Quarter 1 Project Proposal: Chlorophyll-*a* Concentrations in Water Bodies Predictive Model

Group Members:  
Jacob Dipasupil, Petr Kisselev, Nikhil Alladi

# 

# Table of Contents (WIP)

[Table of Contents (WIP) 2](#_wqse8o593t1m)

[Data Information 3](#_zemelijkcb7q)

[Model and Rationale 3](#_3ef5sa3dh4wg)

[Preprocessing 4](#_5ctj50sz9a9h)

Dataset: [Estimates of lake nitrogen, phosphorus, and chlorophyll-a concentrations to characterize harmful algal bloom risk across the United States](https://catalog.data.gov/dataset/estimates-of-lake-nitrogen-phosphorus-and-chlorophyll-a-concentrations-to-characterize-har)

# Data Information

Our dataset contains data compiled by the U.S. Environmental Protection Agency on certain characteristics of lakes. The dataset has 67 different attributes, some of which include lake name, date of sampling, total phosphorus concentration, area of lake surface, monthly and yearly average precipitation across the watershed, annual average nitrogen from human waste, lake depth, log of chlorophyll-*a* concentration, and more. The meaning of every attribute is contained in the data dictionary found in the provided link to the dataset. Since there are 67 different attributes, and we are classifying a class attribute of our choosing, our dataset has a dimensionality of 66. There are 2,226 instances with 45 missing values for lake name, 41 missing values for both nitrogen concentration and phosphorus concentration, 75 missing values for depth, and 132 missing values for the log of chlorophyll-*a* concentration, our class. Since we are trying to classify chlorophyll-*a* concentration, we will have to remove those 132 instances where the values are missing. The distribution of data is slightly right skewed with a mean of 1.053 and a standard deviation of 0.563 ranging from 0.029 to 2.941. Since the log of chlorophyll-*a* is a numerical variable, we will discretize the data into three bins: low, medium, and high. The class distribution is quite heavily skewed to the right, with 1,084 instances in low (-inf-0.999795], 864 in medium (0.999795-1.970205], and 146 in high (1.970205-inf).

# Model and Rationale

Our model will use data on lakes to predict if the concentration of chlorophyll-*a* is high, medium, or low in order to give us information about the state of the lake ecosystem. Chlorophyll-*a* concentrations can be used as a measure of the amount of algae growing in a water body and give us information on the trophic condition of a waterbody. High levels of chlorophyll-*a* concentrations and the subsequent algae growth can lead to harmful algal bloom, characterized by excessive algae growth producing toxins in water bodies, and hypoxia, which is when oxygen concentrations are too low for most organisms to survive in. Both of which are detrimental to the organisms living in and drinking from water bodies and can have harmful effects to the surrounding ecosystem. Being able to predict chlorophyll-*a* concentrations before permanent damage is done can help save some of these ecosystems.

# Preprocessing

The first step of our preprocessing was done in Google Sheets. Many values in our dataset caused errors when trying to open in Weka. In order to allow Weka to open the dataset, all apostrophes in the *LAKENAME* attribute values were replaced with spaces. Additionally, there were 287 cells that contained one of the following values: “#NUM!”, “#DIV/0!”, “#VALUE!”. These obvious error values lead Weka to decide that certain attributes are string when they should be numeric attributes. To fix this we simply converted all data cells with those values into empty cells.

In the future, we will reduce the dimensionality of the dataset to ease space and time costs for the training of our model. Some possible attribute selection algorithms that we may use to aid us in preprocessing the data are learner-based feature selection, information gain based feature selection, correlation based feature selection, principal component analysis, and the intuitive approach to removing attributes that logically seem irrelevant to the trends we are trying to capture. All of these, except for the hand-picked feature selection, will be implemented through algorithms included in Weka. Furthermore, as stated in the Data Information section, we will remove the instances that have a missing value for the class attribute, *logchl\_A*.