**K-Means Clustering Analysis Report**

**Introduction**

K-Means is an unsupervised clustering algorithm that partitions data into k clusters based on feature similarity. In this analysis, K-Means was applied to the Iris dataset, a classic benchmark dataset, to explore how well clustering can group iris species based on sepal and petal measurements.

**Problem Statement**

The objective of this analysis is to group iris flowers into clusters and compare them with the actual species labels (Setosa, Versicolor, Virginica). The key questions addressed are:

* How well can K-Means cluster iris species without supervision?
* How closely do clusters align with the actual species labels?
* Which features contribute most to cluster separation?

**Task 1: Data Preparation & Analysis**

**Dataset Overview:**

* Samples: 150
* Features: 4 (sepal length, sepal width, petal length, petal width)
* Classes: 3 species

**Preprocessing Steps:**

* Features were standardized using StandardScaler.
* No label encoding was required since clustering is unsupervised.
* Data was used directly for clustering after normalization.

**Task 2: Model Training (K-Means)**

**Model Configuration:**

* Number of Clusters (k): 3 (chosen to match the number of species)
* Initialization Method: k-means++
* Number of Initializations: 10
* Maximum Iterations: 300
* Random State: 42

The K-Means algorithm iteratively assigned samples to clusters and updated cluster centroids until convergence was achieved.

**Task 3: Results**

**Performance Metrics:**

* Adjusted Rand Index (ARI): ~0.73 (indicating substantial alignment with actual species)
* Silhouette Score: ~0.55 (showing moderately well-separated clusters)

**Cluster Analysis:**

* Cluster 1: Mostly Setosa (well-separated)
* Cluster 2 & 3: Mixed Versicolor and Virginica, showing overlap due to feature similarity

**Feature Contribution:**

* Petal length and petal width played the largest role in separating clusters.
* Sepal features contributed less effectively to clustering.

**Task 4: Insights & Applications**

* K-Means successfully separated Setosa as a distinct cluster, confirming its strong feature separation.
* Versicolor and Virginica overlap, highlighting the biological similarity between these species.
* Petal features dominate clustering, consistent with their importance in supervised classification tasks.
* K-Means can be a useful tool for exploratory data analysis when labels are unavailable.

**Challenges Faced**

* K-Means requires predefining the number of clusters (k), which may not always be known.
* Sensitive to initialization; different random seeds can yield different results.
* Struggles with clusters of varying shapes and densities.
* Overlap between Versicolor and Virginica reduces clustering accuracy.

**Conclusion & Recommendations**

K-Means achieved reasonable clustering performance (~73% ARI, 0.55 silhouette score) on the Iris dataset. It clearly identified Setosa as a distinct cluster, while Versicolor and Virginica showed significant overlap. For small, structured datasets, K-Means can provide valuable insights, but its limitations must be considered.

**Recommendations:**

* Use K-Means for exploratory analysis when class labels are unknown.
* Experiment with different values of k and evaluate clustering metrics (silhouette, ARI).
* Compare with hierarchical clustering or Gaussian Mixture Models for potentially better separation.
* Visualize clusters with PCA or t-SNE to interpret results effectively.