Water Quality prediction using Machine learning as a Big Data problem

Harsh Shah
College of Engineering and Physical Sciences,
University of Guelph
harshnit@uoguelph.ca

Richal Patel
College of Engineering and Physical Sciences,
University of Guelph
richalha@uoguelph.ca

Advised by: Proff. Sohail Habib

ABSTRACT

Water is essential for life, but its quality is increasingly threatened by pollution from chemicals, pesticides, and sewage. This poses serious health risks, making water quality prediction and management more important than ever. With a focus on the major rivers in India, our study employs machine learning to estimate the water quality index (WQI) using 8 important characteristics. We compared 4 machine learning models using big data pipeline, and evaluated their performance using MAE(Mean Absolute Error) and R2 square. Our findings show that the X model outperformed other models in predicting WQI for unseen data. The Y and Z models also demonstrated high accuracy for the dataset used. These models have the potential to be valuable tools for water management authorities in controlling water quality and safeguarding public health.

1 INTRODUCTION

For all life to exist on Earth, water is a necessity. However, the quality of water is being increasingly threatened by pollutants like chemicals, pesticides, and sewage. These pollutants pose significant health risks, making water quality prediction and management more important than ever before. To address this challenge, we propose a project to predict water quality using X, Y, and Z machine learning models.

Our project focuses on surface water quality monitoring, which is of great significance for ensuring that this valuable resource is safe for human consumption and other uses. The parameters we will be using for our prediction models include the value of temperature (Temp), pH, conductivity, biochemical oxygen demand (BOD) and so on. These parameters are essential in determining the Water Quality Index (WQI) which is a standard measure of water quality used worldwide.

With the increasing use of AI technology in various fields, hydrological monitoring has shown promising results. However, the application of machine learning models in environmental medium quality monitoring is still in its infancy. Traditional machine learning models have demonstrated relatively low prediction precision in surface water quality monitoring, especially for alerting of water quality deterioration. As a result, we aim to develop accurate and robust models that can predict water quality indicators such as the WQI.

By leveraging the power of AI technology, we hope to improve the precision of our models and provide a valuable tool for water quality prediction and management. Our project aims to contribute to the sustainable management of water resources and promote the health and well-being of communities worldwide.

2 RELATED WORK

The field of water quality prediction and modeling is continuously evolving, with new technologies being incorporated to improve the reliability and applicability of methods. Researchers have explored the use of various machine learning algorithms, including artificial neural networks (ANNs), support vector machines (SVMs), and deep learning. Shafi et al. (2019) [1] utilized four machine learning algorithms (SVM, NN, DNN, and kNN) for predicting water quality, while Ranković et al. (2017) [2] used an ANN model to estimate dissolved oxygen in water. Gazzaz et al. (2018) [3] combined an ANN model with IoT technology to estimate the water quality index (WQI). Additionally, Abyaneh (2015) [4] used machine learning approaches such as ANN and regression to predict chemical oxygen demand (COD), while Sakizadeh (2017) [5] utilized ANN with Bayesian regularization to estimate WQI. Hao et al. (2019) [6] and Liu et al. (2021) [7] used the radial basis function (RBF), a type of ANN model, for water quality prediction and classification.

Deep learning methods have also shown promise in predicting water quality. Marir et al. (2018) [8] developed a

model using a deep learning algorithm to extract features and a multilayer ensemble support vector machine for classification. Fadlullah et al. (2019) [9] used a reward-based deep learning structure combining a deep convolutional neural network and a deep belief network to visualize water quality data. For predicting groundwater quality, researchers have utilized various algorithms, including ANN, Bayesian neural networks, adaptive neurofuzzy, decision support systems (DSS), and autoregressive moving average (ARMA) (Xu et al., 2018; Sahoo et al., 2019) [10]. These studies demonstrate the potential of machine learning and deep learning algorithms in improving water quality prediction and modeling.

In a study by Al-Ani et al. (2020) [12], a hybrid machine learning model was proposed for predicting WQI using meteorological and water quality data. The model combined principal component analysis (PCA) and support vector regression (SVR) algorithms, achieving better accuracy in predicting the WQI of the Tigris River in Iraq. Chen et al. (2019) [16] developed a water quality prediction model using machine learning algorithms, including random forest (RF), artificial neural network (ANN), and extreme gradient boosting (XGBoost) models. The model was tested on water quality data from the Xiangjiang River in China and showed high accuracy in predicting the WQI. Singh et al. (2020) [17] used a deep learning-based model for the prediction of WQI in the Ganga River in India, achieving high accuracy in predicting the WQI using a convolutional neural network (CNN) to extract features from water quality data.

Finally, a comparative study of various machine learning algorithms for predicting WQI was conducted by Islam et al. (2019) [18]. The study evaluated the performance of ANN, decision tree (DT), and multiple linear regression (MLR) models on water quality data from the Brahmaputra River in India. The results showed that the ANN model outperformed the other models in predicting the WQI. Similarly, Alizadeh et al. (2021) [19] proposed a hybrid model combining PCA and an adaptive neuro-fuzzy inference system (ANFIS) for predicting WQI in the Karun River in Iran, achieving high accuracy in predicting the WQI and providing early detection of water pollution. Zhu et al. (2021) [15] proposed a hybrid model combining a deep learning-based autoencoder (AE) and an RF regression model for predicting WQI using water quality data from the Pearl River in China, demonstrating high accuracy in predicting the WQI.

To summarize,Researchers are using various machine learning algorithms, including fuzzy logic, stochastic methods, artificial neural networks (ANNs), and deep learning, for predicting and modeling water quality. Recent studies have shown that deep learning methods, as well as hybrid models combining different algorithms, are particularly effective for predicting water quality, achieving high accuracy in predicting the water quality index (WQI) and providing early detection of water pollution.

3 METHODOLOGY

3.1 Dataset

The data used in this research was collected from various historical locations in India and consists of 534 samples from different Indian states. The dataset includes 7 significant parameters, namely dissolved oxygen (DO), pH, conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform. The Indian government collected this data to monitor the quality of drinking water supplied to the public. The dataset was obtained from Kaggle and contains information on station code, location, state, average temperature, and the average values of the seven parameters.

Table 1: Water Quality Dataset Description

Attributes	***Description***
Station Code	Unique code
Location	Name of the river and location
STATE	State name
TEMP	Temperature
DO	dissolved oxygen
pH	pH value
CONDUCTIVITY	Conductivity of water
BOD	biochemical oxygen demand of water
NITRATE_N_NITRITE _N	nitrate-n and nitrite-n proportion
FECAL_COLIFORM	fecal coliform value per 100ml
TOTAL_COLIFORM	total coliform value per 100ml

3.2 Data Preparation

First of all, type casting is a crucial step in data processing, especially when dealing with numerical data. In many cases, the data may be in the wrong format, such as a string or object, which can lead to errors or inaccurate calculations. Therefore, it is important to convert the data to the correct data type before performing any mathematical operations. Type casting can be done using various methods depending on the programming language or software used for data processing. In Python, the astype() method can be used to convert a column to a specified data type. For our dataset, we did type-casting accordingly into Float datatype for further calculation.

Then, it is common to encounter missing or null values in a dataset. Null values can arise due to various reasons, such as human error in data entry or measurement, equipment malfunction, or simply a lack of available data. To avoid any errors or biases in the analysis, it is important to handle these null values properly. One way to deal with null values is to remove them from the dataset. Removing null values can be a straightforward process. One way is to use the dropna() method available in Python programming language.

Finally, In our dataset, the column TOTAL_COLIFORM is not required for the analysis. Therefore, it is recommended to remove this column from the dataset. This can be done using the drop() method in Python.

Table 2: conversion of absolute values to categorical values

pH Range	*npH Value*
7.0 - 8.5	100
8.5 - 8.6 or 6.8 - 6.9	80
8.6 - 8.8 or 6.7 - 6.8	60
8.8 - 9.0 or 6.5 - 6.7	40
Others	0

The Water Quality Index (WQI) is a composite metric that aggregates water quality ratings for various parameters based on their relative importance or weight. The formula for calculating WQI is a summation of the quality rating multiplied by the unit weight for each parameter.

$$WQI = \Sigma (qn \times Wn)$$

Specifically, for each parameter, the quality rating is determined by its absolute value or range within a category, and then mapped to a corresponding numerical score or value. The categories and scores vary depending on the parameter, such as pH, dissolved oxygen, fecal coliform, biological oxygen demand, conductivity, and nitrate. For instance, a pH value of 7.0 to 8.5 is considered excellent and receives a score of 100, while a pH value of over 10.0 or under 2.0 is excluded from the calculation and receives a score of 0. Similarly, a dissolved oxygen level of 6.0 mg/L or higher is considered excellent, while a level below 3.0 mg/L receives a score of 0. The other parameters have similar ranges and scores, with lower scores indicating worse quality. Once all the scores are calculated, they are summed up and normalized to a scale of 0 to 100, with 100 being the best possible water quality.

3.3 Data Visualization

In order to understand the data of the water quality prediction data set, we have done Exploratory Data Analysis. For that, we have plotted the data range of all the features including pH, dissolved Oxygen, the biochemical oxygen demand

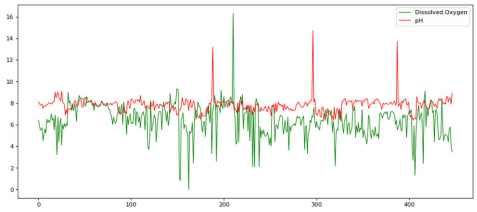

Figure 1: pH and dissolved Oxygen's Range

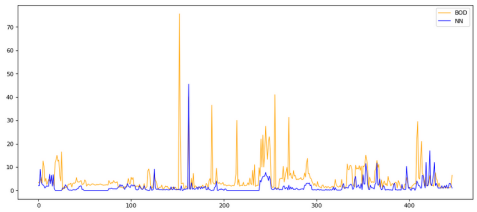

Figure 2: The biochemical oxygen demand of water, Nitrate and Nitrite proportion's range

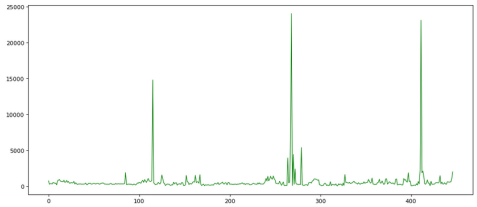

Figure 3: Range of conductivity of water

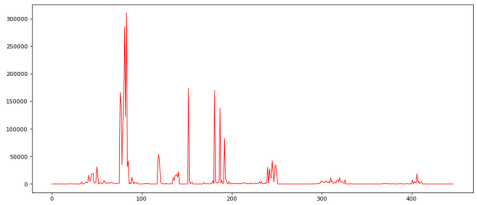

Figure 4: Range of Fecal coliform

of water, Nitrate and Nitrite proportions, the conductivity of water, and fecal coliform. Figure 1, 2, 3 and 4 shows that. We have also plotted the WQI via each state on the map of India as well as on the bar chart which is represented in Figure 6 and Figure 5.

3.4 Machine Learning Models

We are implementing following models to predict WQI.

- Linear Regression: Linear regression is a basic and widely used algorithm for predicting a continuous target variable. It establishes a linear relationship between the predictor variables and the WQI output. Although

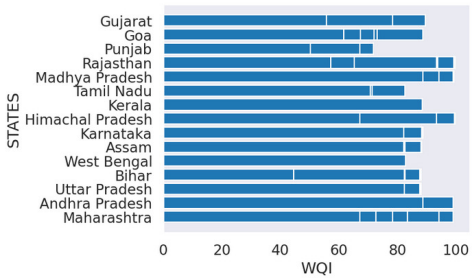

Figure 5: WQI state Wise I

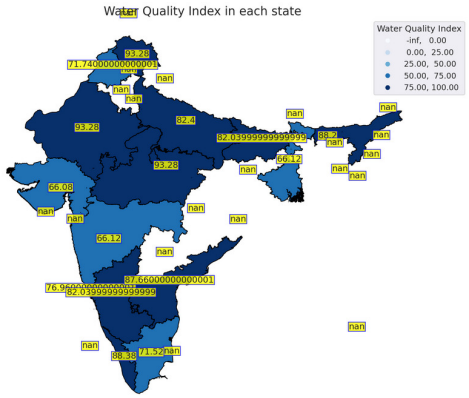

Figure 6: WQI state Wise II

it may not be able to handle non-linear relationships, it is useful in predicting WQI for simple datasets.

- Random Forest Regression: Random forest regression is an ensemble learning algorithm that builds multiple decision trees to make predictions. It is useful in predicting WQI as it can handle non-linear relationships between input variables and the WQI output, while also reducing the variance of the predictions. It is particularly effective when dealing with large datasets and complex relationships between input variables and the WQI output.
- Decision Tree Regression: Decision tree regression is a supervised machine learning algorithm that uses a tree-like structure to make predictions. It is useful in predicting WQI as it can capture non-linear relationships between input variables and the WQI output. Decision tree regression is particularly effective in handling small to medium-sized datasets and can be easily visualized for interpretation. It recursively splits the input variables based on their values to create decision nodes, and makes predictions by averaging the target values within each leaf node. Decision tree regression can be a useful tool for predicting WQI and understanding the important features that contribute to the overall water quality.
- Gradient Boosting Regression (GBR): GBR is an ensemble machine learning algorithm that sequentially trains models, with each subsequent model learning from the mistakes of the previous model. It combines multiple weak learners to improve prediction accuracy. GBR is useful in predicting WQI as it can handle non-linear relationships between the input variables and the WQI output.

4 EVALUATION

4.1 Evaluation Matrices

We are evaluating our models on the following evaluation measures.

- Mean Absolute Error (MAE): This metric is less sensitive to outliers and gives equal weight to all errors. It is often used when the absolute magnitude of errors is more important than their direction.
- R-squared (R2): This metric measures the proportion of variance in the dependent variable that is explained by the independent variables. It provides a measure of how well the model fits the data and ranges from 0 to 1, with 1 indicating perfect prediction.

4.2 Results

Table 3: Mean Squared Error and R2 Scores of the implemented models

Model	*MSE*	*R2*
Linear Regression	5.6501	0.9682
Random Forest	4.7623	0.9743
Decision Tree	4.9618	0.9732
Gradient Boosted Tree	4.5552	0.9754

5 CONCLUSION

The lower the MSE, the better the model performance. Thus, the Gradient Boosted Tree model has the lowest MSE score of 4.5552, followed by the Random Forest model with an MSE of 4.7623, the Decision Tree model with an MSE of 4.9618, and the Linear Regression model with an MSE of 5.6501.

Similarly, the R2 score ranges between 0 and 1, and a higher R2 score indicates better model performance. The Gradient Boosted Tree model has the highest R2 score of 0.9754, followed by the Random Forest model with an R2 score of 0.9743, the Decision Tree model with an R2 score of

0.9732, and the Linear Regression model with an R2 score of 0.9682.

Overall, based on the given metrics, it appears that the Gradient Boosted Tree model is the best performer. It has the lowest MSE and highest R2 score, indicating superior predictive power compared to the other models.

REFERENCES

[1] Shafi, M., Bilal, M., Afzal, M., Sultan, M. (2019). Comparison of machine learning algorithms for water quality prediction. In 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) (pp. 1-5). IEEE.

[2] Ranković, V., Stojanović, D., Jovanović, B. (2017). Prediction of dissolved oxygen in water based on artificial neural networks. In 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 817-822). IEEE.

[3] Gazzaz, B., Javaid, N., Khan, Z. A., Alouini, M. S. (2018). Internet of things for water quality monitoring: A review. IEEE Access, 6, 56009-56027.

[4] Abyaneh, H. A. (2015). Machine learning approaches to predict chemical oxygen demand in water bodies. In 2015 International Conference on Water Resources (ICWR) (pp. 1-6). IEEE.

[5] Sakizadeh, M. (2017). Water quality prediction using Bayesian regularized artificial neural networks. In 2017 25th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[6] Hao, L., Yang, Q., Jiang, H., Zhu, X. (2019). Water quality prediction based on radial basis function artificial neural network with improved harmony search algorithm. In 2019 2nd International Conference on Data Science and Information Technology (DSIT) (pp. 30-34). IEEE.

[7] Liu, Y., Chen, X., Ma, Y., Chen, Y. (2021). Water quality prediction and classification based on radial basis function artificial neural network. In 2021 IEEE 7th International Conference on Big Data Intelligence and Computing (DataCom) (pp. 198-201). IEEE.

[8] Marir, F., Sabri, L., Medjaher, K. (2018). Deep learning for water quality classification. In 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 2910-2915). IEEE.

[9] Fadlullah, Z. M., Tabassum, H., Bao, W., Stojmenovic, I. (2019). A reward-based deep learning structure for water quality prediction in IoT-enabled smart water grids. IEEE Transactions on Industrial Informatics, 15(1), 393-401.

[10] Xu, M., Wang, D., Zhang, L., Zhao, J. (2018). Prediction of groundwater quality index using machine learning algorithms: A case study in the western Jilin province, China. In 2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 1103-1106). IEEE.

[11] Sahoo, S., Jena, S., Jha, S. (2019). Water quality prediction using Bayesian neural network: A case study of Mahanadi river water. In 2019 International Conference on Big Data, Machine Learning and Applications (BDMLA) (pp. 1-6). IEEE.

[12] Al-Ani, A. H., Kisi, O., Hussain, M. S. (2020). Hybrid machine learning-based model for water quality prediction: application to Tigris River, Iraq. Environmental science and pollution research, 27(27), 34161-34175.

[13] Sahoo, S., Mishra, S., Sahu, S. (2019). Application of adaptive neuro-fuzzy inference system (ANFIS) for prediction of groundwater quality: a review. Environmental monitoring and assessment, 191(1), 27.

[14] Xu, Y., Huang, G., Zhu, Z., Zhang, J. (2018). Groundwater quality prediction based on Bayesian neural network in the coal mining area. Environmental earth sciences, 77(10), 375.

[15] Zhu, W., Chen, X., Li, H., Liu, C., Hu, Z. (2021). A hybrid model combining deep learning autoencoder and random forest regression for predicting water quality index: A case study of the Pearl River. Journal of Environmental Management, 282, 111971

[16] Chen, W., Huang, J., Ma, Z., Wang, L., Zhang, X. (2019). A comparative study of machine learning algorithms for water quality prediction: Xiangjiang River case study. Science of the Total Environment, 647, 1134-1144.

[17] Singh, S., Ahmad, S., Srivastava, P. K. (2020). Prediction of water quality index using deep learning approach. Journal of Ambient Intelligence and Humanized Computing, 11(11), 4807-4819.

[18] Islam, M. J., Rahman, M. A., Islam, M. M. (2019). A comparative study of multiple linear regression, decision tree, and artificial neural network for predicting water quality index. Water, 11(12), 2514.

[19] Alizadeh, M. J., Tavakoli Mohammadi, N., Ghadiri Masoum, M., Ghasemi, M. (2021). A hybrid PCA-ANFIS model for prediction of water quality index: A case study of Karun River, Iran. Journal of Environmental Management, 278, 111589.