MA5755 – Data Analysis & Visualization in R/Python/SQL
Assignment 4

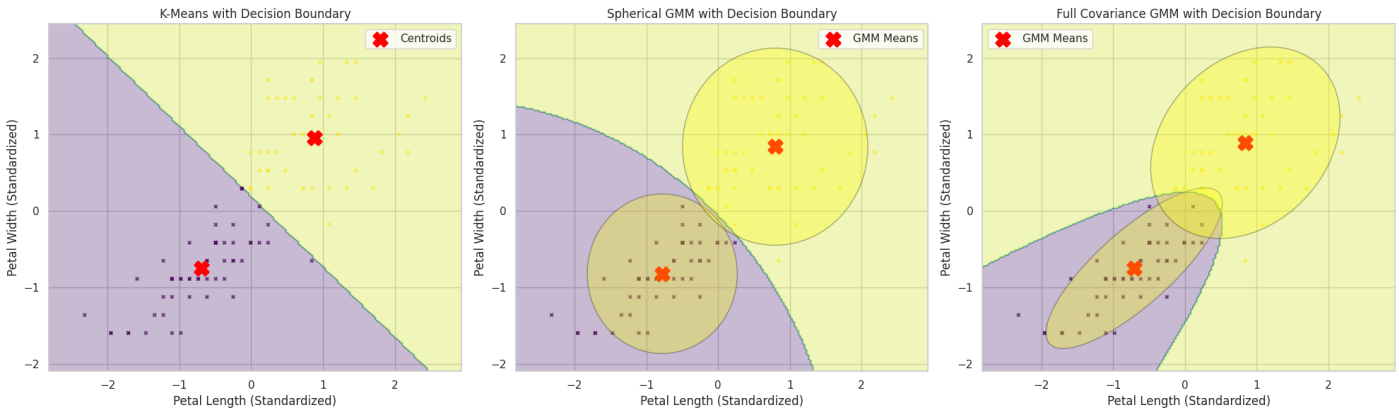
Anurag Dhiman - MA25M004

TASK 1

On executing code for visualizing three algorithms k-means, gaussian mixture and full gaussian mixture, I have found that k means clustering (k= 2) , spherical gmm clustering (k=2) have separating data points in similar looking clusters but full covariance gmm (k=2) outperforms the both. In K-Means clustering K-means identifies spherical clusters, with each cluster represented by a single centroid (marked as 'X'). These centroids are the mean of the data points assigned to that cluster. Also, covariance structure in K-means inherently assumes clusters of equal variance, meaning it cannot capture elongated or arbitrarily oriented clusters. The plots reflect this by showing roughly groupings around the centroids. If i look into spherical GMM cluster representatives like k-means, spherical GMMs also use means ('X') as cluster representatives. However, they assign a probability of belonging to each cluster, allowing for soft assignments. This means the principal axis of the ellipses are aligned with the coordinate axes, and the variances along these axis are equal, resulting in circular ellipses. The yellow ellipses in the plot visually confirm this assumption, showing equally sized and oriented circular shapes around the means.
In last, full covariance GMM cluster representatives also use means ('X') for each Gaussian component. This model offers the most flexibility by allowing each cluster to have its own arbitrary covariance matrix. This enables it to model elliptical clusters of varying sizes, orientations, and eccentricities. The ellipses in this plot are elongated and angled, better reflecting the natural spread and correlation between 'Petal Length' and 'Petal Width' within each cluster. This flexibility often leads to a better fit for data with complex, non-spherical distributions, as evidenced by its higher adjusted rand index (ARI) compared to k-means and spherical GMM in this example.

Table 1: Adjusted Rand Index (ARI) Comparison with True Species Labels

Clustering Method	ARI Score
k-means	0.7721
Spherical GMM	0.7721
Full Covariance GMM	**0.8448**



TASK 2

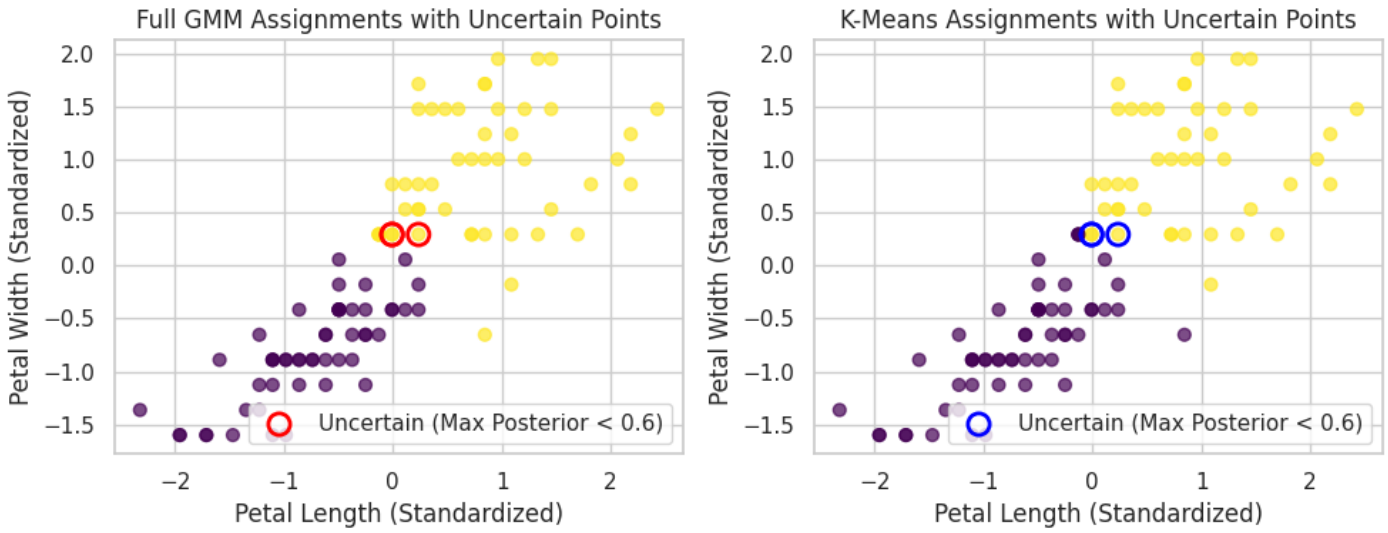
After fitting the all the algorithms and calculating the posterior probabilities for each data point. These probabilities indicate the likelihood that a given data point belongs to each of the two Gaussian components identified by the full covariance GMM. For instance, a point with probabilities $[0.95, 0.05]$ is strongly assigned to the first cluster, while a point with probabilities $[0.55, 0.45]$ has a more ambiguous assignment. While the raw posterior probabilities are not directly plotted, they are fundamental to identifying the "uncertain" points and making the hard assignments.

To identify observations with low maximum posterior probability (ambiguous points near decision boundaries). The visual output identified 3 uncertain points (indices 73, 77, and 99). These are data points whose highest posterior probability of belonging to any single cluster was less than the 0.6 threshold, making their cluster assignment less confident. Visually, in both subplots of the generated figure (full GMM assignments and K-Means assignments), these 3 points are distinctly highlighted (in red circles for GMM, blue circles for K-Means). Their location on the plots clearly shows them situated in the region where the two clusters overlap, near the boundary between them. This visual placement confirms their ambiguous nature and their proximity to the decision boundary.

Interestingly, for all three identified "uncertain" points, both the K-Means and the full covariance GMM models assigned them to the same cluster (cluster 1). Despite their ambiguity (low max posterior probability), both algorithms found the same most probable cluster for these specific points. Visually, this consistency is apparent, as highlighted uncertain points are colored according to cluster 1 in both the GMM and K-Means plots, further reinforcing that even for these boundary cases, the algorithms agreed on their hard assignment.

Table 2: Uncertain Observations Identified by Full Covariance GMM

Number of Uncertain Points: 3			
Index	KMeans Cluster	GMM Cluster	Max Posterior
73	1	1	0.579
77	1	1	0.579
99	1	1	0.591



Finally, on deep analysis through numerical outputs and graph visulaizations revealed that the full covariance gaussian mixture model provided the best fit to the Iris dataset, achieving the highest adjusted rand index. This is attributed to its flexibility in modeling arbitrarily shaped, elliptical clusters, unlike K-Means and spherical GMM which assume spherical clusters. If we do final overall analysis, the full GMM's ellipses accurately captured the correlated features.