DSA 210: Introduction to Data Science
Fall 2025 - 2026

Exploratory Analysis and Machine Learning of Algal Growth Under Environmental Conditions

Prepared by: Başar Kaya
E-Mail Address: basark@sabanciuniv.edu
Submission Date: 09 January 2026

Abstract

This study explores algal growth dynamics using two complementary datasets: a controlled laboratory experiment on algae growth and the COIL 1999 environmental water-quality dataset. The project applies core data science methodologies, including data cleaning, exploratory data analysis (EDA), outlier detection, feature transformation, hypothesis testing, and machine learning, to investigate how environmental variables influence algal systems.

EDA reveals that the laboratory dataset exhibits strong non-linear behavior, particularly a unimodal relationship between light intensity and algal population, while showing negligible linear correlations across environmental variables. In contrast, the COIL dataset displays substantial variability, heavy right-skew in nutrient concentrations, and moderate correlations between nutrients and chlorophyll concentration, necessitating log transformation for meaningful analysis. A Pearson correlation test confirms the absence of a statistically significant linear relationship between light and algal population in the laboratory dataset, supporting the conclusion that algal growth in controlled conditions is governed by non-linear ecological responses rather than simple linear effects.

Building on these findings, supervised machine learning models are developed using the COIL dataset to predict chlorophyll concentration and mean algal abundance from environmental parameters. Model performance comparisons demonstrate that non-linear models outperform linear baselines and that chlorophyll, as a biochemical proxy for biomass, is more predictable than aggregated algal abundance. Overall, the results highlight the importance of non-linearity, proper preprocessing, and model selection when analyzing biological and environmental data.

Introduction

Algae play a central role in aquatic ecosystems and biotechnology, responding sensitively to environmental factors such as light availability, nutrient concentrations, and physicochemical water properties. Understanding how these variables influence algal growth is important not only for ecological monitoring and water-quality assessment, but also for applications in bioengineering, aquaculture, and environmental management. Due to the inherent complexity of biological systems, algal growth often exhibits non-linear responses and interactions that challenge simple analytical approaches.

From a data science perspective, algal datasets provide an opportunity to apply exploratory data analysis, statistical testing, and machine learning to real biological problems. However, such datasets frequently contain skewed distributions, extreme values, and heterogeneous measurement scales, requiring careful preprocessing and interpretation before meaningful conclusions can be drawn.

In this project, two distinct algae-related datasets are analyzed. The first is a controlled laboratory experiment designed to measure algal population responses under systematically varied environmental conditions. The second is the COIL 1999 environmental dataset, which captures real-world variability in water chemistry and algal indicators across multiple aquatic sites. Analyzing these datasets together allows for comparison between controlled experimental behavior and complex environmental dynamics.

The primary motivation of this study is twofold. First, it aims to investigate how exploratory data analysis and hypothesis testing can be used to identify linear and non-linear patterns in biological data. Second, it evaluates the effectiveness of machine learning models in predicting algal growth indicators from environmental parameters, comparing biochemical proxies such as chlorophyll concentration with community-level biological measures. By combining statistical analysis with predictive modeling, this project seeks to demonstrate a structured data science workflow applied to biologically meaningful problems.

Data Sources

This project uses two publicly available datasets related to algal growth under different conditions.

Dataset 1: Research on Algae Growth in the Laboratory
Source: Kaggle
Link:
https://www.kaggle.com/datasets/rukenmissonnier/research-on-algae-growth-in-the-laboratory/data
Description: This dataset represents a controlled laboratory experiment designed to study algal population responses under systematically varied environmental conditions. It contains approximately 9,800 observations with continuous measurements of Light intensity, nutrient concentrations (Nitrate, Iron, Phosphate), Temperature, pH, CO_2 levels, and resulting algal Population. Each environmental variable is sampled at fixed treatment levels, producing a full-factorial experimental structure. The dataset is synthetic in nature and intended to illustrate experimental responses rather than capture natural environmental variability.

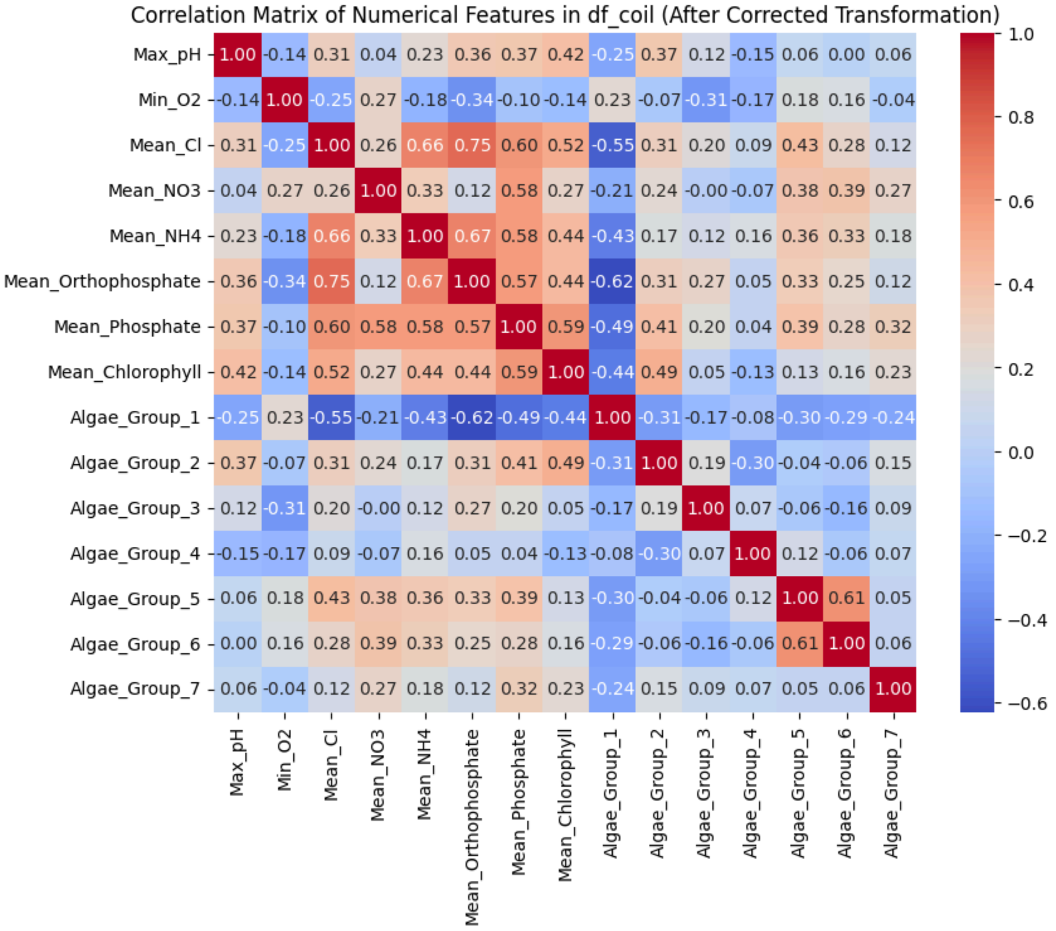
Dataset 2: COIL 1999 Competition Data
Source: UCI Machine Learning Repository
Link: https://archive.ics.uci.edu/dataset/118/coil+1999+competition+data
Description: The COIL 1999 dataset contains environmental and biological measurements from European rivers and streams. It includes physicochemical variables such as nutrient concentrations (NO_3, NH_4, Phosphate, Orthophosphate, Chloride), pH, dissolved oxygen, chlorophyll concentration, and the abundances of seven algal groups. Additionally, categorical descriptors (Season, Size, River_Size) characterize sampling conditions. Unlike the laboratory dataset, COIL reflects real-world environmental heterogeneity and contains substantial skewness and extreme values typical of ecological field data.

Data Description

The two datasets differ fundamentally in structure and purpose. The laboratory dataset is highly controlled, with evenly spaced experimental conditions and a single biological response variable, making it suitable for analyzing mechanistic growth patterns under idealized settings. In contrast, the COIL dataset captures complex environmental dynamics, where algal indicators emerge from the interaction of multiple uncontrolled factors. Key differences include:

- Scale: Laboratory dataset (approximately 9,800 rows) vs. COIL dataset (200 rows).
- Nature: Synthetic experimental design vs. real environmental observations.
- Targets: Direct population counts vs. biomass proxies (chlorophyll) and community-level abundance measures.

Because the datasets describe fundamentally different systems and do not share a common spatial or temporal key, they are analyzed separately throughout the project. The laboratory dataset is primarily used for exploratory analysis and hypothesis testing, while the COIL dataset serves as the basis for machine learning modeling. To summarize the relationships among physicochemical variables in the COIL dataset, a correlation heatmap based on log-transformed features is presented at the end of this section.



Methodology

This study follows a structured data science workflow consisting of data cleaning, exploratory data analysis (EDA), feature transformation, hypothesis testing, and machine learning. The methodology is designed to ensure interpretability, avoid data leakage, and appropriately handle the non-linear and skewed characteristics of environmental data.

Data Cleaning and Preparation

For both datasets, initial checks were performed to verify data types, missing values, and plausible ranges.

- In the laboratory dataset, no missing values were observed; rows with zero algal population were retained as valid biological outcomes.
- In the COIL dataset, numeric variables were converted to appropriate numeric types and categorical variables (Season, Size, River_Size) were encoded for modeling. No rows were removed.

Outlier Assessment

Outliers were assessed using boxplots and interquartile range (IQR) criteria.

- The laboratory dataset showed no problematic outliers under IQR rules.
- The COIL dataset exhibited extensive right-skew and a high proportion of extreme values across nutrient and algal variables. Because these extremes reflect real environmental variability, no rows were removed. Instead, transformation-based handling was preferred.

Feature Transformation

To reduce skewness and stabilize variance in the COIL dataset, log transformations were applied to heavily skewed nutrient variables (e.g., nitrate, ammonium, phosphate, chloride), chlorophyll, and algal abundance measures. Variables such as pH and dissolved oxygen, which already exhibited near-symmetric distributions, were left untransformed. This step improved interpretability and suitability for correlation analysis and machine learning.

Exploratory Data Analysis

EDA focused on understanding distributions, relationships among variables, and structural differences between the datasets.

- In the laboratory dataset, correlation analysis revealed near-zero linear correlations, motivating further investigation of non-linear patterns.
- In the COIL dataset, correlation analysis on log-transformed features revealed moderate associations among nutrients and between nutrients and chlorophyll. A correlation heatmap summarizing these relationships is included at the end of the Data Description section to motivate feature selection for modeling.

Preparation for Hypothesis Testing and Machine Learning

Based on EDA findings, the laboratory dataset was used for hypothesis testing focused on linear versus non-linear relationships, while the COIL dataset was selected for machine learning due to its real-world variability. For machine learning, algal group variables were excluded from the feature set to prevent data leakage, and two prediction targets were defined: chlorophyll concentration and mean algal abundance. Cross-validation procedures were used to ensure consistent and fair model evaluation.

Hypothesis Testing

The laboratory algae growth dataset was used to formally test whether algal population exhibits a statistically significant linear relationship with environmental variables, focusing specifically on light intensity. Exploratory analysis suggested that algal population responds non-linearly to light, motivating a hypothesis test to evaluate whether a linear association exists.

Hypothesis Formulation

The following hypotheses were defined:

- Null Hypothesis (H_0): There is no statistically significant linear correlation between light intensity and algal population.
- Alternative Hypothesis (H_1): There is a statistically significant linear correlation between light intensity and algal population.

Method

The Pearson correlation coefficient was selected to test linear dependence between light intensity and algal population. Pearson correlation is appropriate for assessing linear relationships and provides a clear statistical framework for hypothesis testing through correlation coefficients, p-values, and degrees of freedom.

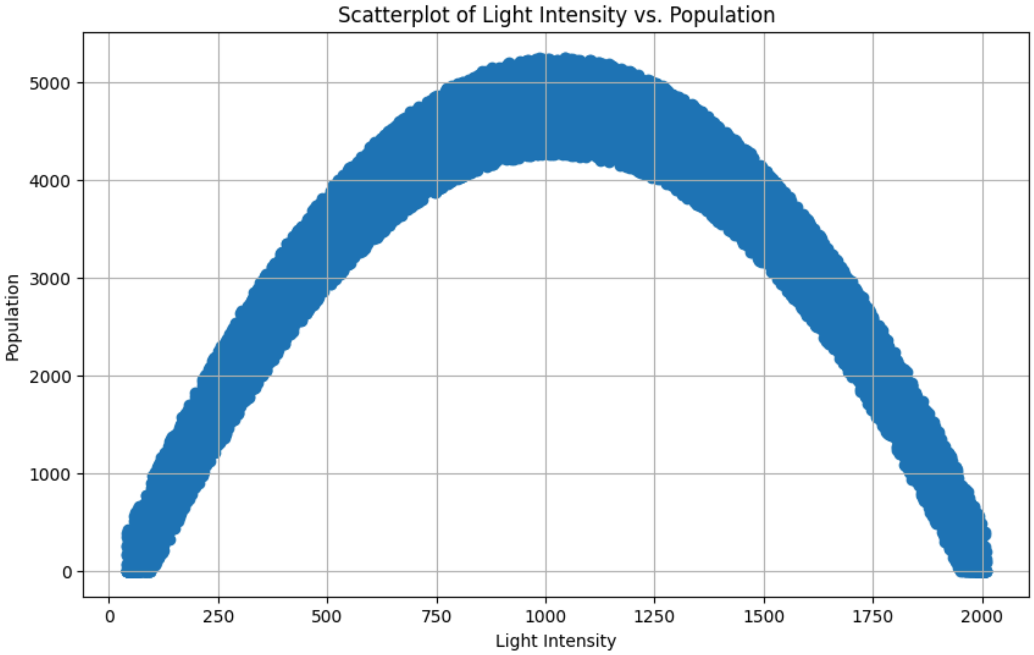
Results

The Pearson correlation analysis produced a correlation coefficient close to zero and a non-significant p-value ($p > 0.05$), indicating no evidence of a linear relationship between light intensity and algal population. Given the large sample size of the dataset, this result is robust and not due to lack of statistical power.

Interpretation

Although the Pearson test indicates no linear correlation, visualization of the data reveals a clear unimodal (U-shaped) relationship between light intensity and algal population, with growth increasing up to an optimal light level and decreasing beyond it due to photo-inhibition. This non-linear biological response explains why the Pearson correlation coefficient approaches zero: positive and negative trends cancel out when measured linearly.

The results therefore support the null hypothesis, not because light has no effect on algal growth, but because its effect is non-linear rather than linear. This finding highlights the importance of combining statistical tests with visual exploration and reinforces the limitations of linear correlation methods when applied to biological systems.



Machine Learning Implementation

Machine learning models were developed using the COIL 1999 environmental dataset to evaluate the extent to which algal growth indicators can be predicted from physicochemical environmental variables. This dataset was selected for modeling because it reflects real-world environmental variability and provides suitable continuous targets for supervised learning.

Problem Definition

Two regression tasks were defined to compare model performance across different biological targets:

- Chlorophyll Prediction: Chlorophyll concentration was used as a biochemical proxy for total algal biomass.
- Mean Algal Abundance Prediction: A second target was defined as the mean frequency of all algal groups, representing a community-level biological indicator.

These two targets allow comparison between predicting a biochemical outcome and a more aggregated biological response.

Feature Selection

The feature set consisted exclusively of environmental variables, including log-transformed nutrient concentrations, physicochemical parameters (pH and dissolved oxygen), and encoded categorical descriptors (Season, Size, River_Size). Individual algal group variables and chlorophyll were excluded from the feature matrix to prevent data leakage and ensure valid predictive modeling.

Modeling Approach

Multiple regression models were evaluated for each target:

- Linear Regression as a baseline model to assess linear predictability.
- k-Nearest Neighbors (kNN) to capture local non-linear patterns.
- Random Forest Regressor to model complex non-linear relationships and interactions among features.

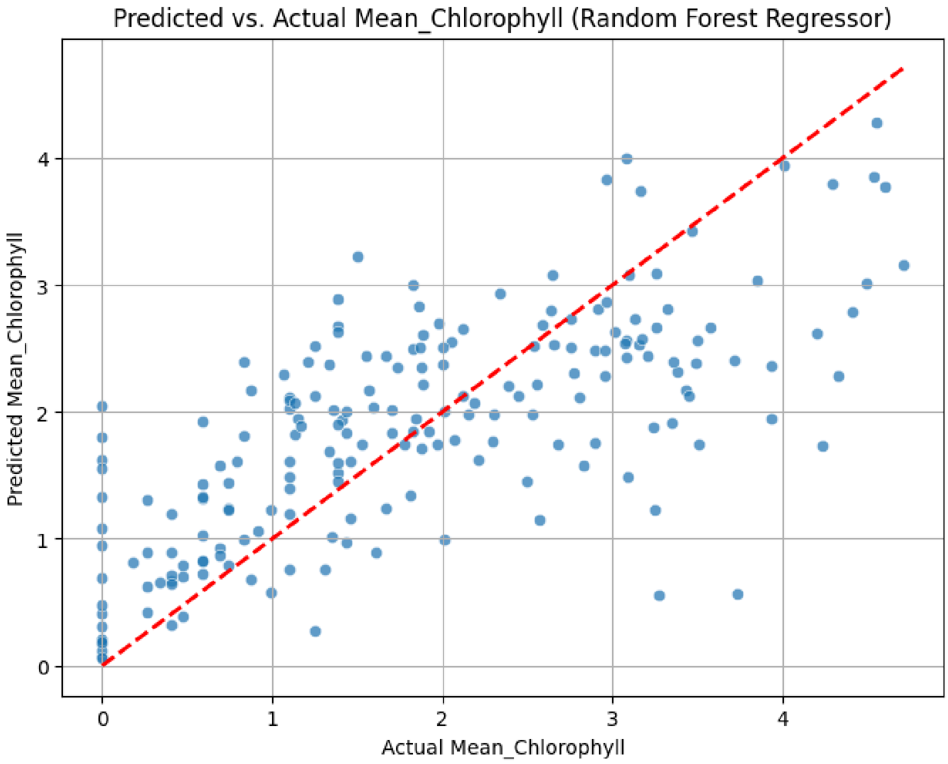
To ensure fair comparison, a 5-fold cross-validation strategy was used with consistent fold assignments across both targets. Feature scaling was applied where required to support distance-based and linear models.

Evaluation Metrics

Model performance was evaluated using:

- R² (coefficient of determination) to measure explained variance,
- RMSE (root mean squared error) to quantify average prediction error, and
- MAE (mean absolute error) to assess typical absolute deviations.

These metrics provide complementary perspectives on predictive accuracy and model robustness.



Results

This section presents the quantitative outcomes of the machine learning models developed to predict chlorophyll concentration and mean algal abundance from environmental variables in the COIL dataset. Model performance is evaluated using cross-validated metrics and compared across targets to assess predictability.

Chlorophyll Prediction Results

Among the evaluated models, the Random Forest regressor achieved the best overall performance in predicting chlorophyll concentration. It explained approximately 47% of the variance ($R^2 \approx 0.47$), outperforming both Linear Regression and k-Nearest Neighbors models. The relatively low standard deviation across cross-validation folds indicates stable performance and good generalization. Error metrics (RMSE and MAE) were consistently lower for the Random Forest model, confirming its superior predictive accuracy.

These results suggest that chlorophyll concentration is moderately predictable from physicochemical environmental parameters and that non-linear relationships and feature interactions play an important role in determining algal biomass.

Mean Algal Abundance Prediction Results

Predicting mean algal abundance proved more challenging. All models exhibited substantially lower performance compared to the chlorophyll prediction task. The Random Forest regressor again performed best, but with a considerably lower explained variance ($R^2 \approx 0.22$). Linear Regression and k-Nearest Neighbors models showed similar or weaker performance, indicating limited predictive structure in the data for this target.

The higher error values and lower R^2 scores reflect the increased biological complexity and noise associated with aggregated community-level abundance measures.

Comparison of Targets

Comparing the two prediction tasks reveals a clear contrast. Models consistently performed better when predicting chlorophyll concentration than when predicting mean algal abundance. This indicates that biochemical biomass proxies are more directly associated with environmental drivers than community-level biological measures, which are influenced by additional ecological processes not captured by the available features.

Overall, the results demonstrate that non-linear machine learning models improve predictive performance over linear baselines, but also highlight intrinsic limits to predictability in complex ecological systems.

Discussion

The results of this study highlight the importance of combining exploratory analysis, statistical testing, and machine learning when working with biological and environmental datasets. Across all stages of the analysis, non-linearity and complexity emerged as defining characteristics of algal growth dynamics.

Exploratory data analysis revealed fundamental differences between the two datasets. The laboratory dataset, despite its large size, showed almost no linear correlations among variables. However, visual inspection demonstrated a clear unimodal relationship between light intensity and algal population. This pattern, confirmed through hypothesis testing, illustrates how reliance on linear statistics alone can be misleading when analyzing biological systems. The failure to detect a significant Pearson correlation does not imply the absence of a relationship, but rather reflects the non-linear nature of algal responses to environmental conditions.

The COIL dataset presented a contrasting scenario. As a real-world environmental dataset, it exhibited substantial variability, skewed distributions, and moderate correlations among physicochemical variables and chlorophyll concentration. Log transformation proved essential for meaningful interpretation and downstream modeling, reinforcing the importance of preprocessing decisions in environmental data science workflows.

Machine learning results were consistent with insights gained from EDA and hypothesis testing. Non-linear models, particularly Random Forests, outperformed linear baselines, confirming that algal growth indicators depend on complex interactions and threshold effects rather than simple linear relationships. The superior performance observed when predicting chlorophyll concentration suggests that biochemical proxies of biomass are more directly linked to environmental drivers than aggregated measures of algal community abundance. In contrast, the relatively poor predictability of mean algal abundance reflects the influence of unobserved ecological factors such as species interactions, grazing pressure, and hydrological variability.

Taken together, these findings demonstrate that while machine learning can capture meaningful patterns in environmental data, its performance is constrained by both data quality and ecological complexity. The alignment between EDA observations, hypothesis testing outcomes, and machine learning results strengthens the validity of the conclusions and illustrates a coherent, end-to-end data science approach applied to a biological problem.

Limitations and Future Work

While this project demonstrates a complete data science workflow applied to algal growth datasets, several limitations should be acknowledged. First, the laboratory dataset is synthetic and based on a controlled factorial design. Although it is useful for illustrating non-linear biological responses, it does not capture the full complexity and stochasticity of natural ecosystems. As a result, findings from this dataset should be interpreted as illustrative rather than representative of real environmental dynamics.

Second, the COIL dataset, despite being environmentally realistic, is relatively small in size and limited to a specific set of measured variables. Important ecological drivers such as light history, hydrological flow, grazing pressure, and seasonal succession are not explicitly captured, which constrains the predictive performance of machine learning models. Additionally, chlorophyll and algal abundance measures represent aggregated indicators and do not account for species-level differences in growth behavior.

Future work could address these limitations in several ways. More advanced non-linear modeling approaches, such as generalized additive models (GAMs), could be applied to explicitly model unimodal and threshold-based relationships observed in the laboratory dataset.

For the COIL dataset, expanding the feature set with additional environmental or temporal variables may improve predictive performance. Machine learning models could also be extended to multi-output or hierarchical frameworks to capture relationships among different algal groups more explicitly.

Finally, future studies could explore transferability between controlled experiments and environmental observations by comparing model behavior across datasets. Such work would contribute to a deeper understanding of how experimental insights translate to real-world ecological systems and further demonstrate the potential of data science methods in environmental and biological research.