Quarter 1 Project Intermediate Report: 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 4](#_3ef5sa3dh4wg)

[Preprocessing 4](#_5ctj50sz9a9h)

[Attribute selection 7](#_jevk7ap5o5ih)

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

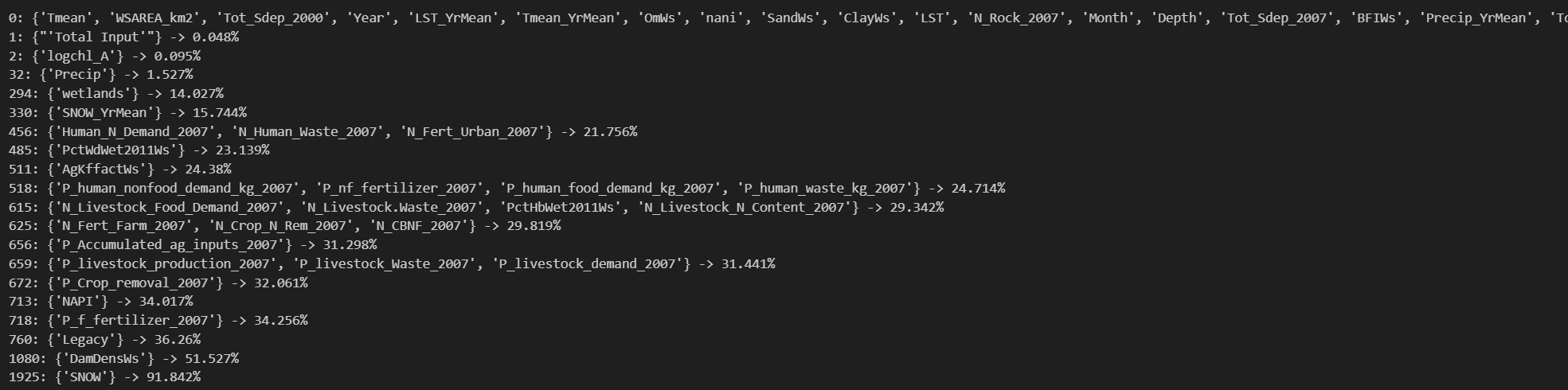
Our model will use data on lakes to predict if the concentration of chlorophyll-*a* is high, medium, or low in order to give us information about the state of the lake ecosystem. Chlorophyll-*a* concentrations can be used as a measure of the amount of algae growing in a water body and give us information 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.

Pushing this data into WEKA, some further steps for preprocessing present themselves. To begin with, 130 instances in our dataset are missing values for our assigned class attribute *logchl\_A*. As supervised learning requires labeled class attributes, we removed these instances from our dataset (note that due to the values here being positive decimals, we set the split point to be above the maximum of 2.941 in this attribute so that the filter did not inadvertently remove valid instances as well). We also removed attributes that can be clearly reasoned to have no relation to the class attribute of any kind, including *LAKENAME, Survey Number, and SITE\_ID.*

Looking at this data in a spreadsheet view, we noticed that some attributes had a notably high amount of the value 0 in them. Due to their numerical basis, we took this 0 to be a default value, and analyzed the amount of zeros per attribute. Doing so, we came to this result, seen below. 

We also analyzed the number of 0s per instance, resulting in this sort:



We chose to drop the *SNOW* attribute from this analysis as it had more than 70% of its values missing, and kept the instances intact.

Due to the extreme variance in magnitude of the data, we decided to normalize all attributes in the dataset. Some attributes had notable outliers, so we used z-score normalization for these. These attributes were as follows: *wsarea\_km2, lake\_area\_ha, fire, fire\_yrmean, lst, lst\_yrmean, precip\_yrmean, tmean, tmean\_yrmean, atmo\_pdep\_2002, atmo\_pdep\_2007, n\_cbnf\_2007, n\_crop\_n\_rem\_2007, n\_fert\_farm\_2007, n\_livestock.waste\_2007, n\_livestock\_n\_content\_2007, p\_crop\_removal\_2007, p\_livestock\_demand\_2007, p\_livestock\_waste\_2007, p\_livestock\_production\_2007, p\_nf\_fertilizer\_2007, p\_human\_food\_demand\_kg\_2007, p\_human\_nonfood\_demand\_kg\_2007, p\_human\_waste\_kg\_2007, p\_accumulated\_ag\_inputs\_2007, napi, total input'* [note that the ‘ here is not an accidental typo and is included in the name of the attribute], *legacy, damdensws, pcthbwet2011ws,* and *p2o5ws.* For all other attributes, we used min-max normalization. A quick plot of *logchl\_A* shows that the data is generally uniformly distributed (which is likely due to the log scale applied in this dataset), so we can use random sampling to split the dataset. We chose a split of 80/10/10 for this dataset, as we believe that ~200 validation/test instances can properly test the nuances of the dataset, while still leaving room for plenty of training data.