Prediction of Water Quality Index Using Machine Learning

1. Introduction

Water is essential for human well-being and socio-economic development [1]. According to the 2021 World Water Development Report by UNESCO, global freshwater usage has increased sixfold in the last century and continues to grow by 1% annually [2]. This increase is driven by factors like population growth, economic development, and evolving consumption habits [2]. However, this increased water consumption poses significant challenges in maintaining water quality. Activities such as industrialization, agriculture, and urbanization have led to environmental degradation and water pollution, with profound implications for human health and sustainable development [3]. Thus, it's imperative to monitor and assess water quality rigorously to protect both human health and the environment, facilitating effective and sustainable water management [3].

Water Quality Index (WQI) is a mathematical model by converting large quantities of water quality data into a single value, making it easier to understand and monitor the overall quality of water resources [4, 5]. The WQI comprises four processes that include parameter selection, sub-Index calculation, weighting factor determination, and sub-Indices aggregation. First process involves choosing which water quality parameters to include in the evaluation. Once the parameters are selected, actual water quality data is collected. For each chosen parameter, the concentrations found in the water samples are converted into a single numerical value. These values help create sub-indices, which are essentially scores or ratings for each parameter. Third, weighting factor is assigned to each parameter to reflect their significance. Parameters with more impact on water quality receive higher weights in the calculation. Finally, an aggregation function combines all the sub-indices by using the assigned weighting factors for each parameter. This process generates a final single numerical value - the WQI. [6-9].

WQI might not provide an entirely accurate assessment of water quality. The reason being that each index is typically designed for certain locations or types of water, making it biased towards those specific conditions. Additionally, these indices can be highly sensitive to specific parameters concentrations or heavily reliant on the assigned weights for these parameters, which can affect the overall assessment of water quality [4, 6]. To address these problems, there's a critical need for an alternative approach that ensures both computational efficiency and accuracy in estimating the WQI [10]. In recent years, machine learning (ML) techniques provides an alternate method to estimate the WQI based on existing data [4, 11].

The primary goal of this study is to develop an efficient method for predicting the WQI, a single metric that encompasses both the Water Quality Classification (WQC) and the overall quality of water. Our aim is to streamline the process of water quality assessment, thereby saving time and resources typically required for manual measurements. We achieved this by utilizing regression learning techniques such as Linear Regression, Random Forest Regression, XGBoosting regression. These techniques applied to predict WQI based on key parameters including dissolved oxygen, nitrite, nitrate, total nitrogen, total phosphorus, total suspended solids, biological oxygen demand, and turbidity.

2. Data Wrangling

The original water quality dataset was loaded from a CSV file into a Pandas DataFrame named q_data. It contains 18148 rows and 48 columns with varying data types, missing values, and unclear column names. To prepare for WQI calculation, relevant columns were selected and renamed for clarity. Numeric conversions were applied to certain columns, such as 'TP (mg/L)', to facilitate analysis. Additionally, adjustments were made to the 'Date' column, converting it to datetime format for temporal analysis.

The modeling process specifically targeted river water quality, so the analysis was streamlined to exclusively include records pertaining to river water samples, which consist of 1068 rows and 15 columns. Addressing missing data, calculations were performed to determine the number and percentage of missing values for each column. Columns with a high percentage of missing values, such as 'Chla (mg/L)' (76.50%), were removed from the dataset to enhance data reliability. Furthermore, zero values in the 'Temp (°C)' and 'pH' columns were replaced with NaN to rectify potential inaccuracies. Erroneous TN, TP, and Turbidity values were identified and corrected, bolstering the dataset's accuracy. Imputing NaN values with the median for selected columns further optimized data quality. Distribution of data before and after processing is provided in Figure 1 and 2. Figure 1 illustrates the distribution of data before processing, where no adjustments were made to missing values or outliers. Figure 2 displays the distribution of data after preprocessing, where missing values were handled, outliers were addressed, and necessary transformations were applied.

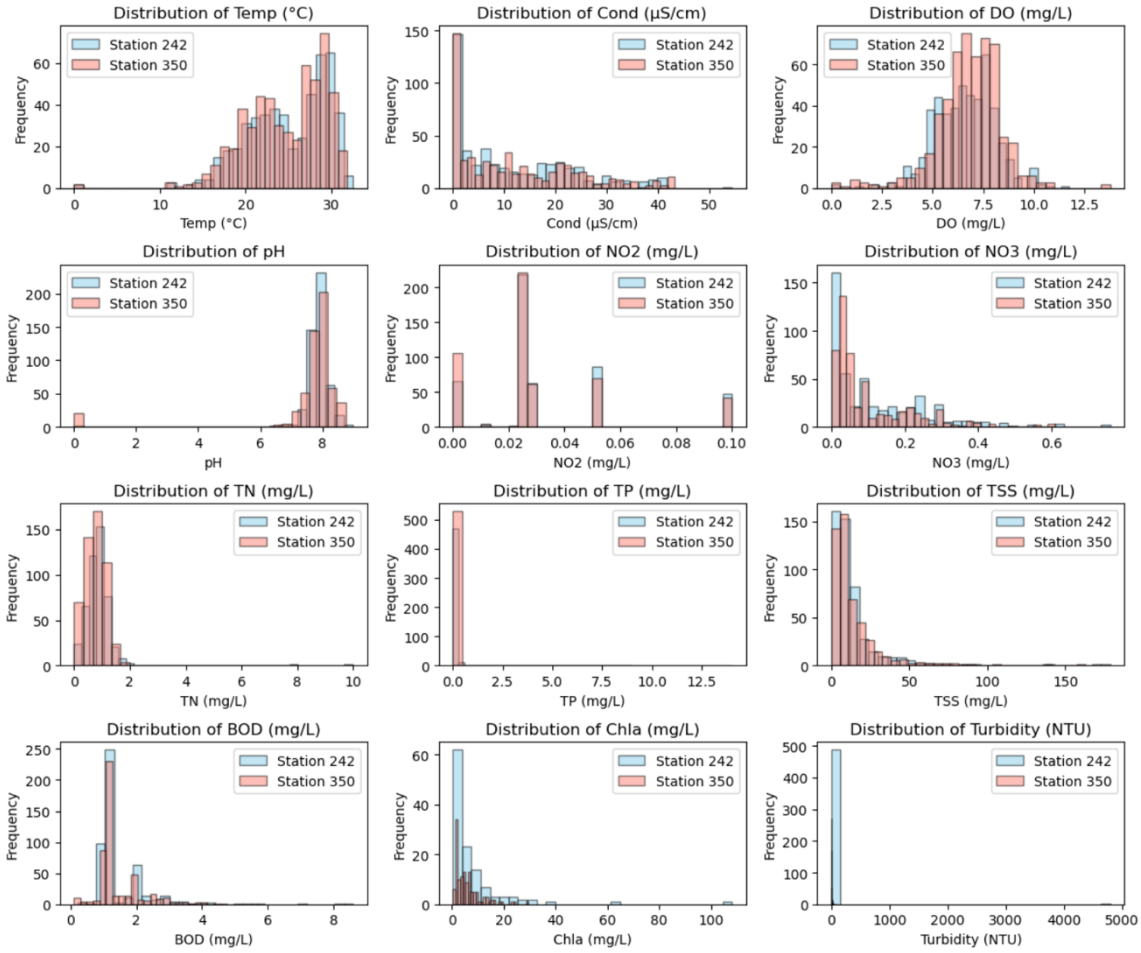

Figure 1. Distribution of data before processing.

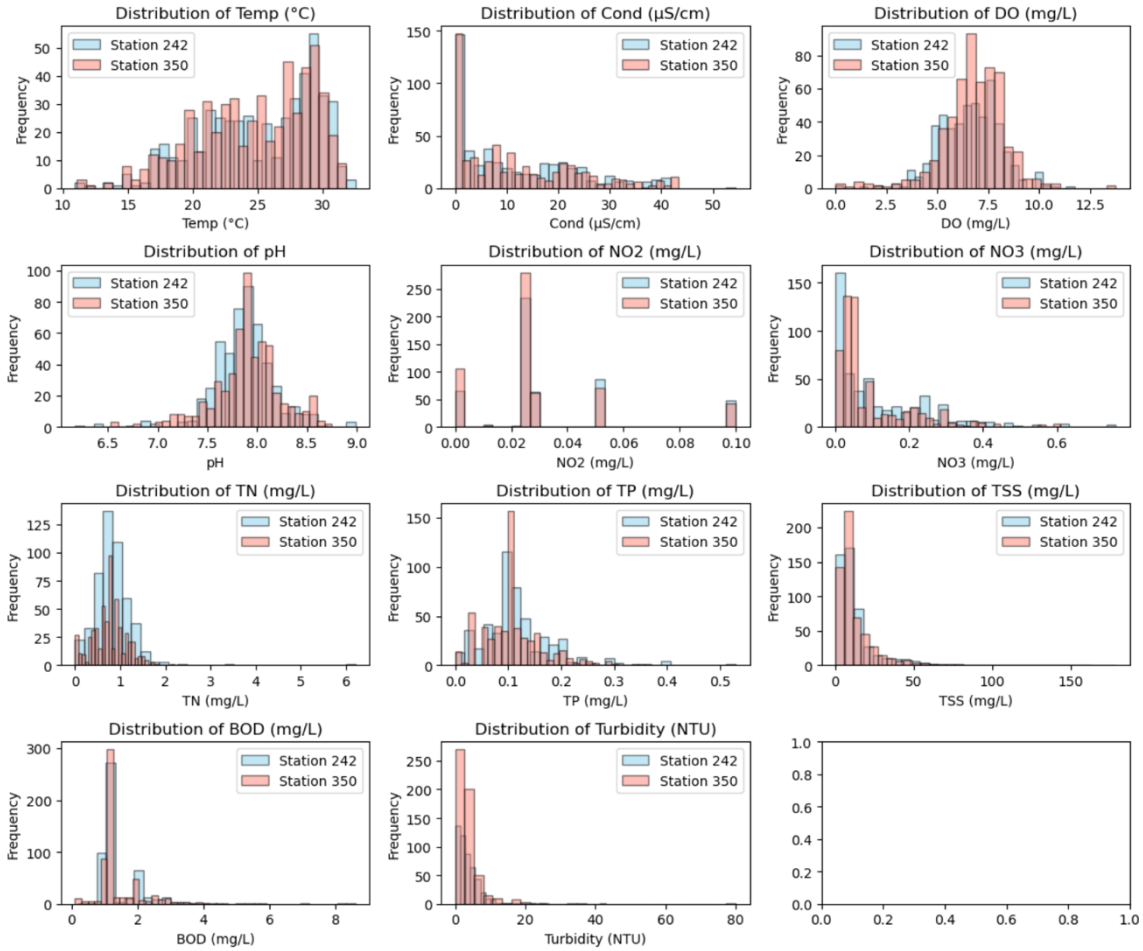

Figure 2. Distribution of data after processing.

The WQI was computed using weighted arithmetic methods, incorporating standard desirable and ideal values for various parameters. This facilitated the quantification of water quality for each station, providing valuable insights into overall water health. Classification functions were applied to categorize WQI into distinct water quality classes, aiding in interpretation and decision-making.

Lastly, the processed data, focusing on river water samples, was stored in a CSV file, named "water_quality_index.csv," ensuring accessibility for future analyses.

3. Exploratory Data Analysis (EDA)

In the EDA phase, duplicate records were first identified and removed based on a subset of columns relevant to water quality parameters, such as conductivity, dissolved oxygen, pH, nitrite, nitrate, total nitrogen, total phosphorus, total suspended solids, biological oxygen demand, and turbidity. Following the removal of duplicate values, the DataFrame wqi

comprises of 1025 rows and 14 columns. Additionally, outliers were observed in certain numerical features, indicating potential errors in data collection or measurement.

Subsequently, the distribution of numerical features was explored through histograms and kernel density estimation (KDE) plots (Figure 3). Skewed distributions were observed in several features, suggesting a concentration of values on one side of the distribution. Moreover, multimodal distributions were identified in certain features (NO2), indicating the presence of distinct groups or clusters within the data. By examining the KDE of the WQI, it was evident that the distribution of WQI values was right-skewed, with the majority of data concentrated between 35 and 56. The KDE also revealed that the range of WQI values spanned from 12.5 to 343, indicating a wide variability in water quality across the dataset. Additionally, the mean WQI value was approximately 51, suggesting a moderate overall water quality level.

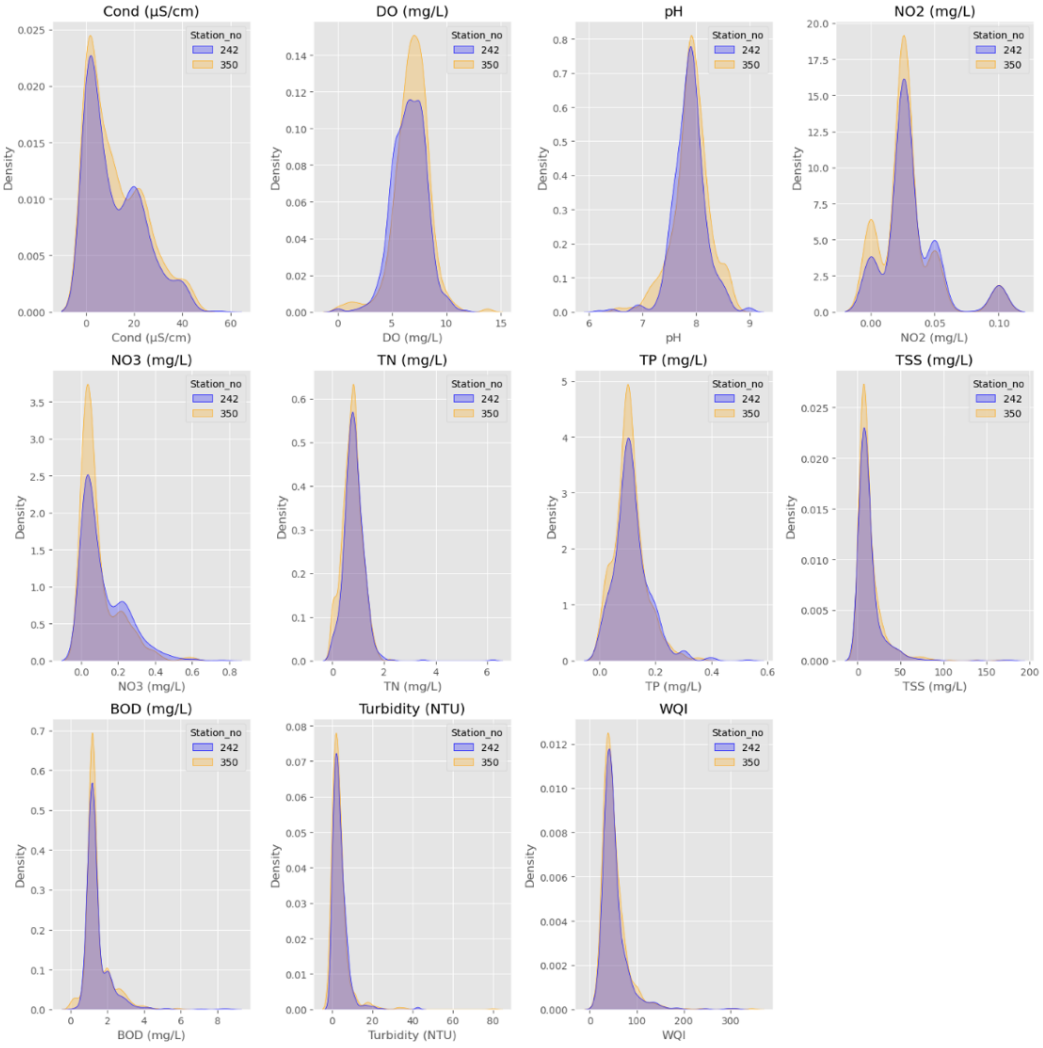

Figure 3. KDE plots for numerical features for each station.

To understand the relationships between numerical features and the WQI, pairplots, scatterplots, and regression plots were generated. Additionally, the correlation heatmap provides an overview of the relationships between different numerical features in the dataset (Figure 4). WQI has a very strong positive correlation with TSS and a moderate positive correlation with Turbidity, indicating that higher values of these parameters are associated with poorer water. The correlations with other variables are weak (Cond, pH, NO2, NO3, BOD) or very weak (DO, TN, TP).

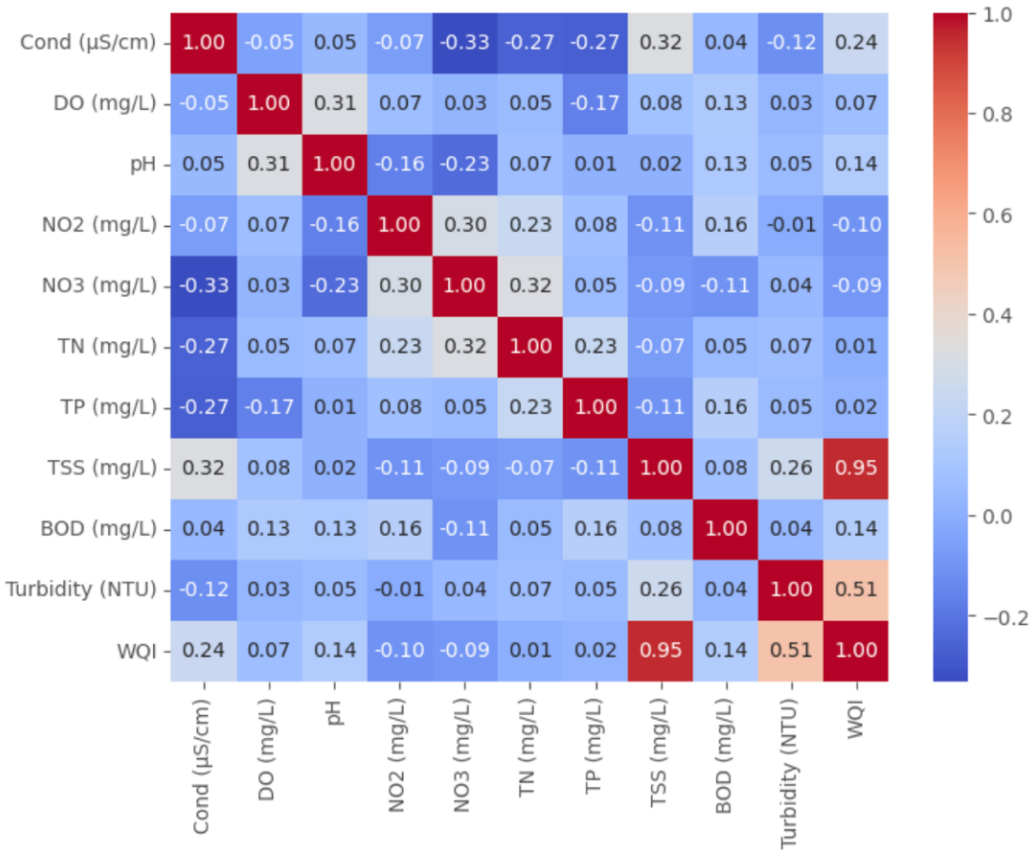

Figure 4. Correlation Heatmap of Numerical Features.

In Figure 5, it is evident that as TSS increase, the WQI tends to increase as well, and vice versa. Despite the presence of outliers in the data, the positive correlation line in the scatter plot suggests that these outliers may not fundamentally alter the relationship between the variables. Knowing the value of TSS can provide valuable information for predicting WQI levels in water samples.

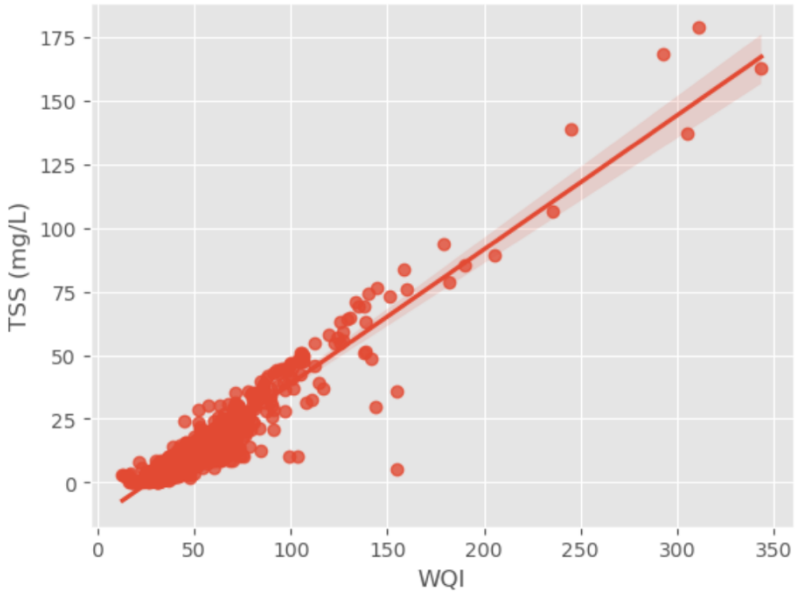

Figure 5. Regression Plot of WQI vs TSS.

Even though the relationship between turbidity and WQI may not be perfectly linear, the general trend indicates that as turbidity increases, the corresponding WQI tends to increase as well (Figure 6). This suggests that higher levels of turbidity are associated with poorer water quality, as indicated by higher WQI values. While the correlation between WQI and turbidity may not be very strong, turbidity can still provide valuable information for predicting WQI levels in water samples.

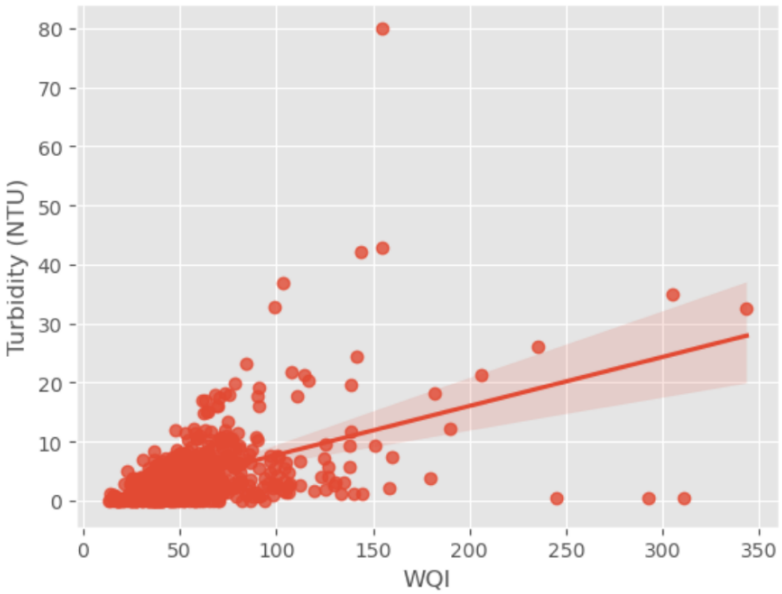

Figure 6. Regression Plot of WQI vs Turbidity.

4. Preprocessing and Training Data Development

In the preprocessing phase, the water quality dataset was initially loaded from a CSV file, and its dimensions were examined to confirm the number of rows and columns. Additionally, the data types were checked, and it was verified that there were no missing values. Subsequently, several transformation steps were applied to prepare the data for analysis. The 'Date' column was converted to datetime format to facilitate time-based analysis, while the categorical variable 'Water Quality Classification' was encoded into binary columns using one-hot encoding. Furthermore, the 'Month' column was extracted from the 'Date' column to investigate potential seasonal patterns in the data.

Following data splitting into training and testing sets with an 80-20 ratio, the features ('X') and target variable ('y') were defined. The next steps focused on feature scaling and transformation to ensure consistency and reduce skewness or outliers in the numerical features. Power Transformation was applied to address skewness and outliers, while MinMaxScaler was used to scale all numerical features to a range between 0 and 1.

To visually assess the effects of preprocessing, histograms of numerical features were plotted before and after scaling (Figure 7 & 8).

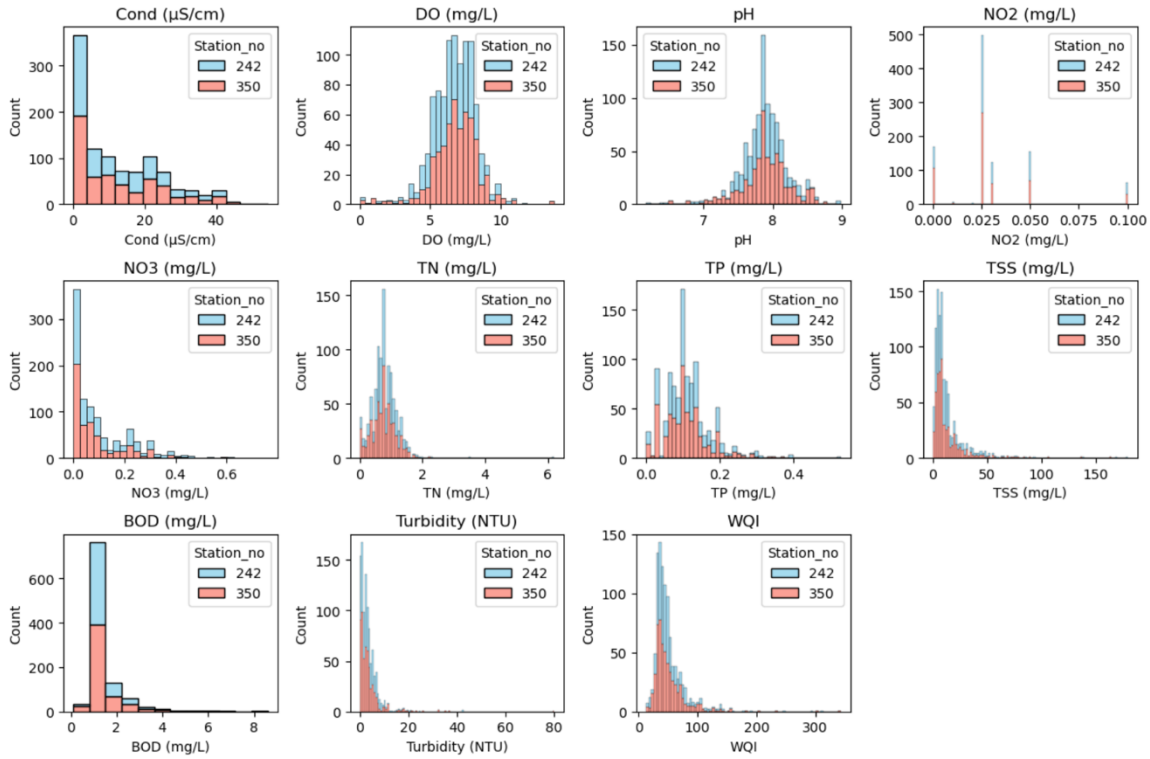

Figure 7. Histograms of numerical features before scaling.

Before scaling, features such as Cond, NO3, TSS, BOD, and Turbidity exhibit skewed distributions (Figure 7). Conversely, DO, pH, TN, and TP demonstrate distributions closer to normal. NO2 appears to have irregular sampling intervals, resulting in gaps between data points. For instance, some data are recorded at intervals of 0, 0.025, 0.050, and 0.1. The scale of each feature varies, with DO measure in mg/L and Cond in µS/cm, contributing to differences in their value ranges. Some features have a narrow range, with closely clustered values, while others exhibit a wider spread, indicating greater variability. Additionally, the central tendency of the features varies, with some showing higher average values than others.

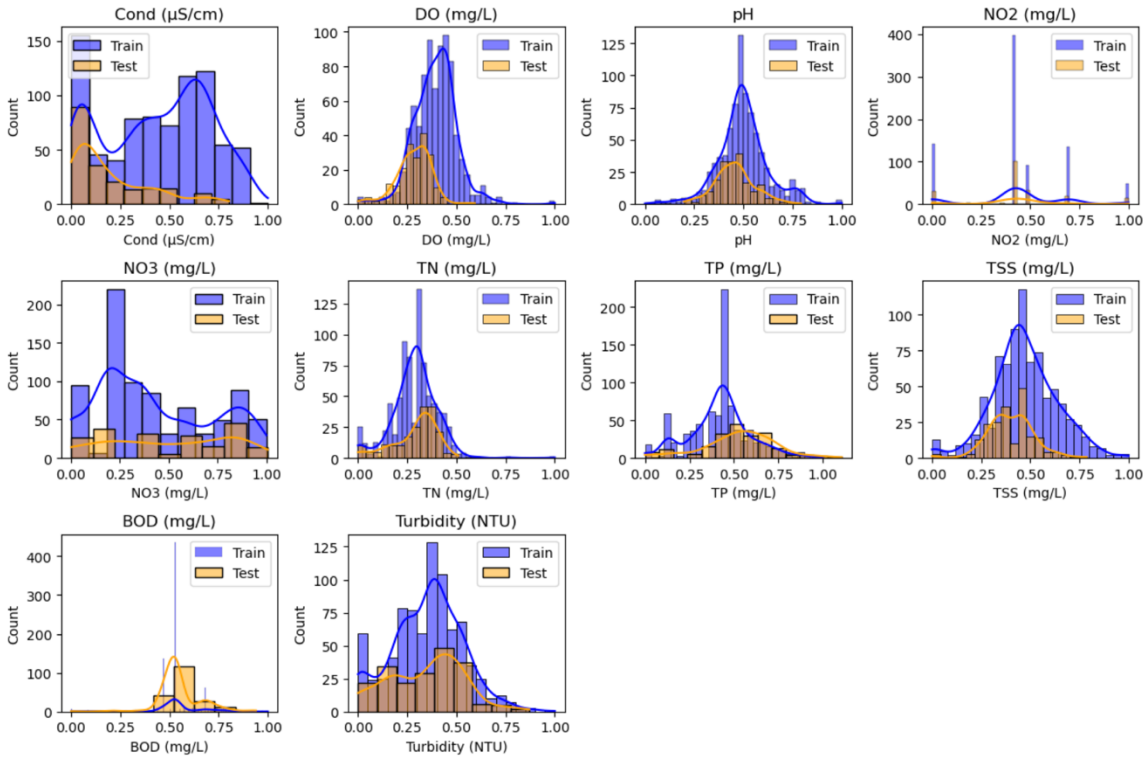

Figure 8. Histograms of numerical features after scaling.

After applying MinMaxScaler to the numerical features in the dataset, we observed that all parameters now exhibit distributions closer to a normal distribution (Figure 8). Additionally, the scaling process has successfully transformed the features to a consistent range between 0 and 1. Additionally, the 'Station_no' and 'Month' columns were concatenated with the transformed dataframes to preserve contextual information for further analysis.

Finally, the transformed training and testing data were saved to CSV files ('X_train_transformed.csv' and 'X_test_transformed.csv'), along with the unchanged target variables ('y_train.csv' and 'y_test.csv').

5. Modeling

Several regression models were selected for predicting the WQI based on the given features. These models included Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regression, and XGBoost Regression. Each model was trained using the training data and evaluated using the testing data to assess its performance. Evaluation metrics such as R-squared (R2) score, Mean Squared Error (MSE), and Mean Absolute Error (MAE) were used to quantify the performance of each model.

To optimize the performance of Random Forest Regression and XGBoost Regression, hyperparameter tuning was performed using GridSearchCV. This involved systematically searching through a range of hyperparameters to find the combination that yielded the best performance for each model. Hyperparameters tuned included the number of estimators for Random Forest and XGBoost, maximum depth of trees, and learning rate for XGBoost.

Cross-validation was employed to assess the models' performance across different splits of the data. This technique helps estimate how the models will perform on unseen data and provides insights into their stability and generalization ability. Various folds (e.g., 3-fold, 5-fold, 7-fold, 10-fold) were used to evaluate the models' performance under different scenarios.

Scatter plots were used to compare predicted vs. actual values, while line plots were used to display cross-validation scores across different folds. These visualizations aided in understanding the models' performance and conveying insights effectively.

The Linear Regression model achieved an **R2** value of approximately **0.477** on the testing set. This indicates that around 47.75% of the variance in the target variable (WQI) is explained by the features in the model. The **MSE** on the testing set is approximately **117.11**, which provides a measure of the average squared difference between the actual and predicted WQI values. The **MAE** on the testing set is approximately **8.07**, representing the average absolute difference between the actual and predicted WQI values.

The scatter plot illustrates the relationship between the predicted and actual values generated by the Linear Regression model (Figure 9). Each point on the plot represents a pair of actual and predicted values for the WQI. The red dashed line represents a perfect fit, where the predicted values would perfectly align with the actual values. However, the scatter of points deviates from this ideal line, indicating that the model's predictions do not precisely

match the actual values. This suggests that the Linear Regression model may not fully capture the underlying complexities of the data, leading to some degree of prediction error.

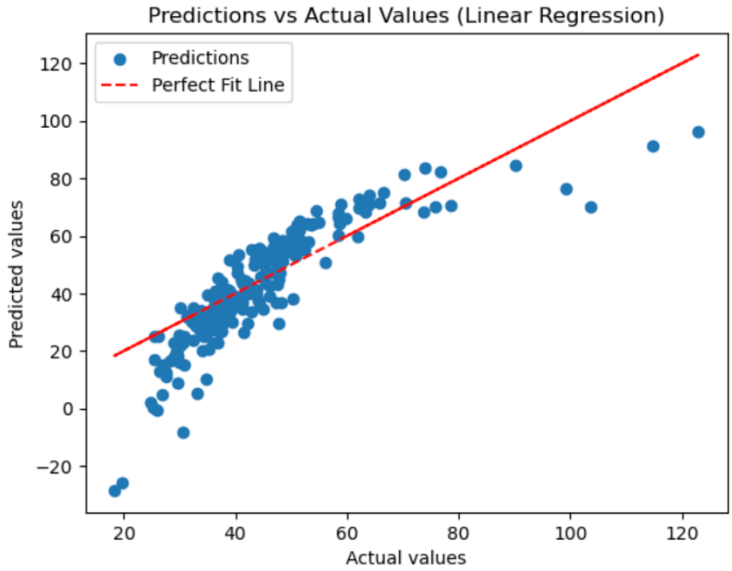

Figure 9. Relationship between the predicted and actual values generated by the Linear Regression model.

The cross-validation results for the Linear Regression model, indicate varying levels of performance across different numbers of splits, is shown in Figure 10. The Linear Regression model demonstrates better performance with lower variability in cross-validation scores when the number of splits is smaller, such as 3 or 5. However, as the number of splits increases, the model's performance on the testing set becomes less consistent, indicating potential overfitting to the training data.

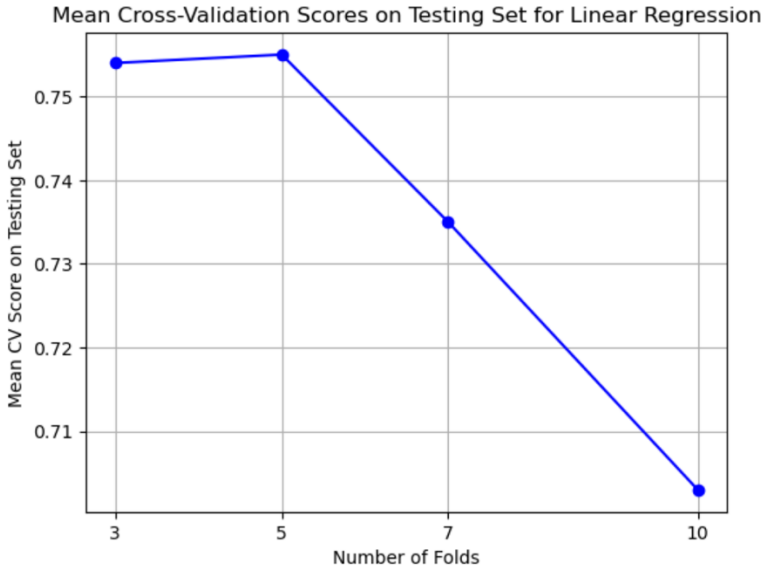

Figure 10. The cross-validation results for the Linear Regression model.

The **Lasso regression** model achieved an **R2** value of approximately **0.536** on the testing set. This indicates that the Lasso model explains about 53.6% of the variance in the target variable, WQI, which is higher than the R2 value obtained from the Linear Regression model. Additionally, the Lasso model yielded a **MSE** of approximately **103.74** and a MAE of approximately **7.44** on the testing set. Both metrics are lower compared to those obtained from the Linear Regression model, suggesting that the Lasso model provides better predictions and has lower prediction errors. Overall, based on these metrics, the Lasso regression model seems to perform better than the Linear Regression model for this dataset.

The **Ridge regression** model achieved an **R2** value of approximately **0.481** on the testing set. This indicates that the Ridge model explains about 48.1% of the variance in the target variable, WQI, which is comparable to the R2 value obtained from the Linear Regression model but slightly lower than the R2 value obtained from the Lasso model. Additionally, the Ridge model yielded a **MSE** of approximately **116.02** and a **MAE** of approximately **8.03** on the testing set. These error metrics are slightly lower than those obtained from the Linear Regression model but slightly higher compared to those obtained from the Lasso model. Overall, the Ridge regression model performs similarly to the Linear Regression model and slightly worse than the Lasso model in terms of R2 value and prediction errors. Therefore, the Lasso model may still be the preferable choice for this dataset based on these evaluation metrics.

The **Random Forest** regression model achieved an impressive **R2** value of approximately **0.933** on the testing set. This indicates that the Random Forest model explains about 93.3% of the variance in the target variable, WQI, which is significantly higher than the R2 values obtained from the Linear Regression, Lasso Regression, and Ridge Regression models. Additionally, the Random Forest model yielded a **MSE** of approximately **14.98** and a **MAE** of approximately **2.61** on the testing set. These error metrics are substantially lower compared to those obtained from the other regression models, indicating that the Random Forest model provides highly accurate predictions for the WQI. Overall, the Random Forest Regression model demonstrates superior performance in terms of R2 value and prediction errors, making it the preferred choice for this dataset.

The scatter plot for Random Forest shows that the predicted values closely follow the trend of the actual values, indicating a strong correlation between the predicted and actual values (Figure 11). Additionally, most of the points are clustered near the perfect fit line, suggesting that the model's predictions are accurate and align well with the ground truth.

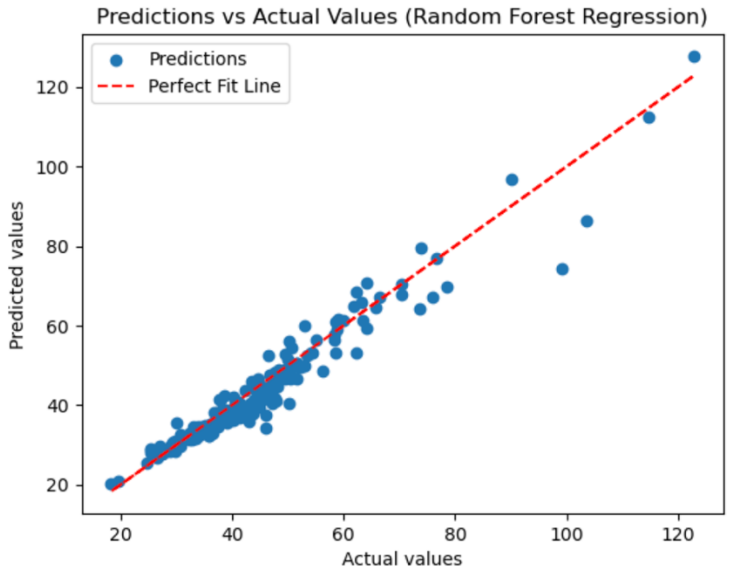

Figure 11. Relationship between the predicted and actual values generated by the Random Forest model.

The Random Forest model performs well across different fold values, with the highest R2 score achieved with 10-fold cross-validation. However, the standard deviation in the scores indicates some variability in model performance, particularly noticeable with fewer folds (5-fold cross-validation) (Figure 12).

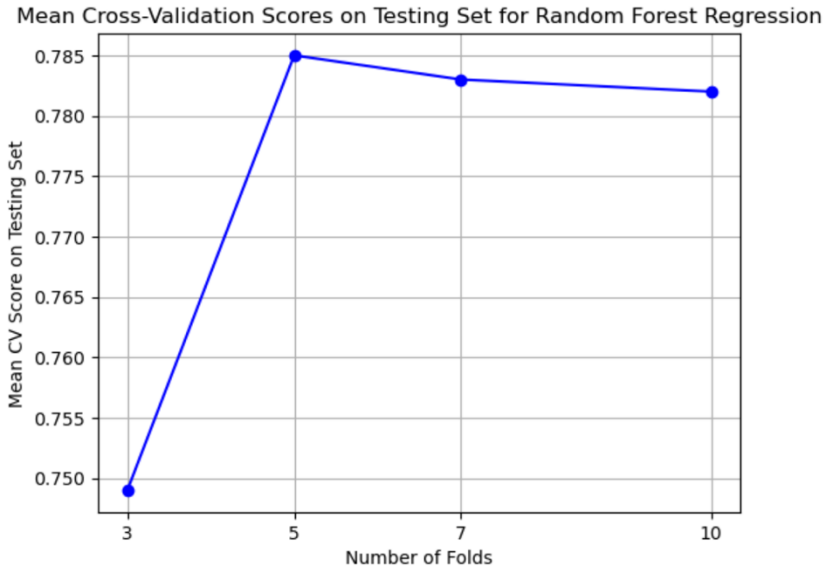

Figure 12. The cross-validation results for the Random Forest Regression.

XGBoost regression model has performed exceptionally well in predicting the WQI. With an **R2** value of approximately **0.962,** the model effectively captures the variance in the target variable, demonstrating a strong fit to the dataset. Furthermore, the low **MSE** and **MAE** values, approximately **8.47** and **2.01**, respectively, highlight the model's accuracy and precision in its predictions. Overall, these evaluation metrics indicate that the XGBoost Regression model performs very well in predicting the WQI based on the given features. As shown in Figure 13, most of the points align closely with the perfect fit line, indicating that the model's predictions are in good agreement with the actual values.

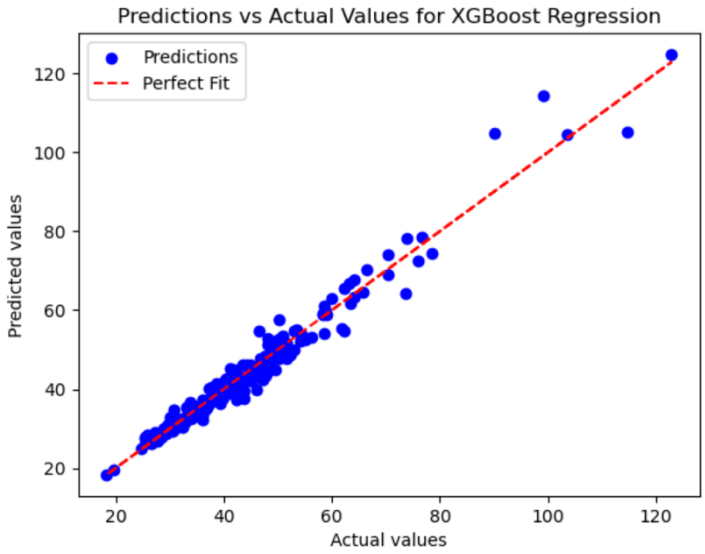

Figure 13. Relationship between the predicted and actual values generated by the XGBoost Regression model.

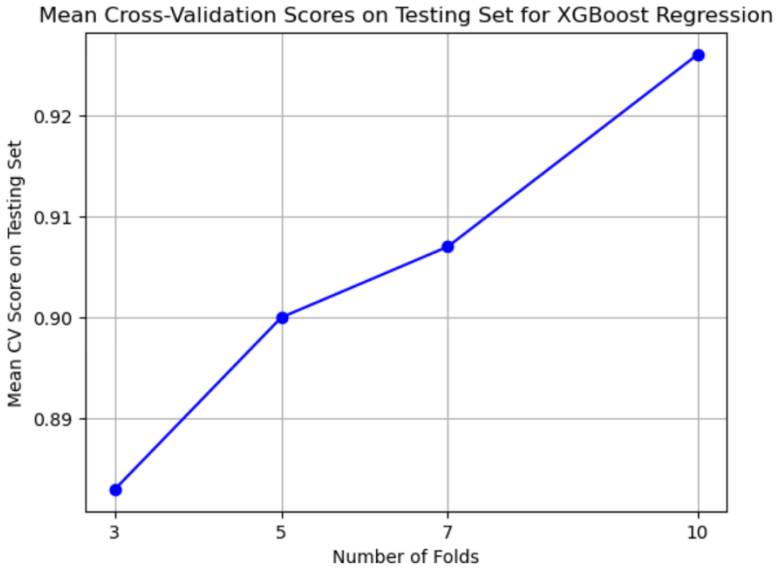

Figure 14. The cross-validation results for the XGBoost Regression model.

Figure 14 illustrates the mean cross-validation scores on the testing set for XGBoost Regression across different numbers of folds. As the number of folds increases from 3 to 10, there is a noticeable improvement in the mean cross-validation score, indicating better generalization performance of the model. The highest mean cross-validation score is achieved with 10-fold cross-validation, reaching approximately 0.926. This suggests that using a larger number of folds leads to better estimation of the model's performance on unseen data, resulting in improved generalization.

6. Results

The performance of each model was analyzed and compared based on evaluation metrics such as R2, MSE, MAE, and mean cross-validation scores. The strengths and weaknesses of each model was identified and determined which model performed best for predicting the WQI (Table 1).

Table 1. Model Comparison

	Model	R-squared (R2)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	Mean Cross-Validation Score on Testing Set
0	Linear Regression	0.477	117.107	8.069	5-fold: 0.755
1	Lasso Regression	0.536	103.736	7.438	-
2	Ridge Regression	0.481	116.024	8.032	-
3	Random Forest Regression	0.933	14.982	2.613	5-fold: 0.785
4	XGBoost Regression	0.962	8.469	2.008	10-fold: 0.926

- XGBoost Regression outperforms all other models in terms of R2 (0.962) on the testing set, indicating its superior ability to explain the variance in the target variable (WQI).
- XGBoost Regression also has the lowest MSE and MAE values on the testing set compared to other models, demonstrating its accuracy in predicting the WQI.
- Among the models compared, XGBoost Regression also exhibits the highest mean cross-validation score on the testing set (0.926) using 10-fold cross-validation, indicating its robustness and consistency in performance across different data splits.
- While Random Forest Regression achieved a high R2 value (0.931) on the testing set, it has a higher MSE and MAE compared to XGBoost Regression.
- Linear Regression, Lasso Regression, and Ridge Regression have lower R-squared values and higher MSE and MAE values compared to both Random Forest Regression and XGBoost Regression, indicating poorer performance in predicting the WQI. However, Lasso Regression shows slightly better performance than Ridge Regression and Linear Regression based on the R-squared value.

7. Conclusion

The modeling phase of the study aimed to develop predictive models for water quality assessment based on a range of environmental features. Through the implementation of various regression algorithms, including Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regression, and XGBoost Regression, we gained valuable insights into the relationship between predictor variables and the WQI.

Our analysis revealed significant variations in the performance of the regression models. While Linear Regression, Lasso Regression, and Ridge Regression demonstrated moderate predictive capabilities, Random Forest Regression and XGBoost Regression outperformed them significantly. Specifically, XGBoost Regression emerged as the top-performing model, achieving an impressive R2 value of 0.962 on the testing set, along with the lowest MSE and MAE values.

Our study showed the importance of selecting appropriate features, models, and evaluation metrics for accurate water quality prediction. Moreover, we identified key factors influencing water quality prediction, providing valuable insights for future research and decision-making processes.

However, our study has some limitations. The data we used might not be perfect, and the assumptions we made during modeling could affect the accuracy of our results. Also, there might be other factors affecting water quality that we didn't consider in our analysis.

8. Recommendations

Based on our findings, there are several promising directions for future research. Firstly, there's room to explore feature engineering further by adding more environmental, geographical, and human-related factors. This exploration could significantly enhance the predictive power of our models by uncovering hidden patterns and relationships within the data.

Furthermore, investigating seasonal variations and long-term trends in water quality could provide valuable insights into how environmental factors impact water quality over time. Utilizing time-series modeling techniques would allow for a comprehensive understanding of these temporal patterns. Employing spatial analysis techniques, such as Geographic Information System (GIS) tools, would enable the exploration of spatial patterns in water

quality, identifying areas of concern and informing targeted interventions and management strategies.

Additionally, exploring the application of deep learning methods, such as neural networks, for water quality prediction could yield significant improvements in predictive accuracy. These models excel at capturing complex relationships in data and may offer novel insights into water quality dynamics. Lastly, validating our models in real-world scenarios through field studies is essential. This ensures their practical usefulness and reliability for decision-makers.

References

1. Rockström, J., et al., *Water resilience for human prosperity*. 2014: Cambridge University Press.
2. UNESCO, *The United Nations World Water Development Report 2021: Valuing Water*. 2021: United Nations.
3. Bui, D.T., et al., *Improving prediction of water quality indices using novel hybrid machine-learning algorithms.* Science of the Total Environment, 2020. **721**: p. 137612.
4. Ahmed, M., R. Mumtaz, and S.M. Hassan Zaidi, *Analysis of water quality indices and machine learning techniques for rating water pollution: A case study of Rawal Dam, Pakistan.* Water Supply, 2021. **21**(6): p. 3225-3250.
5. Lap, B.Q., et al., *Predicting water quality index (WQI) by feature selection and machine learning: a case study of An Kim Hai irrigation system.* Ecological Informatics, 2023. **74**: p. 101991.
6. Ahmed, M., R. Mumtaz, and Z. Anwar, *An Enhanced Water Quality Index for Water Quality Monitoring Using Remote Sensing and Machine Learning.* Applied Sciences, 2022. **12**(24): p. 12787.
7. Uddin, M.G., S. Nash, and A.I. Olbert, *A review of water quality index models and their use for assessing surface water quality.* Ecological Indicators, 2021. **122**: p. 107218.
8. Uddin, M.G., et al., *A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment.* Water Research, 2022. **219**: p. 118532.
9. Uddin, M.G., et al., *A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches.* Water Research, 2023. **229**: p. 119422.
10. Khoi, D.N., et al., *Using machine learning models for predicting the water quality index in the La Buong River, Vietnam.* Water, 2022. **14**(10): p. 1552.
11. Hassan, M.M., et al., *Efficient prediction of water quality index (WQI) using machine learning algorithms.* Human-Centric Intelligent Systems, 2021. **1**(3-4): p. 86-97.