# RT 1: Clustering

To cluster the Covertype dataset subset with k-means and a Gaussian Mixture Model (GMM), I made specific choices regarding scaling, initialization, and other parameters to enhance stability and performance without using any class label information. Here is the rationale behind these choices:

1. **Data Scaling**:  
   I scaled the data using StandardScaler before running k-means and GMM. Since features in the Covertype dataset have varying ranges, standardizing them ensures that each feature contributes equally to distance calculations in both algorithms. This is particularly important for k-means, which relies on Euclidean distances, and beneficial for GMM, as scaling can improve convergence in multimodal distributions.
2. **K-means Initialization and Parameters**:  
   I used the k-means++ initialization for k-means to enhance clustering quality. The k-means++ method spreads out the initial cluster centers, reducing the likelihood of poor convergence and often leading to a more stable result than random initialization. Testing shows that this initialization requires fewer iterations to converge compared to random. Additionally, I set random\_state=42 to ensure reproducibility across runs, which is beneficial for consistent results in coursework while maintaining independence from the actual class labels.
3. **Gaussian Mixture Model Initialization and Parameters**:  
   For GMM, I chose the kmeans initialization for init\_params rather than random to leverage k-means clustering to initialize cluster centers more accurately. kmeans initialization for GMM has shown to improve convergence stability and reduce the number of poor initializations. I set max\_iter=100 to balance sufficient iteration time with computational efficiency, and tol=1e-3 as a convergence threshold to ensure that the model refines its cluster probabilities until there’s minimal improvement. I thought about including warm\_start but I ended up omitting it because, given the scope and scale of this coursework, iterative fitting would not be necessary.
4. **Random State and Reproducibility**:  
   By setting a random\_state=42 for both k-means and GMM, I ensured that the algorithms produce consistent results, allowing for reliable analysis of cluster performance across multiple runs. In a coursework setting, this reproducibility is advantageous as it enables clear assessment and comparison without relying on or affecting class label information.

Through these configurations, I selected initialization methods and parameters that are suited for clustering without label information, thereby achieving reproducibility and stable clustering quality suitable for unsupervised learning tasks.