MACHINE LEARNING REPORT: ANTARCTIC PENGUIN CLUSTERING

Project: Population Segmentation through Unsupervised Learning

Analyst: Jorge C.

Algorithm: K-Means with StandardScaler

1. Executive Summary

This project applies Unsupervised Machine Learning techniques to group a population of Antarctic penguins based solely on their morphological attributes. Using the **K-Means algorithm**, physical patterns were identified to segment individuals into distinct groups, facilitating biological classification without labels.

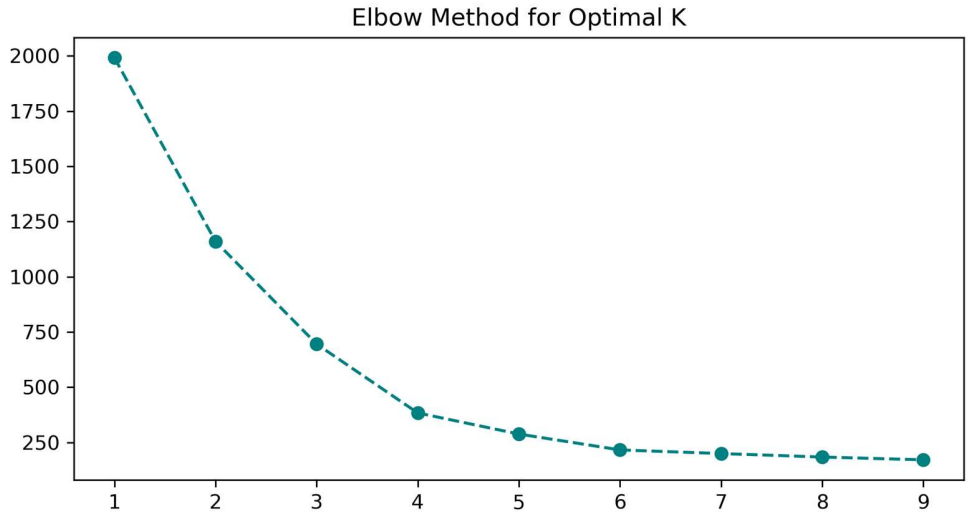
2. Methodology & Preprocessing

To ensure model stability, the following steps were implemented:

- **Cleaning:** Removal of null values to prevent distance calculation errors.
- **Encoding:** Categorical variables were transformed into numerical indicators.
- **Standardization:** All features were scaled to a mean of 0 and variance of 1.

3. Model Optimization: The Elbow Method

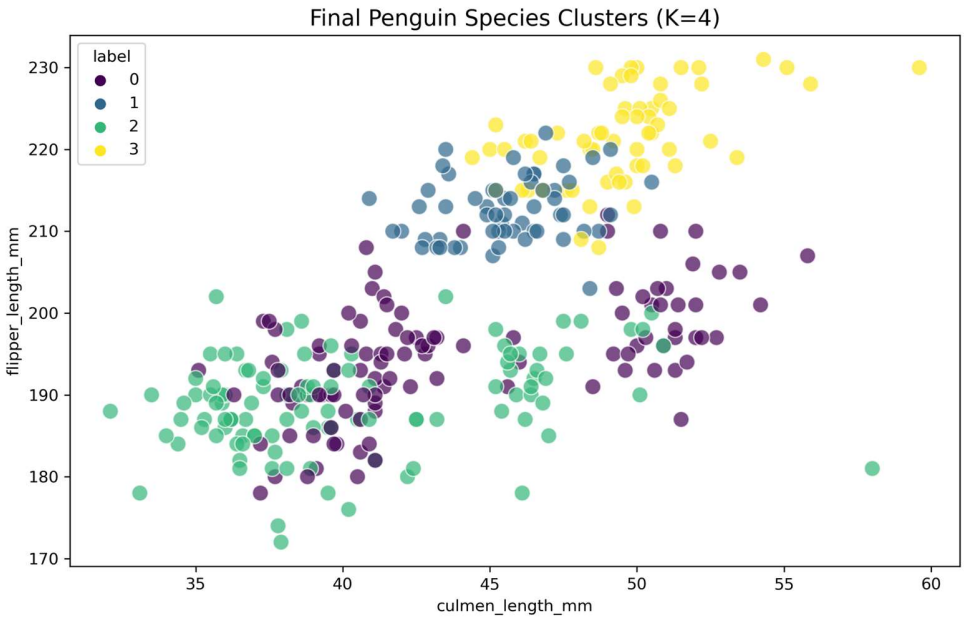
Determining the optimal number of clusters is critical to avoid over-segmentation or lack of detail.



Analysis: The "elbow" is clearly visible at **$k=4$**. Beyond this point, the reduction in inertia (WCSS) becomes marginal, confirming that 4 clusters provide the best balance for this population.

4. Cluster Visualization & Findings

The final model segments the population into four distinct morphological groups.



- **Cluster Separation:** The scatter plot (Culmen Length vs. Flipper Length) shows minimal overlap between clusters, indicating high model performance.
- **Biological Mapping:** By cross-referencing with the stat_penguins data, these clusters accurately represent species-level differences and gender dimorphism.

5. Final Conclusion

The K-Means algorithm effectively "rediscovered" the biological structure of the Antarctic penguin population. This model is now ready to be used as a classification tool for unlabelled field data.