Cluster Analysis of Penguin Morphological Data

Introduction

Clustering analysis is a powerful unsupervised machine learning technique used to group similar data points based on shared characteristics. In this project, we applied K-Means clustering to the `penguins.csv` dataset, which contains morphological measurements of different penguin species, to uncover distinct groupings based on their physical traits.

This analysis helps in understanding natural groupings in the dataset, potentially revealing species classifications or variations based on habitat or evolutionary traits. The optimal number of clusters was determined using the elbow method and silhouette score, resulting in five clusters.

Methodology

Data Preprocessing

1. **Data Cleaning**: Missing values were handled appropriately.
2. **Feature Selection**:
    - Culmen Length (mm)
    - Culmen Depth (mm)
    - Flipper Length (mm)
    - Body Mass (g)
    - Sex
3. **Data Scaling**: Standardization was performed using `StandardScaler` to ensure equal weight across all features.

Clustering Approach

1. **Determining Optimal Clusters**:
    - The elbow method was used to analyze the within-cluster sum of squares (WCSS) and suggested an optimal cluster count between 3 and 5.
    - Silhouette scores were calculated for different values of `k`, with `k=5` yielding a high score (0.520), confirming the best choice.
2. **Applying K-Means**:
    - The K-Means algorithm was run with `k=5` to segment the dataset into meaningful clusters.
    - The centroids were extracted to analyze the cluster characteristics.

Results

Cluster Summary Statistics

The following table shows the mean values of each feature within each cluster:

Cluster	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)
0	40.32	19.01	192.24	4034.64
1	45.56	14.24	212.71	4679.74
2	39.74	17.59	188.86	3410.68
3	50.96	19.20	199.08	3920.62
4	49.47	15.72	221.54	5484.84

These results indicate that Cluster 4 contains the largest penguins in terms of body mass, whereas Cluster 2 represents smaller penguins with shorter flipper lengths.

Cluster Centroids

The centroids represent the central points of each cluster, giving insight into the defining characteristics of each group:

Cluster	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)
0	40.32	19.01	192.24	4034.64
1	45.56	14.24	212.71	4679.74
2	39.74	17.59	188.86	3410.68
3	50.96	19.20	199.08	3920.62
4	49.47	15.72	221.54	5484.84

These centroid values confirm the distinctions among clusters, highlighting variations in body size, flipper length, and culmen dimensions.

Discussion

The clustering results suggest:

- Cluster 2 represents the smallest penguins in terms of body mass and flipper length.
- Cluster 4 has the largest penguins with the longest flippers and highest body mass.
- Culmen depth plays a significant role in distinguishing clusters, with deeper culmens observed in Clusters 0 and 3.

By analyzing these clusters, we can infer species groupings or adaptations based on morphology. Further validation could involve comparing these clusters to known species labels.

Conclusion

This clustering analysis effectively grouped penguins based on morphological measurements, uncovering patterns in their physical traits. The chosen model (`k=5`) was validated using silhouette scores, and the analysis provided insights into the characteristics of each cluster.

Future work could involve:

- Comparing clusters to actual species labels to assess accuracy.
- Incorporating additional environmental or genetic data for deeper insights.
- Exploring alternative clustering methods such as DBSCAN for density-based segmentation.

Appendix: Scaled Centroids

For completeness, here are the centroids after standardization:

Cluster	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)
0	-0.68	0.95	-0.62	-0.21
1	0.28	-1.49	0.84	0.59
2	-0.79	0.22	-0.86	-0.99
3	1.28	1.04	-0.14	-0.36
4	1.00	-0.73	1.47	1.59

These values show how each cluster differentiates after standardization.

Repository Structure

- `penguins.csv` – Raw dataset
- `clustering_analysis.ipynb` – Jupyter Notebook with analysis code
- `README.md` – Project overview