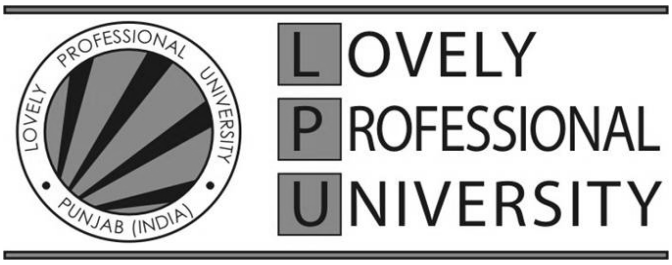


DATA MINING IN HEALTH CARE USING HYBRID APPROACH

A dissertation submitted

BY

Monica Sharma

To

Department of Computer Science and Engineering

In partial fulfillment of requirement for the

Award of the degree of

Master of Technology in Computer Science

Under the guidance of

Rajdeep Kaur

(May 2015)

PAC APPROVAL

LOVELY PROFESSIONAL UNIVERSITY
Transforming Education, Transforming India

School of: Computer Science & Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: Monica Sharma
Registration No: 11306583
Batch: 2013-2015
Roll No. B46
Session: 2014-2015
Parent Section: K2307
Details of Supervisor:
Designation: AP
Name: Rajdeep Kaur
Qualification: M.Tech C.S.E
ID: 16973
Research Experience: 2.5 years

SPECIALIZATION AREA: Database (pick from list of provided specialization areas by DAA)

PROPOSED TOPICS

1. Big data (Big data analytics) (Medical Data)
2. Clustering
3. Predictive Analysis

Signature of Supervisor
16/9/13

PAC Remarks:

first topic approved

Signature:
10/9/13
30/9/14

Date:

APPROVAL OF PAC CHAIRPERSON:

*Supervisor should finally encircle one topic out of three proposed topics and put up for approval before Project Approval Committee (PAC)
*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.
*One copy to be submitted to Supervisor.

ABSTRACT

Modern medicine generates a great deal of information stored in the medical database. Extracting useful data and making scientific decision for diagnosis and treatment of disease from the database increasingly becomes necessary. We propose a Heart diseases Prediction System for the society to prevent the cause of the death. So we are analyzing heart disease patient to identify which treatment is most effective one and provide better result.

CERTIFICATE

This is to certify that Monica Sharma has completed M.Tech dissertation titled “Data mining in health care using hybrid approach” under my guidance and supervision. To the best of my knowledge the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma.

The dissertation proposal is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech computer science and Engineering.

Date: _____

Signature of supervisor:

Name:

UID:

ACKNOWLEDGEMENT

First and foremost I would like to thank almighty for giving me courage to bring up this pre dissertation. Before getting into thick and thin of this pre dissertation I would like to show my gratitude to some of the people who have helped me in this project. Firstly I would like to purpose a word thanks to my mentor RAJDEEP KAUR who has encouraged me to get through this pre dissertation. Secondly I would like to thanks my friends who gave me unending support and helped me in numerous ways from the stage when the idea of the thesis was conceived. I am very thankful to all of them for making my work complete successfully under their guidance.

DECLARATION

I hereby declare that the dissertation entitled, **Data mining in health care using hybrid approach** submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date:

Monica Sharma

RegNo.:11306583

Table of Contents

Topic **Page No.**

LIST OF TABLES

Table name **Page no.**

LIST OF FIGURES

Figure name **Page no.**

Chapter 1

INTRODUCTION

Nowadays healthcare are enough capable to generate and collect large amount of data. Due to the generation of huge amount of data it may require most frequent way to extract this healthcare type of data when needed. Using data mining approaches, it is possible and useful to extract interesting and meaningful information and their regularities. To acquire this type of knowledge using mining it can be used in various areas so as to improve the work efficiency or performance and extend the quality of decision making process. As there is much need for the coming generation related to computer theories and tools to help people in extracting the useful information from the continuously growing of large data.

Day by day information technologies are increasing to implement various problems in healthcare organization in order to take action as needed by the doctors in taking decision making actions usually. Various technologies have developed number of tools of data mining which are useful to manage the limitation of the people as occurred errors due to fatigue and offer them the indication for decision making process. The main goal of data mining technique is to identify relationship, different patterns and model which provide support in medical health. These can be referred as predictive model and they have been included into information system of various hospitals as a representation or prototype as to make decision, reduce content and time for decision making. Using information in healthcare enables the management of medical data and its safe switching between the users and the providers of various healthcare services[10].

With the use of the various technologies, it enables us to eliminate manual task for extracting data from the charts or filling particular questionnaires, retrieving data directly from electronic records, and placing this data safely on electronic system of medical records that can save many lives and also reduces the cost of healthcare, advance collection of data may allow us for early detection of infectious diseases etc thus retrieving information with the use

of computer can help and improve the quality of decision making and avoid human faults .Medical dataset are generally distributed heterogeneous and huge in nature so there is a great need to organize and integrate this dataset with hospital management system.

In today's modern world cardiovascular disease is one of the highest flying diseases. According to world health organization survey more than 12 million people dies every year due to health problem so it need to be diagnose accurately and correctly as there is limitation of medical expert and their unavailability at some location so it can put patient at high risk so it is highly beneficial if this technique i.e. data mining will be integrated with the medical information system.

1.1 Heart diseases patient chosen from healthcare for prediction

Heart Diseases remain the biggest cause of deaths for the last two epochs. Recently computer technology develops software to assistance doctors in making decision of heart disease in the early stage. Diagnosing the heart disease mainly depends on clinical and obsessive data [4]. Prediction system of Heart disease can assist medical experts for predicting heart disease current status based on the clinical data of various patients [2].

In biomedical field data mining plays an essential role for prediction of diseases. For diagnosing, the information which has been provided by the patients may include similar data and interrelated symptoms and shows especially when the patients suffering from more than one type of diseases of the similar category. The physicians are not capable enough to diagnose it correctly.

1.2 Data mining

Data Mining is referred as to extract the valuable or useful data from the huge amount of dataset. It is said that it is defined as to mine the knowledge from complex data and mined information can be used for various applications.

1.3 Need of data mining

Data is most important assets of the organization but finding out useful information from the data is a complex task for finding useful information from given data set we have to apply data mining techniques. Data mining is the field where we study and research techniques for the mining data more effectively and efficiently give more realistic information which provides help in decision planning for example a clothing company preformed mining on his yearly sale and find out in which month he have to provide offers to the consumer which increase overall profit. Today we have lot of unstructured data we need such efficient techniques which help providing efficient analysis of this data [3].

Data mining provides is method which helps us mapping unstructured data to structured data with the help several techniques Data Cleaning, Data Integration, Data Transformation this

Techniques provide us information which helps in various fields of our daily life for decision

making for in all the business and educational field it reduces the un-useful information and provide us authentic information which is necessary for the decision making because storing the data is costly task.

1.4 Data mining applications

Data mining is playing major role in our daily life some of them are explained here

i. Data mining is use the analysing and managing the market trends
ii. Data mining is used for fraud detection
iii. Data mining is used for decisions making
iv. Data mining is used for evaluating risk
v. Data mining is used in medical for developing expert system

1.4.1 Market analysis and management

Data minig is used in these various field [4]:

Customer Profiling – Customer profiling is use to record of buying pattern of the customers and make decision about providing various offers which leads to increase profit of the organizing.

Identifying Customer Requirements - Data Mining use to analyse customer requirement which helps the organization to perform mining of collected requirement and take decision how they can provide better customers satisfactions.

Cross Market Analysis - Data Mining find out similar relation between different items and the association, correlation between the different products and the pattern of sales of that product.

Target Marketing - Data Mining also find out similar type of customers on the basis of different parameter such as (interest, spending and buying pattern) and put them into a cluster.

Cluster is the collection of same types of information.

Determining Customer purchasing pattern - Data mining find out deterging buying pattern for customers which is very necessary for new product launching and marinating demand and supply ratio.

1.4.2 Fraud Detection

Data Mining is also used in fields of credit card services and telecommunication to sense deception. In deception telephone call it helps to find terminus of call, period of call, time day, week. It also analyse the patterns that deviate from an expected norms. Other Applications Data Mining also used in other fields such as sports, fortunetelling and Internet Web browsing Aid.

1.5 How data mining works

By observing and analyzing the data of patient, various data miners experts uncover the hidden patterns or facts .For example in Washington, D.C hospital wants to identify why their patient got soon sick after discharge, after many researches data mining technique reveals the fact that the patient who are staying in the same hospital room afterwards developing the same infection. It was an example to show how various data mining technique helps us to analyze and solve the problems in healthcare. Commonly, data miner experts use the method called cross industry standard process for data mining i.e. CRISP-DM [5].

It involves six steps are as follows:

- **Understanding the business**-Identifying the objective and the requirements from the perspective of business and defining the problem of data mining.
- **Understanding the collected data**-Collecting the raw data, study the data and looking into it for any problem related to data quality.
- **Data preparation-** Build the evaluating dataset from the collected raw data.
- **Modeling**-Different Data mining software were used for the purpose of analyzing.
- **Evaluation**-Evaluate the accomplishment of the objective by comparing data mining model and their result.
- **Deployment**- Implementing the result.

1.6 Benefits of healthcare in Data mining

In healthcare, Data mining has become more and more popular as it offers benefits to patients, health, organization, care provider, researchers, insurers.

- **Patient**-They receive better and more affordable healthcare. Healthcare manager use data mining technique for identifying and tracking the chronicle disease and high risk patient, reduces the number of hospital admission and claiming.

- **Healthcare Organization**-In this mining influence cost revenue and operating efficiency. It provides information by guiding the patient interaction, by giving patient preferences, usage pattern, current and future requirement all these helps in improving patient satisfaction.
- **Care provider**-They have used various data analyzing technique to identify the useful treatment and best practices.
- **Insurers**-They can detect insurance fraud and abuse with the help of using the mining by creating norms and then identify unusual claims pattern.

1.7 Healthcare Management

Major Dimensions in healthcare management are as follows:

- Diagnosing the diseases and treating the patient
- Managing the Healthcare resource
- Customer relationship management
- Fraud and anomaly detection

1.7.1 Diagnosing the diseases and treating the patient

For Doctor medical decision support

i. For diagnose melanoma skin lesions disease digitized images examining.
ii. Ultrasound images are analysed using computer assisted to monitor of tumour reaction for chemotherapy.
iii. To predict the presence of brain neoplasm with spectroscopy

1.7.2 Treatment planning for patient

i. Predominantly treatment plan is provided by data mining when we don't have dispositive evidence.
ii. Cure planning can be recommended to patients, it depend on patient profile, history, physical examination, diagnosis and use earlier treatment pattern.

1.7.3 Managing the Healthcare resource

i. Comparing the hospital information which is based on risk adjusted death within 30 days of non cardiac surgery by using logistic regression model.
ii. To predict disposition in children using neural network for urgent situation with bronchiolitis.

1.7.4 Customer relationship management

i. To improve customer satisfaction identify usage and sell pattern of customer.
ii. Customer can be patients, pharmacists, physicians or clinics.
iii. To reduce overall cost and increase customer fulfillment so there is a need to predict purchasing and usage activities of customer.

1.7.5 Fraud and anomaly detection

i. Early recognition of fraud medical insurance prevented with the help of data mining.
ii. Fraud like prescription fraud and claims to get money grom insurance for medical treatment which is not performed.

1.8 Classification

Classification is a mining utility which are assigning the items in a similar group in order to target the different category or classes. Main purpose of the classification is to predict accurately the main class for each case in the dataset. We can use various classifying model

for identifying the applicants as low, medium, or high credit risks. We are applying classification process on a data set depending upon class assignments that we make. As an example based on historical credit rating, we can develop a classification model which can predict credit risk depending upon the data for many loan applicants over specific period of time. Credit rating would be the target, further elements could be the predictors, and the data for each customer would constitute a case[20].

Classifications are having discrete value and they don't involve sequences. Continuous and floating-point values were indicating a numerical value rather than categorical value. Predictive model with a numerical value referred to use a regression method and not classification algorithm.

The simplest type of classification is binary classification. This classification considers two possible values i.e. high or low. Multiclass has more than two class value such as positive, negative and neutral.

For building a model, a classification algorithm derives relationships between the values of the prediction and the target value. Different techniques have been used by various classification algorithms for finding relationships. We summarize relationships in a model, than we apply it to a different data set in which the class assignments are unknown.

Models for classification are evaluated by comparing the predicted values to known target values for test data. We classify historical data into two data sets: one for building the model and another for testing the model.

Developing a model for classification may results in assigning the class and define the probability in each case. For e.g., a building model that classifies each user as low, medium, or high can also predict the probability of each classification for the user.

Classification is used in various filed such as:

(i) segmenting the customer
(ii) business modeling

(iii) marketing

(iv) credit analysis

(v) In biomedical and drug response modeling

1.9 Classification Methods

Following are some classification methods that we have discuss in brief

Discriminate Models- Case based reasoning, Rule based reasoning, Bayesian learning

Predictive Models- Evolutionary computation, Simple linear regression, Multiple linear regression, Analysis of Variance, Generalized Linear Models, Time series.

1.9.1 Discriminate Models

Case based reasoning- contains instance based learning.

It is defined as learning based on instance: This technique uses previous data for the classification of the new instance of a problem in a predefined set of classes.

Rule based reasoning contains rule based classifier, decision trees, discriminate analysis, support vector machine, regression trees, model trees.

1) Rule-based classifiers: They offer a certain set of rules that can be used afterwards for the evaluation of new cases and classifying them into a predefined set of classes.
2) Decision trees: They are used for the graphical representation of a tree where conditions are defined on nodes. Tree have problem when need to build on very big data sets. Woks well with qualitative variables.
3) Discriminate analysis: They are used for an algebraic discriminate function and a cut-off as the rules as for deciding between two groups for a new instance.
4) Support Vector Machines: They are used for the discriminate functions so as to distinguish between two predefined classes that might be non-linearly separable.

5) Regression-trees: They are used as decision trees for predicting the numerical values. Each leaf consist of numerical value, i.e. calculated as the average of all the training set values that are applied to any leaf, or rule.
6) Model trees: They are used as regression trees that have been combined with regression equations. For these trees, leaves are containing the regression equation in place of single predicted value.

1.9.2 Predictive models

1) Evolutionary computation: Provides the optimization of a certain objective function through the evolution of a population of individuals, which are subjected to several genetic operators. Include techniques simulating the theory of evolution, like genetic algorithms and genetic programming.
2) Simple linear regression: It simply predicts the quantitative variables value for a given instance which is consider as a linear equation of a single numerical variable. Regression may requires the value for normality and linearity.
3) Multiple linear regression: It simply predicts the quantitative variables value for the newly instance as a linear equation of various numerical variables. It needs the value of normality, linearity and independence
4) Analysis of Variance: It simply predicts the quantitative variable value for a newly instance that has a linear combination of one or two qualitative variables. It needs the value of conditional normality, linearity and independence.
5) Generalized Linear Models: It simply predicts the quantitative variable value for the newly instance that has a linear combination of various numerical and qualitative variables. Same hypothesis as previous methods, all of them particular cases of that one.
6) Time series: It simply predict the quantitative variable value for the future instance that has a linear combination of previous values of the similar variable for technical hypothesis required.

Algorithms for classifying the data

I. **Decision Tree**

These tree generate rules automatically, that are consider as conditional statements that are needed for the logic to build the tree.

II. **Naive Bayes**

a. They are based on Bayes' Theorem that evaluates the probability by the total count of the frequency values and combinations of values in the historical data.

III. **Generalized Linear Models**

It uses statistical method for linear modeling. It is used for binary classification and for regression.

IV. **Support Vector Machine**

a. It is a powerful algorithm which is based on linear and nonlinear regression. Mining implement this algorithm for binary and multiclass classification.

The type of the data determines which classifying algorithm give best result for the particular problem. Each algorithm have different accuracy, completion time and transparency. In practical, different models have been developed; we are selecting the best one for each algorithm, and then we are choosing the best among them to solve the problem.

1.10 Predictive Analytics

It is defined as technology which captures the process of data mining in simple routines. Also referred as “one-click data mining”. This analytics make simplification and automates the process of data mining.

It also develops various profiles, discovering the factors which are leading to certain result, predicting the most preferable outcomes, and also find out a degree of confidence in the result of predictions.

Predictive Analytics and Data Mining

Various data mining techniques have been used by Predictive analytics, but useful information of data mining is not necessary for the use of predictive analytics. You do not need to create or use mining models or understand the mining functions and algorithms.

1.11 Naïve bayes Classifier

This classifier probably works on conditional probabilities and it is based on Bayes Theorem, that evaluates the probability by counting the values of frequency and the possible combinations of different values in the historical data.

Bayes' Theorem calculates the probability of an event occurred given that the probability of another event that has already occurred. B represents the dependent event while A represents the prior event.

$$P\left(\frac{B}{A}\right) = P(A \cap B)/P(A)$$

For Example:

We are determining the likelihood for the user under 21 will increase spending. Here, the prior condition A will be under 21 and the dependent condition B would be increase spending.

Consider 100 user in the training set and 25 of them are user under 21 which all have increased their spending. Now calculate the probability:

$P(A \cap B) = 25\%$

Suppose 75 of the 100 user are under 21 so, P(A) will be

P(A) = 75%

Bayes' Theorem will evaluate that 33% of user under 21 are having likelihood to increase their spending (25/75).

Cases where both conditions occurs together are called as pairwise. In this example, pairwise is 25% for all the cases.

Cases where only the prior event where occurred are called as singleton. Here, 75% is singleton for all cases.

Naive Bayes consider the condition that each predictor is independent of the others. For each targeted value, for each predictor the distribution is independent of the other one. The distribution of the predictor is not clarified for the larger population[21].

1.11.1Advantages of Naive Bayes

- This algorithm affords fast, high scalable to build model and scoring. Scaling is done linearly with the number of predictors and rows.
- It can be used for the problems like binary and multiclass classification.

1.11.2 Tuning a Naive Bayes Model

Evaluates the probability by dividing the percentage of occurring of pairwise by the percentage of occurring of singleton. For a specific predictor, if the value of the percentages is small, they will not provide effectiveness to the model. Occurring below a certain threshold cannot be considered.

Two different settings are available for the probability thresholds. They are:

i. **Nabs_Pairwise_Threshold** —It is defined as minimum percentage for the pairwise to be occurred i.e. required to include a predictor in the model.
ii. **Nabs_Singleton_Threshold** – It is defined as minimum percentage for the occurrence of singleton i.e. required to include a predictor in the model.

1.11.3 Preparing Data for Naive Bayes

Automatically, Preparation for data is performed on supervised binning for Naive Bayes. Decision trees have been used by supervised binning for creating the optimal bin boundaries. Categorical and Numerical attributes have been binned.

Naturally, it also handles missing values as missing at randomly. Sparse numerical data have been replaced with zeros and sparse categorical data by zero vectors using this algorithm. If we are managing our own prepared data, generally Naive Bayes requires binning.

Binning should be done on columns so as to reduce the cardinality. Binning on Numerical data can be ranged into values such as low, medium, and high and binning on categorical data into meta-classes for example regions instead of cities. Binning like Equi-width is not suggested, as outliers will cause the data to be concentrated in a few bins, may be in one bin. Resulting the discriminate power of the algorithms will be reduced significantly.

1.11.4 Why we need Bayesian Classification

Probabilistic learning- For most practical approaches to certain types of learning problem,it can calculate explicit probabilities,it can calculate explicit probabilities for assumption.

Incremental-Through training instances it could be incrementally increase/decrease probability that a hypothesis is correct.

Probabilistic Prediction-Several hypothesis can be predicted based on weighted probabilities.

Standard-It provides other standard optimal methods when Bayesian methods gives computationally difficult result.

1.12 Decision Table

- A decision table is used for representing conditional logic by making a list of process which depicts the business level rules. These tables can also be used when constant

numbers of conditions are there which are needed to be calculated and where a exact set of events to be used when the following conditions are to be met.

- These tables are very similar to decision trees except that tables will have the similar number of conditions that needed for evaluation and actions that are needed to be taken. While decision tree, contain one branch with addition of conditions that are necessary to be evaluated than other branches on the tree.

Decision Table

- The main idea of a decision table is of structuring the logic so as to generate rules that are derived from the data which have already been entered into table. A decision table consists of lists causes i.e. business rule condition and effects i.e. business rule action, which have been denoted by the matrix where each column represents a single combination.
- If number of rules are there inside the business which are expressed by the use of some templates and data then we can referred the decision table technique to accomplish the particular task. Individual row in the decision table collects and stores its data uniquely and then bind the data with a particular or customized template to generate a rule. It is not preferable to use decision tables if the rules are not following a set of templates.

1.12.1 Benefits and drawbacks of Decision Tables:

1. A decision table offers structure for a comprehensive and exact statement of processing or decision logic. It powers the programmer to consider every probable condition.
2. It is easy to construct decision table than a flow chart.

3. A decision table is compact and easy to understand which is very effective for communication between analyst or programmers and non technical user. It is useful in documentation.
4. Software packages are available for converting decision tables into program easily.
5. We can check that every test possibility have been taken into consideration.
6. Option are shown alongside to ease exploration of combinations.
7. Cause and effect relationship are reflected by tables.
8. Standardized format is used.
9. It is easy to copy table by typist.
10. We can split complex tables into simpler tables.
11. Little or no computer knowledge is required by table user.

1.12.2 Drawback

1. Decision tables do not have flow sequence as we do have in flow chart.
2. LOGIC-where the logic of a system is simple, flowcharts nearly always serve the purpose better than decision table

Chapter 2

LITERATURE REVIEW

Saba Bashir,Usman Qamar,M.Younus Javed. "Ensemble based decision support framework for intelligent heart disease diagnosis",(2014) presents heart disease prediction system using ensemble based approach. The proposed framework using majority voted based novel classifiers ensemble to combining different data mining classifiers. Dataset for analysis is being collected from UCI repository. Algorithm that is used are of heterogeneous classifiers. Naïve baiyes, Decision table,SVM. Proposed majority vote based ensemble divided into two main parts 1. Produce a training set of individual classifiers 2. For majority voting scheme combine decisions of classifiers to generate new model. Here ,in proposed system,three classifiers are apply to trained the dataset then result of each classifier classify data into two classes i.e patient having heart diseases yes =1 and patient not having heart diseases no=0.Voting method will works as out of three classifiers two classifiers having same result that will be final result taken from heterogeneous classifier then to verify result based on performance of algorithm. Evaluation parameters are measured i.e accuracy,sensitivity and specificity. Finally accuracy of proposed algorithm is 81.82%, Sensitivity 73.68%,Specificity 92.81%. The main goal of this proposed research is to obtain most accurate prediction of heart diseases for patient. [16]

G.Karthiga, C.Preethi, R.Delshi Howsalya Devi."Heart Disease Analysis Sytem Using Data Mining Techniques", (2014) stated that different data mining techniques to be used to predict heart disease in healthcare area. Techniques which to be used for mining are association mining, classification, clustering. Firstly heart disease database is preprocessed so that procedure of mining become well-organized then clustering is done on this preprocessed data. Here MAFIA is used to mine maximal frequent patterns from heart disease database then these frequent patterns can be classified using C4.5 classification algorithm. Classification is an unsupervised learning used to predict class objects whose class label is unknown. It is used to

create classification rules using decision trees from given data set. Prediction accuracy is compared between simple mafia and proposed k-mean based mafia. [4]

Hlaudi Daniel Masthe,Mosima Anna Masethe.“Prediction of heart disease using classification algorithms”,(2014) this paper presents different classification algorithm to predict heart diseases patient and compare the best method for prediction. Algorithm which are used here as classifier are J48,Naïve Bayes , Reptree, cart and Bayes net. Dataset is collected from medical practitioners in south Africa. Weka tool is being used for analysis and predicting model. This research provide important tool for doctor to predict risky cases in put into practice and judgment accordingly. [8]

Lin Li, Saeed Bagheri, Helena Goote.“Risk adjustment of patient expenditures”,(2013) stated that health care require an application which is consist of voluminous patient data contain rich and meaningful insight that revealed using advance matching knowledge algorithms where traditional machine learning tools cannot be applied on this large volume and velocity of such high dimensional data. A risk adjustment model is present over here to forecast health expenditures based on relative absolute risks of a patient. Random forest methodology used here which use divide and conquers strategy to achieve enhanced prediction performance by training an ensemble of decision trees each with a different feature subset and using vote scheme to combine results. In this paper they use Apache Mahout’s random forest implementation to train risk adjustment model on the cluster which build enormous number of decision trees in the model in parallel. Comparison between linear regression and random forest for test data to predicted the inpatient expenditure and actual inpatient expenditure. [9]

Ranganatha S, Pooja Raj H.R, Anusha C, Vinay S.K.” Medical data mining and analysis for heart diseases dataset using classification techniques”, (2013) stated that main aim of this paper is to store medical information of patients who come for hospitalization for heart disease, algorithms run on that information which give result in the form of user

understandable. Methodology explained such as, it firstly we will collect raw data which is to be mined then pre-processed it then after training part of cleaned data is passed to mining algorithms and rules or patterns will extracted based on similarities in the data and passed over classification rule based algorithm, the result accuracy will be checked.ID3 algorithm and Naïve Bayesian algorithm are best suited for heart diseases data set.ID3 algorithm used over here to build decision tree by applying top down approach and greedy search method to test each attribute at each node of the tree. In Naïve Bayesian classifiers, works on probabilistic statistical classifier. It recognizes the characteristics of patients suffering from heart disease.[14]

V.Manikantan & S.Lanthan."Predicting The Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods.", (2013) Stated prediction of heart disease symptoms with the help of classified algorithms which is presented in this paper. Classification is done through MAFIA algorithm which results in accuracy. Here C4.5 algorithm is used for training dataset to show the rank of heart attack with decision tree. Clustering is done with the help of K-Mean Clustering algorithm.C4.5 algorithm and K-Means Clustering algorithm are broadly used to mine the medical data for analysis. The Classification of objects into different groups or divide the dataset into subdivisions so that the data in each of subdivision share a common object with respect to some fixed space known as Clustering-Mean It form clusters of data based on their individual values into k distinct groups. Methodology used for pre analysis of heart disorder, data is grouped with the help of K-Mean algorithm with K values then produce the appropriate cluster data which pass to MAFIA from that select frequent data and classify the data with the help of C4.5 algorithm then show the effective heart attack possibility and accuracy. [19]

R.Chitra and V. Seenivasagam."Review of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques", (2013) "stated the comparisons of different classification algorithm to check performance and accuracy of the algorithm. K.Srinivas presented application using classification algorithm by using Tanagra data mining tool for experimental data analysis, training dataset consists of 3000 cases with 14 different features.

Cases in dataset represent the results of diverse types of testing to predict the accuracy of heart diseases. The performance evaluation was done based on three algorithms and comparison is shown to measure the accuracy. While comparing, it is to be observed that neural network algorithm provide better performance compare to other. [15]

Abhishek Taneja."Heart Disease Prediction System Using Mining Techniques", (2013) stated that in heart disease attack a person in such a condition that it hardly gets any time to get treated with. Diagnose patients correctly and on timely basis is most challenging task for medical. The purpose of this paper is to develop a cost effective treatment using data mining technology for facilitating database decision support system [6] .Here methodology used to develop a prediction model that can predict heart disease cases based on measurement taken from transthoracic echocardiography examination. Various experiments are conducted with different algorithm to measure the accuracy, precision, f-measure for the result generated.[1]

Vikas Chaurasia."Early Prediction of Heart Diseases Using Data Mining Techniques",(2013) stated that here three popular data mining algorithms are used they are CART,ID3 and DT to develop prediction models using a large data set. Most important attributes for heart disease are cp (chest pain), slope(the slope of peak exercise segment), exang(exercise induced angina) & restecg (resting electrocardiographic). This paper describes performance of the each classifiers with accuracy measures and also test the result that done through chi-square, info gain, gain ratio.[18]

Mythili, Dev Mukherji,Nikita Padalia,and Abhiram Naidu."Heart disease prediction model using SVM Decision tree logistic regression",(2013) This paper describes the early prognosis of cardiovascular diseases and due to early detection how a patient can change his lifestyle to reduce difficulty. Here, rule based model is used on support vector machine, decision tree and logistic regression by applying rules to compare accuracy.The approach is used in this paper is partition into six parts comprises of preprocess the data on individual

model. Dataset is arranged from Cleveland heart diseases dataset uci repository.Proposed Framework works as,on patient database preprocessing is to be done first then after this preprocessed dataset is used to build model using classifiers and test this model with the help of parameter analysis. Each classifier is combine with other such as testing apply on SVM-DT rule,SVM-LR rule,DT-LR rule,SVM-DT-LR rule .Result from each classifier is collected based on collection of result, two things will be measure i.e comparison of result and prediction of heart diseases.[13]

David Cornforth,Mika Tarvainen,Herbert F.Jelinek."Computational intelligence methods for the identification of early cardiac autonomic neuropathy",(2013). This paper present how cardiac autonomic neuropathy disease investigated as early as possible. Cardiac autonomic neuropathy disease CAN is seen in patient due to damage in nerves that effects heart rate. To detect this disease in early stage, it could be done with help of HRV i.e Heart Rate Variability. It happens as HRV be continuously analysed with time and frequency technique. Different classifiers are used to evaluate the performance i.e accuracy of each classifier are resulted for comparison.[3]

Hanaa Elshazly,Ahmed taher azar,Abeer el-korany and aboul ella hassanien."Hybrid System for lymphatic diseases diagnosis",(2013) This paper deals with hybrid concept of combining different algorithm to diagnose lymphatic diseases.Algorithm which is used here are genetic algorithm and random forest .Lymphatic disease is responsible for destroying toxins,cancer cells and dead blood cells.Dataset for lymphography is from institute of oncology,Yugoslavia.The result of hybrid shows the accuracy 92%.The number of relevant feature reduced from 18 to 6.It saves the computational time,storage space and increase learning speed.[5]

Ahmed T. sadiq alobaidi,Noor thamer mahmood.Modified full Bayesian network classifiers for medical diagnosis,(2013). In this paper,the state of heart diseases and nervous

diseases patient can be diagnose with the help of modified Bayesian network for full Bayesian classifier(FBC).Database comprise of 2 different types,so there were ten tables in which five tables for heart disease patient and other five for nervous system.Each table indicates a specific disease details with five tables having different types of diseases patient taken for analysis.

Steps for learning FBC,Firstly we will provide a training dataset S then S will divide into c subsets, for each subset Sc, it corresponds to class value c,then construct an FBC for Sc.The structure of this learning model indicates variable to all variables ranked after it.Steps for learn the fast CPT trees,When learning phase of FBC is completed than a CPT tree learned for each variable.C4.5 decision tree algorithm is used to train the model.Proposed model works as first step is to gather the patients files from Iraq hospitals then apply filteration on it and extract the data to take out the major symptoms from history of disease then these data store in repository afterwards to database input learning algorithm is applied to build the model.Output of model shows and gives the probability of each attributes and also provide accuracy.[2]

Hian Chye Koh and Gerald Tan."Data Mining Applications in Health care", (2012) stated that data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Methodology & Technology used in data mining to transforms these huge amount of data into valuable and understandable form to take better decision. In this paper, it discussed about various areas in which mining approach can be applied such as evaluating the cure efficiency, healthcare management,fraud and abuse can be detected. It also provide information for the identification of risk factors associated with onset of diabetes. [7]

M.Akhil jabbar, Priti Chandra, B.L Deekshatulu."Prediction of Risk Score for Heart Disease Using Associative Classification and Hybrid Feature Subset Selection", (2012)stated that Medical applications of data mining include to forecast diabetic and heart disease symptom analysis and best treatment provided to the patient. Associative classification techniques are being used here to achieve high classification accuracy. It adopt search algorithm like apriori, which produces large number of rules from which a set of high quality

of rules are selected to construct effective classifier. Associative Classification contain 2 stages, In first stage association rule mining algorithm like Apriori or FP growth to produce class association rules, which generate large number of rules. If all produced rule are in classifier then accuracy of the classifier should be high but classification process will be slow and time consuming. Here Feature subset selection is discussed which fall into two types. Filter Model and Wrapper Model, Filter model describes the general characteristics of the training data to choose appropriate features without any learning algorithm. Whereas Wrapper model describes one preset learning algorithm for feature selection and check its performance to identify which feature is selected.[11]

Mai Shouman, Tim Turner ,Rob Stocker."Using Data Mining Techniques In Heart Disease Diagnosis And Treatment ", **(2012)** stated that By using hybrid data mining technique which shows most effective results for the diagnosis of heart disease. Researchers are advising that using different mining techniques on patient treatment dataset will advance practitioner performance. Currently researchers taking use of hybrid data mining techniques in the diagnosis of heart diseases. The best accuracy attained by using single data mining techniques is 84.14% as using naive Bayes and in hybrid data mining case accuracy attained is 89.01% by neural network ensemble. Proposed Research Model is given in this paper which explains how it works as single data mining techniques and hybrid data mining techniques apply on heart disease diagnosis data. Then single data mining techniques will establish baseline for the measure the accuracy that passes over to test data for comparison with hybrid mining techniques.[10]

Hezlin Aryani Abd Rahman,Yap Bee Wah."Comparison of predictive models to predict survival of cardiac surgery patients",(2012) This paper presents for prediction of survival patients of cardiac operation which is to be stored and mining for further examination. Three predictive modes used here to develop model and comparison between them results proved the best predictive model after comparison. Dataset of cardiac surgery patient recommended by field experienced doctors of heart treatment hospital from Malaysia .Dataset of cardiac surgery

patient having attributes discussed here. Class attribute is survival status i.e Died=0, Alive=1, Gender, age, Hypertension, Supertype, Chest_reopen, Atrial_fibrilation, Wound_infection, Stroke after surgery. The methodology divides into following parts first phase is analysis of data phase in which study of data is done and preprocessing is carried on. Second phase is data modeling in which undersampled dataset of cardiac surgery patients were modelled. Algorithm which were used to build model are Logistic Regression ,Decision Tree and neural network. Final step is to measure all models based upon its performance to check which one is most excellent model for cardiac patient to predict survival rate of cardiac surgery patient results proves the model which is providing excellent result having accuracy is 88.4% that was artificial neural network.Critical factor that was determined for prediction are chest_reopen, atrial_fibrillation, wound_infection and stroke_after_surgery.[6]

Mohammad Taha Khan,Dr. Shamimul Qamar and Laurent F.Massin."A prototype of cancer/heart disease prediction model using data mining ",(2012) This paper present a model of breast cancer and heart disease using data mining approach.Algorithm which is used to build model are decision tree C4.5.Breast cancer prediction is done in this paper.Data is collected from university of Wisconsin hospital.Total 11 attribute is there in dataset. Here for prediction target class have 2 values i.e 2 for Benign and 4 for malignant. Disease which destroy close by cells and reach to all different part of body that will malignant .Disease which do not spread to different parts of body i.e Benign.Algorithm used i.e decision tree C4.5 to which tree is generated by some parameters and based on generated tree rules are define. Heart disease prediction uses dataset from Cleveland heart diseases database .Here same algorithm is used i.e C4.5.Prediction of presence or absence of heart diseases is based on rules which is generated by algorithm to measure the performance how much accurate result algorithm will produce.[12]

Xiao Fu,Yinzi Guiqiu,Qing Pan."A Computational modelfor heart failure stratification",(2011) In this paper, a model is proposed for early diagnosis and proper treatment of heart failure and to reduce morbidity and mortality rate.Tool which is used to

build a model i.e Monte Carlo simulation(MCS).Three different classifiers are used i.e Naïve bayes, SVM, Radial basis function network.The proposed model is partitioned into 3 parts. Firstly on dataset MCS is apply to determine optimal classification algorithm then characteristics of clinical data are analysed.Lastly model is build based on MCS result and feature of clinical data.The steps which model follows Random input is generated, two class is define i.e group A and group B for the risk of heart failure next step is choose random sample which is generated apply on model using RBF,NBC,SVM.Results of the model will compare with respect to accuracy, sensitivity and specificity for each method.[20]

Chapter 3

PRESENT WORK

3.1 Problem formulation

In today's world people want to live very luxurirous life so they work hard in order to earn lot of money and live comfortable. Due to this, people forget to care of themselves which result in the change in their lifestyle and food habits which leds to high blood pressure,sugar problem at very young age. They don't even worry if they are sick neither go for their own meditation. As a result of these, it leds to major problem called heart diseases. As in human body heart is most essential it may spoil the human health system. Therefore it is very important to diagnose the heart diseases. Due to availability of huge amount of data, the information can't be retrieved easily, so data mining approaches are implemented in order to extract knowledgeable information for the survival of patient or to analyse major cause of disease.

So, we have proposed system in which hybrid method is used which involves combination of naïve bayes and decision table so as to improve the performance i.e accuracy.

3.2 Objectives

- To collect raw dataset from various categories of heart diseases.
- Apply filtration to remove noise and missing values from raw data.
- Implement hybrid classification algorithm on collected data.
- Analyze the performance and compare it with the existing algorithm.

3.3 Methodology

In this proposed work, a heart disease prediction system has to be developed so that it can used for extracting useful information on the basis of symptoms. Due to availability of huge amount of unstructured data on different type of diseases the information can't retrieved easily and also data mining approaches can't apply on all amount of database. The hidden relationship on different diseases and their causes are not easy to extract from unstructured information.

Main steps include:

Data source

Dataset were mainly collected from UCI repository and from various hospitals of heart diseases. Around 1080 patients data were collected which contains 50 attributes but we have selected only 14 of them in order to obtain the accurate results.

Preprocessing

In preprocessing step, it selects an attribute for selecting a subset of attribute so that it can provide good predicted capability. It handles all the missing values and remove them. If an attribute contain more than 5% missing values then the records should not be deleted and it is adviced to put the values where the data is missing using some suitable methods and helps in feature selection and class label.

Classification

Classification is a technique for machine learning by which it is used to predict the grouping membership of different data instances. It will perform the task by which it will generalize the well-known structure so as to apply it on new data. Here naïve bayes classifier has been used quality measurement of dataset will be consider on the basis of percentage of correctly classified instances. For validation phase we use 10 fold cross validation method. Naive bayes classifier helps in identifying the characteristics of patient with heart diseases. It gives the probability of each selected attribute for the predictable state.

Hybrid Approach

In this hybrid approach

Decision table algorithm stores the data based on the preferred set of attributes and using that model as a lookup table while making the predictions for the data. Every entry in the table has class probability related with it. The main challenge of Decision table is to select a part of set that has highly discriminative attributes. Naïve Bayes is depend on Bayes' theorem with independence assumptions between predictors.

1) In proposed approach, Decision table represent the conditional probability table for naïve bayes.

2) Each point in the search, the algorithm estimates the values by isolating the qualities of attributes into two different parts: one for each of naïve Bayes and decision table.

3) Initially every attributes are modeled by the decision table. At every step, forward selection search is used, and the attributes selected from this are build by naive Bayes and the rest by the decision table.

At every point in this exploration it estimates the value which is associated with the attributes splitted into two sets. The class probability estimates must be combined to produce overall class probability estimates.

Split the attributes into two groups based on the searching at each point each selected attribute is modeled using NB and the remaining is modeled by DT. The results of each model are evaluated and probability is calculated.

Evaluation

Calculate the accuracy, precision, recall , f- measure, ROC,TP rate, FP rate. Compare the result for naïve bayes and hybrid approach.

(i) Accuracy-Accuracy is termed as correctly classified instances in percentage.

Accuracy= True positive+true negative/ (true pos+false pos+true neg+false neg).

(ii) Precision -Precision is the fraction of retrieved instances that is measure of exactness[20].

Precision=TP/(TP+FP).

(iii) Recall- Recall is the fraction of relevant instances that is retrieved that is measure of completeness i.e true positive rate of class[20].

Recall=TP/(TP+FP).

(iv) F-measure-F-measure is the harmonic mean of precision and recall[20].

F=2*precision*recall/(precision+recall).

(v) ROC Area-Reciever operating characteristics curve shows the relationship between false positives and true positives[20].

Confusion matrix- Confusion matrix is a matrix include information about actual and predicted classifications[20].

Parameters of confusion matrix

(i) TP rate- It indicates the number of records classified as true though they were actually true. Such as patients having heart disease correctly identified as heart disease.

(ii) FP rate- It denotes the number of records classified as true while they were actually false. Such as people who are healthy are incorrectly identified as heart disease.

(iii) TN- It denotes the number of records classified as false while they were actually false. Such as people who are healthy are correctly identified as healthy.

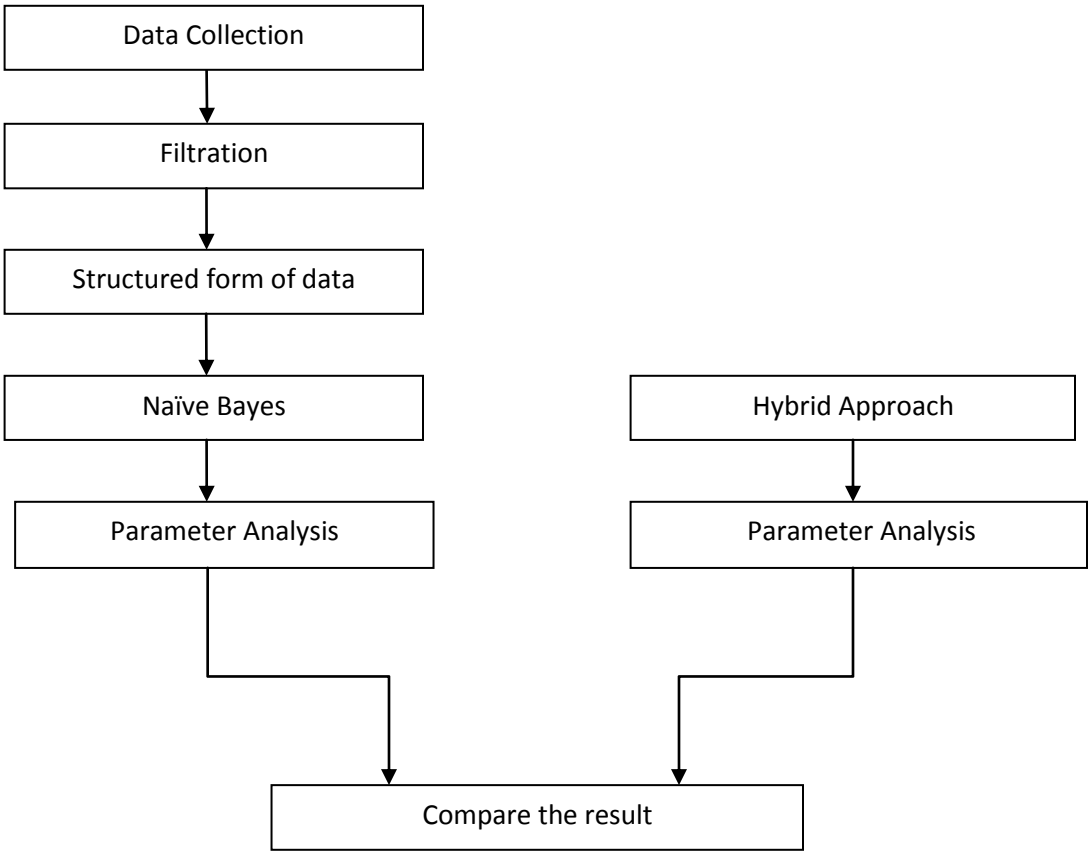
(iv) FN- It denotes the number of records classified as false while they were actually true. Patient having heart disease are incorrectly identified as healthy.

Table 1 Attribute Description:

Name	Type	Description
Age	Continuous	Age of the patient in years
Sex	Discrete	Value for male=1 Value for female=0
Cp	Discrete	Chest pain type typical angina value is 1 atypical angina value is 2 non-angina value is 3 asymptomatic value is 4
Trestbps	Continuous	Resting blood pressure(in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120mg/dl True value is 1 False value is 0
Restecg	Discrete	Resting electrocardiographic results Normal indicate value= 0 Having ST-T wave abnormality indicate value=1 Showing probable or define left ventricular hypertrophy by estes

		criteria value is =2
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina Yes indicates value =1 No indicates value=0
Slope	Discrete	The slope of the peak exercise segment up sloping value=1 flat value=2 down sloping value=3
Oldpeak	Continuous	ST depression induced by exercise
Thal	Discrete	Normal represent value=3 fixed defect represent value=6 reversible defect represent value=7
CA	Discrete	Number of major vessels colored by floursopy(0-3)
Class attribute	Discrete	Diagnosis classes Present= having heart disease Absent=not having heart disease

Fig 1 Flow work of proposed system



Chapter 4

RESULTS AND DISCUSSION.

Graphical representation(GUI) of the system.

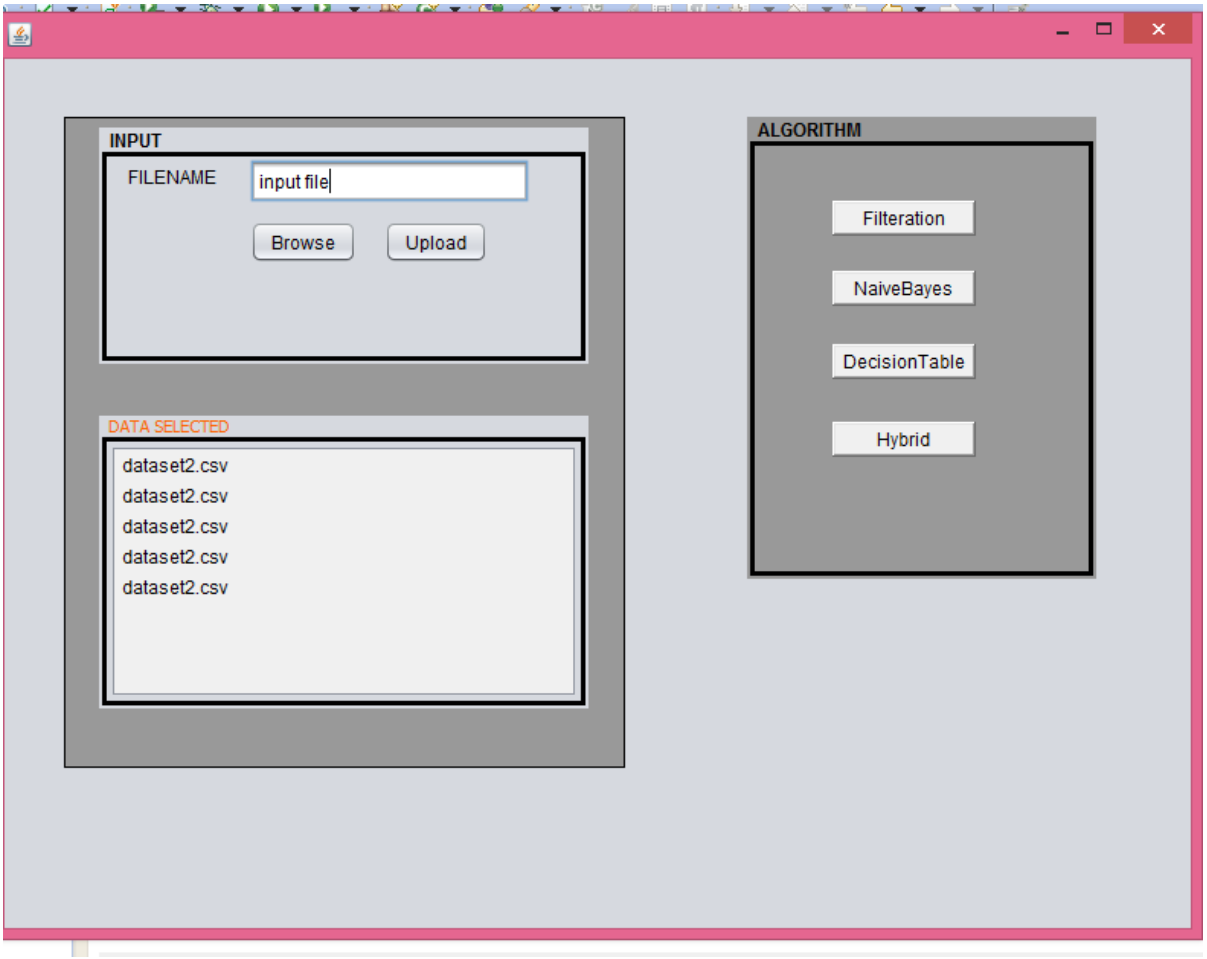


Figure 2 Snapsnot of Gui.

We have build this GUI using Net Beans IDE 7.1.2.

For browsing the dataset in the GUI.

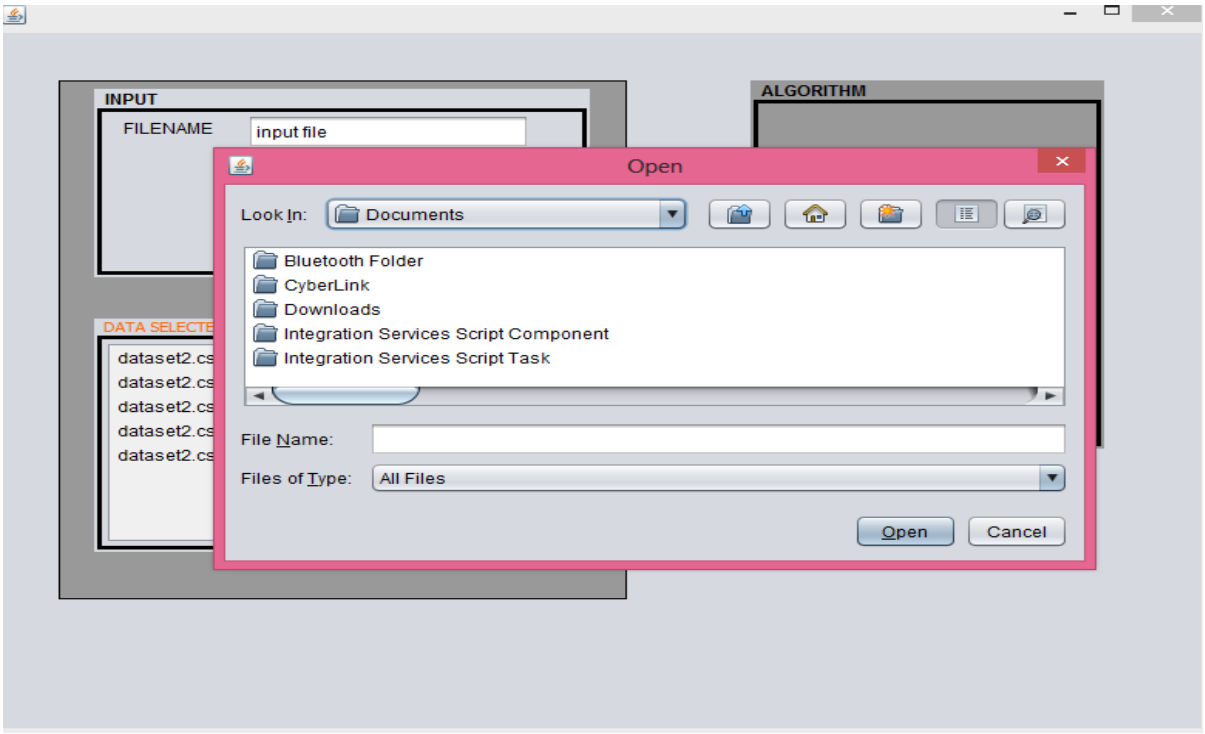


Figure 3 Snapshot of Browsing the Dataset.

As after browsing the dataset upload it.

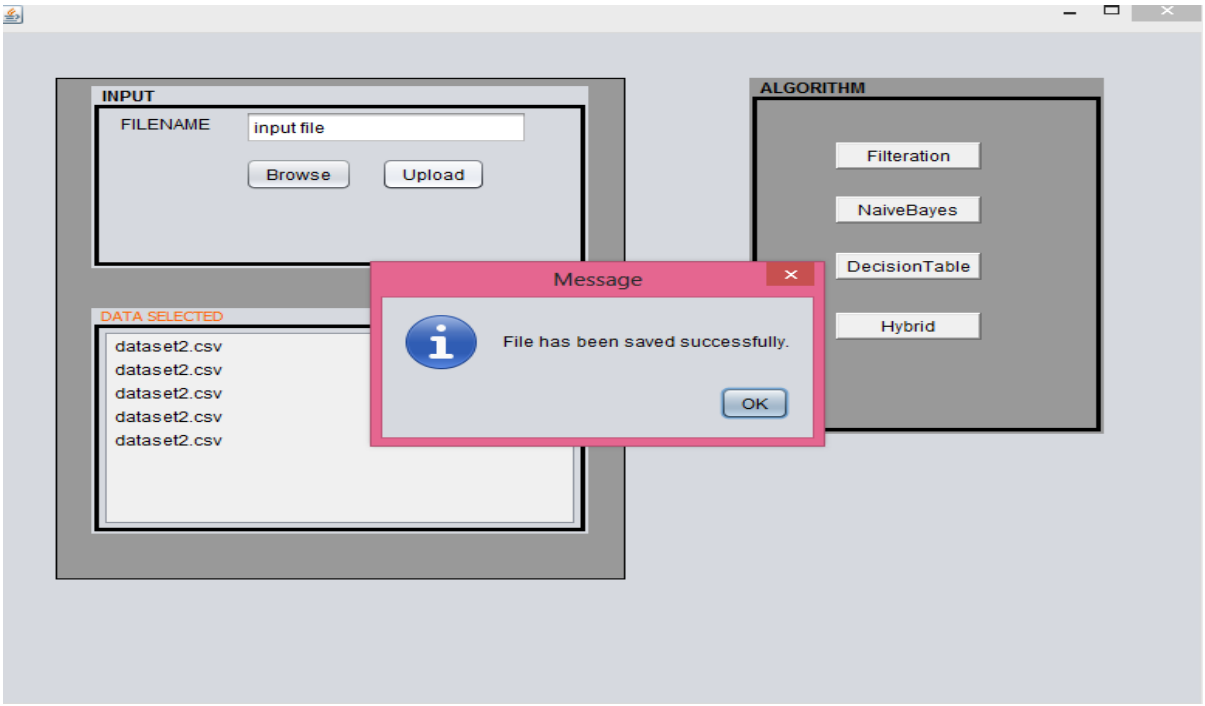


Figure 4 Snapshot of Uploading the Dataset

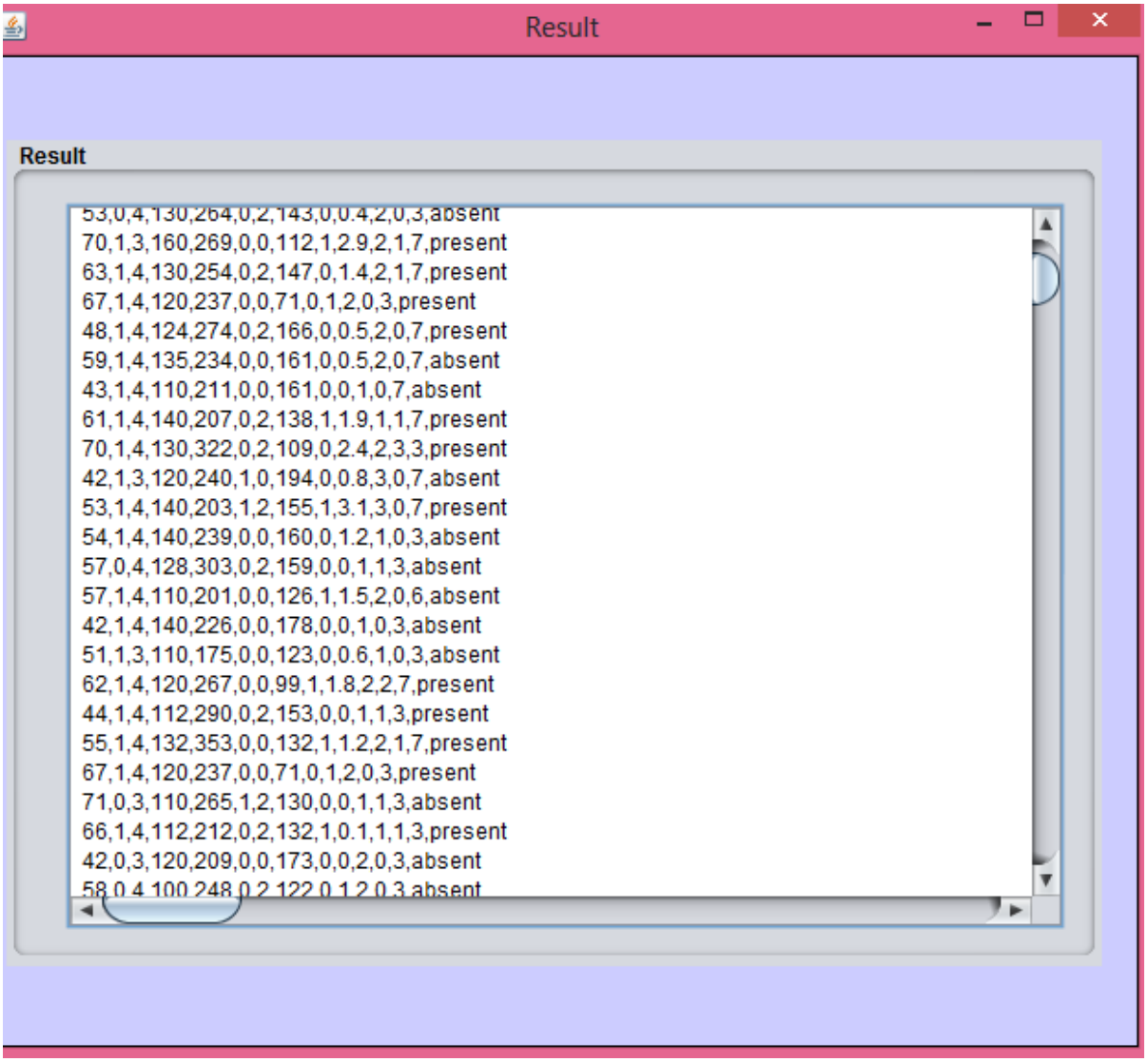


Figure 5 Snapshot of Result after filtering the dataset.

Here, analysis on heart data set is done in WEKA tool based on each attributes and also distributing the values shown as follow.

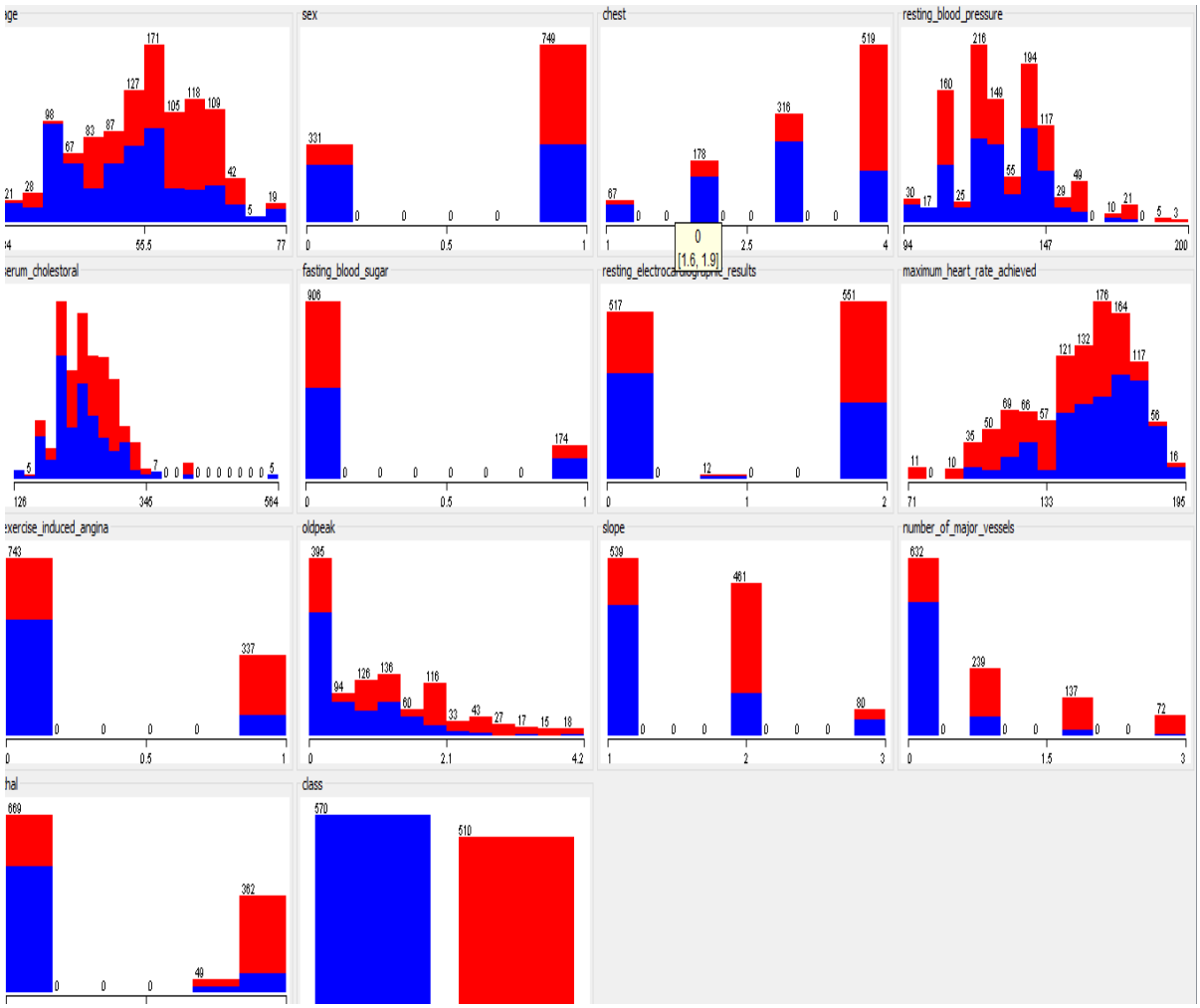


Figure 6 Visualization of the heart patients

Here we have three different algorithms to compare the results that are Naïve bayes, Decision table, hybrid i.e combination of both naïve bayes and decision table. This algorithm is apply on heart dataset inWEKA tool.

Here, we can see the performance of the algorithm i.e how much accurate result is provided by algorithms.

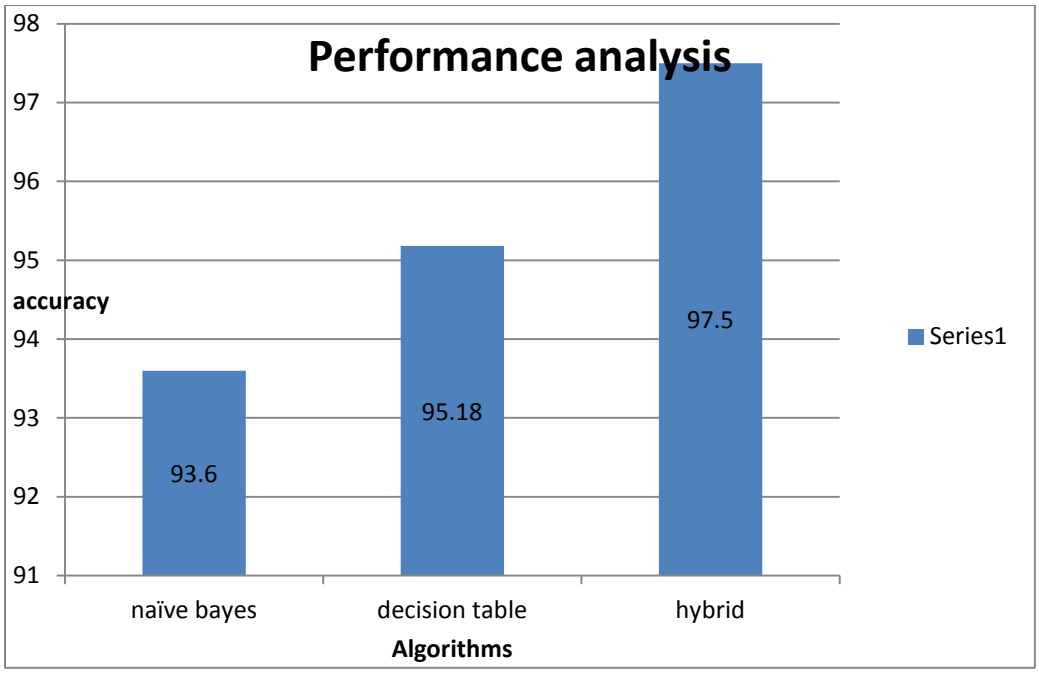


Figure 7 Performance analysis based on accuracy

The dataset contains 1080 instances and on which classification algorithm is applied to measure the performance of algorithm that from these many instances how many are correctly classified instances with the help of WEKA tool.

Table 2 Number of instances that are correctly classified

Name of Algorithm	**Number of correct instances**	**Accuracy**
Naïve bayes	1011	93.6%
Decision table	1028	95.18%
Hybrid	1053	97.5%

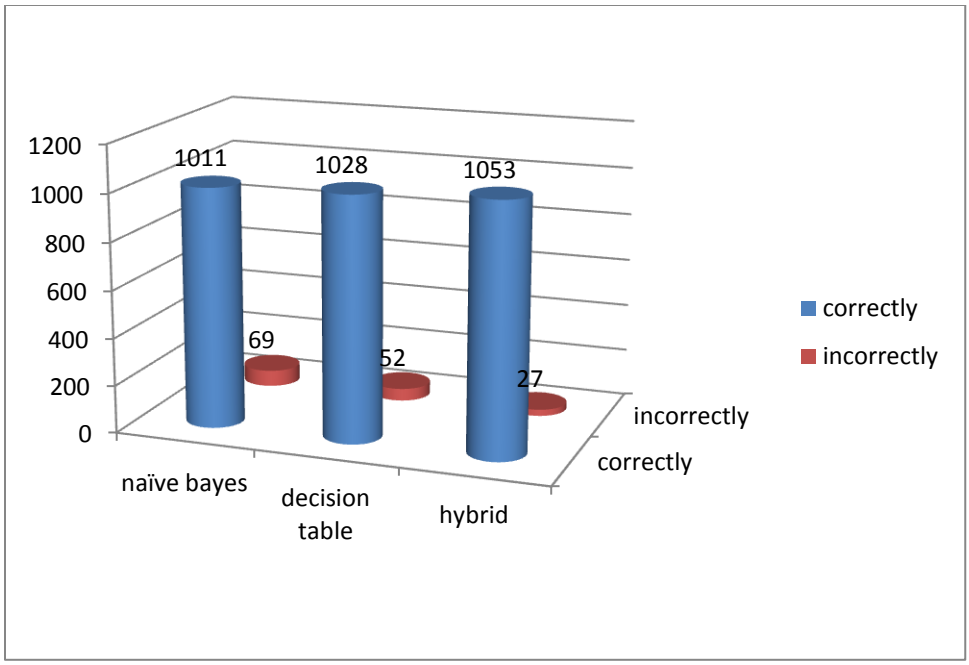


Figure 8 Efficiency of different models based on correctly classified instances.

Table 3 Detailed comparison of accuracy by the class attribute

Evaluation Criteria	**Classifiers**		
	Naïve Bayes	Decision Table	Hybrid
Kappa statistics	0.8715	0.902	0.949
Mean absolute error	0.0821	0.204	0.147
Root mean squared error	0.2209	0.241	0.199
Relative absolute error	0.1646	0.410	0.296
Root relative squared error	0.442	0.483	0.399

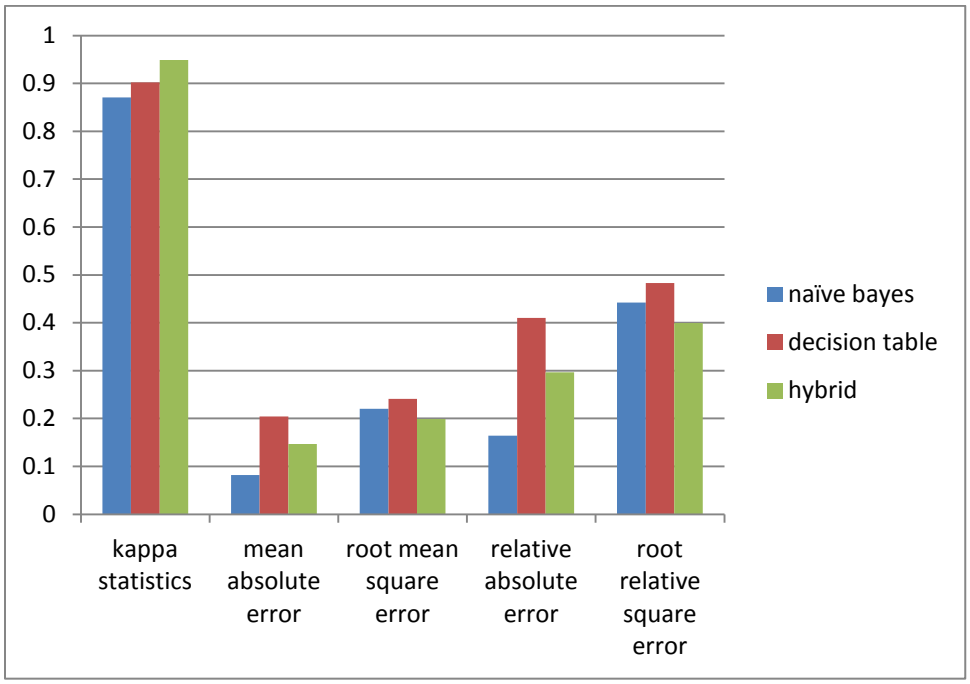


Fig 9 Evaluation of parameter.

For all classifiers detailed accuracy were calculated which includes parameter as TP rate, FP rate, Precision, Recall, F-measure and ROC area.

Table 4 Comparison of accuracy measures.

Classifiers	**TP**	**FP**	**Precision**	**Recall**	**F-measure**	**ROC**	**Class**
Naïve Bayes	0.96	0.09	0.922	0.96	0.941	0.986	Absent
	0.91	0.04	0.953	0.91	0.931	0.986	Present
Decision Table	1	0.102	0.916	1	0.956	0.994	Absent
	0.898	0	1	0.898	0.946	0.994	Present

Hybrid	0.978	0.028	0.976	0.978	0.977	0.996	Absent
	0.972	0.022	0.974	0.972	0.973	0.996	Present

Confusion matrix are very useful for evaluating classifiers.The columns represent the predictions and the rows represent the actual class.confusion matrix of each classifier is shown as follows:

Table 5 Confusion matrix for naïve bayes

		Predicted class	
		Absent	Present
Actual class	Absent	547	23
	Present	46	464

Table 6 Confusion matrix for decision table

		Predicted class	
		Absent	Present
Actual class	Absent	570	0
	Present	52	458

Table 7 Confusion matrix for hybrid

		Predicted class	
		Absent	Present
Actual class	Absent	572	13
	Present	14	481

Chapter-5

CONCLUSION AND FUTURE SCOPE

Day by day healthcare data is increasing and having this huge amount of data that is being difficult to manage so mining techniques apply on it. We proposed a Heart disease prediction system that provides the important tool for physicians to take decisions from this huge and mined data for analysis based on previous data. The research undertake an experiment on application of various data mining algorithms to predict the heart attack and to compare the best method of prediction. Different classification algorithms is used to analyze on heart disease patient data, it will check all the symptoms to predict the presence of heart disease and also measure the accurate result based on the performance of the algorithm. The predictive accuracy determined by Naïve Bayes, Decision Table, Hybrid algorithms are measured and finds that hybrid provides best result while comparing with others. For the future study, analysis of heart disease patient based on the treatment and medicine provided by the doctors to find the best and effective treatment for the risky patient.

Chapter 6

LIST OF REFERENCES

I. Research Papers

[1] Abhishek Taneja (2013)"Heart Disease Prediction System Using Mining Techniques",Oriental Journal of Computer Science and Technology.

[2] Ahmed T. sadiq alobaidi,Noor thamer mahmood(2013)."Modified full Bayesian network classifiers for medical diagnosis*"* IEEE.

[3] David Cornforth,Mika Tarvainen,Herbert F.Jelinek. (2013)"Computational intelligence methods for the identification of early cardiac autonomic neuropathy", IEEE.

[4] G.Karthiga, C.Preethi, R.Delshi Howsalya Devi (2014) "Heart Disease Analysis Sytem Using Data Mining Techniques", International Conference on Innovations in Engineering and Technology IEEE

[5] Hanaa Elshazly,Ahmed taher azar,Abeer el-korany and aboul ella hassanien. (2013) ."Hybrid System for lymphatic diseases diagnosis", IEEE.

[6] Hezlin Aryani Abd Rahman,Yap Bee Wah. (2012)"Comparison of predictive models to predict survival of cardiac surgery patients"

[7] Hian Chye Koh and Gerald Tan (2012) "Data Mining Applications in Health care",Journal of healthcare information management vol.19,No.2

[8] Hlaudi Daniel Masthe,Mosima Anna Masethe. (2014) "Prediction of heart disease using classification algorithms

[9] Lin Li, Saeed Bagheri ,Helena Goote (2013) "Risk adjustment of patient expenditures ,Philips Research North America Briarcliff Manor, US IEEE

[10] Mai Shouman, Tim Turner ,Rob Stocker (2012) "Using Data Mining Techniques In Heart Disease Diagnosis And Treatment", School of Engineering and Information Technology University of New South Wales At the Australian Defence Force Academy Northcott Drive,ACT 2600 IEEE.

[11] M.Akhil jabbar,Priti Chandra,B.L Deekshatulu (2012) "Prediction of Risk Score for Heart Disease Using Associative Classification and Hybrid Feature Subset Selection", Aurora's Engineering College Bhongir A.P,India,Advanced System Laboratory Hyderabad,IDRBT,RBI(Govt of INDIA) Hyderabad, IEEE.

[12] Mohammad Taha Khan,Dr. Shamimul Qamar and Laurent F.Massin. (2012) "A prototype of cancer/heart disease prediction model using data mining "

[13] Mythili, Dev Mukherji,Nikita Padalia,and Abhiram Naidu. (2013) "Heart disease prediction model using SVM Decision tree logistic regression".

[14] Ranganatha S., Pooja Raj H.R,Anusha C, Vinay S.K (2013) "Medical data mining and analysis for heart diseases dataset using classification techniques", Govt .Engineering College, Hassan INDIA,PES Institute of Technology, Banglore, INDIA

[15] R.Chitra and V. Seenivasagam (2013) "Review of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques", Department of Computer Science and Engineering,Noorul Islam Center for Higher Education ,India.ICTACT Journal on Computing.

[16] Saba Bashir,Usman Qamar,M.Younus Javed. (2014) "Ensemble based decision support framework for intelligent heart disease diagnosis", IEEE.

[17] Tina R. Patil,Mrs.S.S.Sherekar (2013) "Performance analysis of Naïve bayes and J48 classification algorithm for data classification".

[17] Vikas Chaurasia (2013) ”Early Prediction of Heart Diseases Using Data Mining Techniques”,Carib.j. Science technology,vol.1.

[18] V.Manikantan & S.Lanthan (2013) ”Predicting The Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods”, Department of Computer Science and Engineering, Mahendra Institute of Technology Tiruchengode, Namakkal, India, ISSN ,Vol -2.

[20] Xiao Fu,Yinzi Guiqiu,Qing Pan. (2011) ”A Computational modelfor heart failure stratification”

II.Websites

[21] http://docs.oracle .com/cd/B28359_01/datamine.111/b28129/algo-decisiontree.htm

#DmC0N019

[22] http://docs .oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#110057465

[23] http://stats.stackexchange.com/questions/23490/why-do-naivebayesian classifiers

[24] http://www.techopedia.com/defination/18829/decisiontabledetab

[25] http://www.cdc.gov/dhdsp/action_plan/pdf/action_planfull.pdf

[26] http://www.nimh.nih.gov/health/publications/depressionandheart-disease/depressionandheartdisease_142318.pdf

[28]http://www.gov.uk/government/uploads/system/uploads/attachment_data/file/217118/93872900853-cvd-outcomes-web1.pdf

Chapter-7

APPENDIX

Abbreviations

WEKA - Waikato Environment for Knowledge Analysis

SVM – Support Vector Machine

HRV-Heart rate variability

BN-Bayesian network

CAN-Cardiac autonomic neuropathy

LR-Logistic regression

CVD-Cardio Vascular disease

DT-Decision table

NB-Naïve Bayes

ID3- Iterative Dichotomiser 3

CART-Classification and regression tree

EBM-Evidence based medicine