

# Efficient Water Quality Prediction Using Supervised Machine Learning

**Abstract:** Water makes up about 70% of the earth's surface and is one of the most important sources vital to sustaining life. Rapid urbanization and industrialization have led to a deterioration of water quality at an alarming rate, resulting in harrowing diseases. Water quality has been conventionally estimated through expensive and time-consuming lab and statistical analyses, which render the contemporary notion of real-time monitoring moot. The alarming consequences of poor water quality necessitate an alternative method, which is quicker and inexpensive. With this motivation, this research explores a series of supervised machine learning algorithms to estimate the water quality index (WQI), which is a singular index to describe the general quality of water, and the water quality class (WQC), which is a distinctive class defined on the basis of the WQI. The proposed methodology employs four input parameters, namely, temperature, turbidity, pH and total dissolved solids. Of all the employed algorithms, gradient boosting, with a learning rate of 0.1 and polynomial regression, with a degree of 2, predict the WQI most efficiently, having a mean absolute error (MAE) of 1.9642 and 2.7273, respectively. Whereas multi-layer perceptron (MLP), with a configuration of (3, 7), classifies the WQC most efficiently, with an accuracy of 0.8507. The proposed methodology achieves reasonable accuracy using a minimal number of parameters to validate the possibility of its use in real time water quality detection systems.

## 1. Introduction

Water is the most important of sources, vital for sustaining all kinds of life; however, it is in constant threat of pollution by life itself. Water is one of the most communicable mediums with a far reach. Rapid industrialization has consequently led to deterioration of water quality at an alarming rate. Poor water quality results have been known to be one of the major factors of escalation of harrowing diseases. As reported, in developing countries, 80% of the diseases are water borne diseases, which have led to 5 million deaths and 2.5 billion illnesses. The most common of these diseases in Pakistan are diarrhea, typhoid, gastroenteritis, cryptosporidium infections, some forms of hepatitis and giardiasis intestinal worms. In Pakistan, water borne diseases, cause a GDP loss of 0.6–1.44% every year. This makes it a pressing problem, particularly in a developing country like Pakistan.

Water quality is currently estimated through expensive and time-consuming lab and statistical analyses, which require sample collection, transport to labs, and a considerable amount of time and calculation, which is quite ineffective given water is quite a communicable medium and time is of the essence if water is polluted with disease-inducing waste. The horrific consequences of water pollution necessitate a quicker and cheaper alternative.

In this regard, the main motivation in this study is to propose and evaluate an alternative method based on supervised machine learning for the efficient prediction of water quality in real-time.

The main contributions of this study are summarized as follows:

- A first analysis was conducted on the available data to clean, normalize and perform feature selection on the water quality measures, and therefore, to obtain the minimum relevant subset that allows high precision with low cost. In this way, expensive and cumbersome lab analysis with specific sensors can be avoided in further similar analyses.
- A series of representative supervised prediction (classification and regression) algorithms were tested on the dataset worked here. The complete methodology is proposed in the context of water quality numerical analysis.
- After much experimentation, the results reflect that gradient boosting and polynomial regression predict the WQI best with a mean absolute error (MAE) of 1.9642 and 2.7273, respectively, whereas multi-layer perceptron (MLP) classifies the WQC best, with an accuracy of 0.8507.

## 2. Literature Review

This research explores the methodologies that have been employed to help solve problems related to water quality. Typically, conventional lab analysis and statistical analysis are used in research to aid in determining water quality, while some analyses employ machine learning methodologies to assist in finding an optimized solution for the water quality problem.

Local research employing lab analysis helped us gain a greater insight into the water quality problem in Pakistan. In one such research study, Daud et al. gathered water samples from different areas of Pakistan and tested them against different parameters using a manual lab analysis and found a high presence of *E. coli* and fecal coliform due to industrial and sewerage waste. Alamgir et al. tested 46 different samples from Orangi town, Karachi, using manual lab analysis and found them to be high in sulphates and total fecal coliform count.

After getting familiar with the water quality research concerning Pakistan, we explored research employing machine learning methodologies in the realm of water quality. When it comes to estimating water quality using machine learning, Shafi et al. [7] estimated water quality using classical machine learning algorithms namely, Support Vector Machines (SVM), Neural Networks (NN), Deep Neural Networks (Deep NN) and k Nearest Neighbors (kNN), with the highest accuracy of 93% with Deep NN. The estimated water quality in their work is based on only three parameters: turbidity, temperature and pH, which are tested according to World Health Organization (WHO) standards. Using only three parameters and comparing them to standardized values is quite a limitation when predicting water quality. Ahmad et al. employed single feed forward neural networks and a combination of multiple neural networks to estimate the WQI. They used 25 water quality parameters as the input. Using a combination of backward elimination and forward selection selective combination methods, they achieved an R<sup>2</sup> and MSE of 0.9270, 0.9390 and 0.1200, 0.1158, respectively. The use of 25 parameters makes their solution a little immoderate in terms of an inexpensive real time system, given the price of the parameter sensors. Sakizadeh predicted the WQI using 16 water quality parameters and ANN with Bayesian regularization. His study yielded correlation coefficients between the observed and predicted values of 0.94 and 0.77, respectively. Abyaneh predicted the chemical oxygen demand (COD) and the biochemical oxygen demand (BOD) using two conventional machine learning methodologies namely, ANN and multivariate linear regression. They used four parameters, namely pH, temperature, total suspended solids (TSS) and total suspended (TS) to predict the COD and BOD. Ali and Qamar used the unsupervised technique of the average linkage (within groups) method of hierarchical clustering to classify samples into water quality classes. However, they ignored the major parameters associated with WQI during the learning process and they did not use any standardized water quality index to evaluate their predictions. Gazzaz et al. used ANN to predict the WQI with a model explaining almost 99.5% of variation in the data. They used 23 parameters to predict the WQI, which turns out to be quite expensive if one is to use it for an IoT system, given the prices of the sensors. Rankovic et al. predicted the dissolved oxygen (DO) using a feedforward neural network (FNN). They used 10 parameters to predict the DO, which again defeats the purpose if it has to be used for a real-time WQI estimation with an IoT system.

Most of the research either employed manual lab analysis, not estimating the water quality index standard, or used too many parameters to be efficient enough. The proposed methodology improves on these notions and the methodology being followed is depicted in Figure 1.

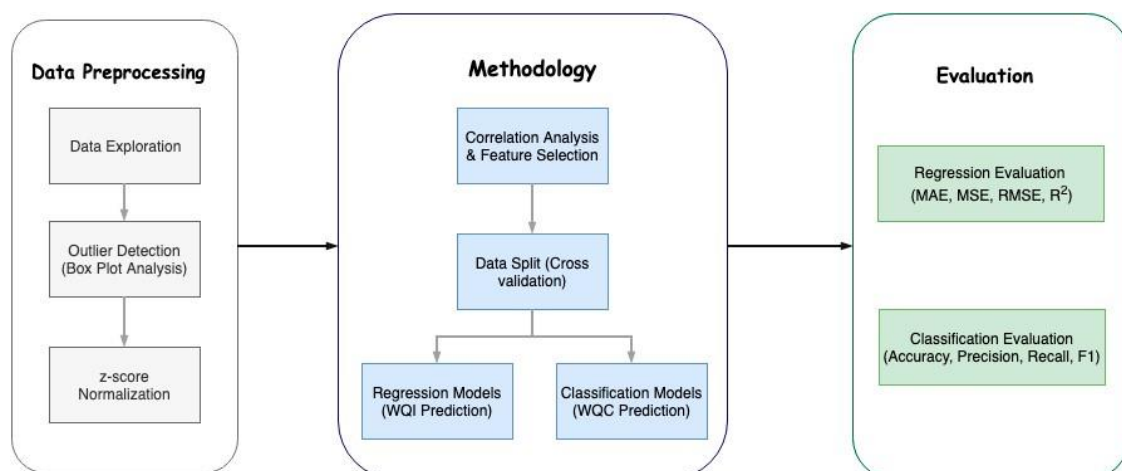


Figure 1. Methodology flow.

### 3. Data Preprocessing

The data used for this research was obtained from PCRWR and it was cleaned by performing a box plot analysis, discussed in this section. After the data were cleaned, they were normalized using q-value normalization to convert them to the range of 0–100 to calculate the WQI using six available parameters. Once the WQI was calculated, all original values were normalized using z-score, so they were on the same scale. The complete procedure is detailed next.

#### 3.1. Data Collection

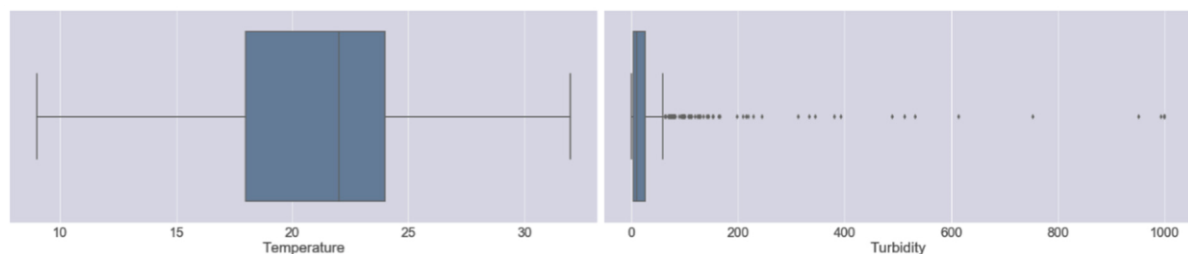
The dataset collected from PCRWR contained 663 samples from 13 different sources of Rawal Water Lake collected throughout 2009 to 2012. It contained 51 samples from each source and the 12 parameters listed in Table 1.

**Table 1. Parameters along with their “WHO” standard limits .**

Parameter	WHO Limits
Alkalinity	500 mg/L
Appearance	Clear
Calcium	200 mg/L
Chlorides	200 mg/L
Conductance	2000 $\mu$ S/cm
Fecal Coliforms	Nil Colonies/100 mL
Hardness as CaCO <sub>3</sub>	500 mg/L
Nitrite as NO <sub>2</sub> <sup>-</sup>	<1 mg/L
pH	6.5–8.5
Temperature	°C
Total Dissolved Solids	1000 mg/L
Turbidity	5 NTU

#### 3.2. Boxplot Analysis and Outlier Detection

We chose boxplot analysis for outlier detection because most of the parameters varied enough and were on the higher end of the values, and a boxplot provides insightful visualization to decide outlier detection threshold values depending upon the problem domain. Boxplot analysis showed that most parameters lied outside the box, deeming outliers normal, so we adapted an upper cap strategy to filter out outliers. We recognized the parameter values that were very different from other values and replaced them with the max threshold value. We set the max threshold value as the parameter value that was just below the outlier values. For example, for turbidity, as reflected in Figure 2, we set the threshold value as the sample value, which was 753, and applied it to all values above 753, so that all the values that lied above 753 were assigned the value of 753. We repeated the same process with all the parameters and manually removed the outliers such as to not risk any data loss at all, given our limited dataset . In addition, we were extremely lenient while choosing the upper threshold of parameters so as not to bias the dataset and just to loosely penalize the values that seemed way out of limits and unlikely to occur.



**Figure 2.** Outlier detection using box plot analysis.

#### 3.3. Water Quality Index (WQI)

Water quality index (WQI) is the singular measure that indicates the quality of water and it is calculated using various parameters that are truly reflective of the water’s quality. To conventionally calculate the WQI, nine water quality parameters are used, but if we did not have all of them, we could still estimate the water quality index with at least six defined parameters. We had five parameters, namely fecal coliform, pH,

temperature, turbidity and total dissolved solids in our dataset. We also considered nitrites as the sixth parameter as the weight and relative importance of nitrites in the WQI calculation is stated to be equal to that of nitrates in multiple WQI studies . Using these parameters and their assigned weightages, we calculated the WQI of each sample as reflected in Equation (1), where  $P_{qvalue}$  reflects the value of a parameter in the range of 0–100 and  $w\_factor$  represents the weight of a particular parameter as listed in Table 2. WQI is fundamentally calculated by initially multiplying the q value of each parameter by its corresponding weight, adding them all up and then dividing the result by the sum of weights of the employed parameters.

$$WQI = \frac{P_{qvalue} \times w\_factor}{w\_factor} P$$

**Table 2. Parameters weights for the WQI calculation.**

Weighing Factor	Weight
pH	0.11
Temperature	0.10
Turbidity	0.08
Total Dissolved Values	0.07
Nitrates	0.10
Fecal Coliform	0.16

### 3.4. Water Qulaity Class (WQC)

Once we had estimated the WQI, we defined the water quality class (WQC) of each sample using the WQI in classification algorithms as shown in Table 3.

**Table 3. Ranges [14,15].**

Water Quality Index Range	Class
0–25	Very bad
25–50	Bad
50–70	Medium
70–90	Good
90–100	Excellent

### 3.5. Q-Value Normalization

Q-value normalization was used to normalize the parameters, particularly the water quality parameters to fit them in the range of 0 to 100 for easier index calculation. Figure 3 shows the q-value charts for six of the water quality parameters. We used them to convert five of these parameters within the range of 0 to 100. For the sixth parameter, namely nitrites, due to unavailability of its q-value ranges, we used the WHO standards to distinctly convert them to the 0–100 range by means of a set of thresholds as follows: assigning 100 if its below 1, 80 if its below 2, 50 if its below 3 and 0 if its greater than 3, reflecting strict penalization. Once the values were q-normalized and were in the range of 0–100, they were used for calculating the WQI of the dataset using (1).

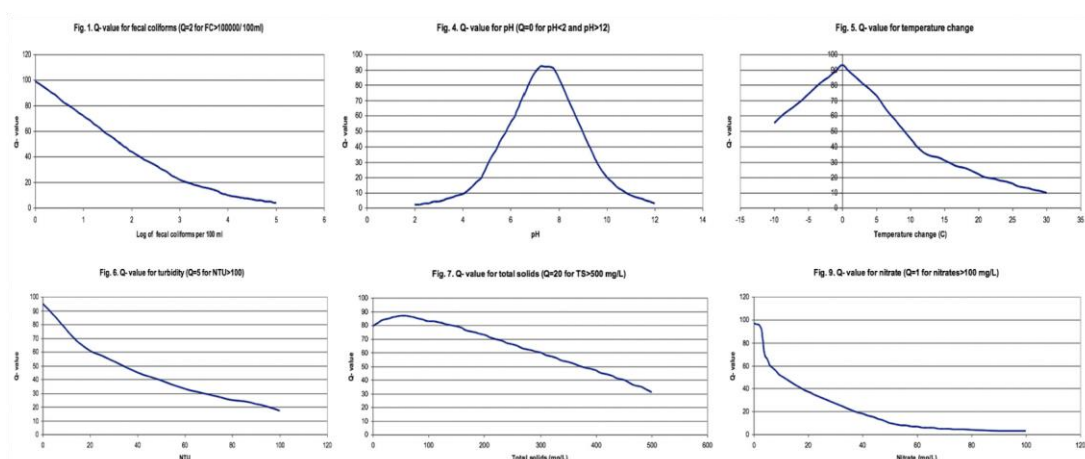


Figure 3. Q-value normalization range charts.

### 3.6. Z-Score Normalization

The z-score is a conventional standardization and normalization method that represents the number of standard deviations; a raw data point is above or below the population mean. It ideally lies between  $-3$  and  $+3$ . It normalizes the dataset to the aforementioned scale to convert all the data with varying scales to the default scale.

To normalize the data using the z-score, we subtracted the mean of the population from a raw data point and divided it by the standard deviation, which gives a score ideally varying between  $-3$  and  $+3$ ; hence, reflecting how many standard deviations a point is above or below the mean as computed by Equation (2), where  $x$  represents the value of a particular sample,  $\mu$  represents the mean and  $\sigma$  represents the standard deviation.

$$z\_score = \frac{(x - \mu)}{\sigma}$$

## 4. Results

In this section, prior to discussing the results, we will describe different measures used to assess the accuracy of the applied machine learning algorithms.

### 4.1. Accuracy Measures

As mentioned earlier, this research employed two types of supervised machine learning algorithms, i.e., regression and classification. The results yielded by both types of algorithms were evaluated differently.

For regression, we used the following measures:

#### (1) Mean Absolute Error (MAE)

Mean absolute error (MAE) is a measure of accuracy for regression. It sums up absolute values of errors and divides them by the total number of values. It gives equal weight to each error value.

The formula for calculating MAE is shown in Equation (5), where  $x_{obs}$  refers to the actual value,  $x_{pred}$  refers to the predicted value, and  $n$  refers to the total number of samples considered.

$$MAE = \frac{\sum (x_{obs} - x_{pred})}{n}$$

#### (2) Mean Square Error (MSE)

Mean square error (MSE) is the sum of squares of errors divided by the total number of predicted values. This attributes greater weight to larger errors. This is particularly useful in problems where there needs to be a larger weight for larger errors. It is measured by Equation (6), where  $x_{obs}$  is the actual value,  $x_{pred}$  is the predicted value, and  $n$  is the total number of samples considered.

$$MSE = \frac{\sum (x_{obs} - x_{pred})^2}{n}$$

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

Root mean squared error (RMSE) is just the square root of MSE and scales the values of MSE near to the ranges of observed values. It is estimated from Equation (7), where  $x_{obs}$  points to the actual value,  $x_{pred}$  points to the predicted value, and  $n$  points to the total number of samples considered.

#### (4) R Squared Error (RSE)

R squared error (RSE), also known as the coefficient of determination, and often denoted as  $R^2$ , determines the goodness of fit of the model. It particularly explains the amount of variance of the dependent variable that is explainable through the independent variable, as shown in Equation (8). Higher RSE values mean that the independent variables largely explain the variance of the dependent variable.

#### Accuracy

Accuracy is the correct number of predictions made by the model over all the observed values. Accuracy is measured by Equation (9), where  $TP$  refers to true positive,  $TN$  refers to true negative,  $FP$  refers to false positive and  $FN$  refers to false negative.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

#### (1) Precision

Precision is the proportion of correctly classified instances of a particular positive class out of the total classified instances of that class. Precision is calculated with the formula shown in Equation (10), where  $TP$  refers to true positive and  $FP$  refers to false positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

#### (2) Recall

Recall is the proportion of instances of a particular positive class that were actually classified correctly. Recall is calculated with the formula shown in Equation (11), where  $TP$  refers to true positive and  $FN$  refers to false negative.

$$\text{Recall} = \frac{TP}{TP + FN}$$

#### (3) F1 Score

As precision and recall, individually, do not cover all aspects of the accuracy, we took their harmonic mean to reflect the F1 score, as shown in Equation (12), which covers both aspects and reflects the overall accuracy measure better. It ranges between 0 and 1. The higher the score, the better the accuracy.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{(2) \text{ Precision} + \text{Recall}}$$

### 4.2. Results for Regression Algorithms

As water quality parameter sensors are expensive, this study aimed to use a minimal number of parameters with cheap sensors to predict water quality. Initially, we used four parameters, namely temperature, turbidity, pH and total dissolved solids. While employing the regression algorithms, we found gradient boosting, having a MAE of 1.9642, MSE of 7.2011, RMSE of 2.6835 and RSE of 0.7485, to be the most efficient algorithm, as shown in Table 5.

**Table 5. Regression results using four parameters.**

Algorithm	MAE	MSE	RMSE	R Squared
Linear Regression	2.6312	11.7550	3.4286	0.6573
Polynomial Regression	2.0037	7.9467	2.8190	0.7134
Random Forest	2.3053	9.5669	3.0930	0.6705
<b>Gradient Boosting</b>	<b>1.9642</b>	<b>7.2011</b>	<b>2.6835</b>	<b>0.7485</b>
SVM	2.4373	10.6333	3.2609	0.3458
Ridge Regression	2.6323	11.7500	3.4278	0.4971
Lasso Regression	3.5850	20.1185	4.4854	-2.9327
Elastic Net Regression	3.6595	20.9698	4.5793	-4.0050

Following that, we tried to reduce more parameters, and so we decided to drop total dissolved solids, which is a little harder to acquire than the others. We found that the polynomial regression, with a MAE of 2.7273, MSE of 12.7307, and RMSE of 3.5680, to be the most efficient algorithm, although linear regression and gradient boosting had the best RSE values, with 0.5384 and 0.5051, respectively, as shown in Table 6. There was an increase in the overall error rate, but the increase was not alarming and still performed well within the limits, given the cost.

**Table 6. Regression results using three parameters.**

Algorithm	MAE	MSE	RMSE	R Squared
Linear Regression	3.1375	15.8321	3.9790	<b>0.5384</b>
<b>Polynomial Regression</b>	<b>2.7273</b>	<b>12.7307</b>	<b>3.5680</b>	0.4851
Random Forest	3.0404	15.2473	3.9048	0.4107
Gradient Boosting	2.8060	13.2710	3.6429	0.5051
SVM	2.8252	13.8546	3.7222	0.1546
Ridge Regression	3.1386	15.8327	3.9790	0.2031
Lasso Regression	3.8800	22.9966	4.7955	-3.6636
Elastic Net Regression	3.9697	24.0678	4.9059	-5.5210

### 4.3. Results for Classification Algorithms

Using classification algorithms, we predicted water quality class (WQC), which was assigned to samples based on their pre calculated WQI. The same parameters, as in previous section, were used for classification as well. Initially, the same four parameters were considered. We found that MLP, in such a setting, performed better than the other algorithms, with an accuracy of 0.8507, precision of 0.5659, recall of 0.5640, and F1 score of 0.5649, as shown in Table 7.

**Table 7. Classification results using four parameters.**

Algorithm	Accuracy	Precision	Recall	F1 Score
<b>MLP</b>	<b>0.8507</b>	<b>0.5659</b>	<b>0.5640</b>	<b>0.5649</b>
Gaussian Naïve Bayes	0.7843	0.4964	0.5491	0.5025
Logistic Regression	0.8401	0.5520	0.5594	0.5548
Stochastic Gradient Descent	0.8205	0.5473	0.5424	0.5443
KNN	0.7270	0.4734	0.4783	0.4750
Decision Tree	0.7949	0.5298	0.5250	0.5268
Random Forest	0.7587	0.5063	0.5011	0.5027
SVM	0.7979	0.5187	0.5327	0.5228
Gradient Boosting Classifier	0.8130	0.5375	0.5376	0.5376
Bagging Classifier	0.8100	0.5410	0.5354	0.5374

In this section, we iterated through our study's results and established that gradient boosting and polynomial regression performed better in predicting WQI, whereas MLP performed better in predicting WQC.

## **5. Discussion**

Water Quality is conventionally calculated using water quality parameters, which are acquired through time consuming lab analysis. We explored alternative methods of machine learning to estimate it and found several studies employing them. These studies used more than 10 water quality parameters to predict WQI. Ahmad et al used 25 input parameters, Sakizadeh used 16 parameters, Gazzaz et al used 23 input parameters in their methodology, and Rankovic et al used 10 input parameters, which is unsuitable for inexpensive real time systems. Whereas, our methodology employs only four water quality parameters to predict WQI, with a MAE of 1.96, and to predict water quality class with an accuracy of 85%. Our results make a base for an inexpensive real time water quality detection system, while other studies, although they use machine learning, use too many parameters to be incorporated in real time systems.

## **6. Conclusions and Future Work**

Water is one of the most essential resources for survival and its quality is determined through WQI. Conventionally, to test water quality, one has to go through expensive and cumbersome lab analysis. This research explored an alternative method of machine learning to predict water quality using minimal and easily available water quality parameters. The data used to conduct the study were acquired from PCRWR and contained 663 samples from 12 different sources of Rawal Lake, Pakistan. A set of representative supervised machine learning algorithms were employed to estimate WQI. This showed that polynomial regression with a degree of 2, and gradient boosting, with a learning rate of 0.1, outperformed other regression algorithms by predicting WQI most efficiently, while MLP with a configuration of (3, 7) outperformed other classification algorithms by classifying WQC most efficiently.

In future works, we propose integrating the findings of this research in a large-scale IoT-based online monitoring system using only the sensors of the required parameters. The tested algorithms would predict the water quality immediately based on the real-time data fed from the IoT system. The proposed IoT system would employ the parameter sensors of pH, turbidity, temperature and TDS for parameter readings and communicate those readings using an Arduino microcontroller and ZigBee transceiver. It would identify poor quality water before it is released for consumption and alert concerned authorities. It will hopefully result in curtailment of people consuming poor quality water and consequently de-escalate harrowing diseases like typhoid and diarrhea. In this regard, the application of a prescriptive analysis from the expected values would lead to future facilities to support decision and policy makers.