In this project, we will use the techniques learned in class to develop models to classify frogs using only their calls. Specifically, we will use the Anuran Calls Dataset,[1] which provides features extracted from the frog call audio as well as the family and species of the frog that produced the call. The tasks detailed below will guide you through the process of training and evaluating classifications models with this dataset.

1. We have split the dataset into a training and test set for you to use. Load the training and testing datasets named `train.csv` and `test.csv`. There are 22 features named MFCCs_1, MFCCs_2, ... MFCCs_22 and two categorical labels named Family and Species.

2. Train classification models for each classification task, model type and multi-classification technique listed below (there should be $2 \times 3 \times 2 = 12$ models). Aim to get as close as possible to (or better than) the target testing errors specified in Table 1 and Table 2. You will likely need to tune the hyper-parameters with grid search, random search, or related techniques.

   (a) *Classification Tasks:* species classification, family classification

   (b) *Model Types:* SVM with the original features, Logistic regression with the original features, SVM with RBF kernel

   (c) *Multi-classification Techniques:* one-vs-one, one-vs-rest

3. Report your testing errors for all trained models in two tables; one for family classification and one for species classification. Present the confusion matrices for your best-performing models for each of the two tasks. Based on these results, discuss the following:

   (a) Which model types (SVM, logistic regression, or kernel SVM) and multi-classification techniques (one-vs-rest or one-vs-one) performed the best and worst? Do you have any intuitions as to why this is the case?

   (b) Was it more difficult to predict the family or species? Why do you think this is the case?

4. Record the computation time required for classifying the test data set using each model. Take the average of five trials and report the results in the form of a table showing the average time for each model. Based on these results, discuss the following:

   (a) Which model types and multi-classification techniques are the least and most computationally intensive to use for prediction? Why do you think this is the case?

   (b) Can you observe any apparent trade-offs between testing accuracy and computation?

[1] You can read about the dataset here: https://doi.org/10.24432/C5CC9H

5. Examine the trained models to evaluate the relative importance of different features. Referring to the dataset website as needed, discuss the following:

   (a) Which features appear to be the most and least important in classifying frogs? Do your observations make practical sense?

   (b) Are some of the models you trained easier or harder to interpret? What characteristics of these models do you think might contribute to their interpretability?

6. Suppose that you are designing a classification model to be implemented in an e-commerce platform that is used by frogs. It is important that the platform can correctly identify the specifies of each frog user to ensure that it is suggesting the best products for that specific species. In particular, misclassifying each frog user costs the platform $x$ dollars (where $x$ is an arbitrary positive real number). Furthermore, the platform has to pay $y$ dollars for each second of computation time (where $y$ is an arbitrary positive number). Assume that your measured testing accuracy represents the ground-truth and that your measured computation time reflects the actual computation time in implementation. For each of your (six) species classification models, determine what range of $x$ and $y$ (if any) results in that model having the lowest expected cost among all models. Discuss your results and any limitations of this model selection approach that you can think of.

| | **One-vs-one** | **One-vs-rest** |
|---|---|---|
| **SVM** | 4% | 7% |
| **Logistic Regression** | 6% | 10% |
| **SVM with RBF** | 1% | 2% |

Table 1: Target testing error for family classification.

| | **One-vs-one** | **One-vs-rest** |
|---|---|---|
| **SVM** | 6% | 9% |
| **Logistic Regression** | 5% | 12% |
| **SVM with RBF** | 4% | 6% |

Table 2: Target testing error for species classification.