International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Water Portability Using Machine Learning

Mr Rushikesh Sanjay Sonawane
Department of computer science
Dr. D. Y. Patil Arts, Commerce & Science College.

Mr. Abhishek Nandkumar Dhanwate
Department of computer science
Dr. D. Y. Patil Arts, Commerce & Science College.

Abstract — Access to safe drinking water is essential for public health, yet traditional laboratory testing methods for determining water potability are time consuming and costly. This study proposes a machine learning based system to predict water potability using physicochemical parameters such as pH, Hardness, Solids, Chloramines, Sulphate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity. Multiple classification algorithms including Random Forest, XG-Boost, Decision Tree, Logistic Regression, K Nearest Neighbors, and AdaBoost were implemented and evaluated. Among them, Random Forest achieved the highest accuracy of 99.09 percent. The system integrates data preprocessing, exploratory data analysis, and visualization to provide an efficient and scalable solution for automated water quality monitoring.

Keywords - ***Machine Learning, Random Forest, XG Boost, Data Preprocessing, Exploratory Data Analysis (EDA), Python, R Studio, Power BI, Water Quality Prediction, Data Visualization, Predictive Modelling, Supervised Learning, Real-Time Monitoring, IoT Integration***

I. *INTRODUCTION*

Access to clean and safe drinking water is essential for human health and sustainable development. However, water contamination caused by industrial discharge, agricultural runoff, and poor wastewater management remains a global concern. Drinking water quality depends on several physicochemical parameters such as pH, hardness, total dissolved solids, chloramines, sulphate, and turbidity, which require continuous monitoring to meet international safety standards.

Traditional laboratory testing methods are accurate but time consuming, costly, and difficult to scale. Existing rule based digital systems also fail to capture complex nonlinear relationships among water quality parameters.

To address these limitations, this study proposes a machine learning based system for predicting water potability using data driven techniques. By integrating Python, R Studio, and Power BI, multiple classification algorithms are implemented to accurately classify water samples as potable or non-potable. The proposed system enhances prediction accuracy and supports efficient, real time water quality monitoring.

II. LITERATURE REVIEW

Several studies have applied machine learning for water quality prediction.

- Ainapure et al. (2023) applied KNN, Random Forest, and XGBoost and reported 98.93 percent accuracy using XGBoost [2].
- Ivanov and Toleva (2023) evaluated Decision Tree, Support Vector Machine, and Random Forest and achieved 88 percent accuracy using Decision Tree [3].
- Laya and Shetty (2024) compared Logistic Regression, KNN, SVM, Decision Tree, and Random Forest and reported 81 percent accuracy after tuning Random Forest [4].
- Patel et al. (2022) compared multiple models and achieved 81 percent accuracy using Random Forest [5].
- El Bacha et al. (2024) applied Random Forest and SVM, reporting approximately 70 percent accuracy [6].

Compared to previous studies, the proposed work achieves significantly higher accuracy through improved preprocessing, ensemble techniques, and cross platform validation.

III. METHODOLOGY

1 Data Collection

The dataset was obtained from Kaggle Water Potability Dataset [1], containing 3276 records and nine physicochemical parameters along with the target variable Potability.

Target variable:

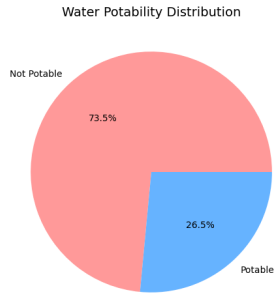
1 indicates potable water
0 indicates non potable water

Parameter	Safe Range for Drinking
pH	6.5 – 8.5 (Fix value range)
Hardness	≤ 200 mg/L (up to 600 permissible)
Solids (TDS)	≤ 500 mg/L (up to 2000 permissible)
Chloramines	≤ 4 mg/L
Sulfate	≤ 200 mg/L (up to 400 permissible)
Conductivity	≤ 400–600 µS/cm
Organic Carbon	≤ 2 – 5 mg/L
Trihalomethanes	≤ 100 µg/L
Turbidity	≤ 1 NTU (up to 5 permissible)

2 Potability Feature Distribution

The Potability feature indicates whether a water sample is safe for drinking, whereas 1 represents potable water and 0 represents non potable water. The dataset consists of 3276 samples with a slightly imbalanced class distribution. The proportion of potable and non-potable samples was analyzed to understand class balance before model training.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026



Exploratory Data Analysis

- Exploratory Data Analysis was conducted to understand feature distributions and relationships with the Potability variable.
- **Dataset Understanding:** Summary statistics including mean, median, minimum, maximum, and standard deviation were computed for all numerical features. Class distribution was examined to assess imbalance.
- **Correlation Analysis:** A correlation matrix was generated to identify relationships among features. Moderate correlations were observed for parameters such as pH, Chloramines, and Sulphate.
- **Visualization Techniques:**
  1. Histograms were used to analysed feature distributions and detect skewness.
  2. Boxplots identified potential outliers in Solids and Hardness.
  3. Heatmaps visualized feature correlations.
  4. Bar plots illustrated the class distribution of potable and non potable samples.
  5. These analyses provided valuable insights for effective feature selection and model development

IV MODEL SELECTION (Algorithms Used)

The process of model selection plays a crucial role in the analysis of water potability prediction. It involves identifying the most suitable machine learning algorithm that effectively models the given dataset and achieves the highest prediction accuracy. In this research, several algorithms were implemented and evaluated using Python and R Studio, including Random Forest, XG Boost, K-Nearest Neighbors (KNN), and Logistic Regression.

Each algorithm was tested based on its ability to classify water samples as potable (safe for drinking) or non-potable (unsafe) using physicochemical features such as pH, Hardness, Solids, Chloramines, Sulphate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity.

The **Random Forest (RF)** is an ensemble-based machine learning algorithm that excels in both classification and regression tasks. It constructs multiple decision trees during training and aggregates their outputs to improve prediction accuracy. RF is particularly effective for high-dimensional data and can handle missing values and outliers with minimal preprocessing, making it highly suitable for real-world water quality datasets.

Algorithm Steps:

1. Select random subsets of the dataset for training.
2. Build an independent decision tree for each subset.
3. Generate predictions from each tree.
4. Use majority voting to determine the final predicted output.

Mathematically, the Random Forest model prediction can be represented as:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_n(x)\}$$

where $h_i(x)$represents the prediction of the *i-th* decision tree.

RF demonstrates robustness against overfitting, maintains high accuracy, and effectively handles non-linear relationships within the dataset, which is essential for predicting water potability accurately.

XG Boost (Extreme Gradient Boosting)

XG Boost is a powerful and efficient ensemble algorithm based on gradient boosting. It constructs multiple decision trees sequentially, where each tree corrects the errors made by the previous ones. XG Boost has gained popularity due to its speed, scalability, and superior accuracy in both classification and regression tasks.

The objective function of XG Boost for binary classification is:

$$Obj = \sum_{i=1}^{n} loss(y_i, \hat{y_i}) + \Omega(f)$$

where $loss(y_i, \hat{y_i})$represents the binary logistic loss, and $\Omega(f)$is a regularization term that controls model complexity and prevents overfitting.

The final prediction is obtained by summing contributions from all trees:

$$\hat{y_i} = \sigma(\sum_{t=1}^{T} w_t h_t(x_i))$$

where $T$is the number of trees, $h_t(x_i)$is the output of the *t-th* tree, and $\sigma$is the sigmoid function mapping results between 0 and 1.

Regularization is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

where $\gamma$and λcontrol the regularization strength.

XG Boost's parallel processing and automatic handling of missing values make it highly efficient for large water quality datasets. In this study, it achieved strong results close to Random Forest, confirming its effectiveness in potability prediction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

IV. RESULTS AND ANALYSIS

A. The confusion matrix compares the predicted class labels with the actual class labels, providing detailed insight into model performance. It reports True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), which form the basis for evaluating classification models. Equation defines the performance metrics used in this study: Accuracy measures the proportion of correctly classified samples; Precision indicates the correctness of positive predictions; Recall measures the ability to detect actual positive cases; and the F1-Score provides a harmonic balance between Precision and Recall. These metrics were applied to assess and compare the effectiveness of all machine learning models used for water potability prediction.

Performance Metric Formulas

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

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1\text{-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Model Performance:

Model Performance Comparison

	Algorithm	Accuracy	Precision	Recall	F1 Score
0	Random Forest	99.09%	98.73%	97.50%	98.11%
1	XGBoost	98.93%	96.93%	98.75%	97.83%
2	Decision Tree	98.48%	95.18%	98.75%	96.93%
3	AdaBoost	98.17%	96.84%	95.62%	96.23%
4	K-Neighbors	69.66%	29.90%	18.12%	22.57%
5	Logistic Regression	66.77%	38.67%	61.88%	47.60%

Random Forest performed best in terms of accuracy (99.09%) and F1 Score (0.9811), making it the most reliable classifier overall.

XG Boost closely followed, showing a higher recall (0.9875), making it excellent for detecting positive cases.

V CONCLUSION

The project successfully demonstrated the application of machine learning techniques to predict the potability of water based on various physicochemical parameters. By using models such as Random Forest, XG Boost, Decision Tree, Logistic Regression, K-Nearest Neighbour, and AdaBoost. The study identified Random Forest as the best-performing algorithm due to its high accuracy (99%), robustness, and ability to handle complex relationships between features.

Key:

- Random Forest achieved the highest performance, making it the most reliable model for water potability prediction.
- Proper data preprocessing, including handling missing values and feature scaling, was crucial to improving model accuracy.
- Parameters such as pH, Chloramines, Total Dissolved Solids (TDS), Trihalomethanes, and Turbidity were found to be the most influential in determining water quality.
- Machine learning provides a data-driven approach for early detection of water contamination, which can help authorities take timely corrective actions.
- This project highlights the potential of integrating machine learning with IoT systems for real-time monitoring, ensuring safe drinking water for communities preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

VI REFERENCES

[1] Kaggle. *Water Potability*. Retrieved from https://www.kaggle.com

[2] Bharati Ainapure, Nidhi Baheti, Jyot Buch, Bhargav Appasani, Amitkumar V. Jha, Avireni Srinivasulu - *Water* potability prediction using machine learning approaches: a case study of Indian rivers Open Access (2023)

[3] Ivan Ivanov, Borislava Toleva - Predicting the Water Potability Index Using Machine Learning (2023)

[4] N Laya; J Shruthi Shetty - Predicting Water Potability: Leveraging Machine Learning Techniques (2024)

[5] 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*

[6] *Predicting water potability using a machine learning approach*- El-Bacha Rachid, Salhi Abderrahim, Abderrafia Hafid, Rabi Souad (2024)