Final Project Report: Can morphology and performance predict frog microhabitat?

Nigel Anderson 12/09/2022

Introduction

A question that has perplexed evolutionary biologists for a long time is why certain morphological features (i.e., body structure measurement such as leg length) repeat in similar habitats. Daniel Moen sought to address this question by studying the evolution of morphology and performance (i.e., physical performance such as jumping or swimming) in certain microhabitats (i.e., small habitat within a larger habitat such as leaf litter or fully aquatic) for frogs. Using mostly museum specimens and some live specimens of 191 different species of frogs, he found that frogs from different microhabitats have unique morphology such as leg length and body mass. He also found that jumping performance is about the same across microhabitats for most frogs. These findings interestingly suggest that jumping performance does not depend on frog morphology. So why do frogs in different microhabitats have different morphologies? This is an interesting question but beyond the scope of what my project aims to accomplish. Rather I am interested in using this phenomenon of frogs having unique morphology to predict what microhabitat the frog exists in. This project has important implications for evolutionary studies and conservation efforts. For example, many frog species that are considered either extremely endangered or extinct, we have minimal data. However, the data we do have on them usually are morphological data. So, if I can build a model to predict microhabitat from morphological measurements, we could predict what likely microhabitats these endangered or extinct species likely occupy.

To build this classification model I am using the dataset from Daniel Moen's paper from 2019. There are 899 data points and nine feature variables. The target variable for the project is going to be microhabitats, which in this dataset is a categorical variable with six categories (arboreal, semi-aquatic, fully aquatic, terrestrial, burrowing, and torrential). All features are continuous variables of either morphological traits or performance measurements. There is snout-vent-length in millimeters (mm) which is the measurement from the nose to the back end of the frog. The leg length in mm. The mass of the frog when it is alive in grams (g). The mass of the frog as a preserved museum specimen in g. The mass of the leg as a preserved museum specimen in g. The max velocity of the frog's jump in meters/second (m/s). The max velocity of the frog's swimming in m/s. And finally, the max acceleration of the frog's swimming in meters/second^2 (m/s^2).

Exploratory Data Analysis

After performing EDA, I find that my target variable, microhabitats is imbalanced. The category arboreal and terrestrial make up 39.4% and 34.7% of the data points respectively (**Figure 1**). Whereas semi-aquatic makes up 10.2%, burrowing makes up 6.67%, torrential makes up 5.12%, and aquatic makes up 3.89% of the data points (**Figure 1**)

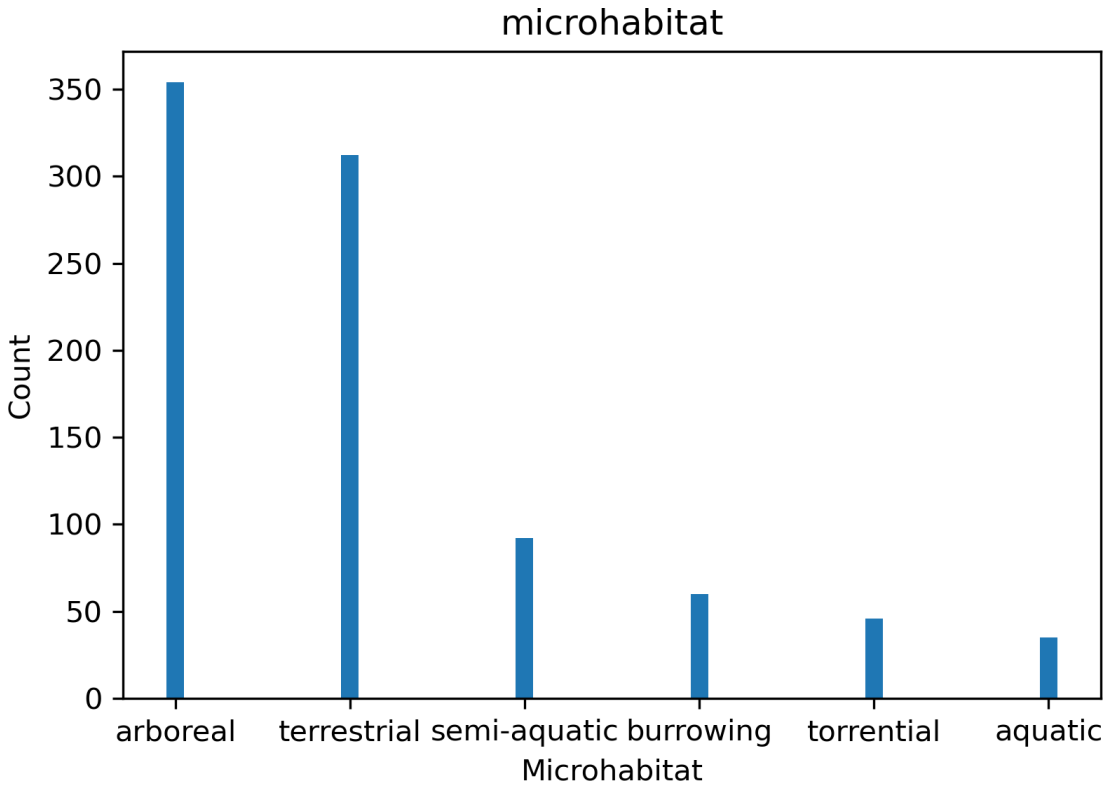

Figure 1: This plot shows the frequency of data points per microhabitat category. On the y-axis we have count, which is the frequency of data points and on the x-axis we have microhabitats.

An interesting result through the EDA is that it seems that frogs occupying environments with water (torrential, aquatic, and semi-aquatic) have longer leg lengths compared to the other microhabitats (**Figure 2**).

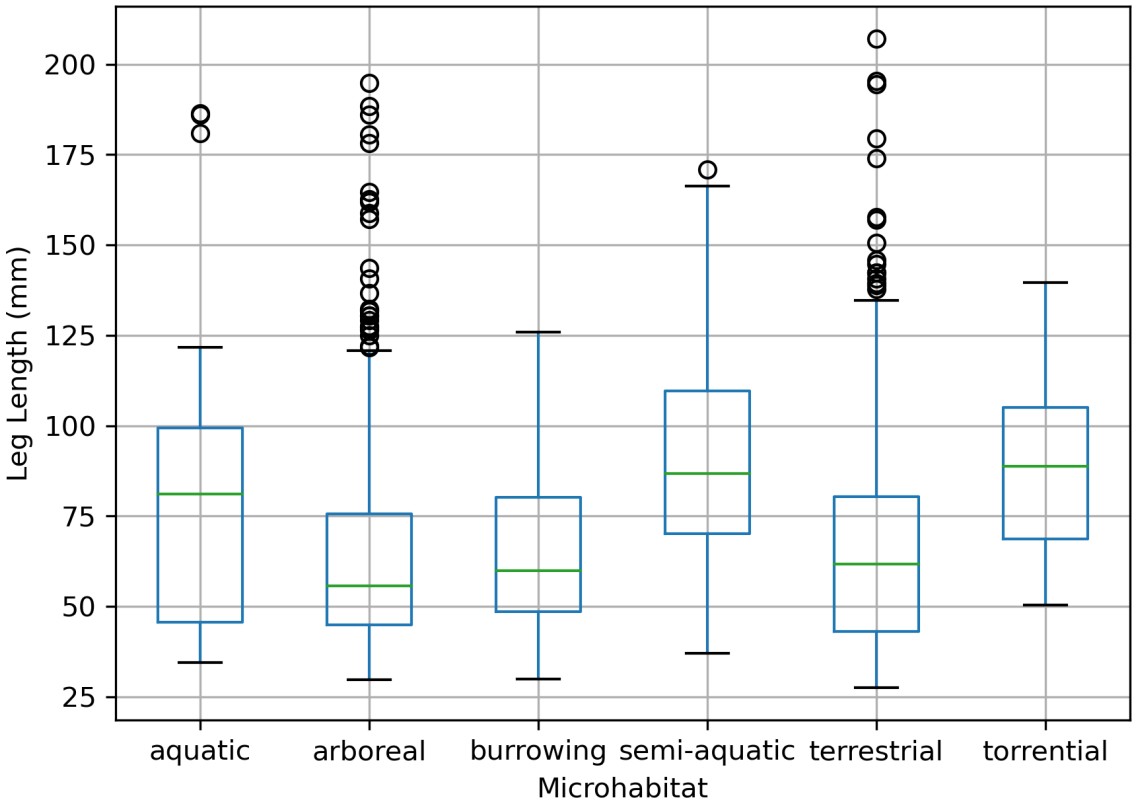

Figure 2: This boxplot shows the relationship between leg length and microhabitat. On the y-axis we have leg length measured in millimeters (mm) and on the x-axis we have microhabitats.

Another interesting result is the amount of variation seen with the snout-vent-length (svl) across microhabitats (**Figure 3**). Additionally, it is interesting that leg muscle mass seems to be different across the microhabitats (**Figure 4**). The species that occupy water-based microhabitats seem to have larger leg muscle mass compared to frogs that occupy terrestrial habitats.

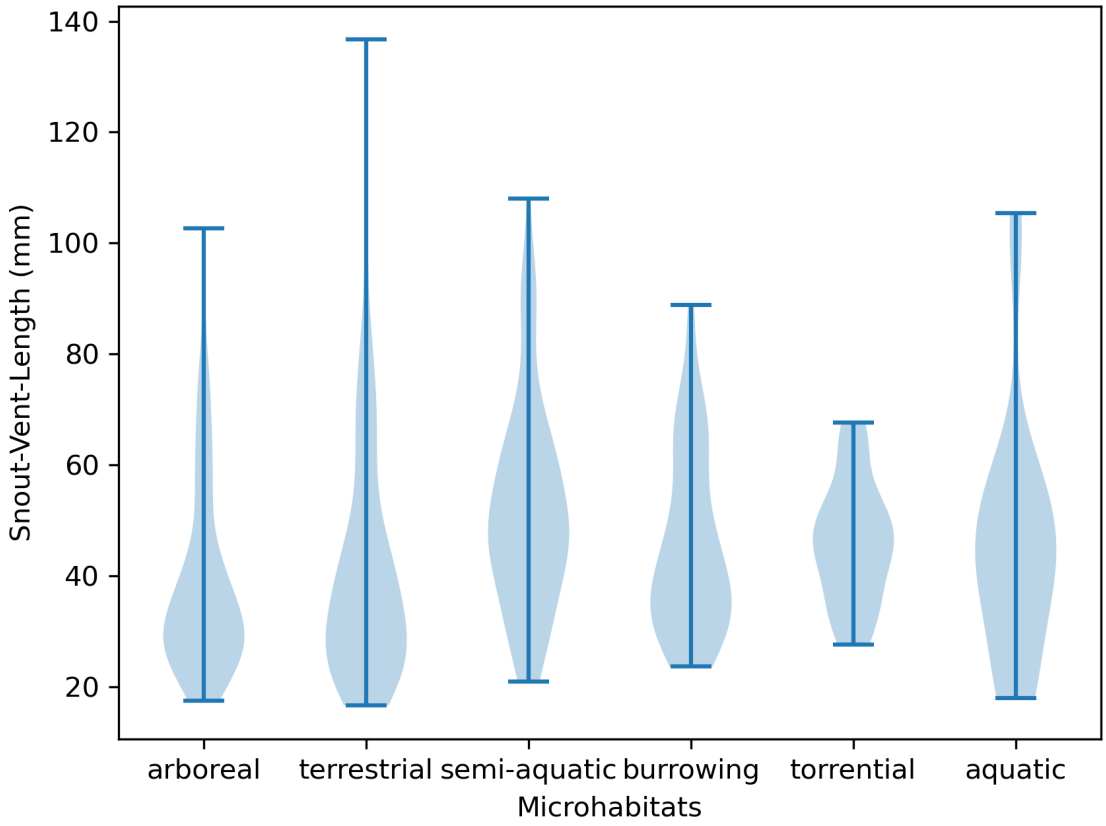

Figure 3: This violin plot shows the relationship between snout-vent-length and microhabitat. On the y-axis we have snout-vent-length measured in millimeters (mm) and on the x-axis we have microhabitats.

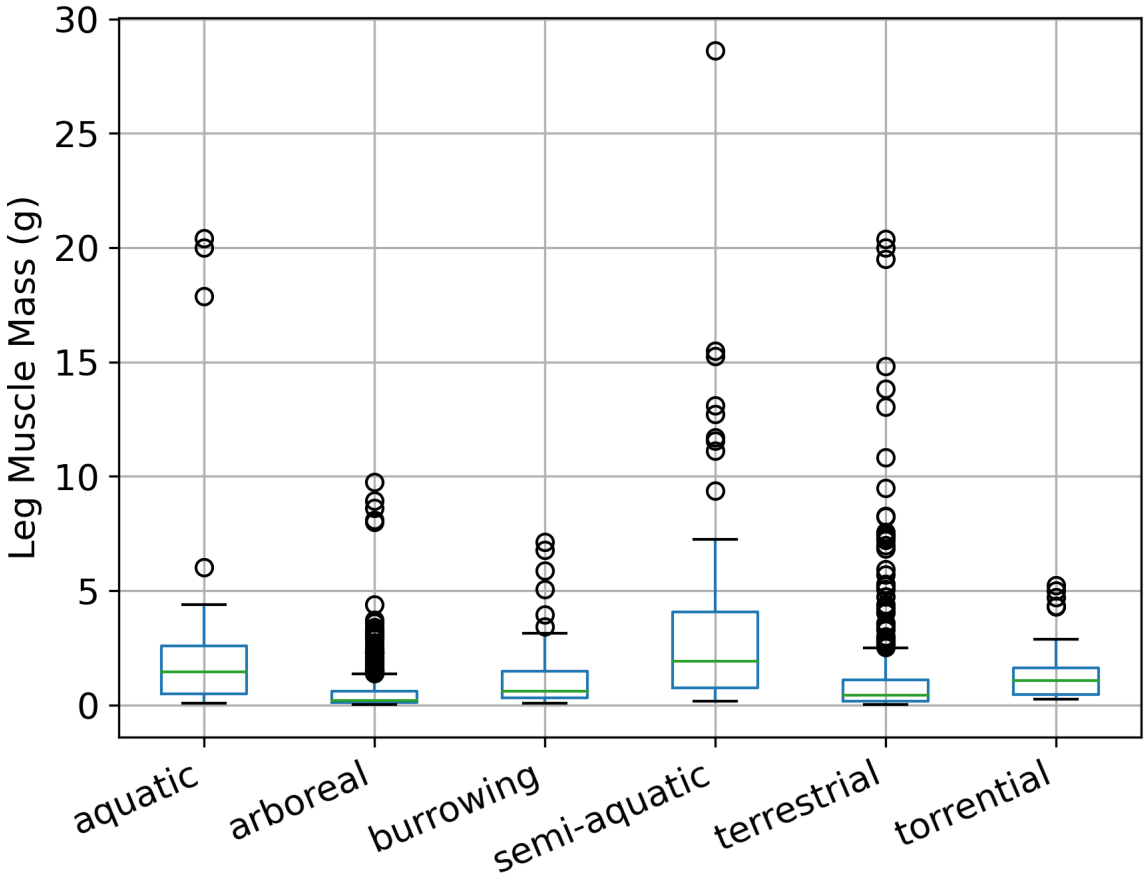

Figure 4: This boxplot shows the relationship between leg muscle mass and microhabitat. On the y-axis we have leg muscle mass measured in grams (g) and on the x-axis we have microhabitats.

Methods

The dataset is non-IID with group structure based on species. So, there are multiple individuals per species. It is not time-series data. Because it is imbalanced, I need to stratify the split. So, I used stratified group k fold to both account for imbalance and for the non-IID structure of the data. I used nsplit equal to 3 to make it 60% training and 40% other. I then used nsplit equal to 2 on the other data to make it 20% validation and 20% testing.

For preprocessing, I used MinMaxEncoder for all my features. This is because they are all continuous features and have a min and max to their potential values. My preprocessed data has 4 features and 899 data points. Five columns were dropped because 75% of the data points for these columns were missing.

Because the target variable is categorical, I decided to use logistic regression, support vector classification (SVC), K neighbors classifier, random forest classifier, and xgboost models in my machine learning pipeline. For all of the models I used accuracy as my test score. I looped through 10 random states and took the mean and standard deviation of the best accuracy scores from the 10 random states to measure uncertainty. For the logistic models, I ran l1 and l2 regularization. For the SVC model I tuned the hyperparameters gamma (1e-3, 1e-1, 1e1, 1e3, 1e5) and C (1e-1, 1e0, 1e1). For the K neighbors classifier model I tuned the hyperparameter n-neighbors (1, 2, 3, 4, 5, 6). For the random forest classifier I tuned the hyperparameter max depth (1, 2, 3, 4) and max features (0.5, 0.75, 1). Finally, for the xgboost model I tuned the hyperparameter max depth (1, 2, 3, 4).

Results

The baseline accuracy for my models is 39.40 % which is the percentage of frogs in the arboreal microhabitat. The models all preformed higher than the baseline accuracy (**Table 1**), however, the logistic regression models all had a standard deviation that put their accuracy range partially below the baseline. The next worst model was SVC with its mean accuracy at 63.83%, its lower range reaching 56.29%, and its upper range reaching 71.37%. The Random Forest Classifier model had a mean accuracy of 69.20%, with its lower range reaching 60.37%, and its upper range reaching 78.03%. The K Neighbor Classifier model had a mean accuracy of 73.52%, with its lower range reaching 66.68%, and its upper range reaching 80.36%. Finally, the best predictive model was the Xgboost model with a mean accuracy score of 88.45% with its lower range reaching 85.65% and its upper range reaching 91.25%.

Table 1: The models used with the mean and standard deviation of their accuracy scores as well as the comparison against the baseline accuracy.

Model	**Mean Accuracy Score**	**Accuracy Standard Deviation**	**Percent Above Baseline**
Logistic Regression	41.09%	6.535%	1.690%

L1 Regularization	41.09%	6.535%	1.690%
L2 Regularization	41.09%	6.535%	1.690%
SVC	63.83%	7.543%	24.43%
K Neighbor Classifier	73.52%	6.839%	34.12%
Random Forest Classifier	69.20%	8.834%	29.80%
Xgboost	88.45%	2.796%	49.05%

Moving forward with the Xgboost model, I calculated feature global importance using permutation importance (**Figure 5**), gain, weight, cover, total gain, and total cover (**Table 2**). I found that leg muscle mass and leg length seem to be the most important features for this prediction. The worst feature for this prediction seems to be preserved total body mass followed by SVL.

Table 2: The global importance measurements along with ranked features.

Global Importance Measurement	**First Feature**	**Second Feature**	**Third Feature**	**Fourth Feature**
Permutation Importance	Leg Muscle Mass	Preserved Mass	Leg Length	SVL
Gain	Leg Muscle Mass	Leg Length	Preserved Mass	SVL
Weight	Leg Length	Leg Muscle Mass	SVL	Preserved Mass
Cover	Leg Length	Preserved Mass	SVL	Leg Muscle Mass
Total Gain	Leg Muscle Mass	Leg Length	SVL	Preserved Mass
Total Cover	Leg Length	SVL	Leg Muscle Mass	Preserved Mass

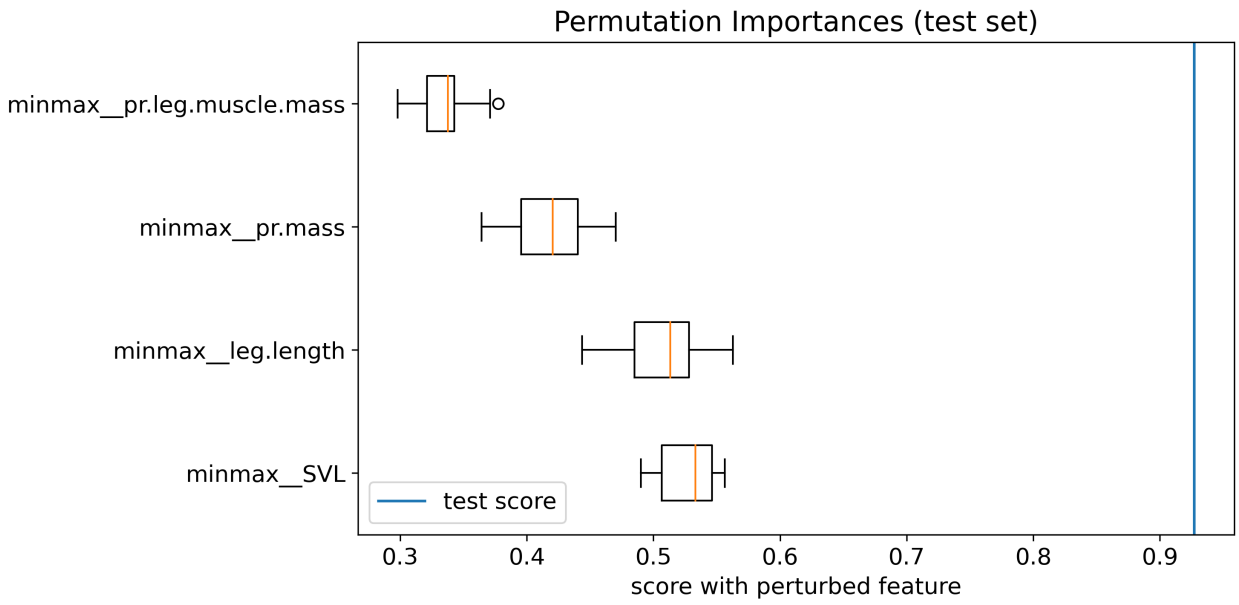

Figure 5: This boxplot shows the permutation importance of the features in the xgboost model. On the y-axis are the scaled features. On the x-axis are the accuracy scores with perturbed feature.

Next, I calculated local importance using SHAP values. I found that for index 0, leg muscle mass and preserved total body mass contribute positively to the prediction. Leg length and svl contribute negatively to the prediction (**Figure 6**). For index 50, leg length and svl contribute positively to the prediction (**Figure 6**). Leg muscle mass and preserved total body mass contribute negatively to the prediction. For index 100, leg length and svl contribute positively to the prediction (**Figure 6**). Leg muscle mass and preserved total body mass contribute negatively to the prediction.

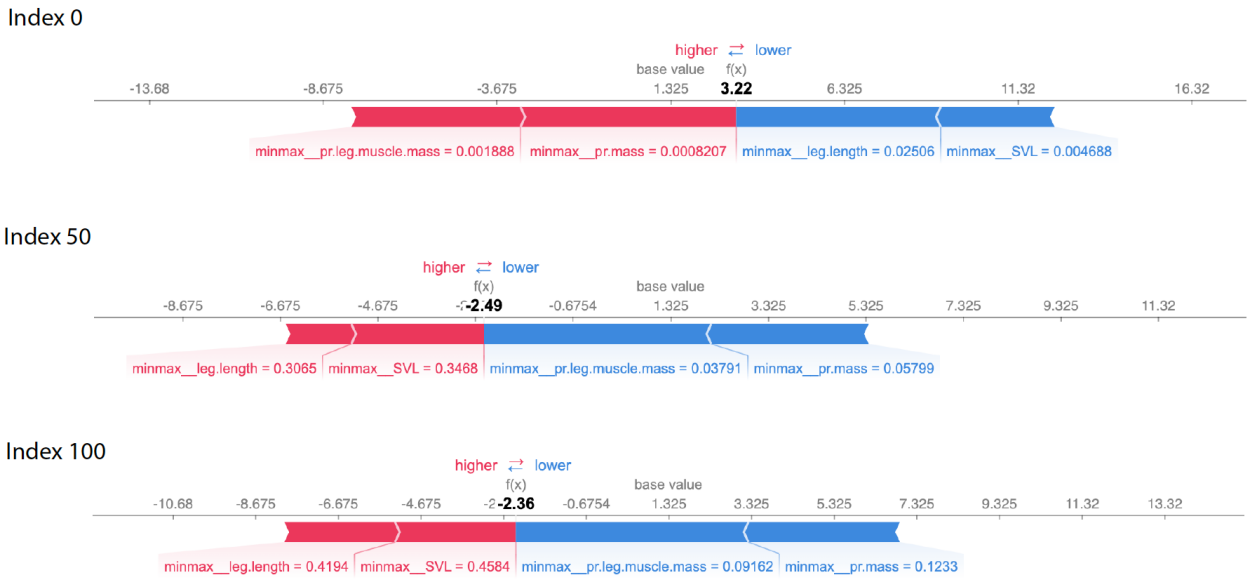

Figure 6: Force plots for indexes 0, 50, and 100. The number line represents SHAP values. The red indicates features that contribute positively to the prediction and blue indicates features that contribute negatively to the prediction.

All in all, my model provided very interesting insight into what morphological features are useful in predicting the microhabitats of frogs. I find it very interesting that leg morphology provides the most information for the prediction, however, it is not surprising. One of the most iconic features of a frog are its hindlimbs. They are the primary tool for their locomotion and predator avoidance. So, I would expect that frogs that use their hindlimbs to swim would have different leg morphology than those that just use them to hop or crawl. It is very cool though to have an accurate model now to predict microhabitat from morphology. Frogs all around the world are disappearing, however, we possess many museum specimens of these disappearing/extinct species that were collected when the species were more abundant. With a model such as the one used in this project, we can predict the ecology and some subsequent behaviors based solely on morphological measurements of the museum specimens. This is a great first step for the animal behavior world in recovering information from recently 'lost' species.

Outlook

The weakest aspect of my model is the number of features. I think it would help the model's predictive power if I added more features such as toe morphology, eye size, and phylogenetic independent contrast values. I also think it would be beneficial to have a more balanced data set. Additionally, I think it would be useful to calculate other test scores besides accuracy such as f1 score.

References.

Moen, Daniel S. (2019), Data from: What determines the distinct morphology of species with a particular ecology? The roles of many-to-one mapping and trade-offs in the evolution of frog ecomorphology and performance, Dryad, Dataset, https://doi.org/10.5061/dryad.n07742q

Moen, Daniel S. (2019). What determines the distinct morphology of species with a particular ecology? The roles of many-to-one mapping and trade-offs in the evolution of frog ecomorphology and performance. *The American Naturalist*. 194(4).

Github repository.

https://github.com/nandersonk/Frog-Microhabitat-Classification