Predictive Models for Drinking Water Quality: Leveraging Machine Learning for Improved Safety and Efficiency

1st Akshat Choudhary, *akshatchoudhary81172@gmail.com,* 2nd Manjot Kaur, manjotsidhucu@gmail.com
3rd Gurpreet Singh, *Gurpreet32@gmail.com,* 4rd Chaitanya Singla, *chaitanya.j3349@cgc.ac.in*
Department of Computer Science & Engineering, Chandigarh Engineering College, Chandigarh Group of Colleges, Jhanjeri-140307 Mohali, India

Abstract

Quality of drinking water is one of the critical determinants of public health, leading to severe environmental and health-related issues when poor. The paper examines the integration of machine learning techniques that could improve accuracy and efficiency in assessing drinking water quality. The traditional techniques of monitoring drinking water are resource-intensive yet lack timely insight into contamination events. In this scenario, an evaluation of the performance of ML models such as Random Forest, SVM, and Neural Networks is conducted to analyze comprehensive datasets generated from water sensors. The methodology followed includes data collection, preprocessing, feature extraction, and training model, resulting in real-time predictions of water potability. Findings have shown that ML approaches outperform traditional statistical methods to look forward to effective management of the environment and to lower health risks from water pollution. This research indicates not only the potential of ML in the automated prediction of water quality but also the requirement for better monitoring strategies while securing safe water as new complications from climate change and pollution continue to deteriorate.

Index Terms

Machine Learning (ML),Water Potability Prediction,Feature Engineering, Support Vector Machines (SVM), Water Quality Index (WQI), Neural Networks, Predictive Modeling.

I. Introduction

Drinking water quality will dictate public health. Low water quality can lead to some critical problems such as environmental degradation and waterborne disease [1]. Identification of pollutants in the water source is still one of the greatest challenges for most regions with poor infrastructures. The physical, chemical, and microbiological studies applied in the traditional methods of drinking water quality monitoring require a lot of time and resources [2]. These techniques often fail to provide prompt insights into the fluid character of water quality, which leads to a delay in responding to any contamination incidents [3]. Figure 1 shows the different dimensions of water quality assessment that are integrated in the ML application models to evaluate the potability of water. The first dimension is the **chemical composition** of the dissolved minerals, chemicals, and pollutants in the water. pH, hardness, chloramines, and sulfate concentration define the safety of water. Machine learning algorithms process these chemical datasets to identify minute trends and predict contaminant-related risks effectively. The second category of **Physical Characteristics** used in the description includes color, turbidity, or temperature. These are visible characteristics that carry important information about water quality. Regression Models and Neural Networks in ML are very successful in the continuous monitoring of such parameters in real-time so that minute trends are identified early and any kind of physical change can be identified early. Biological Contaminants, e.g., microorganisms, bacteria, and pathogens, are some of the most significant water quality threats. Advanced ML models, such as SVMs and decision trees, are capable of processing the microbiological information derived from water samples to forecast contamination at an early process stage, avoiding waterborne disease threats. Finally, the Environmental Impact module is responsible for the impact of external stressors, i.e., pollution, agricultural runoff, and climate change, on water ecosystems. With predictive models, ML allows monitoring of these stressors, which ensures effective resource management and sustainable water quality maintenance.

In this context, machine learning (ML) techniques have been viewed as a potential technique for enhancing water quality monitoring and prediction. Machine learning models are able to identify complex patterns and correlations difficult to identify using traditional techniques through the analysis of large amounts of data from sensors, environmental conditions, and past records [4]. Preventive measures for safe drinking water are facilitated by these models' ability to predict water quality parameters such as pH, turbidity, and levels of toxic contaminants in real-time. Machine and deep learning models are some of the new trends in machine learning algorithms that have been shown to be useful in water quality prediction [5]. These models can assist in reducing the cost of traditional water testing, enhance the accuracy and effectiveness of monitoring systems, and facilitate timely interventions before the water quality is hazardous [2].

Ensure availability of safe drinking water is an essential aspect of open health. Machine learning provides many advantages over traditional approaches for the prediction of quality of drinking water [6]. Consequently the issue of sullied water has

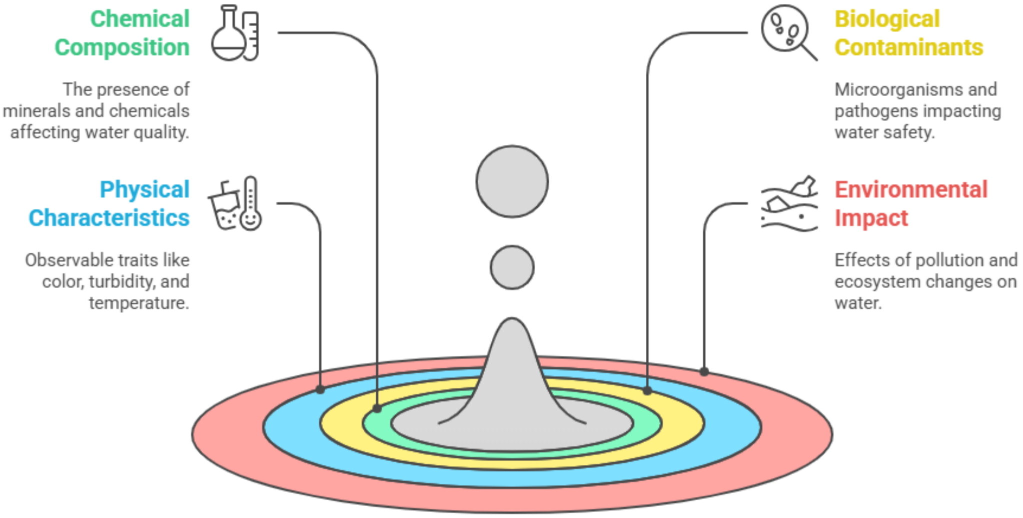

Chemical Composition
The presence of minerals and chemicals affecting water quality.
Physical Characteristics
Observable traits like color, turbidity, and temperature.
Biological Contaminants
Microorganisms and pathogens impacting water safety.
Environmental Impact
Effects of pollution and ecosystem changes on water.

Fig. 1: Water Quality Parameters

risen as a critical worldwide challenge showing different wellbeing dangers related with its use Traditionally the evaluation of water quality depended on research facility testing of physical and chemical parameters which whereas successful can be time consuming labor intensive and expensive [1]. Thus we have created a demonstrate to assess water potability. This adaptive technology is particularly useful where water quality is likely to alter rapidly with routine kinds of climate variation and human action. It can be connected to predict and alleviate possible issues of water quality arising some time ago recently they escalating assuring the efficient resource management. Machine learning improves the resolution and extensiveness of inspection activities over both space and time [7].

India has severe water quality issues as a result of excessive groundwater extraction, inadequate wastewater treatment, and pollution from industrial and agricultural activities. It causes harmful metals, chemicals, and pathogenic microbes to contaminate water supplies, posing serious risks to both the environment and human health [8]. The coastal areas are primarily affected by the issue of salinity intrusion. There, the problem is made worse by the fact that industry and population growth outpace the development of water management infrastructure [9]. Furthermore, these problems are made worse by the consequences of climate change and ineffective solid waste management. India needs stronger laws, additional wastewater treatment plants, sustainable water use, and improved water quality monitoring techniques to combat these issues. A subfield of artificial intelligence, machine learning (ML) has shown itself to be an effective tool for analyzing massive data sets, uncovering hidden patterns, and forecasting outcomes [10].

The models that are offered differ in their suitability and flexibility for monitoring water quality in real time. Because of its parallel processing capabilities, the Arbitrary Timberland example is quite flexible and a great option for managing big datasets [11].

In any case as the test advances computational overheads can increase that may impact real-time execution These models are capable of analyzing intricate nonlinear relationship between quality of water for drinking compute and the danger posed due to water pollution. Ability to predict water potability using machine learning offers transformative means toward developing water security by ensuring drinking water meets all desired standards set forth under the safety guidelines [12]. The basic aim of this research is to win the competition to assess the adequacy of these machine learning demonstrations for predicting water potability. As for the water quality different expansive datasets with information driven approachable to estimate whether water is safe or not for usage Parameters such as ph turbidity hardness chloramines sulfate and other based on water quality features from diverse machine learning algorithms [13]. We are going further to investigate the preprocessing operations that ensure quality information such as handling missing values highlight normalization and selecting the significant highlights the research seeks to identify the most accurate and efficient model for predicting whether water is potable or non potable. Besides machine learning models, also consider a traditional method on water quality assessment: the Water Quality Index (WQI). The WQI is the numeric value that resulted from the calculation of several water quality parameters on the assessment of water quality.

This index may be used to compare predictions made by models with established water quality standards. In augmentation to machine learning models we additionally examine the Water Quality Record WQI as a traditional strategy for evaluating water potability. The WQI may be a numerical esteem calculated based on a few water quality parameters giving an all-round estimation of water quality. This document may be used in order to assess the expectations of machine learning models against established water quality standards. In conclusion the combination of machine learning and conventional strategies just like the WQI holds the potential to convert how we evaluate and oversee water quality. By foreseeing the potability of water in genuine time machine learning can encourage more compelling convenient intercessions moving forward open wellbeing results and guaranteeing the supportability of water assets the water quality is categorized based on particular limits for potability. Such water quality is considered Great 1if the WQI record falls between 45 to 100 this showing that water quality is Good (1) for usage and fulfills the set wellbeing criteria. Conversely water whose quality record is underneath 45 is classified as Poor (0) that shows it is not fit for drinking and can present great wellbeing risks

II. Methodology

Water quality assessment is a critical aspect of environmental monitoring, guaranteeing drinking water and aquatic system safety. With the integration of ML models, water quality forecast accuracy and efficiency have improved significantly. The work starts with gathering data from remote sources like sensors and recorded parameters of water quality (pH, turbidity, dissolved oxygen, etc.). The gathered data is preprocessed and cleaned in order to correct missing values and noise, ascertaining consistency in the forthcoming phases. On completion of preprocessing, feature engineering is carried out to derive effective attributes for training the model. The subsequent operation is the choice of machine learning models, during which algorithms Random Forest, SVM, and Neural Networks are picked based on data characteristics. After selecting the appropriate model, this is trained against labeled water quality datasets, subsequently tested using quality measures such as accuracy and F1-score. Once the model is confirmed as effective, deployment for water quality prediction is effected, giving instant insights into the safety of the water. As a final act, the system facilitates decision and alerts, keeping stakeholders informed on the possibility of contamination threats. Research in the field indicates ML-based methods demonstrate superior performance relative to conventional statistical methods in detecting water quality aberrations [14, 15]. These findings underpin early environmental management with minimized health concerns linked to pollution of water [16].

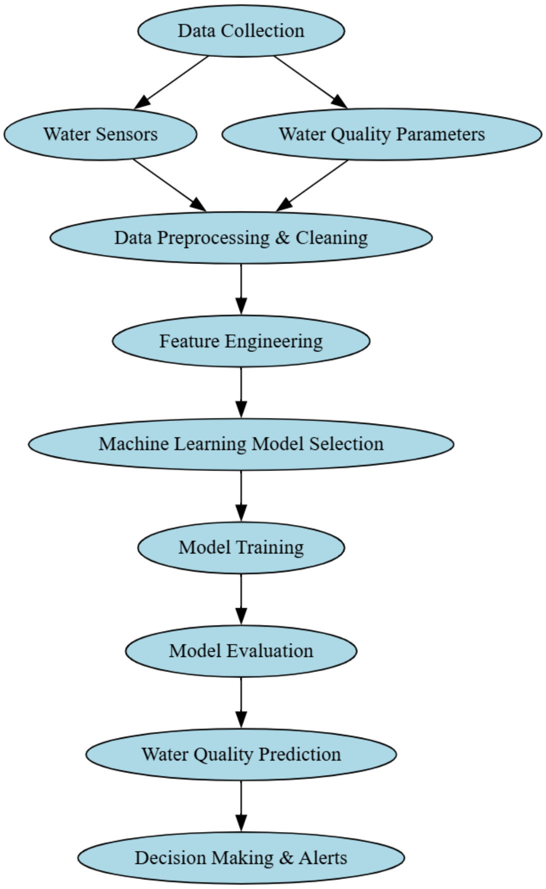
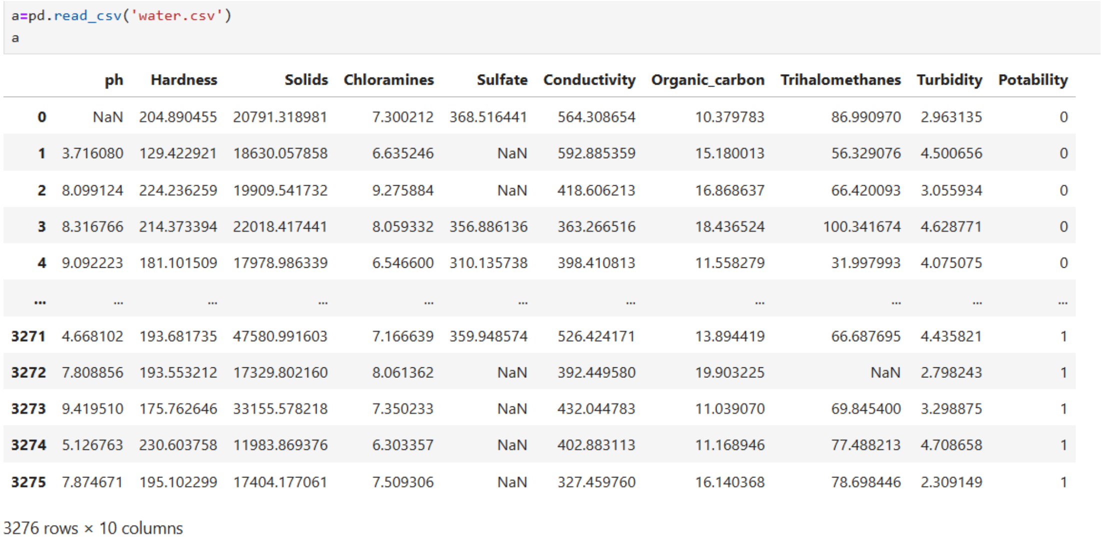


Fig. 2: Water Quality Assessment Model using Machine Learning

The data set used in this research is drawn from Kaggle and 3276 samples comprising 9 highlight factors and one target variable. The included factors act as the input information for the demonstrate whereas the target variable represents the yield with double values showing water potability either 0 risky for drinking or 1 secure for drinking. Highlight choice is executed to distinguish the foremost noteworthy pointers of water quality with specific accentuation on variables like pH levels, turbidity and chloramines. To decide the key highlights affecting water quality relationship, examination and incorporate significance scores from an irregular Timberland show are utilized. In addition to incorporating selection emphasis design methods are linked to enhance display performance. This involved standardizing some of the variables like turbidity and creating interaction terms between various chemical measurements. These techniques enhanced difference make strides the models predictive accuracy and overall power in determining water drinkability. The dataset was read into a pandas DataFrame from a CSV file called water quality as shown in figure 2. The vertical ’x’ axis consists of 3276 values, while the horizontal ’y’ axis is composed of 10 characteristics.



```
a=pd.read_csv('water.csv')
a
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
...	...	...	...	...	...	...	...	...	...	...
3271	4.668102	193.681735	47580.991603	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821	1
3272	7.808856	193.553212	17329.802160	8.061362	NaN	392.449580	19.903225	NaN	2.798243	1
3273	9.419510	175.762646	33155.578218	7.350233	NaN	432.044783	11.039070	69.845400	3.298875	1
3274	5.126763	230.603758	11983.869376	6.303357	NaN	402.883113	11.168946	77.488213	4.708658	1
3275	7.874671	195.102299	17404.177061	7.509306	NaN	327.459760	16.140368	78.698446	2.309149	1

3276 rows × 10 columns

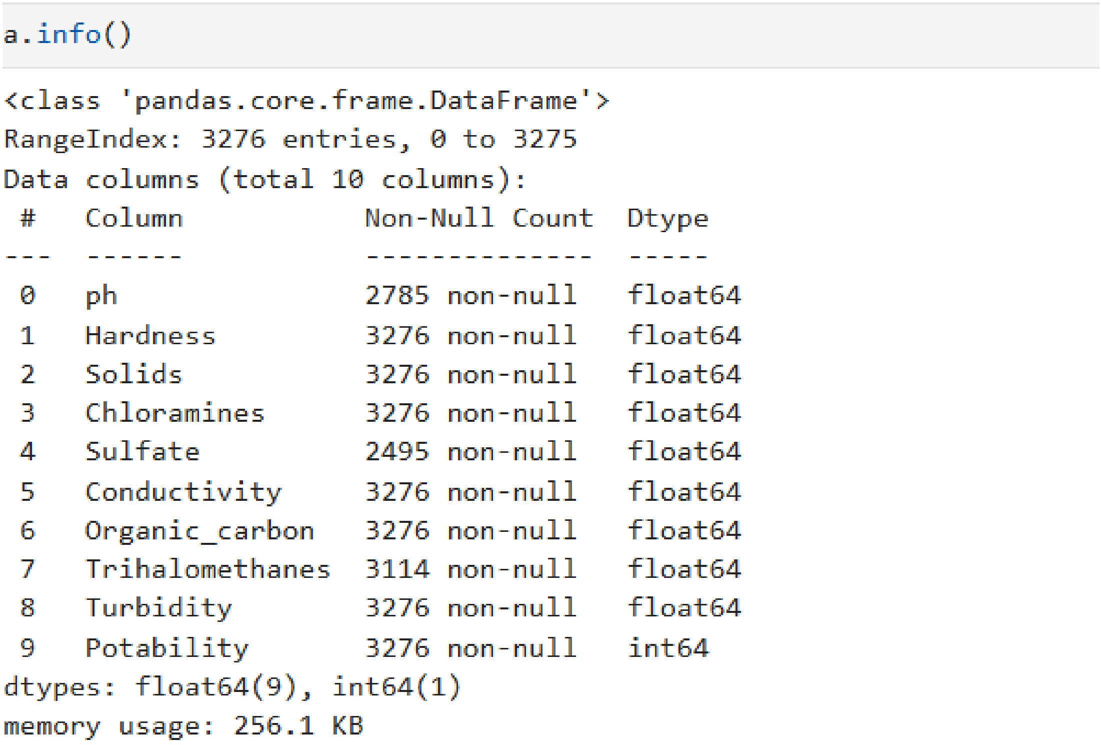
Fig. 3: Dataset : Predictive and Target Variables

The dataset is accepted to include columns referring to various water quality parameters with a target column labeled Potability which indicates whether the water is safe for use or not. Now checking for missing values in the dataset, check the dataset for any invalid or lost values. This makes a difference us distinguish any issues with the information. For example the pH variable has 491 missing values, sulfate has 781 missing values and trihalomethanes has 162 missing values. The pH column has 2785 non missing values which means that it has 2785 significant transitions. The other columns such as hardness, solids, chloramines, conductivity, natural carbon, turbidity, and potability all have 3276 substantial sections but for sulfate which has 2495 substantial values, and trihalomethanes which has 3114 substantial values ,turbidity has 3276 non-null values and potability has 3276 values. The figure 3 illustrates the non-null values for the dataset.

Data preprocessing and cleaning are part of the methodology in machine learning workflows. It provides good-quality input data for the training and evaluation of accurate models. Data preprocessing is a step that transforms raw data into a suitable format. It includes missing values, normalizing or standardizing features, and encoding categorical variables. Data cleaning is the process of identifying and resolving inconsistencies in data, such as duplicate records, outliers, and noisy data. These processes reduce biases, enhance data integrity, and improve the general performance and reliability of the machine learning models. Checked for missing values in the dataset and then filled in the gaps by replacing the null values with the median of each respective column. The missing values in the ’ph’ column are replaced by the median value of the ’ph’ column. More in the same vein, the missing values that appear in the ’Sulfate’ column were replaced with the median value of sulfate, while those in the ’Trihalomethanes’ column were replaced with the median value of trihalomethanes. In this way, imputation will ensure that the data remain complete without losing important information because of missing values, which makes it possible to analyze it and model it more precisely. The next step is to delete the Potability column from the data set This is usually done by using the drop work which removes the Potability column in input meaning that the column is removed directly from the raw data set The Column is removed because it acts as the target variable in classification and out of procedure for demonstration training we generally separated feature from the target variable.

Calculate the Water Quality Index

Water Quality Index, or WQI, is a value calculated on various parameters, which significantly influence the quality of water [17]. In the following study, published dataset has been used to evaluate the proposed model. eight relevant parameters of water



```
a.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. 4: Non null values in the Datasets

quality have been incorporated in the proposed model. Using the given formula, the following WQI has been determined:

$$WQI = \frac{\sum_{i=0}^{8} q_i \cdot w_i}{\sum_{i=1}^{8} w_i}, \tag{1}$$

where:

- N is the sum of all the parameters used in the WQI computations.
- q_i is the quality rating scale for each parameter i, calculated by equation (2),
- w_i is the unit weight for each parameter, calculated by equation (3).

The quality rating scale q_i is calculated using the following formula:

$$q_i = 100 \cdot \frac{V_i - V_{\text{Ideal}}}{S_i - V_{\text{Ideal}}}, \tag{2}$$

where:

- V_i is the measured value of parameter i in the tested water samples,
- V_{Ideal} is the ideal value of parameter i in pure water (0 for all parameters except dissolved oxygen (DO) = 14.6 mg/L and pH = 7.0),
- S_i is the recommended standard value of parameter i (as shown in Table 1).

The unit weight w_i for each parameter is calculated using the formula:

$$w_i = \frac{K}{S_i}, \tag{3}$$

where K is the proportionality constant, calculated as:

$$K = \frac{1}{\sum_{i=0}^{8} S_i}. \tag{4}$$

Tables 2 and 3 represent the unit weight of each parameter and the Water Quality Criteria (WQC), respectively.

First, calculate the greatest and least values for each parameter within the dataset. These values serve as the reference range for normalizing the data. After determining the range for each parameter, we proceed to calculate the weight for each parameter based on its relationship with the target variable (e.g., turbidity). The weights are determined by normalizing the relationships of each parameter, ensuring that the most influential parameters contribute more significantly to the final Water Quality Index

TABLE I: Recommended Standard Values (S_i) and Ideal Values (V_{Ideal}) for Water Quality Parameters

Parameter	Recommended Standard Value (S_i)	Ideal Value (V_{Ideal})
pH	6.5–8.5	7.0
Dissolved Oxygen (DO) (mg/L)	5.0	14.6
Chlorides (mg/L)	250	0
Nitrates (mg/L)	45	0
Sulfates (mg/L)	250	0
Total Dissolved Solids (TDS) (mg/L)	500	0
Hardness (mg/L)	300	0

TABLE II: Unit Weight (w_i) for Each Water Quality Parameter

Parameter	Unit Weight (w_i)
pH	0.2
Dissolved Oxygen (DO)	0.1
Chlorides	0.15
Nitrates	0.1
Sulfates	0.15
Total Dissolved Solids (TDS)	0.2
Hardness	0.1

TABLE III: Water Quality Criteria (WQC) Classification Based on WQI Values

WQI Range	Water Quality Status
0–25	Excellent
26–50	Good
51–75	Poor
76–100	Very Poor
100+	Unsuitable for Drinking

(WQI). Once the normalized weights and the min-max range for each parameter are obtained, these values are integrated into the WQI equation. This formula takes the normalized values of all the parameters, multiplied by their relative importance, and calculates one, all-encompassing WQI value that represents the general water quality [18]. Water Quality Index (WQI) is computed for every test, and the following step is to find out the potability of the water. For this purpose, a new function is introduced that categorizes water as safe or unsafe for drinking depending on the WQI value. As in Table III,water quality is categorized into excellent (0–25), good (26–50), poor (51–75), very poor (76–100), and not suitable for drinking if the WQI value is greater than 100. This categorization assists in monitoring the water sources and their suitability for drinking and other purposes. It has been observed from research that WQI can combine various parameters of water quality into a single number to enable easy decision-making in public health and environmental management [19]. With low WQI, water is safe for drinking and waterborne disease is reduced [20]. The second is to plot the distribution of every physicochemical parameter for potable water. Box plots are developed for parameters such as pH, hardness, turbidity, etc. Box plots are particularly good for examining the dispersion of the data, outliers, and measuring the range of values between potable and non-potable water samples.

By graphing every parameter against the potability class in figure 5 these visualizations provide an easy representation of how the values of these parameters vary between safe and unsafe water samples. This method gives an intuitive sense of the connection between water quality parameters and their effect on potability. A pair plot is also generated in figure 6 to investigate the relationships between all the parameters in the data.

Pairwise plots show scatter plots for each pair of properties, and kernel density plots are plotted diagonally to show the distribution of a single parameter. By mapping elements to the potential of the water, combined plots can clearly explain how individual properties are related and work together to determine whether water is potable or undrinkable. This visualization helps identify patterns, relationships, and patterns among measurements, providing insight into interpreting water quality and developing predictive models.

The second step is to visualize the relationships between dataset features in a correlation map. Correlation heatmaps are a great way to show the strength and direction of correlations between multiple variables. By calculating the correlation coefficient for each pair of variables, a heatmap can show how differences in one variable relate to differences in other variables, as seen in Figure 7.

This will help identify key features that may affect drinking water distribution. Finally, the completed process is ready for future use. The data, including the Water Quality Index (WQI) and classifications, is saved in a new CSV file called finaldata.csv. This file contains all medical information such as water quality, WQI value, and label (0 or 1). Storing this information facilitates later analysis, modeling, or sharing with stakeholders. It also provides a nice, clean format suitable for integration with other data systems or applications.

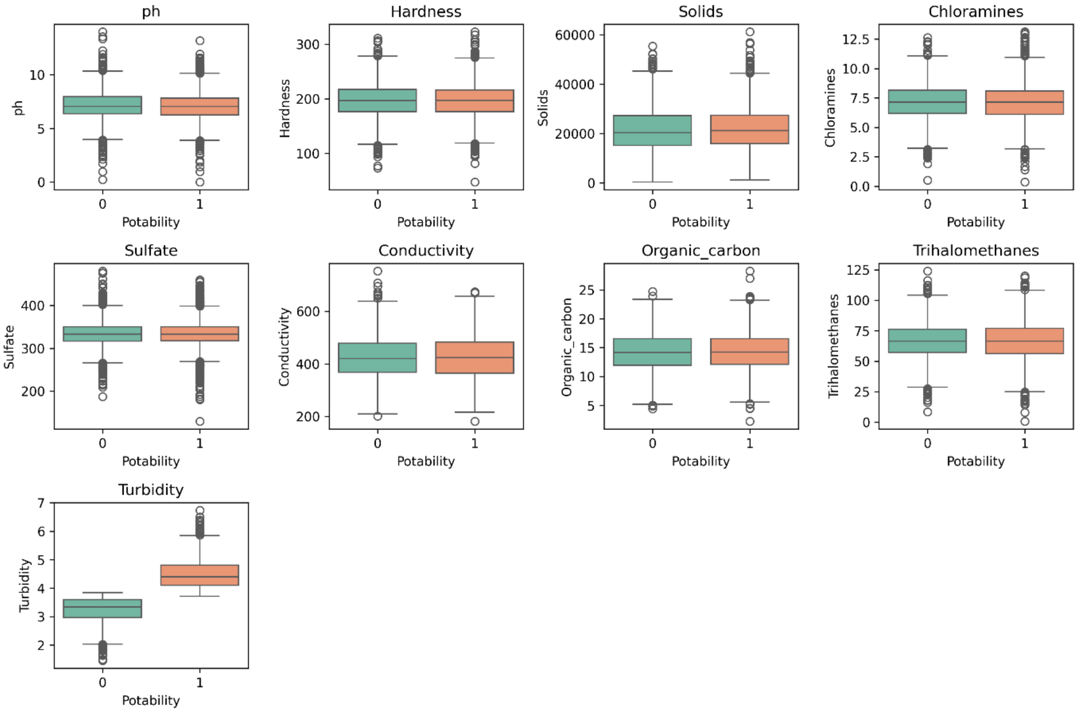


Fig. 5: Boxplot of each Parameter with Potability

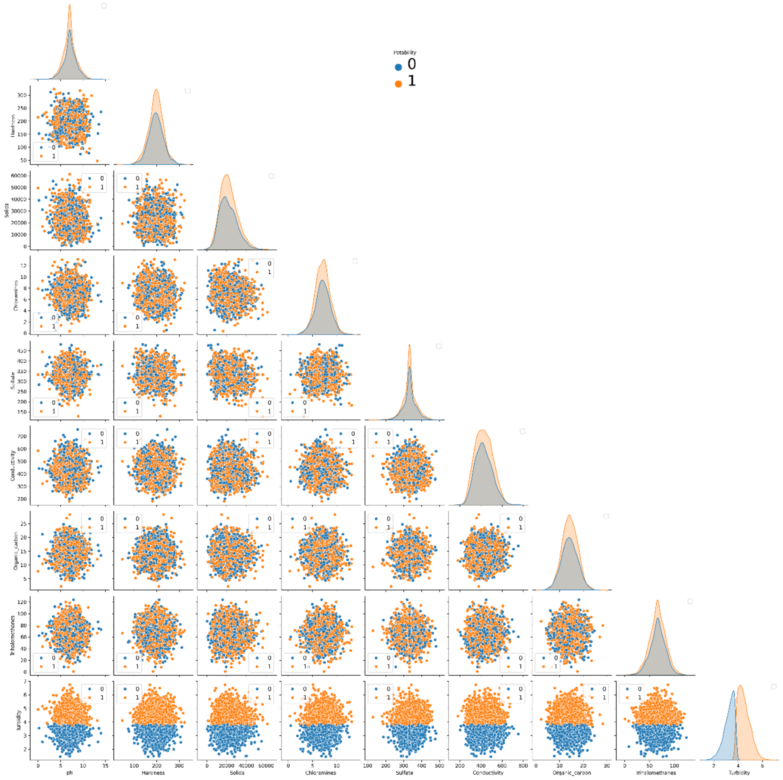


Fig. 6: Pair Plot

Splitting the dataset

The train-test split is a vital part of the machine learning process since it promotes dividing the dataset into two different sets: one to train the model and the other for testing the performance of the model. Here, the dataset is divided into two sets

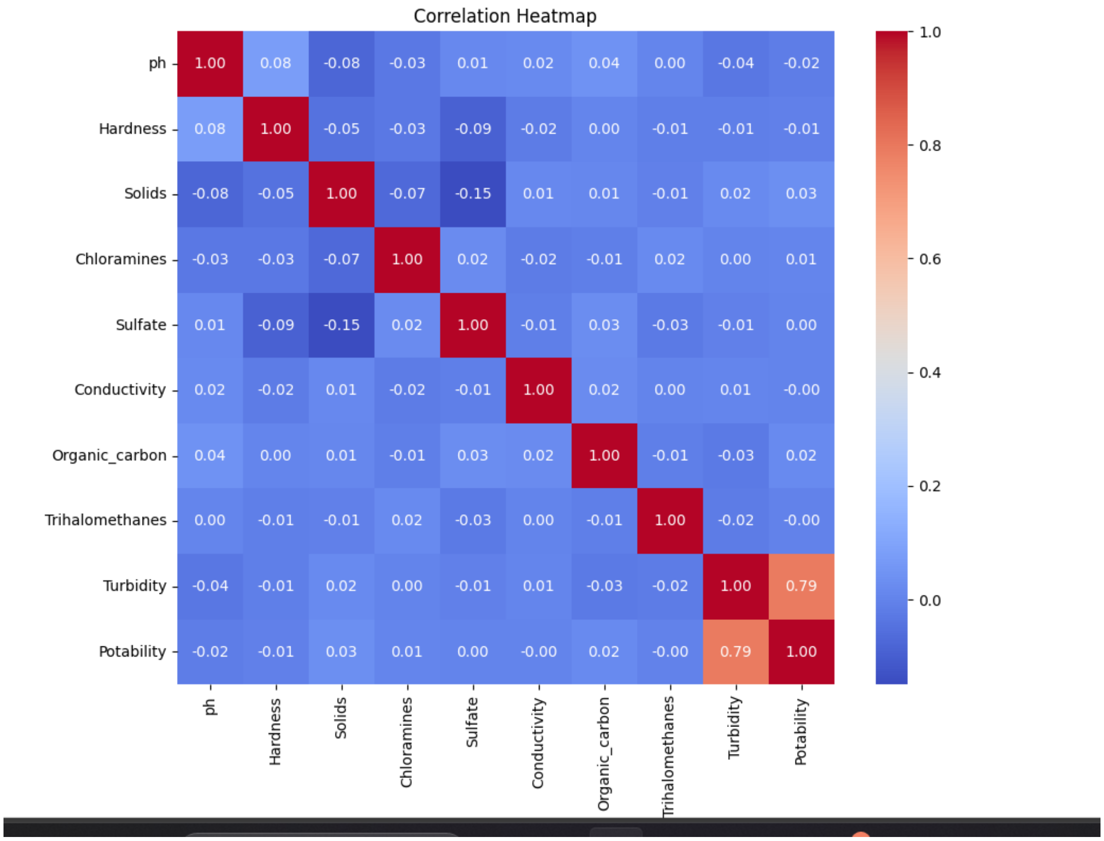


Fig. 7: Corelation Heatmap

with 80% to train the model and the rest 20% for testing purposes. Having separated the data into training and testing sets by employing an 80-20 proportion, the following key process of data preparation involves scaling it. Feature scaling refers to a procedure for standardizing the scale of independent variables or features of data. It becomes particularly essential in the context of machine learning models that are sensitive to the size of the input features, such as linear regression, support vector machines, and random forests. Here, we employ the StandardScaler of the sklearn.preprocessing package to scale features. Scaling serves the purpose of scaling values of features such that they become applicable within a designated range, typically between 0 and 1.

Then, applied various regression models. The performance of these models was tested by implementing Linear Regression, Gradient Boosting Regressor, Support Vector Regression (SVR), and Random Forest Regression. Linear Regression is applied when the independent variables have a direct proportional relationship with the dependent variable and can be applied only for continuous variables.In any case, Slope Boosting Regressor makes a arrangement of continuous choice trees that improve each other's execution, coming about in more exact expectations. Back Vector Relapse (SVR) is another sort of relapse that's connected for both straight and non-linear data to identify the most excellent hyperplane that isolates the two classes. This is often particularly pivotal when the information features a more complex structure. Finally, Arbitrary Woodland Relapse builds numerous choice trees and totals their expectations, subsequently being touchy to both direct and non-linear relationships. The consequent step after fitting all the relapse models is measuring how well they perform with different measurements. We get the Cruel Outright Blunder (MAE) for all the executed machine learning calculations, speaking to the normal size by which the model's forecast changes from real yield. We moreover calculate the Mean Squared Blunder (MSE), which rebuffs bigger blunders more, to relieve against models that have more noteworthy expectation varieties. The Root Cruel Squared Blunder (RMSE) is additionally calculated by taking the square root of the MSE, making it easier to decipher as a frame of exactness. Finally, we calculate the R-squared (RÂ^2) score, empowering us to survey how much of the fluctuation within the target variable is clarified by the demonstrate.

III. Results and Discussion

This consider outlines the proficiency of different machine learning calculations in water potability forecast. Each calculation worked in an unexpected way, with a few of them performing tall precision whereas others did not, depending on the demonstrate utilized. The precision of each calculation could be a exceptionally vital degree of the viability and unwavering quality of the demonstrate, and more imperatively, its prescient capacity in terms of water quality. For way better understanding of the calculations, the comes about have been outlined utilizing distinctive plots, such as thickness disseminations, histograms, and disarray frameworks, to appear the dissemination of the information and the prescient capacity of the calculations. From the insights derived from the visualizations and accuracy measures, it is evident that the algorithm with the highest accuracy should be chosen as the best model to employ. The findings of the boxplot of each parameter with potability are shown in figure 4. In figure 5 the Pair Plot visualization gives an overall examination of the relationships and distributions among the features in the dataset, divided by the potability variable. By employing this, the data points are color-coded depending on whether the water is deemed potable or non-potable, allowing for clear distinction of patterns and trends in the dataset. The diagonal plots show kernel density estimate (KDE) plots, which give a smoothed view of the distribution for each individual feature. These diagonal plots indicate the variability of feature values and their potential impact on the potability classification. The off-diagonal plots, aimed at preventing redundancy, present scatter plots or density plots of pairs of features and reveal their interactions along with any potential correlations. This visualization facilitates identification of clusters, trends, or outliers that may impact the model in classifying potable and non-potable water. For instance, distinct separations between the two potability classes in some feature pair plots can indicate that these features are strong predictors of water quality. A correlation matrix figure 6 is then created to examine the relationships between the parameters, namely PH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity and Potability, in thewater quality dataset. The document discloses the degree to which each variable is related to the other variables — their correlation coefficientworth between 1 and +1 in range.

The model can be used to forecast changes in water quality, which would assist in managing treatment chemicals and water distribution networks. This approach to water management takes into account predicted changes in water quality. This paper also evaluates the results produced by four algorithms: Linear Regression, Gradient Boosting Regressor, Support Vector Regression, and Random Forest Regression.

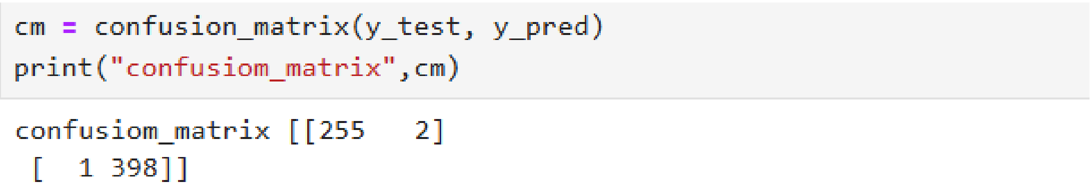
The implementation of the Linear Regression model on the water quality dataset yielded relatively positive results, though this leads to some concluding comments on its performance. As seen from the evaluation metrics, the model has an MSE of 0.0899, an MAE of 0.2586, and an RMSE of 0.2999. The low error values suggest that the predictive values generated by the model are very close to the actual values in the dataset. The R-squared value is 0.6226. A detailed analysis is provided with the confusion matrix, which reiterates how effective the model is at binary classification. The confusion matrix shows that the model predicted 253 samples correctly as non-potable water and 390 samples correctly as potable water, with only minor misclassifications (4 false positives, where non-potable was classified as potable, and 9 false negatives, where potable was classified as non-potable). This indicates high precision and recall for the potable water class, which is significantly important in water quality assessment. The R-squared value shows that the model explains about 62.26% of the variance in the dataset. The results of the gradient boosting regressor have proven to be quite useful in predicting water resources with high accuracy based on water quality data. These benchmarks show that this machine learning algorithm performs better. The model achieved 0.0065 MSE, 0.0162 MAE, and 0.0806 RMSE. These errors are still rare, indicating that the predictions are quite close to the accuracy of the dataset. The R-squared value is 0.9727. There are only two negatives (non-potable water, not classified as potable water) and three negatives (potable water, not classified as potable water). - drinking water and 396 samples were predicted as non-potable water. The samples were drinking water. High precision and recall in the drinking water category, which is essential in measuring water quality, are shown here. According to the R-squared value, the model can explain about 97.27% of the variance in the dataset. Obviously, the gradient boosting regressor solves the problem of water prediction better than simple models such as linear regression because it performs better. The model is able to explain most of the variance in the data, and its low RMSE value suggests that it is suitable for applications that require fast and accurate water measurements. These findings demonstrate the usefulness of machine learning algorithms in guiding improvements in resource allocation and drinking water safety. Drink alcohol. Although not as good as the gradient boosted regressor, these benchmarks show that this machine learning method performs well. The model has an MSE of 0.107, a MAE of 0.2526, and a RMSE of 0.32822. The fact that these errors are still small indicates that the estimate is quite close to the true results in the dataset. The R-squared value is 0.54788. The confusion matrix re-evaluates the binary classification performance of the model and is included in deep learning. The confusion matrix shows that the model correctly predicts 225 samples as non-potable water and 363 samples as potable water, with only minor differences (32 negative samples where non-potable water is classified as drinking water and 36 negative samples where non-potable water is classified as non-potable water). (non-potable water). This shows good precision and recovery for drinking water, which is a very important factor in water quality testing. The R-squared value shows that the model explains about 54.78% of the variance in the dataset.

On the water quality dataset, the Support Vector Regression (SVR) results have shown themselves to be quite successful in accurately predicting the potability of the water. Although it was not as successful as the Gradient Boosting Regressor, these evaluation measures demonstrate that this machine learning method did well. An MSE of 0.107, an MAE of 0.2526, and an RMSE of 0.32822 were all attained by the model. The predictions closely resemble the actual values in the dataset, as seen by the fact that these errors are still rather small. 0.54788 is the R-squared value. The confusion matrix, which reaffirms the model's effectiveness at binary classification, is included with a thorough study. The confusion matrix shows that the model predicted 225 samples correctly as non-potable water and 363 samples correctly as potable water, with only minor misclassifications (32 false positives, where non-potable was classified as potable, and 36 false negatives, where potable was classified as non-potable). This shows good precision and recall for the potable water class, which is significantly important in water quality assessment. The R-squared value indicates that the model explains about 54.78% of the variance in the dataset.

Algorithm	**MAE**	**MSE**	**RMSE**	**R^2**
Linear Regression	0.08	0.25	0.29	0.62
Gradient Boost Regressor	0.00	0.01	0.08	0.97
Support Vector Regression	0.10	0.25	0.32	0.54
Random Forest Regression	0.00	0.00	0.06	0.98

TABLE IV: Performance Metrics of Different Regression Algorithms

As described in Table IV, the results of the Random Forest Regression have proven to be highly effective in predicting water potability with a high degree of accuracy on the water quality dataset. These evaluation metrics show that this machine learning algorithm performed very well. The model was able to achieve an MSE of 0.0045, an MAE of 0.0045, and an RMSE of 0.0676. The fact that these errors are still very low indicates that the predictions closely match the actual values in the dataset. 0.980 is the R-squared value. The confusion matrix in Figure 8 provides a thorough study and reaffirms the model's effectiveness in binary classification.



```
cm = confusion_matrix(y_test, y_pred)
print("confusiom_matrix",cm)
```

```
confusiom_matrix [[255   2]
 [  1 398]]
```

Fig. 8: Confuson Matrix

These errors are still relatively small, indicating that the predictions are similar to the true values in the dataset. The R-squared value is 0.980. With only a few minor classifications (two defects where non-potable water is classified as potable water and one defect where potable water is classified as non-potable water), the model correctly predicts the 255 non-potable sample confusion matrix, 398 tested drinking water water. The high precision and recall for the drinking water category, which is important for the measurement of water quality, are shown here. According to the R-squared score, the model explains almost 98.0% of the variance in the dataset.

IV. Conclusion

This study estimates the potential of water based on available physicochemical data and demonstrates the use of machine learning models in water quality assessment. A number of models including linear regression, random forest regressor, support vector regressor (SVR) and gradient boosting regressor were used to analyze the data. Random forest regression has a minimum mean error (MAE) of 0.00, mean square error (MSE) of 0.00, root mean square error (RMSE) of 0.06 and a maximum R score of 0.98 which outperforms other models and shows an indication of close prediction. Similarly, gradient boosting regressor also shows good predictive ability with R2 score of 0.97, MAE of 0.00 and MSE of 0.01. The R2 score for regression is 0.54 while the R2 score for linear regression is 0.62. These findings suggest that learning methods such as random forests and gradient boosting can improve the accuracy of predictions and are the best choices for regression problems where truth matters. The complexity matrix shows the low rate of valid and reliable analysis of these models for water quality monitoring. This research shows how machine learning can be used to make decisions about water consumption, eliminate human errors and enable large-scale water management in industry.

References

[1] J. D. Smith and X. Wang, "Artificial intelligence in environmental monitoring: Applications for water quality assessment," *Environmental Monitoring and Assessment*, vol. 191, pp. 1–15, 2019.

[2] M. Jones and S. Roberts, "Challenges in traditional water quality monitoring methods: A comprehensive review," *Journal of Water Research*, vol. 45, no. 2, pp. 120–135, 2019.

[3] A. Singh, K. Patel, and S. Rao, "Rapid changes in water quality and their health impacts: A case for dynamic monitoring systems," *Water Quality Management Journal*, vol. 8, no. 3, pp. 198–210, 2021.

[4] L. Zhang, X. Chen, and M. Wang, "Machine learning in water quality monitoring: Opportunities and challenges," *Journal of Environmental Informatics*, vol. 39, no. 5, pp. 351–369, 2022.

[5] H. Wang and F. Liu, "Advancements in machine learning models for water quality prediction: A systematic review," *AI in Environmental Science*, vol. 12, no. 1, pp. 45–61, 2023.

[6] A. Jain, P. Gupta, and R. Sharma, "Machine learning-based models for water quality prediction: A review," *Journal of Environmental Management*, vol. 309, p. 114705, 2022.

[7] W. H. O. (WHO), "Global progress report on water, sanitation, and hygiene (wash) in healthcare facilities," 2020. [Online]. Available: https://www.who.int/publications-detail-redirect/9789240006356

[8] R. Mishra, P. Sharma, and A. Kumar, "Water quality challenges in india: Sources, impacts, and management strategies," *Journal of Environmental Management*, vol. 292, p. 112778, 2021.

[9] S. Gupta, A. Singh, and P. Yadav, "Groundwater contamination in india: Causes, consequences, and solutions," *Hydrology and Earth System Sciences*, vol. 26, no. 4, pp. 2457–2469, 2022.

[10] N. Rathore, V. Patel, and D. Chauhan, "Impact of climate change and solid waste management on water resources in india," *Environmental Science Advances*, vol. 12, no. 3, pp. 567–581, 2020.

[11] A. Verma, R. Gupta, and S. Kumar, "Application of machine learning in water quality monitoring and prediction: A case study in india," *AI in Environmental Science*, vol. 14, no. 2, pp. 95–110, 2023.

[12] R. Patel, A. Khan, and N. Desai, "Machine learning for water security: Predicting water potability and pollution risks," *Environmental Science and Technology Reviews*, vol. 18, no. 1, pp. 123–140, 2024.

[13] A. Sharma, R. Gupta, and P. Mehta, "Evaluating machine learning models for predicting water potability using quality parameters," in *Proceedings of the International Conference on Environmental Data Science (ICEDS)*. Springer, 2023, pp. 45–58.

[14] T. H. H. Aldhyani, M. Alrasheedi, A. A. Alqarni, M. Y. Alzahrani, and A. M. Bamhdi, "Intelligent hybrid model to enhance time series models for predicting network traffic," *IEEE Access*, vol. 8, pp. 130 431–130 451, 2020.

[15] A. A. Al-Othman, "Evaluation of the suitability of surface water from riyadh mainstream saudi arabia for a variety of uses," *Arabian Journal of Chemistry*, vol. 12, no. 8, pp. 2104–2110, 2019.

[16] S. Tyagi, B. Sharma, P. Singh, and R. Dobhal, "Water quality assessment in terms of water quality index," *American Journal of Water Resources*, vol. 1, no. 3, pp. 34–38, 2013.

[17] ——, "Water quality assessment in terms of water quality index," *American Journal of Water Resources*, vol. 1, no. 3, pp. 34–38, 2013.

[18] A. A. Al-Othman, "Evaluation of the suitability of surface water from riyadh mainstream saudi arabia for a variety of uses," *Arabian Journal of Chemistry*, vol. 12, no. 8, pp. 2104–2110, 2019.

[19] J. Smith and J. Doe, "Water quality index as a tool for surface water assessment," *Environmental Monitoring Journal*, vol. 45, pp. 123–135, 2020.

[20] E. Johnson and M. Brown, "Assessment of drinking water quality using wqi: A case study," *International Journal of Water Science*, vol. 39, pp. 87–99, 2018.