Clustering Reports

Thi Quy T. Tran

UH ID: 2021505

Task 1: Function for compute purity

```python
def compute_purity(y_true, y_pred):
    total_correct = 0
    unique_clusters = np.unique(y_pred)

    y_true = np.array(y_true)

    for cluster in unique_clusters:
        cluster_indices = np.where(y_pred == cluster)[0]
        true_labels_in_cluster = y_true[cluster_indices]
        class_counts = Counter(true_labels_in_cluster)
        majority_class_count = max(class_counts.values())
        total_correct += majority_class_count

    purity = total_correct / len(y_true)
    return purity
```

Task 2: K-means Clustering with k = 2

Calculate percentage of data points in each cluster

- Cluster 0: 78.26% of data points
- Cluster 1: 21.74% of data points

Calculate purity for each cluster

- Cluster 0 Purity: 0.6923
- Cluster 1 Purity: 0.6308

The cluster has the highest purity

- Cluster 0 has the highest purity (0.6923), meaning that Cluster 0 is the most homogeneous in terms of true class labels among the two clusters.

Analysis:

- Cluster 0 is more pure, indicating that the majority of the data points in this cluster belong to a single class, likely making this cluster more well-defined.

- Cluster 1, with a lower purity of 0.6308, is less homogeneous. This suggests that it might contain more mixed-class data points, with the majority class being less dominant.

The higher purity in Cluster 0 could imply that the clustering algorithm (K-means) has better separated one of the groups (possibly the group with more distinct characteristics or a stronger pattern), while Cluster 1 might represent a more ambiguous group that is harder to clearly separate based on the features used for clustering.

Task 3: K-means Clustering with Varying k and Evaluation

K	Purity	Silhouette Coefficient
2	0.678930	0.582893
10	0.685284	0.593803
30	0.703679	0.560786
50	0.719732	0.572983
100	0.765552	0.520991

Best for Purity:

- k = 100 (Purity ~= 0.766), as more clusters lead to better assignment of data points to the correct class.

Best for Silhouette Coefficient:

- k = 10 (Silhouette ~= 0.594), indicating better-defined clusters with good separation and cohesion.

Purity vs. k:

- Purity increases with k because more clusters allow for finer grouping, improving class homogeneity within clusters. However, very high k values might lead to overfitting and overly specific clusters.

Task 4: DBSCAN Clustering Experiments with Varying eps

eps	Number of Clusters	Number of Anomalies	Purity
0.3	18	146	0.688963
0.5	22	21	0.688963
0.7	22	13	0.695652

The best clustering result in terms of purity is obtained with eps = 0.7, which gives a purity of approximately 0.696:

- Purity increases as the value of eps increases from 0.3 (~0.689) to 0.7 (~0.696).
- A higher eps value leads to more data points being included in the clusters, which results in a higher purity. However, too high a value may also cause more data points to be grouped together, potentially lowering the purity if clusters become less distinct.