Efficient Water Quality Analysis and Prediction Using Machine Learning

ABSTRACT

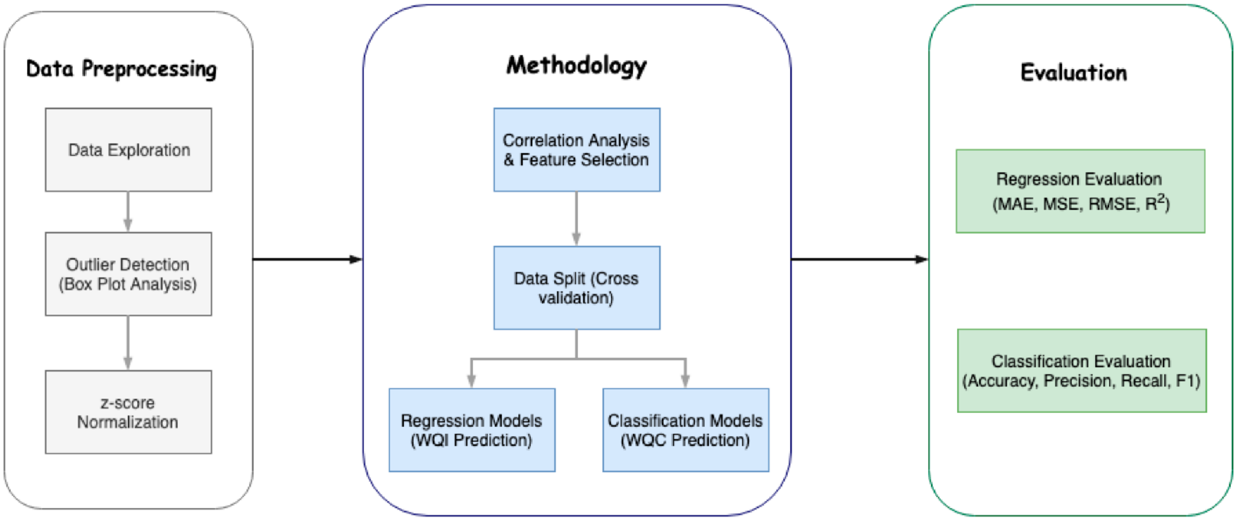
The main purpose of this research is to calculate the water Quality. Water quality that has been conventionally estimated through expensive and time-consuming lab and statistical analyses renders the contemporary notion of real-time monitoring moot. this research explores a series of supervised machine 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 based on the WQI. The proposed methodology Takes input parameters, namely, temperature, turbidity, pH, and total dissolved solids

INTRODUCTION

Water is the most important of source, 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 the deterioration of water quality at an alarming rate. 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 Water Calculation.

LITERATURE REVIEW

This research explores the methodologies that have been employed to help solve problems related to water quality. When it comes to estimating water quality using machine learning, 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.


Data Preprocessing
Data Exploration
Outlier Detection (Box Plot Analysis)
z-score Normalization
Methodology
Correlation Analysis & Feature Selection
Data Split (Cross validation)
Regression Models (WQI Prediction)
Classification Models (WQC Prediction)
Evaluation
Regression Evaluation (MAE, MSE, RMSE, R^2)
Classification Evaluation (Accuracy, Precision, Recall, F1)

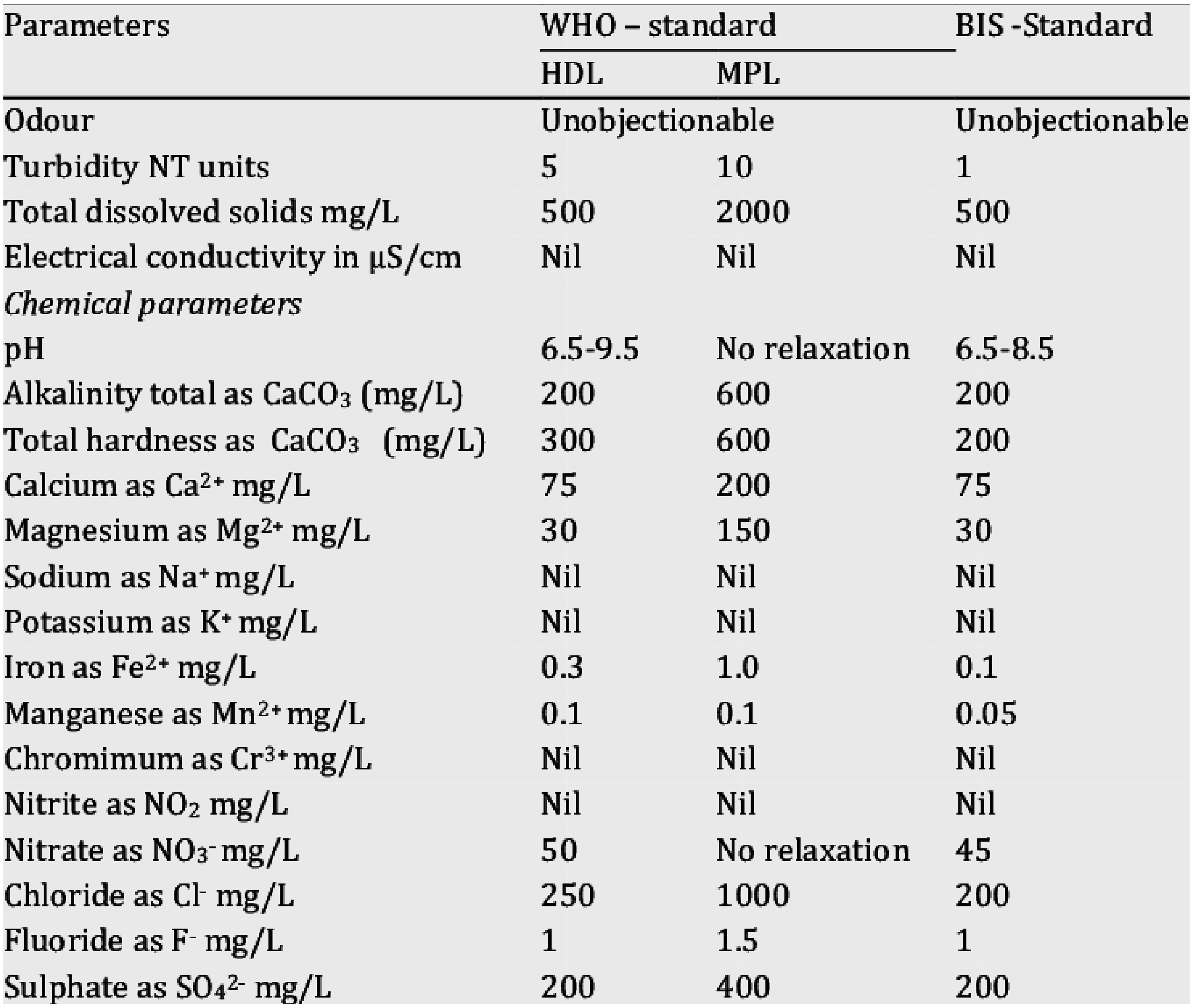
DATA PREPROCESSING

The data used for this research was obtained from Kaggle.com 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 the z-score, so they were on the same scale.

DATA COLLECTION

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



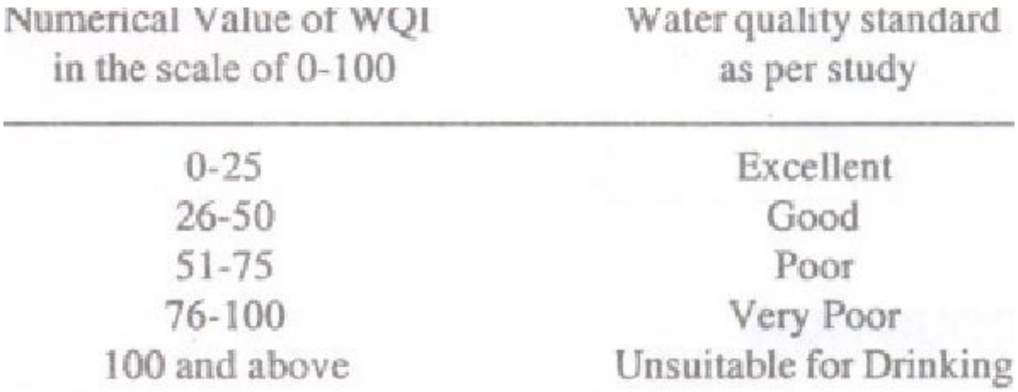
Parameters	WHO – standard		BIS -Standard
	HDL	MPL	
Odour	Unobjectionable		Unobjectionable
Turbidity NT units	5	10	1
Total dissolved solids mg/L	500	2000	500
Electrical conductivity in µS/cm	Nil	Nil	Nil
Chemical parameters			
pH	6.5-9.5	No relaxation	6.5-8.5
Alkalinity total as $CaCO_3$ (mg/L)	200	600	200
Total hardness as $CaCO_3$ (mg/L)	300	600	200
Calcium as Ca^{2+} mg/L	75	200	75
Magnesium as Mg^{2+} mg/L	30	150	30
Sodium as Na^{+} mg/L	Nil	Nil	Nil
Potassium as K^{+} mg/L	Nil	Nil	Nil
Iron as Fe^{2+} mg/L	0.3	1.0	0.1
Manganese as Mn^{2+} mg/L	0.1	0.1	0.05
Chromimum as Cr^{3+} mg/L	Nil	Nil	Nil
Nitrite as NO_2 mg/L	Nil	Nil	Nil
Nitrate as NO_3^{-} mg/L	50	No relaxation	45
Chloride as Cl^{-} mg/L	250	1000	200
Fluoride as F^{-} mg/L	1	1.5	1
Sulphate as SO_4^{2-} mg/L	200	400	200

WATER QUALITY INDEX(WQI)

The 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

WATER QUALITY CLASS(WQC)

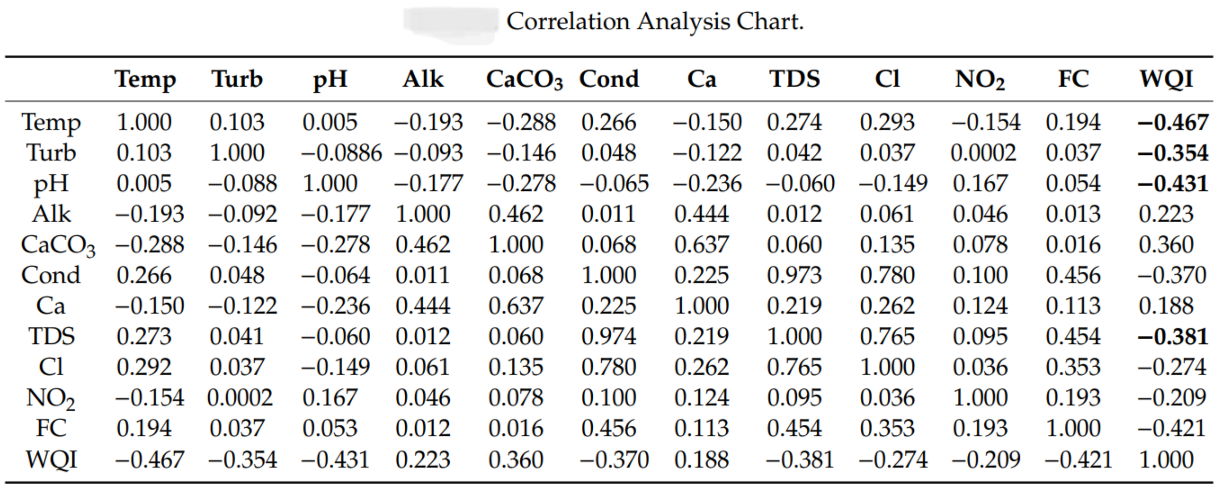
Once we had estimated the WQI, we defined the water quality class (WQC) of each sample using the WQI in classification algorithms.



Numerical Value of WQI in the scale of 0-100	Water quality standard as per study
0-25	Excellent
26-50	Good
51-75	Poor
76-100	Very Poor
100 and above	Unsuitable for Drinking

DATA ANALYSIS

We used the most commonly used and effective correlation method, known as the Pearson correlation. We applied the Pearson correlation on the raw values of the parameters listed in Table 4 and applied it after normalizing the values through q-value normalization as explained in the subsequent section.



[redacted] Correlation Analysis Chart.

	Temp	Turb	pH	Alk	$CaCO_3$	Cond	Ca	TDS	Cl	NO_2	FC	WQI
Temp	1.000	0.103	0.005	−0.193	−0.288	0.266	−0.150	0.274	0.293	−0.154	0.194	**−0.467**
Turb	0.103	1.000	−0.0886	−0.093	−0.146	0.048	−0.122	0.042	0.037	0.0002	0.037	**−0.354**
pH	0.005	−0.088	1.000	−0.177	−0.278	−0.065	−0.236	−0.060	−0.149	0.167	0.054	**−0.431**
Alk	−0.193	−0.092	−0.177	1.000	0.462	0.011	0.444	0.012	0.061	0.046	0.013	0.223
$CaCO_3$	−0.288	−0.146	−0.278	0.462	1.000	0.068	0.637	0.060	0.135	0.078	0.016	0.360
Cond	0.266	0.048	−0.064	0.011	0.068	1.000	0.225	0.973	0.780	0.100	0.456	−0.370
Ca	−0.150	−0.122	−0.236	0.444	0.637	0.225	1.000	0.219	0.262	0.124	0.113	0.188
TDS	0.273	0.041	−0.060	0.012	0.060	0.974	0.219	1.000	0.765	0.095	0.454	**−0.381**
Cl	0.292	0.037	−0.149	0.061	0.135	0.780	0.262	0.765	1.000	0.036	0.353	−0.274
NO_2	−0.154	0.0002	0.167	0.046	0.078	0.100	0.124	0.095	0.036	1.000	0.193	−0.209
FC	0.194	0.037	0.053	0.012	0.016	0.456	0.113	0.454	0.353	0.193	1.000	−0.421
WQI	−0.467	−0.354	−0.431	0.223	0.360	−0.370	0.188	−0.381	−0.274	−0.209	−0.421	1.000

CONCLUSION

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.