HEART DISEASES PREDICTION USING DATA MINING TECHNIQUES

VI SEMESTER

IT8611 MINI PROJECT REPORT

Submitted by

YUVARAJ R (312417205058)

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

St. JOSEPH'S INSTITUTE OF TECHNOLOGY

CHENNAI 600 119

ANNA UNIVERSITY, CHENNAI 600 025

APRIL 2020

ANNA UNIVERSITY: CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report "**HEART DISEASES PREDICT DATA MINING TECHNIQUES**" is the bonafide work of **YUVARAJ R (312417205058)** who carried out the project work under my supervision, for the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Information Technology.

SIGNATURE
Dr. R. PRISCILLA, M.E., Ph.D
Professor
HOD (Lab Affairs)
Department of Information Technology
St.Joseph's Institute of Technology
Old Mamallapuram Road
Chennai-600119

SIGNATURE
Dr.A. MENAKA PUSHPA,B.E,M.E.,Ph.D.
Associate Professor
SUPERVISOR
Department of Information Technology
St.Joseph's Institute of Technology
Old Mamallapuram Road
Chennai-600119

Submitted for the Viva-Voce held on _____

(INTERNAL EXAMINER) **(EXTERNAL EXAMINER)**

CERTIFICATE OF EVALUATION

College Name : St. Joseph's Institute of Technology

Branch & Semester : Information Technology (VI)

S.No.	NAMES OF STUDENTS	TITLE OF THE PROJECT	NAME OF THE SUPERVISOR WITH DESIGNATION
1. 2. 3.	MANOJ M (312417205026) YUVARAJ R (312417205058) ZAKIR HUSSAIN B (312417205059)	HEART DISEASES PREDICTION USING DATA MINING TECHNIQUES	Dr.A.MENAKA PUSHPA,M.E.,Ph.D. Associate Professor

The report of the project work submitted by the above students in partial fulfillment for the award of Bachelor of Technology degree in Information Technology of Anna University were evaluated and confirmed to be reports of the work done by the above students and then evaluated.

(INTERNAL EXAMINER) **(EXTERNAL EXAMINER)**

ABSTRACT

Data Mining is the most popular knowledge extraction method for knowledge discovery (KDD).The healthcare industry contains a huge amount of data. But most of it is not effectively used. Heart disease is one of the main reasons for the death of people in the world. Nearly 47% of all deaths are caused by heart diseases. We use four algorithms including Decision Table Hoeffding Tree, Naïve Bayes and Sequential Minimal optimization to predict heart diseases. Accuracy of the prediction level is high when using more number of attributes. Using ROC curve, the prediction technique is identified effectively. Our aim is to perform predictive analysis using these data mining techniques on heart diseases and conclude which techniques are effective and efficient.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Expansion	Page No
ARFF	Attribute-Relation File Format	6
WEKA	Waikato Environment for Knowledge Analysis	17
SMO	Sequential Minimal Optimization	20
ROC	Receiver Operating Characteristic Curve	30

TABLE OF CONTENT

CHAPTER 1
INTRODUCTION

1.1 DATA MINING

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is an essential process where intelligent methods are applied to extract data patterns. Data mining may be accomplished using classification, clustering, prediction, association and time series analysis. Data mining applications will be used for better health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths.

1.2 PROBLEM STATEMENT

Medical diagnosis plays a vital role and yet a 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, hypertension, family history of heart disease, obesity and lack of physical exercise, fasting blood sugar etc. We use Waikato Environment for Knowledge Analysis (WEKA) as a data mining tool.

1.3 HEART DISEASE

The term Heart sickness alludes to illness of heart & vessel framework inside it. Heart illness is a wide term that incorporates different sorts of illnesses influencing diverse segments of the heart. Heart signifies "cardio." Therefore, all heart sicknesses fit in with the class of cardiovascular ailments.

1.3.1 SYMPTOMS

The most common symptom of coronary artery disease is angina, or chest pain. Angina can be described as a discomfort, heaviness, pressure, aching, burning, fullness, squeezing, or painful feeling in your chest. It can be mistaken for indigestion heartburn. Angina may also be felt in the shoulders, arms, neck, throat, jaw, or back.

Other symptoms of coronary artery disease include:

- Shortness Of Breath .
- Palpitations(irregular heart beats,or a “flip-flop" feeling your chest)
- A fast Heartbeat
- Weakness Or Dizziness
- Nausea
- Sweating
- light-headedness and dizzy sensations
- high levels of fatigue

1.3.2 TYPES OF HEART DISEASES

There are many types of heart disease that affect different parts of the organ and occur in different ways. The below fig 1.1 describes the types of heart diseases.

Coronary Artery Disease (**CAD**)is the most common type of heart disease. In CAD, the arteries carrying blood to the heart muscle (the coronary arteries) become lined with plaque, which contains materials such as cholesterol and fat.

This plaque buildup (called atherosclerosis) causes the arteries to narrow, allowing less oxygen to reach the heart muscle than it needs to work properly. When the heart muscle does not receive enough oxygen, chest pain (angina) or heart attack can occur.

Arrhythmia is an irregular or abnormal heartbeat. This can be a slow heart beat (bradycardia), a fast heartbeat (tachycardia), or an irregular heartbeat. Some of the most common arrhythmias include atrial fibrillation (when the atria or upper heart chambers contract irregularly), premature ventricular contractions (extra beats that originate from the lower heart chambers, or ventricles), and bradyarrhythmias (slow heart rhythm caused by disease of the heart's conduction system).

Heart Failure(congestive heart failure, or CHF) occurs when the heart is not able to pump sufficient oxygen-rich blood to meet the needs of the rest of the body. This may be due to lack of force of the heart to pump or as a result of the heart not being able to fill with enough blood. Some people have both problems.

Heart valve Disease occurs when one or more of the four valves in the heart are not working properly. Heart valves help to ensure that the blood being pumped through the heart keeps flowing forward. Disease of the heart valves (e.g., stenosis, mitral valve prolapse) makes it difficult.

Heart Muscle Disease(cardiomyopathy) causes the heart to become enlarged or the walls of the heart to become thick. This causes the heart to be less able to pump blood throughout the body and often results in heart failure.

Congenital Heart Disease is a type of birth defect that causes problems with the heart at birth and occurs in about one out of every 100 live births. Some of the most common types of congenital heart disease include:

atrial septal defects (ASD) and ventricular septal defects (VSD), which occur when the walls that separate the right and left chambers of the hearts are not completely closed.

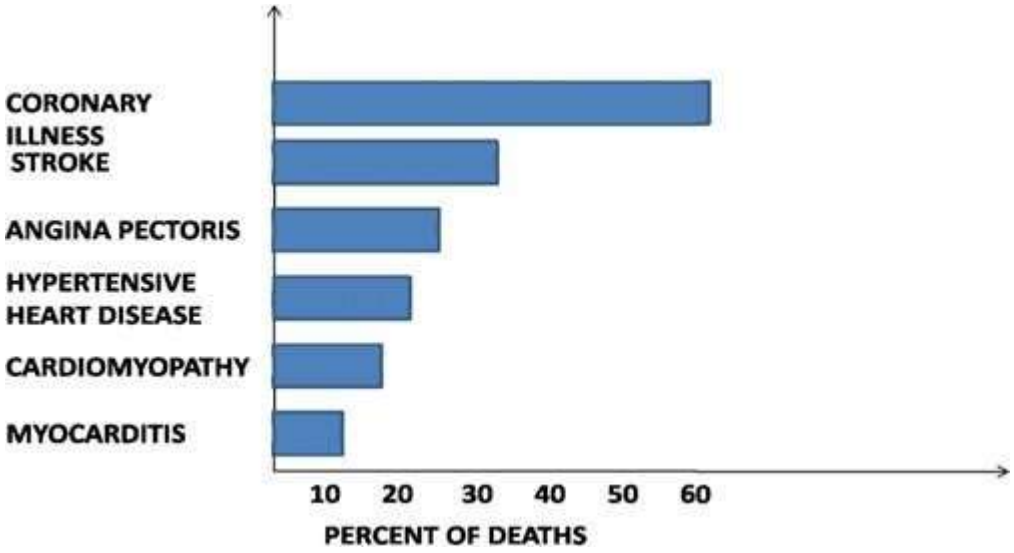

Fig1.1-Types of heart diseases

1.4 OBJECTIVE

1.4.1 General Objective

The main objective is to perform predictive analysis using data mining techniques on heart diseases .

1.4.2 Specific Objective

- Heart disease prediction system can assist medical professionals in predicting heart disease based on the clinical data of patients
- Identifying heart disease patients using various algorithms.
- Simulate the predictive analysis using the WEKA tool and include the data sets in ARFF format.
- Helps avoid human biases
- Reduce the cost of medical tests

1.5 SCOPE OF THE PROJECT

Using different classifiers within the WEKA tool, are deriving the decisions out of it, would help the system to predict the likely presence of heart disease in the patients and will definitely help to diagnose heart disease well in advance and be able to cure it in right time.

1.6 DATA MINING TECHNIQUES

1.6.1 Naive Bayes

Naive Bayes is among one of the most simple and powerful algorithms for classification based on Bayes' Theorem with an assumption of independence among predictors. Naive Bayes model is easy to build and particularly useful for very large data sets. There are two parts to this algorithm:

- Naive
- Bayes

The Naive Bayes classifier assumes that the presence of a feature in a class is unrelated to any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that a particular fruit is an apple or an orange or a banana and that is why it is known as "Naive".

Bayes Theorem

In Statistics and probability theory, Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. It serves as a way to figure out conditional probability.Given a Hypothesis H and evidence E, Bayes' Theorem states that the relationship between the probability of Hypothesis before getting the evidence P(H) and the probability of the hypothesis after getting the evidence P(H|E) is :

$$\mathbf{P(H|E)=P(E|H).P(H)/P(E)}$$

This relates the probability of the hypothesis before getting the evidence P(H), to the probability of the hypothesis after getting the evidence, P(H|E). For this reason, is called the prior probability, while P(H|E) is called the posterior probability. The factor that relates the two, P(H|E) / P(E), is called the likelihood ratio. Using these terms, Bayes' theorem can be rephrased as:**"The posterior probability equals the prior probability times the likelihood ratio."**

Advantages of Naive Bayes

1. When assumption of independent predictors holds true, a Naive Bayes classifier performs better as compared to other models.

2. Naive Bayes requires a small amount of training data to estimate the test data. So, the training period is less.

3. Naive Bayes is also easy to implement.

Disadvantages of Naive Bayes

1. Main imitation of Naive Bayes is the assumption of independent predictors. Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it is almost impossible that we get a set of predictors which are completely independent.

2. If a categorical variable has a category in test data set, which was not observed in the training data set, then the model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency.

To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.

1.6.2 Hoeffding Tree

Hoeffding trees were first proposed by Hulten et al (2001). One of the basic algorithms for stream data classification is the Hoeffding tree algorithm. The below fig 1.3-describes Hoeffding tree structure. It is an incremental, anytime decision tree induction algorithm that is capable of learning from massive data streams assuming that the distribution generating examples does not change over time. It produces decision trees which are similar to traditional batch

learning methods. Hoeffding trees and decision trees are asymptotically related. HT algorithm is based on a simple idea that a small sample can be often sufficient to choose an optimal splitting attribute. The below fig 1.2-describes algorithmic flow.

The key point to be noted here is traditional batch learning methods and also generate decision trees based on splitting attributes. Mathematically, it is proved that HT algorithm uses Hoeffding bound. To understand the meaning of Hoeffding bound few assumptions are made. Suppose we make 'N' independent observations of a random variable 'r' with range 'R', where 'r' is an attribute selection measure. In case of Hoeffding trees 'r' is information gain and if we compute the mean value of r 'rmean' of this sample the Hoeffding bound states that the true mean of 'r' is at least 1–δ where δ is user specified and Performance analysis of Hoeffding trees in data streams .

Algorithm 1 Hoeffding tree induction algorithm.

1: Let HT be a tree with a single leaf (the root)
2: **for all** training examples **do**
3: Sort example into leaf l using HT
4: Update sufficient statistics in l
5: Increment n_l, the number of examples seen at l
6: **if** $n_l \bmod n_{min} = 0$ **and** examples seen at l not all of same class **then**
7: Compute $\overline{G}_l(X_i)$ for each attribute
8: Let X_a be attribute with highest $\overline{G}_l$
9: Let X_b be attribute with second-highest $\overline{G}_l$
10: Compute Hoeffding bound $\epsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n_l}}$
11: **if** $X_a \neq X_\emptyset$ **and** $(\overline{G}_l(X_a) - \overline{G}_l(X_b) > \epsilon$ **or** $\epsilon < \tau)$ **then**
12: Replace l with an internal node that splits on X_a
13: **for all** branches of the split **do**
14: Add a new leaf with initialized sufficient statistics
15: **end for**
16: **end if**
17: **end if**
18: **end for**

Fig1.2- Hoeffding Tree algorithm

Main advantages of HT algorithm are

- It is incremental in nature
- Achieves high accuracy using small sample
- Multiple scans on the same data are never performed.

Main disadvantage is that HT cannot handle concept drift because once the node is created it can never be changed. Wang et al. (2003) explained about handling concept drift using classifiers. The algorithm spends a great deal of time with attributes that have nearly identical splitting quality. In addition, the memory utilisation can be further optimised.

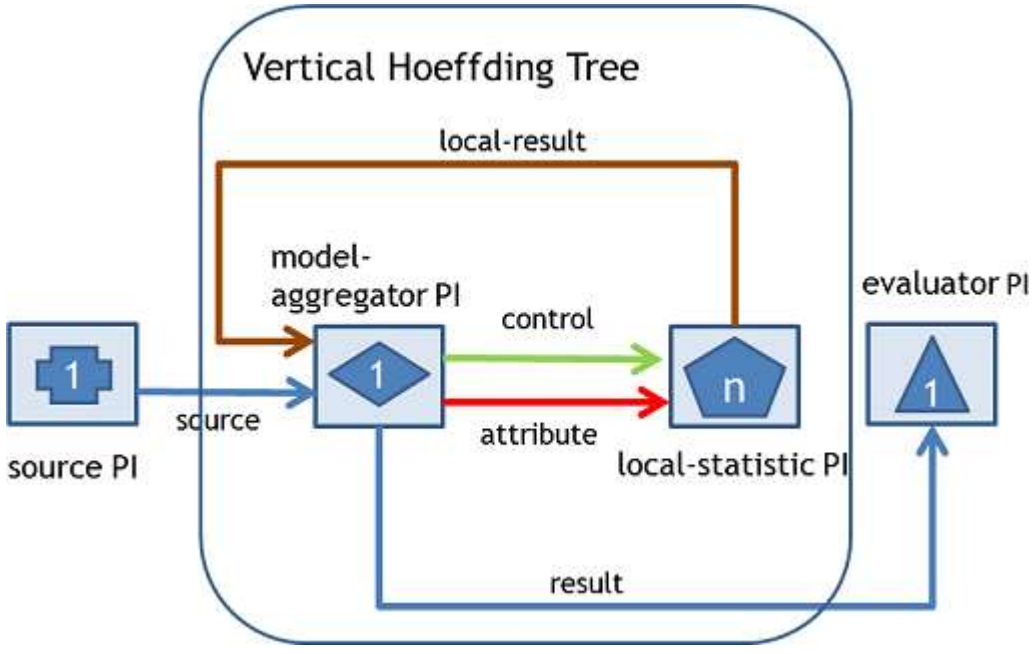

Fig no : 1.3-Vertical Hoeffding tree

1.6.3 Decision Tables

A decision table is a table that indicates conditions and actions in a simplified and orderly manner. By presenting logical alternative courses of action under various operating conditions, a decision table enables an individual to think through a problem and present its solution in compact notation.Decision tables are used to model complicated logic. They can make it easy to see that all possible combinations of conditions have been considered and when conditions are not met, it is easy to see.

A decision table can also be described as a cause-effect table and is the best way to deal with combination inputs with their associated outputs. It is an excellent tool to use in both testing and requirements management. Using decision tables,it becomes easier for the requirements specialist to write requirements which cover all conditions. In a decision table, the logic is well divided into conditions, actions (decisions) and rules for representing the various components that form the logical model.

The general format of a decision table has four basic parts. They include:

Action entry: It indicates the actions to be taken.

Condition entry: It indicates conditions which are being met or answers the questions in the condition stub.

Action stub: It lists statements that describe all actions that can be taken.

Condition stub: it lists all conditions to be tested for factors necessary for making a decision.

Need To Know About Decision Table,

- A decision table is a table of rows and columns separated into four quadrants.
- In a decision table, the inputs are listed in a column, with the outputs in the same column but below the inputs.

- Using decision tables make it possible to detect combinations of conditions that would otherwise not have been found and therefore not tested or developed.
- Decision tables should best be constructed during system design then they become useful to developers, testers and end-users.
- Decision table testing is a black box test design technique to determine the test scenarios for complex business logic.

There are two types of decision tables, that is extended entry table and limited entry table. In the extended entry table both the entry and stub section of any specific condition must be considered together if a condition is applicable to a given rule. In limited entry tables the conditions or actions required are contained within the appropriate stubs. The below fig 1.4-describes Decision Tables.

To build decision tables, the analyst needs to determine the maximum size of the table; eliminate any impossible situations, inconsistencies or redundancies and simplify the table as much as possible.Decision tables can be and often embedded within computer programs and use to drive the logic of the program. A simple example might be a lookup table containing a range of possible input values and a function pointer to the section of code to process that input.

Decision table shows conditions and actions in a simplified and orderly manner. By presenting logical alternative courses of action under various operating conditions, a decision table enables an individual to think through a problem and present its solution in compact notation.

Advantages Of Decision Table

- When the conditions are many then the decision table helps to visualize the outcomes of a situation.
- They are simple to understand and everyone can use this method to design the test scenarios and test cases.
- They are easy to draw.
- They provide more compact documentation.
- Decision tables can be changed easily according to the situation.

- Decision tables summarize all the outcomes of a situation and suggest suitable actions.
- Decision tables have a standard format.

Disadvantages Of Decision Tables

- Decision tables cannot express the complete sequence of operations to solve a problem; it may be difficult for a programmer to translate a decision table directly into a computer program.
- Decision tables do not show the flow of logic for the solution to a given problem.
- When there are many alternatives, decision table cannot list them all.
- Decision tables only present a partial solution.

		Rule 1	*Rule 2*	*Rule 3*	*Rule 4*
IF	Condition 1	Y	Y	N	N
AND	Condition 2	Y	N	Y	Y
AND	Condition 3	-	N	Y	-
AND	Condition 4	-	-	Y	N
THEN	Action 1	X		X	
AND	Action 2	X			X
AND	Action 3		X		
AND	Action 4		X	X	

Fig 1.4-Decision Table

1.6.4 Sequential Minimal Optimization (SMO)

Sequential minimal optimization (SMO) is an algorithm for solving the quadratic programming (QP) problem that arises during the training of support-vector machines (SVM). It was invented by John Platt in 1998 at Microsoft Research. SMO is widely used for training support vector machines and is

implemented by the popular LIBSVM tool. The publication of the SMO algorithm in 1998 has generated a lot of excitement in the SVM community, as previously available methods for SVM training were much more complex and required expensive third-party QP solvers. The below fig 1.5 describes SMO formula.

Consider a binary classification problem with a dataset (x1, y1), ..., (xn, yn), where xi is an input vector and yi ∈ {-1, +1} is a binary label corresponding to it. A soft-margin support vector machine is trained by solving a quadratic programming problem, which is expressed in the dual form as follows:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K(x_i, x_j) \alpha_i \alpha_j,$$

subject to:

$$0 \leq \alpha_i \leq C, \quad \text{for } i = 1, 2, \ldots, n,$$

$$\sum_{i=1}^{n} y_i \alpha_i = 0$$

Fig 1.5-SMO Formula

where C is an SVM hyper parameter and K(xi, xj) is the kernel function, both supplied by the user.

Algorithm 1: Sequential Minimal Optimization Algorithm

Input: training data x_i, labels y_i,
Output: sum of weight vector, α array, b and SV
1: Initialize: $\alpha_i = 0$, $f_i = -y_i$
2: Compute: $b_{high}, I_{high}, b_{low}, I_{low}$
3: Update α_{Ihigh} and α_{Ilow}
4: repeat
5: Update f_i
6: Compute: $b_{high}, I_{high}, b_{low}, I_{low}$
7: Update α_{Ihigh} and α_{Ilow}
8: until $b_{low} \leq b_{up} + 2\tau$
9: Update the threshold b
10: Store the new $\alpha 1$ and $\alpha 2$ values
11: Update weight vector w if SVM is linear

Fig 1.6-SMO Algorithm

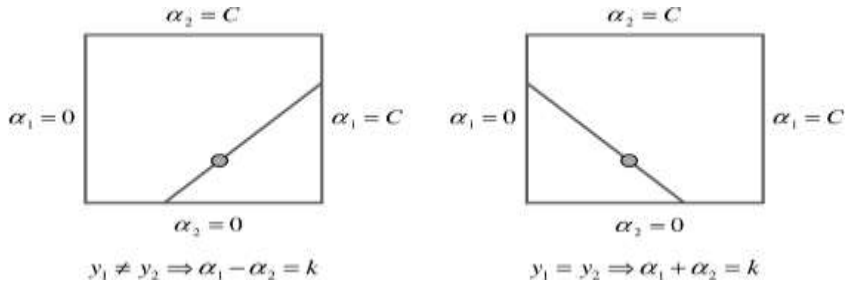

Fig 1.7- SMO margin support

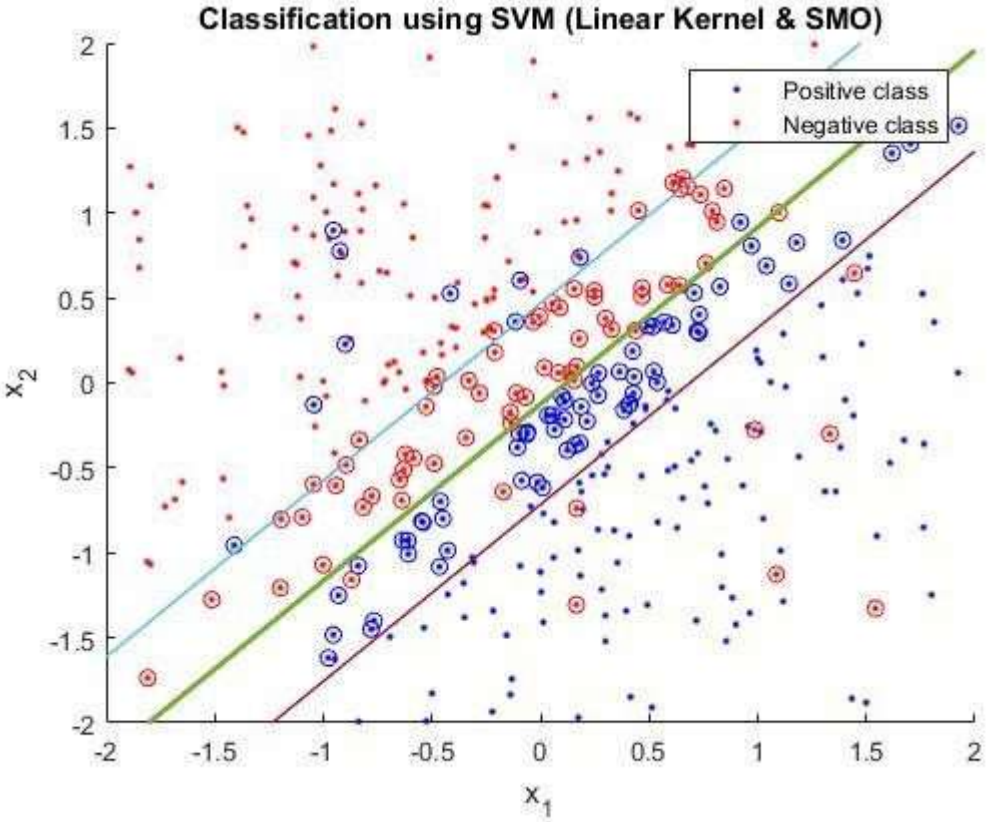

Fig 1.8- SMO Linear Kernal

Benefits and Drawbacks of QP Problems

The SVM training process is slow especially for big data. One of the reasons for being slow is that it requires a solution of an extremely large QP optimization problem. When considering N data points and N is a very large number we have a true big data problem. This is because the QP running time is O(N3) in the worst case. Although this problem cannot directly be solved SMO is one approach of dealing with large quantities of datasets. It requires an amount of memory that is linear in the training set size N.

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 diagnosis of heart disease such as Decision Tree, Naive Bayes, neural network, kernel density, automatically defined groups, bagging algorithm and support vector machine showing different levels of accuracy.

Chang C L and C.- H. Chen et al. in [2] on multiple databases of patients from around the world. One of the bases on which the papers differ is the selection of parameters on which the methods have been used. Many authors have specified different parameters and databases for testing the accuracy.

Michael W.Berry et al. in [3] have 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

Prerana T H M1,SivaPrakash N C2 et al. in [1] implemented several data mining techniques for diagnosis of heart disease such as Decision Tree, Naive Bayes, neural network, kernel density, automatically defined groups, bagging algorithm and support vector machine showing different levels of accuracies

Shilaskar et al. in [4] 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 J4.8 Decision Tree in the diagnosis of heart disease.

CHAPTER 3

SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

Existing systems use the available clinical data for prediction purposes and even if they do, they are restricted by the large number of association rules that apply. Diagnosis of the condition solely depends upon the doctor's intuition and patient's records. Detection is not possible at an earlier stage. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

3.2 PROPOSED SYSTEM

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. It will be also used for better health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital death

3.3 SYSTEM REQUIREMENTS

3.3.1 Hardware Requirements

Operating System: Windows, Linux

Processor: i3 processor

RAM Configuration: 4GB

3.3.2 Software Requirements

- WEKA TOOL

WEKA TOOL

We use WEKA(Waikato Environment for Knowledge Analysis),an open source data mining tool for our experiment. WEKA is developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art tool for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. . The below fig 3.1-describes WEKA GUI.

Fig 3.1-Weka GUI

CHAPTER 4

SYSTEM DESIGN

4.1 ARCHITECTURE DESIGN

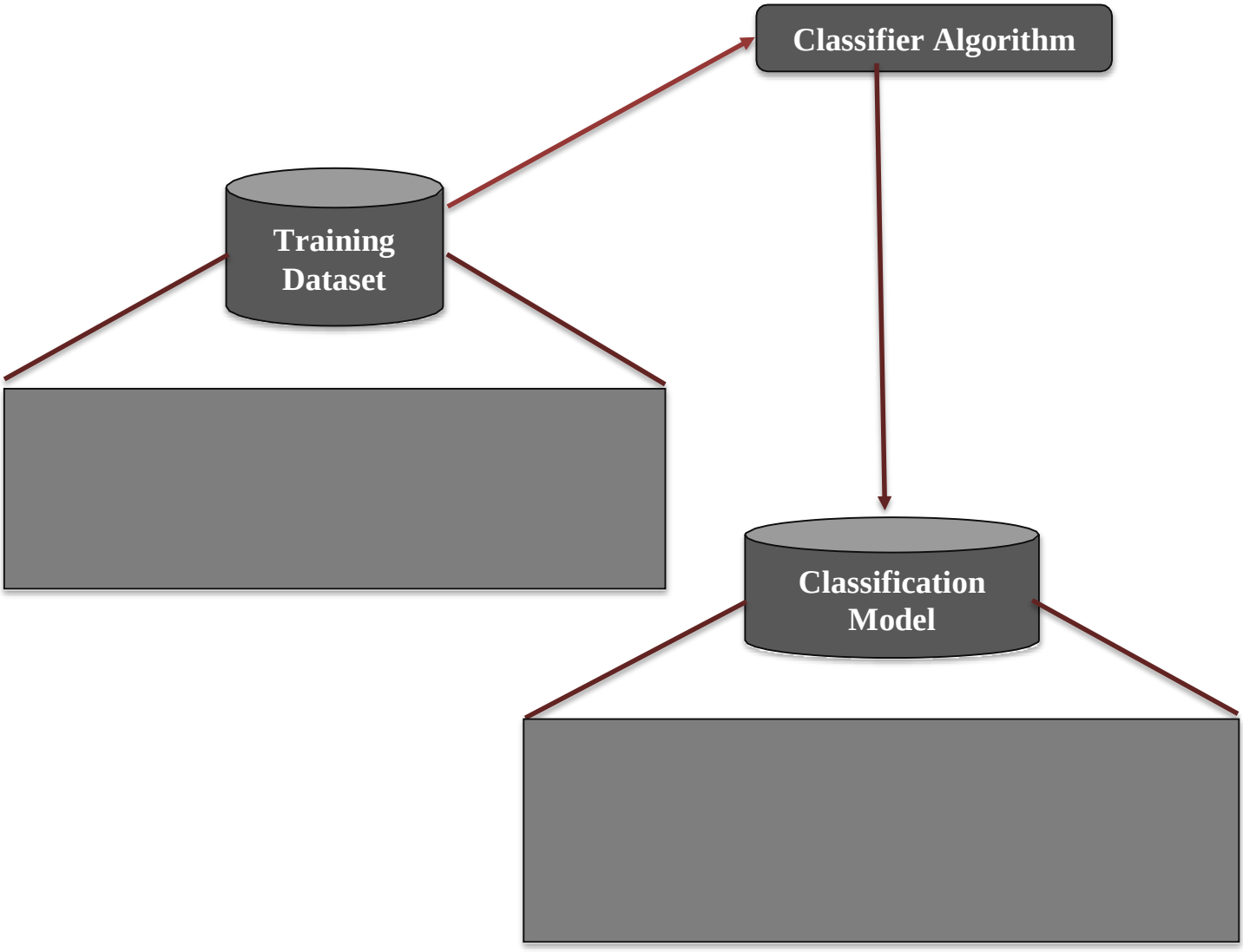

Fig 4.1-TRAINING PHASES

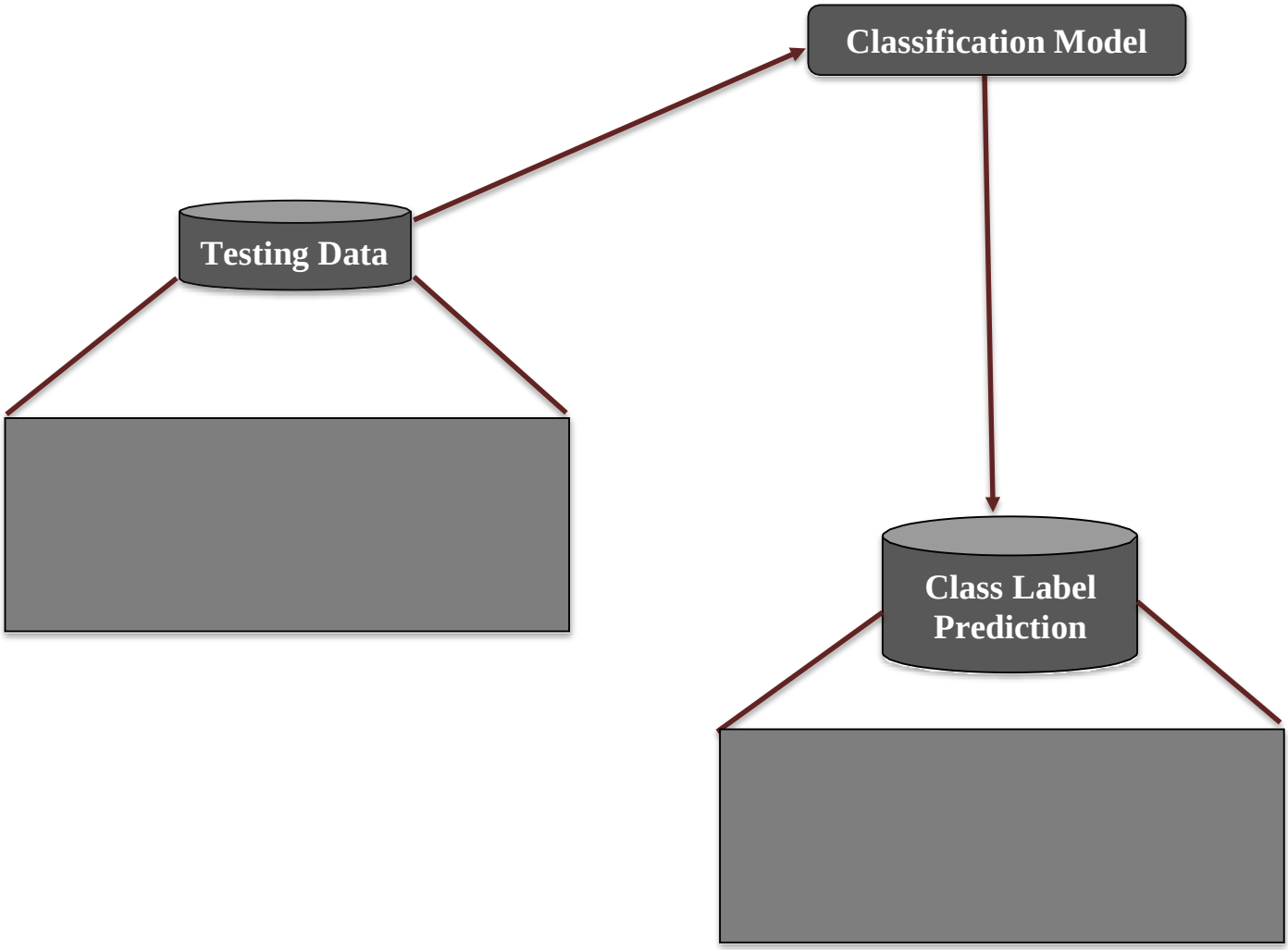

Fig 4.2-TESTING PHASES

CHAPTER 5

SYSTEM IMPLEMENTATION

5.1 LIST OF MODULES

- Training phase.
- Testing Phase.
- Prediction

5.2 MODULE DESCRIPTION

5.2.1 Training Phase

In training Phase, training data such as various attributes are analysed by a using classification algorithms and then the learned model or classifier is represented in the form of classifier rules.

5.2.2 Testing Phase

Test data are used to estimate the accuracy of the classification rule. If the accuracy is considered acceptable, the rule can be applied to the classification of new data tuples. Because the class label of each training tuple is provided, this step is also known supervised learning.

5.2.3 Prediction

Once the testing and training phase has done by classification models (Decision Table, Naive Bayes, SMO , Hoeffding Tree). The high accuracy of class label prediction is achieved by Decision table algorithm.

5.3 Supervised Learning

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

Here, we have used classification type of supervised Learning

Classification: A classification problem is when the output variable is a category, such as “disease present” or “disease absent”.

CHAPTER 6

RESULTS AND DISCUSSION

In this study, the accuracy of three data mining techniques is compared. The goal is to have high accuracy, besides high precision and recall metrics. Although these metrics are used more often in the field of information retrieval, here we have considered them as they are related to the other existing metrics such as specificity and sensitivity. These metrics can be derived from the confusion matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics.

6.1 CONFUSION MATRIX

a = Disease presence b = Disease absence

6.1.1 Confusion Matrix for Decision Table

	a	b
a	133	16
b	24	96

Table 6.1-Confusion Matrix for Decision Table

6.1.2 Confusion Matrix for Hoeffding Tree

	a	b
a	129	21
b	24	96

Table 6.2-Confusion Matrix for Hoeffding Tree

6.1.3 Confusion Matrix for Naïve Bayes

	a	b
a	131	19
b	25	95

Table 6.3-Confusion Matrix for Naïve Bayes

6.1.4 Confusion Matrix for Sequential Minimal Optimization

	a	b
a	131	19
b	24	96

Table 6.4-Confusion Matrix for Sequential Minimal Optimization

6.2 Performance of Different Data Mining Techniques with Accuracy and number of attributes incorporated

	ALGORITHMS	Execution time	kappa statistic	Mean absolute error	Root mean squared error	root relative squared error	correctly classified instances	Incorrectly classified instances
1	Naive Bayes	0.03s	0.6683	0.1835	0.3598	72.4003%	83.7037%	16.2963%
2	Hoeffding tree	0.11s	0.6617	0.178	0.3662	73.6994%	83.333%	16.6667%
3	Decision Tables	0.27s	0.6907	0.2612	0.3588	72.2019	84.8148%	15.1852%
4	Sequential minimal organization	0.22s	0.6762	0.1593	0.3991	80.3119	84.0741%	15.9259%

Table 6.5-Performance of Different Data Mining Techniques

6.2.1 FORMULAE

KAPPA STATISTIC

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

MEAN ABSOLUTE ERROR

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |x_i - x|$$

ROOT MEAN SQUARED ERROR:

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

RELATIVE ABSOLUTE ERROR:

$$U_1 = \frac{\left[\sum_{i=1}^{n}(P_i - A_i)^2\right]^{1/2}}{\left[\sum_{i=1}^{n}A_i^2\right]^{1/2}}$$

ROOT RELATIVE SQUARED ERROR

$$RSE = \frac{\sum_{i=1}^{n}(p_i - a_i)^2}{\sum_{i=1}^{n}(\bar{a} - a_i)^2}$$

6.3 Attribute Information:

S.NO	Attribute Name	Attribute Type
1	age	Continuous
2	sex	Categorical
3	chest pain type	Categorical
4	resting blood pressure	Continuous
5	serum cholesterol	Continuous
6	fasting blood sugar	Continuous
7	resting electrocardiographic results	Categorical
8	maximum heart rate achieved	Continuous
9	exercise induced angina	Categorical
10	oldpeak	Categorical
11	the slope of the peak exercise	Categorical
12	number of major vessels	Categorical
13	thal	Categorical

Table 6.6 Attribute Information

6.4. Attributes Types

Data Type	Values
Real	1,4,5,8,10,12
Ordered	11
Binary	2,6,9
Nominal	7,3,13

Table 6.7 Attributes Types

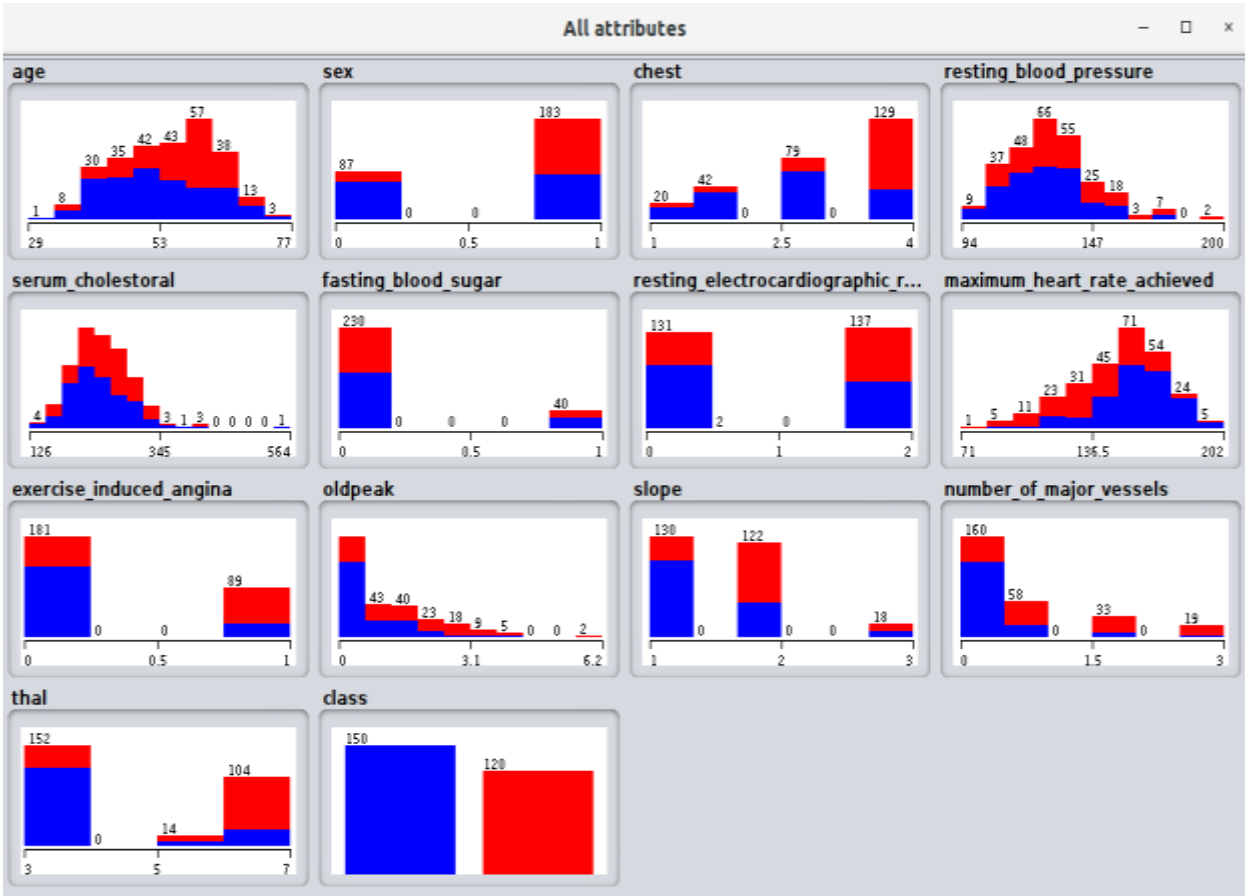

Fig 6.1 Visualizing Attributes

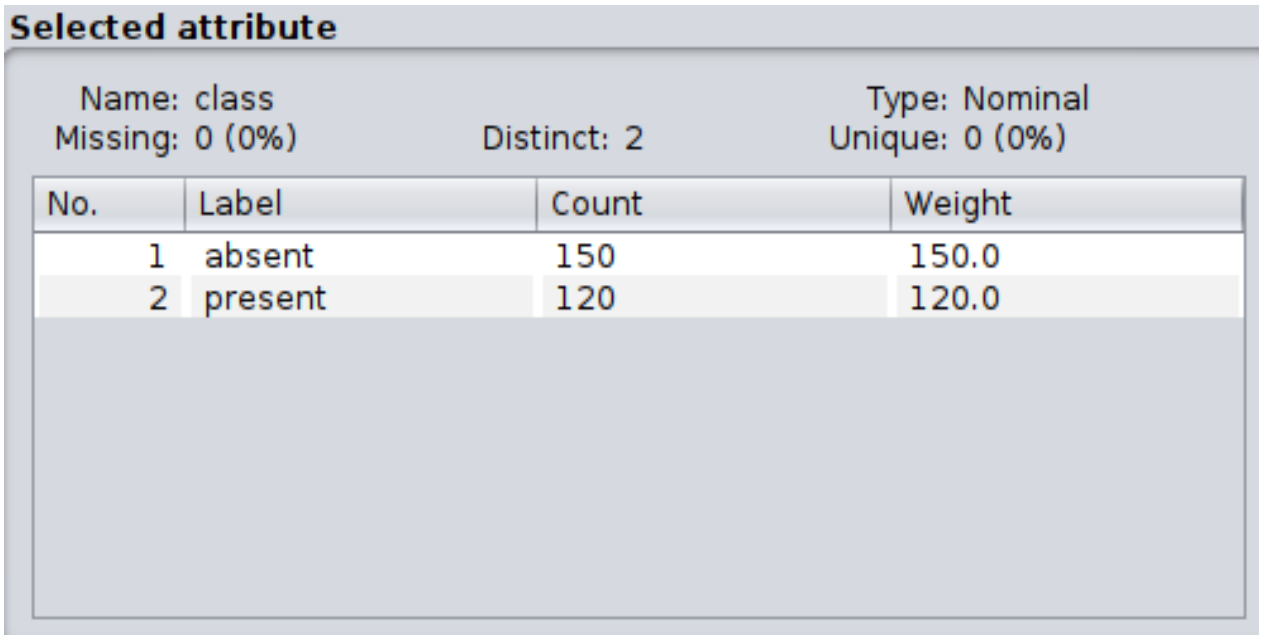

Fig 6.2 Class Attributes

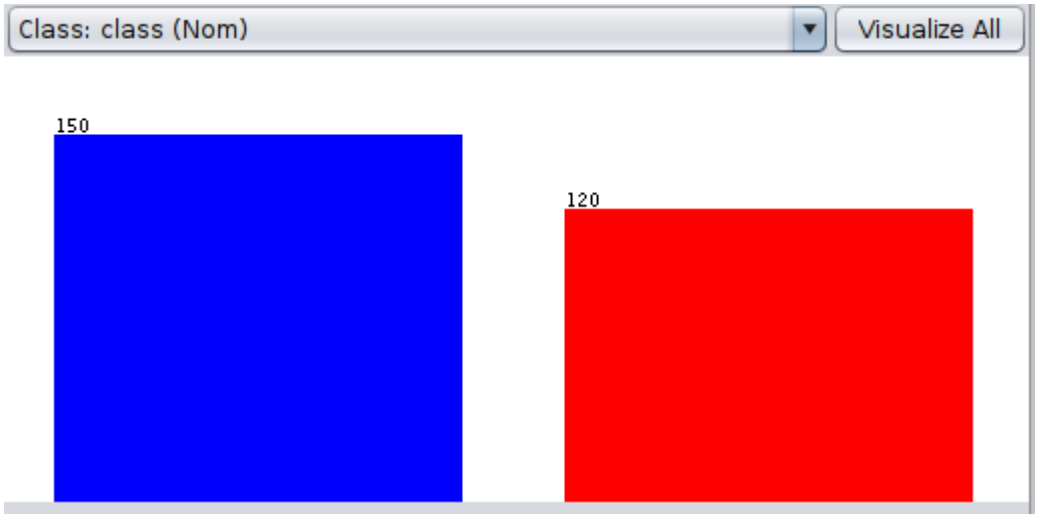

Fig 6.3 Class Visualization

6.5 PICTORIAL REPRESENTATION OF EACH ATTRIBUTES FOR CHOSEN ALGORITHMS

The kappa statistic is frequently used to test interrater reliability. The importance of rater reliability lies in the fact that it represents the extent to which the data collected in the study are correct representations of the variables measured. Measurement of the extent to which data collectors (raters) assign the same score

to the same variable is called interrater reliability. The Fig 6.4 shows kappa statistics chart.

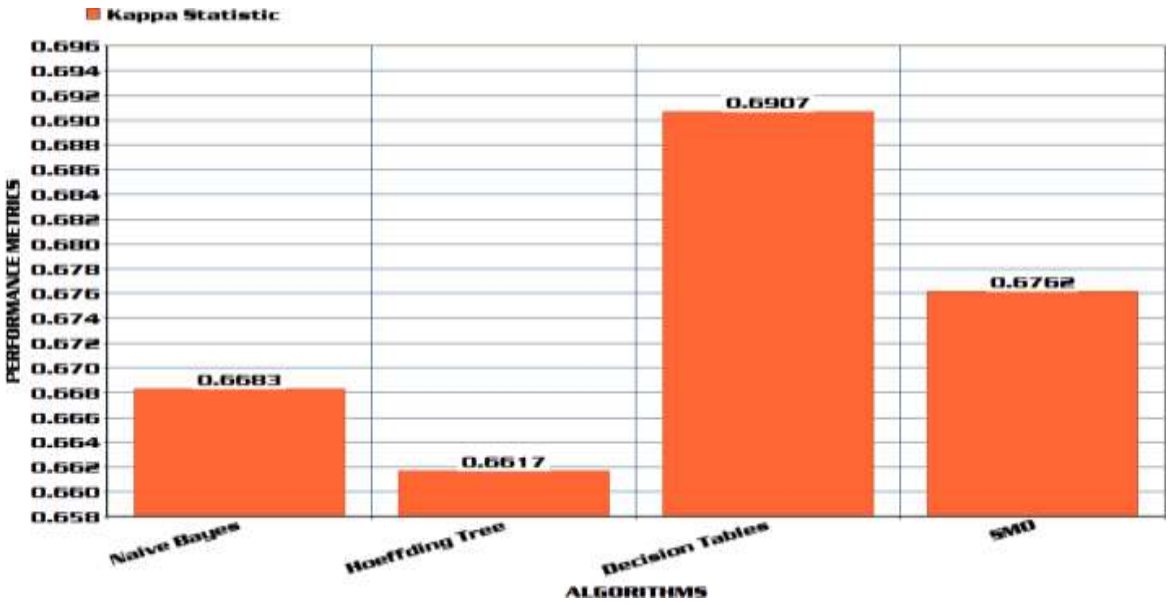

Fig 6.4 Kappa Statistic

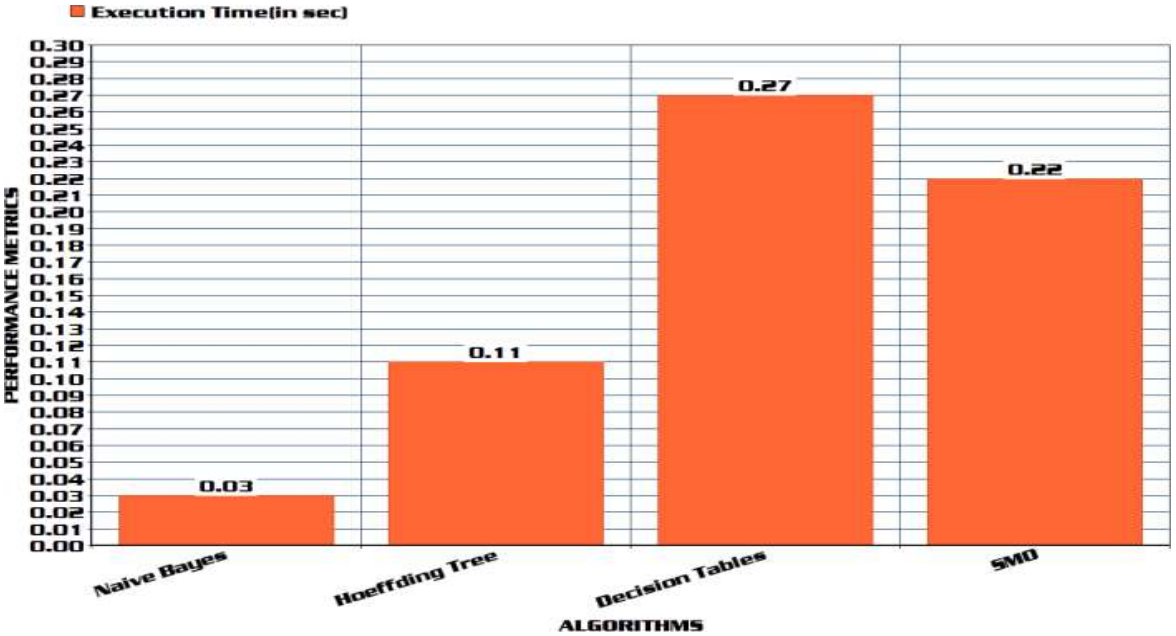

Fig 6.5 Execution Time

Root Mean Square Error (RMSE) measures how much error there is between two data sets. In other words, it compares a predicted value and an observed or known value. The Fig 6.6 shows Root Mean Squared Error chart.

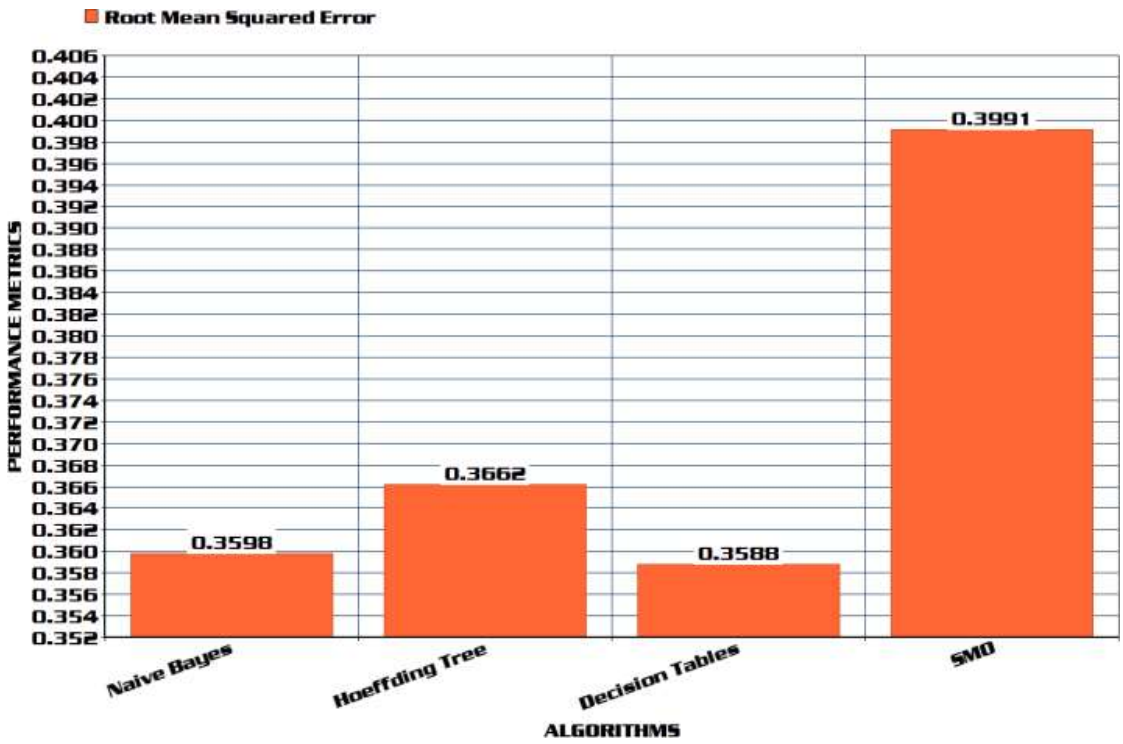

Fig 6.6 Root Mean Squared Error

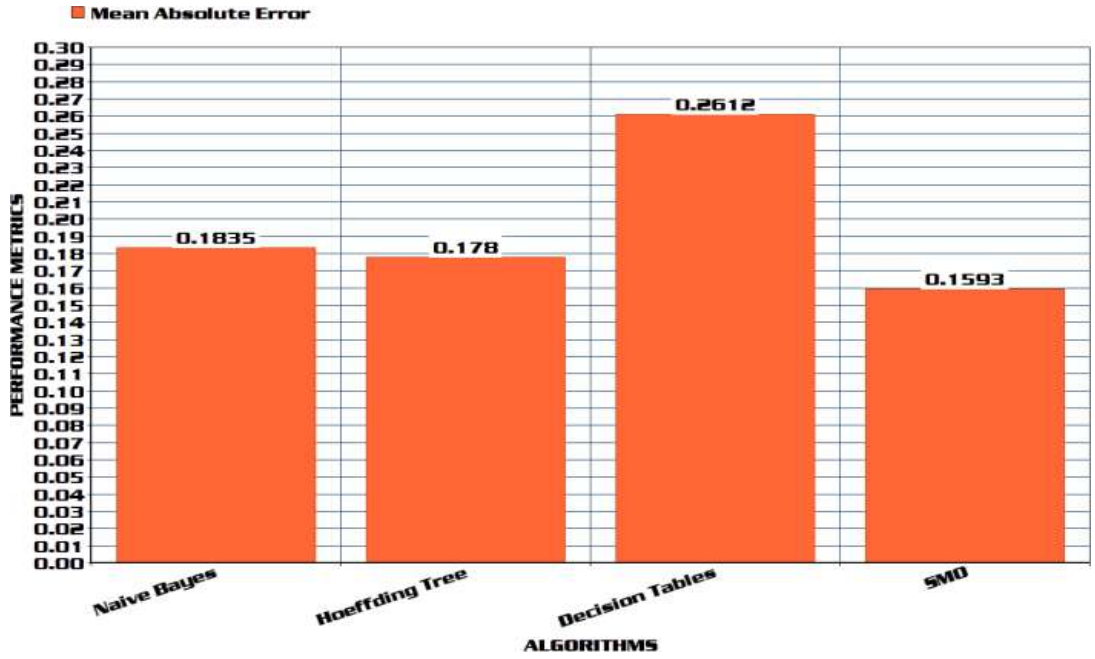

Fig 6.7 Mean Absolute Error

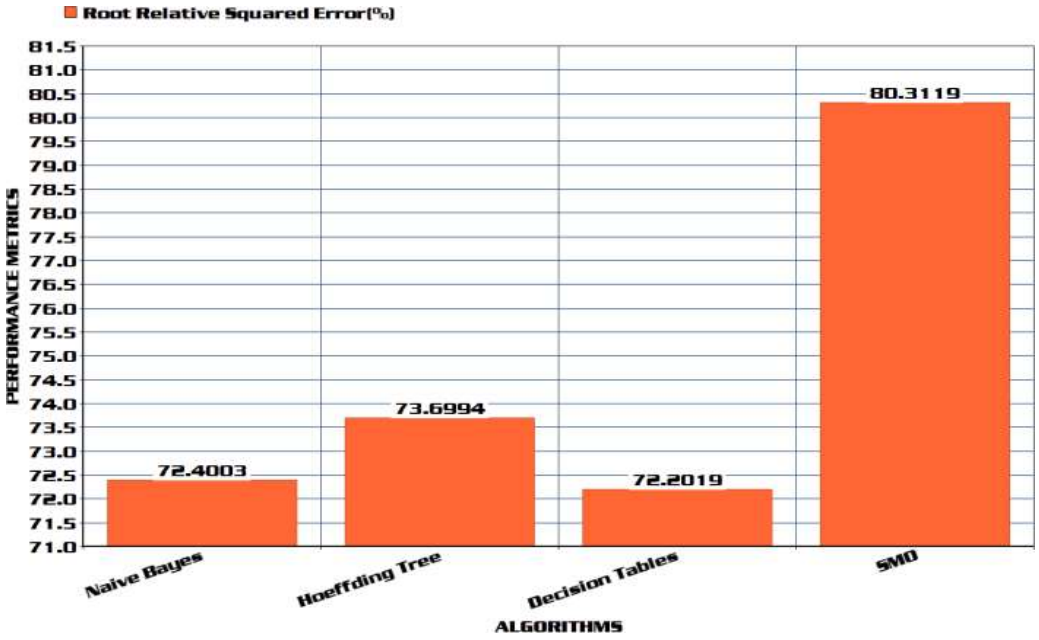

Fig 6.8 Root Relative Squared Error

The labels on the test set are supposed to be the actual correct classification. Performance is computed by asking the classifier to give its best guess about the classification for each instance in the test set. Then the predicted classifications are compared to the actual classifications to determine accuracy. The Fig 6.9 shows Correctly Classified Instances chart.

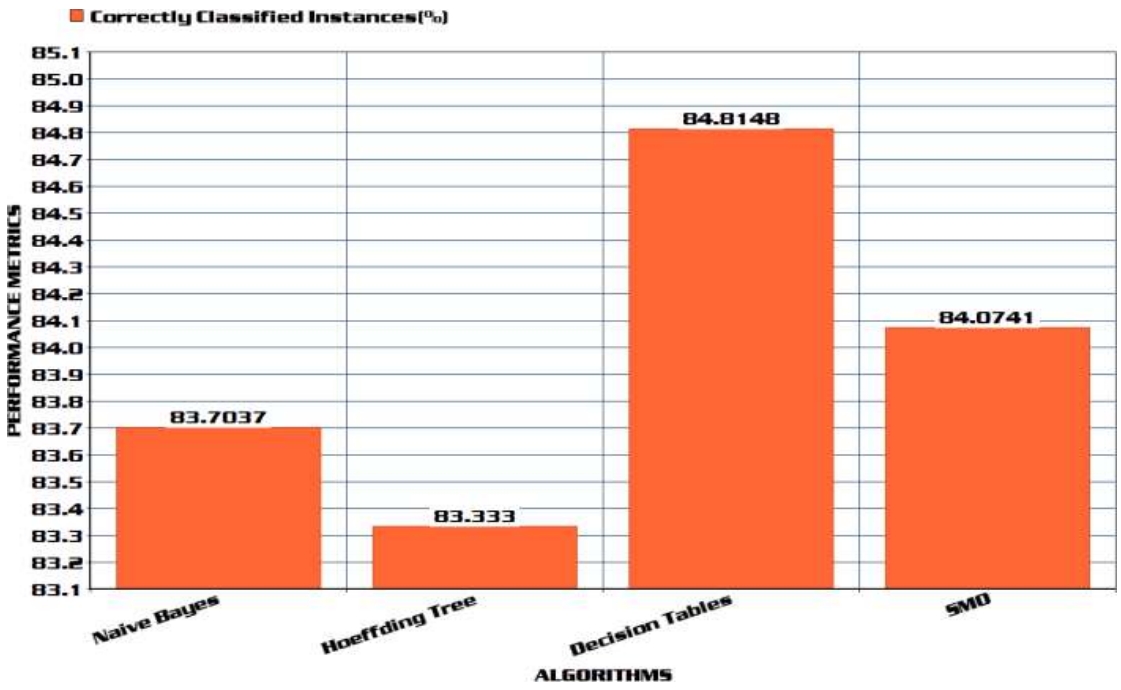

Fig 6.9 Correctly Classified Instances

The labels on the test set are supposed to be the actual incorrect classification. Performance is computed by asking the classifier to give its worst guess about the classification for each instance in the test set. Then the predicted classifications are compared to the actual classifications to determine accuracy. The Fig 6.9 shows Incorrectly Classified Instances chart.

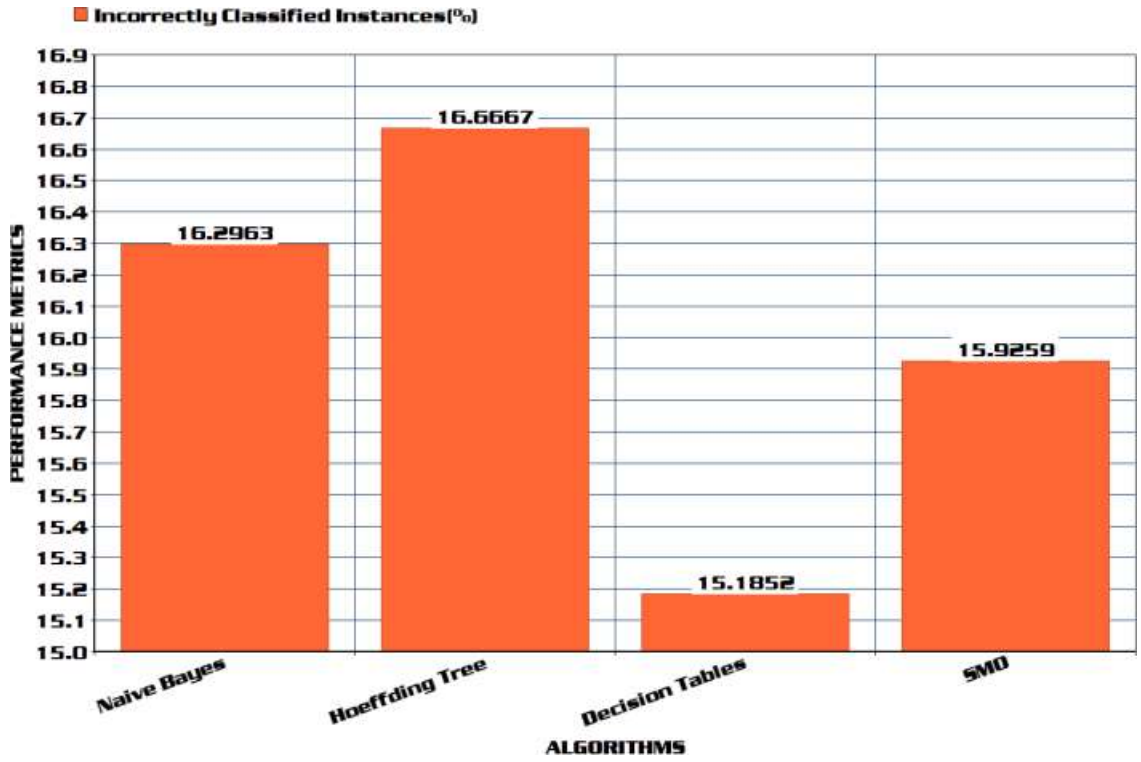

Fig 6.10 Incorrectly Classified Instances

6.6 ROC CURVE

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection[6] in machine learning. The false-positive rate is also known as probability of false alarm and can be calculated as (1 − specificity). It can also be thought of as a plot of the power as a function of the Type I Error of the decision rule (when the performance is calculated from just a sample of the population, it can be thought of as estimators of these quantities).

The ROC curve is thus the sensitivity as a function of fall-out. In general, if the probability distributions for both detection and false alarm are known, the ROC curve can be generated by plotting the cumulative distribution function to the discrimination threshold) of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability on the x-axis.

ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

6.6.1 ROC Curve

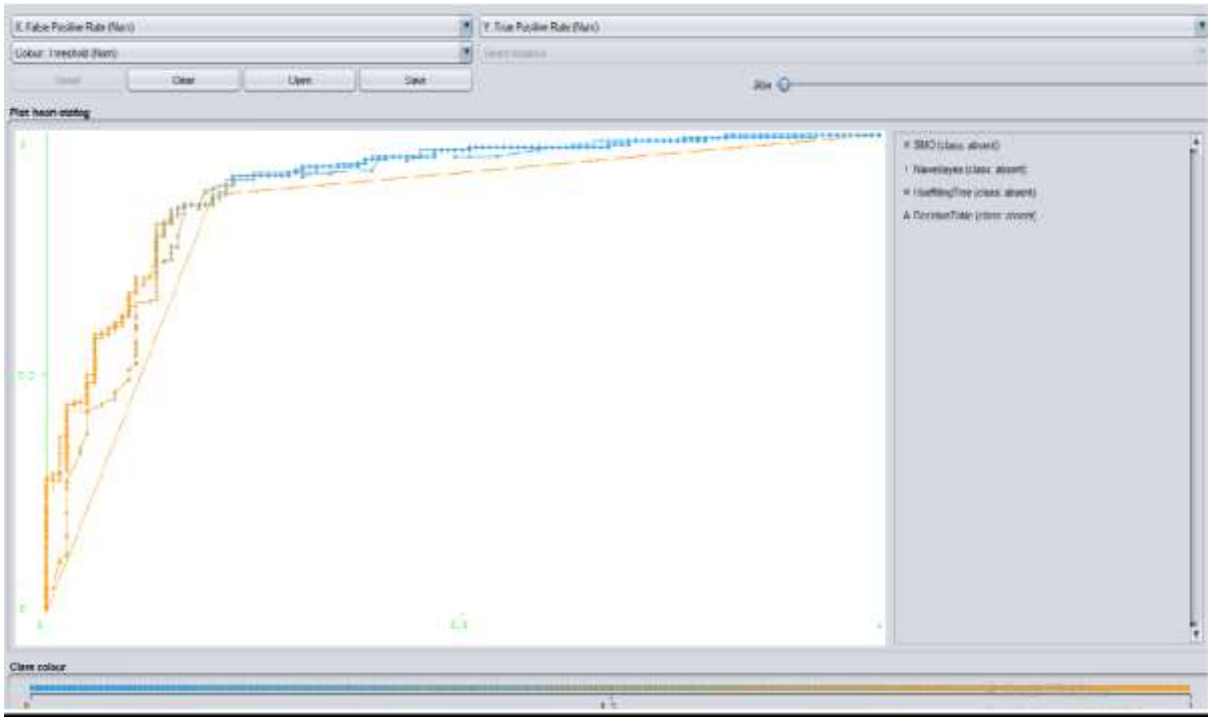

Fig 6.11 ROC Curve

6.6.2 **Knowledge Flow:**

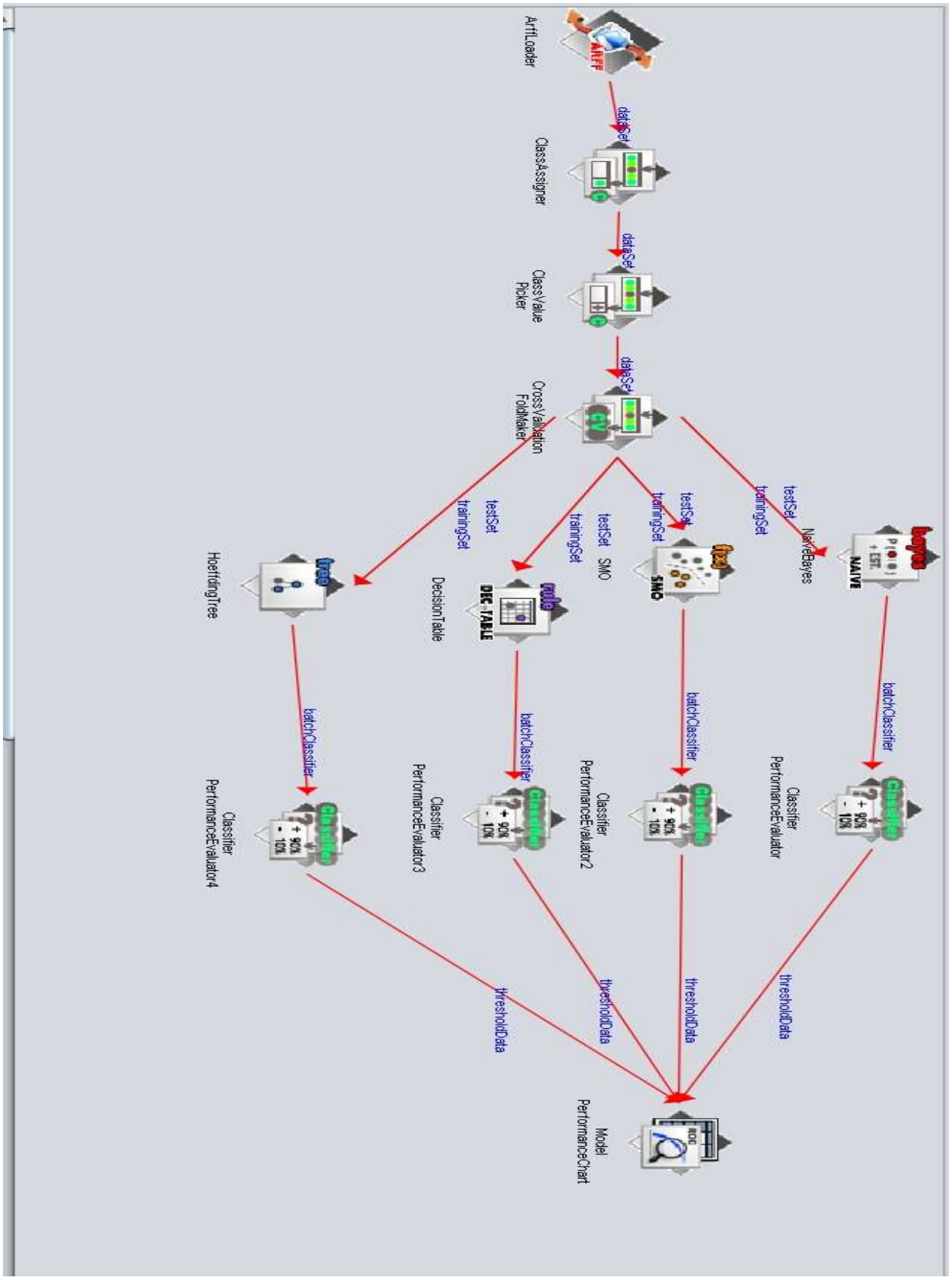

Fig 6.12 Knowledge Flow

CHAPTER 7

CONCLUSION

Finally, we carried out an experiment to find the predictive performance of different classifiers. We select four popular classifiers considering their qualitative performance for the experiment. We also choose one dataset from the heart disease available at kaggle machine learning repository. Decision Table is the best in performance. In order to compare the classification performance of four machine learning algorithms, classifiers are applied on the same data and results are compared on the basis of misclassification and correct classification rate and according to experimental results.

By analyzing the experimental results, it is concluded that the Decision Table technique turned out to be the best classifier for heart disease prediction because it contains more accuracy as it classified 229 correctly classified instances with least incorrectly classified instances over Sequential Minimal optimization, Hoeffding Tree and Naïve Bayes.

7.1 FUTURE SCOPE

The performance of the health's diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In general, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

CHAPTER 8
REFERENCES

[A]. C.-L. Chang and C.-H. Chen, "Applying decision trees and neural networks to increase quality of dermatologic diagnosis," Expert Syst. Appl., vol. 36, no. 2, Part 2, pp. 4035–4041,Mar. 2009.

[B]. Michael W.Berry et.al,"Lecture notes in data mining",World Scientific(2006)

[C]. Prerana T H M1, Shivaprakash N C2 , Swetha N3 "Prediction of Heart Disease Using Machine Learning Algorithms- NaïveBayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS" International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP: 90-99 ©IJSE Available at www.ijse.orgISSN: 2347-2200

[D]. Shilaskar S and A. Ghatol, "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases," Expert Syst. Appl., vol. 40, no. 10, pp. 4146–4153, Aug.2013.