Dheeraj Goud Borlla (dxb160130)

ASSIGNMENT – 1

The User dataset is a measure of a group of Student's Knowledge status about the subject of Electrical Data Machines. The data set has six features; STG, SCG, STR, LPR and PEG.

1. STG (The study of degree time for goal object materials) (input value)
2. SCG (The degree of repetition number of user for goal object materials) (input value)
3. STR (The degree of study time of user for related objects with the goal object) (input value)
4. LPR (The exam performance of the user for related objects with the goal objects) (input value)
5. PEG (The exam performance of the user for goal objects) (input value)

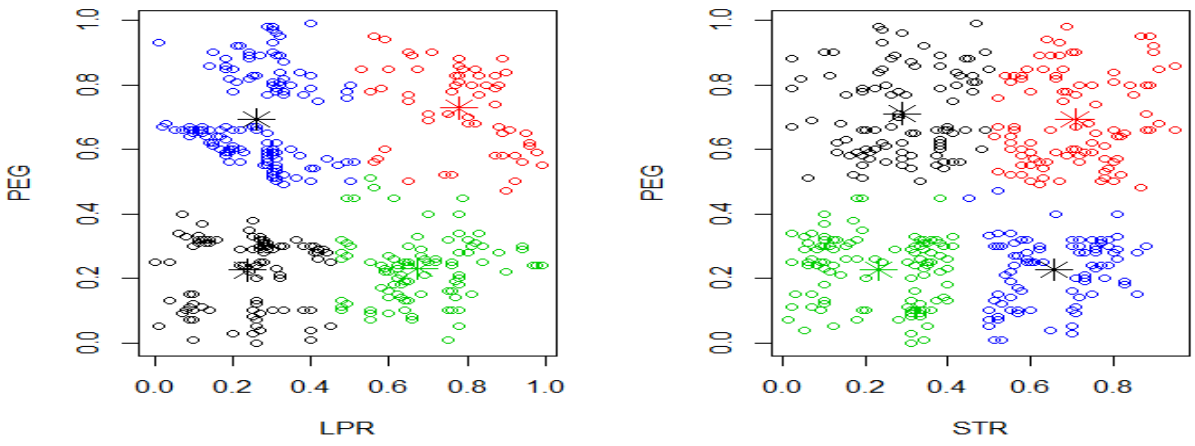
All the features don't seem to be fit for the target value. With this data set, we try to gain understanding of student's behaviours, test performance and their knowledge level about the subject. We will attempt to draw conclusions based on the outcomes obtained from the k-means clustering and hierarchal clustering.

I. K-Means Clustering

When the k-means algorithm is applied to full data set with k=5, the ratio between between_ss and total_ss is about 46.4%. So, we can't conclude that the data set has five clusters. When k is taken a value of 20, then between_ss to the total_ss ratio is 74.6%. The ratio seems to be good, but 20 clusters for the given data set seems big. We will try applying the k-means algorithm to only two features rather than applying to all five of them.

Let us consider two features, LPR ad PEG, and k-means algorithm is applied to those features with k=3,4,5. The result ratio (between_ss / total_ss) is 65.1%, 79.7% and 82.3% respectively. We can see that k=4 is reasonable number of clusters, with four clusters of sizes: 106, 57, 102, 137.

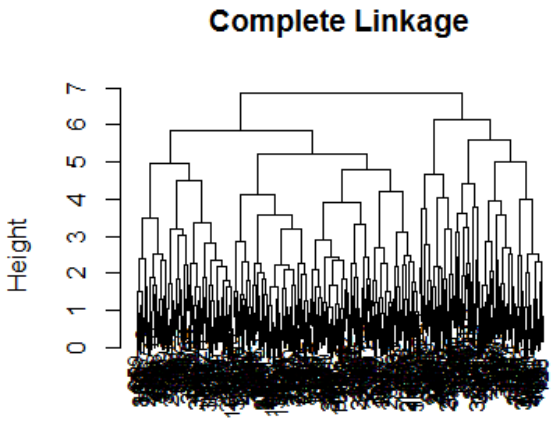
Now we will consider another set of features, STR and PEG, and k-means algorithm is applied to the data with two features. When k is taken values of 3, 4, 5 their respective values for between_ss to the total_ss ratio are 63.2%, 78.4%, 81.7%. Even for this data, four clusters appear to be reasonable for this data set.

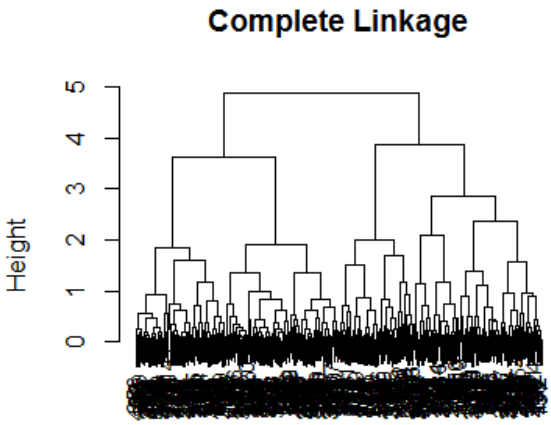


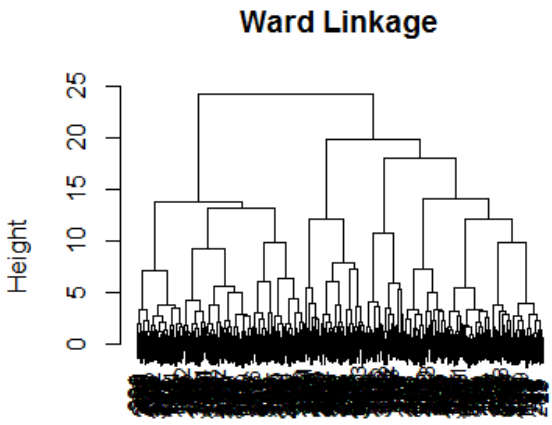
II. Hierarchal clustering

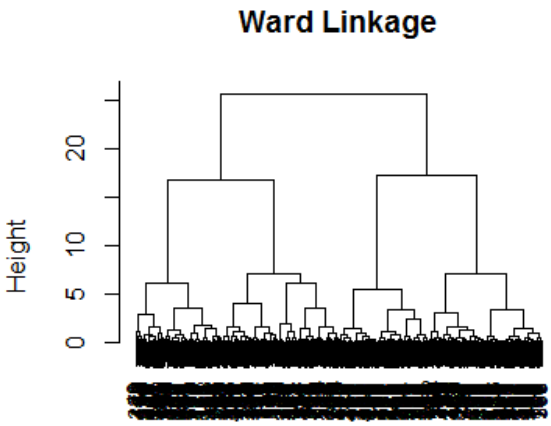
This clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. The results of hierarchal clustering are usually presented in a dendogram.

Now the hierarchal clustering is applied to the complete data set with all features included. The clustering that we obtained in k-means doesn't seems to be reasonable when hierarchal clustering is applied. This happened to be the same case when the k-means is applied to the complete data set. So, we will use fewer variables for clustering the data set.







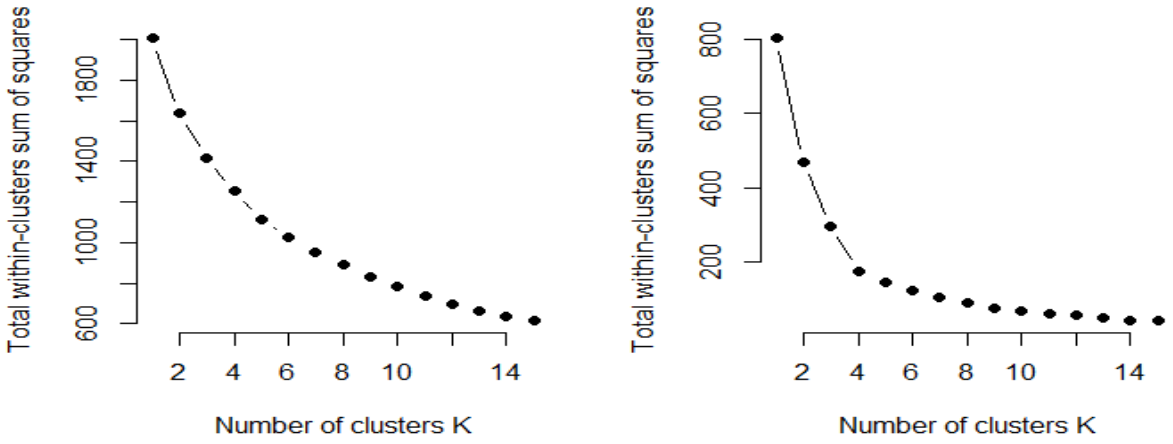


The first dendogram plot in both the figures is obtained when the hierarchal clustering is applied to complete data set. In the second plot, the hierarchal clustering is applied on two variables (STR and PEG). From the plot, we see that if clustering is applied to complete data set, the dendogram seems complicated and becomes hard to divide it into clusters. When the same clustering is applied to two

features of the data set, we can see that from dendogram plot that four clusters seems reasonable for the data.

Elbow Method:

Finally, we verify with elbow method, whether the number of clusters we assumed from k-means and hierarchal clustering seems reasonable or not. The following plots are elbow method applied on complete data set and on only two features from the data set.



From the plots, four clusters seem reasonable for this data.

Conclusion:

From the conclusions made from the K-means clustering method and Hierarchal clustering methods, it seems that the student's knowledge dataset can be divided into four clusters. And these clustering of the data set can be used to build a classification model for the knowledge level of the students. When the clustering methods are applied to the data set with all features included, the results are quite unpredictable, but with few features it was easy to find the clusters within the data. So, the data has four clusters, that is the student's knowledge level on electrical engineering can be divided into four levels, say High, Medium, Below Medium and low.