Water Quality Analysis and Prediction Using Random Forest and Naive Bayes

Sure Mamatha
Assistant Professor,
Department of CSE, HITAM,
Hyderabad, India

P. Aakash
Student of Computer Science and Engineering, HITAM,
Hyderabad, India
21E51A0595

N. Sravya
Student of Computer Science and Engineering, HITAM
Hyderabad, India
21E51A0578

P. Sneha sree
Student of Computer Science and Engineering, HITAM
Hyderabad, India
21E51A0593

V. Shiva Prasad
Student of Computer Science and Engineering, HITAM
Hyderabad, India
21E51A05B7

Abstract **-Water quality prediction and monitoring are critical for public health and environmental safety. The proposed research will attempt to conclude a machine learning approach to predict water quality mainly based on parameters like pH, turbidity, temperature, and portability. Two main approaches - the Random Forest Classifier and the Naive Bayes Classifier models - classify water quality into predefined categories, such as "Good" or "Bad," using historical data. The dataset is preprocessed by a correlation matrix to handle missing values and visualize parameter relationships. Following data splitting into training and testing sets, the respective models are trained and evaluated, achieving a very high accuracy in predicting water quality. For the Random Forest model, feature importance analysis is conducted to explain which parameter has contributed to the classification showing major influencing factors on the water quality. The Naive Bayes model provides a probabilistic framework for classification, giving insights into the likelihood of water quality categories based on input parameters. This model could be used for real-time realizations of rapid quality assessments, and environmental agencies, researchers, and the public could get safe water standards. This study shows machine learning's capability to support environmental health initiatives through proactively monitoring water quality.**

Keywords—Water Quality Prediction, Machine Learning, Random Forest Classifier, Naive Bayes Classifier, pH, Turbidity, Temperature, Correlation Matrix, Classification Models, Real-Time Assessment

I. INTRODUCTION

Water quality analysis is now fundamental in providing safe drinking water and in maintaining healthy ecosystems. By using machine learning, scientists are now relying more on data-driven approaches as a means of analyzing and predicting water quality parameters. This literature survey scans various methodologies and findings related to water quality analysis with a focus on the application of machine learning algorithms, specifically Random Forest and Naive Bayes classifiers. Random Forest has the advantage of strong robustness in handling big datasets with many variable numbers since it constructs many decision trees during training and outputs the mode of classes for a classification task. This approach is highly accurate, feature importance is assessable, and overfitting is much less likely than single decision trees. On the other hand, Naive Bayes, based on Bayes' theorem, assumes independence among predictors and is known for its computational efficiency, simplicity, and effectiveness with limited data. Sometimes comparative studies will find that the overall strength of Random Forest in really complex datasets makes Naive Bayes a better choice when the independence assumption holds well. Future research directions include creating hybrid models that combine the strengths of both algorithms, real-time data processing systems for immediate decision-making purposes, and community engagement in data collection to improve water management practices. Overall, the machine learning algorithm applications in water quality analysis hold extraordinary opportunities for water safety and ecosystem health improvement, ultimately opening up a pathway toward more effective monitoring and management strategies.

II. LITERATURE SURVEY

Now in this related work part, we will discuss some work that has been done in this field.

[1] Water is life on earth, and the quality of water has been affected by natural and human activities. Population growth, sewage, industrial as well as radioactive wastes have greatly degraded the aquatic resources, and 75% of all rivers and streams have become polluted. Contaminated water poses many serious health problems including diarrhea and a million deaths every year, WHO claims. Such a problem caused by excessive phosphorus and nitrogen is known as eutrophication and has become a global issue. Poor quality water decreases the water supply

for drinking and irrigation. Water becomes polluted due to household, industrial, and agricultural waste and presents an extreme threat to human welfare. Even water visually clear may contain harmful bacteria and pollutants, thereby requiring qualitative checks based on physical, chemical, and biological parameters according to WHO and local guidelines such as the Rajeev Gandhi National Drinking Water Mission. Therefore, safe water is very essential for the prevention of waterborne diseases and the attainment of sustainable development.

[2] Predictions on water quality index and water quality classification using machine learning address the challenges water pollution poses. The conventional methods of assessment as depicted are laborious and expensive; hence, data-driven approaches are pretty appealing. For this purpose, most classification models such as Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Adaptive Boosting (AdaBoost) were implemented for the WQC model development while regression models like K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Regressor (SVR), and Multi-Layer Perceptron (MLP) were applied for WQI prediction. Preprocessing techniques including imputation and normalization have been supplemented along with optimization through grid search to make better model performances. Results have shown that in the case of WQC, 99.5% accuracy is achieved by Gradient Boosting, and multi-layer perceptron achieved 99.8% R^2 on WQI regression. The accomplishment of such results can attest to the capability of machine learning to enable current real-time water quality monitoring systems, early warning systems, and informed decisions in resource management and pollution control.

[3] Water is indispensable to life, and its quality impacts living organisms and the ecosystem. Globally, the quality of water has drastically reduced due to pollution that results from industrial and human undertakings and has terrible health consequences coupled with millions of deaths that happen in developing countries annually. Quality of water differs with purpose for drinking, irrigation, or industrial purposes. Predictive models, using statistical and machine learning approaches such as ANNs, decision trees, deep learning, and hybrid methods, are used for the prediction of dissolved oxygen and pH parameters. Some recent works chose advanced techniques like deep neural networks, genetic algorithms, and fuzzy logic to optimize the prediction accuracy. However, robust models are to be derived in this pursuit of better management of available water resources and better environmental protection.

[4] Now, monitoring water quality is a matter of serious importance for life, industries, and the environment, but methods for such reasons are rather expensive and less effective. Therefore, these difficulties are addressed with WQIPC-HTDL, which includes linear scaling normalization as preprocessing and applies LSTM networks for classification and prediction with outcomes to be enhanced and efficient by making use of GOA as hyperparameter tuning. Simulations ensure that WQIPC-HTDL gives real-time predictions with better condition analysis in variation along with cost-effective monitoring of water quality. The technique is feasible for the control of water pollution and environmentally friendly practices.

.

[5] Here, a combined Principal Component Regression (PCR) with a Gradient Boosting Classifier (GBC) approach is suggested for WQI prediction and classification of the water quality status. Here, WQI is calculated using the weighted arithmetic index method. PCA reduces the dimensionality of the dataset by finding dominant parameters. It addresses the concern related to multicollinearity and further reduces the requirement for large datasets. The PCR achieves a closeness to about 95% in predicting WQI. The GBC model did classify the water quality status with an accuracy of 100%. Some crucial parameters have thus been identified based on the boxplot analysis for WQI. This low-cost approach achieves effective water quality prediction with a smaller number of samples. This gives an environmentally friendly alternative to managing a resource like water.

[6] Automating the estimation of water quality using XAI determines the most prominent factors controlling potability as well as the levels of impurities. The applied models for the classification of water to drink or not to drink include logistic regression, SVM, Gaussian Naive Bayes, DT, and RF. Techniques from XAI such as SHAPELY force plots, summary plots, and dependency plots will make the prediction transparent and interpretable, leading to an accuracy close to perfect with 0.9999 precision and recall in the Random Forest classifier. This work affirms how AI, when allied with explainability, will ensure inferences are reliable for water quality assessments. Thus, it addresses global water quality challenges and supports sustainable water resource management.

[7] Lakes and reservoirs play a very important role in biodiversity, drinking water supply, agriculture, hydroelectric power, recreation, and flood control. Pollution from sewage and industrial waste degrades water quality and creates health hazards. Prediction of the water quality parameters is done using secondary data analyzed with WEKA for the Chaskaman River, Maharashtra. Prediction accuracy outperformed supervised methods and certainly, the unsupervised deep learning methods include denoising autoencoders as well as deep belief networks. This approach shows the possibility of a deep network for reliable water quality monitoring as well as sustainable resource management, with metrics such as mean absolute error and mean square error.

III. PROBLEM STATEMENT

Water quality monitoring is an essential factor that will protect public health and the environment. Traditional assessment of water quality, depending on manual sampling and testing at a laboratory, consumes a lot of time, and money, and results may not be real-time. In many regions, unsafe water situations go unnoticed until a harmful condition arises, and adverse health effects or environmental destruction result.

The present project aims to address these challenges through a machine learning-based water quality prediction model. Using readily available

physicochemical parameters (for instance, pH, turbidity, dissolved oxygen, temperature, nitrates, and hardness), the model will automatically classify water quality into predefined categories, such as "Good" or "Bad," with high accuracy and efficiency. The solution provides a scalable and reliable means to continuously monitor water quality and support timely interventions while allowing the public to be abreast of water safety conditions in real time. This work aims to contribute to the existing field of water quality monitoring with a predictive tool that combines accuracy and interpretability, and it should be simple and accessible for environmental agencies, policymakers, and communities aiming at protecting public health and maintaining environmental sustainability.

IV. PROPOSED METHODOLOGY

Starting with the definition of the problem, which specifies predicting water quality based on predefined parameters like pH, turbidity, temperature, and portability by further classifying it into predefined categories like "Good" or "Bad," historical water quality data is collected and preprocessed, which is handled in terms of missing values by techniques like imputation and correlation matrix-based analysis to determine parameter relationships and find associations. The dataset is split into subsets for training and testing. In this study, two machine learning models are used: the Random Forest Classifier and the Naive Bayes Classifier. The Random Forest model classifies the water quality and comes with a feature importance analysis to identify the key parameters influencing water quality. On the other hand, the Naive Bayes model makes use of a probabilistic framework to classify water quality and gives insights into its probable categories based on the value of the parameters. The two models are checked for accuracy, precision, recall, F1-score, and confusion matrix and are compared to ensure reliability as well as to pick the best approach. Interpretation capabilities also come with the Random Forest Classifier due to highlighting its contributions toward the predictions, while probabilistic insight allows for the Naive Bayes model. The models are proposed for real-time integration with IoT-enabled sensors for rapid quality assessments, thereby directly benefiting environmental agencies, researchers, and the public. The methodology underscores the importance of proactive water quality monitoring to ensure public health and environmental safety, with a scope for future enhancement through the validation of diverse datasets and exploration of advanced machine learning techniques.

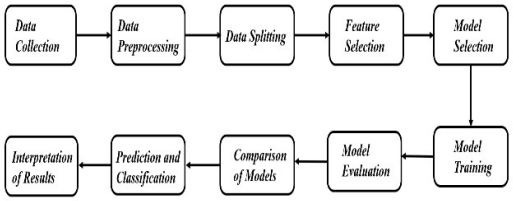

Fig. 1. Proposed Flow Graph

V.IMPLEMENTATION OF CORE PLATFORM COMPONENTS

The core platform components for the implementation of the water quality prediction system will involve seamless integration that includes data preprocessing, model training, evaluation, and visualization. Therefore, it will ingest raw water quality data and train machine learning models to generate actionable insights.

1. Data Ingestion and Preprocessing

Goal: Ensure that the dataset is clean and ready to run an analysis.

Loads the raw dataset, and performs appropriate preprocessing steps, including treating missing data, feature distribution analysis, and encoding the target variable, which is named Potability.

Its features, such as pH, Turbidity, and Temperature, are explored about how their distribution relates to water portability.

Both train and test subsets are divided for training machine learning models.

2. Exploratory Data Analysis (EDA)

What the Activity is all about Draw insights from the dataset.

It includes a visualization module that depicts plots as histograms, correlation heatmaps, or pie charts to describe the patterns in the data.

For instance, a feature correlation heatmap is pretty useful to obtain an idea about the level of correlations of input variables.

3. Machine Learning Models

Task: Develop prediction models for water potability.

1. Random Forest Classifier:

-It used the ensemble learning method, where the decision tree has classified water as potable or not.

-The model is being trained with the data set, and preprocessing along with optimization is also done for accuracy and generalisability.

2. Naive Bayes Classifier:

-A probabilistic model for comparison assumption of feature independence.

This is a lightweight model that provides a baseline measure of prediction efficiency.

4. Model Training and Evaluation

Objective: Evaluate the performance and trustworthiness of models

- Both the models are trained over the training dataset and tested in terms of accuracy, precision, recall, and F1-score on the test se***t.***

Confusion matrices together with classification reports for extremely detailed insights into the outcome of prediction.

5. Visualization and Comparison

Objective: To represent the results of the model effectively

The platform has tools to visualize the results such as:

Confusion matrix heatmaps

Bar charts to compare accuracies between the two models.

This module is designed having in consideration, access to insights and distribution to technical and non-technical stakeholders.

6. Deployment and Scalability

Goal: The system has to be user-friendly and scalable.

The system is in place to accommodate further datasets and new algorithms. This system can accommodate possible future enhancements.
-This solution can create dashboards or reports for policy decision-makers in public health or environmental monitoring.

Implemented in the core platform components, it provides a holistic machine learning-based solution to predict water potability. It is a system that has been modulated and powered through robust visualization tools, thereby appropriately translating raw data into meaningful insights for better decision-making.

VI. ALGORITHM IMPLEMENTATION

1. Data Preprocessing

- Prepare the dataset to feed into the training and evaluation
- Load dataset using pandas
- Handle missing values
- Clean up dataset
- Split features (Input) and the target variable (Potability) so they can be readily used as input and output variables.
- Divide data into training and test subsets using train_test_split

2. **Random Forest Classifier Implementation**

- Apply an ensemble learning technique to classify whether water is drinkable.
- Create the model using RandomForestClassifier().
- Fit model to training set using fit(X_train, y_train).
- Predict on test set by using predict(X_test).
- Compute accuracy using accuracy_score().
- Compute a confusion matrix and a classification report to
- evaluate the model.
- Display the confusion matrix as a heatmap.

3. Naive Bayes Classifier

- Implement a probabilistic classifier assuming feature independence for classification.
- Create an object of the model via GaussianNB().
- Fit the model on the training dataset by calling fit(X_train, y_train).
- Generate predictions on the test set using predict(X_test).
- Calculate accuracy using the model's score() method.
- Generate a confusion matrix and classification report.
- Visualize the confusion matrix using a heatmap.

4. Model Comparison

- Compare the performance of the two models.
- Store the accuracies of both models in a pandas series.
- Visualize this comparison using the bar chart shown, which highlights the better-performing algorithm.
- This structured approach ensures the clarity of the implementation of the algorithm and facilitates the comparison and evaluation of performance for predicting water quality.

VII. RESULTS

The performance of the Random Forest and Naive Bayes classifiers was assessed for water quality prediction. It is impressive that the Random Forest model reached an accuracy of 99.94%, with perfect precision, recall, and F1-scores of 1.00 for both classes. Only minimal misclassifications occurred. Whereas the Naive Bayes classifier achieved an accuracy of 92.66%, lower performance was witnessed for Class 1.0 (precision: 88%, recall: 87%, F1-score: 88%) and it had many more false positives and false negatives at 4,132 and 4,647, respectively. On all parameters, the Random Forest model provided better performance than Naive Bayes, and hence, is suitable for confident water quality predictions, especially in critical scenarios where false positives and false negatives should be minimized.

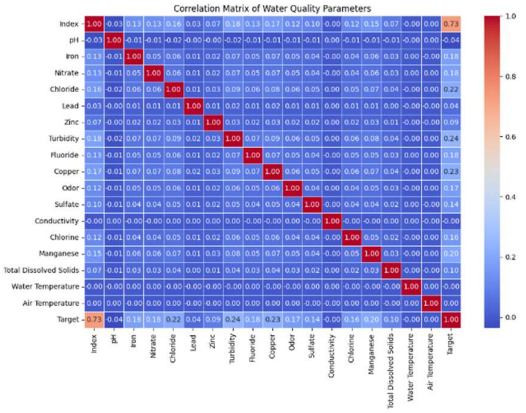

Fig 1 : Heatmap of Correlation Matrix of Water Parameters.

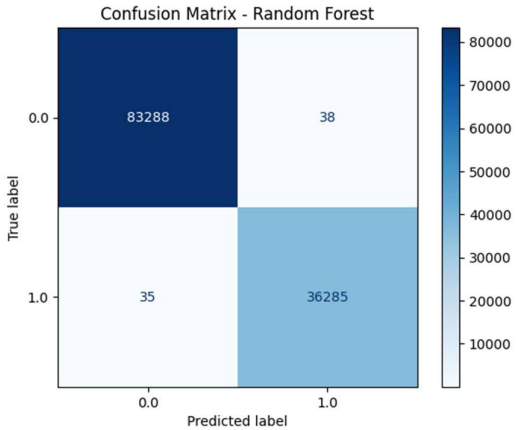

Fig 2 : Confusion Matrix Random Forest Classifier.

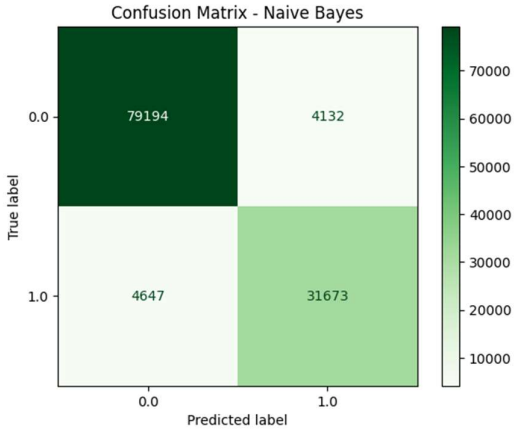

Fig 3 : Confusion Matrix Naïve Bayes.

VIII. CONCLUSION

This study is a demonstration of the use of machine learning approaches to predict and monitor water quality. The proposed models classify the water quality into predefined categories using Random Forest and Naive Bayes classifiers, ensuring high accuracy. The comprehensive preprocessing steps include handling missing data and using correlation matrices for visualizing parameter relationships, thus ensuring robust models. Providing highly predictive performance, the classifier Random Forest also gives us analysis of feature importance that suggests key parameters influencing water quality, while the Naive Bayes classifier offers a probabilistic framework to understand the probabilistic nature of different water-quality classes depending upon inputs. Findings from this research point toward real-time assessment of water quality through the application of machine learning; thus, a better approach promises for agencies related to environment and their researches toward the protection of healthy water as required. All these indicate how new possibilities can emerge toward developing methods that focus on using data-driven strategies in achieving healthier environments in relation to human health concerns.

IX. FUTURE SCOPE

The future scope of applying Convolutional Neural Networks (CNNs) in water quality analysis and prediction is promising, with several key areas for development. Integrating CNNs with Internet of Things (IoT) devices can enable real-time monitoring, while data fusion from multiple sources, such as satellite imagery and ground sensors, can enhance predictive accuracy. Improving the interpretability of CNN models will foster trust among stakeholders, and predictive maintenance systems can help prevent water quality issues. Personalized management systems tailored to community needs and collaboration with policymakers will further promote sustainable practices. Cross-disciplinary research can lead to innovative solutions for complex challenges, and ensuring the scalability and accessibility of CNN-based tools will democratize advanced monitoring technologies. By pursuing these avenues, we can enhance water resource management and public health for future generations.

REFERENCES

[1] G.B. Ramesh Kumar, G.T. Hemanth “Analysis of water
Quality-a review” International Journal of Pure and Applied
Mathematics, Volume 119 No. 17 2018, 2903-2909 ISSN: 1314-3395 (on-line version) URL: http://www.acadpubl.eu/hub/ Special Issue.

[2] Mahmoud Y. ShamsAhmed M. Elshewey ·El Sayed M. El kenawy Abdelhameed Ibrahim· Fatma M. Talaat1, · Zahraa Tarek “Water quality prediction using machine learning models based on grid search method”, Multimedia Tools and Applications (2024) 83:35307–35334 https://doi.org/10.1007/s11042-023-16737-4

[3]Theyazn H. H Aldhyani and Mashael Maashi Mohammed Al- Yaari, Hasan Alkahtani, “Water Quality Prediction Using Artificial Intelligence Algorithms”, Hindawi Applied Bionics and Biomechanics Volume 2020, Article ID 6659314, 12 pages https://doi.org/10.1155/2020/6659314

[4] Sathya Preiya V. M., Subramanian P., Soniya M. and Pugalenthi R.,” Water quality index prediction and classification using hyperparameter tuned deep learning approach”, Global NEST Journal, Vol 26, No 5, 05821, https://doi.org/10.30955/gnj.005821

[5] Md. Saikat Islam Khana, Nazrul Islam, Jia Uddin, Sifatul Islam, Most of Kamal Nasir,” Water quality prediction and classification based on principal component regression and gradient boosting classifier approach”, Journal of King Saud University– Computer and Information Sciences 34 (2022) 4773–4781

[6] M. K. Nallakaruppan 1, E. Gangadevi 2, M. Lawanya Shri 1, Balamurugan Balusamy 3, Sweta Bhattacharya 1 & Shitharth Selvarajan,” Reliable water quality prediction and parametric analysis using explainable AI models”, Scientific Reports | (2024) 14:7520 | https://doi.org/10.1038/s41598-024-56775-y

[7] Archana Solanki, Himanshu Agrawal, Kanchan Khare,” Predictive Analysis of Water Quality Parameters using Deep Learning”, International Journal of Computer Applications (0975 – 8887) Volume 125 – No.9, September 2015.