Clustering Analysis Report

Question 5

1. Provided Clustering outcome

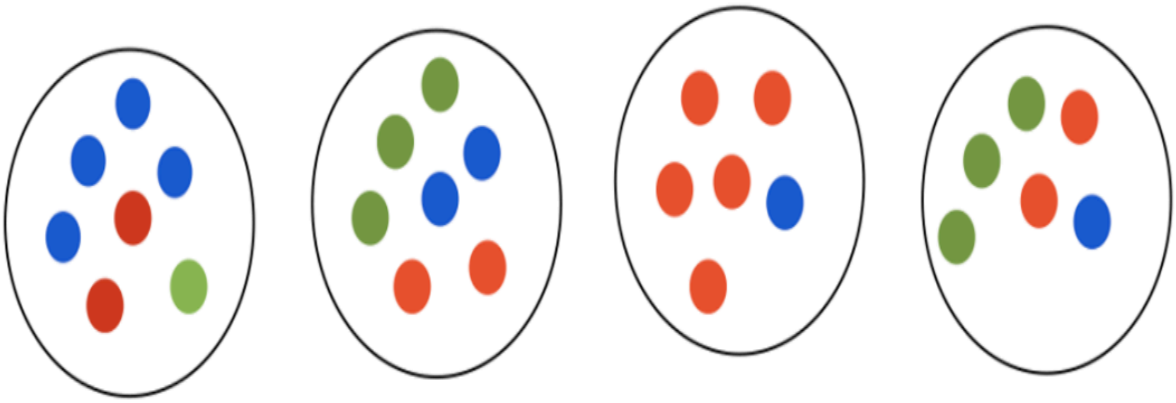
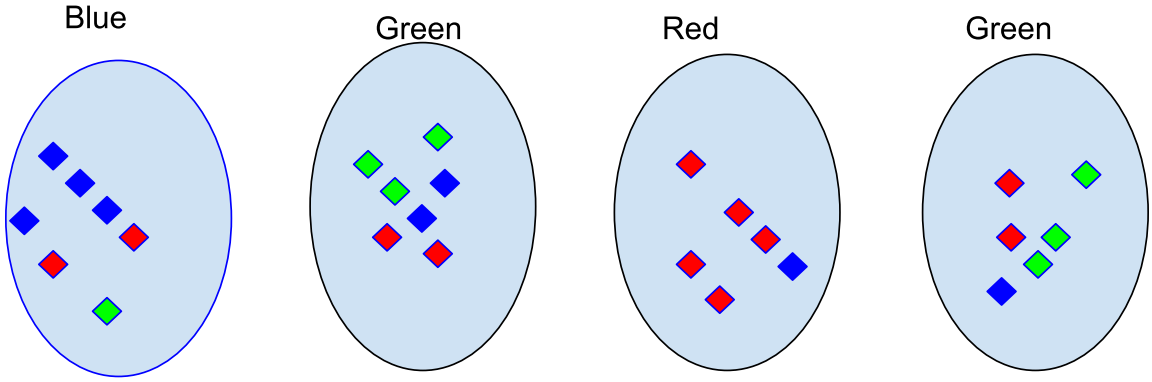


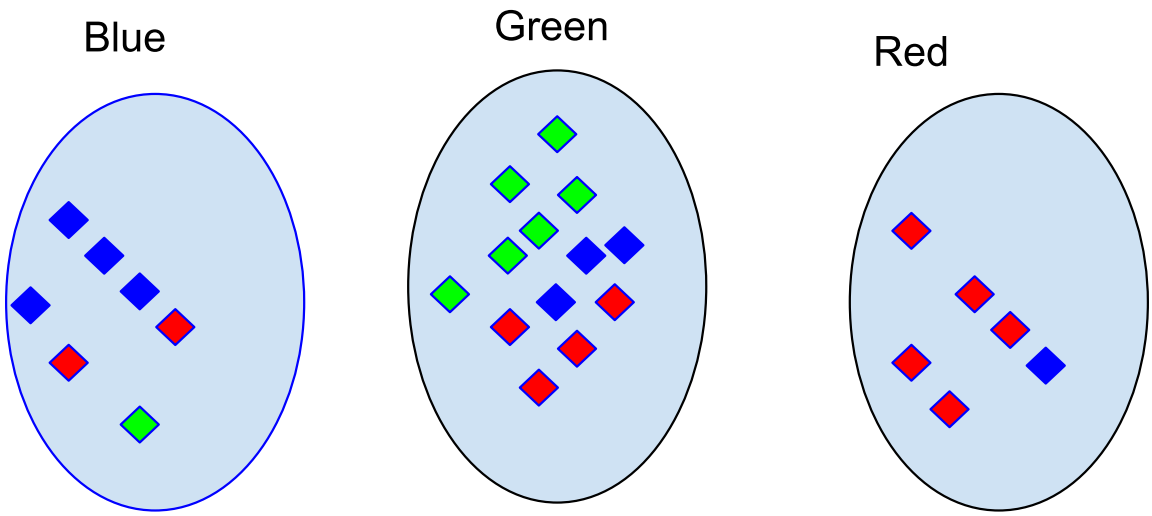
Figure 1: Outcome of a clustring algorithm

To evaluate clustering as classification, I am using extrinsic methods. The method assigns each cluster the label that appears most in the cluster. And merge clusters with the same label. Next step is to measure the precision, recall and F-score for each label type. And then calculate the macro-average, which is the average over all label types(classes) of precision, recall and F-score

2. Assigning each cluster the label:



3. Merging clusters with the same label. Here the green labels appeared twice. Hence, merging into a single cluster.



Confusion Matrix:
The confusion matrix is a matrix of numbers which tells where the number gets confused. Here the columns represent the original(expected) class distribution and the rows represent the predicted(output) distribution.

	Blue(True Label)	Green(True Label)	Red(True Label)
Class 1	4	1	2
Class 2	3	6	4
Class 3	1	0	5

Calculating Precision, recall and F-score for each cluster seperately

$$Precision = True\ Positive \div (True\ Positive + False\ Positive)$$
$$Recall = True\ Positive \div (True\ Positive + False\ Negative)$$
$$F - score = 2 * (Precision * Recall) \div (Precision + Recall)$$
$$Macro\ average\ precision = Average\ of\ precision\ from\ each\ class$$
$$Macro\ average\ Recall = Average\ of\ Recall\ from\ each\ class$$

$$Macro\ average\ F-Score = Average\ of\ F-score\ from\ each\ class$$

Class 1:

Precision = 4 / (4 + 2 + 1) = 0.57
Recall = 4 / (4 + 3 + 1) = 0.50
*F-score = 2 * (0.57 * 0.50) / (0.57 + 0.50) = 0.53*

Class 2:

Precision = 6 / (3+6+4) = 0.46
Recall = 6 / (1+6+0) = 0.85
*F-score = 2 * (0.46 * 0.85) / (0.46 + 0.85) = 0.59*

Class 3:

Precision = 5 / (1 + 5+0) = 0.83
Recall = 5 / (5+4+2) = 0.45
*F-score = 2 * (0.83 * 0.45) / (0.83 + 0.45) = 0.58*

Finally, we take the macro-averaged Precision, Recall, and F-score:

Macro-averaged Precision = (0.57+0.46+0.83)/3=0.62
Macro-averaged Recall = (0.50 + 0.85 + 0.45) / 3 = 0.60
Macro-averaged F-score = (0.53 + 0.59 + 0.58) / 3 = 0.56

Observation from the calculation:

- The macro-averaged precision of 0.62 indicates that, on average, 62% of the instances assigned to each cluster are correct.
- The macro-averaged recall of 0.60 suggests that, on average, 60% of the instances of each true class are correctly identified.
- The macro-averaged F-score of 0.56 combines precision and recall into a single metric, providing a balanced measure of the overall clustering performance.

Question 6

Computing B-cubed Precision, Recall and F-Score based on the clustering outcome figure given.

B-Cubed is an evaluation metric, which can evaluate the precision and recall for every data point in clustering on a given dataset according to ground truth. The idea here is that we evaluate clustering without labeling any clusters.

Equations:

$$Precision = No.\ of\ items\ in\ the\ same\ cluster\ with\ same\ true\ label \div No.\ of\ items\ in\ the\ predicted\ cluster$$
$$Recall = No.\ of\ items\ in\ the\ same\ cluster\ with\ the\ same\ true\ label \div No.\ of\ items\ with\ same\ true\ label$$
$$F - score = 2 * (Precision * recall) \div (Precision + Recall)$$

Based on Figure 1, Total Number of Instances is 26. Below table has the values of each instance's precision, recall and F-score values and their average.

Class	Instances	Precision	Recall	F-score
Class 1	Blue	4/7	4/8	0.52
	Blue	4/7	4/8	0.52
	Blue	4/7	4/8	0.52
	Blue	4/7	4/8	0.52
	Red	2/7	2/11	0.21
	Red	2/7	2/11	0.21
	Green	1/7	1/7	0.14
Class 2	Blue	2/7	2/8	0.26
	Blue	2/7	2/8	0.26
	Green	3/7	3/7	0.42
	Green	3/7	3/7	0.42
	Green	3/7	3/7	0.42
	Red	2/7	2/11	0.21
	Red	2/7	2/11	0.21
Class 3	Red	5/6	5/11	0.59
	Red	5/6	5/11	0.59
	Red	5/6	5/11	0.59
	Red	5/6	5/11	0.59
	Red	5/6	5/11	0.59
	Blue	1/6	1/8	0.13

	Green	3/6	3/7	0.46
	Green	3/6	3/7	0.46
	Green	3/6	3/7	0.46
	Red	2/6	2/11	0.24
	Red	2/6	2/11	0.24
Class 4	Blue	1/6	1/8	0.05
Average Precision	Sum of Precision/Total no. of Instances	0.46		
Average Recall	Sum of Recall/Total no. of Instances		0.34	
Average F-Score	Sum of F-score/Total no. og Instances			0.39

Hence the B-cubed precision is approximately 0.46, the recall is 0.34 and F-score is 0.39