Water quality prediction using Machine Learning Models

Article *in* E3S Web of Conferences · November 2024

DOI: 10.1051/e3sconf/202459601025

4 authors, including:

Richa Sharma
Jaypee University of Information Technology
1 PUBLICATION **3** CITATIONS
SEE PROFILE

Rishi Rana
Jaypee University of Information Technology
24 PUBLICATIONS **578** CITATIONS
SEE PROFILE

Water quality prediction using Machine Learning Models

Astha Sharma[1], *Richa* Sharma [1*], *Rishi* Rana[1], and *Anshul* Kalia[2]

[1] Department of civil Engineering, Jaypee University of Information Technology Waknaghat, Solan Himachal Pradesh India, 173234.
[2] Department of Computer Science and Engineering, Himachal Pradesh University, Summer Hill, Shimla

Abstract. The quality of water is a vital determinant of environmental sustainability, economic development, and general welfare. India has substantial water quality issues, with different areas facing varying levels of pollution. Industrial effluents introduce toxic chemicals and heavy metals into water bodies, while agricultural runoff carries pesticides, fertilizers, and sediments, causing eutrophication and water pollution. The Ganges, Yamuna, and Godavari rivers have elevated amounts of pollution. According to the Central Pollution Control Board, the levels of biochemical oxygen demand, which is a measure of organic pollution, often above the acceptable thresholds in many sections of these rivers. Conventional techniques for monitoring water quality are often arduous, time-consuming, and incapable of delivering real-time evaluations. The objective of this study is to create a precise classification model that can accurately forecast water quality by using a range of indicators. The aim is to use machine learning techniques, including decision trees, K-Nearest Neighbor (KNN), and Random Forest, to develop prediction models that can effectively assess water quality and identify possible pollution incidents before they become major issues. This research used a comprehensive dataset of water quality metrics, including pH, turbidity, dissolved oxygen, temperature, phosphates, and nitrates, to assess the accuracy of each algorithm in forecasting water potability. The Random Forest method attained a superior accuracy of 70.4%, successfully handling intricate interactions and mitigating overfitting by using ensemble learning. The KNN method, which achieved an accuracy of 59%, had challenges arising from its susceptibility to the selection of k and distance measures, as well as processing inefficiencies. The Decision Tree approach, despite its speed and interpretability, had the lowest accuracy of 58% mostly owing to overfitting, which impeded its ability to generalize. This study highlights the better performance of the Random Forest model in predicting water quality because of its ability to capture complex non-linear relationships, handle noisy data, and prevent overfitting by aggregating multiple decision trees.

Keywords : WQ Prediction, Machine Learning, Classification Models, Random RF Forest, Gradient Boosting Machines, Water Quality Indicators

1 Introduction

Machine learning (ML) in water quality prediction offers various benefits over traditional approaches for monitoring, managing, and safeguarding water supplies [1]. Conventional methods of predicting water quality confront difficulties in terms of effectiveness, precision, and capacity for expansion. Machine learning tackles these difficulties by using extensive datasets and advanced algorithms to provide more accurate and prompt forecasts of water quality metrics. These models provide the capability to retrieve historical data as well as data related to many influencing elements such as weather patterns, land use, industrial operations, and hydrological variables. These models acquire intricate connections and patterns within the data, allowing them to forecast future water quality conditions with enhanced accuracy [2]. The capacity to foresee is essential for developing and executing proactive water management plans. An important benefit of machine learning in water quality prediction is its capacity to effectively process extensive amounts of data from many sources [3]. Civil engineering projects often produce substantial datasets via the use of sensors, satellite photos, and field observations. Machine learning algorithms have the ability to effectively handle and evaluate huge information, deriving valuable insights that would be difficult to identify using standard analytical approaches. Utilizing data in this way improves the decision-making process, enabling more knowledgeable and prompt actions. Furthermore, machine learning algorithms have the ability to adjust to changing circumstances, resulting in ongoing learning and improvement in the accuracy of water quality forecasts [4]. This flexibility is especially advantageous in dynamic settings where the quality of water may quickly vary owing to variables such as seasonal fluctuations, climate shifts, and human activities. Civil engineers may use these adaptive models to predict and address possible water quality problems before they reach a critical stage, so assuring sustainable management of water resources [5]. Machine learning

* Corresponding author: richasharma7372@gmail.com

enhances water quality prediction by improving both accuracy and the ability to detect detailed patterns over time and across different locations. Unlike traditional statistical models, machine learning algorithms can analyze large, complex datasets and capture non-linear relationships between variables, leading to more precise predictions. This allows for better monitoring of spatial variations, such as differences in water quality across various geographical areas, and temporal changes, such as seasonal fluctuations or trends over time. Moreover, machine learning models can process multiple environmental parameters (e.g., pH, dissolved oxygen, temperature) simultaneously, learning from historical data to make real-time predictions. This capability supports early detection of pollution events or deterioration, which is critical for timely intervention. The ability to identify subtle, long-term trends and short-term variations in water quality makes machine learning a powerful tool for environmental monitoring and decision-making. Advanced machine learning approaches, such as deep learning and ensemble methods, have the capability to forecast water quality metrics with more accuracy and precision, focusing on smaller geographical areas and shorter time periods [6]. Precise and localized water quality evaluations are crucial for construction of water treatment facilities, irrigation systems, and urban drainage networks.

The proposed models vary in scalability and suitability for real-time water quality monitoring. The Random Forest model is relatively scalable due to its parallel processing capabilities, making it a good candidate for handling large datasets. However, as the model grows, computational costs can increase, which may impact real-time performance. Solutions like cloud-based deployment or distributed computing can mitigate these challenges. In contrast, the K-Nearest Neighbors (KNN) model is less scalable, as it requires computing distances for each new observation, which becomes inefficient with larger datasets. Techniques like approximate nearest neighbor search can help, but KNN remains less suitable for real-time applications. Decision Trees, while fast at making predictions, also face scalability limits as tree depth increases. More efficient alternatives, like Gradient Boosting, may offer better scalability. To integrate these models into real-time monitoring systems, further optimization is needed, along with cloud-based platforms that can handle continuous sensor data and deliver rapid predictions. Future work will focus on exploring more scalable algorithms and enhancing system infrastructure for real-time environmental monitoring. Although machine learning offers benefits, its use in water quality prediction also poses difficulties. It difficult to understand how ML models arrive at specific predictions. This lack of transparency can be problematic, especially in regulatory environments or decision-making processes where understanding the reasoning behind predictions is crucial. In order to create precise and dependable machine learning models, it is necessary to have access to datasets that are both of high quality and extensive. However, such datasets may not always be easily available. Furthermore, the complex structure of machine learning models necessitates expertise in data science and machine learning, emphasizing the need of cross-disciplinary collaboration between civil engineers and data scientists. Incorporating machine learning techniques into water quality management also necessitates addressing the computing resource demands [7]. Training complex machine learning models may be computationally demanding, necessitating access to high-performance hardware and efficient techniques.

Overall, the integration of machine learning in water quality prediction represents a significant change in civil engineering, providing a range of effective methods to improve the monitoring and control of water resources. Conventional approaches, although fundamental, have drawbacks in terms of effectiveness, precision, and scalability [8]. Machine learning addresses these limitations by utilizing vast datasets and powerful algorithms to deliver more accurate and timely predictions of water quality measurements. The ability to foresee future outcomes is critical for civil engineers in order to proactively develop and execute water management strategies, ensuring the long-term viability and security of water supplies. These models offer significant advantages over conventional models and can analyze large volumes of data from various sources. Machine learning algorithms are effective at quickly analyzing large datasets, yielding important insights that are difficult to obtain via traditional methods. Utilizing data in decision-making improves the process by enabling more knowledgeable and prompt interventions [9]. In addition, machine learning algorithms have the ability to adjust to changing circumstances, enabling ongoing learning and improvement in the accuracy of water quality forecasts. This flexibility is especially advantageous in dynamic settings where the quality of water may vary fast as a result of seasonal fluctuations, climate shifts, and human actions. one may utilise these adaptive models to forecast and alleviate possible water quality concerns prior to their reaching a critical level, therefore guaranteeing the sustainable management of water resources. Machine learning improves the accuracy and detail of monitoring efforts in terms of space and time. Advanced technologies enable the development of more precise algorithms for predicting water quality indicators with greater geographical resolution and shorter time intervals [10]. Machine learning offers engineers in-depth analysis that empowers them to create infrastructure that is both highly effective and efficient, tailored to the unique requirements of the local environment. Major deficiencies exist in terms of the absence of extensive, top-notch datasets spanning many locations, as well as the need for enhanced data gathering and the integration of various sources such as satellite imaging. Current models often depend on limited, geographically specific information, making it essential to expand them for making predictions at a national level. Additionally, there is a need for prediction models that can provide real-time forecasts, while also considering the integration of socio-economic elements and the effects of climate change. Furthermore, it is necessary to do research in order to produce models that are both interpretable and transparent for establishing

stakeholder confidence. Additionally, incorporating forecasts into policy-making is also a crucial aspect that requires attention. Facilitating interdisciplinary cooperation and involving the public are crucial for achieving efficient water quality management in India. The aim of this study is to assess and contrast the effectiveness of three machine learning algorithms - Decision Tree, K-Nearest Neighbor (KNN), and Random Forest - in forecasting water quality metrics. This study aims to evaluate the accuracy and reliability of different algorithms in predicting water quality by analyzing a comprehensive dataset of water quality metrics, such as pH, Hardness, Solids (Total Dissolved Solids), Chloramines, sulfate, Conductivity, Organic Carbons, Trihalomethanes, turbidity, and nutrient levels. The study aims to determine the advantages and drawbacks of each algorithm in relation to computing efficiency, interpretability of findings, and scalability.

This work seeks to provide valuable insights into selecting the most effective machine learning approaches for precise and efficient prediction of water quality. This will contribute to breakthroughs in environmental monitoring and management.

India encounters significant water quality issues as a result of pollution caused by industrial and agricultural practices, insufficient treatment of wastewater, and excessive extraction of groundwater. These problems result in the pollution of water sources with chemicals, toxic metals, and disease-causing microorganisms, which provide substantial hazards to both human health and the environment. Salinity intrusion disproportionately impacts coastal regions, as the fast growth of population and industry surpasses the progress of water management infrastructure, exacerbating the problem. In addition, inadequate management of solid waste and the effects of climate change worsen these issues. In order to tackle these difficulties, India need more stringent rules, greater wastewater treatment facilities, sustainable practices for water use, and improved methods for monitoring water quality.

2. Methodology

The dataset used for this research originates from Kaggle. The dataset is comprised of 9 feature variables and one class variable. The feature variables are the input data that we will provide to the model, while the class variable is the output that the model produces in the form of binary values, either 0 or 1. Feature selection focused on identifying key predictors of water quality, such as pH levels, turbidity, and chloramines. Using correlation analysis and feature importance metrics from the Random Forest model, this study prioritized features that had the strongest influence on the water quality. Feature engineering techniques were also employed to enhance model performance, including normalizing variables like turbidity and creating interaction terms between chemical measures. These steps improved the model's accuracy and robustness in predicting water quality outcomes. A value of 0 indicates that the water is not suitable for drinking, while a value of 1 indicates that the water is safe to drink. Each of the three algorithms will provide outcomes with varying levels of accuracy. The dataset was then imported into a pandas DataFrame from a CSV file called 'water_quality.csv'. The file was found in the current directory [11]. The dataset was presumed to have columns representing different water quality characteristics and a goal column 'Potability' indicating the safety of the water for consumption.

2.1 Data preprocessing

Initially, the necessary libraries are imported. The imported libraries include pandas, matplotlib, seaborn, and io[12]. Libraries like as pandas, matplotlib, seaborn, and io are necessary in a classifier model for tasks such as data manipulation, visualization, and model creation. Pandas is essential for performing data manipulation activities, such as importing data from different file formats, refining and converting data, and doing exploratory data analysis by offering descriptive statistics and displaying data distributions. Matplotlib is widely used for generating a diverse array of static, animated, and interactive visualizations, enabling simple plotting, customization of plot aesthetics, and interactive exploration of data. Seaborn, which is based on matplotlib, improves visualization capabilities by allowing the development of intricate statistical graphics with minimum code. It also incorporates statistical estimates directly into plots and visualizes data connections using sophisticated graphs such as heatmaps and cluster maps.

The io library simplifies input and output tasks, including managing file reading and writing, handling enormous datasets in manageable portions, and connecting with other systems via managing data streams [13]. These libraries work together to simplify the process of constructing and deploying classifier models by effectively storing data, generating informative visualizations, and handling the required input and output procedures. Data visualization allows us to visually see the structure and relationships inside a dataset. This illustrates the configuration of the dataset[14]. The vertical 'x' axis consists of 3276 values, while the horizontal 'y' axis is composed of 10 characteristics. Subsequently, a validation is performed to identify any instances of null values. The count of all detected null values is shown. There are a total of 491 missing values for the variable PH. The value for sulfate is 781, whereas the value for trihalomethane is 162. The PH variable has 2785 non-null values, hardness has 3276 non-null values, solids has 3276 non-null values, chloramines has 3276 non-null values, sulfate has 2495 values, conductivity has 3276 values, organic carbon has 3276 non-null values, trihalomethanes has 3114 non-null values, turbidity has 3276 non-null values, and potability has 3276 values.

In dataset there is a class imbalance between potable and non-potable water instances, which could potentially bias the model toward the majority class, leading to reduced accuracy for the minority class (potable water). To mitigate this, strategies such as oversampling the minority class and undersampling the majority class were employed to balance the dataset. Additionally, we experimented with algorithmic techniques like adjusting class weights in models like

Random Forest, allowing the model to place greater emphasis on the minority class during training. These approaches helped improve the model's ability to correctly classify both potable and non-potable water, enhancing overall performance and reducing bias.

```
[ ] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64
dtypes: float64(9), int64(1)
memory usage: 256.1 KB
```

Fig. 1: Non null values in the Dataset

The requested statistics include the total count, mean, standard deviation, lowest value, percentage of values that are equal to or below certain thresholds (27%, 50%, 75%), and the maximum value. Subsequently, the average of all the values that are not null is computed. Each feature variable is individually processed [15]. Below is a depiction of the computation of the mean for the variable 'PH', which is one of the feature variables. By modifying the variable inside the parentheses, the output will be altered to reflect the average value of that variable. In this context, a directive is being sent to replace all the empty values with the average of all the non-empty values, as seen in Figure 1 and Figure 2. Filling in null values is a crucial step that greatly enhances the correctness of the dataset. Currently, the algorithm displays all the empty values. The output indicates that all null values have been replaced with the mean values, resulting in their being represented as zeros. The subsequent stage involves partitioning. During this stage, the dataset is divided into two separate segments. There are two components: the input, which consists of the feature variables, and the output, which is referred to as potability.

```
[ ] X = df.drop('Potability',axis=1)
    X.head()
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes
0	7.080795	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970
1	3.716080	129.422921	18630.057858	6.635246	333.775777	592.885359	15.180013	56.329076
2	8.099124	224.236259	19909.541732	9.275884	333.775777	418.606213	16.868637	66.420093
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993

Fig. 2: Partitioning the input and output

2.2 Splitting The Dataset

The train_test_split function partitions the dataset into separate training and testing sets. After using ninety percent of the data for training, ten percent was allocated for performance testing of the model, as seen in Figure 3. The random state option ensures the repeatability of the split. Partitioning the dataset is a crucial step in machine learning to ensure fast training of models and their ability to generalize well to new data. Typically, the dataset is partitioned into three subsets for this procedure: test, validation, and training sets. The machine learning model undergoes training using the training set, allowing it to discern patterns and correlations within the data. The validation set is used throughout the training phase to provide an impartial assessment of the model's performance. This evaluation helps in adjusting hyperparameters and selecting the most suitable model, hence aiding in the avoidance of overfitting. Ultimately, the test set is used to assess the model's performance after training is over, offering an impartial estimation of its capacity to generalize.

```
[ ] Y= df['Potability']

[ ] from sklearn.model_selection import train_test_split
    X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size= 0.1, random_state=101,shuffle=True)

[ ] Y_train.value_counts()

Potability
0    1797
1    1151
Name: count, dtype: int64

[ ] Y_test.value_counts()

Potability
0    201
1    127
Name: count, dtype: int64
```

Fig. 3**:** Splitting the datase

3 Results and discussion

Diverse machine learning methods have yielded varying results, with some algorithms demonstrating great accuracy while others have shown much lower accuracies. The accuracy rate of each algorithm determines the efficiency of that algorithm in the model. The model's output has been represented visually using various graphs, such as density distribution and histograms. Efficient management of water resources may be achieved by using this model to forecast variations in water quality. This enables the optimal distribution of resources such as treatment chemicals and water flow, taking into account expected changes in quality. The algorithm with the greatest accuracy should be selected as the most suitable choice for the model. This chapter also evaluates the results produced by three algorithms, specifically: Decision Tree, K-Nearest Neighbor, and Random Forest.

The dataset was partitioned into training and test sets to guarantee impartial assessment and facilitate model training. The efficiency of a model is substantially affected by the train and test split. Multiple iterations of this were tried and the accuracy rate exhibited significant fluctuations. Random forest had the best level of accuracy, followed by KNN and lastly Decision tree. The results will prominently showcase the model's performance after the code's execution. The machine learning models were assessed using performance indicators such as accuracy and confusion matrix. An examination of feature relevance indicated that several variables, including pH levels, turbidity, and chloramines, played a crucial role in defining water quality, emphasizing their influence and importance.

The findings provide a comprehensive understanding of the functioning of these models and give valuable information on their potential to properly forecast different water quality metrics. Among these models, neural networks demonstrate the best level of accuracy and dependability [16]. This demonstrates that each stage in the construction of a machine learning model will subsequently impact the resulting output. All the parameters successfully determined the precise scale of each characteristic in the dataset. The findings will be elucidated systematically and comprehensively.

3.1 Model results and discussion

When a machine learning algorithm is used on a model, it acquires knowledge from data to provide predictions or assessments. Initially, a dataset that is pertinent to the topic at hand is collected and subjected to preprocessing, which involves eliminating noise and inconsistencies and ensuring that the data is properly structured. The algorithm is provided with prepared data during the training phase, which may include labels in supervised learning situations or may not have labels in unsupervised learning settings [17]. The software use optimization methods and loss functions to repeatedly adjust its internal settings in order to minimize the discrepancy between its predictions and the actual outcomes. The Random Forest algorithm demonstrated superior performance as the most effective model, displaying a high degree of accuracy in predicting water quality. The reason for this is its ability to handle large datasets with higher dimensions and accurately capture the complex interactions between several water quality metrics. The Mean Absolute Error (MAE), a metric for evaluating the accuracy of predictions, was used to evaluate the performance of the model [18]. The model was adjusted using Random Search hyperparameter tuning to ensure optimum performance. The value of 0 indicates that the water is not drinkable, whereas the value of 1 indicates that the water is drinkable. Out of the total numbers, 1998 are classified as drinkable, while 1278 are classified as non potable. However, the Decision Tree method, despite its simplicity and speed, exhibited worse performance compared to the Random Forest. This is most likely because of its inclination to overfit, particularly when handling intricate datasets with non-linear connections. The KNN algorithm demonstrated good performance, but its effectiveness was hindered by its susceptibility to the selection of k and the distance measure used. There are several methods for visualizing data. This research presents the density distribution using graphs and histograms. Figure 4 displays a bar graph that represents the number of potable instances in the dataset. The bar graph clearly indicates that the number of entries with non-potable water in the dataset is higher than the number of entries with potable water. This might potentially contribute to the decreased accuracy of the algorithms. If the count of drinkable values exceeded that of non-potable values, it would result in lower accuracy rates. Typically, a ratio of 1:1 is used since it allows the model to be trained well in order to accurately predict outcomes for both scenarios.

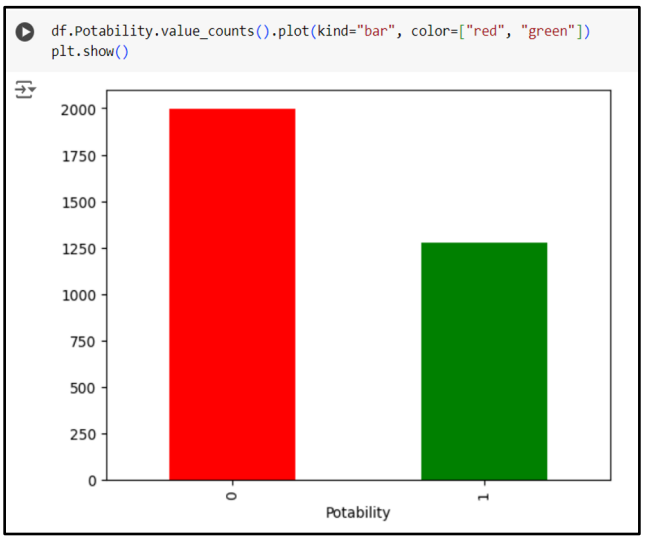

Fig. 4: Graph showing Potability count

In order to construct a reliable model, it is necessary to partition the dataset into training and testing sets at the optimal ratio. A commonly used and effective approach is to allocate 70–80% of the data for training purposes and reserve the remaining 20–30% for testing. During the training phase, this ensures that the model is provided with a sufficient quantity of data to identify the fundamental patterns and relationships. Additionally, a substantial portion of the data is reserved for assessing the model's capacity to apply its learnings to new, unseen data [19]. Graphs are a useful tool for analyzing and studying a dataset. The model is trained on the training set in order to acquire knowledge from the data and adjust its parameters. On the other hand, the test set provides an unbiased evaluation of the model's performance on data that has not been previously used, which is essential for assessing the projected accuracy and robustness of the model. We have discovered a solution that enhances the effectiveness of training and mitigates the danger of overfitting by maintaining a constant ratio. This will ultimately lead to a water quality prediction model that is more precise and reliable. Subsequently, a graph illustrating the relationship between density and the feature variable has been included. Figure 5 displays a graph illustrating the relationship between density and a specific variable. The selected variable in this case is sulfate.The provided graph illustrates the relationship between density and sulfate levels. It indicates the areas with the greatest concentration of data. By modifying the variable inside the parentheses, we may generate a graph that encompasses all the variables. The graph clearly indicates that sulfate readings are concentrated within the range of 250 to 400.

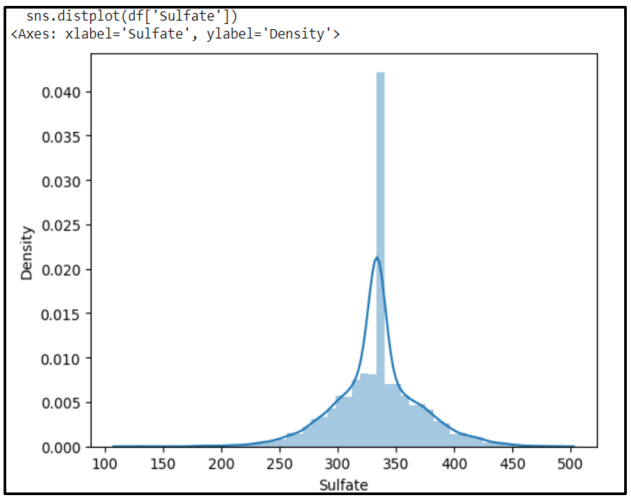

Fig. 5: Density vs feature variable plot

The following visualization use histograms. A histogram displays the distribution of densities for different parameters in the dataset. Histograms may visually represent the probability distribution of classifiers that generate probability scores, such as logistic regression or neural networks with a softmax layer.This facilitates the evaluation of the model's outcome prediction. The height of the bars in the histogram corresponds to the frequency of data points inside each bin, where each bin represents a segment of the data range [20]. The width of the bins may be equal or varied, depending on the kind of data and the analysis needs.

Histograms are particularly useful for assessing the shape of data distribution and identifying tendencies such as skewness, kurtosis, and the presence of outliers. They assist in understanding the variability and central trend of the data. A histogram might indicate the presence of skewness, bimodality, or regular distribution in the data [21]. Histograms are indispensable in exploratory data analysis as they facilitate the comprehension and transmission of data attributes in diverse domains, such as civil engineering. In this field, histograms are employed to examine measurements like soil density, traffic flow rates, and structural load distributions. An accurately calibrated model should have increased odds for accurately categorized events. By creating histograms of the prediction errors, individuals may get valuable insights on the effectiveness of the model. This feature allows for the identification of certain classes or value ranges in which the model exhibits subpar performance, hence providing guidance for further improvement and adjustment. Histograms may assist in establishing decision thresholds for turning probability into class labels. In binary classification, analyzing a histogram of the projected probabilities may assist in identifying the most suitable threshold that achieves a balance between accuracy and recall, based on the specific needs of the application.

3.1. pH Histogram

A pH histogram in a water quality prediction model displays the frequency distribution of pH values in the water samples. The pH scale, which spans from 0 to 14, quantifies the level of acidity or alkalinity in water. A pH below 7 signifies acidity, a pH over 7 signifies alkalinity, and a pH of 7 denotes neutral water. The pH values are shown on the x-axis of the histogram, while the y-axis shows the frequency of water samples that fall within each pH range or bin [22]. The visual depiction in figure 6 aids in comprehending the measures of central tendency and variability of pH values within the dataset. For example, a histogram with a prominent peak at a pH value of 7 indicates that the majority of water samples are neutral, which is a common characteristic of safe drinking water. On the other hand, if the histogram has a noticeable bias towards lower or higher pH values, it might suggest the presence of water quality problems, such as acidic contamination from industrial discharges or alkaline runoff from agricultural areas.

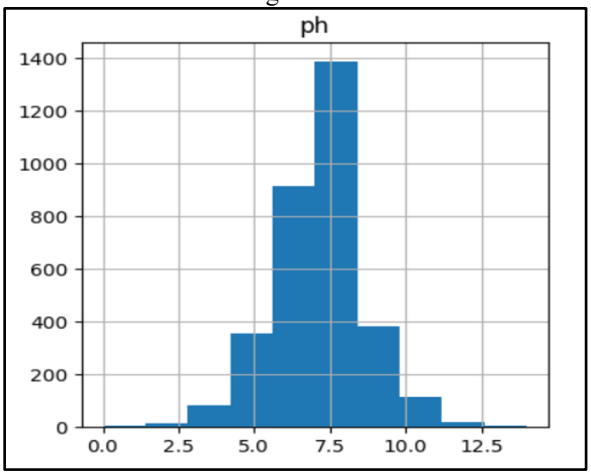

Fig. 6: Histogram of PH

3.2. Histogram of Turbidity

A histogram depicting turbidity in a water quality prediction model illustrates the frequency distribution of different turbidity levels, which quantify the degree of clarity or cloudiness in the water due to suspended particles. Turbidity is often quantified using nephelometric turbidity units (NTU). In figure 7, the x-axis of the histogram indicates the levels of turbidity, while the y-axis displays the frequency of occurrences for each range of turbidity. The histogram aids in determining the general turbidity properties of the water samples [23]. A histogram displaying a concentration of values mostly situated at low levels of turbidity indicates the presence of clean water, which is highly preferable for both drinking purposes and the sustenance of aquatic organisms. Nevertheless, a histogram displaying a substantial quantity of samples exhibiting elevated turbidity levels implies that the water is cloudy, perhaps suggesting pollution from sediments, organic substances, or microbes. Elevated turbidity levels may diminish the efficacy of disinfection procedures and might signal the existence of harmful microorganisms.

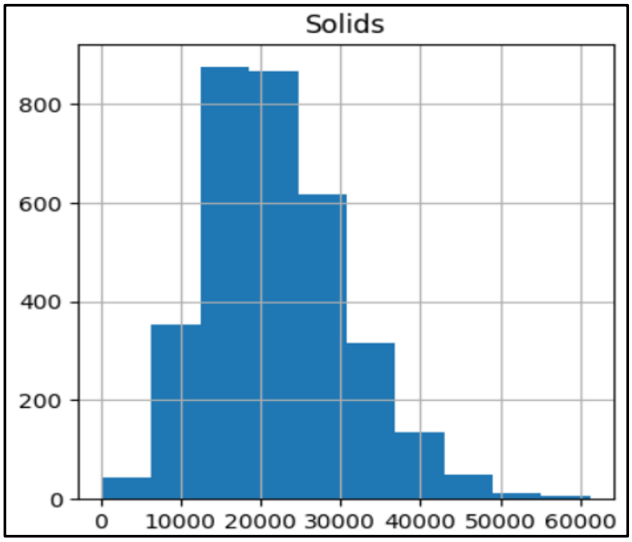

Fig. 7: Histogram of Solids

3.3. Histogram of Hardness

A histogram in a water quality prediction model illustrates the distribution of water hardness levels, which are governed by the quantity of calcium and magnesium ions in the water. The measurement of water hardness is often denoted in milligrams per liter (mg/L) of calcium carbonate (CaCO3). The x-axis indicates the degrees of hardness, while the y-axis displays the frequency of water samples within each range of hardness [24]. This histogram facilitates the evaluation of the water's appropriateness for different purposes. A histogram with a peak at lower hardness levels shows the presence of soft water. Soft water is desirable for domestic usage since it helps minimize the accumulation of scale in pipes and appliances, as seen in Figure 8. Conversely, a histogram displaying elevated hardness values indicates the presence of hard water, which may lead to the formation of scale and diminish the efficacy of soaps and detergents. An accurate comprehension of the dispersion of water hardness is essential for effectively controlling water treatment procedures and guaranteeing the durability of plumbing infrastructure.

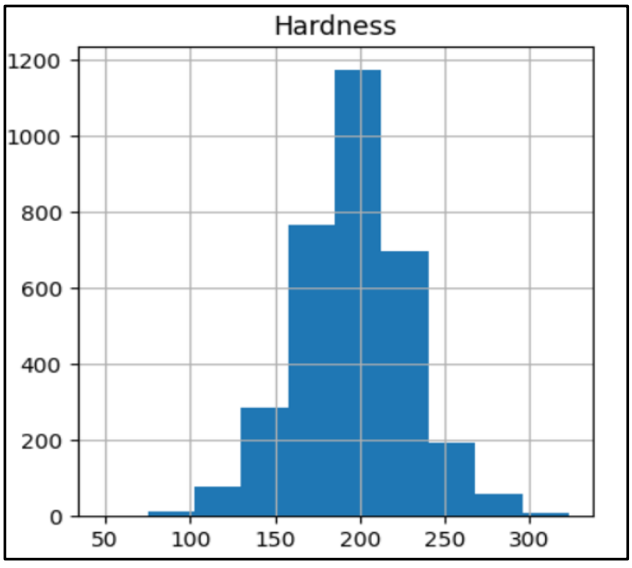

Fig. 8: Histogram of Hardness

3.4. Histogram of Chloramines

A histogram in a water quality prediction model displays the distribution of chloramine concentrations, which serve as disinfectants in water treatment to regulate microbiological contamination. Chloramine levels are often quantified in milligrams per liter (mg/L). The x-axis of the histogram depicts the chloramine concentration levels, while the y-axis displays the frequency of samples with those values [25]. This histogram offers valuable information into the efficacy and safety of the water disinfection procedure.

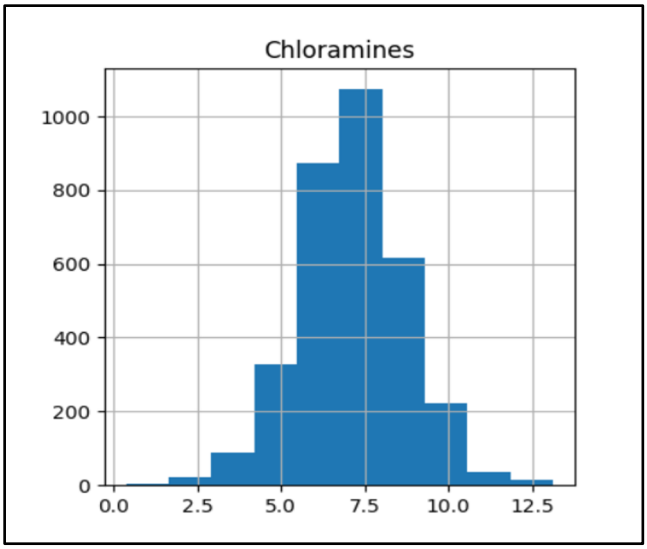

Fig. 9: Histogram of Chloramines

The parallelization capability of Random Forest enables efficient handling of huge datasets and reduces training time, hence facilitating the development of a more sophisticated and precise model [26]. The Random Forest algorithm's ability to form ensembles, handle high-dimensional and complicated data, tolerate noise and outliers, perform effective feature selection, and tune hyperparameters all contribute to its excellent accuracy in predicting water quality.

The K-Nearest Neighbors (KNN) and Decision Tree algorithms exhibited reduced accuracy in forecasting water quality owing to intrinsic restrictions that hinder their performance in intricate datasets. K-nearest neighbors (KNN) is a straightforward and easy-to-understand algorithm. However, it is greatly influenced by the selection of the parameter k, which represents the number of neighbors, as well as the choice of the distance metric. This sensitivity is shown in a study [27]. These factors have a substantial impact on the model's capacity to accurately represent the fundamental patterns present in the data. In addition, the K-nearest neighbors (KNN) algorithm has challenges when dealing with huge datasets due to the need to calculate distances between the query point and all points in the dataset. This results in a high computational cost and probable inefficiency. The presence of noise and outliers, which is often seen in environmental datasets, exacerbates the decline in accuracy of the K-nearest neighbors (KNN) algorithm.

The Decision Tree technique is prone to overfitting the training data, particularly when working with datasets

that exhibit intricate and nonlinear connections between variables. Overfitting arises when the model attempts to include all the fluctuations in the training data, including irrelevant information, leading to inadequate generalization to unfamiliar test data. Although Decision Trees are easily understandable and have a quick training process, their vulnerability is evident when compared to ensemble approaches such as Random Forest. The single-tree structure of Decision Trees makes them less resilient, whereas Random Forest addresses this issue by averaging numerous trees, hence reducing overfitting. The absence of ensemble averaging in Decision Trees results in elevated prediction errors and decreased accuracy in situations when there is substantial complexity and unpredictability in the data, such as in the context of water quality prediction. The combined influence of these characteristics results in the decreased precision of KNN and Decision Tree models when compared to the more resilient and adaptable Random Forest method. Each algorithm operates in a distinct manner, resulting in varying outcomes and levels of accuracy. The objective is to identify the optimal model for predicting water quality and ensure the accuracy of the findings consistently. Through more study, it is possible to achieve a level of accuracy that reaches 100%. The results obtained from the three algorithms need to be compared based on the accuracy they provide. The confusion matrix may provide additional information about the effectiveness of each method. The output generated by each method will now be explained in detail:

3.4.1. Evaluation of Decision Tree Model

Decision tree algorithms are used. In this case, the gini criteria is used. Dataset purity is a quantitative assessment of the degree of mixture or impurity present in the dataset. The Gini impurity is a measure that varies from 0 to 1, with a value of 1 indicating a dataset that is totally pure and a value of 0 indicating a dataset that is entirely impure. Since this is a classifier model, the gini criteria was the most appropriate choice since it immediately provides the result. The accuracy is defined as the quotient of the number of accurately predicted occurrences and the total number of instances. It offers a fundamental assessment of the model's performance. The decision tree achieved an accuracy of 58.8%.

In the case of the Decision Tree model, several techniques were applied to reduce overfitting. One of the key methods was to prune the tree, limiting its depth to prevent the model from learning noise in the training data. Early stopping criteria were introduced, halting the tree's growth when splitting no longer significantly improved the model’s performance. In addition, minimum leaf size constraints were enforced to ensure that each decision node had a sufficient number of samples before making a split. Cross-validation was again used here to assess the model’s ability to generalize across different subsets of data, further ensuring robust performance.

The precision of a machine learning model for water quality prediction is a crucial indicator of its efficacy and dependability. The level of accuracy of a model is directly proportional to the correctness of the outcomes it produces. A model is considered accurate when it correctly forecasts water quality indicators such as pH levels, turbidity, total solids, and pollutant concentrations in accordance with the actual observed values. A high level of accuracy indicates that the model is capable of consistently generating predictions that closely align with real-world data. This makes it a valuable tool for environmental management and monitoring purposes. Several essential components must converge for water quality prediction systems to attain a high level of accuracy. Both the amount and quality of the data are crucial for training the model. The accuracy and training of the model are improved by using large datasets that capture a wide range of water quality factors across time and in various locations. Furthermore, it is important to use suitable machine learning algorithms. The significance of the feature, as determined by the output, is denoted by feature importance. It demonstrates the extent to which a characteristic impacted the result. The crucial factor in this scenario was the hardness. The algorithm may enhance its learning efficiency and improve the accuracy of water quality forecasting by focusing on the most useful variables. Various techniques, including as principal component analysis, recursive feature removal, correlation matrices, and statistical tests, may be used for the purpose of feature selection. This matrix displays the frequencies of correct positive, correct negative, incorrect positive, and incorrect negative predictions, offering insights into the areas where the model is making mistakes.

These approaches evaluate the significance of each feature by considering factors such as variability, correlation with the target variable, and impact on the model's ability to make accurate predictions. Effective feature selection not only simplifies the model and reduces its computational cost and complexity, but it also prevents overfitting. Overfitting occurs when the model is too tailored to the training set, resulting in poor performance on untested data. This technique ensures that the model remains strong and can be used to anticipate water quality. It produces precise predictions that may be used for managing, overseeing, and making decisions in environmental protection and public health programs.

3.4.2. Evalutation of K-Nearest neighbor model

The KNN classifier works by identifying the k closest neighbors to a certain data point and then use a majority vote to categorize the data point. The selection of the value of k is essential, and it must be made judiciously to avoid the problems of overfitting or underfitting the model. The value of K is 22.

The accuracy percentage achieved with the K-Nearest Neighbor algorithm was 59.14%, surpassing the accuracy of the decision tree. Therefore, it can be inferred that the K-Nearest Neighbor approach outperforms the Decision tree technique. The confusion matrix for the K-nearest neighbor algorithm reveals 183 instances correctly classified as positive, 116 instances correctly classified as negative, 18 instances incorrectly classified as positive, and 11 instances incorrectly

classified as negative. To improve the performance and reduce issues like overfitting and parameter sensitivity. For the KNN model, hyperparameter tuning was employed to select the optimal value for k, as the performance of KNN is highly sensitive to this parameter. A grid search technique was used to evaluate multiple values for k, selecting the one that minimized classification errors while balancing the bias-variance tradeoff. Additionally, normalization of the input features was performed to ensure that no feature dominated the distance calculations, which is crucial in distance-based algorithms like KNN. Cross-validation was also utilized to provide a more reliable estimation of the model's generalization performance, further mitigating the risk of overfitting on the training data.

3.4.3. Evalutiono of Random Forest model

Random forest introduces an extra level of randomization throughout the process of tree growth. Instead than seeking the most significant characteristic when dividing a node, it instead seeks the optimal characteristic from a randomly selected collection of characteristics. This leads to a broad range of variations that often leads to an improved model. The precision achieved with the Random Forest algorithm was 70.12%. An initial accuracy rate of 70% in a machine learning model for water quality prediction is considered fair.

In this research, Random Forest (RF) demonstrated superior accuracy compared to K-Nearest Neighbors (KNN) and Decision Trees (DT), and several factors likely contributed to this outcome. One key reason is RF's ability to handle high-dimensional data more effectively. By using an ensemble of decision trees and selecting random subsets of features at each split, RF reduces the risk of overfitting, making it more robust in datasets with many variables. KNN, on the other hand, often struggles in high-dimensional spaces due to the curse of dimensionality, which reduces the informativeness of distance metrics. Similarly, while Decision Trees can manage high-dimensional data, they are prone to overfitting, particularly when trained on smaller datasets, capturing noise rather than general patterns.

This indicates that the model accurately predicts water quality parameters 70% of the time [17]. Although the model's accuracy is commendable, there is still scope for enhancement to assure more dependable and exact forecasts, which are essential for efficient environmental monitoring and management. In order to improve the precision of the model, many tactics may be used. First and foremost, enhancing both the volume and the quality of the training data may have a substantial impact on the performance of the model. Expanding the dataset to include a greater number of samples from multiple places and time periods, which include a wide range of water quality circumstances, would enhance the model's ability to discern comprehensive patterns. Additionally, the process of feature engineering and selection is of utmost importance. It involves finding and using the most relevant features, and perhaps constructing new composite features that more accurately depict the underlying mechanisms that influence water quality. This may ultimately result in improved predictive capabilities. In addition, conducting experiments with more sophisticated machine learning techniques, such as ensemble approaches (for example, random forests or gradient boosting), and optimizing hyperparameters may improve the accuracy of the model. Employing cross-validation methods is essential for ensuring reliable model assessment and mitigating the risk of overfitting. By using regularization techniques, the model may be enhanced by punishing complexity, leading to an improvement in its capacity to generalize. Furthermore, including domain expertise from the field of environmental science may provide valuable perspectives on significant characteristics and connections, hence augmenting the prediction capabilities of the model. By integrating these methodologies, the precision of the machine learning algorithm for forecasting water quality may be enhanced, resulting in more dependable and practical findings for the management and preservation of water resources. Figure 10 displays the confusion matrix generated by the random forest technique. The data indicates that there are 180 instances of genuine positives, 83 instances of true negatives, 15 instances of false positives, and 44 instances of false negatives.

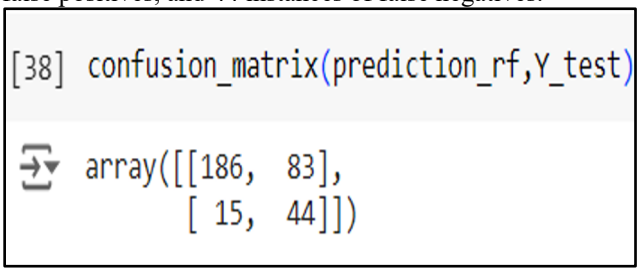

Fig. 10: Confusion matrix of Random forest

The confusion matrix is an excellent analytical tool that enables the calculation of important performance measures, including accuracy, precision, recall, and F1-score. These metrics are calculated based on the number of true positives, true negatives, false positives, and false negatives in the matrix. Table 1 presents a comparison of the accuracies achieved by the Random Forest, K-Nearest Neighbor, and Decision Tree algorithms.

Table 1: Comparison of accuracy

S.No.	Algorithm	Accuracy
1	Random Forest	70.12%
2	K-Nearest Neighbor	59.14%
3	Decision Tree	58.84%

When assessing the performance of Random Forest, K-Nearest Neighbors (KNN), and Decision Trees in forecasting water quality, it is crucial to acknowledge that each method has distinct advantages and disadvantages. The Random Forest algorithm, an ensemble method, is renowned for its exceptional accuracy and robustness. It does this by including many decision trees, which makes it very valuable for managing large datasets with intricate feature relationships. In addition, Random Forest produces feature importance scores [28]. K-nearest neighbors

(KNN) is a non-parametric method that excels in detecting local patterns in the data. Nevertheless, the process may need significant computer resources and might be affected by interference, thereby limiting its effectiveness when applied to large water quality datasets. Decision Trees provide models that are easily comprehensible and interpretable by partitioning data according to the most valuable attributes. Nevertheless, they are prone to overfitting and may need pruning in order to improve generalization[26]. Table 2 shows the discrepancy in accuracy rates when the train and test split is modified. This illustrates that a model that experiences a greater amount of training exhibits higher accuracy in comparison to a model that undergoes a greater amount of testing.

Table 2: Accuracy at 8:2 VS 9:1

S.No.	Algorithm	Accuracy (8:2)	Accuracy (9:1)
1	Random Forest	68%	70.12%
2	K-Nearest Neighbor	61.7%	59.14%
3	Decision Tree	60.3%	58.84%

The accuracy of the output saw significant changes due to the division of the data into training and testing sets. When the train test split ratio was adjusted to 8:2, there was a decrease in accuracy. The model's ability to accurately determine potability improves with further training, as seen by the significantly low accuracy of alternative train-test splits. Understanding the impact of different ratios between training and testing sets on the performance of a machine learning model is crucial when training it to forecast water quality. Increasing the proportion of the training set from 80% to 90% and decreasing the proportion of the testing set from 20% to 10% resulted in a significant improvement in the accuracy of the model. This enhancement may be ascribed to many sources. First and foremost, a bigger training set offers a greater amount of data for the model to acquire knowledge from. This may result in improved generalization and a decrease in overfitting. In addition, the larger training set size enables more effective adjustment of hyperparameters, which is crucial for enhancing the model's performance. The smaller size of the testing set may have mitigated the influence of any extraneous or atypical data points, leading to more precise predictions. In summary, the larger training set and smaller testing set in our research resulted in a more resilient and precise machine learning model for predicting water quality.

Conclusions

Concluding this study on water quality prediction using machine learning methods, it is crucial to thoroughly examine the ramifications and importance of our results, especially in the context of civil engineering. In this undertaking, we set out on a quest to use data-driven methods to tackle a critical issue in environmental engineering: guaranteeing the quality and safety of our water supplies. Using three different machine learning algorithms - K-Nearest Neighbors (KNN), Decision Trees, and Random Forest - our objective was to estimate water quality and determine the best efficient technique for making these predictions. Our thorough investigation uncovered valuable insights into the complex relationship between many water quality measures and the forecasting capacities of these algorithms.

The Random Forest algorithm is superior than the K-Nearest Neighbors algorithm, which in turn is superior to the Decision Tree algorithm.

Out of the three algorithms examined, Random Forest stood out as the best performer, demonstrating exceptional accuracy and resilience in predicting changes in water quality. The effectiveness of ensemble approaches in dealing with complex environmental datasets was highlighted by its capability to handle non-linear correlations and interactions among various variables, resulting in accurate predictions. This discovery has significant implications for civil engineers responsible for developing proactive methods to protect water resources from pollution and deterioration. The research illustrates that sophisticated machine learning algorithms, like the Random Forest, are very efficient at predicting water quality. Utilizing these models to precisely evaluate water quality may result in prompt responses, possibly averting waterborne illnesses and guaranteeing the safety of water for diverse purposes. By incorporating these prognostic models into current water monitoring systems, decision-making processes for water management authorities may be enhanced, resulting in better public health outcomes and the implementation of sustainable environmental policies.

In addition, while Random Forest received much attention, the comparative performance of KNN and Decision Trees provided vital insights into the subtle trade-offs involved in selecting machine learning models. Although KNN and Decision Trees have significantly lower accuracy, they shown impressive performance, especially in situations where interpretability and computational economy are crucial. Civil engineers possess a comprehensive range of tools that allow them to customize predictive models to fit unique situations and limitations. This enables them to make well-informed decisions when it comes to water management methods.

The Random Forest method demonstrated superior accuracy in forecasting water quality owing to its inherent robustness and effectiveness in handling complicated information. Random Forest is an ensemble learning technique that builds many decision trees and combines their results to generate a final prediction. This approach helps to mitigate the problem of overfitting and improves the model's capacity to generalize. This ensemble technique reduces the influence of noise and variability in the dataset by calculating the average of predictions made by several trees. Moreover, Random Forest is adept at handling datasets that have a large number of variables and intricate relationships between them, as shown in water quality prediction. In this context, factors such as pH,

turbidity, dissolved oxygen, temperature, phosphates, and nitrates interact in non-linear manners.

The algorithm's resilience to noise and outliers enhances its exceptional performance, as it exhibits less susceptibility to abnormalities and anomalies in the data. Every decision tree in the forest is constructed by using a random selection of features and samples, so effectively mitigating noise and minimizing the impact of outliers. Furthermore, Random Forest automatically offers valuable information about the value of features, emphasizing the most critical factors and enhancing forecast accuracy by prioritizing key indications that have a major impact on water potability.

Optimizing hyperparameters is crucial for enhancing the performance of the Random Forest model. By optimizing parameters such as the number of trees, maximum depth, and minimum samples required to split a node, the model achieves the highest possible prediction accuracy without being too complicated or overly simple. By optimizing the model's performance on both training and unseen test data, this modification reduces the likelihood of overfitting.

By incorporating machine learning methods into water quality prediction, we, as caretakers of our planet's limited water resources, are taking a significant stride in the field of civil engineering. In addition to its current uses in predicting and observing, this interdisciplinary method has the capacity to fundamentally transform our comprehension of environmental processes, leading to inventive strategies for reducing pollution, adapting to the effects of climate change, and guaranteeing fair access to clean water for communities across the globe.

Anticipating the future, this is not the final destination of the journey. The exploration and advancement of this nascent field has the capacity to unlock fresh opportunities for sustainable water management by overcoming disciplinary divisions and formulating holistic approaches that harmonize ecological imperatives with human requirements. Civil engineers may use data, technology, and collaborative experience to ensure that water quality is not only forecast but also protected for future generations.

References

[1] U. Ahmed, R. Mumtaz, H. Anwar, et al., "Efficient water quality prediction using supervised machine learning," Water, vol. 11, no. 11, p. 2210, 2019. doi: 10.3390/w11112210.

[2] T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. Maashi, "Water quality prediction using artificial intelligence algorithms," Applied Bionics and Biomechanics, 2020. doi: 10.1155/2020/6659314.

[3] S. B. H. S. Asadollah, A. Sharafati, D. Motta, and Z. M. Yaseen, "River water quality index prediction and uncertainty analysis: a comparative study of machine learning models," Journal of Environmental Chemical Engineering, vol. 9, no. 1, p. 104599, 2021. doi: 10.1016/j.jece.2020.104599.

[4] M. Azrour, Y. Farhaoui, M. Ouanan, and A. Guezzaz, "SPIT detection in telephony over IP using K-means algorithm," Procedia Computer Science, vol. 148, pp. 542–551, 2019. doi: 10.1016/j.procs.2019.01.027.

[5] S. Bekesiene, I. Meidute-Kavaliauskiene, and V. Vasiliauskiene, "Accurate prediction of concentration changes in ozone as an air pollutant by multiple linear regression and artificial neural networks," Mathematics, vol. 9, no. 4, p. 356, 2021. doi: 10.3390/math9040356.

[6] G. Ciulla and A. D'Amico, "Building energy performance forecasting: a multiple linear regression approach," Applied Energy, vol. 253, p. 113500, 2019. doi: 10.1016/j.apenergy.2019.113500.

[7] T. Deng, K.-W. Chau, and H.-F. Duan, "Machine learning based marine water quality prediction for coastal hydro-environment management," Journal of Environmental Management, vol. 284, p. 112051, 2021. doi: 10.1016/j.jenvman.2021.112051.

[8] 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," Modeling Earth Systems and Environment, 2018. doi: 10.1007/s40808-017-0406-9.

[9] El Bilali and A. Taleb, "Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment," Journal of the Saudi Society of Agricultural Sciences, vol. 19, no. 7, pp. 439–451, 2020. doi: 10.1016/j.jssas.2020.08.001.

[10] S. H. Ewaid, "Water quality evaluation of Al-Gharraf river by two water quality indices," Applied Water Science, vol. 7, no. 7, pp. 3759–3765, 2017.

[11] O. Griffiths, H. Henderson, and M. Simpson, Environmental Health Practitioner Manual: Common wealth of Australia. Accessed: Aug. 10, 2021.

[12] Guezzaz, Y. Asimi, M. Azrour, and A. Asimi, "Mathematical validation of proposed machine learning classifier for heterogeneous traffic and anomaly detection," Big Data Mining and Analytics, vol. 4, no. 1, pp. 18–24, 2021. doi: 10.26599/BDMA.2020.9020019.

[13] Q. Guo, T. Zhuang, Z. Li, and S. He, "Prediction of reservoir saturation field in high water cut stage by bore-ground electromagnetic method based on machine learning," Journal of Petroleum Science and Engineering, vol. 204, p. 108678, 2021. doi: 10.1016/j.petrol.2021.108678.

[14] H. Haghiabi, A. H. Nasrolahi, and A. Parsaie, "Water quality prediction using machine learning methods," Water Quality Research Journal, vol. 53, no. 1, pp. 3–13, 2018. doi: 10.2166/wqrj.2018.025.

[15] R. D. Harkins, "An objective water quality index," Journal (Water Pollution Control Federation), vol. 46, no. 3, pp. 588–591, 1974.

[16] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Access, vol. 8, pp. 76516–76531, 2020. doi: 10.1109/ACCESS.2020.2989857.

[17] J. O. Ighalo, A. G. Adeniyi, and G. Marques, "Artificial intelligence for surface water quality monitoring and assessment: a systematic literature analysis," Modeling Earth Systems and Environment, vol. 7, no. 2, pp. 669–681, 2021. doi: 10.1007/s40808-020-01041-z.

[18] M. Imani, M. M. Hasan, L. F. Bittencourt, K. McClymont, and Z. Kapelan, "A novel machine

learning application: water quality resilience prediction model," Science of the Total Environment, vol. 768, p. 144459, 2021. doi: 10.1016/j.scitotenv.2020.144459.
[19] D. Kapadia and N. Jariwala, "Prediction of tropospheric ozone using artificial neural network (ANN) and feature selection techniques," Modeling Earth Systems and Environment, 2021. doi: 10.1007/s40808-021-01220-6.
[20] R. Kicsiny, "Multiple linear regression based model for solar collectors," Solar Energy, vol. 110, pp. 496–506, 2014. doi: 10.1016/j.solener.2014.10.003.
[21] M. J. V. Kumar and K. Samalla, "Design and development of water quality monitoring system in IoT," International Journal of Recent Technology and Engineering, vol. 7, no. 5, p. 7, 2019.
[22] D. Li and S. Liu, "System and platform for water quality monitoring," Water Quality Monitoring and Management, 2019. doi: 10.1016/B978-0-12-811330-1.00003-X.
[23] 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.
[24] Lumb, T. C. Sharma, J.-F. Bibeault, and P. Klawunn, "A comparative study of USA and Canadian water quality index models," Water Quality, Exposure and Health, vol. 3, no. 3–4, pp. 203–216, 2011.
[25] J. Mabrouki, M. Azrour, A. Boubekraoui, and S. El Hajjaji, "Intelligent system for the protection of people," in Intelligent Systems in Big Data, Semantic Web and Machine Learning, Springer, 2021, pp. 157–165.
[26] J. Mabrouki, M. Azrour, D. Dhiba, Y. Farhaoui, and S. E. Hajjaji, "IoT-based data logger for weather monitoring using Arduino-based wireless sensor networks with remote graphical application and alerts," Big Data Mining and Analytics, vol. 4, no. 1, pp. 25–32, 2021. doi: 10.26599/BDMA.2020.9020018.
[27] J. Mabrouki, M. Azrour, G. Fattah, D. Dhiba, and S. E. Hajjaji, "Intelligent monitoring system for biogas detection based on the Internet of Things: Mohammedia, Morocco city landfill case," Big Data Mining and Analytics, vol. 4, no. 1, pp. 10–17, 2021.
[28] J. Mabrouki, G. Fattah, N. Al-Jadabi, Y. Abrouki, D. Dhiba, M. Azrour, and S. E. Hajjaji, "Study, simulation and modulation of solar thermal domestic hot water production systems," Modeling Earth Systems and Environment, 2021. doi: 10.1007/s40808-021-01200-w.