In this homework, we are tasked to analyze the olive data using all the clustering methods discussed in the class. Later we would compare the results of this analysis with classification results we acquired in the previous homework as we are not working on supervised classification, the whole dataset used for clustering area and region. Normalization applied on the data to avoid different scalings for variables. However, the data was not centered during the normalization as it would change the clustering analysis. Daisy function has been used to find the distance between the objects while normalizing it by the range of variables. Manhattan distance has been used to compute the distance between different objects. In the following section, hierarchical clustering and Partitioning Around Medoids (PAM) methods have been used.

Hierarchical Clustering (HC):

Using Agnes method, single, complete and average HC methods have been applied. Dendrogram plots of these methods are shown in Figure 1-3.

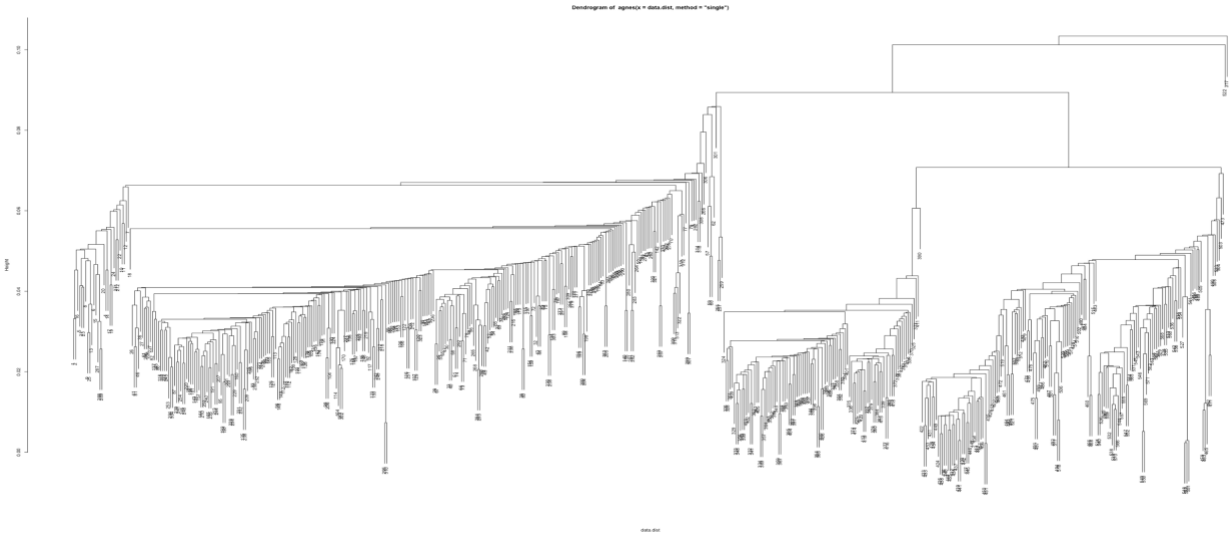

Figure 1:Single HC method Dendrogram tree

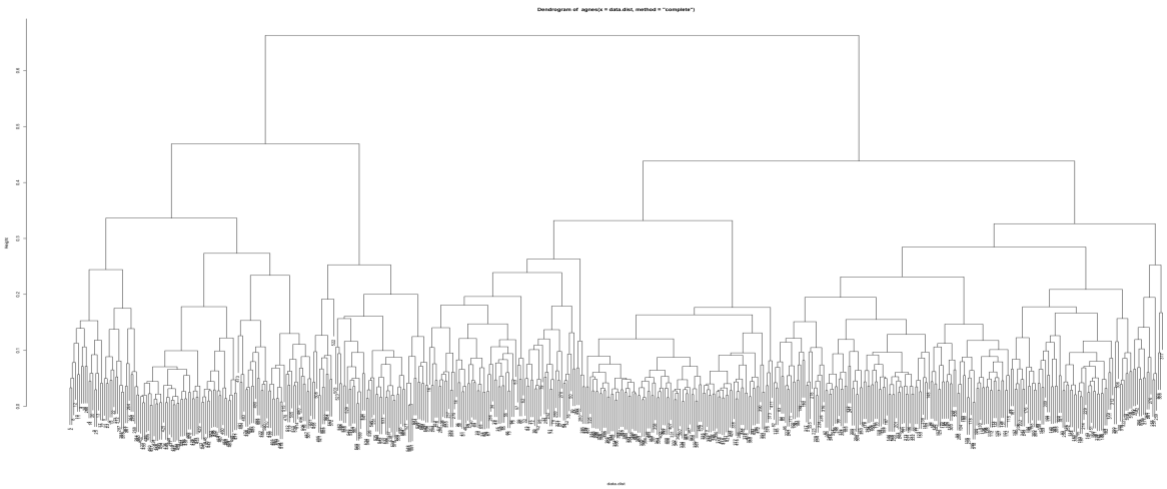

Figure 2:Complete HC method Dendrogram tree

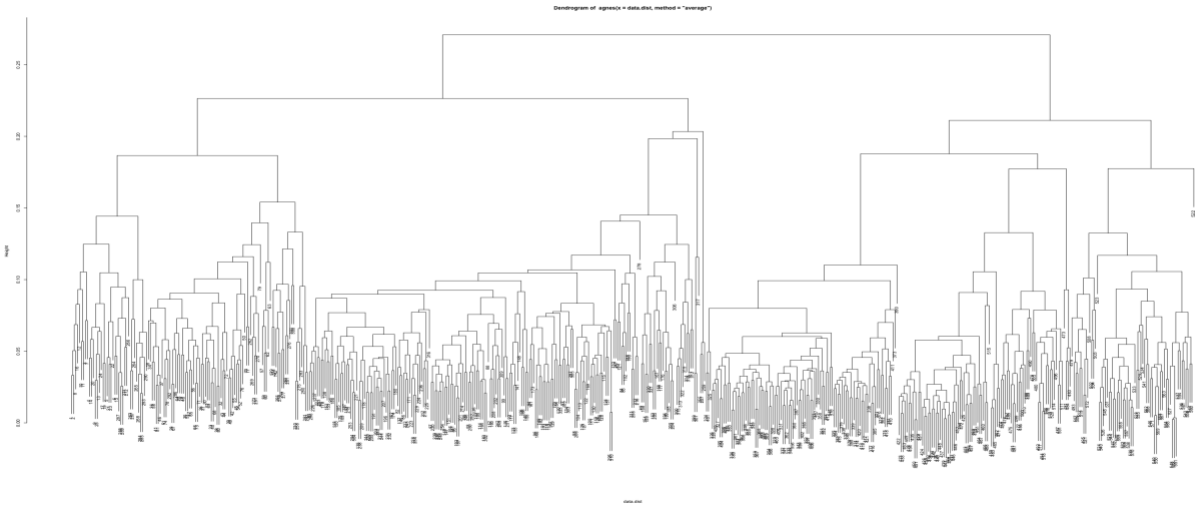

Figure 3: Average HC method dendrogram tree

It is evident in the figures that average and complete methods led to a more balanced tree. The performance of these HC methods for the various number of clusters has been assessed by computing the average silhouette width, and it is summarized in Table 1. According to Table 1, 6, 4 and 2 clusters will lead to the best results for complete, average and single HC method respectively. The silhouette value, ranges from -1 to 1, is a measure of how similar an object is to its cluster (cohesion) compared to other groups (separation). Tables to show how well these methods performed on recognizing different features are in Appendix I. These tables confirm that these methods are not capable of reproducing the same results supervised method acquired before. Moreover, Silhouette plots for k equal to 6,4 and 2 shown in Figure 4-5.

Number of Clusters	Complete	Average	Single
2	0.37	0.368	0.30
3	0.36	0.36	0.10
4	0.32	0.374	0.21
5	0.33	0.34	0.17
6	0.4	0.33	0.09
7	0.37	0.367	0.002
8	0.33	0.3709	-0.009
9	0.32	0.36	-0.02
10	0.31	0.36	0.049

Number of Cluster	PAM	Kmeans
2	0.38	0.35
3	0.29	0.24
4	0.39	0.15
5	0.41	0.18
6	0.40	0.22
7	0.32	0.21
8	0.32	0.17
9	0.30	0.19
10	0.29	0.19

Table 1(left): Hierarchical Clustering methods performance. Table 1(right) PAM and Kmeans performance through different Ks

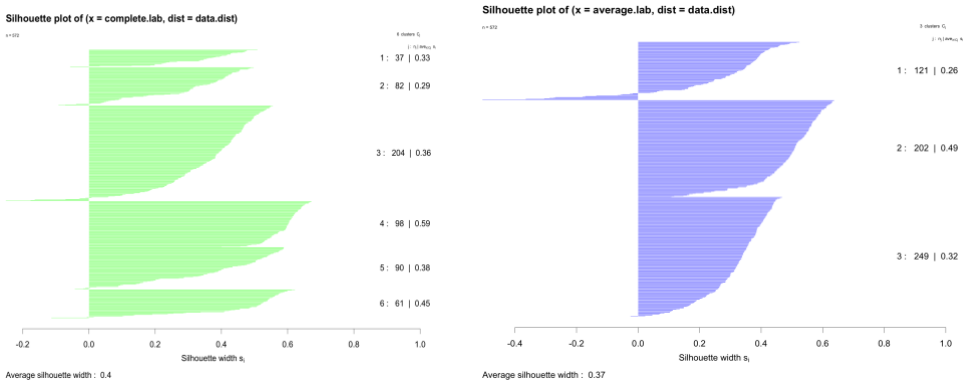

Figure 4: Silhouette plots for k=6 using complete method (Green), for k=4 using average method (blue)

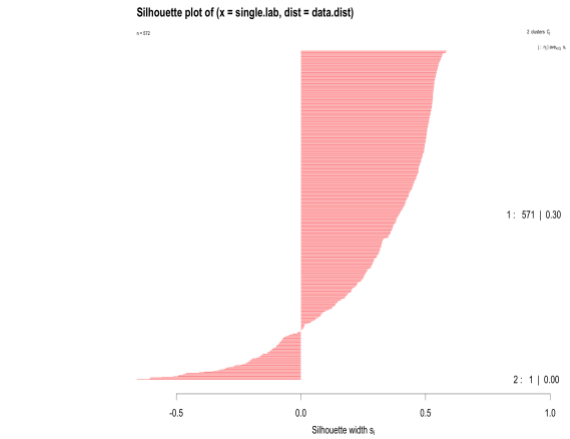

Figure 5: Silhouette of Single HC method

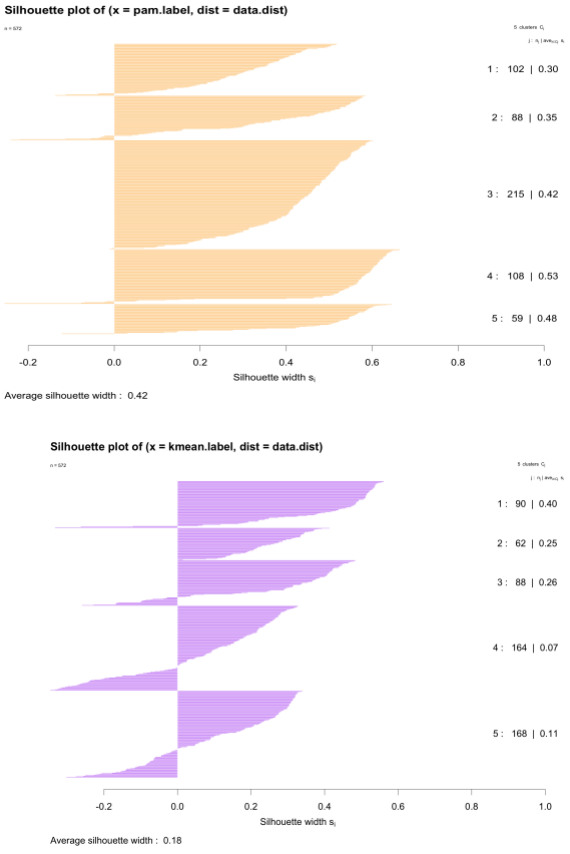

Figure 6: Silhouette plot for PAM(upper one) and K-means(lower one)

The silhouettes plot of the single method shows that there is not a strong cluster structure within this data as all the silhouette scores are less than 0.3. Average and complete methods have some clusters with a score close to 0.5 which are considered as some weak structure. Therefore, we could conclude from these results that hierarchical clustering methods are not able to build the strong structure we need to cluster the instances accurately.

PAM and Kmeans:

PAM is a more efficient version of K-means in which cluster representatives are medoids instead of random mean centers. This slight difference lets PAM become more robust to the noise and outliers. Like HC methods, the performance of PAM and K-means have been assessed through a different number of clusters shown in Table 2. The silhouette plot for PAM has been shown in Figure6. According to Figure 7, PAM gives the best silhouette width with K= 5. Moreover, PAM shows a stronger structure in comparison to K-means according to its silhouettes plot. However, we could not conclude that PAM is capable of finding the strong clustering structure we would need in the data as its silhouette score in the best-case scenario is low.

Clustering Analysis:

Finally, we used the information within each clustering method to form the confusion matrices and analyze how well areas are clustered using these techniques. All the confusion matrices are provided in the Appendix I and we would only discuss PAM results in this section as it outperforms other methods. The confusion matrix for regions vs number of clusters in pam K-means for K=5 has been shown in table 3. We can see that clusters 2 and 5 has been mostly captured data for Centre.North area. Cluster 4 totally captured data for Saridiana area and clusters 1 and 3 captures most of the southern area. Same has been applied for pam K-means with K=5 to show the confusion matrix for region this time (Table4). We can say that areas 1, 3 and 8 have been mostly represented by cluster 1. Region 2 has been captured almost fully by cluster 3. Regions 6 and 7 have been capture 100% by cluster number 4 as region 9 by cluster 2 and Region 5 by cluster 5. Region 4 has been captured mostly but less than others by cluster number 2. It can be seen that it is harder to recover the information about the regions than the areas. Regions within Saradiana area are in one cluster which shows PAM method could understand the underlying patterns for region within that area. However, in other two areas at least one of the regions are in different cluster which confirms that clustering in those areas were not completely successful. In conclusion, we were able to preserve the underlying patterns for regions within same area better in classification methods in comparison to clustering ones.

Appendix I

PAM clustering results for regions and areas are shown in Table 3. (explained in the report)

Cluster/Area	1	2	3	4	5
Centre. North	1	81	0	10	59
Sardinia	0	0	0	98	0
South	101	7	215	0	0

Cluster/Region	1	2	3	4	5
Apulia.north	20	5	0	0	0
Apulia.south	2	0	204	0	0
Calabria	53	0	3	0	0
Liguria.east	1	30	0	9	10
Liguria.west	0	0	0	1	49
Sardiana.coast	0	0	0	33	0
Sardiana.inland	0	0	0	65	0
Sicily	26	2	8	0	0
Umbria	0	51	0	0	0

Kmeans clustering results for regions and areas are shown in Table 4. (Only northern part, marked by blue is clustered completely correct)

Cluster/Area	1	2	3	4	5
Centre. North	17	0	8	0	126
Sardinia	70	0	0	28	0
South	3	62	80	136	42

Cluster/Region	1	2	3	4	5
Apulia.north	0	0	1	0	24
Apulia.south	2	61	15	128	0
Calabria	0	0	51	1	4
Liguria.east	0	0	7	0	43
Liguria.west	17	0	1	0	32
Sardiana.coast	5	0	0	28	0
Sardiana.inland	65	0	0	0	0
Sicily	1	1	13	7	14
Umbria	0	0	0	0	51

Complete HC methods results for regions and areas are shown in Table 5. (Only regions in Sardiana area are classified correctly)

Cluster/Area	1	2	3	4	5	6
Centre. North	0	0	0	0	90	61
Sardinia	0	0	0	98	0	0
South	37	82	204	0	0	0

Cluster/Region	1	2	3	4	5	6
Apulia.north	23	2	0	0	0	0
Apulia.south	0	10	196	0	0	0

Calabria	0	56	0	0	0	0
Liguria.east	0	0	0	0	39	11
Liguria.west	0	0	0	0	0	50
Sardiana.coast	0	0	0	33	0	0
Sardiana.inland	0	0	0	65	0	0
Sicily	14	14	8	0	0	0
Umbria	0	0	0	0	51	0

Average HC methods results for regions and areas are shown in Table 6. (Regions in northern and Saridiana areas are classified correctly)

Cluster/Area	1	2	3
Centre. North	0	0	151
Sardinia	0	0	98
South	121	202	0

Cluster/Region	1	2	3
Apulia.north	25	0	0
Apulia.south	7	199	0
Calabria	55	1	0
Liguria.east	0	0	50
Liguria.west	0	0	50
Sardiana.coast	0	0	33
Sardiana.inland	0	0	65
Sicily	34	2	0
Umbria	0	0	51

Single HC methods results for regions and areas are shown in Table 7. (All the regions almost shown by first cluster which confirms the weak results single method produced)

Cluster/Area	1	2
Centre. North	151	0
Sardinia	98	0
South	322	1

Cluster/Region	1	2
Apulia.north	25	0
Apulia.south	205	1
Calabria	56	0
Liguria.east	50	0
Liguria.west	50	0
Sardiana.coast	33	0
Sardiana.inland	65	0
Sicily	36	0
Umbria	51	0