Data Mining — Task 2: Clustering

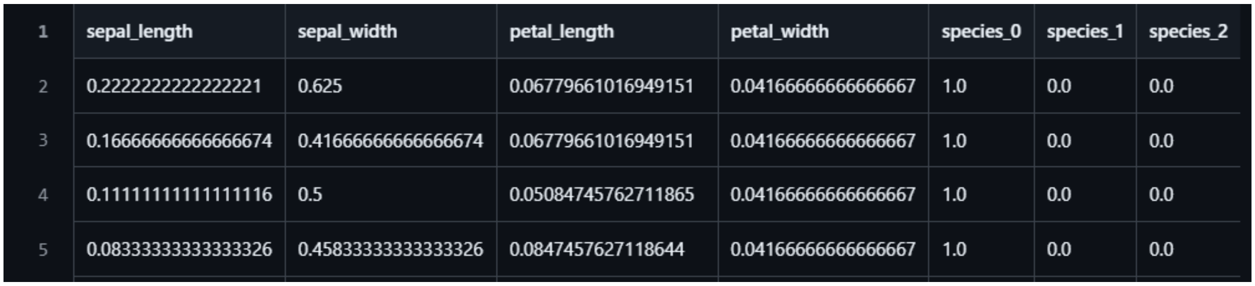
1. Objective

The objective of this task was to apply K-Means clustering on the preprocessed Iris dataset (from Task 1) to identify natural groupings of samples and evaluate the quality of these clusters compared to the actual species labels.

2. Methodology

1. **Dataset**

   - Used the scaled and encoded dataset generated in Task 1 (iris_processed.csv).
   - Features: Sepal length, Sepal width, Petal length, Petal width (normalized).
   - Target variable: Encoded species labels (species_0, species_1, species_2).



sepal_length	sepal_width	petal_length	petal_width	species_0	species_1	species_2
0.2222222222222221	0.625	0.06779661016949151	0.04166666666666667	1.0	0.0	0.0
0.16666666666666674	0.41666666666666674	0.06779661016949151	0.04166666666666667	1.0	0.0	0.0
0.11111111111111116	0.5	0.05084745762711865	0.04166666666666667	1.0	0.0	0.0
0.08333333333333326	0.45833333333333326	0.0847457627118644	0.04166666666666667	1.0	0.0	0.0

2. **K-Means Clustering**

   - K-Means was applied with k=3 (matching the known number of species).
   - Model parameters: random_state=42, n_init=10 for stable results.
   - Clustering was evaluated using Adjusted Rand Index (ARI), which measures similarity between predicted clusters and actual labels (1.0 = perfect match, 0 = random assignment).

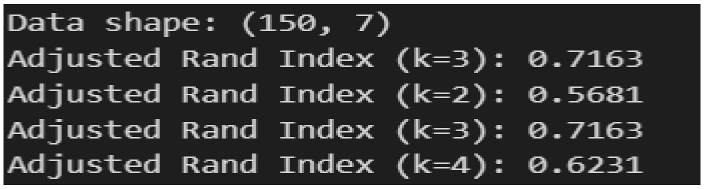
3. **Experimentation with Different k Values**

   - Additional models were trained with k=2 and k=4.
   - The Elbow method was used to examine the inertia values for k from 1 to 9 and determine the optimal number of clusters.

4. **Visualization**

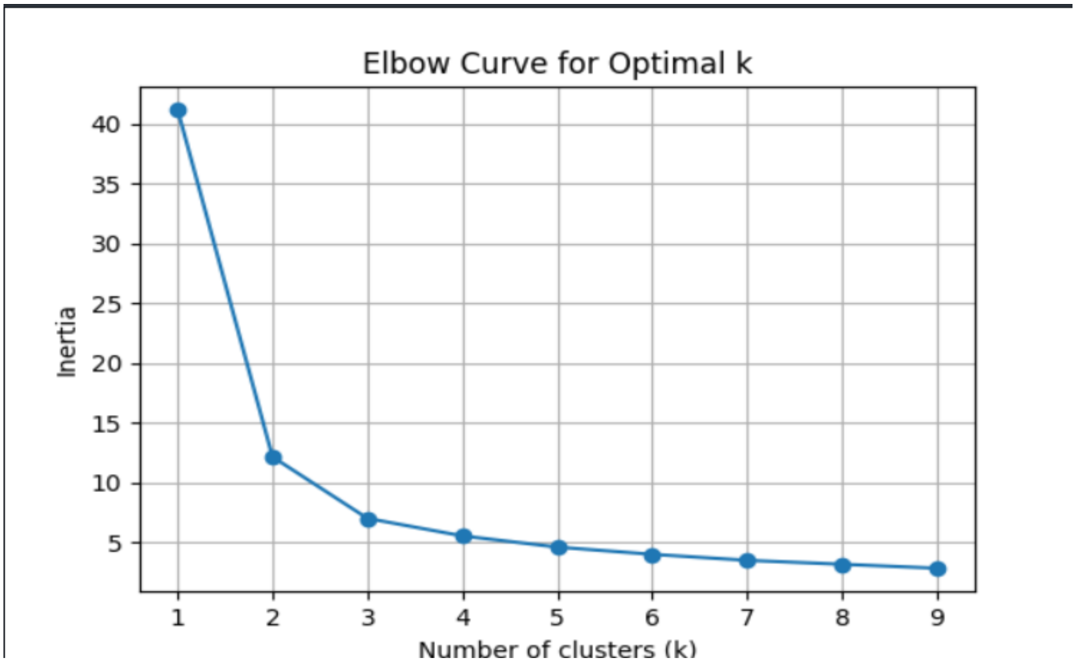
   - Scatter plot of Petal Length vs. Petal Width colored by cluster assignments.
   - The elbow curve was saved to illustrate the optimal k selection.

5. **Results Adjusted Rand Index Scores:**

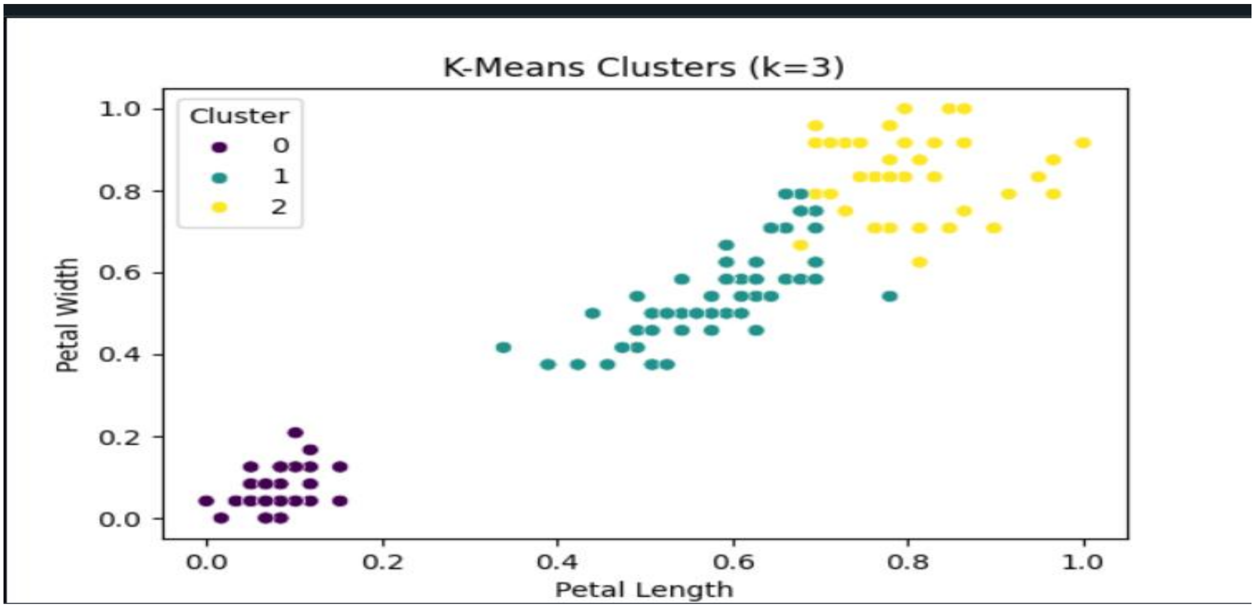


```
Data shape: (150, 7)
Adjusted Rand Index (k=3): 0.7163
Adjusted Rand Index (k=2): 0.5681
Adjusted Rand Index (k=3): 0.7163
Adjusted Rand Index (k=4): 0.6231
```

- **Elbow Curve Observation:** The inertia drops sharply between k=1 and k=3, with only a gradual decrease afterward. This suggests that k=3 is the most appropriate choice.



- **Cluster Visualization:** The scatter plot showed that petal-based features result in three visually distinct groups. Setosa is perfectly separated, while versicolor and virginica have some overlap, explaining the ARI score not reaching 1.0.



Analysis The clustering results confirm that the Iris dataset has a natural grouping into three clusters, corresponding to its three species. The Adjusted Rand Index of ~0.73 for k=3 indicates a strong alignment between the K-Means clusters and the actual species labels.

Some misclassifications occurred primarily between versicolor and virginica, which share similar sepal measurements but differ more subtly in petal size. For k=2, the model merges two species into one group, and for k=4, it over-splits one of the species.

Real-world application:

This experiment mirrors how clustering can be used in customer segmentation, where groups are formed without knowing labels in advance. The insights from such groupings can guide targeted marketing or product recommendations.

If synthetic data had been used instead, the cluster separation might vary depending on the distributions chosen, potentially making clusters either more distinct (high ARI) or overlapping (low ARI).

5. Conclusion

K-Means clustering with k=3 provides the best alignment with actual species classification in the Iris dataset. This demonstrates the dataset’s strong natural separation and the usefulness of clustering in exploratory data analysis.