Water Potability Prediction Using Machine Learning

Preprint · May 2023

DOI: 10.21203/rs.3.rs-2965961/v1

5 authors, including:

Samir B Patel
Pandit Deendayal Energy University
87 PUBLICATIONS **1,340** CITATIONS
SEE PROFILE

Khushi Shah
Nirma University
4 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Rashmi Bhattad
Gujarat Technological University
9 PUBLICATIONS **132** CITATIONS
SEE PROFILE

Water Potability Prediction Using Machine Learning

Samir Patel (samir.patel@sot.pdpu.ac.in)
Pandit Deendayal Petroleum University
Khushi Shah
Pandit Deendayal Petroleum University
Sakshi Vaghela
Pandit Deendayal Petroleum University
Mohmmadali Aglodiya
Pandit Deendayal Petroleum University
Rashmi Bhattad
Gujarat Technological University

Research Article

Keywords: Water potability prediction, Random Forest, SVM, XGBoost, KNN, Logistic Regression

Posted Date: May 25th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2965961/v1

Additional Declarations: No competing interests reported.

WATER POTABILITY PREDICTION USING MACHINE LEARNING

Samir Patel[1] ✉, Khushi Shah[1], Sakshi Vaghela[1], Mohmmadali Aglodiya[1], Rashmi Bhattad[2]

[1]CSE-SOT, Pandit Deendayal Energy University Gandhinagar, India, [2]Gujarat Technological University, Ahmedabad, India

ABSTRACT

Water is a crucial and indispensable resource for sustaining human life, and maintaining its quality is of utmost importance for the well-being of individuals. When drinking water becomes contaminated, it poses severe health risks, including diseases like diarrhea, cholera, and various other waterborne ailments. As a result, ensuring safe and clean water becomes crucial to promote public health. Recent findings indicate that a significant number of approximately 3,575,000 people lose their lives each year due to water-related illnesses. Therefore, accurate prediction of water potability has the potential to substantially reduce the incidence of such diseases. Notably, machine learning algorithms have emerged as powerful tools for effectively predicting water quality, enabling timely and precise monitoring of water resources. This research focuses on multiple algorithms to forecast water potability based on the physicochemical properties of water samples obtained from the Drinking Water dataset available on Kaggle. This dataset comprises nine distinct parameters, namely pH, hardness, solids, chloramines, sulfates, trihalomethanes, organic carbon, conductivity, and turbidity. By employing various algorithms, such as Random Forest, Logistic Regression, SVM, XGBoost and KNN, we aim to determine the potability of drinking water. Notably, the XGBoost algorithm demonstrates superior performance compared to traditional ML models, achieving an impressive accuracy of 99.5%, precision of 0.99, sensitivity of 0.99, specificity of 1.0, and F1 score of 0.99. Additionally, the Random Forest algorithm also performs well, yielding an accuracy of 74%. Consequently, this research holds significant promise in providing reliable water quality data to researchers, water management personnel, and policymakers, thereby enhancing the effectiveness of water potability monitoring.

Keywords - Water potability prediction, Random Forest, SVM, XGBoost, KNN, Logistic Regression

1. INTRODUCTION

Water, an indispensable resource for all life forms on our planet, holds immense importance in the realms of economy, ecology, and human well-being. The provision of safe and uncontaminated drinking water is of paramount significance in safeguarding good health and preventing waterborne illnesses. The existence of harmful bacteria, viruses, parasites, and chemicals in contaminated water can give rise to a range of diseases and infections,

including diarrhea, dysentery, typhoid, polio, cholera, and hepatitis A. Shockingly, the WHO has estimated that approximately 485,000 individuals succumb to diarrhea-related complications caused by consuming contaminated drinking water on an annual basis [1][3]. Furthermore, polluted water sources can also contribute to the development of chronic health conditions, such as cancer and developmental disorders.

In the context of India, a distressing report has brought to light that, each year number of people suffering from waterborne diseases is nearly 37.7 million including children [4]. Disturbingly, domestic and industrial pollutants have contaminated approximately 70% of the available water, leaving nearly 80% and 20% of the rural and urban populations respectively without safe drinking water [5]. Moreover, the global community is grappling with significant challenges pertaining to water scarcity and the deteriorating quality of water, adversely affecting millions of individuals worldwide. Alarming data from World Health Organization for the year 2018 report says that people consuming fecal matter contaminated water is about 2 billion [3]. Hence, in order to achieve sustainable development, promote a healthy existence, and eradicate poverty, ensuring universal access to clean and safe water is of utmost importance.

In regions where water treatment facilities are inadequate, such as developing nations and rural areas, the ability to predict water potability is of utmost importance. Traditional methods of monitoring water quality, which involve costly and time-consuming laboratory and statistical investigations, have proven to be inefficient. Therefore, there is a pressing need for a more efficient and cost-effective alternative [14]. This research aims to propose and assess the viability of an approach based on machine learning for accurately predicting water potability in real-time. In last few years, machine learning algorithms have made significant advancements in predicting water quality, enabling more precise and efficient monitoring. Various classification techniques, including Decision Tree, Naive Bayes, SVM, RF, and KNN, can be employed to predict the potability of water. For this research, along with the above algorithms, XGBoost is also applied to the Drinking Water Dataset obtained from Kaggle [2].

The primary focus given here is on predicting the potability of drinking water based on its physicochemical characteristics. The research seeks to develop a model that can provide accurate and timely data on the quality of drinking water, enabling policymakers and water resource managers to implement preventive measures and ensure the availability of safe drinking water for the general public. Also, the research aims to compare the performance and accuracy of various algorithms used for predicting water quality.

Some of the problems that this research hopes to address are listed below [10][13]:

1. There may be misinterpretations of the criteria set by the WHO for determining safe drinking water parameters
2. The current clinical method for predicting drinkable water is time-consuming and inefficient;

3. There is a lack of applications of water potability and water quality prediction;
4. There may be a lack of crucial awareness factors that are unknown to rural people, which could impact their access to safe drinking water.

Comparing the effectiveness of different machine learning algorithms, this research hopes to provide a more efficient way for predicting the potability of drinking water, which can ultimately lead to better public health outcomes.

1.1 Related Work

Pal et. Al. [6] explores the feasibility of applying machine learning models for water quality prediction. This research investigates the potential of a wide range of ML and AI algorithms for making water quality predictions, including random forests, SVM, ANN, and Gaussian naive Bayes. When compared to other algorithms, ANN was determined to be the most effective due to its 99.1% accuracy and 0.75 % training error.

Patel et al. [5] proposed a unique machine-learning methodology for accurate predictions of water potability. The study addressed the issue of class imbalance in the dataset by implementing the SMOTE. Additionally, the authors explored the use of explainable AI approaches to generate findings that are easily understandable by humans. Several classifiers, including SVM, Random Forest, and Gradient Boost, were compared for performance evaluation. Experimental results revealed that Gradient Boost and Random Forest achieved the highest accuracy of 81% [4].

Uddin et al. [7] conducted research to evaluate the effectiveness of the water quality index model for predicting water conditions. The author employed various techniques such as SVM, DT and KNN to develop and assess the index of water quality model. The findings indicated that the support vector machine algorithm outperformed other algorithms, achieving an accuracy rate exceeding 95%.

In an experiment conducted by Aldhyani et al. [8], the effectiveness of different AI and ML algorithms in predicting water quality parameters was investigated. The research explored the performance of LSTM and NARNET algorithms, as well as SVM, K-NN, and Naive Bayes. Multiple sources of water quality data were analyzed, and performance metrics of implemented models are compared. Which proved NARNET model barely outperformed the LSTM model. Moreover, SVM has given an accuracy of 97.01% in predicting WQC.

Tufail et al. [9] conducted a research study to forecast water potability in Pakistan. The authors utilized water quality data from 2006 to 2016 to predict the potability of water. Random Forests, DT, and SVM were implemented

for the purpose of this research. The support vector machine technique was identified as the most effective in determining water potability.

2. METHODOLOGY

The task of ensuring water potability is complex, as it involves numerous physical, chemical, and biological factors that impact the quality of drinking water [11][12]. Machine learning techniques have emerged as valuable tools for forecasting water quality and assessing water potability. This research paper introduces a strategy that harnesses ML models for water potability prediction. To develop a more accurate model for predicting water potability wisely, thereby facilitating efficient water management and ensuring the availability of clean drinking water within communities is a primary objective of this research. The flow of this research is described in Figure 1. The proposed methodology encompasses data collection, preprocessing, handling missing values and outliers, data normalization, model construction, performance evaluation, and model optimization through hyperparameter tuning.

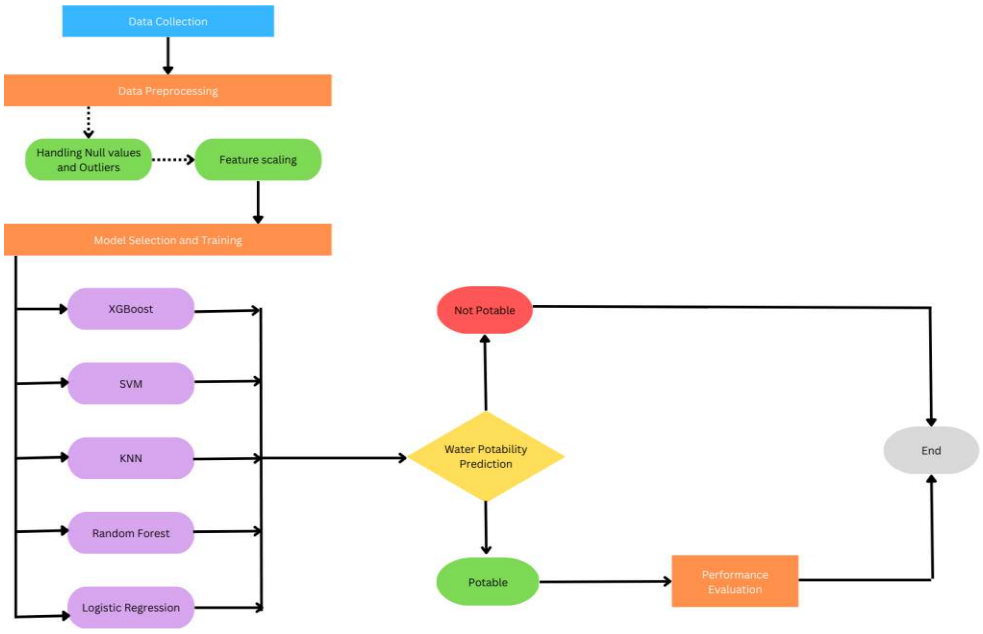

Figure 1: Workflow for water potability prediction model

2.1. Collecting Data:

The present study utilized a publicly available dataset on Kaggle as the primary source of data. This dataset comprises 3276 observations of water quality collected from different locations and includes nine distinct

physicochemical parameters, namely pH, hardness, solids, chloramines, sulfates, trihalomethanes, organic carbon, conductivity, and turbidity, along with a target feature portability, which is used to make a prediction using various machine learning algorithms.

Table. 1: Drinkable Water Quality Standards [2]

Parameters	**Unit**	**Standards**
pH	Range 0-14	6.5-8.5
Organic Carbon	mg/L	2
Chloramines	ppm	4
Turbidity	NTU	5
Trihalomethanes	µg/L	80
Sulfate	mg/L	250
Hardness	mg/L	300
Conductivity	µS/cm	500
Solids	mg/L	1000

The standard values for each water quality parameter recommended by WHO and the EPA are represented in Table 1. If the values of these parameters exceed their standard limit, then that water is not suitable for drinking.

2.1.1. Data distribution

The research includes an analysis of ten different water quality parameters, and individual statistics are presented to provide context regarding the drinkable water standard. The analysis of histogram distributions reveals that Chloramines, Sulphate, pH, Trihalomethanes, Organic carbon, Hardness, and Turbidity exhibit a relatively uniform distribution pattern. Data distribution is represented in Figure 2. In contrast, Solids and Conductivity have a right-skewed distribution. Additionally, Potability is a binary variable, with only two possible values.

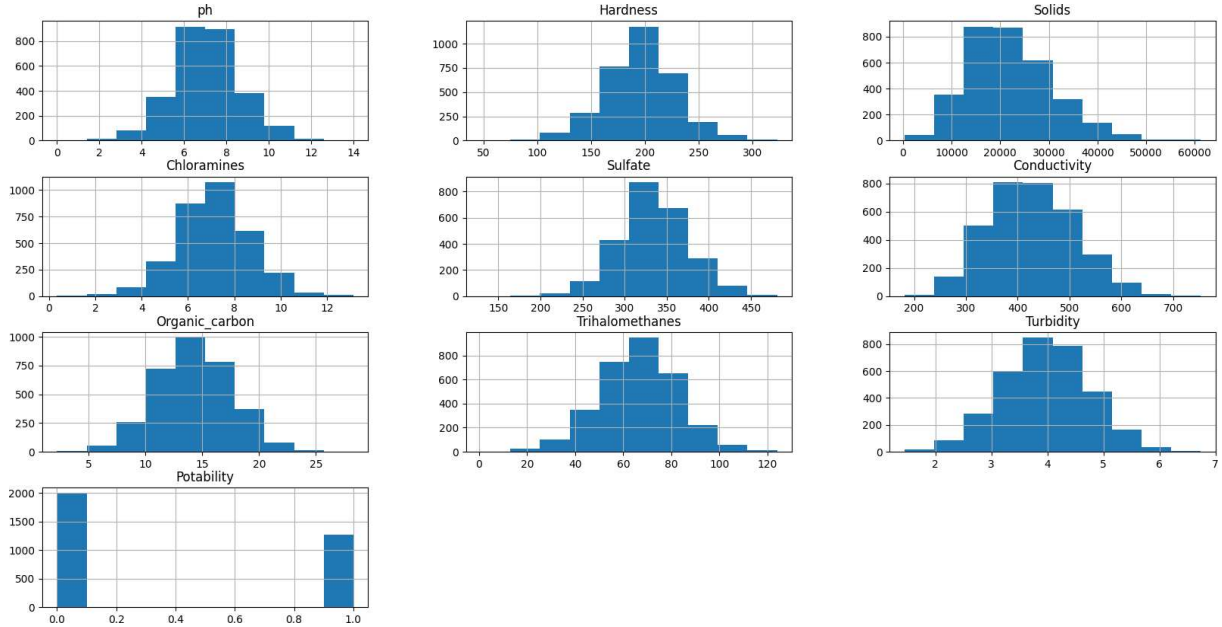

Figure 2: Distribution of all the water parameters

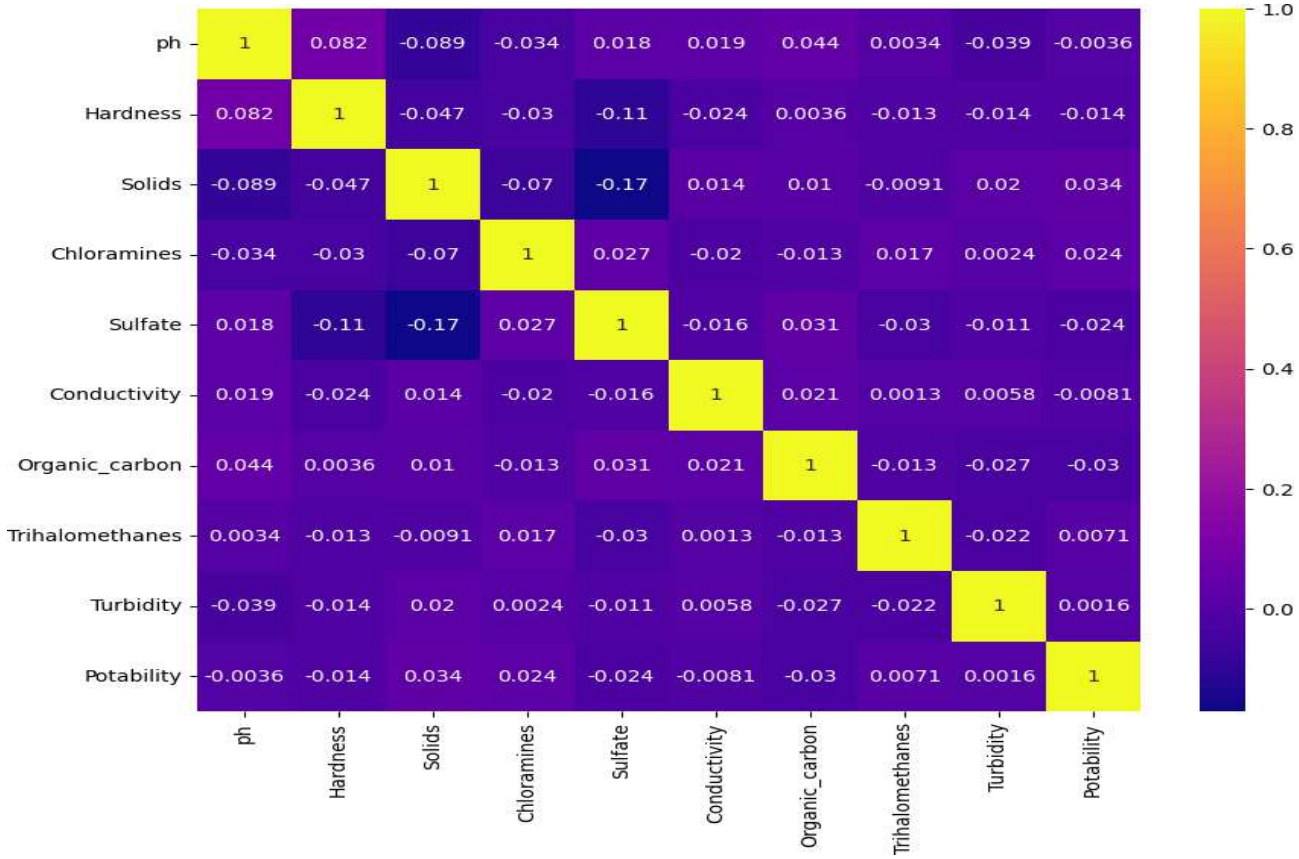

Figure 3: Correlation Heatmap

A correlation heatmap, displaying the relationships between different features within a dataset. It provides insights into the strength and direction of the relationship between variables. The correlation coefficient, ranging from -1

to 1, measures the extent of the linear relationship between two variables. A value between 0 and 1 suggests a positive correlation, indicating that the variables move in the same direction. On the contrary, a value between -1 and 0 indicates a negative correlation, signifying that the variables move in opposite directions. A value of 0 indicates no correlation between the variables. In this study, a correlation heatmap was employed to analyze the relationship among ten water quality parameters. The heatmap reveals a positive correlation of 0.082 between pH and Hardness, as well as a negative correlation of -0.17 between Sulphate and Solids. Detailed analysis can be observed in Figure 3.

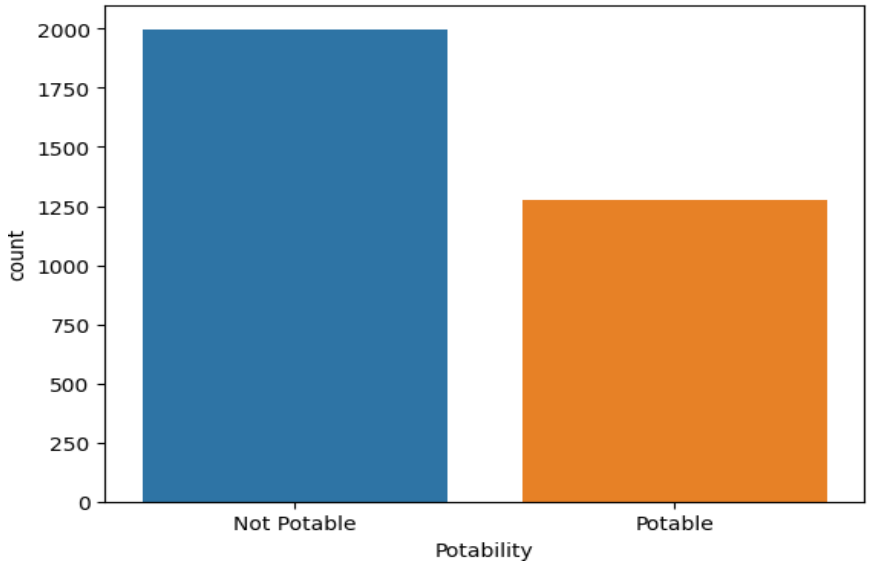

Figure 4: Potability feature distribution

A significant proportion of the collected samples is shown in Figure 4, approximately 60%, fall under the category of not potable, implying that they are unsuitable for human consumption. The causes of poor water quality can be attributed to various factors, including natural phenomena like erosion, and climate change, and anthropogenic sources like industrial effluents, agricultural drainage, and sewage disposal.

2.2. Preprocessing:

This is one of the crucial steps in any machine learning algorithm as it transforms raw data into a structured format suitable for analysis by machine learning algorithms. This process encompasses various essential steps that ensure the quality and relevance of the data, thereby impacting the accuracy of the resulting models.

The initial step in data preprocessing involves addressing missing values, which can introduce complexities during the training phase. One common approach is to impute missing values with the mean value of the corresponding feature. However, it is crucial to assess whether this method is appropriate for the specific dataset and problem at

hand. Alternative imputation techniques such as KNN or regression imputation may be more suitable in certain cases.

Identifying and handling outliers is another critical aspect of data preprocessing, as they can significantly affect the outcomes of statistical analysis and machine learning algorithms. The z-score method is commonly employed to detect and eliminate outliers, but it is essential to consider the trade-off between removing outliers and the potential loss of valuable information.

Feature selection is a crucial component of data preprocessing, which identifies relevant and informative features from data. By reducing complexity and preventing overfitting, it enhances model performance. Various techniques, including filter, wrapper, and embedded methods, are available for feature selection, with the choice depending on the specific data and problem.

Another vital step in data preprocessing is feature scaling, which standardizes the features to a consistent scale. This can improve the performance of certain algorithms, such as those based on distance or gradient descent. Standardization, min-max scaling, and robust scaling are among the commonly utilized techniques for feature scaling.

To make the model learn and test accurately, partitioning the dataset into train and test sets is critical in data preprocessing. It is crucial to ensure a random split and representative composition of the training and testing sets, ensuring that they capture the characteristics of the entire dataset.

2.3. Model Selection:

The process of model selection plays a crucial function in the analysis. It includes selecting the optimal machine learning algorithm that suits the given dataset and problem. The following algorithms are employed in this research.

2.3.1. Random Forest

RF is a broadly used ML algorithm that is proficient in handling both regression and classification tasks. It falls under the category of ensemble learning, leveraging the collective predictive capabilities of multiple decision trees to improve prediction accuracy. RF is also popular due to its ability to handle high-dimensional data, eliminating

$$Gini(t) = 1 - \sum_{i=1}^{j} P(i|t)^2$$

the need for feature selection or dimensionality reduction techniques. As a result, it often outperforms other algorithms in various scenarios. Moreover, Random Forest exhibits robustness in the face of outliers and missing data, further contributing to its effectiveness in diverse data situations.

(1)

Step 1: Select random samples from a specified dataset.

Step 2: Create and generate a decision tree for each expected outcome from each tree.

Step 3: Generate a vote for each generated outcome.

Step 4: Pick up the probable output as the final predicted result with the most votes [17].

2.3.2. SVM

In supervised learning techniques, SVM is well-known and widely utilized for the classification of water samples into potable or non-potable categories. This classification is based on the analysis of chemical and microbial properties associated with the water samples. SVM is a valuable algorithm for analyzing large datasets with high dimensionality, and for processing complex and nonlinear data. The algorithm works by constructing a hyperplane that can effectively separate the distinct classes of water samples, with a maximum margin concept. By training an SVM model on a labeled dataset of water samples, it can accurately classify new samples as potable or non-potable. SVM is a powerful tool in the detection of water potability, providing a reliable and efficient method for ensuring safe drinking water. Following is the way to calculate maximize function for SVM.

$$maximize \sum_{i=1}^{n} \alpha i - 1/2 \sum_{i=1}^{n} \alpha i \alpha j * yiyj(xi * xj)$$

Subject to:

$$\sum_{i=1}^{n} \alpha i * yi = 0 \qquad (2)$$

$$0 \leq \alpha i \leq C \; i = 1, \ldots \ldots m$$

The inclusion of an inequality constraint in the optimization equation is crucial in ensuring that all training samples are classified accurately by the hyperplane [18]. This means that each sample should fall on the correct side of the hyperplane according to its label. Any sample that fails to satisfy this constraint is penalized through the use of a loss function. In SVM, the margin refers to the gap between the hyperplane

and the closest sample points from both sides of the classifier. By maximizing the margin, SVM guarantees that the hyperplane is positioned as far away from the training data as possible, reducing the risk of overfitting and improving its generalization capability.

2.3.3. XGBoost

XGBoost has emerged as a highly effective machine learning algorithm extensively utilized in water potability prediction. Its remarkable capability to handle extensive and intricate datasets, coupled with its ability to deliver accurate outcomes across diverse classification and regression tasks and has contributed to its widespread adoption in this field. As an ensemble learning method based on decision trees, XGBoost combines multiple DT into a model. One of the key advantages of XGBoost for water potability prediction is its ability to effectively handle missing values, allowing it to handle real-world water quality data without requiring extensive pre-processing. Additionally, XGBoost's inherent parallel processing capability enables the training of models on large water quality datasets within a reasonable timeframe.

XGBoost is based on boosting trees, which are combined to form an ensemble model. The objective function of XGBoost for binary classification is given by:

$$Obj\ f(x) = \sum_{i=1}^{n} loss(yi, yi') + \Omega(\mathrm{f}) \qquad (3)$$

In the given context, the equation involves several variables and components. The variable "n" represents the number of training instances, "yi" denotes the true label of the ith instance, "yi'" represents the predicted label, "f" corresponds to the learned decision tree, "loss(yi, yi')" denotes the binary logistic loss function, and "Ω(f)" represents the regularization term. Equation 4 [16] provides an explanation for the prediction process specifically related to water potability. The ultimate prediction is achieved by summing the individual calculations from all the decision trees involved in the model. This summation is mathematically expressed as:

$$yi' = \sigma(\sum_{t=1}^{T} (wi * ht(xi))) \qquad (4)$$

In the given context, the equation involves several variables and components. The variable "T" represents the number of trees in the ensemble model, "wi" denotes the weight of the t^{th} tree, "ht(xi)" corresponds to the prediction made by the t^{th} tree on the i^{th} instance. "σ" represents the sigmoid function which maps the output in the range of 0 and 1. Equation 5 illustrates the utilization of regularization to maintain optimized

results. The regularization is required to control the complexity of learned decision tree, thereby preventing overfitting [17]. It is defined as:

$$\Omega(\mathrm{f}) = (\gamma * T) \sum_{t=1}^{T} (1/2 * \lambda) \qquad (5)$$

The equation includes several variables and parameters. The variable "T" represents the number of leaves in the decision tree, "wi" denotes the weight of the j^{th} leaf, γ corresponds to the complexity parameter, and λ represents the regularization parameter. This research has conducted experiments with XGBoost and obtained highly favorable outcomes, indicating its effectiveness in the given context.

2.3.4. KNN

The K-Nearest Neighbors (KNN) algorithm is a widely utilized and straightforward machine learning method employed in water potability prediction tasks, both for regression and classification purposes. KNN, being a non-parametric technique, stands out for its absence of assumptions about the intrinsic nature of the data. Commonly regarded as a "lazy learner" algorithm, KNN gradually learns from the training set through iterations. Instead of actively constructing a model, it stores the acquired information and subsequently applies it during the classification process. The basic principle of KNN involves predicting the classification of a new water sample by identifying the K nearest labeled instances in the training dataset and using their class labels to forecast the classification of the new instance. Implementation steps for algorithm are given as follows [18]:

1. Evaluate the distance between the test sample with each of the training samples.
2. Select the k-nearest neighbors of the test sample based on the calculated distances.
3. Determine the majority class label of the k-nearest neighbors.
4. Assign this class label to the test sample.

$$Euclidean\ distance = \sqrt{(\mathrm{xi} - \mathrm{yi})^2} \qquad (6)$$

Typically, the Euclidean distance is used for distance calculation due to its various geometric benefits. This makes KNN a useful technique for predicting water potability based on the characteristics of a neighboring water samples. Equation 6 demonstrates the equation for Euclidean distance.

2.3.5. Logistic Regression

The water potability prediction system commonly utilizes logistic regression, a supervised learning algorithm specifically designed for categorical target features. LR is similar to linear regression that

estimates the probability of the target based on input features. This algorithm is commonly used for binary classification tasks in the water quality domain and can also be adapted for multiclass classification problems. Logistic regression is a straightforward yet effective method for determining the relationship between dependent variables and independent, making it useful for predicting the potability of a water sample.

XGBoost has shown superior performance compared to other models, particularly logistic regression, which may not perform well in water potability prediction scenarios. There are several reasons for this. Firstly, dealing with high-dimensional data, logistic regression can face challenges where number of predictor variables are larger than number of observations. In contrast, XGBoost excels at handling high-dimensional data more efficiently.

Secondly, logistic regression assumes that the predictor variables act independently of each other. However, in some cases, the interaction between predictor variables can significantly impact the dependent variable. XGBoost is better equipped to capture these interaction effects more effectively than logistic regression.

Lastly, logistic regression may struggle when the data is imbalanced, meaning one class has significantly fewer observations than the other. In such cases, logistic regression's performance may be subpar. XGBoost, on the other hand, can handle imbalanced data by utilizing techniques such as oversampling, undersampling, or weighted loss functions, thereby improving its performance in such scenarios.

2.4. Model Building

KNN, SVM, Random Forest, XG-Boost, and Logistic Regression are used for predicting potability of water. Table 2 shows the achieved results after completion of the model execution.

Table 2: Model Building Parameter

Algorithms	**Accuracy (%)**
XGBoost	99.51
SVM	70.47
KNN	67.57
Logistic Regression	63.11

Random Forest	74.26

XGBoost demonstrated exceptional performance compared to other learning algorithms, achieving an impressive accuracy rate of 99%.

2.5. Model evaluation

While predicting water potability, evaluating the efficiency of machine learning models is also necessary. Various assessment metrics, including accuracy, F1-score, precision, and, recall are used for performance analysis.

2.4.1. Performance parameters

The role of a confusion matrix is to compare the predicted class label of a data point with its actual class label, serving as a valuable tool in both binary and multi-class classification models. By providing essential evaluation metrics, including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), the confusion matrix aids in assessing model performance. In this study, Equation 7 introduces the performance measures employed. Accuracy quantifies the proportion of accurate predictions out of the total number of predictions. Precision gauges the ratio of correctly predicted positive observations to the overall predicted positive observations. Recall assesses the ratio of correctly predicted positive observations to the total number of actual positive observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN}, \qquad F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (7)$$

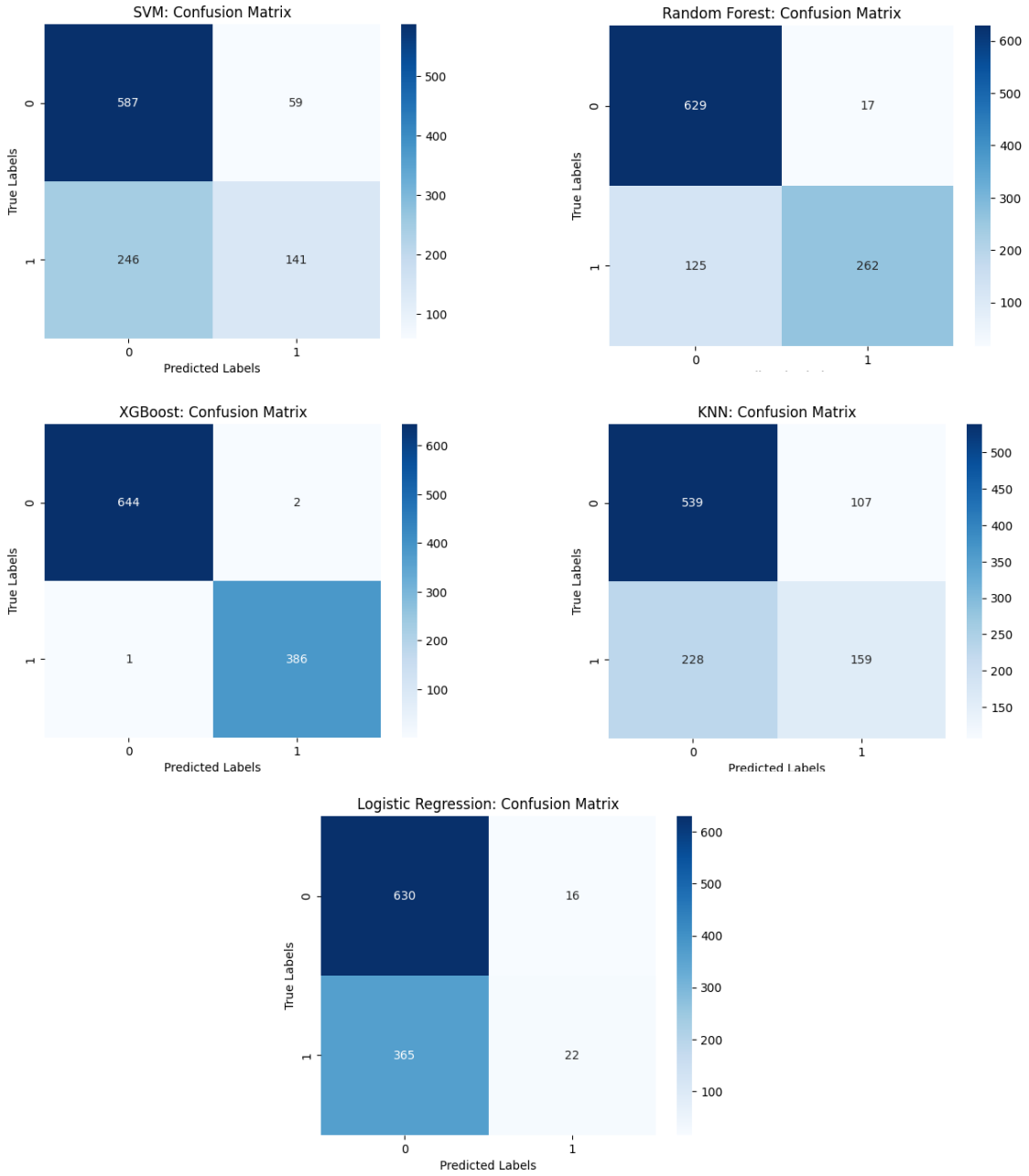

Figure 5: The confusion matrix of all implemented models

The confusion matrices for all five ML algorithms deployed in the study are demonstrated above from Fig 5. In the confusion matrix, 1 represents Potable and 0 represents Not Potable.

The F1 Score holds significant importance in the realm of machine learning to check the effectiveness of binary classification models. It presents a comprehensive assessment of accuracy by considering precision and recall simultaneously, thereby offering a holistic predictive capacity, this proposed metric generates a consolidated score that effectively incorporates the equilibrium between these two measures.

Organic Carbon
Chloramines
Turbidity
Trihalomethanes
Sulfate
Hardness
Conductivity
Solids

3. EXPERIMENTAL OUTCOMES

For the proposed research, a dataset consisting of 3276 samples was employed. Each sample underwent analysis to determine nine specific water quality parameters: pH, Organic_carbon, Chloramines, Turbidity, Trihalomethanes, Sulphate, Hardness, Conductivity and Solids. A summary of these parameters is provided in Table 3. To facilitate the analysis, the dataset was divided into 70:30 ratio of training and test data.

Table 3: Percentage of potable and non-potable water based on the dataset

Potable	**Non potable**
39.01%	60.99%

This study aimed to evaluate the performance of various algorithms such as logistic regression, KNN, RF, XGBoost, and SVM by employing multiple performance metrics via confusion matrix.

Table 4: Proposed algorithms performance analysis

Algorithm	Accuracy (%)	Precision	Recall	F1-score
XGBoost	99.5	1.00	1.00	1.00
Random Forest	74.26	0.68	0.93	0.78
SVM	70.47	0.70	0.91	0.79
KNN	67.57	0.70	0.83	0.76
Logistic Regression	63.11	0.63	0.98	0.77

Based on the findings presented in Table 4, it was evident that the XGBoost algorithm outperformed the other algorithms, demonstrating remarkable results. It achieved an accuracy rate of 99%, precision as 0.99, and F1-score of 1.00, suggesting flawless classification performance without any errors.

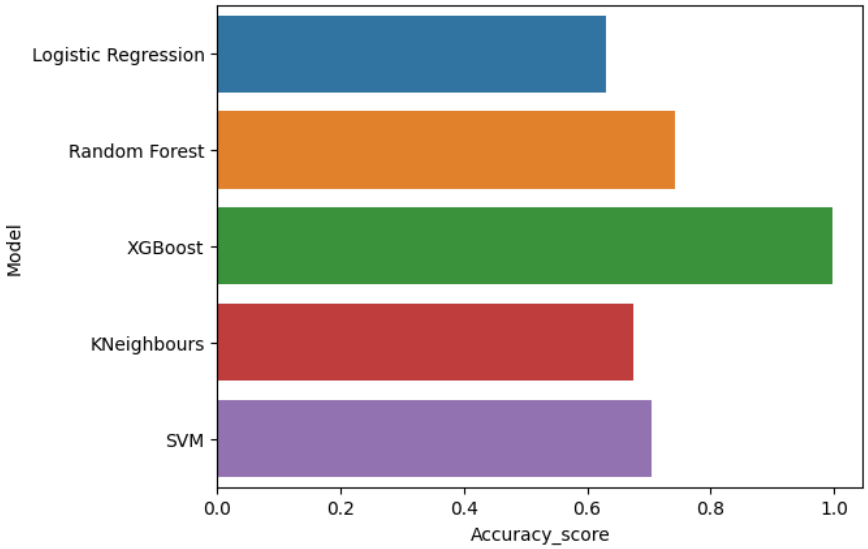

Fig 6: Accuracy comparison among the models

The research presented the accuracy scores of five machine learning algorithms, visualized in the bar graph shown in Figure 6. Based on the graph, it can be inferred that XGBoost achieved the highest accuracy rate, followed by RF, SVM, and KNN. In contrast, the Logistic Regression model demonstrated the lowest accuracy rate among the algorithms.

Table 5: Our approach with other state-of-art approaches

Study	Best Model	Accuracy (%)
Our study	XGBoost	99.57

Pal, O.K. [3]	ANN	98.12
Kurra et al. [8]	KNN	61.70
Patel, J. et al. [4]	Random Forest	81.00

Upon comparing our proposed model with previous studies conducted on the same Drinking Water dataset, it has been observed that our XGBoost model surpasses all other advanced techniques by achieving an accuracy score of 99.57%. As demonstrated in Table 5, our results exhibit significant enhancements in performance.

Following the evaluation of all the models, we have deployed the selected model using Streamlit to develop a web application with identical functionality. Users can access the web application and choose a specific algorithm from the provided options. They can input values for all nine parameters and the application will subsequently predict the potability of water, indicating its suitable for human.

4. CONCLUSION

Ensuring the safety and purity of drinking water is of utmost importance to safeguard human health. Accurate prediction of water potability plays a crucial role in achieving this objective. Access to clean drinking water is a fundamental right for every individual, as it is vital for maintaining overall well-being and preventing waterborne diseases. However, the escalating global population and increasing pollution levels have raised significant concerns about the quality of water sources. Leveraging the power of machine learning techniques can greatly contribute to predicting water potability and implementing necessary measures to enhance water quality, thus ensuring the availability of safe drinking water for the population.

A recent research paper extensively explored various machine learning algorithms and their effectiveness in predicting water potability based on an extensive range of physicochemical factors. This research depicts the potential of all implemeted algorithms as valuable tools for monitoring and managing water quality, which has profound implications for both the water sector and public health.

However, it is important to acknowledge certain limitations of the study. The dataset utilized in the research is relatively small, consisting of only 3276 observations. Consequently, it might be challenging to generalize the findings to larger populations. Additionally, the research focused on a limited set of water quality parameters, and it is advisable for future investigation to consider other relevant factors that could

influence the potability of water.

Author's Contribution:

Samir Patel has done conceptualization, methodology development, and data analysis. Samir played a key role in designing the research framework, formulating the research questions, and developing the methodology. Khushi Shah, Sakshi Vaghela, and Mohmmadali Aglodiya have done data collection, experimental design, and software implementation. They have designed the experiments, implemented the necessary software tools, and carried out data preprocessing and implementation stuff. Rashmi Bhattad has done a literature review, writing, and editing. Rashmi conducted project supervision and review along with Samir Patel.

Conflict of Interest: The authors declare no conflict of interest.

References:

[1] WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene, (2019). Progress on drinking water, sanitation and hygiene: 2017 update and SDG baselines.

World Health Organization (WHO) and the United Nations Children's Fund (UNICEF): https://www.who.int/publications/i/item/9789241512893

[2] https://www.kaggle.com/datasets/balavashan/drinking-water-dataset

[3] "Drinking-water." World Health Organization (WHO), 21 March 2022, https://www.who.int/news-room/fact-sheets/detail/drinking-water

[4] Cabral, João PS. "Water Microbiology. Bacterial Pathogens and Water – PMC." NCBI, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996186/.

[5] Patel, J., Amipara, C., Ahanger, T.A., Ladhva, K., Gupta, R.K., Alsaab, H.O., Althobaiti, Y.S. and Ratna, R., "A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI", Computational Intelligence and Neuroscience: CIN, 2022.

[6] Pal, O.K., "The Quality of Drinkable Water using Machine Learning Techniques", Int. J. Adv. Eng. Res. Sci., 8, p.5. 2021.

[7] Uddin, M.G., Nash, S., Rahman, A. and Olbert, A.I., "Performance analysis of the water quality index model for predicting water state using machine learning techniques", Process Safety and Environmental Protection, 169, pp.808-828, 2023.

[8] Aldhyani, T.H., Al-Yaari, M., Alkahtani, H. and Maashi, M., "Water quality prediction using artificial intelligence algorithms", Applied Bionics and Biomechanics, 2020

[9] Addisie, M.B., "Evaluating Drinking Water Quality Using Water Quality Parameters and Esthetic Attributes", Air, Soil and Water Research, 15, p.11786221221075005, 2022.

[10] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R. and García-Nieto, J., "Efficient water quality prediction using supervised machine learning", Water, 11(11), p.2210, 2019.

[11] Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J. and Zhang, Y., "Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data", Water research, 171, p.115454, 2020.

[12] Kouadri, S., Elbeltagi, A., Islam, A.R.M.T. and Kateb, S., "Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region", (Algerian southeast). Applied Water Science, 11(12), p.190, 2021.

[13] Kurra, S.S., Naidu, S.G., Chowdala, S., Yellanki, S.C. and Sunanda, D.B.E., "Water Quality Prediction Using Machine Learning", International Research Journal of Modernization in Engineering Technology and Science, India, 2022.

[14] Sinha, K.K., Gupta, M.K., Banerjee, M.K., Meraj, G., Singh, S.K., Kanga, S., Farooq, M., Kumar, P. and Sahu, N., "Neural network-based modeling of water quality in Jodhpur", India. Hydrology, 9(5), p.92. 2022

[15] Khambete, Aarti Kelkar., "When water kills." India Water Portal, 9 January 2019, https://www.indiawaterportal.org/faqs/when-water-kills.

[16] Chen, T., & Guestrin, C. "XGBoost: A Scalable Tree Boosting System". In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM. 2016

[17] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction (2nd ed.). Springer.

[18] Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). MIT Press.