Project Evolution

Originally, I wanted to create a bird migration map using an AI website builder, like Lovable, to visualize the path of migratory birds. However, when I searched the web to find a dataset with migration paths, I was unable to find one that worked. I did come across one that had migration start and end locations, which looked promising, but I couldn't download it; it showed that a 15 megabyte file would take two days to download. I ended up giving up looking for a migration dataset, since it was clear I wasn't going to find one.

During my search for a migration dataset, I had also discovered another database called AVONET, which has “morphological, ecological and geographical data” from nearly every single extant bird species, all organized taxonomically. There were three trait datasets all of which had the same traits I could choose from, all of which had the same traits, so I just chose the e-bird one because it had a metadata list and sounded cooler. I then found out about clustering, which seemed interesting, so I decided to explore clustering along with a group of birds in the dataset. I ended up choosing the Caprimulgiformes, partly because it was the largest order after Passeriformes (before, anyway), partly because I had researched hummingbirds and the like when I was bored a while ago, and partly because the great eared nightjar is called the dragon bird.

Progress

I was able to use scikit-learn’s K-means clustering to cluster the birds into different groups. At first, I just tried to cluster the data without scaling it, which resulted in clusters with all the birds with a large mass or all the birds with long tails, rather than each bird with its own family, although the clusters were surprisingly accurate in hindsight. Scaling the data so that the variation of each trait becomes more even allows other, perhaps more important traits in determining a family, to shine through, so now the birds in each family are mostly put together.

Clustering aside, I also used decision trees and normalized cluster plots, courtesy of my mom, to visualize and analyze the data. At first, the decision trees used a train test split, which would be good when trying to predict something. However, this situation called for analysis, not prediction, so I reasoned that a train test split would not be necessary. Using all of the data for the decision tree would provide a more accurate insight into how a dataset was clustered. The decision trees were very helpful for describing each cluster, and were both stable and accurate - no matter the random seed, the decision trees remained nearly identical. There was also a mishap while plotting the normalized cluster plot, which was that the labels were reversed. From this, I learned that sometimes, one should not just trust that the machine is right - using external knowledge can help in making sure that nothing went wrong along the way. In fact, I only discovered it was wrong when I was analyzing the data and realized the cluster plot showed that hummingbirds have tiny beaks (which is obviously incorrect). This is what caused me to take a deeper look at the code and get rid of one, single, pesky line that reversed everything. The correct cluster plots were useful in seeing which traits certain families that had been clustered on their own had, how each family compared to each other, and how some families are very unique.

Next Steps

The next step I would take for this is to cluster the bird families that were in cluster 2, which are the nightjars, owlet nightjars, swifts, and potoos. This way, I can see how those species may be more similar to each other than to the others, even though swifts are most closely related to treeswifts and hummingbirds. Also, I want to cluster all the former Caprimulgiformes/now Strisores (minus the oilbird) into seven clusters, the number of families there are supposed to be, to see if they would all be clustered correctly, or if some species are more similar to families other than their own. One more thing I would like to do is to compare them to owls, which look quite similar to many of the nightbirds, which are the nocturnal birds in the clade Strisores. Beyond this, however, I do not plan to take this any further, since this project is mainly focused on exploring and learning about clustering, and birds of course, rather than being an expanding and developing project to solve an issue or create a functional website. Spending too much time on specifically clustering could be wasting time that can be better spent learning about other, equally as interesting things.

In Retrospect

I would not take on this project any differently than I did. As mentioned before, this project is meant to help me learn about clustering and birds, so every mistake along the way has been valuable in learning what not to do. The learning process involves making mistakes, so doing things right the first time wouldn't help me know what to avoid; the times after that, I could make that mistake anyways and have to figure out what went wrong and how to fix it. Overall, I learned a lot in every step of the process, whether I did it right or wrong.