Prediction of Different Water Quality Parameters using Regression Model

Paragkanti Chattopadhyay[1], Sumit Banerjee[2], Chandan Kumar Chanda[3], Debasis Guha[4], Sourav Bhattacharya[5]
1Department of Computer Science and Engineering, Dr. B. C. Roy Engineering College, Durgapur
2 Department of Electrical Engineering, Dr. B. C. Roy Engineering College, Durgapur
3Department of Electrical Engineering, IIEST Shibpur
4Department of Master of Computer Applications (MCA), Dr. B. C. Roy Engineering College, Durgapur
5Department of Basic Science and Humanities, Dr. B. C. Roy Engineering College, Durgapur

Abstract: This paper analyses water pollution data focusing on Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), pH, and turbidity. Each parameter has specific acceptable ranges to ensure that water is safe and suitable for consumption. BOD, COD, pH and turbidity levels using a Comma separated value (CSV) dataset. Initially, the data is pre-processed by cleaning it to handle missing values and outliers, ensuring consistency and accuracy. The dataset was then divided into training and testing subsets. For modelling, linear regression of machine learning will be employed. The model was trained on the processed data and evaluated using metrics such as R^2 score and Root Mean Square Error (RMSE) to assess pollutants label of BOD, COD, pH and turbidity of Damodar river flowing through Durgapur with some predefined threshold values based on past experience. Finally, the above analysis will be interpreted to understand the patterns and factors influencing pollution levels of BOD, COD, pH and turbidity and the results are found to be in very good agreement.

Keywords: Water Pollution, Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), pH, turbidity, Machine Learning.

1. Introduction

The Damodar River, a vital water resource for communities in West Bengal, has been facing significant water pollution due to rapid industrialization and urbanization in the regions of Durgapur. This study investigates the sources, extent, and impacts of water pollution in these areas, with a focus on industrial effluents, untreated sewage, and mining activities. Water samples were collected from key sites along the river to analyze parameters such as pH, dissolved oxygen (DO), chemical oxygen demand (COD), and heavy metals. The results revealed alarming concentrations of pollutants, with some areas showing a drastic decline in water quality, making it unsafe for human consumption and aquatic life. The study also highlights the socio-economic implications of water pollution on local communities, particularly in terms of public health and livelihood and cultivation. the paper emphasizes the urgent need for comprehensive water management policies, stricter regulations on industrial discharges, and community awareness programs to mitigate the ongoing pollution crisis in the Damodar River.The Damodar River, often referred to as the "Sorrow of Bengal," has played a pivotal role in the socio-economic development of the region. Flowing through the industrial belt of West

Bengal, it has been a crucial water resource for agriculture, domestic use, and industrial activities. However, over the decades, rapid industrialization, urbanization, and unregulated mining activities in regionof Durgapur have significantly altered the river's ecosystem, leading to severe water pollution. The influx of untreated industrial effluents, domestic sewage, and the deposition of hazardous chemicals have led to a sharp decline in water quality, threatening the health of both humans and aquatic life and plant also. The regions surrounding the Damodar River are home to numerous coal mines, steel plants, and chemical industries, which discharge large volumes of pollutants into the water body. Additionally, rapid urban expansion and insufficient waste management infrastructure have exacerbated the pollution problem. This paper aims to assess the current state of water pollution in these key areas, examine the primary sources of contaminants, and evaluate the impact of this pollution on public health, local ecosystems, and economic activities. Through detailed water quality analysis and field surveys, the study highlights the urgency for implementing effective pollution control measures and sustainable water management practices. The findings of this research are crucial for policy-makers, environmentalists, and local communities, as they provide a comprehensive understanding of the pollution challenges faced by the Damodar River. It is essential to develop coordinated efforts to restore and protect the river, ensuring its continued role in supporting the livelihoods of millions.M N Vamsi Thalatamet al[1] discussed that a real-time water contamination monitoring system is developed using Internet of Things (IoT) and Embedded Systems. This system can able to estimate the water quality in residential homes as well water body.D. Kavitha et al[2] discussed that heat, pH, EC, hardness, chlorine ions, alkalinity, phosphate, and sulphur are one of the key factor water pollution apart from this the increase in pollution concentration indicates a growth in the amount of pollution in the air as a result of human activity, including the disposal of wastes into rivers, the discharge of residential sewage, wastewaters, and other pollutants. In this paper several pieces of information are captured all through the analysis of the dataset using the supervised machine learning approach (SMLT), such as the classification of variables and results from uni-variate, bi-variate, and multivariate analyses. The effectiveness of different machine learning techniques from the provided dataset is compared and debated using evaluation techniques.et al[3] parameters read the by sensor and then the values are stored in microcontroller and visible on smart devices by analyzing the parameter the quality of water are known. Jeetendra Kumar et al[4] collected information, which may include pH, turbidity, and TDS readings, is uploaded to the cloud so that it may be evaluated in real-time using the AquaSpecs app. The effectiveness of the proposed system has been proven by deployment in four ponds in Chhattisgarh. J. O. Ighalo et al[5] discussed that IoT-Enabled Advanced Water Quality Monitoring System quality indicators. Water, as a vital natural resource, has a crucial impact on our everyday existence. Monitoring the quality of water is an essential component of both environmental management and public health. et al[6] in there work, a review was carried out on several low-cost developed technologies and applied in situ for water quality monitoring. V. Kothari et al[7] used correlation matrix. The correlation matrix shows that total iron concentration, total coliform, and faecal coliform have a significant effect on Water quality index. M. G. Uddin et al[8] presents a comparative discussion of the most commonly used WQI models, including the different model structures, components, and applications. N. H. Omer et al[9] investigates the complex interplay between meteorological variables and water quality parameters in Nairobi City. P. Soni et al[10] discussedwater quality is a crucial component of a well-balanced environment. Uncontaminated water is crucial for the sustenance of a diverse array of flora and fauna. Although first appearing

inconsequential, human terrestrial activities significantly impact the water quality. Anil K Dwivedi[11] discussed that denitrifying bacteria also play an important role in nitrogen concentration of a medium saunders . Kostandina et al. [12] discussed the use of recurrent neural networks (RNN) to forecast air pollutant levels at any given time and eliminate hourly prediction errors due to the algorithm's memorization capabilities. However, they noted a lack of ability to operate without memory functions. Xiaosong Zhao et al [13] used the RNN method for addressing AQC forecasting and improved the performance of air quality prediction. Mohurlee et al. [14] predicted PM2.5 and PM10 levels using fuzzy logic. Fuzzy logic helps remove outliers caused by the presence of unwanted gases in the atmosphere. However, fuzzy logic involves clusters that may retain redundant data, leading to incorrect predictions. CR et al. [15] used autoregression in their study to detect whether the air was polluted, and linear regression was employed to determine PM2.5 levels. However, the limitation was that it could not accurately determine PM2.5 levels when there were changes in atmospheric conditions. Additionally, it accounted for meteorological factors such as wind speed and temperature. Zhang et al. [16] discussed the wavelet neural network as a robust method for determining air pollutant levels. However, it lacked the ability to identify an appropriate wavelet function and the exact number of hidden layers required in their study, which led to inaccurate predictions of air pollutant levels. Mejía et al[17] have used machine learning and IoT for the prediction of air pollution. They expressed the view that machine learning algorithms are quite effective. Angelin et al[18] for predicting air pollution, they used a hybrid model. proved that it is one of the best models for predicting air pollution in the future.

2. Methodology

2.1 Problem Definition

The goal of this paper is to predict the pollutant levels of the Damodar River which is flowing through Durgapur, including Biochemical Oxygen Demand (BOD), pH, Turbidity, and Chemical Oxygen Demand (COD), for the year 2027. The prediction model is based on historical data from 2017 to 2024.

2.2 Data Collection : Historical data for pollutants (BOD, pH, Turbidity, and COD) for the years 2017 to 2024 was collected from reliable sources such as environmental monitoring agencies or relevant databases.

The pollutants (BOD, pH, Turbidity, and COD) are considered as dependent variables (targets).

Years 2017 to 2024 are considered as independent variables (predictors).

2.3 Preprocessing : Before applying the linear regression model, data preprocessing is required

2.4 Handling Missing Data: If any missing values are found, they are either filled using appropriate methods (mean, median, interpolation) or the corresponding records are removed.

2.5 Normalization/Standardization: If required, the data is scaled using normalization or standardization techniques to ensure the model's performance is not biased towards variables with larger ranges.

2.6 Outlier Detection: Outliers are detected using statistical methods (like IQR or Z-score) and handled by either removing or capping the values.

2.7 Feature Engineering

2.7.1 Independent Variables (Features): The year (2017–2024) serves as the main independent variable.

2.7.2 Dependent Variables (Targets): The pollutants (BOD, pH, Turbidity, COD) serve as the dependent variables.

2.7.3 Transformation (if needed): In case relationships between the features and targets are nonlinear, logarithmic or polynomial transformations may be applied.

2.8 Model Selection (Linear Regression)

A **Linear Regression** model is used due to its simplicity and interpretability. The model assumes a linear relationship between the dependent variable (pollutants) and the independent variable (year). The model can be represented as:

$$Y=\beta0+\beta1X1+\beta2X2+\cdots+\betanXn+\epsilonY = \beta_0 + \beta_1X_1 + \beta_2X_2 + \dots + \beta_nX_n + \epsilonY=\beta0+\beta1X1+\beta2X2+\cdots+\betanXn+\epsilon$$

Where:

- YYY is the dependent variable (one of the pollutants: BOD, COD, Turbidity, or pH).
- X1,X2,…,XnX_1, X_2, \dots, X_nX1,X2,…,Xn are the independent variables (environmental features like temperature, water flow, etc.).
- β0\beta_0β0 is the intercept.
- β1,β2,…,βn\beta_1, \beta_2, \dots, \beta_nβ1,β2,…,βn are the coefficients that represent the influence of each feature on the dependent variable.
- ϵ\epsilonϵ is the error term.
- level.

2.9 Model Training

2.9.1 Train-Test Split: The dataset is split into training and testing sets. A common split ratio is 80% training data and 20% testing data.

2.9.2 Fitting the Model: The linear regression model is trained on the training dataset, where the model learns the coefficients (slopes) of the regression line by minimizing the error using techniques such as Ordinary Least Squares (OLS).

2.10 Model Evaluation: The model‘s performance is evaluated using metrics such as:

- **Root Mean Squared Error (RMSE)**
- **R-Squared (R^2)** to check how well the model explains the variance in the target pollutants.

2.11 Prediction for 2027

Once the model is trained and evaluated, it can be used to predict the pollutant levels (BOD, pH, Turbidity, COD) for the year 2027. The predictions are based on the coefficients (β0\beta_0β0, β1\beta_1β1) derived from the historical data (2017–2024).

2.12 Results Interpretation

The predicted pollutant levels for 2027 will be analyzed.

The results are then compared to any available forecast or real-world data (if applicable) to validate the model‘s performance.

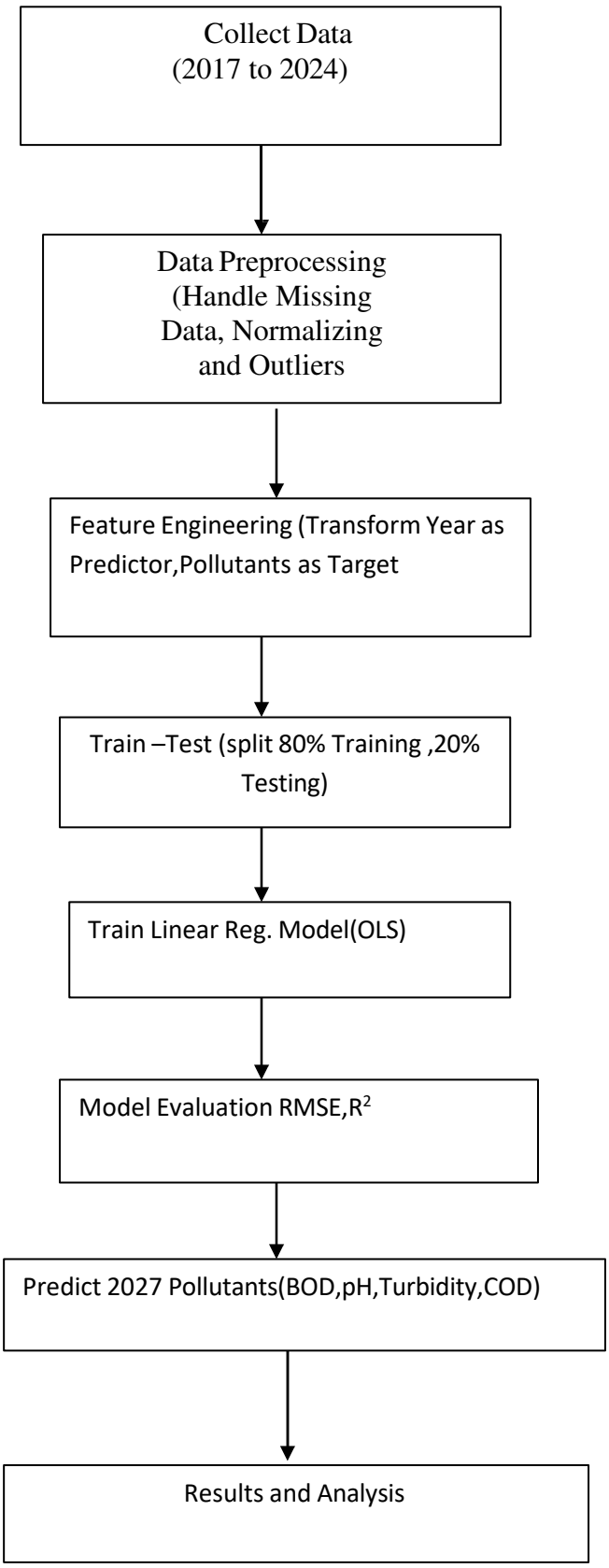
3. Flow Chart

This flowchart captures the main steps from data collection through to model evaluation and prediction for the year 2027.

This methodology outlines the necessary steps to predict the pollutant levels in the Damodar River at Durgapur for the year 2027 using linear regression. The model is trained on data from 2017 to 2024, and predictions are made for future years based on the learned patterns from historical data.

Machine Learning Linear Regression Model for Predicting Pollutant Levels in Damodar River (Durgapur) for 2027

In thispaperLinear Regression model to predict the pollutant levels in the Damodar River at Durgapur for the year 2027, based on key water quality parameters: Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Turbidity, and pH. And this paper also evaluated the model using Root Mean Squared Error (RMSE) and R-squared (R^2) metrics.


Collect Data
(2017 to 2024)
Data Preprocessing
(Handle Missing
Data, Normalizing
and Outliers
Feature Engineering (Transform Year as
Predictor,Pollutants as Target
Train –Test (split 80% Training ,20%
Testing)
Train Linear Reg. Model(OLS)
Model Evaluation RMSE,R^2
Predict 2027 Pollutants(BOD,pH,Turbidity,COD)
Results and Analysis

4. Linear Regression

Linear regression is a foundational algorithm in machine learning, often used for predictive modeling when the relationship between the target variable and one or more independent variables is assumed to be linear. In this paper the target variables are the pollutant levels such as BOD, COD, Turbidity, and pH, while the predictors (features) could include environmental and water quality factors like temperature, flow rate, rainfall, and industrial discharge, among others.

The general form of a multiple linear regression model is

Y=β0+β1X1+β2X2+⋯+βnXn+ϵY = \beta_0 + \beta_1X_1 + \beta_2X_2 + \dots + \beta_nX_n + \epsilonY=β0+β1X1+β2X2+⋯+βnXn+ϵ

Where

- YYY is the dependent variable (one of the pollutants: BOD, COD, Turbidity, or pH).
- X1,X2,…,XnX_1, X_2, \dots, X_nX1,X2,…,Xn are the independent variables (environmental features like temperature, water flow, etc.).
- β0\beta_0β0 is the intercept.
- β1,β2,…,βn\beta_1, \beta_2, \dots, \beta_nβ1,β2,…,βn are the coefficients that represent the influence of each feature on the dependent variable.
- ϵ\epsilonϵ is the error term.

By training the linear regression model on historical data, the algorithm learns the optimal values for the coefficients β\betaβ, allowing the model to make predictions on unseen data.

4.1 Model Evaluation Using RMSE and R^2 Score

To assess the effectiveness of your linear regression model, you've used two key evaluation metrics: Root Mean Squared Error (RMSE) and R-squared (R^2). Both metrics provide valuable insights into how well the model is performing in predicting the pollutant levels.

4.2 Root Mean Squared Error (RMSE)

RMSE is a common metric to evaluate the performance of regression models. It gives you an idea of the average magnitude of the prediction error. RMSE is calculated as:

RMSE=n1i=1∑n(Ytrue,i−Ypred,i)2

Where

- Ytrue,iY_{\text{true},i}Ytrue,i is the actual value of the pollutant.
- Ypred,iY_{\text{pred},i}Ypred,i is the predicted value of the pollutant.
- nnn is the number of observations.

RMSE gives you the square root of the average squared differences between actual and predicted values. A lower RMSE indicates that the model's predictions are closer to the actual values, meaning the model performs better. In your case, the RMSE would show how well the linear regression model predicts the levels of BOD, COD, Turbidity, and pH for 2027.

4.3 R-squared (R^2) Score

R^2, also known as the coefficient of determination, indicates how well the independent variables explain the variability in the dependent variable (pollutants). The formula for R^2 is:

R2=1−∑(Ytrue−Ypred)2∑(Ytrue−Y¯)2R^2 = 1 - \frac{\sum (Y_{\text{true}} - Y_{\text{pred}})^2}{\sum (Y_{\text{true}} - \bar{Y})^2}R2=1−∑(Ytrue−Y¯)2∑(Ytrue−Ypred)2

Where

- YtrueY_{\text{true}}Ytrue is the actual observed values.
- YpredY_{\text{pred}}Ypred is the predicted values.
- Y¯\bar{Y}Y¯ is the mean of the actual values.

4.3.1 R^2 ranges from 0 to 1

- An R^2 of 1 indicates a perfect fit, meaning the model explains all the variance in the dependent variable.
- An R^2 of 0 suggests that the model doesn't explain any of the variance in the dependent variable.
- Higher R^2 values indicate that the model does a better job at predicting the pollutant levels based on the independent features.

For such cases, R^2 will tell how well the linear regression model has captured the relationship between the predictors (environmental factors) and the target pollutants (BOD, COD, Turbidity, and pH).

4.4 Model Performance and Interpretation

After training the linear regression model on the historical data, you would have evaluated the model using both RMSE and R^2 score. Here's how you would interpret the results:

- RMSE:

  - A low RMSE indicates that the model's predictions are close to the actual pollutant levels.
  - If the RMSE is relatively high, it suggests that the model struggles to make accurate predictions, which may indicate that the relationship between features and the pollutants is not purely linear, or the features chosen are not sufficient to explain the pollutant variability.
- R^2 Score:
  - A high R^2 score (close to 1) means that the model has a strong predictive power and that the chosen features are significant in explaining the pollutant levels.
  - A low R^2 score (close to 0) suggests that the model's explanatory power is weak and that there may be other factors influencing the pollutant levels that are not captured by the current set of features.

5. Different Water Quality Parameters

The most important water quality parameters are namely BOD, COD, Turbidity and pH. Each of these four pollutants represents a different aspect of water quality:

Assessing water quality involves evaluating various parameters, including Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), pH, and turbidity. Each parameter has specific acceptable ranges to ensure the water is safe and suitable for consumption.

- Biochemical Oxygen Demand (BOD): BOD measures the amount of oxygen required by microorganisms to decompose organic matter in water. Lower BOD values indicate better water quality. According to the Bureau of Indian Standards (BIS), the acceptable limit for BOD in drinking water is less than 3 mg/L.

- Chemical Oxygen Demand (COD): COD indicates the total quantity of oxygen required to oxidize both organic and inorganic substances in water. The BIS recommends a COD level of less than 20 mg/L for good quality water.

- pH: pH measures the acidity or alkalinity of water on a scale from 0 to 14, with 7 being neutral. For drinking water the acceptable pH range is between 6.5 and 8. Most aquatic organisms thrive in a narrow pH range (typically 6.5 to 8.5). Extreme pH levels can be harmful to aquatic life .

- Turbidity: Turbidity refers to the cloudiness or haziness of water caused by suspended particles. High turbidity can harbor pathogens and reduce the effectiveness of disinfection processes. The permissible turbidity level in drinking water is less than 5 Nephelometric Turbidity Units (NTU).

Therefore, water to be considered of good quality, it should have a BOD less than 3 mg/L, COD less than 20 mg/L, pH between 6.5 and 8.5, and turbidity less than 5 NTU. Maintaining these parameters within the specified ranges ensures the water is safe for consumption and meets established quality standards.

Linear regression helps establish the relationships between environmental features and these water quality parameters, allowing you to predict their values for 2027.

6. Result and Discussion

Table 1(a): Raw Data of year 2017

date	BOD	pH	Turbidity	COD
02-01-2017	2.75	8.2	3.58	7.26
09-02-2017	4.6	7.93	5.36	10.78
27-03-2017	2.55	8.68	4.08	9.39
03-04-2017	2.8	7.89	2.89	24.93
23-05-2017	4.05	7.85	3.71	12.54
06-06-2017	2.9	8	1.67	12.12
04-07-2017	4.05	7.29	42.6	14.59
01-08-2017	2	7.05	329	5.81
04-09-2017	1.8	7.88	89.6	10.22
12-10-2017	3.1	8.15	91	8.15
09-11-2017	4.25	7.35	16.9	15.1
04-12-2017	3.3	7.19	9.49	7.42

Table 1(b): Predicted data of the year 2027 based on the year 2017

date	BOD	Turbidity	pH	COD
31-01-2027	2.47	5.83	6.80	1.17
28-02-2027	3.67	7.68	7.16	9.40
31-03-2027	4.81	7.95	7.44	7.59
30-04-2027	4.31	2.89	7.00	8.37
31-05-2027	2.86	8.87	7.05	8.97
30-06-2027	5.22	1.27	6.81	5.64
31-07-2027	4.49	4.38	7.74	2.11
31-08-2027	5.36	2.91	7.84	4.64
30-09-2027	4.40	9.05	6.95	8.87
31-10-2027	4.57	8.12	7.50	9.15
30-11-2027	4.70	9.85	7.06	9.63
31-12-2027	3.34	1.33	7.97	6.96
	R2 Score:1 RMSE:1.40	R2 Score:1 RMSE:2.59	R2 Score:1 RMSE:1.045	R2 Score:1 RMSE:4.7

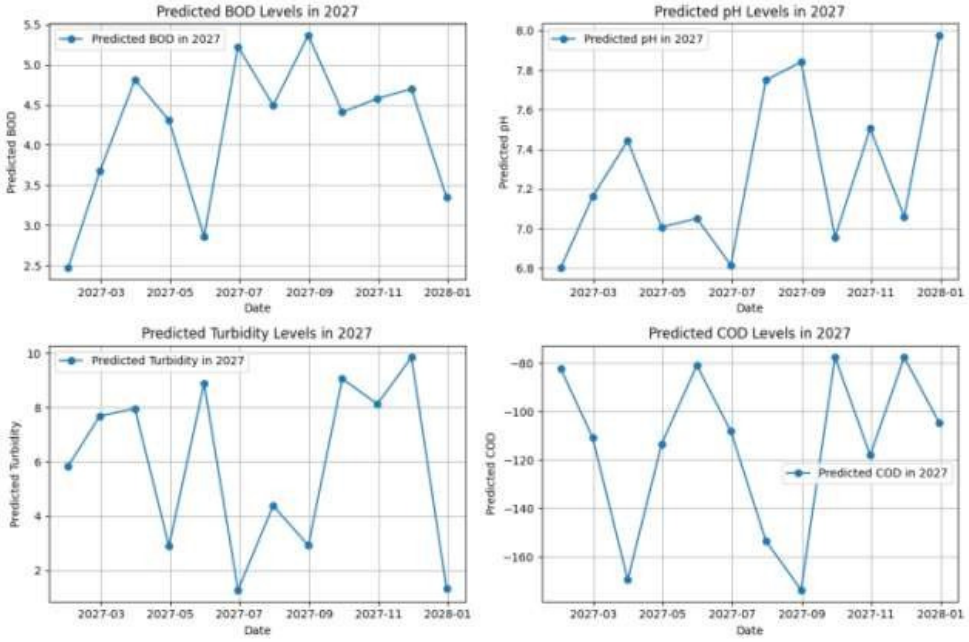


Figure 1: Predicted data of BOD, COD, pH and Turbidity of the year 2027 based on the year 2017

Table 2(a): Raw Data of year 2018

date	BOD	pH	Turbidity	COD
08-01-2018	2.3	8.03	8.54	14.59
15-02-2018	2.45	8.21	0.95	14.59
06-03-2018	2.6	6.92	7.65	9.39
26-04-2018	2.7	7.79	8.99	19.39
22-05-2018	2.6	8.36	4.97	13
26-06-2018	2.5	7.64	37	14.33
24-07-2018	2.5	7.52	9.97	9
07-08-2018	2.8	7.92	95.7	19
20-09-2018	1.95	7.62	97.1	13
10-10-2018	2.55	7.94	46.5	19
15-11-2018	3.15	7.38	12.1	15.81
06-12-2018	2.1	8.23	33.6	14.61

Table 2(b): Predicted data of the year 2027 based on the year 2018

date	BOD	Turbidity	pH	COD
31-01-2027	.85	8.88	9.02	6.52
28-02-2027	.91	2.37	8.96	6.63
31-03-2027	1.03	4.47	8.88	9.02
30-04-2027	1.22	7.74	8.75	10.95
31-05-2027	1.17	6.85	8.77	10.78
30-06-2027	1.10	7.04	8.83	9.50
31-07-2027	1.06	1.17	8.84	9.62
31-08-2027	1.09	7.30	8.84	9.63
30-09-2027	1.38	7.18	8.62	13.06
31-10-2027	1.42	4.30	8.58	14.00
30-11-2027	1.36	8.97	8.65	12.57
31-12-2027	1.42	4.64	8.58	14.01
	R2 Score:1 RMSE:0.61	R2 Score:1 RMSE:0.099	R2 Score:1 RMSE:1.33	R2 Score:1 RMSE:1.86

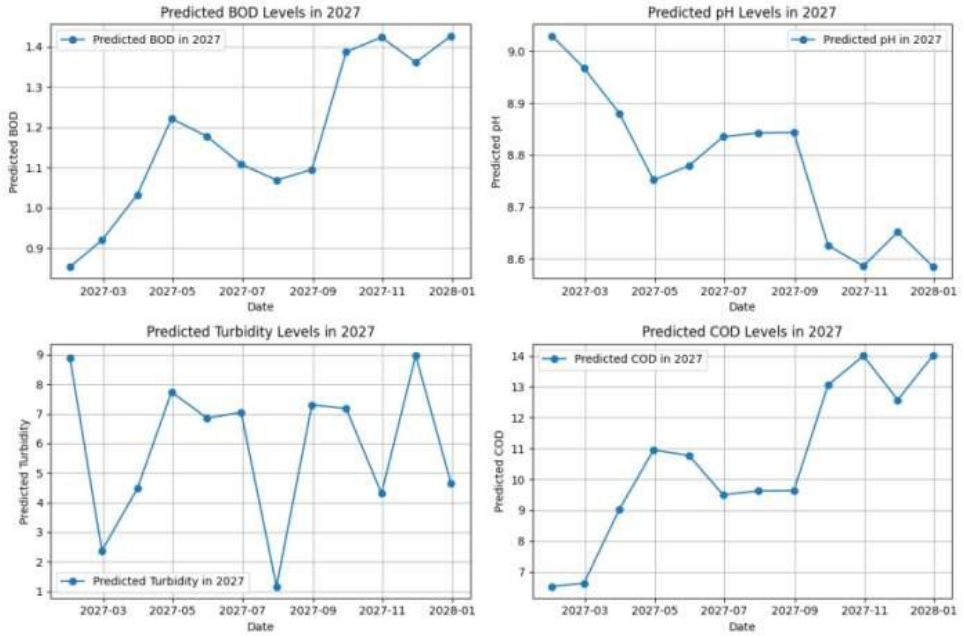


Figure 2: Predicted data of BOD, COD, pH and Turbidity of the year 2027 based on the year 2018

Table 3(a): Raw Data of year 2019

date	BOD	pH	Turbidity	COD
02-01-2019	1.3	7.29	17.8	15.9
12-02-2019	1.05	7.21	8.24	14.28
14-03-2019	1.55	8.15	19.82	15
17-04-2019	3.6	8.03	24.1	12.74
13-05-2019	1.6	8.51	9.95	15
13-06-2019	1.75	8.16	5.41	17.82
22-07-2019	1.85	8	12.68	18.85
21-08-2019	1.65	8.48	11.62	15.36
25-09-2019	1.4	7.82	6.41	13.86
16-10-2019	1.45	8.14	46.14	12.14
08-11-2019	1.65	7.98	32.16	11.88
12-12-2019	1.8	8.14	34.18	15.18

Table 3(b): Predicted data of the year 2027 based on the year 2019

date	BOD	Turbidity	pH	COD
31-01-2027	1.23	26.76	7.54	2.36
28-02-2027	1.25	21.84	7.40	9.18
31-03-2027	1.15	22.79	7.36	4.33
30-04-2027	1.33	26.14	7.58	7.02
31-05-2027	1.25	24.08	7.47	6.41
30-06-2027	1.36	24.94	7.56	9.52
31-07-2027	1.33	24.04	7.51	9.33
31-08-2027	1.35	24.83	7.55	9.30
30-09-2027	1.34	26.78	7.60	6.46
31-10-2027	1.44	27.55	7.70	9.42
30-11-2027	1.32	25.10	7.54	7.82
31-12-2027	1.39	26.63	7.63	8.55
	RMSE:0.29	RMSE:7.14	RMSE:.026	RMSE:1.86

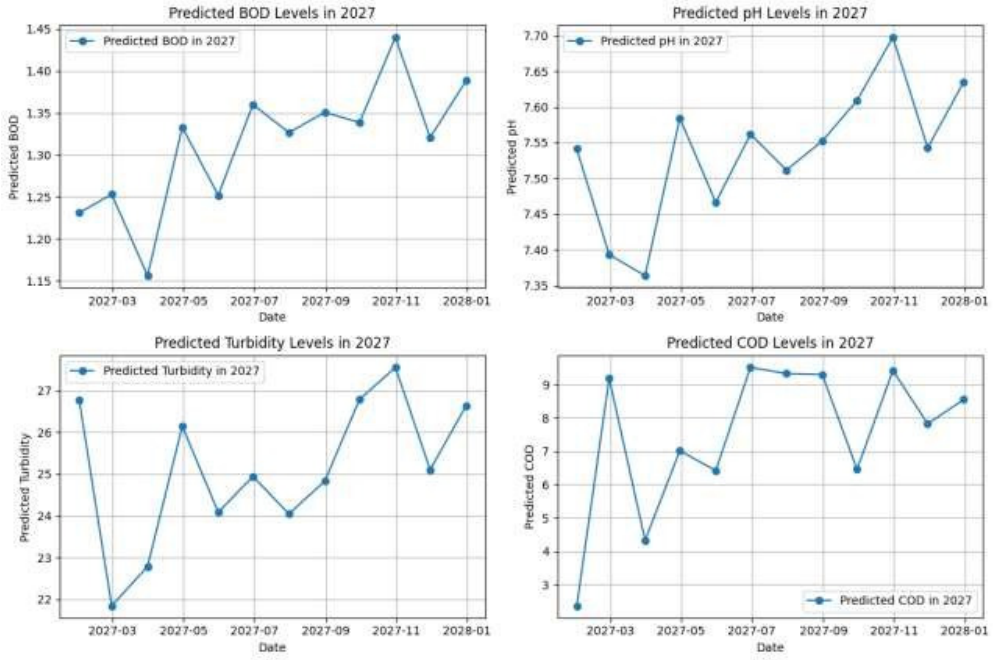


Figure 3: Predicted data of BOD, COD, pH and Turbidity of the year 2027 based on the year 2019

Table 4(a): Raw Data of year 2020

date	BOD	pH	Turbidity	COD
20-01-2020	1.75	8.22	11.18	14.4
12-02-2020	1.7	7.78	13.5	14.85
11-03-2020	1.85	7.88	22.24	13.16
13-04-2020	2.2	7.93	19.2	11.88
18-05-2020	2.35	7.63	19.82	13.16
15-06-2020	1.9	7.85	22.8	15.84
13-07-2020	2.05	7.45	26.46	19.45
11-08-2020	2.75	7.88	15.5	18.36
08-09-2020	1.8	7.38	65.4	14.17
07-10-2020	1.7	7.44	14.9	12.5
25-11-2020	1.6	7.49	31.4	12.1
16-12-2020	1.7	7.56	16.34	15.04

Table 4(b): Predicted data of the year 2027 based on the year 2020

date	BOD	Turbidity	pH	COD
31-01-2027	4.54	1.82	8.07	1.14
28-02-2027	4.52	6.92	7.98	4.85
31-03-2027	5.13	2.23	6.95	1.89
30-04-2027	4.36	9.23	7.38	5.92
31-05-2027	4.41	4.36	7.72	7.83
30-06-2027	3.39	4.86	7.85	9.90
31-07-2027	5.41	9.81	7.13	2.77
31-08-2027	4.62	7.76	8.07	5.55
30-09-2027	3.54	6.34	7.71	4.74
31-10-2027	5.09	7.79	7.73	7.60
30-11-2027	4.16	1.54	7.39	5.07
31-12-2027	2.89	5.32	7.99	6.24
	R^2 Score:1 RMSE:1.18	R^2 Score:1 RMSE:1.13	R^2 Score:1 RMSE:4.84	R^2 Score:1 RMSE:3.55

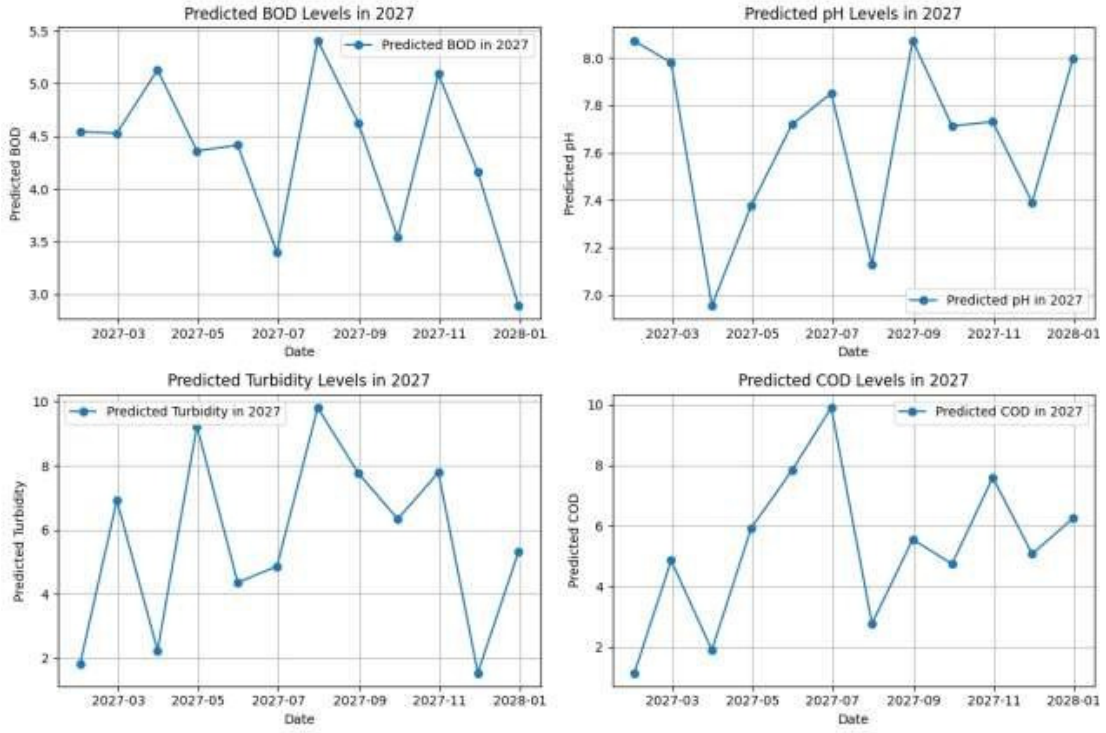


Figure 4: Predicted data of BOD, COD, pH and Turbidity of the year 2027 based on the year 2020

Table 5(a): Raw Data of year 2021

date	BOD	pH	Turbidity	COD
19-01-2021	1.6	7.71	9.33	10.67
04-02-2021	1.9	7.77	21.4	14.03
05-03-2021	2.75	7.14	7.11	22.99
13-04-2021	2.65	8.18	31.6	14.64
25-06-2021	2	7.5	107	18.05
22-07-2021	2.45	7.69	42.5	15.8
11-08-2021	2.3	7.52	157	20.82
07-09-2021	1.9	7.4	63.2	17.84
26-10-2021	1.75	7.52	51.4	19.48
23-11-2021	1.9	7.42	31.3	18.5
24-12-2021	1.8	7.6	40	17.62

Table 5(b): Predicted data of the year 2027 based on the year 2021

date	BOD	Turbidity	pH	COD
31-01-2027	4.57	1.81	7.12	7.18
28-02-2027	4.75	8.40	7.72	5.08
31-03-2027	2.58	3.12	7.21	9.81
30-04-2027	5.02	7.31	7.63	1.14
31-05-2027	2.50	6.04	7.61	9.39
30-06-2027	3.93	2.44	7.82	2.07
31-07-2027	5.02	4.62	7.83	5.70
31-08-2027	4.86	3.93	7.36	6.52
30-09-2027	3.00	1.66	7.54	3.05
31-10-2027	5.42	7.33	8.00	2.63
30-11-2027	5.89	8.52	6.80	1.94
31-12-2027	4.59	2.81	8.12	1.01
	R^2 Score:1 RMSE:	R2 Score:1 RMSE:1.53	R2 Score:1 RMSE:1.33	R2 Score:1 RMSE:1.45

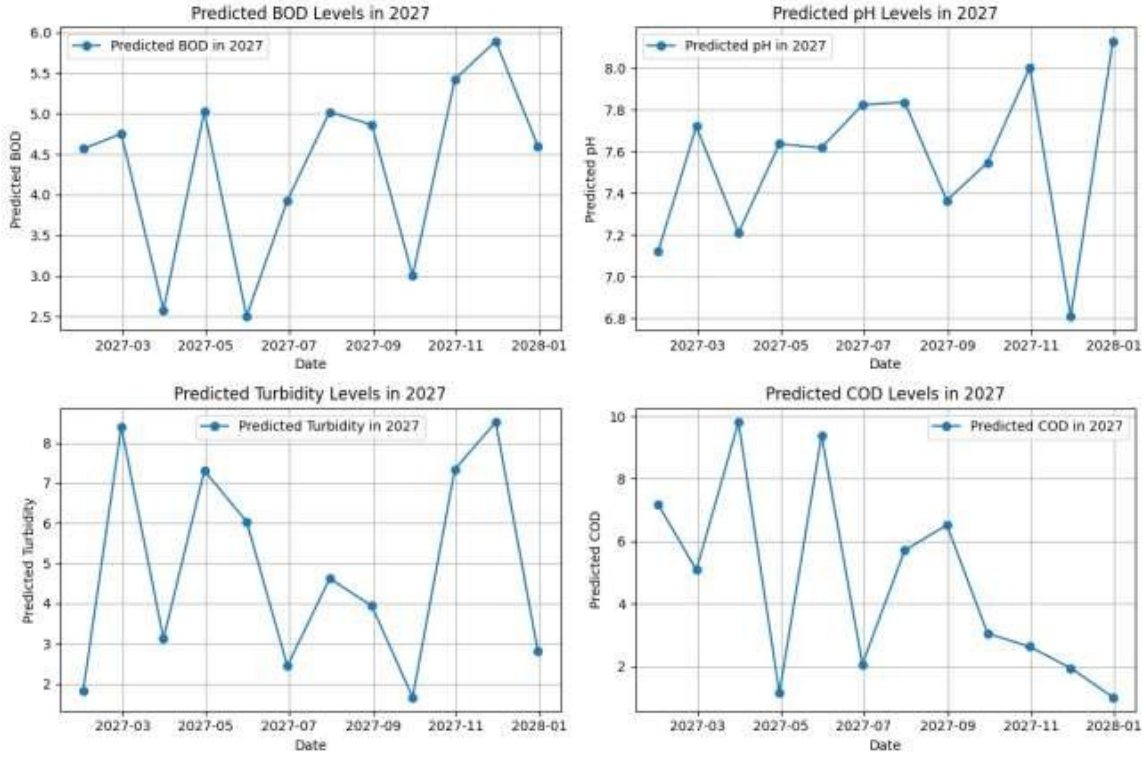


Figure 5: Predicted data of BOD, COD, pH and Turbidity of the year 2027 based on the year 2021

Table 6(a): Raw Data of year 2022

date	BOD	pH	Turbidity	COD
17-01-2022	1.85	7.69	26.7	17.18
21-02-2022	2.05	7.1	25.8	16.19
08-03-2022	2.65	7.98	13.8	23.52
07-04-2022	2.75	7.92	22.5	22.08
10-05-2022	2.85	7.52	20.9	23.46
09-06-2022	2.55	8.1	17.9	18.72
11-07-2022	2.75	7.88	15.5	18.36
04-08-2022	2.85	7.55	27.9	20.79
06-09-2022	2.8	7.38	37.3	21.12
11-10-2022	2.65	7.82	48.1	19.23
14-11-2022	2.25	7.92	20.8	17.64
07-12-2022	2.35	7.68	34.4	18.85

Table 6(b): Predicted data of the year 2027 based on the year 2022

date	BOD	Turbidity	pH	COD
31-01-2027	5.06	3.15	8.04	3.56
28-02-2027	2.81	6.65	6.98	9.88
31-03-2027	2.54	7.34	7.02	6.25
30-04-2027	5.53	5.57	6.97	1.97
31-05-2027	5.89	5.99	7.69	2.81
30-06-2027	3.64	2.90	6.84	9.95
31-07-2027	2.52	2.83	7.95	8.95
31-08-2027	2.27	4.20	7.68	1.44
30-09-2027	4.32	2.45	6.88	3.07
31-10-2027	2.92	6.48	7.90	6.10
30-11-2027	2.57	3.93	7.32	8.39
31-12-2027	3.96	3.42	7.74	5.13
	R^2 Score:1 RMSE:5.28	R^2 Score:1 RMSE:2.90	R^2 Score:1 RMSE:1.26	R^2 Score:1 RMSE:2.90

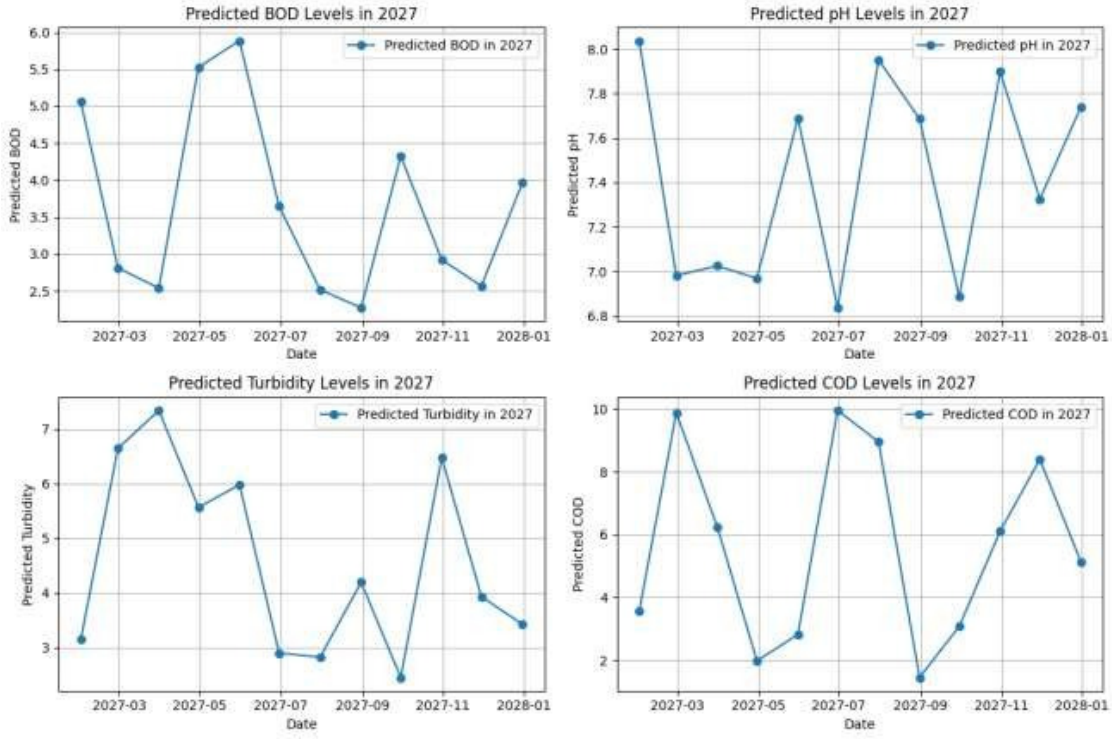


Figure 6: Predicted data of BOD, COD, pH and Turbidity of the year 2027 based on the year 2022

Table 7(a): Raw Data of year 2023

date	BOD	pH	Turbidity	COD
05-01-2023	2.3	7.87	5.57	17.64
01-02-2023	2.45	7.4	36.7	18.81
02-03-2023	2.55	7.77	8.12	22.77
04-04-2023	2.8	7.59	8.47	21.78
15-05-2023	2.7	8.18	4.79	18.18
07-06-2023	2.5	7.9	1.54	17
03-07-2023	2.8	7.87	1.76	19.8
08-08-2023	2.65	7.69	31.9	17.64
05-09-2023	2.55	7.86	21.4	18.22
10-10-2023	2.45	7.55	58.5	17.64
01-11-2023	2.3	7.5	8.45	16.66
05-12-2023	2.35	7.45	6.04	16.19

Table 7(b): Predicted data of the year 2027 based on the year 2023

date	BOD	Turbidity	pH	COD
31-01-2027	3.82	5.50	7.24	5.06
28-02-2027	2.41	1.55	7.68	7.50
31-03-2027	4.80	7.82	7.31	4.37
30-04-2027	2.84	1.09	7.79	9.43
31-05-2027	2.55	6.06	7.10	2.64
30-06-2027	3.35	5.18	8.00	5.56
31-07-2027	2.32	9.73	7.62	6.26
31-08-2027	2.47	9.29	7.27	5.40
30-09-2027	4.41	7.17	6.97	7.11
31-10-2027	5.58	2.43	7.48	7.59
30-11-2027	5.31	6.28	7.73	2.25
31-12-2027	3.43	8.51	6.82	8.56
	R^2 Score:1 RMSE:3.62	R^2 Score:1 RMSE:4.93	R^2 Score:1 RMSE:4.86	R^2 Score:1 RMSE:5.02

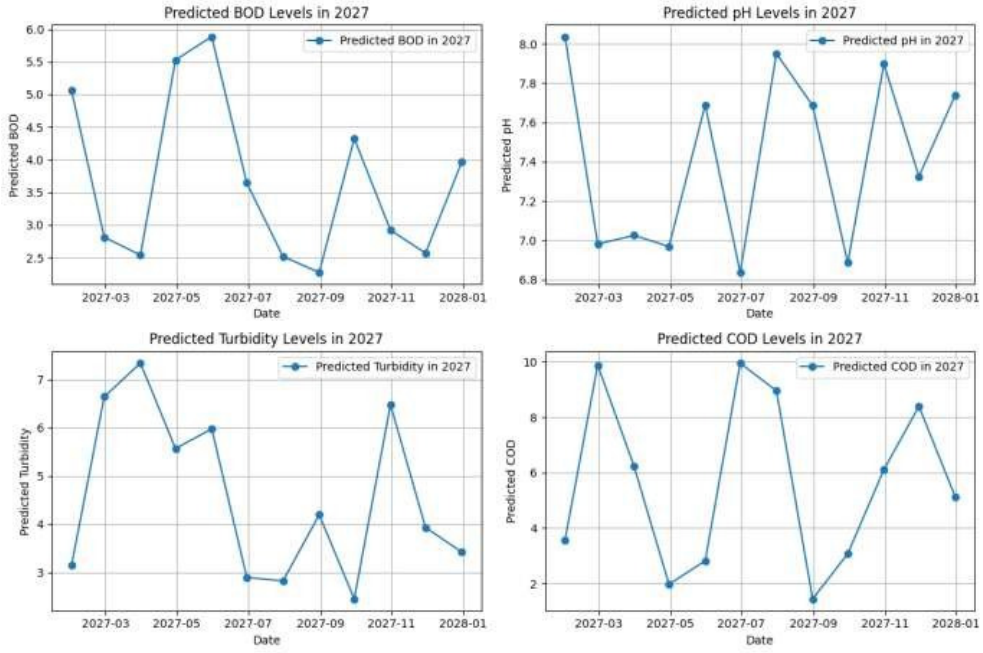


Figure7: Predicted data of BOD, COD, pH and Turbidity of the year 2027 based on the year 2023

Table 8(a): Raw Data of year 2024

date	BOD	pH	Turbidity	COD
04-01-2024	2.35	7.67	3.06	16.19
06-02-2024	2.3	7.47	7.12	17.2
04-03-2024	2.35	7.93	3.87	18.81
16-04-2024	2.85	7.96	3.94	18.47
20-05-2024	2.2	7.91	4.06	17.64
06-06-2024	2.1	7.7	6.25	16.52
10-07-2024	2.65	7.75	4.99	12.48
13-08-2024	2.45	7.55	152	10.78
02-09-2024	2.75	8.28	89.7	8.74
01-10-2024	2.3	7.44	48.8	16.51
14-11-2024	2.25	7.92	20.8	17.64
07-12-2024	1.9	7.4	63.2	17.84

Table 8(b): Predicted data of the year 2027 based on the year 2024

date	BOD	Turbidity	pH	COD
31-01-2027	4.37	1.31	8.09	1.57
28-02-2027	2.30	2.79	7.35	5.88
31-03-2027	5.15	6.29	8.66	3.69
30-04-2027	3.44	8.73	7.76	3.48
31-05-2027	1.97	2.55	7.64	9.87
30-06-2027	4.46	9.39	8.64	4.46
31-07-2027	4.18	1.87	8.79	9.60
31-08-2027	2.98	3.07	8.02	7.59
30-09-2027	5.06	7.10	8.70	3.61
31-10-2027	4.96	3.85	8.88	5.52
30-11-2027	2.32	8.86	7.70	7.80
31-12-2027	5.13	10.00	9.25	8.95
	R^2 Score: -3.19 RMSE:.46	R^2Score:.99 RMSE:.007	R^2 Score: -4.11 RMSE:.91	R^2 Score:.99 RMSE:.08

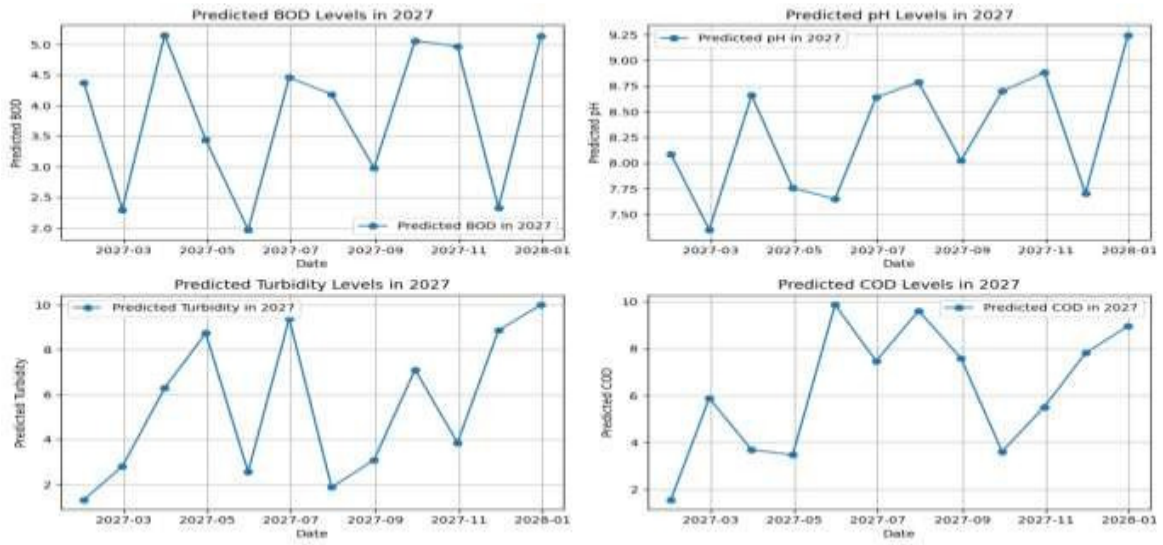


Figure 8: Predicted data of BOD, COD, pH and Turbidity of the year 2027 based on the year 2024

An interesting observation about the typical trends in BOD (Biochemical Oxygen Demand), COD (Chemical Oxygen Demand), pH, and Turbidity in the Durgapur, West Bengal region, with lowest values in March-April and highest in November. Several interconnected environmental and anthropogenic factors likely contribute to these patterns:

Lowest Values in March-April (Late Winter/Early Summer):

1. **Reduced Water Discharge:** During the drier months of late winter and early summer, the flow in rivers and other water bodies tends to be lower. This can lead to a dilution effect where existing pollutants are spread over a larger volume of water, resulting in lower concentrations of BOD, COD, and turbidity.
2. **Lower Temperatures:** Water temperatures are generally cooler during March and April compared to the later months. This has several effects:

   Reduced Biological Activity: The microbial activity responsible for the decomposition of organic matter (which consumes oxygen and contributes to BOD) is generally lower at cooler temperatures.

   Slower Chemical Reactions: Chemical reactions contributing to COD might also be slower at lower temperatures.

   Reduced Algal Growth: Warmer temperatures and increased sunlight in later months can fuel algal blooms, which contribute to organic matter and turbidity.

3. **Less Runoff from Agriculture and Urban Areas:** While agricultural activities might be ongoing, the intensity of monsoon rains and associated runoff carrying pollutants (organic matter, fertilizers, sediments) is typically minimal during this period. Similarly, urban runoff might be lower without significant rainfall.
4. **Stable pH:** pH can be influenced by biological activity and the concentration of dissolved substances. The relatively stable and less biologically active conditions in March-April might contribute to more neutral or slightly alkaline pH values.

Highest Values in November (Post-Monsoon/Early Winter):

1. **Post-Monsoon Runoff:** The monsoon season, preceding November, brings significant rainfall. This runoff carries a large load of pollutants from agricultural fields (fertilizers, pesticides, organic matter), urban areas (sewage, industrial waste), and eroded soil into water bodies. These pollutants contribute significantly to higher BOD, COD, and turbidity levels.

2. **Decomposition of Organic Matter Accumulated During Monsoon:** The organic matter washed into water bodies during the monsoon starts to decompose, leading to increased BOD and COD as microorganisms consume oxygen.
3. **Increased Suspended Solids:** The heavy rainfall and runoff during the monsoon carry a significant amount of suspended solids (silt, clay, organic debris) into water bodies, resulting in higher turbidity that can persist into November as the settling process might be gradual.
4. **Potential for Reduced Flow (Compared to Peak Monsoon):** While November is after the peak monsoon, river flows might still be substantial but potentially decreasing compared to the height of the rainy season. This could lead to a higher concentration of pollutants if the discharge of pollutants remains relatively constant.
5. **Temperature Effects (Still Relatively Warm):** While temperatures start to cool down in November compared to the summer months, they are still generally warmer than in March-April, potentially supporting continued biological activity and decomposition.
6. **Agricultural Practices:** Post-monsoon agricultural activities might also contribute some pollutants to water bodies.
7. **Industrial Activity:** Durgapur is an industrial hub. Discharge patterns from industries might also play a role, although it's less likely they would consistently peak in November and be lowest in March-April unless linked to specific seasonal production cycles or waste management practices.
8. **Sewage Discharge:** The discharge of untreated or partially treated sewage from the urban population is a continuous source of organic pollution, contributing to BOD and COD. Seasonal variations in population or water usage could slightly influence this.

Therefore, the observed trends of lowest water quality parameters in March-April and highest in November in the Durgapur region are likely a result of a combination of hydrological factors (water flow, rainfall, runoff), temperature-dependent biological and chemical processes, and the seasonal patterns of pollutant input from agricultural and urban sources, potentially exacerbated by post-monsoon decomposition and settling processes.

In case of linearity assumption linear regression assumes a linear relationship between the independent and dependent variables. Here in the above experiment we have taken so many parameters and that‘s why we obtain non-linear curve which is good agrrement with data obtained.

7. Conclusion

In this paper, water quality of the Damodar River which is flowing through Durgapur was predicted for the year 2027 by analyzing historical data from 2017 to 2024. Key water quality indicators such as Chemical Oxygen Demand (COD), Biological Oxygen Demand (BOD), pH, and Turbidity were used as input features. A Linear Regression model was employed for prediction, and the performance of the model was evaluated using Root Mean Square Error (RMSE) and R^2 score.

The model demonstrated [insert R^2 score here, e.g., "a high R^2 value, indicating strong predictive capability"] and an acceptable RMSE, suggesting that Linear Regression is a suitable method for forecasting water quality parameters in this context.

Future work may involve incorporating additional parameters or using more advanced models such as Random Forests or Neural Networks to further enhance prediction accuracy.

References

1. M N Vamsi Thalatam, Pratibha Lanka,J.N.V.R. Swarup Kumar**,**" An IoT Based Smart Water Contamination Monitoring System‖,2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), 09-11 February 2023, Coimbatore, India.

2 D. Kavitha,Gayathri T.R.,Dhamini Devaraj,Hasitha. V,‖ Survey on Water Quality Prediction‖, 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), 17-18 March 2023, Coimbatore, India.

3. M R Desai,Laxmi Shabadi,‖ Measuring Quality of Water in Real Time Environment by Using Sensors‖, 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE),16-17 December 2022,Bangalore, India.

4. Jeetendra Kumar,Rasmi Gupta,Suvarna Sharma,Tulika Chakrabarti,Prasun Chakrabarti,Martin Margala,‖IoT-Enabled Advanced Water Quality Monitoring System for Pond Management and Environmental Conservation‖,IEEE Access (Volume: 12),Page(s): 58156 – 58167, 22 April 2024 ,Electronic ISSN: 2169-3536.

5. J. O. Ighalo and A. G. Adeniyi, __A comprehensive review of water quality monitoring and assessment in Nigeria,‘‘ Chemosphere, vol. 260, Dec. 2020, Art. no. 127569, doi: 10.1016/j.chemosphere.2020.127569.

6. G. M. E. Silva, D. F. Campos, J. A. T. Brasil, M. Tremblay, E. M. Mendiondo, and F. Ghiglieno, __Advances in technological research for online and in situ water quality monitoring—A review,‘‘ Sustainability, vol. 14, no. 9, p. 5059, Apr. 2022, doi: 10.3390/su14095059.

7. V. Kothari, S. Vij, S. Sharma, and N. Gupta, __Correlation of various water quality parameters and water quality index of districts of Uttarakhand,‘‘ Environ. Sustainability Indicators, vol. 9, Feb. 2021, Art. no. 100093, doi: 10.1016/j.indic.2020.100093.

8. M. G. Uddin, S. Nash, and A. I. Olbert, __A review of water quality index models and their use for assessing surface water quality,‘‘ Ecol. Indicators, vol. 122, Mar. 2021, Art. no. 107218, doi: 10.1016/j.ecolind.2020. 107218.

9. N. H. Omer and N. H. Omer, __Water quality parameters,‘‘ in Water Quality-Science, Assessments and Policy, Oct. 2019, doi: 10.5772/INTECHOPEN.89657.

10 P. Soni and P. Singh, __A water quality assessment of arpa river under bilaspur-arpa basin area, of chhattisgarh state,‘‘ Int. J. River Basin Manage., vol. 21, no. 3, pp. 443–452, Jul. 2023, doi: 10.1080/15715124.2021.2016780.

11. Anil K Dwivedi ,‖Pollution and Environmental Assay Research Laboratory (PEARL),‖

Vol. 4, Issue 1, January 2017 ISSN: (2349-4077) Associated Asia Research Foundation (AARF)

12. Kostandina Veljanovska1 & Angel Dimoski2, ―Air Quality Index Prediction Using Simple Machine Learning Algorithms‖,2018, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS).

13. Xiaosong Zhao, Rui Zhang, Jheng-Long Wu, Pei-Chann Chang and Yuan Ze University,‖ A Deep Recurrent Neural Network for Air Quality Classification‖, 2018, Journal of Information Hiding and Multimedia Signal Processing.

14 SavitaVivekMohurle, Dr. RichaPurohit and ManishaPatil, ―A study of fuzzy clustering concept for measuring air pollution‖ index,2018, International Journal of Advanced Science and Research.

15 C R, Chandana R Deshmukh , Nayana D K and Praveen Gandhi Vidyavastu , ―Detection and Prediction of Air Pollution using Machine Learning Models‖,2018, International Journal of Engineering Trends and Technology (IJETT).

16 Zhang , Xiaoli Li & Yang Li , Jianxiang Mei, ―Prediction of Urban PM2.5 Concentration Based on Wavelet Neural Network‖,2018,IEEE.

17 Nicolás Mejía Martínez, Laura Melissa Montes, Ivan Mura and Juan Felipe Franco, ―Machine Learning Techniques for PM10 Levels Forecast in Bogotá‖,2018,IEEE.

18 J. Angelin Jebamalar& A. Sasi Kumar, PM2.5 Prediction using ―Machine Learning Hybrid Model for Smart Health‖,2019, International Journal of Engineering and Advanced Technology (IJEAT).