Module 4 Summary and Highlights

Congratulations! You have completed this lesson. At this point in the course, you know:

- Clustering is a machine learning technique used to group data based on similarity, with applications in customer segmentation and anomaly detection.
- K-means clustering partitions data into clusters based on the distance between data points and centroids but struggles with imbalanced or non-convex clusters.
- Heuristic methods such as silhouette analysis, the elbow method, and the Davies-Bouldin Index help assess k-means performance.
- DBSCAN is a density-based algorithm that creates clusters based on density and works well with natural, irregular patterns.
- HDBSCAN is a variant of DBSCAN that does not require parameters and uses cluster stability to find clusters.
- Hierarchical clustering can be divisive (top-down) or agglomerative (bottom-up) and produces a dendrogram to visualize the cluster hierarchy.
- Dimension reduction simplifies data structure, improves clustering outcomes, and is useful in tasks such as face recognition (using eigenfaces).
- Clustering and dimension reduction work together to improve model performance by reducing noise and simplifying feature selection.
- PCA, a linear dimensionality reduction method, minimizes information loss while reducing dimensionality and noise in data.
- t-SNE and UMAP are other dimensionality reduction techniques that map high-dimensional data into lower-dimensional spaces for visualization and analysis.