SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V12I3P101

Volume 12 Issue 3, 1-9, March 2025

Original Article

Analyzing Factors Affecting the Water Quality of the Ganges and Predicting Further Patterns

Arush Nath

Pathways School Gurgaon.

Corresponding Author : arushnath28@gmail.com

Received: 10 January 2025 Revised: 27 February 2025 Accepted: 13 March 2025 Published: 30 March 2025

Abstract *- Observing the degrading water quality in India and the lack of established forecasting methods for the same, this study aimed to use publicly available data from the CPCB website to test different machine learning algorithms and see which one is most viable for prediction of water quality metrics - specifically D.O. This study carried out comparisons using R^2 Score between numerous machine learning algorithms such as XGBoost, Support Vector Regressor, Gradient Boosting, Random Forest, and Linear Regression. The study concluded that the most accurate prediction model was Random Forest, with an R^2 Score of 0.76, making it a viable means of forecasting future patterns in water quality metrics.*

Keywords *- Dissolved Oxygen (D.O.), Machine Learning, CPCB, Hyperparameter, Forecast.*

1. Introduction

Water is fundamental to life, serving as a critical resource for health, agriculture, and the environment. Water has tremendous cultural and religious significance in India, where rivers such as the Ganga, Brahmaputra, Godavari, and Krishna support huge populations, agriculture, and industries [1]. However, the quality of water in India has deteriorated dramatically, with approximately 70% of surface water classified as unsafe for consumption. The Ganges River, which provides a lifeline for over 600 million people, has become the world's sixth most polluted river due to the inflow of untreated wastewater and industrial effluent. This rampant situation highlights the importance of effective prediction and water quality management systems. Even though air quality indices (such as AQI) and predictive frameworks for air pollution are well developed, similar models for water quality have not been as deeply explored. Studies conducted by Michael Wiryaseputra (2022) explored machine learning approaches for predicting water potability, with XGBoost achieving the highest accuracy at 71.23% [2]. Md. Saikat Islam Khan et al. (2022) established a robust prediction model for water quality based on principal component regression and the Gradient Boosting Classifier, with 100% classification accuracy and 95% prediction accuracy on a Gulshan Lake dataset [3]. While not much investigation has been conducted into machine learning-based forecasting algorithms for water quality, such a prediction model could help test the effectiveness of mitigation strategies and predict future issues with water quality. The current work intends to use machine learning approaches to estimate water quality trends in the Ganges. The study uses techniques such as XGBoost, Support Vector Regressor, Gradient Boosting, Random Forest, and Linear Regression to identify the most accurate model for predicting future water quality patterns, providing a data-driven approach to water pollution concerns.

2. Methodology

2.1. Data Collection

The dataset utilized for this study was mainly focused on the River Ganga, flowing through Jharkhand, Uttar Pradesh, Uttarakhand, Bihar, and West Bengal, spanning the years 2012-2021 [4]. It was collected from the Central Pollution Control Board website. It contained 752 samples, and the data has 7 parameters: DO, Temperature, pH, BOD, Conductivity, Nitrate + Nitrite, Fecal Coliform and Total Coliform.

2.1.1. Data Description

Table 1. Water quality indicators

Variable	Description
Temperature	Temperature, an important water quality parameter that determines the kinds of aquatic life present, influences the quantity of dissolved oxygen, and affects the rate of chemical and biological reactions [5].
Dissolved Oxygen	This measurement, which determines the water's oxygen content—a vital element for aquatic life to survive—proves to be an important indicator of water quality [6].
pH	Since it may be influenced by the chemicals present, this indicator of water acidity can reveal chemical changes in the water. The taste, color, and odor of water can all be impacted by pH [7].

Conductivity	When assessing the electrical conduction capacity of water, notable variations in conductivity may signal a decline in the water body's overall health or condition brought on by outside influences [8].
Biological Oxygen Demand	BOD is a crucial sign of low water quality that might reveal fecal pollution or rises in particulate and dissolved organic carbon in a body of water [9].
Nitrate +Nitrite	Nitrates and nitrites have the potential to indicate the presence of excessive algae growth and sewage and manure pollution in water, making them an excellent metric for measuring water quality [10].
Fecal Coliform	A recent fecal contamination inside the water body can be indicated by the presence of fecal coliform in a water sample often. This is also a reliable diagnostic of a specific type of water pollution [11].
Total Coliform	The existence of coliform bacteria may suggest the presence of disease-causing organisms in the water system, but they will not directly cause illness [11].

2.1.2. Data Preprocessing

Since the data was sourced from the CPCB website, it was raw and inconsistent, so preprocessing was required for training and testing the ML model. The following are the steps for processing the raw data.

Temp Min	Temp Max	D.O. (mg/l) Min	D.O. (mg/l) Max	PH Min	PH Max	CONDUCTIVITY Min	CONDUCTIVITY Max	B.O.D. (mg/l) Min	B.O.D. (mg/l) Max	NITRATENAN N+ NITRITENANN (mg/l) Min	NITRATENAN N+ NITRITENANN (mg/l) Max	FECAL COLIFORM (MPN/100ml) Min	FECAL COLIFORM (MPN/100ml) Max	TOTAL COLIFORM (MPN/100ml) Min	TOTAL COLIFORM (MPN/100ml) Max
0.1	0.1	10.4	10.4	7.4	7.4	263.0	263.0	1.0	1.0	0.32	0.32	2.0	2.0	2.0	2.0
8.0	22.0	9.2	10.8	7.0	8.2	102.0	246.0	1.0	1.0	0.32	0.37	2.0	2.0	2.0	2.0
7.0	21.0	9.6	10.4	7.1	8.2	65.0	463.0	1.0	1.0	0.32	0.62	2.0	2.0	2.0	2.0
12.0	20.0	8.8	10.4	6.9	8.2	77.0	230.0	1.0	1.0	0.32	0.34	2.0	2.0	2.0	2.0
11.0	21.0	9.0	10.6	7.2	8.1	78.0	221.0	1.0	1.0	0.32	0.82	2.0	2.0	2.0	2.0

2.1.3. Data Cleaning

Imputation is the process of replacing missing values with appropriate data—this is important to avoid inaccurate results. The column average for that specific station code over all the years was used to replace null values in the dataset.

B.O.D. (mg/l)	NITRATENAN N+ NITRITENANN (mg/l)	FECAL COLIFORM (MPN/100ml)	TOTAL COLIFORM (MPN/100ml)Mean
NaN	NaN	NaN	NaN
0.7	NaN	5000.0	3867.0
4.6	NaN	1100.0	100000.0
0.9	NaN	NaN	NaN
1.1	NaN	200000.0	200000.0

2.2. Feature Selection and Engineering

By calculating the average of the corresponding "Min" and "Max" columns, add a new "Mean" column for all the attributes to provide a measure of central tendency. A "Next Year" column was also created, shifted by one row to compare with the current year.

2.3. Data Splitting

The dataset was split into two categories - 20% for testing and 80% for training. Training with 80% of the data helps retain sufficient unseen data (20%) for a reliable and unbiased performance evaluation.

This ratio is a common standard in the field, ensuring both meaningful learning and a dependable test of generalization.

2.4. Model Building

2.4.1. Linear Regression

A supervised machine learning approach, Linear Regression works by fitting a linear equation between numerous independent factors along with a dependent variable to calculate the linear connection between the data [12].

2.4.2. Random Forest

The random forest tree training technique in ML functions by creating multiple decision trees built using random subsets of the data, measuring random subsets of features in each partition; the randomness increases variability between the decision trees, reducing the risk of overfitting and hence increasing prediction performance [13].

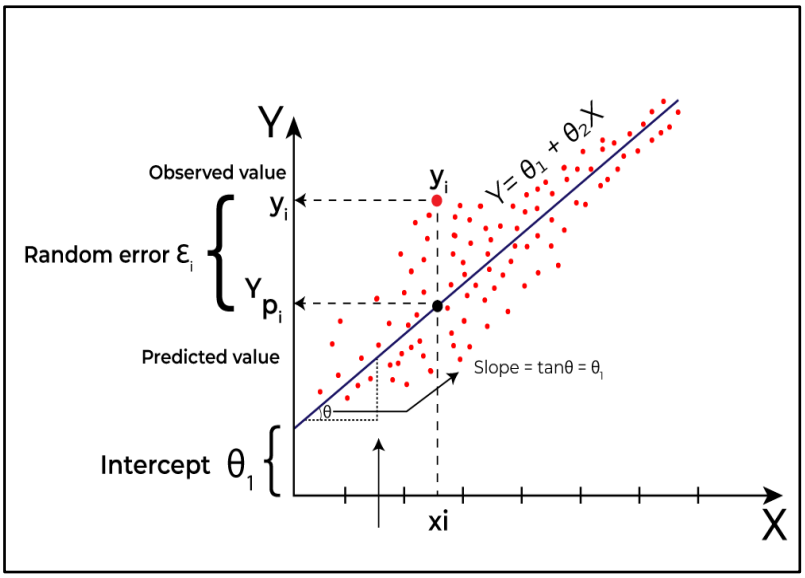

Fig. 1 Linear regression [12]

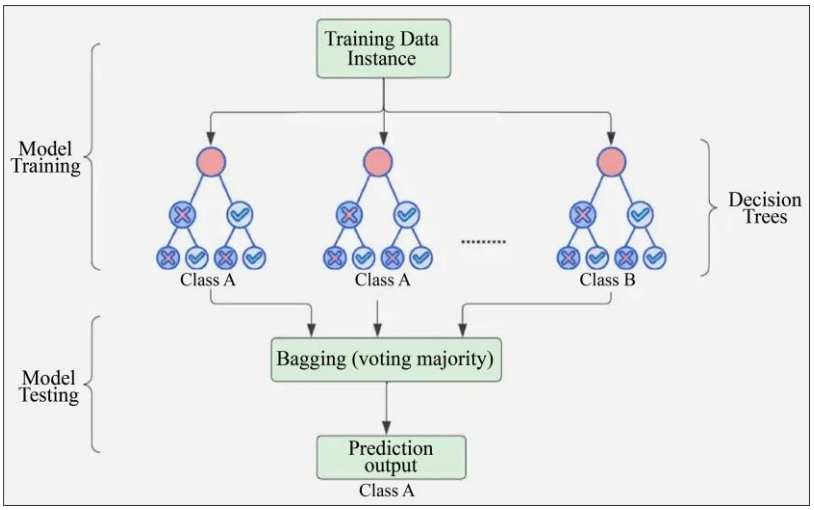

Fig. 2 Random forest [13]

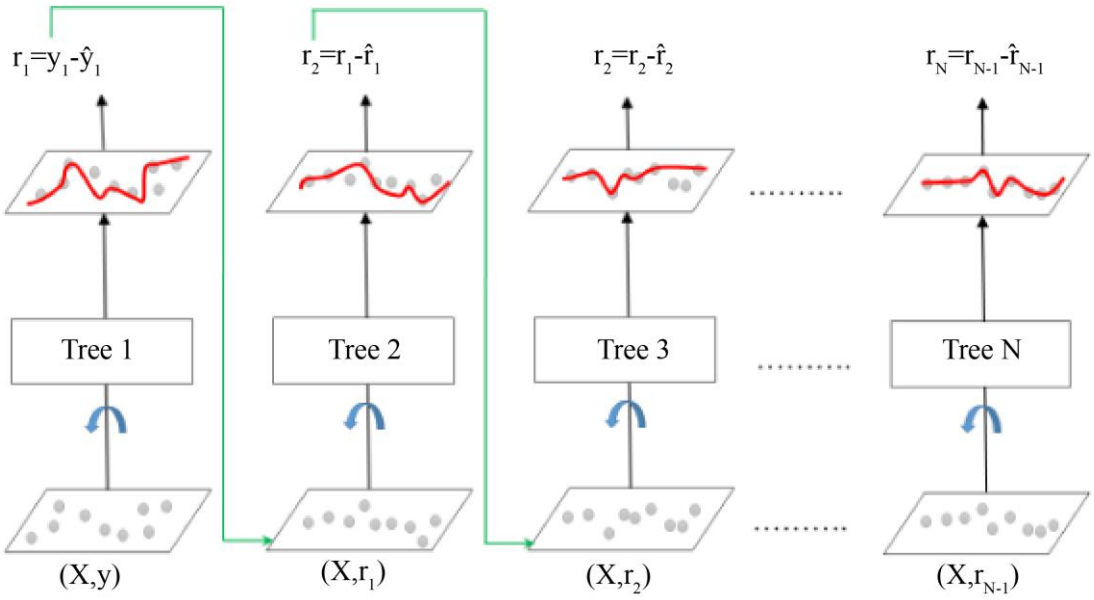

Fig. 3 Gradient boosting [14]

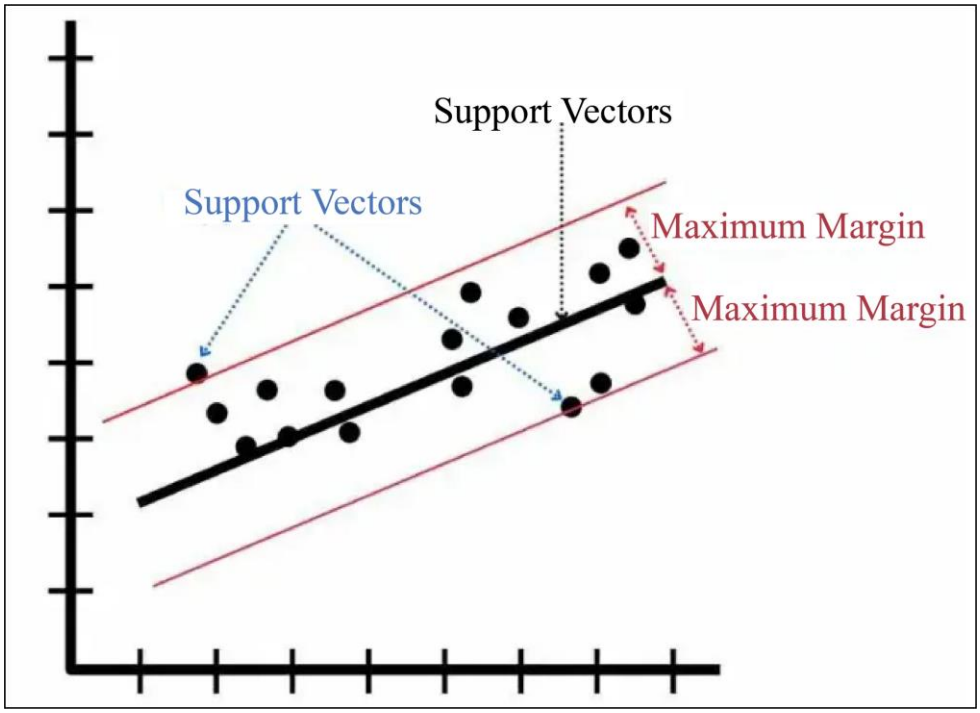

Fig. 4 Support Vector Regressor [15]

2.4.3. Gradient Boosting

A boosting algorithm gradient boosting aims to combine numerous weak learners into strong learners with the most recent model adapting to reduce the loss function from the preceding model [14].

2.4.4. Support Vector Regressor

The SVR (Support Vector Regressor) aims to predict continuous target variables rather than discrete classes by locating the hyperplane that, while maintaining a maximum margin, also best fits the training data [15].

2.4.5. XGBoost

XGBoost (extreme gradient boosting), an optimized distributed gradient boosting technique, is designed for scalable and efficient machine learning model training. XGBoost, an ensemble learning algorithm, combines the predictions of numerous weaker models to create a stronger prediction [16].

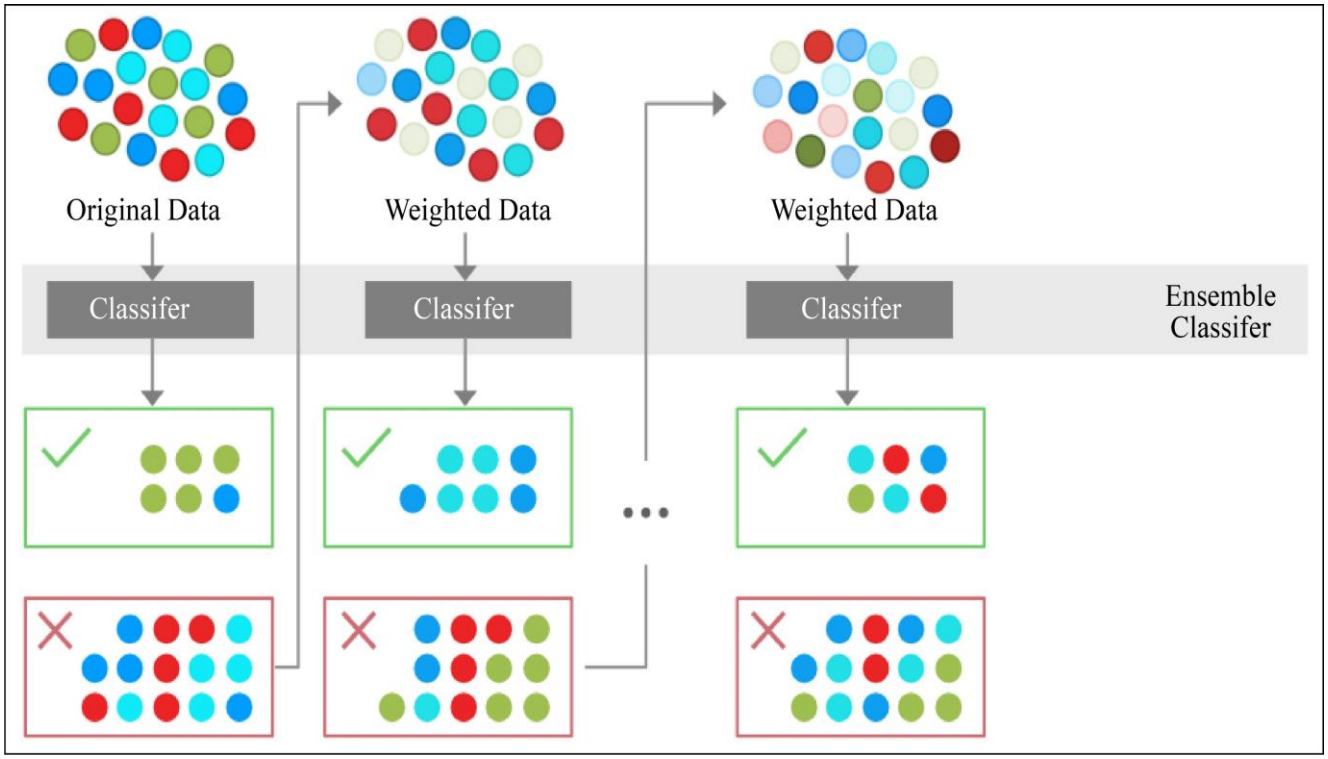

Fig. 5 XGBoost [16]

2.5. R^2 Score

The R^2 Score, also known as the coefficient of determination, indicates how well a machine learning algorithm predicts the relationship between the dependent and independent variables, with values ranging from 0 to 1. A score of 0 indicates that the model explains 0% of the relationship between the dependent and independent variables, whilst a value of 1 indicates that the model explains all of the relationship [17].

The R^2 Score is calculated using the given formula:

$$R^2 = 1 - \frac{Sum\ of\ Squares\ of\ Residuals}{Total\ Sum\ of\ Squares} = 1 - \frac{\sum_{i=1}^{n} \quad (|Y_i - \hat{Y}_i|)^2}{\sum_{i=1}^{n} \quad (|Y_i - \underline{Y_i}|)^2}$$

2.6. Mean Absolute Error

The Mean Absolute Error (or MAE) is calculated by adding the absolute values of the residuals (the difference between actual and projected values) and dividing by the total number of points in the dataset. As a result, the MAE is defined as the absolute average distance between predictions made by a machine learning system (18).

The MAE is calculated using the given formula:

$$MAE = \frac{Sum\ of\ Absolute\ Values\ of\ Residuals}{Total\ Number\ of\ Points\ in\ Dataset} = \frac{1}{n}\sum_{i=1}^{n} \quad |Y_i - \hat{Y}_i|$$

2.7. Best Attribute Selection

The model was tested by using many of the different attributes for prediction to see which would give the highest accuracy in predicting future water quality. From all of these, DO had the best R^2 score and MAE using the Random Forest algorithm, which is why it was chosen as the attribute used for prediction in this current study.

Prediction Attribute	Model Accuracy	
	R^2 Score	MAE
D.O.	0.761792	0.363876
pH	0.549061	0.153220
Conductivity	0.355551	44.202247
Nitrate	-0.361283	0.328380
Fecal Coliform	-6.185962	74578.909300
Total Coliform	-1.724843	151024.642812

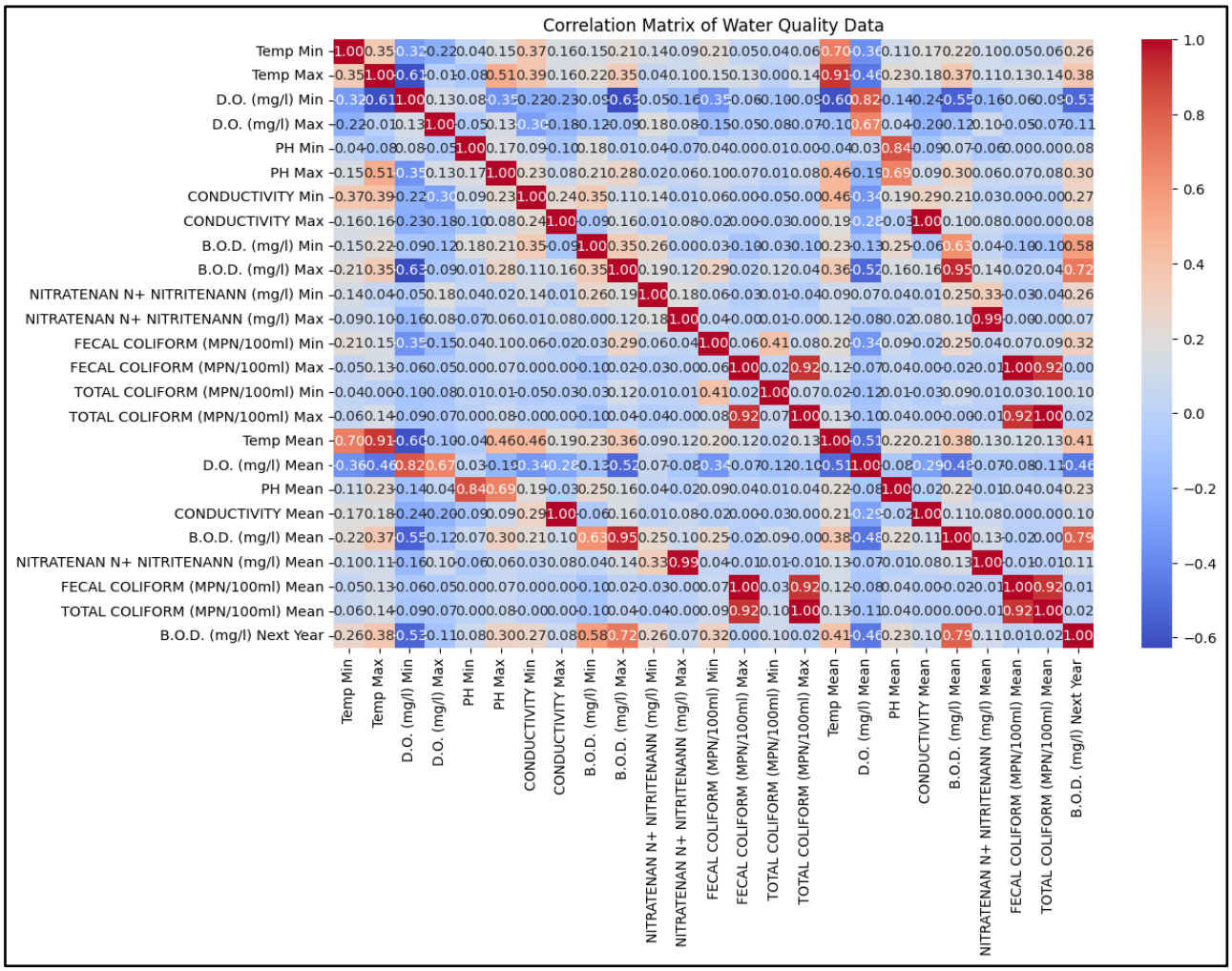

Fig. 2 Map of Correlation for Water Quality Data

This heatmap of the correlation matrix communicates how different variables in the dataset correlate with each other. A value greater than zero signifies a positive correlation, a value of zero means no correlation, and a negative value means that there is a negative correlation between the two variables.

2.8. Best Model Selection

The data was analyzed using multiple regression models, and based on the highest R^2 score, the Random Forest algorithm was selected as the best fit. Since Root Mean Squared Error (RMSE) lacks a baseline comparison and is scale-dependent, it would not prove as useful as the R^2 score for evaluating relative model performance.

2.9. Hyperparameter Tuning

Hyperparameter tuning is a procedure of identifying the best hyperparameters for an ML model to yield the best results. The Random Search CV function was used to identify the optimal hyperparameters imported from the Scikit-learn Library. This algorithm creates a grid of hyperparameter values, trains the model with random combinations, and scores them [19]. The optimal hyperparameters discovered for the RF model using this technique are listed below.

Table 2. Optimal hyperparameters for prediction model

Parameters	Optimal Value
Number of Estimators	300
Minimum Samples Split	5
Minimum Samples Leaf	2
Maximum Features	$log2$
Max Depth	30
Bootstrap	False

Table 3. Evaluation metrics for DO

Machine Learning Models	DO	
	R^2 Score	MAE
RF	0.761792	0.497305
XGBoost	0.713873	0.363876
Gradient Boosting	0.744607	0.384699
SVR	0.260032	0.620532
LR	0.530647	0.386729

3. Results & Discussion

3.1. R^2 Score and MAE Comparison

3.1.1. Variance

R^2 gives an idea of how much variation in the dependent variable is explained by the model, whereas MAE only provides error magnitude.

3.1.2. Comparability

R^2 is unitless and standardized between 0 and 1 (or negative in bad models), making it easier to compare models, while MAE depends on the scale of the data.

3.1.3 Relative Performance

An R^2 score of 0.8 signifies that 80% of the variance is explained, whereas an MAE of 5 does not tell us whether that is good or bad unless we know the data range.

3.1.4. Generalization

A model with a good R^2 value is likely to generalize better, while MAE alone may not capture model quality if the data has high variance.

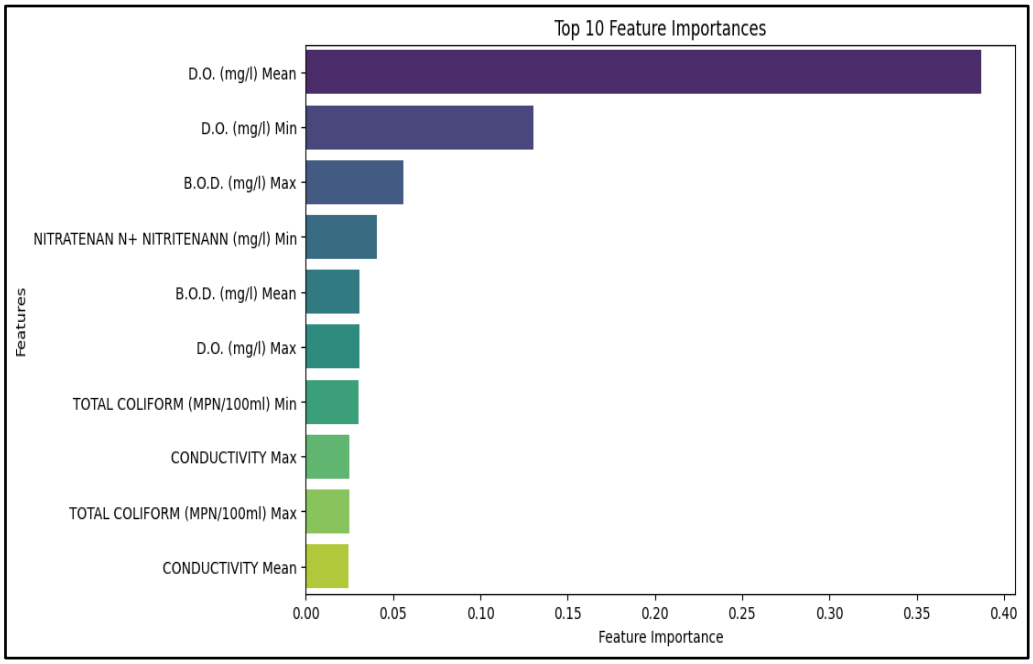

Fig. 6 Top 10 Feature Importances

This bar chart shows which features the model relies on the most; a longer bar means higher importance in predicting the target. We can see that D.O. (mg/l) Mean and Min are the top contributors, indicating that these dissolved oxygen metrics strongly influence the model's predictions.

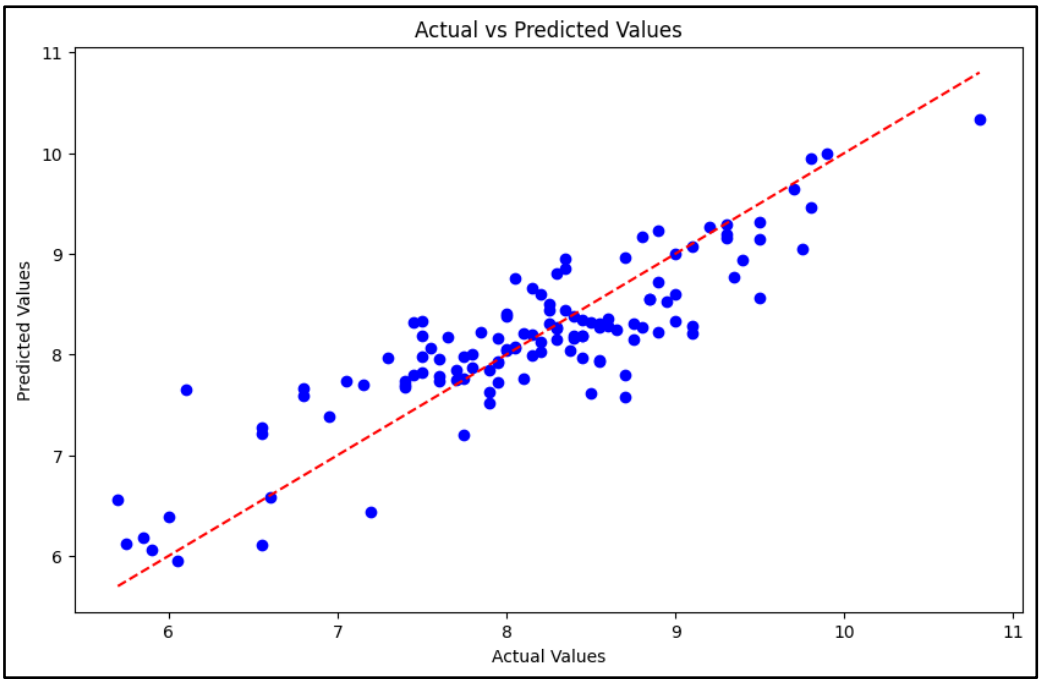

Fig. 7 Actual vs Predicted values

The scatter plot compares the model's predictions (y-axis) to the true values (x-axis). Points clustered around the diagonal (red dotted line) suggest the model performs well, with minimal prediction error.

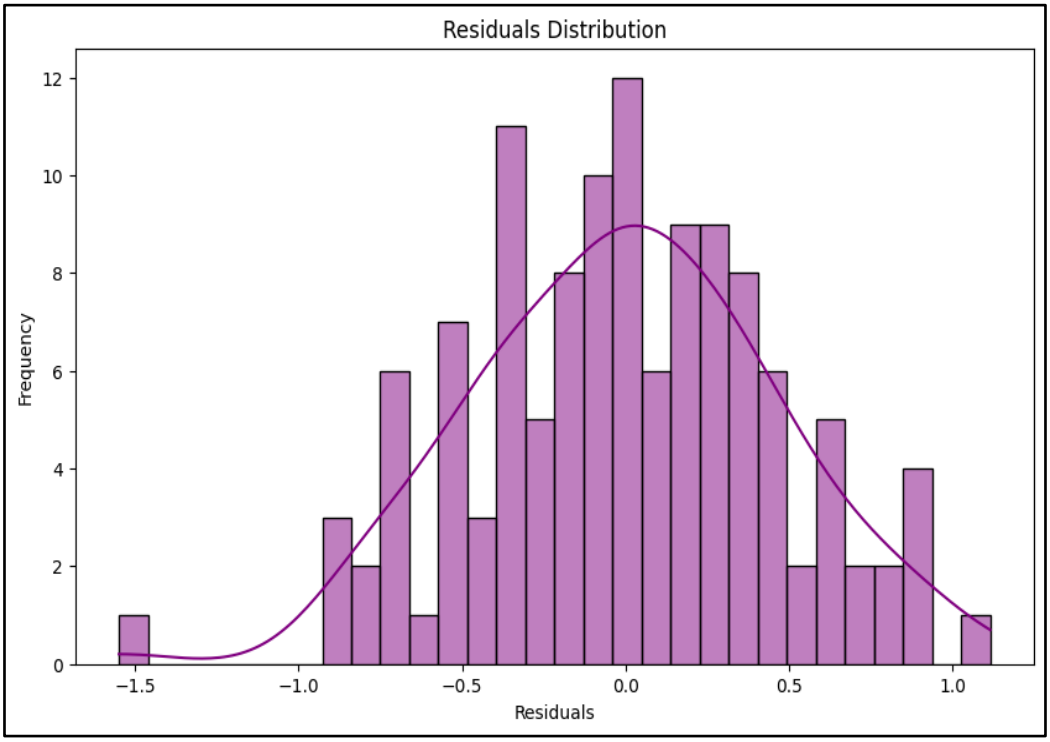

Fig. 8 Residuals Distribution

The histogram appears roughly bell-shaped and centered near zero. This demonstrates that the errors (residuals) are roughly normally distributed, a favorable indication for linear models such as RF.

There is a slight asymmetry (possibly a mild left skew), but it does not look severely skewed or heavily tailed overall.

The points appear fairly randomly scattered around the horizontal line at residual = 0. There is no obvious curvature or "fan" shape, which suggests that heteroscedasticity (unequal variance of errors) is not a large concern here. There does not appear to be a strong pattern (such as a systematic upward or downward trend), meaning the model does not show strong evidence of missing a nonlinear relationship.

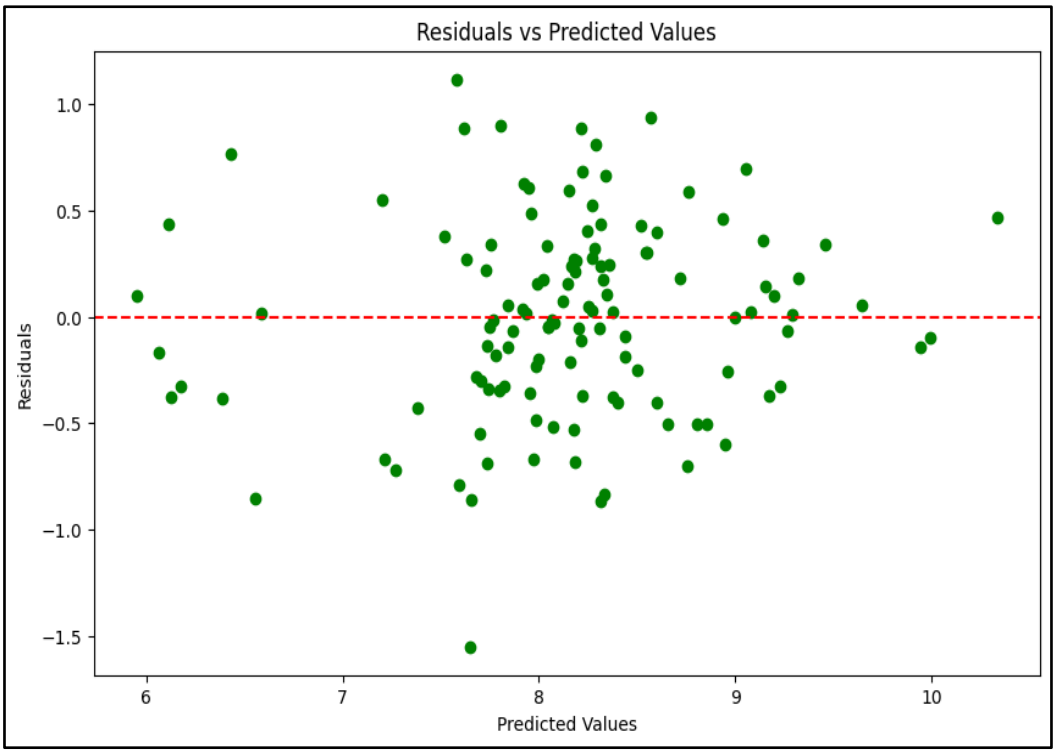

Fig. 9 Residuals vs Predicted Values

4. Conclusion

The study determined that the Random Forest machine learning model was the most reliable in predicting future trends in Dissolved Oxygen levels, with an accuracy of 76%. A limitation of this study is the lack of consistent and abundant data for making accurate predictions. Since this study relies on data collected and reported by government bodies for transparency and credibility, the inconsistencies in the data provided can limit the credibility of the paper's results and conclusions. In the future, the study could be extended by using the shortlisted Random Forest algorithm for making predictions of Dissolved Oxygen levels in the coming years since the study only delves into investigating which model would potentially give the most accurate results when forecasting future water quality metrics. Moreover, the model could be further trained to deal with factors such as seasonal variation. A robust water-quality prediction algorithm could significantly aid policy-making by helping set water-quality goals with future trends in mind.

References

[1] Rajnee Naithani, and I.P. Pande, "Comparative Analysis of the Trends in River Water Quality Parameters: A Case Study of the Yamuna River," *International Journal of Scientific Research Engineering & Technology*, vol. 4, no. 12, pp. 1212-1221, 2015. [Google Scholar] [Publisher Link]

[2] Michael Wiryaseputra, "Banknote Authentication Using Machine Learning Classification Algorithm," *International Journal of Scientific Engineering and Research*, vol. 8, no. 1, 2017. [Publisher Link]

[3] Md. Saikat Islam Khan et al., "Water Quality Prediction and Classification Based on Principal Component Regression and Gradient Boosting Classifier Approach," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 8, pp. 4773-4781, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Central Pollution Control Board, Water Quality Criteria, 2019. [Online]. Available: https://cpcb.nic.in/water-quality-criteria/

[5] Stream Water Quality – Importance of Temperature, Know Your H_2O, Water Research Center, [Online]. Available: https://www.knowyourh2o.com/outdoor-4/stream-water-quality-importance-of-temperature#:~:text=Temperature%20is%20a%20critical%20water,of%20chemical%20and%20biological%20reactions

[6] Dissolved Oxygen, Fondriest Environmental Learning Center. [Online]. Available: https://www.fondriest.com/environmental-measurements/parameters/water-quality/dissolved-oxygen/

[7] pH and Water, USGS. [Online]. Available: https://www.usgs.gov/special-topics/water-science-school/science/ph-and-water#:~:text=pH%20is%20really%20a%20measure,water%20that%20is%20changing%20chemically

[8] Indicators: Conductivity | US EPA. [Online]. Available: https://www.epa.gov/national-aquatic-resource-surveys/indicators-conductivity#:~:text=What%20can%20conductivity%20tell%20us,body%20and%20its%20associated%20biota

[9] Biochemical Oxygen Demand in Water Bodies. [Online]. Available: https://www.un.org/esa/sustdev/natlinfo/indicators/methodology_sheets/freshwater/biochemical_oxygen_demand.pdf

[10] 5.7 Nitrates, Monitoring & Assessment. [Online]. Available: https://archive.epa.gov/water/archive/web/html/vms57.html#:~:text=Nitrates%20from%20land%20sources%20end,nitrite%20methods;%20APHA,%201992

[11] Coliform Bacteria in Drinking Water, Washington State Department of Health. [Online]. Available: https://doh.wa.gov/community-and-environment/drinking-water/contaminants/coliform#:~:text=Coliform%20bacteria%20will%20not%20likely,feces%20of%20humans%20or%20animals

[12] Linear Regression in Machine Learning, GeeksforGeeks, 2025. [Online]. Available: https://www.geeksforgeeks.org/ml-linear-regression/

[13] Random Forest Algorithm in Machine Learning, GeeksforGeeks, 2025. [Online]. Available: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

[14] Gradient Boosting in ML, GeeksforGeeks, 2025. [Online]. Available: https://www.geeksforgeeks.org/ml-gradient-boosting/

[15] Neri Van Otten, Support Vector Regression (SVR) Simplified & How to Tutorial in Python, Spot Intelligence, 2024. [Online]. Available: https://spotintelligence.com/2024/05/08/support-vector-regression-svr/

[16] GeeksforGeeks, XGBoost – GeeksforGeeks, 2025. [Online]. Available: https://www.geeksforgeeks.org/xgboost/

[17] Ihechikara Abba, What is R Squared? R2 Value Meaning and Definition, freeCodeCamp.org. [Online]. Available: https://www.freecodecamp.org/news/what-is-r-squared-r2-value-meaning-and-definition/#:~:text=R-Squared%20values%20range%20from,50%,%20and%20so%20on

[18] Oluniyi Oluniyi, "*Contextual and Ethical Issues with Predictive Process Monitoring*," PhD thesis, University of Westminster School of Computer Science and Engineering Westminster, pp. 1-190, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Will koehrsen, Kaggle. [Online]. Available: https://www.kaggle.com/code/willkoehrsen/intro-to-model-tuning-grid-and-random-search