Penguins in Antarctica

Machine Learning

Final Exam – Ryan Spake

12/12/2021

For this study, we are taking a look at data gathered about penguins at the Palmer Archipelago (Antarctica). This data has been acquired through Kaggle and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER (Long Term Ecological Research Network).

The problem in this study is that as data is collected each year about penguins in antarctica, it is difficult to draw insights from the raw data and compare over time. The objective of this study is to use the data gathered at the Palmer Station to cluster the population of penguins sampled and make it easier to identify clusters and characteristics of these penguins from the data so that insights can be made more efficiently as we continue to sample these penguins in the future. Such characteristics include species, the density of such species on each of three islands that were sampled from, and physical measurements.

To achieve the goal of making it easier to identify the clusters and characteristics of the penguins, our approach uses both K-means and Hierarchical Clustering. This approach provides both a cluster plot (from K-means) and cluster characteristics (from Hierarchical clustering).

While doing K-means clustering, we can clearly illustrate the three different clusters using a cluster plot. This plot allows use to select individual penguins from the data and track which cluster they belong to and relate them back to the cluster characteristics that are developed in Hierarchical clustering.

Using the silhouette approach, we can clearly identify that 3 clusters exist in this dataset. As shown below, three clusters yielded the highest average silhouette width, making it the most optimal number of clusters.

Chart, line chart

Description automatically generated

Now that we have determined the most optimal number of clusters, we can create a cluster plot using K-means to better illustrate the clustering. The graph below shows our three clusters obtained using K-means clustering.

Chart, scatter chart

Description automatically generated

The three clusters above separate the penguins into our three penguin species sampled. These clusters where created using all numerical data from the dataset, and it is clear that each of these three species have very clear statistical distinctions. The characteristics that separate the species into these three distinct clusters can be found in the cluster characteristics we obtain through Hierarchical clustering.

Because we are using a different method of clustering to show characteristics, we want to revalidate our optimal number of clusters for the new clustering algorithm. This was achieved by creating an Agnes dendogram that illustrates our clusters in a tree format. Shown below, our Agnes dendogram illustrates the largest jump in groups at a height of 30, indicating three clusters, affirming the findings from earlier.

Diagram

Description automatically generated with low confidence

Now that we have confirmed that three clusters will be optimal for Hierarchical clustering, we can make our characteristics. The chart below illustrates the values of the centroids of each cluster, for each value they were evaluated on.

Chart, bar chart

Description automatically generated

From the illustration above, we can gain much greater insight into the three clusters of penguins.

From this we can determine that Cluster 1 has the largest number of the Adelie penguin. Cluster 2 has a very large population the Chinstrap penguin. Lastly, cluster 3 has the Gentoo penguins. An interesting point to notice from this information, is that cluster two has very low amounts of the other two species, while the other two clusters have a small gradient as population of the minority species are denser.

Using this information, we can gleam insights into characteristics of the clusters and the species in them. Cluster two has a high density of Chinstrap penguins, as well as the largest Culmen (beak) length and depth and the largest density coming from Dream Island. This can inform us about the characteristics of the chinstrap penguin, such that they live mostly on Dream Island and have long and deep Culmens. Similarly, Gentoo that live on Biscoe Island have the largest flippers and body mass based on the graph.

This methodology allows us to easily cluster the data from our dataset and present them in an easy-to-read fashion. By doing this we can more easily make insights about the lives of these penguins through time. As more data is collected over time, we can compare the graphs then to the graphs we have made now to clearly see changes in these penguins.