International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Special Issue: 1 Volume No.7 April – 2023
DOI:10.46647/ijetms.2023.v07si01.105 ISSN: 2581-4621

WATER QUALITY ANALYSIS AND PREDICTION USING MACHINE LEARNING

S Sundar Pandiyan[1], M Vinutha[2], P Sivasai[3,] M Ravi Prakash[4], K Yaswanth[5]
1,2,3,4,5Department of CSE, Annamacharya Institute of Technology & Sciences, Tirupati-517520

Abstract
With the rapid increase in the volume of data on the aquatic environment, machine learning has become an important tool for data analysis, classification, and prediction. Unlike traditional models used in water-related research, data-driven models based on machine learning can efficiently solve more complex nonlinear problems. Additionally, machine learning can provide solutions for water pollution control, water quality improvement. In this project, we describe the cases in which machine learning have been applied to evaluate the water quality in different water environments such as surface water, groundwater, drinking water, sewage, and sea water approaches.
Keywords: water quality index, water quality class, machine learning algorithms.

Introduction

The rapid industrial development has prompted the decay of water quality at a disturbing rate. Furthermore, infrastructures, with the absence of public awareness, and less hygienic qualities, significantly affect the quality of drinking water . In fact, the consequences of polluted drinking water are so dangerous and can badly affect health, the environment, and infrastructures. As per the United Nations (UN) report, about 1.5 million people die each year because of contaminated water-driven diseases. In developing countries, it is announced that 80% of health problems are caused by contaminated water. Five million deaths and 2.5 billion illnesses are reported annually .Such a mortality rate is higher than deaths resulting from accidents, crimes, and terrorist attacks . Therefore, it is very important to suggest new approaches to analyze and, if possible, to predict the water quality (WQ). It is recommended to consider the temporal dimension for forecasting the WQ patterns to ensure the monitoring of the seasonal change of the WQ . However, using a special variation of models together to predict the WQ grants better results than using a single model. There are several methodologies proposed for the prediction and modeling of the WQ. These methodologies include statistical approaches, visual modeling, analyzing algorithms, and predictive algorithms. For the sake of the determination of the correlation and relationship among different water quality parameters, multivariate statistical techniques have been employed . The geostatistical approaches were used for transitional probability, multivariate interpolation, and regression analysis . Massive increases in population, the industrial revolution, and the use of fertilizers and pesticides have led to serious effects on the WQ environments . Thus, having models for the prediction of the WQ is of great help for monitoring water contamination.

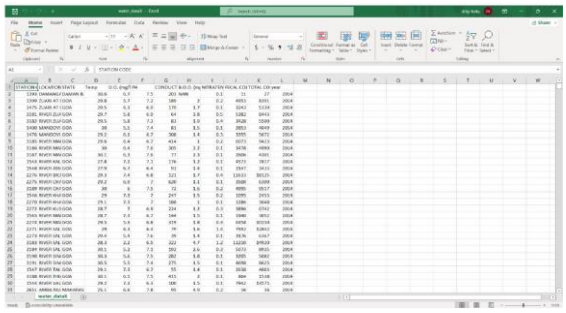
1. Material and Methods:

1.1 Dataset:

This study uses a dataset of medical test results and records from 400 patients, which includes 250 patients with chronic kidney disease (CKD) and 150 patients without CKD. The dataset was prepared in 2015 by Apollo Hospitals, Tamil Nadu, India, and is publicly available at the University of California, Irvine (UCI) machine learning repository. The dataset has 24 factors, such as blood pressure and white blood cell count, and a classification variable indicating whether the patient has CKD or not. It's important to note that the dataset is imbalanced, with more patients having CKD than not.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Special Issue: 1 Volume No.7 April – 2023
DOI:10.46647/ijetms.2023.v07si01.105 ISSN: 2581-4621

Data Collection



1.2 Information Gain:

Feature selection is an important step in building accurate predictive models. By removing attributes that are less useful or unrelated to the target variable, the computational cost is reduced, and the model's performance is improved. In this study, the information gain (IG) technique is used for feature selection. IG is a filter-based method that evaluates the ability of predictor variables to classify the dependent variable

$$IG(X|Y) = H(X) - H(X|Y),$$
$$H(X) = -\sum_{x \in X} P(x) log_2(x),$$
$$H(X|Y) = -\sum_{x \in X} P(x) \sum_{y \in Y} P(x|y) log_2\big(P(x|y)\big),$$

1.3 Water Quality Index

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

Table-1.Parameters weights for the WQI calculation

Weighting factor	weight
pH	0.11
BOD	0.10
Dissolved Values	0.08
Nitrates	0.07

1.4 Water Quality Class

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 2.

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

2. Algorithm

2.1 Random Forest Algorithm

The water quality index is one of the prominent general indicators to assess and classify surface water quality, which plays a critical role in river water resources practices. This research constructs a hybrid

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Special Issue: 1 Volume No.7 April – 2023
DOI:10.46647/ijetms.2023.v07si01.105 ISSN: 2581-4621

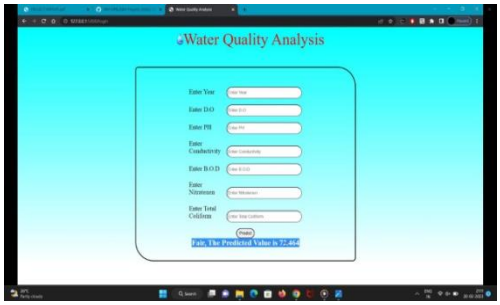
artificial intelligence model namely sequen- tial minimal optimization-support vector machine (SMO-SVM) along with random forest (RF) as a benchmark model for predicting water quality values at the Wadi Saf-Saf river basin in Algeria. The fifteen input water quality datasets such as biochemical oxygen demand (BOD), oxygen saturation (OS), the potential for hydrogen (pH), chemical oxygen demand (COD), chloride (Cl−), dissolved oxygen (DO), electrical conductivity (EC), total dissolved solids (TDS), nitrate-nitrogen (NO3-N), nitrite-nitrogen (NO2-N), phosphate (PO43−), ammonium (NH4+), temperature (T), turbidity (NTU), and suspended solids (SS) were employed for constructing the predictive models. Different input data combinations are evaluated in terms of predictive performance, using a set of statistical metrics and graphical representation. Results show that less than 40% of samples were observed to be poor quality water during the dry season in downstream northeastern part of the basin. The findings also show that the RF model mostly generates more precise water quality index predictions than the SMO-SVM model for both training and testing stages. Although thirteen input parameters attain the optimal predictive performance (R2 testing = 0.82, RMSE testing = 5.17), a couple of five input parameters, e.g., only pH, EC, TDS, T, and saturation, gives the second optimal predictive precision (R2 test = 0.81, RMSE testing = 5.55). The sensitivity analysis results indicate a greater sensitivity by the all input variables chosen except NO2− of the predictive outcomes to the earlier influencing water quality parameters. Overall, the RF model reveals an improvement on earlier tools for predicting water quality index, according to predictive performance and reducing in the number of input variables.

4. Result:

For validating the developed model, the dataset has been divided into 70% training and 30% testing subsets. While the ANN and LSTM models were used to predict the WQI, the SVM, KNN, and Naive Bayes were utilized for the water quality classification prediction.

Performance Measures Results True Positives (TP) are when the model predicts the positive class properly. True Negatives (TN) is one of the components of a confusion matrix designed to demonstrate how classification algorithms work. Positive outcomes that the model predicted incorrectly are known as False Positives (FP). False Negatives (FN) are negative outcomes that the model predicts negative class. Accuracy is the most basic and intuitive performance metric, consisting of the ratio of successfully

predicted observations to total observations. Accuracy=TP+TN/(TP+FP+FN+TN)



5. Conclusion:

In this paper, a new method was suggested to improve the Water quality analysis and prediction using machine learning. Portability determines the quality of water, which is one of the most important resources for existence. Traditionally, testing water quality required an expensive and time-consuming lab analysis. This study looked into an alternative machine learning method for predicting water quality using only a few simple water quality criteria. To estimate, a set of representative supervised machine learning algorithms was used. It would detect water of bad quality before it was released for consumption and notify the appropriate

DOI:10.46647/ijetms.2023.v07si01.105 ISSN: 2581-4621

authorities It will hopefully reduce the number of individuals who drink low- quality water, lowering the risk of diseases like typhoid and diarrhea. In this case, using a prescriptive analysis based on projected values would result in future capabilities to assist decision and policy makers.

6. Future Work:

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 water 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 make

References:

1. P. Zeilhofer, "GIS applications for mapping and spatial modeling of urban-use water quality: a case study in District of Cuiabá Mato Grosso Brazil", *Saúde ...*, vol. 23, no. 4, pp. 875-884, 2007.
2. "Clean water for a healthy world", *Development*, pp. 1-16, 2010.
3. Srivastava, G.; Kumar, P. Water quality index with missing parameters. Int. J. Res. Eng. Technol. 2013, 2, 609–614.
4. PCRWR. Water Quality of Filtration Plants, Monitoring Report; PCRWR: Islamabad, Pakistan,2010.Availableonline:http://www.pcrwr.gov.pk/Publications/Water%20Quality%20Reports/FILTRTAION%20PLANTS%20REPOT-CDA.pdf (accessed on 23 August 2019).
5. Sakizadeh, M. Artificial intelligence for the prediction of water quality index in groundwater systems. Model. Earth Syst. Environ. 2016, 2, 8. [CrossRef]
6. S. Maiti and R. K. Tiwari, "A comparative study of artificial neural networks Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction", *Environ. Earth Sci.*, vol. 71, no. 7, pp. 3147-3160, 2013.
7. Zhao Fu. Jiao Cheng, Mei Yang, Jacimaria Batista, and Yingtao Jiang, "Water quality prediction based on the integration of ANFIS model with intelligence algorithms." Journal paper under preparation.
8. Zhao Fu, Jiao Cheng, Mei Yang, Jacimaria Batista, and Yingtao Jiang, "Wastewater Discharge Quality Prediction using Stratified Sampling and Wavelet De-Noising ANFIS Model," Journal of Computers and Electrical Engineering, Vol. 85, pp. 1-15, Jun. 2020.
9. Zhao Fu, Mei Yang, and Jacimaria Batista, "Using Fuzzy Models and Time Series Analysis to Predict Water Quality," International Journal of Intelligent Systems and Applications (IJISA), Vol. 12, No. 2, pp. 1-10, Apr. 2020
10. Zhao Fu, Jiao Cheng. Mei Yang, Jacimaria Batista, and Yingtao Jiang, "Prediction of Industrial Wastewater Quality Parameters Based on Wavelet De-noised ANFIS Model," IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), 2018, pp. 301-306, Jan. 2018.