Predictive Modeling for Water Quality Assessment

Sai Sushma Maddali

San Jose State University

DATA 245: Machine Learning Technologies

Prof. Vishnu S. Pendyala

October 6, 2025

Abstract

This study presents a machine learning based approach to classify water stations according to their pollution risk levels. Utilizing publicly available water quality datasets from California, the project computes Water Quality Index (WQI) derived from both laboratory and field measurements. Based on WQI, water stations are categorized into three risk levels, low, medium, and high. The classification task is performed using machine learning algorithms such as Random Forest, Decision Trees, and Support Vector Machines. The model further identifies key water quality indicators contributing to contamination risks across stations. The modeling pipeline encompasses data preprocessing, feature engineering, model training, validation and performance evaluation using precision, recall and F1 score metrics. To interpret model outputs and determine influential features SHAP (Shapley Additive explanations) is used. The proposed solution aims to support early detection of unsafe water conditions, thereby contributing to improved environmental health and public safety.

Motivation

Ensuring access to clean water and sanitation is a key objective outlined in the United Nations Sustainability Development Goals. However, water pollution remains a persistent concern, especially in heavily industrialized regions. Its consequences include ecological degradation, proliferation of waterborne diseases and human exposure to hazardous substances such as heavy metals. Traditional monitoring methods are often labor-intensive and limited in scope. This project leverages machine learning to classify pollution risk levels across water stations for faster and more scalable analysis. By enabling early detection, the model supports proactive environmental management aligned with United Nations Sustainability Development Goals.

Literature Survey

Recent research demonstrates the effectiveness of machine learning algorithms, such as Extreme Gradient Boosting (XGBoost), in analyzing water quality (Yuan et al., 2025). The Water Quality Index (WQI) serves as a composite metric for assessing overall water health. Preiya, Subramanian, Soniya, and Pugalenthi (2024) introduced a deep learning approach using Long Short-Term Memory (LSTM) networks and Grasshopper Optimization to enhance WQI prediction. Their model showed improved generalization in temporal datasets. Building on this foundation, the current project proposes a multi-class classification model that uses WQI as a derived label to categorize water stations into low, medium, or high pollution risk. Unlike prior work that treats WQI as continuous output, this approach enables more actionable insights for regulatory agencies.

Methodology

Four water quality datasets from California will be merged using water station identifier. A representative sample will be selected based on specific counties and pollutant parameters, with appropriate missing values and feature scaling. The target variable will be derived by computing the Water Quality Index (WQI), which aggregates multiple parameters into a single score used to classify water stations as low, medium, or high pollution risk. Machine learning models including Random Forest, Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost) will be trained and evaluated using precision, recall and F1 score. Feature importance will be assessed using SHAP (Shapley Additive explanations) to identify key indicators of contamination.

Deliverables

Upon project completion, the following items will be submitted:

- A final report detailing the analysis, methodology, findings, and UI testing instructions.
- Source code and trained machine learning models hosted on GitHub.
- Visualizations illustrating key data insights.
- A deployed Streamlit-based user interface accessible via Streamlit Cloud.

Milestones

- Week 1-2: Data acquisition, cleaning, preprocessing, and feature engineering.
- Week 3-4: Model training and validation using selected algorithms.
- Week5-6: Model interpretation using SHAP, performance evaluation, and documentation

Team members and their roles

The project team comprise of four members: Aakash, Kruthi, Sushma and Vedika. All members will contribute equally and share responsibility for deliverables. Initial tasks include independent data exploration followed by collaborative strategy development for data preparation. During the modeling phase, each member will train models on distinct data subsets to compare performance.

Specific roles are as follows:

- Aakash and Sushma - Lead model development and implement user interface using Streamlit
- Kruthi - Develop data visualizations to communicate key insights.

- Vedika - Manage project documentation and maintain GitHub repository.

All members will participate in writing final report and supporting documentation.

Relevance to the course

The project aligns with course objectives by applying core machine learning techniques such as classification algorithms, model validation, and performance evaluation to a real-world environmental challenge. The problem statement also supports the United Nations Sustainability Development Goals, as emphasized in the course, reinforcing its academic and societal relevance.

Technical difficulty

This project involves several key challenges. First, the water quality data is distributed across four distinct datasets, requiring accurate integration using appropriate join keys to ensure consistency. Second, due to large data volume, a representative sampling strategy must be employed to preserve critical patterns for model training. Third, the target variable must by engineered by computing the Water quality Index (WQI), which demands domain expertise and standardized formulas. Finally, developing a reliable multi-class classification model to assess pollution risk by careful algorithm selection, hyperparameter tuning , and robust validation across all classes.

Novelty

While prior studies have focused on predicting the Water Quality Index (WQI) as a continuous variable using models such as XGBoost and LSTM, this project adopts a distinct approach by utilizing WQI as a derived label for multi-class classification. Water stations are classified as low, medium, and high pollution risk, offering more actionable insights for environmental monitoring. Furthermore, the use of public datasets spanning multiple counties in California enhances spatial generalization, extending the model's applicability beyond specific water bodies.

Impact

Given the increasing global emphasis on water pollution management, this project holds strong potential for publication. The proposed model can assist government agencies in water quality monitoring. Additionally, the findings contribute to ongoing research in environmental monitoring and sustainability.

Heilmeier catechism

1. What are you trying to do? Articulate your objectives using absolutely no jargon.

Analyze water quality data to classify water stations by its pollution risk level.

2. How is it done today, and what are the limits of current practice?

Water quality is currently assessed through manual sampling and periodic testing. Predictive models are not widely implemented, limiting scalability and responsiveness.

3. What is new in your approach and why do you think it will be successful?

The project applies machine learning to historical water quality data to enable early detection and proactive monitoring.

4. Who cares? If you are successful, what difference will it make?

Environmental scientists and public will benefit from improved water quality monitoring. If successful, a better understanding of water quality could lead to more informed decisions to protect water resources.

5. What are the risks?

Key risks include challenges in integrating large datasets, potential data quality issues and accurately computing WQI.

6. How much will it cost?

No financial cost is anticipated, as the project utilizes open-source datasets and tools for development.

7. How long will it take?

The proof-of-concept project is planned for 8 weeks. The first half focuses on data preparation and model training, and the second half on model evaluation, interpretability, UI development, and documentation.

8. What are the mid-term and final "exams" to check for success?

Mid-term success will be measured by successful data integration and WQI computation and a working model with acceptable precision, recall and F1 score. Final success will be determined by deployment on SHAP based interpretability and comprehensive report.

References

Yuan, P., Li, H., Yi, X., et al. (2025). Optimizing water quality index using machine learning: a six-year comparative study in riverine and reservoir systems. *Scientific Reports*, *15*, Article 33919. https://www.nature.com/articles/s41598-025-10187-8

Preiya, S.P.V. M., Subramanian, P., Soniya, M., & Pugalenthi, R. (2024). Water quality index prediction and classification using hyperparameter tuned deep learning. *Global NEST Journal*, *26(4),* 1-12. https://journal.gnest.org/system/files/2024-06/gnest_05821_published.pdf