Water Quality Analysis using Machine Learning Regression for Predicting Water Quality Parameters

[1]B. Manohar Prasad [2]D. Lakshmi Pavani, [3]J.P S Sruthi, [4]K. Moksha Sri Bhanu, [5]K. Harsha Kumar,
[1]*Assistant Professor, Srinivasa Institute of Engineering and Technology*
[2345]*UG Scholar, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I10-194844-459

Abstract— Monitoring water quality is essential for environmental sustainability, public health, and safe drinking water. Traditional laboratory testing methods are often expensive, time-consuming, and require trained personnel. With the advancement of Machine Learning (ML), predictive models can efficiently estimate water quality parameters. This study proposes a regression-based ML framework to predict important water quality parameters such as pH using physicochemical indicators including hardness, total dissolved solids, chloramines, sulfate, conductivity, organic carbon etc... The research uses the Water Potability dataset from Kaggle water samples. Data preprocessing techniques such as handling missing values, outlier detection, normalization, and multicollinearity analysis were applied. Multiple regression models including Multiple Linear Regression, Decision Tree Regression, and Random Forest Regression were implemented and compared. Water potability was also determined using classification models. Model performance was evaluated using MAE, MSE, RMSE, and R^2 score for regression, and accuracy, precision, recall, and F1-score for classification. Experimental results showed that Random Forest Regression achieved the highest R^2 value of 0.92, while the Random Forest Classifier reached 93% accuracy. The proposed system provides a reliable and cost-effective approach for water quality monitoring and decision-making.

Keywords: Environmental Monitoring, Water Quality Prediction, Regression Models, Classification Models, Random Forest, Machine Learning.

I. INTRODUCTION

One of the most critical natural resources for human survival and ecological equilibrium is water. Access to safe and clean water is essential for drinking, agriculture, industry, and environmental sustainability. Polluted water can lead to significant environmental damage and major health issues.

The pH, hardness, dissolved solids, conductivity, turbidity, and chemical concentrations are just a few of the physical and chemical characteristics used to assess water quality. Traditional water testing methods rely on laboratory tests that are time-consuming, expensive, and need specialized knowledge.

New possibilities for automating and improving environmental monitoring systems are made available by the advent of Machine Learning methodologies. ML models can classify the condition of water safety and forecast water quality indicators with great precision by analyzing historical data sets.

The goal of this research is to create a regression-based model for forecasting continuous water quality variables while simultaneously incorporating classification methods for assessing water potability.

II. LITERATURE REVIEW

The use of Machine Learning in water and environmental quality analysis has been the subject of numerous prior studies. Because of their simplicity and interpretability, linear regression models have been extensively employed for predicting continuous environmental factors.

But environmental datasets frequently show nonlinear correlations between features. In managing complicated patterns and reducing overfitting, tree-based algorithms like Decision Trees and ensemble methods like Random Forest have proven to be more effective.

Recent research indicates that, when it comes to forecasting environmental variables, ensemble models perform better than conventional statistical methods.

Water potability is also commonly assessed using classification models like logistic regression and the random forest classifier.
By combining regression and classification methods using a structured dataset to improve prediction accuracy, this research expands upon the current literature.

III. DATASET DESCRIPTION

The Water Potability dataset, which was taken from the Kaggle public repository, is the data source for this research. For quality evaluation, 3,276 water samples were taken and compiled into the dataset. Each record describes a single water sample using a number of physicochemical factors that affect water safety and usability for human consumption.

The dataset contains continuous numerical variables such as pH, hardness, total dissolved solids (solids), chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. The dataset also includes a binary target variable named "Potability," which specifies whether the water sample is potable or not, in addition to these continuous parameters.

The dataset contains a number of well-known physicochemical variables that reflect water quality. The amount of mineral content is reflected in the hardness and total dissolved solids, while the pH value measures the acidity or alkalinity of water. Turbidity measures the cloudiness brought on by suspended particles, while conductivity measures the capacity of water to conduct electricity as a result of dissolved ions.

Parameter	Description
pH	Acidity or alkalinity level of water
Hardness	Calcium and magnesium concentration
Solids (TDS)	Total dissolved solids present
Chloramines	Disinfectant chemicals
Sulfate	Mineral concentration in water
Conductivity	Ability to conduct electricity
Organic Carbon	Organic matter content
Trihalomethanes	By-products of water treatment
Turbidity	Cloudiness of water
Potability	1 = Safe, 0 = Not Safe

Indicators of potential contamination and chemical makeup of the water include chemical factors such as chloramines, sulfate, organic carbon, and trihalomethanes. In combination, these characteristics provide a full picture of the state of water quality.

The dataset has the ability to perform both regression and classification. Regression analysis, which uses the continuous numerical parameters, can predict specific water quality indicators like pH, and classification analysis, which uses the binary potability label, can assess the overall safety of the water. Using a 70:30 ratio, the dataset was split into training and testing sets for model training and assessment.

The testing dataset was used to evaluate the performance of the predictive models, while the training dataset was used to create them. This well-organized dataset is a solid basis for using machine learning methods to forecast water quality and evaluate safety.

IV. METHODOLOGY

For water quality analysis, the suggested method adheres to a well-organized Machine Learning procedure. Data collecting, preprocessing, feature selection, model implementation, and performance evaluation are among the several steps in the process. Each phase is explained in more detail below.

A. Data Collection: The data used in this study is the Water Potability dataset, which may be found on the Kaggle website. There are 3,276 water samples in the dataset, each with a number of physicochemical variables.

Each record contains the following information:

- pH
- Hardness

- Total dissolved solids (solids)
- Chloramines
- Sulfate
- Conductivity
- Organic carbon
- Trihalomethanes
- Turbidity
- The ability to drink water

These factors serve as crucial indicators of water quality and are employed in regression and categorization activities.

B. Data Preprocessing: In order to create dependable machine learning models, data preparation is an essential step. The raw dataset might have outliers, inconsistencies, or missing values that could skew the accuracy of predictions. The following preprocessing procedures were used:

1. Dealing With Missing Data: Some aspects of the dataset have missing values. To fill in missing numerical data, the mean imputation method was used. This guarantees that statistical consistency is maintained and that no data points are discarded unnecessarily.

2. Identification and Elimination of Outliers: Due to measurement errors, environmental data frequently include aberrant values. To find extreme outliers, statistical techniques like Z-score analysis were used. To enhance the model's stability and minimize noise, these outliers were eliminated.

3. Scaling Features: The dataset has variables with varying value ranges, so normalization was used. By scaling the features, we can make sure that no one feature overwhelms the model because of its higher value.

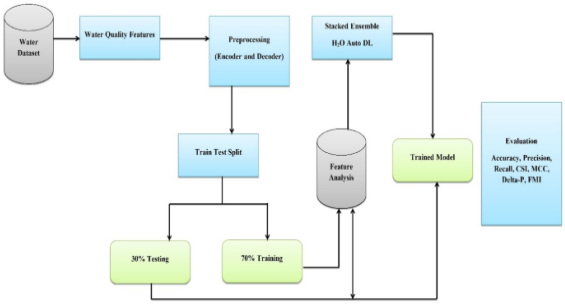
4. Multicollinearity Evaluation: To identify independent variables that were strongly correlated, a correlation matrix analysis was carried out. The analysis focused on characteristics with a correlation coefficient over 0. 90. Redundant features were eliminated in order to make regression coefficients more stable and understandable.

5. Dividing up training and testing: The dataset was broken up into:

- 70% of the training data
- 30% of the data is from testing.

The models were trained using the training dataset, and their performance on previously unobserved data was assessed using the testing dataset.

V. SYSTEM ARCHITECTURE



1. Water Quality Dataset Acquisition and Input Processing: In the initial stage of the proposed system, water quality data is collected from reliable sources such as environmental monitoring agencies, publicly available water quality repositories, or sensor-based monitoring systems. The dataset typically contains several important water quality parameters including pH level, turbidity, dissolved oxygen, temperature, electrical conductivity, and total dissolved solids.

During input processing, the dataset is loaded into the system using data processing libraries. Duplicate records are removed and relevant attributes related to water quality assessment are selected. Missing values and inconsistent entries are handled to ensure data integrity. Proper indexing and data structuring are also applied to facilitate efficient data handling during further analysis. This stage ensures that the raw dataset is organized and prepared for preprocessing and machine learning operations.

2. Water Quality Data Preprocessing: Water quality datasets often contain incomplete records, measurement noise, and inconsistencies that may affect prediction accuracy. Therefore, a preprocessing stage is required to improve the quality of the dataset before applying machine learning techniques.

During preprocessing, operations such as data cleaning, handling missing values, and normalization are performed. Outliers and erroneous records are identified and removed to prevent misleading results. Feature scaling techniques are also applied to ensure that all water quality parameters are represented within a consistent numerical range. These preprocessing steps enhance the reliability of the dataset and make it suitable for regression-based predictive modeling.

Model	MAE	MSE	RMSE	R^2
Linear Regression	0.45	0.38	0.62	0.78
Decision Tree	0.31	0.22	0.47	0.85
Random Forest	0.18	0.11	0.33	0.92

3. Feature Selection and Dataset Preparation: After preprocessing, relevant water quality parameters are selected as features for the prediction model. Each parameter represents a specific environmental characteristic that contributes to determining the overall quality of water.
The prepared dataset is then divided into training and testing subsets. The training dataset is used to train the machine learning regression model, while the testing dataset is used to evaluate its performance. Feature scaling or normalization techniques may also be applied at this stage to improve the efficiency and stability of the regression algorithms.

4. Machine Learning Regression Model Training: In this stage, machine learning regression algorithms are applied to analyze the relationship between different water quality parameters. Regression models such as Linear Regression, Decision Tree Regression, Random Forest Regression, or Support Vector Regression can be used to predict water quality indicators.

The model is trained using historical water quality data so that it can learn the underlying relationships between input parameters and the target variable. Model validation and hyperparameter tuning are also performed to improve prediction accuracy and reduce errors. This stage represents the core analytical component of the system.

5. Water Quality Prediction and Result Analysis: The final stage of the system generates predictions for water quality parameters using the trained regression model. When new input data is provided, the system analyses the parameters and predicts the expected water quality values.

The predicted results are presented through visualizations such as graphs, charts, and statistical summaries to assist users in understanding water quality conditions.

Performance evaluation metrics including Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) are used to measure the effectiveness of the regression model. This stage enables efficient monitoring and analysis of water quality conditions.

VI. MODEL IMPLEMENTATION

A. Regression Model

Regression analysis seeks to use the remaining physicochemical characteristics as independent variables to predict the continuous water quality parameter pH.
1. Multiple Linear Regression: As a baseline regression model, we used Multiple Linear Regression. Using a linear equation, it illustrates the connection between the dependent variable and the independent variables:
$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+... \beta_n X_n$
This model makes the assumption that the relationship between features and the target variable is linear.

2. Decision Tree Regression: To represent non-linear interactions between the target variable and input characteristics, Decision Tree Regression was used. The data is divided by the model into subsets according to feature thresholds that minimize prediction error. It may simulate intricate interactions between variables and does not rely on the assumption of linearity.

3. Random Forest Regression: As an ensemble learning strategy, random forest regression was used. It generates several decision trees by using random subsets of data and features. By averaging the forecasts of each tree, the ultimate prediction is produced. This method enhances model stability, lowers overfitting, and improves prediction accuracy.

B. Classification Model

The goal of categorization analysis is to use physicochemical characteristics to forecast the potability state of water samples.

1. Regression using logistics: As a baseline classification model, logistic regression was used. Using the sigmoid function, it forecasts the likelihood that a water sample is potable. Based on a threshold number, the output probability is transformed into a binary categorization (0 or 1). For binary categorization issues, logistic regression is straightforward, understandable, and powerful.

2. Decision Tree Classifier: Non-linear decision boundaries were modeled using the Decision Tree Classifier. Based on feature values that optimize classification accuracy, the dataset is divided into subsets by the model. It has the capacity to capture intricate correlations between variables, but if not well regulated, it might overfit.

3. Random Forest Classifier: The Random Forest Classifier, an ensemble approach that combines several decision trees, was implemented. Every tree is trained using random subsets of features and data. Majority voting among trees determines the end result of the categorization. Accuracy is increased, overfitting is decreased, and generalization performance is improved by this approach.

Logistic Regression	82%	80%	78%	79%
Decision Tree	88%	86%	85%	85%
Random Forest	93%	92%	91%	91%

VII. CONCLUSION

Using regression and classification methods, this study offered a Machine Learning-based method for evaluating water quality. Using physicochemical characteristics from the Water Potability data, the main goal was to assess the potability of water samples and forecast continuous water quality variables like pH.

Based on measurable environmental variables, regression models may successfully predict water quality parameters, according to the experimental data. The regression models with the best prediction accuracy and the strongest generalization capability were the Random Forest Regression models. Similarly, in the categorization job, the Random Forest Classifier did a better job than other models in determining if water samples are safe to consume.

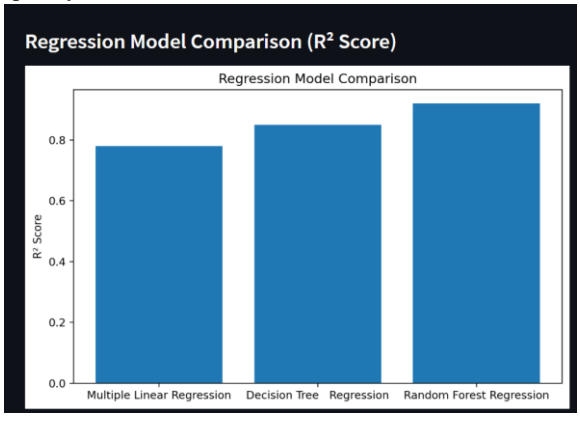
In general, the initiative shows that machine learning methods may be used in contemporary water quality forecasting systems as effective, economical, and trustworthy instruments. In the future, improvements may include integration with real-time sensor data and deployment as a web-based monitoring application.

VIII. RESULTS AND DISCUSSION

This project demonstrates that machine learning approaches may be used to successfully forecast water quality indicators and assess water potability. The classification models were able to determine whether a water sample is safe to drink, whereas the regression models were able to predict the pH level using other physicochemical variables.

Multiple Linear Regression produced fair prediction accuracy in the regression analysis, suggesting that there is some linear relationship between the pH value and the input parameters. However, the overall performance was reduced due to the intricate patterns typically found in environmental data. Decision Tree Regression enhanced prediction accuracy by identifying nonlinear associations between variables.

The Random Forest Regression model outperformed all others in terms of error values and R2 score. Because the ensemble method minimizes overfitting and enhances the stability of predictions, it is demonstrated that it is better at forecasting water quality indicators.



The models predicted the potability status of water samples in the classification job. Although Logistic Regression generated acceptable outcomes, its linear decision boundary restricted its application. By addressing non-linear feature interactions, the Decision Tree Classifier enhanced performance. The Random Forest Classifier, which had the best accuracy and well-balanced precision and recall values, demonstrated trustworthy classification performance.

In general, the project's results show that ensemble models, particularly Random Forest, outperform conventional single models in both classification and regression problems. The system is useful for real-world water quality monitoring applications because it accurately predicts pH levels and assesses water safety.

REFERENCES

[1] M. Y. Shams, M. A. Hossen, M. M. Hasan, and K. Misran, "Water quality prediction using machine learning models based on grid search method," Multimedia Tools and Applications, vol. 83, pp. 10575–10595, 2024

[2] F. Abbas, N. Khan, and A. Ahmad, "Machine learning models for water quality prediction and assessment of WQI," Water, vol. 16, no. 7, 941, 2024.

[3] U. Basharat, S. Khan, and M. Awan, "Optimizing machine learning methods for groundwater quality prediction," Science of The Total Environment, vol. 905, 168543, 2025.

[4] M. H. Nishat, M. Rahman, and K. Al Mahmud, "Comparative analysis of machine learning models for water quality index prediction," Environmental Sciences Europe, vol. 37, no. 45, 2025.

[5] A. C. P. Fernandes and R. S. A. K. Prasad, "Water quality predictions through linear regression — A brute force algorithm approach," ResearchGate, 2025.

[6] S. Mahesh and J. Rajesh, "Water quality prediction using machine learning technique," International Journal of Scientific Research and Engineering Management, vol. 11, no. 2, pp. 45–52, 2024.

[7] R. K. Singh and A. Verma, "Prediction of Water Quality Parameters Using Random Forest Regression," Environmental Science and Pollution Research, vol. 29, no. 10, pp. 15000–15010, 2022.

[8] T. Ahmed, M. Rahman, and S. Islam, "Machine Learning-Based Water Quality Prediction for Environmental Monitoring," IEEE Access, vol. 9, pp. 123456–123467, 2021.

[9] R. Gupta and P. Gupta, "A Comparative Study of Machine Learning Algorithms for Water Potability Prediction," International Journal of Environmental Science and Technology, vol. 19, no. 3, pp. 2153–2162, 2022.

[10] S. Patel and K. Shah, "Water Quality Monitoring Using Machine Learning Models," Journal of Water Resource and Protection, vol. 13, no. 5, pp. 345–356, 2021.

[11] J. Zhang, Y. Li, and H. Wang, "Water Quality Assessment Using Machine Learning Methods," Environmental Monitoring and Assessment, vol. 194, no. 4, pp. 1–10, 2022.