Heart Disease Prediction Using Machine Learning and Data Mining Techniques

Submitted By

Jaymin Patel

12MCEI34

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2015

Heart Disease Prediction Using Machine Learning and Data Mining Techniques

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering

(Information and Network Security)

Submitted By

Jaymin Patel

(12MCEI34)

Guided By

Prof. Tejal Upadhyay

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2015

Certificate

This is to certify that the major project entitled **"Heart Disease Prediction Using Machine Learning and Data Mining Techniques"** submitted by **Jaymin Patel (Roll No: 12MCEI34)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Information & Newtwork Security (CSE),Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this project, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof. Tejal Upadhyay
Guide & Assistant Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Sharda Valeveti
Associate Professor,
Coordinator M.Tech - CSE(INS)
Institute of Technology,
Nirma University, Ahmedabad

Dr. Sanjay Garg
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. K Kotecha
Director,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Jaymin Patel**, Roll. No. **12MCEI34**, give undertaking that the Major Project entitled "**Heart Disease Prediction Using Machine Learning and Data Mining Techniques**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made. It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Signature of Student

Date:

Place:

Endorsed by

Prof. Tejal Upadhyay

(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Prof. Tejal Upadhyay**, Assistant Professor, Computer Science Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank **Dr. Sanjay Garg**, Hon'ble Head of Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr. K Kotecha**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

See that you acknowledge each one who have helped you in the project directly or indirectly.

- Jaymin Patel

12MCEI34

Abstract

Heart disease is that the main reason for death within the world over the last decade. Researchers are victimisation many data mining techniques to assist health care professionals in the diagnosing of Heart disease. However Data mining technique can reduce the number of test that are required. In order to reduce number of deaths from heart diseases there have to be a quick and efficient detection technique. Decision Tree is one in every of the effective data processing ways used. This research compares different algorithms of Decision Tree classification seeking better performance in heart disease diagnosis using R studio. The algorithms which are tested is J48 algorithm,Logistic Model Tree algorithm,Random Forest algorithm,Support vector machine,and K Nearest neighbour. The existing datasets of heart disease patients from Cleveland database of UCI repository is used to take a look at and justify the performance of call tree algorithms.This dataset consists of 618 instances and 76 attributes. Subsequently, the classification rule that has optimum potential are advised to be used in sizeable information. The goal of this study is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases and to predict the presence of heart disease in patients where this presence is valued from no presence to likely presence.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Heart disease is that the leading reason behind death within the world over the past ten years (World Health Organization 2007). The European Public Health Alliance reported that heart attacks, strokes and alternative circulatory diseases account for 41% of all deaths (European Public Health Alliance 2010). Several different symptoms are associated with heart disease, which makes it difficult to diagnose it quicker and better. Working on heart disease patients databases can be compared to real-life application. Doctors knowledge to assign the weight to each attribute. More weight is assigned to the attribute having high impact on disease prediction. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process. It also provides healthcare professionals an extra source of knowledge for making decisions.

The healthcare industry collects large amounts of healthcare data and that need to be mined to discover hidden information for effective decision making.Motivated by the world-wide increasing mortality of cardiovascular disease patients every year and therefore the convenience of giant quantity of patients information from that to extract helpful information, researchers are victimisation data processing techniques to assist health care professionals within the diagnosing of cardiovascular disease (Helma, Gottmann et al. 2000)[1].Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods (Lee, Liao et al. 2000).Data mining techniques can be

used for data selection,finding patterns and predict the diseases using large data Thus data mining refers to mining or extracting knowledge from large amounts of data. Data mining applications will be used for better health political and interference of hospital errors, early detection, interference of diseases and preventable hospital deaths (Ruben 2009). Mining process is more than the data analysis which includes classification, clustering,and association rule discovery.

Heart disease prediction system can assist medical professionals in predicting heart disease based on the clinical data of patients[2]. Hence by implementing a heart disease prediction system using Data Mining techniques and doing some sort of data mining on various heart disease attributes, it can able to predict more probabilistically that the patients will be diagnosed with heart disease. This paper presents a new model that enhances the Decision Tree accuracy in identifying heart disease patients. It uses the different algorithm of Decision Trees.

1.2 Project Definition

Nowadays, Health care industry contains huge amount of heath care data and these healthcare data contains hidden information. This hidden information is useful for making effective decisions using different data mining techniques. We can Develop an efficient decision making for patients who will be suffering from disease and using this we can identify patient for improve their health.

1.3 Objective of Study

Identify the patients who will be suffering from disease using Dataset.For using this find out the different techniques for better efficiency and accurate the prediction.

- To Identify key patterns or features from the dataset.
- To Identify and select attributes that are more relevant in relation to Heart diseases.

Chapter 2

Literature Survey

Prediction of heart disease using data mining techniques has been an ongoing effort for the past two decades. Most of the papers have implemented several data mining techniques for designation of cardiovascular disease like Decision Trees, Neural networks, Regression, Naive Bayesian classifiers,and Support vector machine showing totally different levels of accuracies (Yan, Zheng et al. 2003; Andreeva 2006; Das, Turkoglu et al. 2009; Sitar-Taut, Zdrenghea et al. 2009; Raj Kumar and Reena 2010; Srinivas, Rani et al. 2010) on multiple databases of patients from around the world[3].

One of the bases on which the papers differ are the selection of parameters on which the methods have been used. Many authors have specified different parameters and databases for testing the accuracies. In particular, researchers have been investigating the application of the Decision Tree technique in the diagnosis of heart disease with considerable success. Sitair-Taut et al. used the weka tool to investigate applying Naive Bayes and J48 Decision Trees for the detection of coronary heart disease. Tu et al. used the bagging algorithm in the weka tool and compared it with J48 Decision Tree in the diagnosis of heart disease[4].In the decision making process of heart disease is effectively diagnosed by Random forest algorithm.In supported the likelihood of call support, the center malady is foreseen[5].As a result the author complete that call tree performs well and generally the accuracy is comparable in Bayesian classification.

In year 2013, S. Vijiyarani et. al. [6] performed a work, An Efficient Classification Tree Technique for Heart Disease Prediction. This paper analyzes the classification tree techniques in data mining. The classification tree algorithms used and tested in this paper are Decision Stump, Random Forest and LMT Tree algorithm. The objective of

this research was to compare the outcomes of the performance of different classification techniques for a heart disease dataset.

Table 2.1: Literature Survey

Author	Year	Techniques	Suggestion
M.Akhil jabbar,B.L Deekshatulu,Priti Chandra [3]	2013	Propose a new algorithm which combines KNN with genetic algorithm for effective classification.	Give various weights to the presence of condition for other disease
Jesmin Nahar,Tasadduq Imam,Kevin S. Tickle [5]	2013	Support Vector Machine Algorithm	Apply Support Vector Machine for automated feature selection and a medical knowledge based motivated feature selection process.
Hongmei Yan,Jun Zheng,Yingtao Jiang [2]	2008	Genetic algorithm	Not depends on only one attribute for accurate prediction.
Mu-Jung Huang,Mu-Yen Chen,Show-Chin Lee [6]	2007	Decision Tree Algorithm	Systematical method of integrating DM techniques with Case-based reasoning.
Yanwei Xing, Yonghong Gao [7]	2007	Combination data mining methods	Using that Combination data mining methods the system give good accuracy,
Yuehjen E. Shao, Chia-Ding Hou,Chih-Chou Chiub [1]	2014	Hybrid intelligent modelling schemes	There should be techniques for finding missing value like admission,emergency to better utilized for predict risk
S. Muthukaruppan, M.J. Er [8]	2010	fuzzy expert system	For use only longitudinal data set it can not increase predict power of data.

Table 2.2: Literature Survey

Author	Year	Techniques	Suggestion
Saroj Kr. Biswas , Nidul Sinha, Biswajit Purakayastha, Leniency Marbaniang [9]	2014	Neural network for classification	Used Machine Learning techniques for Diseases
Khaled El Emam ,Luk Arbuckle[9]	2010	Identify the risk factor(high,low) for patients using rule based algorithm	It should be deal with data cleansing and reduction of variables
Maria C,John C [5]	2009	Patients identification based on the Remedial Risk	Not depended only on algorithm also depends on the attributes of the data set.
Alex Bottle,Paul Aylin [3]	2006	Logistic regression analysis for hospital readmission	It is impossible that reduce no. of variables without loss of information so make sure that it contains the maximum amount of information.
Haifeng Xie, Peter H. Millard [10]	2006	Time window based on the risk (high risk,low risk) factor	It should be deal with missing value

2.1 Background

Millions of people are getting some sort of heart disease every year and heart disease is the biggest killer of both men and women in the United States and around the world. The World Health Organization (WHO) analyzed that twelve million deaths occurs worldwide due to Heart diseases. In almost every 34 seconds the heart disease kills one person in world.

Medical diagnosis plays vital role and yet complicated task that needs to be executed efficiently and accurately. To reduce cost for achieving clinical tests an appropriate computer based information and decision support should be aided. Data mining is the use of software techniques for finding patterns and consistency in sets of data. Also, with the advent of data mining in the last two decades, there is a big opportunity to allow computers to directly construct and classify the different attributes or classes.
Learning of the risk components connected with heart disease helps medicinal services experts to recognize patients at high risk of having Heart disease.Statistical analysis has identified risk factors associated with heart disease to be age, blood pressure, total cholesterol, diabetes, hyper tension, family history of heart disease, obesity and lack of physical exercise, fasting blood sugar etc [8].

Researchers are applying totally different information mining Techniques to assist healthful services specialists with progressed exactitude within the judgement of Heart Disease.Neural network, Naive Bayes,Decision Tree etc. are some techniques employed in the identification of Heart Disease.

Applying Decision Tree techniques has shown helpful accuracy within the identification of Heart disease.However aiding health care professionals within the identification of the worlds biggest killer demands higher accuracy. Our analysis seeks to boost identification accuracy to boost health outcomes.

Decision Tree is one of the data mining techniques that cannot handle continuous variables directly so the continuous attributes must be converted to discrete attributes. Couple of Decision Tree use binary discretization for continuous-valued features. Other important accuracy improving is applying reduced error pruning to Decision Tree in the diagnosis of heart disease patients[11]. Intuitively, more complex models might be expected to produce more accurate results, but which techniques is best? Seeking to

thoroughly investigate options for accuracy improvements in heart disease diagnosis this paper systematically investigates comparing multiple classifiers decision tree technique.

This research uses R studio.The information of UCI repository regularly introduced in a database or spreadsheet. In order to use this data for R studio tool, the data sets need to be in the CSV format (Comma-separated values file format). R studio tool is used for to preprocess the dataset. After reviewing all these 76 different attributes, the unimportant attributes is dropped and only the important attributes (i.e. 14 attributes in this case) is considered for analysis to yield more accurate and better results. The 14th one is basically a predicted attribute, which is referred as Class. With thorough comparison between different decision tree algorithms within R studio tool and deriving the decisions out of it, would help the system to predict the likely presence of heart disease in the patient and will definitely help to diagnose heart disease well in advance and able to cure it in right time.

Chapter 3

Approach and Methodology

3.1 Approach and Methodology

The following objectives are set for this Heart Disease prediction.

- The prediction system should not assume any prior knowledge about the patient records it is comparing.
- The chosen system must be scalable to run against large database with thousands of data.

This chosen approach is implemented using R studio tool. R studio is an open source software tool which consists of an accumulation of machine learning algorithms for Data Mining undertakings. It contains apparatuses for information preprocessing, classification, regression, clustering, association rules, and visualization [12]. For testing, the classification tools and explorer mode of R studio are used. Decision Tree classifiers with Cross Validation 10-fold in Test mode is considered for this study.
The following steps are performed in R studio.

- Start the R studio Explorer.
- Open CSV dataset file.
- Write the program for J48.
- Start the program and it gives result.

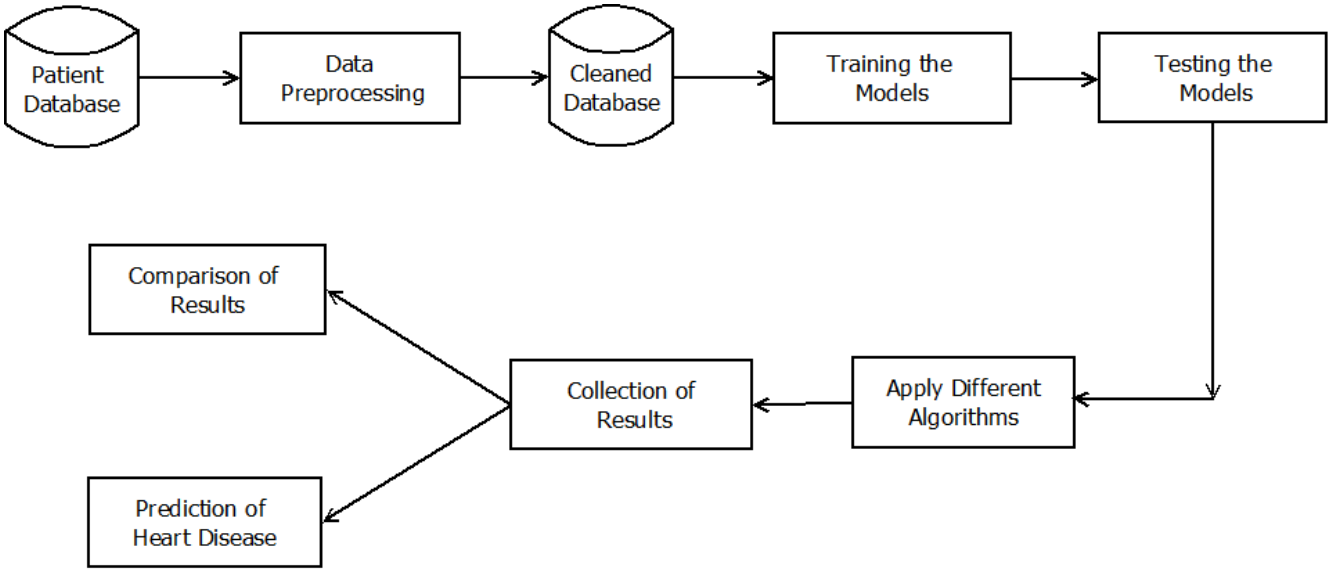

Figure 3.1: Proposed Framework

Our projected approach is use KNN,SVM,Random forest,J48 and LMT algorithmic program to enhance the classification accuracy of heart disease knowledge set. we tend to used genetic search as a goodness live to prune redundant and irrelevant attributes, and to rank the attributes that contribute additional towards classification. Least hierarchical attributes are removed, and classification algorithmic program is constructed supported evaluated attributes. This classifier is trained to classify Heart disease knowledge set as either healthy or sick.

This paper has emphasized specifically on decision tree classifiers for heart disease prediction within R studio. Decision tree was considered here among all types of Data mining techniques due to these below reasons. Decision tree filters are easy to implement and easy to understand. It is a method commonly used in data mining.Decision tree is one of the data mining techniques demonstrating extensive achievement when contrasted with other data mining techniques. It is a decision support system that uses a tree-like graph decisions. Decision trees are the most powerful approaches in knowledge discovery and data mining. Decision trees are highly effective tools in many areas such as data and text mining, information extraction, machine learning, and pattern recognition. It can handle input data like Nominal, Numeric & Text. It is able to process erroneous datasets

or missing values.

A Decision Tree is used to learn a classification function which concludes the value of a dependent attribute (variable) given the values of the independent (input) attributes. This verifies a problem known as supervised classification because the dependent attribute and the counting of classes (values) are given [9]. Tree complexity has its effect on its accuracy. Usually the tree complexity can be measured by a metrics that contains: the total number of nodes, total number of leaves, depth of tree and number of attributes used in tree construction. Tree size should be relatively small that can be controlled by using a technique called pruning [9].

Univariate decision tree approach will be used here. In this technique, splitting is performed by using one attribute at internal nodes. This study can able to distinguish the dominant attributes and provides different labels of Likely Presence for heart disease. In this paper, three decision tree algorithms namely J48 algorithm, logistic model tree algorithm and Random Forest decision tree algorithm are used for comparison. The proposed methodology involves reduced error pruning, confident factor and seed parameters to be considered in the diagnosis of heart disease patients. Reduced error pruning has shown to drastically improve decision tree performance. These three decision tree algorithms are then tested to identify which combination will provide the best performance in diagnosing heart disease patients.

A correlation is predicated on affect ability, specificity and exactitude by real positive and false positive in confusion matrix. to possess an affordable correlation between these algorithms, getting ready time in seconds and tree size proportion for each system is considered with 10-fold stratified cross validation. The general approach took after for Decision Tree classification for satisfying the objective is

$$Training => Algorithm => Model => Testing => Evaluation$$

DATA SOURCE

For comparing various Decision Tree classification techniques, Cleveland dataset from UCI repository is used, which is available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. The dataset has 76 attributes and 618 records. However, only 13 attributes are used for this study & testing as shown in Table 3.1.

Table 3.1: Selected Heart Disease Attributes

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	0 = female 1 = male
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 =asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar >120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic results
Exang	Discrete	Exercise induced angina: 1 = Yes 0 = No
Thalach	Continuous	Maximum heart rate achieved
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3 = down sloping
Ca	Continuous	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6 = fixed defect 7= reversible defect
Class	Discrete	Diagnosis classes: 0 = No Presence 1=Least likely to have heart disease 2= >1 3= >2 4=More likely have heart disease

3.2 Classification Tree Algorithms Used

3.2.1 J48 algorithm:

This algorithm utilizes an avaricious method to make decision trees for classification and uses decreased-error pruning [10]. Decision tree is built by examining information hubs, which are utilized to assess hugeness of existing highlights.J48 algorithm is an extension of ID3 algorithm and possibly creates a small tree. It uses divide and conquers approach to growing decision trees[7].At every node of the tree, the algorithm picks a attribute that can further part the samples into subsets.Every leaf node speaks to a class or decision. Basic steps to construct tree are :

- Check whether all cases belongs to same class, then the tree is a leaf and is labelled with that class.
- For each attribute, calculate the information and information gain.
- Find the best splitting attribute (depending upon current selection criterion).

J48 with Reduced error Pruning:
Pruning is very important technique to be used in tree creation because of outliers. It also addresses over fitting.Datasets may contain little subsets of instances that are not well defined. To classify them correctly, pruning can be used. Separate and Conquer rule learning algorithm is basis to prune any tree. This rule learning scheme starts with an empty set of rules and the full set of training instances. Reduced-error pruning is one of such separate and conquer rule learning scheme. There are two types of pruning i.e.

- Post pruning (performed after creation of tree)
- Online pruning (performed during creation of tree).

After extracting the choice tree rules, reduced error pruning was wont to prune the extracted call rules. Reduced error pruning is one among the quickest pruning strategies and legendary to provide each correct and little call rules (Esposito, Malerba et al. 1997)[13]. Applying reduced error pruning provides a lot of compact call rules and reduces the amount of extracted rules.
The run-time complexity of J48 algorithm matches to the tree depth which is linked to

tree size and number of examples. So their greatest disadvantage is size of J48 trees, which increases linearly with the number of examples. J48 rules slow for large and noisy datasets. Space complexity is very large as we have to store the values repeatedly in arrays.

3.2.2 Support vector machine :

The SVM is a condition of-the-workmanship most extreme edge grouping calculation established in measurable learning hypothesis. SVM performs grouping assignments by amplifying the edge dividing both classes while minimizing the grouping mistakes. We utilized consecutive negligible advancement calculation to prepare the SVM.

A support vector machine could be a variety of model wont to analyse knowledge and see patters in classification and multivariate analysis. Support vector machine (SVM) is employed once your knowledge has specifically two categories.SVM categories knowledge by finding the simplest hyper plane that separates all knowledge points of one category from those of the opposite class. The larger margin between the two categories, the higher the model is a margin should don't have any points in its interior region. The support vectors are the info points that on the boundary of the margin. SVM relies on mathematical functions and wont to model advanced, and planet issues[14]. SVM performs well on knowledge sets that have several attributes, like the CHDD.

Support Vector Machines map the coaching knowledge into kernel area. There are several otherwise used kernel areas linear (uses dot product), quadratic, polynomial, Radial Basis operate kernel, Multilayer Perception kernel, etc. to call a number of. additionally, there are multiple strategies of implementing SVM, like quadratic programming, ordered lowest improvement, and statistical method. The difficult side of SVM is kernel choice and technique choice specified your model isn't over optimistic or pessimistic.

Considering that the CHDD features a sizeable amount of instances yet as options, it's controversial whether or not the kernel chosen is RBF or linear. though the relation between the attributes and sophistication labels are non-linear, owing to the big variety of options, RBF kernel might not improve performance. it's suggested that each kernels be tested and therefore the additional economical one be finally selected .

3.2.3 K nearest neighbour(KNN):

K nearest neighbour(KNN) may be a straightforward formula, that stores all cases and classify new cases supported similarity measure.KNN formula additionally referred to as

- case primarily based reasoning
- k nearest neighbour
- example based reasoning
- instance primarily based learning
- memory based reasoning
- lazy learning.

KNN algorithms are used since 1970 in several applications like applied mathematics estimation and pattern recognition etc.KNN may be a non constant classification technique that is generally classified into 2 types

- structure less NN techniques
- structure based NN techniques.

In structure less NN techniques whole information is assessed into coaching and take a look at sample information.From training purpose to sample purpose distance is evaluated, and therefore the purpose with lowest distance is named nearest neighbor.Structure primarily based NN techniques area unit supported structures of information like orthogonal structure tree (OST), ball tree, k-d tree, axis tree, nearest future line and central line [1].Nearest neighbor classification is employed in the main once all the attributes area unit continuos.Simple K nearest neighbor formula is shown in below

- find the K training instances which are closest to unknown instance
- pick the most commonly occurring classification for these K instances

There square measure numerous ways that of mensuration the similarity between 2 instances with n attribute values. Every measure has the subsequent 3 necessities.Let dist (A, B) be the distance between two points A,B then

- dist(A,B) >= and dist(A,B)=0 if A=B
- dist(A,B) = dist(B,A)
- dist (A,C) <= dist(A,B)+ dist(B,C)

Property 3 is termed as Triangle in equality. It states that the shortest distance between any a pair of points is also a line. commonest distance measures used is mathematician distance .For continuous variables Z score standardization and min soap standardisation unit of measurement used [1].KNN is employed in several applications like

- classification and interpretation
- downside solving
- operate learning and teaching and coaching

KNN suffers from the subsequent drawbacks

- low potency
- dependency on the selection of excellent values for k.

Further analysis is needed to enhance the accuracy of KNN with smart values of K.

3.2.4 Logistic Model Tree Algorithm:

Logistic Model Tree is the classifier for building logistic model trees, which consist of a decision tree structure with logistic regression function at the leaves.The algorithm can oversee parallel and multi-class target variables, numeric and onominal attributes and missing qualities[15]. A combination of learners that place confidence in regression models if solely very little and/or droning information is on the market and add a a lot of advanced tree structure if there's enough information to warrant such a structure. LMT uses cost-complexity pruning. This algorithmic rule is considerably slower than the opposite algorithms[16].

As in decision tree, the tested attributes is associated with every inner node. The attributes with k values, the node has k child nodes for nominal attributes and depending on the value of the attribute, the instances are sorted down. For the attributes of numeric, the node has two child nodes and comparing the attributes of tested value to a threshold (the instances are sorted down based on threshold [17].

Logistic Model Trees have been demonstrated to be extremely exact furthermore, smaller classifiers in diverse examination regions. Their most noteworthy weakness is the computational unpredictability of inciting the logistic regression models in the tree. Anyway the prediction of a model is acquired by sorting it down to a leaf what's more, utilizing the logistic prediction model connected with that leaf[18]. A solitary logistic model is less demanding to translate than J48 trees. However fabricating LMTs takes longer time. It can likewise be demonstrated that trees produced by LMT are much littler than those produced by J48.

To construct a provision model tree by developing a typical classification tree, building provision regression models for all node, pruning a proportion of the sub-trees utilizing a pruning model, and mixing the provision models on how into a solitary model in some manner is performed.

The pruning arrange uses cross-validation to induce a lot of steady pruning results. In spite of the very fact that this expanded the machine many-sided nature, it caused littler and for the foremost half a lot of correct trees. These thoughts result in the subsequent algorithmic program for developing logistical model trees:

Tree developing begins by building a supplying model at the basis utilizing the LogitBoost rule. the amount of cycles (and basic relapse capacities fmj to feature to Fj) is resolved utilizing ten fold cross-validation[19]. During this method the data is an element into getting ready and take a look at set ten times, for every preparation set LogitBoost is rush to a greatest variety of cycles and also the lapse rates on the take a look at set ar logged for every cycle and summed up over the distinctive folds. the amount of emphasess that has the smallest amount whole of blunders is employed to arrange the LogitBoost rule on all the data. this provides the supplying regression model at the bottom of the tree.

Like alternative tree efficacious systems, LMT doesn't oblige any standardisation of parameters. LMT produces a solitary tree containing double elements on numeric properties, multi-route elements on ostensible ones and supply regression models at the leaves, and therefore the rule guarantees that simply applicable attributes are incorporated within the last.

3.2.5 Random Forest Algorithm:

Random forest is an ensemble classifier that consists of many decision trees. The output of the classes is represented by individual trees. It is derived from random decision a forest that was proposed by Tin Kam Ho of Bell Labs in 1995 [17]. This method combines with random selection of features to construct a decision trees with controlled variations. The tree is constructed using algorithm as discussed.

- Let N be the number of training classes and M be the number of variables in classifier.
- The input variable m is used to determine the node of the tree. Note that m<M.
- Choosing n times of training sets with the replacement of all available training cases N by predicting the classes, estimate the error of the tree.
- Choose m variable randomly for each node of the tree and calculate the best split.
- At last the tree is fully grown and it is not pruned. The tree is pushed down for predicting a new sample. When the terminal node ends up, the label is assigned the training sample. This procedure is iterated over all trees and it is reported as random forest prediction.

Multi-classifiers are the after effect of joining a few individual classifiers. Troupes of classifiers towards expanding the execution have been presented[9].
Random Forest(RF) is one of the case of such procedures. RF as a multiclassifier formed by choice trees where each tree h_t had been created from the set of information preparing and a vector θ_t of arbitrary numbers indistinguishably disseminated and free from the vectors. Vectors θ_1, θ_2 ,.., θ_{t-1} used to create the classifiers $h_1, h_2, .., h_{t-1}$.Every decision tree is manufactured from random subset of the preparation dataset.[19] It utilized a random vector that is produced from some altered likelihood dissemination,where the likelihood circulation is shifted to center samples that are difficult to arrange. A Random vector can be joined into the tree-becoming process from various perspectives. The leaf hubs of each one tree are named by evaluations of the back dissemination over the information class names. Every interior hub contains a test that best parts the space of data to be arranged. Another, concealed occasion is ordered by sending it down every

tree and conglomerating the arrived at leaf appropriations.
There are three methodologies for Random Forest, for example, Forest-RI(Random Input choice) and Forest-RC (Random blend) and blended of Forest-RI and Forest-RC.
The Random Forest procedure has some desirable qualities, for example,

- It is not difficult to utilize, basic and effortlessly parallelized.
- It doesnt oblige models or parameters to choose aside from the quantity of indicators to pick at arbitrary at every node.
- It runs effectively on extensive databases; it is moderately strong to anomalies and commotion.
- It can deal with a huge number of information variables without variable deletion; it gives evaluations of what variables are important in classification.
- It has a successful system for assessing missing information and keeps up accuracy when a vast extent of the data are missing, it has methods for adjusting error in class populace unequal data sets.

Chapter 4

Evaluation of Classification Algorithms

4.1 Evaluation of Classification Algorithms

The execution of Classification formula is mostly analyzed by assessing the affectability,specificity, and accuracy of the classification.The sensitivity is proportion of positive instances that square measure properly classified as positive (i.e. the proportion of patients known to have the disease, who test positive for it). The specificity is that the proportion of negative instances that ar properly classified as negative (i.e. the proportion of patients far-famed to not have the illness, World Health Organization take a look at negative for it).The accuracy is that the proportion of instances that ar properly classified.To quantify the responsibility of the execution of planned model, the data is isolated into getting ready and testing data with 10-fold stratified cross validation These values ar outlined as,

Sensitivity = True Positive/(True Positive + False Negative)

Specificity = True Negative/(True Negative +False Positive)

Accuracy = (True Positive + True Negative) / (True Positive + True Negative+ False Negative+ False Positive)

All measures can be ascertained focused around four qualities specifically True Positive, False Positive, False Negative, and False Positive where,

- True Positive (TP) is varied effectively classified that associate degree instances positive.

- False Positive (FP) may be a variety of incorrectly classified that associate degree instance is positive.
- False Negative (FN) could be a range of incorrectly classified that AN instance is negative.
- True Negative (TN) could be a variety of properly classified that associate instance is negative.
- F-Measure could be a approach of mixing recall and preciseness scores into one live of performance.
- Recall is that the quantitative relation of relevant instances found within the search result to the entire of all relevant instances.
- Precision is that the proportion of relevant instances within the results came back.
- Receiver operational Characteristics (ROC) space may be a ancient to plot this same info in a very normalized kind with 1-false negative rate planned against the false positive rate.
- For every algorithm, the test choice cross-validation were utilized.As opposed to holding a vicinity for testing,the cross-validation repeats the coaching and testing method many times with random forest samples.The standard for this is often 10-fold cross-validation.The data is partitioned off willy-nilly into ten sections within which the categories area unit pictured within the same proportions as within the full dataset(stratification).Each one section is command out therefore and therefore the formula is trained on the 9 remaining parts; then its error rate is computed on the holdout set.[20]At long last, the ten error estimates area unit found the center price of to yield Associate in Nursing overall error estimate.For J48 and Random Forest, all the tests were run with 10 completely different random seeds.Choosing the various random seeds is dole out to traditional out applied math variations.

Chapter 5

Results

5.1 Results

The Decision tree classification was performed using J48 algorithm,Logistic model trees algorithm,Support Vector Machine.K-Nearest Neighbour and Random Forest algorithm on UCI repository.The experimental results is beneath the framework of R studio All experiment were performed on Core I3 with 2.4GHz CPU and 4GB RAM. The exploratory results area unit divided into many sub factor for fewer hard to please examination and assessment.

5.1.1 J48 Algorithm

The sample of J48 algorithm is connected on UCI repository and the confusion matrix is produced for class having 5 conceivable qualities are demonstrated in below.The confusion matrix is imperative viewpoint to be considered.From this matrix,classifications can be made.The results of the J48 algorithm are indicated in Table 2.

———Confusion Matrix———

a	b	c	d	e	
146	8	4	6	0	| a = 0
31	9	9	6	0	| b = 1
9	5	13	8	1	| c = 2
11	7	10	4	3	| d = 3
2	5	3	3	0	| e = 4

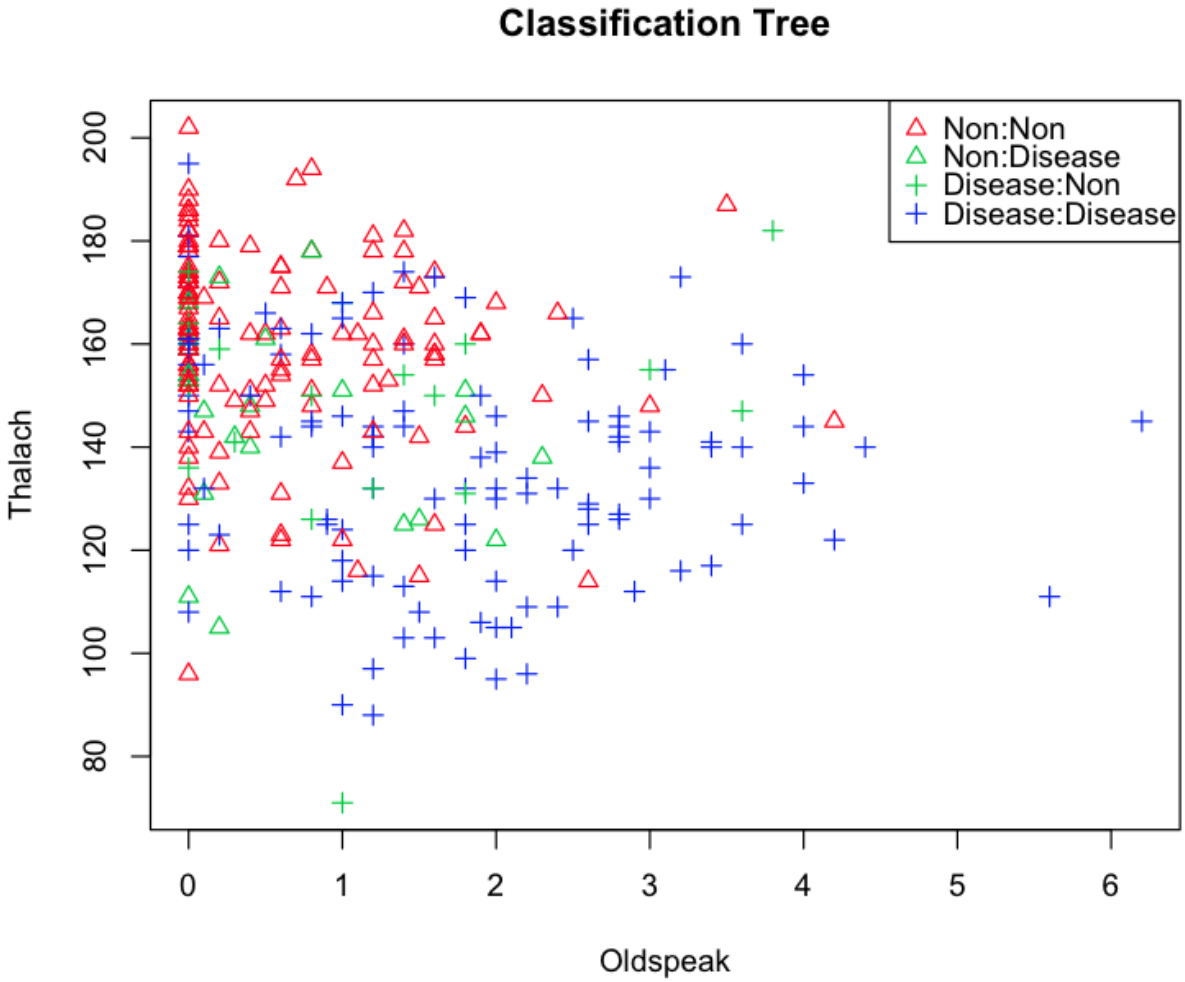

Figure 5.1: classification Algorithm

Table 5.1: Classification Result For J48

	Train Error	Test Error
J48	0.1423221	0.1666667

J48 model sacrifices error rate for a clearer decision process and as a result the error is acceptable.

5.1.2 Random Forest algorithm

The example of Random Forest rule is connected on UCI repository and also the confusion matrix is formed for sophistication having five qualities square measure incontestable in below. The results of the Random Forest rule square measure incontestable in Table 4.

————Confusion Matrix————

a	b	c	d	e	
152	7	2	3	0	| a = 0
34	4	10	5	2	| b = 1
10	11	7	7	1	| c = 2
5	11	12	5	2	| d = 3
1	5	2	3	2	| e = 4

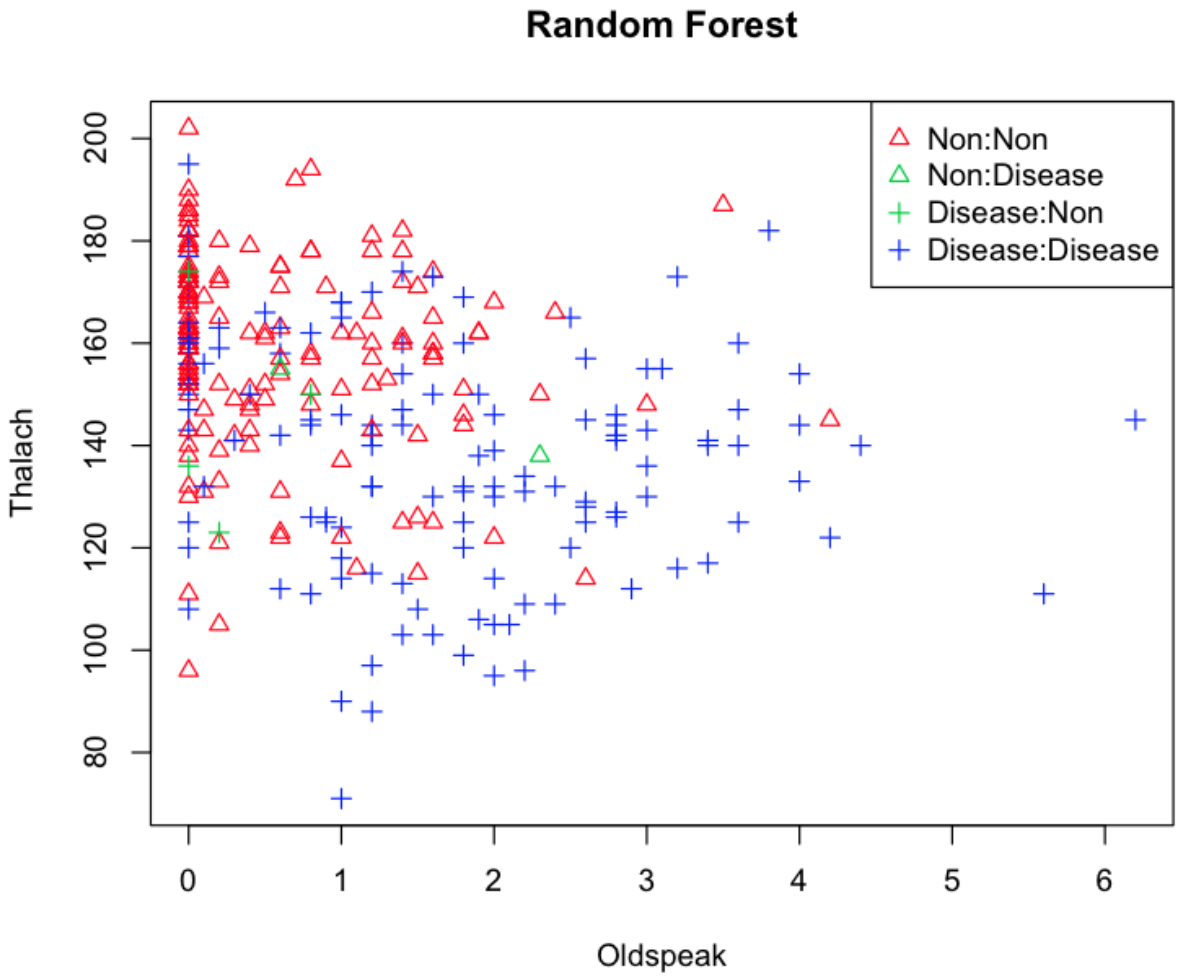

Figure 5.2: Random Forest Algorithm

As is mentioned above, we use random forest to choose key variables to project our data on.Since the model is flexible, the 0 train error is explainable while the 0.2 test error is

Table 5.2: Classification Result For Random Forest Algorithm

	Train Error	Test Error
Random Forest Algorithm	0	0.2

acceptable.

5.1.3 Support Vector Machine

The example of Support Vector Machine rule is connected on UCI repository and also the confusion matrix is made for sophistication having five qualities area unit incontestable in below. The results of the Support Vector Machine rule are unit incontestable in Table 5.

————Confusion Matrix————

a	b	c	d	e	
150	9	3	2	0	| a = 0
30	4	10	7	4	| b = 1
12	11	7	5	1	| c = 2
5	11	12	2	5	| d = 3
1	5	3	2	2	| e = 4

Table 5.3: Classification Result For Support Vector Machine Algorithm

	Train Error	Test Error
Support Vector Machine Algorithm	0.1998502	0.13

Since the linearity does not hold, the radial SVM has better result than original SVM. As SVM can solve classification problem in high dimension case, it gives us lowest test error among all approaches.

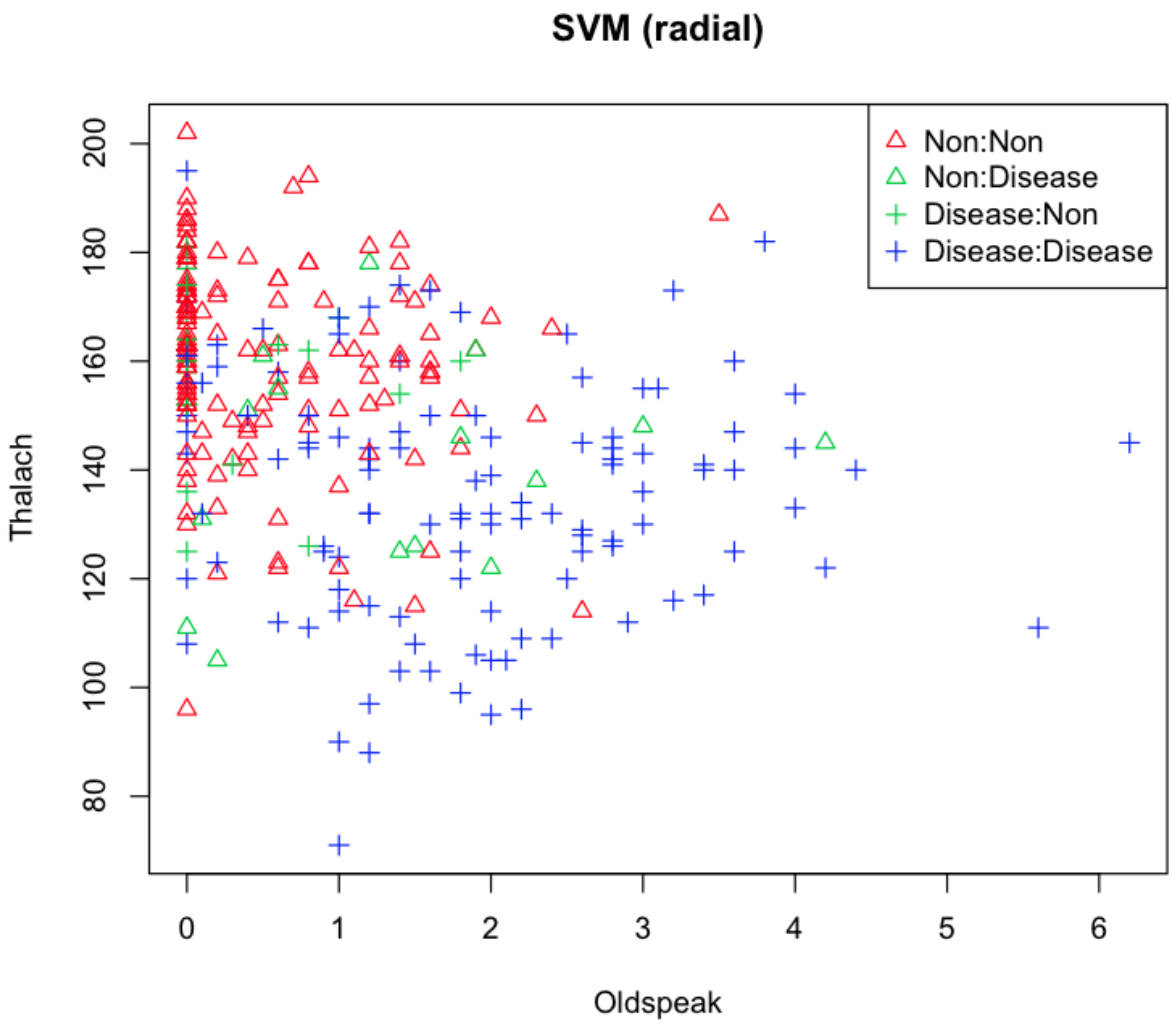

Figure 5.3: Support Vector Machine Algorithm

5.1.4 K Nearest Neighbour

The example of K Nearest Neighbour algorithmic program is connected on UCI repository and also the confusion matrix is made for sophistication having five qualities are incontestable in below. The results of the K Nearest Neighbour algorithmic program are incontestable in Table 6.

————Confusion Matrix————

a	b	c	d	e	
146	9	2	4	1	| a = 0
36	4	8	5	2	| b = 1
14	10	5	6	1	| c = 2
5	11	12	5	2	| d = 3
1	5	2	3	2	| e = 4

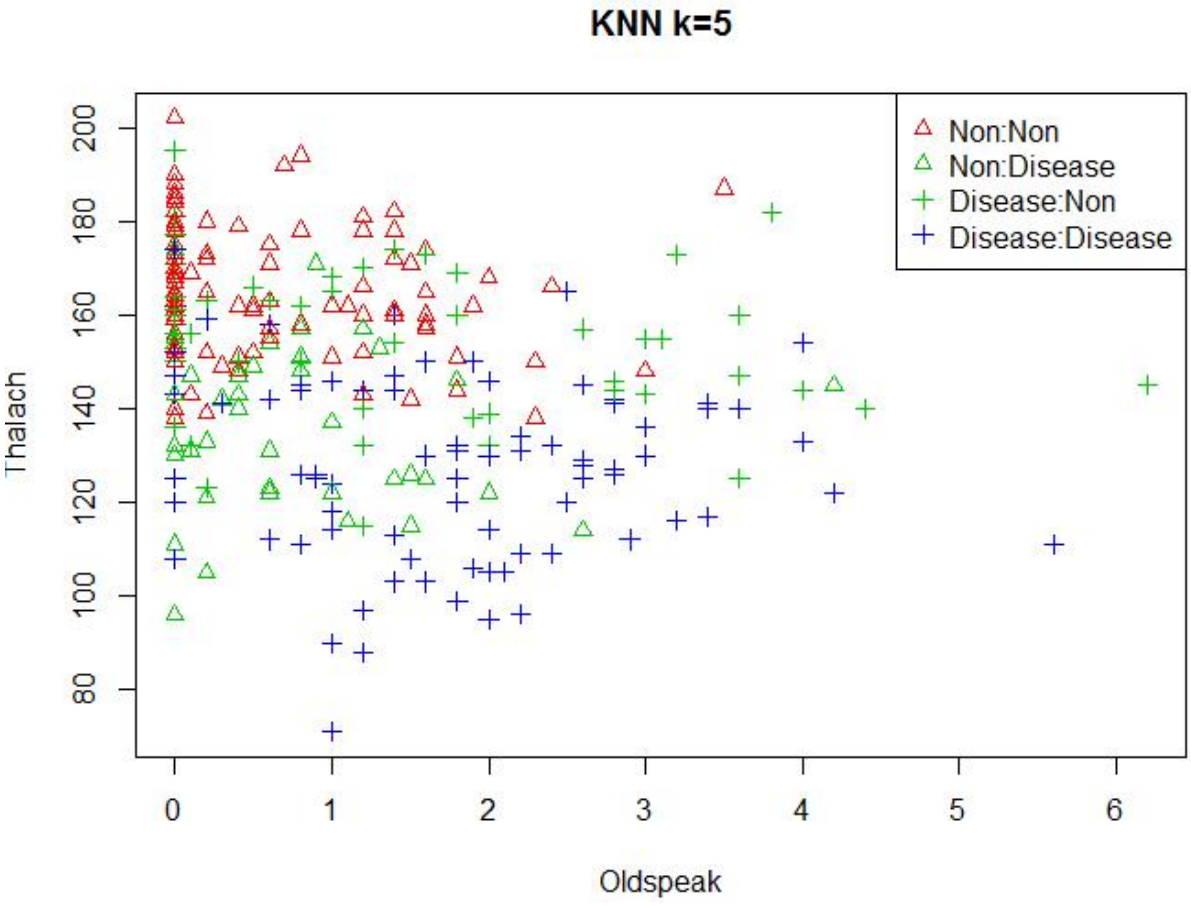

Figure 5.4: K Nearest Neighbour

Table 5.4: Classification Result For K Nearest Neighbour Algorithm

	Train Error	Test Error
K Nearest Neighbour	0.25	0.1666667

For KNN, we compare different errors to set our k-value, which is 5 as a result. Then we project our data on the two most important continuous variables generated from random forest. The increasing green signs shows a relatively high error rate, which means using local information only cannot give us a better classification.

5.1.5 Logistic Model Tree Algorithm

The example of Logistic Model Tree Algorithm is connected on UCI repository and therefore the confusion matrix is created for sophistication gender having 2 conceivable qualities are indicated in below.The results of LMT formula are incontestable in Table 3.

————Confusion Matrix————

a	b	c	d	e	
148	12	2	1	1	| a = 0
31	10	6	8	0	| b = 1
8	12	4	10	2	| c = 2
4	11	11	7	2	| d = 3
0	5	2	6	0	| e = 4

Table 5.5: Classification Result For Logistic Model Tree Algorithm

	Train Error	Test Error
Logistic Model Tree Algorithm	0.1156716	0.137931

Logistic Model Trees have been demonstrated to be extremely exact furthermore, smaller classifiers in diverse examination regions.The result is not that good, since the choice model has only one threshold, which means only one boundary to divide the data space.

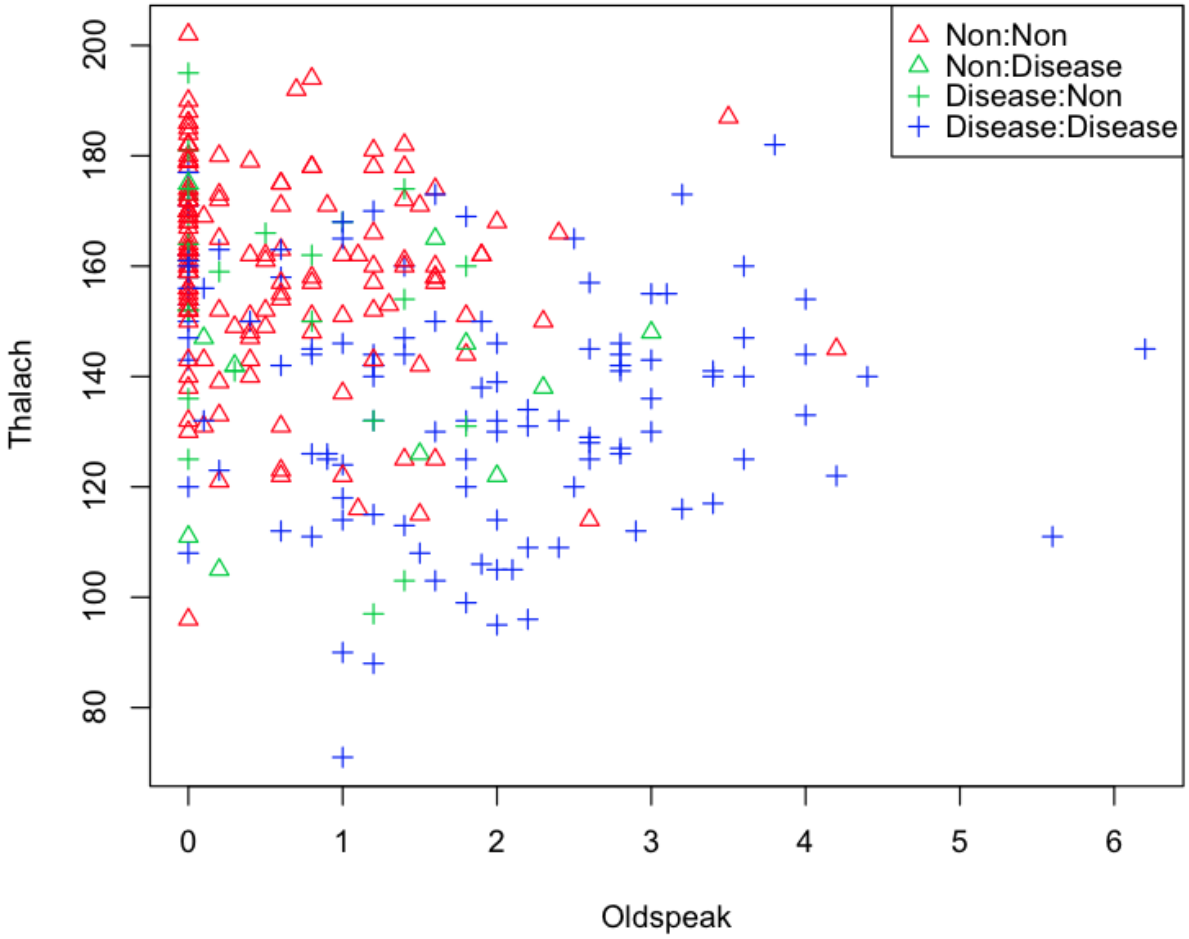

Figure 5.5: Logistic Model Tree

5.2 Comparison of Methodologies

When comparing the results of Random Forest algorithm,J48 algorithm,Logistic Model Tree Algorithm,Support Vector machine,K-NN we achieved higher sensitivity and accuracy using Support Vector machine algorithm than J48 algorithm. So overall from Table, it is concluded that Support Vector machine algorithm has got the best overall performance.

Also, J48 algorithm utilization reduced-error pruning form less number of trees. The LMT algorithmic rule manufactures the smallest trees. this might show that cost-many-sided quality pruning prunes all the way down to littler trees than decreased lapse pruning, however it to boot demonstrate that the LMT algorithmic rule doesn't got to assemble immense trees to cluster the data. The LMT algorithmic rule seems to perform higher on knowledge sets with various numerical attributes, while for good execution for 6 algo-

rithm, the data sets with couple of numerical qualities gave a superior execution. We can see from the outcomes that Support Vector machine algorithm is the best classification tree algorithm among the three with pruning system.

Table 5.6: Comparison of Different Algorithm Results

	J48	Logistic Model Tree Algorithm	Random Forest Algorithm	KNN	SVM
Train Error	0.1423221	0.1656716	0	0.25	0.1235955
Test Error	0.1666667	0.237931	0.2	0.1666667	0.2

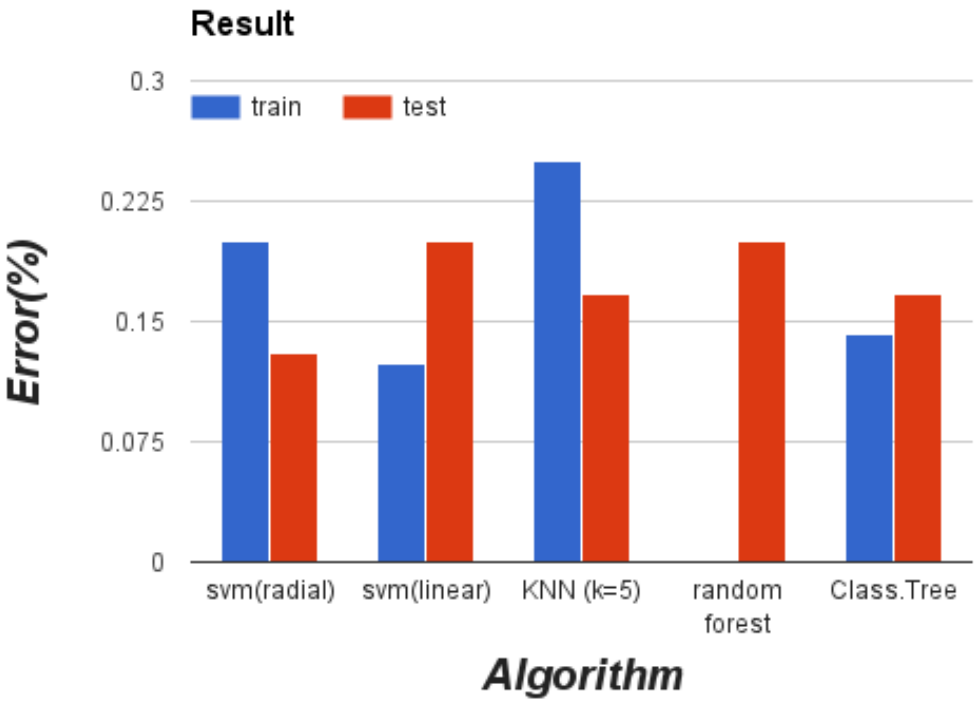

Figure 5.6: Comparison of Different Algorithm Results

Chapter 6

Conclusion and Future Work

6.1 Conclusion

By analyzing the experimental results, it's over that Support Vector machine algorithm tree technique turned out to be best classifier for cardiovascular disease prediction as a result of it contains a lot of accuracy and least total time to make.We can clearly see that highest accuracy belongs to Support Vector machine algorithm with reduced error pruning followed by J48 and Random Forest algorithm respectively. Also observed that applying reduced error pruning to Support Vector machine results in higher performance while without pruning, it results in lower performance. The best algorithm SVM based on UCI data has the highest accuracy i.e. 83.33% while LMT algorithm has the lowest accuracy i.e 76.21%

In conclusion, as identified through the literature review, we believe only a marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of heart disease.

6.2 Future Work

There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. Due to time limitation, the following research/work needs to be performed in the future.

- Like to make use of testing different Data Mining techniques, multiple classifiers Voting technique and different Decision tree types like information gain, gain ratio

and Gini index. Eg. Experiment need to perform on use of Equal Frequency Data Mining Gain Ratio Decision Trees by applying nine Voting scheme in order to enhance the accuracy and performance of diagnosis of heart disease.

- This paper proposes a framework using combinations of support vector machines and decision trees to arrive at an accurate prediction of heart disease. Further work involves development of system using the mentioned methodology to be use for checking the imbalance with other data mining models.

- Like to explore different rules such as Association, Clustering, K-means etc for better efficiency and ease of simplicity.

- To make use of Multivariate Decision Tree approach on smaller and larger amount of data.

References

[1] B. Deekshatulu, P. Chandra, *et al.*, "Classification of heart disease using k-nearest neighbor and genetic algorithm," *Procedia Technology*, vol. 10, pp. 85–94, 2013.

[2] C. S. Dangare and S. S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," *International Journal of Computer Applications*, vol. 47, no. 10, pp. 44–48, 2012.

[3] T. Mythili, D. Mukherji, N. Padalia, and A. Naidu, "A heart disease prediction model using svm-decision trees-logistic regression (sdl)," *International Journal of Computer Applications*, vol. 68, no. 16, pp. 11–15, 2013.

[4] M.-J. Huang, M.-Y. Chen, and S.-C. Lee, "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis," vol. 32, pp. 856–867, Elsevier, 2007.

[5] A. Khemphila and V. Boonjing, "Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients," pp. 193–198, 2010.

[6] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," pp. 108–115, 2008.

[7] Y. Xing, J. Wang, Z. Zhao, and Y. Gao, "Combination data mining methods with new medical data to predicting outcome of coronary heart disease," pp. 868–872, 2007.

[8] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modeling schemes for heart disease classification," *Applied Soft Computing*, vol. 14, pp. 47–52, 2014.

[9] N. Cheung, "Machine learning techniques for medical analysis," *School of Information Technology and Electrical Engineering, BsC thesis, University of Queenland*, vol. 19, 2001.

[10] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach," *Expert Systems with Applications*, vol. 40, no. 1, pp. 96–104, 2013.

[11] S. B. Patil and Y. Kumaraswamy, "Intelligent and effective heart attack prediction system using data mining and artificial neural network," *European Journal of Scientific Research*, vol. 31, no. 4, pp. 642–656, 2009.

[12] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment," pp. 173–177, 2012.

[13] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American journal of cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[14] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert systems with applications*, vol. 36, no. 4, pp. 7675–7680, 2009.

[15] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," vol. 17, pp. 43–48, 2011.

[16] S. K. Biswas, N. Sinha, B. Purakayastha, and L. Marbaniang, "Hybrid expert system using case based reasoning and neural network for classification," *Biologically Inspired Cognitive Architectures*, vol. 9, pp. 57–70, 2014.

[17] I. Kurt, M. Ture, and A. T. Kurum, "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease," *Expert Systems with Applications*, vol. 34, no. 1, pp. 366–374, 2008.

[18] M. Shouman, T. Turner, and R. Stocker, "Using decision tree for diagnosing heart disease patients," in *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*, pp. 23–30, Australian Computer Society, Inc., 2011.

[19] K. Polat, S. Şahan, and S. Güneş, "Automatic detection of heart disease using an artificial immune recognition system (airs) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing," *Expert Systems with Applications*, vol. 32, no. 2, pp. 625–631, 2007.

[20] N. Guru, A. Dahiya, and N. Rajpal, "Decision support system for heart disease diagnosis using neural network," *Delhi Business Review*, vol. 8, no. 1, pp. 99–101, 2007.