Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability

K.R. Lakshmi [*], M.Veera Krishna[**] and S.Prem Kumar[***]

[*] Director, IERDS, Maddur Nagar, Kurnool, Andhra Pradesh, India
[**] Department of Mathematics, Rayalaseema University, Kurnool, Andhra Pradesh, India
[***] Professor & Head, Department of CSE & IT, G. Pullaiah College of Engineering & Technology, Nandikotkur Road, Kurnool, Andhra Pradesh, India.

Abstract- The diagnosis of heart disease is a significant and tedious task in medicine. The healthcare environment is generally perceived as being 'information rich' yet 'knowledge poor'. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. Using medical profile such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. It can serve a training tool to train nurses and medical students to diagnose patients with heart disease. It is a web based user friendly system and can be used in hospitals if they have a data ware house for their hospital. Presently we are analyzing the performances of the ten classification data mining techniques by using various performance measures. For implementation of the work a real time patient database is taken and the patient records are experimented and the final best classifier is identified with quick response time and least error rate. A typical confusion matrix is furthermore displayed for quick check. The study describes algorithmic discussion of the heart disease dataset from Cleveland Heart Disease database, on line repository of large datasets. The Best results are achieved by using Tanagra tool. Tanagra is data mining matching set. The accuracy is calculate based on addition of true positive and true negative followed by the division of all possibilities.

Index Terms- Data mining techniques, SVM, LDA, C4.5, *k*-NN, BLR, MLR, PLS-LDA, *k*-means, EMC and Apriori, Sensitivity and Specificity

I. INTRODUCTION

Knowledge discovery in databases is well-defined process consisting of several distinct steps. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. A formal definition of Knowledge discovery in databases is given as follows: "Data mining is the non trivial extraction of implicit previously unknown and potentially useful information about data". Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. Anticipating patient's future behavior on the given history is one of the important applications of data mining techniques that can be used in health care management. A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Health care data is massive. It includes patient centric data, resource management data and transformed data. Health care organizations must have ability to analyze data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care. The availability of integrated information via the huge patient repositories, there is a shift in the perception of clinicians, patients and payers from qualitative visualization of clinical data by demanding a more quantitative assessment of information with the supporting of all clinical and imaging data. For instance it might now be possible for the physicians to compare diagnostic information of various patients with identical conditions. Likewise, physicians can also confirm their findings with the conformity of other physicians dealing with an identical case from all over the world. Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Wu, et al proposed that integration of clinical decision support with computer based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modelling and analysis tools, e.g.,

data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions. The term heart disease applies to a number of illnesses that affect the circulatory system, which consists of heart and blood vessels. It is intended to deal only with the condition commonly called "Heart Attack" and the factors, which lead to such condition. Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term - cardiovascular disease "includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardiovascular disease (CVD) results in severe illness, disability, and death. Narrowing of the coronary arteries results in the reduction of blood and oxygen supply to the heart and leads to the Coronary heart disease (CHD). Myocardial infarctions, generally known as a heart attacks, and angina pectoris, or chest pain are encompassed in the CHD. A sudden blockage of a coronary artery, generally due to a blood clot results in a heart attack. Chest pains arise when the blood received by the heart muscles is inadequate. High blood pressure, coronary artery disease, alular heart disease, stroke, or rheumatic fever/rheumatic heart disease are the various forms of cardiovascular disease.

Early Signs Of Heart Disease: 1. Dizzy spell or fainting fits. 2. Discomfort following meals, especially if long continued. 3. Shortness of breath, after slight exertion. 4. Fatigue without otherwise explained origin. 5. Pain or tightness in the chest a common sign of coronary insufficiency is usually constrictive in nature and is located behind the chest bone with 6. Radiation into the arms or a sense of numbness or a severe pain in the centre of the chest. 7. Palpitation

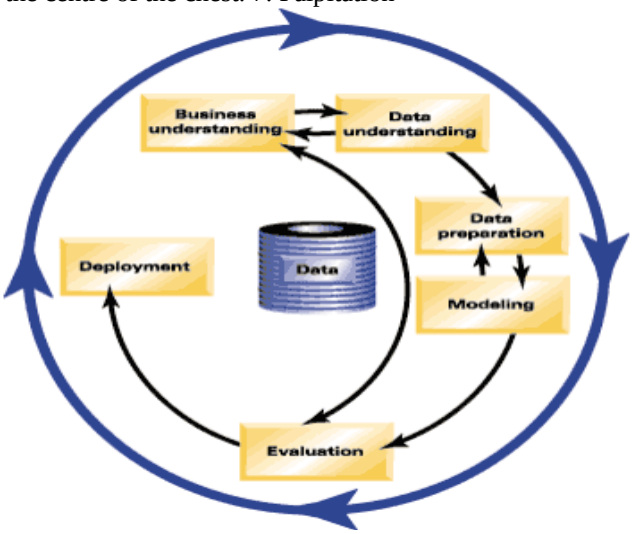

Figure 1: Structure of DM

Data mining is an essential step of knowledge discovery. In recent years it has attracted great deal of interest in Information industry [4]. Knowledge discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. In particulars, data mining may accomplish class description, association, classification, clustering, prediction and time series analysis. Data mining in contrast to traditional data analysis is discovery driven. The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term "cardiovascular disease" includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardio vascular disease results in severe illness, disability, and death. Narrowing of the coronary arteries results in the reduction of blood and oxygen supply to the heart and leads to the Coronary heart disease.A sudden blockage of a coronary artery, generally due to a blood clot results in a heart attack. Chest pains arise when the blood received by the heart muscles is inadequate and inductive logic programming.

II. Review of the Related Literature

The Heart Disease Data Prediction is designed to support clinicians in their diagnosis for heart disease prediction. They typically work through an analysis of medical data and a knowledge base of clinical expertise. The quality of medical diagnostic decisions for heart disease can be increased by improvements to these Predicting systems [4]. Data mining provides a way to get the information buried in the data. Numerous experiments were conducted on linear and nonlinear characteristics of HRV (Heart Rate Variability) indices to assess several classifiers, e.g., Bayesian classifiers [5], CMAR (Classification based on Multiple Association Rules) [10], C4.5 (Decision Tree) [6] and SVM (Support Vector Machine) [2]. SVM surmounted the other classifiers. The problem of identifying constrained association rules for heart disease prediction was studied by Carlos Ordonez [1]. The assessed data set encompassed medical records of people having heart disease with attributes for risk factors, heart perfusion measurements and artery narrowing. Three constraints were introduced to decrease the number of patterns. First one necessitates the attributes to appear on only one side of the rule. The second one segregates attributes into uninteresting groups. The ultimate constraint restricts the number of attributes in a rule. Experiments illustrated that the constraints reduced the number of discovered rules remarkably besides decreasing the running time. Two groups of rules envisaged the presence or absence of heart disease in four specific heart arteries. Data mining methods may aid the clinicians in the prediction of the survival of patients and in the adaptation of the practices consequently. The work of Franck

Le Duff et al. [3] might be executed for each medical procedure or medical problem and it would be feasible to build a decision tree rapidly with the data of a service or a physician. Comparison of traditional analysis and data mining analysis illustrated the contribution of the data mining method in the sorting of variables and concluded the significance or the effect of the data and variables on the condition of the study.The main drawback of the process was knowledge acquisition and the need to collect adequate data to create an appropriate model. In [8] Latha Parthiban et al. projected an approach on basis of coactive neuro-fuzzy inference system (CANFIS) for prediction of heart disease. L. Goodwin et.al [9] discussed different Data mining issues and opportunities for building nursing knowledge. In Lei Yu and Huan Liu [10] introduced a novel concept, predominant correlation, and proposed a fastfilter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. The efficiency and effectiveness of their method is demonstrated through extensive comparisons with other methods using real world data of high dimensionality. A methodology for comparing classification methods through the assessment of model stability and validity in variable selection was proposed by J. Shreve et.al [7]. This study provides a systematic design for comparing the performance of six classification methods using Monte Carlo simulations and illustrates that the variable selection process is integral in comparing methodologies to ensure minimal bias, enhanced stability, and optimize performance. They quantify the variable selection bias and show that, for sufficiently large samples, this bias is minimized so that methods can be compared. Later John peter and Somasundaram [11] discussed the hybrid attribute selection method combining CFS and Filter Subset Evaluation gives better accuracy for classification. Mohammad Taha Khan [12] discussed a prototype model for the breast cancer as well as heart disease prediction using data mining techniques is presented. The data used is the Public-Use Data available on web, consisting of 909 records for heart disease and 699 for breast cancer. Two decision tree algorithms C4.5 and the C5.0 have been used on these datasets for prediction and performance of both algorithms is compared. Adithya Sundar et.al [13] discussed a prototype using data mining techniques, namely Naïve Bayes and WAC (weighted associative classifier). Srinivas et.al [14] briefly examine the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data. Jaya Rama Krishnaiah [15] discussed how the data classification is based on supervised machine learning algorithms which result in accuracy, time taken to build the algorithm. Tanagra tool is used to classify the data and the data is evaluated using entropy based cross validations and partitioned techniques and the results are compared. In this paper we studied describes algorithmic discussion of the heart disease dataset from Cleveland Heart Disease database, on line repository of large datasets. The Best results are achieved by using Tanagra tool.

III. METHODOLOGY

We describe some Classification techniques. Classification is a data mining (machine learning) technique used to predict group membership for data instances. Classification analysis is the organization of data in given class. These approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. Many classification models are used to classify new objects.

3.1 LDA

Linear Discriminant Analysis (LDA) is a supervised learning algorithm. LDA methods are used in statistics, pattern recognition and machine learning to find a linear combination of features. The idea behind LDA is simple, for each class to be identified, calculate linear function of the attributes. The class function having highest score is treated as the predicted class. It is a statistical classification algorithm which is used to classify the values based on the linear combination among values. Linear Discriminant Analysis perfectly handles the data within class frequencies are unequal. LDA also evaluates the performances for randomly generated test data. The LDA Algorithm maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability. The use of Linear Discriminant Analysis for data classification is widely used to classify many biological data sets such cancer, colon cancer, HIV analysis etc. In LDA based classification the data sets can be transformed and test vectors can be classified in two different approaches.

Transformation with class dependency: This type of approach involves maximizing the ratio of between class variance to within class variance. The main objective is to maximize this ratio so that adequate class separability is obtained. The class-specific type approach involves using two optimizing criteria for transforming the data sets independently.

Transformation with class independency: This approach involves maximizing the ratio of overall variance to within class variance. This approach uses only one optimizing criterion to transform the data sets and hence all data points irrespective of their class identity are transformed using this transform. In this type of LDA, each class is considered as a separate class against all other classes.

3.2 SVM

Support Vector Machine is a type of classification method, which estimates the classification function. SVM is a set of related supervised learning methods that analyze data and recognize patterns, used for classification. Support Vector Machine (**SVM**) is a non-linear classifier method which is often reported as producing better classification results compared to other methods. The main idea of SVM is to construct a hyper plane as a decision surface in such a way that the margin of separation between positive and negative examples is maximized. This process non-linearly map the input sample data to some high dimensional space, where the data can be linearly separated, thus providing higher classification (or regression) accuracy. SVMs are rather interesting in that they enjoy both a sound theoretical basis as well as state-of-the-art success in real-world applications, especially in Bioinformatics.

3.3 C4.5

C4.5 algorithm is a greedy algorithm developed by Ross Quinlan, used for the induction of decision trees. C4.5 is a successor of ID3 algorithm. The decision trees generated by C4.5 adopt greedy approach in which decision trees are constructed in top-down recursive divide-and-conquer manner. C4.5 is often referred to as a statistical classifier. Like ID3, C4.5 builds decision trees from training data

set, using the concept of information entropy. The decision tree algorithm C4.5 is developed from ID3 in the following ways: Handling missing data, handling continuous data, and pruning, generating rules, and splitting. For splitting purpose, C4.5 uses the Gain Ratio instead of Information Gain. C4.5 algorithm uses an attribute selection measure to select the attribute tested for each non leaf node in the tree. The highest normalized information gain attribute is chosen to make the decision.

$$\text{Gain Ratio (D, S)} = \text{Gain (D, S)/ Split INFO}$$

$$\text{Where, Split INFO} = -\left(\sum_{i=1}^{s} \frac{D_i}{D} \log_2 \frac{D_i}{D}\right)$$

3.4 *k*-NN :

It is the nearest neighbour algorithm. The *k*-nearest neighbour's algorithm is a technique for classifying objects based on the next training data in the feature space. It is among simplest of all mechanism learning algorithms [30]. The algorithm operates on a set of d-dimensional vectors, $D = \{\mathbf{x_i} \mid i = 1. . . N\}$, where $\mathbf{x_i} \in k^d$ denotes the *i* th data point. The algorithm is initialized by selection *k* points in k^d as theinitial *k* cluster representatives or "centroids". Techniques for select these primary seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data *k* times. Then the algorithm iterates between two steps till junction:

Step 1: Data Assignment each data point is assign to its adjoining centroid, with ties broken arbitrarily. This results in a partitioning of the data.

Step 2: Relocation of "means". Each group representative is relocating to the center (mean) of all data points assign to it. If the data points come with a possibility measure (Weights), then the relocation is to the expectations (weighted mean) of the data partitions.

"Kernelize" *k*-means though margins between clusters are still linear in the embedded high-dimensional space, they can become non-linear when projected back to the original space, thus allowing kernel *k*-means to deal with more complex clusters. The *k*-medoid algorithm is similar to *k*-means except that the centroids have to belong to the data set being clustered. Fuzzy c-means is also similar, except that it computes fuzzy membership functions for each clusters rather than a hard one.

3.5 BLR:

Predictive analysis in health care primarily to determine which patients are at risk of developing certain conditions, like diabetes, asthma, heart disease and other lifetime illnesses. Additionally, sophisticated clinical decision support systems incorporate predictive analytics to support medical decision making at the point of care. Logistic regression is a generalization of linear regression. It is used primarily for predicting binary or multi-class dependent variables.

3.6 Multinomial Logistic Regression (MLR):

A multinomial logit (MNL) model, also known as multinomial logistic regression, is a regression model which generalizes logistic regression by allowing more than two discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable given a set of independent variables (which may be real-valued, binary-valued, categorical-valued, etc.). An extension of the binary logistic model cases where the dependent variable has more than two categories is the multinomial logistic Regression. In such cases collapsing the data into two categories not make good sense or lead to loss in the richness of the data. The multinomial legit model is the appropriate technique in these cases, especially when the dependent variable categories are not ordered. Multinomial regression to include feature selection/importance methods.

3.7 PLS-DA

PLS Regression for Classification Task PLS (Partial Least Squares Regression) Regression can be viewed as a multivariate regression framework where to predict the values of several PLS-LDA (Partial Least squares-Linear Discriminant Analysis target variables (Y_1, Y_2 …) from the values of several input variables (X_1,X_2,…). The algorithm use three axis for the diabetes disease is the following: The components of X are used to predict the scores on the Y components, and the predicted Y component scores are used to predict the actual values of the Y variables. In constructing the principal components of X, the PLS algorithm iteratively maximizes the strength of the relation of successive pairs of X and Y component scores by maximizing the covariance of each X-score with the Y variables. The PLS Regression is initially defined for the prediction of continuous target variable. But it seems it can be useful in the supervised learning problem where we want to predict the values of discrete attributes. In this tutorial we propose a few variants of PLS Regression adapted to the prediction of discrete variable. The generic name "PLS-DA" (Partial Least Square Discriminant Analysis) is often used in the literature. To predict the values of the dependent variable for unseen instances (or unlabeled instances) from the observed values on the independent variables. The process is rather basic if handle a linear regression model. Apply the computed parameters on the unseen instances.

3.8 The *k*-means algorithm:

The k-means algorithm is a simple iterative method to partition a given dataset into a serspecified number of clusters, *k*. This algorithm has been discovered by several researchers across different disciplines. The algorithm operates on a set of *d*-dimensional vectors, $D = \{\mathbf{x_i} \mid i = 1, . . . , N\}$, where $\mathbf{x_i} \in R^d$ denotes the *i*th data point. The algorithm is initialized by picking *k* points in R^d as the initial *k* cluster representatives or "centroids". Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data *k* times. Then the algorithm iterates between two steps till convergence:

Step 1: Data Assignment. Each data point is assigned to its closest centroid, with ties broken arbitrarily. This results in a partitioning of the data.

Step 2: Relocation of "means". Each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a probability measure (weights), then the relocation is to the expectations (weighted mean) of the data partitions.

3.9 Entropy based Mean Clustering (EMC) algorithm:

The Entropy based Mean Clustering algorithm (EMC) is extension to the K mean algorithm, reduces the number of iterations during the clustering process. It works on three phases. In the first phase it computes the min points of the each seed (element or item) in the data set and then arranges the seed elements in the order of their seed entropy (For example 1-10,2-5,3-9,4-6,5-1, then it arranges the data as 1,3,4,2,5 .i.e. data arranged ascending order of the entropy). In the second phase, it makes the candidate set, this candidate set is unique in nature, i.e it does not consisting of duplicated elements. In the third phase the clustering was applied on the Euclidian distances, and remaining elements, which were not in candidate sets were placed in according to the native elements, were resided.

3.10 The Apriori algorithm

One of the most popular data mining approaches is to find frequent itemsets from a transaction dataset and derive association rules. Finding frequent itemsets (itemsets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori is a seminal algorithm for finding frequent itemsets using candidate generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of itemsets, "if an itemset is not frequent, any of its superset is never frequent". By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. Let the set of frequent itemsets of size k be F_k and their candidates be C_k . Apriori first scans the database and searches for frequent itemsets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent itemsets.

1. Generate C_{k+1}, candidates of frequent itemsets of size k +1, from the frequent itemsets of size k.
2. Scan the database and calculate the support of each candidate of frequent itemsets.
3. Add those itemsets that satisfies the minimum support requirement to F_{k+1}.

Function apriori generates C_{k+1} from F_k in the following two step process:

1. Join step: Generate R_{K+1}, the initial candidates of frequent itemsets of size k + 1 by taking the union of the two frequent itemsets of size k, Pk and Qk that have the first k−1 elements in common.

$R_{K+1} = P_k \cup Q_k = \{item_1 , item_2, \ldots , item_{k-1}, item_k , item_{k'} \}$

$P_k = \{item_1 , item_2, \ldots , item_{k-1}, item_k \}$

$Qk = \{item_1 , item_2, \ldots , item_{k-1}, item_{k'} \}$

where, $item_1 < item_2 < \cdots < itemk < itemk'$_ .

2. Prune step: Check if all the itemsets of size k in R_{k+1} are frequent and generate C_{k+1} by removing those that do not pass this requirement from R_{k+1}. This is because any subset of size k of C_{k+1} that is not frequent cannot be a subset of a frequent itemset of size k + 1. Function subset finds all the candidates of the frequent item sets included in transaction t. Apriori, then, calculates frequency only for those candidates generated this way by scanning the database. It is evident that Apriori scans the database at most $k_{max}+1$ times when the maximum size of frequent itemsets is set at k_{max}.

IV. Data Analysis

A total of 2268 records with 15 medical attributes (factors) were obtained from the Cleveland Heart Disease database. The records were split equally into two datasets: training dataset (1857 records) and testing dataset (411 records). The attribute "Diagnosis" was identified as the predictable attribute with value "1" for patients with heart disease and value "0" for patients with no heart disease.

4.1 Predictable attribute 1. Diagnosis (value 0: < 50% diameter narrowing (no heart disease); value 1: > 50% diameter narrowing (has heart disease)) **Key attribute**

1. Patientid – Patient"s identification number

Input attributes 1. Sex (value 1: Male; value 0 : Female)

2. Chest Pain Type (value 1: typical type 1 angina, value2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value2: showing probable or definite left ventricular hypertrophy)
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment (value1: unsloping; value 2: flat; value 3: downsloping)
7. CA – number of major vessels colored by fluoroscopy (value 0 – 3)
8.Thal (value 3: normal; value 6: fixed defect; value7:reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach- maximum heart rate achieved
12. Oldpeak – ST depression induced by exercise relative to rest
13. Age in Year.

V. COMPUTATIONAL RESULTS AND DISCUSSION

The basic phenomenon used to classify the Heart disease classification using classifier is its performance and accuracy. The performance of a chosen classifier is validated based on error rate and computation time. The classification accuracy is predicted in terms of Sensitivity and Specificity. The computation time is noted for each classifier is taken in to account. Classification Matrix displays the frequency of correct and incorrect predictions. It compares the actual values in the test dataset with the predicted values in the trained model. In this example, the test dataset contained 208 patients with heart disease and 246 patients without heart disease.

Predicted	**Classified as Healthy (0)**	**Classified as not Healthy (1)**
Actual Healthy (0)	TP	FN
Actual not Healthy (1)	FP	TN

Table 1: confusion matrix

Table. 1 shows the results of the Classification Matrix for all the ten models. The rows represent predicted values while the columns represent actual values (1 for patients with heart disease, 0 for patients with no heart disease). The left-most columns show values predicted by the models. The diagonal values show correct predictions. For Classification, this work constructed Confusion Matrix for the frequency of correct and incorrect predictions. From the confusion matrix, the Specificity, Sensitivity, Accuracy Rate and Error rate have been calculated. For measuring accuracy rate and Error Rate, the following mathematical model is used.

$$\text{Sensitivity (Recall)} = \frac{TP}{TP+FN}, \quad \text{Specificity} = \frac{TN}{FP+TN}, \quad \text{Accuracy} = \frac{TP+FN}{TP+FP+TN+FN}$$

$$\text{Positive Precision} = \frac{FP}{TP+FP}, \quad \text{Negative Precision} = \frac{FN}{TN+FN}, \quad \text{CVError Rate} = \frac{FP+FN}{TP+FP+TN+FN}$$

The software framework of this work has been developed with Tanagra tool. Tanagra is a data mining suite build around graphical user interface. Tanagra is particularly strong in statistics, offering a wide range of uni and multivariate parametric and nonparametric tests. Equally impressive is its list of feature selection techniques. Together with a compilation of standard machine learning techniques, it also includes correspondence analysis, principal component analysis, and the partial least squares methods. Tanagra is more powerful, it contains some supervised learning but also other paradigms such as clustering, supervised learning, meta supervised learning, feature selection, data visualization supervised learning assessment, statistics, feature selection and construction algorithms. The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development in this domain , and allowing to analyze either real or synthetic data. Tanagra can be considered as a pedagogical tool for learning programming techniques. Tanagra is a wide set of data sources, direct access to data warehouses and databases, data cleansing, interactive utilization. The Entropy based mean Clustering is developed using Advanced Java.

5.1. Performance study of algorithms

The table 2 consists of values of different classification. According to these values the lowest computing time (<590ms) can be determined.

S. No.	Alg.	CT (ms)	TP	FN	FP	TN	Acc. (%)	Spe.	Sen.	CVE rate	P (Prec)	N (Prec)	BVE rate
1	SVM	516	25	34	56	296	78.10	0.8409	0.4237	0.2189	0.6913	0.1030	0.3065
2	LDA	534	6	53	12	340	84.18	0.9659	0.1016	0.1581	0.6667	0.1348	0.3125
3	C4.5	488	14	45	18	334	84.68	0.9488	0.2372	0.1532	0.5628	0.1187	0.2985
4	*k*-NN	582	22	42	24	323	83.95	0.9308	0.3438	0.1605	0.5217	0.1151	0.3221
5	BLR	496	8	49	21	333	82.96	0.9406	0.1403	0.1703	0.7241	0.1283	0.2658
6	MLR	498	15	35	40	321	81.75	0.8892	0.3000	0.1825	0.7273	0.0983	0.2856
7	PLS-DA	432	14	50	7	340	86.13	0.9798	0.2187	0.1386	0.3330	0.1282	0.2558
8	*k*-mean	504	16	37	44	314	80.29	0.8771	0.3019	0.1971	0.7333	0.1054	0.2748
9	EMC	449	9	30	31	341	85.16	0.9166	0.2307	0.1484	0.7756	0.0808	0.2685
10.	Apriori	496	8	32	53	314	80.53	0.8858	0.1162	0.1946	0.8936	0.1043	0.2749

Table 2: Comparison of supervised Algorithms based on performance

Alg-Algorithm names, CT- Computing Time, TP-True Positive,FN-False Negative, FP-False Positive, TN True Negative, Acc-Accuracy, Spec-Specificty, Sen-Senitivity, CVE rate-CrossValidation Error rate, P(Prec)-Positive Precision, N(Prec)-Negative Precision, BVE rate-Bootstrap Validation Error rate.

SVM, LDA, C4.5, *k*-NN, BLR, MLR, PLS-DA, *k*-means, EMC and Apriori in a lowest computing time that we have experimented with a dataset. A distinguished confusion matrix was obtained to calculate sensitivity, specificity and accuracy. Confusion matrix is a matrix representation of the classification results. From the confusion matrix to analyze the performance criterion for the classifiers in disease detection accuracy, precision, recall have been computed for all datasets. Accuracy is the percentage of predictions that are correct. The precision is the measure of accuracy provided that a specific class has been predicted. Recall is the percentage of positive labelled instances that were predicted as positive.

Step 1: The ten algorithms can be filtered by using lowest computing time (<580ms). The ten can be reduced nine algorithms namely (SVM, C4.5, *k*-NN, BLR, MLR, PLS-DA, *k*-means, EMC and Apriori).

S. No.	Alg.	CT (ms)	TP	FN	FP	TN	Acc. (%)	Spe.	Sen.	CVE rate	P (Prec)	N (Prec)	BVE rate
1	SVM	516	25	34	56	296	78.10	0.8409	0.4237	0.2189	0.6913	0.1030	0.3065
2	C4.5	488	14	45	18	334	84.68	0.9488	0.2372	0.1532	0.5628	0.1187	0.2985
3	*k*-NN	542	22	42	24	323	83.95	0.9308	0.3438	0.1605	0.5217	0.1151	0.3221
4	BLR	496	8	49	21	333	82.96	0.9406	0.1403	0.1703	0.7241	0.1283	0.2658
5	MLR	498	15	35	40	321	81.75	0.8892	0.3000	0.1825	0.7273	0.0983	0.2856
6	PLS-DA	432	14	50	7	340	86.13	0.9798	0.2187	0.1386	0.3330	0.1282	0.2558
7	*k*-mean	504	16	37	44	314	80.29	0.8771	0.3019	0.1971	0.7333	0.1054	0.2748
8	EMC	449	9	30	31	341	85.16	0.9166	0.2307	0.1484	0.7756	0.0808	0.2685
9	Apriori	496	8	32	53	314	80.53	0.8858	0.1162	0.1946	0.8936	0.1043	0.2749

Step 2: The above algorithms can filtered by using positive precision values. If the precision value is less than 0.8. We get the eight algorithms namely (SVM, C4.5, *k*-NN, BLR, MLR, PLS-DA, *k*-means and EMC).

S. No.	Alg.	CT (ms)	TP	FN	FP	TN	Acc. (%)	Spe.	Sen.	CVE rate	P (Prec)	N (Prec)	BVE rate
1	SVM	516	25	34	56	296	78.10	0.8409	0.4237	0.2189	0.6913	0.1030	0.3065
2	C4.5	488	14	45	18	334	84.68	0.9488	0.2372	0.1532	0.5628	0.1187	0.2985
3	*k*-NN	542	22	42	24	323	83.95	0.9308	0.3438	0.1605	0.5217	0.1151	0.3221
4	BLR	496	8	49	21	333	82.96	0.9406	0.1403	0.1703	0.7241	0.1283	0.2658
5	MLR	498	15	35	40	321	81.75	0.8892	0.3000	0.1825	0.7273	0.0983	0.2856
6	PLS-DA	432	14	50	7	340	86.13	0.9798	0.2187	0.1386	0.3330	0.1282	0.2558
7	*k*-mean	504	16	37	44	314	80.29	0.8771	0.3019	0.1971	0.7333	0.1054	0.2748
8	EMC	449	9	30	31	341	85.16	0.9166	0.2307	0.1484	0.7756	0.0808	0.2685

Step 3: The above algorithms can filter by using Cross Validation Error rate (< 0.18) i.e. lowest error rate. The above six algorithms can be reduced. We get five algorithms namely (C4.5, *k*-NN, BLR, PLS-DA, and EMC)

S. No.	Alg.	CT (ms)	TP	FN	FP	TN	Acc. (%)	Spe.	Sen.	CVE rate	P (Prec)	N (Prec)	BVE rate
1	C4.5	488	14	45	18	334	84.68	0.9488	0.2372	0.1532	0.5628	0.1187	0.2985
2	*k*-NN	542	22	42	24	323	83.95	0.9308	0.3438	0.1605	0.5217	0.1151	0.3221
3	BLR	496	8	49	21	333	82.96	0.9406	0.1403	0.1703	0.7241	0.1283	0.2658
4	PLS-DA	432	14	50	7	340	86.13	0.9798	0.2187	0.1386	0.3330	0.1282	0.2558
5	EMC	449	9	30	31	341	85.16	0.9166	0.2307	0.1484	0.7756	0.0808	0.2685

Step 4: The above algorithms can filter by using Bootstrap Validation Error rate (< 0.3) i.e. lowest error rate. The above five algorithms can be reduced. We get four algorithms namely (C4.5, BLR, PLS-DA, and EMC)

S. No.	Alg.	CT (ms)	TP	FN	FP	TN	Acc. (%)	Spe.	Sen.	CVE rate	P (Prec)	N (Prec)	BVE rate
1	C4.5	488	14	45	18	334	84.68	0.9488	0.2372	0.1532	0.5628	0.1187	0.2985
2	BLR	496	8	49	21	333	82.96	0.9406	0.1403	0.1703	0.7241	0.1283	0.2658
3	PLS-DA	432	14	50	7	340	86.13	0.9798	0.2187	0.1386	0.3330	0.1282	0.2558
4	EMC	449	9	30	31	341	85.16	0.9166	0.2307	0.1484	0.7756	0.0808	0.2685

Step 5: The above algorithms can filter by using highest accuracy and lowest computing time. The above four algorithms can be reduced to one. We get best one for PLS-DA.

S. No.	Alg.	CT (ms)	TP	FN	FP	TN	Acc. (%)	Spe.	Sen.	CVE rate	P (Prec)	N (Prec)	BVE rate
1	PLS-DA	432	14	50	7	340	86.13	0.9798	0.2187	0.1386	0.3330	0.1282	0.2558

Step 6: Stop the process. We get the best one.

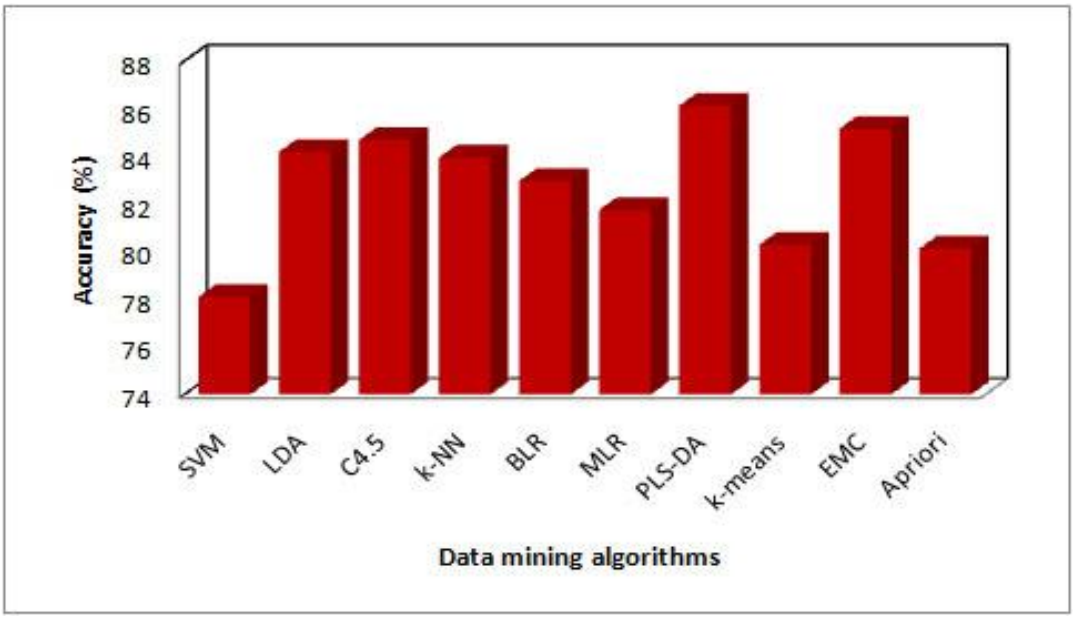

Figure 2: Predicted Accuracy

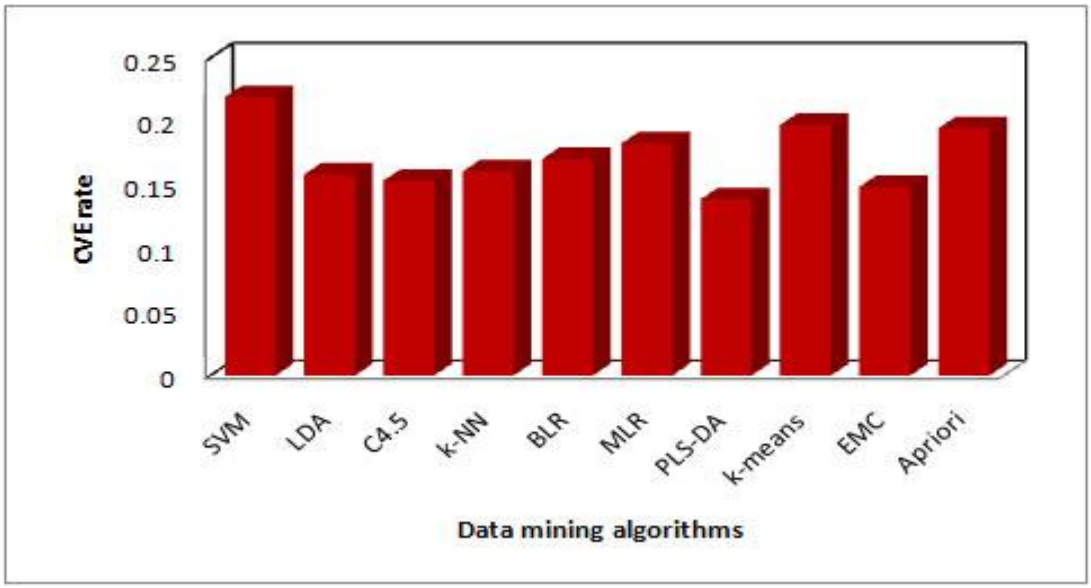

Figure 3: Performance of Cross Validation Error Rate

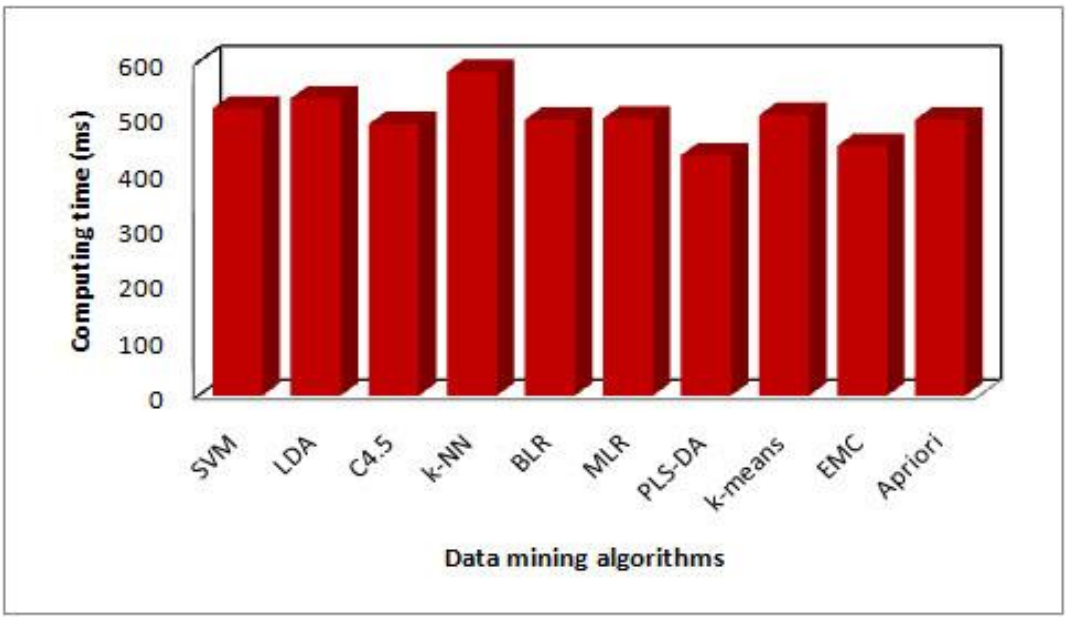

Figure 4: Performance of Computing Time

The step 5 consists of values of different classification. According to these values the accuracy was calculated. From figures (2-4) represents the resultant values of above classified dataset using data mining supervised classification algorithms and it shows the highest accuracy and lowest computing among the three. It is logical from chart that compared on basis of performance and computing time, precision value, Error rate (10 fold Cross Validation, Bootstrap Validation) and finally the highest accuracy and again lowest computing time. PLS-DA algorithm shows the superior performance compared to other algorithms.

VI. CONCLUSION

The main goal medical data mining algorithm is to get best algorithms that describe given data from multiple aspects. The algorithms are very necessary for intend an automatic classification tools. With help of automatic design tools to reduce a wait in line at the experts. The PLS-DA was the best one among ten (five criteria are satisfied). Three axis are used the redundancy cut value is 0.025, positive and negative values are predicted based on the recall and 1-precision values. It can be classified as function as positive and negative and finally constant value of positive and negative. The first one is computing time in 432 milliseconds it is the lowest, second one is Cross Validation error rate is 0.1386, If the precision value is less than 0.8, fourth one Bootstrap Validation error rate is 0.2558 lowest (i.e. repetition is 1, test error rate 0.2245, Bootstrap, Bootstrap+) compare to others and finally three values(Accuracy, Specificity and Sensitivity) are calculated by using formula and the prediction one is Accuracy .Then the Accuracy of PLS-DA is 86.13% from the above results PLS-DA algorithm plays a vital role in data mining techniques. There are so many classifiers are used for gene classification. SVM is the best classifier for gene classification, it shows better results in structural and functional based gene classifications, but for sequence based classification and group classification among sub type it gives minimum results. In that case PLS-DA shows better results. It also results sequence based classification with very least error rate and which increases the accuracy rate. The performance of PLS-DA shows the high level compare with other classifiers. Hence PLS-DA shows the concrete results with different Heart disease of patient records. Therefore PLS-DA classifier is suggested for Heart disease based classification to get better results with accuracy and performance.

ACKNOWLEDGMENT

The authors are thankful to Prof. C.Uma Shankar, Dept. of OR&SQC and Dr. M. Veera Krishna, Department of Mathematics, Rayalaseema University, Kurnool, Andhra pradesh, India, for their valuable guidance and suggestions with thought provoking discussions throughout the period of my research and in the preparation of this paper, and IJSRP Journal for the support to develop this document.

.

REFERENCES

[1] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2004.

[2] Cristianini, N., Shawe-Taylor, J. “An introduction to Support Vector Machines. Cambridge University Press”, Cambridge, 2000.

[3] Frank Lemke and Johann-Adolf Mueller, "Medical data analysis using self-organizing data mining technologies," Systems Analysis Modeling Simulation , Vol. 43, Issue No. 10, 2003, pp. 1399-1408.

[4] Frawley and Piatetsky-Shapiro, Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, MenloPark, C.A, 1996.

[5] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV", LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007.

[6] Hian Chye Koh and Gerald Tan, "Data Mining Applications in Healthcare", Journal of healthcare information management, Vol. 19, Issue 2, Pages 64-72, 2005.

[7] J. Shreve, H. Schneider, O. Soysal, "A methodology for comparing classification methods through the assessment of model stability and validity in variable selection", Decision Support Systems, Vol. 52, pp. 247–257, 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, Vol. 3, Issue No. 3, 2008.

[9] L. Goodwin, M. VanDyne, S. Lin, S. Talbert ,"Data mining issues and opportunities for building nursing knowledge" Journal of Biomedical Informatics, Vol. 36, 2003, pp. 379-388.

[10] Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification Based on Multiple Association Rules. In: Proc. of 2001 International Conference on Data Mining, 2001.

[11] John Peter.T and Somasundaram, "Study and Development of novel feature selection framework for Heart disease prediction", International Journal of Scientific and Research Publications, Volume 2, Issue 10, October 2012, pp. 1-7.

[12] Mohammad Taha Khan, Dr. Shamimul Qamar and Laurent F. Massin, "A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining", International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11, pp. 1-6, 2012.

[13] N. Aditya Sundar, P. Pushpa Latha and M. Rama Chandra, "Performance analysis of classification data mining techniques over heart disease data base", International journal of engineering science & advanced technology, Volume-2, Issue-3, May-June 2012, pp. 470 – 478.

[14] Srinivas.K, B.Kavihta Rani and A.Govrdhan, "Applications of Data mining techniques in health care and Prediction of Heart attacks", International Journal on Computer Science and Engineering, Vol. 02, No. 02, 2010, pp. 250-255.

[15] Jaya Rama Krishnaiah.V.V., D.V.Chandra Sekhar and K.Ramchand H Rao, "Predicting the Heart attack symptoms using Biomedical data mining techniques", The International Journal of Computer Science & Applications, Volume 1, No. 3, May 2012, pp. 10-18.

AUTHORS

First Author – K.R.Lakshmi, M.C.A, B.Ed., e-mail address: krlakshmi_cse@yahoo.com.
Second Author – Dr.M.Veera Krishna, M.Sc.,M.Phil., Ph.D. Rayalaseema University, Kurnool, veerakrishna_maths@yahoo.com.
Third Author –Prof. S.Prem Kumar, M.C.A., M.Tech., Ph.D., G.Pullaiah college of Engineering & Technology, Nandikotkur Road, Kurnool, Andhra Pradesh, India. e-mail: mcahod@gpcet.ac.in.

Correspondence Author – Dr.M.Veera Krishna, veerakrishna_maths@yahoo.com, +91-9849650682.