Prediction of Heart Disease using Hybrid Methodology of Selecting Features

A Thesis

Submitted in partial fulfillment of the
requirements for the award of the degree of

Masters of Engineering

in

Computer Science and Engineering Department

by

Kanika Pahwa
(Roll no: 801532023)

Under the supervision of

Dr. Ravinder Kumar
(Assistant Professor)

Thapar University , Patiala, India
July 2017

Certificate

I hereby certify that the work, which is being presented in the thesis, entitled **Prediction of Heart Disease using Hybrid Methodology of Selecting Features**, in partial fulfillment of the requirements for the award of the degree of **Masters of Engineering** in Computer Science and Engineering and submitted to the institution is an authentic record of my own work carried out during the period **July 2015** to **July 2017** under the supervision of **Dr. Ravinder Kumar**.and refers other researchers work which are duly listed in the reference section.

The matter presented in this thesis has not been submitted elsewhere for the award of any other degree of this or any other University.

Date: 13 July 2017

Kanika

Kanika Pahwa
Candidate

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

Date: 13.07.17

Dr. Ravinder Kumar
(Assistant Professor)
Supervisor

Abstract

Generally Healthcare industry is known to be 'information rich' , but woefully all the data required to discover hidden patterns are not mined .In present time heart disease is most fatal one.This is one of the leading cause of death in countries like UK ,Canada,India,Australia .Attack of heart disease is so abrupt that it rarely gives anyone a time to tackle with it.So detection of disease precisely and timely is complicated and intricate task in field of medical. Medical professionals may take wrong decision during diagnosis which may cause death of patient .

For effective decision making in field of medical ,advanced techniques of data mining are used.

The work represented here mainly focuses on prediction of heart disease using supervised machine learning models. On basis of available features , data is classified into two classes i.e. presence and absence using Random Forest and Naive Bayes . In addition , approach is proposed to select features before classification in order to improve performance of models. Here proposed approach helps to derive importance of features . This is done by applying SVM-RFE and gain ratio algorithms to dataset which in results assigns weight to each feature.This approach helps to improve accuracy and reduce computational time.Experimental results shows that proposed approach of selecting feature increases accuracy for both models.

Keywords: Heart Disease Prediction, Feature Selection , Binary Classification , Machine Learning Models.

Acknowledgements

First, I would like to express my deep gratitude to my supervisor **Dr. Ravinder Kumar (Assistant Professor)** for their invaluable advice and encouragement at every step of my thesis. Without their unfailing support and belief in me, this thesis would not have been possible. Their contribution to this thesis goes well beyond their role as an academic supervisor and includes constant support on a personal level without which this journey may never have been completed. And for this, I am truly grateful. They are great mentor for my life as well.

I would like to express my gratitude to all the faculty members at **Thapar University** for equipping me with the best of knowledge.

Kanika

Kanika Pahwa

Contents

Title No.		Page
Abstract		**ii**
Contents		**iv**
List of Figures		**vi**
List of Tables		**vii**
List of Abbreviations		**viii**
Chapter 1	**Introduction**	**1**
1.1	Research Orientation	2
1.1.1	Research Motivation	2
1.1.2	Key Research Area	3
1.2	Literature Review	4
1.2.1	SVM Learning from imbalanced dataset along with Feature Selection using Gini Index	4
1.2.2	Machine learning models for Classification	6
Chapter 2	**Problem Statement**	**11**
2.1	ALGORITHMS	11
2.1.1	Naive Bayes	11
2.1.2	Random Forest	14
2.1.3	Support Vector Machine -Recursive Feature Elimination	15
2.1.4	Gain Ratio	16
Chapter 3	**Feature selection**	**18**
3.1	Introduction to FS	18
3.1.1	Definition	18
3.1.2	Advantages	19
3.1.3	Characteristics of FS algorithms	19
Chapter 4	**Tools Used**	**22**

4.1 Weka 22
4.2 Eclipse 25

Chapter 5 Implementation 27
5.1 Data Collection and preprocessing 27
5.1.1 Data integration 28
5.1.2 Data transformation 28
5.1.3 Data Cleansing 28
5.2 Feature selection 30
5.2.1 Classification with randomly selected features 32
5.2.2 Proposed approach of selecting features 35

Chapter 6 Evaluation 41
6.1 Confusion matrix 41
6.2 Area Under ROC Curve 44
6.2.1 Area under curve 45
6.3 Cross Validation 45

Chapter 7 Conclusion 47

Chapter 8 Plagiarism and publication list 48

References 50

List of Figures

Figure No.	Title	Page No.
1.1	Neural Network	10
2.1	SVM	15
3.1	Wrapper method	20
3.2	Filter method	20
4.1	data	24
5.1	Flow chart	27
5.2	Integration of data	28
5.3	Weka interface	31
5.4	Opening weka file	33
5.5	Viewer in weka	33
5.6	Randomly chosen attributes for removal	34
5.7	Classification with naive bayes	34
5.8	Classification with random forest	35
5.9	Score calculation	36
5.10	Rank calculation	38
5.11	Result	38
5.12	Naive Bayes	39
5.13	Random Forest	40
5.14	Snapshot of selected features	40
6.1	Roc curve	44
8.1	Plagiarism Report	48
8.2	Paper acknowledgement	49

List of Tables

Table No.	Title	Page No.
5.1	Raw data	29
5.2	Processed data	29
5.3	Instances with missing value	30
5.4	Score	36
5.5	Rank	37
6.1	Confusion matrix	42
6.2	Confusion matrix of naive bayes	43
6.3	Confusion matrix of random forest	43
6.4	Resultant	43
6.5	Cross Validation	46

List of Abbreviations

SVM	Support Vector Machine
FS	Feature Selection
OP	Oldpeak
RFE	Recursive Feature Elimination
LMT	Logistics Model Tree
RBF	Radial Basis Function
MLP	Multi Layer Perceptron
CART	Classification and regression tree
FBS	Fasting Blood Sugar
MLP	Multi Layer Perceptron
LOOCV	Leave One Out Cross Validation

Chapter 1

Introduction

In present time heart disease is most fatal one.This is one of the leading cause of death in countries like UK ,Canada,India,Australia .Attack of heart disease is so abrupt that it rarely gives anyone a time to tackle with it.So detection of disease precisely and timely is complicated and intricate task in field of medical. Medical professionals may take wrong decision during diagnosis which may cause death of patient . Also in some cases treatment of said disease is not affordable specially in India.The main motive of this research is to develop cost effective prediction of said disease using machine learning techniques. As all medical expertise do not possess competence in every field and moreover there is shortage of medical experts at certain places.In that case results of prediction derived by various advanced techniques is used along with medical expertise advice to reduce chances of undesirable results.

Research work presented in this thesis mainly focuses on prediction of heart disease using machine learning methodolgies.Main objective is to predict presence of heart disease on basis of given symptoms of patients.Different machine learning models are available like Decision Tree ,Random Forest ,Logistics Regression,Naive Bayes , K-nearest neighbours.For all models prediction is divided into two phases ,one is training phase and other is testing phase.Training phase is used to train model using training dataset and testing phase is phase where trained model is tested with never seen inputs.Performance of model is decided by comparing results of test phase with actual results. Higher the match means better is the model. Other parameters are also used to compare performance of different models like RMSE,confusion matrix,ROC,AUC depending upon model . Here in this work before classifying presence or absence of disease , features are prioritized .So that lower valued features are seperated from higher valued features of dataset . Higher valued are only used for classification which further improve accuracy of model and reduce computational time.Proposed methodology of feature selection is hybrid of Gain ratio and SVM-RFE (Support Vector Machine - Recursive Feature Elimination).Subset of features are used to classify data using Random forest and naive bayes classifier.

All the above brief descriptions are discussed in subsequent chapters.

1.1 Research Orientation

This section narrates the motivation of this research and research areas chosen for this work.

1.1.1 Research Motivation

Heart disease is result of alteration of functionality and structure of heart .This leads to inadequate pumping due to which organs and tissues receives insufficient amount of oxygen for their metabolic needs .Tissues could become necrotized due to lack of oxygen and may cause death.Heart disease was the crucial cause of casualties in the different countries .In US heart disease kills one individual in every 34 second .On basis of medical professionals knowledge and experience, analysis is often made but in some cases this may bring about undesirable results as all doctors do not possess competence in every field.Also most of hospitals use information system to maintain their patient data.But results of this system is huge and is in form of chart ,text,image and numbers , which doesn't help much in making clinical decision. Diagnosing a disease correctly on time is one of the most difficult task.

So there is need to develop a prototype to extract knowledge with respect to said disease using past heart disease records.This research is required to classify heart disease data with high accuracy by predicting presence or absence of said disease.

Various machine learning algorithms available for classification [1] and clustering [2] .Classification is to assign each observation to predefined class while clustering is to assign each observation to one of cluster on basis of their nature. Different machine learning algorithms have been successfully implemented. Model is applied to available dataset in two stages .In first stage classifier is built with labelled samples and in second stage after classifier is successfully built , never seen inputs are used to test effectiveness of classifier. Of all available classification models , an efficient classifier need to be implemented or chosen to correctly classify heart disease so that heart patients are correctly diagnosed.In this research we have used naive bayes and random forest to classify data into positive or negative result.

Medical experts can use these results along with their knowledge and experience to be more confident on their decisions to avoid undesirable results.

1.1.2 Key Research Area

The key research area focuses on the machine learning algorithms for features selection using hybrid approach and then classifying data on basis of features. Following are the key research areas:

1. The inherent drawback with large number of features is that it may contain irrelevant or redundant features and also removal of such features will not cause much loss of information.Main objective of selecting relevant features from given dataset is :

   (a) Less features allows machine learning algorithm to learn faster

   (b) If correct features are chosen then this may improves accuracy of machine learning model

   (c) It also helps to reduce overfitting

   (d) Less features reduces model's complexity

   In this research we have proposed a new approach of selecting features which is hybrid of two algorithms named as SVM-RFE and GAIN-RATIO. Results of both algorithms are consolidated to select relevant features from dataset.Detailed description is explained in subsequent chapters.

2. Furthermore, classifier chosen also affects results.Selection of classifier depends upon size, nature,quality of data.Machine learning algorithms can be supervised and unsupervised

   (a) Supervised learning algorithms : In this we have input variable as well as output variable and algorithm is used create a function from input to output known as mapping function. Main objective of said type[3] is to approximate the mapping function in best way that can predict output values for never seen input variables.

   (b) Unsupervised learning algorithm:In this type of algorithms we are provided with input variables only , no output variables .

   In this research we have used supervised learning algorithms for classification named as NAIVE BAYES , RANDOM FOREST. Said algorithms are applied to subset of features to produce results.

1.2 Literature Review

1.2.1 SVM Learning from imbalanced dataset along with Feature Selection using Gini Index

There are number of techniques proposed for feature selection by ranking features on basis of correlation or information ranking criteria. Approach of selecting features before classification improves classification performance along with it provides a better understanding of process that generates data.Also imbalance dataset is one of the other issue in machine learning. Here imbalanced dataset means one class have more number of instances as compared to other class and classifier may be biased towards majority class.This may degrades performance of machine learning methods.

Different approaches of selecting features are available like correlation approach , single variable approach and information theoretic ranking.But correlation approach calculates linear dependency between feature and predictive variable(target) while single variable approach calculates importance of feature on basis of predictive capability of each feature.In information theoretic ranking mutual information is only parameter to check degree of information that one feature contain of other feature.This is also known as information gain.Approach used in this work is similar to information gain approach.

The author Nanning Zheng and Jiayi Wu in research paper [4] performs microaneurysm detection which is considered to be as first sign of diabetic retinopathy.In this study relevant features are selected on basis of weights generated by gini index and then modified svm classifier is applied to selected features to classify data in two groups.

Gini Index method is used to compute goodness of all the relevant features.Gini index of given dataset A is calculated as follows:

$$Gini(A) = 1 - \sum_{k=1}^{K} p_k^2 = 1 - \sum_{k=1}^{K} (\frac{C_k}{D}^2) \tag{1.1}$$

where K is number of classes and Ck is number of instances that belong to k class and pk is relative frequency.Also if dataset A is divided into two groups A1 and A2 by splitting value which belong to feature X.Partitioned data is used to compute index as

follow:

$$Gini(A,X) = (\frac{z1}{z}Gini(A1)) + (\frac{z2}{z}Gini(A2)) \tag{1.2}$$

Here z1 and z2 are instances that belongs to subset A1 and A2 respectively.Also feature having minimum gini is considered to be a best feature for classification purpose.

Work flow of computing gini index of training dataset D is explained below : Input: Training dataset D and F is Feature Set

1. Calculate Gini(D,A) : For each feature A in feature set F ,if dataset D is divided into two groups D1 and D2 by dividing value a of feature A and index of partitioned data is computed using formula 1.2
2. Choose minimum index of feature as best seperation feature
3. Remove best seperation feature from set of feature F
4. Repeat above steps until no feature left in F

After removing irrelevant features from feature set modified version of svm is applied.This modified approach is to make classifier to work well even for imbalanced data set. As learning methods work well on balanced dataset but not in imbalanced .

SVM is technique of training classifiers on basis of radial function , neural network or polynomial functions.Therefore standard svm will divide classes on basis of hyperplane .Optimal hyperplane is the one which have maximum margin from the nearest instance But due to imbalanced instances ,standard svm will be biased towards majority class to reduce total error .

To avoid this author in research paper [4] proposed a method to adjust a hyperplane by simply mapping outputs of SVM to probabilities and use a threshold prob. for classification of candidates.This is done as follows:

$$t_i = (\frac{y_i + 1}{2}) \tag{1.3}$$

where ti is target probability.

In testing phase after probability is assigned to each candidate of lesion , then this is compared to t to predict candidate is lesion or not.After comparison of both probabilities hyperplane will be set and this will improve accuracy of model .

This improves accuracy of classifier even for imbalanced dataset.As imbalanced dataset may cause performance loss as in standard SVM minority points lie farther from ideal line.

1.2.2 Machine learning models for Classification

Machine learning[5] performs data analysis to explore meaningful information from huge amount of data.Numerous machine learning algorithms are available that learn from available data and helps to find hidden patterns.One of the important aspect of machine learning is data mining [6] and industries dealing with big data have realised the value of machine learning technology.

Like a human being machine learning has capability to do certain activities that need us to make decision.Like ,While reading inbox of our account, we make decision by marking email as spam on basis of its subject or sender like ”Win gold on clicking here” .In similar way machine learning provides bundle of algorithms that helps a system to do task that a human mind does naturally.

Real life application of machine learning are

1. Healthcare:The technology can help medical professionals to analyze data in better way and explore hidden pattern which helps medical expert to take better decisions at time of treatment.

2. Marketing and sales: Items recommended on a website is done by machine learning technology by analyzing shopping history.Amazon also uses said technology in analyzing purchasing of product and then forecasting demand accurately.

3. Financial Services : Financial institutes use this technology to prevent fraud by recognizing users with high risk profiles.

Machine learning have following advantages :

1. Machine learning makes fast processing and predictions in real time : The way machine learning analyse big data allows to result in boom trend and which makes further real time predictions.Example : It can be used to decide various offers in grocery shop for customer to increase income .This is done on basis of past records of customers. In simple words ,this have capability to identify hidden patterns from large database.

2. Improves decision making: Machine learning have a capability to improve deci-

sion making by making priority of features and removing irrelevant feature which affect performance of models

3. Ability to modify:One algorithm is implemented , we can change algorithm to optimize or improve performance using tolls and different data structures.

4. Learning on basis of features:One of the important benefit of machine learning is the capability to get trained on few dataset given for specific task.In earlier days , human experts are used for selecting features from large dataset. This took number of months or year too for making a synchronization between parameters to get right results.So now machine learning is used to choose relevant features from all available dataset including relevant and redundant features too.Training a model on basis of only relevant features also improves performance of model in most of the case.

5. Ability to learn from past records : Other remarkable feature of machine learning is ability to make rules from past predictions and have capability to improve prediction on basis of new available data .Example On particular day and particular time what type of customers arrive.

6. Consumption of large data : Also machine learning have ability to consume huge set of data and from huge set identify relevant features .Along with this it can also used in marketing by consuming large amount of data and modify sales and making different strategies based on behaviour of customer.

7. Quick and efficient:Machine learning allows to get thing done quickly and efficiently.Machine learning have ability to automate task which is otherwise done by a human expert like checking balance of account, important decision making.

To implement this technology number of machine learning models are available .Models are just like a black box which is used to solve same problem by applying its algorithm which vary from model to model.Broadly machine learning methods are categorized into supervised and unsupervised learning methods.

1. Supervised methods: In supervised learning methods we are provided with desired output in training dataset .Training dataset is divided into two vectors ,one input vector and other as desired output value.In training phase , model is trained with known values and in testing phase model predict output of input vector on basis of training and then results are compared with actual results .On basis of this comparison performance of model is calculated , this is also known as accuracy of model.

2. Unsupervised methods : In unsupervised learning methods [7] we are not provided with output or we can say that we are provided with unlabelled data.One of the known unsupervised learning method is clustering.In this method input is divided into clusters on basis of their characteristics.

There are number of machine learning models.Few of them are discussed below :

1. LMT: In LMT[8] both tree induction classification and linear logistics regression are combined. Primary benefit of logistics regression is class estimate are also generated along with classification. LogitBoost algorithm is divided in two phases a. Logistic regression model is produced for each node. b. Using C4.5 criterion ,node is split. With each invocation of LogitBoost , parent node results are employed

2. J48: This algorithm results in decision tree where each node is splitted using attribute with highest information gain and then repeat this for smaller sublist.

3. MLP: This network uses backpropagation for classification of instances. It is divided in three layers input ,hidden , output layer. Building block of this network is node or neuron[9].Each layer consists of number of interconnected nodes. Pattern are passed to input layer which is further processed to the hidden layer and decision is passed to output layer.

4. SVM: Support Vector Machine (SVM)[10] can be used for both regression or classification challenges. In this ,each data item is plotted in x dimensional space(Here x is no. of features in dataset).Then classification is performed by finding optimal hyperplane that best divides instances of input by their class in way that maximize the distance of points from linear decision boundary and expected risk is minimized.

   The optimal hyperplane is the one which have maximum margin between hyperplane and closest instance .

5. RBF: Radial Basis Function network [11]uses radial function as its activation function. This neural network consist of input layer ,hidden layer ,output layer with one node for each class. Number and position of neuron in hidden layer affects the network , if number of neuron is high then it may affect capability of network either by overlearning situation or by poor generalization and if number of neuron is low then network may not learn data adequately. Primarily benefit of using RBF is that this network brings robustness to our prediction

6. BAYESIAN NETWORK: Bayesnet is probabilistic model for the reason that

they use probability distribution to build model and laws of probability to make decision[12].This network is directed acyclic graph which consist of nodes and links , where each node represents variable which can be continuous or discrete and link between two nodes represents direct dependencies between variables.

7. Naive Bayes: Naive Bayes is powerful supervised learning approach that uses probability of each attribute to make decisions.This algorithm [13] is based on Bayes theorem.This is preferable for high dimensions of input and for complex real world problems as it works effectively for large range of problems.

   In Naive Bayes presence of one feature does not affects other features. In other words this theorem assumes independence among predictors.

8. CART: Classification and regression tree model is represented as a binary tree[14]. Input variables are taken as root variable and output variable are at leaf nodes which is used to make predictions. Once model is trained , tree is traversed according to inputs to reach leaf node .Output variable contained in leaf node is the prediction made by model.

9. RANDOM FOREST: Random forest [15] model constructs number of decision trees and find mode of all classes output by individual tree as a final output.There may be high variance in single tree but RF lighten the problem of high variance.

   Error rate of random forest depends upon two factors are :

   (a) Correlation between two instances of forest(trees).As correlation increases error rate also increases

   (b) Another factor is strength of each instance of forest.With increase in strength ,error rate of forest decreases.

10. RANDOM TREE: This model is used to construct single tree of all trees constructed by random forest model.

    In random tree during training , error is estimated so no need to apply any procedure to estimate accuracy like bootstrap or cross validation.

11. NEURAL NETWORK : NN is collection of layers ,first layer is input layer which passes its output to hidden layers and then to output layer [2]. Each interconnected point on layer is node which also contains activation function. Each layer is connected and weights are assigned to each connection .

    Pattern is passed to input layer and each node of input layer further communicates

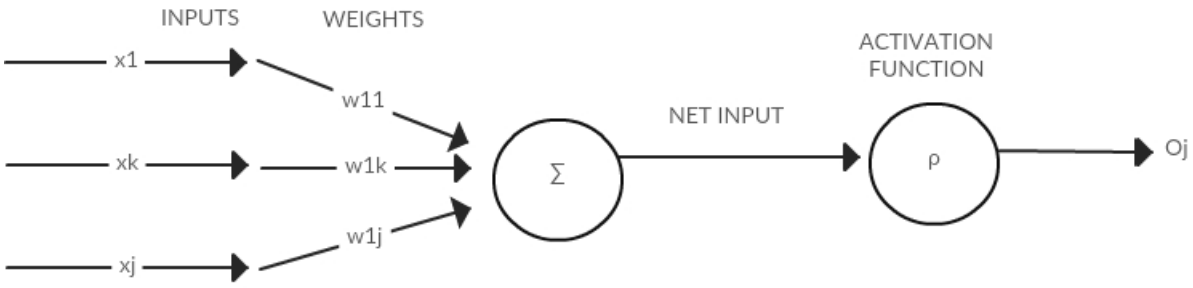


Figure 1.1: Neural Network

to hidden layers by dividing inputs into neurons of hidden layer. Hidden layer can be one or more. After processing with weighted connection of system by adding input signal with weight associated with connection from previous layer, output of processing is passed to output layer as shown in Fig 1.1. The output

$$Y_k = \sum_i (w_{ki} * x_i) \tag{1.4}$$

where wi is weight associated on connections from one layer to other and xi is input vector passed to input layer.

Chapter 2

Problem Statement

We are going to predict heart disease on basis of past records .There is need to develop a prototype to extract knowledge with respect to said disease using past heart disease records to help medical experts for decision making .This research is required to classify heart disease data with high accuracy by predicting presence or absence of said disease.

In this we will apply classifiers for prediction after filtering feature on basis of their relevance by excluding irrelevant features . On subset of features naive bayes and random forest are applied for prediction.For feature selection we have proposed a different approach ,which consolidates result of svm-rfe and gain ratio algorithm .Combined result of said algorithms are used to select features.Classifiers applied on selected features by proposed approach gives the better results as compared to previous approach.

2.1 ALGORITHMS

For prediction of heart disease following algorithms are used

1. Naive Bayes
2. Random Forest
3. SVM-RFE
4. Gain Ratio

2.1.1 Naive Bayes

Naive bayes is one the most important approach of classification. This algorithms assumes all features are independent of each other.Changing value of one doesnot affect other features.Also naive bayes doesnot require much data to estimate outcome. The only basis of Naive Bayes[16] is Bayes theorem

2.1.1.1 Bayes theorem

Bayes theorem uses prior knowledge to computes conditional probability of event .Basically conditional probability is the one which reflects happening of one event on probability of other event[17].Terms related to bayes theoren are:

Prior probability : This is the original probability of an event before referring to any additional information obtained.

Posterior probability : This is probability computed on basis relevant information. It is written as :

$$P(X|Y) = \frac{P(Y|X) * P(X)}{P(Y|X) * P(X) + P(Y|\neg X) * P(\neg X)} \tag{2.1}$$

where

P(X) and P(Y) are prior probabilities of X and Y respectively

P(Y/X) is posterior probability of Y given X

P(X/Y) is posterior probability of X given Y

$P(Y/\neg X) is probability of Y given X is false$

$P(\neg X) is probability of X being false$

For feature vector x1,x2....xn and classes C1,C2,... ,Ck , bayes theorem can be written as:

$$P(C_j|x_1, x_2, ...x_n) = \frac{P(x_1, x_2, ...x_n|C_j) * P(C_j)}{Px_1, x_2....x_n} \tag{2.2}$$

Using assumptions of independence

$$P(x_i|C_j, x_1, \ldots, x_n) = P(x_i|C_j), \tag{2.3}$$

for all i .Further this ca be simplified as:

$$P(C_j|x_1, x_2, ...x_n) = \frac{P(C_j)\prod_{i=1}^{n} P(x_i|C_j)}{P(x_1, x_2, \ldots, x_n)} \tag{2.4}$$

Further P(x1,x2.....xn) is constant and equation is written as:

$$P(C_j|x_1, x_2, ...x_n) \propto P(C_j) \prod_{i=1}^{n} P(x_i|C_j) \tag{2.5}$$

Here we are calculating conditional probability of object with given feature vector x1,x2.....xn with respect to particular class.

Naive Bayes is used for prediction in following way :

1. On basis of given features ,probability is computed for each class using equation (2.6).
2. Class with higher probability will be chosen as a most likely class for that particular instance.

2.1.1.2 Types of naive bayes

Naive Bayes is categorized in three types:

1. Bernoulli Naive Bayes :This type of naive bayes is used for features having binary value i.e 0s or 1s.Few applications of bernoulli are text classification , disease prediction where 1 indicates presence and 0 indicates absence.
2. Gaussian Naive Bayes : This is approach which is followed for continuous values of attributes.For this case value related to each class follows gaussian or normal distribution.
3. Multinomial Naive Bayes :This approach[18] is related to discrete count.Lets consider text classification problem .Here we consider bernoulli trial and in place of " presence or absence of word " in document, we have " count of frequency of word " on basis of is occurence.

Benefits of using naive bayes are listed below:

1. Naive bayes have capability to get trained even on smaller datasets.
2. Implementation is easy
3. Also naive bayes can be used for multi class classification

2.1.2 Random Forest

Random forest[18] is supervised learning technique used for both classification and regression. Random forest as name suggest generates number of trees to make a forest.Unlike decision tree which generates only one tree,random forest generates number of trees. Also approach used in random forest to choose root node is different then decision tree which uses information gain or gini index ,but here in random forest root node is chosen randomly.

Reasons of using random forest are:

1. One algorithm used for both regression and classification.
2. Also compatible with categorical values.
3. Also have capability to deal with missing values.
4. More number of trees in forest means more will be the accuracy.

Random forest classifier works in two stages:

1. Creation of random forest using dataset:This is done by following below steps
   (a) Choosing random set of j features out of i features where j¡i
   (b) Choose root node among all chosen k features on basis of its split approach
   (c) Choose child nodes from remaining features from set of k features
   (d) Repeat above steps until we are left with target node which will be leaf node of tree.
   (e) All above steps are to create decision tree .Above steps are repeated k times to generate k trees.
2. Use random forest for prediction : After generation of random forest next phase is to predict outcome using generated forest in following way:
   (a) Choose random features set of all available features
   (b) Use rules generated by k trees in phase 1 to predict outcome.Store k outcomes generated by k trees.
   (c) Final outcome is considered to be highly voted outcome.This concept of majority is known as majority voting.

2.1.3 Support Vector Machine -Recursive Feature Elimination

SVM is machine learning technique used for both classification and regression.Here each instance is plotted on n dimensional space where n is number of features.

Idea of SVM[19] is to find optimal hyperplane that best divides instances into different classes.Here hyperplane is line or surface that linearly divides dataset or classify into its classes.

2.1.3.1 Choosing Best Hyperplane

Optimal hyperplane is the one that have maximum margin from hyperplane to nearest instance . Here maximum margin means distance between plane and nearest instance on either side of hyperplane.

As shown in figure 2.1 green and red points are support vector and line between them classifies the two classes in best possible way.

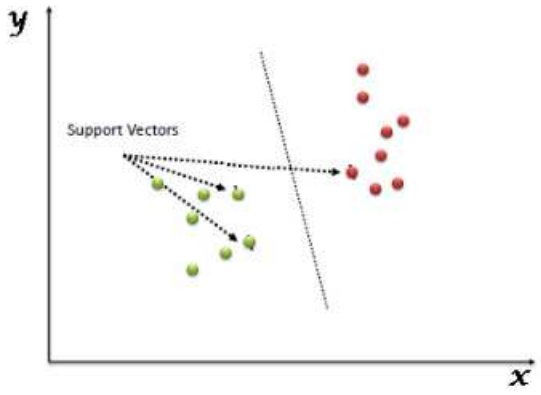


Figure 2.1: SVM

Maximum margin across linear line make model more robust.

2.1.3.2 SVM-RFE

Support vector machine-recursive feature elimination[20] is a wrapper method of SVM. SVM-RFE is efficient approach of selecting features. This wrapper method select features from dataset with k features is done as follows:

1. Linear svm[21] is applied to all features of feature set and assigns weight to each features.

2. Percentage of worst ranked features are eliminated from features set on basis of weight.

3. Svm is trained again with remaining attributes.

4. Repeat above steps until all features are removed from feature set

This approach of finding relevant subset of feature is known as greedy optimization .All features are ranked on basis of weight.

2.1.4 Gain Ratio

Information gain and gain ratio are approaches to choose best node of available nodes for construction of decision tree.Term related to these approaches are:

Entropy : Entropy is measure of uncertainty which is computed using below formula:

$$entropy = -\sum_{k=1}^{n} p(x_k) log_b(x_k) \tag{2.6}$$

Information gain[22] works well for all cases unless we have any feature with multiple values .But information gain is biased towards multi valued attribute.To remove this limitation a modified version is used for multi valued attributes that is gain ratio[23]. Gain ratio have following features:

1. Gain ratio is not biased towards multi valued attribute.

2. For choosing attribute as best split it considers size and number of branches.

3. Gain ratio introduced the concept of split information .

The split information value is information generated by dividing training data A into k partitions, corresponding to k outcomes on attribute X.

$$SplitInfo_X(A) = -\sum_{j=1}^{k} (\frac{|A_j|}{|A|}) * log_2(\frac{|A_j|}{A}) \tag{2.7}$$

Further gain ratio is computed by dividing gain by split info as given below :

$$gainratio(X) = \frac{Gain(X)}{SplitInfo(A)} \tag{2.8}$$

Here Gain is computed by subtracting information before split and after split. In next section we will discuss implementation of above algorithms to classify heart disease on basis of features.

Chapter 3

Feature selection

3.1 Introduction to FS

For high dimensional dataset feature selection is important term .In machine learning , subset of features are used from available dataset for learning process.This is approach of choosing best subset of features which contain least number of dimension to increase accuracy of algorithm in best possible way by discarding leftover attributes [4]. Further features can be categorized into following types:

1. Relevant Features:There are features whose presence improve accuracy and task cannot be completed by rest of features.
2. Irrelevant Features:Those features whose presence doesn't improves accuracy of model but in some cases may degrade performance.Values of such features are randomly generated.
3. Redundant Features:As name suggest when feature take role of other feature then redundancy exists in a dataset.

3.1.1 Definition

It is approach of choosing best attributes from the all available features as all features are not useful in performing task such as classification or clustering.This is so because some features may be redundant or irrelevant which doesn't contributes anything to learning process.

Feature selection [24] or subset selection is approach followed in machine learning where subset is used for applying learning algorithm.The main objective of feature selection is filter a minimum subset from all features while preserving accuracy of model by representing relevant features.Therfore best subset is set of those features which have least dimensions that contribute most to improve performance.

This is crucial phase which involves preprocessing of features before classification or clustering depending on task.This is important because of presence of abundant of irrelevant and redundant features.To be sure in subset of features there are number of iterations to choose best combination of features which sometimes becomes infeasible as for n features there are 2n subsets.

3.1.2 Advantages

Feature selection has numerous advantages.Few of them are:

1. Helps to remove features whose presence doesn't contribute to accuracy.In other words it removes noisy data.
2. Improves quality of data
3. It reduces dimensionality and requires limited storage.
4. Limited features helps to increase algorithm speed by reducing complexity of data .
5. Better data understanding
6. Improves accuracy of machine learning model.

3.1.3 Characteristics of FS algorithms

Feature selection algorithms search best subset of features through space of features.Following are issues that affect the search of features.

1. Initial point:Startng point of search from feature space may affect search.This can be done in different ways.One approach is to begin with no feature and successively adding on basis of its relevancy.Alternative approach is to start will all features of feature space and then removing them on basis of its irrelevancy.
2. Search organization:Heuristic search strategy is better then exhaustive search strategy as this will generate hood result .
3. Strategy involved for evaluation : One of the important task in area of feature selection is to evaluate features selected.One approach of feature selection is by applying filter method independent of learning approach ,this will remove undesirable features before learning .These type of algorithms use information re-

lated to features to evaluate selected subset.Other evaluation strategy is by using induction algorithm and cross validation technique for evaluating subset of features.This approach is known as wrapper method.

Wrapper method is slower then filter due to high interaction with induction algorithms and repetitive iterations .Wrapper method and filter is shown in figures 3.1 and 3.2 respectively.

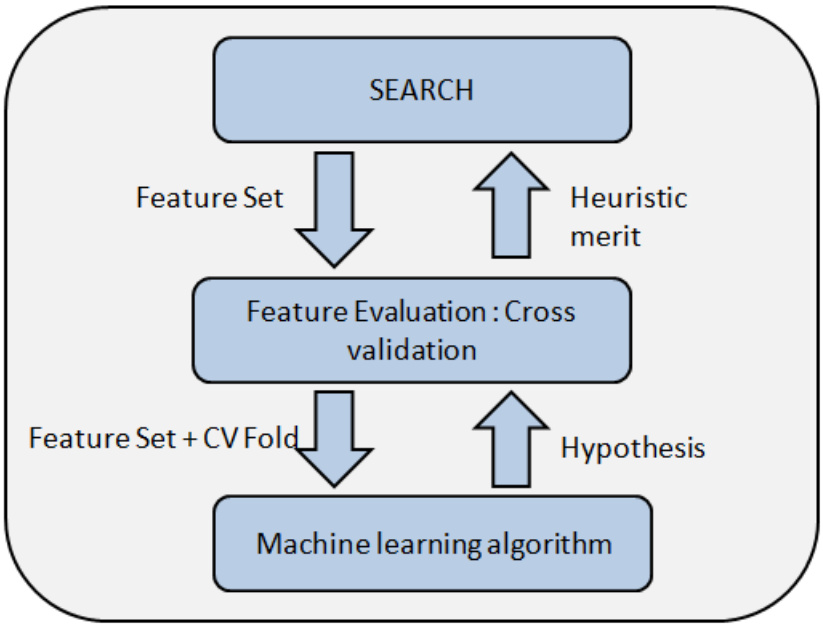


Figure 3.1: Wrapper method

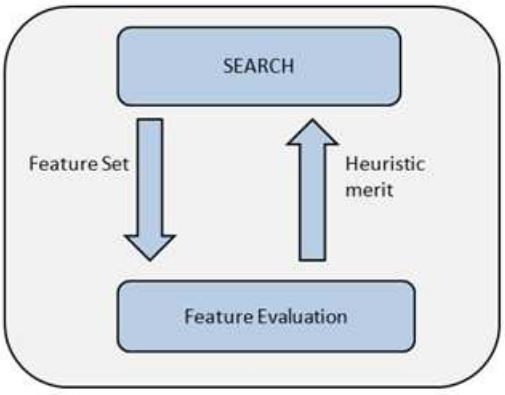


Figure 3.2: Filter method

4. Criteria to stop: It must be clear to feature selector when to stop.Generally feature selector will stop adding feature or discarding feature if it is not making any change to accuracy of model or no other better alternative is available.Other stop-

ping criteria is to continue searching until feature selector reaches opposite point of space and then choosing best subset of features after evaluating all of them.

Chapter 4

Tools Used

Tools used in this research is weka and eclipse.

4.1 Weka

One of the most common and popular tool for data mining task is weka.This is software written in java language.This contains bundle of machine learning algorithms for data mining.Also this software have capability to be called from java code itself.

Weka have capability to perform number of tasks:

1. Clustering
2. Classification
3. Preprocessing
4. Visualization
5. Association

Weka is also suitable for generating new machine learning scheme.Features of weka are:

- Weka tool have graphical user interface
- Evaluation methods are also available to evaluate learning of model is accurate or not.
- Preprocessing tools are also available in weka.

Weka accept input in number of ways .It can be in form of .csv (comma seperated values), .arff(attribute relation file format), .bsi (binary serializes instances)Among them .arff is one of the convenient form.Also input given in this research is also in .arff form.

Arff format files looks like as below: .arff file is divided into two sections ,First is header section and other is body part.

- Header section section define name of the attributes used along with their datatype .This is written in following way:

  @relation ¡name of relation¿ : This is included in beginning of every .arff file .Here name is also written along with relation keyword which specifies name of the relation.

  @attribute :This keyword is written along with each attribute name and data type.Here attribute types can be of four types as follow:

  - String :This type of attributes contain data in form of text.
  - Nominal : Values are from set of predefined values.
  - Numeric : This type contains integer or real values.
  - Date :This type is used for dates.

- Body section : This section contains data .Keyword used to specify body section is @data

  Content after @data is used as instance and used for training and testing.Each row after @data represents each record and each comma seperated value in each record is associated with attributes mentioned in header section

  Example of arff file format is :

  @relation cricket

  @attribute playerName string

  @attribute dob date ”YYYY-MM-DD”

  @attribute height numeric

  @data

  Andrew,1980-11-09,6

  In weka[25] attributes are case sensitive.Also weka expect values in data section to be declared in same format as mentioned in attribute section.As shown in above example Andrew is first value of a instance because first attribute name used in header section is player name followed by dob and height.The arff file used in our research file looks as given in figure 4.1 .Here age,sex ,cp , chol ,fbs , restecg

,thalach,oldpeak , slope , ca are features used for prediction. Rows after @data are instances where each row belongs to record of each individual.

```
@attribute 'age' real
@attribute 'sex' real
@attribute 'cp' real
@attribute 'trestbps' real
@attribute 'chol' real
@attribute 'fbs' real
@attribute 'restecg' real
@attribute 'thalach' real
@attribute 'exang' real
@attribute 'oldpeak' real
@attribute 'slope' real
@attribute 'ca' real
@attribute 'thal' real
@attribute 'num' {yes,no}


@data
63.0,1.0,1.0,145.0,233.0,1.0,2.0,150.0,0.0,2.3,3.0,0.0,6.0,no
67.0,1.0,4.0,160.0,286.0,0.0,2.0,108.0,1.0,1.5,2.0,3.0,3.0,yes
67.0,1.0,4.0,120.0,229.0,0.0,2.0,129.0,1.0,2.6,2.0,2.0,7.0,yes
37.0,1.0,3.0,130.0,250.0,0.0,0.0,187.0,0.0,3.5,3.0,0.0,3.0,no
```

Figure 4.1: data

Output of learning process of weka is given in form of few parameters named as :

- Accuracy : One of the metric to measure performance of the model is accuracy.This is given by weka as output parameter which is calculated as follows:

$$Accuracy = (\frac{Instancescorrectlyclassified}{Totalinstances}) * 100 \tag{4.1}$$

- Confusion matrix :It is matrix of X*X where X is number of classes of target variable .Classification problem we are addressing in this paper contains two classes either positive or negative.For this matrix of 2*2 is constructed which shows correct and incorrect classified instances by chosen model with respect to actual outcome(target).Higher number of instances in diagnol of matrix means higher is accuracy of the model.Basic terms related to confusion matrix are:

  True Positive: In this case model predicted yes and actual outcome is also yes.

  True Negative: In this model predicted no same as actual outcome.

  False Positive(Type 1 error):Model predicted yes but actual outcome is no.

False Negative (Type 2 error): Model predicted no but actual outcome is yes.

- Precision :This metric is calculated with respect to each classes.It is calculated as follows:

$$Precision(Y) = (\frac{No.of Instances correctly classified of class Y}{No.of instances which are classified as belonging to class Y}) \tag{4.2}$$

- Recall : This is similar to precision.It is given by:

$$Recall(Y) = (\frac{No.of Instances correctly classified of class Y}{No.of instances of class Y}) \tag{4.3}$$

4.2 Eclipse

Eclipse is one the most popular java[26] IDE(Integrated Development Environment).This also allows to use plug-ins to provide support at run time.

Latest stable version available of eclipse is neon.Mainly it is used to develop java application but it also provide support for other languages too like C , Ruby , Perl ,python.

SDK(software development kit) of eclipse is available which contains development tools of java.One can also extend its features by installing different plug-ins according to need.This kit is open source i.e. available to everyone for use.

Eclipse has following features :

1. Stability and robustness:
2. Automatic error reporting:
3. Performance
4. Support of java tools

Along with this eclipse also have one of the most important feature of debugging mode. Debugging in eclipse allows to make a program interactive by looking variables and loops used in a program.

This is done by introducing few points in a code .These points are known as breakpoint.Breakpoint is point where execution of program stops during debugging. Breakpoint is created by right clicking line where we want to create breakpoint and then by

selecting toggle breakpoint .Other approach is by double clicking on line.

Chapter 5

Implementation

Work of research work is done in few stages as mentioned below in figure 5.1:

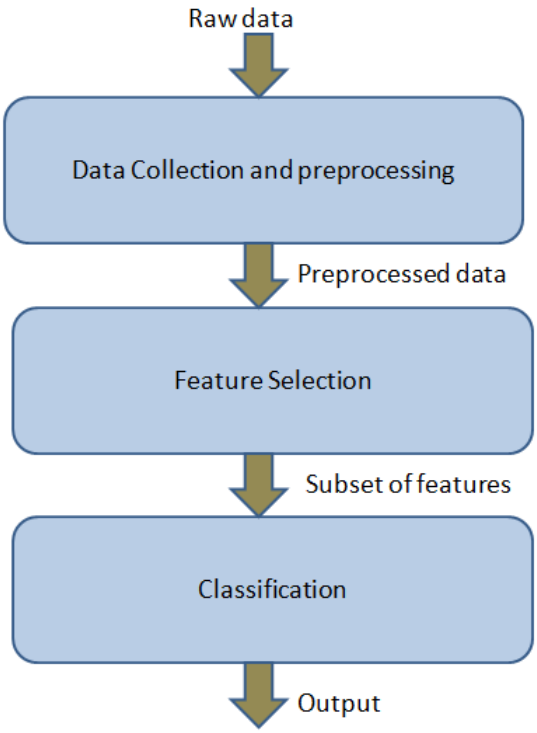


Figure 5.1: Flow chart

5.1 Data Collection and preprocessing

This section is about collection of data and preprocessing.These terms can be written with respect to data transformation , integration and cleansing

5.1.1 Data integration

Data integration is shown in figure 5.2 .Data is collected from multiple sources.For our research work data is collected from various resources.Raw data collected from different sources is shown in table 5.1 .Data shown in table 5.1 is raw

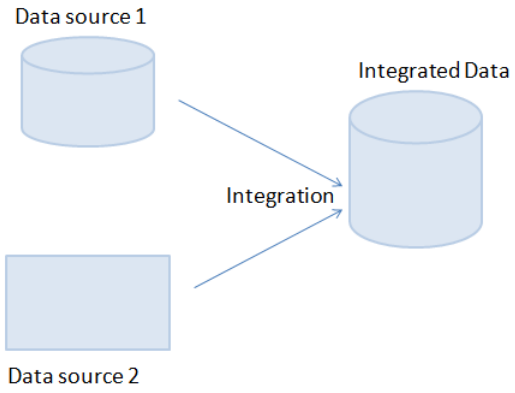


Figure 5.2: Integration of data

5.1.2 Data transformation

After integrating data from multiple sources data further preprocessing is done. Here target variable have 5 possible values i.e. 0,1,2,3,4 which represents absence(0), presence(1,2,3,4).Here 1,2,3,4 have same significance for our research work i.e. presence of heart disease.This raw data is processed to 2 classes 0 and 1 .Here 0 represents absence and 1 represents presence.Here 1 is integrated result of 1,2,3,4.This process is known as data transformation.

Data after transformation is shown in table 5.2 where last feature named as num is transformed to two classes i.e. presence or absence in term of 1 , 0 respectively.

5.1.3 Data Cleansing

Cleansing of data is known to be as crucial part of data science.One who have capability of extracting useful data from noise is said to a true data scientist.As data present in real world is dirty.This raw data can be of following types:

1. Noisy : Data is noisy if contain outliers or errors in one or other form.

Table 5.1: Raw data

Age	Sex	Cp	Trestbps	chol	fbs	ecg	thalach	exang	o.p	slope	ca	thal	num
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
53	1	3	130	250	0	0	187	0	3.5	3	0	3	0

Table 5.2: Processed data

Age	Sex	Cp	Trestbps	chol	fbs	ecg	thalach	exang	o.p	slope	ca	thal	num
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	1
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	1
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	1
53	1	3	130	250	0	0	187	0	3.5	3	0	3	0

2. Incomplete : In incomplete data few features in data doesn't contain values .
3. Inconsistent : Inconsistent data is data with multiple value for same attributes.

5.1.3.1 Dealing with missing values

There are numerous ways to deal with missing values .Few of them are listed below:

1. By ignoring the tuple : This approach is usually followed when class label is missing or number of miss for each attribute doesn't varies.
2. Manually filling missing values :This approach is very tedious.
3. Global constant is used to replace missing value :This is the case where instead of missing value "none" or "unknown" keywords are used

In our work raw data also contain missing values .In place of missing values ”?” is placed.This is shown in below table 5.3 .Here in 9th and 10th instance 12th and 2nd attribute are missing . So tuple containing missing data is ignored using approach first.

Table 5.3: Instances with missing value

Age	Sex	Cp	Trestbps	chol	fbs	ecg	thalach	exang	o.p	slope	ca	thal	num
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	1
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	?	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	1
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	?	7	1
53	?	3	130	250	0	0	187	0	3.5	3	0	3	0

5.2 Feature selection

This stage includes selection of features from available dataset.As discussed in chapter 3 about feature selection improve accuracy by removing redundant or irrelevant features .

There are number of approach of selecting feature.To implement feature selection number of algorithms are available like chi square , info gain ,gain ratio. But of the available approach we proposed a new approach known to be as hybrid approach .Here we consolidated results of two different algorithms and used then to derive results.

Here we will compare accuracy of two different models with features selected randomly to accuracy of features selected by proposed approach using weka.

Figure 5.3 represents interface of weka tool and figure 5.4 shows view of loading arff file in weka .Also 5.5 represents view of heartdisease.arff file in weka.

Here input file named as ”heartdisease.arff” contains:

Relation : heartdisease

Attributes : 14

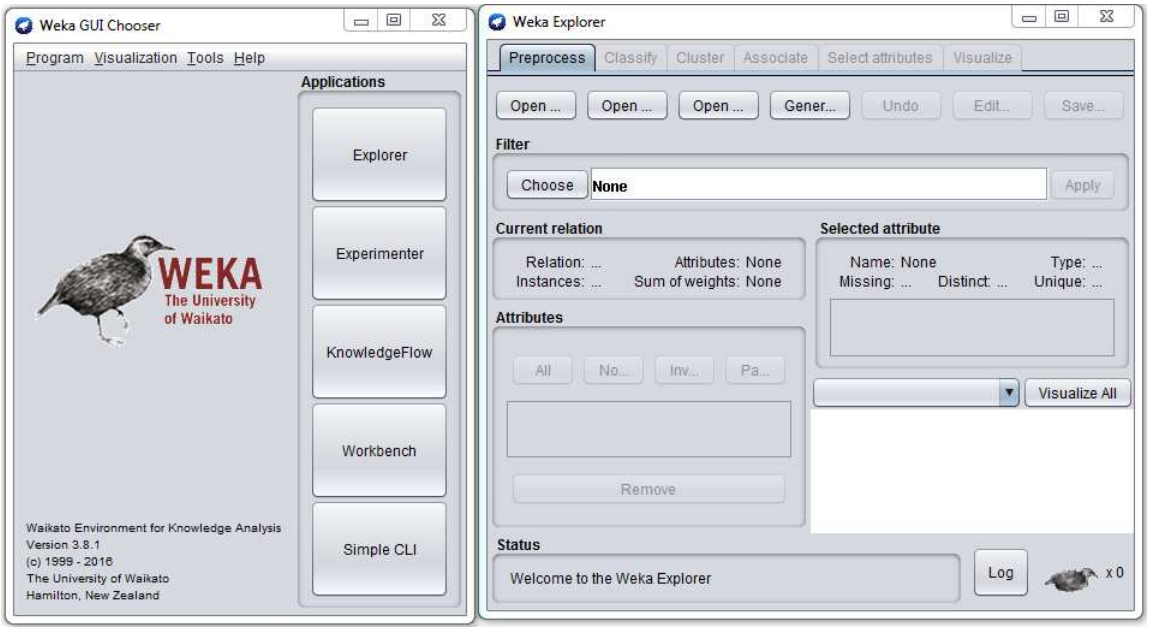


Figure 5.3: Weka interface

- Age : This attribute represents age in years
- Sex : Two possible values are 0 or 1 .Here 1 represents male and 0 female.
- Cp : This attribute is related to chest pain and categorized into 4 possible values :
    1. value 1 represents typical angina
    2. value 2 represents atypical angina
    3. value 3 related to non anginal pain
    4. value 4 related to asymptomatic
- Trestbps : This term is related to blood pressure
- Chol : This represents cholestoral[27] in mm/dl.
- Fbs : This is check fasting blood sugar is greater then 120 mg/dl or not .If true then 1 else 0.
- Restecg : This stands for resting electrocardiographic results.This attribute have 3 possible values :
    1. value 0 represents normal
    2. value 1 means having ST-T problems
    3. value 2 means having left ventricular hypertrophy

- Thalach : This value is maximum value of heart rate achieved.
- Exang : This is true if exercise induced angina otherwise no and 1 for true and 0 for false.
- Oldpeak : This is related to ST depression which may be induced by exercises.
- Slope : This is related to slope of peak exercise and have 3 possible values :
    1. value 1 means upsloping
    2. Here 2 means flat
    3. Value 3 means downsloping
- Ca : This represents number of major vessels which are colored by process of fluoroscopy(This is similar to x-rays.).Here it can have 4 possible values of this attribute 0-3.
- Thal : Thal is the attribute having three possible values:
    1. Value 3 : This means normal
    2. Value 6 : This represents fixed defect
    3. Value 7 : While this is related to reversible defect.
- Num : This is target variable of data set.

5.2.1 Classification with randomly selected features

In this section of available dataset , few feature are removed randomly using remove feature of weka. Total 14 attributes are available in dataset anf from then randomly[28] few of them are removed as shown in figure 5.6.

Randomly chosen features are used for classification using naive bayes model as shown in figure 5.7.Similarly this process is repeated with random forest classifier as shown in figure 5.8

As shown in figure 5.7 and 5.8 accuracy of randomly chosen attributes with naive bayes classifier and random forest is 78% and 84%.

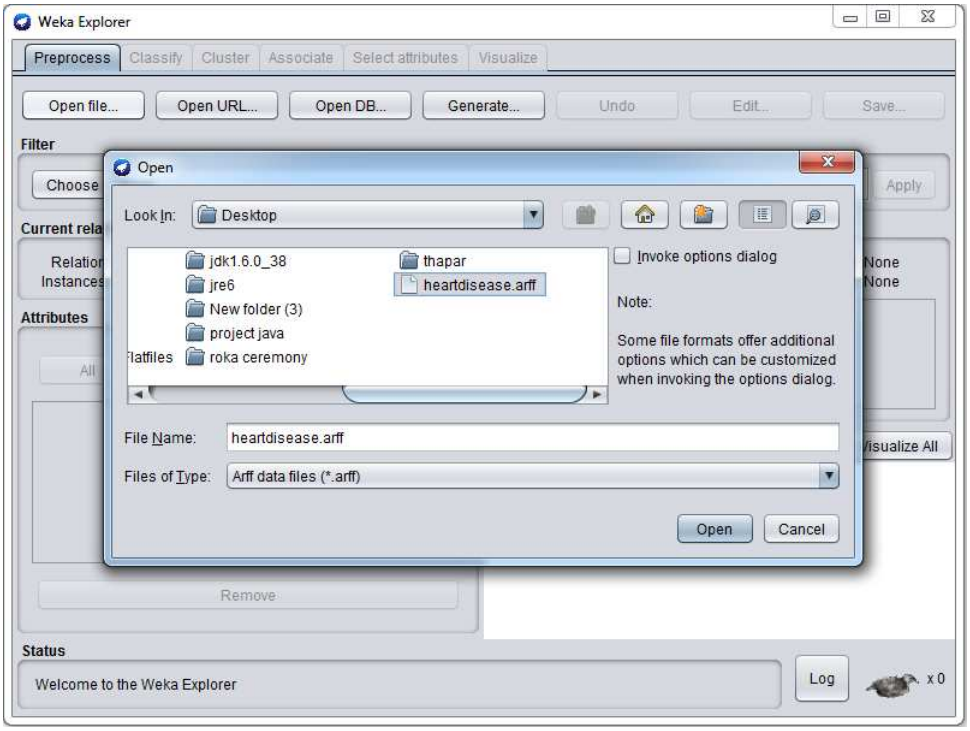


Figure 5.4: Opening weka file

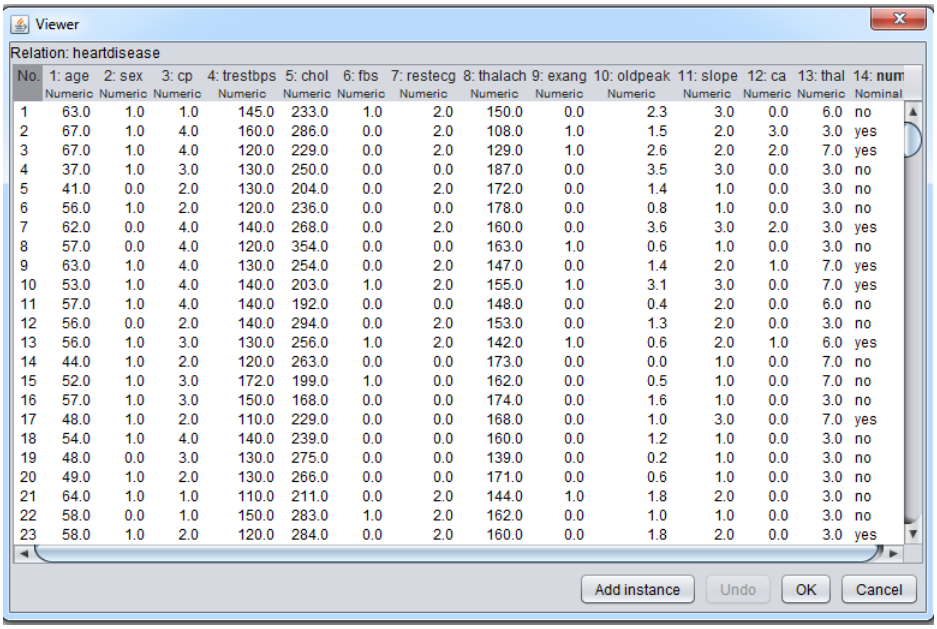


Figure 5.5: Viewer in weka

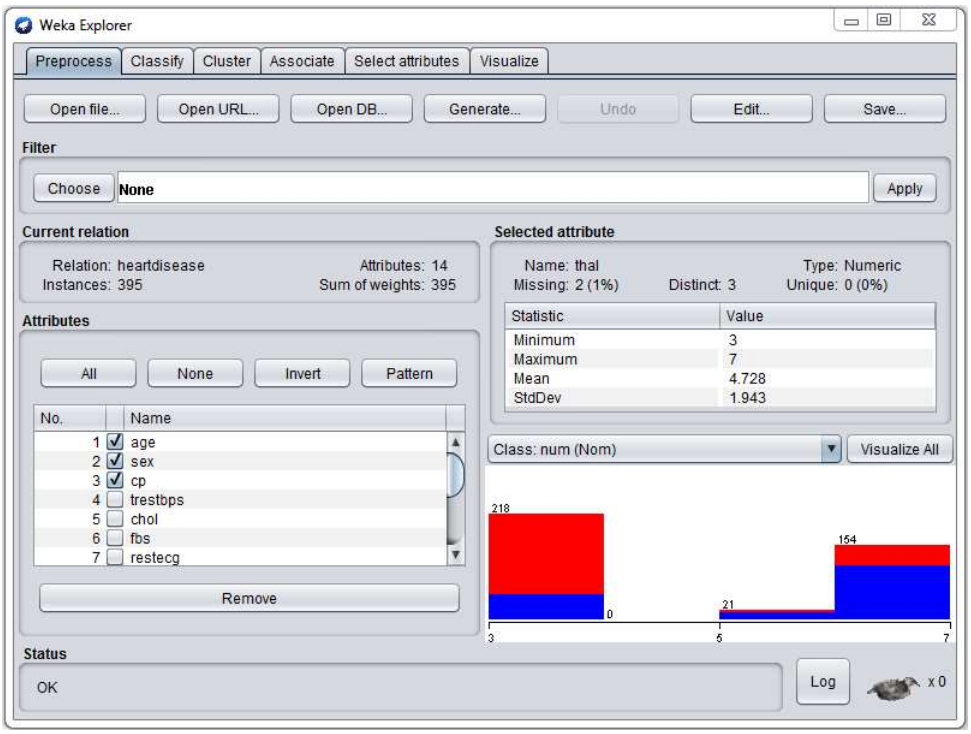


Figure 5.6: Randomly chosen attributes for removal

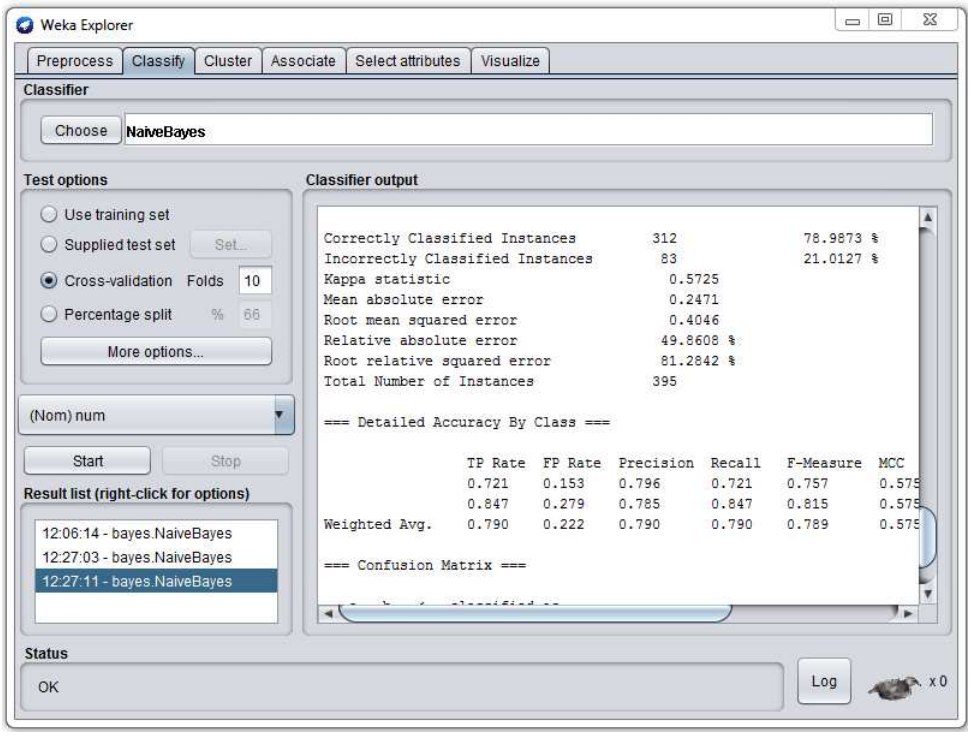


Figure 5.7: Classification with naive bayes

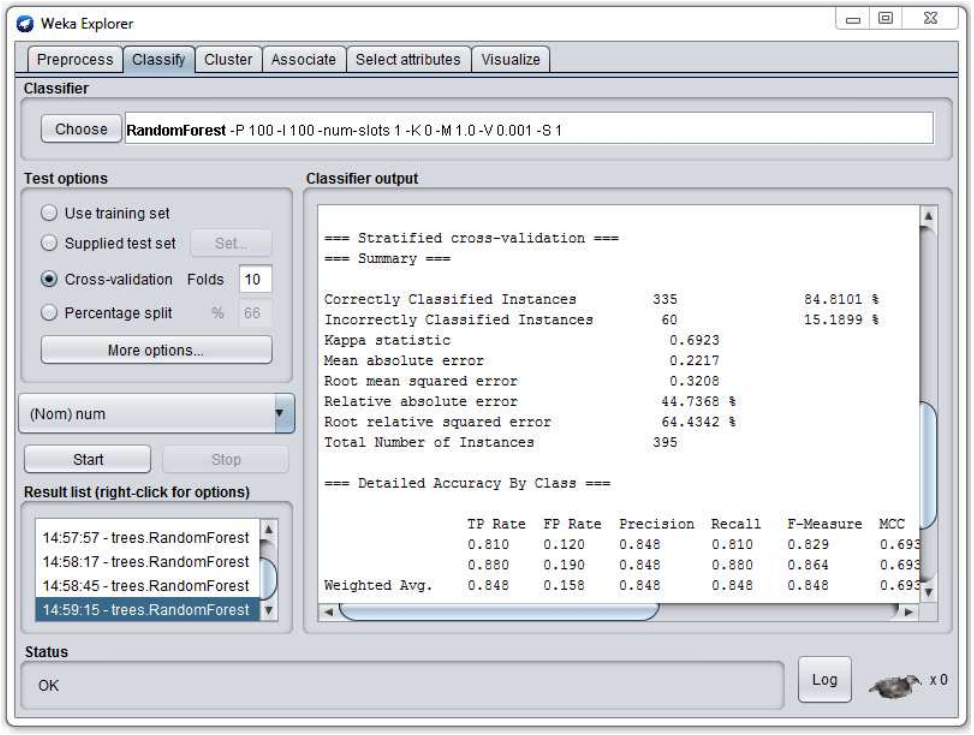


Figure 5.8: Classification with random forest

5.2.2 Proposed approach of selecting features

For feature selection two different algorithms are used for ranking and evaluation and manually both the results are consolidated to derive the final results.As a result features are filtered out on basis of weight assigned[29] . In this section we will explain how features are selected by implementing algorithms.

5.2.2.1 Score

Here we used gain ratio[30] algorithm for calculating score of each feature.As we know for choosing an attribute this algorithm considers number and size of branch .It is improved version of info gain algorithm as it is not influenced towards a multi valued attribute .Gain ratio applies normalization by using split information value. The split information value is information generated by dividing training data A into k partitions, corresponding to k outcomes on attribute X.

$$SplitInfo_X(A) = -\sum_{j=1}^{k}(\frac{|A_j|}{|A|}) * log_2(\frac{|A_j|}{A}) \tag{5.1}$$

Gain Ratio of attribute is gain divided by split info of attribute. Here to implement gain ratio Java machine learning library is used.In this library we are given with gain ratio algorithm which is implemented by a wrapper method shown in figure 5.9

Here score of each attribute is calculated which is shown in table 5.4

Table 5.4: Score

Sno.	Attribute name	Score
1	Age	0.0274962499058939
2	Sex	0.063297055456306
3	Cp	0.118028619017533
4	Trestbps	0.0132773408936235
5	Chol	0.0154015946305047
6	Fbs	0.00075790953402917
7	restEcg	0.0221938071759965
8	Thalach	0.0494500027934771
9	exang	0.152639384089965
10	oldpeak	0.0743596413697911
11	slope	0.086879961531789
12	ca	0.116583963111329
13	thal	0.165872829980505

Here different attributes have different score computed using gain ratio algorithm as shown in table 5.4 . Attribute with higher score will have high preference.As shown in table 5.4 exang have higher value and fbs (fasting blood sugar) have least value.

```
import net.sf.javaml.core.Dataset;
import net.sf.javaml.featureselection.scoring.GainRatio;
import net.sf.javaml.tools.data.ARFFHandler;
import net.sf.javaml.tools.data.FileHandler;

public class score {
    /**
     * Shows the basic steps to create use a feature scoring algorithm.
     *
     * @author Thomas Abeel
     *
     */
    public static void main(String[] args) throws Exception {
        /* Load the iris data set */
        //Dataset data = FileHandler.loadDataset(new File("C:/Users/DELL/workspace/kanika_eclipse/nww/heartdisease.arff"), 1);
        Dataset data = ARFFHandler.loadARFF(new File("C:/Users/DELL/workspace/kanika_eclipse/nww/heartdisease.arff"),13);
        System.out.println("entered");
        GainRatio ga = new GainRatio();
        /* Apply the algorithm to the data set */
        ga.build(data);
        /* Print out the score of each attribute */
        for (int i = 0; i < ga.noAttributes(); i++)
            System.out.println(ga.score(i));
    }

}
```

Figure 5.9: Score calculation

5.2.2.2 Rank

After computing score of features now we will compute rank of features individually.Wrapper method of computing rank is shown in figure 5.10.Here to implement rank of features we will use Support vector machine recursive feature elimination(SVM-RFE) algorithm is used.With each feature a rank is associated and attribute with minimum rank is most preferable while attribute with rank is least preferable.

SVM-RFE is a efficient approach of selecting features .SVM-RFE ranks features according to weight assigned by linear SVM. Here linear SVM is applied to dataset and percentage of worst ranked attributes are eliminated and svm is trained again with remaining attributes. This process is repeated until we are left with only one attribute. Attribute retained till last is lowest ranked attribute as attribute with lowest rank is better then high rank attributes and attribute eliminated first is highest ranked attribute. Similarly rank is assigned to all attributes.

Here different attributes will have different rank computed using svm-rfe algorithm as shown in table 5.5.Attributes with minimum rank will have high preference and maximum value attribute have least preference.As shown in table 5.5 Age has minimum rank while oldpeak have maximum rank.

Table 5.5: Rank

Sno.	Attribute name	Rank
1	Age	1
2	Sex	2
3	Cp	3
4	Trestbps	4
5	Chol	6
6	Fbs	8
7	restEcg	10
8	Thalach	11
9	exang	9
10	oldpeak	13
11	slope	12
12	ca	7
13	thal	5

After computing rank and score of each features using svm-rfe and gain ratio respectively ,now result of both the algorithms are consolidated to compute final weight of each feature .This is done as follows :

- We will compute weight of each attribute using formula which is given as below:

$$Weight(A) = (\frac{score_A}{rank_A + 1}) * 1000 \tag{5.2}$$

  where A is attribute of available dataset and score(A) is score corresponding to attribute A and rank(A) is rank corresponding to attribute A.

- Here we will substitute value of rank and score of each feature .This will give weight of each feature.

- In this way we will choose subset of feature on basis of weight assigned to them

```
public class rank {
    /**
     * Shows the basic steps to create use a feature ranking algorithm.
     *
     * @author Thomas Abeel
     *
     */
    public static void main(String[] args) throws Exception
    {
        /* Load the iris data set */
        //Dataset data = FileHandler.loadDataset(new File("devtools/data/iris.data"), 4, ",");
        Dataset data = ARFFHandler.loadARFF(new File("C:/Users/DELL/workspace/kanika_eclipse/nww/heartdisease.arff"));
        /* Create a feature ranking algorithm */
        RecursiveFeatureEliminationSVM svmrfe = new RecursiveFeatureEliminationSVM(0.2);
        /* Apply the algorithm to the data set */
        System.out.println("inside");
        svmrfe.build(data);
        /* Print out the rank of each attribute */
        for (int i = 0; i < svmrfe.noAttributes(); i++)
            System.out.println(svmrfe.rank(i));
    }
```

Figure 5.10: Rank calculation

	A	B	C	D	E
1	attributes	rank	score	score /(rank +1)	Weight*1000
2	'age' real	1	0.02749625	0.013748125	13.74812495
3	'sex' real	2	0.063297055	0.021099018	21.09901849
4	'cp' real	3	0.118028619	0.029507155	29.50715475
5	'trestbps'	4	0.013277341	0.002655468	2.655468179
6	'chol' real	6	0.015401595	0.002200228	2.200227804
7	'fbs' real	8	7.58E-04	8.42122E-05	0.08421217
8	'restecg' r	10	0.022193807	0.002017619	2.017618834
9	'thalach' r	11	0.049450003	0.004120834	4.120833566
10	'exang' re	9	0.152639384	0.015263938	15.26393841
11	'oldpeak'	13	0.074359641	0.005311403	5.311402955
12	'slope' re	12	0.086879962	0.006683074	6.683073964
13	'ca' real	7	0.116583963	0.014572995	14.57299539
14	'thal' real	5	0.16587283	0.027645472	27.64547166
15					
16					

Figure 5.11: Result

From figure 5.11we can arrange feature according to weight and use them to perform classification

As shown in figures 5.13 and 5.12 accuracy obtained in said figures is better then accuracy obtained in previous approach .

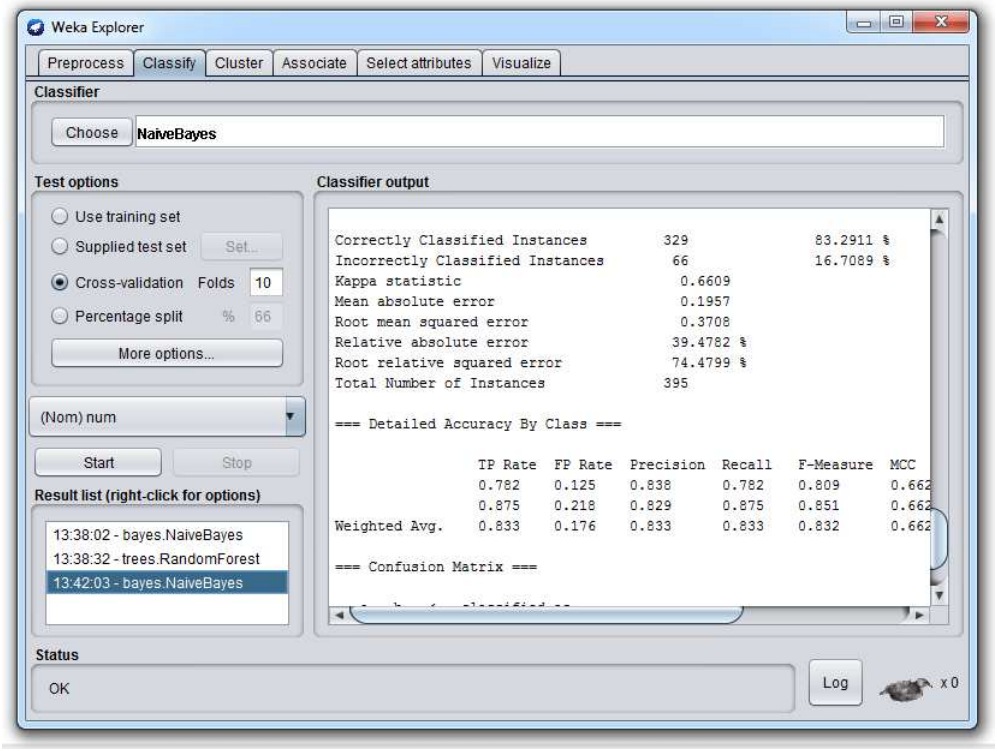


Figure 5.12: Naive Bayes

As shown in figure 5.11 we have listed all features and along with each feature rank and score are listed .Using rank and feature of each feature weight is computed using equation 5.2 as shown in figure 5.11

Features selected by proposed approach is listed in figure 5.14 .Feature which are included are listed below

- Age
- Sex
- Cp
- thalach
- exang
- old peak
- slope
- ca
- thal

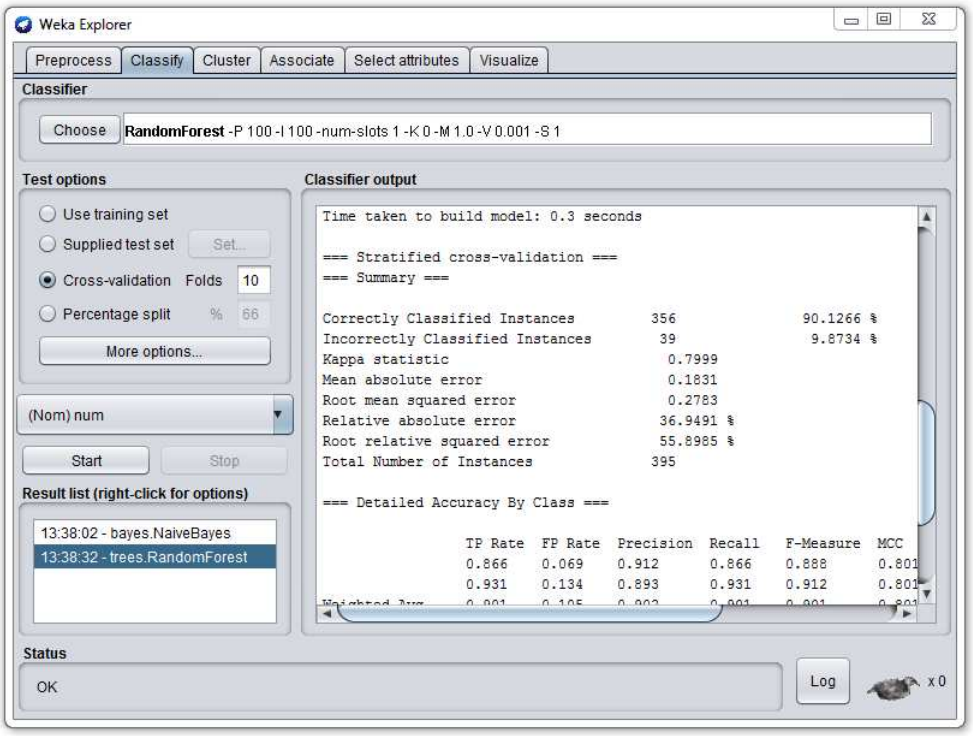


Figure 5.13: Random Forest

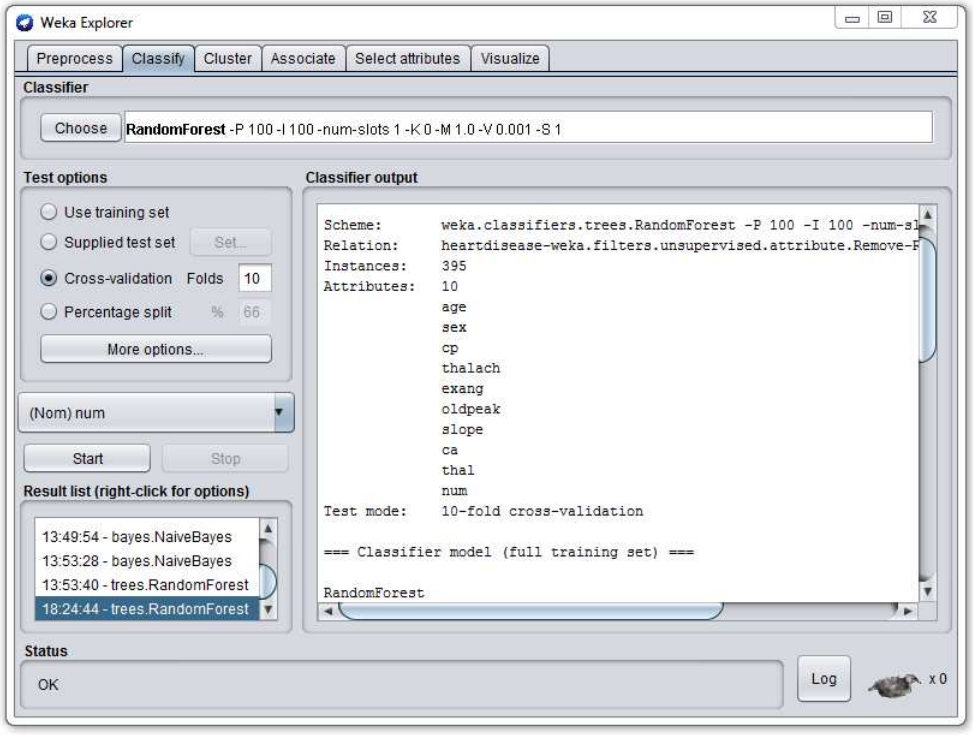


Figure 5.14: Snapshot of selected features

Chapter 6

Evaluation

Here in this section we will discuss metrics to evaluate performance of different classifiers.In other words evaluation metric is tool which is used to measure the performance of different classifiers.

Different criteria are available for evaluation :

- Speed : One of the important criteria is measuring speed for evaluation by calculating total computational cost consumed in generating and using the model.
- Robustness : This criteria is to check ability of making correct predictions even after presence of missing values and noise.
- Predictive accuracy : Other criteria is to check how well model classifies never seen instances.
- Scalability :This criteria is related to ability to generate model for large set of data as a input.

Further evaluation parameters [5] used in this work for evaluating classifiers are as follows :

6.1 Confusion matrix

One of the common and important metric is confusion matrix.As the name suggest confusion matrix is matrix of n x n where n is total number of classes of a target variable.Also this is matrix between actual classes to predicted classes. Here row represents predicted classes while column represents actual classes.

Confusion matrix[31] is shown in table 6.1 and terms related to confusion matrix are :

- tp , tn : These terms represents positive instances that are correctly classified too respectively.

- fp ,fn : These terms are related to positive and negative instances that are are misclassified.

Table 6.1: Confusion matrix

	Actual Yes	Actual No
Predicted Yes	true positive	false positive
Predicted No	false negative	true negative

Using 6.1 we can derive number of other evaluation metrics listed as follows :

1. Accuracy : This is computed by dividing correctly classified instances to total instances.This is given as :

$$Accuracy = \frac{tp+tn}{tp+fp+tn+fn} \tag{6.1}$$

2. Error rate : This is another metric of evaluation which computes misclassification which is ration of misclassified to total number of instances.This is given as :

$$Err = \frac{fp+fn}{tp+fp+tn+fn} \tag{6.2}$$

3. Specificity : This metric is used to measure true negative rate.In other words it is also known as true negative rate. This is expressed as :

$$Specificity = \frac{tn}{fp+tn} \tag{6.3}$$

4. Sensitivity : This metric is used to measure true positive rate.This is expressed as :

$$Sensitivity = \frac{tp}{tp+fn} \tag{6.4}$$

5. Precision : This is used to compute the true instances to total number of instances of positive class .This is expressed as :

$$Precision = \frac{tp}{fp+tp} \tag{6.5}$$

6. Recall : This is given as ratio of true positive instances to total number of correctly

classified instances.This is expressed as :

$$Recall = \frac{tp}{tn+tp} \tag{6.6}$$

7. F-Measure : This is harmonic mean of recall and precision which is expressed as :

$$F-M = \frac{2*p*n}{p+n} \tag{6.7}$$

Two classifiers are used in our research work including Naive Bayes and Random forest.Here confusion matrix obtained for naive bayes and random forest for proposed methodology are shown in table ... and ... respectively.

Table 6.2: Confusion matrix of naive bayes

	Actual Yes	Actual No
Predicted Yes	140	27
Predicted No	39	189

Table 6.3: Confusion matrix of random forest

	Actual Yes	Actual No
Predicted Yes	155	15
Predicted No	24	201

Values of all above metrics for classifier naive bayes and random forest are given in table

Table 6.4: Resultant

Sno.	Metric	Random Forest	Naive bayes
1	Accuracy	0.901	0.832
2	Error Rate	0.098	0.167
3	Specificity	0.9305	0.876
4	Sensitivity	0.8659	0.782
5	Precision	0.911	0.838
6	Recall	0.435	0.425

6.2 Area Under ROC Curve

ROC stands for Receiver Operating Characteristics .Roc is one of the approach to visualize performance of a classifier .This is 2 dimensional graph which is generated between false positive and true positive rate.

1. True Positive Rate :This is given by positives correctly classified to total positives.
2. False Positive Rate: This is given by negatives incorrectly classified to total negatives.

More the curve towards (0,1) coordinate of ROC space better is the model.Also (0,1) point of ROC space is known as perfect classification

Roc curve of naive bayes and random forest methods used in this paper is shown in 6.1

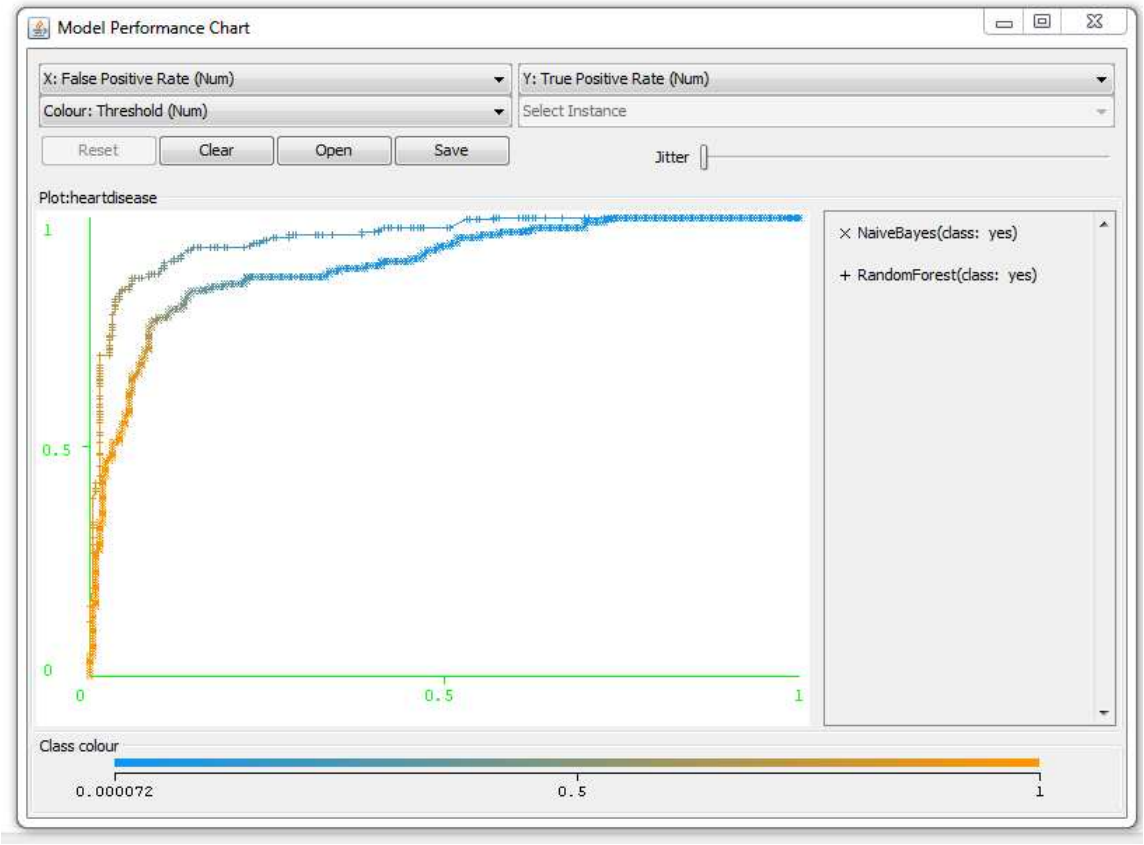


Figure 6.1: Roc curve

6.2.1 Area under curve

One of the metric used to measure quality of classification model is Area under ROC [32] curve(AUC).A classifier is said to be perfect if it has AUC equals to 1 while a random classifier has area 0.5. AUC for naive bayes is 0.8859 and for random forest is 0.9621 which means cases chosen randomly from group with target value equal to yes has larger score as compared to cases chosen from group with value equal to NO in 88% of the time for naive bayes and 96% for Random forest.

6.3 Cross Validation

One of the approach of evaluating the performance of classifier is by training model with given set and then testing classifier by never seen input.This is done by process of cross validation .Requirement of cross validation is to check the stability of model .As in some cases better predictions as a result may be the outcome of over-fitting which may behave differently for different inputs or never seen inputs.So to avoid problem of over-fitting cross validation is used .This is done by reserving some instance which are not used for training and then later used that instances for testing.

This is done as follows :

1. Reserve few of instances for testing
2. Use remaining dataset for training a model
3. Use reserved dataset for testing .This will know actual performance of model

Cross validation can be applied in number of ways as follows :

1. Holdout method : In this approach dataset is divided into two set ,One is training dataset and other is testing dataset.Using training dataset a function is fitted and that function is used to predict output of target variable from testing dataset.This approach is totally dependent on data instances chosen for training.
2. K-Fold Cross Validation :This method is improvement of hold out method .Here dataset is divided into k sets and hold out process is repeated for k times.In each iteration one of the k set is used for testing and remaining k-1 sets are used for training.This process is repeated k times.So that every instance is once used for

testing.

$$ErrorRate = \sum_{i=1}^{k} \frac{e_i}{k} \tag{6.8}$$

3. Leave One Out Cross Validation : This method is also known as LOOCV method.This is special case of K-fold Cross Validation.As the name suggest in this approach we use one instance for testing and remaining data set are used for training .This process is repeated for each instance. Benefit of using this dataset is that this approach is tested for every single instance of available dataset.But issue related to this approach is that for larger number of instances it will take larger time.Therefore this result in higher execution time. Here error rate of n instances is given as :

$$ErrorRate = \sum_{i=1}^{n} \frac{e_i}{N} \tag{6.9}$$

Here in our work k fold cross validation is used for evaluation .Here k varies to compute accuracy with different sets .Random Forest and Naive Bayes is shown in below table 6.5

Table 6.5: Cross Validation

Sno.	Value of k	Random Forest	Naive bayes
1	14	90.6329	83.038
2	10	90.1266	83.2911
3	6	87.5949	83.038
4	2	83.5443	81.519

Chapter 7

Conclusion

The main motive of this research is to classify the data in two classes either in positive or in negative result for heart disease.

In this a hybrid approach is used for selecting subset of features from available set.Algorithms used for this research are:

1. SVM-RFE and gain ratio for feature selection by computing rank and score.
2. Naive Bayes and random forest are used for training model and then testing for unseen data.

A hybrid approach of feature selection is adopted to optimize the classification problem ,consolidated results of SVM-RFE and Gain-ratio are used to get subset of features and remove irrelevant or redundant feature.

On subset of features naive bayes and random forest are applied to classify them into presence or absence of disease.It has been shown in results that accuracy improved for both classifiers when applied to selected features. Proposed approach of feature selection not only reduced size of dataset but also enhanced the performance of both the classifiers models.

As this work is limited to binary class problems .In other words proposed methodology is bounded for binary class problem like in this work target variable is classified into either presence or absence.We can further categorize presence of disease with the amount of blockage of heart and predict time to recover from disease depending on degree of blockage.

Also an interactive system can be designed which can further reminds patient or medical professionals for their pending process of diagnosis .

Chapter 8

Plagiarism and publication list

Plagiarism Report

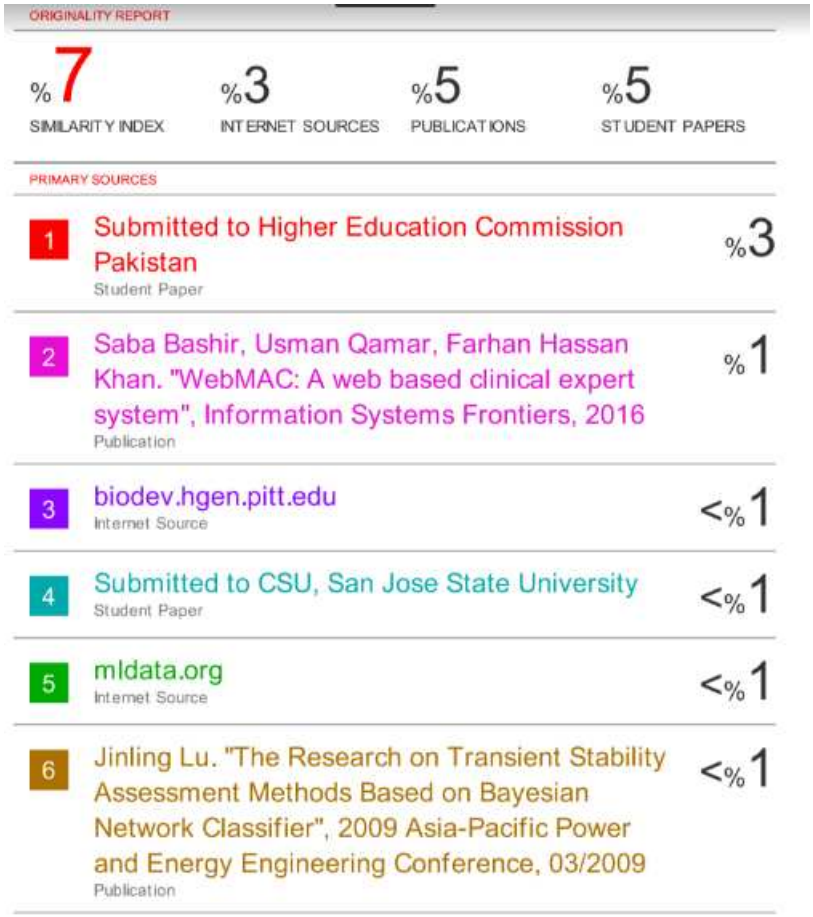


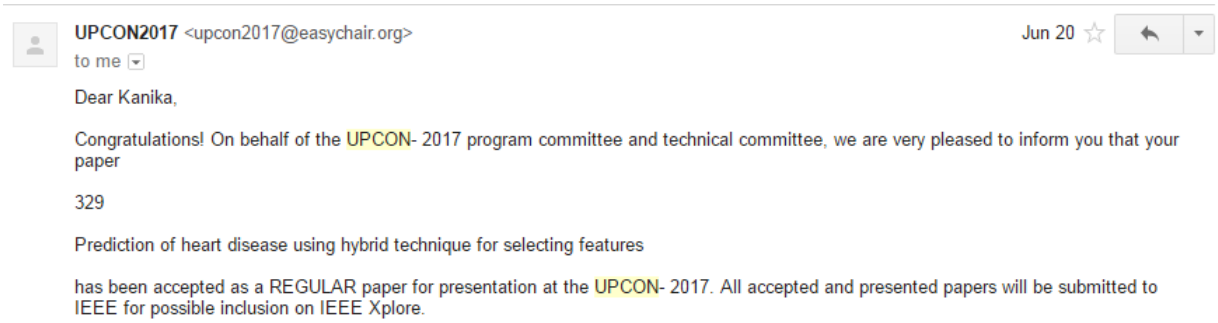
Figure 8.1: Plagiarism Report

Video Link

https://youtu.be/E38L7l-Jixc

Publications

Paper will be published in IEEE Conference :Prediction of heart disease using hybrid technique for selecting features



UPCON2017 <upcon2017@easychair.org> Jun 20

to me

Dear Kanika,

Congratulations! On behalf of the UPCON- 2017 program committee and technical committee, we are very pleased to inform you that your paper

329

Prediction of heart disease using hybrid technique for selecting features

has been accepted as a REGULAR paper for presentation at the UPCON- 2017. All accepted and presented papers will be submitted to IEEE for possible inclusion on IEEE Xplore.

Figure 8.2: Paper acknowledgement

References

[1] Edoardo Pasolli, Farid Melgani, Devis Tuia, Fabio Pacifici, and William J Emery. Svm active learning approach for image classification using spatial information. *IEEE Transactions on Geoscience and Remote Sensing*, 52(4):2217–2233, 2014.

[2] Xuemei Li, Lihong Wang, Yibin Song, and Xianjia Zhao. A hybrid constrained semi-supervised clustering algorithm. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*, volume 4, pages 1597–1601. IEEE, 2010.

[3] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

[4] Jiayi Wu, Jingmin Xin, and Nanning Zheng. Svm learning from imbalanced microanuerysm candidate datasets used feature selection by gini index. In *Information and Automation, 2015 IEEE International Conference on*, pages 1637–1641. IEEE, 2015.

[5] Nigel Williams, Sebastian Zander, and Grenville Armitage. A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification. *ACM SIGCOMM Computer Communication Review*, 36(5):5–16, 2006.

[6] Zhi-Xin Yu, Jing-Ran Chen, and Tian-Qing Zhu. A novel adaptive intrusion detection system based on data mining. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 4, pages 2390–2395. IEEE, 2005.

[7] Feng Zhang, Ya-Jun Zhao, et al. Unsupervised feature selection based on feature relevance. In *Machine Learning and Cybernetics, 2009 International Conference on*, volume 1, pages 487–492. IEEE, 2009.

[8] Niels Landwehr, Mark Hall, and Eibe Frank. Logistic model trees. *Machine Learning*, 59(1-2):161–205, 2005.

[9] Mohamad H Hassoun. *Fundamentals of artificial neural networks*. MIT press, 1995.

[10] Alex M Andrew. An introduction to support vector machines and other kernel-based learning methods by nello christianini and john shawe-taylor, cambridge university press, cambridge, 2000, xiii+ 189 pp., isbn 0-521-78019-5 (hbk,£

27.50)., 2000.

[11] J-SR Jang and C-T Sun. Functional equivalence between radial basis function networks and fuzzy inference systems. *IEEE transactions on Neural Networks*, 4(1):156–159, 1993.

[12] Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.

[13] Swati Dhankhar, Himani Rastogi, and Misha Kakkar. Software fault prediction performance in software engineering. In *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on*, pages 228–232. IEEE, 2015.

[14] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees belmont. *CA: Wadsworth International Group*, 1984.

[15] Ratna Astuti Nugrahaeni and Kusprasapta Mutijarsa. Comparative analysis of machine learning knn, svm, and random forests algorithm for facial expression classification. In *Technology of Information and Communication (ISemantic), International Seminar on Application for*, pages 163–168. IEEE, 2016.

[16] Irina Rish. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM, 2001.

[17] Harry Zhang. The optimality of naive bayes. *AA*, 1(2):3, 2004.

[18] Ashraf M Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. Multinomial naive bayes for text categorization revisited. In *Australian Conference on Artificial Intelligence*, volume 3339, pages 488–499. Springer, 2004.

[19] Michael E Mavroforakis and Sergios Theodoridis. A geometric approach to support vector machine (svm) classification. *IEEE transactions on neural networks*, 17(3):671–682, 2006.

[20] Kai-Bo Duan, Jagath C Rajapakse, Haiying Wang, and Francisco Azuaje. Multiple svm-rfe for gene selection in cancer classification with expression data. *IEEE transactions on nanobioscience*, 4(3):228–234, 2005.

[21] Yin-Wen Chang and Chih-Jen Lin. Feature ranking using linear svm. In *Causation and Prediction Challenge*, pages 53–64, 2008.

[22] Laura Elena Raileanu and Kilian Stoffel. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1):77–93, 2004.

[23] Jianhua Dai and Qing Xu. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *Applied Soft Computing*, 13(1):211–221, 2013.

[24] X. Wang, XZ Gao, and SJ Ovaska. A simulated annealing-based immune optimization method. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, Porvoo, Finland*, pages 41–47, 2008.

[25] Swasti Singhal and Monika Jena. A study on weka tool for data preprocessing, classification and clustering. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2(6):250–253, 2013.

[26] Gail C Murphy, Mik Kersten, and Leah Findlater. How are java software developers using the elipse ide? *IEEE software*, 23(4):76–83, 2006.

[27] Scandinavian Simvastatin Survival Study Group et al. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the scandinavian simvastatin survival study (4s). *The Lancet*, 344(8934):1383–1389, 1994.

[28] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

[29] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 856–863, 2003.

[30] Asha Gowda Karegowda, AS Manjunath, and MA Jayaram. Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2):271–277, 2010.

[31] HG Lewis and M Brown. A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, 22(16):3223–3235, 2001.

[32] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.