Name: Rohit Kulkarni

USC ID: 5402749044

1.b.i)

Hamming loss: The Hamming loss is the fraction of labels that are incorrectly predicted. In multiclass classification, the Hamming loss correspond to the Hamming distance between y_true and y_pred which is equivalent to the subset zero_one_loss function. The Hamming loss is upperbounded by the subset zero-one loss. When normalized over samples, the Hamming loss is always between 0 and 1. In multilabel classification, the Hamming loss is different from the subset zero-one loss. The zero-one loss considers the entire set of labels for a given sample incorrect if it does entirely match the true set of labels. Hamming loss is more forgiving in that it penalizes the individual labels.

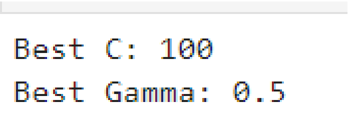
Exact match: In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y_true.

1.b.ii) SVM for each of the labels, using Gaussian kernels and one versus all classifiers.

Below details are for raw attributes and features

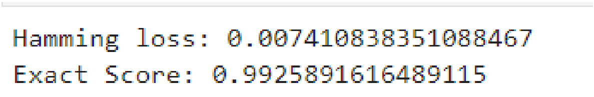
For label Family:

Best C and Gamma:



```
Best C: 100
Best Gamma: 0.5
```

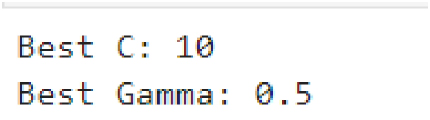
Hamming Loss and Exact Score:



```
Hamming loss: 0.007410838351088467
Exact Score: 0.9925891616489115
```

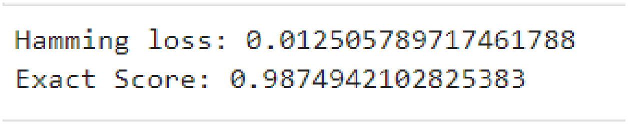
For label Genus:

Best C and Gamma:



```
Best C: 10
Best Gamma: 0.5
```

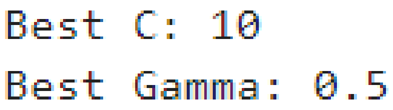
Hamming Loss and Exact Score:



```
Hamming loss: 0.012505789717461788
Exact Score: 0.9874942102825383
```

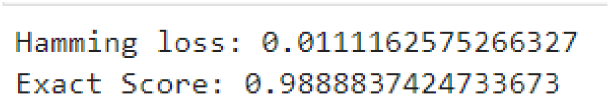
For label Species:

Best C and Gamma:



```
Best C: 10
Best Gamma: 0.5
```

Hamming Loss and Exact Score:

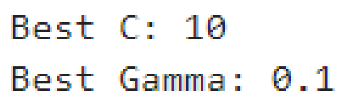


```
Hamming loss: 0.0111162575266327
Exact Score: 0.9888837424733673
```

Now for standardized data and features

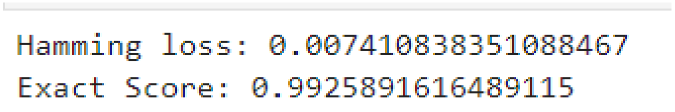
For label Family:

Best C and Gamma:



```
Best C: 10
Best Gamma: 0.1
```

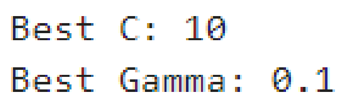
Hamming Loss and Exact Score:



```
Hamming loss: 0.007410838351088467
Exact Score: 0.9925891616489115
```

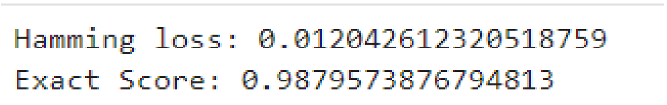
For label Genus:

Best C and Gamma:



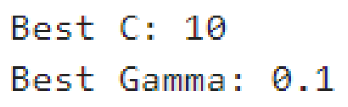
```
Best C: 10
Best Gamma: 0.1
```

Hamming Loss and Exact Score:



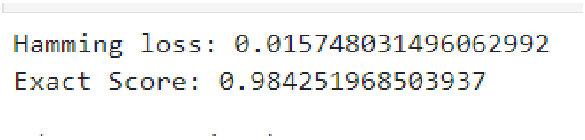
```
Hamming loss: 0.012042612320518759
Exact Score: 0.9879573876794813
```

For label Species:



```
Best C: 10
Best Gamma: 0.1
```

Hamming Loss and Exact Score:

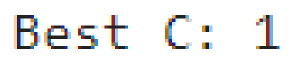


```
Hamming loss: 0.015748031496062992
Exact Score: 0.984251968503937
```

1.b.iii) L1 penalized SVM

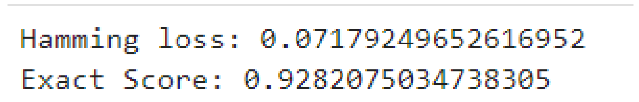
For label Family:

Best C:



```
Best C: 1
```

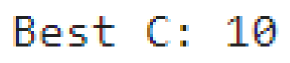
Hamming Loss and Exact Score:



```
Hamming loss: 0.07179249652616952
Exact Score: 0.9282075034738305
```

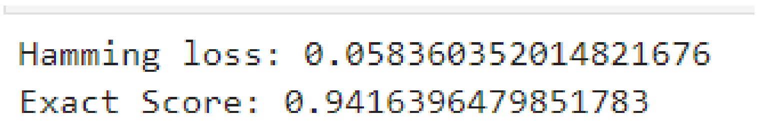
For label Genus:

Best C:



```
Best C: 10
```

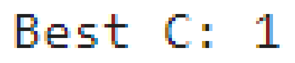
Hamming Loss and Exact Score:



```
Hamming loss: 0.058360352014821676
Exact Score: 0.9416396479851783
```

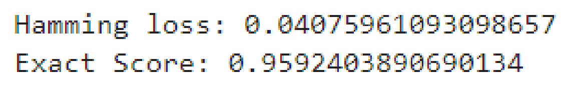
For label Species:

Best C:



```
Best C: 1
```

Hamming Loss and Exact Score:

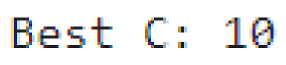


```
Hamming loss: 0.04075961093098657
Exact Score: 0.9592403890690134
```

1.b.iv) SMOTE and L1 penalized SVM

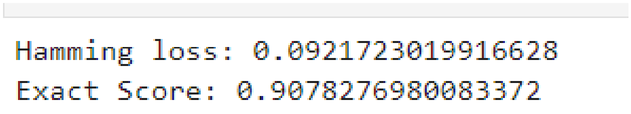
For label Family:

Best C:



```
Best C: 10
```

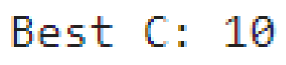
Hamming Loss and Exact Score:



```
Hamming loss: 0.0921723019916628
Exact Score: 0.9078276980083372
```

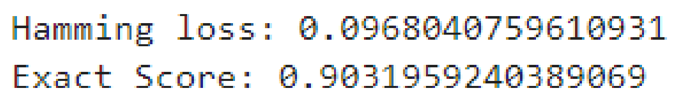
For label Genus:

Best C:



```
Best C: 10
```

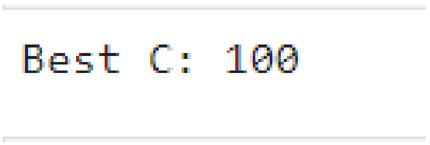
Hamming Loss and Exact Score:



```
Hamming loss: 0.0968040759610931
Exact Score: 0.9031959240389069
```

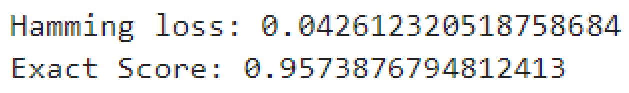
For label Species:

Best C:



```
Best C: 100
```

Hamming Loss and Exact Score:

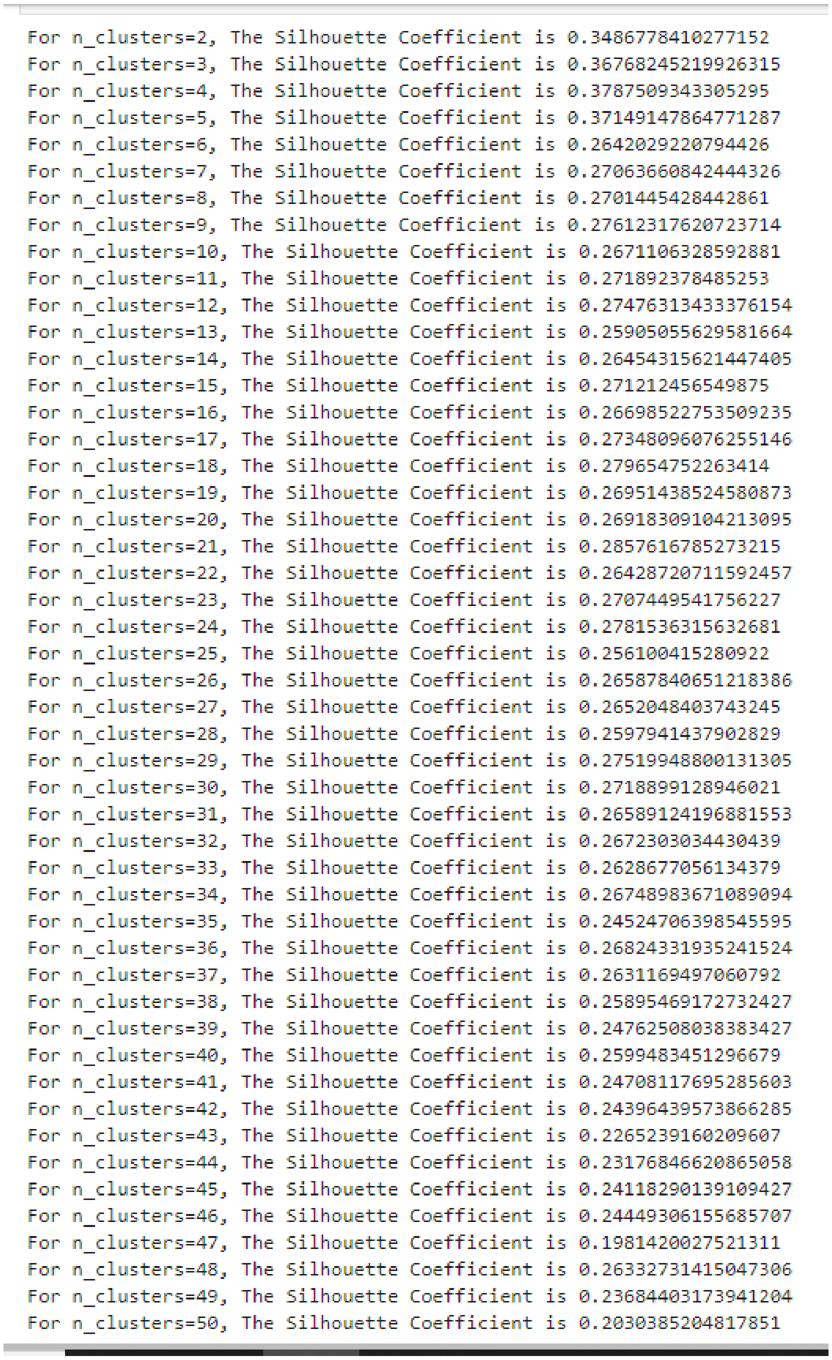


```
Hamming loss: 0.042612320518758684
Exact Score: 0.9573876794812413
```

Exact score is better for each label in SVM when compared to L1 and SMOTE methods, which means that the accuracy is better as well.

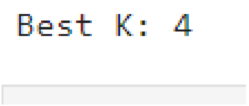
2).K-Means Clustering on a Multi-Class and Multi-Label Data Set

2.a) K-Means clustering with k={1,2,3.....50}



```
For n_clusters=2, The Silhouette Coefficient is 0.3486778410277152
For n_clusters=3, The Silhouette Coefficient is 0.36768245219926315
For n_clusters=4, The Silhouette Coefficient is 0.3787509343305295
For n_clusters=5, The Silhouette Coefficient is 0.37149147864771287
For n_clusters=6, The Silhouette Coefficient is 0.264202922079442б
For n_clusters=7, The Silhouette Coefficient is 0.27063660842444326
For n_clusters=8, The Silhouette Coefficient is 0.2701445428442861
For n_clusters=9, The Silhouette Coefficient is 0.27612317620723714
For n_clusters=10, The Silhouette Coefficient is 0.2671106328592881
For n_clusters=11, The Silhouette Coefficient is 0.271892378485253
For n_clusters=12, The Silhouette Coefficient is 0.27476313433376154
For n_clusters=13, The Silhouette Coefficient is 0.25905055629581664
For n_clusters=14, The Silhouette Coefficient is 0.26454315621447405
For n_clusters=15, The Silhouette Coefficient is 0.271212456549875
For n_clusters=16, The Silhouette Coefficient is 0.26698522753509235
For n_clusters=17, The Silhouette Coefficient is 0.27348096076255146
For n_clusters=18, The Silhouette Coefficient is 0.279654752263414
For n_clusters=19, The Silhouette Coefficient is 0.26951438524580873
For n_clusters=20, The Silhouette Coefficient is 0.26918309104213095
For n_clusters=21, The Silhouette Coefficient is 0.2857616785273215
For n_clusters=22, The Silhouette Coefficient is 0.26428720711592457
For n_clusters=23, The Silhouette Coefficient is 0.2707449541756227
For n_clusters=24, The Silhouette Coefficient is 0.2781536315632681
For n_clusters=25, The Silhouette Coefficient is 0.256100415280922
For n_clusters=26, The Silhouette Coefficient is 0.26587840651218386
For n_clusters=27, The Silhouette Coefficient is 0.2652048403743245
For n_clusters=28, The Silhouette Coefficient is 0.2597941437902829
For n_clusters=29, The Silhouette Coefficient is 0.27519948800131305
For n_clusters=30, The Silhouette Coefficient is 0.2718899128946021
For n_clusters=31, The Silhouette Coefficient is 0.26589124196881553
For n_clusters=32, The Silhouette Coefficient is 0.2672303034430439
For n_clusters=33, The Silhouette Coefficient is 0.2628677056134379
For n_clusters=34, The Silhouette Coefficient is 0.26748983671089094
For n_clusters=35, The Silhouette Coefficient is 0.24524706398545595
For n_clusters=36, The Silhouette Coefficient is 0.26824331935241524
For n_clusters=37, The Silhouette Coefficient is 0.2631169497060792
For n_clusters=38, The Silhouette Coefficient is 0.25895469172732427
For n_clusters=39, The Silhouette Coefficient is 0.24762508038383427
For n_clusters=40, The Silhouette Coefficient is 0.2599483451296679
For n_clusters=41, The Silhouette Coefficient is 0.24708117695285603
For n_clusters=42, The Silhouette Coefficient is 0.24396439573866285
For n_clusters=43, The Silhouette Coefficient is 0.2265239160209607
For n_clusters=44, The Silhouette Coefficient is 0.23176846620865058
For n_clusters=45, The Silhouette Coefficient is 0.24118290139109427
For n_clusters=46, The Silhouette Coefficient is 0.24449306155685707
For n_clusters=47, The Silhouette Coefficient is 0.1981420027521311
For n_clusters=48, The Silhouette Coefficient is 0.26332731415047306
For n_clusters=49, The Silhouette Coefficient is 0.23684403173941204
For n_clusters=50, The Silhouette Coefficient is 0.2030385204817851
```

Best K:

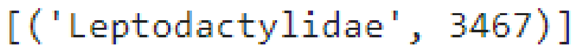


```
Best K: 4
```

2.b) Majority for each label

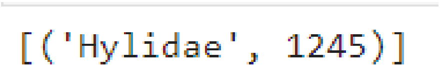
Majority for label Family:

1st cluster:



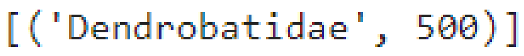
```
[('Leptodactylidae', 3467)]
```

2nd cluster:



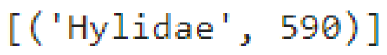
```
[('Hylidae', 1245)]
```

3rd cluster:



```
[('Dendrobatidae', 500)]
```

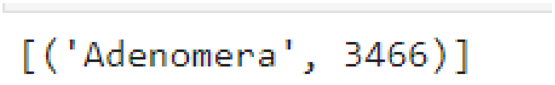
4th cluster:



```
[('Hylidae', 590)]
```

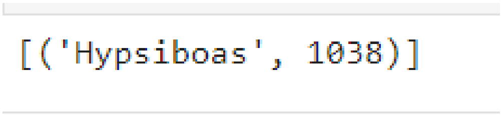
Majority for label Genus:

1st cluster:



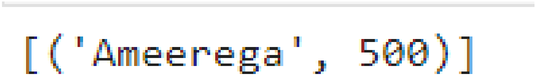
```
[('Adenomera', 3466)]
```

2nd cluster:



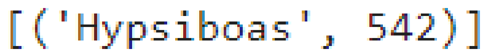
```
[('Hypsiboas', 1038)]
```

3rd cluster:



```
[('Ameerega', 500)]
```

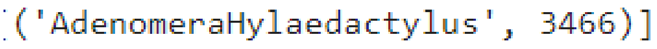
4th cluster:



```
[('Hypsiboas', 542)]
```

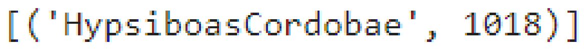
Majority for label Species:

1st cluster:



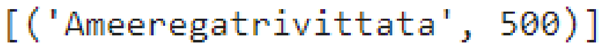
```
('AdenomeraHylaedactylus', 3466)]
```

2nd cluster:



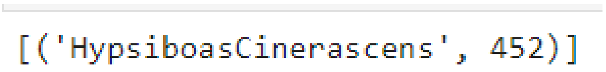
```
[('HypsiboasCordobae', 1018)]
```

3rd cluster:



```
[('Ameeregatrivittata', 500)]
```

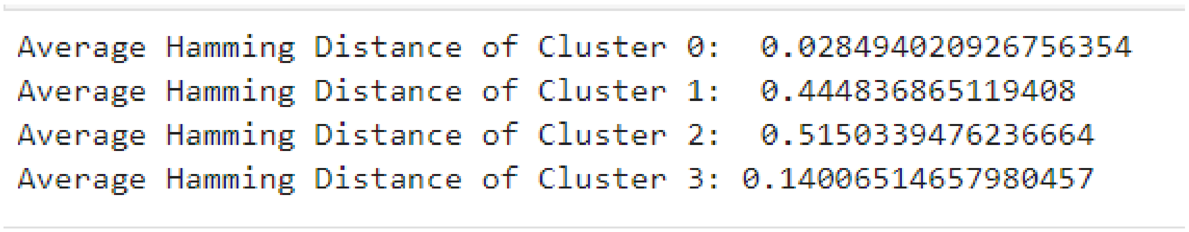
4th cluster:



```
[('HypsiboasCinerascens', 452)]
```

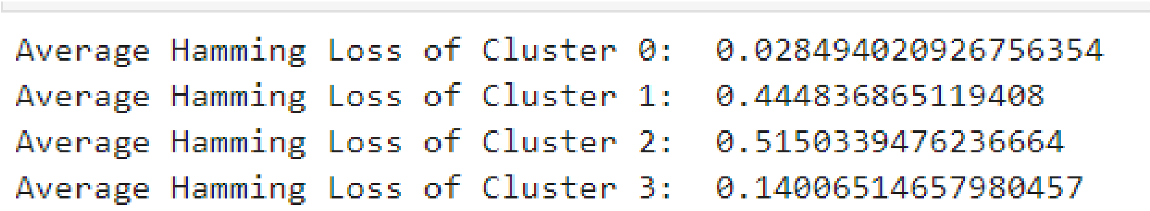
2.c) Hamming distance, hamming loss and hamming score for each cluster.

Hamming Distance:



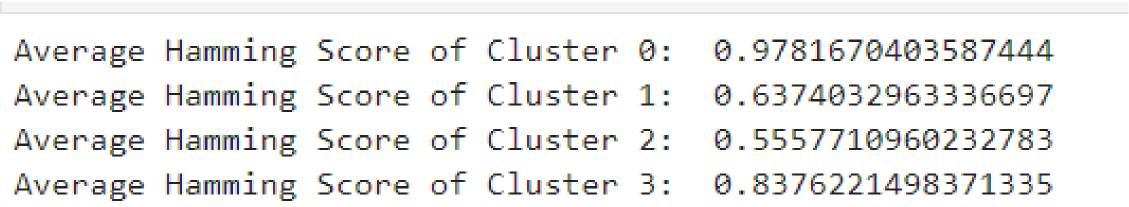
```
Average Hamming Distance of Cluster 0:  0.028494020926756354
Average Hamming Distance of Cluster 1:  0.444836865119408
Average Hamming Distance of Cluster 2:  0.5150339476236664
Average Hamming Distance of Cluster 3: 0.14006514657980457
```

Hamming Loss:



```
Average Hamming Loss of Cluster 0:  0.028494020926756354
Average Hamming Loss of Cluster 1:  0.444836865119408
Average Hamming Loss of Cluster 2:  0.5150339476236664
Average Hamming Loss of Cluster 3:  0.14006514657980457
```

Hamming Score:



```
Average Hamming Score of Cluster 0:  0.9781670403587444
Average Hamming Score of Cluster 1:  0.6374032963336697
Average Hamming Score of Cluster 2:  0.5557710960232783
Average Hamming Score of Cluster 3:  0.8376221498371335
```

10.7.2): Suppose that we have four observations, for which we compute a dissimilarity matrix, given by

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}.$$

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

(a) On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

Ans:

10.7.2)

a) We will use Algorithm 10.2 to explain the different steps that lead to dendogram

Step 1:- We have

$$\begin{pmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{pmatrix}.$$

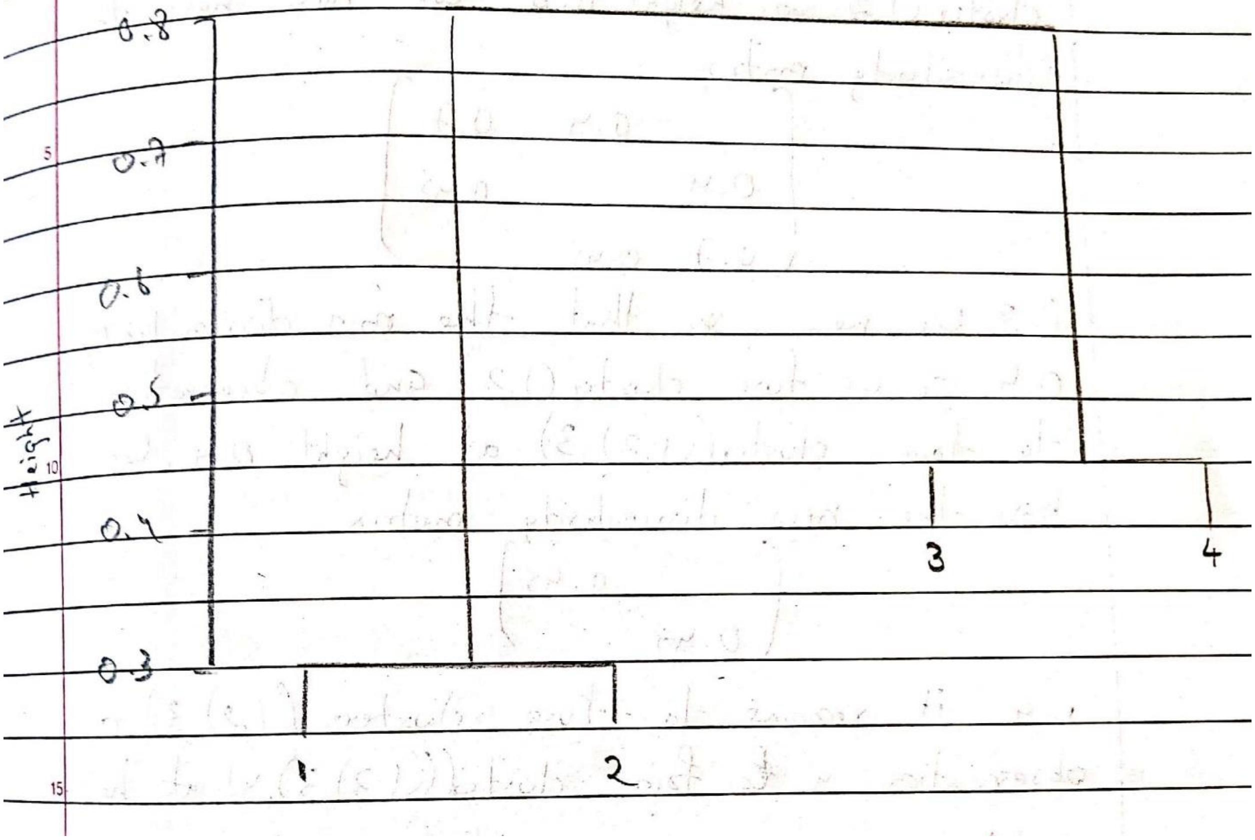
Step 2:- i=4 We may see that 0.3 is the min dissimilarity, so we fuse observations 1 & 2 to form cluster (1,2) at height 0.3. We now have the new dissimilarity matrix

$$\begin{pmatrix} & 0.5 & 0.8 \\ 0.5 & & 0.45 \\ 0.8 & 0.45 & \end{pmatrix}$$

i=3 We now see that the min dissimilarity is 0.45, so we fuse observations 3 & 4 to form cluster (3,4) at height 0.45. We now have the new dissimilarity matrix

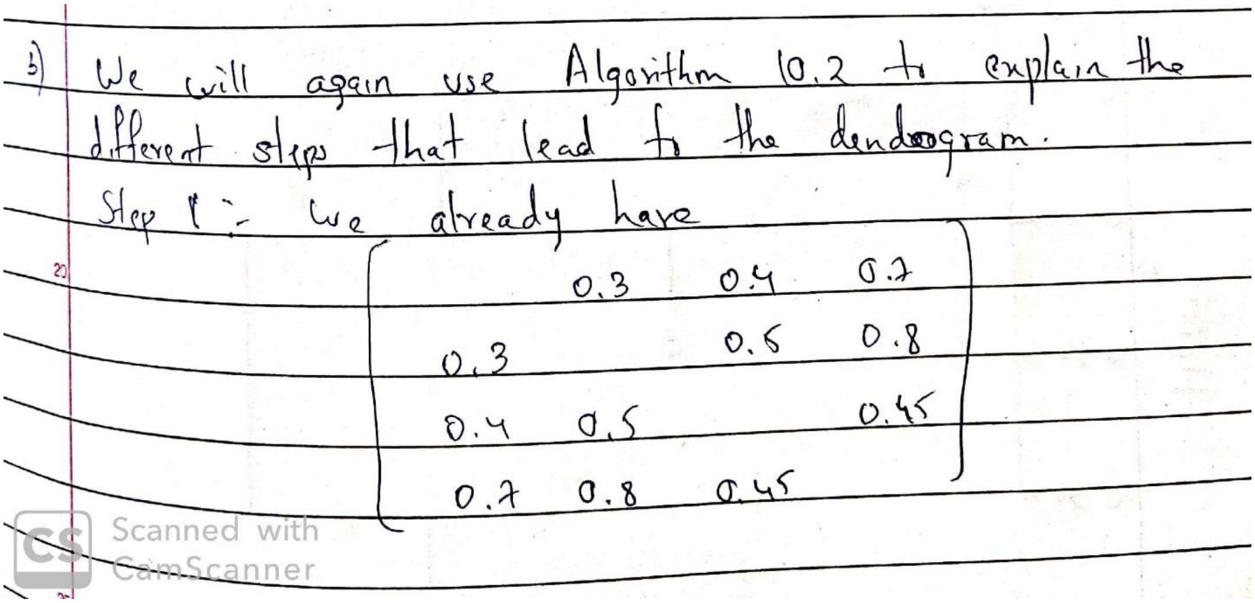
$$\begin{pmatrix} & 0.8 \\ 0.8 & \end{pmatrix}$$

i=4 it remains to fuse clusters (1,2) & (3,4) to form cluster ((1,2)(3,4)) at height 0.8.



(b) Repeat (a), this time using single linkage clustering.

Ans:



b) We will again use Algorithm 10.2 to explain the different steps that lead to the dendrogram.

Step 1 :- We already have

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

(c) Suppose that we cut the dendogram obtained in (a) such that two clusters result. Which observations are in each cluster?

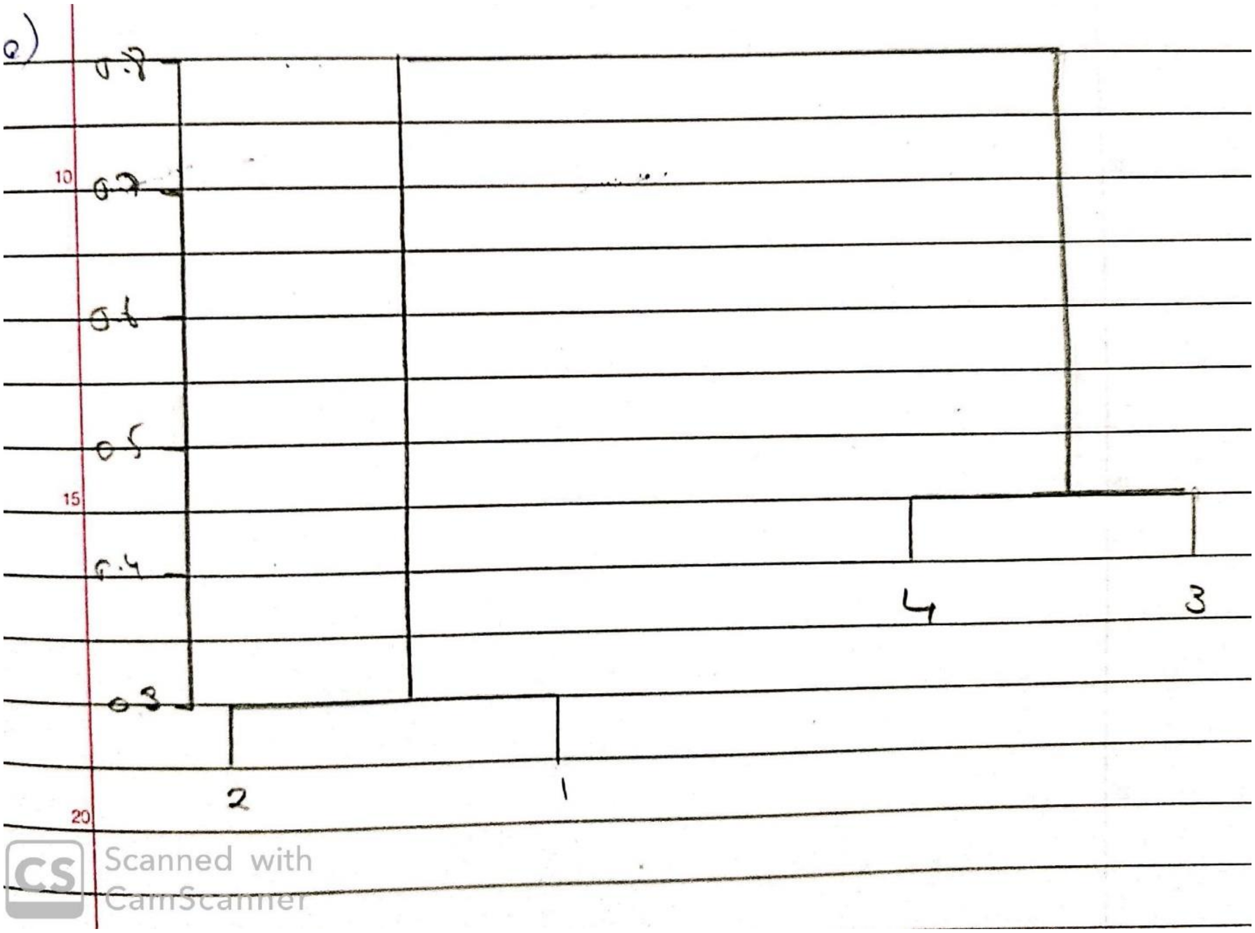
Ans: In this case, we have clusters (1,2) and (3,4).

(d) Suppose that we cut the dendogram obtained in (b) such that two clusters result. Which observations are in each cluster?

Ans: In this case, we have clusters ((1,2),3) and (4).

(e) It is mentioned in the chapter that at each fusion in the dendrogram, the position of the two clusters being fused can be swapped without changing the meaning of the dendrogram. Draw a dendrogram that is equivalent to the dendrogram in (a), for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same.

Ans:


e)
0.8
0.7
0.6
0.5
0.4
0.3
2
1
4
3