EWE

Environment and Water Engineering

Homepage: www.jewe.ir

ISSN: 2476-3683

Predictive modeling of water quality index (WQI) using regression techniques: a comparative analysis

N. S. Prema [1], B. M. Shashikala [2]✉, M.Veena [3], and K. G. Chaithra [4]

[1]Department of Information Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, India
[2]Department of Master of Computer Application, SJCE, JSS Science and Technology University, Mysuru, India
[3]Department of Computer Science and Engineering, PES College of Engineering, Mandya, India
[4]Department of Information Science and Engineering, Maharaja Institute of Technology, Mysuru, India

ARTICLE INFO

Paper Type: Research Paper

Received: 14 February 2025
Revised: 01 April 2025
Accepted: 11 June 2025
Published: 01 September 2025

Keywords
Decision Tree
Gradient Boosting
Linear Regression
Random Forest
Regression Analysis
WQI

Corresponding author:
B M. Shashikala
✉ shashibmk@sjce.ac.in

ABSTRACT

Water quality is a critical determinant of ecological and public health, making its regular assessment essential for sustainable development. This study aims to estimate the Water Quality Index (WQI) using multiple water parameters—pH, temperature, dissolved oxygen (DO), conductivity, faecal coliform, and nitrate-nitrite nitrogen. The dataset, sourced from Kaggle, comprises water samples collected across 18 Indian states. A weighted arithmetic WQI approach is employed to compute the index values. To forecast WQI, four regression models, linear regression, decision tree, random forest, and gradient boosting, are applied. Model performance is evaluated using the coefficient of determination (R^2). Among all models, gradient boosting achieved the highest prediction accuracy, with an R^2 value of 0.94, significantly outperforming the others. The results highlight the effectiveness of machine learning in modelling complex environmental parameters and forecasting water quality. This study demonstrates that data-driven approaches can support timely decision-making for water resource management and public health interventions.

Highlights

- Gradient Boosting achieved the highest WQI accuracy with the lowest errors.
- BOD, DO, and conductivity were the most influential WQI predictors.
- Removing missing data outperformed mean and random value imputation.
- The dataset covered 18 Indian states with seven key water quality parameters.

Citing:
Prema, N S., Shashikala, B. M., Veena, M., & Gaithra, K. G. (2025). Predictive modeling of water quality index (WQI) using regression techniques: a comparative analysis. *Environment and Water Engineering*, *11*(3), 295-304. https://doi.org/10.22034/ewe.2025.506453.2003

1. Introduction

The reduction in water quality has a major effect on the supply of pure fresh water for the people's use, agriculture, and aquatic life ecosystems. According to current estimates, 8442 km^3 of water are needed annually to support human residential, agricultural, and industrial activities, making rivers a significant supply of fresh water (Wu et al., 2022). The Indian government has launched several initiatives to enhance the ecological status and clean up the country's river systems. With rising concentrations of physicochemical characteristics, trace, and heavy elements, the Ganga is one of India's most polluted rivers (Richards et al., 2023; Sharma et al., 2022).

The rapid economic expansion of developing countries has the potential to negatively affect the environment, and every development has an adverse effect on the environment life. Pollutants carried by rivers harm ecosystems and threaten human health as they flow into lakes and oceans. Thus, effective water management for sustainability and the preservation of environmental and human health depends on the water quality assessment and monitoring. WQI is the numerical measure used to check the quality of drinking water. $Ca^{2+,}$ Mg^{2+}, NO_3^-, and other factors, such as temperature, pH, dissolved oxygen, and total suspended particles, are commonly used to predict the WQIs (Rufino et al., 2019).

The WQI is very helpful in directing decision-makers' choices and activities. However, because sub-indices are calculated within WQI equations, the calculation of WQI is complex. The computation of WQI has the disadvantages of being time-consuming, tedious, difficult, and inconsistent because WQIs usually include separate equations. By collecting field data, statistical models seek to infer general principles from experimental findings. Analytical approaches for analysing and validating data and hypotheses must be carefully chosen as part of the statistical modelling and evaluation process. These characteristics and the variables used to anticipate the quality of water have a complicated, non-linear connection; therefore, applying statistical techniques frequently yields low accuracy (Quinn et al., 2021).

The application of machine learning (ML) algorithms for the prediction of the quality of the water will yield faster and accurate results. ML algorithms analyse the patterns and identify relationships among the data; based on that, the quality of water is predicted (Mohammadpour et al., 20150). In surface water quality studies, ML has become a hot topic (Sharma et al., 2021; Tung and Yaseen, 2020). The quality of surface water can be assessed and predicted using many different methods.

An essential initial phase in developing machine learning models is gathering data. Water systems management can use ad hoc and integrated water quality monitoring results as benchmarks. Conventional environmental monitoring methods are frequently employed by conservation authorities. However, practical difficulties border the use of traditional methods for in-site monitoring (Li et al., 2020). Deep learning and conventional machine learning methods have been extensively applied to water quality analysis. The findings indicate that traditional deep learning models outperform standard machine learning approaches in this domain. Specifically, traditional deep learning models achieved an accuracy that was 1.8% higher when analyzing two-layer water data and 1% higher for multi-layer water data. These results underscore the effectiveness and reliability of deep learning techniques in water quality assessment and emphasize the importance of selecting the most suitable method based on the specific requirements and complexity of the data involved (Prasad et al., 2022). Machine learning classification algorithms were included in the study (Uddin et al., 2023) to identify the best classification method for the prediction of the water quality categories. The results exhibited that the k-nearest neighbour (KNN) and XGBoost algorithms performed well, with 100% and 99.9% classification accuracy, respectively. The study (Karangoda and Nanayakkara, 2023) used data on 10 quality of water characteristics gathered from 50 different groundwater sources to analyse the quality of groundwater using a variation of statistical and graphical methodologies.

The present work's goal was to use soft computing techniques to forecast a supply system's water quality. The work's objective was to suggest a trustworthy technique for precisely forecasting the quality of water with machine learning technology.

2. Materials and Methods

2.1 Dataset

The dataset, which was obtained via Kaggle (Kaggle, 2020), offers comprehensive details on the water quality metrics of different Indian rivers. It consists of about 534 instances and seven essential attributes, each of which represents the average values recorded over a given time. Temperature, pH, conductivity, DO, BOD, nitrate-nitrite nitrogen (NO_3-NO_2), faecal coliform, are among the parameters. The water quality data is from diverse locations in 18 states of India; the sample count is shown in Fig. 1. The statistical details of the data set are tabulated in Table 1. The dataset shows that conductivity and faecal Coliform have substantial variability, with extreme ranges and standard deviations. In total, the data reveals significant differences in water quality around metrics in addition to constant estimates.

Fig. 1 The frequency of water samples collected from different states

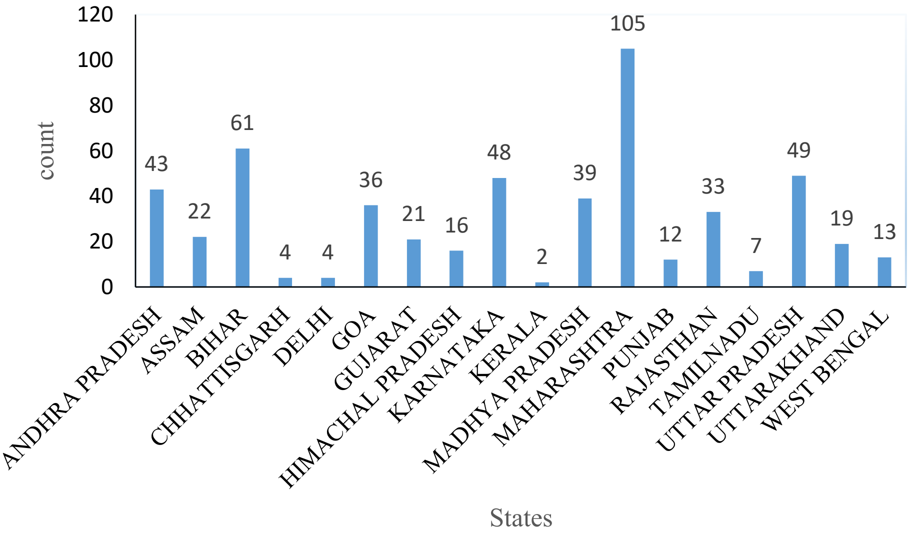


Table 1 The statistical details of the water quality attributes

Attribute/ Parameter	Temperature	DO	pH	Conductivity	BOD	Nitrate N	Faecal Coliform
Mean	25.5	6.44	7.81	626	4.64	1.41	6720
Median	25.7	6.6	7.9	366	2.7	0.6	386
Mode	27	7.7	7.9	290	1.5	0	4
Standard deviation	3.31	1.5	0.69	1772	6.39	2.92	28567
Minimum	10.5	0	6.3	39	0.2	0	0
Maximum	33.8	16.3	14.7	24062	75.6	45.5	310417
Skewness	-0.826	-0.221	4.42	11.1	5.17	8.81	7.16
Kurtosis	2.1	5.28	40.9	135	40.9	119	58.9

2.1.1 Data pre-processing

The data was downloaded from Kaggle (Kaggle, 2020), and based on the Weighted Arithmetic Water Quality Index Method, the WQI was calculated. The following steps were taken to calculate WQI using the weighted arithmetic index method:

i. Assume that there are "n" distinct WQ parameters. The n^th^ parameter's quality rating (Q_n)is a numerical value that represents the degree to which the parameter has deviated from its typical allowed value in the contaminated water. Eq. 1 can be used to get values for Q_n.

$$Q_n = 100 \frac{(V_n - V_i)}{(V_s - V_i)} \qquad (1)$$

where V_i stands for the ideal value, V_s for the standard value, and V_n for the observed value.

ii. In most cases, $V_i$equals 0 except for important parameters like pH and dissolved oxygen. Eqs. 2 and 3 can be used to determine the pH and DO quality rating ($V_i \neq 0$).

$$Q_{pH} = 100 \frac{(V_{pH} - 7.0)}{(8.5 - 1.0)} \qquad (2)$$

$$Q_{DO} = 100 \frac{(V_{DO} - 14.6)}{(5.0 - 14.6)} \qquad (3)$$

iii. Calculating unit weight: Eq. 4 shows that the unit weight $w_n$for each of the many water quality measurements is inversely connected with the recommended needs for each of those parameters.

$$w_n = \frac{k}{S_n} \qquad (4)$$

w_n = unit weight for n^th^ parameter

S_n = standard acceptable value for the nth parameter

k = proportionality constant.

WQI can be derived from Eq. 5 (Reference?).

$$\mathrm{WQI} = \sum_{n=1}^{n} q_n w_n \Big/ \sum_{n=1}^{n} w_n \qquad (5)$$

The following Table 2 lists summaries of whether WQI levels are appropriate for human drinking.

Table 2 Water quality classification (Fathi et al. 2024)

WQI	Quality classification
0 to 25	Excellent
26 to 50	Good
51 to 75	Bad
76 to 100	Very Bad
100 and above	Unfit for usage

2.2 Methodology

The dataset had about 3.1% of missing value instances, and it is handled by using different imputation techniques. This study estimated the WQI using four regression techniques. Forecasts are based on the following regression models, which are separated into 80% training data and 20% data for testing. A 10-fold cross-validation technique is incorporated to check the validity of the models.

2.2.1 Linear Regression

A linear regression model is an easy-to-understand machine learning model for regression tasks. It makes predictions about a target variable by fitting a linear relationship between the independent and dependent variables. It may have trouble handling complex or nonlinear data patterns, but it performs well when the features are linearly related to the target (Gorgan-Mohammadi et al., 2023).

2.2.2 Decision Tree Regressor

The decision tree regressor creates a prototype based on recursive division of data into subsets according to feature values. A decision node is created by splitting the data into leaves that match the expected values. If the data is not tuned correctly, decision trees can be overfit despite their versatility and capability to reproduce nonlinear interactions (Gorgan-Mohammadi et al., 2023).

2.2.3 Random Forest Regressor

Combining many decision trees improves prediction accuracy and robustness using a random forest regressor. It builds many trees and averages the predictions using random data picks. This makes the model more stable than a single decision tree by reducing overfitting and improving generalisation by taking the average out the errors of individual trees (Borup et al., 2023).

2.2.4 Gradient Boosting Regressor

Another powerful ensemble approach is the gradient boosting regressor that constructs trees one after the other intending to modify the errors in the earlier ones. By focusing on residual errors, gradient boosting can achieve good projected accuracy even with complex data. However, because of its high computational cost and sensitivity to hyperparameter modification, its performance may decrease if it is not properly tuned (Demir and Sahin, 2023).

2.2.5 Performance Measures

The metric used to assess regression models is the error rate. If the discrepancy between actual and predicted values for the train, validation, and test data sets is negligible and unbiased, the regression model is good. The following metrics are used to gauge how well regression models perform:

Mean Absolute Error MAE is the computation method employed in Eq.6, is the average of the absolute difference between the actual and predicted values.

$$\mathrm{MAE} = (\Sigma|\mathrm{y} - \bar{\mathrm{y}}|)/\ \mathrm{N} \qquad (6)$$

where $|\mathrm{y} - \bar{\mathrm{y}}|$ represents the absolute magnitude of the difference between the expected and actual values, and y stands for the true value, the predicted value, and the difference. N stands for the quantity of sample points. Root Mean Square Error (RMSE) is like MAE, with the exception that, as indicated in Eq.7, we square the error rather than using the absolute value to eliminate the sign from individual errors. (Because the square of a negative integer is positive.)

$$\mathrm{RMSE} = \sqrt{\frac{\Sigma(\mathrm{y}-\bar{\mathrm{y}})^2}{\mathrm{N}}} \qquad (7)$$

R-squared (R^2) is indicated in Eq.8, it determines how well the regression line fits the data in comparison to the mean line. It represents the percentage of a dependent variable's variance that a regression model's independent variable or variables can account for.

$$R^2 = \frac{var(mean) - var(line)}{var(mean)} \qquad (8)$$

The hyperparameters for the regression models are as follows: Gradient Boosting uses 100 trees with a learning rate of 0.1, a maximum tree depth of 3, and stops splitting nodes with a minimum of 2 instances. Linear Regression is applied without regularization while fitting the intercept. Random Forest consists of 10 trees with an unlimited number of considered features and tree depth, without replicable training, and stops splitting nodes with at least 5 instances. Decision Tree applies pruning with a minimum of 2 instances in leaves, at least 5 in internal nodes, a maximum depth of 100, and follows a binary tree structure.

3. Results and Discussion

3.1 Correlation analysis

The correlation analysis of the attributes and obtained WQI was done using Pearson correlation analysis (Panigrahi et al., 2023) as shown in Fig. 2. The correlation matrix sheds light on how different water quality measures relate to one another.

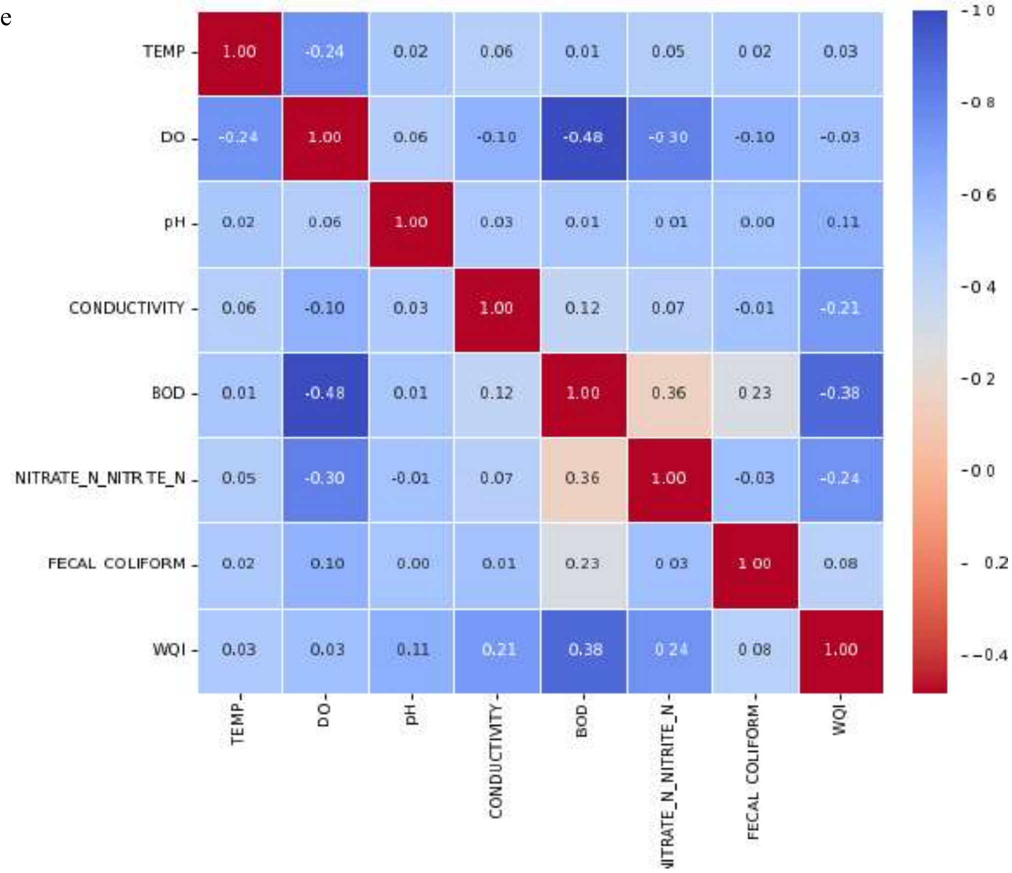


Fig. 2 Correlation between the water quality parameters

The temperature shows little linear interaction with conductivity (0.06) and pH (0.02), both of which have weak positive associations. Higher oxygen levels are linked to less organic pollution, according to the moderately negative connection between dissolved DO and BOD (-0.48). There is also a modest negative association between DO and nitrate and nitrite levels (-0.30). There are very few relationships between pH and other factors. Conductivity may have an impact on the general quality of the water since it has a weakly negative correlation (-0.21) with the WQI but a strongly positive correlation (0.07) with nitrate and nitrite. There is a somewhat positive association (0.36) between BOD and nitrate-nitrite levels, indicating that higher levels of organic pollution are associated with higher amounts of these chemicals. Lastly, there is a moderately negative association between the WQI and both BOD (-0.38) and nitrate/nitrite (-0.24), indicating that worse water quality is linked to higher pollution levels. This matrix facilitates comprehension of the intricate connections among different environmental indicators.

3.2 Performance of the models

The results in Table 3 show that Gradient Boosting consistently outperforms other models, achieving the lowest MSE, RMSE, and MAE, with the highest R^2 across all missing value handling methods. Removing missing instances yields the best performance (MSE = 4.68, R^2 = 0.94) for Gradient Boosting. Linear Regression performs the worst, with significantly high MSE (>60) and low R^2 (~0.2) in all cases. Random Forest and Decision Tree show moderate performance, with Random Forest consistently better than Decision Tree. Replacing missing values with the mean provides similar results to using raw missing data, while random value replacement degrades performance.

Table 3 Performance of the model with 10-fold cross-validation for different types of missing value imputed data

Type of data	Model	MSE	RMSE	MAE	R2
With Missing values	Gradient Boosting	10.91	3.30	2.33	0.90
	Linear Regression	81.20	9.01	6.90	0.22
	Random Forest	22.59	4.75	3.34	0.78
	Decision Tree	48.96	7.00	5.03	0.53
Removing missing value instances	Gradient Boosting	4.68	2.16	1.62	0.94
	Linear Regression	62.29	7.89	6.10	0.25
	Random Forest	11.32	3.36	2.52	0.86
	Decision Tree	21.91	4.68	3.48	0.73
Replace with the average value	Gradient Boosting	10.88	3.30	2.34	0.90
	Linear Regression	81.22	9.01	6.90	0.22
	Random Forest	22.62	4.76	3.40	0.78
	Decision Tree	41.72	6.46	4.71	0.60
Replace with a Random value	Gradient Boosting	19.32	4.40	2.98	0.81
	Linear Regression	81.77	9.04	6.92	0.21
	Random Forest	27.13	5.21	3.67	0.74
	Decision Tree	45.90	6.78	4.95	0.56

The performance of the Gradient Boosting Regressor regression models is also tested by dividing the data into 80% training and 20% testing. The results are tabulated in Table 4. With the lowest MAE (1.49), RMSE (1.92), and greatest R^2 (0.94) of the models examined, the Gradient Boosting Regressor performs the best, indicating that it makes the most accurate and trustworthy predictions.

Table 4 The performance of the regression models for the testing data

Model	MAE	MSE	RMSE	R^2
Linear Regression	5.54	46.47	6.82	0.23
Decision Tree Regressor	3.48	22.90	4.79	0.62
Random Forest Regressor	1.95	6.13	2.48	0.90
Gradient Boosting Regressor	1.49	3.68	1.92	0.94

Interpretability Analysis of Gradient Boosting Using LIME and SHAP is demonstrated in Figs. 3 and 4. The graph shows a LIME (Local Interpretable Model-Agnostic Explanations) explanation for a single prediction, highlighting how each feature influences the model's output. Green bars (like DO and BOD) indicate a positive impact on the prediction, while red bars (like TEMP and FECAL_COLIFORM) show a negative impact. Longer bars reflect stronger influence, helping to understand which feature values contributed most to the decision. The SHAP summary plot shows how each feature impacts the model's predictions across many samples. Each dot represents one prediction, with color indicating the feature value (red = high, blue = low). Features like BOD, DO, and CONDUCTIVITY have a strong influence, as indicated by the wide spread of SHAP values. High BOD and DO values push predictions higher, while low values tend to reduce them. This plot helps visualize both the importance and the effect direction of features in the model.

Fig. 3 Interpretability analysis of gradient boosting using LIME

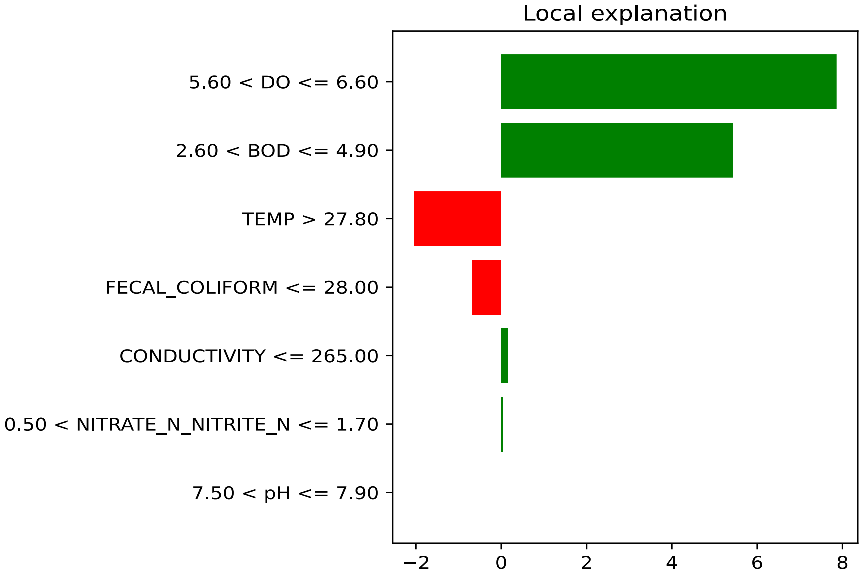
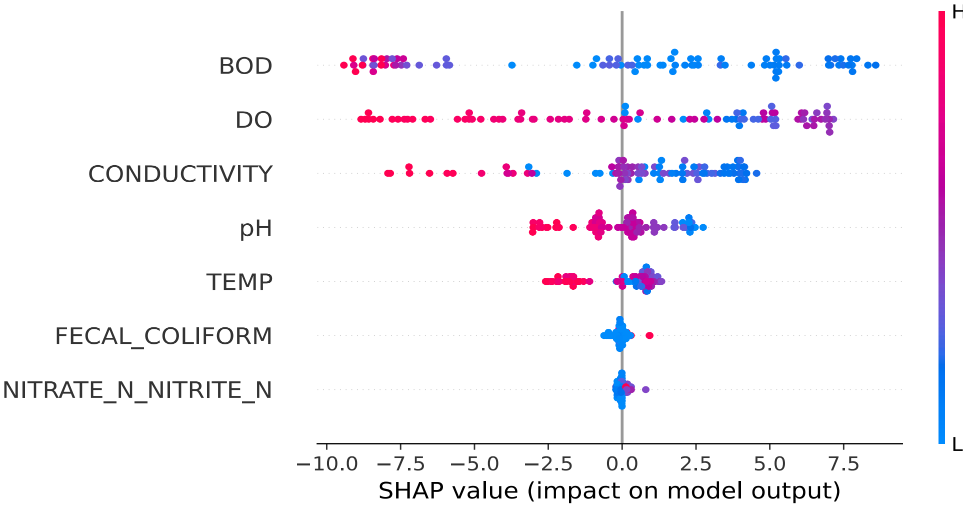


Fig. 4 interpretability analysis of gradient boosting using SHAP



The comparison of experimental and actual values was done for the four regression models as shown in Fig. 5. A low R^2 of 0.23 and high MAE and RMSE indicated a poor fit, while linear regression showed the biggest differences between real and projected values. Though there were still noticeable variations in the predictions, the Decision Tree Regressor demonstrated superior accuracy with reduced residuals and a higher R^2 of 0.62. With substantially smaller MAE and RMSE values that were more in line with the actual values and an R^2 of 0.90, the Random Forest Regressor further decreased prediction errors. Lastly, the Gradient Boosting Regressor showed the best alignment, showing the most accurate predictions with an R^2 of 0.94 and the smallest MAE and RMSE. Gradient Boosting excels by sequentially correcting errors, reducing bias and variance, and optimizing hyperparameters, making it highly effective for Water Quality Index prediction.

The learning curve is plotted for the regression model as shown in Fig. 6. Learning curves are used to assess the model's performance, which shows how changes in training and validation errors occur when the model is subjected to more data. They assist in identifying overfitting and underfitting. These curves lead to the outcomes of model corrections, ensuring better generalization (Loureiro et al., 2021).

The study includes Linear Regression alongside other regression models, providing a baseline for comparison. While more advanced machine learning models are explored, Linear Regression serves as a traditional statistical method, helping to assess the added value of complex models over simpler, interpretable approaches.

The deep learning models are less effective compared to conventional methods, considering factors such as limited dataset size, overfitting risks, and higher computational requirements. Studies have shown that for structured tabular data, traditional models like Gradient Boosting often outperform deep learning due to their ability to handle small-to-medium datasets efficiently (Grinsztajn et al., 2022; Shwartz-Ziv and Armon, 2022).

Fig. 5 Experimental vs Predicted values for regression models: a) Linear regression, b) Decision tree, c) Random Forest, and d) Gradient boosting

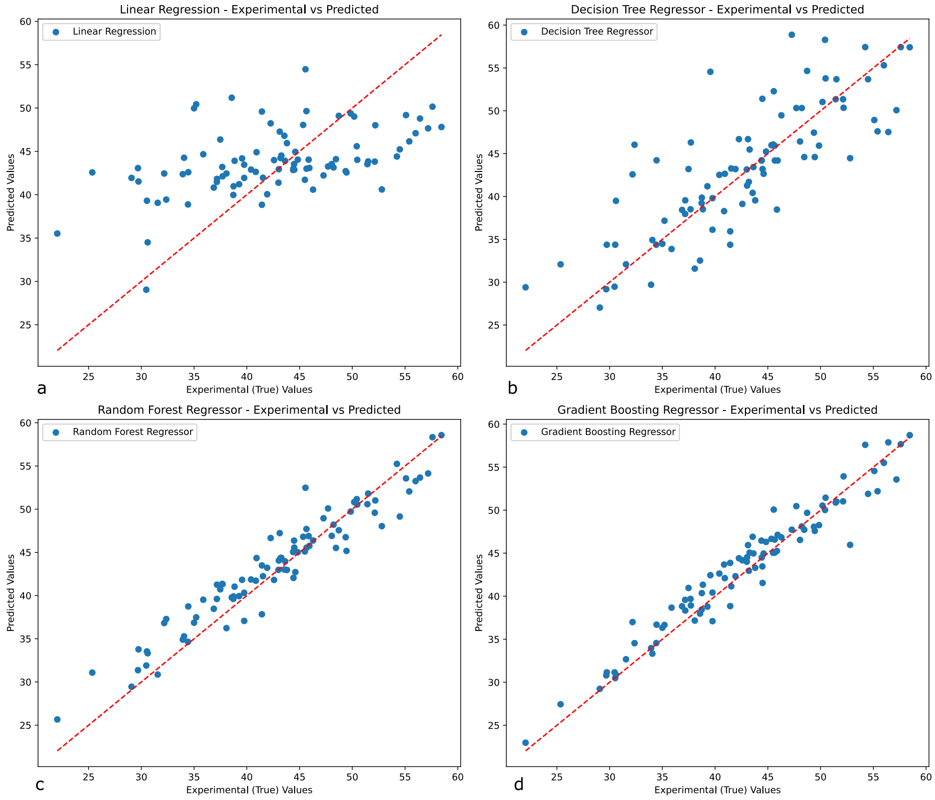
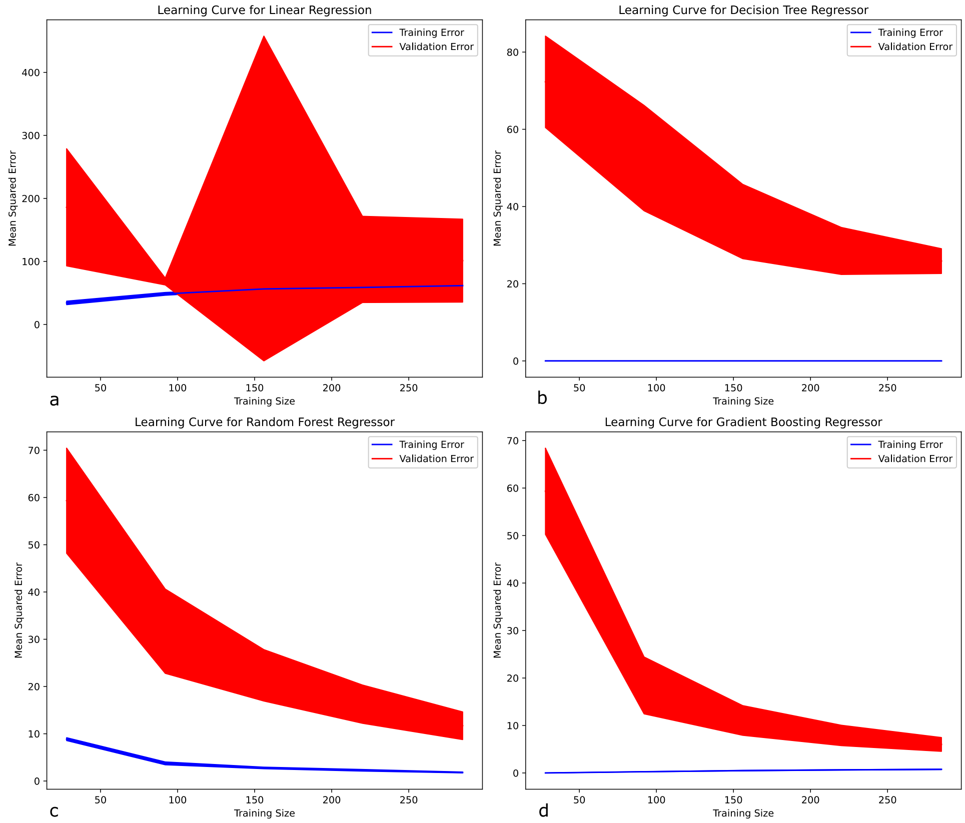


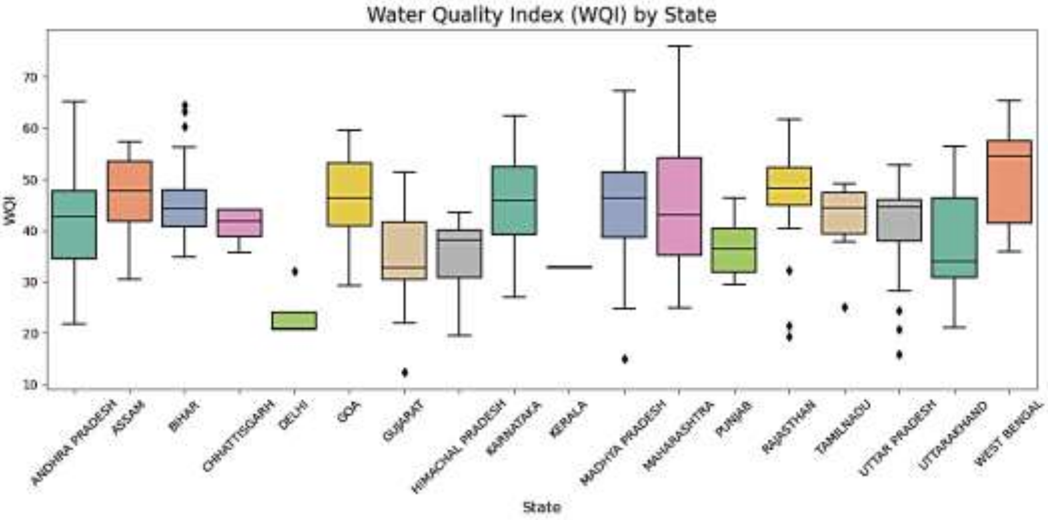
Fig. 6 Learning Curve Plot of the models: a) Linear regression, b) Decision tree, c) Random Forest, and d) Gradient boosting



3.3 WQI of the states

Data on the WQI illustrate that there are considerable differences in water quality amongst Indian states (Nallakaruppan et al., 2024). Fig. 7 is used to perform the WQI analysis of the different Indian states. West Bengal stands out with the highest mean (50.9) and median (54.5), indicating higher water quality, whereas Delhi has the lowest mean (23.7) and median (21), signifying inferior water quality. States like Madhya Pradesh and Maharashtra have large standard deviations (10.9 and 12.3, respectively), suggesting significant fluctuation in water quality within these regions, but Kerala exhibits little variability (standard deviation of 0.156), indicating stable water quality. The range between the minimum and maximum values also demonstrates disparities, with Kerala having the lowest value (33), and Maharashtra having the highest maximum value (76.1). The data also reveals that West Bengal exhibits a right-skewed distribution, indicating some high-quality outliers, while Delhi has a left-skewed distribution (mean lower than median), showing a concentration of worse water quality. In summary, the WQI data reveal regional variations, with some states dealing with more persistent and superior water quality and others dealing with more severe variations.

Fig. 7 The WQI of different states



4. Conclusion

This study applied machine learning techniques to predict Water Quality Index (WQI) using a dataset of key water quality parameters across various Indian states. The main findings can be listed as:

1. To sum up, the WQI regression analysis using seven attributes, temperature, pH, conductivity, DO, BOD, nitrate-nitrite nitrogen, and faecal coliform has demonstrated variable prediction accuracy among models.

2. With the lowest MAE and RMSE as well as the highest R^2, the Gradient Boosting Regressor produced the best results, demonstrating its ability to capture the intricate correlations between the characteristics and WQI.

3. The WQI signifies disparities in state-wise water quality, with West Bengal showing notable water quality, while Delhi exhibits reduced conditions, and Madhya Pradesh and Maharashtra show significant variations. The results also depend on the number of data samples available, with Delhi, Kerala, and Chhattisgarh having very few samples, which may impact the accuracy and reliability of predictions.

This study is limited by the uneven distribution and small number of samples in some regions, which may affect prediction accuracy. Additionally, only seven parameters were considered, excluding other important factors like turbidity and total dissolved solids. Seasonal and regional variations were also not fully captured, suggesting the need for more comprehensive data in future research. Future work can focus on incorporating additional water quality parameters, such as turbidity and dissolved solids, to improve predictive accuracy. Integrating spatiotemporal analysis could help capture regional and seasonal variations, while expanding the dataset across more locations would enhance model generalizability.

Statements and Declarations

Data availability

The datasets used for the study are publicly available on Kaggle.

Conflicts of interest

The author of this paper declared no conflict of interest regarding the authorship or publication of this paper.

Author contribution

Prema N S: Data preprocessing and WQI estimation and proofreading; Shashikala B M: Regression model implementation; Veena M: Writing and literature survey; Chaithra K G: Reviewing and Editing.

AI Use Declaration

During the preparation of this manuscript, the authors used ChatGPT for language translation. All content has been carefully reviewed and revised by the authors, who take full responsibility for the final version of the manuscript.

References

Borup, D., Christensen B. J., Mühlbach N. S., &. Nielsen M. S (2023). Targeting predictors in random forest regression.

International Journal of Forecasting, 39(2), 841-868. DOI: https://doi.org/10.1016/j.ijforecast.2022.02.010

Demir, S., & Sahin E. K. (2023). An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost. *Neural Computing and Applications*, 35(4), 3173-3190. DOI: https://link.springer.com/article/10.1007/s00521-022-07856-4

Gorgan-Mohammadi, F., Rajaee, T., & Zounemat-Kermani, M. (2023). Decision tree models in predicting water quality parameters of dissolved oxygen and phosphorus in lake water. *Sustainable Water Resources Management*, *9*(1), 1. DOI: https://link.springer.com/article/10.1007/s40899-022-00776-0

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, *35*, 507-520. DOI: https://dl.acm.org/doi/10.5555/3600270.3600307

Abdullah Ababakr, F., Shakeri, S., Tand, E., & Kazemi, S. (2024). Interpolation Approaches to Groundwater Quality Mapping: Trends and Techniques in Erbil City. *Advances in Civil Engineering and Environmental Science*, *1*(1), 48-62.DOI: https://doi.org/10.22034/acees.2024.475804.1007

Kaggle (2020), Indian water quality data, edited. Retrieved [December 21, 2024, Available at: https://www.kaggle.com/datasets

Karangoda, R., & Nanayakkara, K. (2023), Use of the water quality index and multivariate analysis to assess groundwater quality for drinking purpose in Ratnapura district, Sri Lanka. *Groundwater for Sustainable Development*, 21, 100910. DOI: https://doi.org/10.1016/j.gsd.2023.100910

Li, W., Fang H., Qin G., Tan X., Huang Z., Zeng F., Du H., & Li, S. (2020). Concentration estimation of dissolved oxygen in Pearl River Basin using input variable selection and machine learning techniques. *Science of Total Environment*, 731, 139099. DOI: https://doi.org/10.1016/j.scitotenv.2020.139099

Loureiro, B., Gerbelot C., Cui, H., Goldt S., Krzakala F., Mezard, M., & Zdeborová, L. (2021). Learning curves of generic features maps for realistic datasets with a teacher-student model. *Adv. Neural Inform. Process. Systems*, 34, 18137-18151. DOI: 10.1088/1742-5468/ac9825

Mohammadpour R., Shaharuddin, S., Chang, C. K, Zakaria, N. A., Ghani, A. A., & Chan, N. W. (2015). Prediction of water quality index in constructed wetlands using support vector machine. *Environmental Science and Pollution Research*, 22, 6208-6219. DOI: https://link.springer.com/article/10.1007/s11356-014-3806-7

Nallakaruppan, M., Gangadevi, E., Shri, M. L., Balusamy, B., Bhattacharya, S., & Selvarajan, S. (2024). Reliable water quality prediction and parametric analysis using explainable AI models. *Scientific Reports*, 14(1), 7520. DOI: https://www.nature.com/articles/s41598-024-56775-y

Panigrahi, N., Patro S. G. K, Kumar R., Omar M., Ngan T. T., Giang N. L., Thu B. T., & Thang N. T. (2023). Groundwater quality analysis and drinkability prediction using artificial intelligence. *Earth Science Informatics*, *16*(2), 1701-1725. DOI: 10.1007/s12145-023-00977-x

Prasad, D. V. V., Venkataramana L. Y., Kumar P. S., Prasannamedha G., Harshana S., Srividya S. J, Harrinei K., & Indraganti S. (2022). Analysis and prediction of water quality using deep learning and auto deep learning techniques. *Science of the Total Environment*, 821, 153311. DOI: https://doi.org/10.1016/j.scitotenv.2022.153311

Quinn, N. W., Tansey, M. K., & Lu, J. (2021). Comparison of deterministic and statistical models for water quality compliance forecasting in the San Joaquin River basin. *Cal. Water*, 13(19), 2661. DOI: https://doi.org/10.3390/w13192661

Richards, L. A., Guo, S., Lapworth, D. J., White, D., Civil, W., Wilson, G. J., Lu, C., Kumar, A., Ghosh, A., & Khamis, K. (2023). Emerging organic contaminants in the River Ganga and key tributaries in the middle Gangetic Plain, India: Characterization, distribution & controls. *Environmental Pollution,* 327, 121626. DOI: https://doi.org/10.1016/j.envpol.2023.121626

Rufino, F., Busico G., Cuoco E., Darrah T. H., & Tedesco D. (2019). Evaluating the suitability of urban groundwater resources for drinking water and irrigation purposes: an integrated approach in the Agro-Aversano area of Southern Italy. *Environmental Monitoring and Assessment*, *191*, 1-17. DOI: https://link.springer.com/article/10.1007/s10661-019-7978-y

Sharma, N., Sharma, R., & Jindal, N. (2021). Machine learning and deep learning applications vision. *Global Transitions Proceedings*, *2*(1), 24-28. DOI: https://doi.org/10.1016/j.gltp.2021.01.004

Sharma, R., Kumar, V., Sharma, D. K., Sarkar, M., Mishra, B. K., Puri, V., Priyadarshini, I., Thong, P. H., Ngo, P. T. T., & Nhu, V. H. (2022). Water pollution examination through quality analysis of different rivers: a case study in India. Environment, *Development and Sustainability*, *Dordrecht*, 24(6), 7471-7492. DOI: https://doi.org/10.1007/s10668-021-01777-3

Shwartz-Ziv R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84-90. DOI: https://doi.org/10.1016/j.inffus.2021.11.011

Tung, T. M., & Yaseen, Z. M. (2020). A survey on river water quality modelling using artificial intelligence models: 2000–2020. *Journal of Hydrology*, *585*, 124670. DOI: https://doi.org/10.1016/j.jhydrol.2020.124670

Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023). Performance analysis of the water quality index model for predicting water state using machine learning techniques.

Process Safety and Environmental Protection, *169*, 808-828. https://doi.org/10.1016/j.psep.2022.11.073

Wu, B., Tian, F., Zhang, M., Piao, S., Zeng, H., Zhu, W., Liu, J., Elnashar, A., & Lu, Y. (2022). Quantifying global agricultural water appropriation with data derived from earth observations. *Journal of Cleaner Production*, *358*, 131891. https://doi.org/10.1016/j.jclepro.2022.131891