**Module 4 – Data Mining Assignment**

**Name: Sunil Raj Thota**

**NUID: 001099670**

**Date: 02/08/2021**

**Title: Data Mining Assignment**

**ALY 6015 – Intermediate Analytics**

**Prof. Roseanna Hopper**

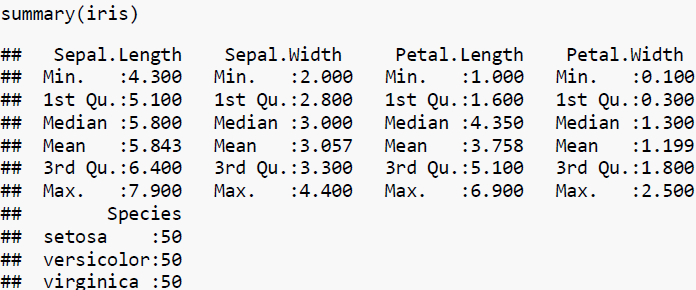
**Introduction**

In this task, I used R and R studio to perform Descriptive Statistics and Data Mining techniques on the preloaded dataset provided by CRAN. We can notice a list of all data sets in R using the function called data().

The dataset used for the analysis is the “iris” to discover the k means clustering, density-based clustering, and hierarchical clustering on the Sepal.width, Petal.width, Petal.length, Sepal.Length, and Species variables. Also, used additional packages such as factoextra, NbClust, party, caret, and fpc libraries.

**Analysis**

**Problem 1**

****

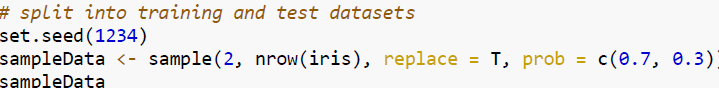
Let's perform some Exploratory Data Analysis and Data Mining techniques using "iris" data set. To get this data set we need to install the 'data sets' package from the packages tab which is right side to the work space in R Studio. Or we can also install the packages by using install.packages("package name") command. Once it is loaded, we can use it in the code for further analysis and calculations.

Loaded the "data sets" library into the work space. Loaded the iris Data set into the Environment. I have also installed factoextra', 'party', 'caret', 'fpc', and 'NbClust' to perform data mining analysis in R. Let's install all the above packages.

To View the diabetes Data set we use View() command, To observe the structure of the Data set we use str() command, and head () and tail() shows first and last few rows in the Data set. Summary() Provides the Descriptive Stats of the iris columns. We noticed 5 variables from the statistics given in the summary.

**Problem 2**

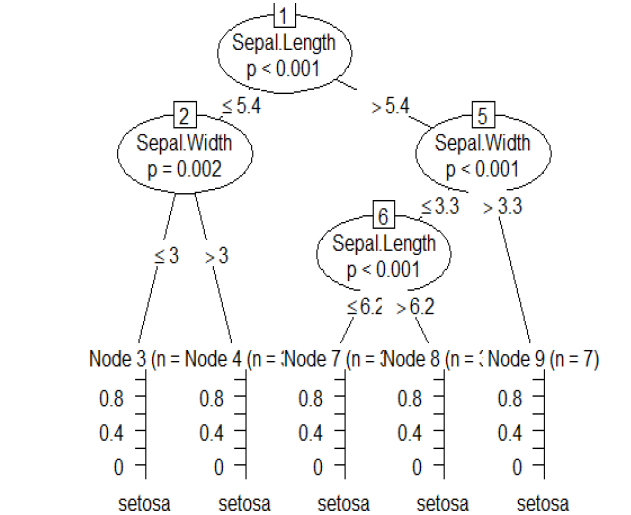
To perform Data Mining, we are taking 'iris' data set which has 150 records and 5 variables. The columns that we are working now are Sepal.Length, Petal.Length, Sepal.Width, Petal.Width, and Species of the ‘iris’ data set. To generate a sequence of random numbers we use set.seed() in R.



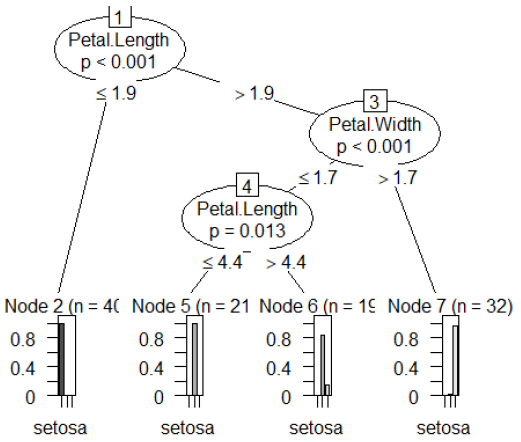
With the help of sample() method, we can split the data set into two parts i.e., training and testing data sets which will be splitted into 70: 30 ratios of the main data set ('iris'). To store these splitted data sets we require two different variables that needs to be assigned as 'trainData' and 'testdata'.

**Exercise 3**

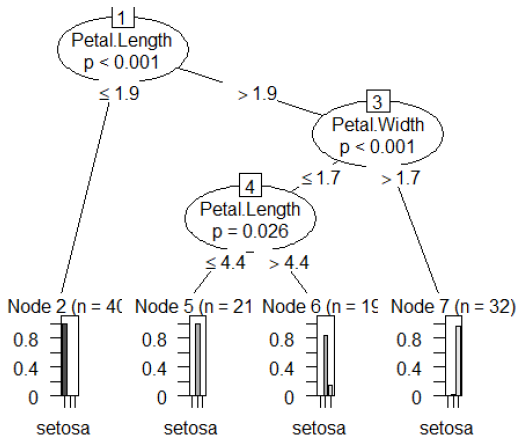
To apply Decision Tree Model to our data set, we have to use "ctree()" method and include our Training Data set. Here we are performing by taking various independent variables and named as irisSepal, irisPetal, and irisAll.



The above plot is between the Species ~ the Sepal.Width and Sepal.Length



The above plot is between the Species ~ the Petal.Length and Petal.Width

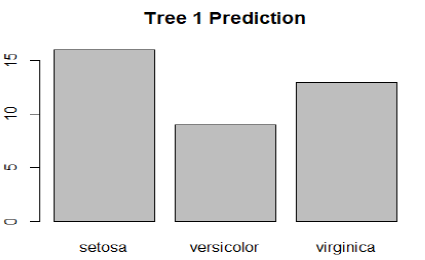


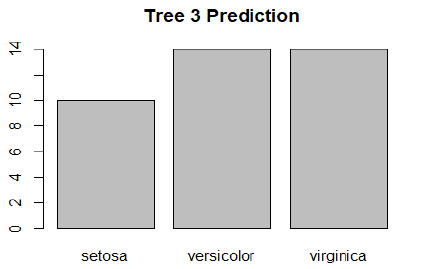
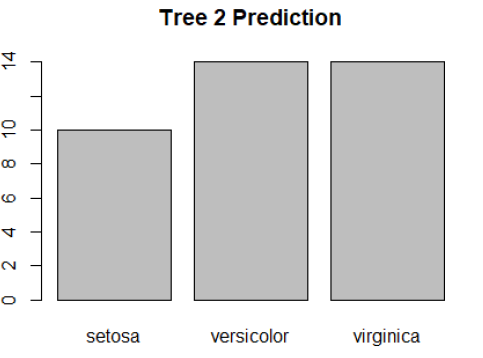
The above plot is between the Species ~ the Petal.Length, Sepal.Length, Petal.Width, and Sepal.Width

Let us also delve deep into the analysis by plotting these Decision Trees models to compare and observe the differences in each plot. From the plots, we can depict that there is a change in the root nodes and leaves. Primary node is called as the root node of the tree and the bottom ones are known as leaves. In this model, the conclusion makers are leaves. The final result is validated by a leaf node where the model terminates and has a value on the leaf

**Problem 4**

Let's predict our model with the help of testing data set which was initially splitted as testData with 30% of main data set. For this, we have to use predict() method on the testing data set. In this, lets mix and match various columns to implement several various decision tree models and compare the prediction results of each.

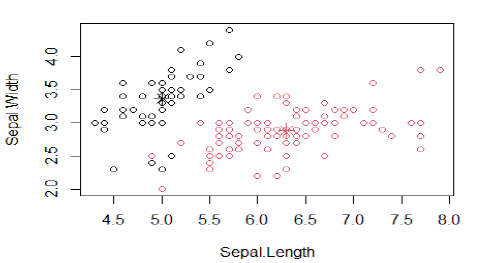




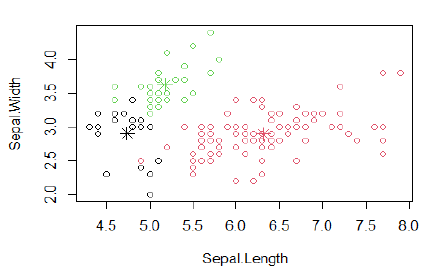
From the plots, we can depict that there is change in each prediction with respect to the columns. If the independent variable is tweaked then the final result will also get altered with respect to the columns

**Exercise 5**

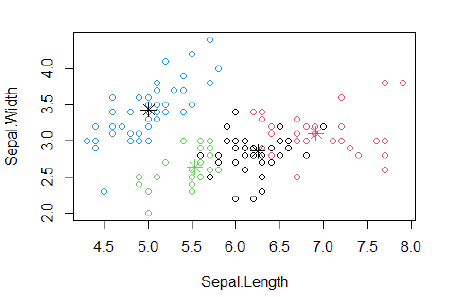
We have set the seed to 9000 and duplicated the data set with irisData2. Then removed the Species column by assigning it to the NULL. Let's perform some K-Means Clustering model also known as K-Means/ K Nearest Neighbors which segregates the the whole data set into various clusters/ groups/ partitions. After this, it further splits the data into alike clusters. groups/ partitions as nearer as available and cluster as apart as possible.



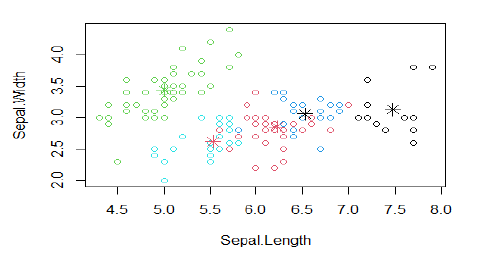
The above plot is for K-means clustering with 2 clusters of sizes 53, 97



The above plot is for K-means clustering with 3 clusters of sizes 21, 96, 33



The above plot is for K-means clustering with 4 clusters of sizes 40, 32, 28, 50



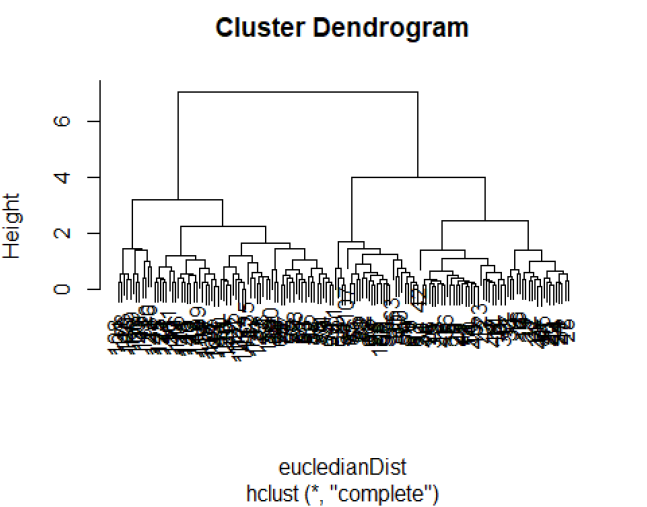
The above plot is for K-means clustering with 5 clusters of sizes 12, 37, 50, 24, 27

We don't know the data sets features and parameters they show. And , also the clustering technique will use numerical data and drop non-numeric columns. After that, we will perform clustering by using the kmeans() method with 2 clusters with our normalized data set. Once it is performed, we again feed with 3 clusters, and then 4, and then 5.

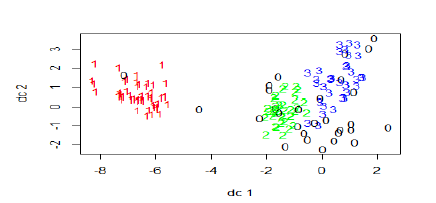
By doing like so, we can observe the change in the distances between them. The Euclidean Distances becomes lesser with increase in the clusters. Here, we can plot to see the clusters.

**Exercise 6**

In this, we perform Hierarchical Clustering and it is not necessary to specify the number of clusters required to get the best result. This is known as a ‘Dendogram’ and it looks like an inverted tree structure.



**Exercise 7**

****

In this, let us remove the Species ID's and use dbscan() method for clustering with the irisData2 data set. After that, let us compare with the original Species ID's in a table and plot the clusters. We can see few 0's in the graphs which are looked as outliers.

**Conclusion**

To perform multiple analytical operations, we use R on a wider range of data like various datasets. R is a very powerful tool to perform analysis which is mainly built by researchers, statisticians, and developers. In the beginning, we did Descriptive Statistics Analysis, Exploratory Data Analysis, and continued with the k means clustering, density-based clustering, and hierarchical clustering Data Mining Techniques to get understandings from the provided data. Plotted relevant parameters, to helps us to understand the data in a clear way. Used factoextra, NbClust, party, caret, and fpc libraries.

**References**

[1] Larry Alton (December 22, 2017) The 7 Most Important Data Mining Techniques was retrieved from <https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques>

[2] Hand, D. J., & Adams, N. M. (2014). Data mining. Wiley StatsRef: Statistics Reference Online, 1-7

[3] Learntek (February 8, 2019) Data Mining Examples and Techniques was retrieved from <https://www.learntek.org/blog/data-mining-examples-and-techniques/>

[4] Hand, D. J. (2007). Principles of data mining. Drug safety, 30(7), 621-622