Efficient water quality analysis and prediction using machine learning

Literature Survey

- Deepthivarsha E G
- Basireddygari Dhavala
- Anusha R
- Abinaya Kamatchi S

IV year – Computer Science and Engineering, R.M.K Engineering College

Problem statement:

To develop a web application that efficiently analyses the quality of water and predicts whether it is useful for further usage or not using machine learning techniques.

Introduction:

Water is the most essential and precious natural resource. As water is used for various purposes its usability of the water must be checked before using it. The poor condition of water bodies is not only an indicator of environmental degradation but is also a threat to the ecosystem. In industries, improper quality of water may cause hazards and severe economic loss. Thus, water quality analysis is essential for using it for any purpose.

What is Water Quality?

Water Quality can be defined as the chemical, physical and biological characteristics of water, usually concerning its suitability for the designated use.

What is Water Quality Analysis?

Water quality analysis is also called hydro-chemical analysis. That is to use chemical and physical methods to determine the content of various chemical components in water.

Why Water Quality Analysis is required?

Water quality analysis is required mainly for monitoring purposes. Some importance of finding water quality includes:

1. To ensure that it is safe to use as drinking water.

2. To ensure that it can be used for industrial purposes.

3. To ensure whether it is useful for other domestic use.

Literature Survey:

Survey 1:

Water Quality Factor Prediction Using Supervised Machine Learning:

This paper "Water Quality Factor Prediction Using Supervised Machine Learning" proposed a model to explore the prediction accuracy of water quality factors, with techniques and algorithms in machine learning.

The two algorithms Support Vector Regression (SVR) and eXtreme Gradient Boosting (XGBoost) are used. These algorithms predict nine different water quality factors. The factors like dissolved oxygen, pH balance, chlorophyll, temperature, specific conductivity, turbidity, cyanobacteria (blue-green algae), nitrate, and fluorescent dissolved oxygen matter (fDOM). Each factor is individually predicted in both algorithms using the eight other factors. With the implementation of these methods, it is possible to improve the quality of the water parameter data by changing the methods by which the sensors are monitored.

The models used for analysis are:

- Support Vector Regression(SVR):

  Support Vector Regression(SVR) is a supervised learning technique. This method works on the principle of the Support Vector Machine. SVR is a regressor that is used for predicting continuous ordered variables.
- eXtreme Gradient Boosting(XGBoost):

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides a parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

The estimated water quality based on parameters is tested according to World Health Organization (WHO) standards.

Outcomes:

The best outcomes of both algorithms were the prediction of temperature. SVR also predicted dissolved oxygen with good results. Some factors varied by more than 5% between the two algorithms, the most severe difference is seen in the prediction of turbidity, while cyanobacteria and fDOM also had larger gaps in factor prediction. In general, out of the three factors that varied the most, XGBoost favored the more accurate predictions in turbidity and cyanobacteria while SVR had a better accuracy prediction for fDOM.

Advantages:

- XGBoost performs fast, and overall had good predictions.
- SVR boasted great prediction accuracy for two of the nine factors.

Disadvantages:

- XGBoost though performed fast but great results for many of the water quality factors were not predicted.
- SVR was much slower than XGBoost

Both algorithms had a decent performance in most of the factors. It would be a matter of adjusting parameters within both algorithms to see what, if any, of the factors could be better predicted.

Survey 2:

Data Analysis, Quality Indexing and Prediction of Water Quality for the Management of Rawal Watershed

The author developed the model based on the average Linkage (Within Groups) method of Hierarchical Clustering using Euclidean distance is an accurate unsupervised learning technique. Similarly, for classifications, Multi-Layer Perceptron (MLP) has been found to be more accurate supervised learning technique. Higher values of fecal coliforms were found in the months of March, June, July, and October. Some of the possible reasons are land-covers especially scrub forest and rain-fed agriculture areas, poultry farms, and population settled around the streams.

In Part I, the author presented initial month wise quality parameters trends.

In Part II, parametric satisfactory analysis is given in which source-wise and month wise analysis is performed against World Health Organization (WHO) quality standards.

In Part III, we have pre-processed our data and removed all the outliers. This was followed by the development of regression models to check the seasonal water quality trends based on monthly and quarterly datasets.

In Part IV, we found the best quality index using different clustering techniques. In the last phase, we have found the months in which fecal coliforms contamination is high, the streams / sources which have high contamination, and the possible reasons for this higher contamination. Moreover, we know of no previous work whereby data mining has been applied to Rawal watershed data.

Advantages:

WASA is interested in studying the behavior of Bacterio-logical Parameters especially Fecal Coliform which always violate the WHO range as discussed in Parametric Satisfactory Analysis. For that purpose, the author grouped the fecal coliforms data by considering months and sources. These grouped data helped them to find the months and sources in which fecal coliforms are high.

Disadvantages:

The forecasting of fecal coliforms in S1, S9, S11, and S7 streams uses only single time series forecasting models.

Survey 3:

Author: Sillberg

The author have developed a machine learning-based approach integrating attribute-realization (AR) and support vector machine (SVM) algorithm to classify the Chao Phraya River's water quality. The AR has determined the most significant factors to improve the river's quality using the linear function. In the categorization, the most contributing characteristics were: NH3 -N, TCB, FCB, BOD, DO, and Sal, boosting the contributed values in the range of 0.80–0.98, vs 0.25–0.64 for TDS, Turb, TN, SS, NO3-N, and Cond. The SVM linear method has enabled the best classification results represented as the accuracy of 0.94, a precision average of 0.84, recall average of 0.84, and F1-score average of 0.84. The validation showed that AR-SVM was a powerful method to identify river water quality with 0.86–0.95 accuracy when applied to three to six characteristics.

Advantages:

By classifying the water according to their quality, the usage of water will be more efficient.

Disadvantages:

The waters in many supply systems have to be allocated based on past availability or existing consumer demand. The practice does not necessarily mean the allocation is proper. In fact, some supply systems can get overly crowded.

Survey 4:

Author: Yilma

The author have used an artificial neural network to simulate the Akaki River's WQI. The twelve water quality indicators from 27 dry and wet season sample locations were utilized to calculate the index. Except for one upstream location, all forecast results have shown low

water quality. Here, the number of hidden layers (2–20), hidden layer neurons (5, 10, 15, 20, 25), transfer, training, and learning functions were used to train and verify the neural network model through 12 inputs and one output. Their study has revealed that an artificial neural network with eight hidden layers and 15 hidden neurons accurately predicted the WQI with an accuracy of 0.93.

Advantages:

Aquatic life preservation practical

Disadvantages:

There are no specific management plans or sanctions on water extractions in many areas, such as pumping groundwater or rivers. These have caused less water to be soluble and even led to the mining of that resource in some respects. This hampers the water levels and increases the risk of contaminated water.

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. A set of representative supervised machine learning algorithms were employed to estimate WQI. This showed that polynomial regression with a degree of 2, and gradient boosting, with a learning rate of 0.1, outperformed other regression algorithms by predicting WQI most efficiently, while MLP with a configuration of (3, 7) outperformed other classification algorithms by classifying WQC most efficiently.

References:

1. Efficient Water Quality Prediction Using Supervised

Machine Learning Umair Ahmed 1, Rafia Mumtaz 1, Hirra Anwar 1, Asad A. Shah 1, Rabia Irfan 1 and José García-Nieto 2

2. Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms

Md. Mehedi Hassan1, Md. Mahedi Hassan2, Laboni Akter3, Md. Mushfiqur Rahman4, Sadika Zaman1, Khan Md. Hasib5, Nusrat Jahan6, Raisun Nasa Smrity2, Jerin Farhana7, M. Raihan1, Swarnali Mollick8

3. Sakizadeh, M. Artificial intelligence for the prediction of water quality index in groundwater systems.

Model. Earth Syst. Environ. 2016, 2, 8. [CrossRef]

4. Data Analysis, Quality Indexing and Prediction of Water Quality for the Management of Rawal Watershed