Week 8 Summary: k-Means, Hierarchical Clustering, and Validation

Z3003 Data Mining: Scaling to Large Data · 9–10 November 2026 · Dr. Innocent Nyalala

Lecture 22: Unsupervised Clustering via k-Means

After log-transforming skewed columns and standardizing via z-scores, k-means alternates assigning each record to its nearest centroid and updating centroids to cluster means until convergence. Run multiple starts (k-means++ / `n_init=10`) to avoid local optima, and guide k via the within-cluster SSE elbow and silhouette score $s(i) = (b(i) - a(i))/\max(a(i), b(i))$. On $10,000$ Zanzibar farms, $k = 4$ yields actionable profiles (irrigated achievers, fertilized rain-fed, mid-size without inputs, small without inputs). One extreme outlier (100 ha) can distort means badly.

Lecture 23: Hierarchical Clustering and Scale Limits

Agglomerative clustering merges closest clusters bottom-up into a dendrogram using single, complete, average, or Ward linkage. Cutting across a wide vertical gap chooses k post hoc. Because storing pairwise distances requires $n(n-1)/2$ entries (0.4 GB at 10^4 rows, 40 GB at 10^5 rows, 4 TB at 10^6 rows), run hierarchical clustering on a small sample to pick k, then scale with k-means or BIRCH.