Classifying Fish Species Using Sonar Data

Vishu Wadhawan

Department of Statistical Sciences, University of Toronto

Course: STA 2453

Instructor: Dr. Vianey Leos Barajas

April 16, 2025

1 Abstract

This project explores the possibility of using sonar data to classify three fish species: trout, whitefish, and pass. The main objectives were to develop a classification model and pinpoint the most important features for predictive modeling; we used the frequencies from the sonar instrument as predictors for our models. We applied dimension reduction and shrinkage techniques to the data, employed feature selection, and then fed the reduced feature sets into K-nearest neighbors (KNN) and Support Vector Machine (SVM) classifiers. The analysis produced underwhelming results with relatively low accuracy rates. However, the gaussian kernel proved to increase prediction accuracy, suggesting that non-linear relationships may exist within the data. Overall, the underlying patterns of the sonar data are difficult to interpret. Future work should focus on capturing the non-linear patterns with advanced modeling techniques. This may improve classification accuracy and provide deeper insights into the relationships between sonar and fish species.

2 Introduction

2.1 Motivation

Water ecosystems are considered vital to the preservation of global health but are notoriously difficult to explore. In an ideal economy, scientists would be able to monitor these ecosystems using underwater laboratories, however, this has proven to be both costly and dangerous. Thus, recent research has developed new ways to learn about water ecosystems using sonar. Sonar (**S**ound **N**avigation and **R**anging) instruments emit sound waves into the water; if an object in the path of these pulses, the sound bounces off the object and the reflected waves can be measured. The returning signal strength (in decibels) is traditionally measured by the following equation:

$$dB = 10 \times log_{10} \frac{P_{received}}{P_{reference}} \tag{1}$$

where $P_{received}$ is the power of the reflected sounds and $P_{reference}$ is the omitted sound from the instrument.

2.2 Description of Data set

The sonar data was collected by a group of researchers at the Algonquin Provincial Park. Their processed data set included 14575 observations and 484 features. Each row (i.e observation) corresponds to a single fish from a particular species which equates to 7248 trout, 4285 whitefish, and 3042 small mouth bass. Some notable features of the data include weight, girth, length, and the reflected signal strength (dB) from an emitted sound frequency. The emitted pulses from from the sonar instrument are labelled "F258" and "F259" (F for frequency) for example and they range from 46 KHz to 259.6 KHz; amounting to 425 different frequencies. One important aspect of this data set is that each fish is similar in size with respect to their specific species. As per figure 1, the length of trout ranges from 343 mm to 648 mm, the length of the whitefish ranges from 220 mm to 446 mm, and finally the bass lengths range from 190 mm to 479 mm. Moreover, there are a significant number of missing responses for certain sound frequencies, though no columns have entirely missing data. Specifically, the trout species has up to 2400 missing entries for several sound frequencies while both whitefish and bass have < 1000 missing entries in only a few categories.

2.3 Project Description

This project will aim to classify between trout, whitefish, and bass using the processed sonar data. The objective is to develop classification models with reasonable accuracy and pinpoint the most significant features.

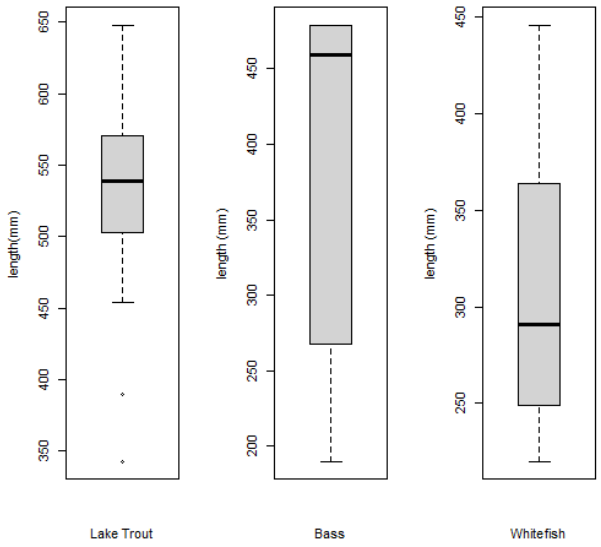

Figure 1: Showcasing the lengths of each fish species using boxplots in R.

3 Explatory Data Analysis

3.1 Data Pre-Processing

For the analysis, only the frequencies emitted by the sonar instrument were retained as predictor variables. The other variables were excluded, as they appeared to be more closely related to the fish's physical characteristics. Given that the fish within each species shared similar phenotypes, it was anticipated that the remaining variables would not significantly influence the strength of the signal reflected back to the instrument. This decision reduced the dataset from 484 features to 425. Moreover, all the missing values were removed from the data which reduced the number of observations to 10333. This modified data set had 4101 trout, 2736 bass, and 3496 whitefish.

3.2 Identifying Trends

The classification model for this project needed to have two key characteristics: 1) The model needed to accurately predict a particular fish species given a new sample of sonar data 2) The model needed to identify strong associations between certain frequencies and

a particular species of fish. For example, do trout have a unique signal at a frequency of 45 KHz? Do whitefish have a unique signal at 247 KHz? These are only examples of questions that we hope the predictive model can answer. However, for the preliminary data exploration aspect, the goal was to discover any trends between the frequencies and the reflected signals, hoping to identify differences between the three fish species. Note that each individual fish has repeated measurements in this dataset. This means that a single fish was 'blasted' with sonar from 45Hz to 247Hz a handful of times, and the reflected signal strength was recorded for each event. In order to visualize any potential trends, the data was manipulated acccordingly:

- Filtered the original set by species and created 3 new datasets (one for each species)
- For each new data set we used the 'groupby' function to group the data set by its individual fish ID number, we then computed the average reflected signal strength for these 'groups'
- We then took the transpose of the aforementioned dataset and computed the average reflected signal strength: this computed the mean signal strength across a single frequency. For example, there were 21 individual trout's that were measured and each trout was 'blasted' with a sound pulse of 45Hz. The reflected signal from these blasts were recorded and we computed the average signal strength across the 21 individuals. Note that missing entries were removed from the calculations beforehand
- There were now three new data sets: one for each species. The rows are the frequencies ranging from 45Hz to 247Hz and the variable of importance is 'Average Frequency' (i.e the average signal strength across all the individual fish belonging to a particular species).

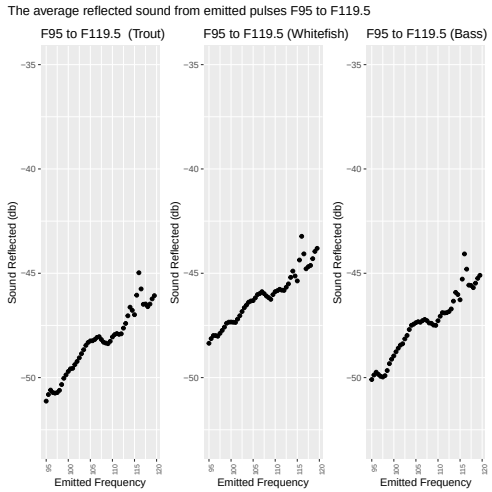

Figure 4: Comparing the reflected sound pulse from each fish between F95 - F119.5

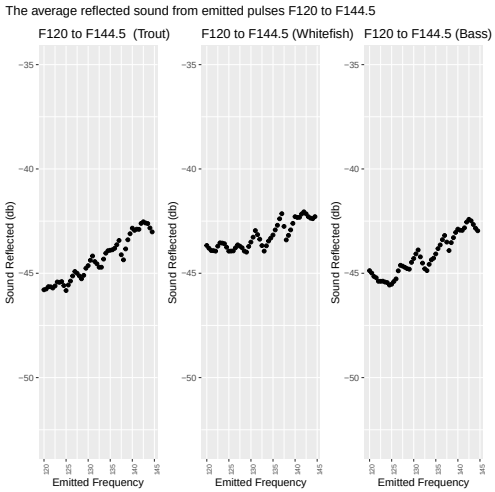

Figure 5: Comparing the reflected sound pulse from each fish between F120 - F144.5

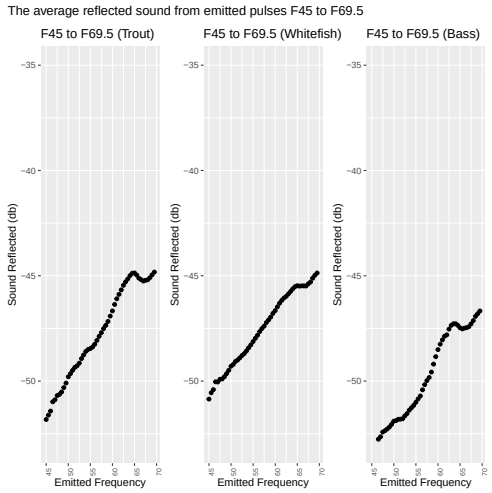

Figure 2: Comparing the reflected sound pulse from each fish between F45 - F69.5

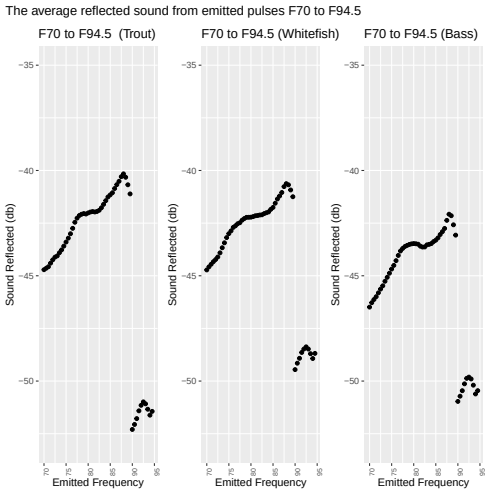

Figure 3: Comparing the reflected sound pulse from each fish between F70 - F94.5

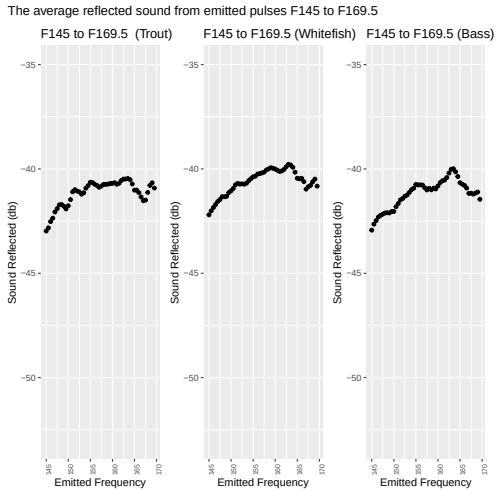

Figure 6: Comparing the reflected sound pulse from each fish between F145 - F169.5

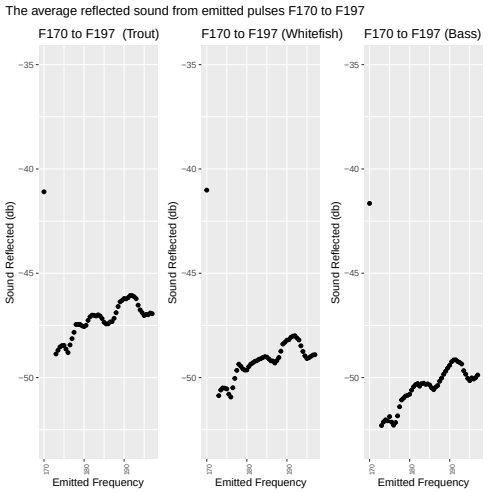

Figure 7: Comparing the reflected sound pulse from each fish between F170 - F197

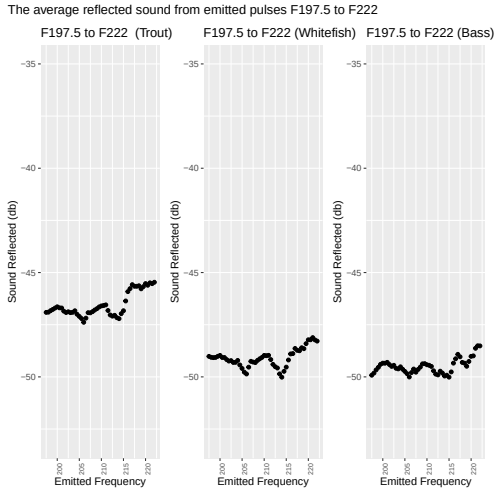

Figure 8: Comparing the reflected sound pulse from each fish between F197.5 - F222

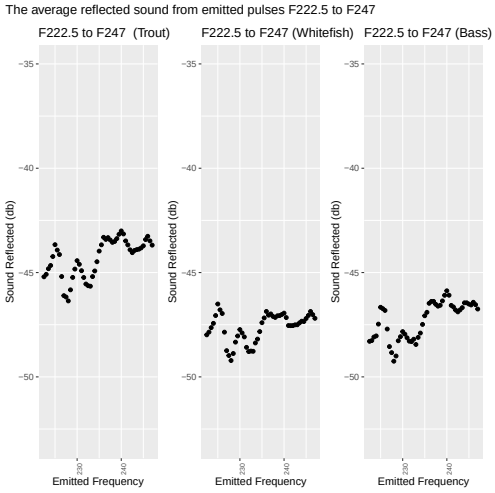

Figure 9: Comparing the reflected sound pulse from each fish between F22.5 - F247

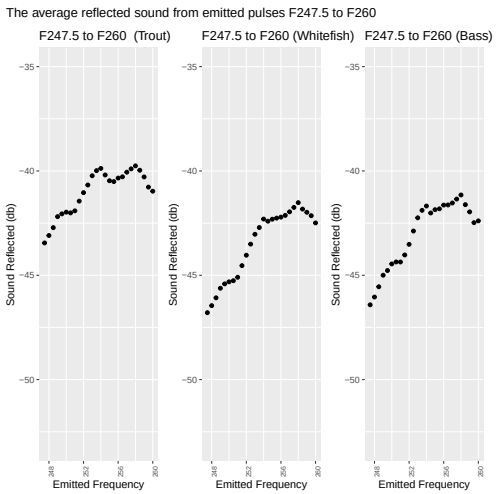

Figure 10: Comparing the reflected sound pulse from each fish between F247.5 - F260

In the figures 2-10, the average reflected signal strength for all three species are compared against one another for a range of frequencies. To begin with, the regions where the plots are identical to each other are trivial at this point considering we cannot differentiate between different species. Yet, in figure 2, there is a stark contrast between the three species along the 90-95 Hz range. The trout clearly have the lowest sound strength here with the bass having the second lowest and the whitefish the highest. This trend seemingly continues in figure 3 within the 95Hz to 119.5 Hz range but the discrepancies seem to dwindle ass the Hz increase. Another notable area of interest is shown in figure 6 within the 170 Hz to 197Hz range. In this plot, the trout now have the highest average sound strength while the whitefish have the second highest and the bass the lowest. The trout continue to have the highest sound strength in the rest of the figures (i.e. 197.5 Hz - 260 Hz), however, the averages between the bass and whitefish are almost identical. Overall, there are two frequency ranges that could be critical in for the future prediction model. The 90-95 Hz range and the 170-197Hz are seemingly the only two regions where all three species are clearly differentiable among each other with respect to their average reflected signal strength.

3.3 Time Series Component

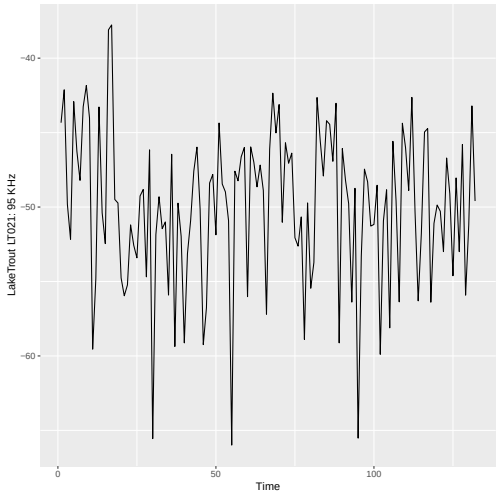

Figure 11: Signal Strength of LakeTrout 21 at 95 Khz Overtime

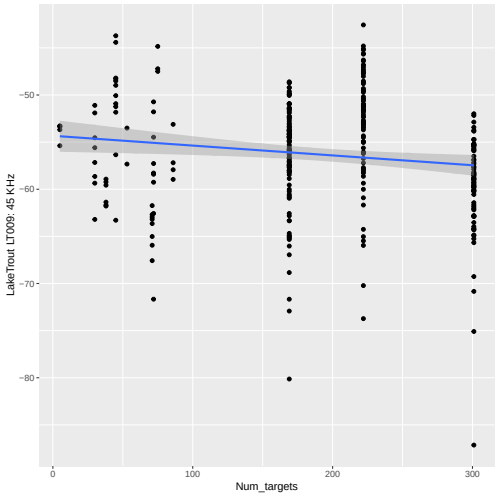

Figure 12: Signal Strength of LakeTrout 09 at 45 Khz for different number of Targets

In the last step of the exploratory data analysis, we visualized the pattern of the reflected signal overtime. Figure 11 shows how the signal strength of a particular laketrout changes after successive measurements at the 95 Khz frequency range. There is clearly a periodic pattern in the signal strength which suggests a consistent behavioral response; these fluctuations are most likely indicative of regular swimming patterns. In contrast, figure 12 shows how the signal strength differs across different number of targets at the 45 Khz. The number of targets refer to the amount of objects that the sonar instrument detects when it emits its pulses. Yet, there is no visible difference in the signal strength as the number of targets increases. Hence, we again chose to exclude this variable from our analysis.

4 Methodology

The exploratory data analysis showed several key findings that ultimately shaped our understanding of the dataset and informed the direction of subsequent modeling and interpretation. The time series component in particular is the most important characteristic since the swimming patterns of each fish clearly determine the strength of signal that is reflected. Each

individual fish has a unique time series, and the species as a whole may follow a general pattern overtime. Thus, the underlying distribution of the dataset has a periodicity component that needs to be accounted for during predictive modeling. Consequently, we approached this study in two separate phases while keeping in mind the irregular underlying distribution of the data. In the first phase we applied dimensionality reduction techniques, and in the second we applied two classification algorithms. Note that all statistical programming and graphical representations were conducted in Rstudio.

4.1 Dimensionality Reduction and Shrinkage

We applied principal component analysis (PCA), Kernel PCA, Lasso Regression, and a Random Forest to the dataset. To begin with, PCA in an unsupervised learning approach that serves as a method for dimensionality reduction and data visualization. PCA projects data onto a 'principal component' which captures as much variance as possible. Each principal component is a linear combination of a set of features and all principal components are orthogonal to one another. This provides a potential low-dimensional representation of the data, which can make it easier to identify trends and clusters. Similarly, Kernel PCA is also an unsupervised learning technique that reduces the dimension of the dataset and makes data visualization more interpretable. Kernel PCA applies a kernel function to the data (e.g, gaussian, polynomial, spline, etc.) and then projects the data onto the orthogonal principal components. The radial basis function in particular (or BRF kernel) is a popular kernel function that is used to help map data points to a higher-dimensional space, allowing algorithms to handle non-linear relationships. The RBF kernel is defined as

$$K_{RBF}(x, x') = exp[-\gamma||x - x'||^2] \tag{2}$$

Often, a Gaussian kernel is used to smooth data and improve interpretability. It is widely

considered as "one of the best performing kernels" and is often used in support vector machines for time series analysis. Next, lasso regression works similarly to linear regression, but instead assigns a penalty to the linear regression coefficients based on their individual L1 norm. In this way, many of the coefficients 'shrink' to 0 and the set of predictors are reduced to smaller subset. Finally, a random forest is traditionally a model classifier that builds decision trees through bootstrap samples. Random forests do not make assumptions about the underlying distribution and can evaluate the significance of each feature. The Gini index in particular reflects the purity of a node; therefore, variables with higher mean decreases Gini index contribute more to node purity, making them more important for prediction.

The dimensionality reduction techniques we selected collectively assume various underlying distributions of the data. Our approach was to apply each method to the dataset, extract a subset of relevant predictors, and then feed the reduced feature sets into classification algorithms. We would then compare the resulting prediction accuracies to assess the effectiveness of each technique in preserving informative structure while reducing dimensionality.

Prior to applying these techniques, we shrunk the data set from 10333 observations to exactly 6000 rows by randomly selecting 2000 fish from each species. In this way, the dataset would be balanced and not skew the prediction results.

4.2 Classification

The classification models we selected were K-nearest neighbours (KNN) and a support vector machine. KNN assigns a data point to a group (i.e cluster) based on its euclidean distance in space. This algorithm assumes no underlying distribution about the data and is best suited for data sets with continous variables. Moreover, a support vector machine is a classification technique that divides points in space with hyperplanes and predicts the accuracy of these boundaries. Support vector machines (SVM) can also map data points using kernel functions (gaussian, polynomial, spline, etc.) which make them effective for data that are not linearly

separable.

We applied the subset of predictors to each of the two classification techniques and then measured the accuracy of the algorithms using a prediction matrix. We used 50% of the data as the training set and the other half as test data. In the KNN algorithm, we chose k=3 to be the number of neighbors because of the three distinct groups of fish species. For the support vector machine, we compared prediction accuracy between gaussian and linear kernels.

4.3 Approach

1) We sampled a random subset of 2000 individuals from each of the fish species and combined them into one data frame

2) We then applied PCA, Kernel PCA with a radial kernel function, Random Forest, and Lasso Regression to the data frame.

- For PCA we examined the variability explained by the components, the projected data, and the loadings for the first few PC's. We conducted a similar procedure for kernel PCA.
- For Lasso Regression, we used 5-fold cross validation in order to compute the optimal tuning parameter λ. After training the model with the optimal tuning parameter , we calculated the predictive accuracy of the regression model. Finally, we evaluated the coefficients by comparing them between the three species.
- We applied a random forest to the training set and examined the Mean Decrease Gini index for each of the variables for feature selection.

3) After having identified valid subsets of features, we applied KNN, SVM with linear kernel, and SVM with gaussian kernel to the data set with the variables of interest. We then apply a 50/50 train-test split and calculated prediction accuracy using a confusion matrix.

5 Results

5.1 PCA

	PC1	PC2	PC3
Proportion of Variance	0.352	0.057	0.02
Cumulative Proportion	0.352	0.41	0.458

Table 1: Proportion of variance explained by the first 3 Principal Components

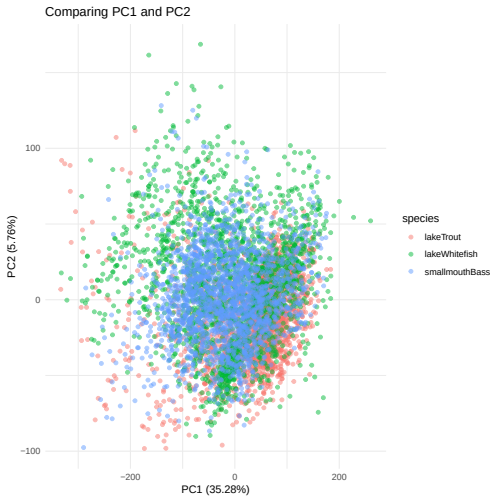

Figure 13: Projected data points onto 2 Dimensions spanned by PC1 vs. PC2

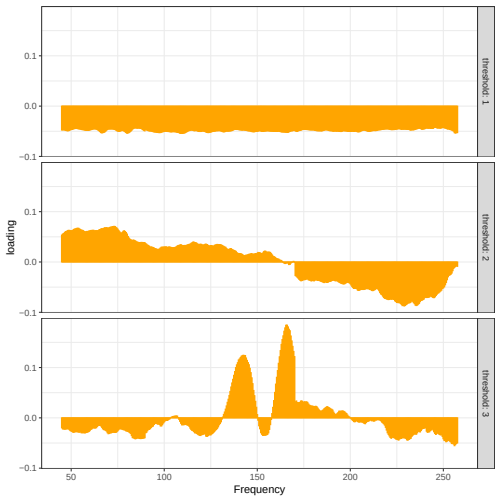

Figure 14: Principal Loadings for PC1,PC2, and PC2

5.2 Kernel PCA

Kernel PCA was not able to run in R given that it was extremley computationally intensive.

5.3 Lasso Regression

Prediction/Reference	Trout	Whitefish	Bass
Trout	691	222	243
Whitefish	127	575	268
Bass	172	204	498

Table 2: Prediction results after lasso regression on a 50/50 train test split

After training half the data set using lasso regression, the other half was used as a test set

and the subset of features computed from lasso regression had a 58.5% prediction accuracy rate.

5.4 Random Forest

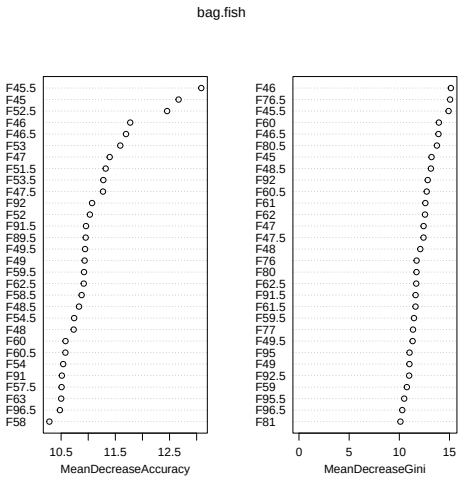

Figure 15: Mean Decrease Gini Index and Mean Decrease Accuracy for the first 25 Frequencies

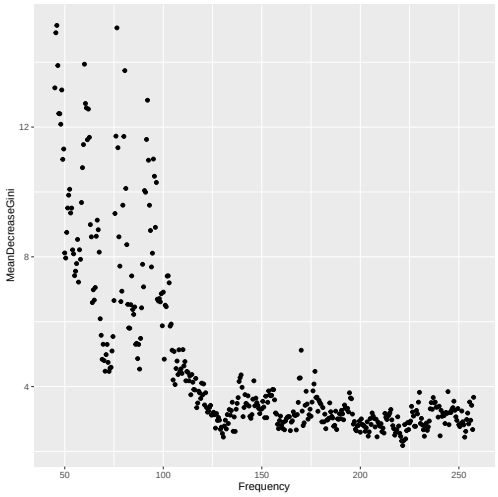

Figure 16: Visualizing the change in the average Gini index across every frequency

Prediction/Reference	Trout	Whitefish	Bass
Trout	666	55	269
Whitefish	137	794	70
Bass	95	59	855

Table 3: Prediction results from the random forest on a 50/50 train test split

After building a random forest on half the dataset, the other half was used as a test set and the model produced a 77.2% prediction accuracy rate.

5.5 Classification Results

Model	Accuracy
KNN (n = 3)	65.3%
SVM(linear)	55.1%
SVM(radial)	56.2%

Table 4: Prediction results from the SVM and KNN for the predictors with MeanDecrease Gini ≥ 6

Model	Accuracy
KNN (n = 3)	67.9%
SVM(linear)	54.6%
SVM(radial)	64%

Table 5: Prediction results from the SVM and KNN for the predictors with MeanDecrease Gini ≥ 8

Model	Accuracy
KNN (n = 3)	67.5%
SVM(linear)	54.5%
SVM(radial)	69.23%

Table 6: Prediction results from the SVM and KNN for the predictors with MeanDecrease Gini ≥ 10

Model	Accuracy
KNN (n = 3)	63.6%
SVM(linear)	53.8%
SVM(radial)	65.7%

Table 7: Prediction results from the SVM and KNN for the predictors with MeanDecrease Gini ≥ 12

6 Conclusion

In this project, we made an effort to highlight the important features in classifying fish species with sonar data. The dimension reduction techniques for the most part showed poor results to say the least. Principal component analysis did not show any significant clusters or trends in the reduced dimensions. The first principal component explained around 35% of the variability in the model which was 20% better than the subsequent PC. In addition, the loadings for the first principal component were nearly identical across all frequencies. This suggests that a linear combination of the features may not be a suitable model for the

data. These findings were further reinforced by the results from Lasso Regression since the out of sample prediction accuracy was slightly above 50%. Overall, we can conclude from these results thats the underlying distribution of the data is non-linear in its entirety.

Kernel PCA proved to be too computationally intensive, but the random forest was able to produce some viable results with regards to feature importance. Using the 'Mean decrease Gini index' we were able to visualize the frequencies that were of the utmost importance to the forest. We extracted four different subsets of predictors using this metric: features with mean decrease in Gini$> 12, > 10, > 8, and > 6$. However, one concerning aspect about the model was the fact that lower frequencies had a higher mean decrease in Gini Index, and there seems to be an inverse linear relationship between frequency and this metric; this suggests some potential correlation issues among the features. The model may be splitting on one frequency and ignoring the importance of the other which is resulting in biased results.

Nonetheless, using the feature subsets selected by the random forest yielded mixed results during classification. The SVM with a linear kernel achieved an accuracy of approximately 54% across all cases, showing limited sensitivity to environmental variation. In contrast, the SVM with a Gaussian (RBF) kernel significantly improved performance, increasing accuracy by nearly 10 percentage points. The KNN algorithm had the highest accuracy, peaking at 68% with the subset of predictors with mean Gini index > 8.

Although the prediction accuracies were underwhelming, the increase in accuracy from the gaussian (RBF) kernel is somewhat promising. Since the frequencies follow a periodic time series pattern, the gaussian (RBF) kernel might possibly be smoothing out the function and thus making prediction more accurate. It is important to note that we were not able to optimize the parameters in the SVM algorithm because of cost, this could have increased the accuracy slightly more.

7 Limitations and Future Work

Even though our methods showed a number of appealing results, much more still needs to be done. The main limitations of this data set include the abundance of missing data and the irregular time series. These two characteristics made it difficult to capture the variability in the data without overfitting, yet, the Gaussian (RBF) kernel did yield some promising results. Future work should build upon this foundation by focusing on rigorous hyperparameter optimization, and developing models that are better suited to handling irregular, real-world time series data. We encourage future research to prioritize the identification of key frequency ranges that contribute most significantly to species differentiation, even if this comes at the expense of classification accuracy. In this way, SONAR instruments can be utilized more effectively and efficiently, enabling broader application across a wider range of species and larger aquatic environments.

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer, 2009. URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

[2] Gareth James et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. URL: https://faculty.marshall.usc.edu/gareth-james/ISL/.

[3] Arun Pandey et al. "Multi-view kernel PCA for time series forecasting". In: *Neurocomputing* 554 (2023), p. 126639. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2023.126639. URL: https://www.sciencedirect.com/science/article/pii/S0925231223007622.

[3]

[2]

[1]