Comparative Analysis of Classification and Regression Algorithms using R

Talha Hanif Butt, L18-1864, FAST-NUCES, Lahore

October 14, 2018

Abstract

This report presents a comparative analysis of three clustering and three classification algorithms using results produced on three different datasets using R.

1 Introduction

There has been an increase in the use of machine learning algorithms in the past 5 years and this is because of the performance of certain algorithms like CNNs, RNNs, LSTMs on tasks such as face recognition, speech recognition and autonomous driving. Classification and regression algorithms have been employed on various tasks related to medical, business, education and many other industries and this widespread use allows scientists to easily apply their methods on such problems. This report presents a study of different classification and regression algorithms on different datasets using R as a means of learning a new and useful language.

2 Datasets

Three datasets have been used in this report for comparing the performance of different algorithms, the details of which are as under:

2.1 Glass

It has 214 examples of the chemical analysis of 7 different types of glass. The problem is to forecast the type of class on basis of the chemical analysis. The study of classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence (if it is correctly identified!).

2.2 Iris

Iris is a dataset with 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.

2.3 PimaIndiansDiabetes

A dataset with 768 observations on 9 variables.

3 Algorithms

Six algorithms have been used for comparison in this report namely kNN, SVM, Random Forest, K-Means, Hierarchical Agglomerative and K-Medoids.

3.1 kNN

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.

3.2 SVM

A support vector machine (SVM) is machine learning algorithm that analyzes data for classification and regression analysis. SVM is a supervised learning method that looks at data and sorts it into one of two categories. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible.

3.3 Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

3.4 K-Means

K-Means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-Means clustering aims to

partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

3.5 Hierarchical Agglomerative

Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC.

3.6 K-Medoids

The K-Medoids algorithm is a clustering algorithm related to the K-Means algorithm and the medoidshift algorithm. Both the K-Means and K-Medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the K-Means algorithm, K-Medoids chooses datapoints as centers (medoids or exemplars) and works with a generalization of the Manhattan Norm to define distance between datapoints.

4 Experimental Setup

For the experiments, built-in packages in R with their default setup for each algorithm have been used except for the number of classes in case of classification and the number of clusters in case of clustering algorithms which is 2, 7 and 3 for PimaIndiansDiabetes, Glass and Iris dataset respectively.

5 Data Visualization

5.1 Glass

Figure 1: Barplot for class breakdown

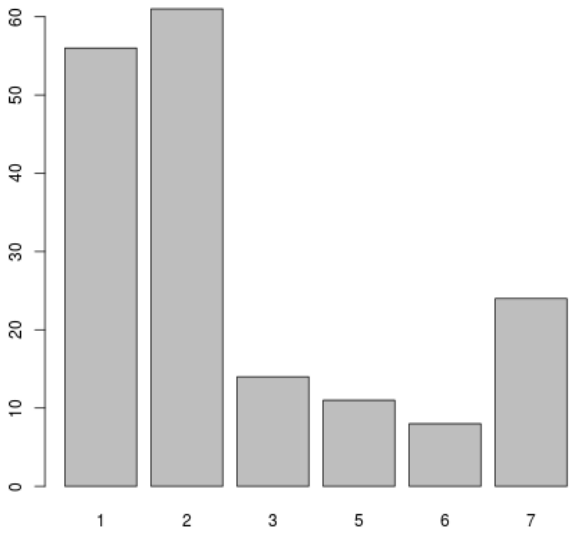


Figure 2: Box and whisker plots for each attribute

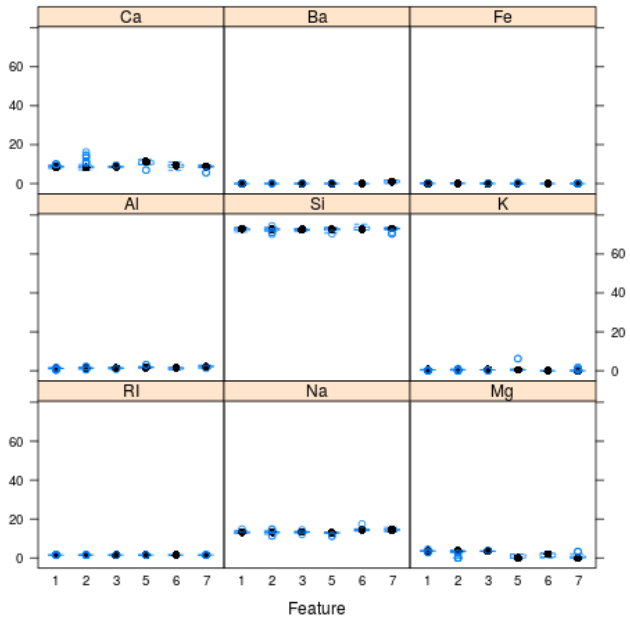


Figure 3: Scatterplot matrix

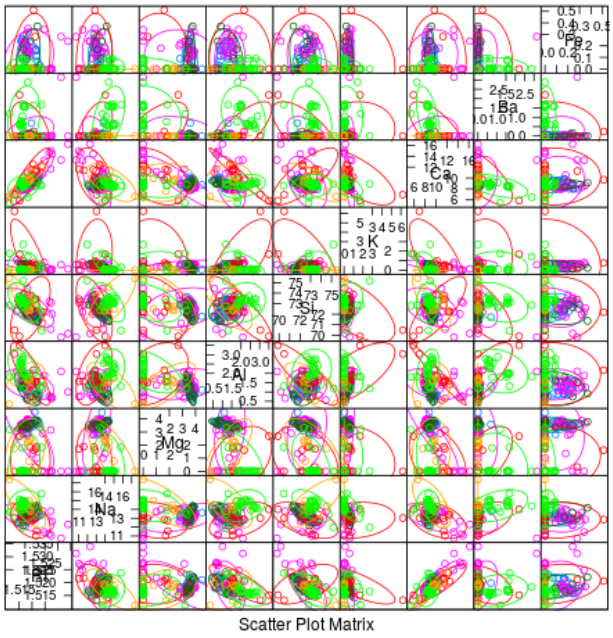
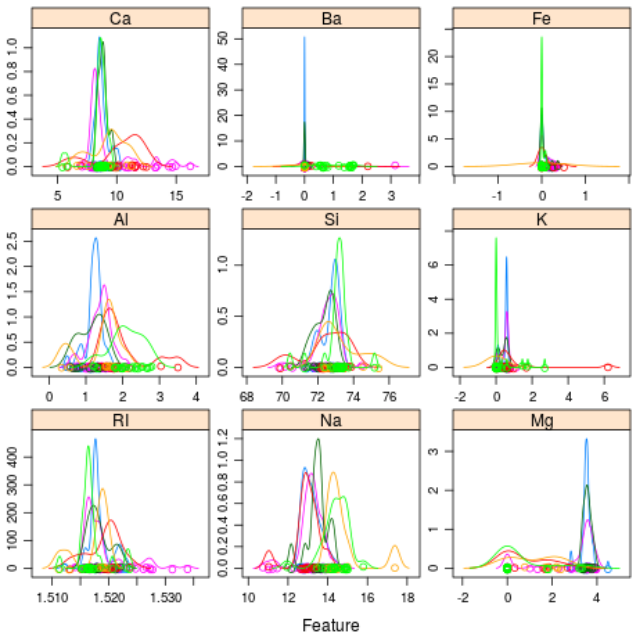


Figure 4: Density plots for each attribute by class value



5.2 Iris

Figure 5: Barplot for class breakdown

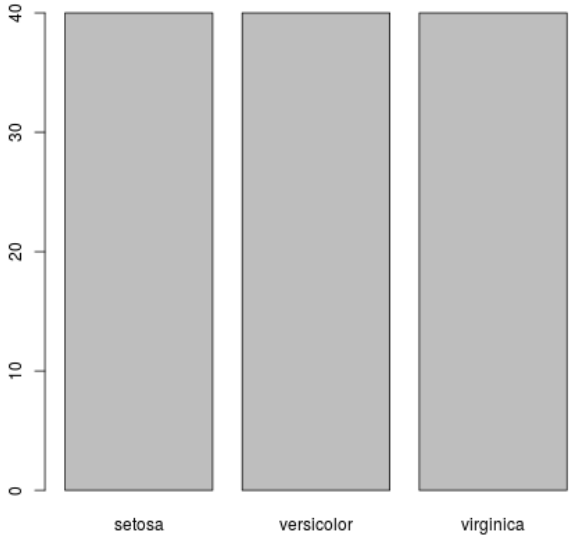


Figure 6: Box and whisker plots for each attribute

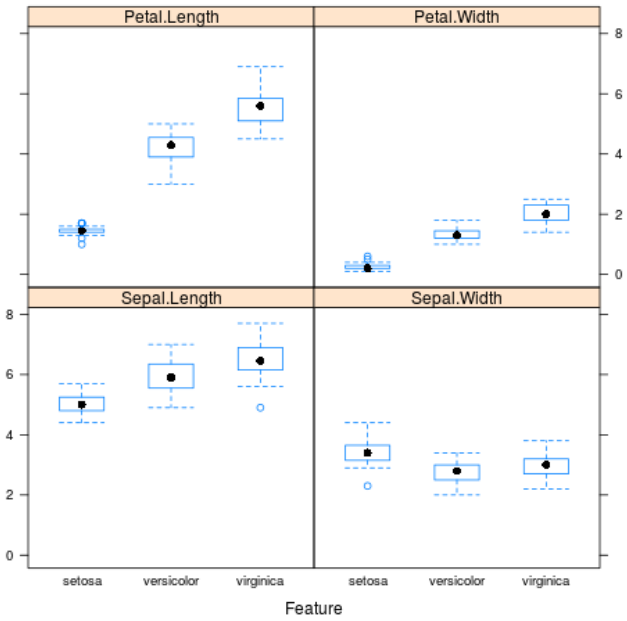


Figure 7: Scatterplot matrix

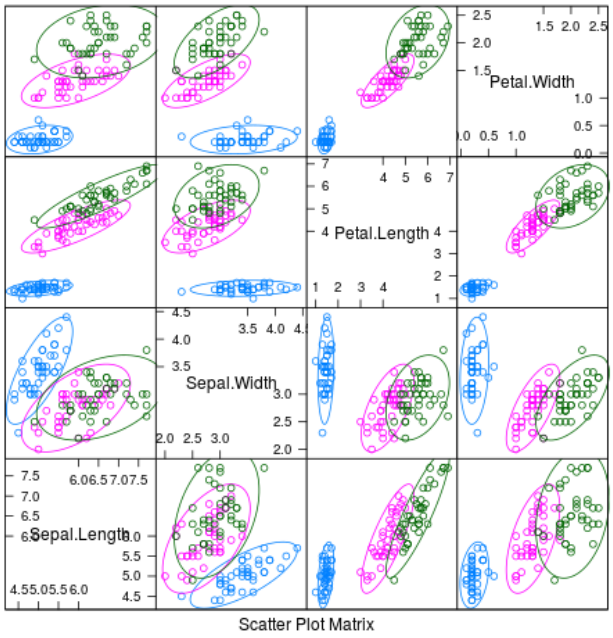
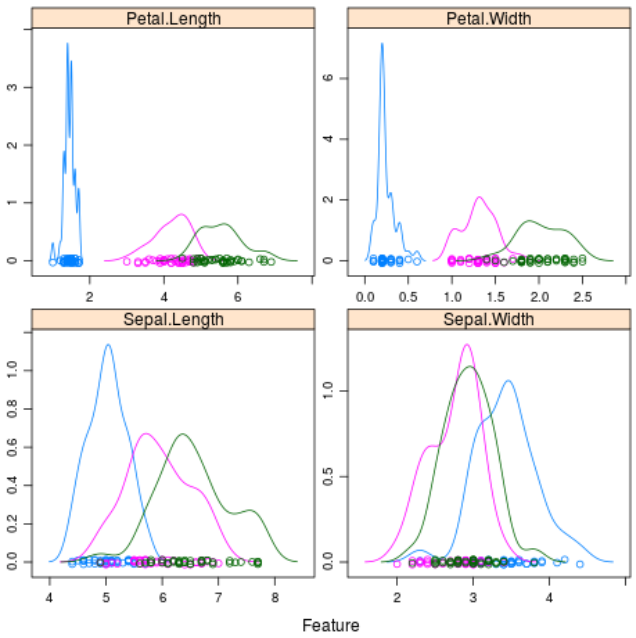


Figure 8: Density plots for each attribute by class value



5.3 Pima Indians Diabetes

Figure 9: Barplot for class breakdown

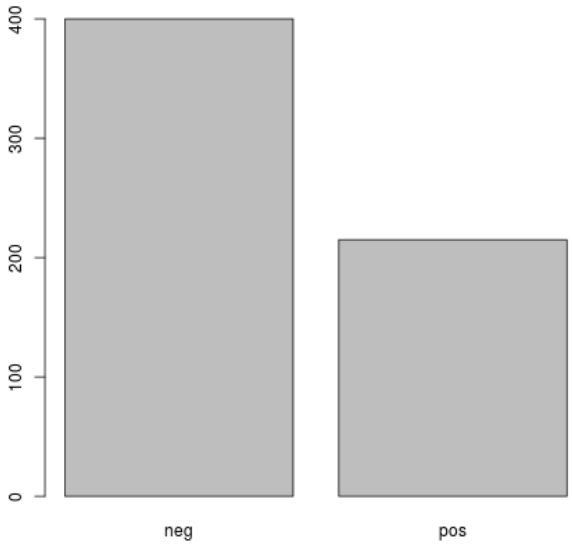


Figure 10: Box and whisker plots for each attribute

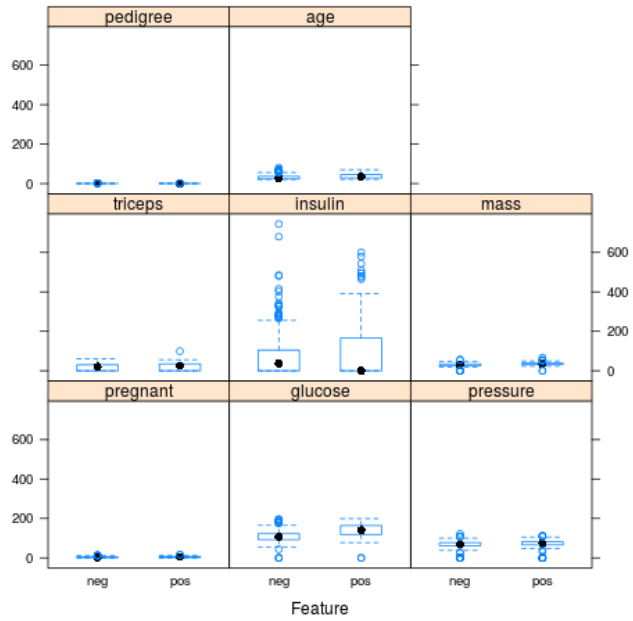


Figure 11: Scatterplot matrix

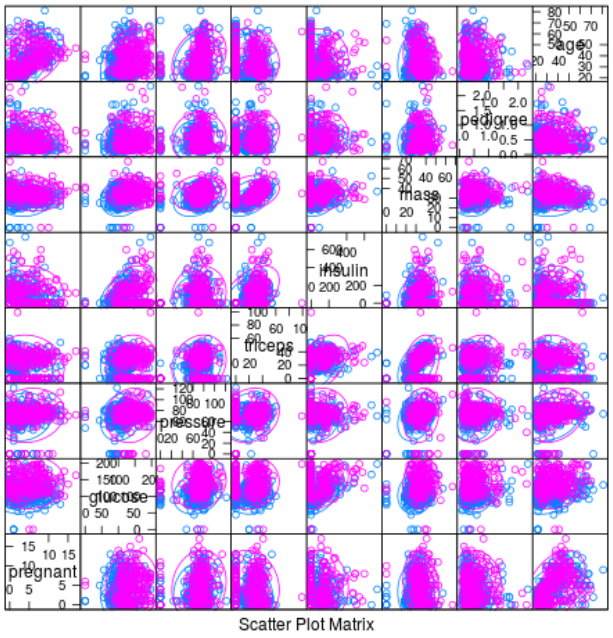
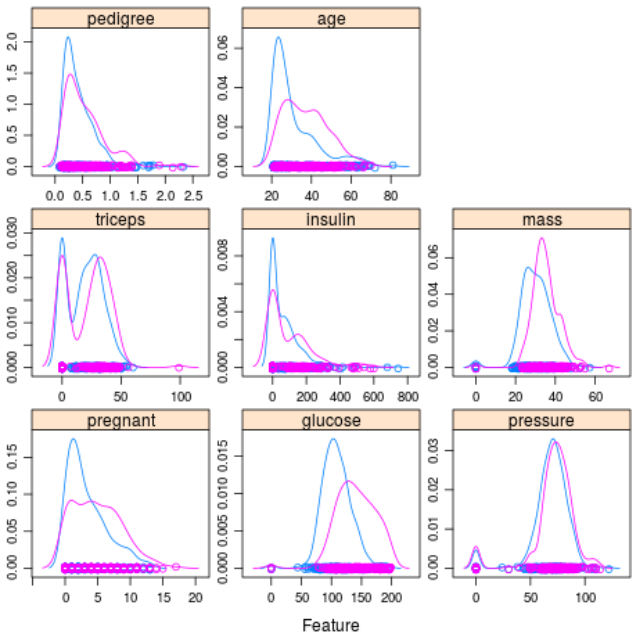


Figure 12: Density plots for each attribute by class value

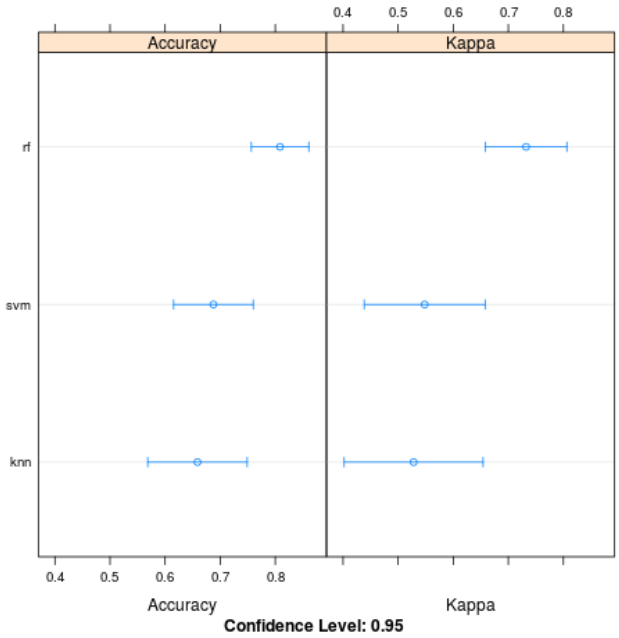


6 Results

6.1 Classification

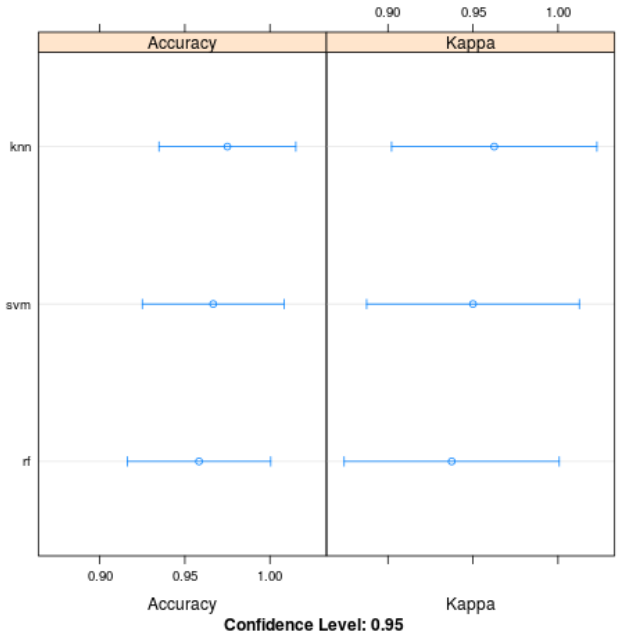
Glass

Figure 13: Accuracy comparison of models



Iris

Figure 14: Accuracy comparison of models



Pima Indians Diabetes

Figure 15: Accuracy comparison of models

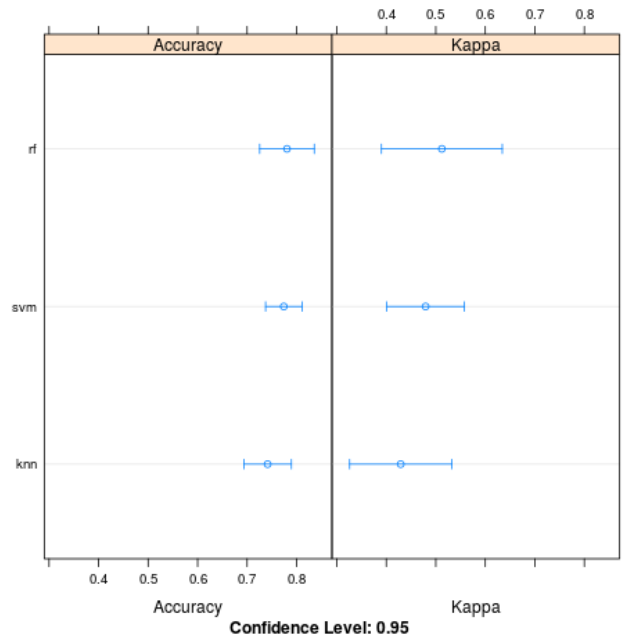
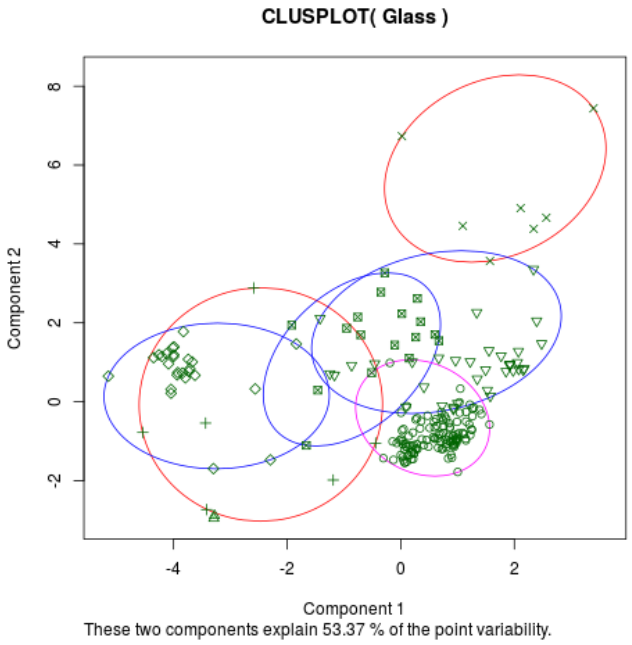


Figure 17: K-Means Clusterplot



6.2 Regression

Glass

Figure 16: Hierarchical Agglomerative Cluster Dendrogram

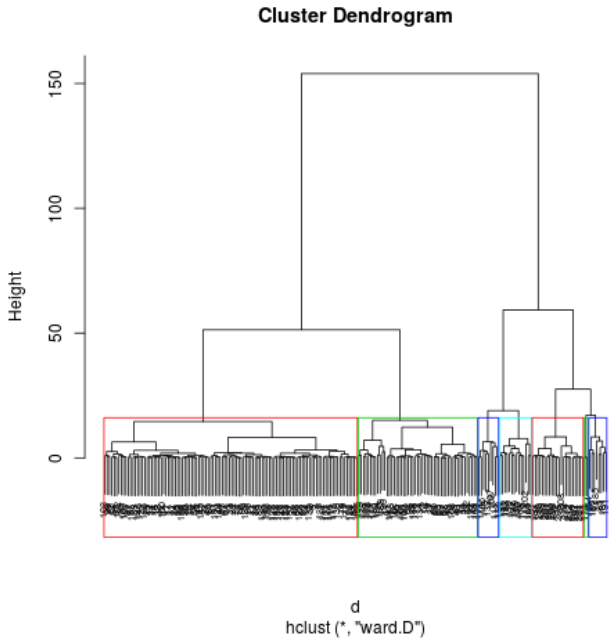
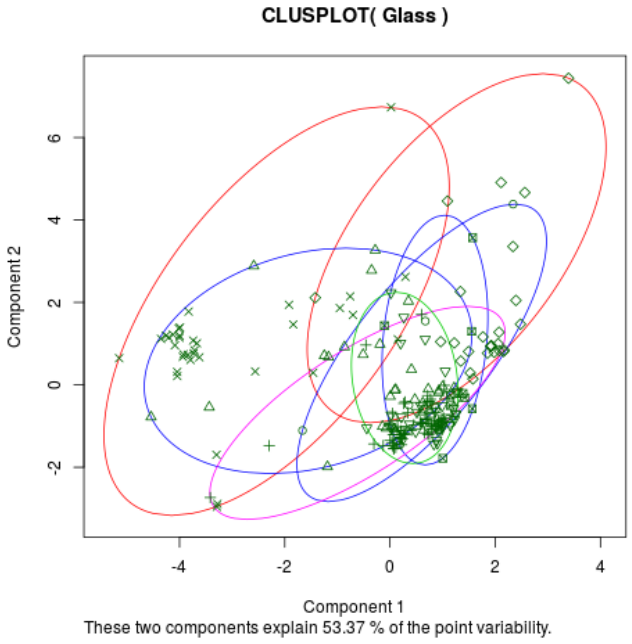


Figure 18: K-Medoids Clusterplot



Iris

Figure 19: Hierarchical Agglomerative Cluster Dendrogram

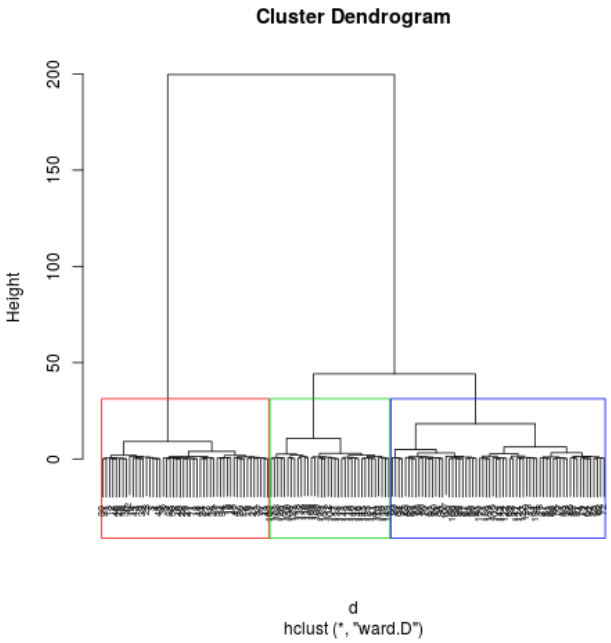


Figure 20: K-Means Clusterplot

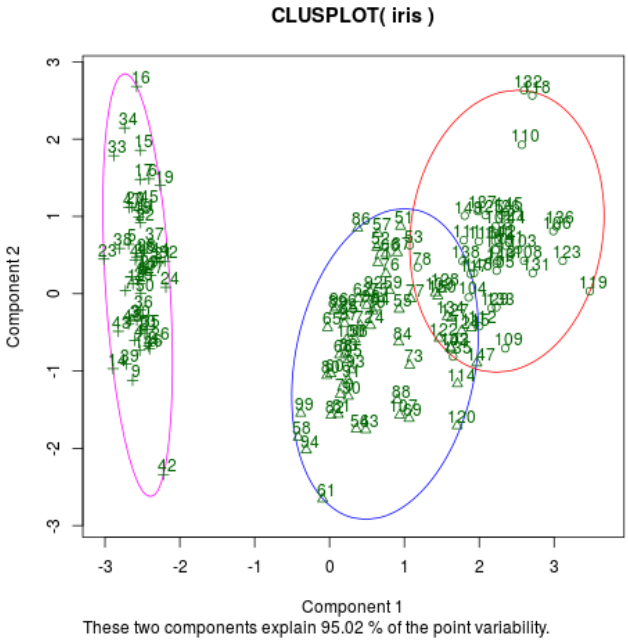
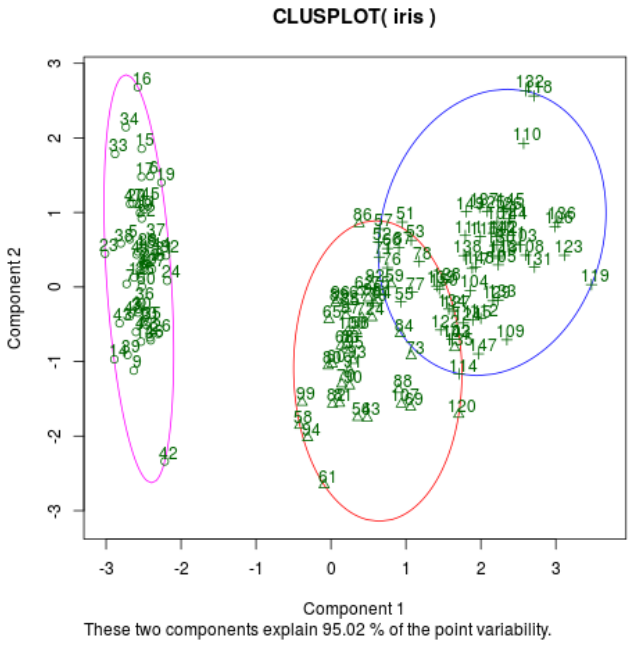


Figure 21: K-Medoids Clusterplot



Pima Indians Diabetes

Figure 22: Hierarchical Agglomerative Cluster Dendrogram

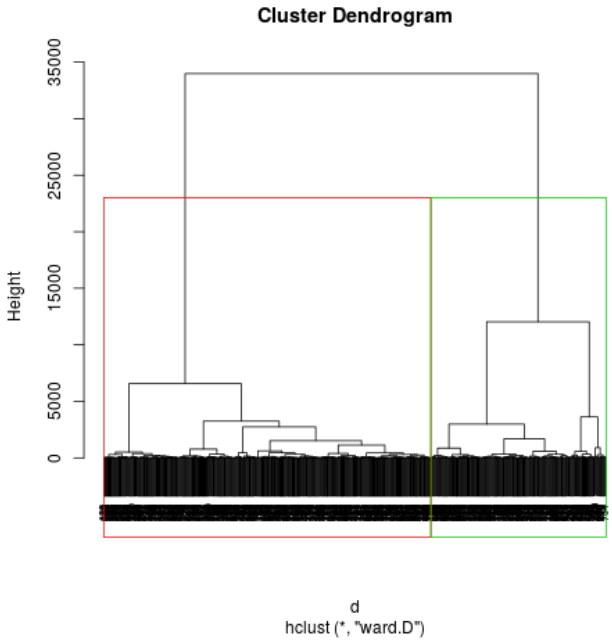


Figure 23: K-Means Clusterplot

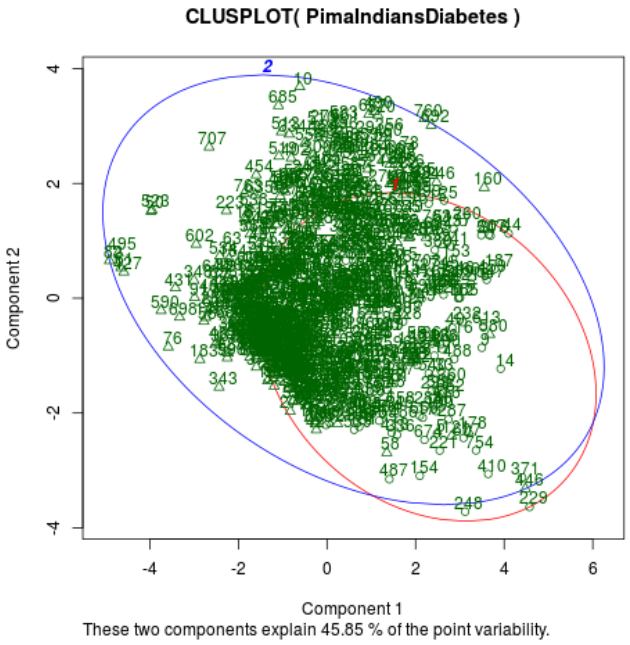
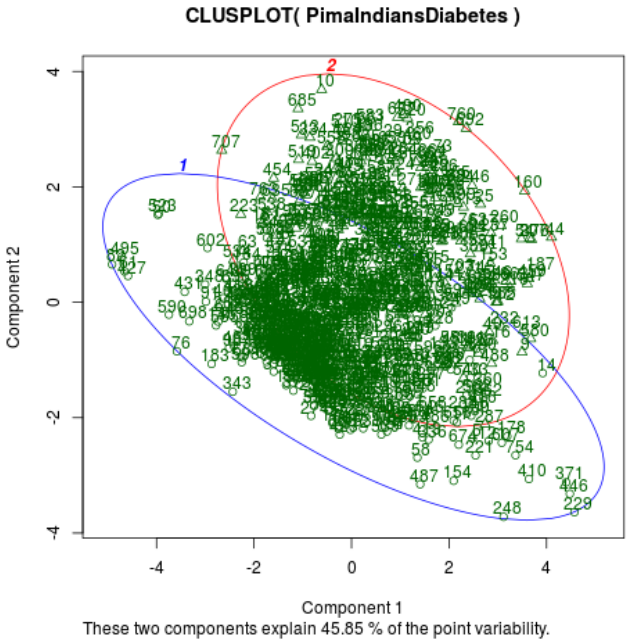


Figure 24: K-Medoids Clusterplot



7 Conclusion

It is evident from Figure 13,14 and 15, that Random Forest has performed better on Glass and Pima Indians Diabetes datasets while SVM has outperformed kNN and Random Forest on Iris dataset as far as classification is concerned. From Figure 16,19 and 22, it can be said that among K-Means, K-Medoids and Hierarchical Agglomerative Clustering, Hierarchical Agglomerative Clustering has separated the classes better in all three cases.