CS 510 Assignment

Name:	Devansh Maurya
Roll No.:	B16CS024
Date:	April 20, 2020

Assignment Report

This is the report of the assignment and the analysis performed on the data. The codes are attached with the report

Dataset

The dataset selected is the Graduate Admission dataset. The dataset contains the prediction result of admission in foreign universities based on certain attributes. The file consists of 100 records.

Contents

1. Serial No. (unique for every record)
2. GRE Scores (out of 340)
3. TOEFL Scores (out of 120)
4. University Rating (out of 5)
5. Statement of Purpose and Letter of Recommendation Strength (out of 5)
6. Undergraduate GPA (out of 10)
7. Research (either 0 or 1 for no experience or experience)
8. Chance of Admit (ranging from 0 to 1)

For the analysis in this assignment, I have taken all the attributes except:

- **Serial No.:** Not relevant as its unique for every record.
- **Research:** Not being considered as it is a class attribute with classes as 0 and 1.
- **Chance of Admit:** Because it is predicted based on the other attributes.

Clustering Methods

First, the ‘scale’ preprocessing mechanism is applied to the data to normalise all the values.

K-Means Clustering

The app provides the following methods to initialize the initial centroids:

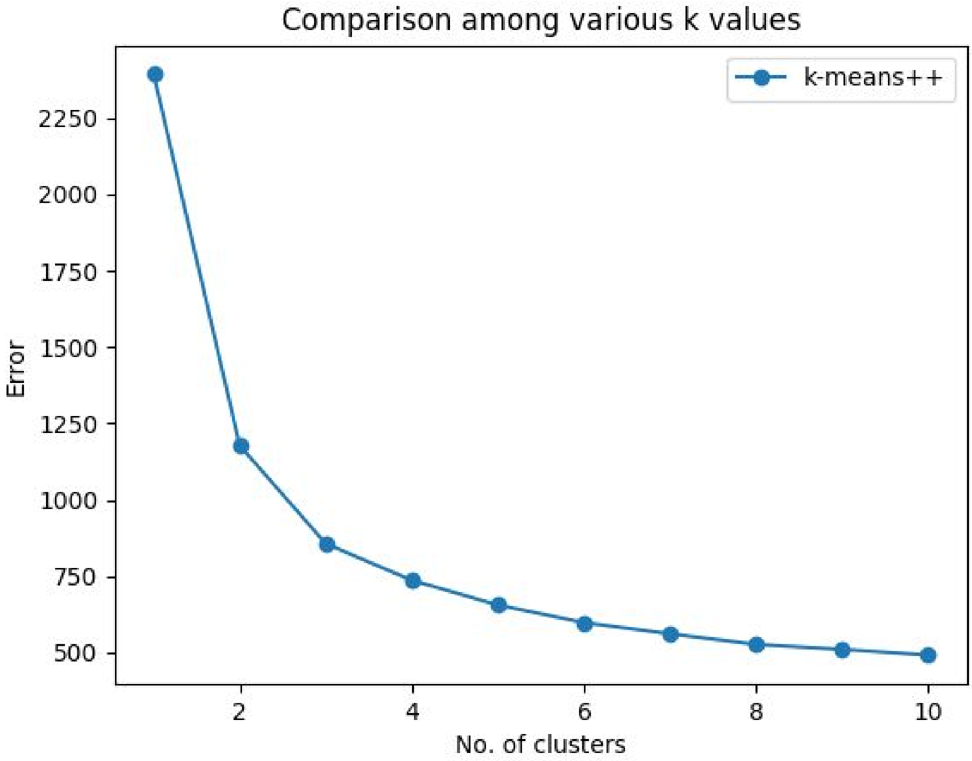
1. K-Means++
2. Random
3. Compare both the methods

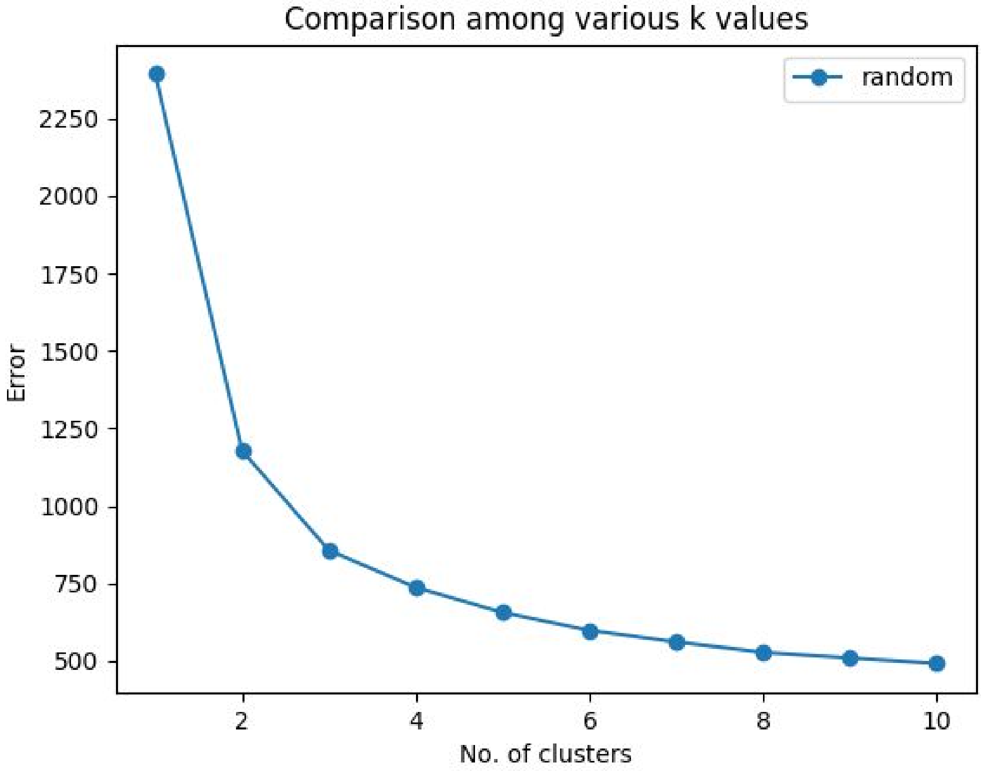
The app provides the following options for providing k values:

1. Enter a value manually
2. Run for all k values from 1 to 10

When the algorithm is run for all k values from 2 to 10, an elbow graph is created. An optimum value of k is then selected from that graph.

The graphs formed by selecting initialisation methods as K-Means and Random are shown below:





The graphs are plotted based on the SSE (Sum of Squares Error) for every k value. By observing the graph, we can see that the elbow shape forms at around k = 3 or 4. So we can select either of these values as the optimal values for the number of clusters in the data.

Agglomerative Hierarchical Clustering

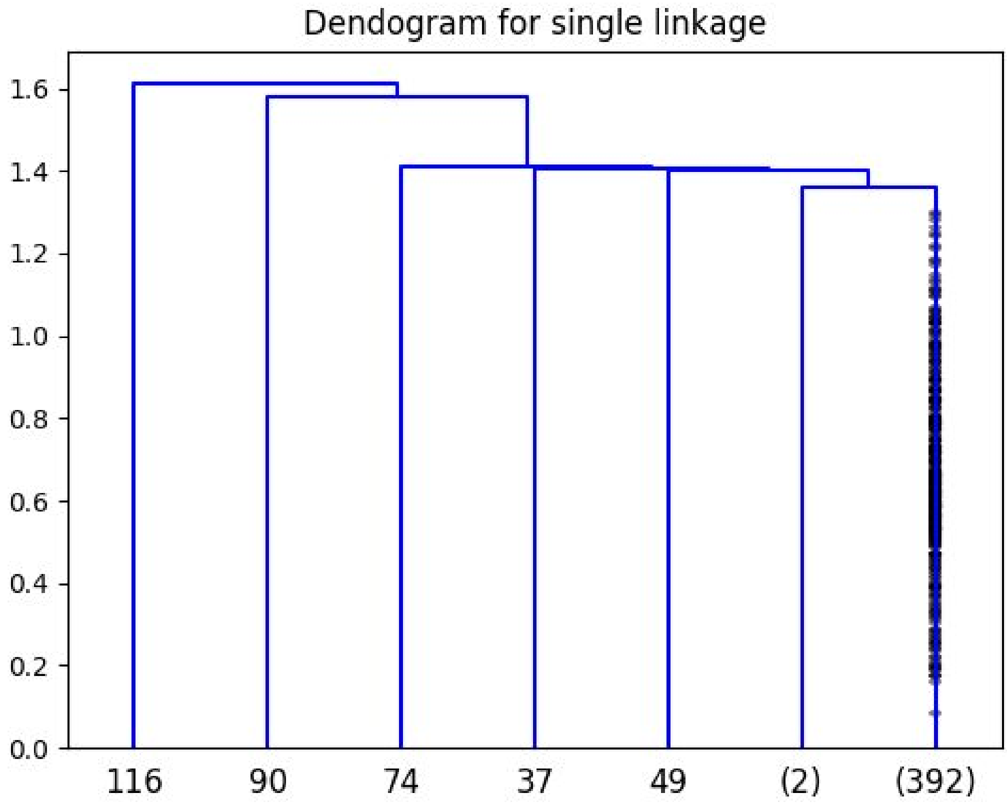
The app provides the following linkage methods:

1. Single linkage
2. Complete linkage
3. Average linkage

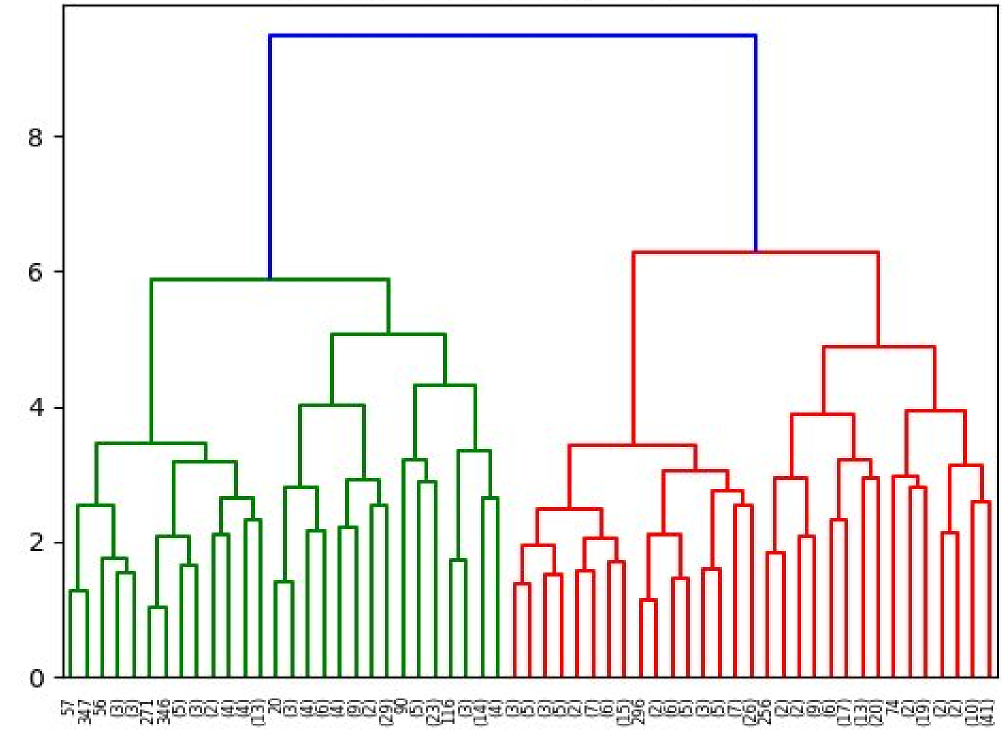
The app provides the following options for providing cluster counts:

3. Enter a value manually
4. Run for all cluster counts from 2 to 10

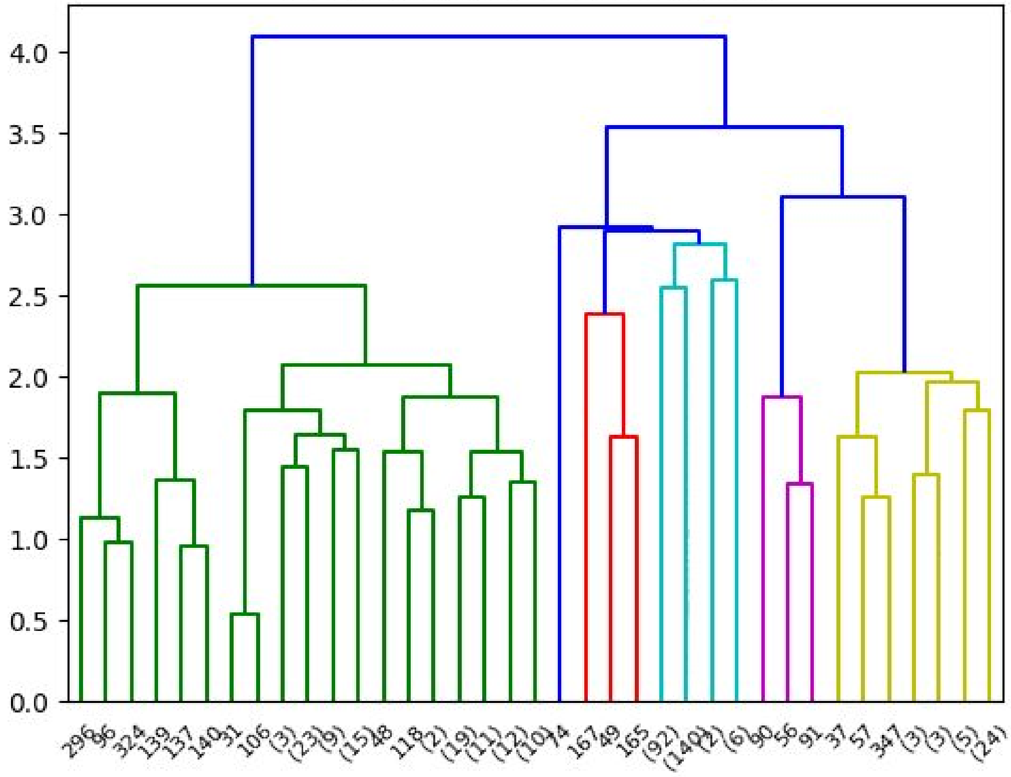
When the algorithm is run for all cluster counts from 2 to 10, an elbow graph is created. An optimum value of k is then selected from that graph.



Dendogram for complete linkage

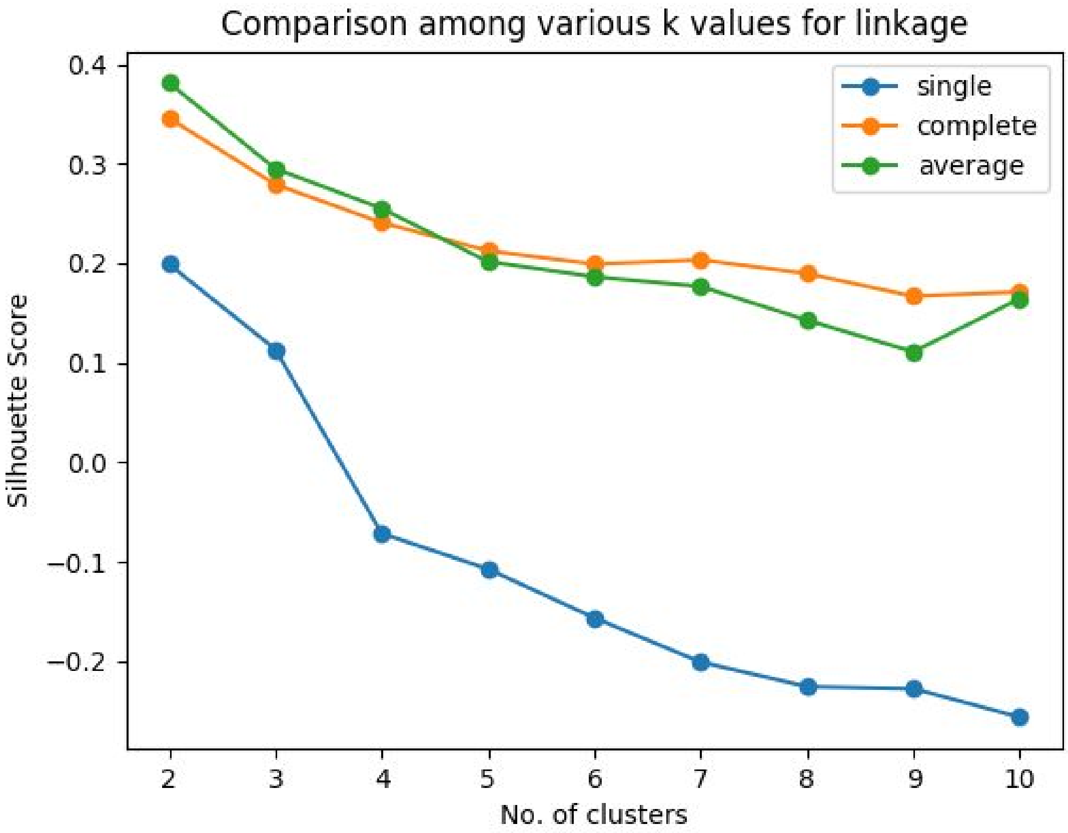


Dendogram for average linkage



All three dendrograms are shown only up to level 5 as with a large number of records, initially, we have far too many clusters which are not much relevant. The dendrograms' vertical lines also show contraction markers because of truncation.

When the algorithm is run for all k values from 2 to 10, for all the above-mentioned linkage methods, the following graph is obtained:



We can see that as the number of clusters increase, the silhouette score shows different trends for different linkage mechanisms.

By observing the dendographs above, we can observe the following k values for different linkage methods:

1. **Single Linkage:** 3
2. **Complete Linkage:** 2
3. **Average Linkage:** 3