Exploratory Data Analysis - Zooplankton Species Classification Project

Data Cleaning

This Exploratory data analysis looked into three zooplankton datasets (mosaic .tif files, .csv export files, and MasterTable) collected from Lake Huron and Lake Simcoe. The mosaic .tif files generated by FlowCam provide high-resolution images that capture zooplankton features, such as shape and size. The .csv export files contain quantitative particle data and metadata linking the particle to its position in the mosaic. The Master Table connects the two datasets by incorporating important sampling contextual information.

The first step of the data cleaning process was to retain only observations corresponding to the two target zooplankton classes. Second, to incorporate environmental sampling variables, a left join to the csv files was performed using the ‘csv file name’ and ‘image file name’ as the unique key, merging the Master Table’s variables into the datasets. During this process, it was identified that 'YPerchDen' contained different values for the same unique key. Due to the lack of information on which value was more reliable, this variable was dropped at this stage. Eventually, one dataset for each lake was obtained, including the particle and environmental sampling data. The Lake Huron dataset consists of 51,429 observations, while the Lake Simcoe dataset contains 390,530 observations.

Data Visualization

The Lake Huron dataset contains 137 columns while the Simcoe dataset has 97. 54 variables were identified to be feature related and potentially contribute to species classification. This EDA focuses on analyzing these variables, examining their distributions, exploring the correlation between them, and assessing the linearity in the classification task.

Histogram, Density Plot, and Box Plot

Preliminary investigation using box plots showed that outliers exist in the majority of the variables. Therefore, log transformation was applied to all numerical variables except for ‘LON0’ and ‘LON1’ which contained negative input values. After the transformation, histograms, density plots, and box plots were used to identify key patterns and relationships between variables.

Three typical distribution patterns were observed. First, some variables exhibited a clear separation between the two target classes with distinct distributions. Both histogram and density plots revealed a bimodal distribution, indicating that the two classes occupy different regions within the feature space. Meanwhile, the box plot showed that the median, 25th percentile, and 75th percentile differ significantly between the two classes, suggesting that these variables could be strong predictors for classification. Second, some variables showed almost complete overlap between the two classes in distribution, with nearly identical medians and interquartile ranges. These features likely do not contribute to classification since their distributions do not vary between classes. Lastly, certain variables presented highly similar distributions, indicating a high correlation among them. This suggests that only one representative variable might be needed from such correlated groups to avoid redundancy, reduce multicollinearity, and maintain model efficiency.

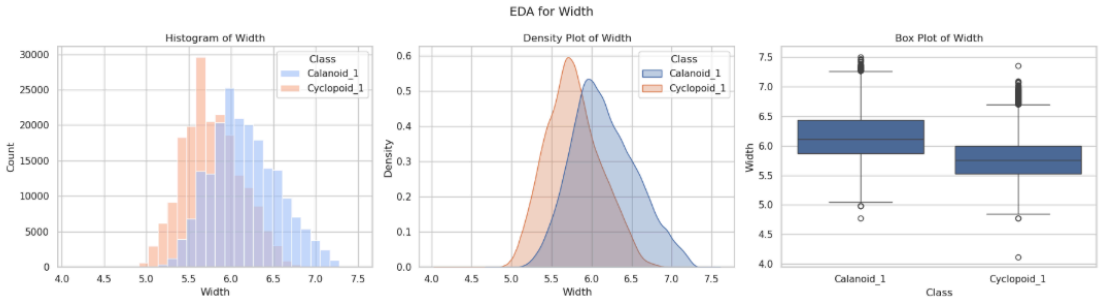

Figure 1: Case 1 - Distinct Sperate Distribution by Classes (Width - Simcoe Example)

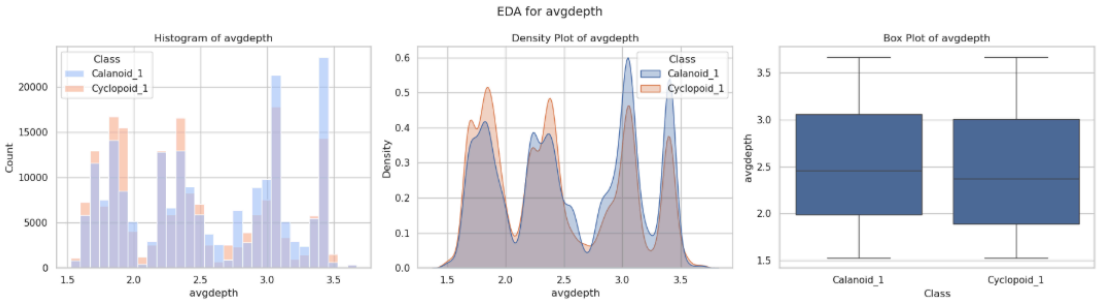

Figure 2: Case 2 - Overlapping Distribution by Classes (avgdepth - Simcoe Example)

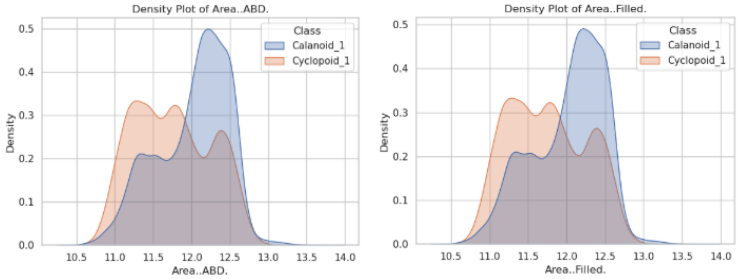

Figure 3: Case 3 - Highly Correlated Variables (AreaABD & AreaFilled - Simcoe Example)

Clustering Analysis

Hierarchical clustering using Ward's method was applied to further investigate case 3 shown above, where variables showed similar distributions. The clustering dendrogram presented the clusters of variables with similar statistical distributions. By analyzing the hierarchical results, groups of variables that are closely related were identified. Additionally, clusters with a distance threshold of less than 0.01 and 0.05 were filtered out for further exploration. These variables demonstrate strong correlations and to balance information retention and model efficiency, the next step plan is choosing one representative variable from each cluster as the predictor to be included in the final classification model.

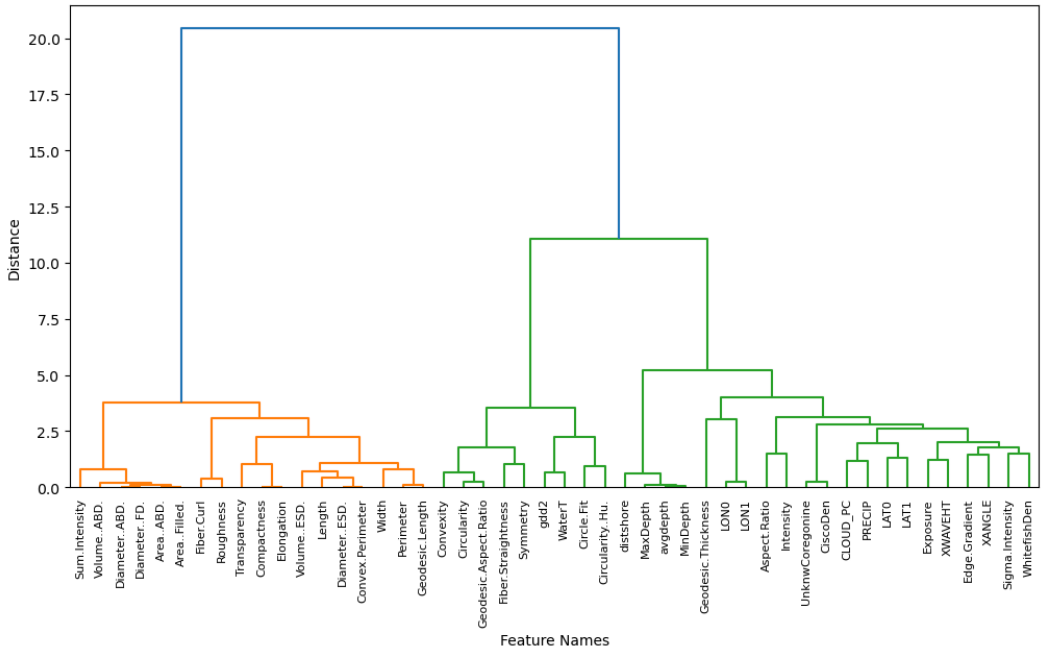

Figure 4: Hierarchical Clustering Dendrogram (Simcoe Example)

Linear Models

To evaluate the significance of different variables in classifying the two target classes, Ordinary Least Squares (OLS) regression and Logistic Regression were applied. During data preprocessing, it was identified that some variables contained missing values. Since the proportion of missing values was relatively low, those rows were excluded before performing regression analysis. The response variable for both models is the two target zooplankton classes. By examining the p-value and coefficient magnitude, variables with significant impact on classification were identified. These results can provide valuable insights to select the most relevant predictors for the final classification model.

```
Total number of rows in the dataset:
390530
Missing values before cleaning:

distshore                50174
Exposure                 11102

WhitefishDen              6446
UnknwCoregonine           6446
CiscoDen                  6446
```

OLS Regression Results:

	Variable	P-Value	Coefficient
0	const	6.715162e-83	1.219659e+02
1	Area..ABD.	4.830084e-245	-4.253469e-05
2	Area..Filled.	0.000000e+00	4.723524e-05
3	Diameter..ABD.	0.000000e+00	6.302844e-02
4	Diameter..ESD.	3.319862e-40	-1.200078e-02
5	Diameter..FD.	0.000000e+00	-4.920837e-02
6	Length	3.452565e-302	1.300382e-03
7	Width	0.000000e+00	1.395899e-03
8	Perimeter	2.570882e-01	7.434256e+00
9	Volume..ABD.	2.011114e-27	-4.974134e-09
10	Volume..ESD.	0.000000e+00	1.018124e-09

Figure 4: Dataset Overview and Partial OLS Regression Results (Simcoe Example)

Mosaic .tif Files Analysis

Convolutional Neural Network (CNN) was initially considered as a potential classification model using the mosaic .tif images. However, several challenges were observed making it challenging to use the image data to train the CNN model. First, each image contains multiple biological particle captures which cause difficulties for CNN to learn meaningful features. Second, the particle captures vary in size, position, and orientation, making it complicated to extract features during training. Lastly, the images contain large

black background areas, which may lead the CNNs to learn irrelevant features from these regions. Further work should be made to explore appropriate methods in cropping and scaling the images for the CNN model training purpose.

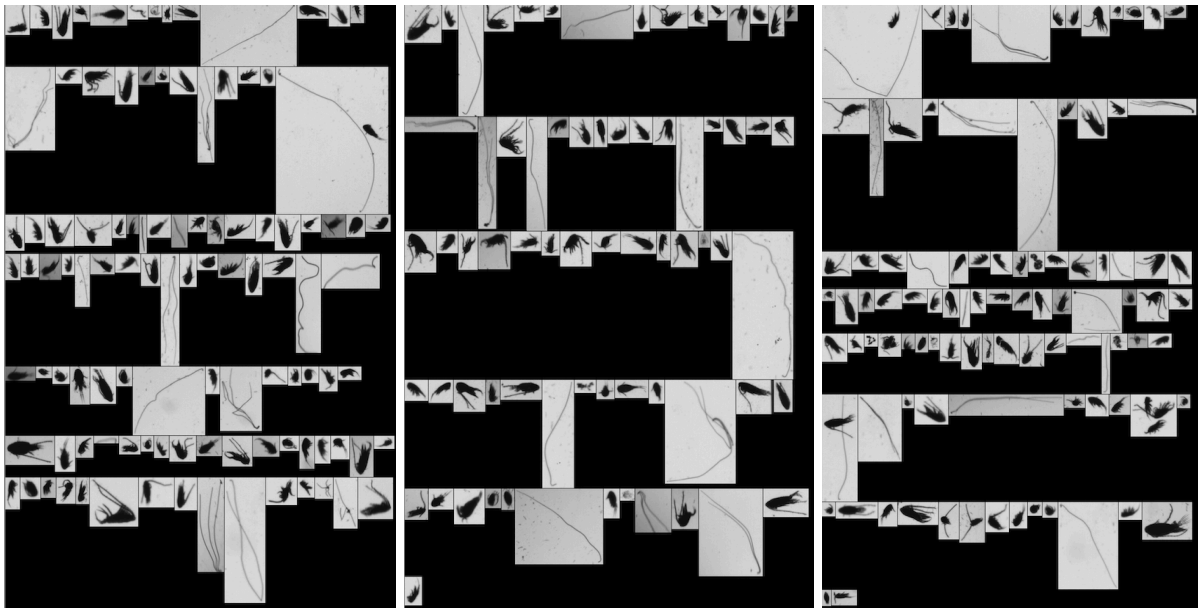

Figure 5: Comparison of Mosaic Images (Simcoe Example)

Conclusion

This Exploratory Data Analysis (EDA) investigated the distribution of various variables within the two target classes to help determine significant predictors for the classification model. Additionally, it explored whether certain variables exhibited high similarity or correlation within a single class, which may indicate feature redundancy. These findings provide practical insights for selecting the most relevant predictors for the classification model. The next step will focus on predictor selection decisions for model training and exploring effective methods for image data adjustments.