Final Project Report

Name: Tanzima Sultana
UTA ID – 1001759430

How to run:

1. Dataset – seeds_dataset.txt
2. On command line run: python ML_Final_Project.py Seed " Clustering_algo" No_of_cluster K
   Ex. python ML_Final_Project.py 1 "single" 3 3
   Where,
   Seed = 1
   Clustering_algo = single
   No_of_cluster = 3
   K = 3
3. All the outputs are saved in – *Output* folder and all scatter plots for cluster saved in – *ScatterPlots* folder.
4. script.sh contains all the commands I have run. To run the script, on command line – sh script.sh

Description:

1. Hierarchical Clustering
   a. Single Linkage Clustering – the two clusters with the smallest *minimum* pairwise distance.
   b. Complete Linkage Clustering - smallest *maximum* pairwise distance.
2. K Nearest Neighbor (KNN) – Choose K neighbors based on Euclidean distance.
3. UCI seeds dataset.
   a. 3 different varieties of wheat: Kama, Rosa and Canadian.
   b. 70 elements in each group. Total = 210.
   c. Class labels are represented as number [1, 2, 3]. Items are categorized sequentially. Kama – 1 (1-70), Rosa – 2 (71-140) and Canadian – 3 (141-210).

Implementation:

1. Split dataset (Train – 190, Test - 20).
2. Hierarchical Clustering - Training
   a. Calculate Euclidean distance between all the nodes of train dataset.
   b. At each step, merge two nodes or clusters with the smallest minimum/maximum pairwise distance. Minimum for Single Linkage and maximum for Complete Linkage clustering.
   c. Repeat these steps until all the nodes are divided into given number of clusters.
3. KNN – Testing
   a. For each test node, calculate Euclidean distance with all the nodes from training phase.
   b. Find K nearest neighbors for each test node.

    c. Then, assign the test node with a cluster id based on majority of its cluster belongs to which cluster.
4. Accuracy Calculation
    a. Compare assigned/predicted cluster label with the ground truth cluster label from dataset.
    b. Calculate accuracy – (correct predicted class label / total test dataset length) * 100.
    c. Compare accuracy with Sklearn library.
5. Scatter plots – generate scatter plots showing different clusters with the test dataset.

Implementation shown with an example:

For seed = 2, clustering algorithm = "single", no of cluster = 4, K = 5.

1. Hierarchical Clustering – Training.

Showing minimum distance at some steps and merged clusters.
Cluster idx = index from array where the clusters are saved.
Cluster id = actual cluster label.

Clustering algorihtm : single , No of cluster : 4
Min value : 1.018160640567096
Merged clusters -
Cluster 1 : [149]
Cluster 2 : [145]
Min value : 1.022232282800734
Merged clusters -
Cluster 1 : [101]
Cluster 2 : [50]
Min value : 1.023376885609598
Merged clusters -
Cluster 1 : [38]
Cluster 2 : [8]
.
.
.
Merged clusters -
Cluster 1 : [175, [98]]
Cluster 2 : [180, [125, [173, [168], [42]]]]
Min value : 1.8826211408565474
Merged clusters -
Cluster 1 : [186]
Cluster 2 : [158, [37]]
Min value : 1.8871721940511939
Merged clusters -
Cluster 1 : [146, [137], [121]]
Cluster 2 : [104]

Min value : 1.9291916986136963
Merged clusters -
Cluster 1 : [112]
Cluster 2 : [21]
Min value : 1.9564056864566708
Merged clusters -
Cluster 1 : [108, [80], [122, [44]]]
Cluster 2 : [175, [98], [180, [125, [173, [168], [42]]]]]
.
.
.
Min value : 3.588217397260093
Merged clusters -
Cluster 1 : [163, [115, [63, [46], [119, [116, [103, [161, [72]]], [86, [133, [84]], [62]], [96, [58, [12]]], [151, [69, [5]]]]]], [162, [57, [29], [124, [9]], [3]]]]
Cluster 2 : [2, [113, [164, [154, [110]]], [13, [17, [95, [97, [114, [76]], [108, [80], [122, [44]], [175, [98], [180, [125, [173, [168], [42]]]]], [20]], [187, [132], [71, [126, [88, [109, [68], [64]]]], [89, [49, [41, [102, [18]]]]]]], [1]]]]]]
Min value : 3.6033251990349138
Merged clusters -
Cluster 1 : [55]
Cluster 2 : [163, [115, [63, [46], [119, [116, [103, [161, [72]]], [86, [133, [84]], [62]], [96, [58, [12]]], [151, [69, [5]]]]]], [162, [57, [29], [124, [9]], [3]]], [2, [113, [164, [154, [110]]], [13, [17, [95, [97, [114, [76]], [108, [80], [122, [44]], [175, [98], [180, [125, [173, [168], [42]]]]], [20]], [187, [132], [71, [126, [88, [109, [68], [64]]]], [89, [49, [41, [102, [18]]]]]]], [1]]]]]]]
Min value : 3.6701302374711444
Merged clusters -
Cluster 1 : [146, [137], [121], [104], [178, [179, [139], [52], [56, [45]]]]]
Cluster 2 : [188, [105, [92], [36, [34], [32, [27]]]], [131, [75, [51], [74, [28, [38, [8]]], [7]]]], [11, [142, [111, [82]], [120, [30]], [22, [6]]], [118, [152, [106], [25, [24]]], [153, [182, [147, [100, [16]]]], [183, [61, [148, [4]]]], [185, [167, [60], [181, [65], [141, [130, [40]]], [170, [171, [39]]]]], [176, [54], [0]]]]]]]
Min value : 3.7843652585341174
Merged clusters -
Cluster 1 : [166]
Cluster 2 : [146, [137], [121], [104], [178, [179, [139], [52], [56, [45]]]], [188, [105, [92], [36, [34], [32, [27]]]], [131, [75, [51], [74, [28, [38, [8]]], [7]]]], [11, [142, [111, [82]], [120, [30]], [22, [6]]], [118, [152, [106], [25, [24]]], [153, [182, [147, [100, [16]]]], [183, [61, [148, [4]]]], [185, [167, [60], [181, [65], [141, [130, [40]]], [170, [171, [39]]]]], [176, [54], [0]]]]]]]
Min value : 3.8537733508861165
Merged clusters -
Cluster 1 : [99, [59], [140, [83, [26]], [107, [127, [90], [73, [155, [135, [77, [53]]]]], [78, [101, [50]]], [94, [189, [184], [186, [158, [37]]]], [157, [87], [112, [21]]]]], [67, [81, [79, [66, [174, [129], [136, [33]]]]], [48, [35, [160, [15]]]]]], [177, [47], [172, [138, [93]], [91], [134, [143, [43], [156, [31]]]], [159, [150, [123]], [117, [85]], [144, [14]]]]]]

Cluster 2 : [165, [149, [145], [19]], [169, [128], [70], [10]]]
Min value : 3.949304622335431
Merged clusters -
Cluster 1 : [99, [59], [140, [83, [26]], [107, [127, [90], [73, [155, [135, [77, [53]]]]], [78, [101, [50]]], [94, [189, [184], [186, [158, [37]]]], [157, [87], [112, [21]]]]], [67, [81, [79, [66, [174, [129], [136, [33]]]]], [48, [35, [160, [15]]]]]], [177, [47], [172, [138, [93]], [91], [134, [143, [43], [156, [31]]]], [159, [150, [123]], [117, [85]], [144, [14]]]]], [165, [149, [145], [19]], [169, [128], [70], [10]]]]
Cluster 2 : [55, [163, [115, [63, [46], [119, [116, [103, [161, [72]]], [86, [133, [84]], [62]], [96, [58, [12]]], [151, [69, [5]]]]]], [162, [57, [29], [124, [9]], [3]]], [2, [113, [164, [154, [110]]], [13, [17, [95, [97, [114, [76]], [108, [80], [122, [44]], [175, [98], [180, [125, [173, [168], [42]]]], [20]], [187, [132], [71, [126, [88, [109, [68], [64]]]], [89, [49, [41, [102, [18]]]]]]], [1]]]]]]]]

---------- Clusters : Train Data ----------

Cluster : 1 , Cluster id : 1.0 , size : 62
[44. 41. 25. 10. 13. 23. 35. 24. 29. 28. 6. 12. 18. 5. 2. 3. 14. 53.
54. 65. 9. 66. 45. 64. 48. 42. 57. 11. 8. 21. 1. 36. 62. 0. 27. 20.
30. 17. 59. 55. 16. 61. 68. 56. 19. 60. 34. 40. 26. 32. 39. 49. 50. 52.
46. 69. 67. 58. 33. 38. 51. 4.]

Cluster : 2 , Cluster id : 3.0 , size : 63
[200. 207. 140. 161. 143. 158. 169. 188. 182. 201. 141. 183. 197. 192.
146. 180. 179. 142. 157. 156. 147. 173. 175. 160. 178. 149. 185. 189.
153. 155. 176. 184. 202. 159. 187. 198. 181. 151. 190. 174. 206. 209.
186. 166. 152. 164. 191. 144. 194. 172. 171. 154. 163. 205. 150. 145.
165. 208. 196. 195. 199. 193. 170.]

Cluster : 3 , Cluster id : 2.0 , size : 64
[135. 71. 118. 139. 112. 130. 84. 126. 119. 106. 85. 89. 115. 91.
98. 114. 122. 120. 134. 101. 99. 90. 127. 128. 131. 78. 109. 94.
79. 82. 108. 113. 92. 100. 111. 77. 74. 93. 133. 123. 87. 97.
125. 81. 129. 86. 70. 137. 138. 76. 96. 83. 110. 105. 73. 117.
136. 132. 80. 121. 88. 103. 107. 102.]

Cluster : 4 , Cluster id : 3.0 , size : 1
[203.]

Accuracy : Train Data
Accuracy : 100.0 %

2. <u>KNN – Testing.</u>

Showing K nearest neighbors list for each test item and selecting majority cluster as cluster index.
Then retrieve the original cluster label as cluster id.

Nearest Neighbor Algorithm : 5
Node : 167.0 , nn_list : [208. 205. 198. 192. 194.]
node : 208.0 , cluster : 1
node : 205.0 , cluster : 1
node : 198.0 , cluster : 1
node : 192.0 , cluster : 1
node : 194.0 , cluster : 1
Majority cluster idx : 1
Node : 37.0 , nn_list : [79. 36. 136.   8. 74.]
node : 79.0 , cluster : 2
node : 36.0 , cluster : 0
node : 136.0 , cluster : 2
node : 8.0 , cluster : 0
node : 74.0 , cluster : 2
Majority cluster idx : 2
Node : 116.0 , nn_list : [125. 102. 126. 118. 111.]
node : 125.0 , cluster : 2
node : 102.0 , cluster : 2
node : 126.0 , cluster : 2
node : 118.0 , cluster : 2
node : 111.0 , cluster : 2
Majority cluster idx : 2
Node : 124.0 , nn_list : [135. 36. 44. 139. 138.]
node : 135.0 , cluster : 2
node : 36.0 , cluster : 0
node : 44.0 , cluster : 0
node : 139.0 , cluster : 2
node : 138.0 , cluster : 2
Majority cluster idx : 2
Node : 148.0 , nn_list : [198. 160. 69. 26. 199.]
node : 198.0 , cluster : 1
node : 160.0 , cluster : 1
node : 69.0 , cluster : 0
node : 26.0 , cluster : 0
node : 199.0 , cluster : 1
Majority cluster idx : 1
Node : 31.0 , nn_list : [135. 44. 38. 138.   6.]
node : 135.0 , cluster : 2
node : 44.0 , cluster : 0
node : 38.0 , cluster : 0
node : 138.0 , cluster : 2
node : 6.0 , cluster : 0
Majority cluster idx : 0
Node : 63.0 , nn_list : [197. 19. 29. 163. 12.]

node : 197.0 , cluster : 1
node : 19.0 , cluster : 0
node : 29.0 , cluster : 0
node : 163.0 , cluster : 1
node : 12.0 , cluster : 0
Majority cluster idx : 0
Node : 47.0 , nn_list : [44. 38. 48. 56. 6.]
node : 44.0 , cluster : 0
node : 38.0 , cluster : 0
node : 48.0 , cluster : 0
node : 56.0 , cluster : 0
node : 6.0 , cluster : 0
Majority cluster idx : 0
Node : 204.0 , nn_list : [194. 19. 205. 147. 163.]
node : 194.0 , cluster : 1
node : 19.0 , cluster : 0
node : 205.0 , cluster : 1
node : 147.0 , cluster : 1
node : 163.0 , cluster : 1
Majority cluster idx : 1
Node : 95.0 , nn_list : [74. 136. 76. 107. 139.]
node : 74.0 , cluster : 2
node : 136.0 , cluster : 2
node : 76.0 , cluster : 2
node : 107.0 , cluster : 2
node : 139.0 , cluster : 2
Majority cluster idx : 2
Node : 177.0 , nn_list : [193. 175. 189. 187. 174.]
node : 193.0 , cluster : 1
node : 175.0 , cluster : 1
node : 189.0 , cluster : 1
node : 187.0 , cluster : 1
node : 174.0 , cluster : 1
Majority cluster idx : 1
Node : 162.0 , nn_list : [182. 150. 157. 186. 181.]
node : 182.0 , cluster : 1
node : 150.0 , cluster : 1
node : 157.0 , cluster : 1
node : 186.0 , cluster : 1
node : 181.0 , cluster : 1
Majority cluster idx : 1
Node : 7.0 , nn_list : [28. 2. 21. 5. 56.]
node : 28.0 , cluster : 0
node : 2.0 , cluster : 0

node : 21.0 , cluster : 0
node : 5.0 , cluster : 0
node : 56.0 , cluster : 0
Majority cluster idx : 0
Node : 104.0 , nn_list : [92. 103. 117. 118. 91.]
node : 92.0 , cluster : 2
node : 103.0 , cluster : 2
node : 117.0 , cluster : 2
node : 118.0 , cluster : 2
node : 91.0 , cluster : 2
Majority cluster idx : 2
Node : 75.0 , nn_list : [71. 80. 100. 122. 74.]
node : 71.0 , cluster : 2
node : 80.0 , cluster : 2
node : 100.0 , cluster : 2
node : 122.0 , cluster : 2
node : 74.0 , cluster : 2
Majority cluster idx : 2
Node : 43.0 , nn_list : [132. 10. 134. 133. 122.]
node : 132.0 , cluster : 2
node : 10.0 , cluster : 0
node : 134.0 , cluster : 2
node : 133.0 , cluster : 2
node : 122.0 , cluster : 2
Majority cluster idx : 2
Node : 22.0 , nn_list : [25. 4. 46. 17. 1.]
node : 25.0 , cluster : 0
node : 4.0 , cluster : 0
node : 46.0 , cluster : 0
node : 17.0 , cluster : 0
node : 1.0 , cluster : 0
Majority cluster idx : 0
Node : 72.0 , nn_list : [71. 70. 74. 76. 107.]
node : 71.0 , cluster : 2
node : 70.0 , cluster : 2
node : 74.0 , cluster : 2
node : 76.0 , cluster : 2
node : 107.0 , cluster : 2
Majority cluster idx : 2
Node : 15.0 , nn_list : [50. 12. 6. 52. 32.]
node : 50.0 , cluster : 0
node : 12.0 , cluster : 0
node : 6.0 , cluster : 0
node : 52.0 , cluster : 0

node : 32.0 , cluster : 0
Majority cluster idx : 0
Node : 168.0 , nn_list : [172. 206. 154. 185. 191.]
node : 172.0 , cluster : 1
node : 206.0 , cluster : 1
node : 154.0 , cluster : 1
node : 185.0 , cluster : 1
node : 191.0 , cluster : 1
Majority cluster idx : 1

---------- Clusters : Test Data ----------

Cluster : 1 , Cluster id : 1.0 , size : 6
[31.0, 63.0, 47.0, 7.0, 22.0, 15.0]

Cluster : 2 , Cluster id : 3.0 , size : 6
[167.0, 148.0, 204.0, 177.0, 162.0, 168.0]

Cluster : 3 , Cluster id : 2.0 , size : 8
[37.0, 116.0, 124.0, 95.0, 104.0, 75.0, 43.0, 72.0]

Accuracy : Test Data
Accuracy : 90.0 %

Scatter Plots:

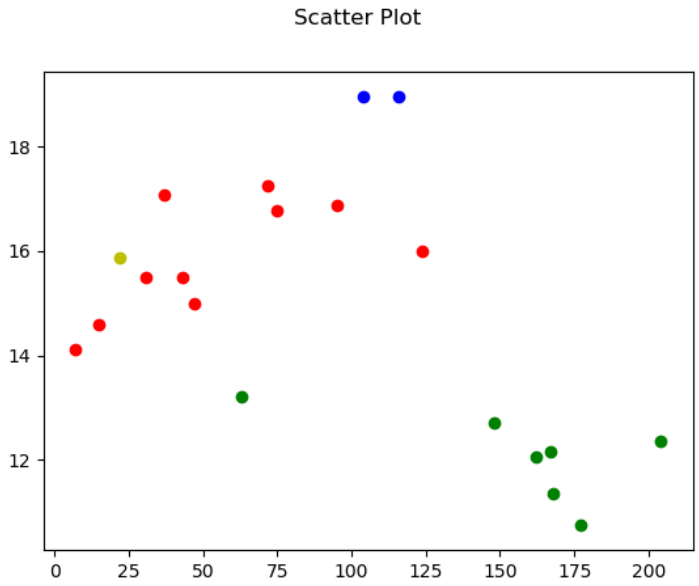

Fig 1: Scatter plots for seed = 2, clustering algorithm = “single”, no of cluster = 4, K = 5

Results:

Seed = [1, 2, 3]
Clustering algorithm – "single" and "complete"
No of clusters = [3, 4, 5]
K for KNN algorithm = [3, 5]

Table 1 and 2 shows train and test accuracy (%) for Single and Complete Linkage clustering respectively.
Table 3 and 4 shows accuracy (%) on test data of my implemented clustering algorithm and Sklearn library of the same clustering algorithm.
Table 5 shows accuracy comparison between Single and Complete Linkage clustering.

Table 1: Accuracy (%) for Single Linkage Clustering

Seed	Single Linkage clustering						KNN
	No of clusters						
	3		4		5		
	Train	Test	Train	Test	Train	Test	
1	100	80	100	85	100	85	3
	100	80	100	85	100	85	5
2	66.84	55	100	85	100	85	3
	66.84	60	100	90	100	90	5
3	96.84	85	96.84	85	96.84	85	3
	96.84	85	96.84	85	96.84	85	5

Table2: Accuracy (%) for Complete Linkage Clustering

Seed	Complete Linkage Clustering						KNN
	No of clusters						
	3		4		5		
	Train	Test	Train	Test	Train	Test	
1	94.73	70	94.73	70	94.73	70	3
	94.73	75	94.73	80	94.73	80	5
2	76.31	70	81.57	85	92.63	85	3
	76.31	80	81.57	90	92.63	90	5
3	96.84	80	96.84	85	96.84	85	3
	96.84	85	96.84	85	96.84	85	5

Table 3: Accuracy (%) for Single Linkage and Sklearn Clustering

Seed	No of clusters			KNN
	3	4	5	

	Clustering	Sklearn	Clustering	Sklearn	Clustering	Sklearn	
1	80	85	85	90	85	95	3
	80	85	85	90	85	95	5
2	55	50	85	75	85	75	3
	60	50	90	75	90	75	5
3	85	80	85	80	85	90	3
	85	80	85	80	85	90	5

Table 4: Accuracy (%) for Complete Linkage and Sklearn

Seed	No of clusters						KNN
	3		4		5		
	Clustering	Sklearn	Clustering	Sklearn	Clustering	Sklearn	
1	70	85	70	85	70	85	3
	75	85	80	85	80	85	5
2	70	85	85	85	85	85	3
	80	85	90	85	90	85	5
3	80	90	85	90	85	90	3
	85	90	85	90	85	90	5

Table 5: Accuracy (%) comparison of Single and Complete Linkage clustering

Seed	No of clusters						KNN
	3		4		5		
	Single	Complete	Single	Complete	Single	Complete	
1	80	70	85	70	85	70	3
	80	75	85	80	85	80	5
2	55	70	85	85	85	85	3
	60	80	90	90	90	90	5
3	85	80	85	85	85	85	3
	85	85	85	85	85	85	5

Scatter Plots:

Some example scatter plots.

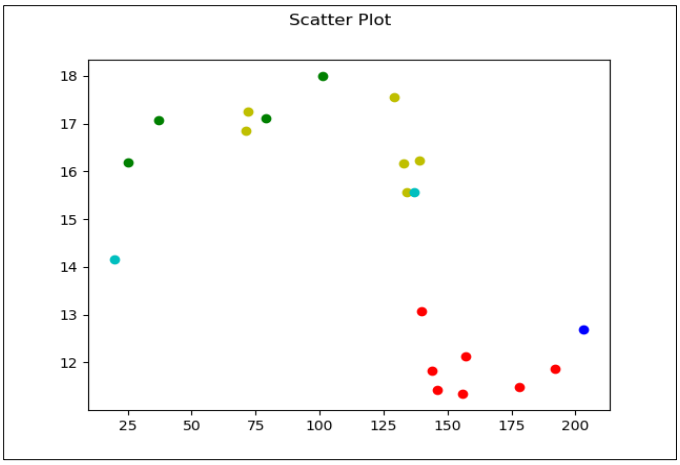

Fig 2: Complete Linkage, cluster = 5, K = 3

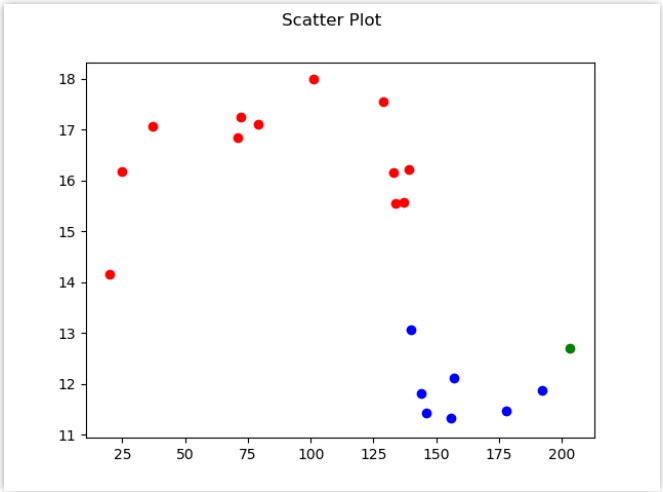

Fig 3: Single Linkage, cluster = 3, K = 3

Dendrogram:

On Test data. Using Sklearn.

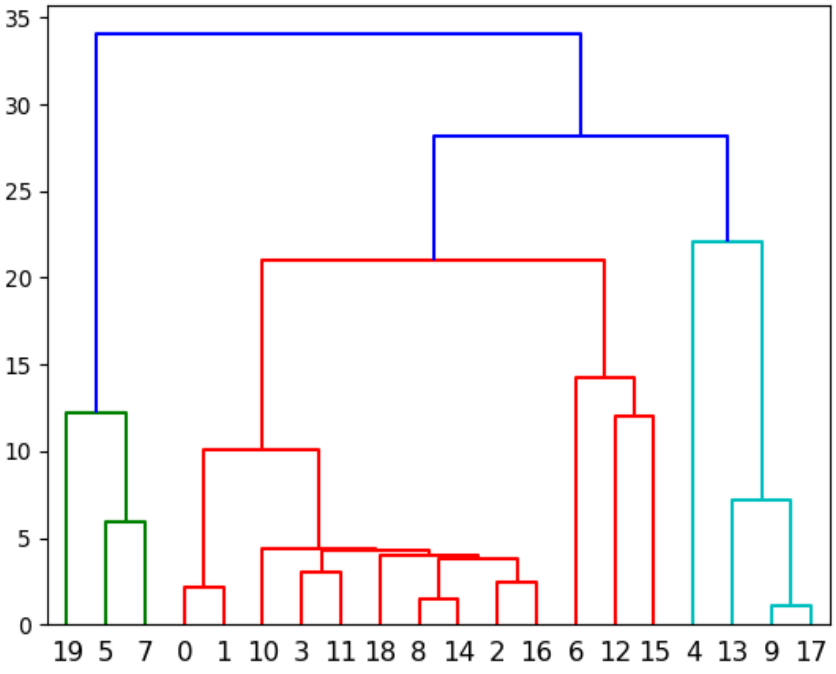

Fig 4: Single Linkage clustering

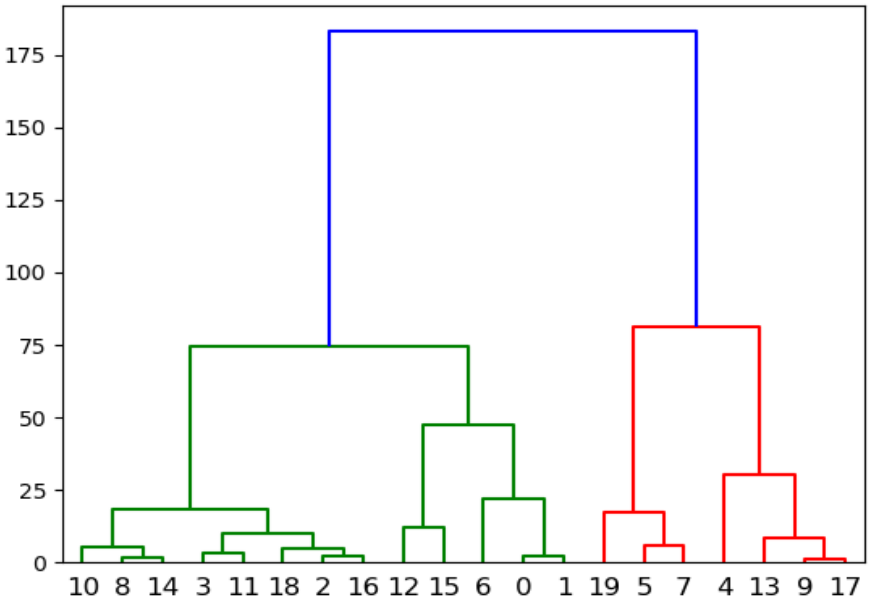

Fig 5: Complete Linkage clustering

Conclusion:

1. For different no of seed, different train-test accuracy.
2. Most of the cases, increasing number of clusters, increases accuracy.
3. For seed = [2], cluster = [4, 5], K = [5], train accuracy = 100%, test = 90%.
4. Most of the cases, both single and complete performs same. Only in 2 cases, complete performs better.
5. Sklearn performs better.