# Predicting and Evaluating Water Quality Using Machine Learning in Maharashtra, India

**Sanaya Kotwal**

Symbiosis International School, Maharashtra, India

**Abstract**

Water quality is critical for public health and environmental sustainability, necessitating effective monitoring to prevent contamination. In this study, we focus on predicting and evaluating water quality in Maharashtra, India, using machine learning techniques. Groundwater contamination in Maharashtra is a significant issue due to poor waste management, yet research in this area is limited. Traditional water quality monitoring methods involve complex calculations based on fixed parameters, which can lead to errors. This study aims to streamline the monitoring process by identifying the most significant features, thereby saving time, money, and energy. We calculated the Water Quality Index (WQI) using the Weighted Arithmetic Mean method, analyzing data from 2012 to 2022 from the National Water Monitoring Program in India. The analysis identified three key parameters, BOD, pH, and Fecal Coliform, as most correlated with the WQI. Machine learning techniques, including regression and classification, were employed to predict WQI and Water Quality Classification (WQC). The results indicate that Polynomial Regression and Ridge Regression achieved high accuracy in predicting the WQI, while the Decision Tree classifier excelled in WQC classification. This research demonstrates the potential of machine learning to enhance water quality monitoring, offering a cost-effective solution for managing water resources in Maharashtra.

**Keywords:** Water Quality Evaluation, Maharashtra, India, Water Quality Prediction, Supervised Machine Learning

## 1. Introduction

Water is an essential resource for life on Earth, covering two-thirds of the planet's surface. However, most of this water is not fit for human consumption or agricultural use. Therefore, monitoring water quality is crucial to ensure the safety of consumable water and to prevent the spread of waterborne diseases (Nasir et al., 2022).

Water quality is assessed using the Water Quality Index (WQI), which classifies water based on its physical, chemical, and biological characteristics. The threat of water contamination is a significant global issue, affecting millions of people. Poor water quality endangers human health by spreading diseases and posing risks to businesses and marine biodiversity. Key parameters of water quality include temperature, pH, conductivity, biological oxygen demand (BOD), nitrate, nitrite, faecal coliform, and total coliform.

Groundwater contamination, in particular, has been severely impacted by agricultural, domestic, and industrial activities. This contamination occurs when harmful substances such as chemicals, viruses, bacteria, heavy metals, and dyes are introduced into the water through human activities (Ravindiran et al., 2023). Such contamination can have detrimental effects on human health, economic growth, vegetation, and marine ecosystems.

In India, these problems are exacerbated by industrial pollution, agricultural waste, poor sanitation, and improper waste disposal. The country's water crisis results from mismanagement of water resources, insufficient rainwater collection, and excessive groundwater use. Pollution and ineffective agricultural irrigation techniques further reduce water availability and quality (Singh, 2023). The National Water Monitoring Program provides extensive data on water quality, making it a valuable resource for research and model development. Utilising this data for machine learning to predict water quality factors can significantly aid in managing water quality and pollution control.

Measuring the numerous physical, chemical, and biological factors affecting water quality is costly, time-consuming, and inefficient. Traditional methods of determining water quality involve complex calculations based on numerous fixed parameters, which are prone to errors (Abbas et al., 2024). To overcome these challenges, modern computational techniques like machine learning can be employed to identify the most influential parameters, allowing for focused and efficient monitoring.

Machine learning, a branch of artificial intelligence, is highly effective for data analysis and prediction. It can handle complex, multi-dimensional problems such as water quality analysis and predictions with high accuracy and adaptability (Zhu et al., 2022). There are four types of machine learning: supervised, unsupervised, reinforcement, and semi-supervised learning. This paper utilises supervised learning to identify the most significant contributors to the water quality index (WQI). The algorithms evaluated included classification machine learning algorithms such as SVM, Random Forest, Decision Trees, XGBClassifier, Logistic Regression, and KNN, and the regression algorithms were Polynomial Regression, Lasso Regression, Ridge Regression, Elastic Net Regression, SVM, Random Forest, and Gradient Boosting.

Research in various regions has demonstrated the accuracy of machine learning models in water quality assessment. For instance, Yafra Khan and Chai Soo See (2016) used data from the National Water Information System (NWIS) to construct artificial neural networks and time-series analysis models for evaluating water quality in Island Park village, South-Western Nassau County (Khan and See, 2016). Nida Nasir et al. (2022) employed classifiers such as SVM, Random Forest, Logistic Regression, Decision Tree, CATBoost, XGBoost, and MLP to classify water quality based on a dataset (Nasir et al., 2022). Umair Ahmed et al. applied gradient boosting and multilayer perceptron techniques to estimate the WQI in Pakistan (Ahmed et al. 2019). These studies align with the objectives of this research, which aims to develop effective predictive models for assessing water quality parameters.

The main objective of this research is to develop a machine learning model for predicting and evaluating water quality in Maharashtra, India, which will be instrumental in addressing the challenges of water contamination, resource management, and pollution control. This will involve assessing the accuracy of various machine learning algorithms, such as SVM, Random Forest, Decision Trees, XGBClassifier, Logistic Regression, KNN, Polynomial Regression, Lasso Regression, Ridge Regression, and Elastic Net. Regression, SVM, Random Forest, and Gradient Boosting, to identify the most significant factors contributing to the Water Quality Index (WQI).

This study is novel in its application of machine learning techniques to groundwater quality assessment in Maharashtra, an under-researched region facing critical contamination issues. Unlike traditional water quality monitoring methods, which are time-consuming and error-prone due to complex calculations based on numerous parameters, this research streamlines the process by identifying the most significant indicators—Biological Oxygen Demand (BOD), pH, and Fecal Coliform—using machine learning. By focusing on these key parameters, the study reduces the complexity of water quality evaluation, offering a more efficient, cost-effective solution. Additionally, it leverages a decade-long dataset (2012-2022) from the National Water Monitoring Program to provide a comprehensive analysis of water quality trends over time. The study further distinguishes itself through the use of multiple machine learning models, such as Polynomial Regression, Ridge Regression, and Decision Trees, rigorously comparing their performance in predicting Water Quality Index (WQI) and Water Quality Classification (WQC). This approach not only enhances prediction accuracy but also offers practical, scalable solutions for improving water quality management in resource-constrained regions.

## 2. Data and Methods

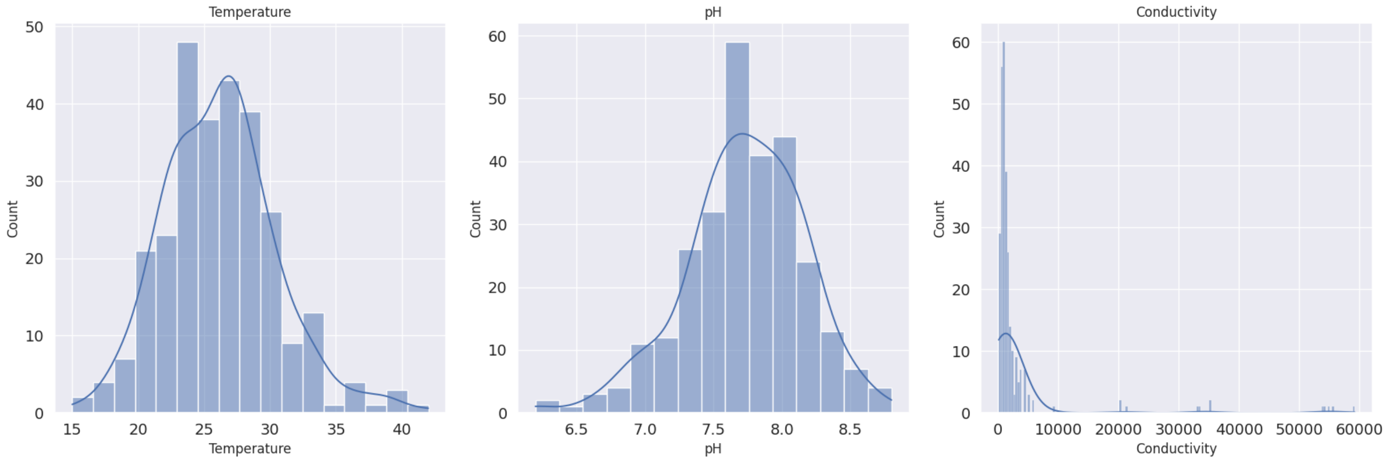
### 2.1 Dataset

The data for this study is sourced from the National Water Monitoring Program (NWMP), specifically focusing on the Maharashtra region in India. This dataset encompasses a comprehensive range of water quality parameters essential for evaluating groundwater quality in the region. The parameters include pH levels, dissolved oxygen, contaminant concentrations, temperature, biological oxygen demand (BOD), nitrate, nitrite, faecal coliform, and total coliform, among others.

This dataset, spanning from 2012 to 2022, provides a decade-long perspective on water quality trends and variations in Maharashtra. The extensive time range and diverse set of parameters make this data highly relevant for the study's objectives, allowing for a robust analysis and prediction of water quality. The detailed information on various contaminants and the physical properties of water is crucial for developing an accurate and reliable machine learning model. By leveraging this dataset, the research aims to identify the most significant factors affecting the Water Quality Index (WQI) and predict water quality efficiently, thus addressing the critical challenges of water contamination and resource management in the region.

### 2.2 Data Preprocessing

The initial step involved standardising the data from 2012 to 2022 to ensure consistency. This was achieved by converting all the PDF files into Excel format. Once converted, these Excel files were concatenated into a single comprehensive file. This consolidated Excel file was then imported into Python and processed using Pandas, a powerful data analysis library in Python. The different variables are shown in Figure 1.



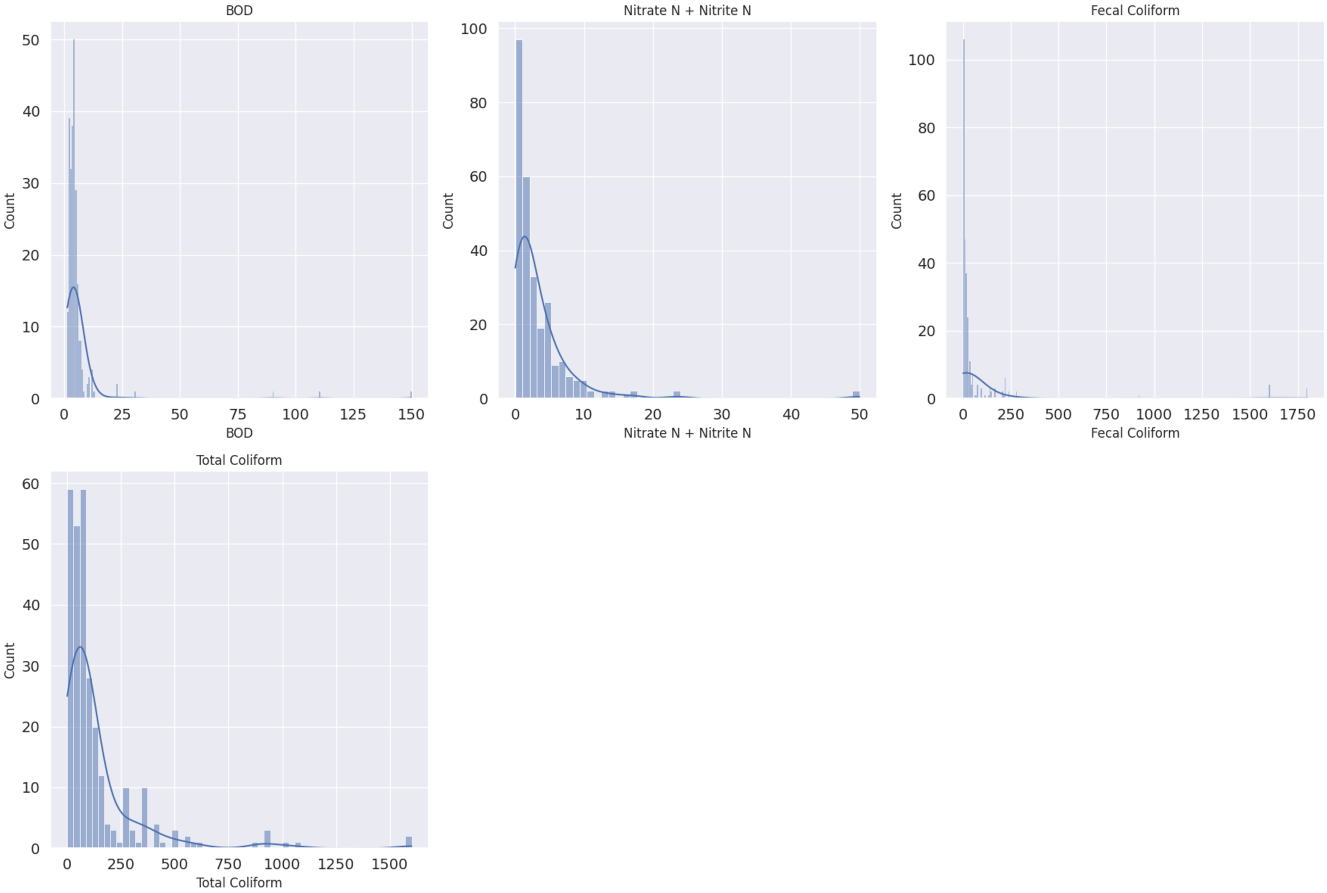
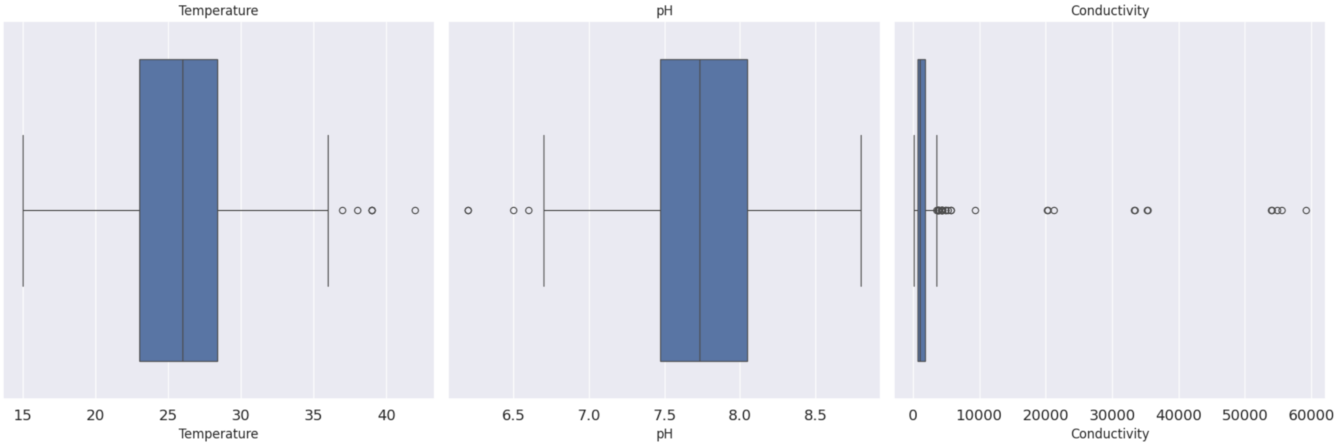


**Figure 1. Histograms for all the features**

Before training the machine learning models, the data undergoes several processing steps to ensure quality and suitability for analysis. Outliers are identified using the boxplot method as shown in figure 2 and removed to have good model performance, which improves model accuracy by eliminating data points outside the interquartile range. Standardisation, through Standard scaling, ensures that all features contribute equally to the model by scaling the data to a range between 0 and 1. Additionally, correlation analysis helps identify redundant features and understand relationships between parameters, reducing multicollinearity and enhancing model performance. These preprocessing steps ensure the dataset is clean, standardised, and ready for effective machine learning model training to predict water quality in Maharashtra. The following is the image of the boxplots for the data:



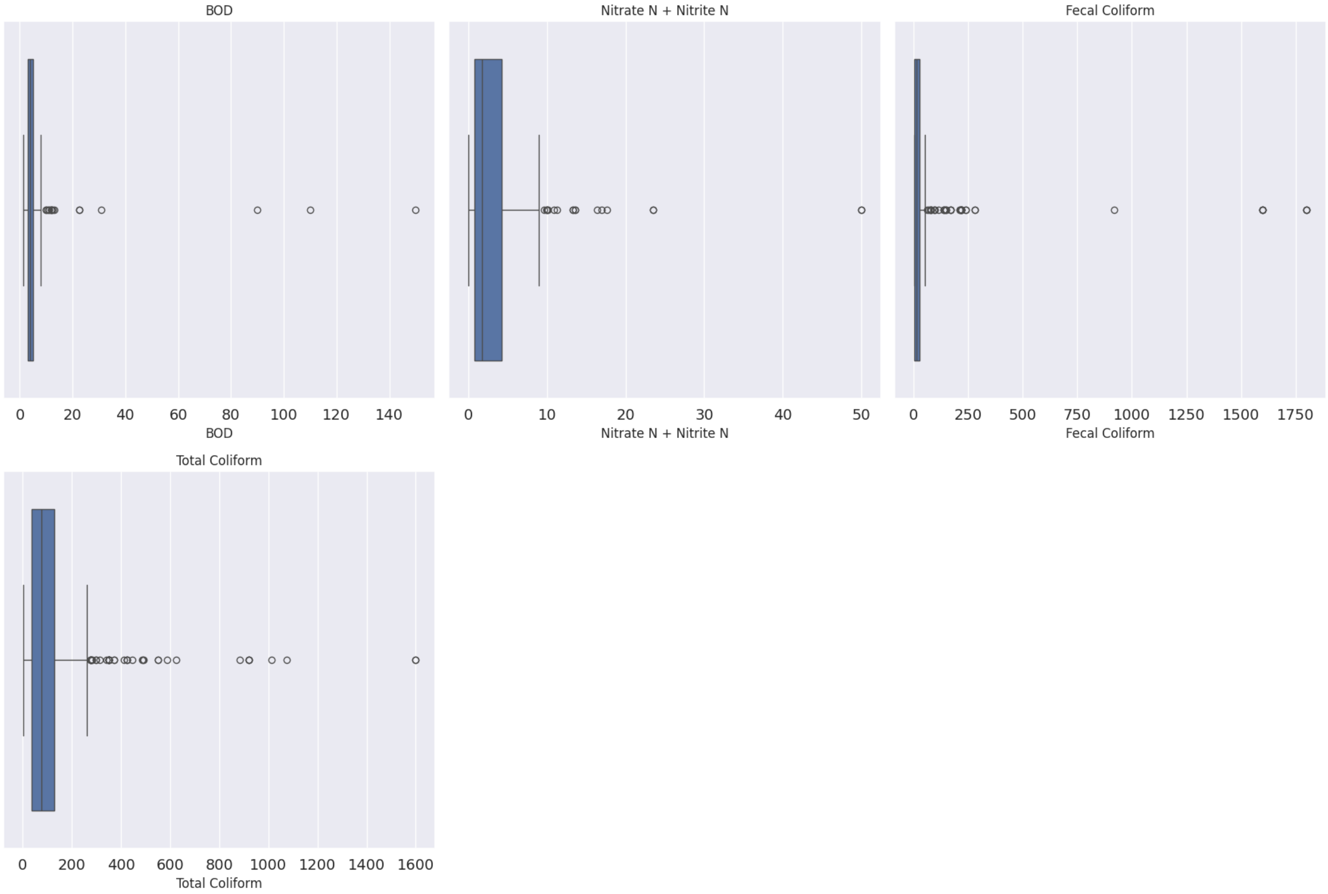
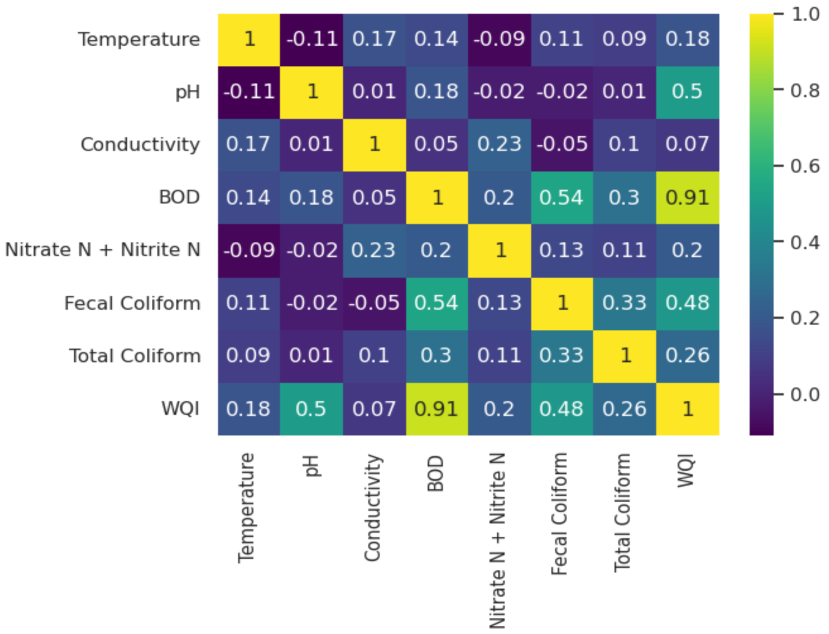


**Figure 2. Boxplots for all the features in the data**

After the data has been processed, the features that have the highest correlation with WQI have to be calculated. This was achieved by creating a covariance matrix. The covariance matrix is shown in Figure 3:



| | Temperature | pH | Conductivity | BOD | Nitrate N + Nitrite N | Fecal Coliform | Total Coliform | WQI |
|---|---|---|---|---|---|---|---|---|
| Temperature | 1 | -0.11 | 0.17 | 0.14 | -0.09 | 0.11 | 0.09 | 0.18 |
| pH | -0.11 | 1 | 0.01 | 0.18 | -0.02 | -0.02 | 0.01 | 0.5 |
| Conductivity | 0.17 | 0.01 | 1 | 0.05 | 0.23 | -0.05 | 0.1 | 0.07 |
| BOD | 0.14 | 0.18 | 0.05 | 1 | 0.2 | 0.54 | 0.3 | 0.91 |
| Nitrate N + Nitrite N | -0.09 | -0.02 | 0.23 | 0.2 | 1 | 0.13 | 0.11 | 0.2 |
| Fecal Coliform | 0.11 | -0.02 | -0.05 | 0.54 | 0.13 | 1 | 0.33 | 0.48 |
| Total Coliform | 0.09 | 0.01 | 0.1 | 0.3 | 0.11 | 0.33 | 1 | 0.26 |
| WQI | 0.18 | 0.5 | 0.07 | 0.91 | 0.2 | 0.48 | 0.26 | 1 |

**Figure 3. Covariance matrix of features. The parameters BOD, pH, and Faecal Coliform have a high correlation with WQI.**

As seen in the covariance matrix, the features with the highest correlation to WQI were BOD, pH, and Feacal Coliform. Hence, these features were chosen for the machine learning algorithms in order to accurately predict the WQI and WQC (Water Quality Class) using the least number of features.

### 2.3 WQI Calculation and WQC Classification

The Water Quality Index is calculated using a formula involving all 7 of the parameters. It can be calculated using the Weighted Arithmetic Mean method (Srivastava 2024).

$$WQI = \frac{\Sigma q_i w_i}{\Sigma w_i} \qquad (1)$$

Here, $w_i$ represents the unit weight of the i-th parameter towards overall water quality and $q_i$ is the quality estimate of the ith parameter. They can be calculated as follows:

$$q_i = \frac{V_i - V_{ideal}}{S_i - V_{ideal}} * 100 \qquad (2)$$

$$w_i = \frac{K}{S_i} \qquad (3)$$

K is the constant of proportionality and is defined by:

$$K = \frac{1}{\Sigma S_{i}\blacksquare} \qquad (4)$$

$S_i$ is the standard permissible value of the $i$-th parameter.

WQC is the Water Quality Class and is determined using the WQI. Table 1 shows the ranges of WQI and classification based on the range:

**Table 1. WQC Classification Table**

| WQI | Classification |
|---|---|
| < 50 | Excellent |
| 50-100 | Good |
| 100-200 | Poor |
| 200-300 | Very Poor |
| > 300 | Unsuitable for drinking |

### 2.4 Machine learning algorithms

This paper utilises classification and regression machine learning techniques in order to accurately predict the WQI and evaluate WQC using the parameters that have a high correlation with the WQI. The classification machine learning algorithms used in this paper are SVM, Random Forest, SDG, Decision Trees, XGBClassifier, Logistic Regression, and KNN, and the regression algorithms are Multiple Linear Regression, Polynomial Regression, Lasso Regression, Ridge Regression, Elastic Net Regression, SVM, Random Forest, and Gradient Boosting.

#### 2.4.1 Support Vector Machine

A Support Vector Machine (SVM) is a supervised algorithm used mainly to classify, but it can also be used for regression. Support Vector Machines (SVM) have been widely used in water quality prediction due to their ability to handle complex, nonlinear relationships in data. It has been used to predict water quality indices by analysing various environmental parameters such as pH, temperature, and dissolved oxygen levels (Kisi & Shiri, 2012)

### 2.4.2 Random Forest

Random Forest is a machine learning algorithm that can be used for both classification and regression. Random Forest belongs to the ensemble learning category, where multiple models are combined to solve a problem, typically resulting in improved performance compared to a single model.

### 2.4.3 Decision Tree

A Decision Tree is a supervised learning algorithm that is used for classification and regression modelling. They are easy to interpret, making them popular for both data analysis and machine learning.

### 2.4.4 XGBoost

Extreme gradient boosting, an optimised distributed gradient boosting toolkit, is a machine learning model that may be trained effectively and is scalable. By combining the predictions of several weak models, this ensemble learning technique generates a stronger prediction (Jain, 2023).

### 2.4.5 Logistic Regression

Logistic regression is one of the most popular machine learning algorithms and is a part of supervised learning. With a given collection of independent factors, it is used to predict the categorical dependent variable (Kanade 2024).

### 2.4.6 K-Nearest-Neighbours

K-Nearest-Neighbors is a non-parametric, supervised learning classifier that uses proximity to classify or predict how a single data point will be grouped. It is among the most widely used and straightforward regression and classification classifiers in machine learning today (Srivastava, 2024).

### 2.4.7 Multiple Linear Regression

Linear regression is a supervised learning technique that serves to forecast a variable's value depending on the value of another variable (Dasgupta, 2024).

### 2.4.8 Polynomial Regression

Polynomial regression is a specific kind of linear regression in which the dependent and independent variables have a curvilinear connection and the polynomial equation is fitted to the data (Stojanovski and Guide 2020).

### 2.4.9 Ridge Regression

Ridge regression is a model-tuning technique that is applied to any data that exhibits multicollinearity. This technique carries out L2 regularisation. Predicted values deviate significantly from real values when multicollinearity is present, least-squares are impartial, and variances are high (Guide and Ashok 2024).

### 5.11 Lasso Regression

Lasso regression operates on the same principles as ridge regression; the way they penalise their erroneous coefficients is the sole distinction. Lasso penalises the sum of squared coefficients but not the number of absolute mistakes (Ahmed et al., 2019). Lasso utilises L1 regularisation.

### 5.12 Elastic-Net Regression

Elastic Net regression is a type of linear regression that combines the penalties of both the Lasso and Ridge regression methods to create a more robust model. It is especially useful when dealing with datasets that have many features or when those features are highly correlated. It uses Lasso's ability to select important features and Ridge's ability to handle multicollinearity, resulting in better predictive performance (Ahmed et al., 2019).

## 3. Results

The machine learning algorithms mentioned in Section 2 have been used to predict WQI and evaluate WQC, and their performance has been evaluated. The accuracy measures to evaluate the machine learning models are described below. These evaluation metrics were used to assess the performance of both the regression and classification algorithms used to predict the WQI as well as the WQC for the dataset.

### 3.1 Accuracy Measures

#### 3.1.1 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions without considering their direction. It is calculated as the average of the absolute differences between predicted and actual values, providing a straightforward indication of model accuracy in terms of absolute units (Harrell 2001).

#### 3.1.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is the square root of the average of squared differences between predicted and actual values. This metric gives higher weight to larger errors, making it useful for identifying models that are sensitive to outliers (James et al. 2013).

#### 3.1.3 Mean Squared Error (MSE)

The Squared Error (MSE) represents the average of the squares of the errors, calculated by squaring the difference between each predicted and actual value and then averaging those squares. It is commonly used for regression tasks as it penalises larger errors more than smaller ones (Wasserman 2004).

#### 3.1.4 R-Squared ($R^2$)

R-Squared ($R^2$), also known as the coefficient of determination, indicates how well the variability of the dependent variable is explained by the independent variables in a regression model. A value closer to 1 signifies a better fit, meaning the model explains a large portion of the variance (Weisberg 2013).

#### 3.1.5 Precision

Precision is a measure of the accuracy of positive predictions, defined as the ratio of true positive results to the total predicted positives. It is crucial in contexts where the cost of false positives is high, such as medical diagnosis (Manning, Raghavan, and Schütze 2008).

#### 3.1.6 Recall

Recall, also known as sensitivity, measures the proportion of actual positives that are correctly identified by the model. It is particularly important in scenarios where capturing all positive cases is critical, such as in disease outbreak detection (Fawcett 2006).

### 3.2 Regression Results

The regression algorithms predicted the WQI based on the 3 factors most related to the WQI, found using the covariance matrix, which were BOD, pH, and Fecal Coliform, while the classification algorithms predicted the WQC using the same features. This study seeks to evaluate the efficiency of machine learning algorithms to predict WQI and WQC, eliminating the need for expensive sensors to constantly monitor these physical, chemical, and biological factors.

The regression algorithms used MSE, MAS, RMSE, and $R^2$ as evaluation metrics to find the best one.

The results for the regression algorithms are showcased below.

**Table 2. Evaluation metrics for regression algorithms**

| Regression algorithm | MSE | MAE | RMSE | $R^2$ |
|---|---|---|---|---|
| **Polynomial regression** | **3.04e-11** | **5.51e-06** | **2.46e-11** | **0.99** |
| Lasso | 12.05 | 8.27 | 3.47 | 0.90 |
| **Ridge** | **0.02** | **0.01** | **0.14** | **0.98** |
| Elastic Net | 38.41 | 18.54 | 6.19 | -1.87 |
| SVM | 22.66 | 16.81 | 4.76 | 0.00038 |
| Random Forest | 6.16 | 2.64 | 2.48 | 0.93 |
| Gradient Boosting | 4.85 | 2.23 | 2.20 | 0.95 |

As seen in the results, the Polynomial Regression algorithm performed the best with an $R^2$ value of 0.99 and the lowest errors, while the Elastic Net algorithm had the lowest performance with a negative $R^2$value, indicating that the algorithm performed extremely poorly. Most algorithms performed sufficiently well, such as Ridge, Random Forest, and Gradient Boosting, with $R^2$ values above 0.90.

### 3.3 Classification Results

The classification algorithms used accuracy, precision, and recall to assess their performance. The results for the classification algorithms are showcased below.

**Table 3. Evaluation metrics for classification algorithms**

| Classification algorithm | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| SVM | 0.93 | 0.62 | 0.63 | 0.62 |
| Random Forest | 0.92 | 0.82 | 0.85 | 0.83 |
| **Decision Trees** | **0.98** | **0.99** | **0.89** | **0.93** |
| XGBClassifier | 0.94 | 0.83 | 0.87 | 0.84 |
| Logistic Regression | 0.80 | 0.93 | 0.53 | 0.59 |
| KNN | 0.82 | 0.84 | 0.61 | 0.67 |

For the classification algorithms, the Decision Trees algorithm performed the best with an accuracy score of 0.98 and F1 score of 0.93; however, the Logistic Regression Algorithm had the least accuracy score of 0.80 and F1 score of 0.59. Almost all algorithms performed admirably, with SVM, Random Forest and XGBClassifier with an accuracy score above 0.90.

## 4. Discussion

The results of this study demonstrate the considerable potential of machine learning algorithms, particularly Polynomial Regression and Decision Trees, in accurately predicting water quality with minimal input parameters. The use of Biochemical Oxygen Demand (BOD), pH, and Fecal Coliform as primary predictors simplified the model, while still maintaining high prediction accuracy, evidenced by the $R^2$ value of 0.99 for Polynomial Regression and the 0.98 accuracy score for the Decision Tree Classifier in classifying Water Quality Classes (WQC). These results suggest that these models are well-suited for environments where comprehensive sensor networks are either unavailable or prohibitively expensive.

However, while the simplicity of using only three key parameters offers practical benefits in terms of reduced costs and model efficiency, there is a potential limitation in the exclusion of other factors that influence water quality. Variables such as Dissolved Oxygen (DO), Total Dissolved Solids (TDS), and Turbidity may provide additional insights into water quality and could be incorporated into future iterations of the model. This trade-off between model simplicity and the risk of missing critical information highlights a key challenge in environmental monitoring: achieving the balance between practicality and comprehensiveness. While the selected algorithms demonstrated robustness, the inclusion of more parameters could potentially improve the model's adaptability to different environmental conditions across diverse regions.

Another important takeaway from the results is the relative performance of machine learning models compared to conventional water quality prediction methods. Machine learning offers significant advantages in handling large datasets, modeling complex nonlinear relationships, and providing real-time predictions. This efficiency is crucial, especially in the context of Maharashtra, where water contamination is an ongoing issue and timely interventions are necessary. However, it is important to acknowledge the limitations of machine learning models, such as their dependence on high-quality data and the challenge of interpretability. In this study, models like Polynomial Regression and Decision Trees performed well, but more sophisticated methods like Elastic Net and XGBoost did not show similar results, likely due to overfitting or the specific characteristics of the dataset.

To address these limitations, integrating machine learning models with conventional monitoring approaches may offer the most robust solution. Machine learning can serve as an efficient tool for initial screening and rapid assessment, while more traditional methods can provide the granular, transparent insights needed for critical decision-making. This hybrid approach could ensure accuracy while reducing the operational costs associated with continuous water quality monitoring.

## 5. Conclusion

Water quality is critical to the health and livelihoods of millions in Maharashtra, and ensuring its safety is a top priority. Conventional water quality testing, which relies on expensive and time-consuming lab analysis, is unsustainable for long-term monitoring, especially in resource-limited settings. This study explored the use of machine learning to develop an alternative, cost-effective method for predicting water quality indices (WQI) and water quality classes (WQC) using minimal parameters.

The practical applications of this research are significant for improving water quality management in regions like Maharashtra, where contamination is a pressing issue. By utilizing machine learning models to predict the Water Quality Index (WQI) and Water Quality Classification (WQC) based on key parameters, the study offers a more efficient, cost-effective method for monitoring water quality. This approach can be implemented by local governments, environmental agencies, and water management authorities to quickly assess water safety, identify contamination hotspots, and make informed decisions about resource allocation and pollution control, ultimately safeguarding public health and ensuring sustainable water use.

The study's key findings showed that Polynomial Regression and Ridge Regression excelled in predicting WQI with $R^2$ values of 0.99 and 0.98, respectively, while Decision Tree Classifiers achieved an accuracy score of 0.98 in classifying WQC. These results highlight that a carefully selected set of predictive variables, combined with machine learning algorithms, can effectively monitor water quality without the need for complex and costly sensor networks. This research suggests that focusing on a few highly correlated water quality parameters could revolutionize the field, making continuous monitoring more feasible.

While this study offers promising insights, future research could enhance model performance by including additional water quality parameters, such as Turbidity, Dissolved Oxygen, and Heavy Metal Concentrations. Additionally, expanding the geographic scope to other regions with different water quality challenges would help validate the generalizability of the model. Incorporating real-time data acquisition systems could further increase the responsiveness of the system to dynamic changes in water quality, enabling more proactive management.

Ultimately, machine learning offers a scalable, efficient solution for water quality prediction, but the full benefits can only be realized through continuous refinement of the models and close integration with existing monitoring frameworks. Future efforts should aim to enhance the transparency of these models, ensuring that they are interpretable and trusted by environmental agencies responsible for safeguarding public health.

## Data Availability

The data used in this study is available at https://cpcb.nic.in/nwmp-data/ and the accompanying code is available at GitHub following this link:
https://github.com/sanayakotwal/WQI_Prediction_and_Evaluation.git

## Funding

This research received no external funding.

## Conflicts of Interest

There are no competing interests to declare.

## Acknowledgements

I want to thank the National Water Monitoring Program (NWMP) for providing the data. I also acknowledge the support I got from Tesfahiwet Yemane and Yamini Joshi.

## References

1. Abbas, Farkhanda, Zhihua Cai, Muhammad Shaoib, Javed Iqbal, Muhammad Ismail, Arifullah, Abdulwahed F. Alrefaei, and Mohammed F. Albeshr. 2024. “Machine Learning Models for Water Quality Prediction: A Comprehensive Analysis and Uncertainty Assessment in Mirpurkhas, Sindh, Pakistan.” *water* 16, no. 7 (March): 941. https://doi.org/10.3390/w16070941.
2. Ahmed, Umair, Rafia Mumtaz, Hirra Anwar, Asad A. Shah, Rabia Irfan, and José García-Nieto. 2019.

"Efficient Water Quality Prediction Using Supervised Machine Learning." *Water* 11, no. 11 (October): 6. https://doi.org/10.3390/w11112210.

3. Dasgupta, Atrayee. 2024. "What is Linear Regression?" Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/.
4. Fawcett, Tom. "An introduction to ROC analysis." *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.
5. Gazzaz, N.M.; Yusoff, M.K.; Aris, A.Z.; Juahir, H.; Ramli, M.F. Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. Mar. Pollut. Bull. 2012, 64, 2409–2420.
6. Guide, Step, and Prashanth Ashok. 2024. "What is Ridge Regression?" Great Learning. https://www.mygreatlearning.com/blog/what-is-ridge-regression/.
7. Harrell Jr., Frank E. "Regression Modeling Strategies." Springer Series in Statistics, Springer, 2001.
8. Jain, Sandeep. 2023. "XGBoost." GeeksforGeeks. https://www.geeksforgeeks.org/xgboost/.
9. Jain, Sandeep. 2024. "ML | Stochastic Gradient Descent (SGD)." GeeksforGeeks. https://www.geeksforgeeks.org/ml-stochastic-gradient-descent-sgd/.
10. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. "An Introduction to Statistical Learning: With Applications in R." *Springer*, 2013.
11. Kanade, Vijay. 2024. "Everything You Need to Know About Logistic Regression." Spiceworks. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/.
12. Kisi, O., and J. Shiri. "Prediction of Long-Term Monthly Groundwater Levels Using Different Artificial Intelligence Methods." *Computers & Geosciences* 43 (2012): 144-155.
13. Lohani, B. N., and Norhayati Mustapha. 2008. "Stochastic water quality index." *Environmental Technology* 3, no. 1 (December): 1-11. https://doi.org/10.1080/09593338209384157.
14. Nasir, Nida, Afreen Kansal, Omar Alshaltone, Feras Barneih, Mustafa Sameer, Abdallah Shanableh, and Ahmed Al-Shamma'a. 2022. "Water quality classification using machine learning algorithms." *Journal of Water Process Engineering* 48, no. 20 (May): 17. https://doi.org/10.1016/j.jwpe.2022.102920.
15. Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. "Introduction to Information Retrieval." *Cambridge University Press*, 2008.
16. Parra, M.; Jimenez, J.M.; Lloret, J.; Parra, L. Description of the Processing Technique for the Monitoring of Marine Environments with a Sustainable Approach Using Remote Sensing. In Water, Land, and Forest Susceptibility and Sustainability: Insight Towards Management, Conservation, and Ecosystem Services: Volume 2: Science of Sustainable Systems; Elsevier: Amsterdam, The Netherlands, 2023; Volume 2, pp. 165–188. ISBN 9780443158476.
17. Ravindiran, Gokulan, Sivarethinamohan Rajamanickam, Balamurugan K. Sathaiah, Gobinath Ravindran, Senthil K. Muniasamy, and Gasim Hayder. 2023. "A Review of the Status, Effects, Prevention, and Remediation of Groundwater Contamination for Sustainable Environment." *water* 15, no. 20 (October): 3662. https://doi.org/10.3390/w15203662.
18. Singh, Kuldeep. 2023. "Water Crisis." Water Crisis - Issues & Problems of Water Crisis in India. https://www.wateraid.org/in/blog/water-crisis.
19. Srivastava, Tavish. 2024. "Guide to K-Nearest Neighbors (KNN) Algorithm [2024 Edition]." Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-

algorithm-clustering.

20. Stojanovski, Pepi, and Step Guide. 2020. “Understanding Polynomial Regression!!! | by Abhigyan | Analytics Vidhya.” Medium. https://medium.com/analytics-vidhya/understanding-polynomial-regression-5ac25b970e18.
21. Wang, X., Zhang, F. & Ding, J. Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. Sci Rep 7, 12858 (2017). https://doi.org/10.1038/s41598-017-12853-y
22. Wasserman, Larry. "All of Statistics: A Concise Course in Statistical Inference." Springer Texts in Statistics, *Springer*, 2004.
23. Weisberg, Sanford. "Applied Linear Regression." Wiley Series in Probability and Statistics, Wiley, 2013.
24. Y. Khan and C. S. See, "Predicting and analysing water quality using Machine Learning: A comprehensive model," *2016 IEEE Long Island Systems, Applications, and Technology Conference (LISAT)*, Farmingdale, NY, USA, 2016, pp. 1-6, doi: 10.1109/LISAT.2016.7494106.
25. Zhu, Mengyuan, Jiawei Wang, Xiao Yang, Yu Zhang, Linyu Zhang, Hongqiang Ren, Bing Wu, and Ling Ye. 2022. “A review of the application of machine learning in water quality evaluation.” *Eco-Environment & Health* 1, no. 2 (June): 107-116. https://doi.org/10.1016/j.eehl.2022.06.001.