International Journal for Modern Trends in Science and Technology
Volume: 03, Issue No: 08, August 2017
ISSN: 2455-3778
http://www.ijmtst.com

A Heart Disease Prediction Model Using Quality Management and SVM with Logistic Regression

K.Parish Venkata Kumar[1] | B.D.C.N Prasad[2]

[1]P.hD Scholar, Dept of Computer Science and Engineering PP.COMP.SCI&ENG.0165C Rayalaseema University, Kurnool.
[2] Dept of Computer Science and Engineering PSCMR College of Engineering & Tech, Vijayawada, India.

To Cite this Article
K.Parish Venkata Kumar, and B.D.C.N Prasad, "A Heart Disease Prediction Model Using Quality Management and SVM with Logistic Regression", International Journal for Modern Trends in Science and Technology, Vol. 03, Issue 08, August, pp.-191-199.

ABSTRACT

we explore the expectation and issues of data condition in this appropriate domain afford primary exploration indication and desire about this restrictive. We underline the agreement of the analysis of data aspect on e-Health operation, exclusively respecting remote monitoring and allowance of patients with chronic action. Secondly The early prediction of cardiovascular diseases can aid in making arrangement to lifestyle changes in high risk patients and in turn diminish their complexity. Research has pursue to pinpoint the most dominant factors of heart disease as well as exactly anticipate the overall risk using homogenous data mining techniques. Recent analysis has delved into admix these techniques using access such as hybrid data mining algorithms. This paper nominate a rule based model to compare the efficiency of applying rules to the particular results of abutment vector machine, decision trees, and logistic relapse on the Cleveland Heart Disease Database in order to present an authentic model of predicting heart disease.

Keywords: *Heart disease, support vector machine (SVM), logistic relapse, decision trees, rule based access.*

I. INTRODUCTION

Data mining (DM) is the eradication of useful advice from large data sets that results in anticipate or describing the data using approach such as classification, clustering, cooperative, etc. Data mining has found expanded appropriateness in the healthcare commerce such as in classifying excellent treatment methods, anticipate disease risk factors, and finding adequate cost arrangement of patient care. Research using data mining figure have been practiced to diseases such as diabetes, asthma, cardiovascular diseases, AIDS, etc. Various capability of data mining such as naïve Bayesian allotment, artificial neural networks, abutment vector machines, decision trees, logistic relapse, etc. have been used to advance models in healthcare exploration.

An predicted 17 million people die of cardiovascular diseases (CVD) every year [1]. Although such diseases are tractable, their early prognosis and a patient's appraise risk are necessary to curb the high mortality rates it instant. Common cardiovascular diseases combine coronary heart disease, cardiomyopathy, hypertensive heard disease, heart failure, etc. Common explanation of heart diseases combine smoking, diabetes, lack of physical enterprise, hypertension, high cholesterol diet, etc.

Research in the field of cardiovascular condition using data mining has been an growing effort involving indicator, treatment, and risk score reasoning with high levels of efficiency. Multiple CVD surveys have been control with the most arresting one being the data set from the Cleveland Heart Clinic. The Cleveland Heart Disease Database (CHDD) [2] as such has been treated the de facto database for heart disease exploration. Recommending the framework from this database, this paper introduce a framework to apply logistic relapse, support vector apparatus, and decision trees to attain individual forecasting which are in turn used in guideline based algorithms. The result of each rule from this organization is then

correlated on the basis of efficiency, sensitivity, and particularity.

The methodology aims to achieve of two goals: the first is to primarily present a anticipating framework for heart disease, and the second is to compare the adaptability of merging the reaction of numerous models as antithetical to using a single model.

2. Literature Survey

Prediction of heart disease using data mining approach has been an ongoing attempt for the past two decades. Most of the papers have implemented approach such as SVM, neural networks, regression, agreement trees, naïve Bayesian classifiers, etc. on numerous databases of patients from around the world.

One of the paltry on which the papers contradict are the selection of criterion on which the methods have been enforced. Many authors have stated different framework and databases for testing the accuracies. Xing et al. [3] attend a survey of 1000 patients, the results of which showed SVM to have 92.1% accuracy, unreal neural networks to have 91.0% and agreement trees with 89.6% using TNF, IL6, IL8, HICRP, MPO1, TNI2, sex, age, smoke, hypertension, diabetes, and continuity as the parameters. Similarly, Chen et al. [4] correlated the certainty of SVM, neural networks, Bayesian classification, agreement tree and logistic backsliding. Considering 102 cases, SVM had the highest efficiency of 90.5%, neural networks 88.9%, Bayesian 82.2%, agreement tree 77.9%, and logistic backsliding 73.9%.

Comparing the efficiency across multiple data sets with different framework arrives at disparate results which do not administer a just basis for connection. Realizing this, Soni et al. [5] listed most dominant parameters as gender, smoking, corpulent, alcohol intake, high salt diet, high sodden fat diet, action, sedentary lifestyle, hereditary, cholesterol, blood squeeze, fasting blood sugar, and heart rate. More freshly, Shouman et al. [6] cited the statistically analyze risk influence to be age, blood pressure, cholesterol, smoking, total cholesterol, diabetes, hypertension, hereditary, obesity, lack of physical enterprise. The same paper also conferred the Cleveland Heart Disease Database as the accepted database for heart disease analysis as it has been widely approved. As such, the CHDD has been employed for the method expected in this paper and details of the criterion it involves have been elaborated in further category.

Work on the CHDD can be dated since 1989, when Detrano et al. [7] used logistical backsliding to obtain 77% accuracy of prediction. The efficiency of different models on the CHDD have been tabulated in Table 1. A analogous graph of the same has been conferred in Figure 1.

What's noteworthy of these milestones is the advance in the efficiency when using hybrid techniques such as that in Polat et al. [8], Ozsen and Gunes [9], Das et al. [10], and Muhammed's [11] CLIP4 ensemble. The advance was also pointed out in the Shouman et al. [6], who also highlighted a necessity to management further exploration on hybrid techniques.

Table 1: Earlier methodologies and their accuracies applied to CHDD

Author (Ref #)	Technique	Accuracy
Detrano et al. [7]	Logistic regression	77%
Cheung [12]	C4.5 Naïve Bayes BNND BNNF	81.11% 81.48% 81.11% 80.96%
Polat et al. [13]	AIS	84.5%
Polat et al. [8]	Fuzzy-AIRS-KNN	87.0%
Ozsen and Gunes [9]	GA-AWAIS	87.0%
Tu et al. [14]	J4.8 Decision Tree Bagging Algorithm	78.9% 81.41%
Das et al. [10]	ANN ensembles	89.01%
Shouman et al. [15]	Nine Voting Equal Frequency Discretization Gain Ratio Decision Tree	84.1%
Muhammed [11]	CLIP3 CLIP4 CLIP4 ensemble	84.0% 86.1% 90.4%

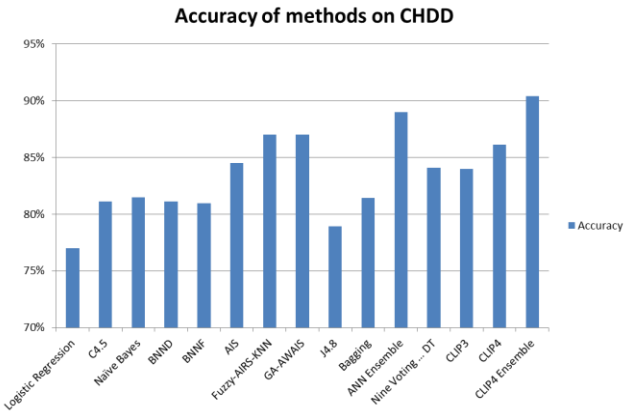

Figure 1: Performance Analysis

There is a need for forecasting based on more practical data mining models. A rule based access is a frequently used technique that associate the results of multiple models. Rule occupying models

such as C4.5 have been enforced before, but never as a combination of multiple anticipating models. As such, this paper presents a unique model to analogously study the utilization of rule based algorithms to combinations of SVM, agreement trees, and logistical backsliding.

For assessment of risk of heart disease using a sequence of models, this paper nominate the scheme shown in Figure 2 on the next page. This access is divided into six schedule involving preprocessing, training, testing with particular models, application of rules, and finally, contrast of results and the forecasting of heart disease. The schedule have been chronicle below.

3. Patient Database

Patient database is datasets possessed from Cleveland Heart Disease Dataset (CHDD) applicable on the UCI Repository [11]. The 13 attributes treated are age: age, sex, chest pain type, trestbps (resting blood pressure), chol (serum cholesterol in mg/dl), FBS (abstain blood sugar > 120 mg/dl), restecg (resting electrocardiographic results), thalach (maximal heart rate achieved), exang (exercise induced angina), oldpeak (ST depression convinced by activity analogous to rest), slope (the slope of the peak exercise ST division), and CA (number of major vessels (0-3) colored by fluoroscopy). There are a total of 303 patient records in the database.

3.1 Data Preprocessing

This phase combine abstraction of data from the Cleveland Heart Disease Dataset (CHDD) in a uniform composition. The step involves convert the data, which associate removal of missing fields, normalization of data, and deportation of outliers. Out of the 303 available records, 6 tuples have missing aspect. These have been eliminate from the data set. For SVM, data points were automatically centered at their mean and scaled to have unit accepted deviation. No adjustment need be made to the data sets for decision trees or logistic relapse.

3.2 Quality Management

It is very important to commence the quota of Data Quality most crucial to your institution. This is required to authorize a control for the condition of your data and to monitor the growth of your DQM action.

The other foundational comprise of the Data Quality Cycle convenient to Discover, Profile, Establish Rules, Monitor, Report, Remediate, and constantly improve Data Quality are construe in the next section.

3.3 Components of DQM

Once in place, these key ingredient grant robust, recyclable and highly effective DQM efficiency that can be leveraged across the company:

• Data Discovery: The process of conclusion, association, organizing and coverage metadata about your data (e.g., files/tables, record/row explanation, field/column definitions, keys)

• Data Profiling: The process of examine your data in detail, correlate the data to its metadata, attentive data statistics and reporting the allotment of condition for the data at a point in time

• Data Quality Rules: Based on the employment demand for each Data Quality allotment, the business and industrial rules that the data must discover to in order to be inspected of high quality

• Data Quality Monitoring: The ongoing check of Data Quality, based on the results of disqualify the Data Quality rules, and the relation of those results to defined error access, the creation an storage of Data Quality restriction and the generation of applicable proclamation

• Data Quality Reporting: The coverage, dashboards and yellow pages used to address and trend ongoing Data Quality allowance and to drill down into accurate Data Quality omission

• Data Remediation: The ongoing adjustment of Data Quality omission and concern as they are described

Each of these DQM components is describe in greater detail in terms of roles and responsibilities, course, automation and business assistance in the sections that follow.

3.4 Data Discovery

Roles and Responsibilities

Data discovery is generally the authority of IT. However, tech-savvy business users/managers may also perform data discovery when user-friendly data discovery tools are accessible.

Processes

Data discovery should be an computerized course using a robust data discovery tool. The data domains and physical database servers and/or file systems in quantity must first be identified, and read-only security acknowledgment to those database servers and/or file organization must be achieve in order to behead the discovery development.

The discovery tool will gather all of the applicable metadata and store it in a detection metadata archive where it can then be inquire and analyzed. The metadata grab typically catalogue database schema/file catalogue names, table/file names and definitions, column/field names and explanation, and any construe database or file accord (e.g., primary/foreign key relationships). Technologies

Results:-

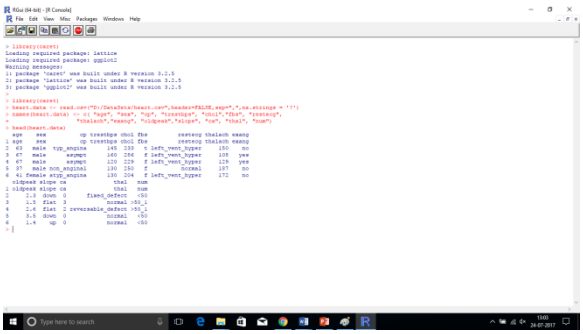

Figure 2 : Read Dataset

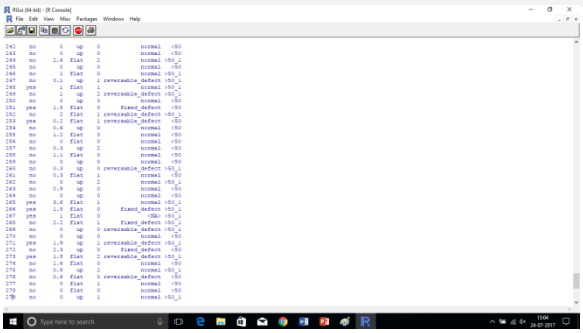

Figure 3: Actual Data

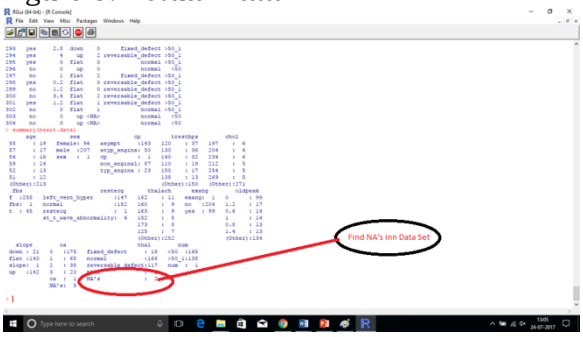

Figure 4: Summarized Data and finding the missing values

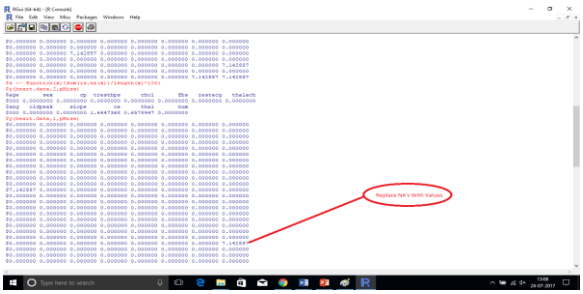

Figure 5: replaced by with their relevant values on the patient data by statistical approach.

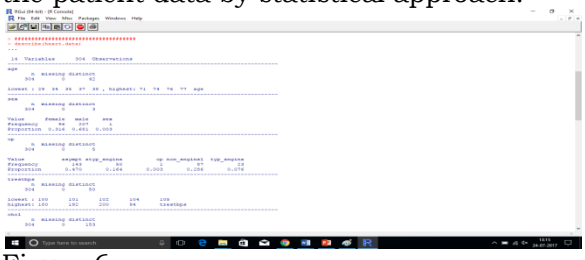

Figure 6:

describe <- function(...) {
 Hmisc::describe(...)
}

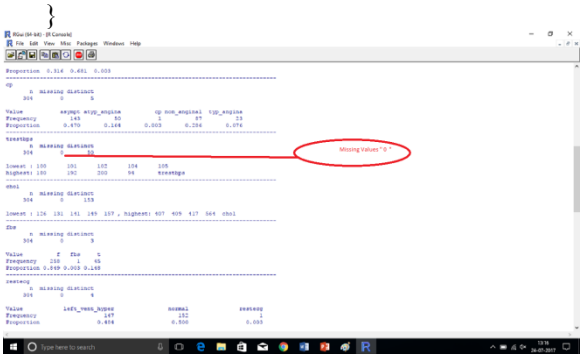

Figure 7 : replace the missing values.

Figure 8: Data Visualization

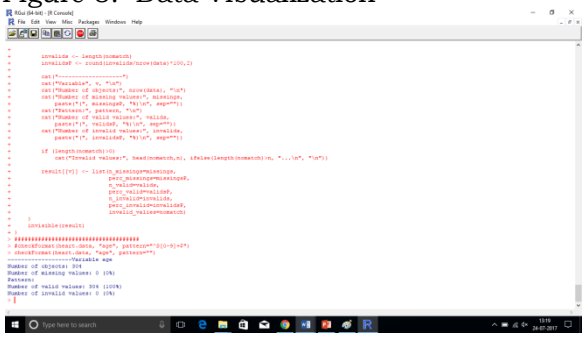

Figure 9: Accuracy (Authenticity) Check Data Format

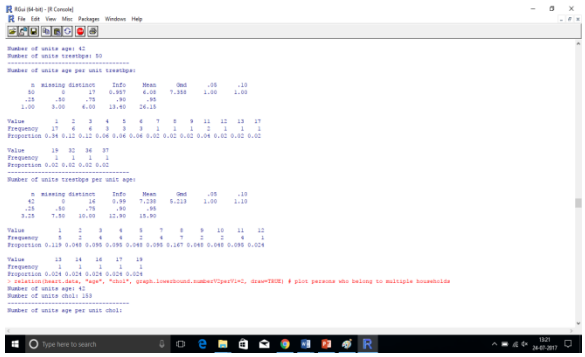

Figure 10: Inconsistent objects & 2.3 Dubious objects

#overview of 1-to-n and m-1 relations between two variables.

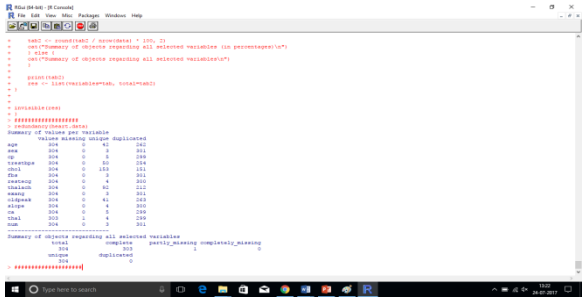

Figure 11. Result here by using my research we show that data was not qualified for the further analysis.

Preprocessing

3.5 Training the Models Each of the three models has been competent using different methods. For decision trees, a node disband criterion is compulsory. The best split is one that splits the data into

definite groups. Purity is a measure used to calculate a hidden split. A split that divides an aspect into two noticeable classes is the most pure.

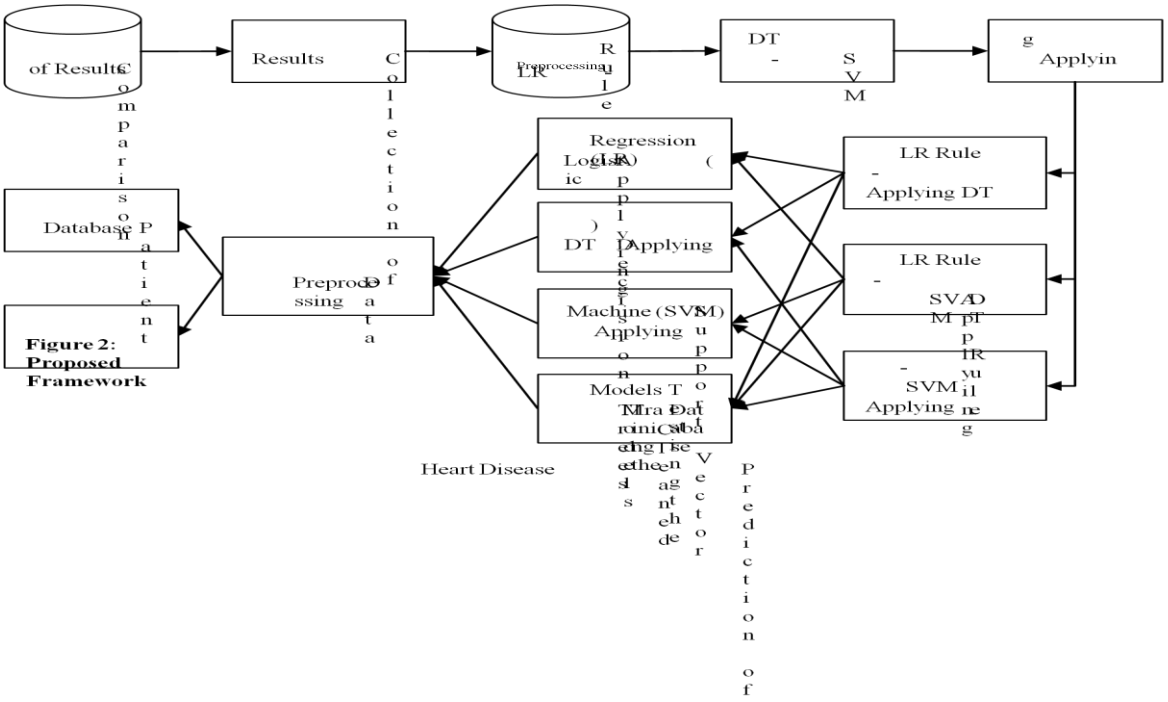

Figure 2: Proposed Framework

There are many different aberrancy criterions that can be used, such as the Gini Coefficient, or using statistical aberrance. Gini Coefficient is the most generally used splitting precedent which works on the population assortment of the aspect and thus splits it. As such, it is endorsed that it be used as the
splitting precedent for decision trees, admitting others may be used as well appear in different efficiency.
For support vector machines, agreeably the most adequate training

Preprocessing
Method is complete using K-fold cross acceptance, wherein the data can be trained set by carve it into k blocks and balance the results of the
blocks. This approach uses all the tuples to train the data, and then examination the data using one of the chunk. Generally, a
10-fold cross affirmation is used for training.
For logistic relapse, the first step to training is to find the significant characteristic by calculating their individual P-values. As a rule of thumb, if it is below 0.05, only then is the attribute important. The Hosmer-Lemeshow test is also appropriate to check for goodness fit of the miniature. The corresponding P-value must abide by a 5% level of implication in order to be a good fit model. Only after the model has been traditional is the data ready for testing.

3.6 Testing the Models
3.6.1 Support Vector Machine

A support vector machine is a type of model used to evaluate data and design patters in allotment and regression reasoning. Support vector machine (SVM) is used when your data has altogether two classes. An SVM analyze data by finding the best hyper plane that disconnect all data points of one class from those of the other class. The larger edge between the two classes, the better the copy is. A margin must have no points in its internal region. The agency vectors are the data points that on the confines of the margin. SVM is based on analytical functions and used to model complicated, and real world problems. SVM achieve well on data sets that have many aspect, such as the CHDD.

Support Vector Machines map the exercise data into kernel space. There are many individually used kernel spaces – linear (uses dot product), rectangular, polynomial, Radial Basis activity kernel, Multilayer perspiring kernel, etc. to name a few. In addition, there are numerous methods of achieve SVM, such as quadratic compute , sequential minimal optimization, and least squares. The impose aspect of SVM is kernel selection and approach selection such that your model is not over assured or depressed.
seeing that the CHDD has a large number of instances as well as appearances, it is arguable whether the kernel chosen is RBF or linear. Although the association between the aspect and class labels are nonlinear, due to the large number of features, RBF kernel may not improve

achievement. It is approved that both kernels be tested and the more efficient one be finally preferred.

3.7 Decision Trees

A decision tree is a tool that uses analysis or regression to predict a return to data. Classification is used when the features are arrange, and regression is used when the data is continued. Decision tree is one of the main data mining approach. A decision tree is made of a root node,division, and leaf nodes. To classify the data, follow the procedure from the root node to reach a leaf node.

Decision trees must be conceive using a purity index which will split the bump as discussed in the disciplines section. For the CHDD, each of the 297 tuples is check out down the decision tree and appear at a positive or negative decision for heart disease. These are related to the original decision limitation in the CHDD to check for false positives or false negatively giving us the accuracy, particularity , and sensitivity of the model. The splitting criterion used is also suggestive of the importance of each attribute.

3.7.1 Logistic Regression

Logistic relapse is a type of regression analysis in statistics used for indicator of outcome of a absolute dependent fluctuating (a dependent variable that can take a limited number of values) from a set of predictor or separate variables. In logistic regression the dependent variable is always binary (with two categories). Logistic lapse is mainly used to for prediction and also considerate the probability of success. Logistic backsliding involves fitting an comparison of the form to the data:

Y = ß0 + ß1x1 + ß2x2 + … + ßnxn – eq. 1

The regression concerto are usually predicted using maximum likelihood evaluation. The maximum likelihood ratio helps to determine the statistical implication of independent variables on the reliant variables. The likelihood-ratio tests determine the contribution of individual diviner (independent variables). Then the probability (p) of each case is determined using odds ratio, P/(1-P) = eY – eq. 2

From this p–value is found out. This gives the chance or chance for the individual to have cardiac arrest heart disease.

3.8 Rule Based Algorithm

Rule based systems are actually decision trees that use a small number of aspect for decision making. These are simple systems which are usually used to increase awareness of knowledge patters. Rule based algorithms are suggestive of trends in the features they contemplate and thus provides us with logical conclusion

Rules are used to support managerial in classification, regression, and association tasks. Depending on the data, there are disparate types of rules that can be achieve such as classical proposition logic (C-rules), association rules (Arules), fuzzy logic (F-rules), M-of-N or verge rules (Trules), similarity or model -based rules (P-rules). It is approved that classification rule (C-rule) be used for this model. C-Rules are of the form of if-else extent, and provide the simplest and most intelligible way of expressing knowledge. These rules will detail the result of each of the individual approach based on weight of the model which is reliant on the accuracy, particularity and sensitivity complete. It is consider that a result with higher awareness and particularity but lower accuracy will be complete from the results of this model which is in itself, a highly efficient model.

3.9 Comparison of Results

The results obtained after exercise the C-rule will be analyzed on the basis of sense, specificity, and accuracy. From these,closure to the most effective model, the efficacy of conjoint modes and the final accuracy of the comprehensive model can be drawn

Results:-

Figure 13: shows how many had heart attack, women or men, age? Values of num > 0 are cases of heart disease

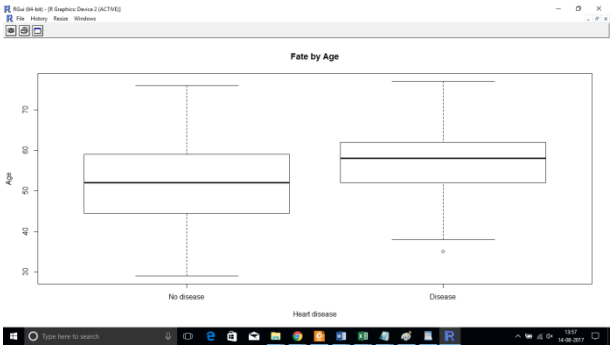

Figure 14: Fate by Age

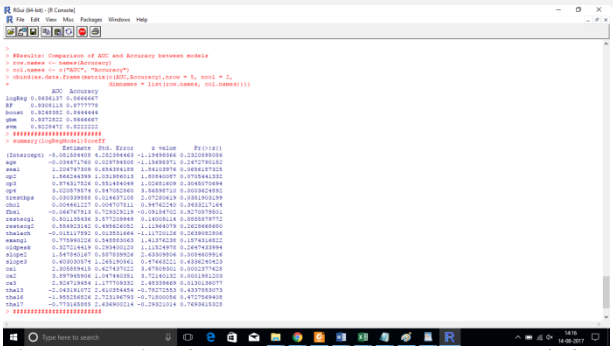

Figure 15: **logistic regression model and importance of variables from boosted tree**

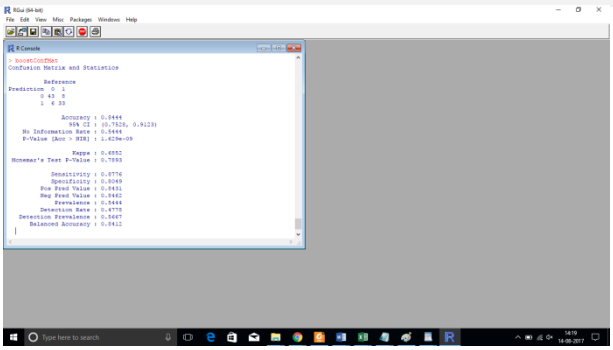

Figure 19: Boosted Tree matrix

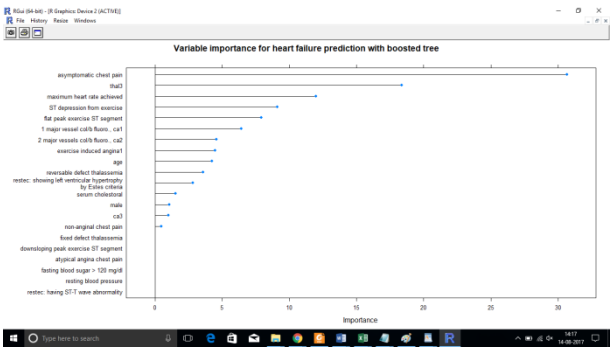

Figure 16: heart failure prediction using boosted tree

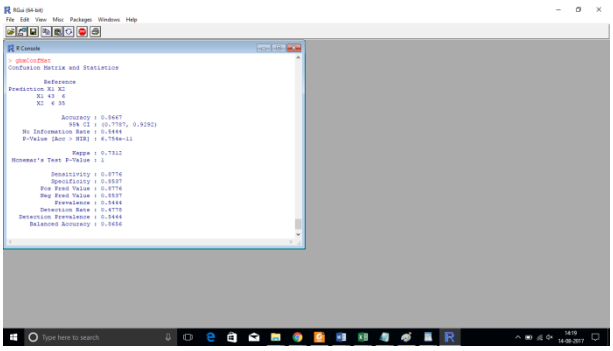

Figure 20: Confusion matrix and statistics

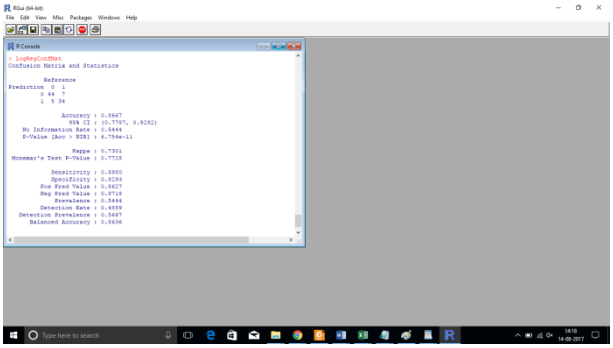

Figure 17: Logistic Regression matrix

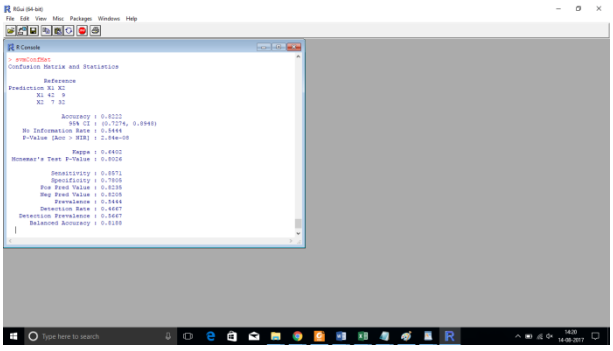

Figure 21: SVM matrix

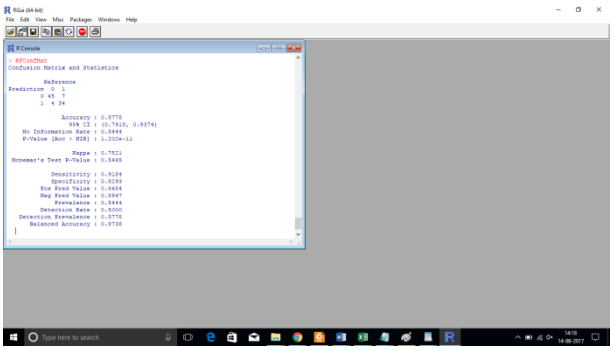

Figure 18: Random Forest matrix

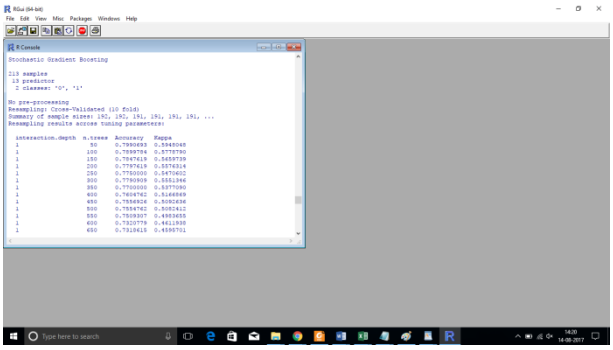

Figure 22: Stochastic gradient boosted tree

Method	AUC	Accuracy
logReg	0.9161585	0.8651685
RF	0.8953252	0.8089888
boost	0.9095528	0.8426966
gbm	0.9070122	0.8426966
svm	0.882622	0.7977528

Table: 2 Performance Analysis

4. Conclusion

In conclusion, as identified through the literature review, there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of cardiovascular diseases.

This paper proposes a framework using combinations of support vector machines, logistic regression, and decision trees to arrive at an accurate prediction of heart disease. Using the Cleveland Heart Disease database, this paper provides guidelines to train and test the system and thus attain the most efficient model of the multiple rule based combinations. Further, this paper proposes a comparative study of the multiple results, which include sensitivity, specificity, and accuracy. In addition, the most effective and most weighed model can be found. Further work involves development of the system using the mentioned methodologies and thus training and testing the system.

Acknowledgment

I would like to thank Dr.B.D.C.N Prasad for his guidance and support for preparing this paper.

5. References

[1] Mackay,J., Mensah,G. 2004 "Atlas of Heart Disease and Stroke" Nonserial Publication, ISBN-13 9789241562768 ISBN-10 9241562765.

[2] Robert Detrano 1989 "Cleveland Heart Disease Database" V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

[3] Yanwei Xing, Jie Wang and Zhihong Zhao Yonghong Gao 2007 "Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease" Convergence Information Technology, 2007. International Conference November 2007, pp 868-872.

[4] Jianxin Chen, Guangcheng Xi, Yanwei Xing, Jing Chen, and Jie Wang 2007 "Predicting Syndrome by NEI Specifications: A Comparison of Five Data Mining Algorithms in Coronary Heart Disease" Life System Modeling and Simulation Lecture Notes in Computer Science, pp 129-135.

[5] Jyoti Soni, Ujma Ansari, Dipesh Sharma 2011 "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" International Journal of Computer Applications, doi 10.5120/22372860.

[6] Mai Shouman, Tim Turner, Rob Stocker 2012 "Using Data Mining Techniques In Heart Disease Diagnoses And Treatment" Electronics, Communications and Computers (JECECC), 2012 Japan-Egypt Conference March 2012, pp 173-177.

[7] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, Victor Froelicher 1989 "International application of a new probability algorithm for the diagnosis of coronary artery disease" The American Journal of Cardiology, pp 304-310.15

[8] Polat, K., S. Sahan, and S. Gunes 2007 "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing" Expert Systems with Applications 2007, pp 625-631.

[9] Ozsen, S., Gunes, S. 2009 "Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems" Expert Systems with Applications, pp 386-392.

[10] Resul Das, Ibrahim Turkoglub, and Abdulkadir Sengurb

2009 “Effective diagnosis of heart disease through neural networks ensembles” Expert Systems with Applications, pp 7675–7680.

[11] Lamia Abed Noor Muhammed 2012 “Using Data Mining technique to diagnosis heart disease” Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on March 2012, pp 173-177.

[12] Cheung, N 2001 “Machine learning techniques for medical analysis” School of Information Technology and Electrical Engineering, B.Sc. Thesis, University of Queensland

[13] Polat, K., Sahan, S., Kodaz, H., Günes, S. 2005 “A new classification method to diagnose heart disease: Supervised artificial immune system”. In Proceedings of the Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN), pp 2186-2193.

[14] My Chau Tu, Dongil Shin, Dongkyoo Shin 2009 “Effective Diagnosis of Heart Disease through Bagging Approach” Biomedical Engineering and Informatics, 2009. BMEI '09. 2nd International Conference, pp 1- 4.

[15] Shouman, M., Turner, T. and Stocker.R 2011 “Using Decision Tree for Diagnosing Heart Disease Patients" Australasian Data Mining Conference (AusDM 11) Ballarat 2011, pp 23-30.

[16] ABC: A knowledge Based Collaborative framework for e-health

Flora Amato; Giovanni Cozzolino; Alessandro Maisto; Antonino Mazzeo; Vincenzo Moscato; Serena Pelosi; Antonio Picariello; Sara Romano; Carlo Sansone

2015 IEEE 1st International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)
Year: 2015
Pages: 258 - 263, DOI: 10.1109/RTSI.2015.7325107
IEEE Conference Publications

[17] A Novel Storage Architecture for Facilitating Efficient Analytics of Health Informatics Big Data in Cloud
Manish Kumar Pandey; Karthikeyan Subbiah
2016 IEEE International Conference on Computer and Information Technology (CIT)
Year: 2016
Pages: 578 - 585, DOI: 10.1109/CIT.2016.86
IEEE Conference Publications

[18] Hospital occupation rate analysis of the Brazilian federal university hospitals through business intelligence
Carlo Alessandro Melo Noce; Georges Daniel Amvame Nze; Lourdes Mattos Brasil
2017 12th Iberian Conference on Information Systems and Technologies (CISTI)
Year: 2017
Pages: 1 - 6, DOI: 10.23919/CISTI.2017.7975882