Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

JOURNAL OF INFORMATION SYSTEM AND TECHNOLOGY MANAGEMENT (JISTM)
www.jistm.com

ANALYSIS OF MACHINE LEARNING MODELS FOR EFFICIENT WATER QUALITY PREDICTION

Mahi Aliyu[1], Mansir Abubakar[2*], Armaya'u Z. Umar[3], Alwatben Batoul Rashed[4], Rukayya Musa Ismail[5]

[1] College of Computing and Information Science, Al-Qalam University Katsina, Nigeria
Email: mahigital@gmail.com
[2] Department of Computer Science, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (UiTM), Shah Alam, 40000 Selangor, Malaysia
Email: mansir@uitm.edu.my
[3] College of Computing and Information Science, Al-Qalam University Katsina, Nigeria
Email: azumar@auk.edu.ng
[4] Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia
Email: batool.alwtban@gmail.com
[5] College of Computing and Information Science, Al-Qalam University Katsina, Nigeria
Email: rukayyamismail@auk.edu.ng
[*] Corresponding Author

Article Info:

Article history:
Received date: 26.05.2025
Revised date: 16.06.2025
Accepted date: 10.08.2025
Published date: 14.09.2025

To cite this document:

Aliyu, M., Abubakar, M., Umar, A. Z., Rashed, A. B., & Ismail, R. M. (2025). Analysis Of Machine Learning Models for Efficient Water Quality Prediction. *Journal of Information System and Technology Management, 10* (40), 136-151.

DOI: 10.35631/JISTM.1040010

Abstract:

Water plays a vital role in every aspect of human life, including the metabolism of organisms, industrial manufacturing of goods, and so on. Since water is essential to humanity and is used to improve our way of life, it will be a front-line challenge to humanity if it is heavily contaminated by their activities. In this study, a system that determines an efficient Machine Learning (ML) Model for better water quality prediction is proposed. The performance of the proposed model is evaluated in two different directions: the classification accuracy and the model accuracy in terms of quality (precision) and quantity (recall) of the prediction output. It is identified that the classification accuracy of the Random Forest algorithm appears to be the best model for water quality prediction on the obtained dataset. Random Forest outperforms other algorithms with the best accuracy score of 0.87 as against the XGBoost with 0.86. The model will be useful for treatment plants by automating and training the procedure of determining the quality of the water sample since the water quality of the water sample must be identified to determine the right amount of chemicals to be introduced for the water sample to be potable. In the near future, this work is aimed to be deployed with Flask to provide an interactive interface for users with a non-technical background to use the standard parameters in assessing the quality of water being used for daily activities.

Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

Keywords:

KNN, Logistic Regression, Machine Leaning, Prediction, Random Forest, Water Quality, XGBoost

Introduction

The crucial role of water as the trigger and sustainer of civilizations has been witnessed throughout human history. Human activity on water, such as domestic use and agricultural production, causes water quality deterioration and impacts the aquatic ecosystem and portable water for human consumption. It is estimated that around one billion people lack access to potable water, while 2.5 billion people do not have adequate sanitation (Khatri, N., & Tyagi, S. 2014). Conventional methods for measuring water quality have their drawbacks such as being time-consuming, expensive and inefficient when manual analyses are carried out in a laboratory (Ahmed et, al. 2020), (Dhany Sutadian et al.)

Water and Oxygen are considered vital natural resources for which the living organisms depend; their absence will mean the extinction of the living. However, (Abbasi,T., & Abbasi, S. A. 2012). There are organisms, such as anaerobes, which can survive without oxygen. But no organism can survive for any length of time without water. Water plays a vital role in every aspect of organisms which includes the metabolism of organisms, industrial manufacturing of goods etc. Since water is essential to humanity and is used to improve our way of life it is heavily contaminated by human activities. This also affects humans, animals, plants, the environment and aquatic life (Hounslow, A. W., 1995). The crucial role of water as the trigger and sustainer of civilizations has been witnessed throughout human history, (Abbasi, T., & Abbasi, S. A. 2012). Listing the quantities of every ingredient a water sample contains is one technique to characterize its quality. This list would be created based on the quantity of constituents examined, which may range from hundreds to 20 odd common constituents.

Moreover, only knowledgeable specialists in water quality will find any value in such a list. According to (Tao et al., 2014) when dealing with environmental water pollution or trying to assess the integrity/quality of water, it is necessary to integrate Inorganic water geochemistry and organic geochemical water quality for it to be successful. Geochemistry has contributed greatly to understanding the substances that contribute to the contamination of water sources. (Hounslow, A. W., 1995) In the past, considerable work has been conducted relative to the discovery of metalliferous ore deposits by using geochemical prospecting methods. Now the emphasis is on the transport and fate of various toxic trace metals from known sources. To analyze water quality, it is imperative to understand the inorganic geochemistry and organic geochemistry that influence water sources. Many pollutants in water are constituents of inorganic compounds such as trace metals and industrial waste, hence, the quality of water and its attributes is largely affected by the inorganic compound.

Literature Review

The major contributors to the contamination of water sources are water swages, leakages of petroleum refinery products in storage facilities and toxic trace metals in water sources. The rural populations depend on untreated water sources due to the lack of facilitation to provide adequate water for drinking and usage. Thus, with no other option, the rural population uses the available water sources for consumption and usage. Water sources are mostly contaminated

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010
by impurities and accumulate in victims' systems which may lead to diseases such as kidney stones and Cancer (Li, 2014). Although the quality might be acceptable for drinking, it is unfit for use as a coolant in a commercial setting. While it might work well for some crops, it might not work well for others. It might work well for cattle, but not for raising fish. Geochemists use the term "geochemical spheres" to refer to the many regions of the world that they study, because water is an essential component of the planet. They include the lithosphere (rock), Pedosphere (soil), Biosphere (living organism), Atmosphere(air), Hydrosphere(water), and Anthrosphere (man's effect on the other sphere). The lithosphere significantly affects water quality depending on the environment's permeability. Rocks will determine the permeability of pollutants into a water source, also the decay of soils contributes to the large input of C02 into the atmosphere.

In general, if a specific part of the geochemical sphere is polluted it may affect the quality of water. In other words, water quality refers to the chemical, physical, biological, and radiological characteristics of water. Different parameters are used to assess water quality, and they can be categorized into several key aspects. Regular monitoring of these parameters helps authorities and environmental scientists understand the quality of water in a particular area, detect pollution sources, and implement measures to protect and improve water quality. Regulatory agencies often set standards and guidelines for acceptable levels of various water quality parameters to ensure the safety and sustainability of water resources, but the serious challenge is that how does the daily users of the water comply and get use of these standards in measuring the quality of water the often use every minute. This and other open questions need to be answered in order to save humanity from the danger associated with contaminated water. In this paper, a Machine Learning Model is proposed to predict a water quality, analyses a water sample and outline the contributors affecting the water quality. The proposed model uses data from the United States Geological Survey (USGS) to test the prediction power of the algorithms.

Machine learning prediction models have revolutionized numerous industries by providing powerful tools for prediction and decision-making. Machine learning prediction models have seen significant advancements in recent years, with diverse applications and innovative methodologies pushing the boundaries of what ispossible (Singh, A., R., & Li, X., 2024). In the realm of materials science, machine learning models are being used to predict properties of solid electrolytes based on lattice dynamics. Researchers have developed logistic regression classifiers and random forest regression models that incorporate phonon-related descriptors to predict ionic conductivity. These models have shown high accuracy and potential in identifying promising materials for super-ionic conductors, highlighting the role of machine learning in accelerating computational materials design (Yu, X., et al., 2024) and (Kim, J. et al. 2024).

Recent advancements in machine learning models for water quality prediction have significantly improved the accuracy and efficiency of monitoring water resources (Guan G., 2022). Traditional methods, such as multiple linear regression and auto-regressive integrated moving average (ARIMA), often struggle with the nonlinear and non-stationary nature of water quality data. In contrast, machine learning techniques like artificial neural networks (ANNs), support vector machines (SVMs), and ensemble models such as Random Forest and Gradient Boosting have demonstrated superior performance in capturing complex patterns in water quality datasets.

Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

Deep learning methods, particularly those involving recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, have shown remarkable success in predicting water quality variables like dissolved oxygen, nitrogen, and phosphorus levels. These models are capable of processing temporal sequences and handling the dynamic variations in water quality data, leading to more reliable predictions. For instance, LSTM-based models have been used to develop early warning systems for water pollution risks, showcasing their practical application in environmental monitoring and management (Islam N., Irshad K. (2022).

In practical applications, machine learning models have been employed to predict the Water Quality Index (WQI) in various regions. Studies conducted in areas like Mirpurkhas in Sindh, Pakistan, have utilized extensive datasets and machine learning algorithms, including Random Forest, Gradient Boosting, and SVM, achieving high accuracy rates in WQI prediction (Kayalvizhi S., Jiavana K., Suganthi K., Malarvizhi S., 2023).. These models leverage tools like Python and specialized libraries such as scikit-learn and XGBoost to preprocess data, build, and evaluate classifiers, demonstrating their potential to enhance water quality assessment methods proposed (Irwan D., et al., 2023) significantly. These advancements illustrate the diverse and rapidly evolving nature of machine learning prediction models, with significant implications for both practical applications and theoretical developments in the field as well as across different domains.

Over the years, scholars have strives to find solutions that would support humanity in terms of the quality of water they associate themselves with every minute. Many studies suggest that the quality of water may be good enough for drinking but not suitable for use as a coolant in an industry. It may be good for irrigating some crops but not well for other crops. It may be suitable for livestock but not for fish culture (Altansukh & Davaa, 2011), (Shams et al., 2023). Water is an integral part of the earth, and the geochemical spheres term used by geochemists to describe the various parts of the earth being studied, they include: lithosphere (rock), Pedosphere (soil), Biosphere (living organism), Atmosphere (air), Hydrosphere (water), and Anthrosphere (man's effect on the other sphere), in which Hydrosphere (water) has significant effect on the existence of all other spheres,(Nasir et al., 2022) A study revealed that three factors were determined to be responsible for 66.88% of total variances of water quality in Parsuk River using multivariate statistical analysis including principal component analysis, factor analysis and clusters (Yerel, S. 2010), (Nasir et al., 2022).

The work of Tripetchkul et al., (2019) launched a program to examine the Obulavaripalli Mandal YSR district's drinking water quality using the Water Quality Index (WQI). Twenty groundwater samples and several physio-chemical characteristics were collected to assess WQI in the research region. Thirty percent of groundwater samples meet the excellent classification, forty percent fall into the good category, and thirty percent fall into the poor category according to WQI data. According to the study, the general quality of groundwater is unfit for human consumption. Water pollution sources mainly include phosphate, nitrate, and other chemical pollution copper, cadmium, lead and other heavy metals pollution, even microplastics can be found in contaminated water. Predicting water quality is essential for controlling and preventing pollution of the aquatic environment. Because they don't consider the similarities between parts, this approach cannot consider the association between the water qualities in each segment.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

In a similar study that uses the chemical oxygen demand (COD) of the water environment in the Lanzhou section of the Yellow River, the research object is to build a multi-task deep learning-based water quality prediction model. The model retains its heterogeneity while sharing and learning various correlation sections on water quality. Better information mining from local to whole time series features of water quality is possible with the hybrid CNN-LSTM model. Research demonstrates the root mean square error (RMSE) and mean absolute error (MSE) of the model of the predicted value (Kumar & Dua, 2009). The method performs better in terms of time stability and generalization. The study also discusses the use of fuzzy linear regression in predicting dissolved oxygen levels in a river environment in Calgary, Canada, and the Chambal River Health Index in India.

The work of (Shams et al., 2023) and (Vijay Anand et al., 2023) employ machine learning models based on grid search techniques to forecast water quality. For the health of people, animals, plants, industries, and the environment, water quality is essential. In the last few decades, pollution and contamination have had a major effect on water quality. Anticipating the Water Quality Index (WQI) and Water Quality Classification (WQC) is a difficulty, as WQI is a crucial signal for the validity of the water. The study optimizes and fine-tunes the parameters for four regression and classification models using grid search. There are 1991 cases and seven features in the dataset that was used. The study by (Vijay Anand, Sohitha, Saraswathi, & Lavanya, 2023) in the Journal of Physics: Conference Series discusses the use of Machine Learning for water quality prediction as used in different scenarios (Rashed et al., 2020) The researchers found that the color of water is influenced by the interaction of solar radiation with water level concentration and elements. The alteration of water's color is indicative of the water's properties and its suitability for use. The study highlights the increasing problem of water pollution, which causes 40% of deaths worldwide. Prediction of water quality can be done before consumption using various methods, including filtration and IoT. The study mainly focuses on the primary level of water prediction using Machine Learning. The model is trained using Tensorflow, Keras, and is cost-effective and efficient, and can be used as an immediate and initial level of water quality check. The model can be checked using mobile captured and Google Earth images of water samples.

Many of these models predicts water quality well with the use of some parameters. However, many fail to meet the global standard of measuring the quality of water, and the question is ow the daily consumers of water would comply and get used to the standards in measuring the quality of water they often use every minute. A system is required to appropriately answer this question by providing timely and easy responses to users whenever the need arises.

Method

In this study, analytical and machine learning techniques as used in (Abubakar et al., 2019)is employed to build a water quality model and in the work of (Nurul Shahira et al., 2022) for Predicting Life Expectancy for Asian Population. Similar method is also used in the prediction of Malaysian Women Divorce (Nazim A. et al., 2022). Machine learning becomes remain a technique that is all-encompassing method of prediction across all domain of knowledge (Abdelkader, E. M., Al-Sakkaf, A., & Ahmed, R., 2020), (Guzella, T. S., & Caminhas, 2009). A Comprehensive Comparative Analysis of Machine Learning Models for Predicting Heating and Cooling Loads. Decision Science Letters, 9(3), 409-420. Using analytical methods helps us to get inferences about the data and prepare the data for building prediction models conveniently. Figure 1 shows an overall prediction process flow of our proposed model.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

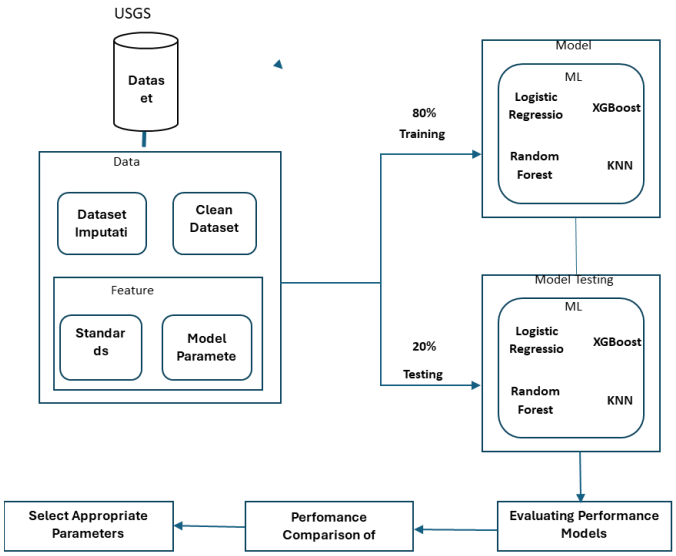

Figure 1: Prediction Process Flow

Dataset

The data used in this study is downloaded from the United States Geological Survey. The U.S. Geological Survey (USGS) Produced Waters Geochemical Database, which contains geochemical and other information for 114,943 produced water and other deep-formation water samples of the United States is a provisional, updated version of the 2002 USGS Produced Waters Database (Breit and others, 2002). The USGS collects the data from 1-1-1905 to 28-02-2014. The data was clean and processed, and the parameters for water quality prediction are selected using the recommended standard provided by the CCME. Features for water quality are extracted from the model to make predictions that will determine the water quality. The data is collected from the 8 regions of the United States: Alaska, West Texas and Eastern New Mexico, Eastern, Mid-Continent, Colorado Plateau and Basin and Range, Gulf Coast, Rocky Mountains and Northern Great Plains, and Pacific Coast.

Preprocessing

In the data set, the main parameters used for water quality are missing, with about 99% missing values, which include turbidity, temperature and total suspended solids attributes. PH has 70% of values since pH is crucial for determining water quality as usual. The threshold for selecting parameters missing values has been set to be 40%. Based on the analysis conducted, the missing value is Missing Completely at Random (MCAR). MCAR is a concept in statistics and data analysis, particularly in the context of handling missing data. When data is missing completely at random, it means that the probability of a particular data point being missing is unrelated to both observed and unobserved data. In other words, the missing data points are a result of a completely random process, and there is no systematic reason for their absence. This randomness implies that the missing data are not systematically related to any variables, observed or unobserved, and therefore, the missing data can be considered as occurring by chance. In statistical terms, if data is missing completely at random, the analysis of the remaining observed data is not biased due to the missing values. Input for Algorithm 1 is raw data from the USGS where an output of clean data is produced. The algorithm's objective is to impute the raw data for preprocessing. It is important to note that the assumption of MCAR is sometimes difficult to verify in practice. However, if this assumption holds, it simplifies the

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

analysis of missing data because it implies that any analysis conducted on the observed data can be applied without introducing systematic bias.

Feature Extraction

Machine learning algorithms are used to predict a certain entity using some set of variables. For this study, we will use water quality parameters to determine the portability of a water source. The data set provided does not have a potable variable, hence, the study cannot continue in the absence of portability. Therefore, using a recommended standard of water parameters that indicate potable water, we can determine the portability of a water sample from the data sets. Figure 2 extracts the feature required for the proposed model. Using the NFSWQI model the algorithm determines the weights of water parameters for water quality such as PH, TDS, Ca, C, Na, SO4, and weightage are used in the NFSQWI model to determine the portability of water samples.

Input: Clean Data After Preprocessing
Output: A dataset with a portability feature for model training

Function weightages(*PH, TDS, Ca, Na, Cl, SO4*):
ViPH $\leftarrow$ 7.5
ViTDS $\leftarrow$ 500
ViCa $\leftarrow$ 1000
ViCl $\leftarrow$ 250
ViNa $\leftarrow$ 200
ViSO4 $\leftarrow$ 500
$k \leftarrow 1/(\frac{1}{\text{ViPH}} + \frac{1}{\text{ViTDS}} + \frac{1}{\text{ViCa}} + \frac{1}{\text{ViCl}} + \frac{1}{\text{ViNa}} + \frac{1}{\text{ViSO4}})$
for each parameter (PH, TDS, Ca, Cl, Na, SO4) **do**
    $Wi \leftarrow k/\text{ViParameter}$
end for
WQI $\leftarrow$ (PH $\times$ WiPH) + (TDS $\times$ WiTDS) + (Ca $\times$ WiCa) + (Cl $\times$ WiCl) + (Na $\times$ WiNa) + (SO4 $\times$ WiSO4)
return WQI
Function quality(x):
if $x \leq 50$ **then**
    return 1
else
    return 0
end if
for each row in dataset **do**
    Calculate WQI using *weightages* function for parameters in the row
    Add WQI to vector *val*
end for
for each WQI value in vector *val* **do**
    Evaluate potability using *quality* function
    Add potability value to the dataset under the column 'potability'
end for

Figure 2: Feature Extraction Algorithm

In the data set, the main parameters used for water quality contains some missing values, these include turbidity, temperature and total suspended solids. PH has 70% of values since pH is crucial for determining water quality for this there will be an exception 40% per cent of missing values has been set as the threshold for selecting parameters. Based on the analysis conducted the missing value has been found to be MCAR (missing completely at random) i.e there is no reason for the missing value.

Parameter Setting

Selecting the right parameters to determine the water quality requires domain knowledge of water quality. The domain knowledge of any field is required to perform machine learning prediction. For this analysis, we have gathered the recommended standard for each parameter in order to accurately ascertain the quality of water with much precision. The percentage of missing values for selected parmeters is shown in table 3.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

Table 3: Missing Values of Selected Parameters

Parameter	Missing Values (%)
PH	24
Total Dissolved Solids	14
Calcium	6
Chloride	5
Sodium	16
Sulfate	19

Recommended Standard

CCME is the primary minister-led intergovernmental forum for collective action on environmental issues of national and international concern based in Canada. is an institution that provided standards for environmental phenomena such as air and water which the living depends on to survive. The recommended standards for the variable acquired is:

Table 4: Recommended Parameters by CMEE

Parameter	Standard (mg/l)
PH	7.5
Total Dissolved Solids	500
Calcium	1000
Chloride	250
Sodium	200
Sulfate	500

Water Quality Index

Experts in the field of chemistry, hydrology and other related field have developed water quality indices (WQI) to determine the quality of water. An individual with no domain knowledge to determine the water quality is provided with several WQI techniques. The standard WQI is shown in table 2.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

Table 5: Water Quality Index (NSFWQI)

S/N	Parameters	Weight
1	Dissolved Oxygen	0.17
2	Fecal Coliform	0.15
3	PH	0.12
4	BOD	0.10
5	N03-	0.10
6	P043-	0.10
7	Temperature	0.08
8	Turbidity	0.08
9	Dissolved Solids	0.18

A commonly used water quality index (WQI) was developed by the National Sanitation Foundation (NSF) in 1970 (Brown and others, 1970). The NSF WQI was developed to provide a standardized method for comparing the water quality of various bodies of water. NFSWQI is a popular WQI, derived from Horton's WQI. This model is used to determine the potability of water in this study.

$$NSFWQI = WiQi \quad (1)$$

Assignment of Weightage

To calculate the water quality index, we need to know how much weight each factor has. Higher allowable limit factors are less dangerous because they can still degrade river water quality in extremely large concentrations. Therefore, the factor's weight and allowable limits are inversely related. (Kumar & Dua, 2010). Therefore;

$$Wi \propto 1Vi \quad (2)$$

Or

$$Wi \propto kVi \quad (3)$$

Were,

$k = cons\tan t\ \ of proportionality$

$Wi = unit\ weight\ factor$

Vi = maximum permissible limits (as recommended by Canadian Council of Ministers of the Environment (CCME)

Value of k is calculated as:

$$k = \sum_{i=1}^{6} \frac{1}{Vi} \quad (6)$$

Rating Scale

The rating of water quality varies from 0 to 100 and is divided into five intervals. The rating Vr = 0 implies that the parameter present in water exceeds the standard maximum permissible limits and water is severely polluted. On the other hand, Vr = 100 implies that the parameter present in water has the most desirable value. The other ratings fall between these two extremes and are Vr = 40, Vr = 60 and Vr = 80 standing for excessively polluted, moderately polluted and slightly less polluted respectively. This scale is a modified version of the rating scale given by (Vijay Anand et al., 2023). Water Quality Index Calculation Essentially, a WQI is a

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

compilation of several parameters that can be used to determine the overall quality of water sample. The parameters involved in the WQI have pH, total dissolved solids, Calcium, Chloride, Sodium, and Sulfate. The numerical value is then multiplied by a weighting factor that is relative to the significance of the test on water quality. The sum of the resulting values is added together to arrive at an overall water quality Index (Kumar & Dua, 2009)

Evaluation Metrics

Accuracy

One popular evaluation criterion for evaluating the effectiveness of classification models is accuracy. It calculates the percentage of cases in the dataset that are correctly classified relative to all instances in the dataset. In terms of math, accuracy is computed as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100\% \quad (7)$$

Precision

The percentage of accurate positive predictions among all positive predictions the model makes is called precision, and it is a classification evaluation statistic. It frequently works in tandem with recollection to offer a more thorough comprehension of the classifier's performance. In mathematics, accuracy is determined as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (8)$$

Precision focuses on the accuracy of positive predictions. It tells us how many of the instances predicted as positive are actually positive. A high precision value indicates that the classifier is making fewer false positive predictions.

Recall

Another crucial categorization assessment statistic is recall, which is often referred to as sensitivity or true positive rate. Out of all actual positive occurrences, it quantifies the percentage of true positive instances that the model properly identifies. In mathematics, recall is determined as follows:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (9)$$

The ability of the model to capture every good experience is the focus of recall. It provides us with the number of real positive examples that the classifier can accurately recognise. Recall values that are high mean that the classifier is doing a good job of reducing false negatives.

F1-Score

Combining recall and precision into a single statistic that strikes a balance between the two is called the F1-score. It is helpful when you wish to consider both false positives and false negatives in your evaluation. It is the harmonic mean of precision and recall. In terms of math, the F1-Score is determined as follows:

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

$$F1 = 2 \times \frac{\text{Precision x Recall}}{\text{Precision + Recall}} \quad (10)$$

where:

A score of 1 indicates perfect precision and recall while a score of 0 indicates that either precision or recall (or both) is 0.

When considering circumstances where both false positives and false negatives are significant, like binary classification problems with imbalanced classes, the F1-score offers a fair evaluation of a classifier's performance. It's critical to consider the criteria and goals of your classification task when interpreting the F1-score. If recall and precision are equally important, the F1-score offers a single statistic to assess the classifier's overall performance. In our situation, we evaluated the accuracy and recall independently because they may be more important for your task, and you should analyse them individually in addition to the F1-score.

Prediction Model

Water quality models are used to simulate the transport and transformation of pollutants in water bodies, and they typically involve a range of physical, chemical, and biological processes as discussed in the introduction section. Several metrics can be used to evaluate the performance of a water quality model. Classification is employed in this paper to test the capability of Machine Learning in addressing this problem. The pseudocode of the model used in training and evaluating the classifiers is presented below.

Input: *Extracted features from Algorithm 1 (training data: Xtrain,ytrainX_{\text{train}}, y_{\text{train}}, test data: Xtest,ytestX_{\text{test}}, y_{\text{test}})*

1. ***Initialize classifiers***:
   - *Logistic Regression (LR)*
   - *K-Nearest Neighbors (KNN)*
   - *Decision Tree (DT)*
   - *Random Forest (RF)*
   - *AdaBoost (ADA)*
   - *XGBoost (XGB)*
2. *Define hyperparameter grids for KNN, DT, RF, ADA, and XGB*
3. *Perform hyperparameter tuning using GridSearchCV or RandomizedSearchCV*
4. *Train a Bagging Classifier with DecisionTree (entropy criterion)*
5. *Train each classifier using training data*
6. *Make predictions on the test set*
7. *Evaluate performance using accuracy score*
8. *Print accuracy results for each classifier*
9. *Specifically evaluate RF model:*
   - *Compute classification report*
   - *Compute precision score*
   - *Compute recall score*

Output: *Trained models and evaluation metrics*

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

Features extracted from algorithm 1 is used to perform model training, which is the objective of algorithm 3. The algorithm starts by splitting the data, 80 % for training and 20% for testing, four classifiers were initialized Logistic regression, XGBoost, Random Forest and K nearest Nearbours. Hyperparameter tuning is performed using grid search to find out the optimal number 25 of parameters of a classifier for a given data set, the models are now trained using the parameters suggested by hyperparameter tuning. Bagging and boosting are applied to the classifiers to reduce variance and bias in finding the best parameters as shown in the excerpt below for all algorithms:

Best parameters for KNN: {'n_neighbors': 7}
Best parameters for Decision Tree: {'criterion': 'entropy', 'max_depth': 29, 'min_samples_leaf':1}
Best parameters for Random Forest: {'min_samples_leaf': 2, 'n_estimators': 200}
Best parameters for AdaBoost: {'learning_rate': 0.2, 'n_estimators': 50}
Best parameters for XGBoost: {'n_estimators': 100, 'learning_rate': 0.2}

These sets of parameters represent the optimal configurations found through some form of hyperparameter tuning for different machine learning algorithms. K-Nearest Neighbors (KNN), Decision Trees, Random Forest, AdaBoost, and XGBoost. KNN is a simple yet effective algorithm that classifies data points based on the majority class among their nearest neighbors. The parameter `n_neighbors` determines how many neighbors to consider when making predictions. In this case, the optimal number of neighbors is found to be 7. Decision Trees partition the feature space into regions, making decisions based on simple rules inferred from the data. `criterion` refers to the function used to measure the quality of a split, `max_depth` controls the maximum depth of the tree to avoid overfitting, and `min_samples_leaf` is the minimum number of samples required to be at a leaf node. In this case, the tree uses the entropy criterion, a maximum depth of 29, and a minimum leaf sample size of 1.

Random Forest is an ensemble learning method that constructs a multitude of decision trees at training time and outputs the mode of the classes as the prediction. `n_estimators` represents the number of trees in the forest, and `min_samples_leaf` is the minimum number of samples required to be at a leaf node. In this case, the optimal parameters include 200 estimators and a minimum leaf sample size of 2. AdaBoost is another ensemble method that focuses on correcting the errors of weak classifiers by iteratively training new models on the misclassified samples. `n_estimators` represents the number of base estimators, and `learning_rate` controls the contribution of each model. In this case, the optimal learning rate is 0.2, and there are 50 estimators. XGBoost is an efficient and scalable implementation of gradient boosting machines. It builds multiple trees in a sequential manner, where each new tree aims to correct the errors made by the previous ones. `n_estimators` represents the number of boosting rounds, and `learning_rate` controls the contribution of each tree. In this case, the optimal learning rate is 0.2, and there are 100 boosting rounds. These parameter configurations are optimized for their respective algorithms and datasets to achieve better performance in terms of accuracy, precision, or other evaluation metrics. However, it is essential to note that the best parameters may vary depending on the problem under investigation and the nature of the dataset.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

Results and Discussion

The performance of the proposed model is evaluated in two different directions, thus, the classification accuracy and the Model accuracy in terms of quality (precision) and (quantity) recall of the prediction output. As presented in table 6, Logistic regression is a simple linear model commonly used for binary classification tasks. An accuracy score of 0.73 indicates that the logistic regression model correctly classified about 73% of the instances in the dataset. While logistic regression is interpretable and efficient, it may struggle with capturing complex relationships in the data compared to more flexible models like random forests or XGBoost. KNN is a non-parametric classification algorithm that classifies instances based on their similarity to neighboring instances. The accuracy score of 0.74 suggests that the KNN model achieved slightly better performance than logistic regression. KNN's performance heavily depends on the choice of the number of neighbors (K) and the distance metric used for calculating similarity.

The Random Forest model with accuracy 0.87 outperforms both logistic regression and KNN significantly. Random forests are known for their ability to handle complex relationships and high-dimensional data, making them a popular choice for classification tasks. XGBoost is another ensemble learning technique that boosts the performance of decision trees. The accuracy score of 0.86 suggests that the XGBoost model achieved slightly lower performance compared to the random forest model but still outperformed logistic regression and KNN as visualized in Figure 4. XGBoost is known for its scalability, efficiency, and effectiveness in handling a variety of data types and structures.

Table 6: Models Classification Accuracy

Model	Accuracy Score
Logistic Regression	0.73
KNN	0.74
Random Forest	0.87
XGBoost	0.86

The results indicate that both ensemble methods, random forest and XGBoost, outperform simpler models like logistic regression and KNN in terms of classification accuracy. It is important to consider other factors besides accuracy in evaluating the model's performance, as such, we further examined the precision, recall, and F1-score to get a comprehensive evaluation of the models regarding water quality prediction.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

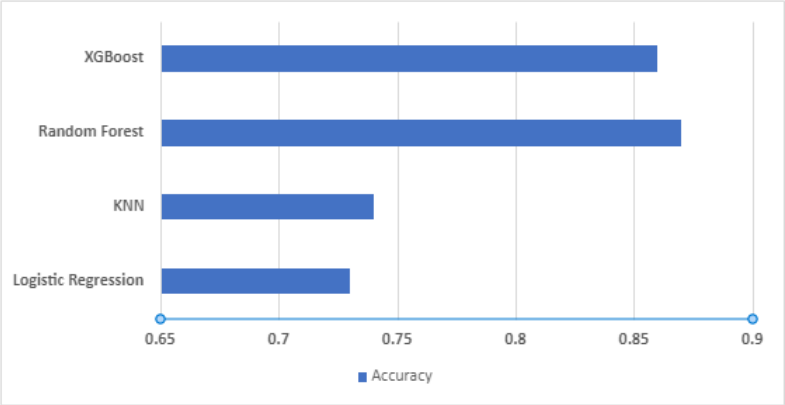

Figure 4: Accuracy Graph

Table 7 shows the performance of the proposed model in terms of precision, recall, and F1-score based on two instance generalization scenarios, referred as Portability. In this context, portability may indicate how well the model performs for class 0 instances and class 1 instances. A precision of 0.91 for class 0 indicates that when the model predicts an instance as class 0, it is correct about 91% of the time, while a precision of 0.76 suggests that when the model predicts an instance as class 1, it is correct about 76% of the time. A recall of 0.92 for class 0 suggests that the model correctly identifies about 92% of all actual class 0 instances in a recall of 0.72 indicates that the model correctly identifies about 72% of all actual class 1 instances.

Table 7: Model Prediction Accuracy

Portability	Precision	Recall	F1-Score
0	0.91	0.92	0.91
1	0.76	0.72	0.74

An F1-score of 0.91 indicates that the model achieves a good balance between precision and recall for class 0 instances while an F1-score of 0.74 suggests that the model achieves a good balance between precision and recall for class 1 instances, though it is slightly lower compared to class 0. However, these metrics provide a more basic understanding of the model's performance beyond accuracy and can help identify areas for improvement, such as addressing imbalanced classes or fine-tuning the model's parameters.

Conclusion

This paper proposed a water quality prediction model using machine learning to accurately determine the portability of water sources. The data has been collected from the USGS and preprocessed the proposed model, trained and tested on the preprocessed data using various machine learning techniques. It has been observed that being water as a vital resource for the survival of living. However, people and other natural constituents affect the quality of the water source, which may result in a water source not being suited for human drinking or other purposes. We obtained data from the United States Geological Survey for water quality to conduct analysis and build a model to predict the quality of water using different machine learning algorithms based on the water parameters used in the system model. Finally, using the classification matrix, we have identified that the random forest model appears to be the best for water quality prediction on the obtained data set.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010

This model is potential for treatment plants by automating and training the procedure of determining the quality of the water sample, as the water quality must be identified to determine the right amount of chemicals to be introduced to make the water sample potable. In the near future, this work is aimed to be deployed with a *Flask* interface to provide an interactive interface for users with a non-technical background to use standard parameters in assessing the quality of water being used for daily activities.

Acknowledgement
The authors would like to acknowledge the support of the College of Computing and Information Science, Al-Qalam University Katsina, Nigeria. They also like to thank the College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia, for providing enabling environment to prepare this article.

References

Abubakar, M., Hamdan, H., Mustapha, N., & Aris, T. N. M. (2019). Bootsrapping instance-based ontology matching via unsupervised generation of training samples. *Journal of Theoretical and Applied Information Technology*, *97*(6).

Abdelkader, E. M., Al-Sakkaf, A., & Ahmed, R. (2020). A comprehensive comparative analysis of machine learning models for predicting heating and cooling loads. *Decision Science Letters, 9*(3), 409–420.

Afthanorhan, A., Mahmud, A., Sapri, A., Aimran, N., Aireen, A., & Rambli, A. (2022). Prediction of Malaysian women divorce using machine learning techniques. *Malaysian Journal of Computing, 7*(2), 1149–1161. doi: https://10.24191/mjoc.v7i2.17077.

Alsaqqar, A. S., Khudair, B. H., & Hasan, A. A. (2023). Application of water quality index and water suitability for drinking of the Euphrates River within Al-Anbar Province, Iraq. *Journal of Engineering, 19*(12), 1619–1633. https://doi.org/10.31026/j.eng.2013.12.10.

Altansukh, O., & Davaa, G. (2011). Application of index analysis to evaluate the water quality of the Tuul River in Mongolia. *Journal of Water Resource and Protection, 3*(6), 398–414. https://doi.org/10.4236/jwarp.2011.36050.

Sutadian, A. D., Muttil, N., Yilmaz, A. G., & Perera, B. J. C. (2016). Development of River Water Quality Indices—A Review. Environmental Monitoring and Assessment, 188(1), Article 58. https://doi.org/10.1007/s10661-015-5050-0.

Guan, G., Wang, Y., Yang, L., Yue, J., Li, Q., Lin, J., & Liu, Q. (2022). Water-quality assessment and pollution-risk early-warning system based on web crawler technology and LSTM. *International Journal of Environmental Research and Public Health, 19*. doi: https://10.3390/ijerph191811818.

Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. *Expert Systems with Applications, 36*, 10206–10222.

Irwan, D., Ali, M., & Ahmed, A. N. (2023). Predicting water quality with artificial intelligence: A review of methods and applications. *Archives of Computational Methods in Engineering, 30*, 4633–4652. doi: https://10.1007/s11831-023-09947-4.

Islam, N., & Irshad, K. (2022). Artificial ecosystem optimization with deep learning enabled water quality prediction and classification model. *Chemosphere, 309*, 136615. doi: https://10.1016/j.chemosphere.2022.136615.

Kayalvizhi, S., Jiavana, K. F. K., Suganthi, K., & Malarvizhi, S. (2023). Prediction of groundwater quality in western regions of Tamil Nadu using deep autoencoders. *Urban Climate, 49*, 101458. doi: https://10.1016/j.uclim.2023.101458.

Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 10 Issue 40 (September 2025) PP. 136-151
DOI: 10.35631/JISTM.1040010
Kim, J., Lee, D., Lee, D., Li, X., Lee, Y., & Kim, S. (2024). Machine learning prediction models for solid electrolytes based on lattice dynamics properties. Retrieved from https://paperswithcode.com/paper/machine-learning-prediction-models-for-solid.

Kumar, A., & Dua, A. (2009). Water quality index for assessment of water quality of River Ravi at Madhopur (India). *Global Journal of Environmental Sciences, 8*(1).

Li, P. (2014). Abbasi T and Abbasi SA: Water quality indices. *Environmental Earth Sciences, 71*(10), 4625–4628. https://doi.org/10.1007/s12665-014-3141-9.

Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., & Al-Shamma'a, A. (2022). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering, 48*, 102920. https://doi.org/10.1016/j.jwpe.2022.102920.

Rashed, A. B., Hamdan, H., Sharef, N. M., Sulaiman, M. N., Yaakob, R., & Abubakar, M. (2020). Multi-objective clustering algorithm using particle swarm optimization with crowding distance (MCPSO-CD). *International Journal of Advances in Intelligent Informatics, 6*(1). https://doi.org/10.26555/ijain.v6i1.366.

Shahira, P. N., Abdul-Rahman, S., Hanafiah, M., & Kamarudin, S. I. (2022). Prediction of life expectancy for Asian population using machine learning algorithms. *Malaysian Journal of Computing, 7*(2), 1150–1161.

Shams, M. Y., Elshewey, A. M., El-Kenawy, E. S. M., Ibrahim, A., Talaat, F. M., & Tarek, Z. (2023). Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-16737-4.

Stojnic, A., Singh, R., & Li, X. (2024). ZLaP: Zero-shot classification with label propagation for vision-language models. *KDnuggets*. Retrieved from https://www.kdnuggets.com/2024/04/5-machine-learning.

Tao, L., Xiaogu, Z., Yongjiu, D., Chi, Y., Zhuoqi, C., Shupeng, Z., Guocan, W., Zhonglei, W., Chengcheng, H., Yan, S., & Rongwei, L. (2014). Mapping near-surface air temperature, pressure, relative humidity and wind speed over Mainland China with high spatiotemporal resolution. *Advances in Atmospheric Sciences, 31*(5), 1–9. https://doi.org/10.1007/s00376.

Tripetchkul, S., Prakirake, C., & Chaiprasert, P. (n.d.). Development of specific water quality index for water supply in Thailand. In *Songklanakarin Journal of Science and Technology, 31*(1). Retrieved from https://www.researchgate.net/publication/26627556.

Vijay Anand, M., Sohitha, C., Saraswathi, G. N., & Lavanya, G. V. (2023). Water quality prediction using CNN. *Journal of Physics: Conference Series, 2484*(1), 012051. https://doi.org/10.1088/1742-6596/2484/1/012051.