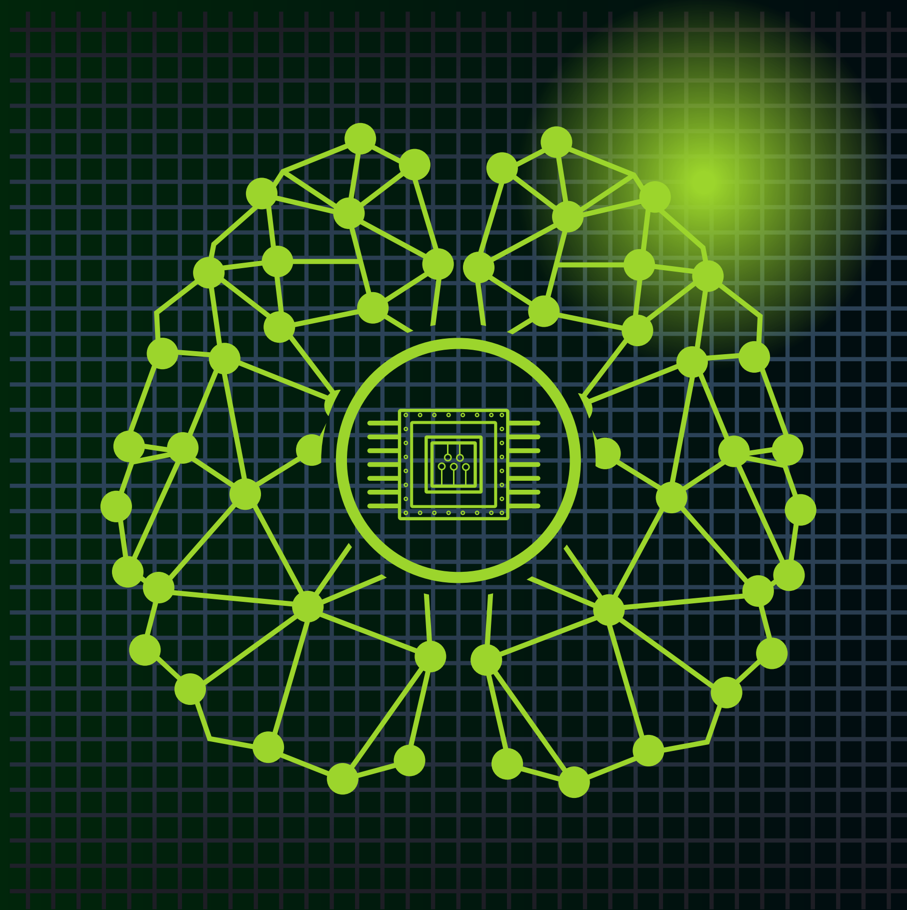
MACHINE LEARNING

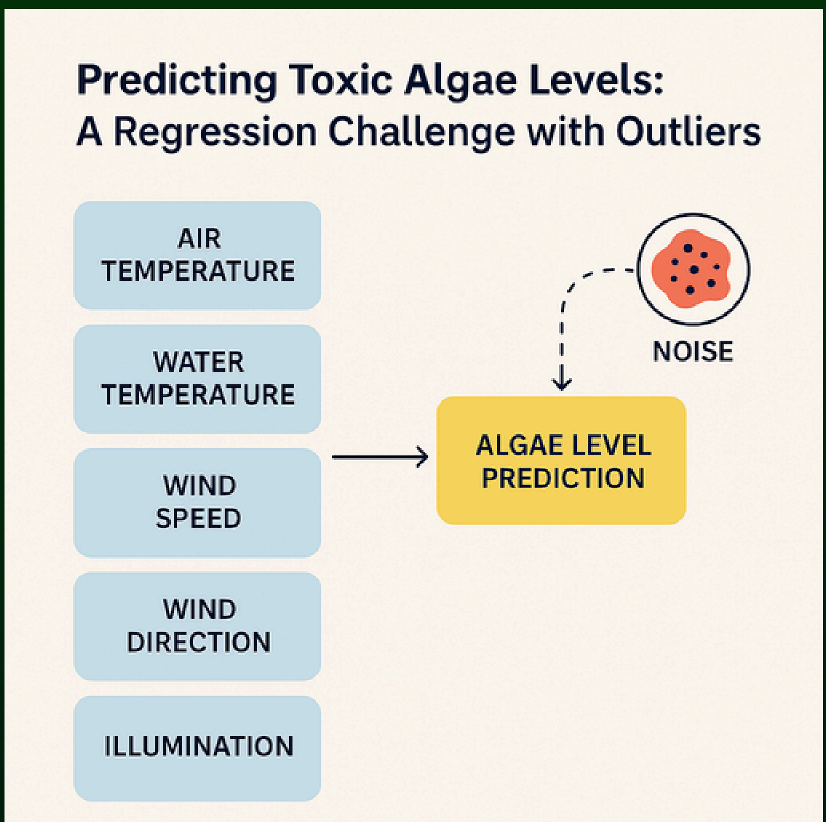
PREDICTING TOXIC ALGAE LEVELS: A REGRESSION CHALLENGE WITH OUTLIERS



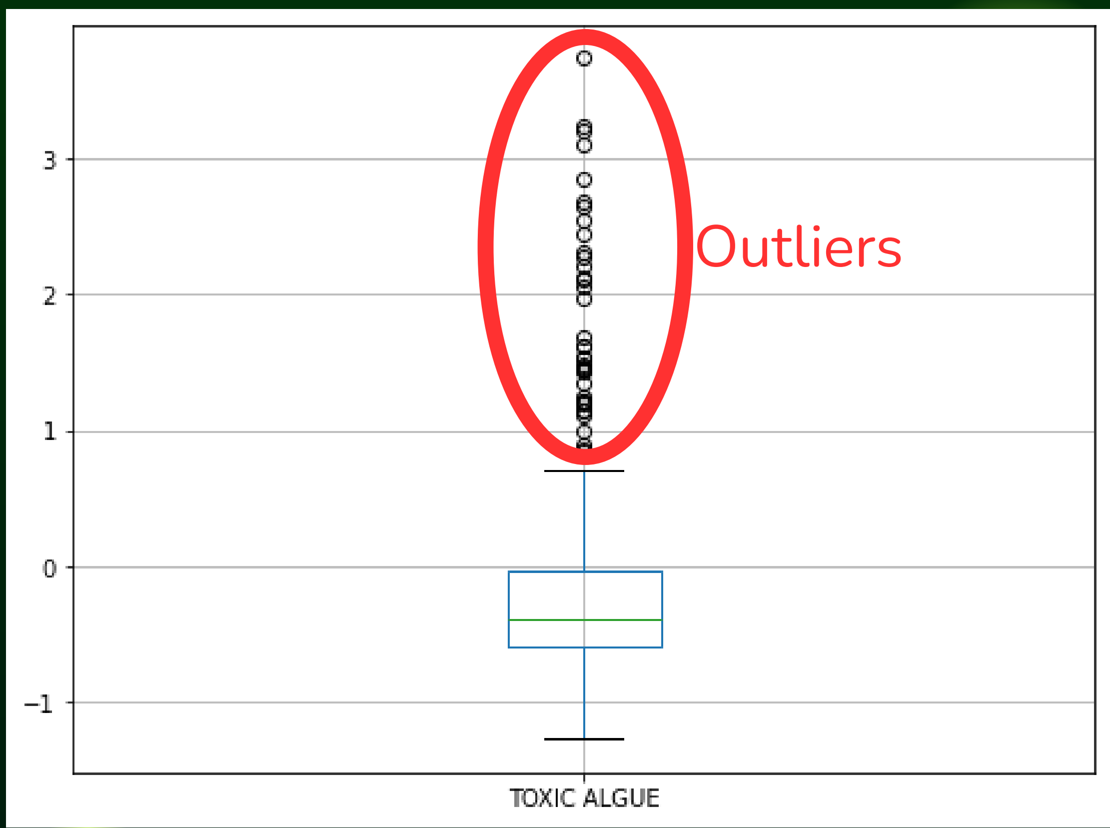
PROBLEM INTRODUCTION

Given environmental data (air/water temperature, wind, illumination), predict toxic algae concentration using linear regression.

Data is contaminated with instrument noise (Gaussian) and human error (~25% of y values) causing outliers.



Problem diagram.



Boxplot for dependent variable.

Objective: Estimate algae level (y) from sensor inputs (X) despite noisy readings.

To address the outlier challenge, we explored two iterative regression approaches: Iterative Outlier Removal (IOR) and RANSAC Regression.

APPROACH 1 — ITERATIVE LASSO + OUTLIER REMOVAL

We remove outliers step by step (Goal: Improve regression robustness by iteratively removing outliers):

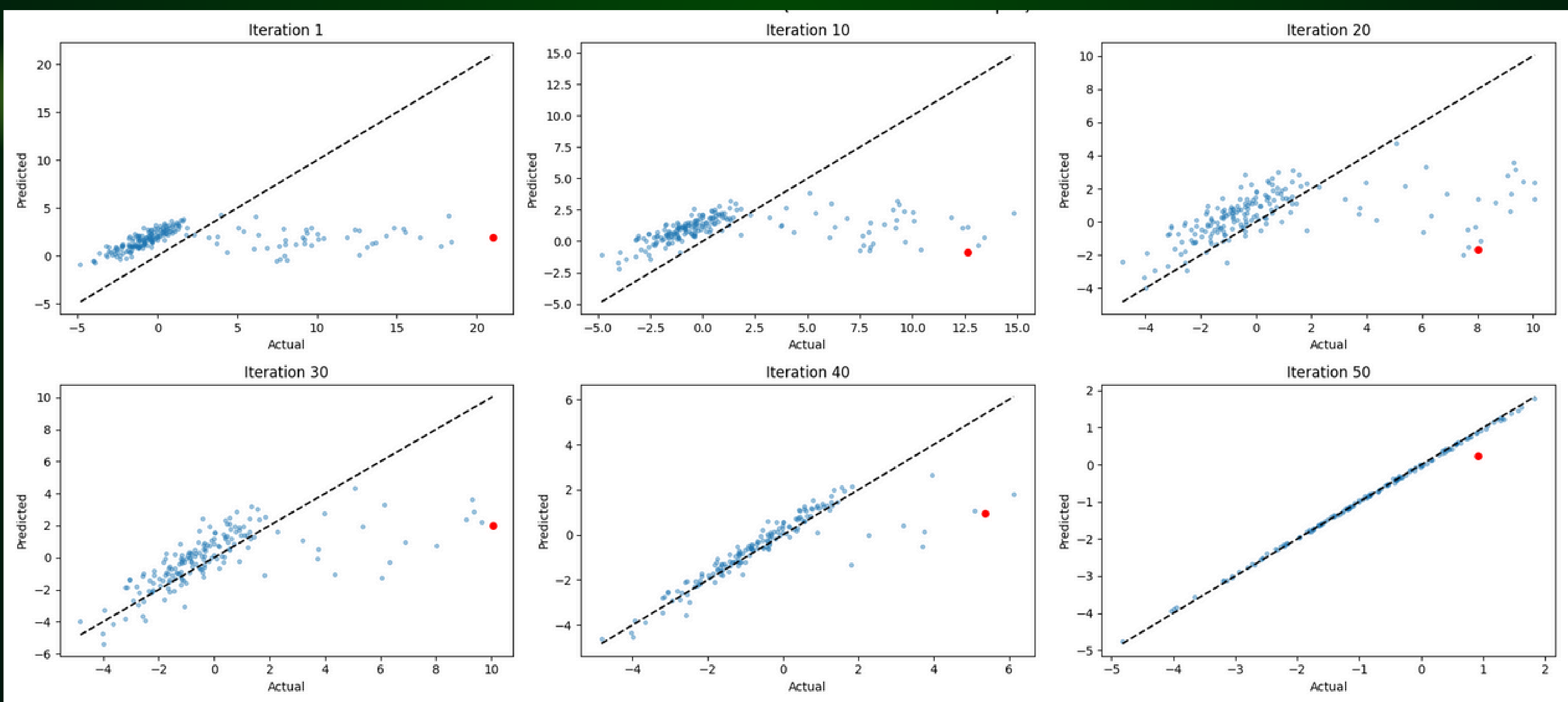
01 **Fit a Lasso regression model**

02 **Identify the sample with the highest error**

03 **Remove it**

04 **Repeat for 50 iterations**

Final model is trained on cleaned data, improving robustness.

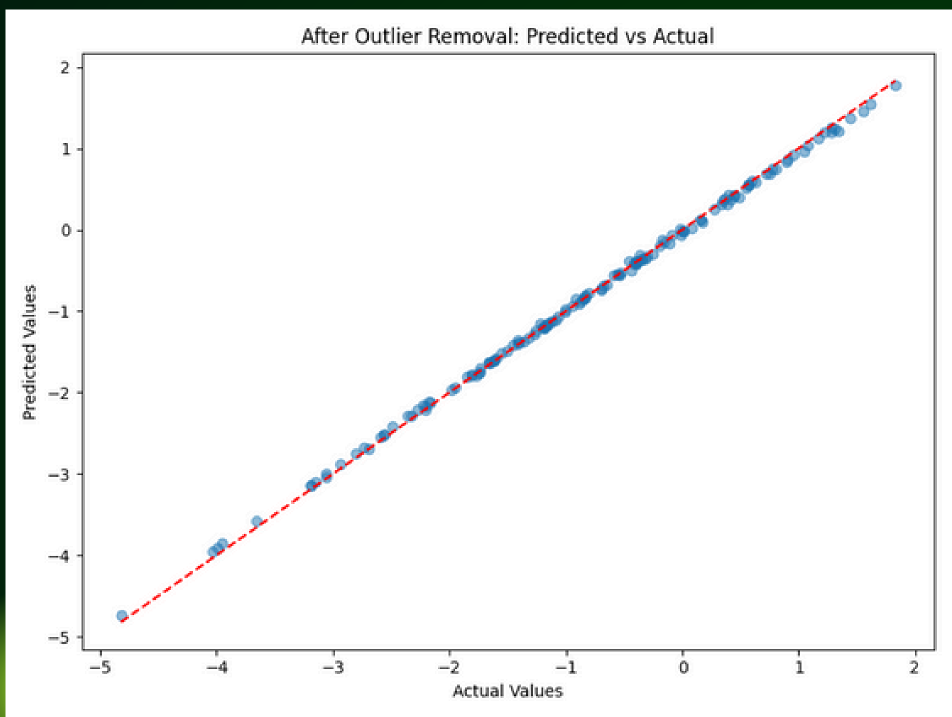


Predicted vs. Actual Values - Iterations

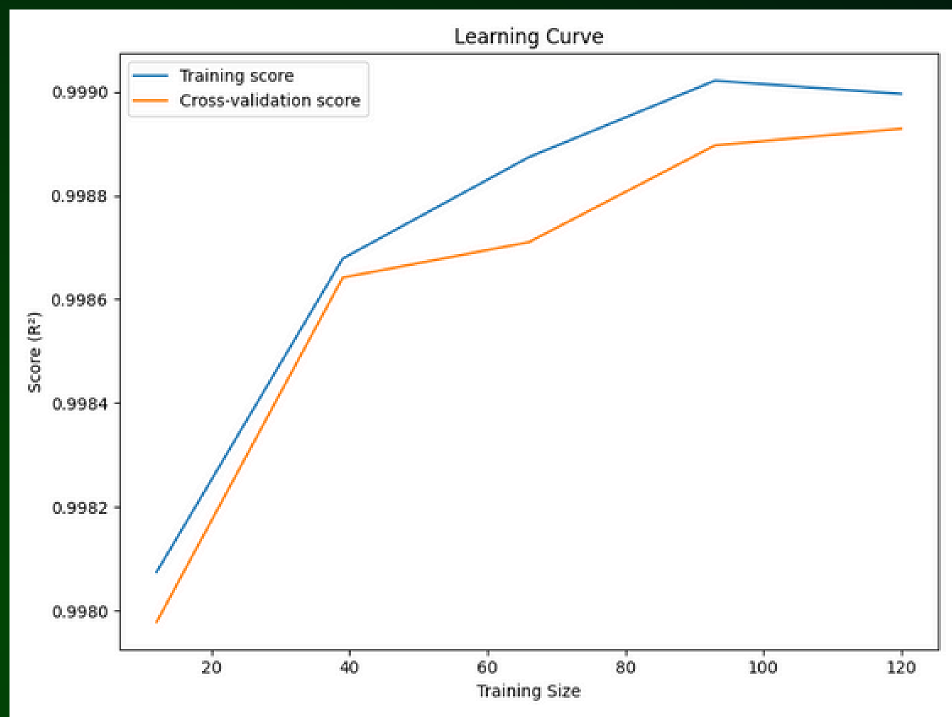
APPROACH 1 — ITERATIVE LASSO + OUTLIER REMOVAL

RESULTS

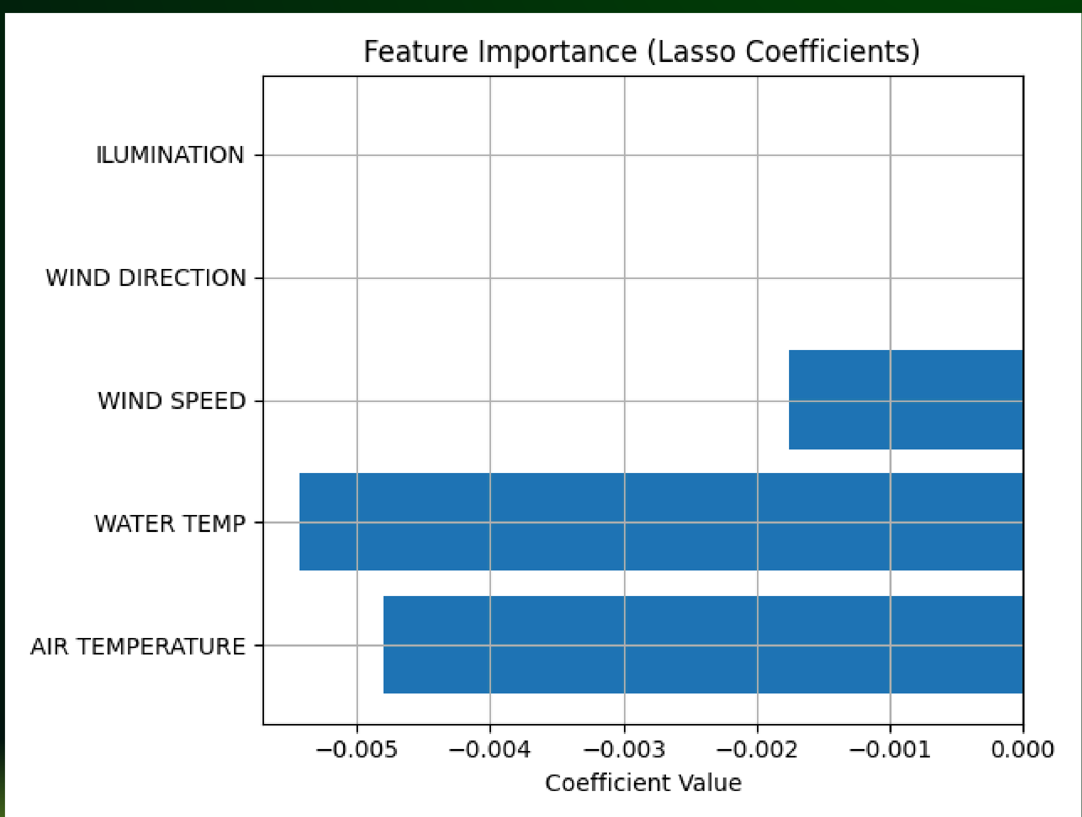
Better fit after eliminating high-error samples.
Evaluated using cross-validation (R^2 scores)
and learning curves.



Predicted vs. Actual Values - After IOR



Learning Curve



Feature Importance Graph

"Water temperature and air temperature are the most influential predictors — matching domain expectations."

Coeficientes	
β0 (intercept)	0.701985
β1 (AIR TEMP)	-0.0048
β2 (WATER TEMP)	-0.00543
β3 (WIND SPEED)	-0.00175
β4 (WIND DIR)	0
β5 (ILUM)	0

Cross Validation R2	
Average	0.9988

APPROACH 2 – RANSAC FOR ROBUST FITTING

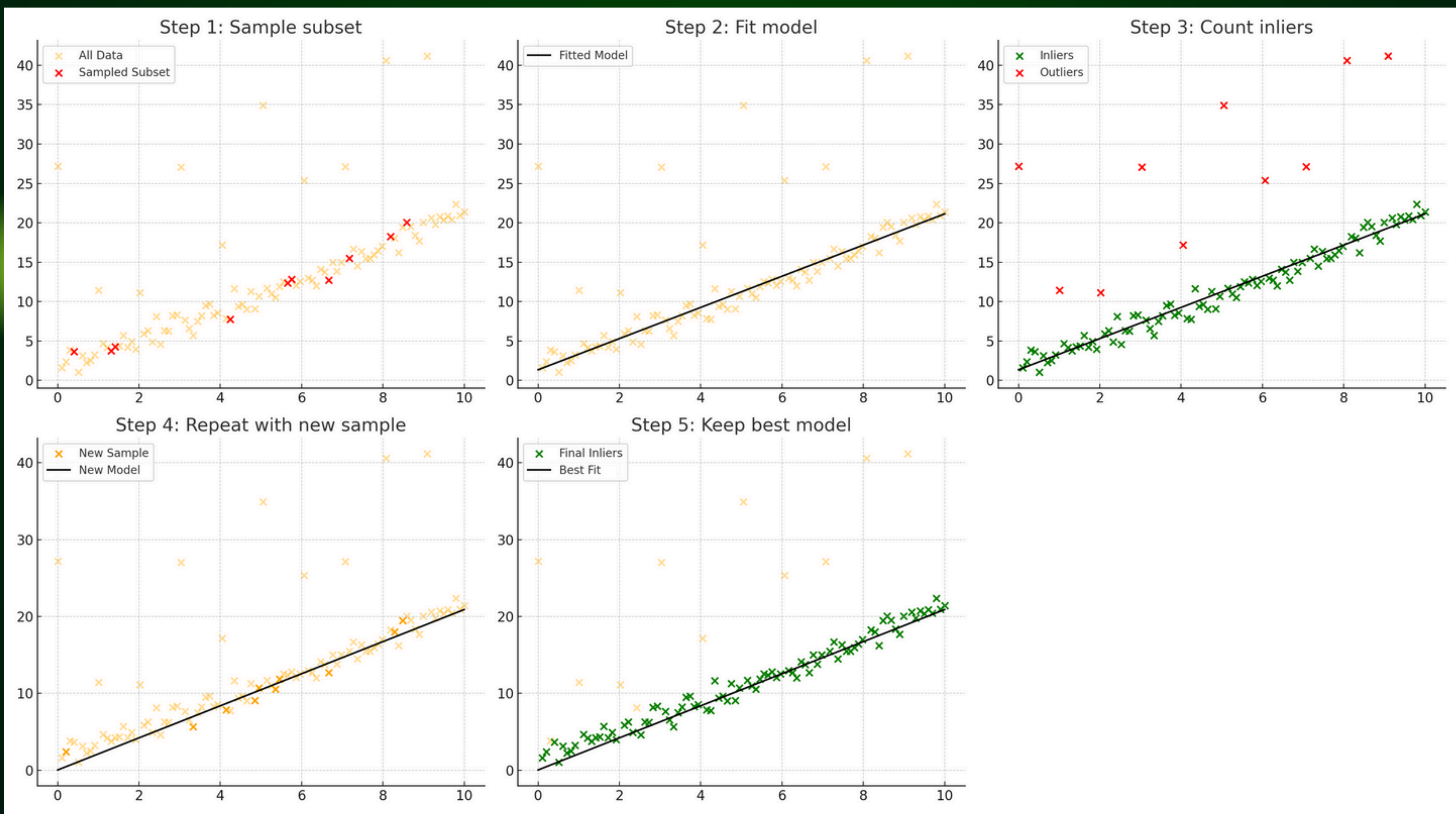
The steps to follow are:

01 **Select a random subset of data points.**

02 **Fit a Lasso regression model on this subset**

03 **Count how many full-data points are inliers (within residual threshold).**

04 **Repeat for 100 iterations and keep the model with the most inliers.**

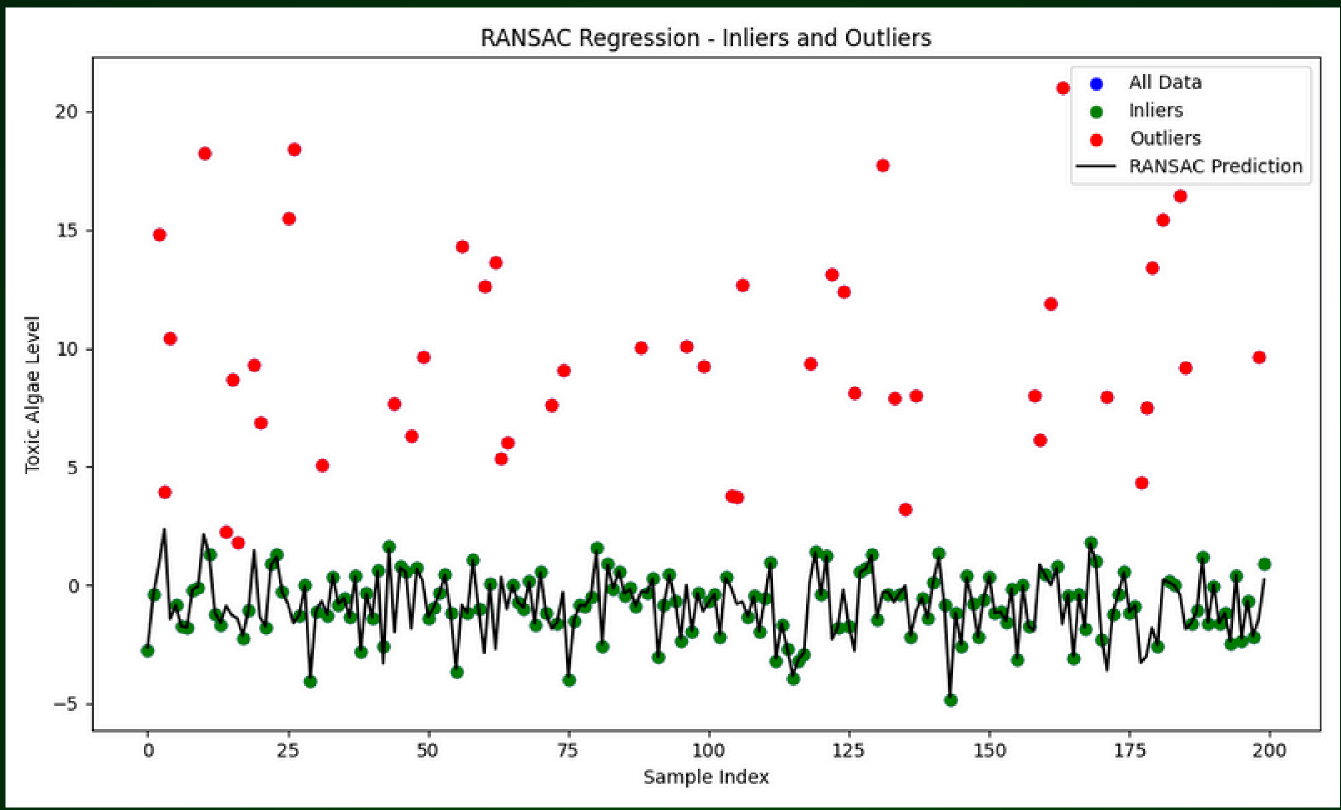


Explanation Graph of RANSAC Iterative Process

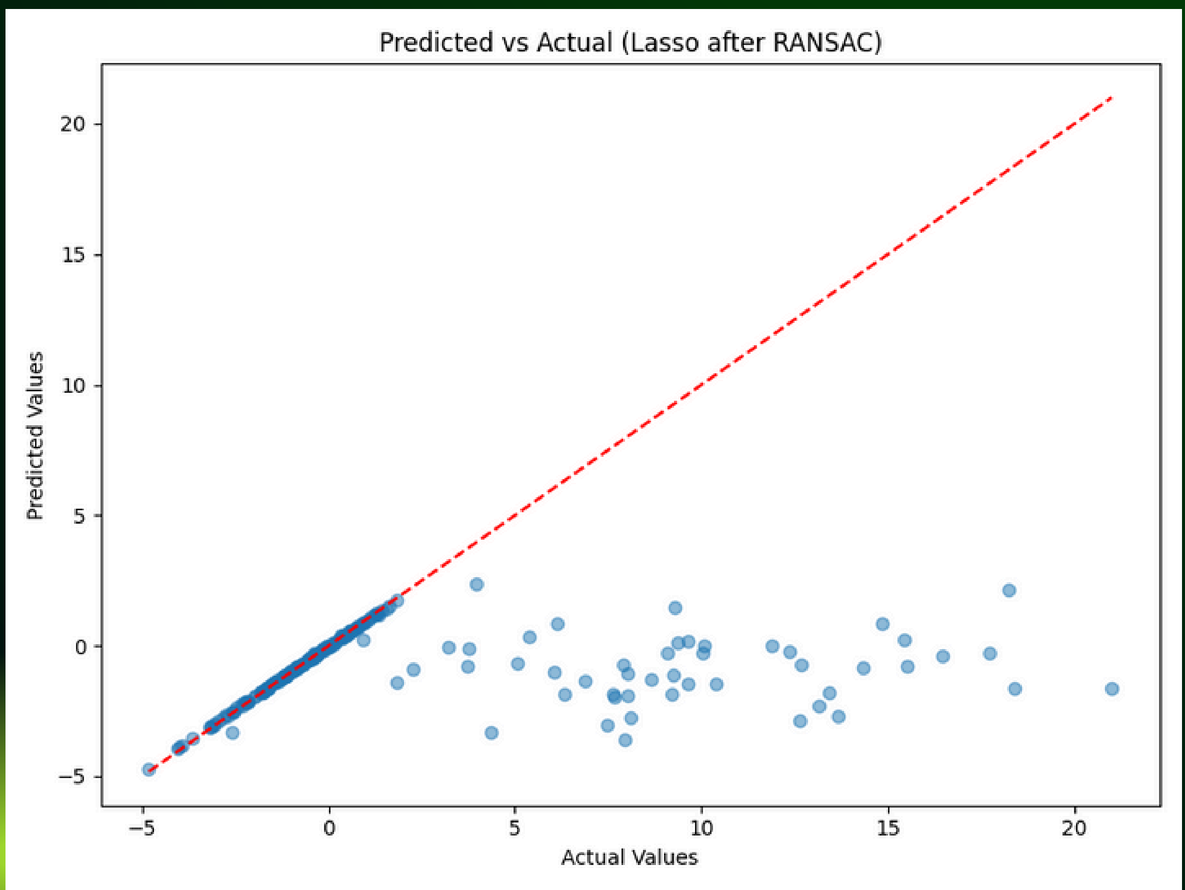
APPROACH 2 – RANSAC FOR ROBUST FITTING

RESULTS

Model performance evaluated with cross-validation R^2 and MSE.
Identified and visualized inliers/outliers effectively.
RANSAC effectively isolates extreme values, training a clean model on inliers only. This improves generalization and stabilizes feature interpretation.



RANSAC Inliers and Outliers



Predicted vs Actual (After RANSAC)

Coeficientes	
β0 (intercept)	0.70662
β1 (AIR TEMP)	-0.00478
β2 (WATER TEMP)	-0.00533
β3 (WIND SPEED)	-0.00179
β4 (WIND DIR)	0
β5 (ILUM)	0

Cross Validation R2	
Average	‘0.99403

CONCLUSION – ROBUST REGRESSION FOR NOISY DATA

- Outliers significantly affect regression models, leading to poor generalization and biased coefficients.
- Two iterative techniques — IOR and RANSAC — were applied to handle outliers effectively, improving robustness and interpretability.
- Both approaches achieved excellent performance, with an average cross-validation $R^2 \approx 0.99$ on a dataset of 200 samples.
- These results highlight the importance of outlier handling in small, real-world datasets commonly affected by noise or human error.
- Robust regression techniques like these are essential tools in applied machine learning pipelines where data quality is not guaranteed.

Robust regression methods significantly improve model reliability when dealing with noisy or contaminated data.