Efficient Water Quality Analysis and Prediction using Machine Learning

Introduction

Water is the most crucial resource because it is necessary for all forms of life to exist, but it is also constantly in danger of being contaminated by those same lives. One of the most effective communication tools with a wide range is water. As a result of rapid industrialisation, the quality of the water is rapidly declining. One of the main causes of the spread of terrible diseases is recognised to be poor water quality. According to reports, 2.5 billion people have been unwell and 5 million have died as a result of water-borne diseases, which account for 80% of illnesses in underdeveloped nations. The most prevalent illnesses include giardiasis intestinal worms, diarrhoea, typhoid, gastroenteritis, cryptosporidium infections, several types of hepatitis, and typhoid fever. Since water is a very contagious medium and time is of the essence if water is contaminated with disease-inducing waste, the current method for estimating water quality involves costly and time-consuming lab and statistical analyses that require sample collection, transportation to labs, and a significant amount of time and calculation. This method is quite ineffective. There must be a speedier and less expensive solution given the terrible effects of water pollution. In this research, we investigate various machine learning approaches for predicting water quality with few parameters and talk about the outcomes of regression and classification algorithms in terms of classification accuracy and error rates.

Literature Survey

[1] Water sample collection and laboratory analysis are time and resource intensive processes. Numerous machine learning methods, including multivariate linear regression (MLR) and artificial neural network (ANN) models, have been suggested in the last ten years to solve the issue. It has been demonstrated that the adaptive neuro-fuzzy inference system (ANFIS) is a useful tool for extracting the complex linear and non-linear relationships concealed in datasets. Despite having good performance in predicting water quality, the ANFIS model. Stratified sampling and wavelet denoising techniques are used, and the results are presented together with a comparison of the deep prediction performance of the MLR, ANN, and ANFIS models.
Disadvantages: Because of the complex linear and nonlinear interactions in the water quality dataset, simple linear regression analysis cannot predict water quality with any degree of accuracy. When the input parameters are uncertain, ANN models are unable to articulate the non-linear relationship concealed in the dataset.

[2] In order to establish a reliable strategy for forecasting water quality as accurately as feasible, various AI algorithms are evaluated to handle Water Quality data collected over an extended period of time. The Water Quality data was classified using the Water Quality Index by a number of machine learning classifiers and their stacking ensemble models (WQI). Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), CATBoost, XGBoost, and Multilayer Perceptron were among the classifiers that were examined (MLP). 1679 samples and associated

meta-data were collected over a nine-year period and were part of the dataset used in the study. Additionally, Receiver Operating Characteristic curves (ROC) and precision-recall curves were utilised to evaluate the effectiveness of the different classifiers. **Disadvantages:** In comparison to linear regression, random forests do not offer as much insight into the coefficients. Since the required training time is longer when a large data collection is found, SVM does not perform effectively. The feature selection process for decision trees is automatic, and the model is non-parametric. It costs money because DT takes time to process and train the model. DT is also insufficient when using regression and forecasting continuous values. LRM should not be utilized if there are less observations than features because this could lead to overfitting.

[3] The best model and algorithms for pattern extraction and prediction were evaluated and categorised using the water quality results using Knowledge Discovery in Databases (KDD). The goal of the study is to use data mining techniques to extract information from the dataset for assessing and categorising the water quality based on various characteristics. Four well-known data mining techniques, such as CBA, SVMs, NB, and KNN, are used in a sequence of phases that include data selection of water quality metrics, cleaning, and normalisation. For model prediction and pattern extraction, many methods were used as classifiers. These included Gradient Boosted Trees, Decision Trees, Random Forests, Generalized Linear Models, Naive Bayes, and the Deep Learning algorithm.

[4] The evaluation and forecasting of water quality research findings are reviewed in this publication. In order to analyse water quality, the article categorises and contrasts big data analytics methodologies in use and big data-based prediction models. The structures and networks in the human brain are attempted to be simulated by an artificial **neural network model**. Nodes in the design of neural networks either produce a signal or don't, often according to a sigmoid activation function. Radial basis functions are used in radial basis function network models as activation functions. Radial basis functions of the inputs and neuron parameters are combined linearly as the network's output. The accuracy, robustness, issue kinds, sample size, efficiency, and simplicity of **RBFN** are its overall strengths. An unsupervised learning method called a **Deep Belief Network (DBN)** has many layers of hidden units. Although the units are not connected, the layers are. When trained in an unsupervised manner, DBN can learn to probabilistically recreate its inputs. One of the best methods for knowledge discovery and data mining is the **decision tree** model. Contrary to the prior model, this one employ supervised learning, which links observations about an item to predictions about its intended value. Artificial neural networks and the decision tree method are combined in the **improved decision tree model**. The grouping of data is this approach's key benefit. The **least squares support vector machine model** offers a supervised learning method that involves solving a group of linear equations to arrive at the solution.

Disadvantages: These models have a number of drawbacks, including problems with data quality and validation, the need for research on big data quality assurance, real-time monitoring of water quality, and supervision of water resources. Research and development of real-time water quality monitor and evaluation systems that support water quality evaluation and analysis on various levels, as well as challenges with big data modelling for dynamic water quality monitor and analysis at the different levels for smart cities.

[5] A variety of models, including the Adaptive Neuro-Fuzzy Inference System (ANFIS), Radial Basis Function Neural Networks (RBF-ANN), and Multi-Layer Perceptron Neural Network (MLP-ANN), were used to set up a water quality prediction model for better water resource management. Two scenarios—Scenario 1 and Scenario 2—were presented during these procedures. In Scenario 1, a prediction model is built for each station's water quality characteristics; in Scenario 2, a prediction model is built based on the value of the same parameter at the preceding station. **ANFIS**, which allowed for the realisation of a highly non-linear mapping, is regarded as being more effective than conventional linear approaches for producing non-linear time series. ANFIS is a multi-layer feed-forward network that uses fuzzy reasoning and neural network learning methods to help map the input space to the output space. The **MLP-ANN**, which has multiple layers of neurons, is a feed-forward network because the output of one neuron is transmitted to its neighbouring layer's input. As an alternative, the **RBF-ANN**, which has capabilities comparable to those of the MLP-ANN, is frequently used for stringent interpolation problems in space with multiple dimensions.

[6] This study's objective is to create a water quality prediction model utilising Artificial Neural Networks (ANN) and time-series analysis to incorporate water quality parameters. Historical data on water quality are used in this study. Mean-Squared Error (MSE), Root Mean Squared Error (RMSE), and Regression Analysis are the performance evaluation metrics used to gauge how well the model is doing. ANN has received widespread recognition as a tool for classifying complicated information, including those pertaining to environmental dynamics. It can effectively explain the non-linear relationship between the intricate water quality statistics.

[7] The objective of the study is to analyse and predict the quality of water using machine learning algorithms such as Linear Regression and Stacked Denoising Autoencoder as well as using neural networks such as Deep Belief Networks and Multi Layer Perceptron Network. The data collected for the study are pH, dissolved oxygen, turbidity, chlorine, etc which is collected from Krishna river basin near Chaskaman. After data collection, three clusters were created for 3 seasons: Summer, Winter, Monsoon. On experimenting, turbidity showed to have the highest variability among all the parameters. It is affected during the monsoon season most. pH does not have much variation in data and hence, it is stable compared to dissolved oxygen and turbidity.

[8] The popularity of deep learning (DL), an advanced branch of machine learning (ML) for artificial intelligence, has grown recently. Convolutional neural networks (CNN) and long short-term memory (LSTM) are two popular DL models. The LSTM is a subtype of recurrent neural network (RNN) that stores, processes, and represents extended sequential data in hidden memory. Each neuron in a layer of the CNN is connected to a tiny local region of neurons in the input data by means of a series of convolutional layers. This is accomplished by sliding a filter, which is a weight matrix.

In recent years, hybrid neural networks—which combine the benefits of multiple networks—have drawn more and more attention. Examine how well DL models—LSTM, CNN, and hybrid CNN-LSTM models—can predict variables related to water quality in comparison to more conventional ML models (decision tree and support vector regression). Compare the effectiveness of the LSTM and CNN approaches for predicting short-term changes in water quality. Create a hybrid (CNN-LSTM) model that combines the benefits of the CNN and LSTM approaches.

REFERENCES:

[1] Water quality prediction based on machine learning techniques - Zhao Fu

[2] Efficient Water Quality Prediction Using Supervised Machine Learning - Umair Ahmed.

[3] Water quality classification using machine learning algorithms -

[4] Data-driven Water Quality Analysis and Prediction: A Survey - Gaganjot Kaur Kang, Jerry Zeyu Gao, Gang Xie.

[5] Machine learning methods for better water quality prediction - Ali Najah Ahmeda

[6] Predicting and analyzing water quality using Machine Learning: A comprehensive model - Yafra Khan, Chai Soo See.

[7] Predictive Analysis of Water Quality Parameters using Deep Learning - Khanchan Khare

[8] Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model - Rahim Barzegar, Mohammad Taghi Aalami, Jan Adamowski.