International Journal of Advanced Technology in Engineering and Science
Vol. No. 11, Issue No. 04, April 2023
www.ijates.com
ijates
ISSN 2348 - 7550

Water Quality Prediction Using Machine Learning

K.Kavitha, M. Pavani Sarayu, M.Devadasu , M.Venkata Sekhar

*Under the guidance of Mr. **S. Anil Kumar**, Associate Professor, Project Coordinator,*
Department of Computer Science Engineering, Tirumala Engineering College, Narasaraopet, Andhra Pradesh

ABSTRACT

Potable or drinking water is a daily life necessityfor humans. The safety of this water is a concern in many regions around the world, since polluted waters are increasing and causing the spread of disease among populations. Continuous management and evaluation of the water which is meant for drinking is very essential and must be taken seriously. Often, the quality of water is evaluated through regular laboratory testing and analysis which can be tiresome and time consuming. On the other hand, advanced technologies using big data with the help of machine learning can have better results in terms ofpotability evaluation. For this reason, several studies have been conducted on predicting the quality of water and the severalfactors and classification that affect the prediction model. In this study, a random forest model was developed using PySpark classification to predict the potability of river water by relying onten different features: pH, hardness, presence of solids, presence of chloramines, presence of sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and finally potability. In addition, Thedeveloped model was able to predict water potability classificationwith a 1.0 accuracy, and 1.0 F1-score.

Keywords—Big data; machine learning; classification; random forest; water quality; PySpark

INTRODUCTION

When there is no water, there's no life. Freshwater is the most essential natural resource without which life, all forms of life, would not exist. Humans of all the other living organisms rely on the water not just for drinking but also for various aspects of their lives such as bathing, cooking, and watering their agricultural fields. In fact, there's even increased demand for water due to the increase in wide spreading urbanization, the development and expansion of the economic movement, and the general rapid increase in human population.

However, the water quality and its safety for use for different purposes is a complex issue. The overuse of the water both underground and on the surface in addition to other factors are causing the deterioration of water quality. One of the factors that are having a significant impact on water is the global climate change since it doesn't just affect the availability of water resources but also affects their future quality. Add to that the dangerous pollution resulting from the human activities where individual humans don't only dump their waste into rivers and wells, but also large factories could pollute rivers and underground water as a result of their chemical wastes. As a matter of fact, the poor-quality water is the source of many water-bourne illnesses such as diarrhea. This means that using non-clean water especially for drinking raises health issues that can be avoided but choosing the appropriate water to drink.evaluated to answer the following research questions for a The most common estimation of water quality has been the laboratory analysis which is time-consuming,

expensive, and not very practical. The laboratory analysis of water requires the collection of water samples from different areas over a period of time, then transporting these samples in suitable conditions before they can be analyzed. Of course, this method is still being applied, but with the current development of technology, these processes can be made much more efficient by applying machine learning and big data tools. Machine learning ML is a method of programming software in a way that allows them to learn from historical data and adapt accordingly such that they learn, assess their performance, and improve. Machine learning algorithms are often used to detect patterns in data and the non-visible behavior of data. There are several algorithms already in ML divided into classes: unsupervised, semi-supervised, supervised algorithms.

Random Forest RF is one of the machine learning algorithms through which several decision trees are merged together to achieve more accurate results. The term random forest also corresponds to the randomness of the method where the choice of samples is random. More specifically, a number of samples are randomly chosen from the training dataset in order to form what is called the "root node" samples. Furthermore, the choice of attributes in RF is also random, where the candidate attributes are selected at random, and after that the most suitable attribute is picked to be the "split node". The RF model starts with shuffles input sample data, creates many training sets that make up the decision trees, and finally chooses the output prediction results based on the majority of votes from the collection of decision trees.

In this paper, PySpark for the classification is utilized to evaluate water potability using a well-known Water Quality dataset. The Random Forest Classifier was used to build a model that asses various properties, including temperature, acidity, turbidity, and hardness, to arrive at an accurate decision. The developed model is better understanding of the presented work.

RELATED WORK

Modeling the quality of water resources is vitally important for water scheduling and management. In the past, scientists regularly sampled the water in water quality monitoring stations and assessed the components in the water sample in a lab. However, this process takes a long time, and thus, the detected results are not timely. With the emergence of artificial intelligence (AI) techniques since the last decade, researchers have begun to adopt multivariate linear regression (MLR), artificial neural networks (ANN), adaptive neuro-fuzzy inference system (ANFIS), and Fuzzy time series (FTS) model to predict water quality by exploring the linear and non-linear relationships residing in water quality datasets. In addition, the wavelet denoising method and intelligent algorithms are also proposed to combine with machine learning techniques to enhance the prediction accuracy. In the following, we will review these related work in four categories of machine learning methods.

Multivariate Linear Regression (MLR):

MLR is a kind of statistical analysis method which is used to estimate the target value based on given values collected from a set of independent variables. It is adopted to predict the water quality because of its speed and simplicity. In the MLR model is used to predict biochemical oxygen demand (BOD) and chemical oxygen demand base on four independent variables, temperature, pH, total suspended solid, and total suspended. The system quickly receives relatively good result in BOD prediction with a correlation coefficient value of 0.5.

MLR model has also been used to predict the water quality index in and found to be reliable in formulating the relationship excluding the parameter chloride. However, the MLR model can only be used to formulate linear relationship. It is likely to have a large prediction error if the MLR model is used to predict non-linear relationship.

Artificial Neural Networks (ANN):

Various ANN models have been designed to predict water and wastewater discharge quality based on previous existing datasets. A two-layer ANN model has been applied to predict the DO concentration in the Mathura River, and the experimental result showed that the ANN model worked well. In various neural network types are compared in predicting water temperatures in streams. A radial basis function neural network has also been proposed to describe the water quality parameters. The summary of the experiment result shows the model outperforms the linear regression model in conductivity, turbidity, and total dissolved solids prediction. A time series prediction model, namely the autoregressive integrated moving average, was integrated with the ANN model to improve the prediction performance. The experimental results showed that the hybrid model provided better accuracy than ARIMA and ANN models. Additionally, a comprehensive comparison between ANN and MLR models in biochemical oxygen demand and chemical oxygen demand prediction has been performed. The experimental results show that a three-layer neural network model outperforms an MLR model. The other comparison between ANN and MLR models in water quality index prediction furtherly proves that the ANN model is a better option. Although ANN models can effectively improve the prediction accuracy of water quality parameters, shortcomings still exist. Especially in some scenarios where the input parameters are ambiguous, neural networks struggle to formulate a non-linear relationship. In wavelet transformation was applied to the ANN model to improve the prediction accuracy of a variety of ocean water quality parameters. An integration of a particle swarm optimization algorithm with ANN models has also been investigated to improve the forecasting performance. In 9120 data samples, collected from 2002 to 2012, are used to verify whether the integration of fuzzy logic and ANN models can improve the water quality prediction performance. The experimental results confirm that the proposed method works.

Fuzzy time series (FTS):

A water quality data is a kind of time series dataset which is likely to have complicated linear and nonlinear relationships. The Fuzzy time series (FTS) model was first proposed by Song and Chissom in 1993 to address an enrollment prediction problem [34]. Chen improved this model by replacing complicated max-min composition operations with simplified arithmetic operations. In a Heuristic Gaussian cloud transformation was integrated with an FTS model to forecast water quality. The experimental results showed that the proposed model significantly improved the prediction accuracy. However, there were only 520 water quality samples available to build the cloud, and thus, the model was not reliable or robust. Time series analysis is also 12 proposed to address dissolve oxygen prediction, and the experimental results show that the proposed analysis method can find out valuable knowledge from water quality historical timeseries data. In this dissertation, MLR, ANN, ANFIS, and FTS models are integrated with statistical analysis, wavelet denoising, and intelligence algorithm to explore the prediction of water quality.

Adaptive Neuro-Fuzzy Inference System (ANFIS):

Many studies have proven that ANFIS, which can integrate linear and non-linear relationships hidden in the dataset, is a better option in this scenario. The experimental results as show that an ANFIS model works much better than an ANN model in predicting dissolved oxygen, even though there are only 45 data samples available. An ANFIS model with eight input parameters is used to predict total phosphorus and total nitrogen, the experiment result based on 120 water samples shows the proposed model is reliable. The ANFIS model has also been applied to estimate the biochemical oxygen demand in the Surma River . The testing results from 36 water samples confirmed that the ANFIS model could accurately formulate the hidden relationship and correlation analysis can improve the prediction accuracy. Two different kinds of ANFIS model, fuzzy c-means and subtractive clustering-based was compared with the experiment result shows the ANFIS model built by fuzzy c-means provides more accurate prediction result. In the ensemble models of wavelet ANNs are found to be superior to the best single model for forecasting chlorophyll and salinity concentrations in coastal water. An ensemble of ANN and ANFIS is proposed in to improve the prediction performance of the ANN and ANFIS model, the test result shows there is a significant improvement in the Ensemble ANN-ANFIS model. According to the developer of the ANFIS model, the size of the training dataset should be no less than the number of training parameters. In the aforementioned papers, though the ANFIS models have received higher prediction accuracy, the sizes of the training datasets are 10 me scenarios, especially when the input data have a large value range and there exist some extreme data value points, an out-of-range error is likely to occur, which happens when the testing dataset cannot find any insight from the training model. A few out-of-range errors can cause a very large prediction error, even though the model can accurately predict most of the data samples. In a dataset collected from 122 wells in Mashhad plain (Iran) is used to investigate the performance of ANFIS, ANN, and geostatistical models in groundwater quality prediction. The experimental result shows that the ANFIS model has poor performance in the testing stage because the limited training dataset cannot build a robust or reliable model.

On the other hand, many researchers have also tried to integrate intelligence algorithms with the ANFIS model to improve the performance of the proposed model. An application of genetic algorithm (GA), ant colony optimization for continuous domains, and differential evolution is introduced in to improve the performance of the ANFIS model in predicting parameter electrical conductivity, sodium absorption ratio, and total hardness. The experiment result confirms that the proposed model can improve the performance of the ANFIS model for predicting EC and pH and the root mean square error (RMSE) value of the proposed model in the testing stage is 73.03 and 49.55, respectively. In the genetic algorithm and particle swarm optimization (PSO) algorithm are integrated with the ANFIS model to optimize the threshold bank profile prediction. This method is also used in precipitation modeling. The experimental result indicates that the integrated ANFIS models with hybrid GA/PSO achieve better accuracy than the simple ANFIS model.

METHOD

Supervised learning is the type of machine learning in which machines are trained using well "labelled" training data, and on the basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the sameconcept as a student learns in the supervision of the teacher. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable.

Random Forest is a supervised learning algorithm. It is an extension of machine learning classifiers which include the bagging to improve the performance of Decision Tree. It combines tree predictors, and trees are dependent on a random vector which is independently sampled. The distribution of all trees are the same. Random Forests splits nodes using the best among of a predictor subset that are randomly chosen from the node itself, instead of splitting nodes based on the variables. The time complexity of the worst case of learning with Random Forests is O(M(d*n*log(n))), where M is the number of growing trees, n is the number of instances, and d is the data dimension.

It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest consists of trees. It is said that the more trees it has, the more robust a forest is. Random Forests create Decision Trees on randomlyselected data samples, get predictions from each tree and select the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

Assumptions:

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

1) There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.

2)The predictions from each tree must have very low correlations.

This system is implemented using the following modules.

1) Collection of Dataset

2) Selection of attributes

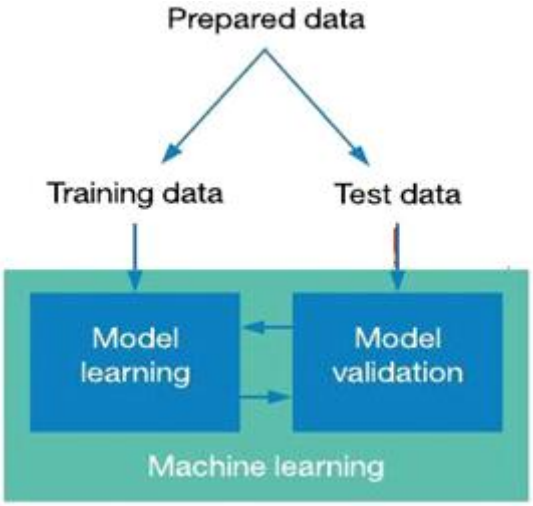
3) Data Pre-Processing

4) Balancing of Data

5) Quality Prediction

COLLECTION OF DATASET:

Initially, we collect a dataset for water quality Assessment. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 80% of training data is used and 20% of data is used for testing. The dataset used for this project is Water Portability. The dataset consists of 8 attributes which used for the system.



SELECTION OF ATTRIBUTE:

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the water are selected for the prediction. The Correlation matrix is used for attribute selection for this model pH and hardness testing data Solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, turbidity, and portability are selected for the prediction. The Correlation matrix is used for attribute selection for this model.

PRE-PROCESSING OF DATA:

Data pre-processing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Preprocessing of data is required for improving the accuracy of the model.

QUALITY PREDICTION:

Various machine learning algorithms like Decision Tree, Random Tree, KNN are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for Water Quality Assessment.

DATA

The chosen dataset comprises a total of ten features that will be used to predict the quality of water and whether itis good for drinking or not.shows the ten values that describe the quality of the water, which are: pH, hardness, presence of solids presence of chloramines, presence of sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and finally potability. The corresponding definitions and values of these features are described below.

pH. The pH metric is an evaluation of the concentrationof hydrogen ions within a solution, and it allows the differen- tiation between acid media, basic media, and neutral mediaas water . The pH recommended by the world health organization determines that the pH of drinkable water must range between 6.52 and 6.83.

Hardness. Hardness is the resultant of both magnesiumand calcium salts that deposit from the geologic surroundingof running water . The period of time in which the wateris in contact with hardness-producing material determines howmuch hardness there is in raw water.

Total Dissolved Solids. Dissolved solids in water referto the salts that can be present including potassium, magne- sium, calcium, bicarbonates, sodium, chlorides, etc. The presence of these dissolved solids in water leads to changingits flavor in addition to affecting its safety . The ideal concentration for TDS is 500 mg/l and should not go above 1000 mg/l for drinking water.

Chloramines. Chloramine alongside chlorine is often used for the treatment of water and disinfecting it from bacteriaand other microorganism. For safety, the amountof chloramine in drinkable water should not exceed 4 mg per liter.

Sulfate. Sulfates are natural elements present in the soil, minerals, food, groundwater, plants, and rocks. Yet they are heavily used in the chemical industry. The sulfate concentra- tion in freshwater should be between 3 and 30 mg per liter .

Conductivity. Electric conductivity is a measure of con- ducting electricity through water. Pure water does not conduct electricity, rather it is considered an insulator . However, ionic water has an increased electric conductivity as a result ofthe ionic compounds in it . The safe electric conductivity level should be less than

400 *µS/cm.*

Organic Carbon. Total organic Carbon TOC resembles the total quantity of carbon from organic matter within the water . This organic carbon can originate from either the decay of natural organic matter or from an unnatural synthetic source. The normal values of organic carbon should be less than 2 mg per liter for drinkable water, and less than 4 mg per liter for the water to be treated.

Trihalomethanes. Trihalomethanes are referred to as THMs in short, and these are molecules abundant in the case of chlorine treatment of water . The factors that affect the amount of THMs are the temperature of treated water, the required chlorine concentration, and the level of organic matter within the water . In order for water to be drinkable, the THM value must be below 80 ppm.

Turbidity. Turbidity is a description of the state of water and whether solids are suspended in it or not . The turbidity of water can be calculated by the light emitting characteristics of water, which represents the quality of waste discharge in re- gard to the colloidal matter. The turbidity value recommended by the World Health Organization is turbidity=5.00 NTU.

Potability. Potability is a term given to describe whether the water is safe for human consumption or drinking or not . In fact, it should also be considered if the same water is good for watering plants. If the given value=1 then the water is potable or drinkable, whereas value=0 means the water is not suitable for consumption.

CONCLUSION

Water Quality monitoring is very much needed as it is consumed by residents. Traditional water Quality monitoring and some of the technology-based Water Quality got lot of challenges. In addition, there is no intelligence in existing water Quality Monitoring for analysis and prediction.

Potability 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 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.

All the seven machine learning methods accuracies are compared based on which one prediction model is generated. Hence, the aim is to use various evaluation metrics like confusion matrix, accuracy, precision, recall, and f1-score which predicts the disease efficiently. Comparing all seven the extreme gradient boosting classifier gives the highest accuracy of 81%.

The safety of our drinking water is a very essential matter, which should be monitored and managed effectively due to its importance. The quality of water that we used for drinking or cooking has a direct effect on our own health, which is why having perfectly safe water is not only a right for humans but also extremely critical. Several protocols and assessment criteria were developed to keep an eye on the safety and potability of water of

different origins (underground, surface, inshore waters, etc.).

Using machine learning and PySpark classification for collection, storage, and analysis of water samples is a much more effective and efficient method for water quality evaluation than regular laboratory tests. This motivated us to create a machine learning model based on the Random Forest algorithmto evaluate the quality of river water based on 10 distinc-tive features: pH, hardness, presence of solids, presence of chloramines, presence of sulfate, conductivity, organic carbon,trihalomethanes, turbidity, and finally potability. The obtained results show that the developed RF model is capable of predicting whether the collected water sample is potable ornot with a 100% accuracy and 1.0 F1-score.

The water quality prediction water quality of their local waterways. This method can be used to collect the necessary data using sensors to perform the prediction. The next step in the development of the water quality prediction model will be to collect the stream data necessary to perform the prediction. This method will be carried out through a dynamic update of the model.

REFERENCES

[1]. M. Kachroud, F. Trolard, M. Kefi, S. Jebari, and G. Bourrié, "Water quality indices: Challenges and application limits in the literature," *Water*, vol. 11, no. 2, 2019. [Online]. Available: https://www.mdpi.com/2073-4441/11/2/361

[2]. O. T. Opafola, K. T. Oladepo, F. O. Ajibade, and A. O. David, "Potability assessment of packaged sachet water sold within a tertiary institution in southwestern nigeria," *Journal of King Saud University*

- *Science*, vol. 32, no. 3, pp. 1999–2004, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1018364720300537

[3]. U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and

[4]. J. Garc´ıa-Nieto, "Efficient water quality prediction using supervised machine learning," *Water*, vol. 11, no. 11, 2019. [Online]. Available:

[5]. https://www.mdpi.com/2073-4441/11/11/2210

[6]. D. Poudel, D. Shrestha, S. Bhattarai, and A. Ghimire, "Comparison of machine learning algorithms in statistically imputed water potability dataset," *preprint*, 2022.

[7]. J. H. Lee, J. Y. Lee, M. H. Lee, M. Y. Lee, Y. W. Kim, J. S. Hyung,

[8]. K. B. Kim, Y. K. Cha, and J. Y. Koo, "Development of a short-term water quality prediction model for urban rivers using real-time water quality data," *Water Supply*, vol. 22, no. 4, pp. 4082–4097, 02 2022. [Online]. Available: https://doi.org/10.2166/ws.2022.038

[9]. M. Kejariwal, C. Patil, A. B. Tiwari, and S. K. Sahani, "Water potability testing case study of mumbai region," *Journal of Harmonized Research in Applied Science*, 2018.

[10]. D. T. Burns, E. L. Johnston, and M. J. Walker, "Authenticity and the Potability of Coconut Water - a Critical Review," *Journal of AOAC INTERNATIONAL*, vol. 103, no. 3, pp. 800–806, 03 2020.

[Online].

[11]. Available: https://doi.org/10.1093/jaocint/qsz008

[12]. E. Ochungo, G. Ouma, J. Obiero, and N. Odero, "Water quality index for assessment of potability of groundwater resource in langata sub county, nairobi-kenya," *American Journal of Water Resources*, vol. 7, no. 2, pp. 62–75, 2019.

[13]. Z. Jamshidzadeh and M. T. Barzi, "Groundwater quality assessment using the potability water quality index (pwqi): a case in the kashan plain, central iran,"