**Prediction of Heart Disease Using Machine**

**Learning Algorithms and Data Mining Techniques**

Ninad S Shetty (Sec ‘A’ - 140905103) Prajwal P (Sec ‘A’ - 140905123)

6th Semester, CSE, Manipal Institute of Technology

14TH APRIL, 2017

# INTRODUCTION

Heart disease is the leading cause of death in the world over the past 10 years (World Health Organization 2007). The European Public Health Alliance reported that heart attacks, strokes and other circulatory diseases account for 41% of all deaths (European Public Health Alliance 2010). Several different symptoms are associated with heart disease, which makes it difficult to diagnose it quicker and better. Working on heart disease patients databases can be compared to real-life application. Doctors knowledge to assign the weight to each attribute. More weight is assigned to the attribute having high impact on disease prediction. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the Diagnosis process. It also provides healthcare professionals an extra source of knowledge for making decisions.

There are many scientific technologies which help doctors in taking clinical decisions but they might not be accurate. Heart disease prediction system can assist medical professionals in predicting state of heart, based on the clinical data of patients fed into the system. Doctors may sometimes fail to take accurate decisions while diagnosing the heart disease of a patient, therefore heart disease prediction systems which use machine learning algorithms assist in such cases to get accurate results. There are many tools available which use prediction algorithms but they have some flaws. Most of the tools cannot handle big data. The advancement of information technology, system integration as well as software development, techniques have shaped a innovative generation of multifaceted computer systems. Information technology researchers have been offered several challenges by these systems. An instance of such system is the healthcare system. Newly, there has been an enlarged awareness to make use of the advancement of data mining technologies in healthcare systems. Consequently, the objective of the present effort is to explore the aspects of making use of health data for the assistance of humans by means of new machine learning and data mining techniques. The thought is to recommend a computerized method for diagnosing heart diseases based on prior data and information.

Data mining is a discipline to realize knowledge from databases. The database contains a set of instances (records or case). Machine learning can be defined as a scientific field so as to plan and develop algorithms that let computers to enhance acquaintance of real time problem based on earlier statistics, and perform to resolve a real time problem beneath definite instructions and rules. At hand there are numerous presentations of machine learning; data mining is the largely used application of machine learning. Every illustration used by machine learning and data mining algorithms is formatted by means of same set of fields (features, attributes, inputs, or variables). When the instances contain the correct output (class label) then the learning process is called the supervised learning.

On the other hand, the process of machine learning without knowing the class label of instances is called unsupervised learning. Clustering is a common unsupervised learning method (some clustering models are for both). The objective of clustering is to describe data. On the other hand, classification and regression are predictive methods. In the present research, my focus is on supervised machine learning. This thesis proposes new methods intended for investigating feature selection techniques as well as develop new machine learning algorithms designed for providing automatic computer aided analysis and decision support system for heart disease diagnosis. The aim is to build up an integrated structure with a righteous workflow (constructing missing features values, feature selections, and classification algorithms). In requisites of features selection techniques, the research decided on features selection technique as a process to increase high superiority attributes to improve the mining process. In regards to analysis approach the present work projected a new means for diagnosis based on a combination of learning algorithm and feature selection technique. The thought is to get hold of a hybrid incorporated approach so as to merge the most excellent performing learning algorithms and the finest performing feature selection technique by means of an experimental estimate on a dataset.

# SCOPE

Algorithms play a significant role in prediction and data exploration. The healthcare industry collects huge amounts of healthcare data which is not feasible to handle manually. With the tremendously growing population, the doctors and experts available are not in proportion with the population. Also, symptoms of heart disease may not be significant and thus are often neglected. The main objective of this research project is predicting the heart disease risk level of a patient using machine learning algorithms - Nearest neighbor, decision tree, Naïve Bayes, logistic regression and comparing them to find the most optimal, accurate algorithm for the prediction of the heart disease.

Getting the correct result is very important in any algorithm. In the medical world a false analysis may cost a life. Hence the scope of this project is to find the most optimal algorithm, that is the one which gives correct result for most of the cases.

In this project we use a dataset that already has the results, which is called supervised learning. We apply Nearest neighbor, decision tree, Naïve Bayes, logistic regression to the dataset and find out which dataset gives the best results.

Comparative study of the machine learning algorithms is done and graphical representation of the results is provided for easier understanding. This project can be extended for prediction risk level of other similar diseases like Cancer, Brain tumor etc.

# SURVEY

References:

1. Heart Disease Prediction Using Machine learning and Data Mining Technique Jaymin Patel, Prof.TejalUpadhyay, Dr. Samir Patel Department of Computer Science and Engineering, Nirma University, Gujarat, India Principal, Grow More Faculty of Engineering, Ahmedabad, Gujarat, India
2. Development and Validation of a Logistic Regression-Derived Algorithm for Estimating the Incremental Probability of Coronary Artery Disease Before and After Exercise Testing ANTHONY P. MORISE, MD, FACC, ROBERT DETRANO, MD, PHD,\* MARCO BOBBIO, MD,t GEORGE A. DIAMOND, MD, FACC Morgantown, West Virginia; Torrance and Los Angeles, California; Turin, Italy
3. THE IMPLEMENTATION OF SUPPORT VECTOR MACHINES USING THE SEQUENTIAL MINIMAL OPTIMIZATION ALGORITHM by Ginny Mak Supervisors: Professor G. Ratzer, Professor H. Vangheluwe ,April 2000 School of Computer Science McGill University, Montreal, Canada
4. Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS Prerana T H M1 , Shivaprakash , Swetha , Samsung R&D Institute, Bangalore , Indian Institute of Science Bangalore, Computer Science and Engineering ,Vivekananda Institute of Technology, Bangalore
5. Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients G. Parthiban Research Scholar, Dr. MGR Educational Research and Institute, Maduravoyal,Chennai, India. S.K.Srivatsa Phd, Sr. Professor, Dept of E & I, St.Joseph’s College of Engineering, Chennai
6. for Machine Learning, Allison Chang.

# GAPS

# Support Vector machine using sequential minimal optimization could not be implemented as we could not understand it from the survey.

# OBJECTIVE

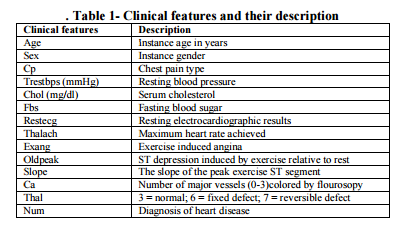
To extract hidden patterns and relationships from large data bases, Data mining merges statistical analysis, machine learning and database technology Diabetes is a chronic disease which causes serious health complications including heart disease, kidney failure and blindness. It is a major risk factor for cardiovascular disease (disease of the heart and circulatory system). Diabetes also increases the risk of micro-vascular damage and macrovascular complications. Thus diabetes is found to be one of the leading causes of global death by disease. Around 366 million people have diabetes world wide according to statistics taken in the year 2011. Also it has been projected that the people with diabetes will increase to around 552 million by the year 2030. Heart disease is a term for variety of disease that affecting the heart such as chest pain, shortness of breath, heart attack and other symptoms. It encompasses the diverse diseases that affect the heart. Chest pains arise when the blood received by the heart muscles is inadequate. Heart disease refers to numerous problems that distress the heart and the blood vessels in the heart. The term ‘cardiovascular disease’ that represents a category of heart disease comprises a broad variety of conditions that upset the heart and the blood vessels and the way in which blood is pumped and circulated in the body . Heart disease is the leading cause of death in the world over the past 10 years. The World Health Organization reported that heart disease is the first leading cause of death in high and low income countries . There are several methods in the literature individually to diagnosis diabetes or heart disease. There is no automated diagnosis method to diagnose Heart disease for diabetic patient based on diabetes diagnosis attributes to our knowledge. This research paper is related to our previous work, diagnosis of heart disease for diabetic patients using Naïve bayes method and Diagnosing Vulnerability of Diabetic Patients to Heart Diseases using Support Vector Machines to predict the heart disease for diabetic patients using diabetic diagnosis attributes.

# METHODOLOGY

# 

We apply supervised learning algorithms. The dataset already contains the results. We apply the algorithms and check with the results to find the performance of the algorithms.

1. Data sources In this paper, we use the heart disease data from machine learning repository of UCI [11]. We have total 303 instances of which 164 instances belonged to the healthy and 139 instances belonged to the heart disease. 14 clinical features have been recorded for each instance.
2. Features description Table 1 shows the 14 clinical features and their description.



We use the following algorithms :

**Decision tree**

The decision tree type used in this research is the gain ratio decision tree. The gain ratio decision tree is based on the entropy (information gain) approach, which selects the splitting attribute that minimizes the value of entropy, thus maximizing the information gain . Information gain is the difference between the original information content and the amount of information needed. The features are ranked by the information gains, and then the top ranked features are chosen as the potential attributes used in the classifier. To identify the splitting attribute of the decision tree, one must calculate the information gain for each attribute and then select the attribute that maximizes the information gain. The information gain for each attribute is calculated using the following formula



Where k is the number of classes of the target attributes Pi is the number of occurrences of class i divided by the total number of instances (i.e. the probability of i occurring). The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values . Gain Ratio adjusts the information gain for each attribute to allow for the breadth and uniformity of the attribute values.

Gain Ratio = Information Gain / Split Information Where the split information is a value based on the column sums of the frequency table.

**K-Nearest Neighbor**

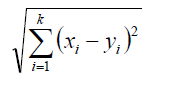
K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970’s as a non-parametric technique.

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

The best value of k is the square root of number of instances. In our dataset there are 303 instances hence the best value of k is 17.

We have used two methods to find the distance function namely Euclidean distance and Mahalanobis distance.

Euclidean distance is given by

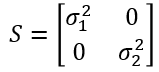


Mahalanobis distance is given by

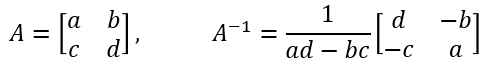


Mahalanobis distance equation reduces to the simple two dimensional example from early in the post when there is no correlation.

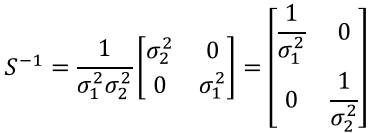
Assuming no correlation, our covariance matrix is:

[](https://chrisjmccormick.files.wordpress.com/2014/07/2d_covariance_nocorrelation.png)

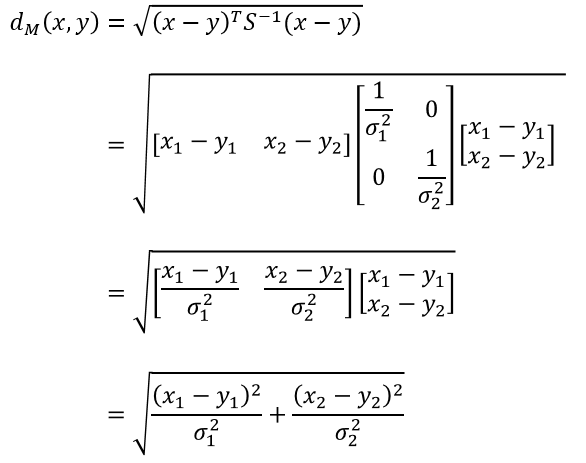
The inverse of a 2x2 matrix can be found using the following:

[](https://chrisjmccormick.files.wordpress.com/2014/07/2x2matrixinverse.png)

Applying this to get the inverse of the covariance matrix:

[](https://chrisjmccormick.files.wordpress.com/2014/07/covarianceinverse.png)

Now we can work through the Mahalanobis equation to see how we arrive at our earlier variance-normalized distance equation.

[](https://chrisjmccormick.files.wordpress.com/2014/07/mdist_wo_correlation.png)

The above equation is for 2 attributes. By modifying the above equation we can do it for 13 attributes.

**Naïve Bayes**

A conditional probability is the likelihood of some conclusion say C, given some evidence/observation, E, where a dependence relationship exists between C and E.

This probability is denoted as P(C |E) where



The naive Bayesian classifier, or simple Bayesian classifier, works as follows:

1. Let D be a training set of tuples and their associated class labels as Ca and Cp. As usual, each record is represented by an n-dimensional attribute vector, X=(x1, x2…,xn-1, xn), depicting n measurements made on the tuple from n attributes, i.e. A1 to An.

2. Suppose that there are m number of classes for prediction, C1, C2… Cm. Given a record, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple x belongs to the class Ci if and only if

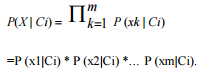
P (Ci|X)>P (Cj|X) for 1≤ j≤m and j ≠ i

Thus we maximize P(Ci|X). The class Ci for which P (Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes’ theorem



3. As P(X) is constant for all classes, only P (X|Ci)\* P (Ci) need be maximized. If the class prior probabilities are not known, then it is often assumed that the classes are equally likely, that is, P (C1) =P (C2) =…P(Cm-1)=P (Cm) and we would therefore maximize P (X|Ci). Otherwise, we maximize P (X|Ci) P (Ci). Note that the class prior probabilities may be estimated by P (Ci) =|Ci, D|/|D|, where |Ci, D| is the number of training tuples of class Ci in D.

4. Given data sets with many attributes, it would be extremely computationally expensive to compute P(X|Ci). To reduce computation in evaluating P(X|Ci), the naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple,Thus



We can easily estimate the probabilities P (x1|Ci), P (x2|C i)… P (xm|Ci) from the database training tuples. Recall that here xk refers to the value of attribute Ak for tuple X. For each attribute, we will see that whether the attribute is categorical or continuous-valued. For instance, to compute P (X|Ci), we consider the following:

(a) If Ak is categorical, then P (Xk|Ci) is the number of tuples of class Ci in D having the value xk for Ak, divided by |Ci, D|, the number of tuples of class Ci in D.

5. In order to predict the class label of X, P(X|Ci )P(Ci ) is evaluated for each class Ci. The classifier predicts that the class label of tuple X is the class Ci if and only if

P(X|Ci) P(Ci) > P(X|Cj) P(Cj) for 1 ≤ j ≤ m, j ≠ i

In other words, the predicted class label is the class Ci for which P (X|Ci) P (Ci) is the maximum.

For continuous attributes we have used two methods to convert them to categorical data : quantitative approach , k-means clustering

In quantitative approach the dataset is equally divided with respect to range and categorized.

In k-means clustering approach we categorize them using the following algorithm:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

**Logistic regression**

The Logistic Regression is a regression model in which the response variable (dependent variable) has categorical values such as True/False or 0/1. It actually measures the probability of a binary response as the value of response variable based on the mathematical equation relating it with the predictor variables.

As we are using R language for the analysis, we use function glm .The basic syntax for **glm()** function in logistic regression is −

glm(formula,data,family)

Following is the description of the parameters used −

* **formula** is the symbol presenting the relationship between the variables.
* **data** is the data set giving the values of these variables.
* **family** is R object to specify the details of the model. It's value is binomial for logistic regression.

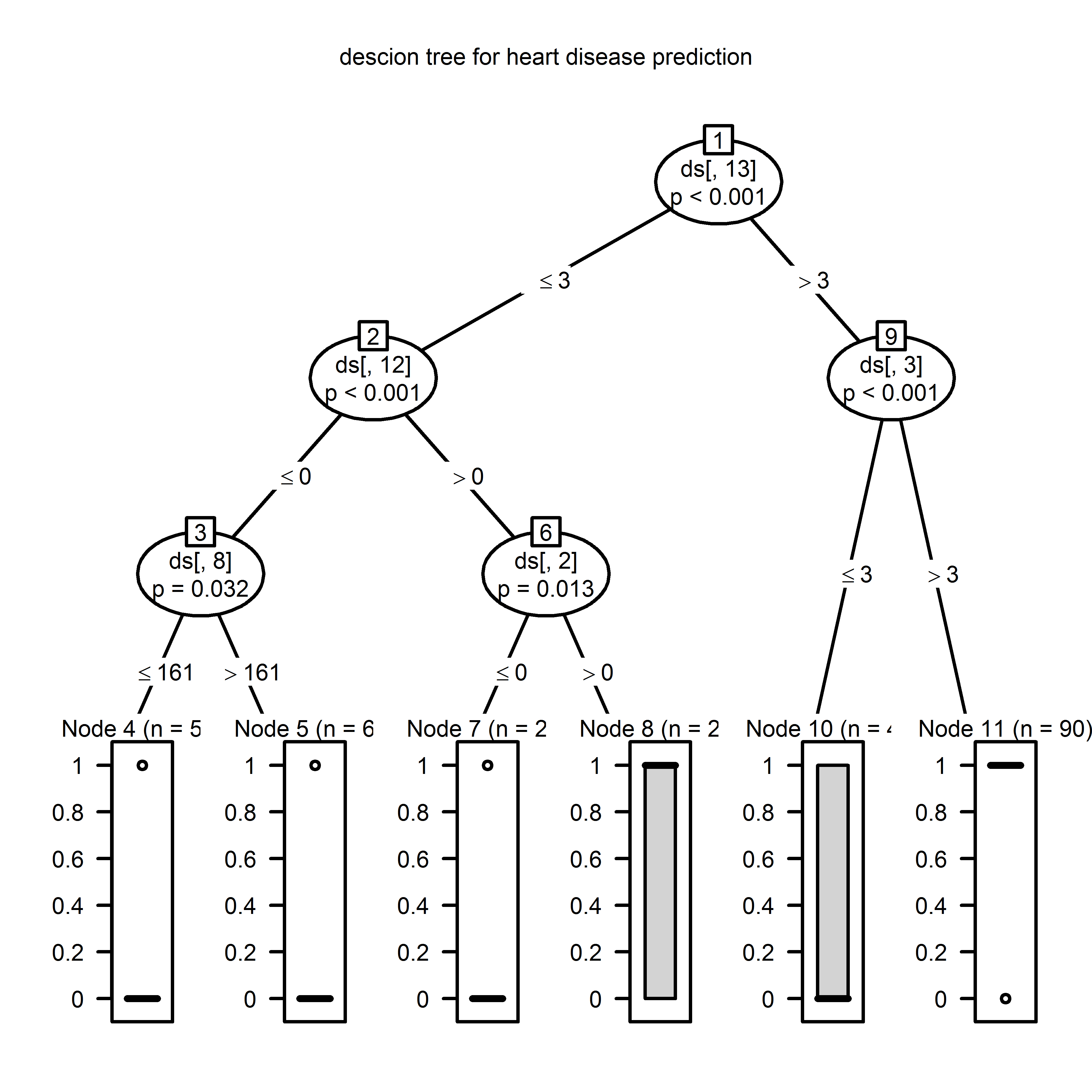
By using glm we find a equation and predict data for the training set.

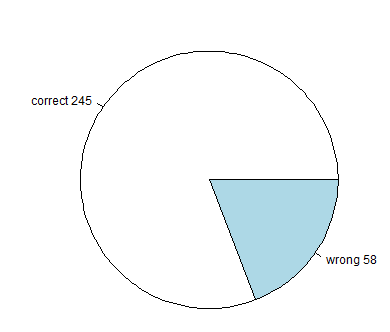
**RESULTS AND ANALYSIS**

The following are the results of the algorithms.

**Decision tree:**

The decision tree first was drawn .Later with the decision tree, decision rules are formed. The decision tree is shown below .with this tree decision rules are formed.

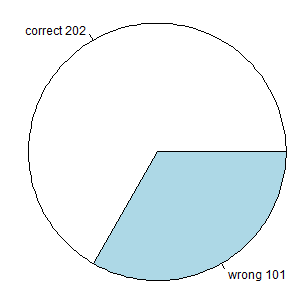




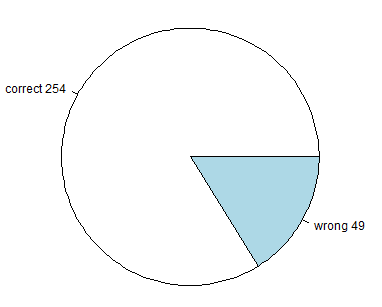
For 303 instances the efficiency is : 85.85809%

**K-nearest neighbor :**

With Euclidean Distance :



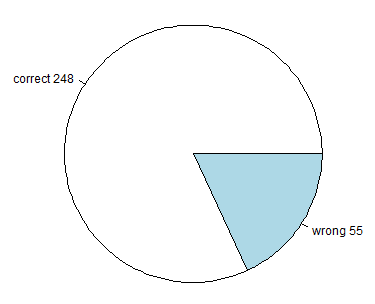
With Mahalanobis distance :



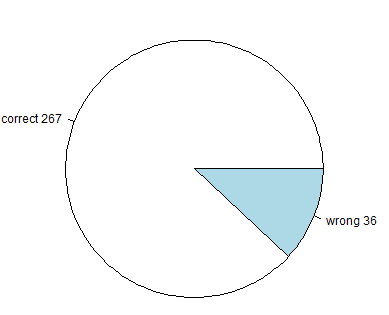
|  |  |
| --- | --- |
| **Efficiency of KNN** | **In percentage** |
| Euclidean distance | 66.66666 |
| Mahalanobis distance | 82.83828 |

**Naïve Bayes :**

With quantitative approach :



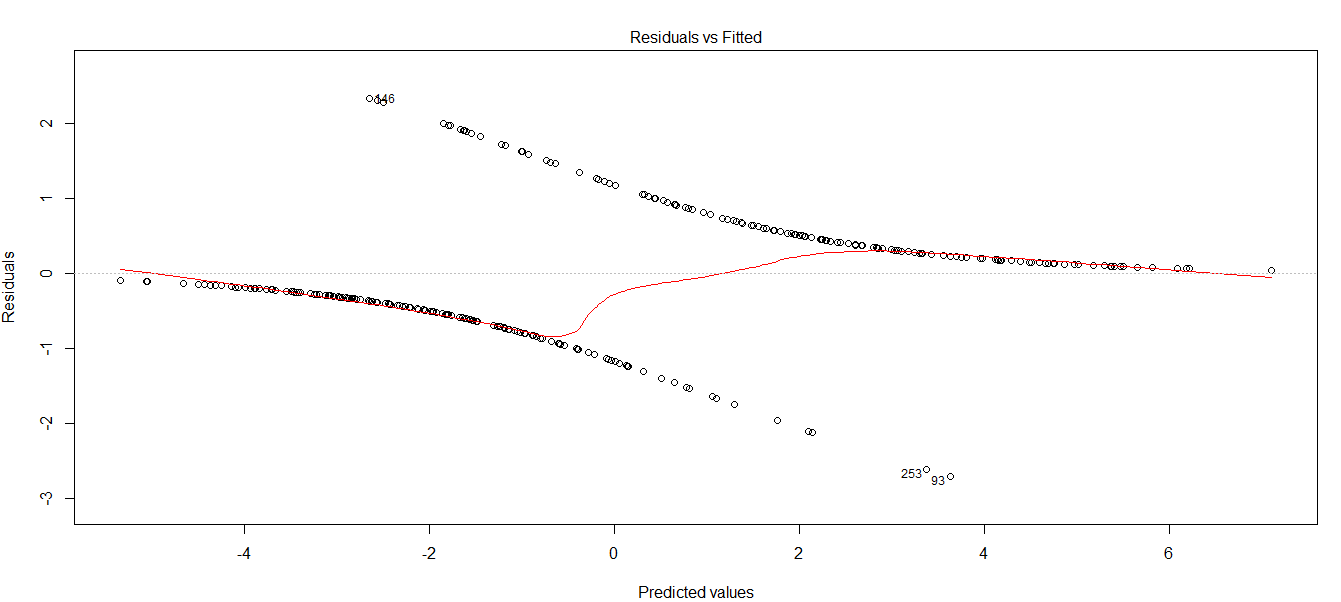
With k-means clustering approach :

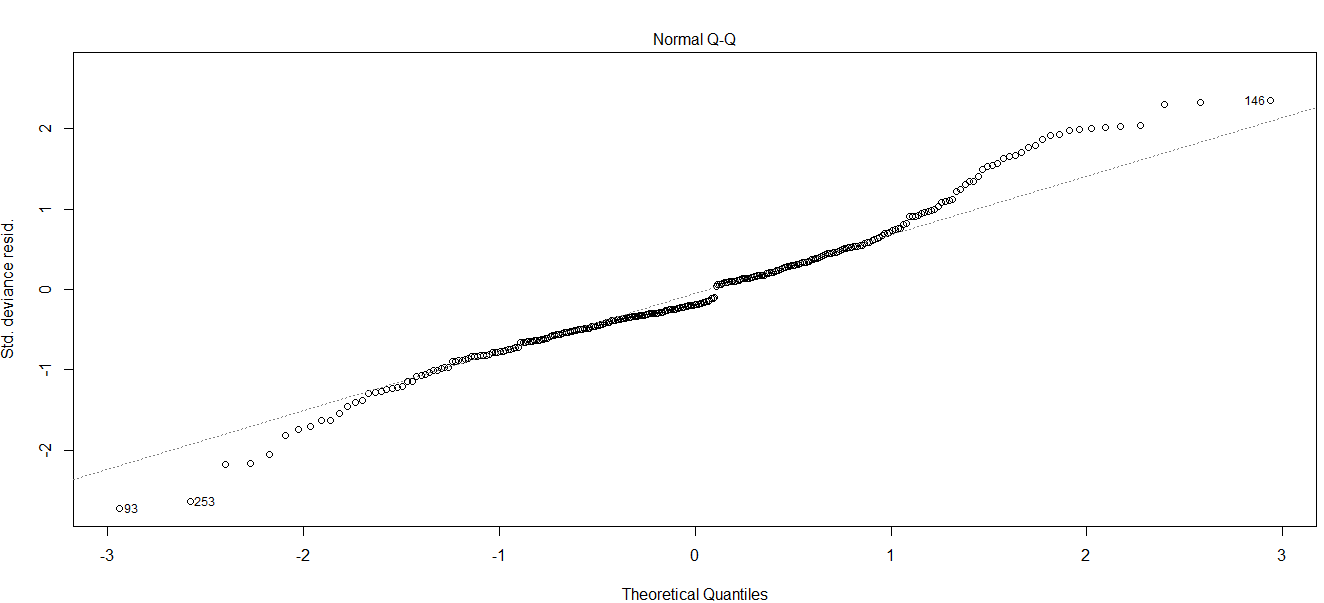


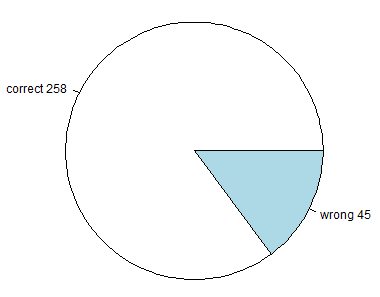
|  |  |
| --- | --- |
| **Efficiency of Naïve Bayes** | **In percentage** |
| Quantitative approach | 81.84818 |
| K-means clustering approach | 88.06436 |

**Logistic Regression:**

The model is plotted and shown below







the accuracy comes out to be: 85.1485%

# Conclusion

Out of all the algorithms tested the **Naïve Bayes with k-means clustering** has the highest performance with efficiency 88.064.

Thus we conclude for this particular data set **Naïve Bayes with k-means clustering** is the best option.

# Scope of future work

There are many classifier algorithms for supervised data like support vector machine, random forest, multiple regression. We aim to implement many supervised classifier algorithms as possible and find the best algorithm with highest efficiency because the correct results are very important in medical world.