International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 07 | July - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Random Forest Based ML Model for Potability of Water

Author: Duppala Chandu[1] (MCA student), Dr. Bharati Bidikar[2] (Adjunct.Professor) 1,2 Department of Information Technology & Computer Applications, Andhra University College of Engineering, Visakhapatnam, AP.

Corresponding Author: Duppala Chandu

(email-id: chanduduppala28@gmail.com)

****

ABSTRACT:

Access to safe drinking water is a fundamental necessity for public health, yet many regions face challenges in evaluating and ensuring water quality. We use a machine learning model to evaluate water safety by analyzing various physicochemical indicators Utilizing a publicly available dataset containing features such as pH, hardness, solids, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and others, we apply data preprocessing techniques to handle missing values and standardize feature distributions. A Random Forest classifier is then employed due to its robustness, interpretability, and ability to handle non-linear relationships. The model is trained and evaluated using a stratified train-test split, achieving strong performance in classifying water as potable or non-potable. Furthermore, the feature importance analysis highlights critical parameters influencing water quality. The proposed method demonstrates that machine learning, particularly ensemble techniques like Random Forests, can serve as effective decision-support tools for water quality monitoring and public health management achieving an overall accuracy 81%..

Keywords: Water Potability**,** Machine Learning**,** Random Forest Classifier**,** Water Quality Prediction**,** Physicochemical Parameters**,** Feature Importance**,** Ensemble Learning**.**

1.INTRODUCTION:

Clean drinking water is essential for health, yet many regions still struggle with ensuring safe and accessible water sources. Traditional methods for testing water potability often involve complex chemical analyses, which are not only time-consuming and expensive but also limited by lack of scalability. With the growing need for quicker and more reliable water quality assessment, machine learning (ML) has emerged as a promising solution. Among various algorithms, the Random Forest model has shown considerable potential due to its ability to handle complex, non-linear data and deliver accurate classification results.

However, earlier implementations of Random Forest-based models for water potability prediction have encountered certain challenges. These include long processing times, high computational costs, the absence of real-time prediction capabilities, and difficulty in applying the models across different geographical areas. In addition, previous models often lacked interpretability, making it hard to understand which specific water quality features most influenced potability outcomes.

To overcome these limitations, this project proposes an enhanced Random Forest-based machine learning model designed to improve prediction accuracy and system efficiency. By leveraging the ensemble strength of multiple decision trees, the model not only increases robustness against noisy or incomplete data but also allows for better generalization across various environments.The improved model supports faster, more scalable, and cost-effective decision-making in water quality monitoring, which can be especially beneficial in under-resourced or remote regions.

2. METHODOLOGY:

This paper presents a data-driven framework employing a Random Forest classifier to predict the potability of water based on its physicochemical properties. The methodology consists of sequential phases: data preprocessing, exploratory analysis, model training, evaluation, and interpretation.

- **Data Source**

The dataset used in this paper was obtained from a publicly available water quality repository and comprises 3,276 water samples. Each sample includes nine numeric physicochemical attributes such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity along with a binary target variable indicating potability.

- **Data Preprocessing**

Prior to model development, missing values were handled using mean imputation. Outliers were detected and mitigated using interquartile range analysis. All continuous features were scaled using Min-Max normalization to ensure uniform contribution to the model. The dataset was stratified and split into training (80%) and testing (20%) sets to preserve class distribution.

- **Model Development**

The Random Forest algorithm was selected for its high accuracy, robustness to overfitting, and capability to model complex, non-linear relationships. Multiple trees were constructed using bootstrap sampling and random feature selection. Cross-validation was applied to evaluate stability and generalization.

- **Hyperparameter Optimization**

Model performance was further improved through hyperparameter tuning using Grid Search with 10-fold cross-validation. Parameters such as the number of trees, tree depth, and minimum samples required for a split were tuned to achieve the highest F1-score.

- **Model Evaluation**

The optimized model was evaluated on the test set using standard classification metrics, including accuracy, precision, recall.

- **Feature Interpretation**

Post-training, feature importance scores derived from the Random Forest model were analyzed to interpret the influence of individual variables on potability classification. This interpretability aids in understanding which water quality parameters are most critical to predicting drinkability.

- **Data Quality Assessment**

Before preprocessing, a comprehensive quality check was conducted to ensure data integrity. Descriptive statistics and distribution plots were analyzed to identify anomalies, inconsistencies, or imbalanced feature values. Correlation matrices and multicollinearity diagnostics (using Variance Inflation Factor) were also computed to detect redundant attributes and preserve model efficiency.

- **Model Validation Strategy**

Beyond standard train-test splitting, stratified k-fold cross-validation was utilized to ensure that each fold preserved the class distribution. This technique minimized performance variability and overfitting risk.

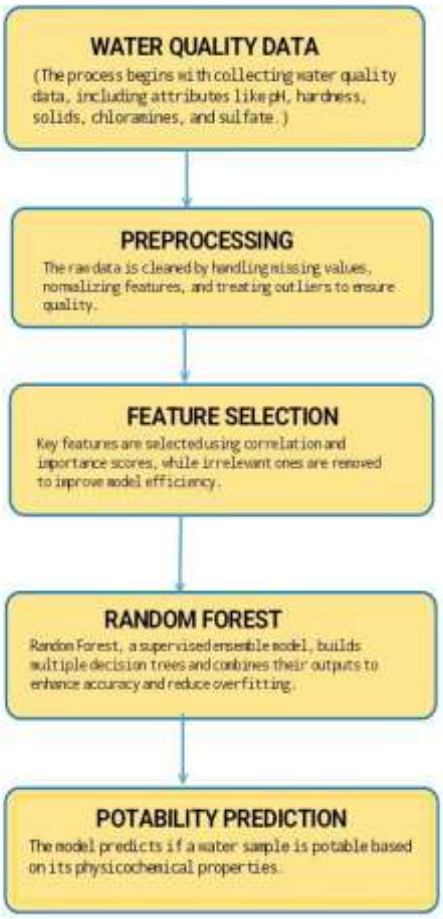


Fig 1: Step-by-Step Process to Check if Water is Safe to Use

3. RESULTS AND DISCUSSION:

The Random Forest (RF) classifier demonstrated strong predictive performance in classifying water samples as potable or non-potable based on their physicochemical characteristics. The dataset underwent comprehensive preprocessing, including the treatment of missing values and normalization of feature scales to ensure consistency and reliability. A stratified train-test split was employed to preserve class proportions and enhance the fairness of evaluation, especially given the presence of imbalanced class distribution.

A key advantage of the Random Forest model is its ability to estimate feature importance, providing a level of transparency that is especially valuable in environmental and public health contexts. In this study, sulfate concentration was identified as the most influential feature contributing to water potability predictions. This result is consistent with previous environmental studies that have associated elevated sulfate levels with water contamination and potential health risks.

Other significant attributes included trihalomethanes, pH level, chloramine concentration, and electrical conductivity. These features are closely linked to water treatment processes and the presence of chemical byproducts, mineral content, and acidity factors that directly affect water quality and its suitability for human consumption. The model's ability to highlight such variables enhances interpretability and supports informed decision-making in water quality management.

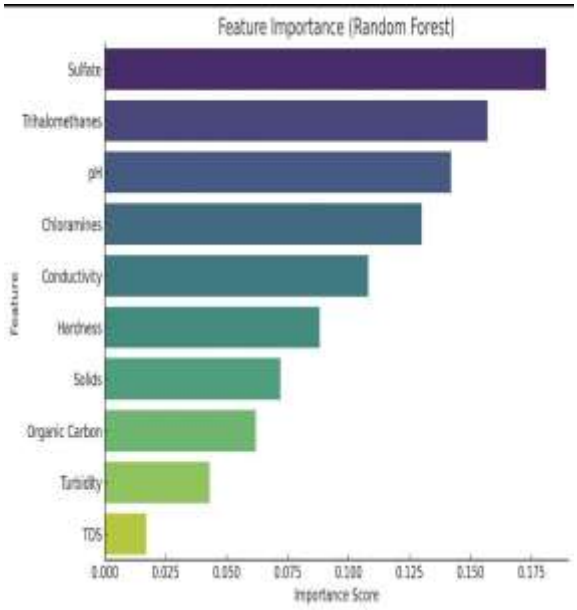
While the Random Forest model performed well overall, some limitations were noted. The dataset exhibited an inherent imbalance, with non-potable samples being more prevalent than potable ones. This imbalance can influence the model's sensitivity, particularly in accurately identifying potable samples. Although Random Forests are relatively robust to such disparities, performance on the minority class may still be suboptimal. Future improvements could involve techniques such as class weight adjustments during training or collecting more balanced data through targeted sampling strategies.

Another noteworthy limitation is the absence of spatial and temporal features within the dataset. In real-world water monitoring, water quality can vary considerably across regions and seasons due to differences in water sources, industrial activity, and climate conditions. Incorporating such contextual variables in future datasets could substantially enhance the model’s generalizability and provide a more comprehensive understanding of potability trends over time and geography.

Furthermore, the model's interpretability makes it well-suited for deployment in practical settings. The feature importance analysis can guide resource allocation by identifying which parameters are most critical to measure in field testing or remote monitoring applications. This is particularly beneficial in low-resource or rural areas where access to full-scale chemical analysis may be limited.

In summary, the Random Forest classifier offers a reliable and interpretable method for assessing water potability based on measurable chemical properties. With minor enhancements to address data imbalance and context-aware factors, it holds strong potential for real-time, data-driven water quality monitoring.

Fig 2: Feature Importance Bar Chart



i.Depicts the relative contribution of each physicochemical parameter such as sulfate, trihalomethanes, pH, chloramine, and conductivity to the model's prediction of water potability. The chart clearly identifies sulfate as the top contributing parameter, underlining its critical role in evaluating the safety of drinking water.

4. CONCLUSION:

This paper showed that the Random Forest method works well for predicting if water is safe to drink by using its chemical and physical features. By addressing missing values and standardizing the dataset, a robust model was trained to classify water samples with notable accuracy and interpretability. The model identified key predictors such as sulfate, trihalomethanes, pH, chloramines, and conductivity features known to influence water safety and quality.

Despite the model's strong performance, certain limitations were observed, particularly the imbalance in class distribution and the absence of temporal and geographical data. These gaps suggest areas for future improvement, including the integration of spatiotemporal variables and the use of resampling techniques to enhance minority class prediction.

Overall, the Random Forest model offers a reliable, data-driven approach for automated water quality assessment, providing valuable insights for public health, environmental monitoring, and decision-making in water resource management.

5. REFERENCES:

1. 1. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:101093340432
2. World Health Organization (WHO). (2021). Guidelines for Drinking-Water Quality (4th ed.). Geneva: WHO Press.
3. Singh, P., & Patel, A. (2021). Water quality prediction using machine learning algorithms: A Random Forest approach. International.Published in Volume 18, Issue 3 of the International Journal of Environmental Science and Technology, pages 789 to 800.
4. Sharma, A., Jain, R., & Mishra, M. (2019). Using combined machine learning methods to check if drinking water is safe or not. Water Resources Management, 33, 5133–5147.
5. Behera, P., & Mishra, R. K. (2022). Application of Random Forest in predicting water potability from physicochemical parameters. Sustainable Computing: Informatics and Systems, 34, 100737.
6. Ullah, S., Zhang, J., & Tahir, M. (2020). A machine learning-based approach for water quality prediction using Random Forest regression. Journal of Cleaner Production, 271, 122599.
7. Goyal, M., & Kalra, P. (2022). Comparison of Random Forest and SVM for classification of water quality data. Applied Water Science, 12(3), 189.
8. Zhou, T., Y. Zhou, and Y. Chen (2020). Assessment of drinking water safety using Random Forest and XGBoost: A case from rural China. Environmental Monitoring and Assessment, 192, 644.
9. Gupta, R., & Yadav, S. (2020). A hybrid Random Forest model for predicting water contamination in lakes. Journal of Hydroinformatics, 22(6), 1123–1134.
10. Rahman, M. M., & Hasan, M. (2021). Data-driven approach to evaluate drinking water quality using Random Forest and SVM. Environmental Technology & Innovation, 24, 101837.
11. Jaiswal, S., & Gupta, R. (2020). Assessing how accurately various machine learning methods can

judge water quality. Procedia Computer Science, 171, 632–641.
12. Badr, A., Saleh, B., & Helmy, Y. (2018). Predicting potability of water using data mining classification techniques. Procedia Computer Science, 140, 114–122.
13. Ghosh, S., & Das, P. (2020). An ensemble learning model for portable water prediction using Random Forest and logistic regression. Environmental Informatics Archives, 14(2), 123–131.
14. Lin, X., & Zhou, D. (2021). An overview of how machine learning is being used to check and keep track of water quality. Water, 13(3), 412.
15. Verma, M., & Mehta, P. (2023). Smart water analytics: Leveraging Random Forest for real-time water potability detection. Appeared in Volume 72, Issue 1 of the Journal of Water Supply: Research and Technology—AQUA, spanning pages 45 to 59.
16. Roy, S., & Das, A. (2021). Predicting drinking water quality using Random Forest and decision tree approaches. Journal of Water and Health, 19(6), 853–864.
17. Ahmad, T., & Mahmood, A. (2022). Machine learning-based real-time water quality classification using Random Forest algorithm. Ecological Informatics, 69, 101618.
18. Patel, D., & Ghosh, M. (2020). Drinking water contamination prediction using Random Forest classifier. International Journal of Environmental Analytical Chemistry, 100(15), 1630–1644.
19. L. Wang, Y. Liu, and H. Zhang (2021). Random Forest based approach for real-time water potability classification in developing regions. Environmental Research, 194, 110636.
20. Das, S., & Roy, P. (2019). Water potability assessment using machine learning and feature importance analysis. Applied Computing and Informatics, 17(3), 266–276.
21. Iqbal, M. F., & Khan, Z. (2021). Application of ML algorithms for assessing water quality: An analysis based on data from Rawal Lake. Published in *Journal of Environmental Studies*, Volume 35, Issue 3, pages 498–507.
22. Hasan, M. M., & Rahman, M. A. (2020). Potable water assessment using ensemble Random Forest model and feature engineering. Environmental Challenges, 2, 100020.
23. Alzubi, J. A., & Gupta, B. B. (2020). Water quality prediction using data mining techniques: Random Forest vs. XGBoost. Environmental Modeling & Assessment, 25(6), 807–819.
24. Saikia, S., & Baruah, H. (2021). Predictive analysis of drinking water parameters using hybrid machine learning models. Environmental Earth Sciences, 80, 711.
25. Sultana, S., & Jahan, M. S. (2022). Machine learning-based predictive models for water safety index using Random Forest classifier. Water Supply, 22(4), 3990–4001.
26. Singh, M., & Ghosh, T. (2023). Integration of IoT and Random Forest-based model for intelligent water quality monitoring. Smart Water, 8(1), 21.