OPEN ACCESS

Volume: 11

Special Issue: 1

Month: July

Year: 2023

E-ISSN: 2582-0397

P-ISSN: 2321-788X

Impact Factor: 3.025

Received: 05.05.2023

Accepted:12.06.2023

Published: 01.07.2023

Citation:
Darshan, P., and T. Subburaj. "Evaluation of the Quality of Water Considering SVM and the XGBoost Technique." *Shanlax International Journal of Arts, Science and Humaniteis*, vol. 11, no. S1, 2023, pp. 65–70.

DOI:
https://doi.org/10.34293/sijash.v11iS1-July.6317

Evaluation of the Quality of Water Considering SVM and the XGBoost Technique

Darshan P
Department of Master of Computer Applications
Raja Rajeswari College of Engineering

Dr. T. Subburaj
Department of Master of Computer Applications
Raja Rajeswari College of Engineering

Abstract

Various contaminants have been threatening aquatic excellence for decades. As a result, forecasting and modelling water quality has become serious to reducing water pollution. This study created a classification process for foreseeing water quality (WQC) classifying. By means of a Sustenance Trajectory Appliance and Extreme Gradient Increasing (XGBoost), the WQC is determined by analysing a dataset's water quality indicator (WQI), which is derived from seven parameters. The suggested model's outputs may properly classify in accordance with the qualities of the water. The threesearches of this research showed that the application of X algorithm outscored the SVM the structure which had an accuracy rating of 94% but only 67% specificity. Additionally, the error rate for misclassification for the XGBoost was only 6% as opposed to 33% for the SVM. Additionally, XGBoost regularly outperformed SVM in the 5-fold validation, with an regularaccurateness of 90% as opposed to 64% for SVM.Given its improved presentation, XGBoost is believed to be more effective at classifying the cleanliness of water.

Keywords: Water Quality Prediction, XGBOOST, SVM, Machine Learning.

Introduction

Water is the most imperative resource for life, essential for the survival of most extant species and humans. Living creatures require sufficient quality water to survive. There are some pollution levels that water species can tolerate.

Exceeding these restrictions has an impact on these organisms' existence and risks their life. Most bodies of ambient water, such as rivers, lakes, and streams, have quality criteria that reflect their quality. Furthermore, water requirements for different applications/usages have their own norms. Irrigation water, for example, must not be overly saline or include hazardous elements that might be passed to plants or soil, harming ecosystems. Water eminence for industrial tradition necessitates a variety of qualities reliant on on the unique industrial operations.

Natural water resources include some of the maximum affordable sources of fresh water, such as ground and surface water. Human/industrial activity and other natural processes, however, can contaminate such resources.

As outcome, fast industrial expansion has caused water quality to deteriorate at an alarming rate. Furthermore, facilities with low public awareness and poorer sanitary attributes have a substantial impact on drinking water quality. In reality, the repercussions of dirty drinking water are quite harmful, negatively impacting health, the environment, and infrastructure. Conferring to the World Health Organisation, around 1.5 millionpersons die apieceday as a result of illnesses caused by polluted water. It is estimated that polluted water causes 80% of health problems in underdeveloped nations.

Every year, five million fatalities and 2.5 billion illnesses are documented. It is advised that the temporal dimension be considered when anticipating Water Quality (WQ) trends to ensure the monitoring of seasonal changes in WQ. employing a unique variety of models together to prediction the WQ, on the additional hand, yields better outcomes than employing a single model.

Literature Survey

Prediction of Water Excellence This study discusses the Using artificial intelligence (AI) techniques for predictive modelling of water excellence developed by Theyazn H, H Aldhyani, Mohammed AI-Yaari, Hasan Alkahtani, and MashaelMaashi. The Water Quality Index (WQI) and Water Quality Classification (WQC) algorithms, as employed in our advanced technology, were utilised in this work. Deep learning algorithms such as Nonlinear Autoregressive Neural Network (NARNET) and long Short-Term Memory (LSTM) were utilised. Machine learning methods such as SVM, KNN, and Nave Bayes are also utilised to categorise the WQI. The dataset contains 7 copies that envisage water eminence grounded on improved resilience. With a little modification in the regression coefficient, this approach attained comparable accuracy during the testing phase[1].

Comparative study of hybrid autoregressive neural networks TugbaTaskaya-Temizel and colleagues proposed. Many academics argue in this publication that integrating many models for forecasting harvestsrecoveringguesses than single time series models. It was just shown that integrating the features of every model in a hybrid building design, for instance, made up of a self-regressive mixed marching averages framework (ARIMA) and a neural network that produces accurate predictions. However, by assuming that individual forecasting approaches are acceptable, say, for modelling the residuals, this assumption risks underestimating the link between the model's lined and non-linear components[2].

Ina Khandelwa and colleagues proposed. In this work, we discuss how the The use of disconnected wavelets transforms (DWT) in science as well as engineering has recently skyrocketed. In this study, we demonstrate how DWT might progress the accurateness of prediction of time series. This paper proposes a unique forecasting methodbuilt on separating a stretch series datasets into linear and nonlinear components using DWT. The directional (detailed) and unpredictable (approximate) components of the length of the series' in-sample training set are first separated using DWT. The model of artificial neural networks (ANN) and ARIMA (Autoregressive Integrated Moving Averages) approaches are then utilised to recognise and forecast the rebuilt detailed and approximation components [3].

PulverizedLiquid Quality Prediction Using Machine Learning Algorithms in R by S.Vijay and Dr.Kamaraj, the study details the bore wells from which the samples were taken and how they are broadly utilised for drinking. Water quality metrics counting PH, TDS, EC, Chloride, Sulphate, Nitrate, Carbonate, Bicarbonate, metal ions, and trace elements have been calculated. In the Vellore region, there are two chief types of water contamination: high and low. This research focuses on forecasting water quality with high exactness and efficiency utilising Machine Learning classifier algorithm C5.0, Nave Bayes, and Random forest as leaner for water quality prediction [4].

Jianping Huang et al. proposed this. In this paper, the OTOXI and EUTRO modules of the WASP7.2 model stood used to perform forecast scrutiny on the attenuation condition of the chief water quality control factors within the second-level protection area delineated by experience value, using the reality water volume of 2007 as the validating condition and the forecast water volume of 2010, 2020 as the validating condition, demonstrating the rationality of the second-level protection area delineated by experience value method [5].

Authors Md. Saikaut Islam Khan, Nauzrul Islam, JiaaUddinnsifatull Islam, and Mustafa Kamal Nasirremployed a shift boosted classifiers and the main aspect extraction to predict and categorise the water's condition. The main component regression approach is used in this work to predict water quality. First, the weighted arithmetic index approach is use to calculate the water quality index (WQI). Second, the dataset is subjected to primaryelement analysis (PCA), and the most important WQI parameters are retrieved. Finally, numerousdeteriorationpractices are used to the PCA data to predict the WQI. Finally, the Slope Boosting Classifier is used to organize the national of the water quality. The suggested method is established on a dataset connected to Gulshan Lake [6].

Existing Model

The current method classifies WQC by considering an aquatic quality measure (WQI) from 7 characteristics in a dataset using the Support Vector Machine (SVM) and thrillingincline acceleration (XGBoost). The study's findings confirmed that the XGBoostprototypical outperformed the SVM framework, previously offered only a 67% accuracy, with a precision of 94%[1].

SVM and XGBoost have been used to organize water superiority in the present system; nevertheless, their implementation is difficult. The constraint of restricted water quality limitations and the algorithm's resilience while coping with sounds are currently being researched[2].

SVMs, which were initially created for binary classification issues, employ hyperplanes to construct decision limitations between data points of distinct classes. The hyperplanes are decision functions that differentiate between positive and negative data and have indicated the maximum margins. SVM is regarded a valid classifier for the Hughes effect due to its low sensitivity to feature space dimensions; cataloguing using SVM has little impact on the outcome[3].

XGBoost is a technique that builds a huge number of shallow decision trees, and merging all of them yields a high prediction accuracy. The XGBoost algorithm's decision trees not only minimise an objective function by accounting for the loss function, but they also safeguard the tree from overfitting by employing a regularisation procedure[4].

While SVM and XGBoost are widely employed in machine learning, they may not continuallyproduce correct results for water quality classification. The system's accuracy is restricted by the quality and amount of the data set utilised in the model[5].

SVM and XGBoost demand a large amount of computer power and time to develop and train the model. When dealing with a huge volume of water superiority data, this might be a drawback.

Difficulty in Interpreting Results: Because SVM and XGBoost models are sophisticated and difficult to read, it might be difficult to comprehend how the model classified the water superiority. When attempting to pinpoint particular variables leading to poor water quality, this might be a serious constraint [6].

Limited Feature Selection: Feature selection is an essential part of model construction since the inclusion or absence of specific characteristics can affect the model's accuracy. Because SVM and XGBoost models have restricted feature selection capabilities, they may produce inferior results.

SVM and XGBoost models are known to be sensitive to data imbalance, which means that if one class of water quality is considerably more abundant in the dataset, it may dominate the model's classification, resulting in poor results for the underrepresented classes.Overall, while SVM and XGBoost are prominent machine learning algorithms, their limitations may make them not always the ideal choice for water quality arrangement. It is critical to analyse the application's unique requirements and carefully choose datasets effectively [7].

Proposed Methodology

Using data mining, we create a categorization of the water's puritywith InclineImproving Classifiers in the suggested system. The dataset for this homework was collected from the Kaggle website, which was derived from an Indian government website. The data is relevant for the currenteducation endeavour since it has the parameters needed to develop the water quality catalogue. The marine imminence index may be used to generate a water quality categorization. Data must be pre-processed before being used for training. Data pre-processing refers to finding and fixing mistakes in the datasets that may negatively effect a predictivemodel. The dissolved oxygen (DO), pH, conductive, biologically oxygen demand (BOD), nitrate, faecal coliform, and over-all coliform were used to produce the water quality index (WQI).

The Gradient Improving Classifier model is then trained using the specified features and the computed WQI. A portion of the water quality data is utilised to train the model, while the remainder is used for testing.

The model's accuracy is tested using measures such as Train Accurateness, Test Accuracy, Precisions, Recall, and F1 Score. A confusion matrix is used to assess the effectiveness of water quality categorization. Because this study comprises four water classification classes, a confusion matrix for multi-class classification is used to depict the real classes of the data sets.

Implementations

The main idea behind this algorithm is to build models sequentially and these subsequent models try to reduce the errors of the previous model. But how do we do that? How do we reduce the error? This is done by building a new model on the errors or residuals of the previous model.

When the target column is continuous, we use Gradient Boosting Regressor whereas when it is a classification problem, we use Gradient Boosting Classifier. The only difference between the two is the "Loss function".

Methodology

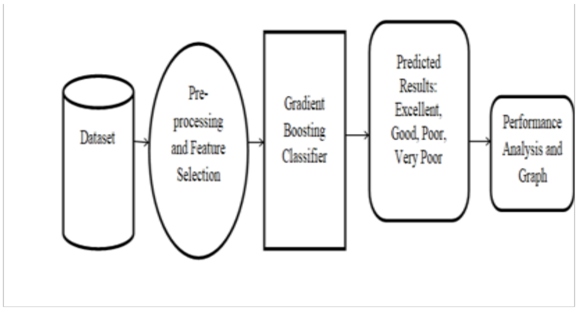

Dataset
Pre-processing and Feature Selection
Gradient Boosting Classifier
Predicted Results: Excellent, Good, Poor, Very Poor
Performance Analysis and Graph

Figure 1 Methodology

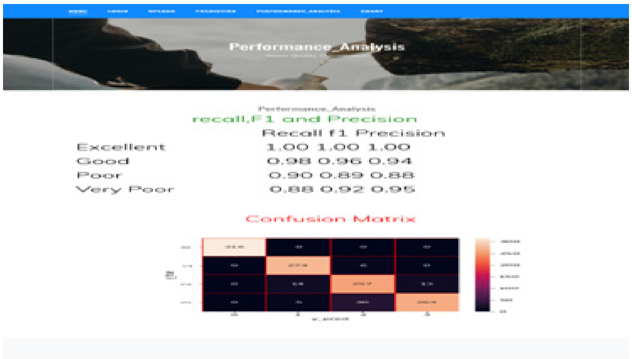
The objective here is to minimize this loss function by adding weak learners using gradient descent. Since it is based on loss function hence for regression problems, we'll have different loss functions like Mean squared error (MSE) and for classification.

Results

Efficient Feature Selection: To properly categorise water quality, the suggested system contains a complete collection of water quality limitations and features, such as pH, liquefied oxygen, disease, turbidity, and electrical conductivity. The method employs feature selection algorithms to discover the furthermostimperative features, ensuring that the model is accurate.

Real-time Water Quality Monitoring: The suggested system may be utilised for real-time water quality nursing, allowing for early detection of changes in water quality and proactive water resource management.

Overall, the suggested method provides a more accurate, efficient, and interpretable approach to water quality categorization, making it an excellent instrument for monitoring and control of water quality.



Performance_Analysis
recall,F1 and Precision

```
                        Recall f1 Precision
Excellent               1.00 1.00 1.00
Good                    0.98 0.96 0.94
Poor                    0.90 0.89 0.88
Very Poor               0.88 0.92 0.95
```

Confusion Matrix

Figure 2 Performance Analysis

This Fig shows High Accuracy: The suggested method achieves 98% Train Accuracy and 94% Test Accuracy. This shows that the system can effectively categorise water into dissimilar quality groups, making it a dependable tool for liquid superiority management.

Conclusions

Water quality modelling and prediction are critical for environmental protection. To predict future water quality, powerful artificial intelligence algorithms can be employed to create a model. The sophisticated artificial intelligence algorithms KNN Algorithm, XG BOOST, Logistic Regression, Random Forest, Decision Tree, and SVM are used in this suggested technique. The system design is sound, the construction is solid, and the functionalities are flawless. A new low-cost, energy-saving, low-power-consumption, adaptable, expandable, and convenient operation and management monitoring system has been deployed.

In future work, we suggest incorporating the conclusions of this study into a large-scale IoT-based online monitoring system employing only the essential parameter sensors. Based on real-time data from the IoT system, the tested algorithms would estimate water quality instantly. The suggested IoT system would use pH, turbidity, temperature, and TDS parameter sensors and communicate those data through an Arduino microcontroller and ZigBee transceiver. It would detect low-quality water before it was released for human consumption and notify the appropriate authorities. It will presumably result in fewer people eating contaminated water, hence de-escalating dreadful illnesses like typhoid and diarrhoea. In this context, a prescriptive analysis based on predicted values might be useful.

References

1. P. Zeilhofer, L. V. A. C. Zeilhofer, E. L. Hardoim, Z. M. Lima, and C. S. Oliveira, "GIS techniques for imaging and spatial modelling of urban-use quality of water: a case study in Cuiabá, Mato Grosso, Brazil," Cadernos de Sade Pblica, vol. 23, no. 4, pp. 875-884, 2007.
2. United Nations Water, "Clean Water for a Healthy World," Technical Report, Development, 2010.
3. Water Quality & Monitoring, K. Farrell-Poe, W. Payne, and R. Emanuel, University of Arizona Repository, 2000, http://hdl.handle.net/10150/146901.
4. Taskaya-Temizel, T., and M. C. Casey, "A comparative study of autoregressive neural network hybrids," Neural Networks, vol. 18, no. 5-6, 2005, pp. 781-789.Visit the Publisher's Website | Scholar on Google
5. C. N. Babu and B. E. Reddy, "A hybrid ARIMA-ANN model based on a moving-average remove for predicting time series data," Applied Soft Computing, vol. 23, pp. 27-38, 2014.
6. Y. C. Lai, C. P. Yang, C. Y. Hsieh, C. Y. Wu, and C. M. Kao, "Evaluation of non-point source contamination and river water quality," C. M. Kao, and C. M. Lai, "Using a multimedia two-model system to improve water quality," Journal of Hydrology, vol. 409, no. 3-4, pp. 583-595, 2011.