Report 3: Clustering

Chengze Liu 2316609

Task 1

Test purity function with examples.

Example 1:

```
y_pred = np.array([2, 2, 1, 2, 2, 2, 0, 0, 0, 1, 2, 1, 1, 1, 1, 1])
y_true = np.array([0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2])
```

```
Purity calculated by hand: (4+3+5)/16=0.75
Purity by compute_purity(): 0.75
```

Example 2:

```
y_pred = np.array([0, 0, 0, 1, 1, 2, 2, 2])
y_true = np.array([1, 1, 1, 2, 2, 0, 0, 0])
```

```
Purity calculated by hand: (3+2+3)/8=1.0
Purity by compute_purity(): 1.0
```

Example 3:

```
y_pred = np.array([0,0,0,1,1,1,1,2,2,2,2,2])
y_true = np.array([0,0,1,2,0,1,1,2,2,3,3,3])
```

```
Purity calculated by hand: (2+2+3)/12=0.58
Purity by compute_purity(): 0.58
```

The purity function produces the same result as hand calculation in all examples.

Task 2

Test SSE function with Iris dataset.

```
sse_kmeans = km.inertia_ / iris_data.shape[0]
sse = compute_sse(iris_data, km.labels_)
```

```
kmeans sse: 0.94740
compute_sse(): 0.94740
```

The SSE function produces the same result as the KMeans model.

Task 3

KMeans with k=4, all other parameters on default.

What percentage of the data points were assigned to each of the four clusters?

```
cluster 0: 32.84%
cluster 1: 18.48%
cluster 2: 7.32%
cluster 3: 41.36%
```

Compute the purity of the clustering result.

```
purity: 0.8158
```

Compute the purity of the clustering result for each of the four clusters. Which cluster has the highest purity?

```
cluster 0 purity: 0.8283
cluster 1 purity: 0.5974
cluster 2 purity: 0.4645
cluster 3 purity: 0.9657

cluster 3 has the highest purity
```

Task 4

KMeans with k= 2, 3, 4, 10, 20, 30. For each k, run KMeans 10 times and compute average purity and SSE. Create a table.

Table: K-Means Experiment

k	Purity	SSE
2	0.59644	5.489387
3	0.76348	4.472919
4	0.80914	3.891763
10	0.87202	2.973289
20	0.89018	2.407288
30	0.89578	2.153647

Which value of k gives the best clustering result in terms of purity?

k=30 gives the best clustering result in terms of purity. Higher purity indicates a better clustering result, since the clusters contain more homogenous data points. k=30 has the highest purity. Though we barely see purity gain from k=20 to k=30, suggesting that the model is overfitting.

Which value of k gives the best clustering result in terms of SSE?

k=30 gives the best clustering in terms of SSE. Low SSE indicates more compact clusters and a better clustering result. k=30 has the lowest SSE.

How does purity change w.r.t. the value of k? Please explain why purity changes in this way w.r.t. the value of k.

Purity increases as k increases.

When k is small (k=2,3,4,10), we see rapid increase in purity as we increase k. This is because the clustering matches the natural pattern of the data better at each step. However, as we increase k when k is larger (k=20,30), the gain in purity slows down significantly. At this point, the model starts to overfit and break down well-defined clusters into smaller ones. Purity still increases because smaller clusters are more likely to contain data points with a single class label. In the extreme case where each cluster only has one data point the purity would be 1.

Task 5

DBSCAN with eps=0.8, 0.9, 1.0, 1.1, 1.2, minPts=5, metric="euclidean", all other parameters on default. Count the total number of clusters and the total number of anomalies generated by DBSCAN, calculate the purity, SSE and Silhouette coefficient. Create a table.

Table: DBSCAN Experiment

eps	Number of Clusters	Number of Anomalies	Purity	SSE	Silhouette Coefficient
0.8	12	2656	0.7901	2.9224	-0.0093
0.9	11	2071	0.6664	3.6124	0.0073
1.0	8	1532	0.5712	4.5179	0.1002
1.1	10	1136	0.5210	5.1103	0.0974
1.2	4	854	0.4824	5.6817	0.2271

Which value of eps gives the best clustering result in terms of purity?

eps=0.8 gives the best clustering result in terms of purity. Higher purity indicates better clustering, and eps=0.8 has the highest purity.

Which value of eps gives the best clustering result in terms of SSE?

eps=0.8 gives the best clustering result in terms of SSE. Lower SSE indicates better clustering, and eps=0.8 has the lowest SSE.

Which value of eps gives the best clustering result in terms of Silhouette coefficient?

eps=1.2 gives the best clustering result in terms of Silhouette coefficient. A Silhouette coefficient closer to 1 indicates more intracluster compactness, more intercluster separation, and in summary, better clustering. eps=1.2 has the Silhouette coefficient closest to 1.

How do purity, SSE and Silhouette coefficient, respectively, change w.r.t. the value of eps?

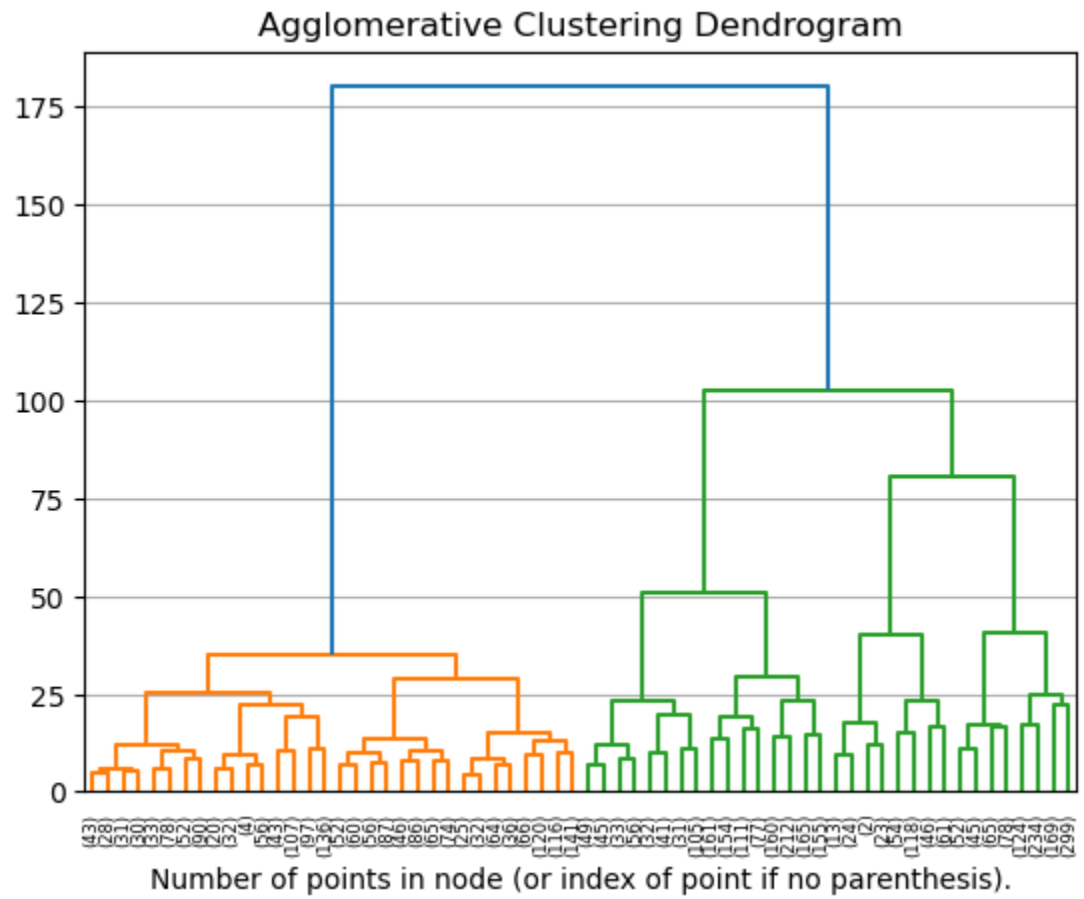
Purity decreases as eps increases. Small eps values create many small, dense clusters that tend to contain points from the same class, resulting in high purity. As eps grows, clusters expand and merge, causing more label mixing and lower purity.

SSE increases as eps increases. When eps is small, clusters are compact with low SSE. When eps grows, clusters become large and sparse with high SSE.

The Silhouette coefficient generally increases as eps increases. Small eps values produce many tiny clusters, leading to poor separation and low Silhouette scores. Larger eps values form more coherent clusters, improving the Silhouette coefficient. The slight decrease of Silhouette coefficient at eps=1.1 is probably due to the natural shape of the data that pulls noise points into the clusters when clusters expand.

Task 6

Agglomerative Clustering with distance_threshold=25, 75, 125, n_cluster=None, all other parameters on default. Draw a dendrogram.



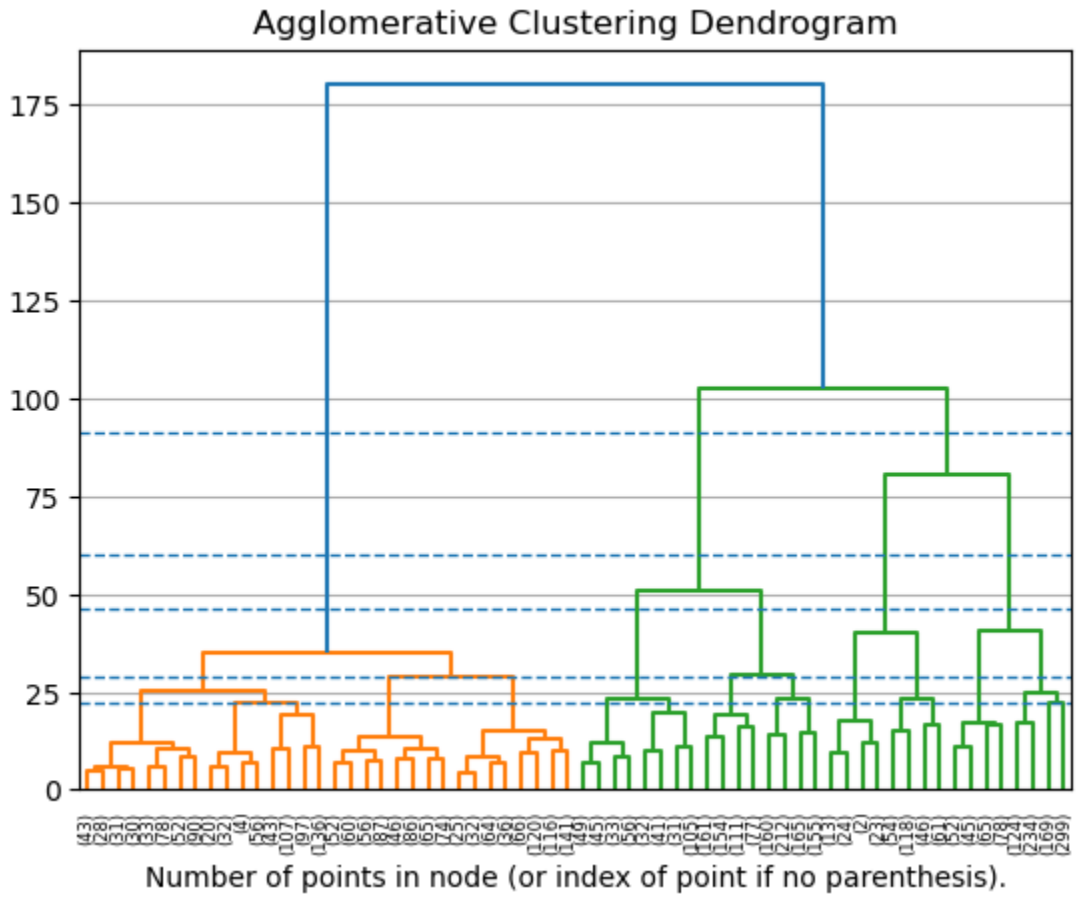
Count the total number of clusters. Calculate purity and SSE. Create a table.

Table: Hierarchical Clustering Experiment

Distance Threshold	Number of Clusters	Purity	SSE
25	11	0.8818	3.1350
75	4	0.8450	4.0730

125	2	0.6940	5.7627

Use the dendrogram to select which distance_threshold value among [22, 29, 46, 60, 91] should be used if we want to detect 5 clusters and why to select this distance_threshold value?



From this annotated dendrogram, we can find the number of clusters by counting the number of intersections between the dendrogram and the thresholds (dashed horizontal lines).

distance_threshold=22: 15+ cluster

distance_threshold=29: 9+ clusters

distance_threshold=46: 5 clusters

distance_threshold=60: 4 clusters

distance_threshold=29: 3 clusters

We use distance_threshold=46 if we want to detect 5 clusters.