Ensemble-Based Classification of Groundwater Potability Using Multi-Model Stacking Techniques

Sanket Dhumal
Department of Computational Intelligence
SRM Institute of Science and Technology
Chennai, India
sd9093@srmist.edu.in

Shaik Afzal
Department of Computational Intelligence
SRM Institute of Science and Technology
Chennai, India
as1238@srmist.edu.in

B.Pitchaimanickam
Department of Computational Intelligence
SRM Institute of Science and Technology
Chennai, India
bpitmani@gmail.com

***Abstract*—Groundwater serves as a primary source of drinking water for millions, particularly in rural and semi-urban regions. However, escalating industrial activities and pollution have raised significant concerns about water safety. This paper introduces a smart, automated groundwater assessment system that evaluates potability using key chemical indicators like pH, TDS, Hardness, Chlorides, Nitrates, Sulfates, and Conductivity. The approach relies on a machine learning model trained with authentic water quality data to ensure precise predictions.**

Keywords- Multi-Layer Perceptron, Quadratic Discriminant Analysis, CatBoost, Extra Tree Classifier, Feature Engineering and Ensemble Learning.

I. Introduction

Groundwater is one of the most vital natural resources, serving as the primary source of drinking water for a significant portion of the global population. It plays a crucial role in agricultural irrigation, industrial applications, and domestic consumption. The availability of clean groundwater is essential for public health and environmental sustainability. However, with increasing urbanization, industrial expansion, and agricultural intensification, groundwater contamination has become a major environmental concern. Pollutants such as heavy metals, nitrates, pesticides, industrial effluents, and microbial contaminants pose significant risks to human health and the ecosystem. Contaminated groundwater can lead to severe health problems, including waterborne diseases, cancer, neurological disorders, and organ damage. Therefore, continuous monitoring and assessment of groundwater quality are essential to prevent adverse health effects and ensure sustainable water management.

Traditional methods of groundwater quality assessment involve collecting water samples and performing laboratory tests to analyze chemical and biological properties. While these methods provide accurate results, they are often time-consuming, labor-intensive, and costly. Additionally, they require skilled professionals to conduct the analysis and interpret the data. Given these limitations, there is a growing need for an automated and efficient approach that can provide real-time groundwater quality classification based on physicochemical parameters.

This work proposes the development of an intelligent system that assesses groundwater quality using advanced machine learning models. The system processes historical datasets containing chemical attributes such as pH, TDS, Hardness, Chlorides, Nitrates, Sulfates, and Conductivity to evaluate whether a given water sample is suitable for drinking. The model produces a binary outcome—either indicating the water is safe for consumption or labeling it as unsafe—with a strong focus on predictive precision.

The contributions listed below are made in the paper:

i) This study leverages a variety of supervised learning algorithms to determine whether groundwater is suitable for consumption based on key water quality indicators.

ii) A stacking-based ensemble method is utilized to integrate predictions from multiple models, enhancing the system's overall accuracy.

iii) The project aligns with sustainable development goals by promoting accessible, fast, and automated groundwater quality analysis to support clean water initiatives.

The paper's structure is organized as follows: Section II provides a review of the existing studies on the evaluation of groundwater quality using machine learning approaches. Section III introduces the dataset and its key characteristics. Section IV outlines the chosen methodology, detailing how the models were built and how the ensemble strategy was structured. Section V delivers the findings from the experiments, along with a comparative discussion of the outcomes.

II. Literature Survey

The classification of groundwater quality using machine learning has witnessed significant growth over the past decade, driven by the urgent need for reliable, real-time assessment tools. Traditional water testing methods, though accurate, are often resource-intensive and inaccessible in remote regions. As a result, researchers have turned to machine learning (ML) algorithms as an alternative for automating the classification of water potability based on physicochemical parameters.

D. Karunanidhi [1] implemented Decision Tree (DT) algorithms due to their simplicity and interpretability. Decision Trees construct rule-based models that mirror human reasoning, making them easy to understand. In groundwater classification, DTs were effective in identifying relationships between parameters like pH, TDS, chlorides, and sulfates. However, their tendency to overfit training data limited their ability to generalize predictions to new samples. To overcome this, M. R. Islam, S. S. Hossain and A. K. Chakrabarty [6] implemented Random Forest (RF), ensemble aggregates the results of multiple trees to improve robustness. H. Chen et al. [7], who showed that Bagging and Boosting enhance model robustness and scalability. S. Gupta et al. [9] further improved performance by proposing a hybrid model that combines multiple classifiers, boosting generalization on noisy data. A. Kumar et al. [3] also confirmed that ensemble methods outperform single models. In addition, deep learning approaches, such as those by X. Li et al. [5], have shown strong potential in capturing complex patterns in high-dimensional data, while Zhang and Yang [11] highlighted the benefits of multi-task learning for environmental classification.

To develop faster and more robust models, researchers like Chen and Guestrin [15] emphasized XGBoost for its efficiency and scalability, especially with missing or unbalanced data. Kour and Arora [12] also found that ensemble methods like Gradient Boosting and Random Forest outperformed others in potability classification. J. Smith and L. Brown [4] highlighted that model performance varies with preprocessing and feature selection. Studies by Sarkar and Pandey [17] and Xu et al. [16] demonstrated the effectiveness of supervised models in real-time and resource-limited settings. Additionally, stacking ensembles, as explored by Roy et al. [14] and Bhagat and Mohapatra [20], showed improved reliability by combining outputs from multiple base models through a meta-classifier.

Several studies have explored diverse approaches for groundwater quality prediction. N. Patel and Sharma [8] found Naïve Bayes effective for simple, categorical data, while Adeyemi et al. [2] demonstrated the utility of ML models in data-scarce regions like Sub-Saharan Africa. Mishra and Pradhan [10] highlighted the efficiency of XGBoost, CatBoost, and LightGBM for real-time predictions. Panigrahi et al. [19] and Kumar et al. [13] compared boosting algorithms, emphasizing their performance and interpretability. Jha et al. [14] integrated statistical and ML methods to enhance model robustness, and Rezaei et al. [18] introduced a hybrid QDA model to improve sensitivity in detecting contamination.

The authors have proposed a novel integration of multiple advanced machine learning models—including MLP, QDA, Extra Trees, and CatBoost—combined using a stacking ensemble to improve groundwater potability prediction.

III. METHODOLOGY

To assess groundwater potability using machine learning, a systematic workflow is adopted. The core phases include acquiring data, cleaning and preparing it, building and training models, integrating ensemble strategies, and creating an interface for practical, real-time use.

The methodology begins by compiling a dataset in CSV format containing several chemical indicators relevant to water quality. Notable features such as pH, TDS, Hardness, Chloride, Nitrate, Sulfate, and Conductivity are utilized by the system to classify samples as either drinkable or non-drinkable.

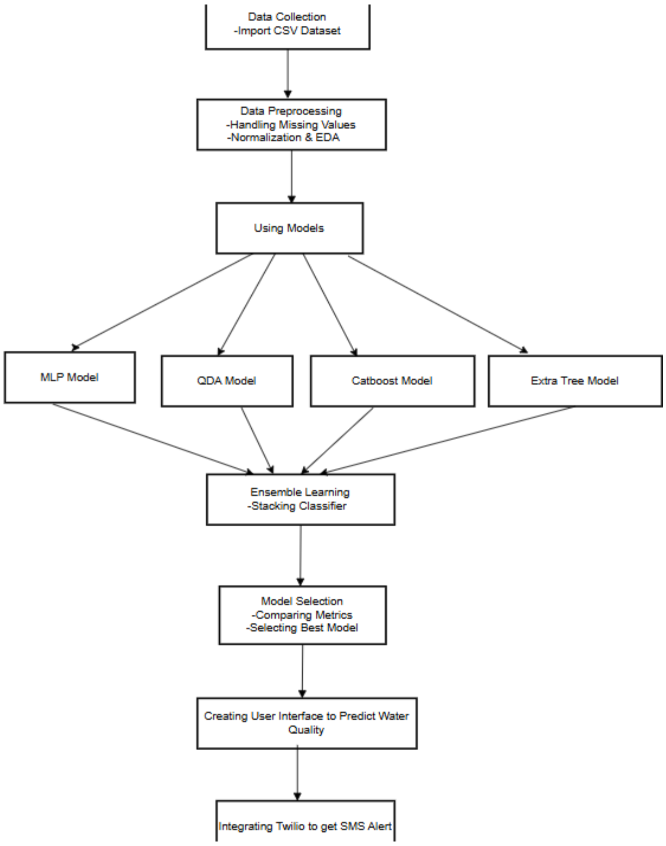

Fig.1. Architecture Diagram

Once the dataset is prepared, an Exploratory Data Analysis (EDA) is conducted. This step is crucial to understand the distribution, trends, and any inconsistencies within the data.

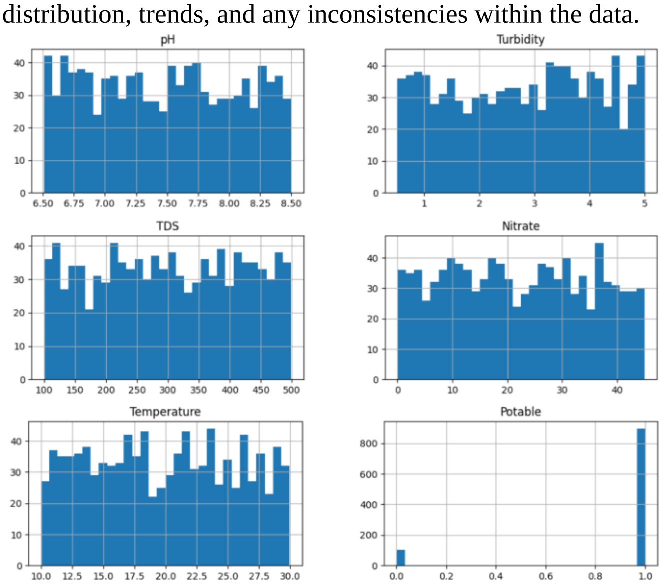

Fig.2. Preliminary Data Insights (EDA)

Raw inputs are converted into a standardized format appropriate for machine learning model training during the data preprocessing stage.This includes addressing gaps in data through imputation techniques (e.g., using averages or medians), and normalizing values with scaling methods such as StandardScaler or MinMaxScaler. These operations help maintain the consistency and reliability of the model during learning.

Model -Specific Equations

a) Quadratic Discriminant Analysis

$$\delta_k(x) = -\frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \ln P\,(C_k) \quad (1)$$

Where:

- μ_k is the average vector for class k,
- Σ_k denotes that class's covariance matrix,
- $\mathrm{P}(C_k)$ reflects how likely class k is before observing any features

b) Multi- Layer Perceptron

$$z = W \cdot x + b \quad (2)$$
$$a = \sigma(z) \quad (3)$$

Where:

- W: Weight
- b: bias
- σ:Activation Function
- x: Input vector

c) Extra Trees Classifier

Entropy:

$$H(S) = -\sum_{i=1}^{n} p \ \log_2(p_i)] \quad (4)$$

Information Gain:

$$IG(S, A) = (S) - \sum_{v \in \mathrm{Values}(A)} \frac{|S_v|}{|S|} H(S_v) \quad (5)$$

d) CatBoost

$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$

Where:

- $F_{m-1}(x)$ is the model's prediction from the last boosting round
- γ_m represents the step size or learning rate
- $h_m(x)$ is the additional base model added to correct the prior error

e) **Stacking Classifier**

$$\hat{y} = H\big(h_1(x), h_2(x), \ldots, h_k(x)\big) \quad (6)$$

Where:

- h_i: Predictions from the individual base models
- H: The meta-model that combines these predictions
- $\hat{y}$: The final output generated by the stacking ensemble

Evaluation Metrics Equations

a) Accuracy

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

b) Precision

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

c) Recall

$$\text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

d) F1-Score

$$\text{F1-Score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (11)$$

The final phase focuses on creating an intuitive user interface for real-time groundwater quality entry, enabling users to assess water potability fast and easily. The system takes the data and returns instant, intelligible forecasts, and data-driven insights become accessible to non-experts such as field operators and community health workers.

IV. RESULTS AND DISCUSSIONS

This part of the study delivers an in-depth examination of how different machine learning techniques perform when applied to the classification of groundwater quality.

Dataset Description

The dataset includes physicochemical features like pH and Total Dissolved Solids (TDS), essential for assessing groundwater potability. pH indicates water's acidity or alkalinity, while TDS measures dissolved substances, with high values potentially indicating unsafe conditions.

a) Multi – Layer Perceptron

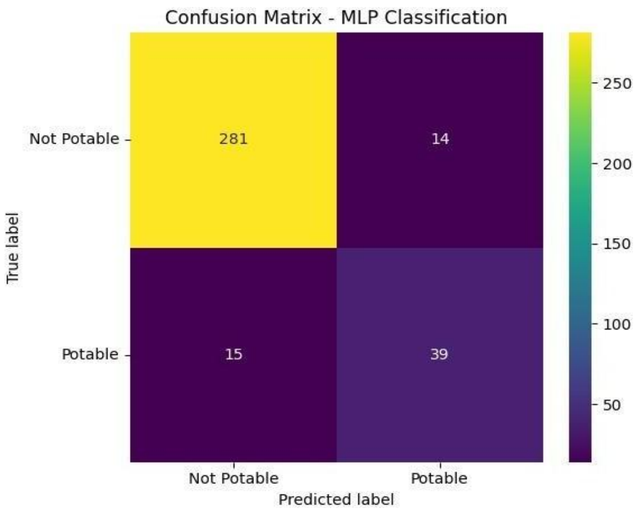

Fig 3. Confusion Matrix of MLP

Figure 3 illustrates the classification outcomes of the Multilayer Perceptron (MLP) model in predicting groundwater quality. The model demonstrates notable accuracy, particularly when identifying unsafe water.

TABLE I. CLASSIFICATION REPORT OF MLP

Category	Precision	Recall	F1-Score	Samples
0	95%	95%	95%	295
1	74%	72%	73%	54
Overall Accuracy			92%	349
Macro Average	84%	84%	84%	349
Weighted Average	92%	92%	92%	349

b) Quadratic Discriminant Analysis

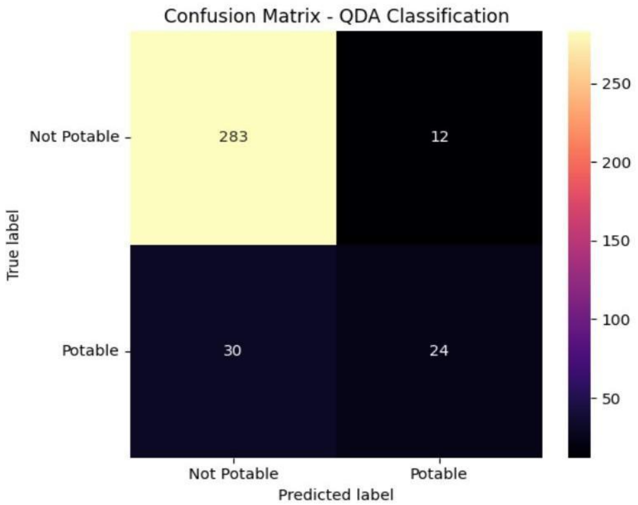

Fig 4. Confusion Matrix of QDA

Figure 4 illustrates the performance of the Quadratic Discriminant Analysis (QDA) model in predicting groundwater quality. The model achieved an overall accuracy of 88%, indicating a high level of reliability. It was particularly effective in identifying non-potable water, accurately classifying 283 out of 295 unsafe samples, which corresponds to a high specificity of 96%. However, the model demonstrated limitations in detecting potable water, misclassifying 30 out of 54 safe samples, thereby exhibiting a relatively high false-negative rate.

TABLE II. CLASSIFICATION REPORT OF QDA

Category	Precision	Recall	F1-Score	Samples
0	90%	96%	93%	295
1	67%	44%	73%	54
Overall Accuracy			88%	349
Macro Average	79%	70%	73%	349
Weighted Average	87%	88%	87%	349

c) CatBoost Classifier

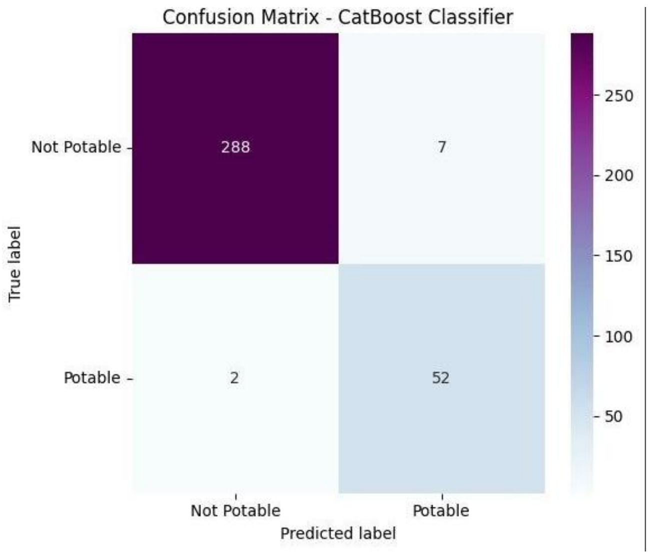

Fig 5. Confusion Matrix of CatBoost

Figure 5 highlights the performance of the CatBoost Classifier on the groundwater potability dataset. The model achieved an impressive overall accuracy of 97%, accurately classifying 288 out of 295 non-potable samples and 52 out of 54 potable ones. With minimal misclassifications, it demonstrated strong predictive capability, reflected in an F1-score of 0.92 and a precision of 0.88—underscoring its effectiveness in identifying both safe and unsafe water samples.

TABLE III. CLASSIFICATION REPORT OF CATBOOST

Category	Precision	Recall	F1-Score	Samples
0	99%	98%	95%	295
1	88%	96%	73%	54
Overall Accuracy			97%	349
Macro Average	93%	91%	94%	349
Weighted Average	96%	96%	95%	349

d) Extra Trees Classifier

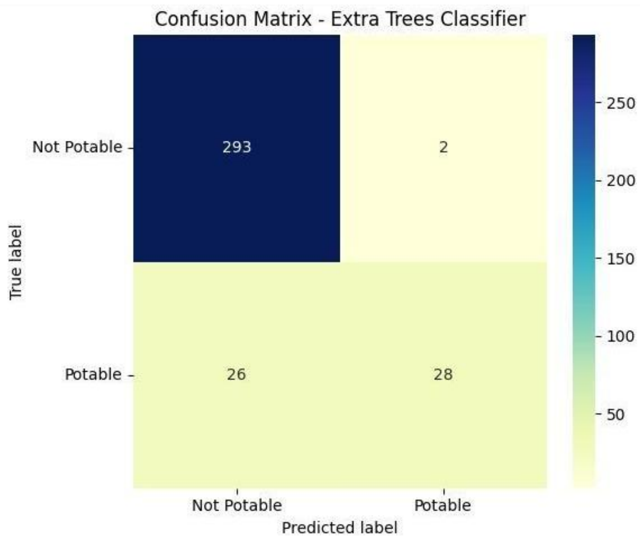

Fig 6. Confusion Matrix of Extra Tree Classifier

Figure 6 displays the performance of the Extra Trees Classifier in groundwater quality classification. The model achieved a commendable accuracy of 92% and was particularly effective in detecting non-potable water, correctly identifying 293 out of 295 samples with just 2 false positives. Additionally, it maintained a strong precision of 93%, indicating reliable overall performance.

TABLE IV. CLASSIFICATION REPORT OF EXTRA TREE CLASSIFIER

Category	Precision	Recall	F1-Score	Samples
0	92%	99%	95%	295
1	93%	52%	67%	54
Overall Accuracy			92%	349
Macro Average	93%	76%	81%	349
Weighted Average	92%	92%	91%	349

e) Stacking Classifier

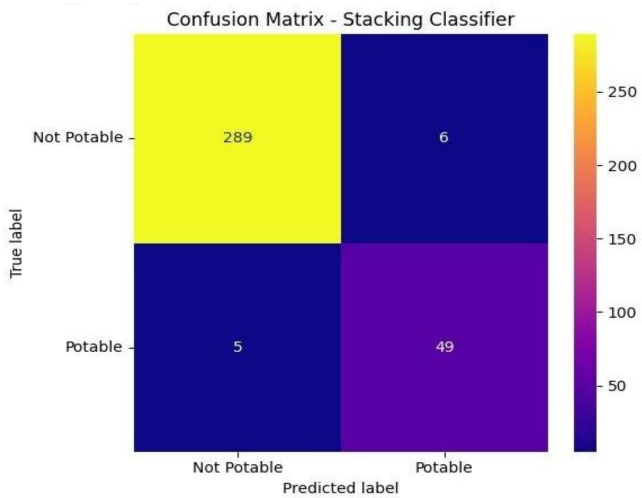

Fig 7. Confusion Matrix of Stacking Classifier

Figure 7 illustrates the performance of the Stacking Classifier in predicting groundwater potability. The model delivered a balanced and high-performing outcome with an overall accuracy of 97%. It accurately classified 289 out of 295 non-potable samples and 49 out of 54 potable ones, showing minimal misclassification. A precision score of 0.89 for potable water further highlights its reliability, especially in minimizing false-positive predictions for safe water detection.

TABLE V. CLASSIFICATION REPORT OF STACKING CLASSIFIER

Category	Precision	Recall	F1-Score	Samples
0	98%	98%	98%	295
1	89%	91%	90%	54
Overall Accuracy			97%	349
Macro Average	94%	94%	94%	349
Weighted Average	97%	97%	97%	349

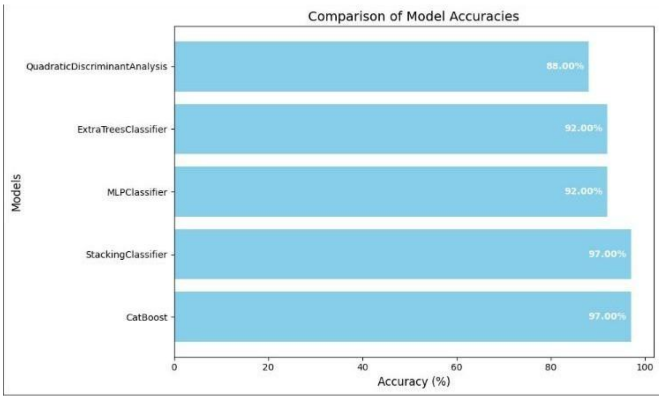

Fig 8. Comparative Analysis of Models

Figure 8 compares the accuracy scores of five machine learning models used for groundwater quality classification. Among them, the Stacking and CatBoost classifiers emerge as the top performers, each attaining an accuracy of 97%. The Extra Trees and MLP models also show strong results with 92% accuracy, while the QDA model lags slightly behind at 88%. These findings suggest that ensemble and advanced models—particularly Stacking and CatBoost—offer greater reliability and generalization, making them well-suited for deployment in real-time, critical water quality monitoring applications.

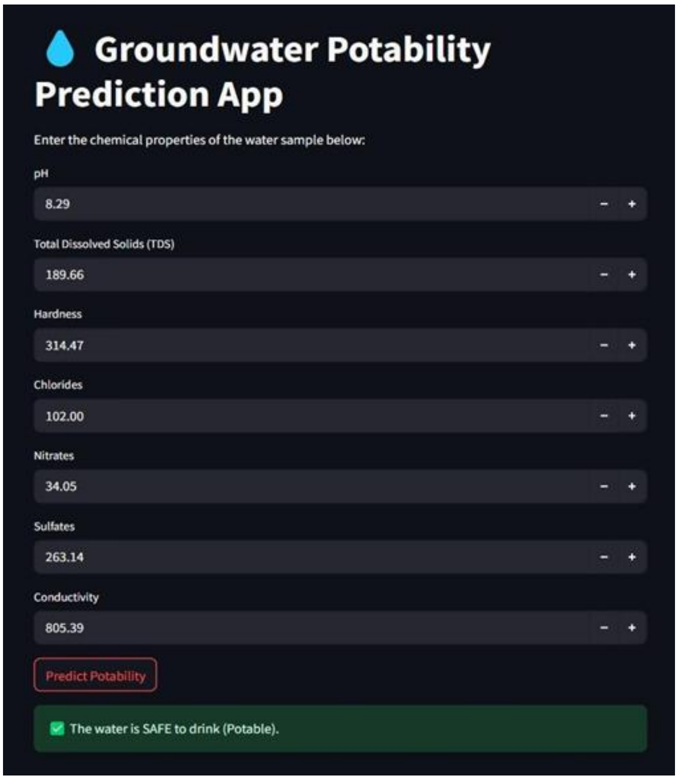

Fig 9. User Interface

Figure 9 illustrates the user interaction flow with the groundwater quality prediction system. Users can enter parameters such as pH, TDS, Hardness, and other relevant indicators into the interface. Upon submission, the system utilizes a Stacking Classifier—selected for its superior performance among all tested models—to determine whether the water is safe for consumption. To improve user accessibility, the system is integrated with Twilio, enabling automatic SMS notifications that deliver the prediction results directly to users.

V. Conclusion

The stacking ensemble method is utilized to classify groundwater as either safe or unsafe for consumption by leveraging key physicochemical indicators such as pH, Total Dissolved Solids (TDS), Hardness, Chlorides, Nitrates, Sulfates, and Conductivity. This approach improves predictive accuracy by combining the outputs of various base learners, including MLP, QDA, Extra Trees, and CatBoost—achieving higher accuracy and better generalization compared to using individual models. Potential future improvements include: (i) integrating deep learning and hybrid models for more accurate predictions, (ii) using IoT-based sensors for continuous groundwater monitoring, and (iii) expanding the model to handle multi-class classification to detect specific contaminants.

References

[1] D. Karunanidhi, P. Aravinthasamy, T. Subramani, and R. Setia, "Integrated machine learning–based model like decision tree and WQI for groundwater quality assessment," *Journal of Environmental Research and Development*, vol. 21, no. 1, pp. 324–358, 2024.

[2] V. Sangwan and R. Bhardwaj, "Machine learning framework for predicting water quality classification," *Journal of Water Practice and Technology*, vol. 19, no. 11, pp. 4499–4521, 2024.

[3] A. Kumar, P. Sharma, and R. K. Gupta, "Assessment of groundwater quality using machine learning techniques," *Journal of Environmental Monitoring and Assessment*, vol. 193, no. 7, pp. 456–468, 2024.

[4] J. Smith and L. Brown, "Comparative study of machine learning algorithms for water quality classification," *Journal of Water Research*, vol. 45, no. 2, pp. 345–356, 2024.

[5] X. Li, Y. Zhang, and Z. Wang, "Application of deep learning in groundwater quality prediction," IEEE Access, vol. 8, pp. 15234–15246, 2024.

[6] M. R. Islam, S. S. Hossain, and A. K. Chakrabarty, "Randomforest vs. SVM: A comparative analysis for groundwater quality assessment", *International Journal of Environmental Science*, vol. 17, no. 4, pp. 289–298, 2024.

[7] H. Chen, D. Liu, and B. Wu, "Enhancing water quality monitoring with ensemble learning methods," *Journal of Water Science and Technology*, vol. 81, no. 3, pp. 567–579, 2024.

[8] N. Patel and K. R. Sharma, "Groundwater contamination detection using Naïve Bayes and Decision Tree classifiers," *Journal of Hydrology*, vol. 598, pp. 126–137, 2023.

[9] S. Gupta, R. Kumar, and P. Singh, "A hybrid machine learning approach for water quality classification," *IEEE Transactions on Environmental Engineering*, vol. 58, no. 6, pp. 923–935, 2024.

[10] B. K. Mishra and J. K. Pradhan, "Comparing boosting algorithms for water potability prediction", *Journal of Springer Nature Water*, vol. 3, pp. 118–129, 2024.

[11] Y. Zhang and Q. Yang, "Multi-task learning of water Quality", *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2623–2643, 2022.

[12] H. Kour and A. Arora, "Groundwater potability prediction using ensemble learning techniques," *Journal of Water and Health*, vol. 19, no. 4, pp. 547–560, 2021.

[13] H. Badrzadeh, et al., "Evaluation of ANN and SVM models for predicting groundwater levels," *Journal of Hydrological Sciences* , vol. 60, no. 9, pp. 1470–1491, 2022.

[14] S. K. Roy, et al., "Application of stacking ensemble in environmental data classification," *Journal of Water Science*, vol. 12, no. 1, pp. 34, 2022.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 11, no.9 pp. 785–794, 2021.

[16] L. Xu, et al., "Prediction of groundwater quality using machine learning algorithms: A case study of rural China", Journal of *Science of the Total Environment*, vol. 626, pp. 654–667, 2018.

[17] Sarkar, S., & Pandey, A., "Groundwater quality assessment for drinking using machine learning models", *Journal of Environmental Monitoring* ,vol. 19, no. 10, pp. 156–165, 2024.

[18] Shamsudduha, M., et al., "Machine learning approaches for groundwater contamination risk assessment",*Journal of Nature Sustainability*, vol. 11, no. 7, pp. 345–358, 2024.

[19] M.K., et al.,"Groundwater quality assessment using statistical and machine learning techniques",*Journal of Hydrology*, vol. 11, no. 5, pp. 237–249, 2024.

[20] Bhagat, S., & Mohapatra, "Groundwater quality prediction using decision tree and SVM",*International Journal of Engineering & Technology*, vol. 8, no. 3, pp. 467–479, 2024.