ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

Surveillance of Heart Disease using Data Mining Technique

Jyotismita Talukdar

Project Engineer, Dept. of Instrumentation & USIC, Gauhati University, Assam, India

ABSTRACT : Diagnosis of diseases from the database available is one of the vital and intricate jobs in medicine. With the advent of time, people are becoming more and more vulnerable to several diseases due to several reasons. One of the most frequently found disease all across the globe is the **heart disease.** Almost 60% of world population become victim of this disease. In this paper, we are trying to find the most probable factors that may be responsible for a person suffering from heart disease. The whole process of data mining is being carried out on the data available for the patients suffering from heart disease. **Rattle data mining** tool is being used for performing the tasks of analyzing the data of the heart patients. The data is being partitioned into training and testing datasets. The next steps namely clustering and modeling is performed on the training datasets .The testing dataset is used to obtain the unbiased errors. We also find out the correlation of the attributes being used in the present study. After finding the relationship of several attributes of the datasets of the heart patients we give a detailed explanation through the use of rattle data mining tool. Finally, the optimal heart parameters related to heart problem are found out for quick and correct diagnosis.

Keywords: Data mining, Rattle data mining tool, Heart disease, Correlation, Clustering, Apriori algorithm, Decision tree and Ada - Boost modeling.

LIST OF ABBREVIATIONS

CHB	Complete Heart Block
CAD	Coronary Artery Disease
SSS	Sick Sinus Syndrome
UA	Unstable Angina
PS	Pulmonary Stenosis.
AWMI	Anterior Wall Myocardial Infarction
AF with SVR	Atrial Fibrillation with Slow Ventricular Rate
LBBB	Left Bundle Branch Block
MI	Myocardial Infarction
IWMI	Inferior Wall Myocardial Infarction
Rattle	R analytical tool to Learn Easily

I.INTRODUCTION

The concept of extracting useful information and making concrete decisions to support medical organizations has existed from centuries when the data was comparatively low. **John Snow[1]** is considered to be the father of modern epidemiology. He tried to analyze the source of cholera using early form of bar graphs in early **1854** and prove that it was transmitted through the water supply. **Florence Nightingale[2]** invented polar-area diagrams in 1855 which helped to trace the deaths of the army due to unsanitary clinical practices. She explained all these with the help of diagrams to convince the policy makers to reduce the number of deaths. However the concept of

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

using data mining in the health sector is not dated long back but just in the early 2003. In 2003,**Wilson et al** [3] scanned cases where Knowledge discovery database and data mining techniques were applied in health databases. **Witten and Frank** in 2005[4] used the concept of data mining to find the trends of the hospital data and also define several patterns of these data. **Cheng, et al in 2006[5]** stated that the enormous amount of data in the databases would make it quite difficult for the people to discover some useful knowledge. **Shillabeer and Roddick in 2007[6]** proposed the concept of using data mining in the health sector to deal with the enormous amount of data to find out some useful information and take fruitful decisions. According to the Health Grades Hospitals Study in 2007 "about 87% of hospital deaths in the United States could have been prevented, had hospital staff (including doctors) been more careful in avoiding errors". **Lavrac et al. in 2007** took help of the geographic information system along with data mining to analyze the similarities of the community health centres in Slovenia. Due to the lack of strict sanitization and sterilization measures in Philippines at the Rizal Medical Center in Pasig City in October 2006, lead to the death of several new born babies due to several bacterial infection. This was because nobody tried to see the factors of death and also the patterns. **Kou et al. in 2004[7]** used the concept of data mining and KDD to discover the faults in credit cards and insurance claims. This greatly helped the health sector insurance policies to be setup strictly to benefit the patients. Cheng, et al stated that the use of data mining in the analysis of the heart diseases of human beings could highly help in the early detection of the heart problems. **Cao et al in 2008[8]** showed how the use of data mining could aid in tracking the changing trends in the cancer vaccines. **Kellogg et al. in 2006** used the concept of spatial modeling, spatial data mining and simulation to show the characteristics of disease outbreak in several places. **Wong et al. in 2005[9]** introduced WSARE, an algorithm to detect disease outbreaks in their early stages. **Thangavel et al in 2006[10]** used the **K-Means clustering algorithm** to predict the outbreak of cervical cancer. **Gorunescu in 2009** explained how computer-aided diagnosis (CAD) and endoscopic ultrasonographic elastography (EUSE) in addition to data mining would create a new area of cancer detection.

II.ROOT CAUSES OF THE DISEASE

Heart diseases are one of the most prominent reasons for the deaths all around the globe till date. Heart disease is a broad term that encompasses several abnormalities on the heart effecting different components of the heart. Heart can also be referred to as cardio, so all diseases relating to heart are known as cardiovascular diseases. According to several surveys, it is found that India has the largest number of heart patients all around the world. According to the California-based CADI, India will have 62 billion heart patients by 2015. 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:

- Coronary heart disease.
- Angina pectoris.
- Congestive heart failure.
- Cardiomyopathy.
- Congenital heart disease.
- Arrhythmias.
- Myocarditis.

Among all of them, Coronary heart disease is the most common heart disease in the world. Also known as coronary artery diseases (CAD), in this type of the coronary blood vessels are blocked by the plaque deposits leading to a reduced supply of blood and oxygen to the heart. For any diagnosis it is very important to have the approximate number of patients available for any time duration. The coronary heart disease can be further divided into several classes. They are:

Unstable Angina (UA): Unstable Angina is a category of heart disease where the heart does not get enough oxygen and blood flow. It is generally resulted from the weak heart muscles also referred to as "Myocardium". Some other risk

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

factors of Unstable Angina heart disease are: high blood pressure, high LDL cholesterol, low HDL cholesterol, obesity, older age, smoking etc

Complete heart Block (CHB): Complete heart block occurs when the electrical signals of the heart cannot pass from the heart's upper chambers to the lower chambers. Some of the symptoms of CHB are as follows:

- Dizziness.
- Palpitation.
- Fatigue.
- Chest pressure or pain.
- Shortness of breath.

CHB is also referred to as the third degree heart block. It refers to the failure of the heart's normal rhythm control called arrhythmia.

Myocardial Infarction (MI): This type of heart disease refers to the problem of malfunctioning of the heart and problem in blood supply to the other parts of the heart

Some of the symptoms of Myocardial Infarction are:

- Feeling of Squeezing in the heart.
- Chest pain
- Fatigue
- Shortness of breath.
- Anxiety.

III.PROBLEM STATEMENT

With the increasing number of patients of heart diseases it has become very difficult to bring out any informative knowledge from the data available for heart patients. As a result, people are not able to make any informative decision regarding the primary causes and relationships of the several factors of the heart diseases. It is very necessary to bring out some informative knowledge that would help the doctors and patients to carry on their research and diagnosis much more easily. Data mining helps in this regard. However, the users are not much aware of the data mining tools. As we are primarily focusing on the Rattle data mining tool we need to see the pros and cons of this tool. Also we need to check its adaptability on this research

IV DATA COLLECTION

The data used in the present study collected from five Govt. and private Hospitals of the North Eastern region of India. In the present study data of 5000 heart patients (Male and Female)have been used. The authenticity of the data have been examined by the Cardiologists of the concern Hospitals.

V.DATA MINING IN HEALTHCARE SECTOR

With the increasing number of heart patients globally,the healthcare industry or the hospitals produces a large amount of complex data on patients records. These are mostly based on the respective hospital resources, diagnosis methodology followed, hospital instruments, and doctor's expertise etc. However, the required expertise and technology is yet a far cry for most of the countries. Further, the lack of effective analytical tools to bring out useful information on the patient data is major bottle neck . In addition, the processing of large and complex date sets by traditional methods is another big hindrance against the correct diagnosis. For this case, data mining offers as an effective method to discover very useful information from these databases. Data mining can provide the healthcare professionals an additional source of knowledge for making decisions. Data mining is also

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

useful in the health sector to detect fraud and abuse of the data and also helps the patients to receive better care and services at affordable prices.In healthcare sector the use of data mining has considerably increased due to the following reasons

Data overload

Evidence-based medicine and prevention of hospital errors

Early detection and/or prevention of diseases:

Fraud and Anomaly Detection

Diagnosis and Treatment

Customer Relationship Management

Rattle data mining tool

The abbreviation of Rattle is the **R** Analytical Tool to Learn Easily. Rattle is a graphical data mining tool that provides a pathway to R (**Williams, 2009b**). Rattle has been developed using the **Gnome (1997)** toolkit with the **Glade (1998)** graphical user interface (GUI) builder. Gnome is independent of any programming language; however the graphical user interface of rattle uses python programming language. Rattle uses the Gnome graphical user interface as provided through the RGtk2 package. One of the most important point of Rattle is that it runs under various operating systems namely MS/Windows, GNU/Linux, and Macintosh OS/X. R is used as data mining tool in Rattle. It is one of the open source data mining toolkits. Rattle is based on an extensive collection of R packages. These packages are all available from the Comprehensive R Archive Network (CRAN)

VI.PROPOSED METHODOLOGY

The main aim of this paper is to find out useful relationships of the data being available and also exploit the data knowledge. The steps that we need to follow for this process are as follows:

Load the dataset: Rattle can load the dataset from various sources. It can accept files from CSV (comma separated data), text file, **.xls** files. It can accept data through ODBC connections, ARFF format also. For the present study , we shall be dealing with the .**xls** file that contains the heart data of the patients suffering from heart disease.

Explore: This special tab has several options. It allows us to explore the data that is being available and also view detailed description about the various attributes.. It has the option of find the histograms, box plots, dot plots for the data available It helps us to find the mean, median, quartile, range, variance, co-variance, correlation, skewness, and kurtosis

VII. ANALYSIS AND RESULTS

In the exploration tab, the several options are: **Summary, Basics, Distributions, Correlation,** and **Principal components**. So we described each and every option in detail. We initially find the probability of the chances of the occurrences of heart disease taking into consideration the **Age factor** as the **target variable**.

A.*SUMMARY*

It gives a basic detail of the dataset. It gives information about the mean, median, first quartile and third quartile. It also gives information about the minimum and maximum values of each attribute of the dataset. The results are as shown in Table-1.0 : It gives a basic detail of the dataset.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

Table 1.0

Sex	**Systolic BP**	**FBS**
Mean: 0.7365	Mean: 138.6	Mean: 351
Median:1.0000	Median:135	Median:133
Maximum:1	Maximum:196	Maximum:227
Minimum:80	Minimum:40	Minimum:0
1st Quartile:0	1st Quartile:120	1st Quartile:98
3rd Quartile:1	3rd Quartile: 150	3rd Quartile:220
Chest Pain	**Thalach**	**Family History**
Mean: 3.65	Mean: 79	Mean: 0.3793
Median:4	Median:78	Median:0
Maximum:5	Maximum:190	Maximum:1
Minimum:1	Minimum:40	Minimum:0
1st Quartile: 67	1st Quartile: 2	1st Quartile: 0
3rd Quartile:5	3rd Quartile: 88	3rd Quartile: 1
Symptoms	**Co-morbidity**	**Exchang**
Mean: 2	Mean: 2.554	Mean: 0.2663
Median:1	Median:2	Median:0
Maximum:5	Maximum:8	Maximum:1
Minimum:1	Minimum:1	Minimum:0
1st Quartile:1	1st Quartile: 2	1st Quartile: 0
3rd Quartile:3	3rd Quartile:4	3rd Quartile: 1
Rest-ECG	**Cholestral**	**Age**
Mean: 0.382	Mean: 234.3	Mean:57.05
Median:0	Median:230	Median:59
Maximum:270	Maximum:1	Maximum:98
Minimum:0	Minimum:24	Minimum:8
1st Quartile:220	1st Quartile: 0	1st Quartile: 51
3rd Quartile:1	3rd Quartile: 250	3rd Quartile: 65

Table 1 gives information about the mean, median, first quartile and third quartile. It also gives information about the minimum and maximum values of each attribute of the dataset. The results are as follows:

B.BASICS:

It gives the information about the mean, median and all the details given by the Summary tab, including the skewness, kurtosis and variance of the values. They are shown as follows:

Table 2.0 :

Sex	**Chest Pain**	**Systolic BP**	**FBS**
Variance: 0.194221	Variance: 1.991355	Variance: 488.084	Variance: 441.527
Skewness: -1.0724	Skewness: -0.403539	Skewness: 0.662051	Skewness: 34.07858
Kurtosis: -0.850504	Kurtosis: -1.521033	Kurtosis: -0.168617	Kurtosis: 1160.56255

Thalach	**Family History**	**Symptoms**	**Co-Morbidity**
Variance: 513.138	Variance: 0.235629	Variance: 1.717466	Variance: 1.553768
Skewness: 1.671943	Skewness: 0.496959	Skewness: 1.039847	Skewness: 0.529550
Kurtosis: 4.619513	Kurtosis: -1.754532	Kurtosis: -0.128345	Kurtosis: -0.765066

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

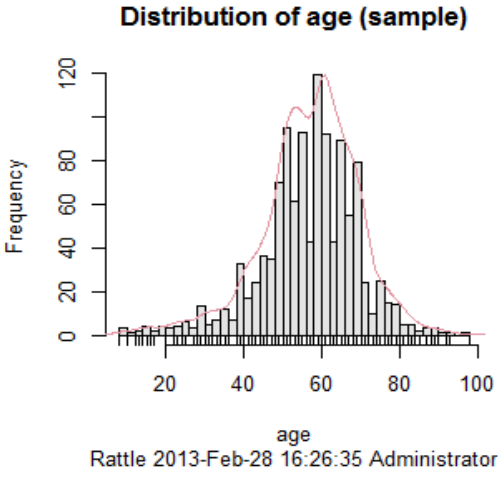
(An ISO 3297: 2007 Certified Organization)

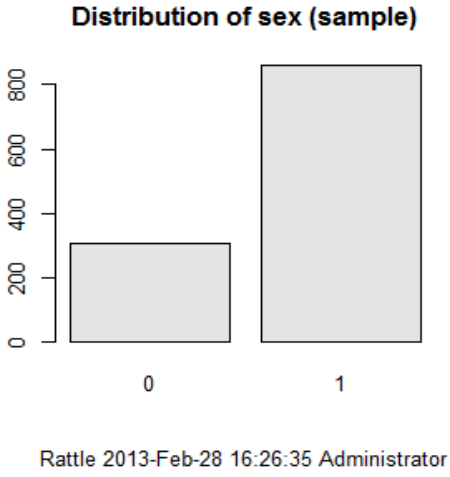
Vol. 2, Issue 8, August 2013

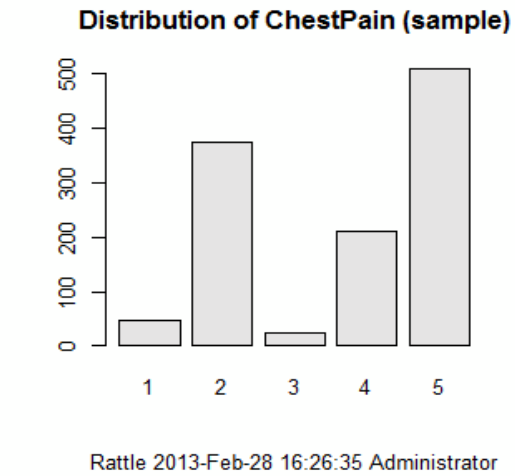
Exchang	Rest ECG	Cholesterol	Age
Variance: 0.195536	Variance: 0.236568	Variance: 322.186990	Variance: 154.760627
Skewness: 1.056244	Skewness: 0.479732	Skewness: -1.23709	Skewness: -0.614224
Kurtosis: -0.885105	Kurtosis: -1.771370	Kurtosis: 14.833724	Kurtosis: 1.39200
Weight	**Disease**		
Variance: 71.634512	Variance: 0.217539		
Skewness: -0.345691	Skewness: .774131		
Kurtosis: 1.622876	Kurtosis: -1.401922		

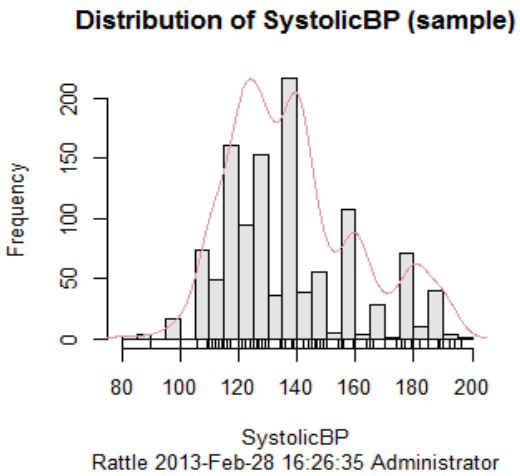
C.DISTRIBUTIONS

It helps to generate several plots to explore the various distributions of the data. It gives representations using **Histograms** and **Cumulative distribution** and **Ben ford** graphs.







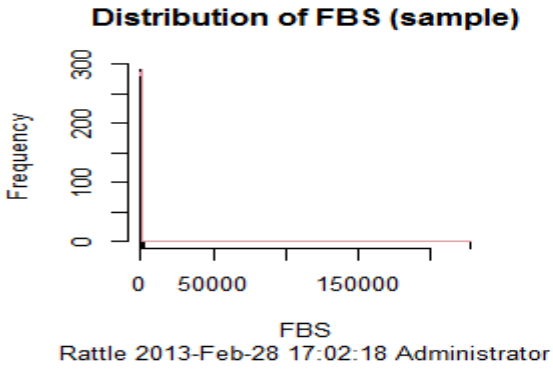


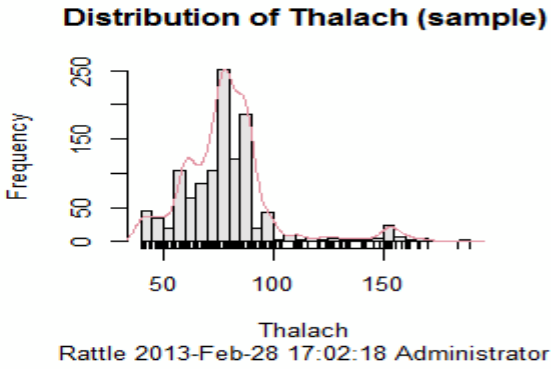
ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

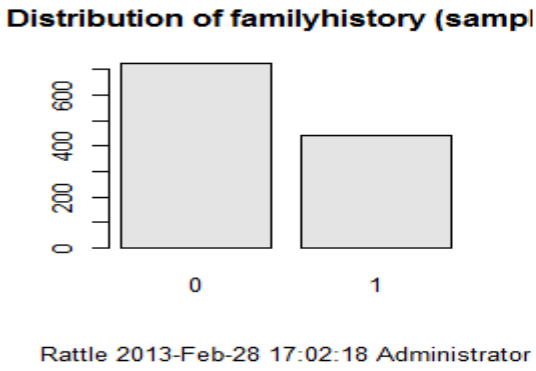
International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

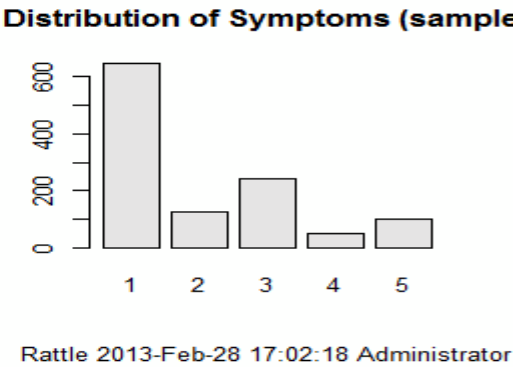
(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013









ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

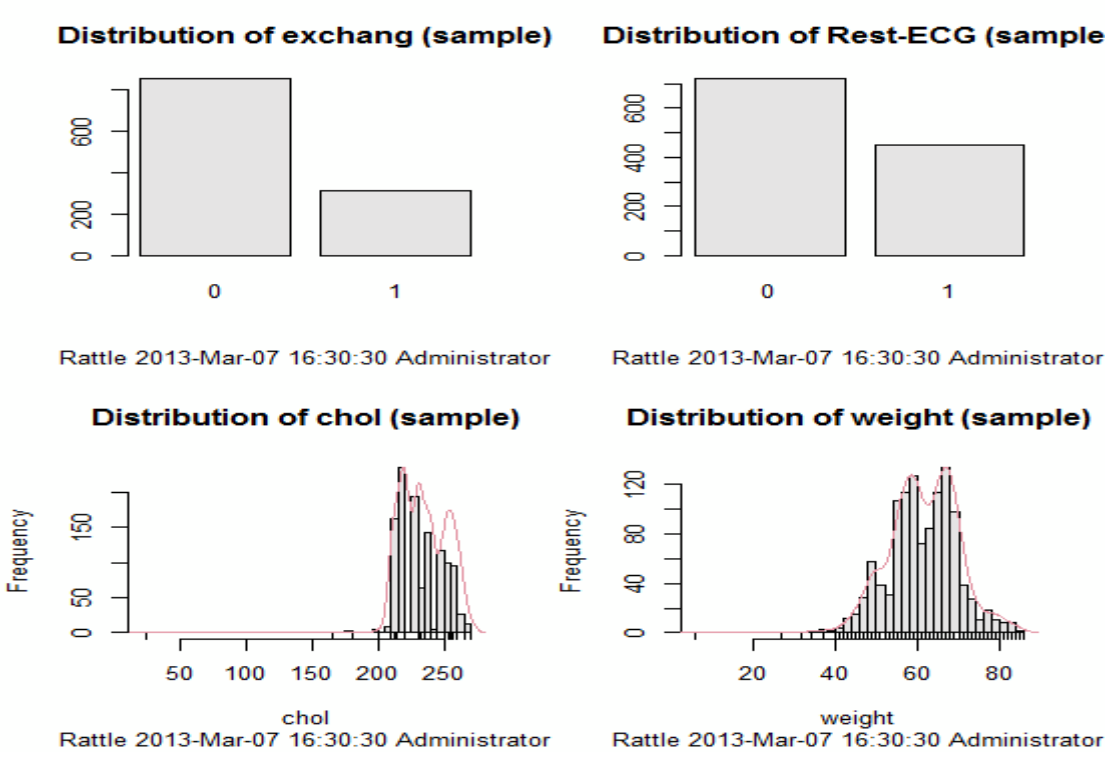


Fig. 1 Frequency Histogram of all variables.

The fig. 1 represents the frequency distribution of all the variables considered for the present study.

D.CUMULATIVE DISTRIBUTION:

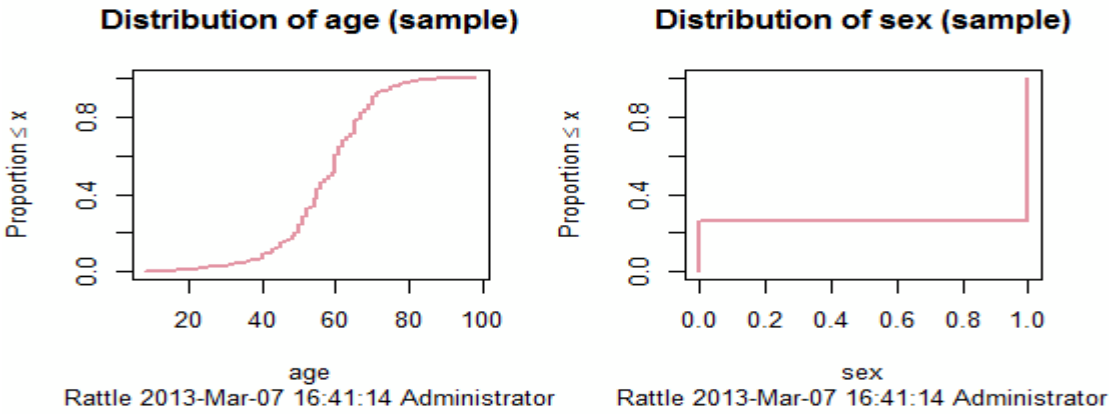
The cumulative distribution function generally finds the amount of probability that a random variable "X" will be found at a probability distribution less than or equal to x.

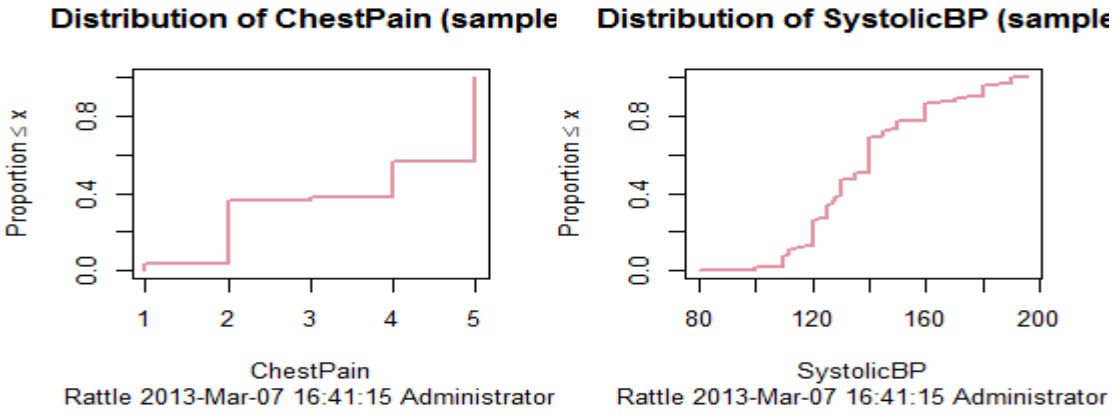
ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

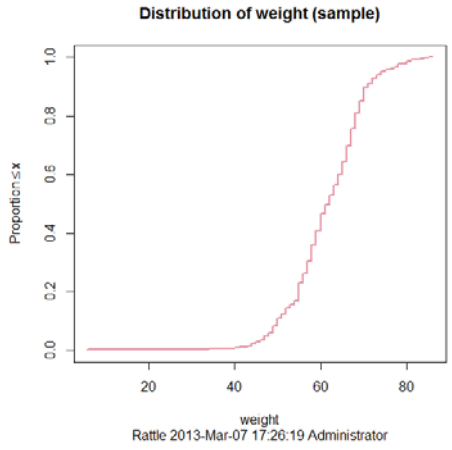
International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013





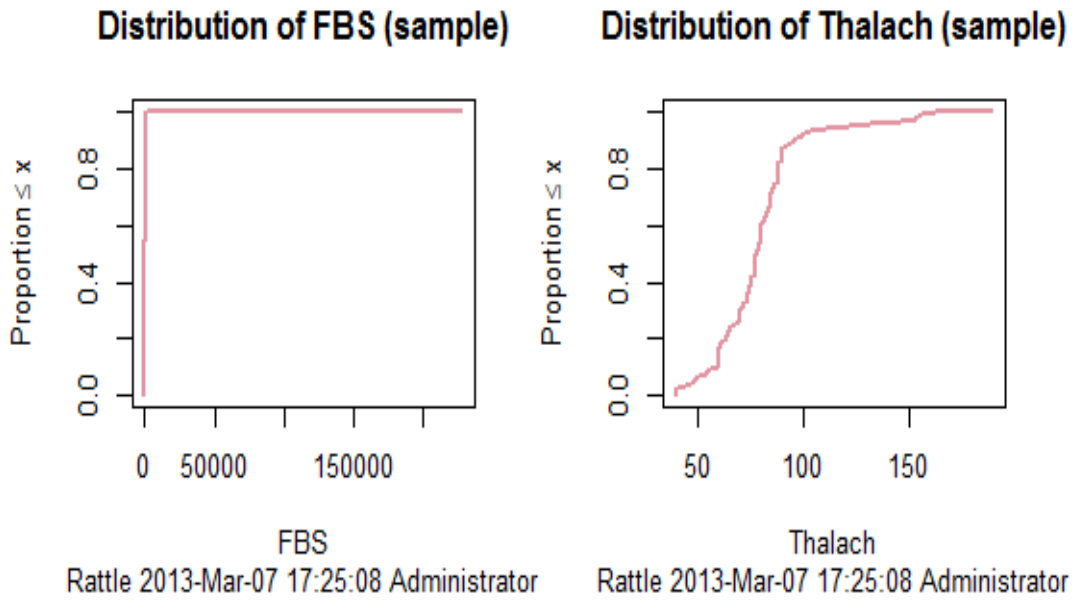


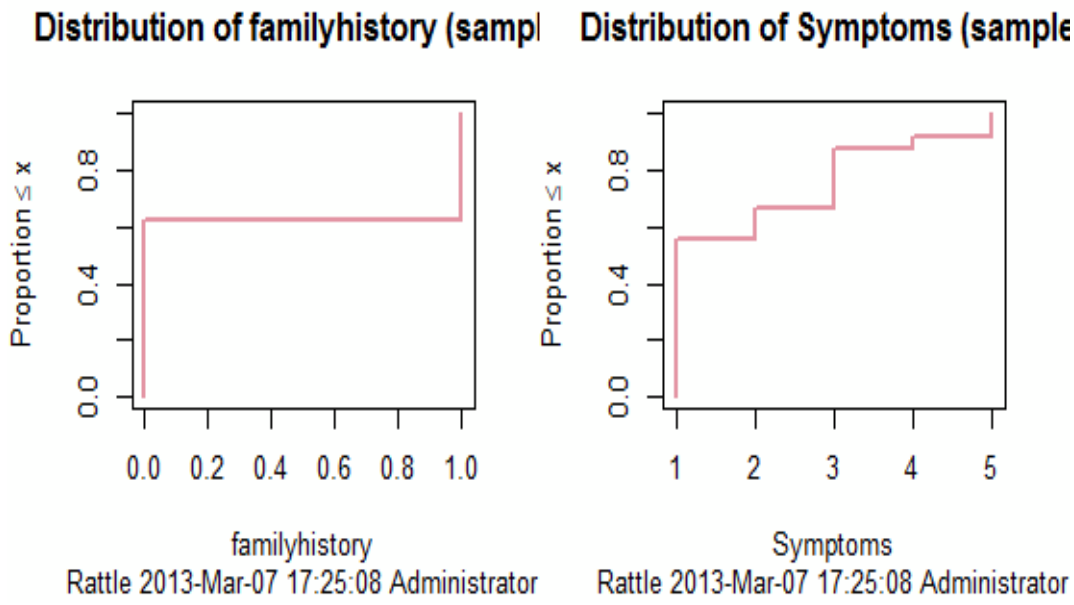
ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013





ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

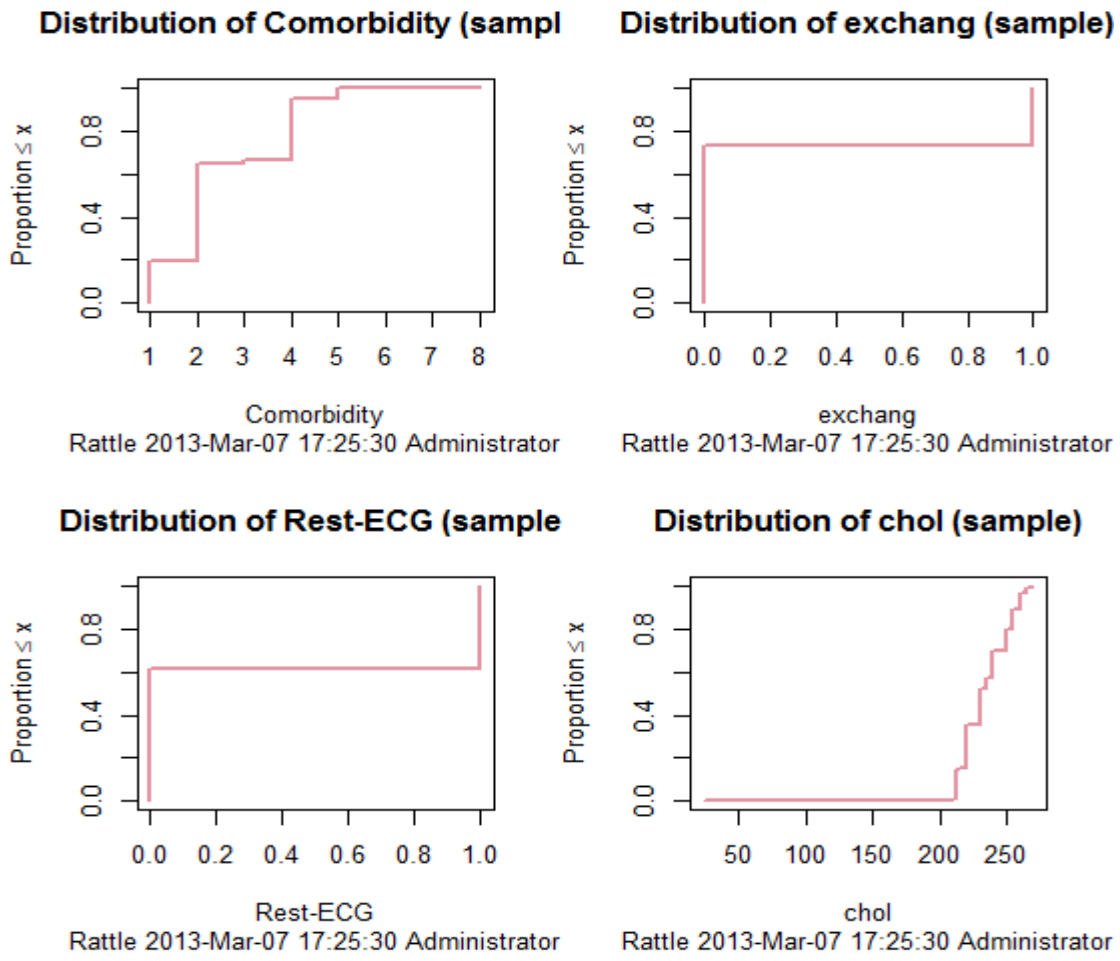


Fig.2 Cumulative distribution of all variables

Fig. 2 represents the cumulative distribution of all the variables/parameters considered for the present study.

VII.CORRELATION

It refers to the dependency of two variables. It can be done in various methods namely: **Pearson, Spearman, Kendell**. In the Rattle data mining tool, it supports these three methods to find the correlation among the various variables. These are shown graphically below : We can understand the order and degree of correlation by looking at the shape and color of the graphic elements as given in **fig. 2.0**. If any two variables are perfectly correlated to each other with the value of correlation equal to 1, then it is represented by a line as shown in the above figure. However, a perfect circle in the figure indicates that there is almost negligible correlation between the variables or we can say that there is no correlation between them. The circles turn into straight lines gradually as the correlation between the variables goes on increasing. Again we see that the direction of alignment of the ellipses indicates whether the correlation is positive or negative. If the ellipse is aligned anti clock wise then we conclude that the correlation is negative. Whereas, if the ellipse is aligned clockwise, then the correlation is said to be positive. The color of the ellipses and circles indicates the degree of the correlation. For zero correlation, the color of the circle is white which becomes darker as the degree of correlation increases. When the variables are perfectly correlated, then the color of the line is perfectly black. Thus we can say that as the degree of correlation increases, the color of the ellipse goes on

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

becoming deeper. Red shades are used for positive correlation whereas blue shades are used for negative correlation.

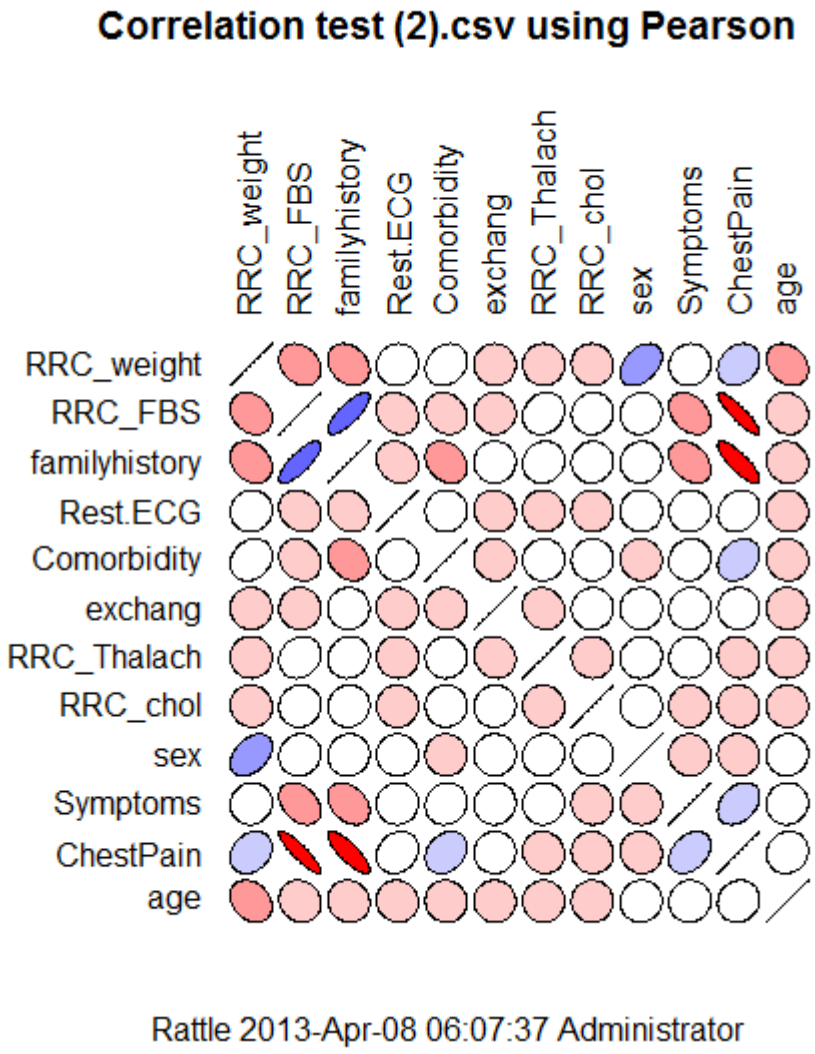


Fig.3 Correlation using Pearson method .

Fig. 3 represents he dependency of two variable out of the 12 variables considered for the present study. So one can find the correlation of the 12 different variables considered for the present study just by seeing the shapes and orientation of the circles/ellipse .

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

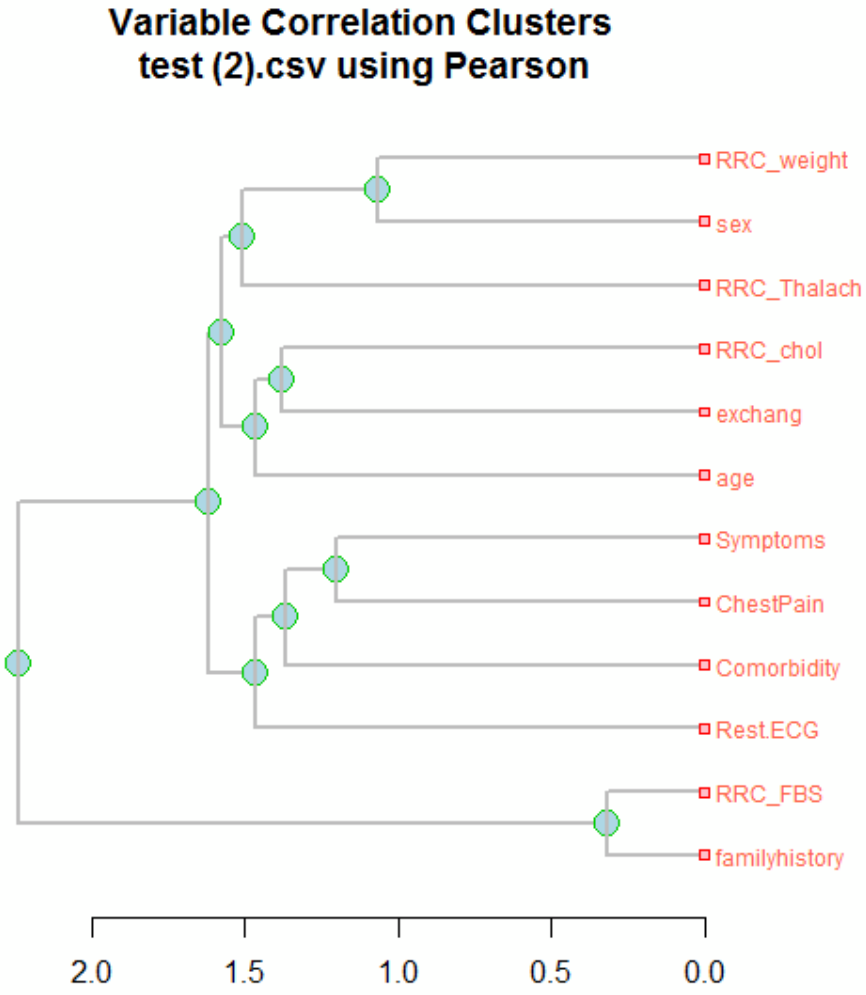


Fig. 4 Variable correlation clusters

The fig. 4 represents the Variable correlation clusters of the total 12 parameters considered for the present study.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

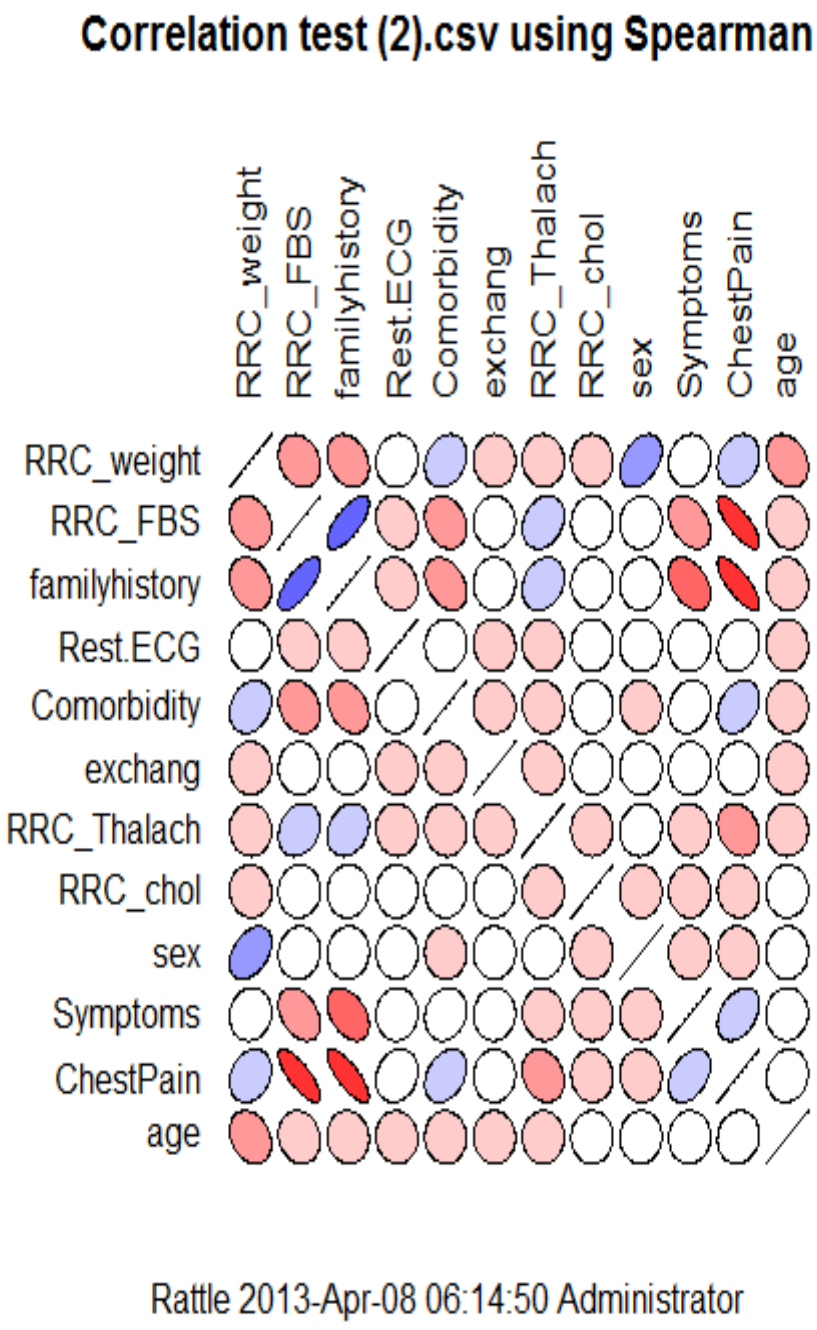


Fig. 5 correlation using Spearman method

The fig. 5 represents theSpearman method of determining the correlation among the different parameters considered for the proper diagnosis of any heart related problem. The shapes , perfectly circle or elliptical with orientations

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

indicates the magnitude of correlation amongst the parameters as described in the text..

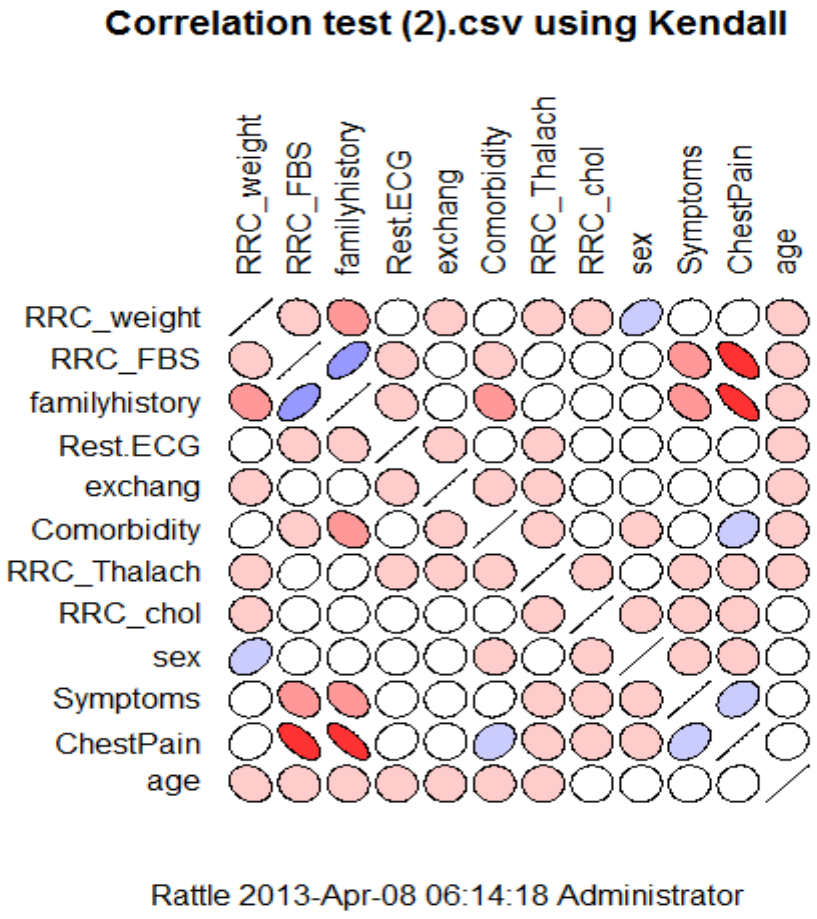


Fig.6 correlation using Kendell method

The fig. 6 represents the Kendell method of determining the correlation among the different parameters considered for the proper diagnosis of any heart related problem. The shapes , perfectly circle or elliptical with orientations indicates the magnitude of correlation amongst the parameters as described in the text..

For linear relationship between the variables, Pearson correlation method is better. Whereas, for nonlinear relationship, Spearman correlation method is better. Now looking into the correlation using hierarchical method, we get the correlation between the variables in the form of a dendogram. The variables are plotted in the right side. The variables are linked to the dendogram according to how well they are correlated. The x axis is a measure ranging from 0 to 2. The length of the lines within the dendogram gives an indication of the level of correlation between the variables. The correlation of the variables is now shown with the help of a dendogram as follows:

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

VIII.MODELLING

A. DECISION TREE

We can build a decision tree using Rattle's Model tab's Tree option. After loading our dataset, and identifying the input variables and the target variable, we select the Model tab's Decision Tree option. Decision tree is a very convenient way for the efficient representation of knowledge. Let's look at the structure of the tree which is presented in the text view.

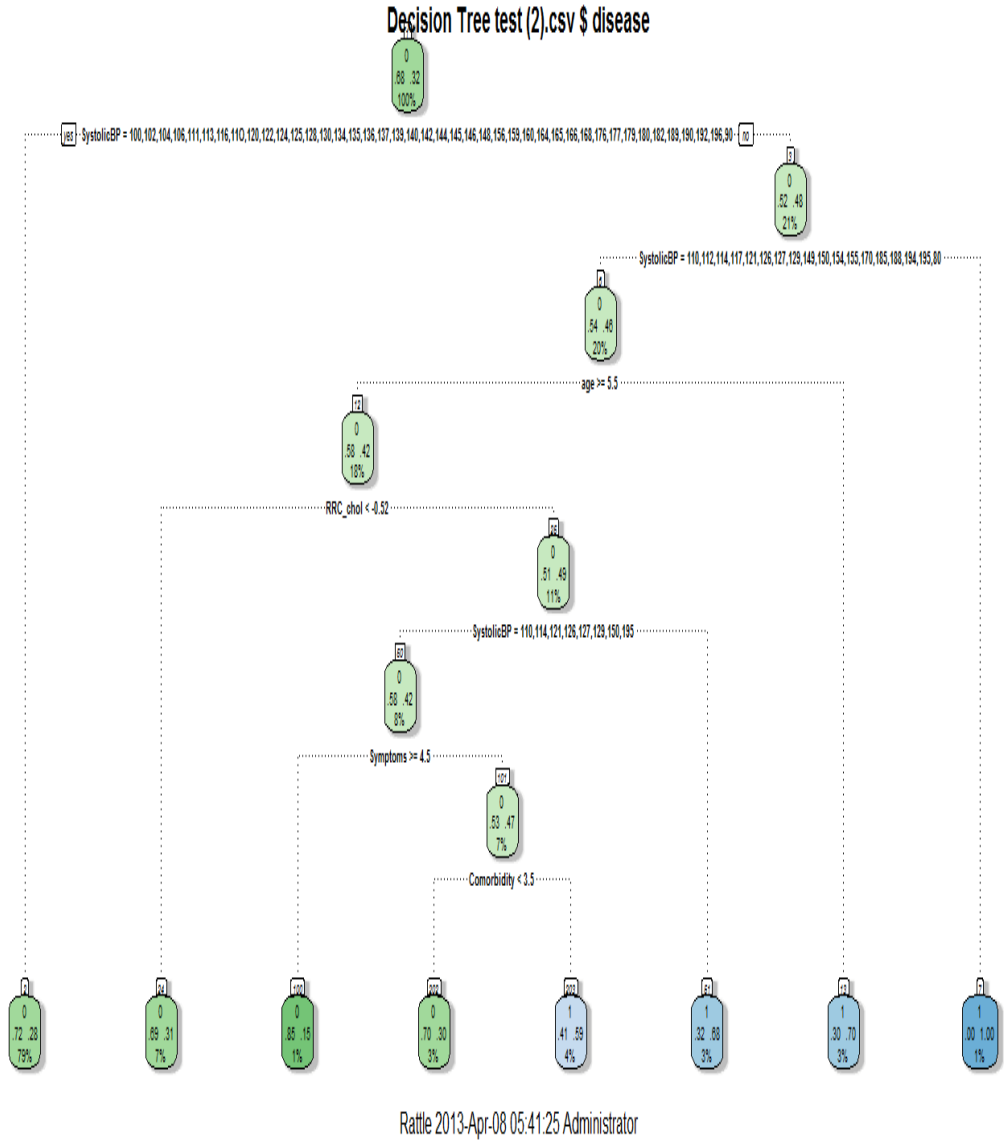


Fig. 7 Decision Tree of the model

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

As represented in fig. 7, the decision tree model is very convenient way for the efficient representation of knowledge. Let's look at the structure of the tree which is presented in the text view.

B. ERRORS

This option helps to plot the error rate progressively for the number of trees being built. In our case the total number of trees being built is 50. However, it is to be noticed that the whole error rate is determined in the training dataset. The relative error rate is shown as follows:

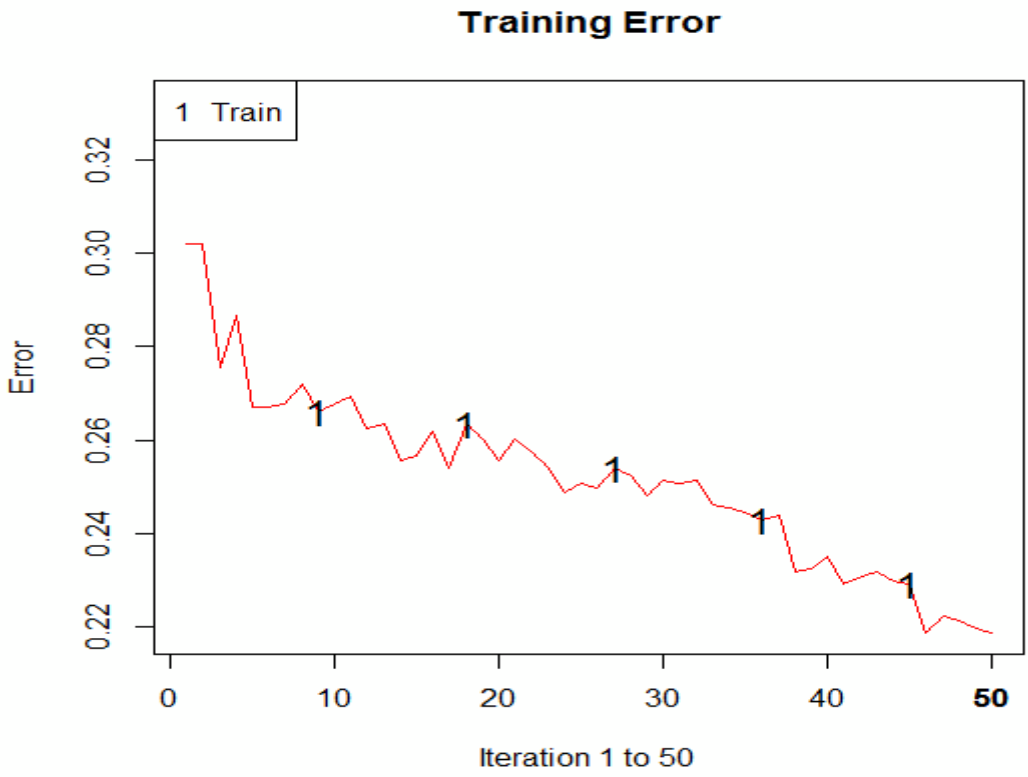


Fig.8 Training Error plot

This plot, fig.8, helps us to visualize the error rate encountered with the training dataset. This error rate has to be considered while making use of the optimized parameter for proper diagnosis.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

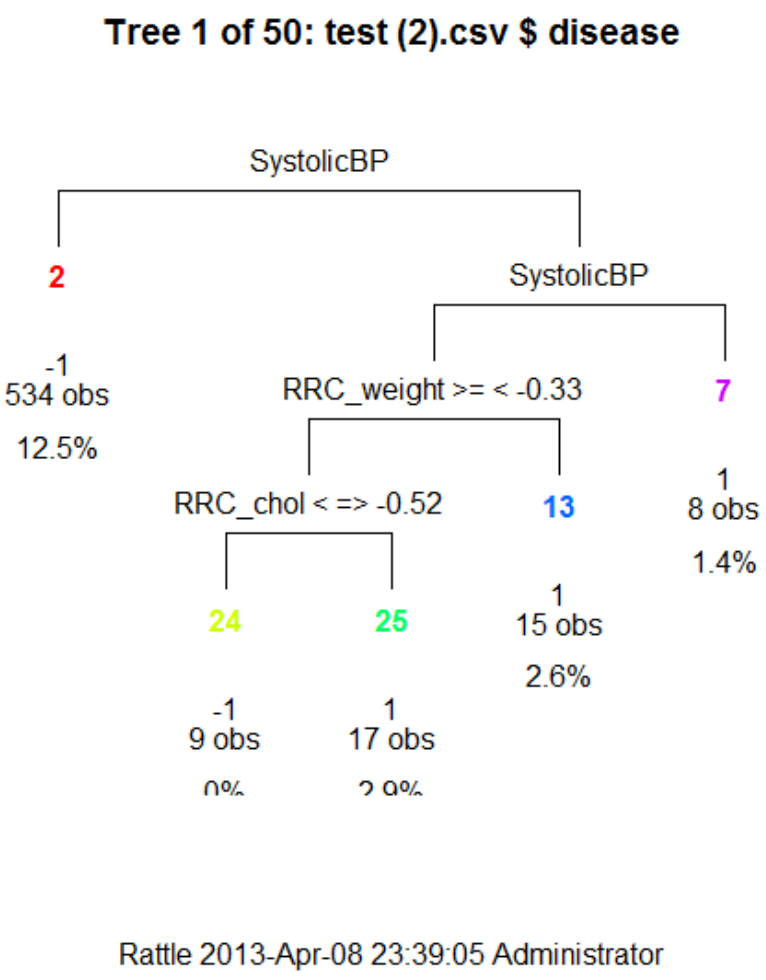


Fig.9 Decision Tree using Ada Boosting algorithm

IX. CONCLUSION

In this paper we are trying the isolate the most common factors of heart disease using Rattle data mining tool and computed the relationship among the several attributes. The dataset that is being taken into consideration consists of 16 variables. However, out of these 16 attributes, two attributes were ignored as it does not contribute in any way in the analysis of our data. The attribute" **disease**" is taken as the target and "**id**" is taken as the **iden**t variable. Before performing any analysis we preprocess the data as it contained many abnormalities. The deleted values were ignored. The attributes "**cholesterol**", "**Weight**", "**Thalach**" and "**FBS**" were normalized and then the values of "**Cholesterol**" and "**Weight**" are categorized. Thus from the overall analysis of the data we can come to a very important conclusion which would really benefit the patients in need. Out of the several variables that are present in the dataset, we see that only **five variables** really contribute to the determination of the probability of the occurrences of heart disease for a patient. Thus it is really helpful for a patient to find out if he/she is susceptible to heart disease or not.

This work can be further extended if more variables are available to find the adaptability of the variables with each

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2013

other and also with other diseases. We can extend our observations and findings considering impact of environmental conditions in day to day life.

REFERENCES

[1] Steven Berlin Johnson " The Ghost Map : The story of London's most Terrifying Epidemic – and how it changed science, cities and the modern world ". 2006

[2] Audain,C. Florence Nightingale Online:
http://www.scottlan.edu/riddle/women/nitegale.htm,2007.

[3] Wilson,R;Hemsley.Brown,J.;Easton,C;Sharp,C;: Using research for School Improvement : The LEA's Role , National Foundation for Education and Research(NFER), Slough.2003.

[4] Witten,I.H and Frank . E (2005) : Data Mining : practical machine learning tools and techniques.Morgan Kaufmann series in data management systems. Morgan Kaufmann. Wilson A.,Thabane, I, Holbrook A (2003) " Application of Data Mining techniques in pharmacovigilance ". British Journal of Clinical Pharmacology. (57)2, 127-134.

[5] Cheng, T .H.,Wei,C.P.,Tseng,V.S. Feature Selection for Medical Data Mining : Comparisons of Expert Judgment and Automatic Approaches. Proceedings of the 9th IEEE Symposium on Computer- Based Medical Systems (CBMS'06),2006.

[6] Shillabeer ,A. and Roddick, J, Establishing a lineage for Medical Knowledge Discovery. ACM International Conference Proceding Series . (311)70,29-37.2007.

[7] Kou,Y.,Lu,C.-T.,Sirwongwattana,S.,and Huang,Y.-P. Survey of fraud detection techniques . In networking , Sensing and Control,2004 IEEE International Conference on Networking, Sensing and Control. (2) 749-754.,2004.

[8] Cao,X.,Maloney,K.B. and Brusic,V. Data Mining of Cancer vaccine trials : a bird's –eye view. Immunome Research,4:7.DOI:10.1186/1745-7580-4-7.2008.

[9] Wong,W.K,Moore,A.,Cooper,G. and Wagner,M, What's Strange About Recent Events(WSARE) : An algorithm for the early Detection of Disease Outbreaks. Journal of Machine Learning Research . 6, 1961-1998.,2005.

[10] Thangavel,K., Jaganathan,P.P. and Easmi, P.O.
:Data Mining Approach to Cervical Cancer Patients Analysis Using Clustering Techniques . Asian Journal of Information Technology (5) 4 ,413-417.2006.

BIOGRAPHY

Ms Jyotismita Talukdar , born in July, 1989, did her B.Tech(Computer Science) under Gauhati University in 2010. She completed her M.Engineering (Computer Science and Information Management)) from the Asian Institute of Technology (AIT), Thailand in 2013. Presently she is working as a project fellow in the Dept. of Instrumentation & USIC,Gauhati University. She is one of the core member of the Speech research group of the Gauhati University. Her interested field of research are application of Data Mining in health sector, Speech processing and Image Water Marking..