Clustering Species Report

This project involves working with a dataset containing observations of different species based on 15 covariates. The data does not include any pre-defined response variable, suggesting that we're dealing with an unsupervised machine learning problem. The objective is to accurately classify our observations into different species, with the possibility of having 4 species.

The method utilized in this analysis is k-means clustering, which is an unsupervised machine learning algorithm. The objective of k-means clustering is to partition a dataset into distinct clusters where each observation belongs to the cluster with the nearest mean. This is accomplished through an iterative process where each observation is assigned to the nearest centroid, and then the centroids are updated based on the new groupings. This process continues until the centroids stabilize.

Given that I don't have pre-defined class labels for our observations, I leverage certain samples provided to label the resultant clusters. The labels are based on 4 clusters identified in the dataset. The dataset contains an identifier for each observation and 15 covariates (locus_1 to locus_15). These loci are presumably features that can help distinguish between different species. The data was cleaned by standardizing all covariate values, ensuring that the algorithm treats all features equally, and the results are not skewed by the different scales of the original data

Initially, a k-means model was built with k=4, representing four clusters. The output of this model was used to assign the cluster labels to the dataset based on the predetermined samples for each species. As per the given conditions, the cluster of sample_3 was labelled as species1, the cluster of sample_9 as species2, and the cluster of sample_6 as species3. Any remaining clusters were labelled as species4.

After labeling, the cluster labels were mapped to the corresponding species in the submission dataset and then exported to a .csv file for evaluation.

In this project, I applied the k-means clustering algorithm to a dataset consisting of observations with 15 covariates, without any pre-defined response variable, to classify the observations into species. The dataset was standardized to ensure all features were treated equally, and the algorithm was initiated with 4 clusters. The clustering results were labeled as species based on pre-determined samples, after which they were mapped to the submission dataset. The model performed reasonably well, with a Kaggle score of 0.85777. This project underscores the applicability of unsupervised learning, particularly k-means clustering, in species categorization tasks based on multiple covariates.