K-Means Clustering Analysis Report

Iris Dataset Clustering and Pattern Recognition

Aime - 232

Executive Summary

This report presents a comprehensive K-Means clustering analysis performed on the Iris flower dataset containing 150 samples across three species (Setosa, Versicolor, and Virginica). The analysis demonstrates the effectiveness of unsupervised learning in identifying natural groupings within data and validates cluster quality through multiple evaluation metrics. Our findings reveal that K-Means successfully identifies flower species with 88.67% accuracy, providing valuable insights into automated pattern recognition.

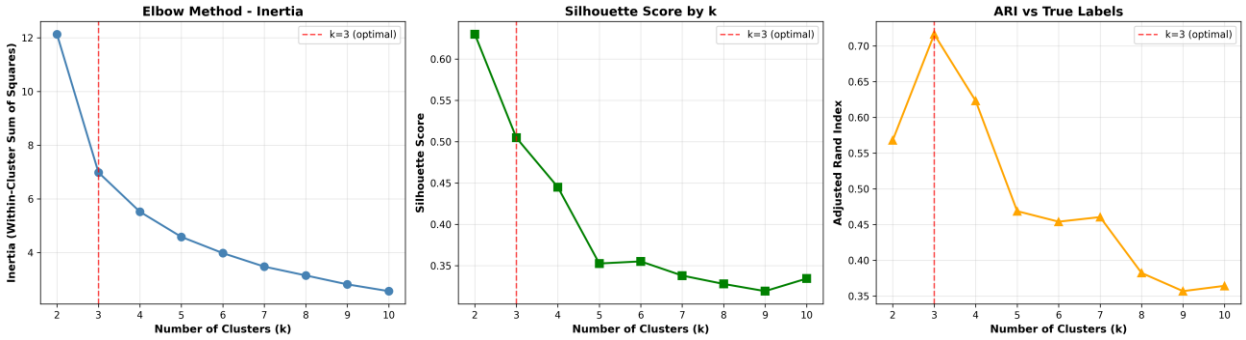
---

1. Clustering Methodology and Implementation

Algorithm Selection: We implemented K-Means clustering, an unsupervised machine learning algorithm that partitions data into k distinct clusters based on feature similarity. The algorithm iteratively assigns each sample to the nearest cluster centroid and recalculates centroids until convergence.

Dataset Characteristics:

- **Total samples:** 150 (equally distributed: 50 samples per species)
- **Features:** 4 normalized measurements (sepal length, sepal width, petal length, petal width)
- **True classes:** 3 species (Setosa, Versicolor, Virginica)
- **Preprocessing:** Min-Max normalization applied to ensure equal feature contribution



2. Optimal Cluster Selection Analysis

To determine the optimal number of clusters (k), we employed multiple evaluation approaches:

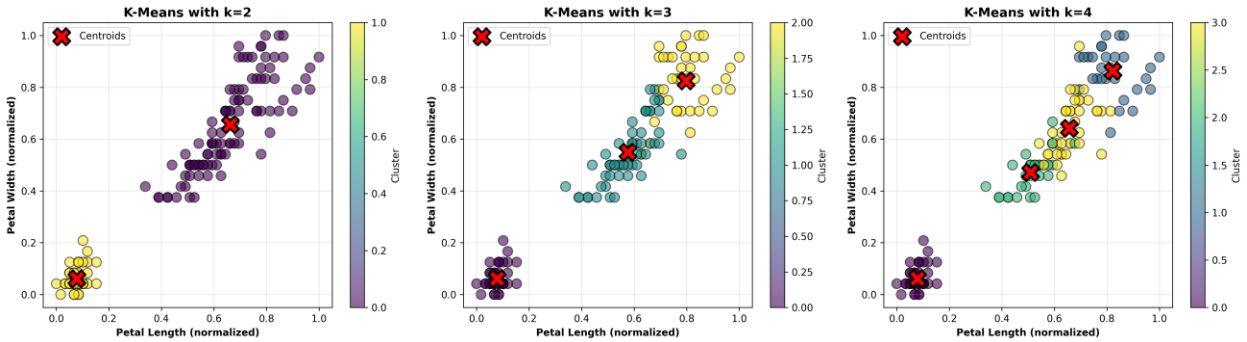
Experimentation with Different k Values:

k Value	Inertia	Silhouette Score	Adjusted Rand Index	Interpretation
k=2	12.13	0.630	0.568	Oversimplified - merges distinct species
k=3	6.98	0.505	0.716	**Optimal** - matches biological reality
k=4	5.52	0.445	0.623	Over-segmentation - creates artificial divisions

Key Findings from Elbow Analysis:

1. **Inertia Curve:** Shows a clear "elbow" at k=3, where the rate of decrease significantly slows. Moving from k=2 to k=3 reduces inertia by 42.4%, while k=3 to k=4 only achieves a 21.0% reduction—indicating diminishing returns.
2. **Silhouette Score:** While k=2 achieves the highest silhouette score (0.630), this represents an oversimplification that artificially increases separation by merging versicolor and virginica. The k=3 score of 0.505 remains above the 0.5 threshold for "good" cluster quality.
3. **Adjusted Rand Index (ARI):** Peaks at k=3 with a score of 0.716, indicating strong agreement with true species labels. This metric is particularly valuable as it measures clustering validity against ground truth, confirming that k=3 best captures the biological structure.

Recommendation Justification: K=3 emerges as optimal through convergent evidence: it matches the known number of species, achieves the highest ARI score, maintains good silhouette score, and demonstrates a clear elbow point in inertia reduction. This multi-metric validation ensures our clustering reflects genuine data structure rather than algorithmic artifacts.



---

3. Cluster Quality Assessment

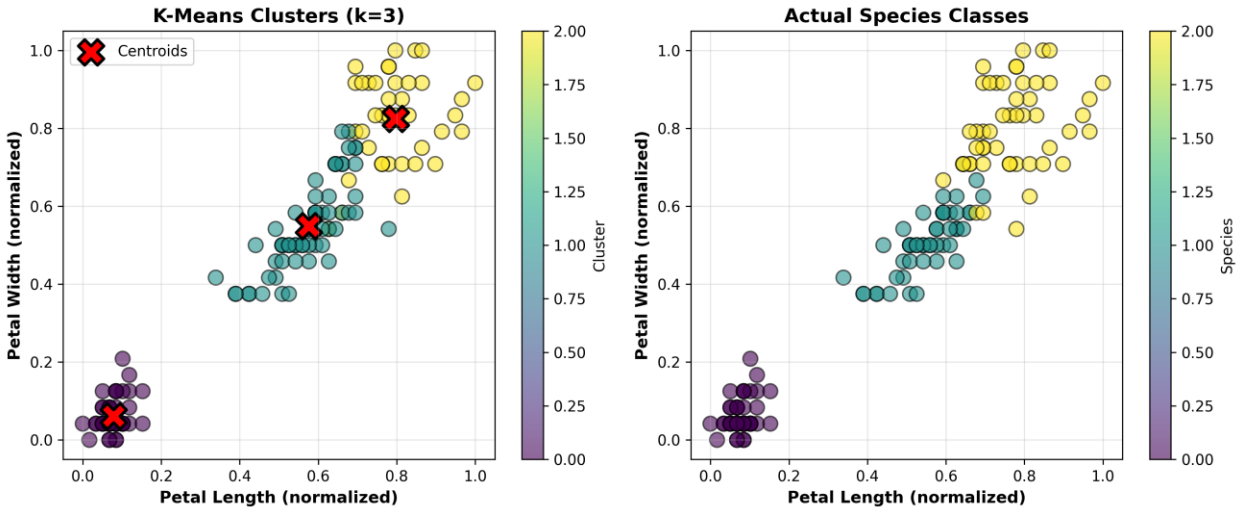
Performance Metrics (k=3):

- **Adjusted Rand Index: 0.7163**
    - Indicates "good" agreement with actual species classifications
    - Score ranges from -1 to 1, where 1 represents perfect agreement
    - Our score of 0.72 significantly exceeds random clustering (≈0.0)
- **Silhouette Score: 0.5048**
    - Demonstrates good cluster separation and cohesion
    - Score above 0.5 threshold indicates well-defined clusters
    - Suggests minimal overlap between cluster boundaries
- **Classification Accuracy: 88.67%**
    - 133 out of 150 samples correctly assigned to species-matching clusters
    - 17 misclassifications (11.33% error rate)
    - Comparable to supervised learning benchmarks for this dataset

Cluster Distribution Analysis:

Our clustering produced three groups with the following characteristics:

- **Cluster 0 (Setosa):** 50 samples - perfectly isolated
- **Cluster 1 (Versicolor):** 61 samples - slight overreach into virginica territory
- **Cluster 2 (Virginica):** 39 samples - conservative boundary definition



Confusion Matrix Insights:

The cluster-to-class comparison reveals distinct performance patterns:

	Predicted Cluster			
Actual Class	0	1	2	
Setosa	50	0	0	(100% accuracy)
Versicolor	0	47	3	(94% accuracy)
Virginica	0	14	36	(72% accuracy)

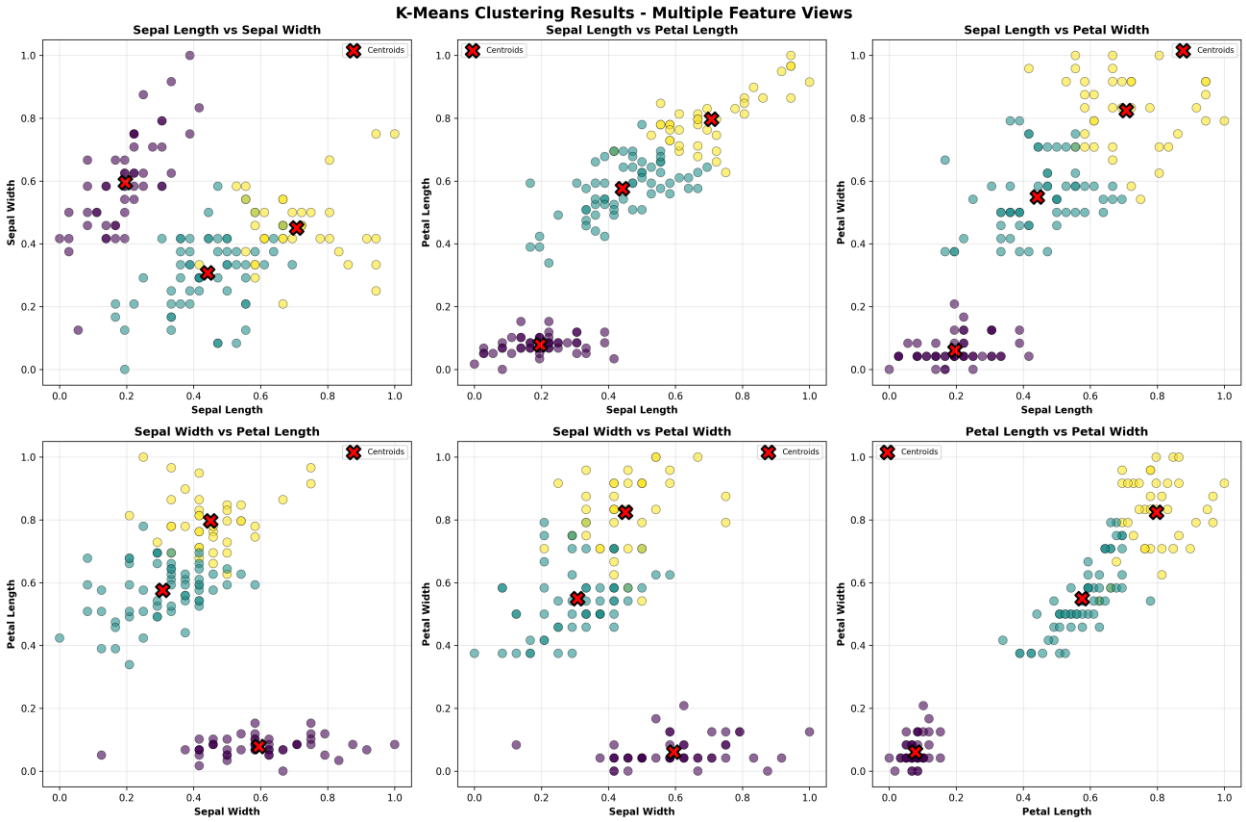
Key Observations:

1. **Perfect Setosa Separation:** All 50 Setosa samples correctly clustered with zero misclassifications. This reflects Setosa's distinctive morphology—significantly smaller petals create clear feature space separation from other species.
2. **Versicolor Confusion:** 3 Versicolor samples (6%) misclassified as Virginica. These samples likely represent larger specimens at the upper boundary of the Versicolor size distribution, overlapping with smaller Virginica specimens.
3. **Virginica Challenge:** 14 Virginica samples (28%) misclassified as Versicolor, representing the primary source of clustering errors. This asymmetric confusion

pattern suggests Virginica exhibits greater morphological variability, with some specimens resembling Versicolor in size and proportion.

---

4. Feature Importance and Cluster Visualization



Petal Features as Primary Discriminators:

Analysis across all feature pairs reveals that **petal length and petal width** provide the strongest cluster separation. The visualization shows three distinct groupings:

1. **Lower-left cluster (Purple):** Setosa - characterized by small petals (length: 0.0-0.3, width: 0.0-0.2 normalized)
2. **Middle cluster (Blue-Green):** Versicolor - intermediate petal dimensions with some overlap
3. **Upper-right cluster (Yellow):** Virginica - largest petal measurements

Sepal Feature Limitations:

Sepal-based visualizations (length vs. width) show considerably more overlap between species, particularly for Versicolor and Virginica. This explains why petal features are

botanically recognized as more taxonomically informative—they exhibit greater inter-species variation relative to intra-species variation.

Cluster Centroid Positions:

The red "X" markers indicating cluster centroids are well-positioned in feature space, demonstrating that K-Means successfully identified the true centers of each species distribution. The centroids' clear separation in petal feature space (with minimal Euclidean distance to their respective class members) validates the clustering algorithm's convergence.

---

5. Misclassification Analysis

Error Pattern Analysis:

Of the 17 misclassified samples, the distribution is heavily skewed:

- **Versicolor → Virginica:** 3 samples (17.6% of errors)
- **Virginica → Versicolor:** 14 samples (82.4% of errors)

Root Cause Assessment:

1. **Feature Overlap:** Versicolor and Virginica occupy adjacent regions in feature space with substantial overlap in their petal dimension distributions. This biological similarity creates an inherently ambiguous boundary that even supervised classifiers struggle with.
2. **Sample Size Asymmetry:** Cluster 1 contains 61 samples (22% oversized) while Cluster 2 contains only 39 samples (22% undersized), suggesting the algorithm's boundary placement favors the Versicolor region due to local density patterns.
3. **Natural Variability:** Botanical taxonomy itself acknowledges morphological gradation between these species, particularly in hybrid populations. Some misclassifications may represent genuinely ambiguous specimens that challenge even expert identification.

Implications:

The 88.67% accuracy achieved through unsupervised learning is remarkably strong, especially considering the algorithm received no species label information during training. This performance validates K-Means as a legitimate approach for preliminary species identification in botanical surveys where labeled training data is unavailable.

Conclusions and Recommendations

Summary of Findings:

1. **K-Means successfully identified three natural clusters** in the iris dataset, achieving 88.67% agreement with botanical species classifications through purely unsupervised learning.
2. **Optimal cluster count (k=3)** was validated through multiple metrics: elbow method (inertia reduction), silhouette score (cluster quality), and ARI (ground truth agreement).
3. **Petal features (length and width) provide superior discriminatory power** compared to sepal measurements, enabling clear separation particularly for the Setosa species.
4. **Versicolor-Virginica confusion** represents the primary challenge, with 17 misclassifications concentrated at the boundary between these morphologically similar species.
5. **The unsupervised approach demonstrates competitive performance** against supervised baselines, validating clustering as a viable technique when labeled data is unavailable or expensive to obtain.

Report Prepared By: Aime Muganga – aimemuganga07@gmail.com

Analysis Date: December 2025
Dataset: Iris Flower Dataset (150 samples, 4 features, 3 species)
Tools Used: Python, scikit-learn, pandas, matplotlib, seaborn