Kavya Puranam
12/13/2025
GGIS 527
Final Project Report

Forecasting Hypoxia in Temperate Lakes Using Ensemble Machine Learning Methods

Introduction

In this study, we investigate deoxygenation in lake ecosystems to forecast hypoxic events in temperate lakes using the "Widespread deoxygenation of temperate lakes: companion dataset 1980-2017" (Jane et al., 2020). The primary research question is whether ensemble machine learning methods such as random forest, AdaBoost, or gradient boosting can accurately predict seasonal hypoxia using dissolved oxygen concentrations across temperate lakes. Additionally, we aim to identify which environmental predictors contribute the most to hypoxia predictions.

Background

"Deadzones" are areas of low oxygen within bodies of water which cause major environmental issues in aquatic ecosystems. Hypoxia refers to lack of oxygen within bodies of water, and it is a common phenomenon due to seasonality and stratification where lower oxygen levels are found in denser and lower layers of a body of water. It is measured by Dissolved Oxygen concentration levels, where lower concentrations indicate hypoxia. Recently, hypoxia has been occurring more due to climate change. Increase in global temperatures leads to warmer water which results in decreased solubility of oxygen which increases stratification where oxygen stays in less dense warmer water near the top of the body of water and do not mix in deeper layers (Hinson et al., 2024). Without oxygen in the body of water organisms will die off which is where the term "deadzone" comes from.

This project aims to address a research gap where ensemble machine learning methods are applied for forecasting surface (Epilimnion) and deep water (Hypolimnion) dissolved oxygen (DO) concentrations using multiple years and locations worth of data. While AI models have been used to predict hypoxia (Ou et al., 2025), random forest, gradient boost, and AdaBoost ensemble methods have not been used yet. Ensemble methods can be useful because they have higher accuracy, reduce overfitting and variance, and are better at complex patterns. Complex deep learning models might be powerful and able to solve complex problems, but they require hyperparameter tuning in order not to overfit and use an unnecessarily larger computational power than we need (Mohammed et al., 2023). Here we want to be able to capture the complexities of the many variables affecting deoxygenation with better model fit for a smaller dataset and less computational need, therefore ensemble methods are best for our needs in analyzing those relationships.

Methods

The dataset used is the "Widespread deoxygenation of temperate lakes: companion dataset 1980-2017" (Jane et al., 2020). The area of the geographic location is worldwide, but lakes are mostly located in North America and Africa (Figure 1). This dataset contains surface and depth temperatures in Celsius and

dissolved oxygen concentration in milligrams per Liter across 400 temperate zone lakes. In order to analyze each of the Epilimnion and Hypolimnion datasets on the ensemble models, we needed to build each of the tables separately. We first imported Python libraries such as Scikit-learn for the machine learning models and visualization libraries such as Matplotlib and Plotly which is interactive. Each of the datasets needed to be preprocessed, by updating column names and merging both the Epilimnion and Hypolimnion datasets with the lake information dataset by two primary keys, Lake ID and Lake Name, then plotting the preliminary data visualizations to assess the shape of the data.

Then we ran each of the models (Random Forest, Gradient Boost, and AdaBoost) using Scikit Learn library for ensemble models in Python through Google Colab on both the Epilimnion and Hypolimnion datasets. The predictors were lake ID, year, mean temperature, unique data contributor, latitude, longitude, maximum depth, surface area, chlorophyll-a concentration, state or providence, and the response variable was continuous Dissolved Oxygen concentration. If the training time took too long after attempting to run a single model, I planned to limit the dataset to just North American lakes and then training each of the models on this data. However, the models were trained and predicted in a normal amount of time with the full datasets. Then I visualized the feature importance for each of the ensemble methods as a histogram using Matplotlib and plotted the model predicted Dissolved Oxygen content against the observed Dissolved Oxygen content and created a line with the minimum and maximum predicted values of each model as the predicted regression line. I also calculated the Root Mean Square Error (RMSE) as a metric of model accuracy and performance using the following equation:

$$RMSE = \sqrt{\frac{\Sigma (y_i - \hat{y}_i)^2}{N - P}}$$

y_i: *Actual Observed Values*

$\hat{y}_i$: *Model Predicted Values*

n: *Total Data Points*

Limitations: There is no current data past 2017 which we will need to be aware of since global temperatures have risen significantly since then.

Delimitations: Based on multicollinearity we removed values that were strongly correlated with our response variable dissolved oxygen content in milligram per liter. We found that there is multicollinearity between the response variable and mean percent of saturation for dissolved oxygen, both are measures of deoxygenation (Figure 14, 15). We used mean dissolved oxygen in mg/Liter and dropped the mean percent of saturation for dissolved oxygen column from both the Epilimnion and Hypolimnion datasets. At this point, I also dropped the Row ID as it is a unique identifier and not a predictor that we want to assess.

Results

Epilimnion Table

The Root Mean Squared Error (RMSE) scores from each of the models for the Epilimnion dataset show that the model with the best performance was Gradient Boost, with an RMSE of 1.20, as it was the lowest and closest to 0, followed by Random Forest at 1.21 and finally AdaBoost at 1.48. The feature importance table showed similar features as most important across all three models. For the Gradient Boost model, year was found to be the most important feature, followed by mean temperature in Celsius and then latitude (Figure 4). For the Random Forest model, mean temperature was found to be the most important feature, with year being nearly as important, and latitude being 0.10 lower in importance and therefore

less influential (Figure 2). Finally, the AdaBoost model identified mean temperature as the most important feature, followed by chlorophyll concentration, with year being the next most important (Figure 6).

We also examined the model predicted values against the actual observed values to assess model performance in addition to the RMSE scores through visual analysis. For the Epilimnion dataset, the Gradient Boost results were more tightly clustered around the y = x line and evenly split on either side, indicating evenness in model fitting (Figure 5). The Random Forest model showed similar results (Figure 3). However, the AdaBoost model was more spread out from the y = x line, with more points above the curve than below and visible bands of points, indicating overfitting at values less than 7, 8, and 9 (Figure 7).

Hypolimnion Table

The Root Mean Squared Error (RMSE) scores from each of the models for the Hypolimnion dataset show that the best-performing model was Gradient Boost, with an RMSE of 1.16, as it was the lowest and closest to 0. This was followed by Random Forest at 1.19 and finally AdaBoost at 2.01. The feature importance table showed similar features as most important across all three models, though with different ranges. For the Gradient Boost model, chlorophyll was found to be the most important feature, with a score of 0.6, while the remaining features were close in score at approximately 0.1 (Figure 10). For the Random Forest model, chlorophyll concentration was also the most important feature, with the remaining features again having scores close to 0.1 (Figure 8). Finally, the AdaBoost model identified chlorophyll concentration as the most important feature at 0.3, followed closely by temperature, maximum depth, and year (Figure 12).

We also examined the model predicted values versus the actual observed values to assess model performance in addition to the RMSE scores through visual analysis. For the Hypolimnion dataset, the Gradient Boost results were more tightly clustered around the y = x line, with more data points concentrated near (0.0, 0.5). The data points were evenly split around the y = x line, indicating good model fit (Figure 11). The Random Forest plot was more spread out from the y = x line, although there was still evenness, with data points split around the line, indicating reasonable model fit (Figure 9). In contrast, the AdaBoost model showed signs of underfitting, as the data points were spread across values above 7, and the cluster of actual values between (0.0, 0.5) appeared to be significantly misfit, with points lying far from the prediction line (Figure 13).

Discussion and Conclusion

To compare the results across the Epilimnion and Hypolimnion datasets, Gradient Boosting emerged as the most accurate regression model for both, aligning with the predictions vs actual visualizations. However, the Hypolimnion dataset achieved a lower RMSE of 1.16, indicating higher predictive accuracy than the Epilimnion dataset. This finding is interesting because we initially expected deeper water layers to be more difficult to predict, as dissolved oxygen concentrations at depth are more sensitive to external factors, which can increase variability and reduce predictability (Ma et al., 2025).

Consistent with the Hypolimnion results, chlorophyll-a concentration was identified as the most important feature contributing to model predictions. This aligns well with established ecological understanding, as high chlorophyll levels, associated with increased algal biomass, can settle into the hypolimnion where

decomposition consumes oxygen and increases hypoxia (National Academies of Sciences, Engineering, and Medicine, 2022).

For the Epilimnion dataset, feature importance included temperature, which also follows expected environmental dynamics, since higher water temperatures reduce oxygen solubility and can contribute to hypoxic conditions in surface waters. The model also identified year as an important predictor, highlighting the role of seasonal variability, as lake stratification patterns are strongly influenced by temporal dynamics (Bouffard et al., 2013).

Overall, the models consistently relied on scientifically meaningful variables to generate predictions across both datasets. This study demonstrates that ensemble methods can effectively forecast hypoxia in temperate-zone lakes by capturing complex interactions within multivariate environmental data. Applying AI models to geospatial datasets offers the potential for faster, more efficient conservation decision making, ultimately supporting efforts to reduce hypoxia and restore lake ecosystems.

Works Cited

Altieri, Andrew H., and Keryn B. Gedan. 2015. "Climate Change and Dead Zones." Global Change Biology 21 (4): 1395–1406. https://doi.org/10.1111/GCB.12754.

Arend, Kristin K., Dmitry Beletsky, Joseph v. Depinto, Stuart A. Ludsin, James J. Roberts, Daniel K. Rucinski, Donald Scavia, David J. Schwab, and Tomas O. Höök. 2011. "Seasonal and Interannual Effects of Hypoxia on Fish Habitat Quality in Central Lake Erie." Freshwater Biology 56 (2): 366–83. https://doi.org/10.1111/J.1365-2427.2010.02504.X.

Bouffard, Damien, Josef Daniel Ackerman, and Leon Boegman. 2013. "Factors Affecting the Development and Dynamics of Hypoxia in a Large Shallow Stratified Lake: Hourly to Seasonal Patterns." Water Resources Research 49 (5): 2380–94. https://doi.org/10.1002/WRCR.20241.

Hinson, Kyle E., Marjorie A.M. Friedrichs, Raymond G. Najjar, Zihao Bian, Maria Herrmann, Pierre St-Laurent, and Hanqin Tian. 2024. "Response of Hypoxia to Future Climate Change Is Sensitive to Methodological Assumptions." Scientific Reports 2024 14:1 14 (1): 1–18. https://doi.org/10.1038/s41598-024-68329-3.

"Hypoxia." n.d. Accessed November 10, 2025. https://oceanservice.noaa.gov/hazards/hypoxia/.

Jane, Stephen F, Gretchen J. A. Hansen, Kraemer Benjamin, Peter R. Leavitt, Joshua L. Mincer, Rebecca L. North, Rachel M. Pilla, et al. 2020. "Widespread Deoxygenation of Temperate Lakes: Companion Dataset 1980- 2017." https://doi.org/10.6073/PASTA/AC8B05BB0DA19032B3DF3EFC21F83874.

Ma, Bing, Fei Dong, Wenqi Peng, Xiaobo Liu, and Aiping Huang. 2025. "Dynamics of Oxygen Evolution in a Thermally Stratified Reservoir under Climate Warming." Scientific Reports 2025 15:1 15 (1): 40419-. https://doi.org/10.1038/s41598-025-13432-2.

Mohammed, Ammar, and Rania Kora. 2023. “A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges.” *Journal of King Saud University - Computer and Information Sciences* 35 (2): 757–74. https://doi.org/10.1016/J.JKSUCI.2023.01.014.

Ou, Yanda, Z. George Xue, Supratik Mukhopadhyay, Magesh Rajasekaran, and Dylan Wichman. 2025. “Forecasting Coastal Hypoxia Using a Blend of Mechanistic and Artificial Intelligence Models.” Scientific Reports 2025 15:1 15 (1): 1–18. https://doi.org/10.1038/s41598-025-17053-7.

Tellier, Joshua M., Nicholas I. Kalejs, Benjamin S. Leonhardt, David Cannon, Tomas O. Höök, and Paris D. Collingsworth. 2022. “Widespread Prevalence of Hypoxia and the Classification of Hypoxic Conditions in the Laurentian Great Lakes.” Journal of Great Lakes Research 48 (1): 13–23. https://doi.org/10.1016/J.JGLR.2021.11.004.

National Academies of Sciences, Engineering, and Medicine. 2022. “The Future of Water Quality in Coeur d’Alene Lake.” The National Academies Press, December. https://doi.org/10.17226/26620.

Figures and Tables

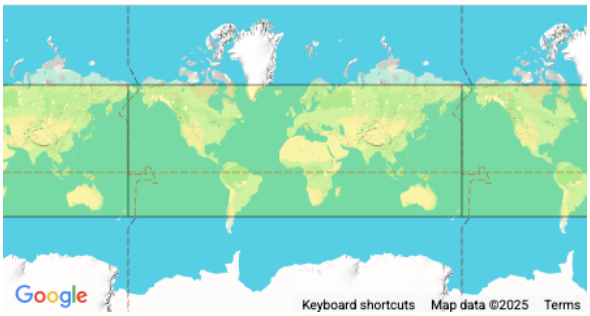

Figure 1. Worldwide geographic region but primarily in North America and Europe with bounding coordinates of N: 68.2962 degrees, S: -42.6175 degrees, E: 180 degrees, W: -180 degrees

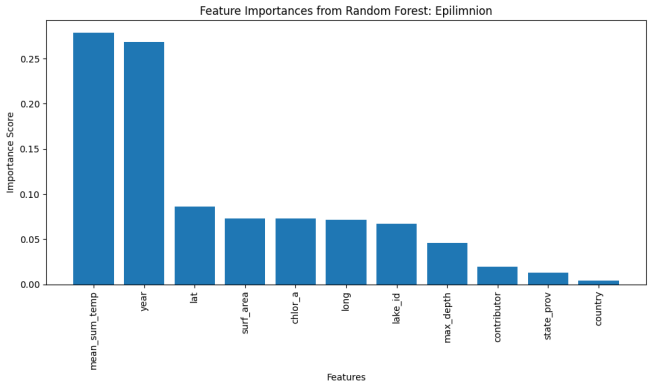

Figure 2. Feature Importance Graph for Surface Level Epilimnion Predictors in Random Forest Model

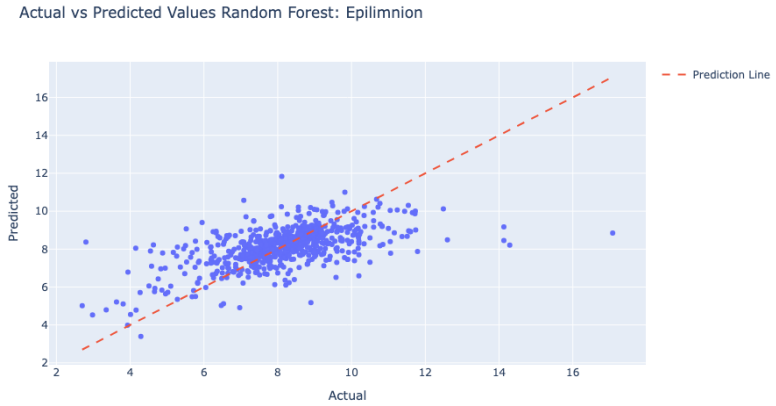

Figure 3. Actual vs. Predicted Scatter Plot for Random Forest Epilimnion Dissolved Oxygen content predictions.

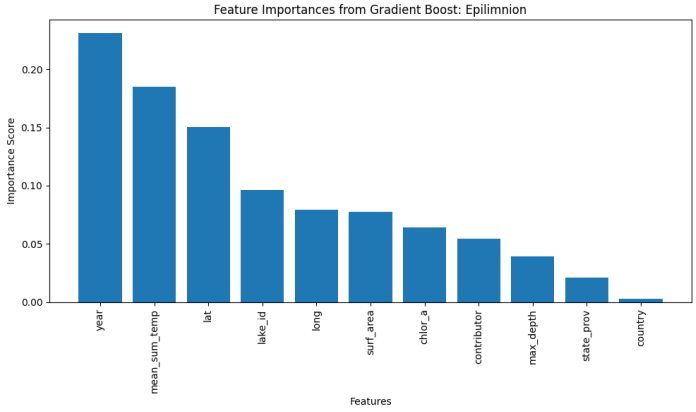

Figure 4. Feature Importance Graph for Surface Level Epilimnion Predictors in Gradient Boost Model

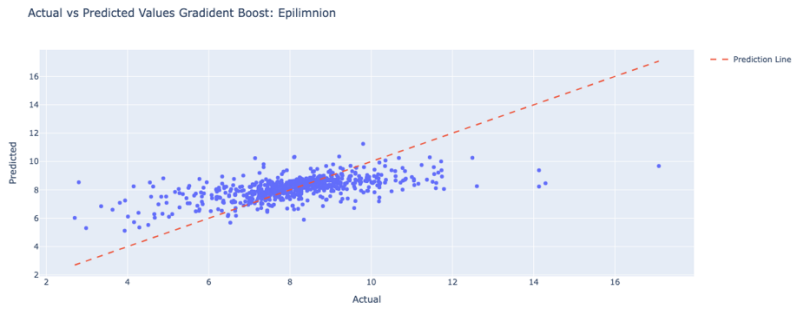

Figure 5. Actual vs. Predicted Scatter Plot for Gradient Boost Epilimnion Dissolved Oxygen content predictions.

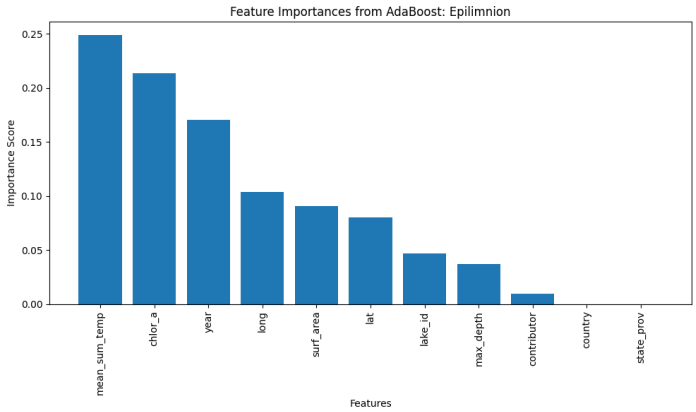

Figure 6. Feature Importance Graph for Surface Level Epilimnion Predictors in AdaBoost Model

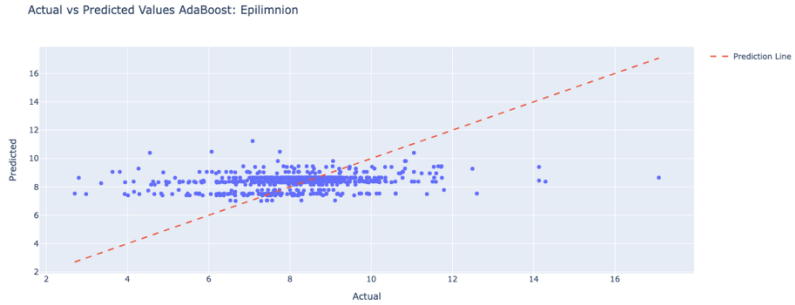

Figure 7. Actual vs. Predicted Scatter Plot for AdaBoost Epilimnion Dissolved Oxygen content predictions.

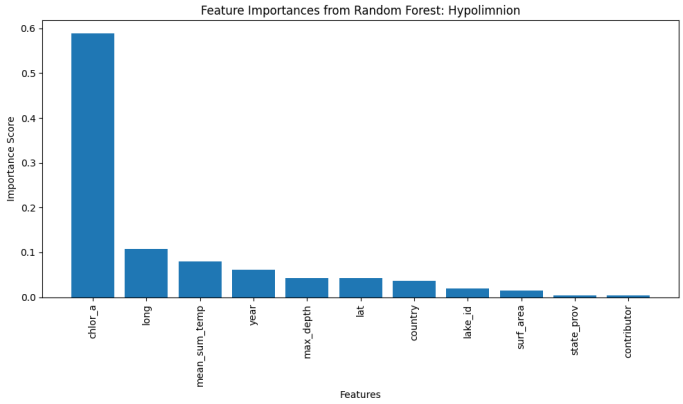

Figure 8. Feature Importance Graph for Bottom Levels Hypolimnion Predictors in Random Forest Model

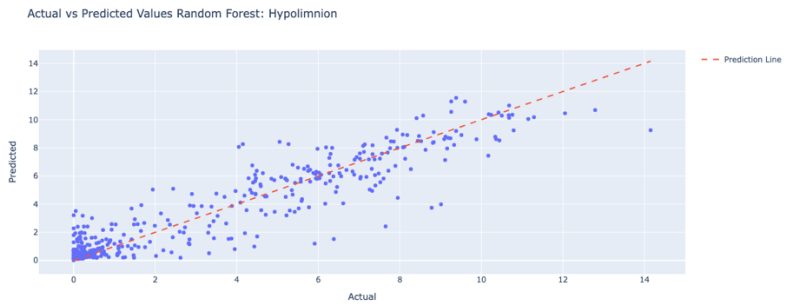

Figure 9. Actual vs. Predicted Scatter Plot for Random Forest Hypolimnion Dissolved Oxygen content predictions.

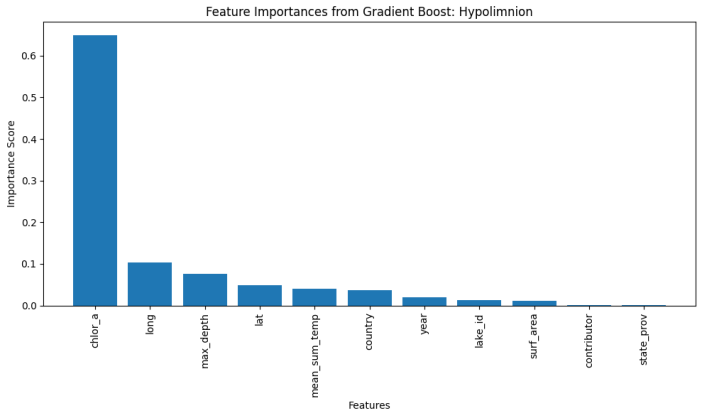

Figure 10. Feature Importance Graph for Bottom Levels Hypolimnion Predictors in Gradient Boost Model

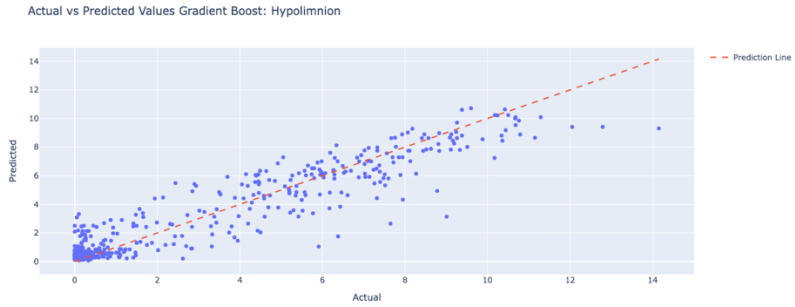

Figure 11. Actual vs. Predicted Scatter Plot for Gradient Boost Hypolimnion Dissolved Oxygen content predictions.

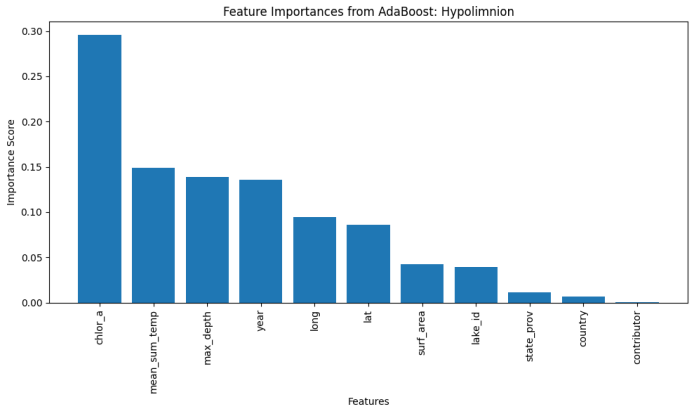

Figure 12. Feature Importance Graph for Bottom Levels Hypolimnion Predictors in AdaBoost Model

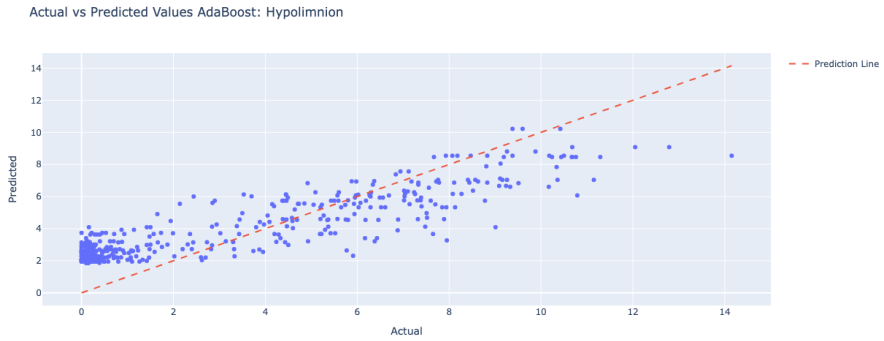

Figure 13. Actual vs. Predicted Scatter Plot for AdaBoost Hypolimnion Dissolved Oxygen content predictions.

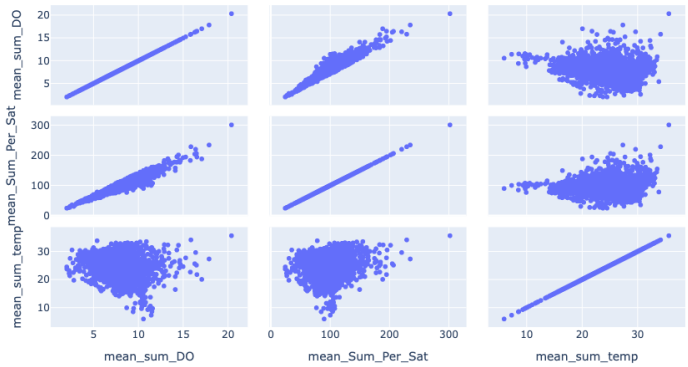

Figure 14. Visualization of multicollinearity between mean dissolved oxygen concentration, mean percent of saturation for dissolved oxygen, and mean temperature with the Epilimnion dataset.

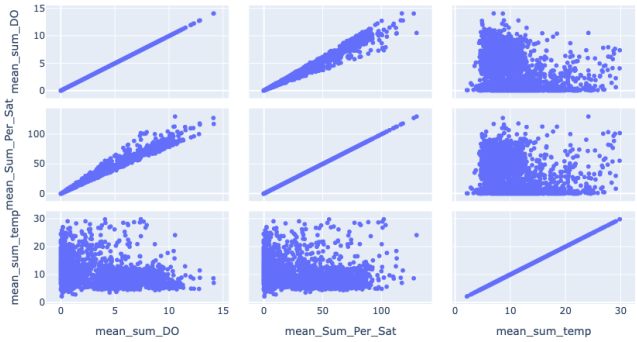

Figure 15. Visualization of multicollinearity between mean dissolved oxygen concentration, mean percent of saturation for dissolved oxygen, and mean temperature with the Hypolimnion dataset.

RUBRIC FOR GRADING

Criteria	Mastery (Full Credit)	Excell	Meet (50% credit)	Below Expectation	Fail (No credit)
Professional Presentation of Material	Well written, few to no grammatical errors, clear details of your work	Well written, some grammatical errors, clear but insufficient level of detail	Well written, some grammatical errors, minimum lack of clarity and detail	Poorly written, grammatical errors throughout, lacking clarity and detail	Did not use proper written communication, incomprehensible writing, ambiguous details
Choice of Research Problem	Research problem is challenging but completable for the time given	Research problem is average difficultly but completable	Research problem is not difficult and completable	Research problem is not achievable in the time provided	It is not possible to understand what your research problem is from what you have written
Choice of Data	You have chosen to work with multiple GeoData sets that requires data synthesis	You have chosen to work with multiple data files that do not require synthesis	You have chosen a GeoData Source that is particularly well suited to your goal	You have chosen a generic geospatial dataset	You have not chosen a GEOSPATIAL dataset
Proposed Analysis	The geospatial analysis you suggest is appropriate, clearly described, and completely detailed	The geospatial analysis you suggest is appropriate, clearly described, but incomplete in detail	The geospatial analysis you suggest is appropriate, but the description leaves out details that	The geospatial analysis you suggest is appropriate, but is not described well or completely	You did not choose an appropriate method or have not written about it in enough detail to judge whether it is appropriate

			makes it difficult to know If it is complete		