Forecast of Heart Disease using K-means and SVM

1Pooja Gupta, 2Pritesh Jain, 3Upendra Singh
1M.Tech Scholar, 2Assistant Professor
1,2Patel College of Science and Technology, Indore

ABSTRACT: In medical sciences prediction of Heart disease is most difficult task. In India, main causes of Death are due to Heart Diseases. The deaths due to heart disease in many countries occur due to work overload, mental stress and many other problems. It is found as main reason in adults is due to heart disease. Thus, for detecting heart disease of a patient, there arises a need to develop a decision support system. Data mining classification techniques, namely Modified K-means and SVM are analyzed on Heart Disease is proposed in this Paper.

KEYWORDS: Data Mining, Heart Disease, Machine Learning, Decision Support, Modified K-Means, SVM.

I. INTRODUCTION

In this world people want to live a very luxurious life so they work like a machine in order to earn lot of wealth. At very young age, this type of lifestyle doesn’t take rest for themselves, which results in diabetics and blood pressure. It is a world known fact that heart is the most essential part in human body if that heart gets affected then it also affects the other parts of the body. Therefore it is essential for people to go for a heart disease diagnosis. People go to healthcare checkup but the prediction made by them is not 100% accurate.

Today, healthcare industry generates large amount of data about patients, disease diagnosis etc. Diagnosis is important task and complicated that needs to be executed accurately and efficiently. Based on doctor’s experience &
Knowledge, the diagnosis is often made. This leads to unwanted results & excessive medical costs of treatments Provided to patients. Quality of service is a major challenge facing Healthcare industry. Quality of service guarantee
Diagnosing disease correctly & to provides effective treatments to patients.

II. RELATED WORK

To have focus on diagnosis of heart disease different studies have been done. Different data mining techniques has been used by them for diagnosis & achieved different probabilities for different methods. Using data mining techniques an Intelligent Heart Disease Prediction System (IHDPS) is developed. Sellappan Palaniappan et al [14] proposed Naïve Bayes, Neural Network, and Decision Trees. For appropriate results each method has its own strength. Hidden patterns and relationship between them is used to build this system. It is user friendly, expandable & web-based.

Niti Guru et al [7] proposed the prediction of Blood Pressure, Sugar and Heart disease with the aid of neural networks. The record of 13 attributes in each was used in the dataset. For training and testing of data, the supervised networks i.e. Neural Network with back propagation algorithm is used.

Heon Gyu Lee et al. [5] proposed a novel technique, to develop the multi-parametric feature with linear and nonlinear characteristics of HRV (Heart Rate Variability) several classifiers e.g. Bayesian Classifiers, CMAR (Classification based on Multiple Association Rules), C4.5 (Decision Tree) and SVM (Support Vector Machine) has been used by them. To measure the impurity of a partition or set of training tuples [2], CART uses Gini index. High dimensional categorical data can handle by it.

III. HEART DISEASE

The heart is important part of our body. Our life is totally dependent on efficient working of heart. If operation of heart is not proper, it will affect the other parts of body such as kidney, brain, etc. It is nothing more than a pump, which pumps blood through the body. Death occurs within minutes, if circulation of blood is inefficient. The term Heart disease refers to blood vessel system within it and disease of heart.

There are number of factors which increase the risk of Heart disease [4].

- Smoking
- Family history of heart disease
- Poor diet
- High Blood Pressure
- Physical inactivity
- Hyper tension
- Obesity
- Cholesterol

Factors like these are used to analyze the Heart disease. In many cases, diagnosis is generally based on doctor’s experience and patient’s current test results. Thus the diagnosis is a complex task that requires much experience & high skill.

IV. DATA SOURCE

Dataset with input attributes is obtained from Cleveland Heart Disease database. With the help of recordset, the heart attack predictions with significant patterns are extracted. The attribute “Diagnosis” with value “1” is identified as Heart Disease prediction and value “0”is identified as no Heart disease prediction for patients. Here key attribute is “PatientId” and other attributes are used as input.

Predictable attribute

1. Diagnosis (value 0: <50% diameter narrowing (no heart disease); value 1: >50% diameter narrowing (has heart disease))

Key attribute

1. PatientId – Patient’s identification number

Input attributes

1. Sex (value 1: Male; value 0: Female)
2. Age in Year
3. Oldpeak – ST depression induced by exercise
4. Restecg – resting electrographic results (value 0:normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl)
6. Slope – the slope of the peak exercise ST segment (value 1:unsloping; value 2: flat; value 3: downsloping)
7. Exang - exercise induced angina (value 1: yes; value 0: no)
8. Serum Cholesterol (mg/dl)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
11. CA – number of major vessels colored by fluoroscopy (value 0-3)
12. Thalach – maximum heart rate achieved
13. Chest Pain Type (value 1:typical type 1 angina, value 2: typical type angina, value 3:non-angina pain; value 4: asymptomatic)

IV. PROPOSED ALGORITHM

Today, many hospitals manage healthcare data using healthcare information system; as this system contains huge amount of data, and it is used to extract hidden information for medical diagnosis. The main objective of this system is to build Heart Disease Prediction System using historical heart database that gives diagnosis of heart disease. To build this system, medical terms such as blood pressure, sex, cholesterol, sugar etc 13 input attributes are used. Data mining techniques such as clustering, Classification is used in extracting knowledge from database.

A. Modified K-means:

The proposed modified algorithm proves to be a better method to determine the initial centroids and it is easy to implement. By eliminating one of its drawbacks, this modified K-means tries to enhance the k means clustering Algorithm. K-means was used to apply on numerical data only. But, we encounter both numerical and categorical combination data values.
This algorithm does not require number of clusters (k) as input is described below. By choosing two initial centroids, two clusters are created initially, which are farthest apart in the datasets. It can create two clusters with the data members at the initial steps, which are most dissimilar ones.

Input:

D: The set of n tuples with attributes Al, A2, , Am. All attributes are numeric, (where m = no. of attributes)

Output:

With n tuples suitable number of clusters distributed properly

Method:

1) To find the points in the data set which are farthest apart, compute sum of the attribute values of each tuple
2) As initial centroids take tuples with maximum and minimum values of the sum.
3) Using Euclidean distance create initial partitions (clusters) between the initial centroids and every tuple

4) From the centroid find distance of every tuple in both the initial partitions. Take other than zero. d=minimum of all distances.
5)) For the partitions created in step 3,compute new means (centroids)
6) From the new means (cluster centers) compute Euclidean distance of every tuple.
and depending on the following objective function, find the outliers: If Distance of the tuple from the cluster mean>dthen only it is an Outlier.
7) New centroids of the clusters can be computed
8) From the new cluster centroids , calculate Euclidean distance of every outlier and find the objective function in step 6. outliers is not satisfying
9) Let the set of outliers obtained in step 8 is B={ Yl,Y2,. ....Y p} (Where value of k is depends on number of outliers).
10) Repeat the steps until I (B==<D)
a) By taking mean value of its members as centroid, create a new cluster for the set B,
b) Depending on the objective function in step 6, find the outliers of this cluster,
c) Check if no. of outliers = p then
i) Test every other outlier for the objective function as in step 6 after creation of a new cluster with one of the outliers as its member
ii) If there is any outfliers find it
d) From the centroid of the existing clusters, calculate the distance of every outlier. If the existing clusters which satisfy the objective function in step 6. Then adjust the outliers
e) The new set of outliers be B={ ZI,Z2 .... Zq}. (Where value of q is depends on number of outliers)

4.2 Flow chart of proposed work

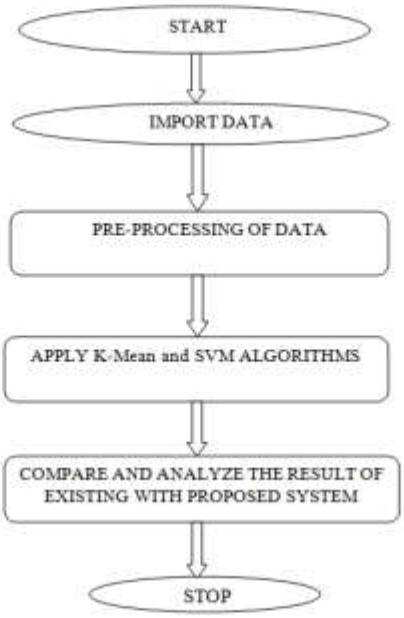
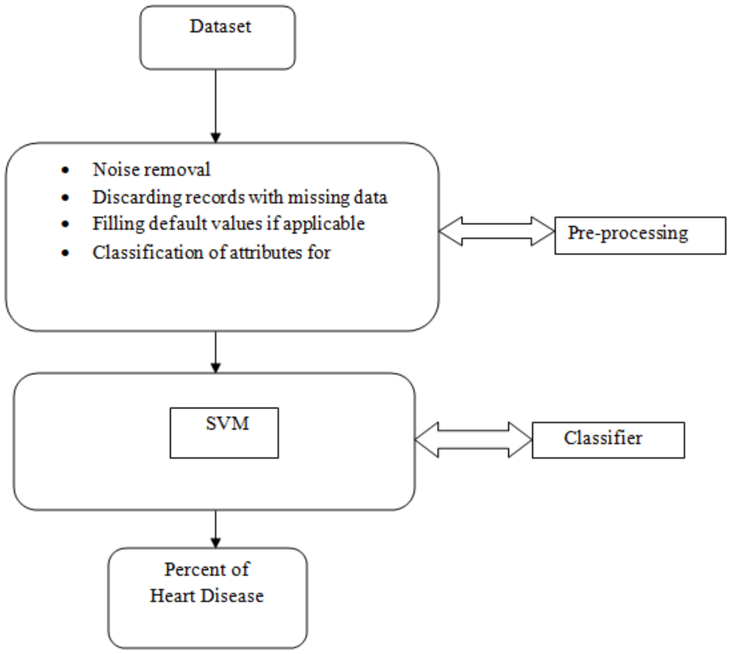

START
IMPORT DATA
PRE-PROCESSING OF DATA
APPLY K-Mean and SVM ALGORITHMS
COMPARE AND ANALYZE THE RESULT OF EXISTING WITH PROPOSED SYSTEM
STOP

Figure 4.1: Flow chart of proposed work.

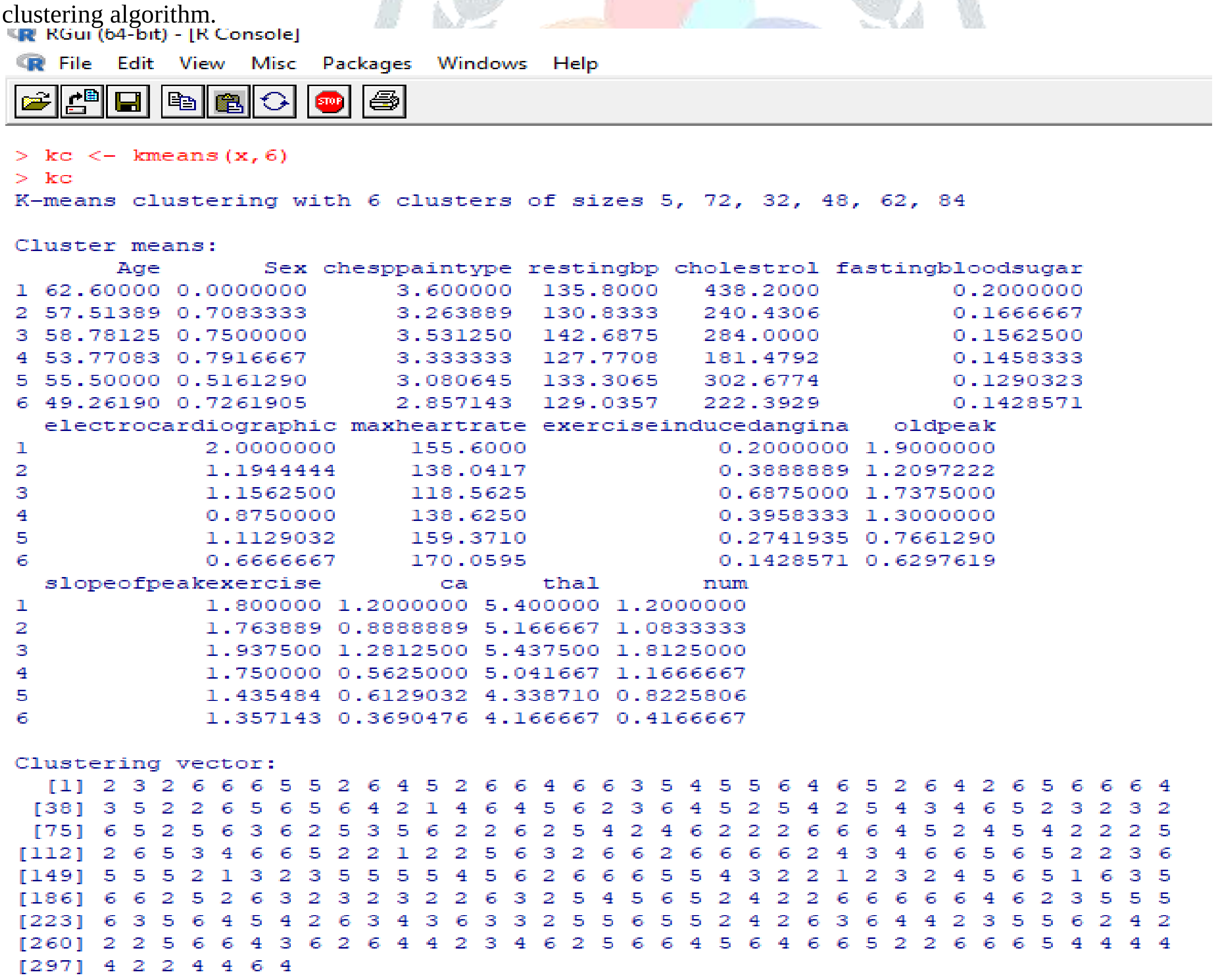
4.3 Flow Chart of SVM Working Process


Dataset
Noise removal
Discarding records with missing data
Filling default values if applicable
Classification of attributes for
Pre-processing
SVM
Classifier
Percent of
Heart Disease

4.2 Flow Chart of SVM Working Process.

V. EXPERIMENT RESULTS

A total of 308 records with 14 attributes were used from the Cleveland Heart database [1].User enter values in medical attributes like sex, age, etc. This model predicts that patient is having heart disease or not depending on this value, doctors would recommend to go for further heart examination. Fig 1.are used to load UCI repository dataset for testing purpose using K-Mean clustering algorithm.



```
> kc <- kmeans(x,6)
> kc
K-means clustering with 6 clusters of sizes 5, 72, 32, 48, 62, 84

Cluster means:
       Age       Sex chesppaintype restingbp cholestrol fastingbloodsugar
1 62.60000 0.0000000      3.600000  135.8000   438.2000         0.2000000
2 57.51389 0.7083333      3.263889  130.8333   240.4306         0.1666667
3 58.78125 0.7500000      3.531250  142.6875   284.0000         0.1562500
4 53.77083 0.7916667      3.333333  127.7708   181.4792         0.1458333
5 55.50000 0.5161290      3.080645  133.3065   302.6774         0.1290323
6 49.26190 0.7261905      2.857143  129.0357   222.3929         0.1428571
  electrocardiographic maxheartrate exerciseinducedangina   oldpeak
1            2.0000000     155.6000             0.2000000 1.9000000
2            1.1944444     138.0417             0.3888889 1.2097222
3            1.1562500     118.5625             0.6875000 1.7375000
4            0.8750000     138.6250             0.3958333 1.3000000
5            1.1129032     159.3710             0.2741935 0.7661290
6            0.6666667     170.0595             0.1428571 0.6297619
  slopeofpeakexercise        ca     thal       num
1            1.800000 1.2000000 5.400000 1.2000000
2            1.763889 0.8888889 5.166667 1.0833333
3            1.937500 1.2812500 5.437500 1.8125000
4            1.750000 0.5625000 5.041667 1.1666667
5            1.435484 0.6129032 4.338710 0.8225806
6            1.357143 0.3690476 4.166667 0.4166667

Clustering vector:
  [1] 2 3 2 6 6 6 5 5 2 6 4 5 2 6 6 4 6 6 3 5 4 5 5 6 4 6 5 2 6 4 2 6 5 6 6 6 4
 [38] 3 5 2 2 6 5 6 5 6 4 2 1 4 6 4 5 6 2 3 6 4 5 2 5 4 2 5 4 3 4 6 5 2 3 2 3 2
 [75] 6 5 2 5 6 3 6 2 5 3 5 6 2 2 6 2 5 4 2 4 6 2 2 2 6 6 6 4 5 2 4 5 4 2 2 2 5
[112] 2 6 5 3 4 6 6 5 2 2 1 2 2 5 6 3 2 6 6 2 6 6 6 6 2 4 3 4 6 6 5 6 5 2 2 3 6
[149] 5 5 5 2 1 3 2 3 5 5 5 5 4 5 6 2 6 6 6 5 5 4 3 2 2 1 2 3 2 4 5 6 5 1 6 3 5
[186] 6 6 2 5 2 6 3 2 3 2 3 2 2 6 3 2 5 4 5 6 5 2 4 2 2 6 6 6 6 6 4 6 2 3 5 5 5
[223] 6 3 5 6 4 5 4 2 6 3 4 3 6 3 3 2 5 5 6 5 5 2 4 2 6 3 6 4 4 2 3 5 5 6 2 4 2
[260] 2 2 5 6 6 4 3 6 2 6 4 4 2 3 4 6 2 5 6 6 4 5 6 4 6 6 5 2 2 6 6 6 5 4 4 4 4
[297] 4 2 2 4 4 6 4
```

Figure 5.1 K-Mean Algorithm divided into 6 Cluster in whole Data Set.

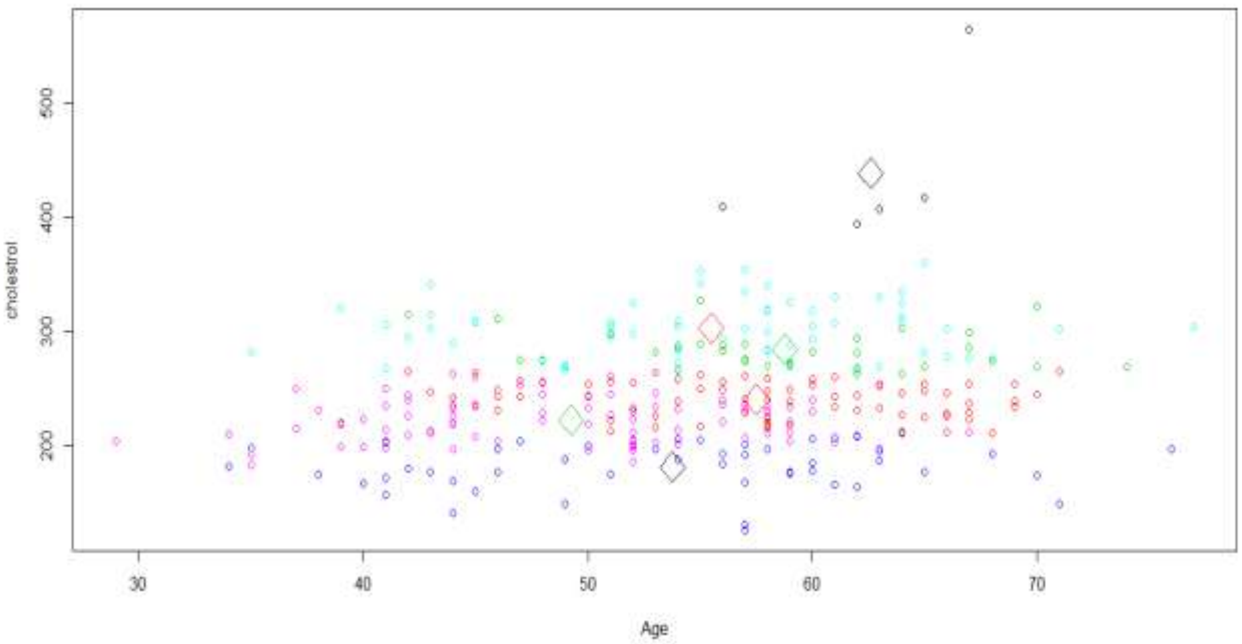


Figure 5.2 K-Mean Algorithm plot Graph Age verses Cholesterol.

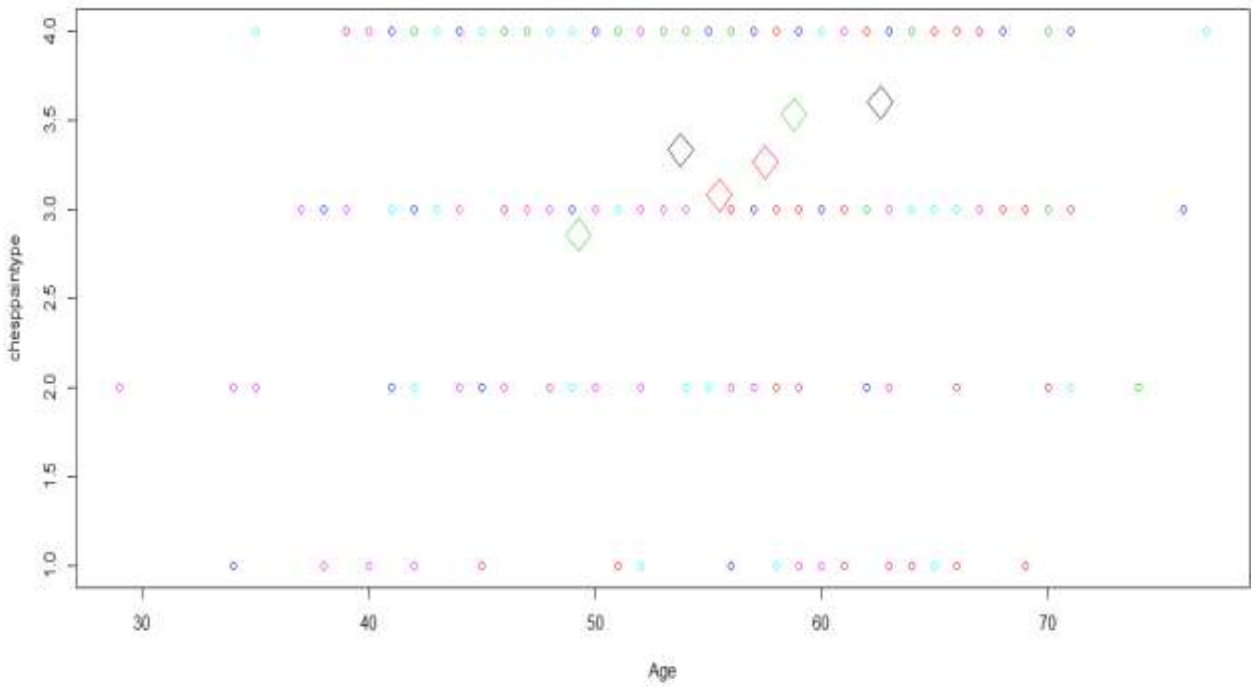


Figure 5.3 K-Mean Algorithm plot Graph Age verses Chest Pain Type.

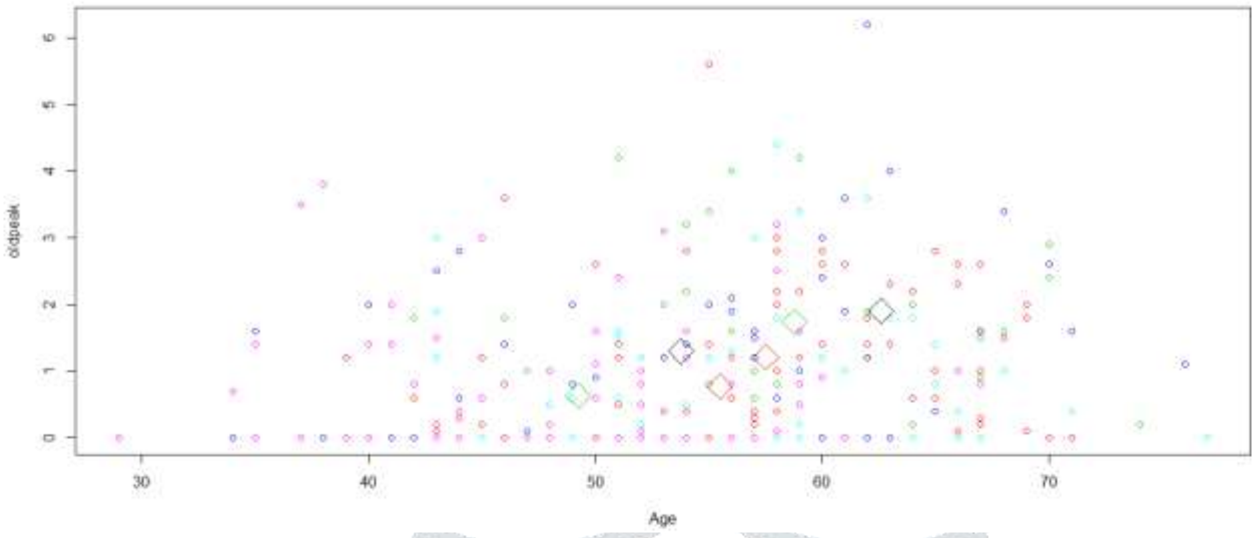
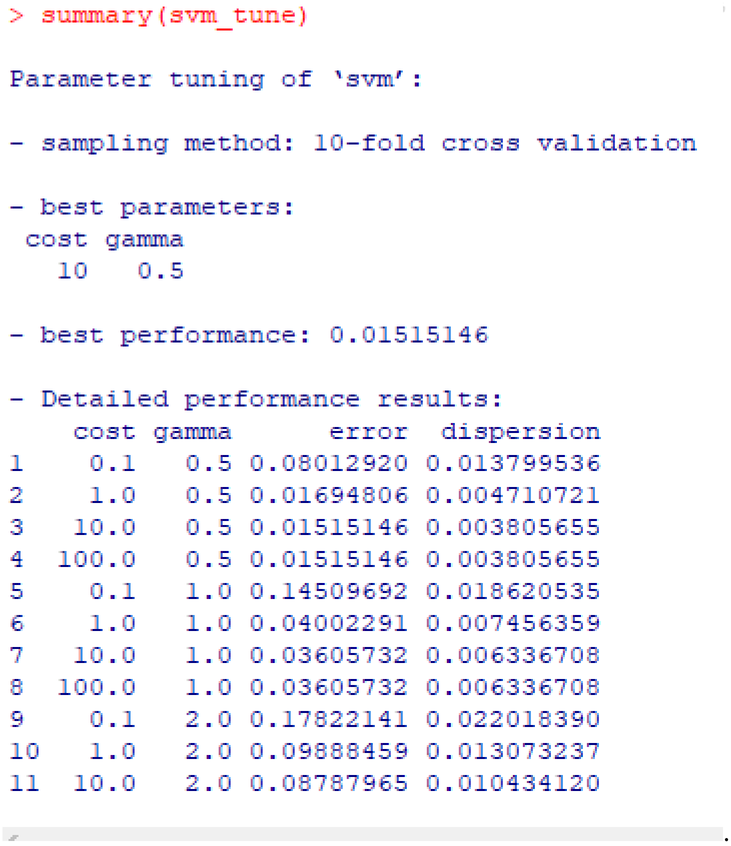


Figure 5.4 K-Mean Algorithm plot Graph Age verses Old Peak.



```
> summary(svm_tune)

Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost gamma
   10   0.5

- best performance: 0.01515146

- Detailed performance results:
    cost gamma      error  dispersion
1    0.1   0.5 0.08012920 0.013799536
2    1.0   0.5 0.01694806 0.004710721
3   10.0   0.5 0.01515146 0.003805655
4  100.0   0.5 0.01515146 0.003805655
5    0.1   1.0 0.14509692 0.018620535
6    1.0   1.0 0.04002291 0.007456359
7   10.0   1.0 0.03605732 0.006336708
8  100.0   1.0 0.03605732 0.006336708
9    0.1   2.0 0.17822141 0.022018390
10   1.0   2.0 0.09888459 0.013073237
11  10.0   2.0 0.08787965 0.010434120
```

.

Figure 5.5 SVM Algorithms.

VI Conclusion and Future Work

The main aim of our project is to predict more accurately the presence of heart disease. With less number of attributes is a challenging task in Data Mining, Instead of going for a number of tests. Two data classification techniques were applied namely modified K-means and SVM. In this paper a modified K means algorithm is proposed which tries to remove one of the major limitations of basic K-means algorithm, which requires number of clusters as input. This system can be further used in Future work as, for eg. It can incorporate other medical attributes besides the above list. To mine large amount of unstructured data, the Text mining can be used, available in healthcare industry database.

REFERENCES

1 Blake, C.L., Mertz, C.J.: “UCI Machine Learning Databases”, http://mlearn.ics.uci.edu/databases/heartdisease/,2004.

2 Han, J., Kamber, M.: “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2006.

3 Mrs.G.Subbalakshmi, “Decision Support in Heart Disease Prediction System using Naive Bayes”, Indian Journal of Computer Science and Engineering

4 Yanwei, X.; Wang, J.; Zhao, Z.; Gao, Y., “Combination data mining models with new medical data to predict outcome of coronary heart disease”. Proceedings International Conference on Convergence Information Technology 2007, pp. 868 – 872.

5 Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, “Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV,” LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007

6 Statlog database: http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart

7 Niti Guru, Anil Dahiya, NavinRajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1 (January - June 2007).

8 B M Ahamed Shafeeq, K S Hareesha, " Dynamic Clustering of Data with Modified K-Means Algorithm", International Conference on Information and Computer Networks (ICICN 2012), IPCSIT, vol. 27,pages 221-225,2012

9 Mohamed Abubaker, Wesam Ashour, "Efficient Data Clustering Algorithms: Improvements over K-means", International Journal of Intelligent Systems and Applications, vol. 5,issue 3, pages 37-49, 2013

10 Mohammed EI Agha, Wesam M. Ashour, " Efficient and Fast Initializtion Algorithm for K-means Clustering" ,1.1. Intelligent Systems and Applications, vol. 4, issue 1, pages 21-31, 2012.

11 Kaur, H., Wasan, S. K.: “Empirical Study on Applications of Data Mining Techniques in Healthcare”, Journal of Computer Science 2(2), 194-200, 2006.

12 MA.Jabbar,B.L Deekshatulu,Priti chandra,”Prediction of Risk Score for Heart Disease using Associative classification and Hybrid Feature Subset Selection”,In .Conf ISDA,pp 628-634,IEEE(2013)

13 MA.Jabbar,B.L Deekshatulu,Priti chandra,”Knowledge discovery from mining association rules for heart disease prediction” pp45-53,vol 41,no 2 ,JATIT(2013

14 SellappanPalaniappan, RafiahAwang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008