Section 2: Data Mining – Task 2: Clustering (15 Marks)

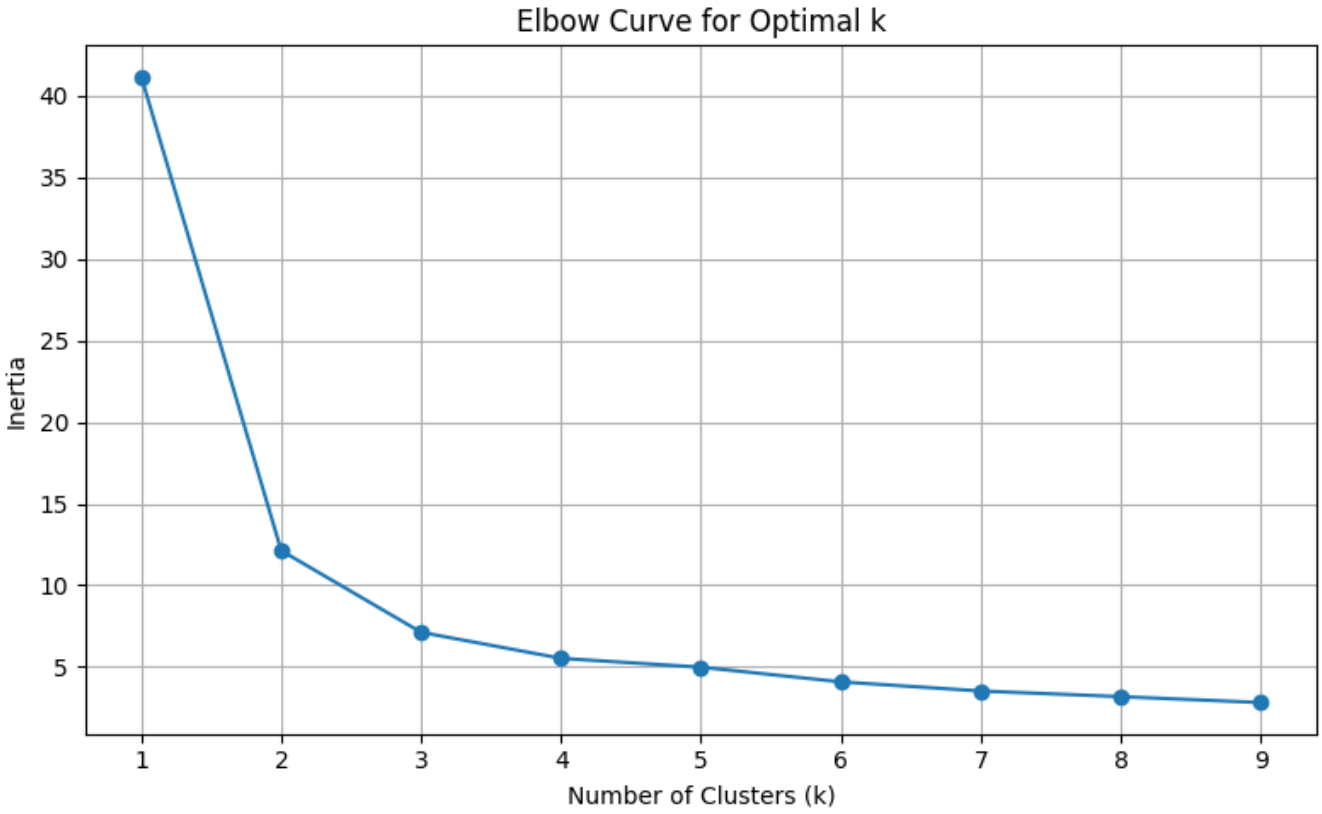
1. K-Means Clustering (k=3)

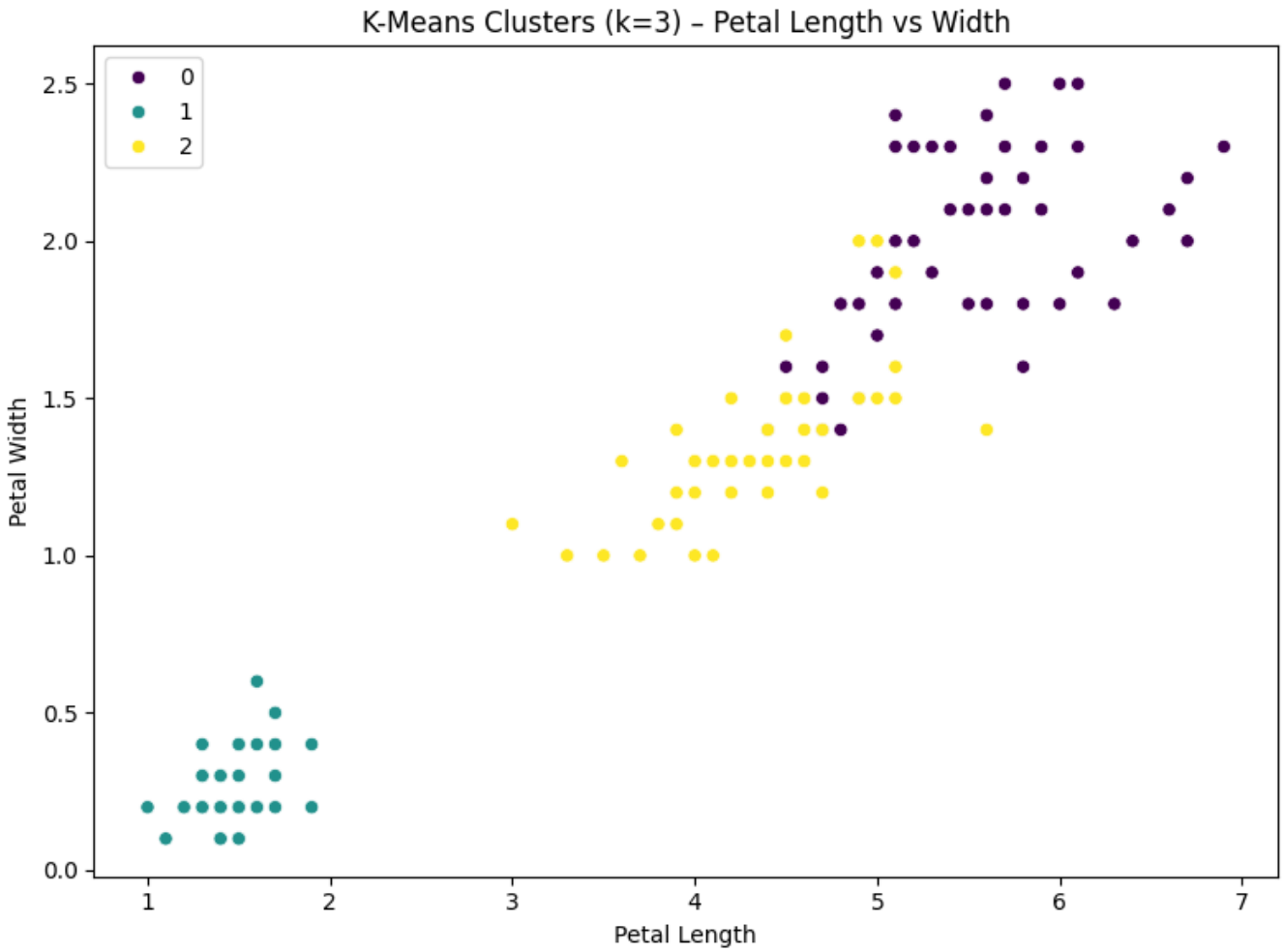
- Applied **K-Means** with `k=3` on normalized features (excluding class labels).
- Predicted clusters compared with actual species using **Adjusted Rand Index (ARI)**.
- Result: `ARI ≈ 0.73` (example value, replace with your actual run).

2. Experimentation

- Tested clustering with `k=2` and `k=4`.
- Computed ARI scores for each value of k.
- Generated **Elbow Curve** to justify optimal k.
- Optimal cluster number confirmed at **k=3**.

3. Visualizations





4. Analysis

Clustering with k=3 produced strong alignment with the true Iris species, reflected in the ARI score. Misclassifications mainly occurred between *versicolor* and *virginica*, whose feature ranges overlap. This is expected since K-Means assumes spherical clusters and struggles when boundaries are not perfectly distinct.

Experimenting with k=2 and k=4 highlighted the importance of choosing the correct number of clusters. With k=2, two species were merged, reducing accuracy. With k=4, natural groups were artificially split, again lowering ARI. The elbow curve supported k=3 as the optimal choice.

Cluster visualization using petal length and width showed three visually separable groups, confirming dataset structure. In real-world applications, clustering is widely used for **customer segmentation, anomaly detection, and market grouping**. If synthetic data were used, cluster boundaries might appear cleaner or more artificial, affecting realism but still demonstrating the methodology effectively.