DOI:

Journal of Computers and Intelligent Systems
Vol. 2, No. 1, 2024, pages 15 - 30
Journal Homepage:
https://journals.iub.edu.pk/index.php/JCIS/

Prediction Of Water Quality Using Effective Machine Learning Techniques

Uzma Aman[1], Dr. Fariha Ashfaq[1]
[1]Department of Computer Science, Islamia University Bahawalpur, Pakistan

ARTICLE INFO

Article History:

Received:	April	24, 2024
Revised:	April	30, 2024
Accepted:	May	03, 2024
Available Online:	May	05, 2024

Keywords:

Machine learning
Water quality
Environmental sciences
Prediction
Comparative Analysis

Classification Codes:

Funding:

This research received no specific grant from any funding agency in the public or not-for-profit sector.

ABSTRACT

One of the most vital natural resources for all earth's living things is water. Life's fundamental need is access to clean water. Water quality has substantially declined over the previous few decades as a result of pollution and numerous other problems. In this study, machine learning (ML) algorithms are developed to predict water quality and water quality classification (WQC). For the prediction of water quality classification, six machine learning algorithms Naïve Bayes, Random Forest (RF), Gradient Boosting (GBoost), K-nearest neighbor (K-NN),Logistic Regression (LogR), and Decision Tree (DT), have been used. The models were evaluated based on 16 parameters. The machine learning model's result demonstrates the Random Forest model out performed than the other models.

Corresponding Author's Email: fariha.ashfaq@iub.edu.pk

Citation:

1. Introduction

The most prevalent chemical that is continually recycled inside the human body is water [1] . Water is the most crucial resource because all forms of life must exist, but it is also constantly in danger of being contaminated by those same lives[2] [3].The standard of people's drinking water has a significant impact on their health [4]. The availability of water for drinking both for home and industrial use determines a country's level of development [5]. Pollution from both domestic and industrial sources had a greater effect on water quality[6]. Policymakers and public health authorities who conduct activities for the prevention of water pollution and the preservation of public health may find this study interesting [7]. Many people in developing nations are now more likely to suffer from water-related illnesses [8]. Contaminated drinking water may cause very serious consequences that harm people's health, the environment, and infrastructure. A United Nations (UN) estimate [9] states that illnesses brought on by tainted water kill up to 1.5 million people each year. Eighty percent of health problems in developing countries are reportedly caused by contaminated water. Five million deaths and 2.5 billion illnesses are reported annually. Thus, it is critical to propose novel methods for

assessing and, if feasible, forecasting the water quality (WQ)[9]. The assessment and improvement of water sources are of particular importance because of serious health issues connected to the quality of drinking water [10]. Furthermore, early control of intelligent aquaculture requires the ability to forecast changes in water quality [11].Having water of good quality means reducing costs associated with using it for drinking, industrial purposes, and enhancing agricultural output [12]. The wider understanding of this problem is still inadequate, especially in nations where water supply outages are relatively modest[13]. Numerous water management groups that control water have installed monitoring stations to track changes in the water quality status up until the present [14]. It might be feasible to accurately determine the degree of contamination in a water resource using real-time observation of the ensuing variation in water quality [15]. An efficient treatment method should meet several characteristics, such as an easy-to-install, operate, and maintain, low investment operation, maintenance cost, and success in improving water quality [16].To manage a significant amount of missing data and estimate the quality of the water in real-time, Prediction using machine learning (ML) is one of the different methods. When a computer program is tasked with performing a set of tasks, machine learning experts claim that the machine has learned from its experience if its measured performance on those tasks improves over time [17]. A few hybrid approaches and other popular methods allow machine learning problem types to naturally expand [18].The effectiveness of (ML) models that predict water quality may rely on both the models themselves and the parameters in the data set that were selected for developing the machine learning model[19].The use of machine learning (ML) models for mapping hardness susceptibility has not been investigated because of the significance of groundwater quality modeling [20]. Data scientists want to show that no single algorithm is effective in every circumstance[21].The effectiveness of (ML) models that predict water quality may rely on both the models themselves and the parameters in the data set that were selected for developing the machine learning model [19]. Several techniques have been put in for the WQ prediction and modeling. Here, we attempted to assess the six machine learning models' presentations of groundwater prediction: the random forest (RF), Naive Bayes, Gradient Boosting, Decision Tree (DT), Logistic Regression, and the k-nearest-neighbors (KNN models. Several real-world applications of the Naive Bayes (NB) have proven effective, including the classification of medical diseases, consumer credit scores, and weather prediction services[22].

The Random Forest is a multi-classifier ensemble that uses historical data to anticipate class label values[23].The Powerful Gradient-boosting machines are a class of machine learning techniques that have significant success in a range of real-world applications[24].A classification algorithm is the K-nearest neighbor algorithm. The (KNN) technique is to perform feature analysis when it is unclear or difficult to acquire precise parameterization approximations of probability densities. [25].To address a categorization problem, Logistic Regression (LogR)is carried out. Logistic regression is used to address a classification problem, which represents the probability that an occurrence will happen or not, according to the contents of the input parameters in terms of 0 and 1 [17]. A Decision Tree is a Tree-Based method. A classification tree's decision variable is categorical as opposed to a regression tree's continuous decision variable (the outcome takes the form of a Yes/No) [26].

2. Related Work

According to [27], machine learning is an approach to evaluating data that tries to make the assessment model more automatic. Machine learning is a very important tool for the analysis of data, prediction, and classification[28] [29]. Water sinks are ranked according to their level of resilience using a fuzzy analytical hierarchy procedure, and it is used to determine which ones require early refurbishment plans [30] focused. In their work, a model based on three various algorithms for using machine learning for the classification of WQI, the use of ANN will categories water quality data accuracy of 85.11% and a precision of 89.01%. Deep Neural Networks (Deep NN), Support vector machine (SVM), neural networks (NN), and K-Nearest Neighbors (KNN) were used to evaluate water quality, with Deep NN having the highest accuracy (93%) [31]. [32] have proposed a model for forecasting the components of water quality that is based on support vector machines (SVM) and artificial neural networks (ANN). The four water characteristics that make up the suggested methodology are pH, turbidity, total dissolved solids, and temperature. [33] use three models—the SVM, K-nearest neighbor (K-NN), and probabilistic neural network—were created in order to classify the water quality state. Consequently, despite only having 4 water characteristics while input, it has an accuracy rate of 88.37% and an error rate of just 11.63%.

[34] Conductivity and pH are the only two water factors used in the constructed models. Furthermore, to estimates the water quality class, 3 machine learning algorithms, primarily K-NN, Naive Bayes, and SVM, are built[35]. Moreover

[36] created 2 hybrid machine-learning algorithms to forecast the water's short-term quality. The Random Forest and Extreme Gradient boosting models serve as the foundation for the two hybrid models. [37] Use DL, a novel algorithm that can forecast drinking water quality. The [38] directly contrasted their approach for evaluation using 3 machine learning models: ANN, XGBoost, and random forest.

[39] compared a particularly extreme gradient-boosted tree-based ensemble model with other tree-based learning techniques like Random Forest (RF), the CatBoost, and Classifier tree, the results of this study showed that the Light-GBM method for classification performed on par with these techniques, with an accuracy of 85%. [40] gathered samples from bodies of water and examined in labs utilizing a lab methodology. [41] determines the levels of groundwater arsenic pollution in Jharkhand, India, this research provides a machine learning methodology. Additionally, at just 90.11% accuracy, the random forest method performed the best overall. [42] emphasized on the purpose of work is to forecast several aspects of water quality utilizing artificial intelligence (AI) approaches, such as MLP, SVM, and GMDH. Analyzing the results for ANN and SVM showed that both models are effective at foretelling the different components of water quality. The SVM had the greatest result, although the results indicate that the acceptable models worked effectively in predicting the different aspects of water quality. The strategy of [30] is based on just four water quality criteria, as opposed to some suggested works that would employ more than ten to anticipate WQI. Temperature, pH, turbidity, and coliforms are the four water characteristics on which the strategy we suggest is based. They may note that the ANN model has a higher accuracy ranges from 85 to 90% for both training and validation trials.

In a paper [43], the categorization of water samples has been investigated using a variety of tree-based models for machine learning, including SPARC, Optimal Forest, CS Forest, REP Tree algorithms, and random forest. With a low FPR of 73.33%, high accuracy of 80.64%, the precision of 80.70%, recall of 97.87%. [44] debated, the NB model's variables are conditionally independent, making it simple to update, add, or remove data from the network. In this study's model, 64 out of 68 cases were correctly predicted, and it correctly predicts the overall quality class even though some data are absent, which is crucial.

[45] suggested to use recently established regression analysis to estimate the difficult-to-measure characteristics from the simple-to-measure ones. The effectiveness of a support vector machine (SVM) approach was satisfactory in the instance of COD estimation. The paper's primary [46] focus is on the binary classification of water portability using a variety of XG-Boost, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Gaussian Naive Bayes, Random Forest, Light GBM, Bernoulli Nave Bayes, and Gradient Boost. Such popular machine learning methods focus on selected features from the dataset using correlation matrices. Compared to all other machine learning methods, Decision Tree performed the best, achieving an accuracy level of F1 score of 0.9358, AUC-ROC score of 0.9220, as well as F1 score of 0.9374 for the categorization of water portability.

[47], develops accurate and reliable machine-learning methods for irrigation parameters. To reach the assessment, three machine learning (ML) models have been trained: multi-linear regression (MLR), long-short-term memory (LSTM) (ANN), and artificial neural networks (ANN). In a study, [48] examined two distinct classification techniques: the decision tree (DT) as well as K-Nearest Neighbor. KNN scores 61.7% accuracy while decision trees score 58.5%.

3. Used Approach

The current study's suggested methodology is shown in Figure 1, including the opted phases. Each phases is explained in following sections.

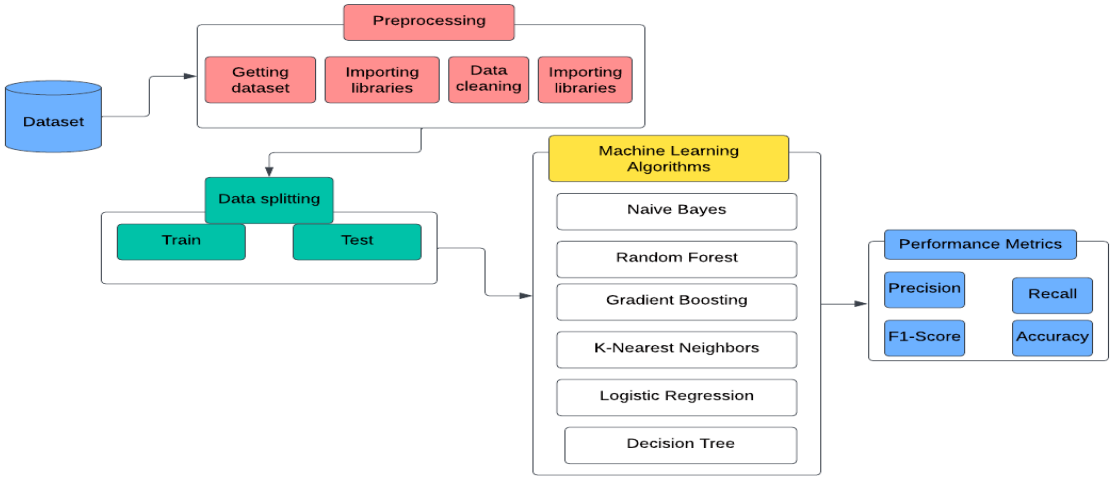

Figure 1: Framework of proposed methodology

3.1. Dataset

We used data that was available to the public on Kaggle.com. The dataset includes multiple (independent) and one (dependent) variable that are connected to the prediction of water quality.

The dataset consists of Water Quality classification which is manually annotated. The dataset is consisting of 7999 instances with 21 parameters. Nominal data types exist for all features. All Features are uniquely identified, and some features have null values, the null values are eliminated and the data is clean after the preparation stage. A supervised ML technique is typically used to obtain the target variable. Depending on the goal and the available data, the target variables may change. Here the target variable showing the water is safe or not safe for health. About the water quality dataset, safe and unsafe data is shown in figure 2.

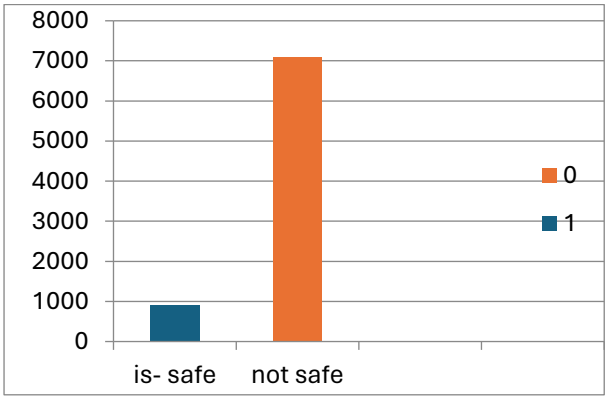

Figure 2: Class attributes

Figure 2 shows the target variables in our dataset, it has 909 safe variables and 7084 are not-safe variables. The dataset features and their dangerous values are displayed in Table 2. Table 2 shows all dataset attributes as well as their dangerous values, indicating that the water is unhealthy and should not be consumed (It means if the values of the features increase from the given range water will be dangerous). (Table 1 information taken from (MSSMSRTYPANTS) where the dataset is downloaded, the link is placed here,

https://www.kaggle.com/datasets/mssmartypants/water-quality)

Table 1: All attributes and their dangerous values

Attributes name	Dangerous values (dangerous if greater than)
Arsenic	0.01
Aluminum	2.8
Cadmium	0.005
Chromium	0.1
Barium	2
Ammonia	32.5
Fluoride	1.5
Copper	1.3
Chloramines	4
Bacteria	0
Lead	0.015
Viruses	0
Bacteria	0
Mercury	0.002
Radium	5
Nitrites	1
Silver	0.1
Perchlorate	56
Nitrates	10
Uranium	0.3
Selenium	0.5
Target variable (Is_safe)	attribute (0 being unsafe, 1 b safe)

3.2. Data Preprocessing

Data preparation, a critical stage in machine learning, enhances data quality and promotes the finding of unobserved patterns and insights. For building and training machine learning models, unstructured data is cleaned up and organized using a process known as data preprocessing. Data preprocessed data increased our machine learning model's accuracy. In the first step, according to the requirement, we collected the Water Quality dataset in a CSV file to perform the machine-learning model. It can use large datasets and makes use of them in programs. It can be downloaded from kaggle.com. Due to Python's popularity and recommendation among data scientists, we have imported Python libraries for use during data preparation for machine learning; here five predefined libraries are used:

1. **Time library**: Obtain real-world time and carry out numerous actions associated with it. By using this module we can manipulate execution time, due to the Time library here not needing to install pip separately. It imported as below:

import time

start_time =time.time()

2. **Pandas library**
   Panda library is an open-source library for data analysis and manipulation. The dataset is imported and managed by this library. It is imported below:

Import pandas as pd

3. **The Matplotlib package** is utilized in the Python code to plot graphs and other types of charts. It imported as below:

Import matplotlib.pyplot as plt

4. **Seaborn library** is used for the visual representation of data. It imported as below:

Import seaborn as sns

5. **Numpy library** is used to perform any numerical calculation and mathematical operation in code. It imported as below:

importnumpy as np

3.3 Importing the datasets

In this phase, the dataset is imported and used the read_csv function as Figure 3.

```
data = pd.read_csv("C:/Users/Uzma Aman/Desktop/uzma/waterQuality1.csv")
```

Figure 3: Importing the dataset

Here the "data" is the name of the variable where data is stored inside the function. After execution of this line, the dataset is imported successfully. The dataset is read using the method of the Jupiter notebook. The Water Quality Evaluation dataset was uploaded. Figure 4 shows the data from the Water Quality evaluation dataset's first five rows.

a)

```
In [5]: data.head(5)
```

Out[5]:

	aluminium	ammonia	arsenic	barium	cadmium	chloramine	chromium	copper	flouride
0	1.65	9.08	0.04	2.85	0.007	0.35	0.83	0.17	0.05
1	2.32	21.16	0.01	3.31	0.002	5.28	0.68	0.66	0.90
2	1.01	14.02	0.04	0.58	0.008	4.24	0.53	0.02	0.99
3	1.36	11.33	0.04	2.96	0.001	7.23	0.03	1.66	1.08
4	0.92	24.33	0.03	0.20	0.006	2.67	0.69	0.57	0.61

5 rows × 21 columns

bacteria	...	lead	nitrates	nitrites	mercury	perchlorate	radium	selenium	silver	uranium	is_safe
0.20	...	0.054	16.08	1.13	0.007	37.75	6.78	0.08	0.34	0.02	1
0.65	...	0.100	2.01	1.93	0.003	32.26	3.21	0.08	0.27	0.05	1
0.05	...	0.078	14.16	1.11	0.006	50.28	7.07	0.07	0.44	0.01	0
0.71	...	0.016	1.41	1.29	0.004	9.12	1.72	0.02	0.45	0.05	1
0.13	...	0.117	6.74	1.11	0.003	16.90	2.41	0.02	0.06	0.02	1

b)

Figure 4: First five rows of the dataset

3.4 Data Cleaning

To begin, we first ingest our dataset and perform some initial cleaning. When we examine data, the first thing we see is a record that is incorrectly labeled and several null values. We discard the entire row with null values because there aren't many missing entries in the dataset.

3.5 Dataset Splitting into the Training Set and Test Set

The splitting process is used to improve our machine learning model's performance. Here we divided the dataset into two parts, the first one is the Training set and the second one is the Test set. Here we use train size 70 and test size 30.

3.6 Machine Learning Algorithm

Data that have been divided into testing and training, the dataset has eventually been submitted to the application of machine learning algorithms. We used six distinct machine-learning algorithms in this investigation.

Every model is put to use. Because the model has a significant impact on prediction accuracy, it is crucial to choose one that works for the dataset. The classifiers given, as described in the introduction, are employed in this research,

1. Naïve Bayes algorithm
2. Random Forest algorithm
3. Gradient-Boosting algorithm
4. K-Nearest Neighbors algorithm
5. Logistic Regression algorithm
6. Decision Tree algorithm

The sequence of algorithms is performed the same as in Figure 5.

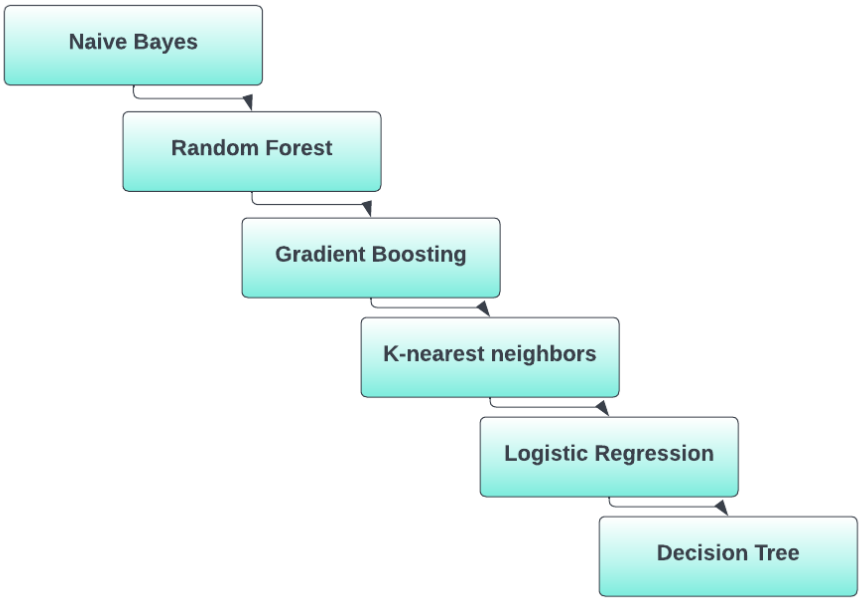

Figure 5: Classification Algorithms

3.7 Performance metrics

Performance metrics like recall, f1-score, accuracy, and precision score are utilized to evaluate how well machine learning models work. These performance indicators are described as follows:

Precision

The model's precision rate is the proportion of labels for which it correctly predicts the outcome. The positive predictive value of a test or prediction indicates how accurately it achieves its goal (also known as precision).Score for precision = TP (FP+TP). True positives and false positives are denoted by TP and FP, respectively, in the equation above [49].

Recall

A model's capacity to differentiate between positive and negative data is measured by the recall score. The true positive rate, which is also frequently referred to as sensitivity, is another recall word. If the recall score of the model is high, the model is successful in choosing the right samples.

Recall Score = TP / (FN + TP)[49].

F-score

Precision and Recall are given equal weight when measuring the accuracy performance of machine learning models using the F-score performance measure (rather than insisting on knowing the total number of observations).

The Score of F1 = 2* Precision Score * Recall Score/ (Precision Score + Recall Score/)[49].

Accuracy score

Model accuracy, which is determined by dividing the total number of positive and negative events by the ratio to accurate classifications, is the statistic used to evaluate the algorithm's performance when it employs a machine learning model.

Accuracy Score is calculated = (TP + TN)/(TP + FN + TN + FP)[49].

4. Results and Discussion

Within the scope of this research, we have utilized a total of six different machine-learning strategies. The accuracy of these six separate approaches remained quite distinct from one another. Data gathered from Kaggle was utilized in the training of models.

Following the training phase, the models are then effectively deployed for the prediction of Water Quality, at which time each model displays its findings.

Implementation of ML models and Confusion matrix

This study's goal is to outline the use of six different strategies that were taken into consideration when conducting the study and the outcomes that were attained using those strategies. Examine the models' instructive examples.

A machine learning classification performance measurement method is a confusion matrix. We used a confusion matrix to evaluate the effectiveness of categorization models given a set of test data. It determines the true values for test data that are known. By this, we know the performances of all classification algorithms. The accuracy view and performance matrix of all classification models is shown.

1 Naïve Bayes

We apply the Naive Bayes classifier as one of our initial models on the Water Quality dataset. Naïve Bayes is binary classification model. The NB Classifier is then trained using a collection of data to predict the Water Quality using the target variable (Is_safe).

The naïve displays an accuracy of 83.43%, In the Naive Bayes performance matrix, the macro averages for accuracy, recall, and f1-score are 0.65, 0.73, and 0.68 respectively, while the weighted averages for precision, recall, and f1-score are 0.88, 0.83, and 0.85 respectively, on the training dataset, The accuracy view and performance matrix of the Nave Bayes classifier are shown in Figure 6.

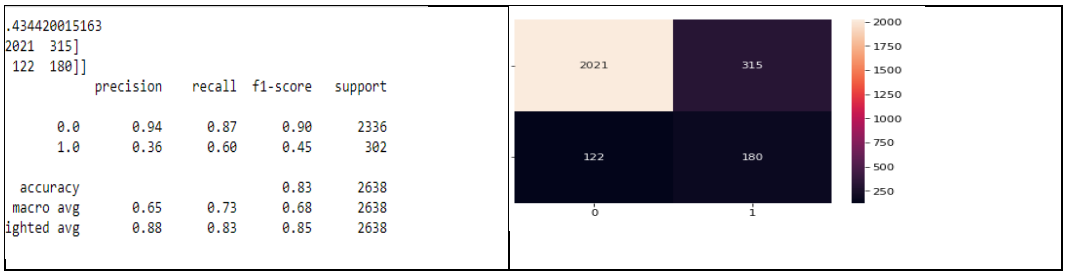

Figure 6: Accuracy view and performance matrix of naive bayes

2 Random Forest

The Random Forest method is the second model we employed for our prediction models. It is a good machine-learning model that provides good results. The Random Forest provides the most accurate results. The Random's precision is 0.94.01. In the Random Forest performance matrix, the macro averages for accuracy, recall, and f1-score are 0.89, 0.79, and 0.83 respectively. The weighted averages for these three metrics are 0.94, 0.94, and 0.94 for precision, recall, and f1-score, correspondingly, on the training dataset. The accuracy view and performance matrix of the Random Forest is shown in Figure 7.

.

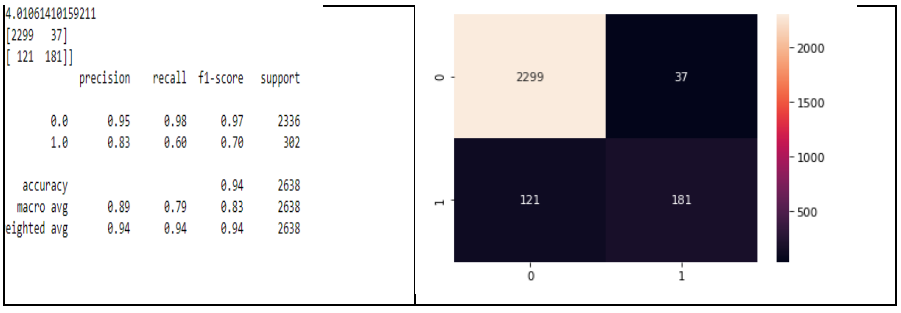

Figure 7: Accuracy view and performance matrix of random forest

3 Gradient Boosting

We used gradient Boosting as our third model. It is quite a good model of machine learning. At 93.85% accuracy, gradient boosting is also quite accurate. The accuracy, recall, and f1-score macro averages in the Gradient Boosting performance matrix are 0.88, 0.79, and 0.83, correspondingly.

The weighted averages for these metrics are 0.93, 0.94, and 0.93 for precision, recall, and f1-score, respectively, on the training dataset. The gradient-boosting performance matrix is shown in detail in Figure 8.

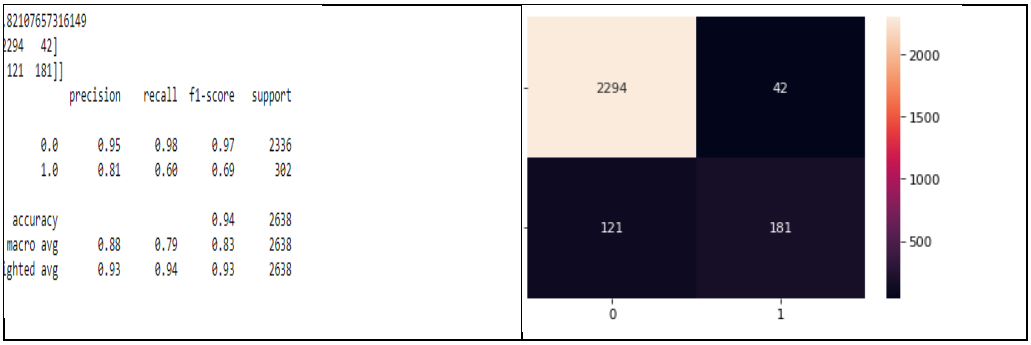

Figure 8: Accuracy view and performance matrix of Gradient Boosting

4 K-nearest neighbors

We used the KNN model after the gradient boosting. KNN is a classification model of machine learning. The K-nearest neighbors' accuracy is 90.44%. In the performance matrix of K-nearest neighbors, the macro averages of precision, recall, and F1 score are 0.80, 0.64, 0.68, and 0.89, 0.90, and 0.89, respectively, when weighted and perform well on the trained dataset. The view accuracy and performance matrices of the K-Nearest Neighbors are shown in Figure 9.

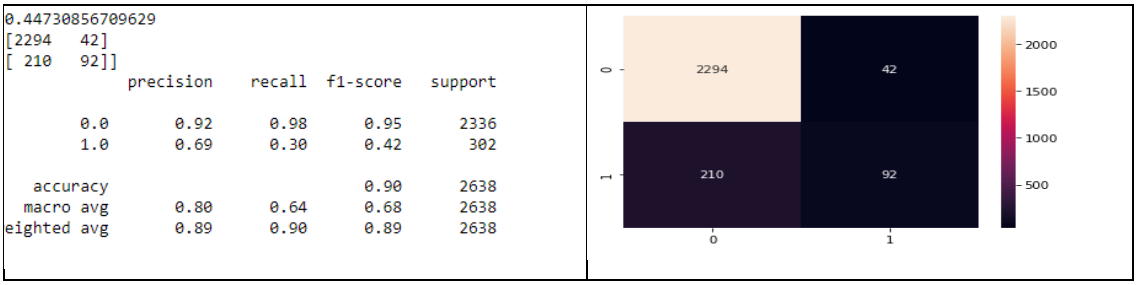

Figure 9: Accuracy view and performance matrix of K-Nearest Neighbors

5 Logistic Regression

We apply logistic regression in our prediction model. The accuracy of the Logistic Regression is 90.25. Inside the performance matrix of logistic regression, the weighted average of precision, recall, and f1-score is 0.89 and these three metrics' average values are 0.79, 0.64, and 0.68 perform admirably using the training dataset. Figure 10 displays the Logistic Regression's performance matrix.

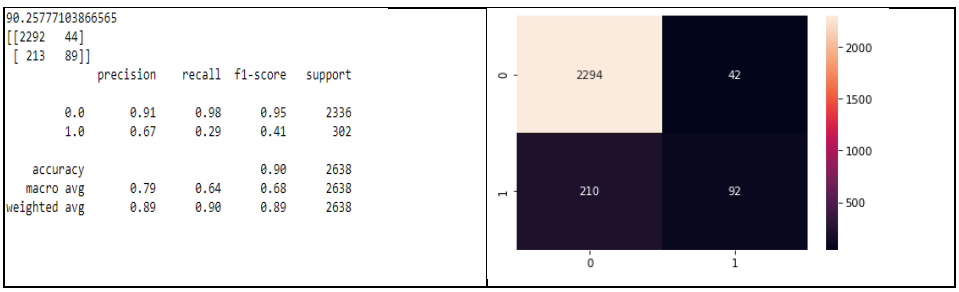

Figure 10: Accuracy view and performance matrix of Logistic regression

6 Decision Tree

At the conclusion of a list lies the Decision Tree. The Decision Tree's accuracy in the Logistic Regression performance matrix is 91.96%. The f1-score is 0.81, recall is 0.83, and precision is 0.80 on the macro averages. The precision, recall, and f1-score weighted averages of 0.92, 0.92, and 0.92, respectively, demonstrate strong performance on the training dataset. Figure 11 displays the accuracy and performance matrix of the Decision Tree viewpoint.

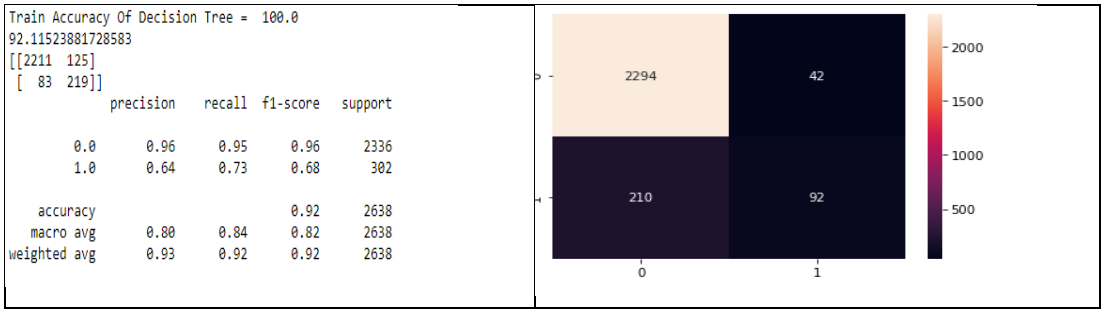

Figure 11: Accuracy view and performance matrix of Decision Tree

4.1. Comparison of different ML models

We have used all six different machine-learning approaches within the confines of this investigation. Throughout the entire procedure, the accuracy of the six separate methodologies remained relatively distinct from one another. Data gathered from the Kaggle "Water Quality" dataset was used to train the models. After the training phase, the models are successfully used to predict the water quality label. Each model now informs the audience of its findings. We examined numerous prediction models in this part based on their accuracy, and precision, then recall f1 scores, and we defended our decision to choose the model that we consider to be the best. The predicting accuracy that is provided by every single one of the above models is side-by-side compared in the accompanying graph. To compare the performance metrics with different classifiers, with the help of Logistic Regression, Gradient Boosting, Naive Bayes, Random Forest, K-nearest Neighbors, and Decision Tree, we used them with a test size of 0.30 and training size at 70. The figures display the experimental results. Figure 12 results of all algorithms shown; the accuracy is explained in each part separately in this figure.

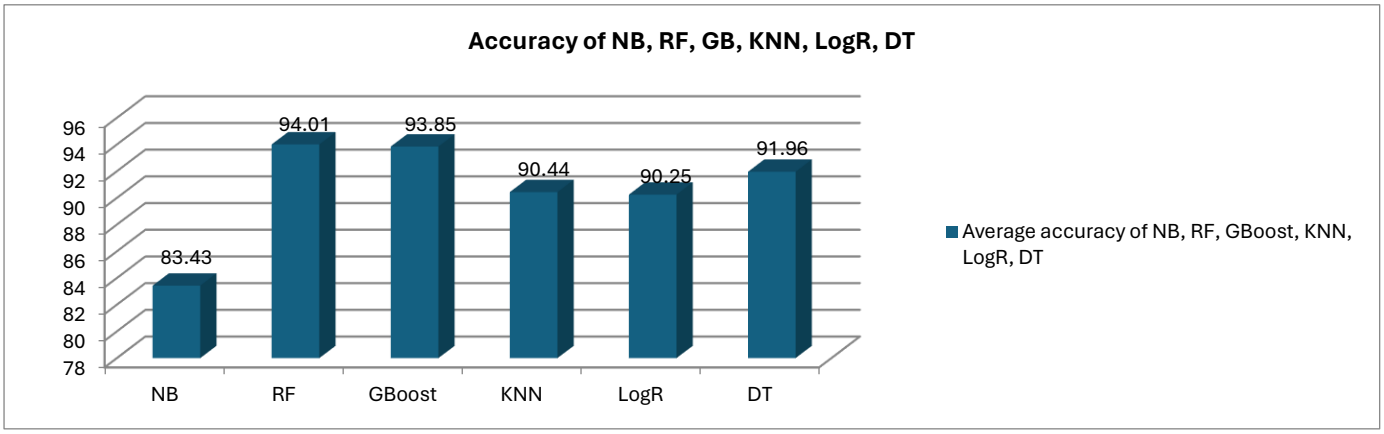

Figure 12: Comparison of different Ml models

4.1.1. Multiple Model's Precision Comparison

Figure 13 depicts a graph of the precision comparison between multiple models. Shows that RF attained 95% Precision and DT produced 96% with a test size of 0.30 and train size of 70.

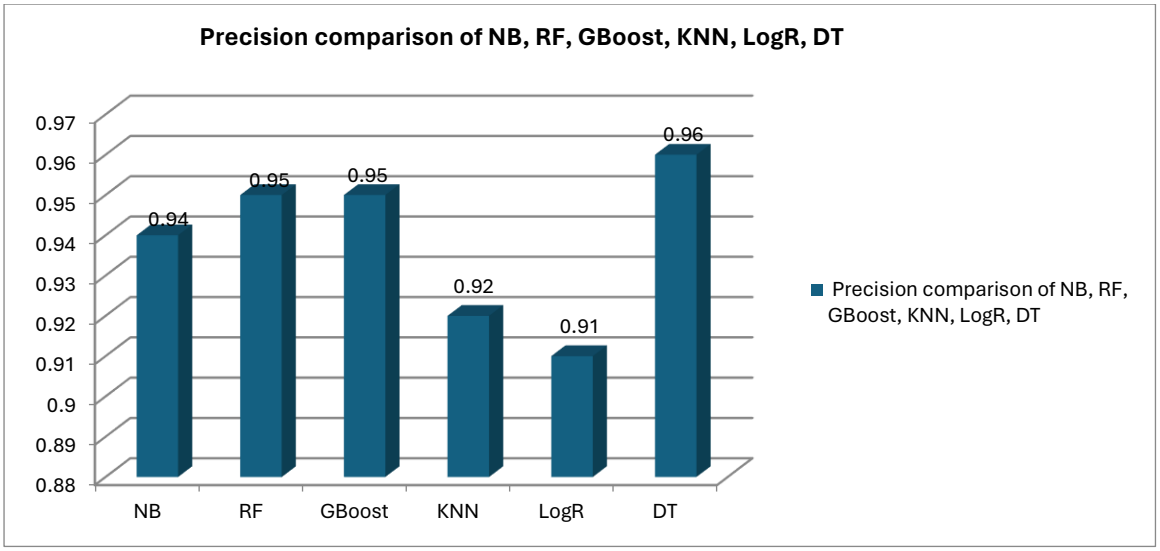

Figure 13: Multiple Models' Precision Comparison

4.1.2. Multiple Model's Recall Comparison

Recall the comparison of different models shows. In Figure 14 RF achieved a recall of 0.99 % at test size 0.30 and train size 70, which yields the ideal results.

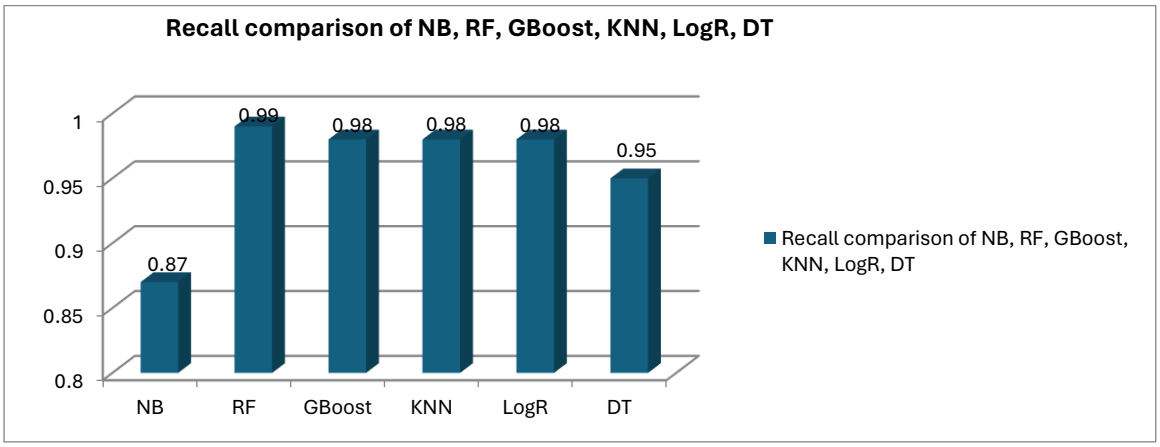

Figure 14: Multiple model Recall Comparison

4.1.3. Multiple Model's F1 score comparison

K-nearest Neighbors, Naive Bayes, Gradient Boosting, Random Forest, Logistic Regression, and Decision Tree have all been compared using the F1 score. Figure 15 shows demonstrate that, with a test size of 0.30, RF and Gradient Boosting provide the best F1 score, which is quite high in comparison to other classifiers. When compared to other classifiers, RF offers the most encouraging results in terms of Recall, the F1 score, and precision.

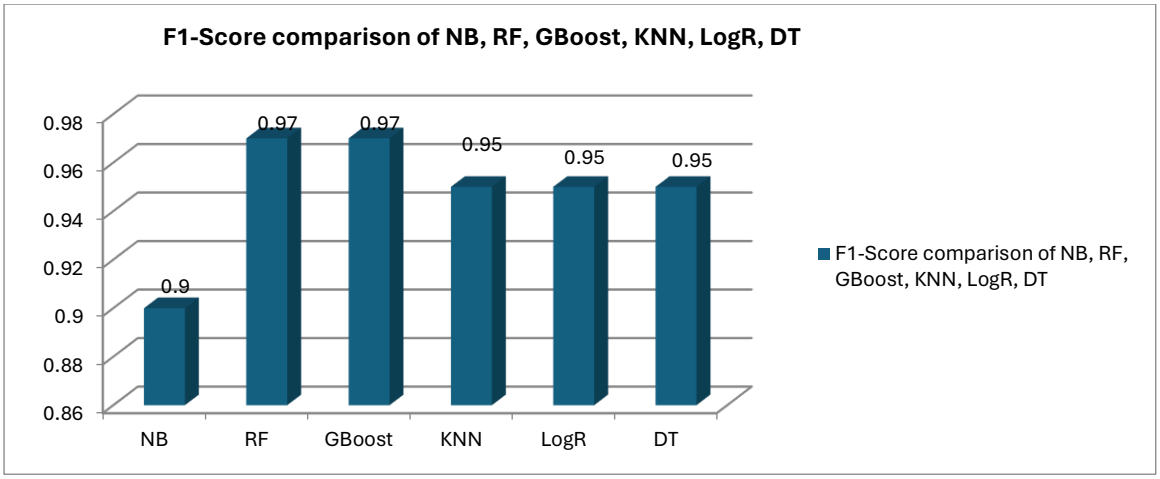

Figure 15: Multiple Models' F1-Score Comparison

4.2. Comparison of parameters and model performance with previous studies

Table 2 compares the model performance and parameters with those from earlier studies, which used some parameters and obtained different levels of accuracy. Our study employed sixteen parameters, which is a significant departure from previous research. Compared to other studies, this one had a greater accuracy rating.

Table 1 Comparison of parameters and performance models with the previous study

References	**Parameters detected**	**Model used**	**Highest Accuracy%**
[31]	Turbidity, PH, Temperature	Deep NN	93
[30]	Temperature, Turbidity, PH,	ANN	85.11
[32]	Temperature, solid, PH, and Turbidity	Multi-layer perceptron	85
[39]	Turbidity	XGBoost	85
[40]	Chloride, sulfate, hardness, alkalinity	ANN	87.91
Our Work	Nitrates, arsenic, barium, mercury, bacteria, cadmium, chloramines, perchlorate, chromium, radium, silver, copper, viruses, lead, aluminum, and nitrites.	RF	94.01

DOI:

Journal of Computers and Intelligent Systems
Vol. 1, No. 1, 2023, pages xx - xx
Journal Homepage:
https://journals.iub.edu.pk/index.php/JCIS/

4.3. Result in comparison with benchmark and this study

[41] Used machine learning techniques inspired by the groundwater contamination caused by arsenic in Jharkhand, India.

To categorize data as safe or unsafe, three algorithms for machine learning trained and evaluated Decision Tree, Random Forest, and Multilayer Perceptron. Based on the results, the Random Forest model outperformed other algorithms the best.

(Proposed work) In this research, a machine learning algorithm based on the Target variable "Is_safe" which is a dependent variable on sixteen parameters namely aluminum, Six machine learning algorithms are used to forecast the quality of water using nitrates, arsenic, barium, mercury, bacteria, cadmium, chloramines, perchlorate, chromium, radium, silver, copper, viruses, lead, aluminum, and nitrites. To distinguish between safe and unsafe data, Gradient Boosting, Naive Bayes, Random forest, KNN, and LogR are trained and tested. This study's accuracy was higher than that of other state-level studies. In addition, this study performed more accurately when compared to research conducted at the county level.

Table 3 displays a comparison between the findings of this study and earlier research.

Table 3: Result in comparison with benchmark and this study

Machine Learning algorithms	[41] Accuracy Findings	(Proposed work) Accuracy Findings
Random Forest	90.11%	94.01%
Decision Tree	84.65%	91.96%
Naïve Bayes		83.43%
Multilayer Perceptron	82.77%	
Gradient Boosting		93.85%
K-Nearest Neighbors		90.44%
Logistic Regression		90.25%

Table 3 shows the comparison result for the benchmark, the article "Assessment of Groundwater Arsenic Contamination Level in Jharkhand, India Using Machine-Learning" used four Machine learning models including RF, DT, MLP, and NB, and the highest accuracy is Random forest 90.11. Six models—NB, RF, GBOOST, KNN, LogR, and DT—were employed in "The Prediction of Water Quality Using Effective Machine Learning Techniques." Random forest has the best accuracy, at 94.11%. In this study, we find better accuracy as compared to other studies by using different datasets and different machine learning techniques.

5. Conclusion

The results based on this study, factors including such nitrates, barium, mercury, arsenic, bacteria, cadmium, chloramines, perchlorate, chromium, radium, copper, viruses, silver, lead, aluminum, and nitrites, to categorize data that is safe or not safe six machine learning algorithms are used to predict water quality as well as evaluation techniques for precision, recall, accuracy, and F1 score measure. A performance metric is significantly impacted by highly correlated characteristics. The most accurate result of the testing was Random Forest, which had an accuracy of 94.01 percent in this area.

5.1. Advantages of the proposed research

In this research, a machine learning algorithm based on the Target variable "Is_safe" which is a dependent variable on sixteen parameters. This study's accuracy was higher than that of other studies. In addition, this study performed more accurately when compared to research conducted at the county level.

5.2. Limitation of the proposed research

Due to lack of time, six machine learning techniques are used; furthermore machine learning techniques can be used for better results. Results may vary with the change of dataset.

6. Future Work

We recommended future research can use various models with various water datasets, like introducing additional parameters to the model or applying that much more-advanced deep learning algorithm, to enhance the classification of water quality.

In addition, we recommend an Internet of Things-based monitoring system that simply uses sensors to collect the necessary metrics. Tested algorithms would estimate the water quality in real-time using data from the IoT system

With any luck, fewer people will consume tainted water, lessening the severity of horrible illnesses like typhoid and diarrhea. In this way, the application of an advisory analysis was predicated on the values that were expected to lead to the creation of instruments in the future that would support policy- and decision-makers.

7. References

[1] S. A. Kavouras and C. A. Anastasiou, "Water physiology: Essentiality, metabolism, and health implications," *Nutr. Today*, vol. 45, no. 6 SUPPL., pp. 27–32, 2010, doi: 10.1097/NT.0b013e3181fe1713.

[2] J. Hoslett *et al.*, "Surface water filtration using granular media and membranes: A review," *Sci. Total Environ.*, vol. 639, pp. 1268–1282, 2018, doi: 10.1016/j.scitotenv.2018.05.247.

[3] M. A. El-Alfy, A. F. Hasballah, H. T. Abd El-Hamid, and A. M. El-Zeiny, "Toxicity assessment of heavy metals and organochlorine pesticides in freshwater and marine environments, Rosetta area, Egypt using multiple approaches," *Sustain. Environ. Res.*, vol. 29, no. 1, pp. 1–12, 2019, doi: 10.1186/s42834-019-0020-9.

[4] P. Li and J. Wu, "Drinking Water Quality and Public Health," *Expo. Heal.*, vol. 11, no. 2, pp. 73–79, 2019, doi: 10.1007/s12403-019-00299-8.

[5] H. S. Majdi, M. S. Jaafar, and A. M. Abed, "Using KDF material to improve the performance of multi-layers filters in the reduction of chemical and biological pollutants in surface water treatment," *South African J. Chem. Eng.*, vol. 28, no. January, pp. 39–45, 2019, doi: 10.1016/j.sajce.2019.01.003.

[6] F. Fx *et al.*, "$ OJRULWKP FRXSOHG $ UWLILFLDO 1HXUDO 1HWZRUN EDVHG."

[7] T. Benameur, N. Benameur, N. Saidi, S. Tartag, H. Sayad, and A. Agouni, "Predicting factors of public awareness and perception about the quality, safety of drinking water, and pollution incidents," *Environ. Monit. Assess.*, vol. 194, no. 1, 2022, doi: 10.1007/s10661-021-09557-2.

[8] T. Ahmed, M. Zounemat-Kermani, and M. Scholz, "Climate change, water quality and water-related challenges: A review with focus on Pakistan," *Int. J. Environ. Res. Public Health*, vol. 17, no. 22, pp. 1–22, 2020, doi: 10.3390/ijerph17228518.

[9] N. Norfolk and D. Council, "Annual Monitoring Report 2005 - 2006 Annual Monitoring Report 2005 - 2006 All of the LDF Documents can be made available in large print or in other languages .," 2006.

[10] O. A. Adesina, F. Abdulkareem, A. S. Yusuff, M. Lala, and A. Okewale, "Response surface methodology approach to optimization of process parameter for coagulation process of surface water using Moringa oleifera seed," *South African J. Chem. Eng.*, vol. 28, no. August 2018, pp. 46–51, 2019, doi: 10.1016/j.sajce.2019.02.002.

[11] E. Fijani, R. Barzegar, R. Deo, E. Tziritis, and K. Skordas, "Science of the Total Environment Design and implementation of a hybrid model based on two-layer decomposition method coupled with extreme learning machines to support real-time environmental monitoring of water quality parameters," *Sci. Total Environ.*, vol. 648, pp. 839–853, 2019, doi: 10.1016/j.scitotenv.2018.08.221.

[12] F. Westall and A. Brack, "The Importance of Water for Life," *Space Sci. Rev.*, vol. 214, no. 2, pp. 1–23, 2018, doi: 10.1007/s11214-018-0476-7.

[13] L. Bross, J. Bäumer, I. Voggenreiter, I. Wienand, and A. Fekete, "Public health without water? Emergency water supply and minimum supply standards of hospitals in high-income countries using the example of Germany and Austria," *Water Policy*, vol. 23, no. 2, pp. 205–221, 2021, doi: 10.2166/wp.2021.012.

[14] S. Pasika and S. T. Gandla, "Heliyon Smart water quality monitoring system with cost-effective using IoT," *Heliyon*, vol. 6, no. May 2019, p. e04096, 2020, doi: 10.1016/j.heliyon.2020.e04096.

[15] H. Lu and X. Ma, "Chemosphere Hybrid decision tree-based machine learning models for short-term water quality prediction," *Chemosphere*, vol. 249, p. 126169, 2020, doi: 10.1016/j.chemosphere.2020.126169.

[16] E. Guchi, "Review on Slow Sand Filtration in Removing Microbial Contamination and Particles from Drinking Water," *Am. J. Food Nutr.*, vol. 3, no. 2, pp. 47–55, 2015, doi: 10.12691/ajfn-3-2-3.

[17] S. Ray, "A Quick Review of Machine Learning Algorithms," *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 35–39, 2019, doi: 10.1109/COMITCon.2019.8862451.

[18] S. Sah, "Machine Learning: A Review of Learning Types," *ResearchGate*, no. July, 2020, doi: 10.20944/preprints202007.0230.v1.

[19] K. Chen *et al.*, "Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data," *Water Res.*, vol. 171, p. 115454, 2020, doi: 10.1016/j.watres.2019.115454.

[20] U. Ensemble and M. Learning, "Susceptibility Prediction of Groundwater Hardness," pp. 1–17.

[21] B. Mahesh, “Machine Learning Algorithms - A Review,” no. January 2019, 2020, doi: 10.21275/ART20203995.

[22] F. J. Yang, “An implementation of naive bayes classifier,” *Proc. - 2018 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2018*, pp. 301–306, 2018, doi: 10.1109/CSCI46756.2018.00065.

[23] N. M. Abdulkareem and A. M. Abdulazeez, “Machine Learning Classification Based on Radom Forest Algorithm : A Review,” pp. 128–142, 2021, doi: 10.5281/zenodo.4471118.

[24] A. Natekin and A. Knoll, “Gradient boosting machines , a tutorial,” vol. 7, no. December, 2013, doi: 10.3389/fnbot.2013.00021.

[25] K. Taunk, S. De, S. Verma, and A. Swetapadma, “A brief review of nearest neighbor algorithm for learning and classification,” *2019 Int. Conf. Intell. Comput. Control Syst. ICCS 2019*, no. Iciccs, pp. 1255–1260, 2019, doi: 10.1109/ICCS45141.2019.9065747.

[26] I. D. Mienye, Y. Sun, and Z. Wang, “Prediction performance of improved decision tree-based algorithms: A review,” *Procedia Manuf.*, vol. 35, pp. 698–703, 2019, doi: 10.1016/j.promfg.2019.06.011.

[27] S. Bekesiene and I. Meidute-kavaliauskiene, “Accurate Prediction of Concentration Changes in Ozone as an Air Pollutant by Multiple Linear Regression and Artificial Neural Networks,” 2021.

[28] M. Zhu *et al.*, “A review of the application of machine learning in water quality evaluation,” *Eco-Environment Heal.*, vol. 1, no. 2, pp. 107–116, 2022, doi: 10.1016/j.eehl.2022.06.001.

[29] M. Imani, M. M. Hasan, L. F. Bittencourt, K. McClymont, and Z. Kapelan, “A novel machine learning application: Water quality resilience prediction Model,” *Sci. Total Environ.*, vol. 768, p. 144459, 2021, doi: 10.1016/j.scitotenv.2020.144459.

[30] M. Azrour, J. Mabrouki, G. Fattah, A. Guezzaz, and F. Aziz, “Machine learning algorithms for efficient water quality prediction,” *Model. Earth Syst. Environ.*, vol. 8, no. 2, pp. 2793–2801, 2022, doi: 10.1007/s40808-021-01266-6.

[31] U. Shafi, R. Mumtaz, H. Anwar, A. M. Qamar, and H. Khurshid, “Surface Water Pollution Detection using Internet of Things,” pp. 92–96, 2018.

[32] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, and R. Irfan, “E ffi cient Water Quality Prediction Using Supervised,” pp. 1–14, 2019.

[33] D. Dezfooli, S. M. Hosseini-Moghari, K. Ebrahimi, and S. Araghinejad, “Classification of water quality status based on minimum quality parameters: application of machine learning techniques,” *Model. Earth Syst. Environ.*, vol. 4, no. 1, pp. 311–324, 2018, doi: 10.1007/s40808-017-0406-9.

[34] A. El Bilali and A. Taleb, “Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment,” *J. Saudi Soc. Agric. Sci.*, vol. 19, no. 7, pp. 439–451, 2020, doi: 10.1016/j.jssas.2020.08.001.

[35] T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. Maashi, “Water Quality Prediction Using Artificial Intelligence Algorithms,” *Appl. Bionics Biomech.*, vol. 2020, 2020, doi: 10.1155/2020/6659314.

[36] H. Lu and X. Ma, “Hybrid decision tree-based machine learning models for short-term water quality prediction,” *Chemosphere*, vol. 249, p. 126169, 2020, doi:

10.1016/j.chemosphere.2020.126169.

[37] S. Singha, S. Pasupuleti, S. S. Singha, R. Singh, and S. Kumar, “Prediction of groundwater quality using efficient machine learning technique,” *Chemosphere*, vol. 276, p. 130265, 2021, doi: 10.1016/j.chemosphere.2021.130265.

[38] A. Danades, D. Pratama, D. Anggraini, and D. Anggriani, “Comparison of accuracy level K-Nearest Neighbor algorithm and support vector machine algorithm in classification water quality status,” *Proc. 2016 6th Int. Conf. Syst. Eng. Technol. ICSET 2016*, pp. 137–141, 2017, doi: 10.1109/FIT.2016.7857553.

[39] L. Li *et al.*, “Interpretable tree-based ensemble model for predicting beach water quality,” *Water Res.*, vol. 211, p. 118078, 2022.

[40] N. M. Ragi, R. Holla, and G. Manju, “Predicting Water Quality Parameters Using Machine Learning,” *2019 4th IEEE Int. Conf. Recent Trends Electron. Information, Commun. Technol. RTEICT 2019 - Proc.*, pp. 1109–1112, 2019, doi: 10.1109/RTEICT46194.2019.9016825.

[41] S. Kumar and J. Pati, “Assessment of groundwater arsenic contamination level in Jharkhand, India using machine learning,” *J. Comput. Sci.*, vol. 63, no. July, p. 101779, 2022, doi: 10.1016/j.jocs.2022.101779.

[42] A. H. Haghiabi, A. H. Nasrolahi, and A. Parsaie, “Water quality prediction using machine learning methods,” *Water Qual. Res. J.*, vol. 53, no. 1, pp. 3–13, 2018, doi: 10.2166/wqrj.2018.025.

[43] S. Kumar and J. Pati, “Assessment of groundwater arsenic contamination using machine learning in Varanasi, Uttar Pradesh, India,” *J. Water Health*, vol. 20, no. 5, pp. 829–848, 2022, doi: 10.2166/WH.2022.015.

[44] M. Ilić, Z. Srdjević, and B. Srdjević, “Water quality prediction based on Naïve Bayes algorithm,” *Water Sci. Technol.*, vol. 85, no. 4, pp. 1027–1039, 2022, doi: 10.2166/wst.2022.006.

[45] M. Najafzadeh and S. Niazmardi, “A Novel Multiple-Kernel Support Vector Regression Algorithm for Estimation of Water Quality Parameters,” *Nat. Resour. Res.*, vol. 30, no. 5, pp. 3761–3775, 2021, doi: 10.1007/s11053-021-09895-5.

[46] D. Urud *et al.*, “$ QDO \] LQJ WKH 3RWDELOLW \ RI : DWHU XVLQJ 0DFKLQH / HDUQLQJ $ OJRULWKP,” pp. 250–256, 2022, doi: 10.1109/CCiCT56684.2022.00054.

[47] S. Kouadri, C. B. Pande, B. Panneerselvam, K. N. Moharir, and A. Elbeltagi, “Prediction of irrigation groundwater quality parameters using ANN, LSTM, and MLR models,” *Environ. Sci. Pollut. Res.*, vol. 29, no. 14, pp. 21067–21091, 2022, doi: 10.1007/s11356-021-17084-3.

[48] B. Desai and R. K. Sungkur, “Water Quality Prediction Using Machine Learning,” *Lect. Notes Networks Syst.*, vol. 393, no. 05, pp. 401–411, 2022, doi: 10.1007/978-3-030-94191-8_32.

[49] B. H. Reddy and P. R. Karthikeyan, “Classification of Fire and Smoke Images using Decision Tree Algorithm in Comparison with Logistic Regression to Measure Accuracy ,” pp. 0–4, 2022.