European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 7, Issue 5, 2020

PERFORMANCE EVALUATION OF MACHINE LEARNING ALGORITHMS IN FORECASTING WATER QUALITY INDICES: STUDY IN TAMILNADU WATER BODIES

A.Rama[1], S Rajakumari[2],P.Selvamani[3]
[1]Assistant Professor, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences,Chennai
[2]Assistant Professor ,Department of Chemistry, Rajalakshmi Institute of Technology, Kuthambakkam, Chennai, India. E-mail:rajakumari.s@ritchennai.edu.in
[3]Assistant Professor, Department of Computer Science and Engineering, Sri Venkateswara College of Engineering,Chennai. Email-pselvamani@svce.ac.in

ABSTRACT

Water quality prediction play an essential role in aqua environment management. The demand for accurate water quality prediction techniques for efficient water resources management. Currently, the Indian pollution control board has set up various monitoring stations to measure water quality frequently. However, the forecast for water quality is currently not being carried out. In this work, machine learning models have been implemented to predict the indices of water quality. The efficiency of logistic Linear regression and AdaBoostRegressor in the prediction of seven major water quality parameters were evaluated. The Tamil Nadu water quality dataset is used in this analysis. The parameters such as pH value, the quantity of oxygen dissolved, total coli form, B.D.O, electric conductivity, the quantity of phosphorus, and nitrate are considered. The assessed error-index value of the applied models showed that the AdaboostRegressor obtains a lesser error-index and it can consider being a more accurate model than the Linear regression model. The entire methodology proposed here is in the context of water quality is based on numerical analysis. While investigating the outcomes of the implemented machine learning models, it is demonstrated that they have nearly over-estimation properties. The proposed models are assessed using the metrics Mean Square Error and R^2 score the results reflect that AdaboostRegressor predicts the (Water Quality Indices) WQI rate with a Mean Square Error value of 0.8, and R^2 score rate is 0.41, whereas AdaBoostRegressor with a obtains Mean Square Error (MSE) rate as 0.74 and R^2 score rate as 0.44.

KEYWORDS: Water quality prediction, Linear Regression, AdaBoostRegressor.

1. INTRODUCTION

Water is the most essential substance of life and it is utilized for various practices, such as drinking, washing, irrigation, industries, etc... Among various sources of water supply, rivers and lakes are considered most common for the development of humankind because of its ease of use. Several kinds of research associated with analyzing and forecasting the quality indices for river water have been conducted around the world and the corresponding area has been proposed with the name as river engineering [1]. Water quality prediction helps in controlling water pollution and protects living beings. In this work, we use using machine learning techniques to estimate the quality indices of water in various water stations of Tamil Nadu. In particular, this machine learning model consists of two methods, Linear Regression and AdaBoostRegressor. In general, the experiments are done using real-world water quality datasets of Tamil Nadu to validate the effectiveness of our approach [5]. Water is an essential molecular substance for the functioning of the human body, it is composed of an uncommon chemical and physical properties. Some factors lead to water irritancy including the amount of pH in water, the hardness of water, temperature, and other chemical elements. The increase in calcium salts ratio in water will leads to irritation in the skin more easily [6]. Substantially in developing countries, people are afraid of contaminated water from a tap or underground water. Most of the time use purified water for their daily usage. Thus, the main objective of this study is to analyze the pH, dissolved oxygen, total coliform, B.D.O (Biochemical Oxygen Demand, electrical conductivity, nitrate, and phosphorus of water from various stations that are utilized by people for their daily activities. In this study, the performance of machine learning algorithms for predicting the water quality components is investigated. The water sources of various stations located in the Tamil Nadu state of India are considered for the experiment for analysis of the performance of machine learning techniques.

1.1 SYSTEM DESIGN

The methodology design flow is as follows.

This work starts with Data acquisition (the collected and available data are then pre-processed; later machine learning models are constructed using various techniques to perform prediction. The input data is fed into the proposed machine learning model and therefore obtained the prediction value, later the performance of machine learning algorithms is evaluated based on prediction.

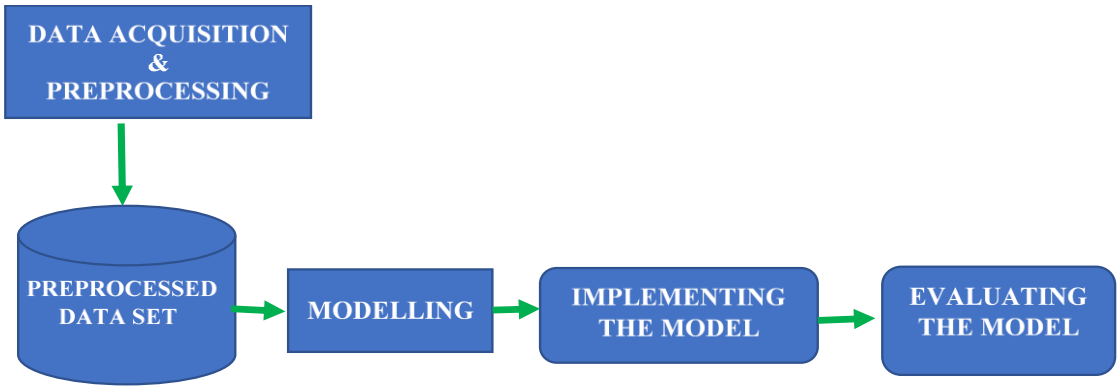

Fig 1: Steps involved in proposed water quality prediction

1.2 WATER QUALITY INDICATORS

Water quality is analyzed by using several indicators, which are categorized based on chemical, physical, biological, radioactive indicators, etc. [8] water has a different purpose, and based on the purpose the evaluation indicator value changes. The dataset considered here is to estimate the suitability of drinking water and irrigation water quality that are used in various regions of Tamilnadu, India. Water Quality Index (WQI) is a numeric value that signifies the usability of drinking or irrigation water. Therefore, for experiment WQI value is considered and it is calculated based on various parameters such as pH value, dissolved oxygen, phosphate, potassium, BDO, Total Coliform, and Nitrate. Considering the irrigation indices, most of the water and locations are mainly used for irrigation purposes. The dataset considered for this experiment shows the details of water quality indices and how it can help the policymakers to maintain and handle the water resources efficiently to satisfy the needs of society at large. [4]. pH is a measure of acidic/basic in water, its value should be greater than or equal to seven as it can be used for either drinking or irrigation. If it is less than seven then the water is acidic. Like the same phosphate is also not toxic if its value ranges between 0,005 to 0.05 mg/L else it becomes toxic for living beings.

The remaining session of this paper is organized such that, Section 2 offers a study of literature concerning the domain. Section 3, provides details about materials and methods that support this work. Section 4, demonstrates the experiment and the discussion of the results based on adaboostregressor and Linear regression. In Section 5, The paper is concluded with the future scope of the work

2. LITERATURE REVIEW

[3] Shuangyin Liu et.al proposes a hybrid RGA–SVR approach for water quality forecasting. where a suitable parameter for SVR is chosen by a real-valued genetic algorithm. The result

demonstrated that the AI techniques are more suitable for performing forecast functions of any nonlinear problems. The proposed RGA–SVR technique avoids economic losses to a certain extent. [9] Jafari, H., Rajaee et al, proposed a hybrid technique of (WGP)Wavelet-Genetic Programming for Prediction of water quality based on the parameter biochemical oxygen request (BOD), which works in as the fundamental water contamination markers in different freshwater assets.Later evaluation of the proposed method is used to demonstrate the efficiency of the five various machine learning algorithms consisting of WANN, ANN, GP, DT, and BN. Examination of the outcomes shows that the WGP model is better than every other model dependent on information from Dam stations. [10] Miklas Scholz et al., proposes a template-based fuzzy inference system (TFIS) to stimulate the content of dissolved oxygen (DO). Later Stepwise Linear regression analysis is implemented to choose the best independent parameter as an input parameter to feed to the regressive model. Temperature and conductivity are considered as an effective parameter compared to pH and DO; therefore these two parameters are considered as an input for the Linear regression model and later the model is evaluated. As a result, TFIS techniques are demonstrated as a superior method compared to the SR model. [11] DiWu et al, propose an intelligent data-driven method to forecast water quality. The model includes Adaptive learning and random forest. Later the performance of these two models is evaluated based on accuracy and time cost. [12] Pietro Boccadoro et al., proposes a WaterS system to enhance the forecast model dependent on the work of neural systems. The model has been proven as a better water quality analysis model by achieving MAE as low as 0.20, an MSE of 0.092 respectively, and a CP equal to 0.94. [13] Yunrong Xiang et.al proposes a forecasting model to predict water quality. the model is constructed using LS-SVM algorithm for analyzing the water quality of the Liuxi River in Guangzhou. They have implemented least squares support vector machines along with particle swarm optimization for time series prediction. The result of this experiment shows improvement in the efficiency of prediction. Also demonstrated his efficient prediction value using simulation testing the model.

3. MATERIALS & METHODS

3.1 Dataset

Data is acquired from the TamilNadu Pollution Control Board, it monitors the quality of water under two programs namely Indian National Aquatic Resources and Global Environment Monitoring Systems from 1984. it monitors 32 stations along with four major rivers (Cauvery, Tamirabarani, Palar, and Vaigai) and seven lakes (Kodaikanal, Porur, Poondi, Redhills, Veeranam, Yercaud, and Pulicat). The water quality dataset consists of thirty-two parameters,

where twenty-eight parameters are Physico-chemical parameters that are measured for all the monitoring stations. Its parameters include pH (a measure of acidic/basic water).

	river	duration	station	DO	pH	Conductivity	BOD	Nitrate	Nitrite	Sulphate	...	nco	nna	nec	wph	wdo	wbdo	wec	wna	wcc
2	CAUVERY	Jan	Musiri ferrygate	8.3	545.0	2.0	0.500	0.066	170.0	51.0	...	60	100	100	0.0	28.1	23.4	0.9	2.8	16.86
3	CAUVERY	Jan	Bathira\nkaliamman koil	7.5	217.0	2.0	0.097	0.056	79.0	15.0	...	60	100	100	0.0	28.1	23.4	0.9	2.8	16.86
4	CAUVERY	Jan	Sirumugai	7.0	263.0	3.0	0.104	0.063	63.0	18.0	...	60	100	100	0.0	28.1	23.4	0.9	2.8	16.86
5	CAUVERY	Jan	Bhavani sagar	7.0	159.0	2.0	0.045	0.009	43.0	9.0	...	60	100	100	0.0	28.1	23.4	0.9	2.8	16.86
6	CAUVERY	Jan	Bhavani	7.5	423.0	2.0	0.056	0.055	63.0	29.0	...	60	100	100	0.0	28.1	23.4	0.9	2.8	16.86
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
515	CAUVERY	Dec	Thirumukkudal	8.3	755.0	2.0	0.129	0.005	170.0	80.0	...	60	100	100	0.0	28.1	23.4	0.9	2.8	16.86
516	CAUVERY	Dec	Trichy U/S	8.0	483.0	2.0	0.743	0.004	46.0	43.0	...	60	100	100	0.0	28.1	23.4	0.9	2.8	16.86
517	CAUVERY	Dec	Trichy D/S	7.7	570.0	2.4	0.864	0.603	63.0	52.0	...	60	100	100	0.0	28.1	23.4	0.9	2.8	16.86
518	CAUVERY	Dec	Grand Anaicut	7.7	542.0	3.0	0.786	0.352	93.0	44.0	...	60	100	100	0.0	28.1	23.4	0.9	2.8	16.86
519	CAUVERY	Dec	Coleroon	7.9	528.0	2.1	0.292	0.008	170.0	49.0	...	60	100	100	0.0	28.1	23.4	0.9	2.8	16.86

518 rows × 34 columns

Fig 2:Normalized Attribute.

3.2MACHINE LEARNING METHODS

Linear regression

Predictive modelling is mainly concerned with making more accurate predictions as possible with minimized error rate, one of the most well-known statistics used for prediction is linear regression. Linear regression is represented as an equation, where a line that fits the relationship between an input variable A and an output variable B with the corresponding coefficients as X. it can be represented in the equation as.

$$B = X_0 + X_1 * A \qquad 1$$

An AdaBoostregressor

An AdaBoost regressor is a meta-estimator of a regression algorithm. It initially fits the original dataset and then the additional copies of the regressor. It automatically updates weights of instances according to the current prediction error. The Adaboost is an efficient statistical method in estimating the quantity from a data sample.

Algorithm of Adaboost regressor

1. Input (Sequence of m values, number of iterations, threshold value).
2. Initialize

Iteration as j

Distribution Dr (j) = 1/m

Error rate Er

3. Iterate

Construct regression model: f (A) -> B for regression problems

4. Error Rate of fr(A):

Er=fr ai-bibinDr(j) 2

3.3 The Procedure for predicting water quality can be summarized as

Step 1: Data acquisition and pre-processing of water quality dataset. Data during 2019 are acquired from the Tamil Nadu pollution control board. The collected data is further pre-processed. Pre-processing includes replacing missed values with zero value and all the metrics are changed to numeric values.

Step 2: building the water quality estimation model using linear-regression and Adaboostregressor. Python code is used for the construction of machine learning models for Linear - regression, and Adaboostregressor.

Step 3: The dataset is dividedrandomly into a trainingand testing test with a ratio of 80% and 20% respectively.

Step 4: The training set and testing set are Fed into the machine learning model

Step 5: Analysing and estimating the implemented technique based on the correlations of the data. The implemented techniques are estimated using the metrics Mean Square Error and R^2 Score.

Step 6: Evaluating the validating the model

3.4 METICS FOR EVALUATING THE MODEL

Mean Squared Error

It is the square average of the difference between the predicted and the Actual values.

$$MSE = \frac{1}{n}\sum_{j=1}^{N}(predicted - Actual)^2 \quad 3$$

R^2 Score

It is statistical measure that indicates the amount of variation of a dependent variable that fits is described by an independent variable(s) in a regression model.

$$R^2 \text{ Score} = 1 - \frac{unexplained\ Variation}{Total\ Variation} 4$$

4. EXPERIMENTAL RESULTS AND DISCUSSION

To demonstrate the performance of the regressor model, an experiment was done using both training as well as testing datasets. One-step ahead of prediction, comparison of the performance of the regression, and adaboostregressor methods are also adopted. During the experiments, WQI

(Water Quality Indicator) is computed considering the parameters pH, dissolved oxygen, total coliform, conductivity, nitrate, B.D.O(Biochemical Oxygen Demand) and Phosphate. Parameter values are normalized for the range required for drinking/irrigation water. WQI value for every month of the year 2019 is computed as shown in fig3.

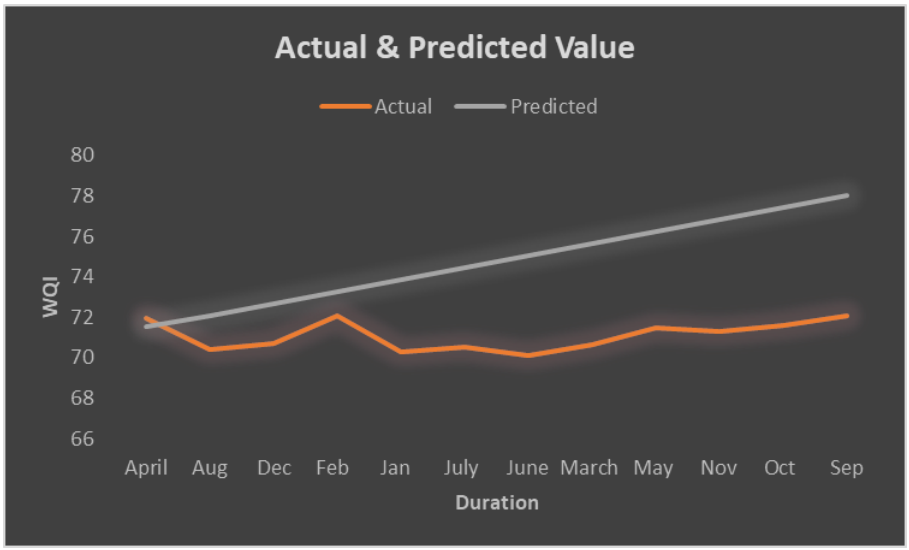

Fig 3: Actual and Predicted WQI value with respect to duration

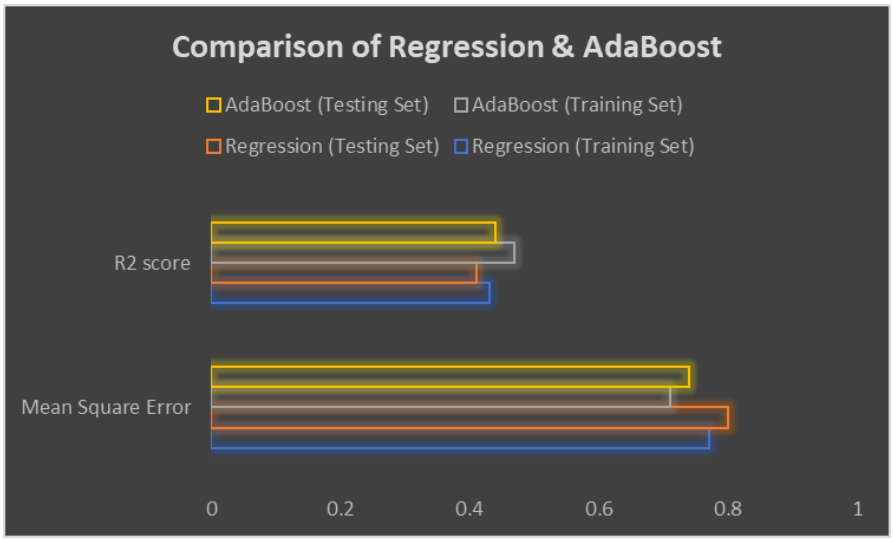

Fig 4: Comparison of Mean Square error value and R^2 Score value for Regression and Adaboostregressor

As shown in Fig 3 the WQI value is higher in September while lower in April.Theoutputs of the implemented techniques are assessed using Mean Square Error and R^2 score the results reflect that AdaboostRegressor predicts the (Water Quality Indice) WQI rate with a Mean Square Error value of 0.8, and R^2score rate is 0.41, whereas AdaBoostRegressor with a obtains Mean Square Error

(MSE) rate as 0.74 and R^2 score rate as 0.44. As shown in Fig4, the performance of AdaBoostRegressor is proved to have better efficiency compared to the Linear regression model in water quality prediction.

5. CONCLUSION& FUTURE WORK

In this work, two machine learning methods Linear regression and adaboostregressor are used for forecasting the quality of water. For the experiment the Tamil Nadu water quality dataset (open source) is considered as the research subject, where the water quality data belongs to the year 2019, Machine learning models are used to predict WQI value based on the parameter dissolved oxygen, pH value, phosphate, nitrate, conductivity, etc.. and the error rate of predicted output are compared and assessed to determine the finer water quality prediction, model. The results indicate: that AdaboostRegressor predicts the (Water Quality Indice) WQI rate with a Mean Square Error value of 0.8, and the R^2 score rate is 0.41, whereas AdaBoostRegressor with a obtains Mean Square Error (MSE) rate as 0.74 and R^2 score rate as 0.44.

In future various machine learning models can be considered therefore to get more accurate prediction results. Also planned to identify and consider other factors that can have the possibility of affecting the quality of water.

6. REFERENCES

[1] Amir Hamzeh Haghiabi,et. al “Water quality prediction using machine learning methods”, Water Quality Research Journal, doi: 10.2166/wqrj.2018.025, 2018

[2] Dominic A. Libera, et al., "A non-parametric bootstrapping framework embedded in a toolkit for assessing water quality model performance", Environmental Modelling & Software, Volume 107, 2018, Pages 25-33.

[3] ShuangyinLiu,et al., “ A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction“, Mathematical and Computer Modelling, Volume 58, Issues 3–4, August 2013, Pages 458-465.

[4] Balamurugan P, et al., “Dataset on the suitability of groundwater for drinking and irrigation purposes in the Sarabanga River region”, Tamil Nadu, India. Data Brief. 2020 Feb 7;29:105255. doi: 10.1016/j.dib.2020.105255. PMID: 32099882.

[5] Ye Liu, Yu, et.al., “Urban Water Quality Prediction based on Multi-task Multi-view Learning”, Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI 2016, June 2016,

[6]Kulthanan, K et al., “The pH of water from various sources: an overview”, recommendation for patients with atopic dermatitis. *Asia Pacific allergy, 3*(3), 155–160. https://doi.org/ 10.5415/apallergy.2013.3.3.155.

[7] Hongfang Lu, Xin Ma, et al., “Hybrid decision tree-based machine learning models for short-term water quality prediction, Chemosphere, Volume 249, 2020.

[8] M Valdivia-Garcia, et al, “predicted Impact of Climate Change on trihalomethanes Formation in Drinking Water treatment”, Scientific reports, 2019

[9] Jafari, H., Rajaee, et al, “Improved Water Quality Prediction with Hybrid Wavelet-Genetic Programming Model and Shannon Entropy”, *Nat Resource Res* (2020).

[10] Sahoo, G. B., Ray, et al., “Application of artificial neural networks to assess pesticide contamination in shallow groundwater” .*Science of the Total Environment,367*, 234–251.

[11] Di Wu, Hao Wang, Razak Seidu, “Smart data driven quality prediction for urban water source management”,Future Generation Computer Systems,Volume 107,2020,Pages 418-432.

[12]Pietro Boccadoro, et al., “Water Quality Prediction on a Sigfox-compliant IoT Device: The Road Ahead of WaterS”,arXiv:2007.3436**, 2020.**

[13] Y. Xiang and L. Jiang, "Water Quality Prediction Using LS-SVM and Particle Swarm Optimization," *2009 Second International Workshop on Knowledge Discovery and Data Mining*, Moscow, 2009, pp. 900-904.

[14] Yan, J et.al “A Prediction Model Based on Deep Belief Network and Least Squares SVR”, Applied to Cross-Section Water Quality, *Water* **2020**, Vol *12*, PP.1929.

[15] K.S. Kasiviswanathan, et al., “Constructing prediction interval for artificial neural network rainfall runoff models based on ensemble simulations”,Journal of Hydrology,Volume 499,2013,Pages 275-288.