart Disease include:

- ***Heart Attack (Myocardial Infarction)***

“Myo” means muscle and cardial refers to the heart. Infarction means death of tissue due to lack of blood supply. Heart attack is myocardial infarction, and it causes permanent damage to the heart muscle.

- ***Heart Failure***

Heart failure is a condition in which your heart can't pump enough blood to meet your body's needs. Heart failure doesn't mean that your heart has stopped or is about to stop working. It means that your heart can't cope with the demands of everyday activities.

- ***Arrhythmia***

An arrhythmia is a problem with the rate or rhythm of the heartbeat. During an arrhythmia, the heart can beat too fast, too slow, or with an irregular rhythm.
Some other heart disease types are abnormal heart rhythms (arrhythmias), heart failure, heart valve disease, heart muscle disease, and congenital heart disease.

1.1.2 CORONARY ARTERY DISEASE (CAD)

CAD is the most common heart disease type. In this disease, the arteries supplying blood to the heart muscle (the coronary arteries) become lined with plaque, which contains materials such as cholesteroland fat. This plaque buildup (called atherosclerosis) causes the arteries to narrow, allowing less oxygen than required to reach to the heart muscle. When the heart muscle does not receive enough oxygen, chest pain (angina) or heart attack can occur [33].

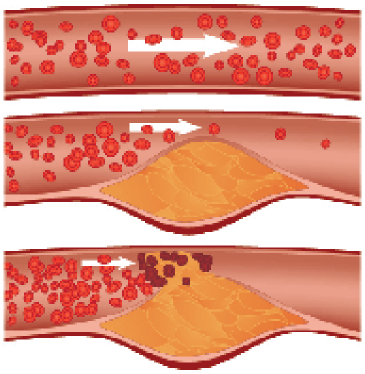


Figure1.1.Arteries supplying blood to heart muscle

Three types of CAD are:

- SVCAD
- DVCAD
- TVCAD

1.1.3MAJOR SYMPTOMS OF HEART DISEASE& ATTACK

Some of the major symptoms of this threat revealed by different studies are [2] [5]:

- Angina (chest pain)
- Dyspnea (shortness of breath)
- Fatigue
- Anxiety/nervousness
- Increased irregular heart rate

1.1.4RISK FACTORS OF HEARTATTACK &DISEASE

There are multiple risk factors that enhance the chances of getting a heart attack or heart disease. Some of the major revealed by different studies are [2] [5]:

- **Age, Male gender and Family history**

  These are the factors that can’t be controlled or prevented. This disease attacks male gender more than female. Besides family history of this disease increase the hazard of premature death.

- **Diabetes mellitus, Hypertension, Smoking, High cholesterol level, Physical inactivity, Stress/Depression and Unhealthy diets [6]**

  These all are the important factorsthatmay seriously be unsafe for your heartbut can be controlled by taking precautionary measures. Diabetes and Hypertension are revealed responsible for this risk in many studies. Besides one should avoid smoking, high cholesterol diets and other unhealthy foods to prevent himself from this disease. Doctors suggest that even 30 minute exercise/physical activity per day reduces the chance of getting heart disease many times.

1.1.5DATA MINING APPROACH FOR HEALTH CARE DATA

The effective selection and extraction of information enriched data from a large collection of unstructured data is very significant for valuable decision making. This process known as ‘Knowledge Discovery from Data’ is iterative and interactive and contains many subtasks and decisions. The core procedure of Knowledge Discovery is to transform data into knowledge for the purpose of decision making which is called Data Mining[1]. By applying data mining techniques valuable knowledge can be extracted from health care data [9].

Data mining syndicates database technology, machine learning and statistical analysis to mine unseen patterns and associations from huge databases [2]. Data Mining has its utilization in various fields i.e. marketing, customer relationship, management, engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing. In medicine the researchers can diagnose and predict various diseases of patient with the help of data mining techniques. In this research [35] various diseases like Diabetes, Hepatitis, Cancer and are diagnosed using data mining techniques.

Data mining uses two strategies: supervised and unsupervised learning. In supervised learning (like in regression model), a training set is used to learn model parameters whereas in unsupervised learning no training set is used (e.g., k-means clustering is unsupervised) [11]. Medical treatment can be made more effective and cheap with the help of automated medical diagnosis system. The significance of data mining techniques role in efficient pattern and knowledge extraction increases with the size of dataset [5].

1.2 NUCLEAR MEDICINE DEPARTMENT, ARMED FORCES INSTITUTE OF CARDIOLOGY (AFIC)

The Nuclear Cardiology Department, AFIC operates different patients of cardiology every day. This 250-bedded hospital is a major tertiary cardiac care center in Rawalpindi, Pakistan, dedicated to the cardiac patients from all over the country including Armed Forces, Federal Government employees and civil population.Patients from all over Pakistan come to AFIC to seek diagnosis and treatment related to different cardiac problems. Reports of all these patients are stored in an unstructured database. A record of all these reports is also kept in hard copies within the department.

1.2.1 MEDICAL PROCEDURES IN AFIC

The different medical procedures and diagnosis performed in AFIC include different essential tests which lead to the diagnosis of a patient’s problem. Once results of these tests are obtained, the decision about further procedure/treatment to be undertaken can be made. These testsinclude:

1. 2D Echo Test
2. Blood Tests
3. Angiography
4. CT Angiography
5. ECG
6. Thallium-201

7. ETT Test
8. Radiology
9. Cardiac CT Scan

1.2.2 MEDICAL TREATMENTS IN AFIC

Once the diagnosis is made, necessary and complete treatment is given immediately for that. There are several treatments carried out in hospital depending upon the nature of the patient problem.Some are:

1. Percutaneous Transluminal Coronary Angioplasty(PTCA)
2. Coronary Artery Bypass Grafting (CABG)
3. Percutaneous Coronary Intervention (PCI)

1.3 RESEARCH MOTIVATION

According to [8] the database at health care organizations is gigantic database that keeps patients records and their medical history. This data is considered by its intricacy and heterogeneity with respect to the data type. It is highly dimensional, indefinite, dispersed and raucous with improper, unacceptable or omitted values [4]. It is believed that there is a lot of hidden information in those records. This raises a significant problem: "How can we extract and transform the data into expedient information that can allow healthcare experts to make intellectual clinical judgments?" This is the main motivation for this research. A well-known solution for this problem is data mining.Data mining is extraction of implicit formerly unknown and possibly valuable information about data, so by applying data mining techniques, trends and patterns identification process can be made simple which can positively impact the quality of medical decision making.[27].

1.4PROBLEM DEFINITION

The basic step in the data mining is to get a deep understanding of the problem to be solved. The existing information system in AFIC works under the supervision of the IT Department. Data regarding the clinical history, procedural statistics, scan findings and impressions of a patient is stored in the existing system but in an unstructured format which cannot be used for data mining and analysis purpose. It can neither be integrated with any other system for further use. It merely stores reports in an unstructured textual format which can't be used for further processing. Clinical decisions thus are made depending only on doctor's perception and experience which leads toundesirable biases, mistakes and unnecessary medical costs which affects thequality of medical service.This huge ‘knowledge-enriched data’ data needs to be processed in order to extract useful information from it as this is very useful in making valuable medical decisions.However, there is a lack of effective analysis framework to discover hidden patterns/relationships in data. So by applying Biostatistics on these cardiac procedures, better visualization can be provided to a cardiologist about what goes on about when a patient is undergone with different tests or treatments or when a new patient comes with specific complains or diseases which increases the quality of clinical decisions.

1.5RESEARCH CONTRIBUTION

The contributionof this research is to proposea framework for cardiac data mining that mines the historical unstructured data of heart patients and extract significant features and patterns which will not only enable doctors to predict heart condition but also provide in-depth insight for better diagnosis and prescription for patients in future. This work is based on a large amount of unstructured data in the form of patients medical reports collected from AFIC. Firstly data preparation is done in which the unstructured (textual) data of heart patients is converted to structured (tabular) form and then pre-processed to make it suitable to apply different data mining techniques. After data preprocessing, unsupervised learning techniqueis used in which clustering and correlation techniques are applied on dataset to discover knowledge/ hidden patterns related to heart patients. These patterns can then be used for heart condition prediction besides helping medical practitioners in making intelligent verdicts or at least a second opinion. Finally, performance evaluation of the selected k-Means algorithm is performed with other clustering algorithms on the basis ofsome internal evaluation indexes.The system is a theoretical study which will propose implementation of an expert system.

1.6 THESIS OUTLINE

Section 2 presents the literature review explaining the works of different researchers in biomedical data mining. Section 3 describes the methodology followed in this research and section 4 contains the details of data preparation and pre-processing. Section 5 presents correlation and clustering application on the data set, detailed analyses of results and significant extracted patterns. Comparison and performance evaluation of k-means with other algorithms is done in section 6. Conclusion and future work are explained in the last section 7.

Chapter 2

LITERATURE REVIEW

Data mining in bio-medical is finding useful information from a large collection of bio medical data.Data mining uses its strong predictive models and algorithms in exploring, selecting and discovering the unknown/hidden information from a set of large data [7]. According to [8] the literature reports that to predict heart diseases and to make heart disease decision support systems, developer/researches use predictive models of data mining.There is a lot of work in literature that is related with pattern extraction and heart disease prediction using data mining techniques.

2. RELATED WORK

2.1 FEATURE SELECTION AND PATTERN EXTRACTION

To extract the significant patterns from Coronary heart disease database forefficient prediction of heart attack, authors of [9], have presented a dexterous approach.A rough set technique associated to dynamic programming is suggested to abridge high interest features. In this research study authors have used Random Forest(RF) decision tree to classify the perilous heart disease cases. Three different approaches of random forests are used in validation of results accuracy. It has been noted that Forest-RI is the best among the different techniques. The classification results using RFs are obtained from ten-fold cross-validation. However, the conclusion says that, by using RFs multi-classifier technique, the initial number of attributes is reduced from 19 to 9. The relevant features are only taken into consideration which leads to reduce the complexity of the proposed model by focusing the study based on reduced features. The experimental results reveal that the cardiac level of risk is predicted efficiently by the proposed system. Some important steps include:

- Provides level information about Coronary heart disease risk.
- Identify patient’s early, asses risk accurately, develops perception of patient about risk.
- Evaluation of the proposed model

The study is based a huge amount of data gathered from various clinical institutions. Besides, some expert knowledge is taken into account for description of disease, risk factors and relations between medical factors. For this a computer-aided system based on population of 525 adults is developed. The analysis and evaluation of the proposed model is done on the basis of set ofbenchmark procedures.Another efficient approach was proposed by [10] for heart attack prediction by extraction of significant patterns from the heart disease warehouse. The approach has utilized ‘Clustering’ and ‘Frequent pattern mining’ techniques. The heart disease DWH consists of both numerical and categorical attributes. Initially, data was preprocessed by removing duplicate records and supplying missing values to make it compatible for mining process. After, K-Meansclustering algorithm was applied on the preprocessed data with K=2 to get useful clusters. In one cluster there is data that is most pertinent to heart attack which is then input to the MAFIA algorithm and other consists of all remaining data.

The algorithm named MAFIA was applied to get common patterns from the extractedclustered dataset that are related to heart attack. Then the noteworthy weightagewas calculatedfor the frequent patterns based on each attribute weightage that exist in pattern and frequency of every pattern. The formula used for significantweightage (S_{wi}) calculation is as under:

$$S_{wi} = \sum_{i=1}^{n} (W_i f_i) \qquad (1)$$

$W_i$denotes the weightage of each attribute and $f_i$represents the frequency of each rule.

Supplementary, on the basis of calculated weightage some patterns important to heart attack prediction were chosen. The system was implemented in Java. The dataset contains attributes like BP, chest pain, cholesterol, maximum heart rate etc.In the future they will developHeart attack prediction system on the basis of these significant patternsusing artificial intelligence techniques.

2.2 HEART DISEASE PREDICTION SYSTEMS

Author in [11] used data mining techniques e.g. Naïve Bayes, Decision Trees and Neural Network to develop a prototype of IHDPS. IHDPS has the capability to extract hidden patterns and relationships related to heart disease. IHDPS can respond to simple as well as complex’ what if’ queries thus allow medical persons in making wise clinical decisions. Besides it provides effective and cheap treatment and improves visualization and understanding.Among the three models the most efficient in predicting heart disease patients comes to be Naïve Bayes followed by Neural Network and Decision Trees.

IHDPS has used the CRISP-DM [17] methodology to make three models and Data Mining Extension language is used to create, train, predict and access model content.Classification Matrix and Lift Chart methods are used to check which model gave maximum percentage of right predictions. Five mining goals were set and assessed with respect to three trained models. These are:

- Predict those patients who have chances to get heart disease based on their medical profile.
- Findout the important influences and relationships between medical inputs and medical attributes related to the predictable state.
- Find out heart patient characteristics.
- Define attribute values that discriminate nodes favoring and disfavoring the predictable conditions.

Naïve Bayes appears most efficient by answering four out of the five goals and by identifying all important medical predictors; Decision Trees answered three and its results are easier to interpret; Neural Network two and in this the correlation between attributes is hard to interpret.IHDPS is based on 15 medical attributes, 909 records and only categorical data but it can be expanded to include more medical attributes, more techniques like Clustering, Time Series and Association Rules and continuous data as well. Another approach for Heart Disease Prediction system (HDPS) was proposed by [12] using CANFIS. In proposed model Neural network adaptive capabilities are combined with fuzzy logic qualitative approach which is further integrated with genetic algorithm for heart disease diagnosis. The results of the study show that the proposed model has great capability to predict the heart disease.

Jyoti in [13] has proposed an Intelligent and Effective Heart Disease Prediction System using CAR rules generated by Weighted Association rule based Classifier(WAC) for prediction of heart disease. The system is Web-based, user-friendly, scalable and reliable. A GUI based Interface is designed where patient record is entered and prediction is done by mininghistorical data of patient. In WAC different attributes are given different weights based on their predicting capability. The results show that WAC is very efficient in extraction of important patterns from data which are further stored in rule base as Prediction rules. Associative Classifiers have better performance than decision tree and rule Induction besides providing improved accuracy. Currently it is using Cleveland Heart Disease database which consists of 303 records and 14 attributes but it can be easily expanded for new training dataset.

A Decision Support System is proposed by Rajeswari in [14] to predict the patient heart risk in following Years by determining the risk score. This will assist patients to take preventive measures by taking suited diet and medicine that may enhance the patient's life.The proposed system is based on nineteen features and 125 records of patient.The selected features for prediction are based on extensive literature review and expert knowledge of medical practitioners. The features include Age, Sex, Height, Weight, BMI, Menopause details, Waist Circumference, Systolic BP, Diastolic BP, Sugar , Thyroid, Total Cholesterol, LDL Cholesterol, HDL Cholesterol, Smoking, Genetic aspects, Family History, Type A personality, Sleep Disruption. Feature reduction is done among the 19 features on the basis of Genetic algorithm.The Risk factor can be determined by doing sum of several features risk score. With the help of this algorithm, the association patterns having high influence on heart disease identification along with optimal values of considered attributes is determined. The fitness function given below is expected to determine the optimal values related to each attribute.

$$f\text{itness} = \sum_{i=1}^{m} \frac{1}{m}\left(freq(A_{v(i)} * (A_{v(i)} - N_{v(i)}) \right) \qquad (2)$$

However, the system is a theoretical study proposing implementation of algorithms for Machine intelligence.

2.3 NEURAL NETWORKS FOR HEART DISEASE DIAGNOSIS

Neural networks have the capability to solve too complex problems that can't be solved by the conventional technologies. Neural networks proved very good in pattern recognition and forecasting, the problems that are well solved by humans rather than computers aid. It has the capability to generate the desired output even with the inaccurate input. Neural networks are used for heart disease diagnosis in [11][12].

In[36] author has predicted heart disease, BP and Diabetics with the help of Neural network technique. Their dataset consists of 78 records with 13 input variables on which various experiments are conducted and the system is trained. For heart disease diagnosis author has suggested supervised network which is trained using Back Propagation algorithm.When the system is input with new data, it will find it from the trained data and generate list of probable diseases that may be occurred by the patient.The success rate of the system to give desired output from the inaccurate input is 100%. The results of the study show that Neuralnetworks has the extreme capability to be used as an indexing function.It is used for modeling and prediction of experimental data, thus is a fast substitute to classical statistical techniques. The system can avoid human error. Thus, the system is reliable and assists medical practitioners in making accurate decision. Certainly, expert mind can’t be replaced by neural networks since expert is more reliable but it can assist human experts by cross checking their diagnosis. In CANFIS approach [12], the neural network adaptive capabilities are combined with fuzzy logic qualitative approachwhich is further integrated with genetic algorithm for diagnosis of heart disease. The Cleveland heart-disease database which comprises of 303 cases is used in the study. The proposed CANFIS model performance is evaluated with respect to training performances and classification accuracies. The results of the study show that the proposed model has great capability to predict the heart disease.

3. ANALYSIS OF EXISTING WORK FOR MEDICAL DATAMINING

In countries like China, India and Malaysia much work has been done in medical data mining and also specifically in cardiac data mining[9][10][11][12] on the basis of real and artificial data sets. All the research discussed above is based on either of these countries.Besides, most of the medical data mining work discussed above focuses on either clustering, classification or association rules miningwhile in some [9] [10] the details of results are not discussed and visualized properly.

However, there islittle research seen in Pakistan in building a framework specifically for cardiac data mining on the basis of real data obtained from some renowned cardiac hospital.Also there is need of framework that unifies the data mining tasks from data preparation to data visualization and discovery of knowledge. In our work analysis is based on results of data mining techniques like clustering, correlation as well as association rule mining with better and complete visualization of results.Further, the generated rules in our work are validated against completeness, interestingness and comprehensibility along with these three rudimentary quality measures like support, confidence and lift. Thus providesmore robust rule mining results.

Chapter 3

METHODOLOGY AND DATA PREPARATION

3.1 METHODOLOGY

In our research study of heart disease features and patterns extraction, we have used the well-known CRISP-DM methodology. The CRISP-DM was called the "de facto standard for developing data mining and knowledge discovery projects” in 2009. CRISP-DM is successful because it is based on practical and real world experience of people conducting data mining projects. Its methodology consists of set of tasks defined at four levels of abstraction: phase, generic task, specialized task, and process instance.Data mining process model defines the approaches and methods used by data mining experts to deal with difficulties [17].

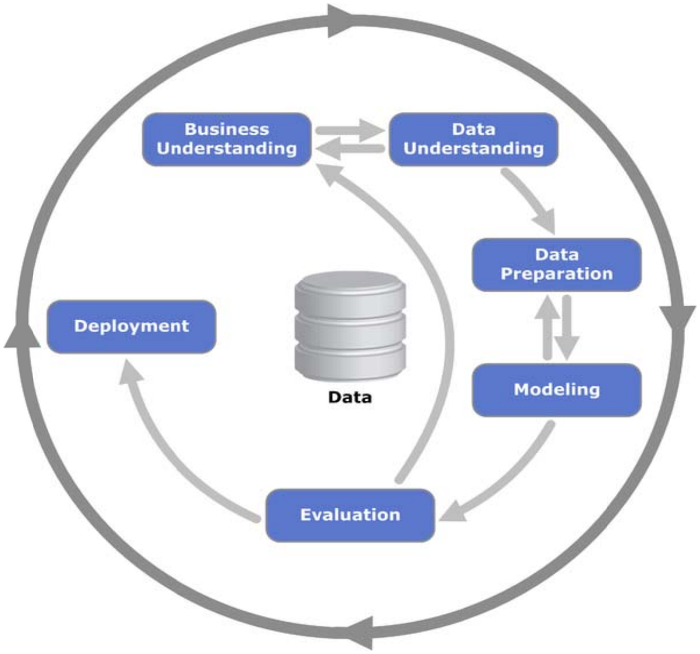


Figure 3.1: CRISP-DM Process Model [17]

The CRISP-DM reference model in figure 3.1 presents the whole data mining project life cycle.This model is divided into six phases. Each phase has its particular inputs, tasks and outputs. . The figure above shows the phases of the model whose sequence is not rigid. The decision about next phase or task is determined on the basis of output of the previous phase. The arrows between the phases show the imperative and recurrent dependencies.

3.2 MODELING TOOLS

Rapid Miner® 5.3, Microsoft SQL Server R2 2008, Crystal Reports, Microsoft Excel 2010.

- Rapid Miner® by Rapid-I was used as primary analytical tool for data pre-processing and algorithms implementation and other data mining related tasks.
- Microsoft SQL Server 2008 R2 was used as a database tool for data preparation, transformation into numerical form and restructuring after getting text reports from AFIC.
- Crystal Report is also used for some reporting purposes.
- Microsoft Excel 2010 pivot table is used for clustering results analysis and for better visualization of clusters MS Office charts utility is utilized.

3.3 DATA PREPARATION

3.3.1 DATA ACQUISITION AND UNDERSTANDING

The data used in this research project was obtained from AFIC, Rawalpindi, Pakistan which comprised of 300 historical heart patient reports in unstructured (textual) form. In relevant information extraction more than 80 medical attributes were extracted from each report manually during conversion from unstructured reports to structured format and then entered into the database. In order to extract important features and patterns these extracted attributes in database were then preprocessed by using Rapid Miner.

The heart patient's reports in AFIC have five main portions which are as follows,

a. Profile
b. Past History
c. Procedure
d. Scan Findings
e. Impressions

A brief explanation of the content within these portions is given below,

- **Patient profile**

This heading contains the information related to patient profile. Some important attributes extracted from this portion are name, age, gender, contact info, Ref no. and order and signed date of the report.

- **History**

This portion contains the history information of patients. The attributes extracted from this portion are BMI, current and past diseases, past treatments and scans, results of past test and his current complains.

- **Procedure**

This portion comprises the details of the current procedures applied on the patient. The procedures currently applied are divided into two Protocols which are:

1. Adenosine Infusion Protocol
2. Bruce Protocol

Each protocol containsdifferent procedures (particular tests and results) carried out on patients to diagnose his actual problem.For the decision of protocol implementation, a patients past history which includespast diseases, scans, past and current biological history and complains are taken into account. After this, diagnosis is done using a series of steps involved in each type of protocol as shown in figure 3.2.

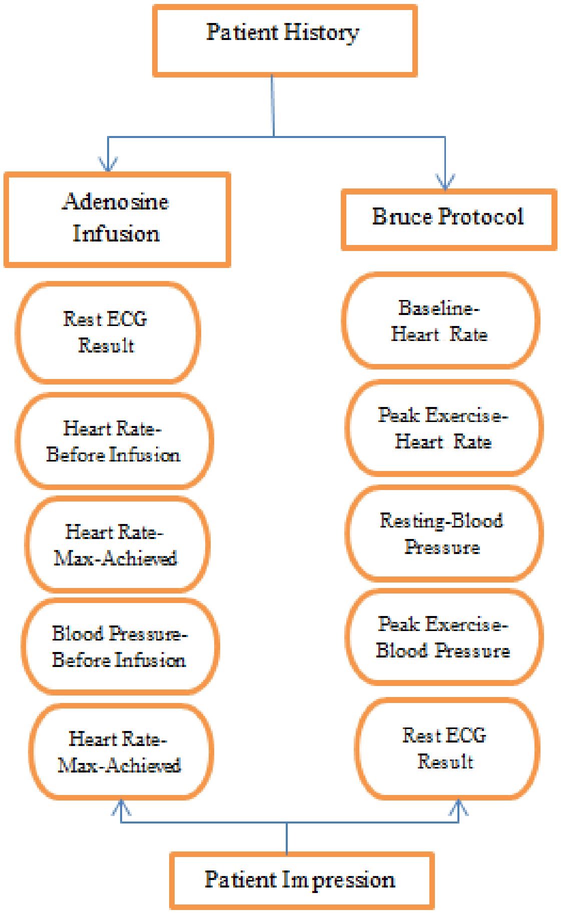


Figure 3.2: Steps of Procedure/Diagnosis of patient condition

- **Scan Findings**

This portion includes the details of the findings which the cardiologist gets from various visualizations of the patients heart through different computer aides and high tech machines. These visualizations include the cine mode, the tracer uptake images and oral observations of the cardiologist.

- **Impressions**

This portion comprises the impressions deduced by the cardiologist by examining the patient history, the results of the procedures and the scan findings. This includes detail about the patient's condition as concluded by the cardiologist. The cardiologist specifically points out the affected regions of the heart whether they are defected or ischemic etc. The defect area size is also specified here and whether the defect is viable or non-viable. Besides, the risk factors and regions are alsoidentified in this portion.The important attributes extracted from this portion are:

1. LVEF
2. Viable/Non-viable affected Regions
3. Heart part condition (Defected or ischemic etc.)

3.3.1.1 Business need of mining the unstructured data

According to a source [18], more than 80 % of the medical data that could be patients reports, lab results, doctors reviews etc. are in unstructured form. All the unstructured data sets in healthcare organizations are not always compatible with one another. So extracting useful information from such data sets is an imperative issue for data scientists. If the organizations are not able to extract meaningful and useful information from the dataset then there is a probability that they miss out some important information, chances to develop a better medical decision support system and enhance patient care and operational efficiency [19].

3.3.1.2 Drawbacks of Unstructured data

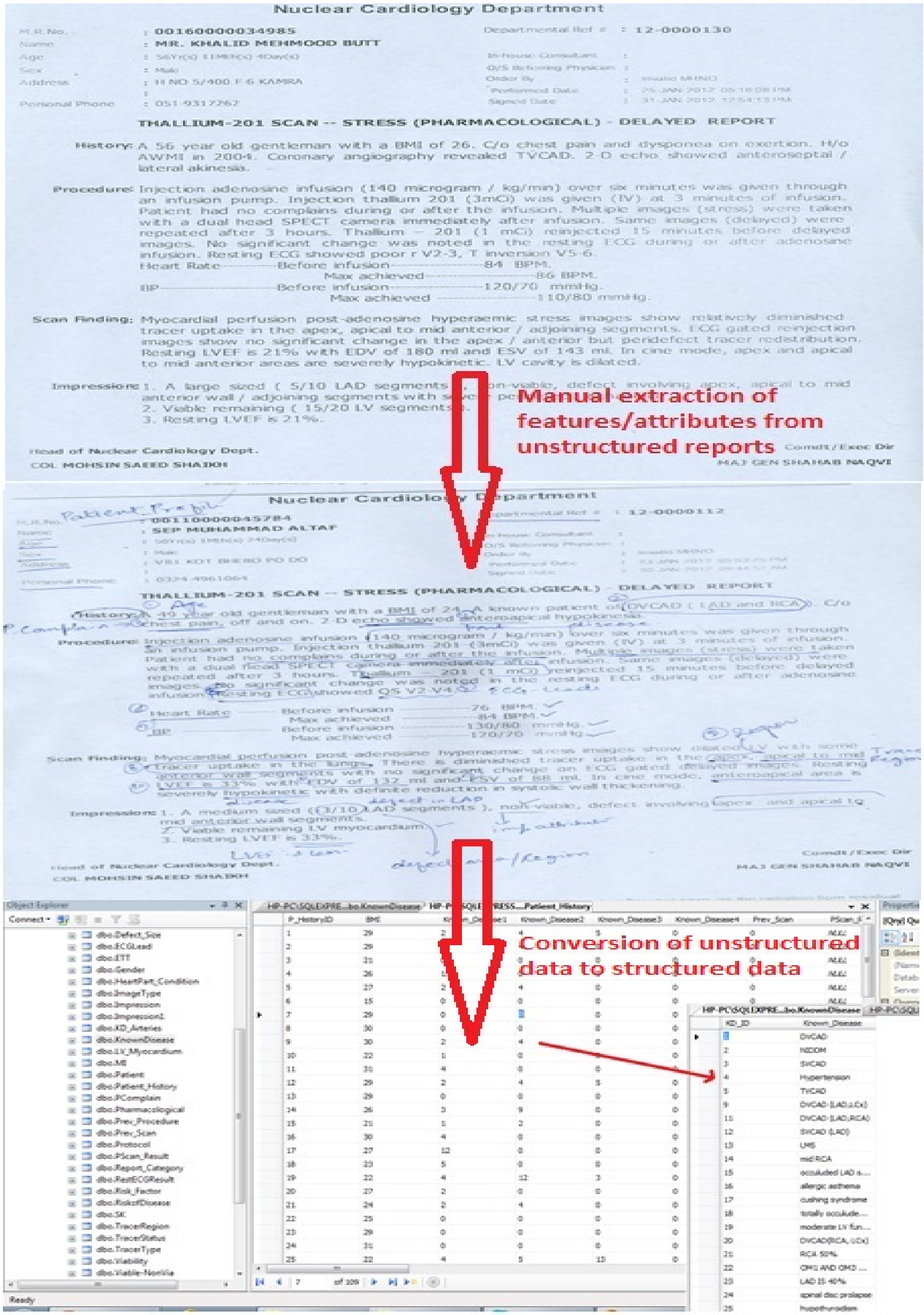
Human minds are good at analyzing and dealing with unstructured data in small amounts. However, human minds fail to handle huge amount of data. Then we need automated systems to manage and process large size data and for those automated systems, we need structured data [20].

3.3.1.3. Benefits of Structured data

When the unstructured data is organized and placed in a database, it is called structured data. The structured data is easily understandable by computers and is more convenient for simple machine learning models. Besides, it also allows using wide variety of algorithms that are available for Data Mining.

3.3.2 HANDLING NON-STANDARD DATA

The collected unstructured reports are gone through extraction process by selecting all the noteworthy attributes/features with the help of expert doctor opinion. After, these attributes are entered into MS SQL Server database which is used to store them for further processing.


Manual extraction of features/attributes from unstructured reports
Conversion of unstructured data to structured data
Patient Profile
① Age
P. Complain
Known disease
ECG - Lead
Region
Tracer Region
defect in LAD
Imp attribute
LVEF
defect area/Region

Nuclear Cardiology Department

M.R.No.	: 00160000034985	Departmental Ref #	: 12-0000130
Name	: MR. KHALID MEHMOOD BUTT		
Age	: 56Yr(s) 11Mth(s) 4Day(s)	In-house Consultant	:
Sex	: Male	O/S Referring Physician	:
Address	: H NO 5/400 F 6 KAMRA	Order By	: Invalid MRNO
	:	Performed Date	: 25-JAN-2012 05:18:08 PM
Personal Phone	: 051-9317262	Signed Date	: 31-JAN-2012 12:54:13 PM

THALLIUM-201 SCAN -- STRESS (PHARMACOLOGICAL) - DELAYED REPORT

History: A 56 year old gentleman with a BMI of 26. C/o chest pain and dysponea on exertion. H/o AWMI in 2004. Coronary angiography revealed TVCAD. 2-D echo showed anteroseptal / lateral akinesia.

Procedure: Injection adenosine infusion (140 microgram / kg/min) over six minutes was given through an infusion pump. Injection thallium 201 (3mCi) was given (IV) at 3 minutes of infusion. Patient had no complains during or after the infusion. Multiple images (stress) were taken with a dual head SPECT camera immediately after infusion. Same images (delayed) were repeated after 3 hours. Thallium – 201 (1 mCi) reinjected 15 minutes before delayed images. No significant change was noted in the resting ECG during or after adenosine infusion. Resting ECG showed poor r V2-3, T inversion V5-6.

Heart Rate----------Before infusion--------------------84 BPM.
Max achieved--------------------86 BPM.
BP--------------------Before infusion--------------------120/70 mmHg.
Max achieved --------------------110/80 mmHg.

Scan Finding: Myocardial perfusion post-adenosine hyperaemic stress images show relatively diminished tracer uptake in the apex, apical to mid anterior / adjoining segments. ECG gated reinjection images show no significant change in the apex / anterior but peridefect tracer redistribution. Resting LVEF is 21% with EDV of 180 ml and ESV of 143 ml. In cine mode, apex and apical to mid anterior areas are severely hypokinetic. LV cavity is dilated.

Impression: 1. A large sized (5/10 LAD segments), non-viable, defect involving apex, apical to mid anterior wall / adjoining segments with se[illegible] pe[illegible]
2. Viable remaining (15/20 LV segments).
3. Resting LVEF is 21%.

Head of Nuclear Cardiology Dept.
COL MOHSIN SAEED SHAIKH

Comdt/Exec Dir
MAJ GEN SHAHAB NAQVI

Nuclear Cardiology Department

M.R.No.	: 00110000045784	Departmental Ref #	: 12-0000112
Name	: SEP MUHAMMAD ALTAF		
Age	: 50Yr(s) 1Mth(s) 24Day(s)	In-house Consultant	:
Sex	: Male	O/S Referring Physician	:
Address	: VILL KOT BHERO PO DO	Order By	: Invalid MRNO
	:	Performed Date	: 20-JAN-2012 05:53:25 PM
Personal Phone	: 0324-4961064	Signed Date	: 30-JAN-2012 09:44:52 AM

THALLIUM-201 SCAN -- STRESS (PHARMACOLOGICAL) - DELAYED REPORT

History: A 40 year old gentleman with a BMI of 24. A known patient of DVCAD (LAD and RCA). C/o chest pain, off and on. 2-D echo showed anteroapical hypokinesia.

Procedure: Injection adenosine infusion (140 microgram / kg/min) over six minutes was given through an infusion pump. Injection thallium 201 (3mCi) was given (IV) at 3 minutes of infusion. Patient had no complains during or after the infusion. Multiple images (stress) were taken with a dual head SPECT camera immediately after infusion. Same images (delayed) were repeated after 3 hours. Thallium – 201 (1 mCi) reinjected 15 minutes before delayed images. No significant change was noted in the resting ECG during or after adenosine infusion. Resting ECG showed QS V2-V4.

Heart Rate ---------- Before infusion -------------------- 76 BPM. ✓
Max achieved -------------------- 84 BPM. ✓
BP ---------- Before infusion -------------------- 130/80 mmHg. ✓
Max achieved -------------------- 120/70 mmHg. ✓

Scan Finding: Myocardial perfusion post-adenosine hyperaemic stress images show dilated LV with some tracer uptake in the lungs. There is diminished tracer uptake in the apex, apical to mid anterior wall segments with no significant change on ECG gated delayed images. Resting LVEF is 33% with EDV of 132 ml and ESV of 88 ml. In cine mode, anteroapical area is severely hypokinetic with definite reduction in systolic wall thickening.

Impression: 1. A medium sized (3/10 LAD segments), non-viable, defect involving apex and apical to mid anterior wall segments.
2. Viable remaining LV myocardium.
3. Resting LVEF is 33%.

Head of Nuclear Cardiology Dept.
COL MOHSIN SAEED SHAIKH

Comdt/Exec Dir
MAJ GEN SHAHAB NAQVI

dbo.Defect_Size
dbo.ECGLead
dbo.ETT
dbo.Gender
dbo.HeartPart_Condition
dbo.ImageType
dbo.Impression
dbo.Impression1
dbo.KD_Arteries
dbo.KnownDisease
dbo.LV_Myocardium
dbo.MI
dbo.Patient
dbo.Patient_History
dbo.PComplain
dbo.Pharmacological
dbo.Prev_Procedure
dbo.Prev_Scan
dbo.Protocol
dbo.PScan_Result
dbo.Report_Category
dbo.RestECGResult
dbo.Risk_Factor
dbo.RiskofDisease
dbo.SK
dbo.TracerRegion
dbo.TracerStatus
dbo.TracerType
dbo.Viability
dbo.Viable-NonVia

HP-PC\SQLEXPRE...bo.KnownDisease | HP-PC\SQLEXPRESS...Patient_History

P_HistoryID	BMI	Known_Disease1	Known_Disease2	Known_Disease3	Known_Disease4	Prev_Scan	PScan_R
1	29	2	4	5	0	0	NULL
2	29	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	NULL
3	31	[illegible]	0	0	0	0	NULL
4	26	1	0	0	0	0	NULL
5	27	2	4	0	0	0	NULL
6	15	0	0	0	0	0	NULL
7	29	0	[illegible]	0	0		
8	30	0	0	0	0		
9	30	2	4	0	0		
10	22	1	0	[illegible]	0		
11	31	4	0	0	0		
12	29	2	4	5	0		
13	29	0	0	0	0		
14	26	3	9	0	0		
15	21	1	2	0	0		
16	30	4	0	0	0		
17	27	12	0	0	0		
18	23	5	0	0	0		
19	22	4	12	3	0		
20	27	2	0	0	0		
21	24	2	4	0	0		
22	25	0	0	0	0		
23	29	0	0	0	0		
24	31	0	0	0	0		
25	22	4	5	13	0		

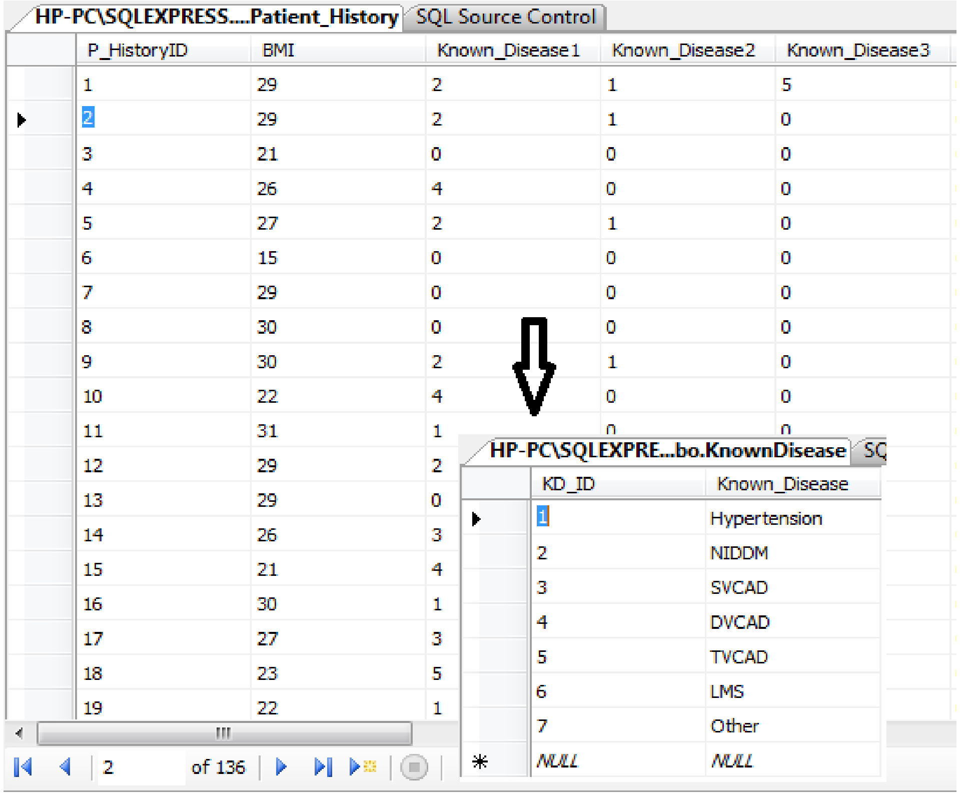
7 of 109

HP-PC\SQLEXPRE...bo.KnownDisease | HP-PC\SQL

KD_ID	Known_Disease
1	DVCAD
2	NIDDM
3	SVCAD
4	Hypertension
5	TVCAD
9	DVCAD (LAD,LCx)
11	DVCAD (LAD,RCA)
12	SVCAD (LAD)
13	LMS
14	mid RCA
15	occluded LAD s...
16	allergic asthema
17	cushing syndrome
18	totally occlude...
19	moderate LV fun...
20	DVCAD(RCA, LCx)
21	RCA 50%
22	OM1 AND OM3 ...
23	LAD IS 40%
24	spinal disc prolapse
25	hypothyroidism

Figure3.3: Manual Feature Extraction and conversionfrom unstructured (text) data to structured (tabular) data

To make stored data compatible with machine learning algorithms like k-means, k -medoids and decision tree models we made mapping tables in SQL Server against different features which helped us to convert structured data to numerical form as shown in figure 4.3 below.



HP-PC\SQLEXPRESS....Patient_History
SQL Source Control

P_HistoryID	BMI	Known_Disease1	Known_Disease2	Known_Disease3
1	29	2	1	5
2	29	2	1	0
3	21	0	0	0
4	26	4	0	0
5	27	2	1	0
6	15	0	0	0
7	29	0	0	0
8	30	0	0	0
9	30	2	1	0
10	22	4	0	0
11	31	1	0	0
12	29	2		
13	29	0		
14	26	3		
15	21	4		
16	30	1		
17	27	3		
18	23	5		
19	22	1		

HP-PC\SQLEXPRE...bo.KnownDisease
SQ

KD_ID	Known_Disease
1	Hypertension
2	NIDDM
3	SVCAD
4	DVCAD
5	TVCAD
6	LMS
7	Other
NULL	NULL

Figure3.4: Mapping Table of ‘Known Disease’ Attribute

In the above figure3.4‘Patients_History’ table is presented that shows how the structured data is mapped with numerical form by replacing each value of different attributes i.e. attribute ‘Known_Disease1’ values Hypertension, Diabetes and SVCAD are replaced by 1, 2 and 3 in the Patient_History table that are taken from ‘Known_Disease’ mapping table.

3.3.3DESCRIPTION OF DATABASE

Theprepared heart disease data set consists of 300 records with 80 medical attributes. The data set consists of mixed attributes comprising both categorical, binomial and numeric data.The dataset was created in MS SQL Server 2008. A view is created joining all relevant tables that includes all important attributes. This view contained 80 attributes which were then imported to Microsoft® Excel 2010 for further processing. The description and values of some of the important attributes are stated in below table.

Table3.1: Description of Heart patients Database

S#	Observation	Description	Values
1	Age	Age in years	Continuous
2	Gender	Subject gender	Male/Female
3	Protocol	Procedure applied on the patient for diagnosis	Adenosine Infusion Bruce protocol
4	BMI	Body Mass Index	Continuous
5	Known_Disease	Disease already present.	Five main types: • Hypertension • diabetes • SVCAD • DVCAD • TVCAD
6	MI_Type	Type of Heart attack in past	Four main types: • Anterior • Inferior • Lateral • Septum
7	Angiography_Result	The disease diagnosed through angiography	Three main types • SVCAD • DVCAD • TVCAD
8	2DEcho_EF	Ejection Fraction	Continuous
9	2DEchoResult_Disease	The disease diagnosed through 2DEcho test.	2 main types: • Hyperkinesia • Akinesia
10	Patient_Complain	Complain of patient	Five main types • Hypertension • chest pain • dyspnea • palpitation • fatigue
11	Patient_Condition	Condition of patient before and after	Two main types

		infusion	• normal • abnormal
12	Resting_ECG_Result	ECG Result	Five main types: • normal limits • T-inversion • QS • RBBB • LBBB
13	Heart_Rate	Heart Ratebefore infusion and maximum achieved, Heart Rate baseline and peak exercise, measured in BPM	Continuous
14	Blood Pressure	Before and after Infusion and Resting and peak exercise, Both upper and lower limit measured in mmHg	Continuous
15	Diagnosed_Complaint	Complaint/situation of patient diagnosed after applying procedure.	Two main types: • Defected • Ischemic
16	LV_Condition	LV myocardium	Four main types: • good function • viable remaining • globally ischemic • poor function
17	Defect size	Defected heart area size	From 1-10
18	Viable/Non-Viable	Either the defect is viable or non-viable	Viable:0 Non-viable:1
19	Defect_Segment	Defected portion of heart	Five types: • lateral • inferior • apex • Septum • Anterior
20	LVEF	Left Ventricular Ejection Fraction	Continuous

Chapter 4

PROPOSED CARDIAC DATA MINING APPROACH

In this chapter we introduce our proposed framework for biomedical (cardiac) data mining: Extraction of significant patterns for predicting heart condition.The main aim of this research is to provide a compact biomedical data miningapproach based on prepared and pre-processed data gathered from patient reports to help determine patterns and trends to improve decision making.So, in this chapter, we present a detailed description of our approach and explain that how we carried out pre-processing, categorization and pattern extraction tasks.

4.1 PROPOSED FRAMEWORK

The proposed framework provides a unifying vision of steps and methods used in cardiac data mining. The basic scheme for proposed framework constitutes three phases which are:

1. Data Preparation
2. Data Pre-processing
3. Medical Data Mining

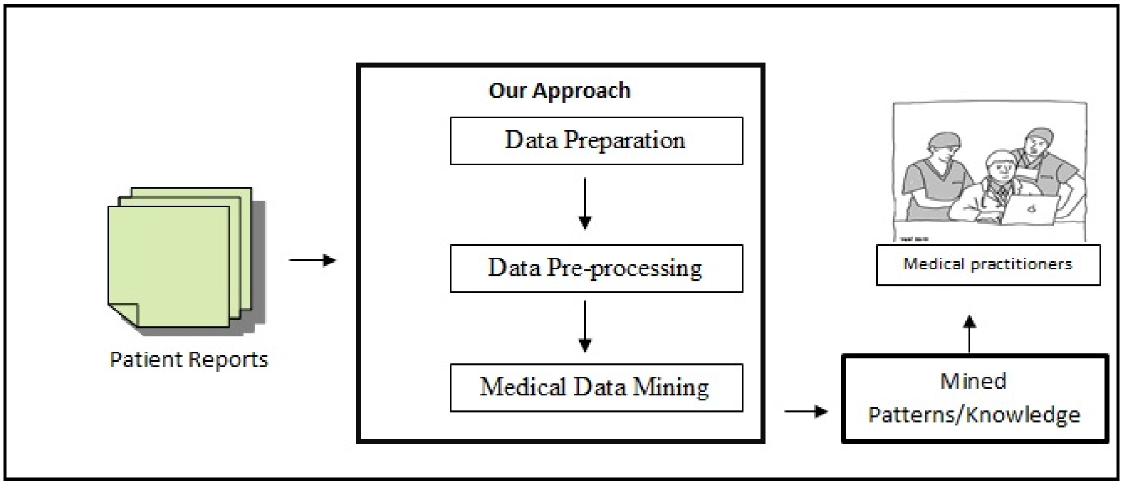


Figure 4.1: Block diagram illustrating architecture of our proposed approach

The figure 4.1 shows the basic architecture of our proposed approach. In first phase, patient medical reports are collected and a profound understanding was developed after detailed study of patient’s medical reports and organizing meetings with medical experts.This step also helps to deeply understand the problem area and this deep knowledge can help the data mining resource to generate better information out of any Data Mining Algorithm. After, attributes/features are accessed and extracted manually keeping expert opinion in mind and the stop criterion conditional step determines whether the retrieved attributes are good/mature enough to proceed and they covers future reports as well. For this the extracted attributes are tested for conformance to an additional 30 new patient reports. In the second phase data pre-processing is done and earlier collected attributes are also refined and important attributes are chosen. After data mining techniques like correlation, clustering and association rule mining is applied on final dataset. The analysis is done and the results are used to extract significant patterns/knowledge relevant to heart condition which can aid medical practitioners in making intelligent verdicts.Further, the selected k-means algorithm is compared with other clustering algorithm and performance is evaluated.

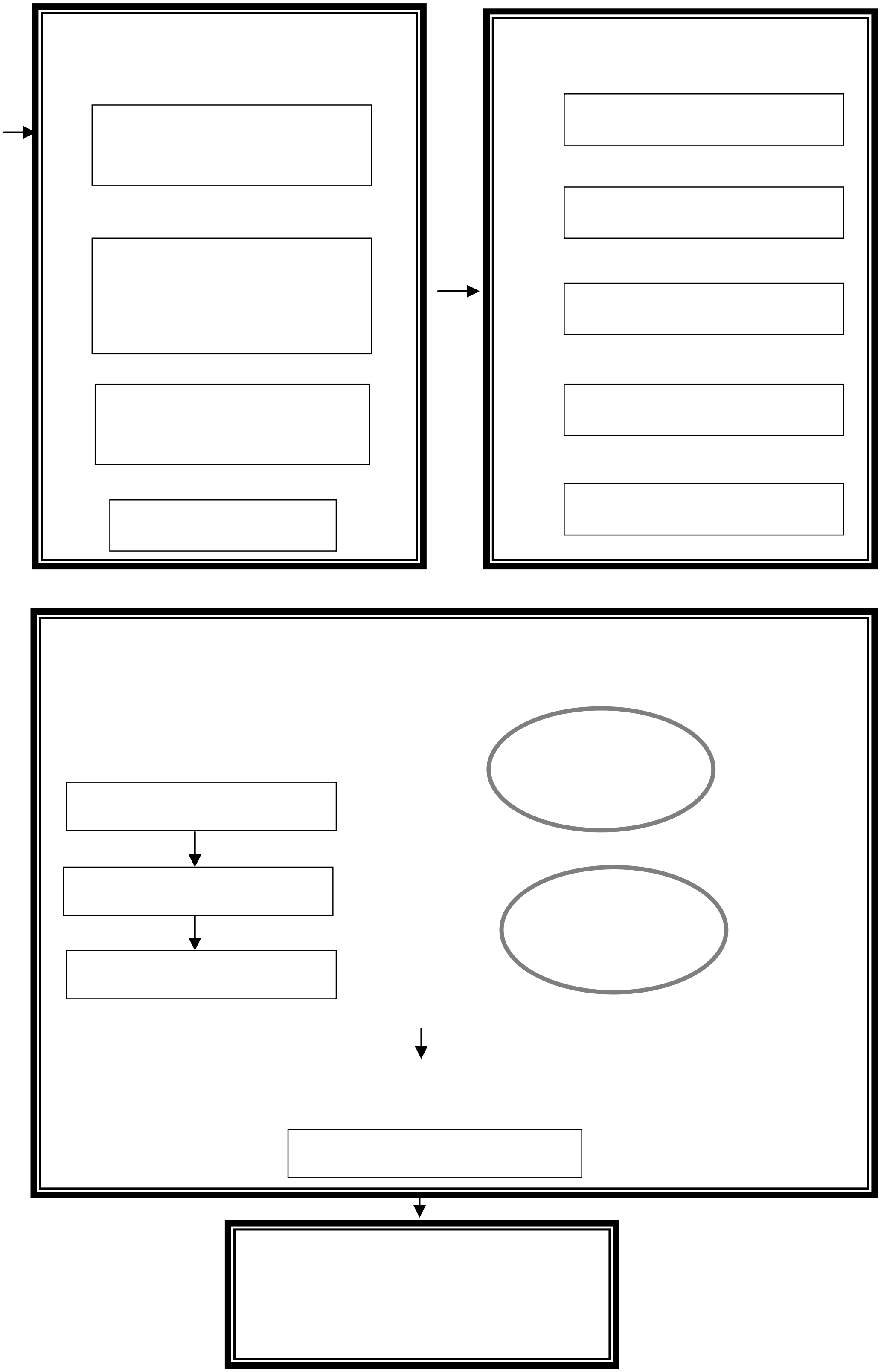


Figure 4.2: Framework of cardiac data mining

The above figure 4.2illustrates steps during each phase of proposed approach. Following sections give a detailed understanding of each phase.

4.2 DATAPRE-PROCESSING

Data Pre-processingis an important step in knowledge discovery process and consider as essential building block of data mining. The real world data is almost useless without this.Data mining project success is dependent on how good the Extraction,Transformation and Loading(ETL) process is carried out.Data Pre-processing which is part of ETL is basically the transformation of source data into a different format which ensure [21]:

- Easy application of data mining algorithms.
- Improves theperformance and effectivenessof mining algorithms.
- Represents the data in easily understandable format for both humans and machines.
- Supports faster data retrieval from databases.
- Makes the data suitable for a specific analysis to be performed.

After converting the data into structured and then numeric form, the collected data undergoes from different steps as there was some missing, erratic, and identical data that needs to go through from scrubbing and sifting in order to avoid the creation of misleading or incongruous rules or patterns. These steps are:

4.2.1 Data Cleaning

Data Cleaningis mandatory because usually data sets don't work without this. Data cleaning is performed in which missing values are replaced because many clustering algorithms don't allow null data. So, the missing values in the structured data are replaced with a suitable value that is zero in our case in order to make the data compatible to data mining algorithm.The zero shows that the data is not present and it makes sure that every blank data element has the same meaning/answer.

4.2.2 Data Type conversion

Some attributes in the data set require data type conversioninto numeric as K-Means algorithm can't handle binomial, polynomial or nominal data. And while nominal to numeric conversion an ordinal approach was considered for all variables mapping i.e. critical end numerical values were assigned to 'alarming situation' that tends towards larger value and less critical end values were assigned to 'less alarming situation' that tends towards smaller value.

4.2.3 Data Normalization

Data normalization is of different types. We have applied Min-Max Normalization to scale the attribute value into the specified range that is [0.0, 1.0]. It transforms a value A to B which fits in the range[C,D].The formula of this is given below:

$$B = \left(\frac{\text{A}-\text{minimum value of A}}{\text{maximum value of A}-\text{minimum value of A}}\right) * (D - C) + C \tag{3}$$

Min-max normalization accomplishes linear transformation on the original data values and reserves the relationships between original data values. It is recommended so that all attributes can have equal impact on the computation of distances (i.e. Euclidean distance) [22].

4.2.4 Remove Duplicates

Removal of duplicate records is applied on the dataset on the basis of comparison of all records with each other. For this purpose, two records are considered duplicate if selected attributes in that contains the similar values.

4.2.5Discretize by Binning

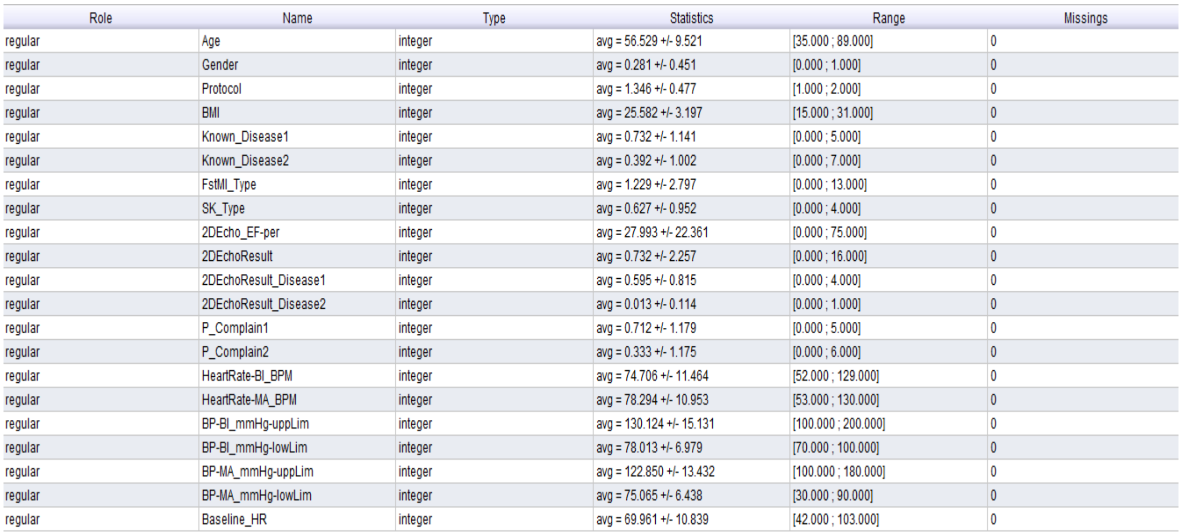
Discretize by Binningis done to transform selected continuous numerical variables into user specified categorical so that the results using binned variables can be easily read and interpreted. For example Age feature was varying from 35 to 90. And after discussion with medical practitioners we divided the Age feature into two different categories. In the same way LVEF range was divided into three categories [23]. The below table shows the binning of continuous variables:

Table 4.1: Binning of continuous numerical variables

S#	Variable Name	Binning range
1.	Gender	0: MALE 1: FEMALE
2.	Age	MATURE AGE<=50 OLD AGE= 50+
3.	LVEF	Normal LVEF: = 55-70% Near Alarming Stage LVEF:<55 Extremely Alarming LVEF:<=35

4.	BP-BI-UPPLIM	NORMAL: 90-130 Near Alarming Stage: 140-160 Extremely Alarming: >160
5.	BP-BI-LOWLIM	NORMAL: 60-80 ABNORMAL: ABOVE 80 AND BELOW 60
6.	HEARTRATE-BI-BPM	NORMAL: 60-100 ABNORMAL: BELOW 60 AND ABOVE 100
7.	RISK OF DISEASE	0: NO RISK 1: RISK
8.	BASELINE_HR	NORMAL: 60-100 ABNORMAL: BELOW 60 AND ABOVE 100
9.	RESTING_BP_UPP	NORMAL: 120-130 Near Alarming Stage: 130-159 Extremely Alarming Stage: >160
10.	RESTING_BP-LOW	NORMAL: 80 and <80 ABNORMAL: >80
11.	Is_Defected	0: NOT DEFECTED 1: DEFECTED

By going through the above steps, the dataset is converted into a standardize form that is suitable for any data mining algorithm. The table below displaying Name, Type, Statistics, Range and Missing columns of attributes in our dataset shows that our data now becomes very clean with no missing, duplicate and inconsistent data so there is no need for further preprocessing on data.



Role	Name	Type	Statistics	Range	Missings
regular	Age	integer	avg = 56.529 +/- 9.521	[35.000 ; 89.000]	0
regular	Gender	integer	avg = 0.281 +/- 0.451	[0.000 ; 1.000]	0
regular	Protocol	integer	avg = 1.346 +/- 0.477	[1.000 ; 2.000]	0
regular	BMI	integer	avg = 25.582 +/- 3.197	[15.000 ; 31.000]	0
regular	Known_Disease1	integer	avg = 0.732 +/- 1.141	[0.000 ; 5.000]	0
regular	Known_Disease2	integer	avg = 0.392 +/- 1.002	[0.000 ; 7.000]	0
regular	FstMI_Type	integer	avg = 1.229 +/- 2.797	[0.000 ; 13.000]	0
regular	SK_Type	integer	avg = 0.627 +/- 0.952	[0.000 ; 4.000]	0
regular	2DEcho_EF-per	integer	avg = 27.993 +/- 22.361	[0.000 ; 75.000]	0
regular	2DEchoResult	integer	avg = 0.732 +/- 2.257	[0.000 ; 16.000]	0
regular	2DEchoResult_Disease1	integer	avg = 0.595 +/- 0.815	[0.000 ; 4.000]	0
regular	2DEchoResult_Disease2	integer	avg = 0.013 +/- 0.114	[0.000 ; 1.000]	0
regular	P_Complain1	integer	avg = 0.712 +/- 1.179	[0.000 ; 5.000]	0
regular	P_Complain2	integer	avg = 0.333 +/- 1.175	[0.000 ; 6.000]	0
regular	HeartRate-BI_BPM	integer	avg = 74.706 +/- 11.464	[52.000 ; 129.000]	0
regular	HeartRate-MA_BPM	integer	avg = 78.294 +/- 10.953	[53.000 ; 130.000]	0
regular	BP-BI_mmHg-uppLim	integer	avg = 130.124 +/- 15.131	[100.000 ; 200.000]	0
regular	BP-BI_mmHg-lowLim	integer	avg = 78.013 +/- 6.979	[70.000 ; 100.000]	0
regular	BP-MA_mmHg-uppLim	integer	avg = 122.850 +/- 13.432	[100.000 ; 180.000]	0
regular	BP-MA_mmHg-lowLim	integer	avg = 75.065 +/- 6.438	[30.000 ; 90.000]	0
regular	Baseline_HR	integer	avg = 69.961 +/- 10.839	[42.000 ; 103.000]	0

Figure 4.3: Meta Data view of our dataset.

4.3RELEVANCE ANALYSIS /FEATURE SELECTION

Feature selection is a task in which some of the original and irrelevant attributes are omitted to decrease computation time and achieve simplicity in representation.For that we analyzed the model produced by Rapid Miner as a result of k-means algorithm and noted the attributes which play the dominant role in results. The figure below shows the Centroid Plot view of the model.

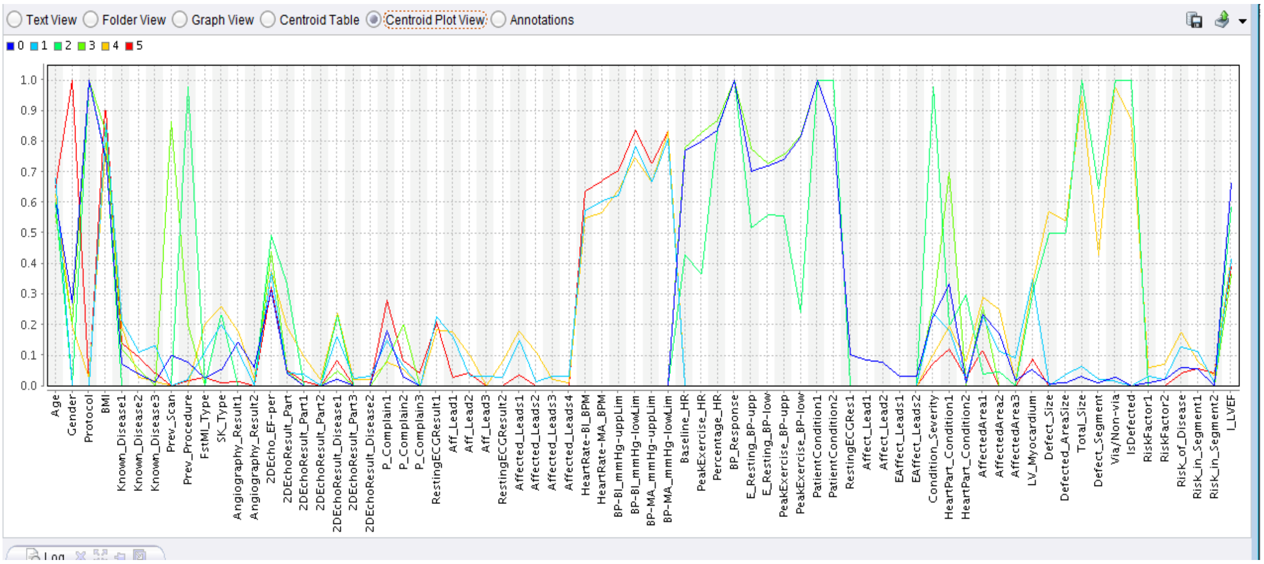
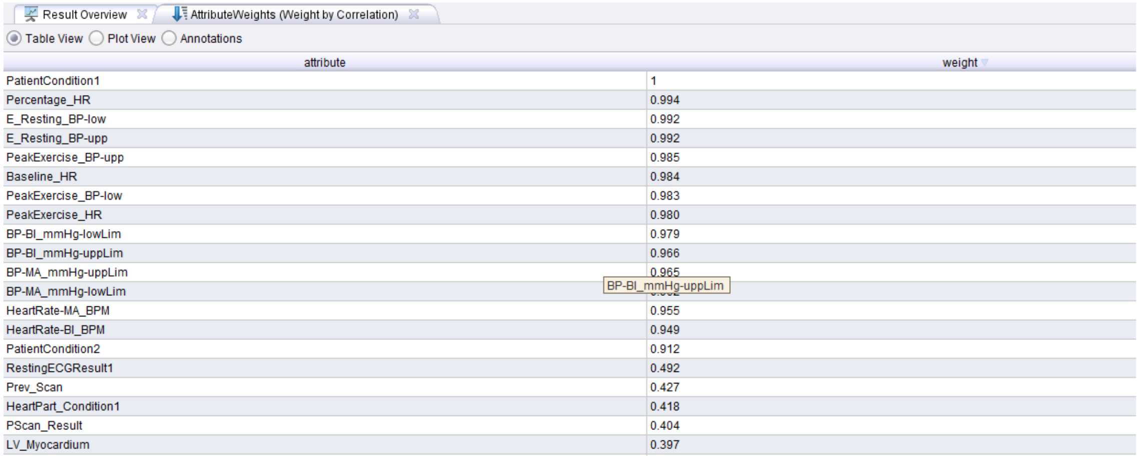


Figure 4.4: Centroid Plot view of K-means algorithm

In addition, ‘weight by correlation’ is generated in Rapid Miner in which the weight of all attributes is computed using correlation. The weights are normalized in the range from 0 to 1. The attributes with higher weights are considered most relevant. The figure below shows the output of this process.


Result Overview
AttributeWeights (Weight by Correlation)
● Table View ○ Plot View ○ Annotations

attribute	weight
PatientCondition1	1
Percentage_HR	0.994
E_Resting_BP-low	0.992
E_Resting_BP-upp	0.992
PeakExercise_BP-upp	0.985
Baseline_HR	0.984
PeakExercise_BP-low	0.983
PeakExercise_HR	0.980
BP-BI_mmHg-lowLim	0.979
BP-BI_mmHg-uppLim	0.966
BP-MA_mmHg-uppLim	0.965
BP-MA_mmHg-lowLim	0.962
HeartRate-MA_BPM	0.955
HeartRate-BI_BPM	0.949
PatientCondition2	0.912
RestingECGResult1	0.492
Prev_Scan	0.427
HeartPart_Condition1	0.418
PScan_Result	0.404
LV_Myocardium	0.397

Figure 4.5: Weight by Correlation

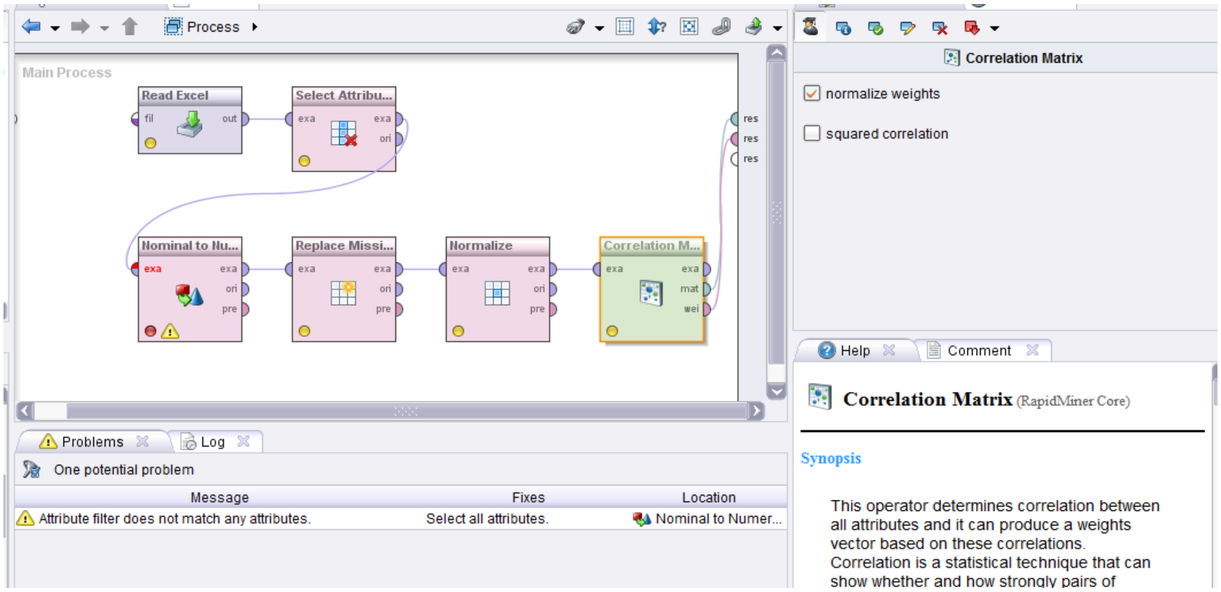
Further, keeping in mind the expert doctor's opinion and extensive study, a comparative analysis of k-means was done with k-medoids, X-means, k-Means (fast) and DBSCAN to observethe leading attributes. The total attributes shortlisted are 24 as shown in table 4.2below:

Table 4.2: Attributes extracted by applying clustering algorithms

Age	BP-MA-upplim	First-MI
Gender	BP-MA-lowlim	BMI
Protocol	Baseline-HR	2D-Echo_EF
Known-Disease	PeakExercise-HR	Heart Part-Condition
HeartRate-BI	Resting-BP-upplim	LV-Condition
HeartRate-MA	Resting-BP-lowlim	Defect-Size
BP-BI-upplim	PeakExercise-BP-upplim	Risk Factor
BP-BI-lowlim	PeakExercise-BP-lowlim	LVEF

4.4CORRELATION- MODELING

Statistically-oriented in nature, correlation has seen increasing use as a data mining technique. The prepared dataset is retrieved in RapidMiner in order to check the correlation between attributes. Correlation isbasically a statistical measure of how strong the relationships are between attributes in a data set.The 'Correlation Matrix Operator' is used to calculate the correlation between all attributes of the input data set.The number of attributes selected to check correlation are 30 among the 80 total attributes. The figure below shows the process of correlation and the operators used to design the model.



Correlation Matrix

☑ normalize weights

□ squared correlation

One potential problem

Message	Fixes	Location
Attribute filter does not match any attributes.	Select all attributes.	Nominal to Numer...

Correlation Matrix (RapidMiner Core)

Synopsis

This operator determines correlation between all attributes and it can produce a weights vector based on these correlations. Correlation is a statistical technique that can show whether and how strongly pairs of

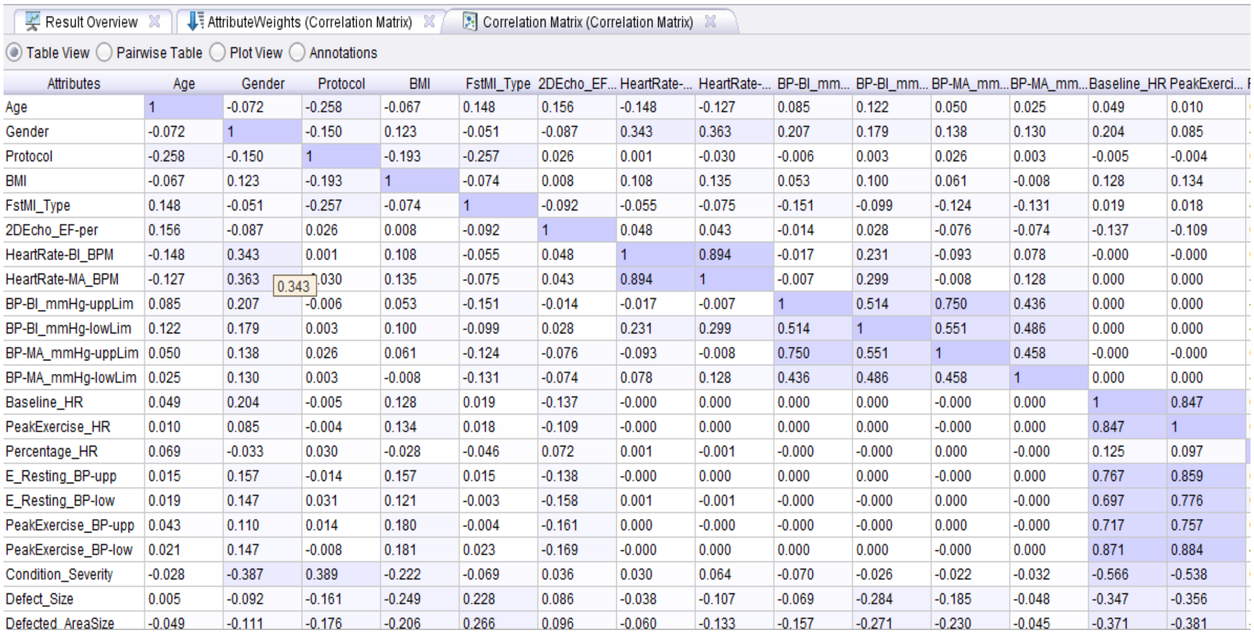
Figure4.6:Correlation matrix process

The correlation coefficient measures the linear relationship between two attributes or columns of data. The value can range from -1 to +1 [24].

A value [24]:

- near 0 indicates little correlation between attributes.
- near +1 or -1 indicates a high level of correlation.
- positive correlation coefficient: an increase in the value of one attribute indicates increase in the value of the other attribute.
- negative correlation coefficient: value less than 0, an increase in value of one attribute shows decrease in the value of other.

When process in Figure was run Correlation matrix was generated as shown in Figure 4.7 below.


Result Overview
AttributeWeights (Correlation Matrix)
Correlation Matrix (Correlation Matrix)
● Table View
○ Pairwise Table
○ Plot View
○ Annotations
0.343

Attributes	Age	Gender	Protocol	BMI	FstMI_Type	2DEcho_EF...	HeartRate-...	HeartRate-...	BP-BI_mm...	BP-BI_mm...	BP-MA_mm...	BP-MA_mm...	Baseline_HR	PeakExerci...
Age	1	-0.072	-0.258	-0.067	0.148	0.156	-0.148	-0.127	0.085	0.122	0.050	0.025	0.049	0.010
Gender	-0.072	1	-0.150	0.123	-0.051	-0.087	0.343	0.363	0.207	0.179	0.138	0.130	0.204	0.085
Protocol	-0.258	-0.150	1	-0.193	-0.257	0.026	0.001	-0.030	-0.006	0.003	0.026	0.003	-0.005	-0.004
BMI	-0.067	0.123	-0.193	1	-0.074	0.008	0.108	0.135	0.053	0.100	0.061	-0.008	0.128	0.134
FstMI_Type	0.148	-0.051	-0.257	-0.074	1	-0.092	-0.055	-0.075	-0.151	-0.099	-0.124	-0.131	0.019	0.018
2DEcho_EF-per	0.156	-0.087	0.026	0.008	-0.092	1	0.048	0.043	-0.014	0.028	-0.076	-0.074	-0.137	-0.109
HeartRate-BI_BPM	-0.148	0.343	0.001	0.108	-0.055	0.048	1	0.894	-0.017	0.231	-0.093	0.078	-0.000	-0.000
HeartRate-MA_BPM	-0.127	0.363	-0.030	0.135	-0.075	0.043	0.894	1	-0.007	0.299	-0.008	0.128	0.000	0.000
BP-BI_mmHg-uppLim	0.085	0.207	-0.006	0.053	-0.151	-0.014	-0.017	-0.007	1	0.514	0.750	0.436	0.000	0.000
BP-BI_mmHg-lowLim	0.122	0.179	0.003	0.100	-0.099	0.028	0.231	0.299	0.514	1	0.551	0.486	0.000	0.000
BP-MA_mmHg-uppLim	0.050	0.138	0.026	0.061	-0.124	-0.076	-0.093	-0.008	0.750	0.551	1	0.458	-0.000	-0.000
BP-MA_mmHg-lowLim	0.025	0.130	0.003	-0.008	-0.131	-0.074	0.078	0.128	0.436	0.486	0.458	1	0.000	0.000
Baseline_HR	0.049	0.204	-0.005	0.128	0.019	-0.137	-0.000	0.000	0.000	0.000	-0.000	0.000	1	0.847
PeakExercise_HR	0.010	0.085	-0.004	0.134	0.018	-0.109	-0.000	0.000	0.000	0.000	-0.000	0.000	0.847	1
Percentage_HR	0.069	-0.033	0.030	-0.028	-0.046	0.072	0.001	-0.001	-0.000	-0.000	0.000	-0.000	0.125	0.097
E_Resting_BP-upp	0.015	0.157	-0.014	0.157	0.015	-0.138	-0.000	0.000	0.000	0.000	-0.000	0.000	0.767	0.859
E_Resting_BP-low	0.019	0.147	0.031	0.121	-0.003	-0.158	0.001	-0.001	-0.000	-0.000	0.000	-0.000	0.697	0.776
PeakExercise_BP-upp	0.043	0.110	0.014	0.180	-0.004	-0.161	0.000	-0.000	-0.000	-0.000	0.000	-0.000	0.717	0.757
PeakExercise_BP-low	0.021	0.147	-0.008	0.181	0.023	-0.169	-0.000	0.000	0.000	0.000	-0.000	0.000	0.871	0.884
Condition_Severity	-0.028	-0.387	0.389	-0.222	-0.069	0.036	0.030	0.064	-0.070	-0.026	-0.022	-0.032	-0.566	-0.538
Defect_Size	0.005	-0.092	-0.161	-0.249	0.228	0.086	-0.038	-0.107	-0.069	-0.284	-0.185	-0.048	-0.347	-0.356
Defected_AreaSize	-0.049	-0.111	-0.176	-0.206	0.266	0.096	-0.060	-0.133	-0.157	-0.271	-0.230	-0.045	-0.371	-0.381

Figure4.7:Correlation matrix generated by RapidMiner

4.5UNSUPERVISED LEARNING STRATEGY- CLUSTERING TECHNIQUE

In Unsupervised Learning Strategy the data is unlabeled (no target attribute) and the purpose is to extract or discover information on the basis of understanding of input data only [25].Clustering is often called an unsupervised learningtask and it is one of the most used data mining techniques. This is used to find patterns in data by dividing the data into homogenous clusters. Clustering quality depends on maximizing the Inter-clusters distance and minimizing the Intra-clusters distance [26].

4.5.1K-MEANS CLUSTERING

K-Means is the most popular unsupervised learning algorithm. It partitions the given data into *k* clusters [4][10][29].

- Each cluster has a cluster center, called centroid.
- K is specified by the user.
- Given k, the k-means algorithm works as follows:
    1) Randomly choose k data points (seeds) to be the initial centroids, cluster centers.
    2) Assign each data point to the closest centroid.
    3) Re-compute the centroids using the current cluster memberships.

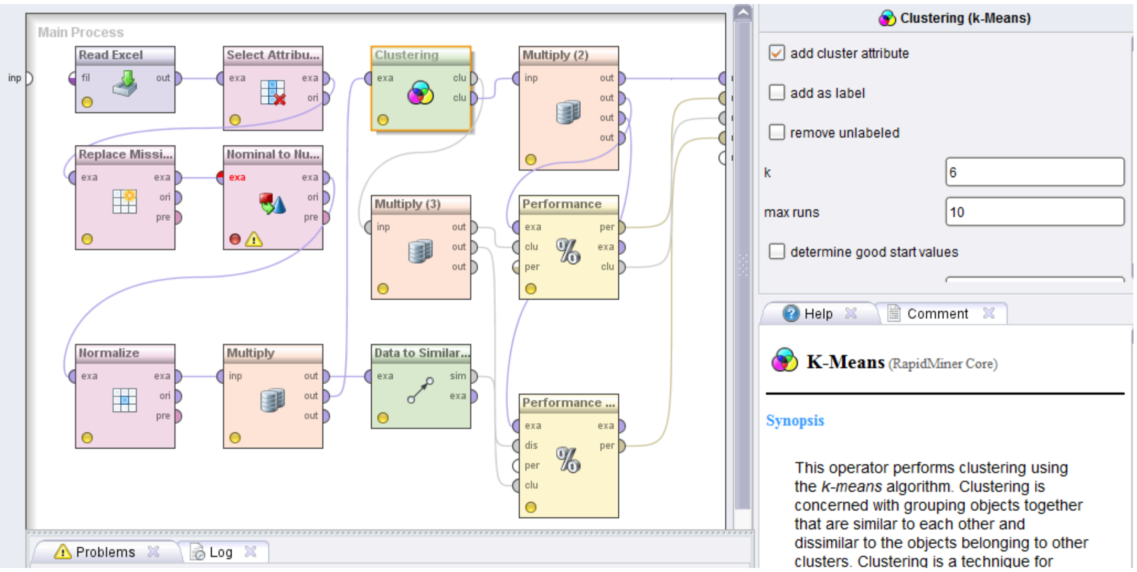
4) If a convergence criterion is not met, go to 2).

4.5.2 ADVANTAGES OF THIS TECHNIQUE

- K-means isfast and easy to understand/implementalgorithm.
- It is widely used in medical, scientific and industrial applications to solve the clustering problem [11] [12].
- This technique is used to get efficient and accurate results in diagnosing heart disease using real and artificial datasets [13].

4.5.3K-MEANS CLUSTERING- MODELING

K-means clustering is performed on our dataset comprising of 300 records with 100 medical attributes.This algorithm will cross-compare the data set in order to group it into specific clusters of related items. Analysis is done with the help of different tables and for better visualization of clusters MS Office charts utility is utilized. This helps to identify the significant patterns for predicting heart disease and also classifying the patients in different groups. The figure4.8 below shows the basic flow of how k-means algorithm is applied in Rapid Miner.



Clustering (k-Means)

☑ add cluster attribute

□ add as label

□ remove unlabeled

k 6

max runs 10

□ determine good start values

K-Means (RapidMiner Core)

Synopsis

This operator performs clustering using the *k-means* algorithm. Clustering is concerned with grouping objects together that are similar to each other and dissimilar to the objects belonging to other clusters. Clustering is a technique for

Figure4.8: K-Means Flow Diagram

In figure4.6 the algorithm is run with k set to 6. K is a primary variable which tells the algorithm to create the specified clusters.In order to ensure better consistency and high quality output results the max run is set to 10 and optimization steps are set to 100. As a result, there are two output connections produced by k-means clustering operator, the first one is the clustering model (K-Means algorithm learning) produced by Rapid Miner and the second is an example set in which cluster number is allocated to each row of data set.

The choice of k value in k-means algorithm is vital in solving clustering problem effectively because the incorrect choice often leads to poor results. Several techniques are there to estimate k value that are: rule of thumb, elbow method, information theoretic method and use Silhouette to estimate k [15].

We selected ELBOW techniqueusing ANOVA [16] to determine the optimal k by plotting no of clusters vs. the percentage of variance explained by the clusters. The number of clusters is chosen at the point where the marginal gain drops. K-mean clustering is performed for each value of k and Percent of variance is calculated. The formula for calculating Percent of variance is as under:

$$percent\ of\ variance = \frac{\sum_{n=1}^{k}\sum_{i=1}^{Nn}(X_{in}-\bar{X})^2 - \sum_{n=1}^{k}\sum_{i=1}^{Nn}(X_{in}-\overline{X_n})^2}{\sum_{n=1}^{k}\sum_{i=1}^{Nn}\left(X_{ij}-\bar{X}\right)^2} \qquad (4)$$

K represents total number of clusters and $\mathrm{N_n}$ is total number of items in the nth cluster.

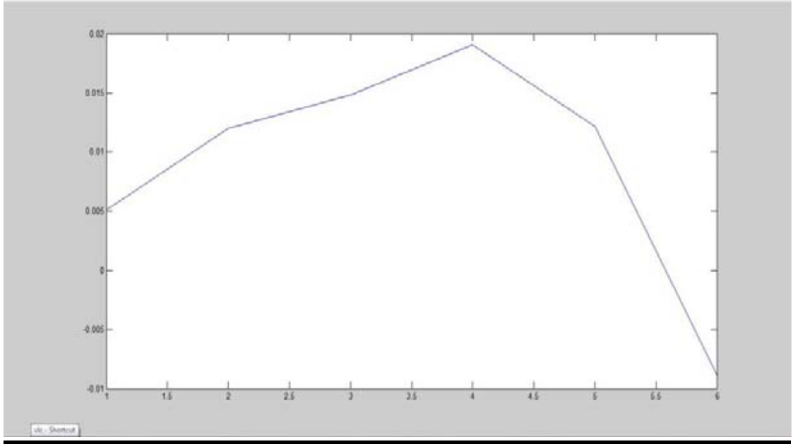


Figure4.9: No. of clusters vs. percentage of variance

The figure4.9 shows the value of k is 6.Moreover, rather relying on ‘just a number’ obtained by elbow technique, the results of k-means are analyzed with different values of k and with k=6 the resulting clusters are found more meaningful with negligible overlapping as shown in figures below:

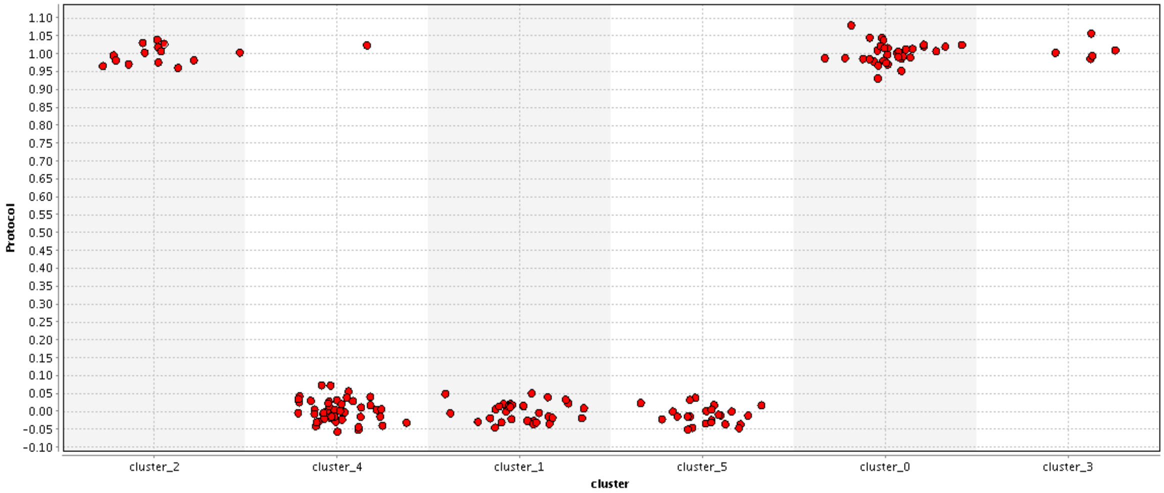


Figure4.10: Cluster vs. Protocol (Adenosine and Bruce)

The x-axis in figure above shows cluster_id and y-axis shows patient protocol (Adenosine and Bruce). It's clear from the figure that Cluster 0, 2 and 3 have patients which are treated with Bruce protocol and Cluster 1, 4 and 5 have all Adenosine induced patients. The figure also shows that Adenosine induced treated patients are more in number than bruce treated patients.

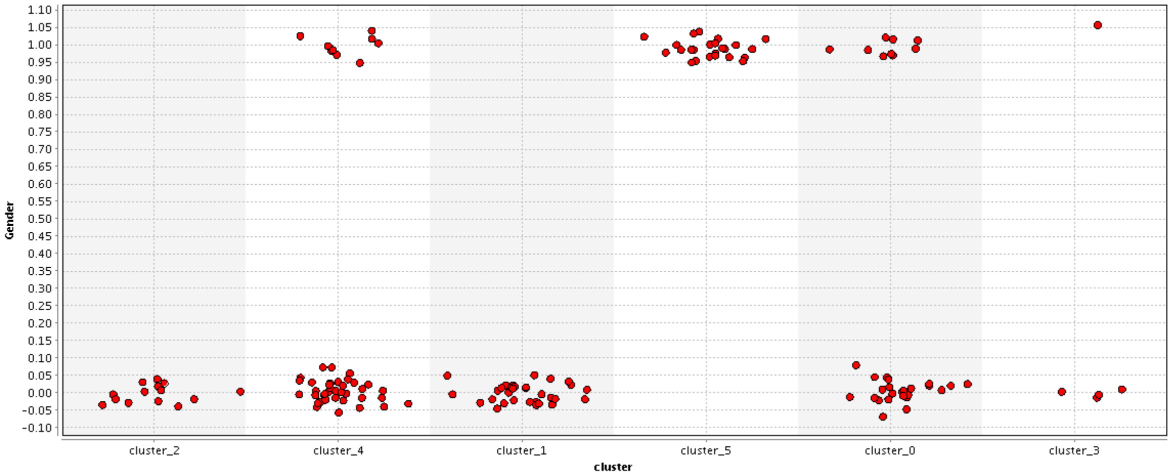


Figure 4.11: Cluster vs. Gender

The x-axis in figure above shows cluster_id and y-axis shows Gender attribute.The figure indicates that Cluster 1, 2 and 3 have only male patients, cluster 5 have only female patients and cluster 0 and 4 have both male and female patients.

Euclidean distance is used to calculate thedistance from one data point to a mean/centroid. The Euclidean distance is the most popular similarity measure and the smaller value of this shows greater similarity [32]. It is basically the length of the line segment connecting two point’s p and q that is(p, q). The formula of this is given below:

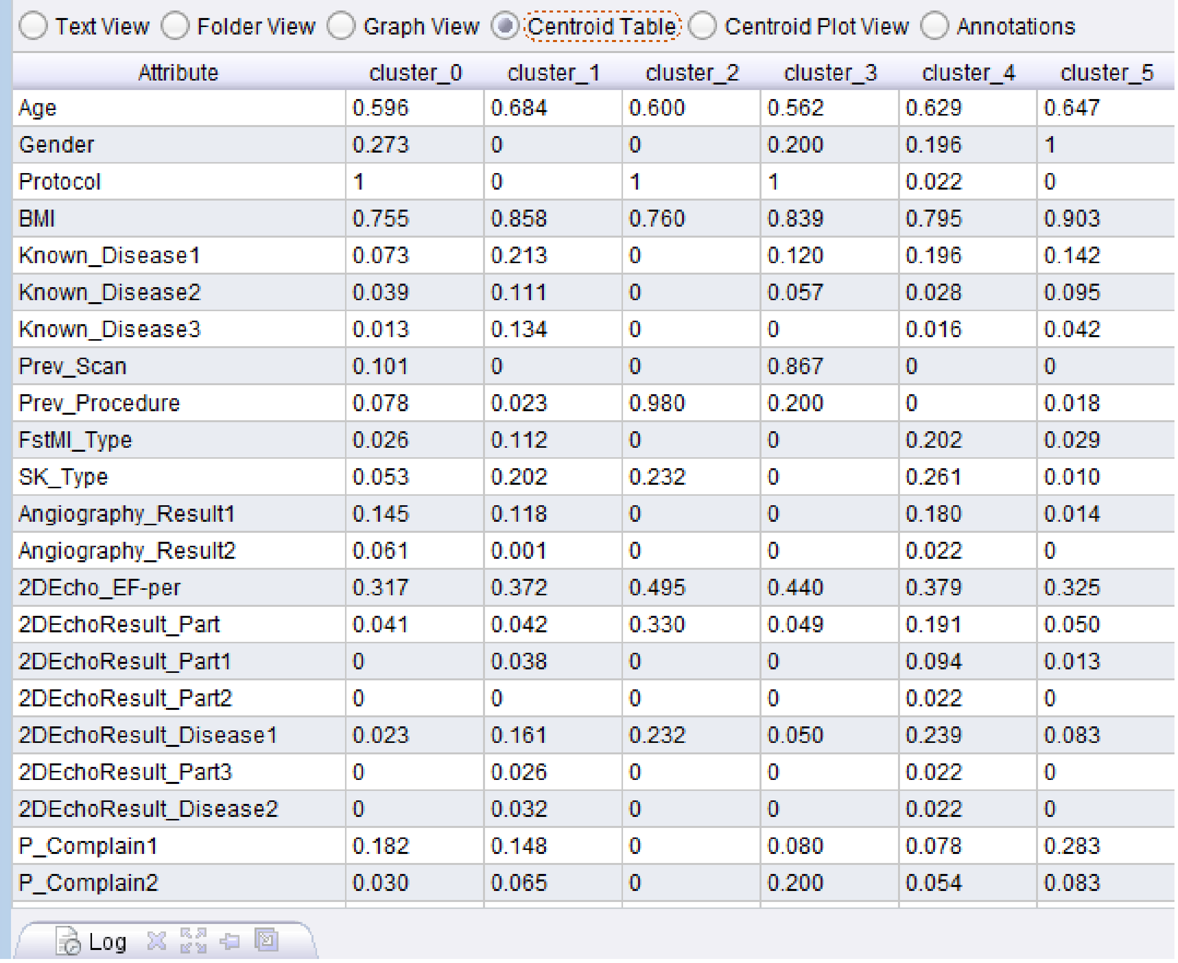
$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (5)$$

Where d(p,q) and d(q,p) is the distance between p and q or q and p respectively.Further, the mean in the Euclidean space of each cluster is computed with the formula below:

$$m_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \qquad (6)$$

Where $|C_j|$ is the number of data points in clusterC_j.

The Centroid table generated by cluster model showing the summary of centroid/mean of all attributes in 6 clusters is given below:



○ Text View ○ Folder View ○ Graph View ● Centroid Table ○ Centroid Plot View ○ Annotations

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
Age	0.596	0.684	0.600	0.562	0.629	0.647
Gender	0.273	0	0	0.200	0.196	1
Protocol	1	0	1	1	0.022	0
BMI	0.755	0.858	0.760	0.839	0.795	0.903
Known_Disease1	0.073	0.213	0	0.120	0.196	0.142
Known_Disease2	0.039	0.111	0	0.057	0.028	0.095
Known_Disease3	0.013	0.134	0	0	0.016	0.042
Prev_Scan	0.101	0	0	0.867	0	0
Prev_Procedure	0.078	0.023	0.980	0.200	0	0.018
FstMI_Type	0.026	0.112	0	0	0.202	0.029
SK_Type	0.053	0.202	0.232	0	0.261	0.010
Angiography_Result1	0.145	0.118	0	0	0.180	0.014
Angiography_Result2	0.061	0.001	0	0	0.022	0
2DEcho_EF-per	0.317	0.372	0.495	0.440	0.379	0.325
2DEchoResult_Part	0.041	0.042	0.330	0.049	0.191	0.050
2DEchoResult_Part1	0	0.038	0	0	0.094	0.013
2DEchoResult_Part2	0	0	0	0	0.022	0
2DEchoResult_Disease1	0.023	0.161	0.232	0.050	0.239	0.083
2DEchoResult_Part3	0	0.026	0	0	0.022	0
2DEchoResult_Disease2	0	0.032	0	0	0.022	0
P_Complain1	0.182	0.148	0	0.080	0.078	0.283
P_Complain2	0.030	0.065	0	0.200	0.054	0.083

Figure4.12: Centroid Table showing mean/centroid of each cluster

The above figure displays the summary of each cluster. It gives the centroid/mean of each cluster and attribute which helps to define the criteria of the clustering algorithm to find the given cluster. This can also help to predict the cluster of an unknown data element besides build a basic understanding of your own data set. Thefigure above shows that the centroids of almost all attributes are close to each other which mean that the data iswell evenly distributed.

Chapter 5

CORRELATION & CLUSTERING – RESULTS AND INTERPRETATIONS

5.1 CORRELATION –RESULTS AND CONCLUSIONS

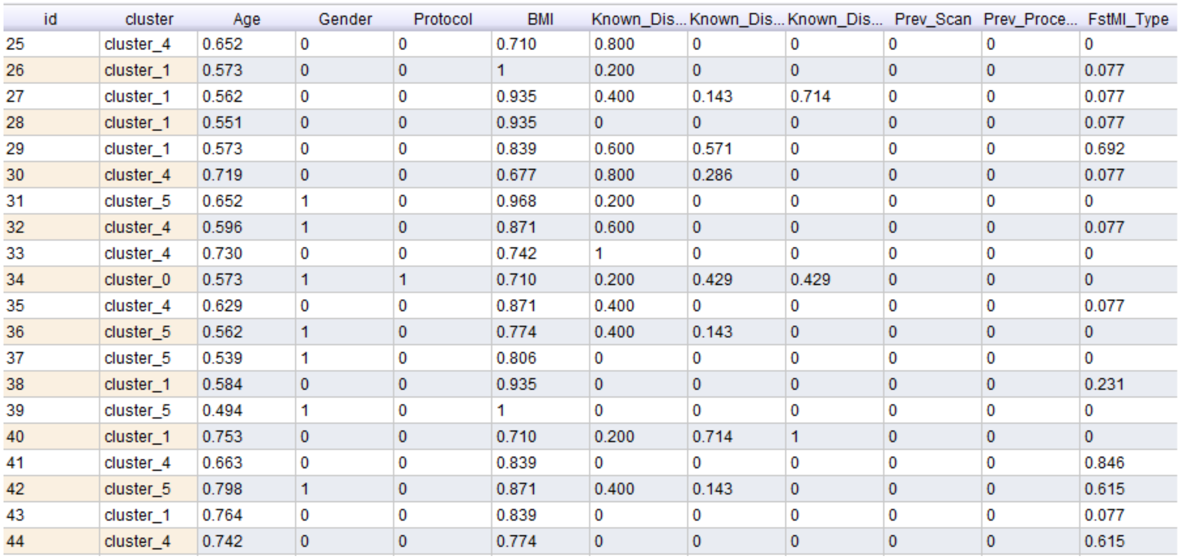
Some important facts revealed from results of ‘Model of Correlation matrix’ are:

- Age and Heart_Rate attributes are slightly negatively correlated with each other which means that with the increase in age the Heart_Rate is going to decrease i.e. tends toward abnormal range that is below 60.
- Age and Blood_Pressure attributes are positively correlated with each other which means that with the increase in age the Blood_Pressureis going to increase i.e. tends toward near alarming and extremely alarming stages.
- Age and Defect_Size attributes are positively correlated with each other which mean that greater age people are showing larger size Defect in some heart portion.
- Age and LVEF attributes are negatively correlated with each other which that means with the increase in age the LVEF is going to decrease i.e. i.e. tends toward near alarming and extremely alarming stages.
- The Gender and BMI attributes are slightly positively correlated with each other which indicate that female patients show somewhat greater BMI as compared to male patients.
- Gender and LVEF attributes are slightly positively correlated with each other which indicate that female patients tend towards increased LVEF i.e. normal stage.
- LVEF and IsDefected attributes are slightly negatively correlated with each other which means that with the increase in LVEF (tends towards normal) the IsDefected(0: not defected and 1: defected) is going to decrease i.e. tends toward not defected. This indicates that people with no defect in heart part has somewhat normal LVEF range.
- Heart_Rate before Infusion (BI) and Heart_Rate Maximum achieved (MA) are highly positively correlated with each other, same is the case with Baseline_Heart _Rate and PeakExercise_Heart_Rate.
- Blood_Pressure before Infusion (BP_BI) and Blood_Pressure Maximum achieved (BP_MA) are highly positively correlated with each other, same is the case with Resting_ Blood_Pressure and PeakExercise_Blood_Pressure.

- Defect_Size, Defect_Via/NonVia, Is_Defected attributes are all negatively correlated with Heart_Rate and Blood_Pressure related attributes which means that defected people tend towards abnormal Heart_Rate and Blood_pressure ranges.

5.2 CLUSTERING –RESULTS, VISUALIZATION AND CONCLUSIONS

After applying K-Means algorithm on the final dataset, Rapid Miner generates an example set (cluster_ID attached with each row of data set) and Cluster Model (report of algorithm learning)as shown in figure below.



id	cluster	Age	Gender	Protocol	BMI	Known_Dis...	Known_Dis...	Known_Dis...	Prev_Scan	Prev_Proce...	FstMI_Type
25	cluster_4	0.652	0	0	0.710	0.800	0	0	0	0	0
26	cluster_1	0.573	0	0	1	0.200	0	0	0	0	0.077
27	cluster_1	0.562	0	0	0.935	0.400	0.143	0.714	0	0	0.077
28	cluster_1	0.551	0	0	0.935	0	0	0	0	0	0.077
29	cluster_1	0.573	0	0	0.839	0.600	0.571	0	0	0	0.692
30	cluster_4	0.719	0	0	0.677	0.800	0.286	0	0	0	0.077
31	cluster_5	0.652	1	0	0.968	0.200	0	0	0	0	0
32	cluster_4	0.596	1	0	0.871	0.600	0	0	0	0	0.077
33	cluster_4	0.730	0	0	0.742	1	0	0	0	0	0
34	cluster_0	0.573	1	1	0.710	0.200	0.429	0.429	0	0	0
35	cluster_4	0.629	0	0	0.871	0.400	0	0	0	0	0.077
36	cluster_5	0.562	1	0	0.774	0.400	0.143	0	0	0	0
37	cluster_5	0.539	1	0	0.806	0	0	0	0	0	0
38	cluster_1	0.584	0	0	0.935	0	0	0	0	0	0.231
39	cluster_5	0.494	1	0	1	0	0	0	0	0	0
40	cluster_1	0.753	0	0	0.710	0.200	0.714	1	0	0	0
41	cluster_4	0.663	0	0	0.839	0	0	0	0	0	0.846
42	cluster_5	0.798	1	0	0.871	0.400	0.143	0	0	0	0.615
43	cluster_1	0.764	0	0	0.839	0	0	0	0	0	0.077
44	cluster_4	0.742	0	0	0.774	0	0	0	0	0	0.615

Figure5.1: Dataset illustrating Row No. with cluster id

The figure 5.1 displays the raw data with the addition of two new columns: ‘id’ and ‘cluster’. The ID column is basically a replicated Row Number column. The key results are stored under cluster column which shows that every row has been classified into a cluster. It’s clear from figure 8 above that actually K Cluster ID’s are assigned to 300 different records where each value of K represents cluster number. Now every row of data set has been classified into a cluster. After this, analysis and visualization techniques can be applied using Rapid Miner.

Now next part is to visualize these clusters and making these results understandable by medical practitioner. There are many ways for visualizing these clusters but we took help of Microsoft excel to visualize hidden patterns.First of all the following cluster frequency distribution was observed after K-Means algorithm scoring over 300 records.

TABLE 5.1: FREQUENCY DISTRIBUTION OF CLUSTERS

Cluster #	PERCENTAGE_DISTRIBUTION
Cluster 0	**12%**
Cluster 1	**3%**
Cluster 2	**15%**
Cluster 3	**17%**
Cluster 4	**31%**
Cluster 5	**22%**
Complete Base	**100%**

It's very clear from the above table that cluster number 3, 4 and 5 are the top three clusters identified by K-Means Clustering.

After that a knowledge enriched report was build which was portraying percentage contribution of each attribute to every cluster. For example in figure5.2 below Cluster 1 and 2 and 3 areBruce protocol treated while Cluster 3 and 4 and 5 areAdenosine protocol treated patients.

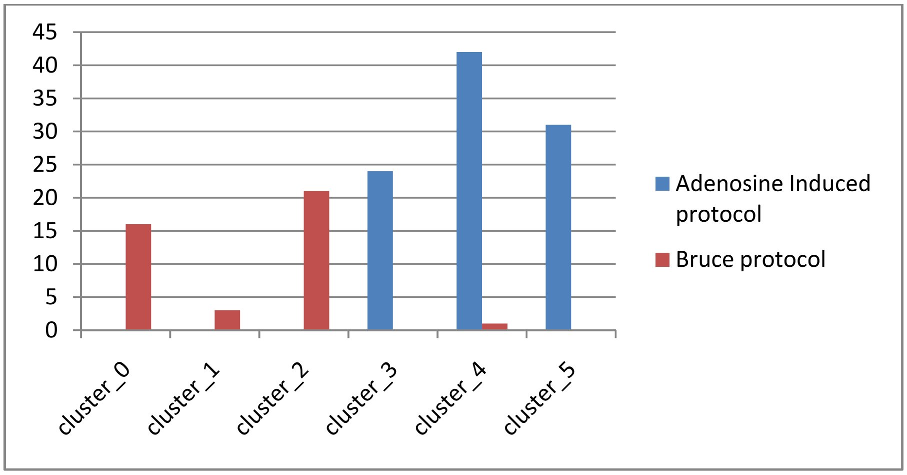


Figure5.2: Percentage of Adenosine vs.Bruce protocol patients

Among the 1500 patient records, the female percent is only 28% while male is of 72% which confirms the risk factor mentioned in [18] that male gender has more chances to get heart problem. Besides, the results in table also show that only 28% are mature patients while 72% are patients with more than 50+ age which shows that the chances of getting heart problem increases with the age [18]. For better understanding of results a sample data is copied below.

TABLE 5.2: PERCENTAGE OF MATURE VS OLD AGE PATIENTS

Cluster #	MATURE AGE	OLD AGE
Cluster 0	44%	56%
Cluster 1	0%	100%
Cluster 2	29%	71%
Cluster 3	17%	83%
Cluster 4	30%	70%
Cluster 5	13%	87%
Complete Base	28%	72%

After making above referenced percentage matrix for all features used in training dataset, Microsoft Charts facility was used to visually explain the clusters so that human mind can easily visualize categories and extract features out of these clusters. There are many type of charts available in Microsoft Office, we used Radar Charts to visualize the clusters. The methodology used to visualize clusters is simple. The percentage distribution of different features within 300 records and within a particular cluster is compared.

In below chart there are two types of behaviors captured one with maroon line and other one with blue line. The blue line describes the behavior of a particular cluster and maroon line represents the behavior of complete training set. It's very clear from visualizing the chart that there is some dimension which strongly deviates from training set percentage contributions. The radar chart below depicting trend in cluster 0 depicts that it contains 44% male and 56% female patients, 31% came with the chest pain complain. This cluster has nominal past heart attack history with no defected and ischemic patients. That's why 81% patients have normal LVEF value so this cluster can be named as Bruce protocol treated normal patients cluster.

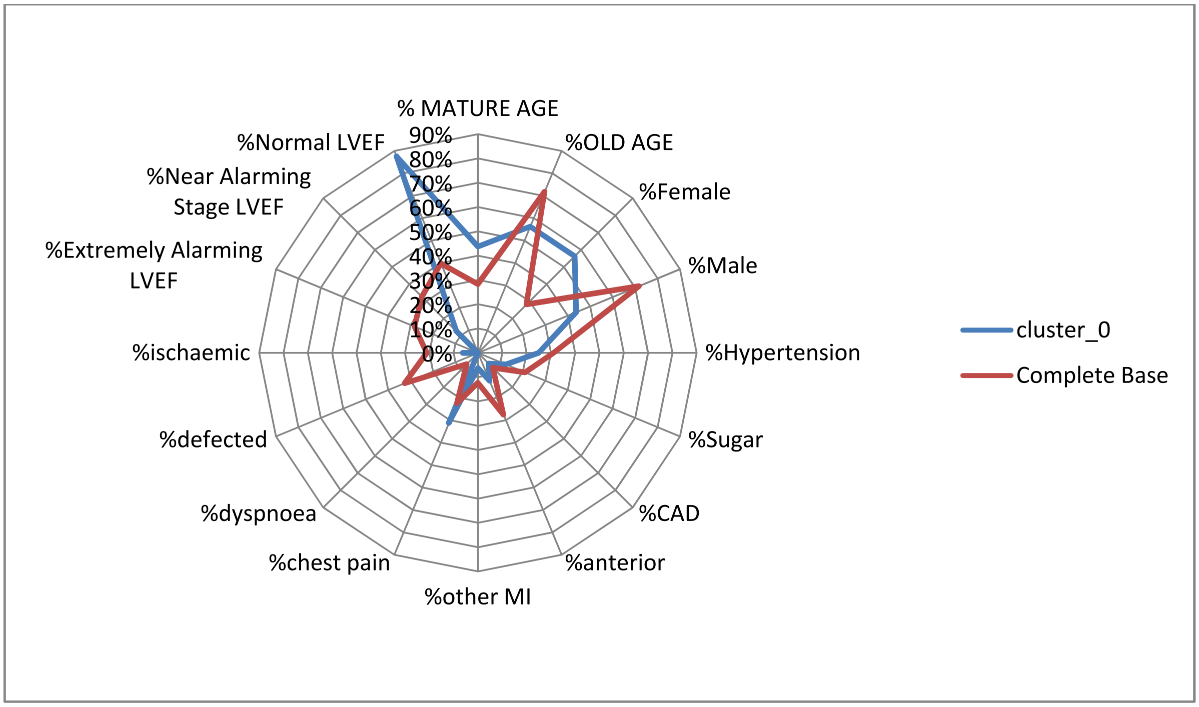


Figure 5.3: Radar Chart showing trend in Cluster 0

The radar chart below showing trend in cluster 1 shows that this is purely old age male patients cluster. 67% patients in this cluster have CAD while the total percentage of CAD is 9%. This cluster has 100% defected people that’s why extremely alarming and near alarming LVEF is dominating in this cluster.

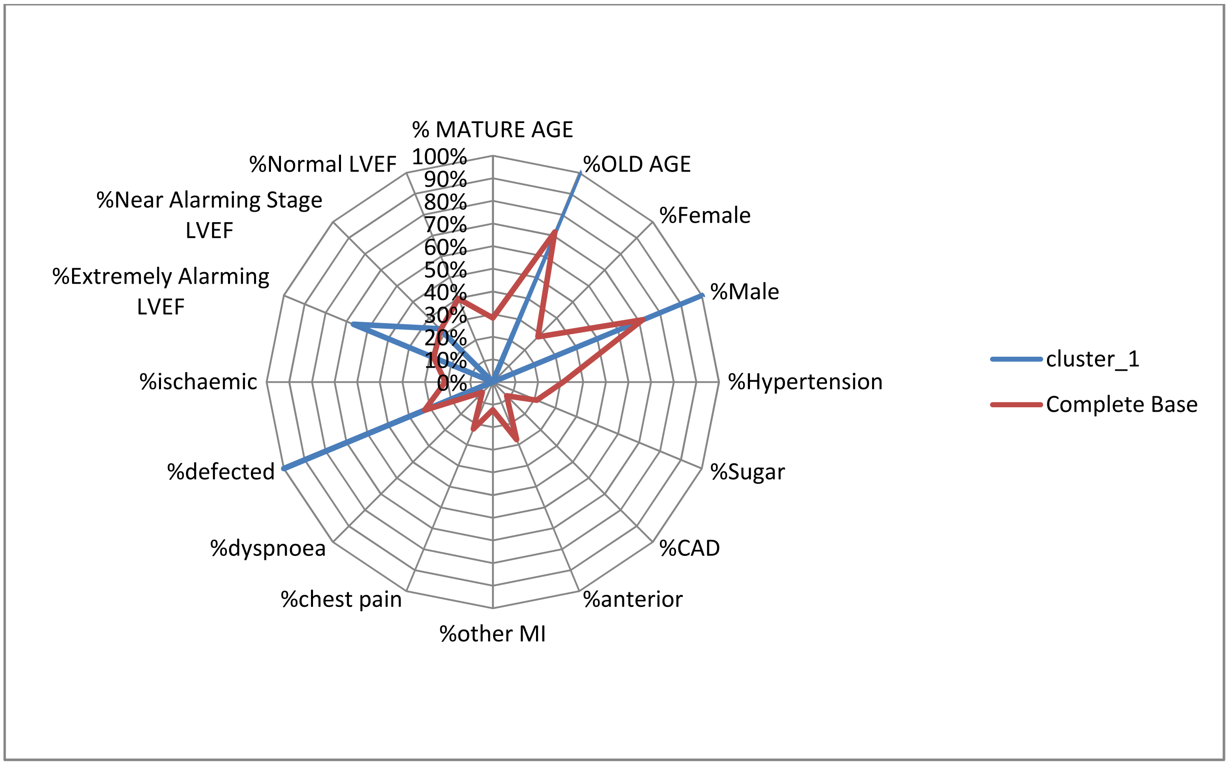


Figure5.4: Radar Chart showing trend in Cluster 1

The radar chart below showing trend in cluster 2 shows that this is purely bruce protocol treated male patients cluster. The old age hypertension patients are dominating here. Chest pain attribute is 33% here while the total percentage of this is 22%. Almost 50% patients are those with ischemia disease but normal and near alarming LVEF are dominating in this cluster.

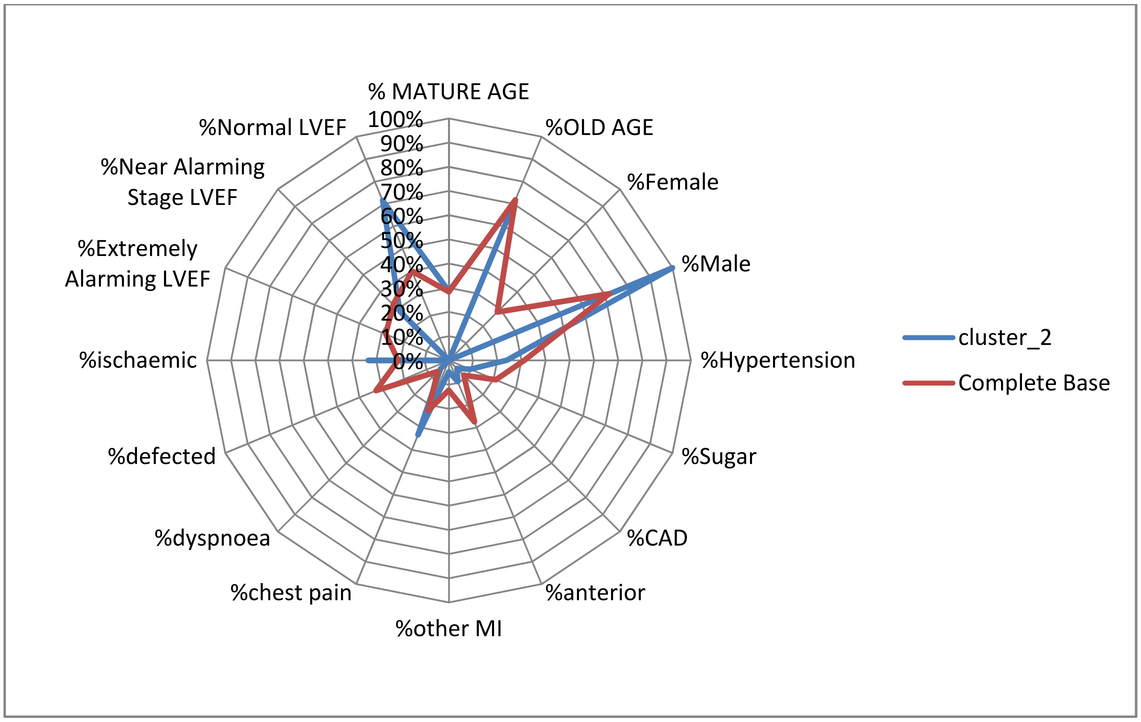


Figure5.5: Radar Chart showing trend in Cluster 2

The radar chart below presenting trend in cluster 3 shows that this is purely adenosine induced protocol treated female patients cluster of which 83% are of old age i.e. above 50 years. The hypertension and chest complain is dominating in this cluster but no ischemia, defected or past heart attack history patients. 75% are patients with normal LVEF and rest has near alarming LVEF. So after reviewing the discussed features we can easily depict that cluster3 only contains Female patients those who have reported about hypertension and chest pain and have Normal Heart Rate. And using these deviation's insights we can also name this cluster to further make the cluster easily understandable by medical practitioners.

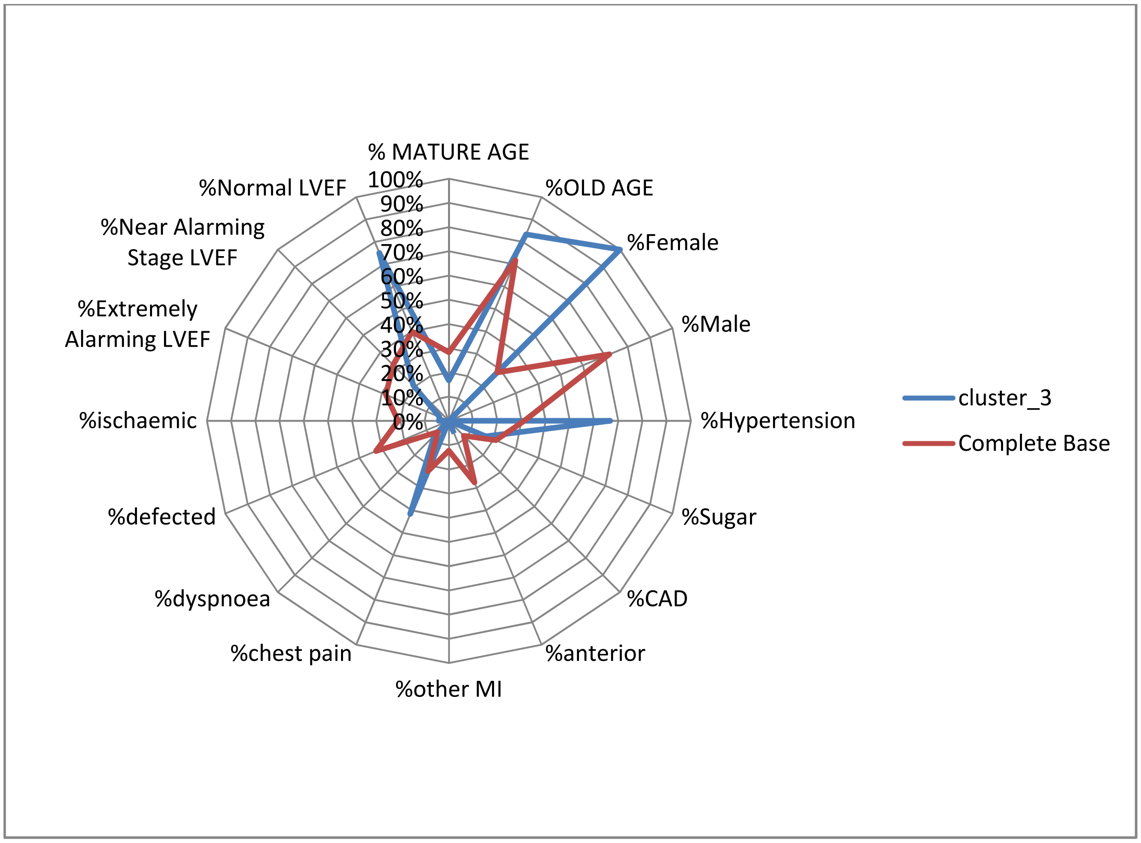


Figure5.6: Radar Chart showing trend in Cluster 3

Similarly, the radar chart next shows that cluster 4 has 86% male and 14% female patients. Also 70% of them are old age and 30% are of mature age.and 76%. The total percentage of anterior MI patients is 25% while in this cluster almost 50% patients have past anterior heart attack history and 26% have other types of heart attack history i.e. inferior, anterolateral, AWMI. The ‘Defected’ feature shows that total defected rate is 33% while this cluster has 98% patients who has either viable or non-viable defect in some portion of heart. That’s why the percentage of extremely alarming and near alarming LVEF is dominating here. This cluster can be called critical patients cluster.

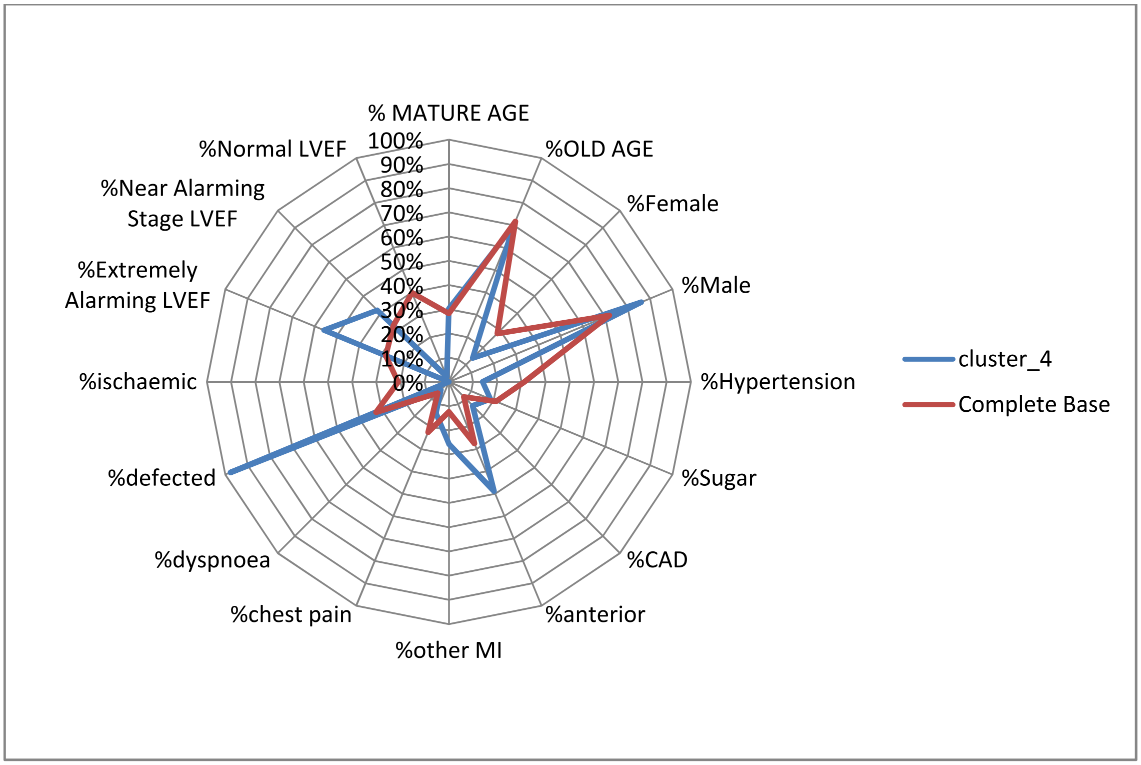


Figure5.7: Radar Chart showing trend in Cluster 4

Similarly, the radar chart next portray that cluster 5 is purely adenosine protocol treated male patients cluster. 87% of them are of old age i.e. above 50 years. The hypertension and diabetesfeature is dominating here with 45% hypertension and 35% diabetespatients. This cluster also has 45% anterior past heart attack feature leading with 65% ischemia disease affected people. This cluster has a mix of extremely alarming, near alarming and normal LVEF people.

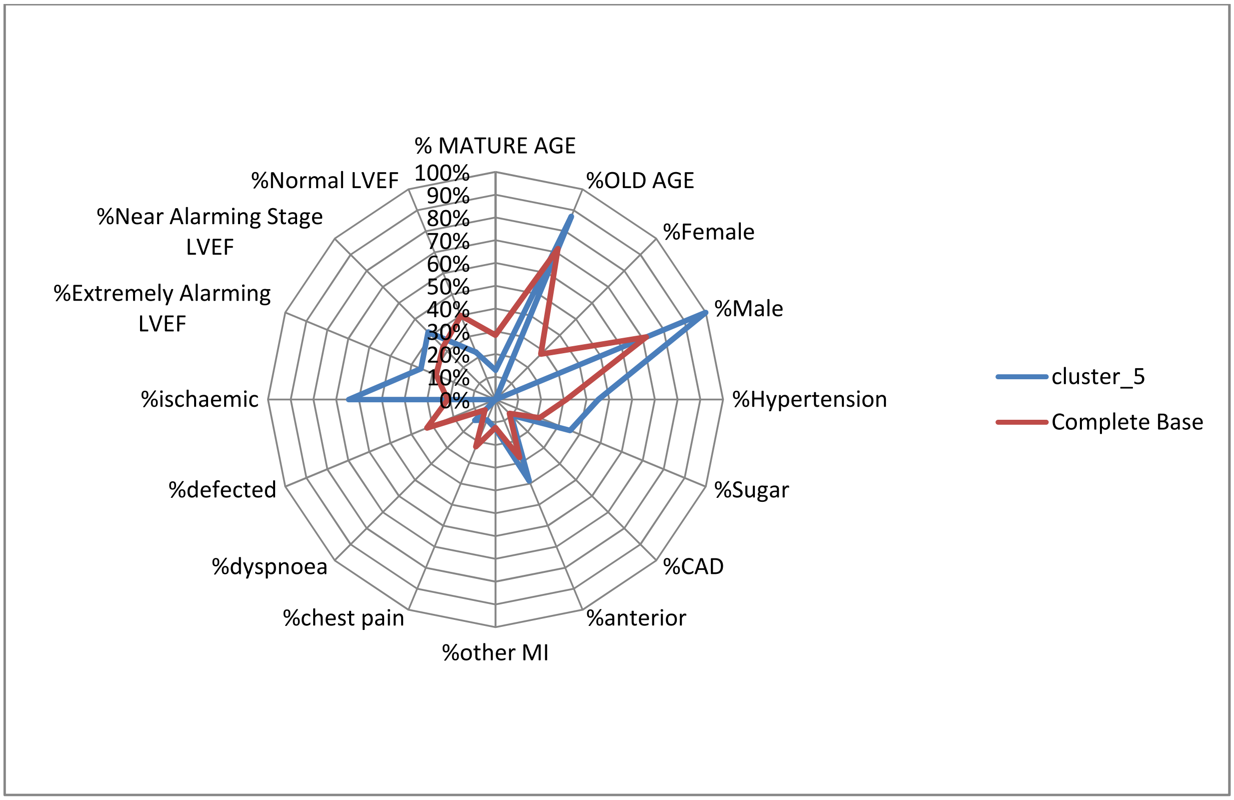


Figure5.8: Radar Chart showing trend in Cluster 5

5.3 CLUSTER LABELING

In the same way, above explained exercise to visualize clusters and to extract the dominated features was applied toall clusters in order to note the trend and give unique descriptive name to those clusters. The below mentioned table describes the summary of complete exercise.

TABLE 5.3: CLUSTER NAMING

Cluster #	CLUSTER LABEL
Cluster 0	Bruce Protocol treated Normal patients
Cluster 1	Bruce Protocol treated old aged Males, Defect in heart part. Low LVEF- Critical patients
Cluster 2	Bruce Protocol treated Males with ischemia disease
Cluster 3	Adenosine Protocol treated old aged Females, Normal Heart rate, Hypertension patients, Chest complain, Normal LVEF
Cluster 4	Adenosine Protocol treated, Past Heart attack history, Defect in heart part, Low LVEF - Critical patients

Cluster 5	Adenosine Protocol treated old aged Males with ischemia disease, Hypertension and diabetespatients, Low LVEF - Critical patients

5.4COMPARISON WITH CLUSTERING ALGORITHMS

After extraction of patterns, a comparison of k-means with other clustering algorithm is done on the basis of various internal evaluation indexes. These indexes/methods can help human mind to interpret best the clustering results.

5.4.1CLUSTERING ALGORITHMS

5.4.1.1 K-MEANS FAST

This algorithm is faster than the standard k-means algorithm especially on data sets with up to 1000 dimensions and larger value of k. When k value becomes greater than 20then this implementation of k-means becomes much faster than the standard k-means algorithm [34].

5.4.1.2 X-MEANS

This algorithm is an extension of k-means with the capability to cater k-means shortcomings of scaling poorly computationally and user overhead to provide no of clusters. This implementation is capable of searching the space of cluster locations efficiently andestimating the no. of clusters itself. The author proved by experiments that this algorithm will estimate the number of clusters correctly and is faster than repetitively applying k-means with different k values [34].

5.4.1.3 K-MEDOIDS

This algorithm is a classical partitioning technique to cluster the data set in to k clusters specified by user. This algorithm has properties similar to k-means but in this each centroid chosen is a data point rather than mean [28].

5.4.1.4 DBSCAN

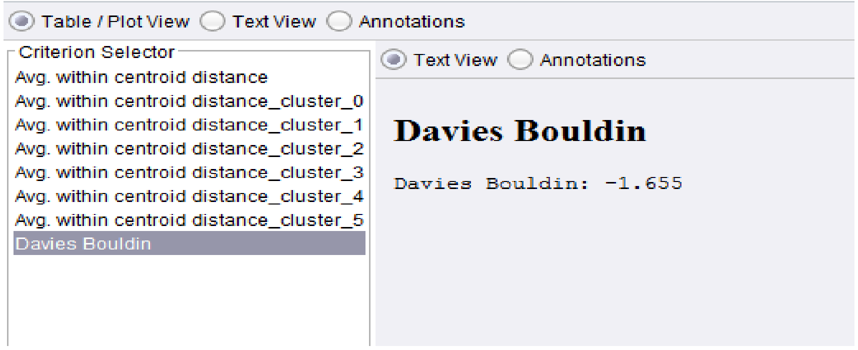
This algorithm is a density based clustering technique. It finds no of clusters starting from assessed density distribution of corresponding nodes. It does not require user knowledge to specify no of clusters [29].

5.4.2 INTERNAL EVALUATION INDEXES

The internal evaluation indexes on the basis of which comparison is done are:

5.4.2.1 DAVIES–BOULDIN INDEX

The smallest value of this index implies low inter-cluster similarity and high intra-cluster similarity. Hence, the algorithm that produces the lower value of this is considered the best on the basis of this criterion [30]. A snapshot of this index calculated in Rapid Miner is attached in figure below:


● Table / Plot View ○ Text View ○ Annotations
Criterion Selector
Avg. within centroid distance
Avg. within centroid distance_cluster_0
Avg. within centroid distance_cluster_1
Avg. within centroid distance_cluster_2
Avg. within centroid distance_cluster_3
Avg. within centroid distance_cluster_4
Avg. within centroid distance_cluster_5
Davies Bouldin
● Text View ○ Annotations

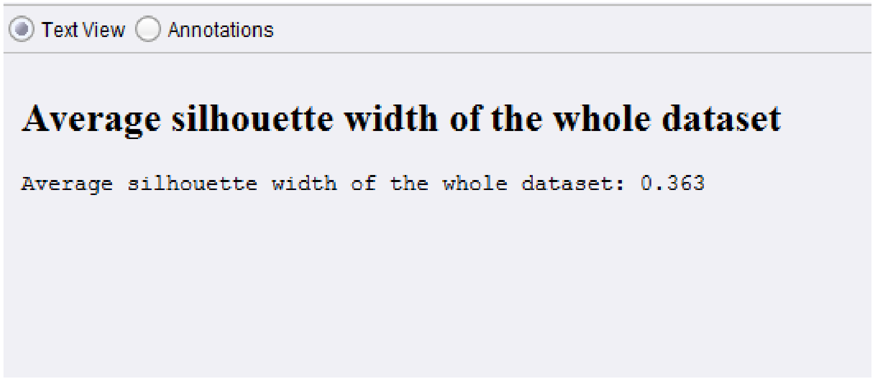
Davies Bouldin

```
Davies Bouldin: -1.655
```

Figure5.9: Davies Bouldin index

5.4.2.2 SILHOUETTE INDEX

Silhouette index value indicates the separation and firmness of the clusters. This value ranges between -1 and 1. A value near 1 show that data is consigned most suitable cluster [31]. A snapshot of this index calculated in Rapid Miner is attached in figure below:



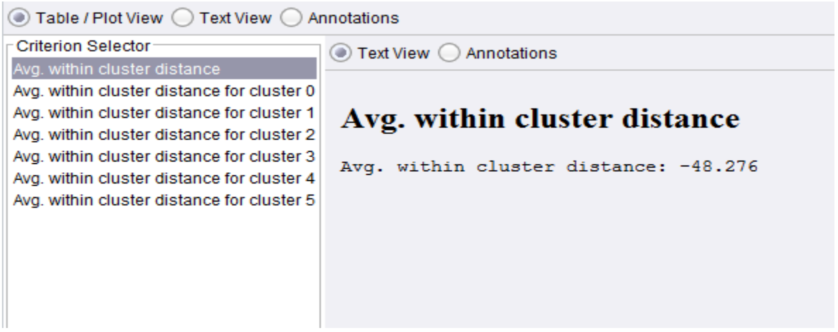
Average silhouette width of the whole dataset

```
Average silhouette width of the whole dataset: 0.363
```

Figure5.10: Average Silhouette index

5.4.2.3 CLUSTER DENSITY OPERATOR

This performance operator is used for evaluation of centroid based clustering algorithms. The smallest value of this operator indicates better clustering solution [30]. A snapshot of this index calculated in Rapid Miner is attached in figure below:


● Table / Plot View ○ Text View ○ Annotations
Criterion Selector
Avg. within cluster distance
Avg. within cluster distance for cluster 0
Avg. within cluster distance for cluster 1
Avg. within cluster distance for cluster 2
Avg. within cluster distance for cluster 3
Avg. within cluster distance for cluster 4
Avg. within cluster distance for cluster 5
● Text View ○ Annotations

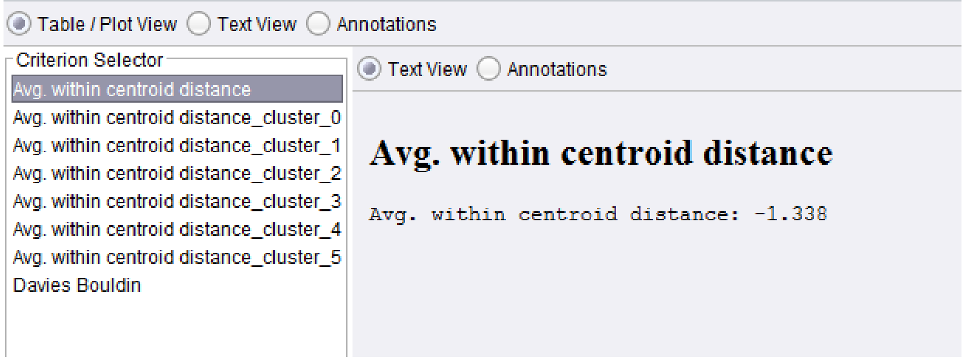
Avg. within cluster distance

```
Avg. within cluster distance: -48.276
```

Figure5.11: Cluster Density Operator

5.4.2.4 AVG. WITHIN CENTROID DISTANCE

The performance is delivered on the basis of cluster centroids by calculating avg. between the centroids and all cluster examples. The smallest value of this shows better clustering results. A snapshot of this index calculated in Rapid Miner is attached in figure below:


● Table / Plot View ○ Text View ○ Annotations
Criterion Selector
Avg. within centroid distance
Avg. within centroid distance_cluster_0
Avg. within centroid distance_cluster_1
Avg. within centroid distance_cluster_2
Avg. within centroid distance_cluster_3
Avg. within centroid distance_cluster_4
Avg. within centroid distance_cluster_5
Davies Bouldin
● Text View ○ Annotations

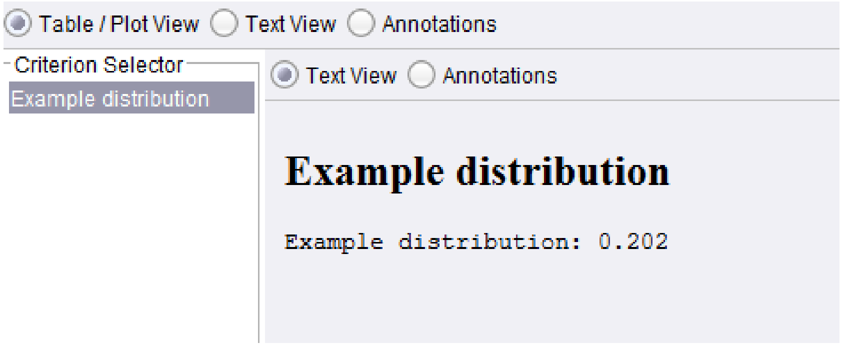
Avg. within centroid distance

```
Avg. within centroid distance: -1.338
```

Figure5.12: Avg. within centroid distance

5.4.2.5 SUM OF SQUARES ITEM DISTRIBUTION

This operator evaluates clustering results based on distribution of examples. When the algorithm gives equal distribution of examples in resulting clusters this value lean towards 1/no. of clusters [31]. A snapshot of this index calculated in Rapid Miner is attached in figure below:


◉ Table / Plot View ○ Text View ○ Annotations
Criterion Selector
Example distribution
◉ Text View ○ Annotations

Example distribution

```
Example distribution: 0.202
```

Figure5.13: Sum of squares item distribution

Table 5.4: Clustering algorithms performance evaluation

Evaluation criteria	k-Means	k-Means(fast)	k-Medoids	X-Means	DBSCAN
Davies–Bouldin	-1.760	-1.760	-1.488	-1.760	N/A
Silhouette index	0.415	0.415	0.420	0.415	0.482
Cluster Density	48.841	48.841	54.370	48.841	169.598
Sum of squares	0.200	0.200	0.221	0.200	0.527
Avg. centroid distance	1.360	1.360	2.075	1.360	N/A

The result of above table 5.4 shows that k-means, k-means (fast) and X-means are similar in performance on our dataset with respect to all above described cluster validity indexes. Also these are better in performance as compared to k-medoids and DBSCAN with respect to all validity indexes except Silhouette index. DBSCAN performs very poor in terms of cluster density and sum of squares item distribution while k-medoids is better than DBSCAN with respect to these two indexes.

5.5 ASSOCIATION MINING –MODELING AND RESULTS

The purpose of the study is to discover the relationship between different attributes to aid in decision making. Now next step is to identify any worthwhile rules that might be helpful in establishing inferences about other attributes.

5.5.1 Data Preparation

In order to dig out any potential rules, we have to rehash our basic data suite for rule-hunting. There are potentially infinite ways we can rehash primary data to explore for hidden rules but considering simplicity and scalability; we will rehash only selected attributes erstwhile.

As our dataset consists of continuousnumerical and categorical attributes so in order to apply association rule modeling the data must be converted to binomial form.For this purpose handling of continuous numeric attributes is done through discretize by binningand converting them into categorical as discussed in heading 4.4. After that categorical attributes are transformed into binomial so that existing rule mining algorithms can be applied to the data set.Categorical attributes are converted into binomial by picking as manynew “items” as the number of distinct attribute-value pairs as discussed in [38] andwith the help of tables 3.1and 4.1.

5.5.1.1 Rehashing Age: Age is a continuous numeric attribute with minimum value of 35 and maximum value of 81 and no missing values. We de-normalized and broke this attribute into two attributes naming: OLD AGE and MATURE AGE.Range qualifying MATURE AGE is 50 and below and OLD AGE is 50 above.

5.5.1.2 Rehashing Gender: Gender is symmetric binary attribute which is handled by replacing itwith two asymmetric binary variables namely Male and Female.

5.5.1.3 Rehashing Protocol : Protocol isa categorical variable with two distinct categories. So we de-normalized and broke this attribute into two attributes naming: Bruce Protocol Treated and Adenosine Protocol Treated.

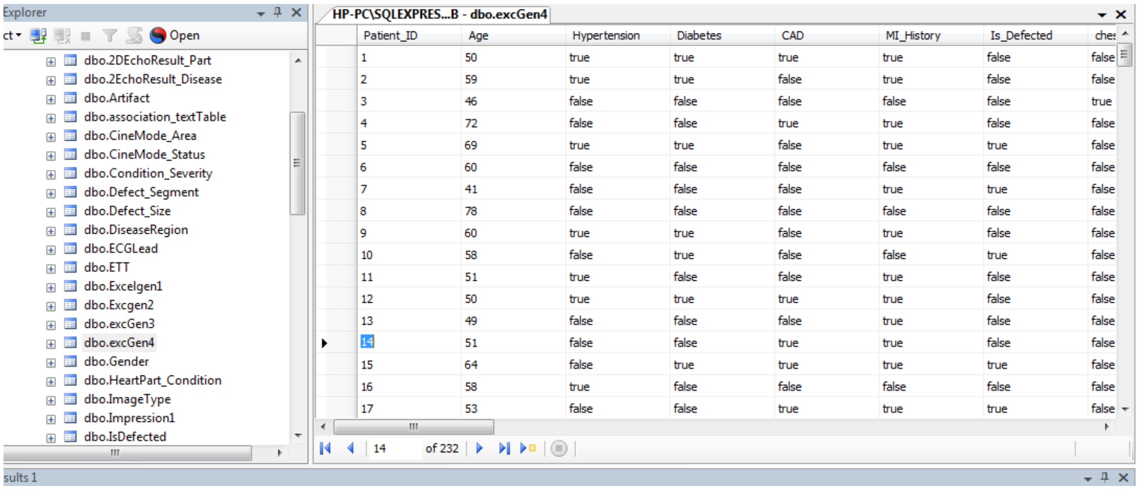
5.5.1.4 Rehashing Known_Disease : Known disease is a categorical variable with three distinct values. It is de-normalized and broken into three attributes naming: Hypertension, Diabetes and CAD.

5.5.1.5 Rehashing Patient Condition: Patient Condition is a categorical variable with three distinct values. It is de-normalized and broken into three attributes naming: is_Ischemic and is_Defected.

5.5.1.6 Rehashing LVEF: LVEF is a continuous numeric attribute with minimum value of 18 and maximum value of 71 and no missing values. We de-normalized and broke this attribute into three attributes naming:LVEF_Normal, LVEF_NearAlarmingand LVEF_ExtremeAlarming.

5.5.2 ASSOCIATION RULES- MODELING & RESULTS

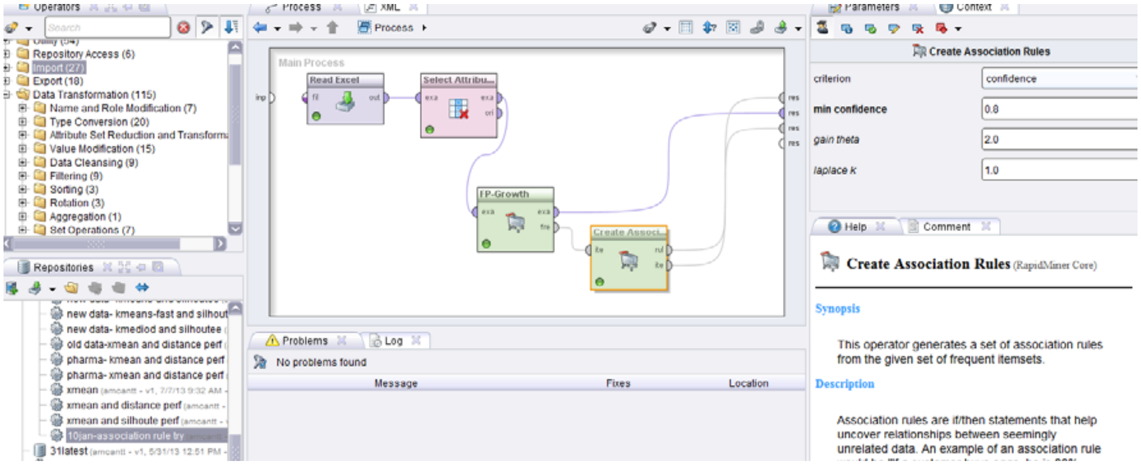
De-normalizing in fashion described yields 20 regular attributes. All attributes are prepared using MS SQL Server 2008 and then imported in MS Excel. Figure below identifies the source for data preparation.



Patient_ID	Age	Hypertension	Diabetes	CAD	MI_History	Is_Defected	ches
1	50	true	true	true	true	false	false
2	59	true	true	false	true	false	false
3	46	false	false	false	false	false	true
4	72	false	false	true	true	false	false
5	69	true	true	false	true	true	false
6	60	false	false	false	false	false	false
7	41	false	false	false	true	true	false
8	78	false	false	false	false	false	false
9	60	true	true	false	true	false	false
10	58	false	true	false	false	true	false
11	51	true	false	false	true	false	false
12	50	true	true	true	true	false	false
13	49	false	false	false	true	false	false
14	51	false	false	true	true	false	false
15	64	false	true	true	true	true	false
16	58	true	false	false	false	false	false
17	53	false	false	true	true	true	false

Figure 5.14: Data preparation in MS SQL Server 2008 for association rule generation

Further, the data is imported and retrieved in Rapid Miner for further analysis. Frequency Pattern Growth (FP-Growth) operator is now used on data stream to identify frequencies with minimum support kept to 0.3 to involve infrequent items also.Frequency pattern analysis is an essential element in mining of association rule. On frequency port data stream we apply '*Create Association Rule*' operator to identify association rules. The item port is also attached to results port in order to review frequent item list. See Figure 5.6 below. For extracted results minimum confidence is set to 0.8.


Operators
Search
Repository Access (6)
Import (27)
Export (18)
Data Transformation (115)
Name and Role Modification (7)
Type Conversion (20)
Attribute Set Reduction and Transforma
Value Modification (15)
Data Cleansing (9)
Filtering (9)
Sorting (3)
Rotation (3)
Aggregation (1)
Set Operations (7)
Repositories
new data- kmeans-fast and silhout
new data- kmediod and silhoutee
old data-xmean and distance perf
pharma- kmean and distance perf
pharma- xmean and distance perf
xmean (amcantt - v1, 7/7/13 9:32 AM -
xmean and distance perf (amcantt -
xmean and silhoute perf (amcantt -
10jan-association rule by (amcantt
31latest (amcantt - v1, 5/31/13 12:51 PM -
Process
XML
Process
Main Process
Read Excel
Select Attribu...
FP-Growth
Create Associ...
inp
fil
out
exa
ori
fre
rul
ite
res
Problems
Log
Message
Fixes
Location
Parameters
Context
Help
Comment

No problems found

Create Association Rules

criterion: confidence
min confidence: 0.8
gain theta: 2.0
laplace k: 1.0

Create Association Rules (RapidMiner Core)

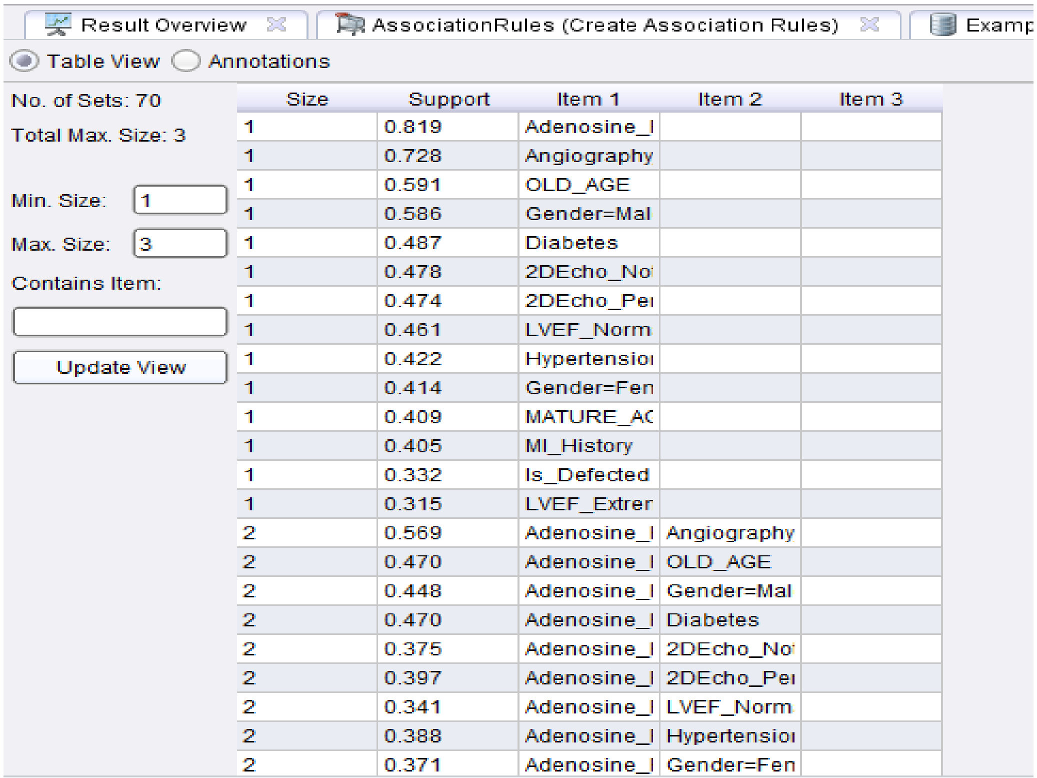
Synopsis

This operator generates a set of association rules from the given set of frequent itemsets.

Description

Association rules are if/then statements that help uncover relationships between seemingly unrelated data. An example of an association rule

Figure 5.15: Association Rule Flow Diagram


Result Overview
AssociationRules (Create Association Rules)
Examp
○ Table View ○ Annotations
Update View

No. of Sets: 70

Total Max. Size: 3

Min. Size: 1

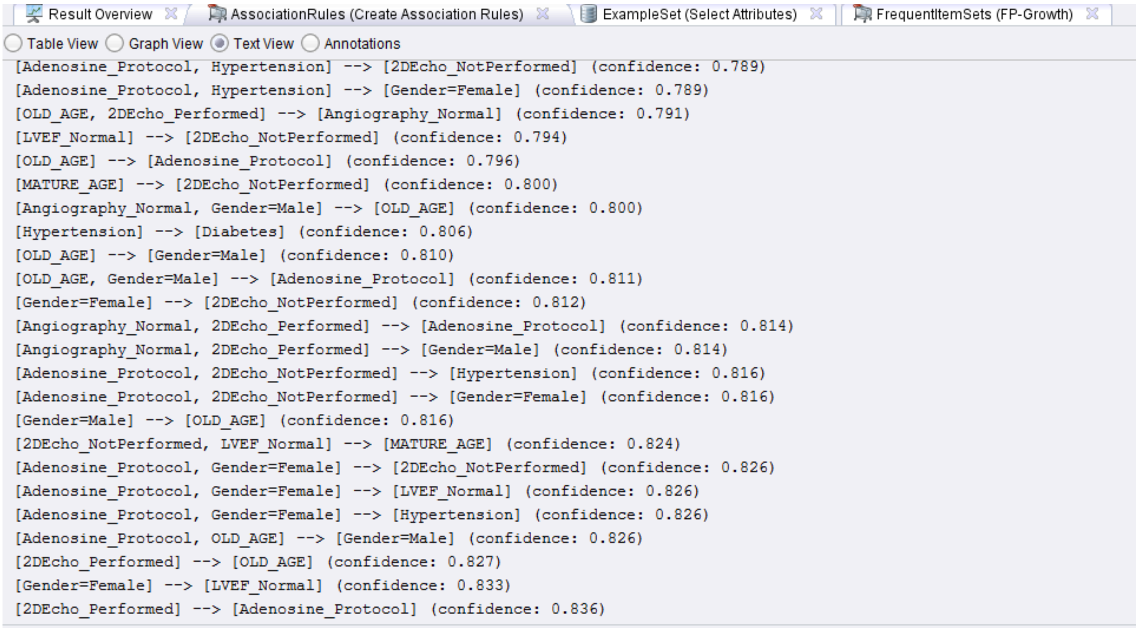
Max. Size: 3

Contains Item:

Size	Support	Item 1	Item 2	Item 3
1	0.819	Adenosine_I		
1	0.728	Angiography		
1	0.591	OLD_AGE		
1	0.586	Gender=Mal		
1	0.487	Diabetes		
1	0.478	2DEcho_No		
1	0.474	2DEcho_Pe		
1	0.461	LVEF_Norm		
1	0.422	Hypertensio		
1	0.414	Gender=Fen		
1	0.409	MATURE_AC		
1	0.405	MI_History		
1	0.332	Is_Defected		
1	0.315	LVEF_Extrer		
2	0.569	Adenosine_I	Angiography	
2	0.470	Adenosine_I	OLD_AGE	
2	0.448	Adenosine_I	Gender=Mal	
2	0.470	Adenosine_I	Diabetes	
2	0.375	Adenosine_I	2DEcho_No	
2	0.397	Adenosine_I	2DEcho_Pe	
2	0.341	Adenosine_I	LVEF_Norm	
2	0.388	Adenosine_I	Hypertensio	
2	0.371	Adenosine_I	Gender=Fen	

Figure 5.16: FrequentItemSet (FP-Growth) output

Figure above shows FP-Growth output in which the No. of sets identified is 70. With a 240 records dataset total of 292 rules are generated with 20 attributes when min. confidence is set 0.8.Same process was repeated several times with variety of confidence and support. Figure 5.11 identifies association rules using same 20 attributes dataset but now the minimum confidence is set to 0.5. It is noted that by decreasing the confidence threshold to 0.5the rules extracted are 1406.


Result Overview
AssociationRules (Create Association Rules)
ExampleSet (Select Attributes)
FrequentItemSets (FP-Growth)
○ Table View ○ Graph View ● Text View ○ Annotations

```
[Adenosine_Protocol, Hypertension] --> [2DEcho_NotPerformed] (confidence: 0.789)
[Adenosine_Protocol, Hypertension] --> [Gender=Female] (confidence: 0.789)
[OLD_AGE, 2DEcho_Performed] --> [Angiography_Normal] (confidence: 0.791)
[LVEF_Normal] --> [2DEcho_NotPerformed] (confidence: 0.794)
[OLD_AGE] --> [Adenosine_Protocol] (confidence: 0.796)
[MATURE_AGE] --> [2DEcho_NotPerformed] (confidence: 0.800)
[Angiography_Normal, Gender=Male] --> [OLD_AGE] (confidence: 0.800)
[Hypertension] --> [Diabetes] (confidence: 0.806)
[OLD_AGE] --> [Gender=Male] (confidence: 0.810)
[OLD_AGE, Gender=Male] --> [Adenosine_Protocol] (confidence: 0.811)
[Gender=Female] --> [2DEcho_NotPerformed] (confidence: 0.812)
[Angiography_Normal, 2DEcho_Performed] --> [Adenosine_Protocol] (confidence: 0.814)
[Angiography_Normal, 2DEcho_Performed] --> [Gender=Male] (confidence: 0.814)
[Adenosine_Protocol, 2DEcho_NotPerformed] --> [Hypertension] (confidence: 0.816)
[Adenosine_Protocol, 2DEcho_NotPerformed] --> [Gender=Female] (confidence: 0.816)
[Gender=Male] --> [OLD_AGE] (confidence: 0.816)
[2DEcho_NotPerformed, LVEF_Normal] --> [MATURE_AGE] (confidence: 0.824)
[Adenosine_Protocol, Gender=Female] --> [2DEcho_NotPerformed] (confidence: 0.826)
[Adenosine_Protocol, Gender=Female] --> [LVEF_Normal] (confidence: 0.826)
[Adenosine_Protocol, Gender=Female] --> [Hypertension] (confidence: 0.826)
[Adenosine_Protocol, OLD_AGE] --> [Gender=Male] (confidence: 0.826)
[2DEcho_Performed] --> [OLD_AGE] (confidence: 0.827)
[Gender=Female] --> [LVEF_Normal] (confidence: 0.833)
[2DEcho_Performed] --> [Adenosine_Protocol] (confidence: 0.836)
```

Figure 5.17: Total no. of rules generated with minimum confidence set to 0.5.

5.5.3 ASSOCIATION RULES – EVALUATIONS AND INTERPRETATION

Tremendous amount of data excavation was done and large numbers of rule cum attribute combinations were generated. However, it further requires a mining process to gain intelligence out of these outputs. So, in order to further enhance rule evaluation quality RakeshAgarwal [37] and J.Malar Vizhi1 [38] work was considered. For rule selection factors like support, confidence, lift, completeness, interestingness and comprehensibility measures were selected as benchmarking tools. Definition and formulas of each of these is included as Appendix A.1. Besides, domain knowledge are also required along with objective factors employed by data mining algorithms to decide whether a rule is genuinely interesting in a specific domain or not.

5.5.3.1 EVALUATING THE QUALITY OF A RULE

As many rules are generated with different confidence and support values but these two factors are not enough to gauge the power of an association rule. So to further shortlist the rules other quality measures such as completeness, interestingness, lift and comprehensibility are taken into consideration.Now our next step is to evaluate rules on the basis of these quality measures and domain knowledge to select top rules.The table below displays the statistics values calculated for different rules using the definitions and formulas in Appendix 1. For quality measures formulascalculation we took help of SQL queries.

Table 5.5: Evaluation of Rules on basis of quality measures

#	Association Rule Antecedent→ Consequent	Accuracy	Support	Lift	Completeness	Interes-tingness	Comprehen-sibility
1	Gender=Male, MI_History →OLD_AGE	86%	32%	1.457	54%	24	0.5
2	Angiography_Normal, Gender=Male →OLD_AGE	80%	33%	1.355	82%	38	0.5
3	Adenosine_Protocol, Hypertension→Diabetes	84%	33%	1.734	67%	33	0.5
4	Adenosine_Protocol, Gender=Female →Hypertension	83%	31%	1.954	76%	33	0.5
5	Gender=Female→MATURE_AGE	73%	30%	1.781	73%	31	0.63
6	MATURE_AGE→ 2DEcho_NotPerformed, LVEF_Normal	74%	30%	2.001	82%	36	0.79
7	Adenosine_Protocol, LVEF_Normal→Gender=Female	90%	31%	2.172	73%	39	0.5
8	Gender=Female→ Adenosine_Protocol, Hypertension	83%	31%	1.906	88%	43	0.79
9	Hypertension→LVEF_Normal	84%	31%	1.615	83%	45	0.63
10	LVEF_Normal→Angiography_Normal	77%	35%	1.052	68%	27	0.63
11	MI_History→ Adenosine_Protocol, OLD_AGE	78%	32%	1.653	74%	34	0.79
12	Gender=Male→OLD_AGE	81%	47%	1.382	81%	31	0.63
13	2DEcho_NotPerformed, LVEF_Normal→Gender=Female	84%	31%	2.047	75%	37	0.5
14	Is_Defected→Gender=Male	93%	31%	1.595	65%	36	0.63

15	LVEF_Normal, MATURE_AGE→ 2DEcho_NotPerformed	95%	30%	2.004	72%	46	0.5
16	Angiography_Normal, Diabetes→ Adenosine_Protocol	96%	36%	1.179	44%	13	0.5
17	Diabetes, Hypertension→ Adenosine_Protocol	96%	32%	1.175	76%	42	0.5
18	Hypertension, Gender=Female →Adenosine_Protocol	94%	30%	1.156	64%	36	0.5
19	2DEcho_NotPerformed, Hypertension→Adenosine_Protocol	94%	30%	1.156	64%	36	0.5
20	2DEcho_NotPerformed, Gender=Female→Adenosine_Protocol	91%	30%	1.111	64%	35	0.5
21	Adenosine_Protocol, Diabetes→ Angiography_Normal	87%	40%	1.071	79%	39	0.5
22	Gender=Female→Hypertension	88%	36%	1.849	86%	45	0.63
23	Hypertension→Diabetes	80%	34%	1.655	69%	32	0.63
24	2DEcho_Performed, MI_History -→LVEF_ExtremeAlarming	71%	30%	2.243	62%	30	0.5
25	Adenosine_Protocol, Gender=Male, Is_Defected→LVEF_ExtremeAlarming	77%	31%	2.257	53%	22	0.43
26	Diabetes, Gender=Female, MATURE_AGE→Adenosine_Protocol	100%	26%	1.221	50%	30	0.43
27	Diabetes, MATURE_AGE→Hypertension	98%	26%	2.330	63%	36	0.5
28	OLD_AGE, Gender=Male, Diabetes, 2DEcho_Performed,LVEF_ExtremeAlarming→Is_Defected, CAD	82%	21%	6.369	93%	24	0.52
29	Diabetes, 2DEcho_NotPerformed, MATURE_AGE→ Hypertension, Gender=Female	100%	25%	3.093	78%	40	0.61
30	Diabetes, LVEF_Normal, Gender=Female→Hypertension	100%	25%	2.367	60%	35	0.43

In the same way, above mentioned measures are calculated for all the remainingrules as well. The above table shows that there are many rules having good accuracy, completeness and lift values. Based on the values of these quality measures only top 10% potentially reliable and interesting association rules having accuracy above 80%, lift above 1, completeness above 60%, support and interestingness above 30% are selected for further interpretation.The interpretation of some among the top 10% rules is:

- **Interpreting Rule 6**

Antecedent A LVEF_Normal, MATURE_AGE

Consequent C 2DEcho_NotPerformed

Statistics Accuracy=96%, Coverage=34%, Lift=2.004, Completeness=72%, Interestingness=41, Comprehensibility=0.5, A-Locations=83, C-Locations=111, AC overlap=80

Interpretation The mature age patients having age less than 50 and LVEF in normal range that is above 55 is likely to have 2D Echo test not performed with an accuracy of 96%.

- **Interpreting Rule 13**

Antecedent A 2DEcho_NotPerformed, Hypertension

Consequent C Adenosine_Protocol

Statistics Accuracy=94%, Coverage=30%, Lift=1.156, Completeness= 64%, Interestingness=36,Comprehensibility=0.5,
A-Locations=75, C-Locations= 110, AC overlap=71

Interpretation Those hypertension patients who have 2D Echo test not performed is likely to be treated with Adenosine Infusion protocol.

- **Interpreting Rule 52**

Antecedent A 2DEcho_Performed, MI_History

Consequent C LVEF_ExtremeAlarming

Statistics Accuracy=71%, Coverage=30%, Lift=2.257, Completeness=62%, Interestingness=30, Comprehensibility=0.5,
A-Locations=41, C-Locations=66, AC overlap=41

Interpretation The patients who don’t have 2D echo test performed and have past heart attack history (interior, inferior etc.) are likely to have LVEF in extreme alarming stage that is lower than 30.

- **Interpreting Rule 113**

Antecedent A Gender=Male, MI_History

Consequent C OLD_AGE

Statistics Accuracy=86%, Coverage=32%, Lift=1.457, Completeness=54, Interestingness=24, Comprehensibility=0.5, A-Locations=86, C-Locations=137, AC overlap=74

Interpretation Those male patients who have past heart attack history (anterior, inferior etc.) are likely to be greater than 50 years in age.

- **Interpreting Rule118**

Antecedent A Adenosine_Protocol, LVEF_Normal

Consequent C Gender=Female

Statistics Accuracy=89%, Coverage=30%, Lift=2.172, Completeness=73%, Interestingness=39, Comprehensibility=0.5, A-Locations=79, C-Locations=96, AC overlap=71

Interpretation Those patients who are treated with Adenosine protocol and their LVEF range is normal that is above 55 are likely to be female patients.

- **Interpreting Rule 98**

Antecedent A Hypertension

Consequent C Diabetes

Statistics Accuracy=80%, Coverage=34%, Lift=1.655, Completeness=69%, Interestingness=32, Comprehensibility=0.63, A-Locations=98, C-Locations=113, AC overlap=79

Interpretation The hypertension patients are also diabetic with the confidence of 80%.

- **Interpreting Rule 366**

Antecedent A Diabetes, LVEF_Normal, Gender=Female

Consequent C Hypertension

Statistics Accuracy=100%, Coverage=25%, Lift=2.367, Completeness=60%, Interestingness=35, Comprehensibility=0.43, A-Locations=59, C-Locations=98, AC overlap=59

Interpretation The patients whose LVEF range is normal that is above 55 their angiography result is also normal.

- **Interpreting Rule 344**

Antecedent A Diabetes, MATURE_AGE

Consequent C Hypertension

Statistics Accuracy=98%, Coverage=26%, Lift=2.330, Completeness=63%, Interestingness=36, Comprehensibility=0.5, A-Locations=63, C-Locations=98, AC overlap=62

Interpretation The mature age (i.e. age less than 50) diabetic patients are likely to have hypertension disease with an accuracy of 98%.

- **Interpreting Rule 189**

Antecedent A Gender=Male, Is_Defected

Consequent C MI_History

Statistics Accuracy=86%, Coverage=26%, Lift=2.125, Completeness=73%, Interestingness=36, Comprehensibility=0.5, A-Locations=72, C-Locations=94, AC overlap=62

Interpretation The male patients having defect in heart part is likely to have MI history in the past with an accuracy of 86%.

It is noted that with artificial datasets lots of rules can be generated with good confidence and support but with real life datasets usually limited numbers of rules are extracted with high or moderate levels of confidence, supportand completeness measures possibly due to lots of missing values in records of dataset.

5.6EXTRACTED PATTERNS

Based on the deep analysis of correlation, clustering and rule mining results, some significant patterns are extracted that can aids in heart condition prediction. These are stated below:

- *The greater no of male patients in dataset indicates that male gender is more prone to get heart attack and heart problem.*
- *The results show that young people who are generally under 35 years are not risky and do not have the chance to get a heart attack.*
- *The greater no of Hypertension and diabetes patients indicate that this disease has an important role in heart problem.*
- *Patients with some kind of heart problem have chest pain and dysphonia as common complain.*
- *Patients with past Heart attack (MI) history=> abnormal BP => defect in heart part=> low LVEF => Risk of CAD => Risk of MI*
- *Patients having previous MI history has more chance to get heart attack.*
- *The most common type of MI found is anterior MI.*
- *Patients with Hypertension and diabetes disease => ischemia disease=>low LVEF=> Risk of CAD => Risk of MI*

Chapter 6

CONCLUSION AND FUTURE WORK

Healthcare organizations have huge datasets that comes from assorted sources. That data is not always suitable in structure or quality. Medical analysis is a significant but difficult task that ought to be accomplished precisely and proficiently and its computerization would be very expedient and beneficial.

6.1 ACHIEVEMENT OF THIS RESEARCH

The achievement of this research is as follows:

- Heart patient dataset design which can be used in further research.
- Applied data mining techniques on thepre-processeddataset to extract useful features and patterns that aids in effective heart condition prediction.
- Reporting of the results is done in which trend in different clusters is noted which can give quick knowledge to doctors in predicting the future behavior of patients. For example, if old age male patient comes with hypertension and diabetes disease, then it is more likely that he may have ischemia disease and a low LVEF value. The chance of him to be critical is more.
- Association Rule mining are applied to determine association between attributes and inferences are made.
- Comparison of selected k-means technique with other clustering algorithms is done with respect to internal evaluation indexes..
- The proposed framework can be utilized for mining unstructured data in other health care centers.

6.2 BENEFITS OF THE RESEARCH

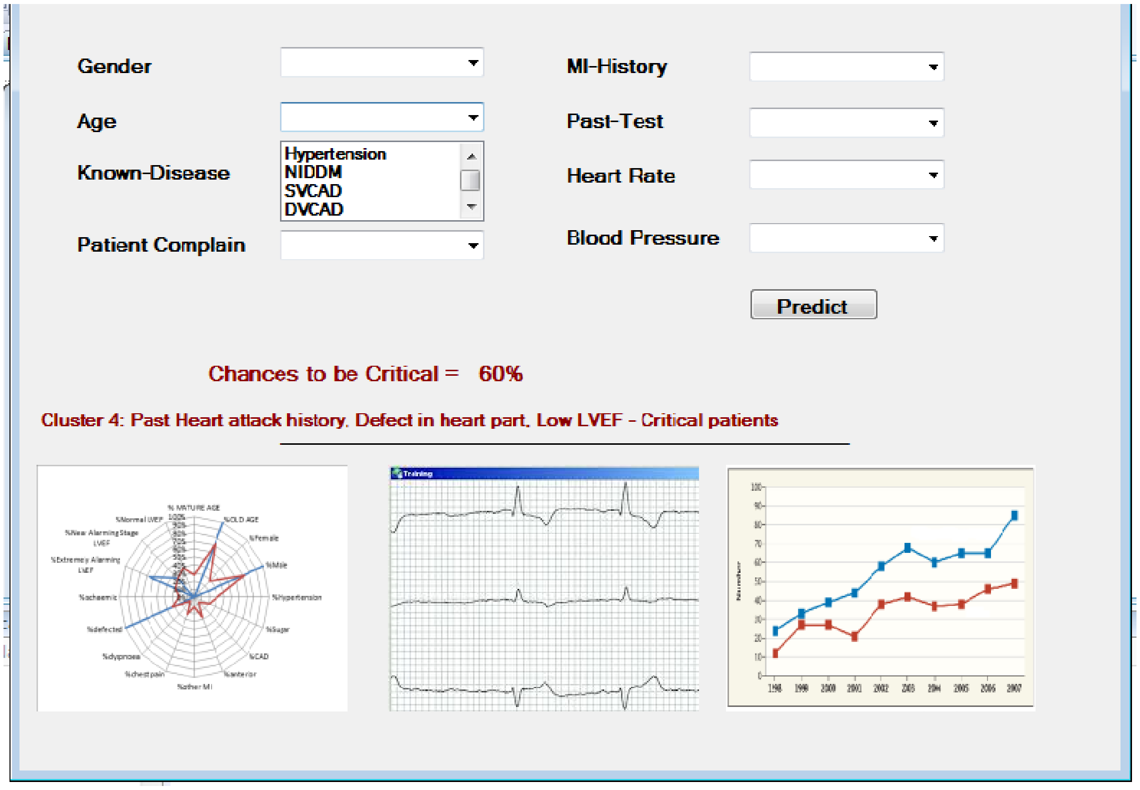
The benefits of this research are as follows:

- The outcomes of the study help medical practitioners and researchers to take precautionary measures to deal with the type of heart attack, disease and complain occurring more in patients.
- Besides, these could be very productive and expedient for doctors to diagnose patients in a better and efficient way and also helps them in the field of medicine research.

6.3 FUTURE WORK

The future work of the research can be:

- The results of the study can be further improved by investigating other clustering algorithms and by improving the data pre-processing techniques.
- Other data mining techniques i.e. Association rules and Time series can also be incorporated.
- Text mining technique can be applied to mine the huge unstructured data in hospitals.
- Integrate data mining and text mining.
- Using the same dataset to explore the reason, solution and precautionary measures of specific type of complain, disease and problem occurring in specific type of patients.
- To implement an expert system that would predict the probability of patient being critical state using regression and neural networks algorithms and these extracted patterns.



Gender

Age

Known-Disease: Hypertension, NIDDM, SVCAD, DVCAD

Patient Complain

MI-History

Past-Test

Heart Rate

Blood Pressure

Predict

Chances to be Critical = 60%

Cluster 4: Past Heart attack history. Defect in heart part. Low LVEF - Critical patients

Figure 6.1: Heart Condition Prediction system as future work

Appendix A.1

PERFORMANCE MEASURES - DEFINITIONS*

Confidence/ Accuracy

Confidence is a measure of strength of the association rules. Let a rule be of the form: IF A THEN C, where A is the antecedent and C is the consequent (predicted class). A very simple way to measure thepredictive accuracy of a rule is to compute the so-called confidence factor (CF) of the rule,defined as:

$$CF = |A \& C| / |A| = AC\text{-}Overlap / A\text{-}Locations * 100$$

Where |A| is the number of examples satisfying all the conditions in the antecedent A and |A & C|is the number of examples that both satisfy the antecedent A and have the class predicted by theconsequent C. For instance, if a rule covers 10 examples (i.e. |A| = 10), out of which 8 have theclass predicted by the rule (i.e. |A&C| = 8) then the CF of the rule is CF = 80%.

Support/ Coverage

Support(s) of an association rule is defined as the percentage/fraction of records that contain X∪Y to the total number of records in the database.

Support = |A & C| / N

Where N is the total number of transactions in the dataset and |A & C|is the number of examples that both satisfy the antecedent A and have the class predicted by theconsequent C.

Lift

The Lift factor, is a ratio of Confidence / Expected Confidence and measures theoverall rule strength.A lift value greater than 1 indicates there is a positive association, whereas a value less than 1 indicates there is a negative association.

Completeness

We can now measure the predictive accuracy of a rule by taking into account not only its CF butalso a measure of how “complete” the rule is, i.e. the proportion of examples is, having thepredicted class C that is actually covered by the rule antecedent. The rule completeness measureis computed by the formula:

Completeness = |A & C| / |C|

Where |C| is the number of examples satisfying all the conditions in the consequent C and |A & C|is the number of examples that both satisfy the antecedent A and have the class predicted by theconsequent C. For instance, if a rule covers 10 examples (i.e. |C| = 10), out of which 5 have theclass predicted by the rule (i.e. |A&C| = 5) then the Completeness of the rule is = 50%.

Interestingness

The ruleinterestingness measure is computed by the formula:

Interestingness= |A&C| - (|A|*|C|)/N= AC-Overlap - (A-Locations * C-Locations / N)

Where |A&C| is the number of examples that both satisfy the antecedent A and have the classpredicted by the consequent C minus the product of |A| is the number of examples satisfying allthe conditions in the antecedent A and |C| is the number of examples satisfying all the conditionsin the consequent and N is the total number of examples in the dataset.

Comprehensibility

In association rule mining if the number of conditions involved in the antecedent part is less, the rule is more comprehensible. We therefore require an expression where the number of attributes involved in both the parts of the rule has some effect. The following expression can be used to quantify the comprehensibility of an association rule

$$\text{Comprehensibility} = \log(1+ | C |) / \log(1+ | A \cup C |)$$

Here, | C | and | A ∪ C | are the number of attributes involved in the consequent part and the total rule, respectively.

REFERENCES

[1] AbuKhousa, E.; Campbell, P., "Predictive data mining to support clinical decisions: An overview of heart disease prediction systems," International Journal of Computer Applications, vol. 17, pp. 267-272, 2012.

[2] K. Aziz, S. Aziz, Evaluation and Comparison of Coronary Heart Disease Risk Factor Profiles of Children in a Country with Developing Economy

[3] Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart disease in low- and middle-income countries. Current problems in cardiology. Current Problems in Cardiology,Volume 35, Issue 2 , pp. 72-115, 2010

[4] Algorithm AS 136: A K-Means Clustering Algorithm, J. A. Hartigan and M. A. Wong Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1, pp. 100-108.

[5] K. Rajeswari, Dr. V. Vaithiyanathan, Dr.P. Amirtharaj, Prediction of Risk Score for Heart Disease in India Using Machine Intelligence, 2011 International Conference on Information and Network Technology, IPCSIT vol.4, pp. 18-22, 2011

[6] Cardiovascular-diseases, http://www.thenews.com.pk/Todays-News-6-134656-Cardiovascular-diseases-claim-200000-lives-annually-in-Pakistan

[7] Giudici, P.: “Applied Data Mining: Statistical Methods for Business and Industry”, New York: John Wiley, 2003.

[8] AbuKhousa, E.; Campbell, P., "Predictive data mining to support clinical decisions: An overview of heart disease prediction systems," Innovations in Information Technology (IIT), 2012 International Conference on , vol., no., pp.267,272, 2012.

[9] WalidMoudani, Dynamic Features Selection for Heart Disease Classification, World Academy of Science, Engineering and Technology Issue 0074 February 2013

[10] ShantakumarB.Patil, Y.S.Kumaraswamy, Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction, International Journal of Computer Science and Network Security ,Vol. 9 No. 2, pp. 228-235, 2009

[11] Ms. Ishtake S.H , Prof. Sanap S.A., Intelligent Heart Disease Prediction System Using Data Mining Techniques, International J. of Healthcare & Biomedical Research, Volume: 1,pp. 94-101, 2013.

[12] LathaParthiban and R.Subramanian, Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm, International Journal of Biological and Life Sciences 3:3,pp 157-160, 2007

[13] JyotiSoni, Uzma Ansari, Dipesh Sharma, Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers, International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 6, pp. 2385-2392,2011

[14] K. Rajeswari; V. Vaithiyanathan, Heart disease diagnosis : an efficient decision support system based on fuzzy logic and genetic algorithm, International journal of decision sciences, risk and management.- Genève : Inderscience, ISSN 1753-7169, ZDB-ID 25122769. - Vol. 3.2011, 1/2, pp. 81-97, 2011

[15] Determining_the_number_of_clusters, http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

[16] Lu, Jie; Goyal, Madhu; Kaur, Preetinder, Pricing Analysis in Online Auctions Using Clustering and Regression Tree Approach, Agents and Data Mining Interaction, pp. 248-257, 2012

[17]CRISP-DM, ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

[18] Unstructured Data Mining: The Tools You Need To Dig The Deep Web, Posted February 13, 2013 @ 3:41 pm by Scott Raspa, http://www.ikanow.com/blog/02/13/unstructured-data-mining-dig-the-deep-web

[19] Unstructured data a valuable resource for healthcare providers, Posted by Digital Reasoning in Industry News on April 11, 2013, http://www.digitalreasoning.com/2013/industry-news/unstructured-data-a-valuable-resource-for-healthcare-providers/

[20] Andrew Harbison&Pearse Ryan, The problem of analyzing unstructured data, The limits of computers in electronic discovery http://www.grantthornton.ie/db/Attachments/Publications/Forensic_&_inve/Grant%20Thornton%20-The%20problem%20of%20analysing%20unstructured%20data.pdf

[21] Anil Jain, KarthikNandakumar, Arun Ross, Score normalization in multimodal biometric systems, Pattern Recognition, Volume 38, Issue 12,pp. 2270-2285, 2005 http://dx.doi.org/10.1016/j.patcog.2005.01.012.

[22] Dr. Luai Al Shalabi and Dr. ZyadShaaban, Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix, Proceedings of the International Conference on Dependability of Computer Systems (DEPCOS-RELCOMEX’06), 2006 IEEE

[23] Ejection Fraction, http://my.clevelandclinic.org/heart/disorders/heartfailure/ejectionfraction.aspx

[24] http://publib.boulder.ibm.com/infocenter/db2luw/v8/index.jsp?topic=/com.ibm.db2.udb.doc/admin/c0006909.htm

[25] Dr. Oliver Nelles, Nonlinear System Identification, pp. 137-155

[26] Cluster analysis, http://en.wikipedia.org/wiki/Cluster_analysis

[27] Feature Selection and Extraction
http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/feature_extr.htm Retrieved on May 5, 2013.

[28] Alan P. Reynolds, Graeme Richards, Vic J. Rayward-Smith, Intelligent Data Engineering and Automated Learning – IDEAL, pp. 173-178, 2004

[29] Maria halkidi and Yannisbatistakis , On Clustering Validation Techniques, Journal of Intelligent Information Systems, 17:2/3, pp. 107–145, 2009

[30] Zahid Ansari &A.VinayaBabu, Quantitative Evaluation of Performance and Validity Indices for Clustering the Web Navigational Sessions, World of Computer Science and Information Technology Journal (WCSIT), Vol. 1, No. 5, pp. 217-226, 2011

[31]RapidMiner_OperatorReference,
http://docs.rapid-i.com/files/rapidminer/RapidMiner_OperatorReference_en.pdf

[32] Mu-Chun Su and Chien-Hsing Chou , A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry, IEEE transactions on pattern analysis and machine intelligence, vol. 23, no. 6, 2001

[33] Coronary artery disease, http://en.wikipedia.org/wiki/Coronary_artery_disease

[34] RuiXu and Wunsch, D., II , Survey of clustering algorithms, Neural Networks, IEEE Transactions on Vol. 16, pp. 645 – 678, 2005

[35] Rafael S. Parpinelli, An Ant Colony Based System for Data Mining:Applications to Medical Data, GECCO, pp. 791-797, 2001

[36] Miss Chaitrali S. Dangare, A data mining approach for prediction of heart disease using neural networks, International Journal of computer engineering & technology, Vol. 3, pp.30-40, 2012

[37]Agrawal, R.; Roberto J. Bayardo Jr. "Mining the most interesting rules". Pro ceedings of the 1999 ACM SIGKDD international conference on knowledge discovery and data mining - SIGKDD '99.pp. 145-154, 1999.

[38] J.MalarVizhiand Dr. T.Bhuvaneswari, Data quality measurement oncategorical data using geneticalgorithm, International Journal of Data Mining & Knowledge Manage ment Process (IJDKP) Vol.2, No.1, 2012