Model comparison for Water Quality Predictions

Omkar Singh[1], Rishidev Mishra[2], Prathamesh Vaidya[3]
[1] H.O.D (Data Science), [2] [3] P.G. Student
[1] [2] [3] Thakur College of Science & Commerce, Mumbai, India
[1]omkarsingh@tcsc.edu.in, [2]rishimishra20014@gmail.com, [3]prathameshsvaidy77@gmail.com

Abstract
This project develops a water quality prediction system using machine learning to assess and classify water quality accurately, efficiently and effectively. It collects data from government sources, preprocesses missing values, and applies indices like NSFWQI. Various machine learning models like linear regression, decision tree regressinr, support vector regressor, random forest regressor and gradient boosting regressor which were used to predict water quality, evaluated using MAE, MSE RMSE and R2. The project demonstrates how AI can enhance water quality assessment for household, industrial, and environmental applications.

Keyword:
Water Quality Prediction, Machine Learning, Water Quality Index (WQI), Feature Engineering, Data Preprocessing, Environmental Assessment.

1. Introduction

Groundwater is a critical resource supporting drinking water, agriculture, and industry, especially in regions with limited surface water access. However, contamination from industrial waste, agricultural runoff, and improper disposal poses significant risks to human health and ecosystems. Traditional monitoring methods, relying on periodic sampling, often fail to capture real-time fluctuations in water quality.

Artificial Intelligence (AI) presents a transformative approach to groundwater quality prediction. By analyzing large datasets—including historical water records, environmental factors, and land use—AI models can identify contamination patterns, predict water quality trends, and assist decision-makers in mitigating risks. Techniques like machine learning enhance the ability to handle complex interactions among multiple factors, improving prediction accuracy.

AI-driven models, especially enables real-time accurate and responses to contamination. Despite challenges like data availability, AI offers a powerful complementary tool for ensuring sustainable groundwater management and public health protection.

2. Literature Review

A literature review on groundwater quality prediction using AI models explores the advancements in predictive techniques, datasets, and methodologies employed in this field. Various studies emphasize the necessity of monitoring groundwater parameters such as pH, Biochemical Oxygen Demand (BOD), Total Dissolved Solids (TDS), and temperature. The integration of AI models has significantly improved the accuracy and efficiency of groundwater quality assessment, reducing dependency on traditional monitoring systems. Machine learning algorithms such as Support Vector Machines (SVM), Random Forest (RF), and Gradient Boosting Models (GBM) have been widely used for groundwater classification and prediction of Water Quality Index (WQI).

The study "Predicting Groundwater Quality Using AI-Based WQI Models" presents an approach that utilizes machine learning models to classify groundwater quality. The research highlights the significance of water quality for human consumption and investigates factors influencing groundwater contamination. Key parameters such as pH, temperature, and TDS are considered for modeling. The study implements an AI-driven WQI framework that categorizes water into excellent, good, fair, and

poor quality. The methodology involves data collection, preprocessing, and training of multiple AI models to enhance prediction accuracy. The research demonstrates the effectiveness of AI in evaluating groundwater safety, benefiting policymakers and environmental agencies. [1]

The study "Clustering-Based Water Quality Analysis in an Agricultural Reservoir" explores the use of clustering techniques to analyze water quality in an agricultural reservoir. The research focuses on IoT-based sensor data to monitor water quality in real time. Parameters like dissolved oxygen (DO), pH, TDS, and conductivity are measured using sensor-equipped buoys. The study applies the K-Means clustering algorithm to segment water quality patterns and assess contamination levels. The elbow method and silhouette coefficient are used to determine optimal clusters, improving monitoring strategies. The results validate the effectiveness of clustering in water quality assessment, aiding aquaculture and irrigation practices. [2]

Wang et al. (2017) introduce a machine learning-based approach for water treatment optimization. The study employs a Support Vector Machine (SVM) model to predict chemical dosing levels in water treatment plants. Parameters such as turbidity, pH, and organic matter concentration are analyzed. The research incorporates Particle Swarm Optimization (PSO) to fine-tune SVM parameters, enhancing real-time assessment capabilities. The study proposes a feedforward-feedback control mechanism to improve treatment efficiency. The findings highlight the potential of AI-driven models in ensuring safe and stable drinking water quality while addressing challenges related to treatment process variability. [3]

Khatri et al. (2020) investigate the transition from MATLAB’s fuzzy logic-based water quality evaluation to Python-based real-time assessments. MATLAB’s fuzzy inference systems are widely used for WQI computation, but their offline nature limits real-time applications. The study implements fuzzy logic on a Raspberry Pi system using Python, enabling continuous water quality monitoring. Key parameters such as pH, turbidity, DO, and TDS are analyzed to compute WQI. The study demonstrates that Python-based fuzzy logic is as accurate as MATLAB but offers enhanced scalability and efficiency, making it suitable for real-time water quality assessments. [4]

Nayan et al. (2020) explore the use of machine learning for river water quality prediction in Bangladesh. The research applies a Gradient Boosting Model (GBM) to analyze water quality trends based on historical data from 2013 to 2019. Parameters like pH, DO, and BOD are considered for modeling. The study finds a strong correlation between predicted and observed water quality levels, demonstrating the effectiveness of GBM in forecasting water conditions. The findings emphasize the importance of predictive analytics in water resource management, assisting authorities in mitigating contamination risks. [5]

Prakash et al. (2018) evaluate classification algorithms for groundwater quality assessment in India. The study compares Decision Tree (DT), K-Nearest Neighbors (KNN), and SVM models to classify water samples based on mineral content and conductivity. Data from 51 districts are analyzed, categorizing samples into good, average, and poor quality. The study utilizes confusion matrices and ROC curves to measure classification performance. Results indicate that SVM achieves the highest accuracy (96.6%), outperforming DT and KNN. The study concludes that SVM is a reliable model for groundwater quality classification and resource management. [6]

Hassan et al. (2020) investigate machine learning applications in water quality assessment, focusing on WQI prediction. The study evaluates models such as Random Forest (RF), Neural Networks (NN), Multinomial Logistic Regression (MLR), and Bagged Tree Models (BTM) on datasets from Indian water bodies. Data preprocessing includes missing value imputation using Random Forest and normalization through min-max scaling. The study finds that MLR achieves the highest prediction accuracy (99.83%), demonstrating the potential of AI models in optimizing water quality assessments for public health management. [7]

Sidek et al. (2024) analyze WQI prediction in the Johor River Basin, Malaysia, using ensemble machine learning. The study highlights BOD, COD, and DO% as primary water quality indicators. Gradient Boosting (GB) and Random Forest (RF) models are employed to predict WQI, achieving R^2 values of 0.85 and 0.86, respectively. The research concludes that using only three parameters, WQI can be accurately estimated, offering cost-effective solutions for water quality monitoring. The study provides valuable insights for environmental management agencies aiming to improve water resource sustainability. [8]

William et al. (2023) explore AI-based monitoring in Water Distribution Systems (WDS). The study addresses the limitations of traditional monitoring by simulating bio-contamination risk propagation using AI models. ANN and SVM techniques are integrated into a smart monitoring system to detect water quality anomalies. The research demonstrates that AI-driven systems can improve real-time water quality control by identifying contaminants early. The implementation of ANN enhances the accuracy of predictions for parameters like DO, pH, and organic carbon, emphasizing AI's role in cost-effective and efficient water quality monitoring. [9]

Nair and Vijaya (2022) examine river water quality prediction using machine learning. The study focuses on WQI-based assessments for eleven sampling stations along the Bhavani River in India. Parameters such as pH, BOD, and turbidity are analyzed using ML algorithms. The research applies GBM and RF models, achieving high prediction accuracy. The study underscores the importance of AI in water quality evaluation and provides insights into model selection for future predictive applications. The findings support the adoption of AI in environmental monitoring to ensure sustainable water resource management. [10]

3. Methodology

To accurately predict wqi, we employ a range of pre-processing methods as well as machine learning algorithms. Which can be stated as median immputations, national sanitation foundation water quality index, linear regression, decision tree regression, support vector regression, random forest regression and gradient boosting regressor. The following sections provide a detailed explanation of each approach.

a. Median Imputation:

Median imputation is a statistical technique used to handle missing data by replacing missing values with the median of the observed values in a dataset. It is particularly effective for numerical data with skewed distributions, as the median is less sensitive to outliers compared to the mean. This method preserves the central tendency of the data without being overly influenced by extreme values. Median imputation is commonly used in machine learning and data preprocessing to maintain data consistency while minimizing bias introduced by missing values.

b. National Sanitation Foundation Water Quality Index (NSF-WQI):

NSF-WQI is a standardized method for assessing water quality based on multiple physical, chemical, and biological parameters. Each parameter is assigned a weight and transformed into a sub-index, which are then aggregated to compute a final score. The index categorizes water quality into five classes, ranging from excellent to poor. NSF-WQI simplifies complex water quality data for public understanding and policy decisions. However, it may not capture local environmental factors and specific contaminant effects accurately.

c. Linear Regression (LR):

Linear Regression is a fundamental statistical method used to model relationships between a dependent variable and one or more independent variables. It assumes a linear relationship and estimates coefficients using the least squares method to minimize errors. LR is widely used in predictive modeling and trend analysis due to its simplicity and interpretability. However, it is

sensitive to outliers and performs poorly when the relationship between variables is non-linear or when multicollinearity is present.

d. Decision Tree Regression (DTR):

Decision Tree Regression is a non-linear machine learning algorithm that splits data into hierarchical decision nodes based on feature values. It recursively partitions data to create a tree-like structure, making predictions by following branches to terminal nodes. DTR is easy to interpret, handles both numerical and categorical data, and requires minimal data preprocessing. However, it tends to overfit on small datasets and is sensitive to noisy data, which can reduce its generalization ability.

e. Support Vector Regression (SVR):

Support Vector Regression is a machine learning algorithm that uses Support Vector Machines (SVM) for regression tasks. It maps input data into a high-dimensional space using kernel functions and finds a hyperplane that best fits the data within a defined margin. SVR is effective in handling non-linearity and works well with small datasets. However, it is computationally expensive, sensitive to hyperparameter selection, and may not scale well to large datasets.

f. Random Forest Regression (RFR):

Random Forest is an ensemble learning technique that combines multiple decision trees to improve classification accuracy and reduce overfitting. It operates by creating a "forest" of decision trees, each trained on a random subset of data and features. The final prediction is determined through majority voting (classification) or averaging (regression) across all trees. The key steps in the Random Forest algorithm include:

1. Bootstrapping the dataset to create multiple training subsets.
2. Training decision trees independently on different feature subsets.
3. Aggregating the predictions from all trees for the final output.

Random Forest is robust, handles nonlinear relationships well, and is less prone to overfitting compared to individual decision trees. However, it can be computationally expensive for large datasets.

g. Gradient Boosting Regressor (GBR):

Gradient Boosting Regressor is an advanced ensemble learning technique that builds multiple weak learners (usually decision trees) sequentially, with each tree correcting the errors of the previous ones. It optimizes performance using gradient descent, making it highly accurate for complex datasets. GBR is effective in handling non-linearity and feature interactions but is prone to overfitting if not properly tuned. Additionally, it is computationally intensive, making training time longer than simpler models.

4. Results & Discussion

To assess the performance of different machine learning models for wqi calculation, we evaluated them using metrics like: ROC AUC, Accuracy, Precision, and Recall.

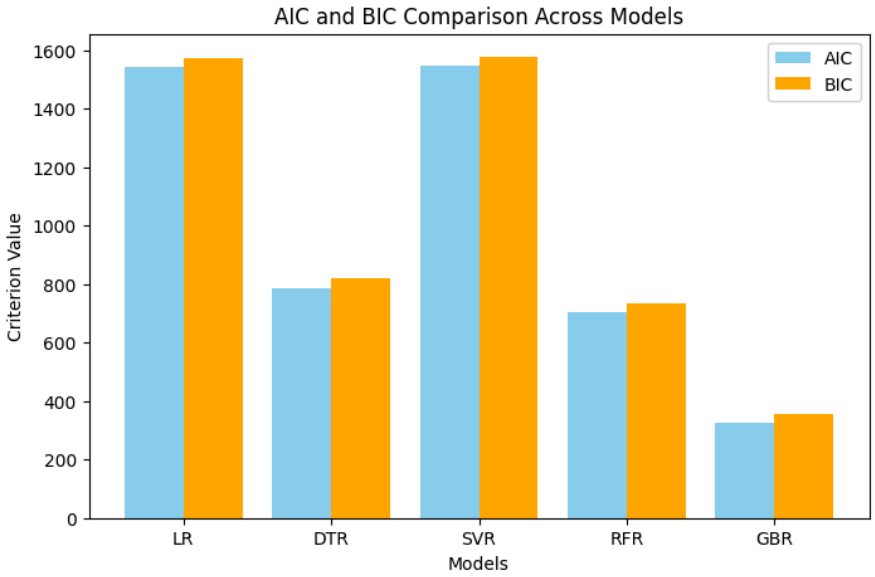
- **MAE (Mean Absolute Error)** measures the average absolute difference between predicted and actual values, indicating overall prediction accuracy.
- **MSE (Mean Squared Error)** calculates the average squared difference between predicted and actual values, penalizing larger errors more than smaller ones.
- **RMSE (Root Mean Squared Error)** represents the square root of MSE, providing a more interpretable measure of error in the same unit as the target variable.
- **R^2 (R-squared)** evaluates how well the model explains variance in the target variable, with values closer to 1 indicating better fit.

	MAE	MSE	RMSE	R^2
Linear Regression	5.12	45.75	6.76	0.74
Decision Tree Regressor	0.82	6.9	2.63	0.96
Support Vector Regressor	4.85	46.21	6.8	0.74
Random Forest Regressor	0.87	5.59	2.36	0.97
Gradient Boosting Regressor	0.75	2.16	1.47	0.99

From the results:

- Gradient Boosting Regressor achieved the lowest error values (MAE: 0.75, MSE: 2.16, RMSE: 1.47) and the highest R^2 score (0.99), indicating the best overall performance.
- Random Forest Regressor also performed well with low MAE (0.87) and RMSE (2.36), and a strong R^2 score (0.97), making it a reliable choice.
- Decision Tree Regressor showed good accuracy (R^2: 0.96) but had slightly higher error values than Random Forest.
- Linear Regression and Support Vector Regressor had higher error values (MAE ~5, RMSE ~6.8) and lower R^2 (0.74), indicating weaker predictive performance.

AIC & BIC Graph:



The AIC and BIC comparison shows that Gradient Boosting Regressor (GBR) has the lowest values, indicating the best model fit with minimal complexity. Support Vector Regressor (SVR) and Linear Regression (LR) have the highest AIC and BIC values, suggesting a poorer fit and higher model complexity. Decision Tree Regressor (DTR) and Random Forest Regressor (RFR) fall in between, with moderate AIC and BIC values. Based on these findings, GBR emerges as the most efficient model due to its balance of predictive performance and simplicity.

5. Conclusions

Our study explores toxic comment classification using machine learning algorithms to enhance online content moderation. Among the evaluated models, Random Forest and SVM demonstrated superior

performance, achieving high accuracy, recall, and precision in detecting toxic content. The study underscores the effectiveness of text preprocessing techniques like Lemmatization and TF-IDF in improving classification accuracy. Despite these advancements, challenges such as sarcasm detection, dataset bias, and contextual interpretation persist, necessitating further research into deep learning and transformer-based models. Future work will focus on enhancing model interpretability and reducing false positives to develop robust, real-time moderation systems.

6. Reference

[1] Alahakoon, T., and Jayarathna, T., "Water Quality Index Based Prediction of Ground Water Properties for Safe Consumption," in 2020 2nd International Conference on Advancements in Computing (ICAC), IEEE, 2020. DOI: 10.1109/ICAC51239.2020.9357146

[2] Ardarsa, P., & Surinta, O., "Water Quality Assessment in the Lam Pa Thao Dam, Chaiyaphum, Thailand with K-Means Clustering Algorithm," in 2021 Research, Invention, and Innovation Congress (RI2C), IEEE, 2021. DOI: 10.1109/RI2C51727.2021.9559811

[3] W. Dongsheng, L. Yongjie and Z. Lei, "Raw water quality assessment oriented to the drinking water treatment based on SVM model," 2017 29th Chinese Control And Decision Conference (CCDC), Chongqing, China, 2017, pp. 6236-6241, DOI: 10.1109/CCDC.2017.7978293.

[4] Khatri, P., Gupta, K. K., & Gupta, R. K. (2020). Water Quality Index Calculation: Switching from MATLAB Fuzzy Toolbox to Python for Real-Time Implementation. 2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI). DOI: 10.1109/ICATMRI51801.2020.9398318

[5] Nayan, A. A., Kibria, M. G., Rahman, M. O., & Saha, J. (2020). River Water Quality Analysis and Prediction Using GBM. 2nd ICAICT 2020, IEEE. DOI: 10.1109/ICAICT51780.2020.9333492

[6] Prakash, R., Tharun, V. P., & Renuga Devi, S. (2018). A Comparative Study of Various Classification Techniques to Determine Water Quality. 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018), IEEE. DOI: 10.1109/ICICCT.2018.8473168

[7] Hassan, M. M., Hassan, M. M., Akter, L., Rahman, M. M., Zaman, S., Hasib, K. M., Jahan, N., Smrity, R. N., Farhana, J., Raihan, M., & Mollick, S. (2021). Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms. Human-Centric Intelligent Systems, 1(3–4), 86–97. DOI: 10.2991/hcis.k.211203.001

[8] Sidek, L.M., Mohiyaden, H.A., Marufuzzaman, M. et al. Developing an ensembled machine learning model for predicting water quality index in Johor River Basin. Environ Sci Eur 36, 67 (2024). DOI: 10.1186/s12302-024-00897-7

[9] P. William, O. J. Oyebode, G. Ramu, M. Gupta, D. Bordoloi and A. Shrivastava, "Artificial Intelligence based Models to Support Water Quality Prediction using Machine Learning Approach," 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 2023, pp. 1496-1501, doi: 10.1109/ICCPCT58313.2023.10245020

[10] Jitha P. Nair and M. S. Vijaya, "River Water Quality Prediction and Index Classification Using Machine Learning," Journal of Physics: Conference Series, vol. 2325, no. 1, pp. 012011, Aug. 2022, doi: 10.1088/1742-6596/2325/1/012011