Water Quality Prediction

Sarthak Arora
2018307
sarthak18307@iiitd.ac.in

Mohammad Sajeel Khan
2018342
sajeel18342@iiitd.ac.in

Shekhar Shukla
2018414
shekhar18414@iiitd.ac.in

Abstract

The main motivation to explore this topic comes from the continuous degradation of the environment, which leads to various diseases and substandard living for people facing that degradation in their local habitats. The water quality of the post-industrial era has been subject to continuous degradation from practises such as manufacturing, industrial mega-structures and plastic dumping. The governments are trying to find ways to check on these projects, so that water degradation could be stopped further. Some of these methods involve simple measurement techniques and take the required steps post recording these depreciated values. The regular methods require lab analyses which prove to be costly and time consuming. Thus, the need to process data smartly becomes an important part of this problem. In our study, our team reviewed multiple researches which were relatable to the Indian subcontinent. Our data collection was majorly based upon Indian water sphere including water bodies like rivers, lakes, ponds, creeks and ponds. After preprocessing the data, we aimed to achieve good predictions with low losses and accurate classifications with different combinations of attributes lesser than recommended by the World Health Organization(WHO). We achieved an R2 value of 0.98 and the classification accuracy of 0.96 with only five features which is one less than the WHO recommendations for which DO was left out. R2 value with only three features came out to be 0.93 with accuracy for the same case as 0.93. Our work can be found on our ***Github.***

1. Introduction

The WHO lays out at least six features by which the water quality of a sample can be decided. The process of collection of water samples and the calculation of these various features is time consuming and expensive. Things like transportation, storage, and sampling for many locations throughout a country used in the process add up to the cost. By introducing machine learning algorithms to this scenario we aim to predict the water quality of these samples using lesser features than recommended by WHO.

2. Literature Review

1. Efficient Water Quality Prediction Using Supervised Machine Learning [1]
   This paper tries to explore various aspects involved in predicting the Water Quality Indexes. The paper explains the process of outlier detection in data related to water quality by following the norms set by WHO. The paper also introduced how Quality values are used from experimental data to analyse the parameters.The research is done based on data collected from Rawal Water Lake, Pakistan.

   - Samples Explored: 663
   - Final Parameter Considered: Temperature, Turbidity, pH and TDS
   - Prediction models : Linear Regression and Polynomial Regression
   - Result Analysis: MSE, RMSE and MSE

2. Comparison of Water Quality Classification Models using Machine Learning [2]
   This paper aims to classify water quality based on Water Class Index. The paper takes a different approach from usual by taking less than suggested number of parameters required for calculating WQI.The dataset used involved used was a mixture of data from surface, ground and wastewater. The origin of this data was India only.

   - Samples Explored: 1991
   - Final Parameter Considered: pH, BOD, DO and EC.
   - Prediction models : SVM, Naïve Bayes and Decision Trees.
   - Result Analysis: Balanced Accuracy Score and Confusion Matrix.

3. Dataset

3.1. Data Description

The dataset was taken from Ministry of Environment and Forests, Govt of India. Water data of rivers, drains, creeks, lakes, ponds, tanks of years 2016 to 2019 chosen. [3] The data had 8 features, namely, Temperature, Dissolved Oxygen(DO), pH, Conductivity, Biochemical Oxygen Demand(BOD), Nitrate, Faecal Coliform and Total Coliform. The number of samples used in this study was 3938.

	Temp	DO	pH	Conductivity
count	1285.000000	1285.000000	1285.000000	1285.000000
mean	25.880000	6.379121	7.709241	2201.455720
std	4.834291	2.215601	0.391944	7313.910823
min	0.000000	0.100000	6.150000	22.000000
25%	24.000000	5.500000	7.500000	236.000000
50%	26.500000	6.600000	7.750000	405.000000
75%	28.000000	7.450000	7.950000	759.500000
max	131.500000	41.750000	9.150000	54340.000000

Figure 1. Data Description-1

BOD	Nitrate	Faecal Coliform	Total Coliform
1285.000000	1285.000000	1.285000e+03	1.285000e+03
8.331241	2.382125	8.881134e+04	1.788309e+05
26.058498	15.245897	1.074042e+06	1.568586e+06
0.000000	0.000000	0.000000e+00	0.000000e+00
1.700000	0.400000	5.350000e+01	3.050000e+02
2.700000	0.950000	6.000000e+02	1.440000e+03
6.050000	1.900000	3.000000e+03	9.200000e+03
559.000000	442.350000	2.710500e+07	3.500600e+07

Figure 2. Data Description-2

3.2. Data Preprocessing

- The data was converted to means as it had the maximum and the minimum values of the features.
- The Faecal and Total Coliform were transformed by log10 transformation due to their excessively large values.
- Due to very large gaps in the minimum and the maximum values of a feature, decision was made to recognise and treat the outliers. This was done by box-plot analysis and upper and lower were decided accordingly.

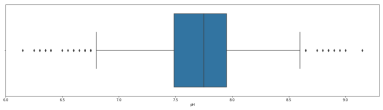

Figure 3. Box Plot of pH

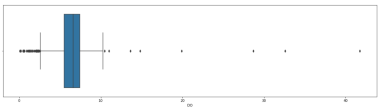

Figure 4. Box Plot of DO

- After all these transformations, the z-score normalization was applied.
- Correlation matrix of the features shows that there is a strong correlation between :
  - Faecal and Total Coliform
  - BOD and Conductivity
  - BOD and DO

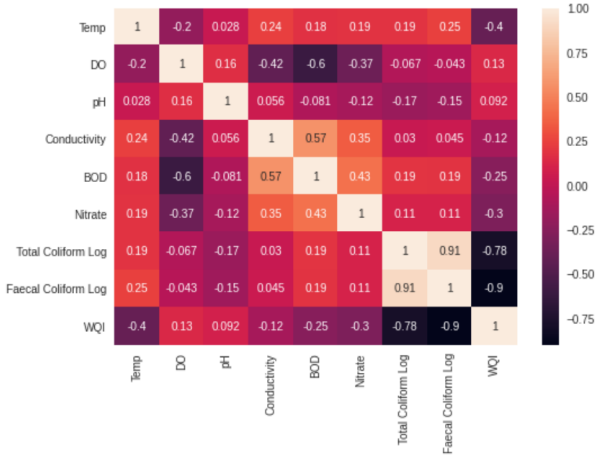

Figure 5. Correlation Matrix

4. Methodology

- Prior to calculating the WQI values, Total Faecal Coliform and Conductivity were dropped from being considered as important aspects for WQI with respect to the results achieved in the correlation matrix.
- Next q values [4] chunks were defined and data was processed to get continuous values for the calculation of WQI with availability of all the required parameters.
- Since the availability of data points for all parameters is not feasible for every place, WQI is predicted using parameters less than the number used for calculating

pH (units)	Q-Value
<2	0
2	2
3	4
4	8
5	24
6	55
7	90
7.2	92
7.5	93 (max)
7.7	90
8	82
8.5	67
9	47
10	19
11	7
12	2
>12	0

Figure 6. Q-Value conversion table

it i.e. 6. Our study included around 41 subsets with number of features 3,4 and 5.

- Next, 5 different methods were used for WQI prediction:
  - Linear Regression
  - Lasso Regression
  - Ridge Regression
  - Support Vector Regressor
  - Random Forest Regressor
- All five methods were performed using K-Fold cross validation.
- R Square, MAE, MSE, RMSE were used for comparison of the results. Observed vs Fitted lines were plotted for the results achieved.
- Water samples were classified into 3 different classes namely bad, mediocre and good.
- Next, 5 methods were used for classification:
  - Logistic Regression
  - Gaussian Naïve Bayes
  - Support Vector Classifier
  - Bagging Classifier
  - Decision Tree
- Accuracy score was calculated for result comparison. Confusion matrix, ROC curves were plotted for the results achieved.

5. Results and Analysis

Results are obtained from running 5 models for prediction and 5 for classification on each of the 41 subsets of the attributes. The best models for both prediction and classification turned out to be Support Vector algorithms.

Highest R square score achieved for predicting WQI was 0.96 which had Temperature, pH, BOD, Nitrate, and Faecal Coliform as the selected features and was using Support Vector Regressor. This combination resulted in a RMSE of 1.52.

	3 Features	4 Features	5 Features
Linear Regression	0.88	0.90	0.91
Ridge Regression	0.88	0.89	0.91
Lasso Regression	0.85	0.85	0.88
Support Vector	.98	0.96	0.98
Random Forest	0.92	0.96	0.97

Figure 7. Model Comparison: Best R square values for various models

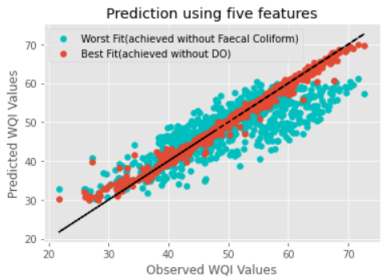

Figure 8. Predicted vs. Observed WQIs, Five features

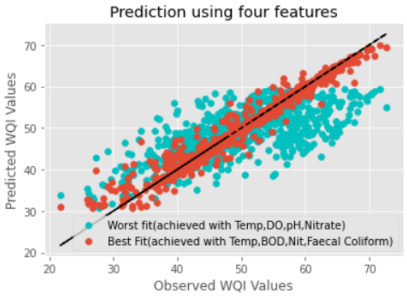

Figure 9. Predicted vs. Observed WQIs, Four features

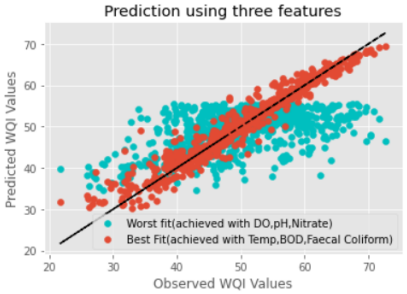

Figure 10. Predicted vs. Observed WQIs, Three features

Highest accuracy achieved for classifying water quality was 0.98 which had Temp, pH, BOD, Nitrate, and Faecal Coliform as the selected features and was using Support Vector Regressor. SVC was also able to achieve accuracy of as high as 0.93 using only 3 features. These three features being Temperature, BOD, and Faecal Coliform.

	3 Features	4 Features	5 Features
Logistic Regression	0.92	0.93	0.94
GNB	0.89	0.89	0.87
Support Vector	0.93	0.95	0.96
Bagging Classifier	0.91	0.94	0.94
Decision Tree	0.90	0.92	0.93

Figure 11. Model Comparison: Best Accuracies for various models

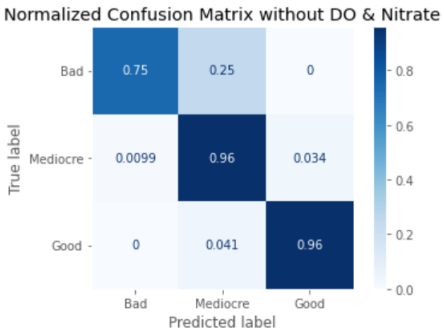

Figure 12. Confusion Matrix, Four features

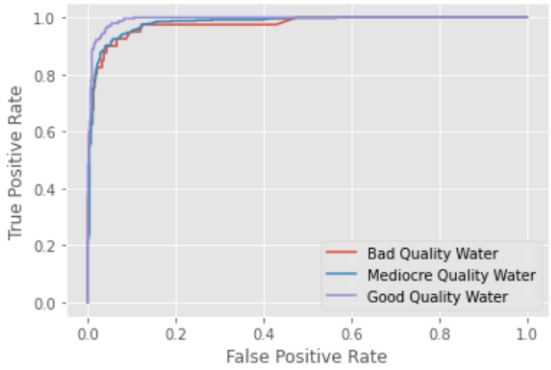

Figure 13. ROC Curves, Three features

According to our analysis, the priority order of attribute selection was established in the following manner:

- In case of five features, the best results were achieved with the exclusion of DO(Dissolved Oxygen).
- In case of four features, the best results were achieved with the exclusion of DO and pH.
- In case of three features, the best results were achieved with the exclusion of DO, pH and Nitrate.

6. Conclusion

These results provide an alternative for calculating WQI effectively. Results show us the importance of various subsets of attributes. The study explains the importance of features such as faecal coliform and BOD as a combination of these two features contributes highly in prediction of WQI and classifying samples to different classes. These results are highly useful in situations where all parameters required for WQI prediction can not be measured due to any reason. This method is not just reliable but it is also a cheaper and effective method. By implementing this, the cost of deploying multiple costly equipment for measuring at least 6 values for water quality metrics can be significantly reduced. This can be accounted to the fact that our study presents a procedure by which water quality can be measured effectively with as low as 3 attributes.

6.1. Contribution

- Sarthak Arora: Papers' analysis, Data collection and conversions, Outlier Detection, Classification codes, Overleaf documentation
- Mohammad Sajeel Khan: Papers' analysis, Data collection and conversions, Preprocessing, Regression and Classification codes
- Shekhar Shukla: Papers' analysis, Preprocessing, Log transformations, Regression Codes, Overleaf documentation, Github Maintenance
- The work was cohesive in nature with every teammate in the know-how at every stage of the project. The conclusive analysis was done collectively by the team.

6.2. Timeline

We were able to cope up with proposed timeline till the mid-semester. Correcting various errors along with proceeding further, out team was able to complete the project within the timeline of the course.

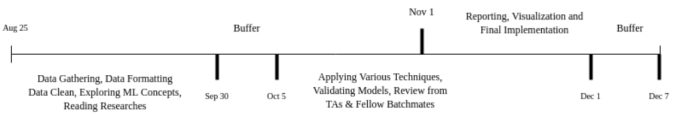

Figure 14. Timeline

References

[1] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, "Efficient Water Quality

Prediction Using Supervised Machine Learning," Water, vol. 11, no. 11, Art. no. 11, Nov. 2019, doi: 10.3390/w11112210.

[2] N. Radhakrishnan and A. Pillai, "Comparison of Water Quality Classification Models using Machine Learning - IEEE Conference Publication." https://ieeexplore.ieee.org/document/9137903?denied= (accessed Nov. 22, 2020).

[3] "CPCB ENVIS— Control of Pollution." http://www.cpcbenvis.nic.in/water$_{q}uality_{d}ata.html$ $(accessedNov.22, 2020)$.

[4] A. Thukral, R. Bhardwaj, and R. Kaur, "Water Quality Indices," 2005, pp. 138–155