International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 17 (2017) pp. 6778-6787
© Research India Publications. http://www.ripublication.com

Analysis on Retrospective Cardiac Disorder Using Statistical Analysis and Data Mining Techniques

Jyotismita Talukdar
Assistant Professor, Department of Computer Science,
University of Technology and Management, Shillong, Meghalaya, Assam, India.

Orcid Id: 0000-0002-5167-7500

Dr. Sanjib Kr. Kalita
Assistant Professor, Department of Computer Science,
Gauhati University, Guwahati, Assam, India

Orcid Id: 0000-0002-0576-3010

Abstract

Abstract: Heart diseases are one of the most prominent reasons for deaths all around the globe till date. According to several surveys, it is found that India has the largest number of heart patients all around the world. CADI in California has confirmed that by 2015, around 62 billion population in India are sole patients of cardiovascular diseases. The most common cause of cardiovascular disease is the inefficiency of the heart to pump blood from heart to the rest of the body and vice versa. In this paper, we present and compare the various statistical and data mining techniques and algorithms in order to predict the specific risk factors of cardiovascular diseases. The different correlations, partial correlations of the risk attributes have been studied and presented in this paper. An attempt has been made to develop a linear model for early prediction of the cardiovascular disease. Further, using data mining techniques, these risk factors are compared to predict the list of risk attributes that are most susceptible to heart disease.

Keywords: cardiovascular disease, k-means, apriori, correlation, CHD, Regression, Rattle.

INTRODUCTION

Cardiovascular disease (CVD) includes heart disease (i.e., myocardial infarction and angina), stroke, hypertension, congestive heart failure (CHF), hardening of the arteries, and other circulatory system diseases. CVD is the number one cause of death in America, responsible for more than 40% of annual deaths. An average of 1 death due to CVD occurs every 33 seconds in the United States [1]. According to the research done by the Registrar General of India (RGI) and the Indian Council of Medical Research (ICMR), people belonging to the age group between 25 years to 69 years contribute to about 25% of mortality rate due to heart diseases[2]. In 2008, five out of the top ten causes for mortality worldwide, other than injuries, were non-communicable diseases; this will go up to seven out of ten by the year 2030. By then, about 76% of the deaths in the world will be due to non-communicable diseases (NCDs) [3] which also includes cardiovascular diseases. In 2010, it was found that of all the widespread diseases around the globe, around 23 million deaths happened only because of cardiovascular diseases (CVDs). In fact, CVDs would be the single largest cause of death in the world accounting for more than a third of all deaths [4]. The most common cause of heart disease is the inefficiency of the heart to pump blood from heart to the rest of the body and vice versa. There are several types of heart disease. Some of them are [5]:

- Coronary heart disease: It also known as coronary artery disease (CAD), it is the most common type of heart disease across the world. It is a condition in which plaque deposits block the coronary blood vessels leading to a reduced supply of blood and oxygen to the heart.
- Angina pectoris: It is a medical term for chest pain that occurs due to insufficient supply of blood to the heart. Angina pectoris, also termed as Angina is basically a warning signal for an early heart attack.
- Congestive heart failure: It is a condition where the heart cannot pump enough blood to the rest of the body. It is commonly known as heart failure
- Cardiomyopathy: Cardiomyopathy is generally caused due to a weak heart which is mainly due to the result of a defective structure of the heart muscle or any changes in the structure of the muscle of the heart. The main cause of Cardiomyopathy is the inadequate pumping of the heart.
- Congenital heart disease: It mainly refers to the formation of an abnormal heart during birth due to a defect in the structure of the heart or its functioning. It is mainly found in children.
- Arrhythmias: Arrhythmias is mainly caused due to improper rhythmic movement of the heartbeat. The heartbeat in this type of disease is either very slow or fast or is generally

irregular. These abnormal heartbeats are caused by a short circuit in the heart's electrical system.

- Myocarditis: Here, the heart muscle suffers inflammation which is usually due to any infection caused by virus, bacteria or any fungus. The symptoms of Myocarditis include joins pain, leg swelling or fever which is not directly related to the heart. However, among all the possible heart diseases, Coronary heart disease is one of the most common heart disease in the world that contributes to almost half of world’s mortality rate. Also known as coronary artery diseases (CAD), here the coronary blood vessels are blocked by the plaque deposits in the arteries, thus reducing the supply of blood and oxygen to the heart. The most important behavioral risk factors of heart disease and stroke are improper diet, lack of physical activity, tobacco consumption and harmful use of alcohol [6]. The results of these risk factors may be prominent in individuals as increased blood pressure, blood glucose, uncontrolled blood lipids, overweight and obesity. These “intermediate risks factors” indicate an increased risk of developing a heart attack, stroke, heart failure and other complications. Data mining can be termed as the process of discovering correlations, patterns and trends by taking into consideration large amount of data stored inn repositories. Several methods can be used to extract useful knowledge from the large data repositories pattern recognition, association analysis as well as statistical and mathematical techniques [7]. It can also be defined as the process of analyzing data from different database and extracting some useful knowledge, pattern, association or relationship out of it. As shown in Figure 1, data mining is a term that describes different techniques used in a domain of machine learning, statistical analysis, pattern recognition, prediction, classification, clustering, visualization, modelling techniques. It has wide applications in all branches of industry such as telecommunications, retail, production, banking, education, and health care management

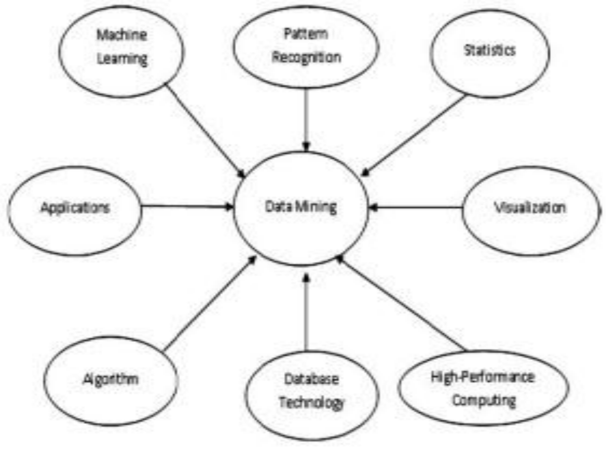


Figure 1: Domains in data mining

Furthermore, statistical analysis is also one of the most trivial data analysis for predicting and comparing data results from a set of data population. Its main goal is to identify the possible trends from a large set of data. Statistical analysis is generally explained using the following five discrete steps:

1. Analysis of the nature of data taken.
2. Exploring the relationship of the data with the given population.
3. Defining a specified model to explain how the data relates to the population.
4. Validation of the model.
5. Predictive analysis using the model to find the possible trends.

PROBLEM STATEMENT

Cardiovascular diseases are considered to be one of the reasons for highest mortality rate across the globe. Early prediction of the risk factors for a person suffering from cardiovascular disease is of utmost importance. However, statistical analysis and data mining techniques can reduce the number of tests that bare generally required to be carried out to predict the occurrence of any cardiovascular diseases. This reduced test set plays an important role in time and performance. Cardiovascular data analysis is important because it allows doctors to see which features or attributes are more important for diagnosis such as age, weight, etc. This will help the doctors diagnose heart diseases more efficiently.

Several techniques are available in the healthcare industry for the early prediction of heart diseases, but research that has to be done to track the performance of various classification techniques, to enable the choice of the best among them can be chosen.

This paper presents a research model predict the heart disease for patients by providing timely response in predicting the disease. This paper mainly focusses on the following topics:

- How various data mining techniques can be used in health care industry and to identify their performance in prediction?

Use of the regression analysis in developing the prediction model to accurately predict the risk factors for patients suffering from heart disease.

Objective

The primary objective of this paper is to develop a predictive model for proper analysis and prediction of the risk factors of cardiovascular diseases using various statistical and datamining techniques. It also shows that data mining can be applied to the medical databases to predict or classify the data with reasonable accuracy.

The following are the objectives leading to achievement of the primary objective mentioned supra:

- To identify the best classification model which can help the physicians in predicting the risk of heart disease using several attributes.
- To recognize and classify patterns in multivariate patient attributes.
- To predict the future outcomes based on previous experiences and present conditions.
- To identify the patients at risk, with the aim of increasing quality of care and to reduce the cost of care.
- To construct a prediction model using several classification techniques such as naïve Bayes, decision trees and support vector machines.

BACKGROUND

Park (2009) in his text book titled 'Preventive and Social Medicine' states that, chronic non-communicable diseases are increasing among the adult population in both developed and developing countries. Cardiovascular diseases and cancer are at present the leading causes of death in developed countries such as Europe and North America. The risk factors that are responsible for morbidity and premature mortality are smoking, alcohol abuse, failure or inability to obtain preventive health services, life-style changes and stress.

Shantakumar et al. (2009) point out that the term heart disease encompasses the diverse diseases that affect the heart and the term cardiovascular disease includes a wide range of conditions that affect the heart and the blood vessels that pump blood throughout the body. It results in disability of several body parts, illness and even death. Myocardial infarctions which is also known as a heart attacks and angina pectoris, or chest pain are encompassed in the CHD. High blood pressure, coronary artery disease, stroke, or rheumatic fever/rheumatic heart disease are the various forms of cardiovascular disease.

Jayshri Sonawane et al. (2013) have illustrated the heart is the organ that pumps blood, with its life giving oxygen and nutrients, to all tissues of the body. If the pumping action of the heart becomes inefficient, vital organs like the brain and kidneys suffer and if the heart stops working altogether, death occurs within minutes.

Latha Parthiban et al. [7] developed an approach taking into consideration the coactive neuro-fuzzy inference system (CANFIS) for early detection of heart disease. The CANFIS model diagnosed the presence of disease by merging the neural network adaptive capabilities and the fuzzy logic qualitative approach and further integrating with genetic algorithm.

Kiyong Noh et al. [8] put forth a classification method for the extraction of multi-parametric features by assessing HRV from ECG, data preprocessing and heart disease pattern. The efficient FP-growth method was the basis of this method which is an associative. They developed a rule to generate the patterns. The multiple rules and pruning, biased confidence (or cohesion measure) and dataset consisting of 670 participants, distributed into two groups, namely normal people and patients with coronary artery disease, were employed to carry out the experiment for the associative classifier.

Akhil Jabbar et al. proposes efficient associative classification algorithm using genetic approach for heart disease prediction.

Hian Chye Koh and Gerald Tan mainly discusses data mining and its applications with major areas like Treatment effectiveness, Management of healthcare, Detection of fraud and abuse, Customer relationship management[9].

Jayanthi Ranjan, in her paper explains how data mining discovers and extracts useful patterns of large data to find useful patterns. This paper describes the uses of data mining to improve the quality of the decision making in Pharmaceutical industry. Issues in the pharma industry are adverse reactions to the drugs [10].

M. Durairaj, K. Meena explains a hybrid prediction system consisting of Rough Set Theory (RST) and Artificial Neural Network (ANN) on medical data. They developed a new data mining technique to assist competent solutions for medical data analysis. [11].

K. Srinivas, B. Kavitha Rani and Dr. A. Goverdhan explains the use of classification based data mining such as Rule based, decision tree, Naïve bayes for medical data. Some vital attributes like age, sex, blood pressure and blood sugar were used in order to predict the likelihood of patients getting a heart attack. [12].

Shweta Kharya discussed various data mining approaches that have been utilized for breast cancer diagnosis and prognosis Decision tree is found to be the best predictor with 93.62% Accuracy on benchmark dataset and also on SEER data set [13].

Elias Lemuye discussed the AIDS is the disease caused by HIV, which weakens the body's immune system until it can no longer fight off the simple infections that most healthy people's immune system can resist. Apriori algorithm is used to discover association rules. WEKA 3.6 is used as the data mining tool to implement the Algorithms. The J48 classifier performs classification with 81.8% accuracy in predicting the HIV status [14].

Arvind Sharma and P.C. Gupta explained the benefits of data mining in medicinal area. Their research area was mainly focused in the blood bank sector. They used the J48 algorithm and WEKA to extract useful knowledge out of the attributes available. The accuracy rates reached to about 89.9% using these algorithms on blood donors. [15].

METHODOLOGY

This paper mainly focusses on the analysis of the cardiovascular diseases using statistical and data mining analysis. This paper explains the statistical analysis of different CHD/CVD risk attributes. The objectives of this chapter are:

- Selecting the risk attributes of CHD/CVD, such as: Fasting Blood Sugar (FBS), Cholesterol (CHOL), Systolic Blood Pressure (BP), Thalach (TH) and age etc.
- To find the correlation (Pearson) or dependency of these risk attributes among themselves.
- To compute the partial correlation coefficients of these risk attributes and identifying the most risky attributes responsible for any possible CHD/CVD disease.
- To develop predictive linear and non-linear statistical predictive models based on the results of correlation (simple and partial) among the selected risk attributes.
- To represent the CHD/CVD victims (male & female), following a pre-defined hierarchy of risk attributes, in the form of tree.

In the present study and analysis of CHD/CVD risk factors the correlations and partial correlation coefficients have been computed on four basic risk attributes , such as : -

- Blood Pressure (BP)
- Fasting Blood Sugar (FBS)
- Cholesterol (CHOL), and
- Thalach (THAL).

The attribute '**age**' has been considered to study the frequency of CHD/CVD victims with respect to age and sex. The Pearson's correlation coefficients have been computed using the relation (3.1)

$$r_{xy} = \frac{\sum_{i=1}^{N}\left[(x_i - \bar{x})(y_i - \bar{y})\right]}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \qquad (3.1)$$

Where $\bar{x} = \frac{\sum x_i}{N}, \bar{y} = \frac{\sum y_i}{N}$ (3.2)

N is the number of sampled data (number of CHD/CVD victims), r_{xy} is the correlation coefficient between x and y. x and y are any two CHD/CVD risk attributes selected from the four risk attributed as mentioned above.

Similarly, the partial correlation coefficients among the four risk attributes (BP, FBS, CHOL and THAL) are computed using the relation (3.3).

$$_{x_i x_j .(y)} = \frac{r_{x_i x_j} - r_{x_i .y} . r_{x_j .y}}{\sqrt{\left(1 - r^2{}_{x_i y}\right)}\sqrt{\left(1 - r^2{}_{x_j y}\right)}} \qquad (3.3)$$

Where, x_i, x_j are the patient's risk attributes for CHD/CVD For example, BP and FBS stand for xi and **x_j** respectively and **y** is representing CHOL (say) or THAL (say).The expression (3.1) represents the partial correlation coefficients between x_i, x_j if they obtain the same score on the variable **y**. For example, in the expression (3.4):-

$$r_{1.2.(3)} = \frac{r_{1.2} - r_{1.3} . r_{2.3}}{\sqrt{\left(1 - r_{1.3}{}^2\right)}\sqrt{\left(1 - r_{2.3}{}^2\right)}} \qquad (3.4)$$

The partial correlation coefficient $r_{1.2.(3)}$ indicates the relationship between variables 1 and 2 when each of them obtained the same score on the variable 3.In case of four variables, the partial correlation coefficients are computed as given in equation (3.5).

$$r_{12.34} = \frac{r_{12.3} - r_{14.3} . r_{24.3}}{\sqrt{\left(1 - r^2{}_{14.3}\right)}\sqrt{\left(1 - r^2{}_{24.3}\right)}} \qquad (3.5)$$

In equation (3.5), the variables 3 and 4 are parallel out. The general form of equations (3.4) and (3.5) is represented as given by equation (3.6).

$$r_{12.34...N} = \frac{r_{12.34...(N-1)} - r_{1N.34..(N-1)} . r_{2N.34..(N-1)}}{\sqrt{\left(1 - r^2{}_{1N.34..(N-1)}\right)}\sqrt{\left(1 - r^2{}_{2N.34..(N-1)}\right)}} \qquad (3.6)$$

The present statistical analysis of CHD/CVD victims of age between 25 years and 100 years is made. The total number of victims analyzed in the present study is 3236.

Data mining comprises of extracting useful knowledge from a set of large database. The process of mining and analyzing data can be carried out in the following steps depending on the type of modelling used for the process, namely predictive or descriptive. A predictive data modelling includes processes like classification, Regression, Time series analysis and then finally prediction. However, a descriptive data modelling includes processes like Clustering, Association, Summarization and Sequence discovery.

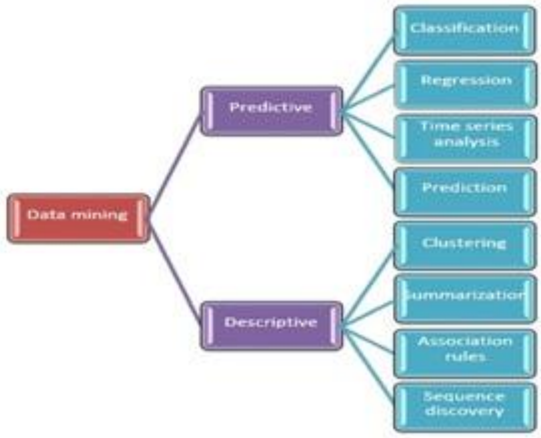


Figure 2: Data mining model and task

This paper specially focusses on the Rattle data mining tool to predict the risk factors for a patient suffering from heart disease. Rattle is a graphical tool that provides a user interface to R programming for analysis and prediction. Rattle data mining tool follows the given steps to extract knowledge from a large data set:

1. **Loading of the data set:** The data is generally in .csv format. Other format allowed are ARFF, .txt, .xls etc.
2. **Exploring the data:** It includes exploring the data in terms of mean, median, variance, range, skewness, kurtosis etc. It also various numeric and graphical tools for the data. It has the option of find the histograms, box plots, dot plots for the data available.
3. **Correlation:** It represents the inter dependence of the input variables present in the dataset. Mathematically, it is defined as the covariance of the dataset divided by the product of their individual standard deviations. The sample correlation coefficient of a dataset where $\mathbf{S_x}$ and $\mathbf{S_y}$ are the sample standard deviation and $\mathbf{S_{xy}}$ is the sample covariance can be defined as follows:

$$S_{xy} = \frac{S_x}{S_x S_y}$$

The values of the correlation factor indicates the strength of the interdependence of the several attributes take. If the value of correlation coefficient is close to 1, it indicates that the variables are closely positively related to each other. Thus the scatter plot falls almost along a straight line with a positive slope. For a negative value of correlation coefficient, variables are linearly negatively related to each other and the scatter plot is linear with a negative slope. However, for value of correlation to be zero, it indicates a very weak relation between the variables. Correlation can be categorized as Pearson, Kendall. Figure 3 shows the interdependence of the risk factors using correlation:

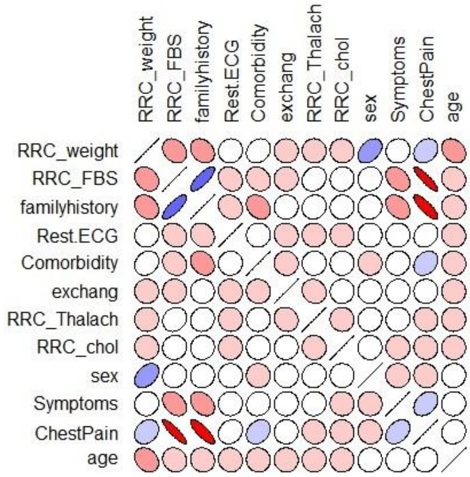


Figure 3: Correlation using Pearson coefficient.

The different shapes and sizes indicate several ranges of correlation values. The diagonal lines indicate perfect correlation between the variables; however it is obvious that there is perfect correlation between the same variables. Zero correlation indicates a perfect circle with white color. For example, there is no correlation between the variables "Sex" and "Cholesterol". Thus it can be well said that cholesterol of a person is completely independent of the gender of a person. The color of the circles has a great relation with the correlation of the variables. The correlation goes on decreasing as the color fades from dark to light. Moreover, the red shades indicate positive correlation whereas the blue shades indicate negative correlation. However, it is seen that for linear data analysis, Pearson provides correlation is the much better than spearman and Kendell correlation. Correlation can also be explained with the help of dendograms.

Clustering is the classification technique for unsupervised data. Clustering allows us to classify data and helps us to consider the cluster set (groups of cluster) by measuring the distance between the clusters. This paper uses the k- means clustering, also called partitioning method where a specific number (say k) of clusters are taken into consideration. Each of the cluster in k- means algorithm analysis uses at least one data object, which is considered to be the basic requirement for clustering process in the partitioning based clustering. It takes into consideration the numeric data. When the number of clusters (say k) is assumed by the user, then a prior clustering method may be formulated to measure the correctness of the algorithm for 'n' objects on 'N' dimensional space among the 'k' cluster groups. K-means algorithm is a partitioning based method that creates k-partitions/clusters from a given set of data. The fundamental requirement for the k- means clustering algorithm is that each cluster must contain at least one data object and each of the data objects must belong to exactly one cluster. The different clusters in k-means partitioning based algorithm is defined by the distance measurement for clustering the data. The most common method to measure the distance for the clustering determination is the Euclidian distance, which is described by the following equation:

$$D_2(x_i,x_j) = \sqrt{\sum_{k=1}^{n}(xi,k - xj,k)^2} = \| xi,xj\|^2$$

The algorithm for k-means position method is as follows:

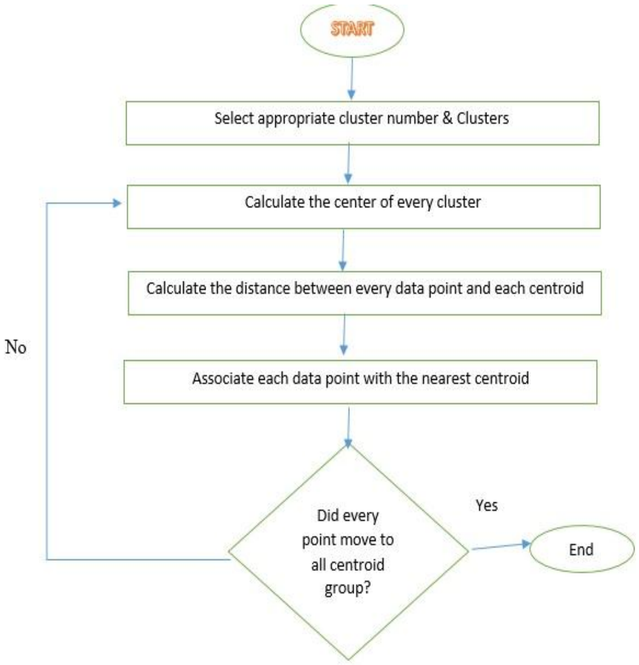

START
Select appropriate cluster number & Clusters
Calculate the center of every cluster
Calculate the distance between every data point and each centroid
Associate each data point with the nearest centroid
Did every point move to all centroid group?
No
Yes
End

Figure 3: K means algorithm

It is an unsupervised partitioning algorithm based on a simple iterative scheme for finding a local minimal solution. It considers a specified number of clusters ‘k’. Let us consider the ‘k’ prototype/clusters be (w1, w2, w3……. Wk) for the ‘n’ input patterns (i1, i2, i3,in) [16]. Thus,

$$W_j = i_{l}, j\varepsilon\{1,2,\ldots..k\}\ l\varepsilon\{1,2,\ldots\ldots n\}$$

The quality of a clustering algorithm is determined by the following error function:

$$E = \sum_{j=1}^{k}\sum_{il\varepsilon Cj}^{k} \|il - Wj\|^2$$

Where, Cj is the jth cluster whose value is a disjoint subset of the input parameter. [16]

The error function can also be explained as following:

$$E = \sum_{i=1}^{c}\sum_{j=1}^{c} \|Xi - Vj\|^2$$

Where, $\|Xi - Vj\|^2$ is the Euclidian distance between Xi & Vj.

“Ci” is the number of data points in the ith cluster.

“C” is the number of cluster centers.

The new cluster is selected using the following formula:

$$Vi = (1/Ci) \sum_{j=1}^{ci} Xi$$

Where, “Ci” represents the number of data points/values in the ith cluster.

The accuracy of the K-Means clustering algorithm can be illustrated as follows:

$$\text{Accuracy (A)} = (t_p + t_n) / \text{total sample.}$$

Where, t_p = true positive

t_n = true negative

The cluster sizes that are considered in the dataset for heart patients are as follows:

189,114,9,237,59,60,178,326,88,19

Runs = 10

Cluster size:

181, 71, 92, 96, 136, 168, 77, 70, 271, 117.

For each pattern ‘X$_i$’, we determine the membership m(C$_j$ / X$_j$) in each cluster C$_j$. The membership function determines the value of ‘X$_i$’ that belongs to the jth cluster ‘C$_j$’. If the pattern ‘X$_i$’ is closest to the jth cluster with similar properties, then it is assumed that:

$$m(C_j / X_i) = 1$$

else

$$m(C_j / X_i) = 0$$

Thus, we can say that the values of m(C$_j$ / X$_i$) is bound to (0,1).

Therefore,

$$m(C_j / X_i)\ \varepsilon\ (0,1)$$

Once the values of the clusters are determined, the centroids of the cluster centers are to be re-computed in order to find the new cluster centers,’V$_j$’ and then calculate the sum of square error ‘E’. The centroids of the re-computed cluster centers are determined as follows:

$$V_j = \frac{\sum_{i=1}^{n}|m(Cj\,|Xi)|\,.\,Xi}{\sum_{i=1}^{n} m(Cj\,|Xi)} \qquad \text{for } j = 1, 2, .k$$

The sum of square error is calculated as follows:

$$E = \sum_{j=1}^{k}\sum_{Xi\varepsilon Cj}^{c} \|Xi - Vj\|^2 \qquad \text{for } i=1\ldots..n$$

$$j=1\ldots..k$$

We repeat the computation of the new cluster centers until convergence.

The data means for K- means clustering analysis is shown as follows:

Table I: Data means of K-Means clustering

Sex	Chest pain	Systolic BP	FBS
0.73	3.64	138.60	0.005

Age	Thalach	Family history	Symptoms
7.891	0.02	0.38	1.94

Comorbidity	Exang	Rest ECG	Chol	Weight
2.59	0.266	0.39	.003	-0.01

Association analysis identifies the relationships or affinities between observations or between variables. These variables/ relationships are expresses as a collection of rules called association rules. Association rule analysis is especially applicable when the size of the transaction database is very large. The association analysis generally comprises of the following two steps:

1. Generation of frequent item sets.
2. Mining the association rules from the frequent item sets [17].

The original concept of association rule mining was brought up by Agarwal who defined the concept of association rule mining [18] [19] as:

Let I = {i_1, i_2, i_3…..i_n} be the set of 'n' binary attributes called "Items"

D = {t_1, t_2, t_3 ….. t_n} be the set of 'n' transactions called "Database".

Each transaction in the database 'D' has a unique transaction id and contains a subset of items in I.

Thus, the association rule is defined as:

$$X \rightarrow Y$$

where,

$X, Y \varepsilon I$

$X \cap Y = ø$

The set of items 'X' is called antecedent

The set of items 'Y' is called consequent.

Apriori algorithm is considered as one of the Boolean association rules for mining frequent item sets. The apriori algorithm was initially proposed by R.Agarwal and R. Shrikanth in the year 1994. The entire Apriori algorithm can be divided into two steps:

Step1: Apply minimum support to find all the frequent sets with 'k' item sets in a database [20].

Step2: Mine the frequent k-item sets to find the frequent sets with (k+1) items

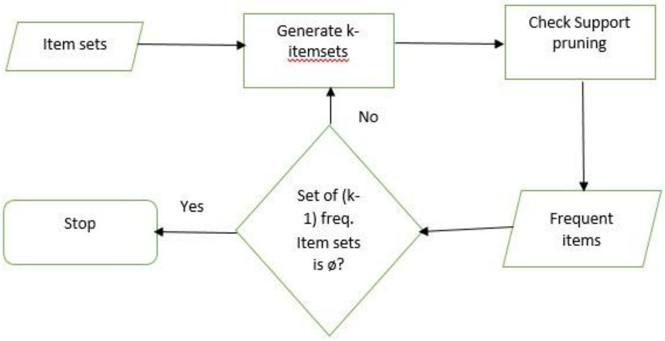


Figure 4: Apriori algorithm

Principal component analysis is the process to find out the different patterns in data, and also finding out the similarities and differences in the data provided. It can found using Singular value decomposition method (SVD) and Eigen value method. The result of principal Component analysis using **SVD** and **Eigen value** method is shown below:

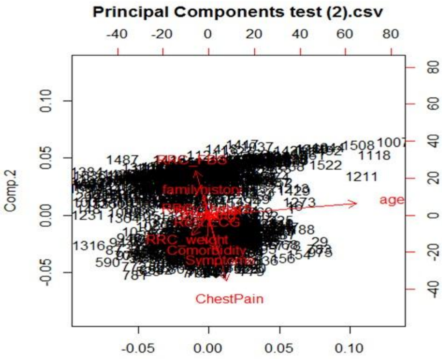


Figure 5: Principal component analysis using Eigen value

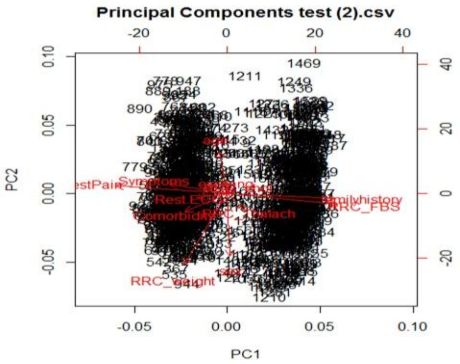


Figure 6: Principal component analysis using SVD

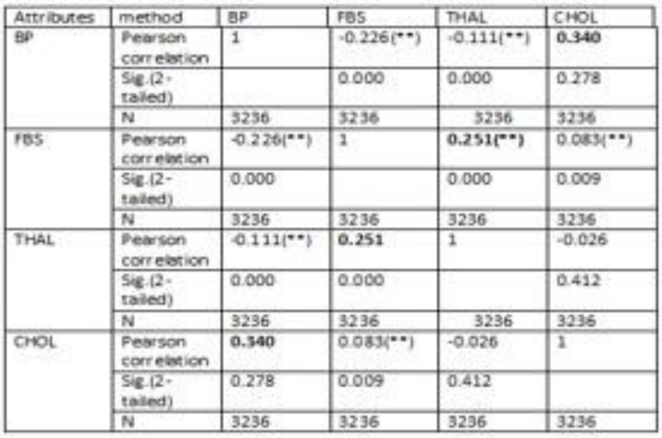
Table II: Standard Deviation Values Using Eigen Method

Comp.1	**Comp.2**	**Comp.3**	**Comp.4**	**Comp.5**	**Comp.6**
6650.446	23.0363	21.7589	17.8627	8.42258	1.65
Comp.7	**Comp.8**	**Comp.9**	**Comp.10**	**Comp.11**	**Comp.12**
1.147458	1.0317	0.48271	0.43233	0.3987077	0.2665

RESULTS AND DISCUSSION

The following table III shows bivariate correlations among BP, FBS, CHOL and THAL.

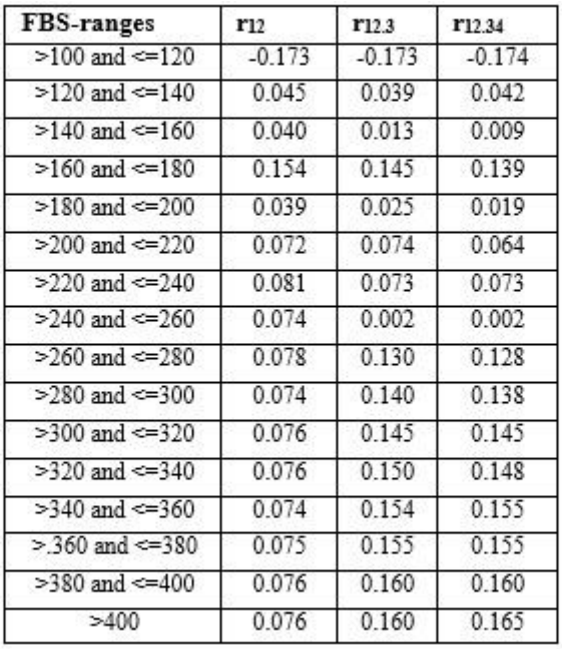
Table III: Correlation among BP, CHOL, FBS and THALACH



Attributes	method	BP	FBS	THAL	CHOL
BP	Pearson correlation	1	-0.226(**)	-0.111(**)	**0.340**
	Sig.(2-tailed)		0.000	0.000	0.278
	N	3236	3236	3236	3236
FBS	Pearson correlation	-0.226(**)	1	**0.251(**)**	0.083(**)
	Sig.(2-tailed)	0.000		0.000	0.009
	N	3236	3236	3236	3236
THAL	Pearson correlation	-0.111(**)	**0.251**	1	-0.026
	Sig.(2-tailed)	0.000	0.000		0.412
	N	3236	3236	3236	3236
CHOL	Pearson correlation	**0.340**	0.083(**)	-0.026	1
	Sig.(2-tailed)	0.278	0.009	0.412	
	N	3236	3236	3236	3236

In the table IV the correlation and partial correlation coefficients between BP and FBS, considering CHOL and THAL as control variables, have been shown

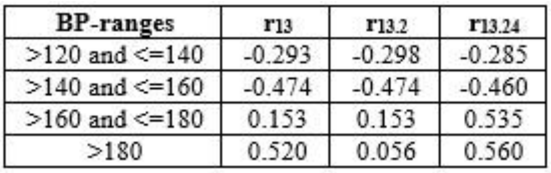
Table IV: Correlation between BP and FBS



FBS-ranges	r_{12}	$r_{12.3}$	$r_{12.34}$
>100 and <=120	-0.173	-0.173	-0.174
>120 and <=140	0.045	0.039	0.042
>140 and <=160	0.040	0.013	0.009
>160 and <=180	0.154	0.145	0.139
>180 and <=200	0.039	0.025	0.019
>200 and <=220	0.072	0.074	0.064
>220 and <=240	0.081	0.073	0.073
>240 and <=260	0.074	0.002	0.002
>260 and <=280	0.078	0.130	0.128
>280 and <=300	0.074	0.140	0.138
>300 and <=320	0.076	0.145	0.145
>320 and <=340	0.076	0.150	0.148
>340 and <=360	0.074	0.154	0.155
>.360 and <=380	0.075	0.155	0.155
>380 and <=400	0.076	0.160	0.160
>400	0.076	0.160	0.165

Table V shows the correlation and partial correlations of BP and THAL, while CHOL and FBS considered as control variables

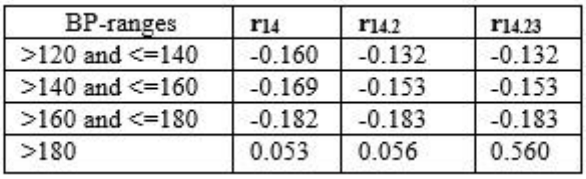
Table V: Correlation between BP and THAL



BP-ranges	r_{13}	$r_{13.2}$	$r_{13.24}$
>120 and <=140	-0.293	-0.298	-0.285
>140 and <=160	-0.474	-0.474	-0.460
>160 and <=180	0.153	0.153	0.535
>180	0.520	0.056	0.560

Table VI shows the correlation and partial –correlation coefficients of BP and CHOL. While THAL and FBS considered as control variables. This has been shown at different BP ranges.

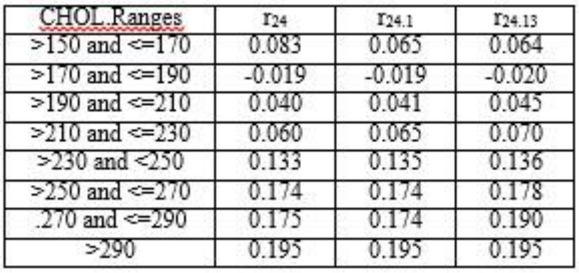
Table VI: Correlation between BP and CHOL



BP-ranges	r_{14}	$r_{14.2}$	$r_{14.23}$
>120 and <=140	-0.160	-0.132	-0.132
>140 and <=160	-0.169	-0.153	-0.153
>160 and <=180	-0.182	-0.183	-0.183
>180	0.053	0.056	0.560

Table VII shows the correlation and partial correlation coefficients between FBS and CHOL, when BP and THAL considered as control variables.

Table VII: Correlation between FBS and CHOL

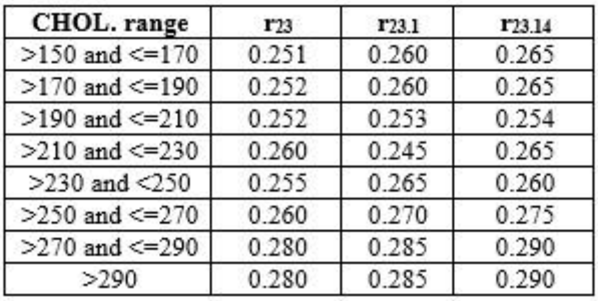


CHOL.Ranges	r_{24}	$r_{24.1}$	$r_{24.13}$
>150 and <=170	0.083	0.065	0.064
>170 and <=190	-0.019	-0.019	-0.020
>190 and <=210	0.040	0.041	0.045
>210 and <=230	0.060	0.065	0.070
>230 and <250	0.133	0.135	0.136
>250 and <=270	0.174	0.174	0.178
.270 and <=290	0.175	0.174	0.190
>290	0.195	0.195	0.195

Table VIII shows the results of correlation and partial correlation coefficients between FBS and THAL while BP and CHOL considered as control variables

.

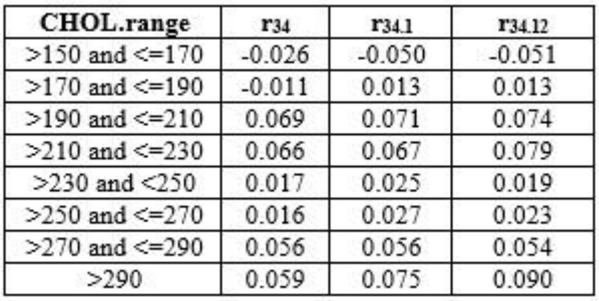
Table VIII: Correlation between FBS and THAL



CHOL. range	r_{23}	$r_{23.1}$	$r_{23.14}$
>150 and <=170	0.251	0.260	0.265
>170 and <=190	0.252	0.260	0.265
>190 and <=210	0.252	0.253	0.254
>210 and <=230	0.260	0.245	0.265
>230 and <250	0.255	0.265	0.260
>250 and <=270	0.260	0.270	0.275
>270 and <=290	0.280	0.285	0.290
>290	0.280	0.285	0.290

Similarly, the table IX depicts the correlation and partial correlation coefficients between THAL and CHOL taking FBS and BP as control variables.

Table IX: Correlation between THAL and CHOL



CHOL.range	r_{34}	$r_{34.1}$	$r_{34.12}$
>150 and <=170	-0.026	-0.050	-0.051
>170 and <=190	-0.011	0.013	0.013
>190 and <=210	0.069	0.071	0.074
>210 and <=230	0.066	0.067	0.079
>230 and <250	0.017	0.025	0.019
>250 and <=270	0.016	0.027	0.023
>270 and <=290	0.056	0.056	0.054
>290	0.059	0.075	0.090

CONCLUSION

In this study, the aim was to design a predictive model for heart disease prediction using statistical analysis and data mining techniques from the 12 attributes dataset that is capable of finding the risk factors in early detection of cardiac disorders. Data collected from Gauhati Medical College and hospital, Hayat Hospital, GNRC Hospital, Downtown Hospital and Narayanan Super-specialty Hospital in India 0f year 2015 was considered and preprocessed for this study. The models were built using the statistical and data mining techniques. Correlation and partial correlation techniques were used to analyze the data set. Data mining used K-Means, Apriori and Decision tree algorithms to detect the most important risk factors for the preprocessing of the data. The performances of the models were evaluated and was the found that Blood Pressure (BP), Fasting Blood Sugar (FBS), Cholesterol (CHOL) and Thalach (THAL) are the most important risk factors for detecting a heart disease. Thus we see that from a total of 12 attributes, 4 attributes proved to be risk factors that were highly relevant. As a future work, I have planned to perform additional experiments with more dataset and algorithms to improve the classification accuracy and to build a model that can predict specific heart disease types.

REFERENCES

[1] "A Literature Review of Cardiovascular Disease Management Programs in Managed Care Populations,SHETA ARA, PharmD, http://www.jmcp.org/doi/pdf/10.18553/jmcp.2004.10.4.326.

[2] Vikas Chaurasia, et al. "Early Prediction of Heart Diseases Using Data Mining Techniques", Carib.j.SciTech, 2013, Vol.1, 208-217.

[3] Preventing Chronic Disease: A Vital Investment. World Health Organization Global Report, 2005.

[4] Global Burden of Disease. 2004 update (2008). World Health Organization.

[5] K.Sudhakar, Dr. M. Manimekalai," "Study of Heart Disease Prediction using Data Mining" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014 ISSN: 2277 128X.

[6] http://www.who.int/mediacentre/factsheets/fs317/en/

[7] Gorunescu, F, Data Mining: Concepts, Models, and Techniques, Springer, 2011.

[8] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences 3; 3, 2008

[9] Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer, Vol:345, pp: 721- 727, 2006.

[10] HianChyeKoh and Gerald Tan, ―Data Mining Applications in Healthcare‖, journal of Healthcare Information Management – Vol 19, No 2

[11] M.Durairaj, K.Meena, ―A Hybrid Prediction System Using Rough Sets and Artificial Neural Networks‖, International Journal Of Innovative Technology & Creative Engineering (ISSN: 2045-8711) VOL.1 NO.7 JULY 2011.

[12] K. Srinivas , B. Kavitha Rani and Dr. A. Govrdhan, ―Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks‖ International Journal on Computer Science and Engineering (2010).

[13] Ming Chuan Hung, Jungpin Wu, Jin-HuaChang, Don lin Yang,"An efficient K-Means clustering algorithm using simple partitioning",Journal of information science and engineering, year 2005.

[14] EliasLemuye, ―Hiv Status Predictive Modeling Using Data Mining Technology.

[15] Rakesh Agaewal, Ramakrishnan Srikanth,"Fast Algorithm for mining association rules inn large databses",Proc 20th International conference very large databases(VLDB), pp-487-499, year 1994

[16] Arvind Sharma and P.C. Gupta ―Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool‖ International Journal of Communication and Computer Technologies Volume 01 – No.6, Issue: 02 September 2012.

[17] Jiao Yabing,"Research of an improved Apriori algorithm in data mining association rules:,IJCCE, Vol.2, No.1, January 2013

[18] Rachna Somkunwar,"A study on various data mining approaches of association rules",IJARCSSE, Vol.2, Issue 9, September 2013.

[19] Jong Soo Park, Mingsyan Cheng, Philip S. Yu,"An effective hash based algorithm for mining association rules",Proc ACM, SIGMOD Conference, PP.175-186, year 1995

[20] Blog.hackereath.com/beginners-tutorial-apriori-algorithm-data-mining-r-implementation.