Unsupervised Learning - Coding Practice Questions

CM52054: Foundational Machine Learning
Practice set with fully worked answers

1) **Implement k-means from scratch.**
Implement k-means clustering with:

- centroid initialisation by sampling K *distinct data points*,
- Euclidean distance,
- stopping when assignments stop changing *or* centroid movement is below a tolerance.

Return `(centroids, labels, wcss)`, where

$$\text{WCSS (Within-Cluster Sum of Squares)} = \sum_{i=1}^{N} \|x_i - c_{\ell_i}\|_2^2.$$

2) **Best-of-n restarts for k-means (mitigate non-determinism).**
Implement a wrapper `kmeans_best_of_n` that:

- runs k-means `n_init` times with different seeds,
- returns the solution with the *lowest* WCSS.

3) **Compute "good clustering" metrics (intra vs inter).**
Implement a function `clustering_quality_metrics(X, labels, centroids)` that computes:

(a) average intra-cluster distance: mean $\|x_i - c_{\ell_i}\|_2$,

(b) minimum inter-centroid distance: $\min_{k \neq k'} \|c_k - c_{k'}\|_2$.

4) **Optional: Agglomerative hierarchical clustering (single linkage) for small datasets.**
Implement agglomerative clustering with single linkage:

- start with each point as its own cluster,
- repeatedly merge the pair of clusters with minimum single-link distance,
- return a merge history list.

Assume N is small; clarity is more important than speed.