Predicting Water Potability: Leveraging Machine Learning Techniques

1st Laya N
Department of Computer Science and Engineering
Nitte Meenakshi Institute of Technology
Bengaluru, India
nlaya681@gmail.com

2nd Shruthi Shetty J
Department of Computer Science and Engineering
Nitte Meenakshi Institute of Technology
Bengaluru, India
shruthi.shetty@nmit.ac.in

2024 Second International Conference on Data Science and Information System (ICDSIS) | 979-8-3503-8166-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDSIS61070.2024.10594228

***Abstract*—In recent times, the spotlight has been on understanding and forecasting water quality, owing to the variety of pollutants that pose potential harm. This research aims to advance strategies for managing and minimizing water pollution risks by examining the application of machine learning models for water quality classification. The algorithms under scrutiny include Logistic Regression, K-Nearest Neighbors Classifier, Support Vector Machine, Decision Tree, and Random Forest Classifier. These models underwent testing on a wate_potability dataset with 10 features, and their performance was assessed using various accuracy metrics, precision, fl_score and recall, where parameter tuning is also applied lastly for increasing better performace. The findings indicate that the proposed models effectively categorize water quality, with the Random Forest Classifier emerging as the most accurate in making predictions with 80% before parameter tuning and with 81% after parameter tuning. The .pkl pickle library has been employed to save the random forest classifier, ensuring its accessibility and usefulness in the future. Through this process, the ultimate model has been stored, facilitating its application in forecasting outcomes on unseen datasets in upcoming scenarios.This research where the prediction is made on the dataset to categorize the water as potable or non-potable.**

Keywords—water quality, machine learning, accuracy metrics, precision, recall, f1-score, water potable, water non-potable, pickle library.

I. INTRODUCTION

Covering 71% of the Earth's surface, water stands as a crucial natural resource vital for the survival of all living organisms. Its importance transcends basic consumption, playing a central role in diverse sectors like industries, agriculture, and global trade facilitated by oceans and seas. With an understanding of the essential role water plays in human existence, research efforts have been directed towards preserving water quality and mitigating pollution in accordance with strict international standards. [1].

A variety of water reservoirs, including groundwater, springs, rivers, lakes, and streams, are subject to particular quality standards that match their intended uses, whether for agriculture, industry, or human consumption.

Securing water quality necessitates a thorough grasp of potential pollution risks. The objective of this research initiative is to confront this challenge by examining the application of machine learning techniques in classifying water quality. By utilizing machine learning algorithms, we can examine and classify water quality based on a diverse set of parameters and attributes.

Water quality, crucial for human welfare, can be substantially and alarmingly influenced by human activities and natural processes [2] [3] [4]. The improper disposal of waste and pollutants resulting from these practices poses significant threats to aquatic ecosystems and human health. Notably, industrial plants and vehicles emerge as major contributors to water pollution, causing effects on both surface water and groundwater.

Industrial practices have the potential to discharge a range of pollutants into water bodies, changing their chemical composition and overall quality. This involves the release of harmful substances that contribute to the acidification of water sources, resulting in decreased pH levels, reduced acid-neutralizing capacity, and increased concentrations of aluminum [2]. The acidification not only impacts the accessibility of clean water but also has detrimental effects on aquatic organisms and their habitats, disrupting the ecological balance and potentially causing the decline of sensitive species.

An array of features, including pH measurements, hardness derived from calcium and magnesium salts, total dissolved solids (TDS), chloramines, sulfate concentrations, electrical conductivity (EC), total organic carbon (TOC), turbidity, and trihalomethanes, are pertinent for evaluating water quality. Concerning the prediction of water quality classification (WQC), machine learning algorithms are crucial for tasks such as data preprocessing, handling missing data, managing feature correlations, utilizing classification methods, and assessing the significance of feature selection[5].

In the process of exploration, five machine learning algorithms, which encompass K-Nearest Neighbors (KNN), Support Vector Machine, Logistic Regression, Decision Tree, and Random Forest Classifier, were applied for the purpose of water quality classification.The dataset utilized for analysis included 10 features relevant to water quality. Various accuracy measures were employed to assess performance of algorithms .

The study's findings illustrated the effective categorization of water quality through the implementation of the proposed machine learning models. Notably, the Random Forest Classifier emerged with the highest accuracy among the tested algorithms, indicating the potential of leveraging machine learning methods for accurate water quality classification.

Moreover, the research plays a role in advancing swifter and more cost-effective approaches for identifying water pollution. The conventional methods involving laboratory procedures and statistical analyses for appraising water quality can prove to be time-consuming and financially burdensome [6]. Leveraging machine learning algorithms facilitates the streamlining of this process, offering efficient solutions for the monitoring and adherence to water quality standards.

Within the document, section two discusses previous works in the field, while section three outlines the methodology employed in this study.The experimental

configuration and results is detailed in section four, succeeded by the exposition of conclusion in section five, where the findings are summarized and their implications are discussed.

II. Related Works

Water quality monitoring, amidst pollution challenges, embraces machine learning innovations. This study, leveraging the PyCaret platform, utilizes quadratic discriminant analysis to anticipate potability based on essential parameters. It explores eight algorithms, identifying the most accurate model.Stressing precise water quality forecasting, machine learning achieves remarkable precision [1-4]. Seven methodologies predict water quality attributes, addressing scarcity and contamination risks .Diverse techniques offer comprehensive evaluation in water resource management.Introducing an innovative approach enhances accurate forecasting.SVM, MLP, and LSTM excel in various aspects. Machine learning's potential in water quality management is underscored, emphasizing its significance [5-8]. Potability assessment is crucial for survival.This study tackles null values, achieving notable accuracies. Ensemble models demonstrate superiority in comparisons .Introducing a high-accuracy model highlights advancements .In water-related research, machine learning showcases versatility [9-14].

TABLE I. Comparative Analysis of Existing Literature

Sl. No.	Authors	Paper Title	Methodology/Proposed System
1.	Vaibhav Singh, Navpreet Kaur Wallia, Animesh Kudake, Aniket Raj	Water Potability Prediction Model Based on Machine Learning Techniques	Artificial Intelligence, Machine Learning, Water Quality, Quadratic Discriminant Analysis.
2.	Priyanshu Rawat, Madhvan Bajaj, Vikrant Sharma, Satvik Vats	A Comprehensive Analysis of the Effectiveness of Machine Learning Algorithms for Predicting Water Quality	Machine Learning, Water Quality, KNN, Decision Tree, Extreme Gradient Boost, Naïve Bayes, AdaBoost.
3.	S. R. SANNASI CHAKRAVARTHY, N. BHARANIDHARAN, VINOTH KUMAR VENKATESAN et al	Prediction of Water Quality Using SoftMax-ELM Optimized Using Adaptive Crow-Search Algorithm	Crow search algorithm, extreme learning machine, neural network, optimization, water quality
4.	Neenu Anil, Anu Ram, M Soumya Krishnan	WATER QUALITY ANALYSIS OF CANALS USING MACHINE LEARNING ALGORITHMS AND HYPERPARAMETER TURNING	NWMP, WHO, Hyperparameter Tuning, AdaBoost Classifier, Bagging Classifier, Gradient Boosting Classifier, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, and KNeighbors Classifier.
5.	Prachi Patel, Anaida Lewis, Bhumika Mange, Harsh Mangukiya, Narendra Shekokar	Water Quality Prediction and Estimate Percentage of Water Generated from a Device to Convert Atmospheric Moisture into Water	water quality, potability, water, AQUOMIST dataset, water generation, percentage of water, atmospheric moisture, Logistic Regression, Decision Tree, Random Forest, XGBoost, K-Neighbours, SVM, AdaBoost
6.	Michelle C. Tanega, Arnel Fajardo, Jomel S. Limbago	Analysis of Water Quality for Taal Lake Using Machine Learning Classification Algorithm"	Machine Learning, Classification Algorithm, WQI, Water Quality Classification, Taal Lake
7.	Xianhe Wang, Ying Li , Qian Qiao , Adriano Tavares,Yanchun Liang	Water Quality Prediction Based on Machine Learning and Comprehensive Weighting Methods	Combines entropy weighting and Pearson correlation. ML Models: SVM, MLP, LSTM.
8.	Adil Masood , Majid Niazkar , Mohammad Zakwan , Reza Piraei	A Machine Learning-Based Framework for Water Quality Index Estimation in the Southern Bug River	Introduces a machine learning-based framework. ML Models: Gaussian Process, XGBoost.
9.	Saqib Alam Ansari, Chetan Sharma, Trapty Agarwal.	Mean and Prediction Imputation-Based Approach for Predicting Water Potability Using Machine Learning	Water Portability, Machine Learning, Mean Imputation, Prediction Imputation, ANN
10.	Elisuba Kuruvilla, Subrahmanya Kundapura	Performance Comparison of Machine Learning Algorithms in Groundwater Potability Prediction	Machine Learning, Ensemble model, Groundwater quality, Potability.
11.	Md. Saikat Islam Khan, Nazrul Islam, Jia Uddin, Sifatul Islam , Mostofa Kamal Nasir	Water quality prediction and classification based on principal component regression and gradient boosting classifier approach	Utilizes principal component regression and gradient boosting classifier. Gradient Boosting Classifier model is used
12.	Mengyuan Zhu, Jiawei Wang, Xiao Yang, Yu Zhang, Linyu Zhang, Hongqiang Ren, Bing Wu , Lin Ye	A review of the application of machine learning in water quality evaluation	Provides a review of machine learning applications.
13.	Jinal Patel,Charmi Amipara et al	A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI	Utilizes Synthetic Minority Oversampling Technique and Explainable AI.
14.	Sai Sreeja Kurra, Sambangi Geethika Naidu et al	WATER QUALITY PREDICTION USING MACHINE LEARNING	AVV,SVM,Dez Catchment,GMDH,Tireh river
15.	Nishant Rawat, Mangani Daudi Kazembe et al	Water Quality Prediction using Machine Learning	AI,MLP,SVM
16.	Osim Kumar Pal	The Quality of Drinkable Water using Machine Learning Techniques	Random forest,ANN,SVM,DNN Gaussian Naive Bayes
17.	Prof. A. J. Kadam , Alex Sunny et al	Water Quality Monitoring Model using Machine Learning	RF,Xgboost,Gradient boosting,Adaboost,KNN,DT,SVR,MLP
18.	Neha Radhakrishnan, Anju S Pillai	Comparison of Water Quality Classification Models using Machine Learning	classification model; decision tree; support vector machine; naïve bayes; water quality index
19.	Fitore Muharemia, Doina Logofătuaand Florin Leon,	Machine learning approaches for anomaly detection of water quality on a real-world data set	SVM,RNN,ANN,DNN,LSTM
20.	Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi,Abbas Parsaie	Water Quality Prediction Using Machine Learning Methods	ann, dez catchment, GMDH, SVM, tireh river

Freshwater quality evaluation is crucial for agriculture and industry, ensuring safe supply .This study employs modern machine learning to predict drinking water quality efficiently.It integrates IoT and Machine Learning for classification [15-17]. Parameters such as pH, turbidity, and DO are scrutinized .Decision Tree achieves high accuracy at 98.50% .Anomalies in water quality time series data are explored [18-20].

The above table I depicts the Comparative Analysis Of Existing Literature where the literature review showcases a diverse range of approaches to predicting and monitoring water quality using machine learning techniques. From traditional algorithms like SVM and decision trees to advanced methods like deep learning and fusion with satellite imagery, researchers are exploring various avenues to improve accuracy and efficiency in water quality assessment.Advantages of these approaches include enhanced accuracy, real-time monitoring capabilities, and the ability to analyze large datasets for comprehensive insights into water quality parameters. These advancements hold promise for effective water management, pollution control, and safeguarding public health.However, there are notable limitations to consider. These include the need for extensive training data, computational complexity, implementation costs, and challenges related to interpreting results, especially in remote or resource-limited areas. Additionally, issues such as dataset imbalance and sensor reliability may impact the reliability of predictions.Overall, while machine learning shows great potential for addressing water quality challenges, further research is needed to overcome these limitations and ensure practical applicability across diverse environmental settings.

III. Proposed Method

Human activities and industrial processes contribute to water pollution, negatively impacting ecosystems and human health. This research advocates for faster, cost-effective water quality assessment using machine learning to ensure environmental sustainability while the existing system make use of traditional laboratory and statistical analyses for assessing water quality can be time-consuming and expensive.

This investigation aimed to apply machine learning methodologies for forecasting water potability, considering diverse attributes such as pH levels, mineral content, and contaminants. Multiple classification algorithms, including K-Nearest Neighbors, SVM, Logistic Regression, Decision Tree, and Random Forest, underwent systematic evaluation to ascertain the most efficient model.The methodology encompassed crucial steps like data collection, preprocessing, visualization, analysis, splitting, training, testing, tuning, and selecting the final model for making predictions on unseen data.

A. Model Architecture

The configuration of the model is fundamental in orchestrating and overseeing the collaboration and supervision of various subsystems within the project, ensuring their seamless operation as shown in figure 1.

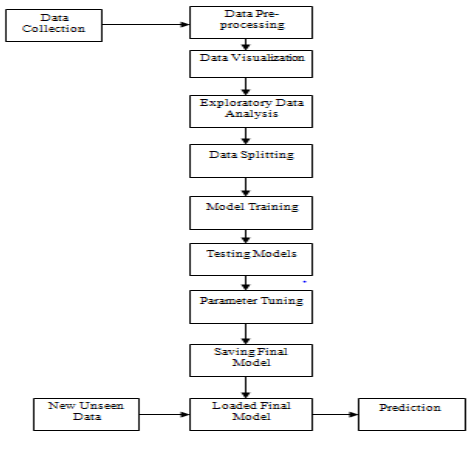

Fig. 1. Proposed Architecture to Predict Water Potability

1) Data Collection: Initial data collection involved obtaining a dataset water_potability_Dataset from kaggle, encompassing diverse water characteristics like pH, mineral content, turbidity, conductivity, organic carbon, hardness, and contaminant presence. Samples were labeled as potable or non-potable based on drinking water standards is shown in figure 2 as potable and non-potable.

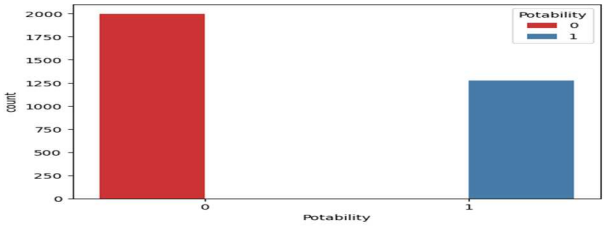

Fig. 2. Water potable v/s non-potable attribute samples.

2) Data Pre-processing: Data preprocessing, essential for machine learning, began with cleaning the dataset—addressing missing values, inconsistent formatting, and potential errors. Strategies like imputation or removal handled missing values, while outlier examination ensured model performance integrity.

3) Data Visualization: Following preprocessing, the data underwent visualization to deepen understanding. Plots and charts revealed patterns, trends, and potential issues, offering valuable insights for informed decision-making in subsequent analysis phases.

4) Exploratory Data Analysis: Rigorous scrutiny of the data was performed to augment understanding of its features and associations. Feature scaling was employed to ensure uniformity in variable scales, facilitating meaningful comparisons.

5) Data Splitting: In evaluating machine learning models, the data is divided into training and testing sets. The training set facilitated model fitting, enabling pattern recognition. The subsequent testing set assessed model performance on unseen data, with an 80-20 split ensuring comprehensive and reliable evaluation.

6) Model Training: After partitioning the data, machine learning models were trained with the assigned set. Five classifier algorithms, KNN, SVM, Logistic Regression,

Decision Tree, and Random Forest, were configured with appropriate parameters for effective learning.

7) Model Testing: Model accuracy was evaluated on testing set using metrics like accuracy, precision, recall, and F1 score, providing insights into their effectiveness.

8) Parameter Tuning: Fine-tuning models involved implementing grid search to systematically adjust parameters, exploring combinations to identify optimal configurations for peak performance through iteration.

9) Selecting The Final Model: Following training, testing, and parameter tuning, the final model selection considered performance metrics like accuracy, precision, recall, and F1 score. Chosen model demonstrated the most effective predictive capabilities for water potability.

10) Saving The Final Model: The final model was stored for future use using the pickle library, enabling serialization and deserialization of Python objects, including machine learning models. Saving in .pkl format ensures easy loading and utilization for future predictions or analysis.

11) Loading The Final Model: After selecting the ultimate model, it was saved for future reference, ensuring easy access to predict outcomes on new data. Loading the model allows utilizing acquired patterns and relationships for precise predictions in practical scenarios related to water potability.

12) Predictions: Subsequently, upon loading the ultimate model, predictions can be generated for unseen data. The dataset is checked with potability issues and predicted if water is potable or not as is it safe to drink.

B. Water Potability Dataset

Initial data collection involved obtaining a dataset water_potability_Dataset from kaggle The dataset attributes visualization is mentioned in below figure 3.The attributes explanations are described as:

1) pH Value: The acidity or alkalinity of water can be determined by its pH value, with the ideal range for drinking water typically falling between 6.5 and 8.5.

2) Hardness: Calcium and magnesium salts dissolved from geological deposits are the primary contributors to water hardness. The acceptable range for hardness typically falls between 120 and 170 mg/L.

3) Total Dissolved Solids (TDS): Total dissolved solids (TDS) indicate water's capability to dissolve inorganic as well as certain organic minerals or salts such as potassium, calcium, sodium, and bicarbonates. The advised limit for TDS in drinking water is 500 mg/l, with a maximum acceptable level of 1000 mg/l.

4) Chloramines: Chlorine and chloramine are frequently employed as disinfectants in public water distribution systems. The concentration of chlorine in drinking water is considered safe when it remains below 4 milligrams per liter (mg/L).

5) Sulfate: Existing naturally in minerals, soil, rocks, groundwater, plants, and dietary items, sulfates serve various industrial purposes. In freshwater sources, the concentration of sulfates typically varies between 3 and 30 mg/L.

6) Conductivity: The ability of water to conduct electrical current is measured through conductivity. The World Health Organization suggests that the electrical conductivity (EC) value should not go beyond 400 µS/cm.

7) Organic Carbon: Total Organic Carbon (TOC) found in water originates from the degradation of natural organic matter and synthetic sources. Keeping TOC levels in water below 0.05 ml/l is advised.

8) Trihalomethanes: Chemicals known as Trihalomethanes (THMs) can be present in water treated with chlorine. It is generally regarded as safe for drinking water to contain THM levels of up to 80 ppm.

9) Turbidity: Turbidity serves as a gauge for the concentration of suspended solids in water. The WHO recommends a maximum turbidity value of 5.00 NTU (Nephelometric Turbidity Units).

Total of 3,276 samples were collected and analyzed, making it a substantial dataset for conducting a robust analysis. The dataset consists of 3,276 water samples, each containing information on nine important parameters. The significance of these parameters in evaluating water quality and ensuring its safety for human consumption discussed. The dataset includes the above hydro-chemical parameters and portability labels which is shown through visualization in the below figure 3.

The below figure 4 depicts that if the features have any co-relation between them. To perform the analysis, the dataset was divided into training and testing sets. Approximately 75% of the samples (2,457) were allocated for training, while remaining 25% (819 samples) were used for testing the models. This division ensures the reliability and generalizability of the results obtained from analysis.

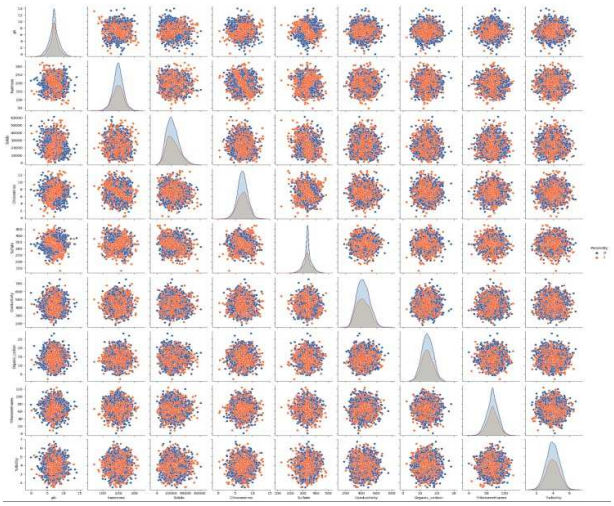

Fig. 3. Various dataset attributes visualization

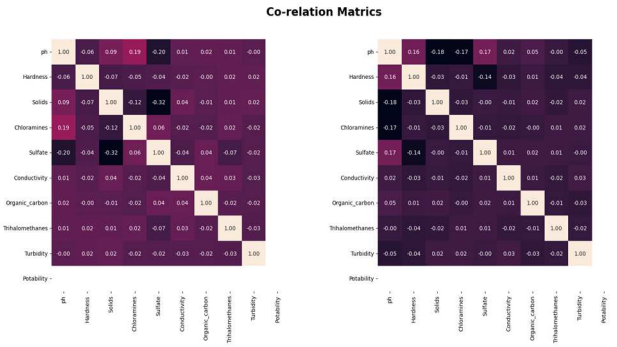

Fig. 4. Co-relation Matrics

In the preprocessing steps, missing values were replaced with the mean values calculated based on the potability category. For instance, the pH, Sulfate, and Trihalomethanes columns had missing values filled with the mean values of their respective potability categories. This ensures a more accurate representation of the data. After preprocessing, exploratory data analysis was conducted, including pair plots, KDE plots, and correlation matrices, to gain insights into the relationships between variables and their impact on water potability. Additionally, categorical features were visualized to understand their distribution among potability classes.

C. Machine Learning Classifiers

The machine learning classifiers used are:

1) Random forest classifier: Random Forest, when used as a classifier, employs multiple decision trees created from distinct portions of the dataset to enhance predictive accuracy by averaging their results. Unlike depending solely on a lone decision tree, this technique involves amalgamating predictions from each tree, and the final prediction is based on the predominant consensus among these individual forecasts.

Eq (1) mathematically represents the prediction of Random Forest as

$$y\text{^}RF = 1/N\sum Ni = 1Ti(x) \quad (1)$$

where y^RF is the predicted output, N is the number of decision trees, and Ti(x) is the prediction of the ithdecision tree.

2) Decision tree classifier: When utilizing a decision tree to predict the class of a given dataset, the algorithm begins its assessment at the root node. Here, it evaluates the values of the root attribute against the corresponding attributes in the dataset and proceeds along the appropriate branch based on this evaluation. At each subsequent node, the algorithm repeats this process of comparing attribute values with those in the dataset, advancing further along the tree's structure. This iterative process persists until the algorithm reaches the terminal leaf node of the tree..Eq(2) represents process of predicting the class.

a) Splitting Criteria: Q(Xm,tm)=Some criterion,Best Split: Xm≤tm and Xm>tm,Recursive Splitting: Repeat splitting on subsets,Leaf Node Prediction: Majority class label in leaf.

3) Model Representation:

$$DecisionTree(X) = \sum Li = 1class(X,i) \times leaf(X,i) \quad (2)$$

Where X - Input feature space,y - Target variable (class labels),N - Total number of samples,D - Set of features,L - Number of leaf nodes,m-node,t-treshold, class(X,i) is the class label associated with the ith leaf node.leaf(X,i) is an indicator function that returns 1 if X belongs to the ithleaf node, otherwise 0.

4) Logistic regression classifier: When addressing binary and linear classification tasks, logistic regression stands out as a straightforward and effective solution. As a classification model, it offers ease of implementation and commendable performance, particularly in scenarios where classes are linearly separable. Its wide adoption underscores its role as a prevalent algorithm for classification purposes. In comparison, support vector machine classifiers. Eq (3) mathematically represents the logistic function as:

$$P(Y = 1 \mid X) = 1/1 + e - \beta X \quad (3)$$

where ,P(Y=1|X) represents the probability of the positive class, X is the input features, and β is the vector of coefficients.

5) SVM classifier: Utilizing a support vector machine (SVM) classifier entails creating a linear boundary between two classes. This results in assigning one class to all data points on one side of the line and the other class to those on the opposite side. Consequently, there exists an unlimited number of potential lines to consider. Eq (4) represents,Binary classification,Decision based on linear separation with hyperplane.

$$y = sign(wTx + b) \quad (4)$$

Where y is the predicted class label,w is the weight vector representing the coefficients of the hyperplane,x is the feature vector of the data point,b is the bias term.

6) K-nearest neighbour classifier: K-nearest neighbors (KNN) represents a form of supervised learning algorithm utilized for regression and classification tasks alike. In KNN, the algorithm endeavors to determine the correct class for test data by evaluating the distance between the test data point and all training points. Following this, it identifies the K nearest points to the test data point for further analysis.Eq(5) represents ,majority voting Class determined by closest neighbor class labels.

$$y\text{^} = mode(yneighbor) \quad (5)$$

Where,y^is the predicted class label,yneighbor is the set of class labels of the nearest neighbors.

D. Objectives

- Machine learning models, including KNN, SVM, Logistic Regression, Decision Tree, and Random Forest, were trained on the dataset to learn patterns to build a training model.
- Model performance was assessed on the testing set, evaluating accuracy and various metrics such as precision, recall, F1-score to evaluate the best model among all the ML models. The best model is evaluated as final model for highest predictive capabilities for determining water potability.

IV. Results and Discussion

The evaluation of the models employed in this work is examined, highlighting their respective performance through detailed classification reports. Furthermore, the time taken for threat detection is deliberated. System prerequisites for this research involve 8GB of RAM, an i5 processor, and jupyter notebook with comprehensive machine learning libraries.

A. Evaluation Metrics

Several essential metrics were employed to evaluate the model's performance, each providing valuable insights into the effectiveness of the anomaly detection approach:

1) True Positives: When the model makes a correct prediction of the positive class, it results in a true positive.

2) False Positives: An incorrect prediction of the positive class characterizes a false positive outcome in the model.

3) True Negatives: A true negative outcome occurs when the model accurately predicts the negative class.

4) False Negatives: In cases of a false negative, the model makes an incorrect prediction of the negative class.

5) Precision: In evaluating precision, the focus lies on discerning the accuracy of predicted positive cases, calculated as the division of true positives by the sum of true positives and false positives. Eq (6) quantifies the proportion of true positive results among all instances predicted as positive.

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

6) Recall: Sensitivity, or recall, assesses the proportion of actual positive cases accurately identified by the model. This is determined by dividing the number of true positives by the sum of true positives and false negatives. Eq(7) computes the ratio of true positives to the sum of true positives and false negatives.

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

7) F1-Score: The F1 score, derived as the harmonic mean of precision and recall, integrates these metrics into a singular value, providing a holistic representation of the model's performance. Its calculated value, as per a specific formula, furnishes a balanced assessment of the model's accuracy, proving advantageous in instances of imbalanced data.F1_score in Eq(8) reperesents the model's performance in the balanced fasion.

$$Fl\text{-}Score = 2 \star \frac{Precision \star Recall}{Precision + Recall} \tag{8}$$

8) Confusion matrics: The confusion matrix summarizes a model's performance, displaying accurate and inaccurate predictions in classification tasks using categorical labels having TP,FP,TN,FN.

9) Accuracy: Accuracy serves as a metric, gauging the frequency of correct predictions in machine learning models, determined by dividing correct predictions by the total made as in Eq(9).

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{9}$$

10) Parameter Tuning: The iterative nature of hyperparameter tuning involves exploring different parameter combinations to optimize a selected metric, often focusing on enhancing accuracy.

B. Result Analysis

Employing Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest, the dataset was subjected to analysis using five distinct classifier algorithms. The performance assessment of each classifier algorithm was conducted based on precision, recall, F1 score, and accuracy. Detailed results for these classifiers on the test dataset are presented in the table II below.

TABLE II. TEST RESULTS FOR EACH CLASSIFIER

Classifier	Precision	Recall	F1-Score	Accuracy
KNN	0.3352	0.2459	0.2837	0.5381
Logistic Regression	0.0000	0.0000	0.0000	0.6280
Support Vector	0.0000	0.0000	0.0000	0.6280
Decision Tree	0.6364	0.6311	0.6337	0.7287
Random Forest	0.8187	0.6107	0.6995	0.8049

Notably, the Random Forest classifier secured the highest F1 score in the assessments, indicating the most favorable overall performance. Alongside it, the Decision Tree also showcased promising performance. Consequently, both algorithms were designated for additional parameter tuning to refine and optimize their effectiveness.

1) Parameter Tuning: Here, Grid search CV is utilized for parameter tuning.

2) Random Forest Classifier: GridSearchCV optimizes RandomForestClassifier's hyperparameters (n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features) using cross-validation. The best parameters results showing 'n_estimators.' as below.

a) Best Estimator: {'n_estimators': 400 , Best Score : 0.7992366412213741

3) Decision Tree Classifier: GridSearchCV is employed to find the optimal hyperparameters for a Decision Tree model. The hyperparameters include 'max_depth,' 'min_samples_split,' 'min_samples_leaf,' and 'max_features.' The grid search explores various combinations of these hyperparameters using a 5-fold cross-validation. The best combination is determined based on accuracy. The final Decision Tree model with the optimal hyperparameters is shown below.

a) Best Estimator: {'max_depth': 10, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 10}, Best Score: 0.7312977099236642

4) KNN Classifier: GridSearchCV is used to optimize hyperparameters (n_neighbors, weights, p) for KNeighborsClassifier using cross-validation. The best parameters and score result is shown below.

a) Best Estimator: {'n_neighbors': 9, 'p': 1, 'weights': 'uniform'}, Best Score: 0.5694656488549619

5) Logistic Regression Classifier: GridSearchCV optimizes LogisticRegression hyperparameters (C, penalty, solver) using cross-validation. The best parameters and score shown in result, including the logistic regression estimator.

a) Best Estimator: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}, Best Score: 0.6061068702290077

6) SVC Classifier: GridSearchCV optimizes Support Vector Classifier (SVC) hyperparameters (C, kernel, gamma,

degree) using cross-validation. The best parameters and score are printed, including the SVC estimator.

Upon fine-tuning the parameters, a reassessment of the performance for both the Decision Tree and Random Forest classifiers was performed since the both had highest accuracy and f1-score when compared to other cassifier, showcasing the outcomes presented in the table III.

TABLE III. TEST RESULTS AFTER PARAMETER TUNING

Classifier	Precision	Recall	F1-Score	Accuracy
Decision Tree	0.6364	0.6311	0.6337	0.7287
Random Forest	0.81	0.63	0.71	0.81

These results show that parameter tuning had a small but positive impact on the performance of the Decision Tree classifier, with no increase in F1 score and accuracy. The performance of the Random Forest classifier has increased in f1-score from 0.69 to 0.71 and increase in accuracy from 0.80 to 0.81. Based on these results, the Random Forest classifier had the highest F1 score and accuracy, indicating the best overall performance. This algorithm was therefore selected as the final model for predicting. The classification report of random forest classifier is shown below figure 5.

	precision	recall	f1-score	support
0	0.81	0.92	0.86	412
1	0.82	0.63	0.71	244
accuracy			0.81	656
macro avg	0.81	0.77	0.79	656
weighted avg	0.81	0.81	0.80	656

Fig. 5. classification report of random forest classifier

The above figure 6 shows that confusion matrix of random forest classifier for n_estimator=400 after parameter tuning for better accuracy.

Fig. 6. Confusion matrix of random forest classifier for n_estimator=400

Lastly, this random forest model is saved in .pkl for future use for predictions where water department can use it for predictions.

Now, the saved model is loaded for predictions to be done on unseen data, which is done as below

```
new_data = pd.DataFrame({
    'ph': [9.5],
    'Hardness': [180.8054024],
    'Solids':[14168.52916],
    'Chloramines':[9.555571086],
    'Sulfate':[377.5833739],
    'Conductivity':[599.659021],
    'Organic_carbon':[7.606396747],
    'Trihalomethanes':[87.57745951],
    'Turbidity':[7.875165247]
})
prediction = rf_classifier.predict(new_data)
# Make predictions on the new data
prediction = rf_classifier.predict(new_data)
# Print predictions for debugging
print("Predictions:", prediction)
# Display the prediction
if prediction[0] == 1:
    print("The water is predicted to be potable.")
else:
    print("The water is predicted not to be potable.")
```

The result is shown below for water potability where 1 binary classification indicates water is potable. This is for one set of attributes.

Predictions: [1]

The water is predicted to be potable.

The second set of attributes where water is non-potable is also evaluated where 0 binary classification indicates water is non-potable.

```
new_data = pd.DataFrame({
    'ph': [11.18028447],
    'Hardness': [227.2314692],
    'Solids':[25484.50849],
    'Chloramines':[9.077200017],
    'Sulfate':[404.0416347],
    'Conductivity':[563.8854815],
    'Organic_carbon':[17.92780641],
    'Trihalomethanes':[71.97660103],
    'Turbidity':[4.370561937],

})
prediction = rf_classifier.predict(new_data)
```

The result is shown below.

Predictions: [0]

The water is predicted not to be potable.

C. Discussion

The study's key findings reveal the Random Forest Classifier's effectiveness in accurately categorizing water quality, achieving an 80% accuracy rate before tuning and improving to 81% post-tuning. This underscores the algorithm's robustness in handling complex datasets.

TABLE IV. COMPARATIVE ANALYSIS OF THE PROPOSED METHOD WITH THE EXISTING LITERATURE

Work#Ref	Methods	Classifier	Accuracy (%)
Sai Sreeja Kurra [14]	Machine Learning	K-Nearest Neighbor	61.7
Nishant Rawat [15]	Machine Learning	Random Forest	67
Priyanshu Rawat [2]	Machine Learning	XG-Boost	69.89
Saqib Alam Ansari [9]	Machine Learning	Artificial Neural Network	70.09
Proposed Method	Machine Learning	Random Forest	81

The interpretation suggests that ensemble learning techniques, like the Random Forest, hold promise in mitigating water pollution risks and ensuring safe drinking water access. However, limitations include the need for extensive training data and challenges in computational complexity and interpretability. Acknowledging these constraints, recommendations entail further exploration of ensemble learning techniques and integration of additional

data sources to enhance model robustness. Comparative analysis shown in table IV indicates that while Random Forest outperforms other algorithms in accuracy, choice should consider application-specific requirements and interpretability concerns. Overall, the study's findings advocate for the adoption of machine learning-based approaches in water quality management, offering potential for more efficient and proactive strategies to safeguard public health and environmental sustainability.

V. Conclusion

Machine learning techniques, analyzing pH level, mineral content, and contaminant presence, predict water potability. Five classifiers were evaluated, with Decision Tree and Random Forest exhibiting superior performance in F1 scores. Further parameter tuning modestly enhanced Random Forest's accuracy, leading to its selection as the final model with an impressive F1 score of 0.71 and 0.81 accuracy. This underscores machine learning's potential in critical water quality assessment, emphasizing ongoing research for model enhancement and real-world application.

In summary, the project employed machine learning for accurate water quality classification, identifying Random Forest as the most effective model. The final model, saved for future use, ensures accessible and precise predictions for water potability, with opportunities for further improvement through algorithmic refinement.

References

[1] Vaibhav Singh, Navpreet Kaur Wallia, Animesh Kudake, Aniket Raj," Water Potability Prediction Model Based on Machine Learning Techniques" 2023 World Conference on Communication & Computing (WCONF),979-8-3503-1120-4, DOI: 10.1109/WCONF58270.2023.10235096.

[2] Priyanshu Rawat, Madhvan Bajaj, Vikrant Sharma, Satvik Vats," A Comprehensive Analysis of the Effectiveness of Machine Learning Algorithms for Predicting Water Quality", 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA) ,979-8-3503-9720-8, DOI: 10.1109/ICIDCA56705.2023.1009996.

[3] S. R. Sannasi Chakravarthy, N. Bharanidharan, Vinoth Kumar Venkatesan, Mohamed Abbas, Harikumar Rajaguru, T. R. Mahesh, Krishnamoorthy Venkatesan ," Prediction of Water Quality Using SoftMax-ELM Optimized Using Adaptive Crow-Search Algorithm",2023, 140900, VOLUME 11.

[4] Neenu Anil, Anu Ram, M Soumya Krishnan," Water Quality Analysis Of Canals Using Machine Learning Algorithms and Hyperparameter Turning", 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) ,979-8-3503-3509-5, DOI: 10.1109/ICCCNT56998.2023.1030793.

[5] Prachi Patel, Anaida Lewis, Bhumika Mange, Harsh Mangukiya, Narendra Shekokar, "Water Quality Prediction and Estimate Percentage of Water Generated from a Device to Convert Atmospheric Moisture into Water", 2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT),979-8-3503-0336-0, DOI: 10.1109/CISCT57197.2023.10351318.

[6] Michelle C. Tanega, Arnel Fajardo, Jomel S. Limbago," Analysis of Water Quality for Taal Lake Using Machine Learning Classification Algorithm", 2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE),979-8-3503-0050-5,DOI: 10.1109/JCSSE58229.2023.10202046.

[7] Xianhe Wang, Ying Li , Qian Qiao , Adriano Tavares,Yanchun Liang ,"Water Quality Prediction Based on Machine Learning and Comprehensive Weighting Methods", Entropy 2023, 25, 1186. https://doi.org/10.3390/ e25081186.

[8] Adil Masood , Majid Niazkar , Mohammad Zakwan , Reza Piraei,"A Machine Learning-Based Framework for Water Quality Index Estimation in the Southern Bug River", Water 2023, 15, 3543. https://doi.org/10.3390/ w15203543.

[9] Saqib Alam Ansari, Chetan Sharma, Trapty Agarwal," Mean and Prediction Imputation-Based Approach for Predicting Water Potability Using Machine Learning", 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 978-1-6654-7433-7,DOI: 10.1109/ICRITO56286.2022.9964809.

[10] Elisuba Kuruvilla, Subrahmanya Kundapura," Performance Comparison of Machine Learning Algorithms in Groundwater Potability Prediction", 2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE),978-1-6654-8910-2, DOI: 10.1109/ICRAIE56454.2022.1005429.

[11] Md. Saikat Islam Khan, Nazrul Islam, Jia Uddin, Sifatul Islam, Mostofa Kamal Nasir," Water quality prediction and classification based on principal component regression and gradient boosting classifier approach", Journal of King Saud University – Computer and Information Sciences 34 (2022) 4773–4781, https://doi.org/10.1016/j.jksuci.2021.06.003 1319-1578.

[12] Mengyuan Zhu, Jiawei Wang, Xiao Yang, Yu Zhang, Linyu Zhang, Hongqiang Ren, Bing Wu, Lin Ye ," A review of the application of machine learning in water quality evaluation"https://www.sciencedirect.com/journal/eco-environment-and-health, volume 1,issue1,June 2022, Pages 107-116.

[13] Jinal Patel, Charmi Amipara,Tariq Ahamed Ahanger , Komal Ladhva,Rajeev Kumar Gupta,Hashem O. Alsaab , Yusuf S. Althobaiti , Rajnish Ratna," A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI", Hindawi Computational Intelligence and Neuroscience Volume 2022, Article ID 9283293, 15 pages https://doi.org/10.1155/2022/9283293.

[14] Sai Sreeja Kurra, Sambangi Geethika Naidu, Sravani Chowdala, Sree, Chithra Yellanki, Dr. B. Esther Sunanda," Water Quality Prediction Using Machine Learning", International Research Journal of Modernization in Engineering Technology and Science (Peer-Reviewed, Open Access, Fully Refereed International Journal) Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com, e-ISSN: 2582-5208.

[15] Nishant Rawat, Mangani Daudi Kazembe, Pradeep Kumar Mishra," Water Quality Prediction using Machine Learning", International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue VI June 2022- Available at www.ijraset.com.

[16] Osim Kumar Pal,"The Quality of Drinkable Water using Machine Learning Techniques", International Journal of Advanced Engineering Research and Science (IJAERS) Peer-Reviewed Journal ISSN: 2349-6495(P) | 2456-1908(O) Vol-9, Issue-6; Jun, 2022 Journal Home Page Available: https://ijaers.com/ Article DOI: https://dx.doi.org/10.22161/ijaers.96.2.

[17] Prof. A. J. Kadam, Alex Sunny, Mitali Admuthe , Atharva Bhosale, Aashay Bhujbal," Water Quality Monitoring Model using Machine Learning",ISSN (Online) 2581-9429 International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) , ISSN (Online) 2581-9429 Volume 2, Issue 1, December 2022.

[18] Neha Radhakrishnan, Anju S Pillai, "Comparison of Water Quality Classification Models using Machine Learning", Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020) IEEE Conference Record # 48766; IEEE Xplore ISBN: 978-1-7281-5371-1.

[19] Fitore Muharemia, Doina Logofătuaand Florin Leon, "Machine learning approaches for anomaly detection of waterquality on a real-world data set", JOURNAL OF INFORMATION AND TELECOMMUNICATION2019, VOL.3, NO.3,294307https://doi.org/10.1080/24751839.2019.1565653.

[20] Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi,Abbas Parsaie," Water Quality Prediction Using Machine Learning Methods",Water Quality Research Journal(2018) 53 (1): 3–13.https://doi.org/10.2166/wqrj.2018.025,vol 53.