IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 2, March-April 2025

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202032

Water Quality Analysis and Potability Prediction Using Machine Learning

Mr. A.DilipKumar, Dr. A C.Sountharraj, M.Phil., Ph.d

II PG, Department of Computer Science with Data Analytics, Dr. N.G.P. Arts and Science College, Coimbatore, India

Department of Computer Science with Data Analytics, Dr. N.G.P. Arts and Science College, Coimbatore, India

ABSTRACT: Monitoring water quality is crucial to guaranteeing sustainable and safe water supplies. This research analyzes water quality parameters and makes real-time predictions about pollution levels using machine learning (ML) and the internet of things (IoT). To as certain important indicators like pH, turbidity, dissolved oxygen, and conductivity, sensor data is gathered and processed. Using historical data, a machine learning model is built to classify water quality, increasing accuracy through ongoing learning. By warning users of possible risks, the system promotes proactive decision-making, improving environmental sustainability and public health. The project provides a scalable and effective way to test the quality of water since it incorporates cloud-based data management and visualization for remote monitoring.

The goal of this project is to integrate machine learning (ML) and the internet of things (IoT) to create an intelligent water quality monitoring system. It gathers sensor data in real time on important aspects of water quality, including conductivity, turbidity, pH, and dissolved oxygen. Machine learning algorithms are used to process the data in order to identify any contaminants and classify the water quality. Proactive water management is made possible by the system's ability to recognize patterns and issue early warnings by utilizing predictive analytics. It is a scalable and effective solution for environmental and public health applications since it integrates cloud-based data storage and visualization tools to enable remote monitoring. By increasing precision, automation, and real-time decision-making, this method improves on conventional water quality monitoring.

KEYWORDS: Water quality monitoring, Machine Learning, Internet of Things, Real-time analysis, Predictive analytics, Cloud-based monitoring, Environmental sustainability, Smart water management.

I. INTRODUCTION

The environment, agriculture, and public health are all impacted by the growing problem of water pollution. Conventional techniques for monitoring water quality are frequently labor-intensive, time-consuming, and devoid of real-time analysis. An effective and automated method of water quality monitoring is offered by this project, ML-Based Water Quality Monitoring, which combines Internet of Things (IoT) with machine learning (ML) technology. The system uses machine learning algorithms to process and analyze real-time sensor data on important parameters like conductivity, turbidity, dissolved oxygen, and pH in order to forecast contamination levels.

II. LITERATURE REVIEW

Due to the labor-intensive, time-consuming, and expensive nature of traditional methods of water quality evaluation, which are mostly based on physical and chemical testing, access to clean and drinkable water continues to be a major global concern. Machine Learning (ML) has become a scalable and effective method for analyzing water quality and predicting potability with the development of data science and increases in processing capacity. In order to determine whether water is safe for human consumption, several studies have used important water quality parameters as input features for machine learning (ML) models. These parameters include pH (acidity or alkalinity), turbidity (water clarity), dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), hardness, total solids, conductivity, chloramines, sulfate, nitrates, and trihalomethanes (THMs).

In these applications, supervised learning models like K-Nearest Neighbors (KNN), Random Forests (RF), Decision Trees (DT), Support Vector Machines (SVM), and Logistic Regression are frequently used. Choubey et al. (2021), for example, used Decision Trees and achieved an accuracy of over 85% in determining whether water was potable or not. By efficiently controlling feature interactions and lowering model variance, ensemble approaches—in particular, Random

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202032

Forests and Gradient Boosting techniques like XGBoost and LightGBM—have demonstrated higher performance. According to Patel & Parmar (2022), Random Forests are effective at reaching accuracy levels of about 90%.

III. METHODOLOGY

System Architecture

From data collection to model prediction, the system design for this ML-Based Water Quality Monitoring project follows an organized path. This is a detailed breakdown of the architecture:

A. Data Acquisition & Pre-processing

Data Acquisition & Pre-processing Dataset Collection

The data was gathered from publicly accessible water quality datasets that included chemical concentrations, pH, turbidity, and total dissolved solids (TDS).

Pictures The information comes from reputable sources such internet repositories for machine learning research, environmental monitoring organizations, and government water quality reports.

Model Selection and Training Regression

- **Logistic Regression**
  Selection: Models the likelihood of a water quality category and is used for binary classification.
  Training: Optimizes coefficients to maximize likelihood by fitting a logistic function to the data.
- **KNN, or K-Nearest Neighbors**
  Selecting the best hyperplane for classification is effective when dealing with high-dimensional data.
  Finding the hyperplane that optimizes the margin between classes is the goal of training.
- **Linear Support Vector Machine**
  Selecting the best hyperplane for classification is effective when dealing with high-dimensional data.
  Training: Determines which hyperplane maximizes the class margin.
- **The Gaussian Naïve Bayes**
  Selection: Effective for high-dimensional data, it assumes features are regularly distributed.
  Training: Assuming feature independence, it computes conditional probabilities using Bayes' theorem.
- **Decision Tree**
  Selection: Creates a tree-like structure to represent decision rules, easy to interpret.
  Training: Recursively partitions data based on feature values to maximize I information gain.
- **Random Forest**
  Selection: An ensemble method that combines multiple decision trees, reducing overfitting and improving accuracy.
  Training: Trains multiple decision trees on random subsets of the data and averages their predictions.

Model Evaluation

Performance Metrics

Accuracy: Measures the overall percentage of correctly classified water quality samples. This indicates how often the model correctly predicts the water's quality category (e.g., "safe," "unsafe").

Precision: Assesses the model's ability to avoid false positives for each water quality category. Specifically, it reveals how often a sample predicted as "unsafe" is truly unsafe.

Recall (Sensitivity): Indicates how well the model can identify real-world occurrences of a specific water quality category. It demonstrates, for instance, how well the model detects all water samples that are actually. When working with imbalanced datasets—where certain water quality categories are more prevalent than others—the F1-score offers a balanced measure of precision and recall.

Confusion Matrix: Shows each water quality category's categorization performance while emphasizing incorrect classifications. It highlights the categories that are frequently mistaken for one another, suggesting possible areas for model enhancement.

Validation & Testing

Training and testing sets are created from the water quality dataset. The models are trained using the training set, and their performance on unseen data is assessed using the testing set.

The model's capacity to generalize in the real world is evaluated using an independent Test Set. This guarantees that overfitting on the training data won't result in an overly optimistic model performance.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202032

To maximize the model's performance and aid in choosing the optimal model, cross-validation. By ensuring that the trained models are dependable and appropriate for practical implementation in water quality monitoring applications, this evaluation procedure produces accurate and trustworthy evaluations of water safety.

IV. RESULTS

Different levels of predicted accuracy were found when the machine learning models were evaluated using the test dataset. In general, the Random Forest Classifier and Support Vector Machine using an RBF kernel performed well, showing that they could identify intricate patterns in the water quality data. The capacity of these models to manage non-linear correlations and interactions among the features probably helped them. As an ensemble approach, Random Forest also reduced overfitting, which strengthened its generality.

On the other hand, the Gaussian Naive Bayes and Logistic Regression models tended to have relatively lower accuracy. This might be explained by the fact that these models' basic presumptions—linearity in the case of logistic regression and feature independence in the case of naive bayes—would not have been entirely consistent with the distribution of the water quality dataset. Although the K-Nearest Neighbours and Decision Tree models performed reasonably well, it appears that they may be more susceptible to noise or outliers in the data, even though they were able to identify some pertinent trends.

V. CONCLUSION

To sum up, this project successfully illustrates how machine learning methods may be applied to water quality monitoring. In order to accurately predict water quality categories based on several physicochemical factors, the research used a variety of classification models, such as Random Forest, Support Vector Machines, and Logistic Regression. Metrics like accuracy, precision, recall, and F1-score were used to evaluate these models, highlighting the advantages and disadvantages of each strategy. Interestingly, ensemble techniques like Random Forest typically performed well, demonstrating their capacity to manage intricate datasets and reduce overfitting.

REFERENCES

1. Choubey, S., Tiwari, A., & Shukla, R. (2021). *Water quality assessment and potability prediction using decision tree algorithm*. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(1), 123-128.
2. Patel, H., & Parmar, J. (2022). *Machine learning based water potability prediction using ensemble models*. International Journal of Innovative Research in Science, Engineering and Technology, 11(4), 567-574.
3. Kumar, R., Sharma, A., & Singh, M. (2020). *Application of artificial neural networks in water quality prediction*. Environmental Monitoring and Assessment, 192(10), 1-13.
4. Rathore, S., Ahmed, A., & Jhanjhi, N. (2022). *IoT-enabled smart water quality monitoring system using cloud-based machine learning model*. Journal of Ambient Intelligence and Humanized Computing, 13(2), 543–556.
5. UCI Machine Learning Repository. (n.d.). *Water Potability Dataset*. Retrieved from https://archive.ics.uci.edu/ml/datasets/Water+Potability
6. Central Pollution Control Board (CPCB), India. (2021). *Water quality monitoring status and reports*. Retrieved from http://cpcb.nic.in
7. National Water Quality Monitoring Council (NWQMC), USA. (2020). *Water quality data portal*. Retrieved from https://www.waterqualitydata.us/
8. Brown, J., & Mihelcic, J. (2020). *Assessing the sustainability of water quality monitoring approaches using IoT and machine learning*. Journal of Environmental Management, 255, 109896.
9. Zhang, Y., Zhang, Y., & Qian, Z. (2021). *A comparative study of machine learning algorithms for water potability classification*. Journal of Water and Health, 19(3), 432-444.
10. Ghosh, A., & Das, S. (2023). *Explainable AI in environmental monitoring: Opportunities and challenges in water quality prediction*. Environmental Informatics Letters, 2(1), 45-53.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975 Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com