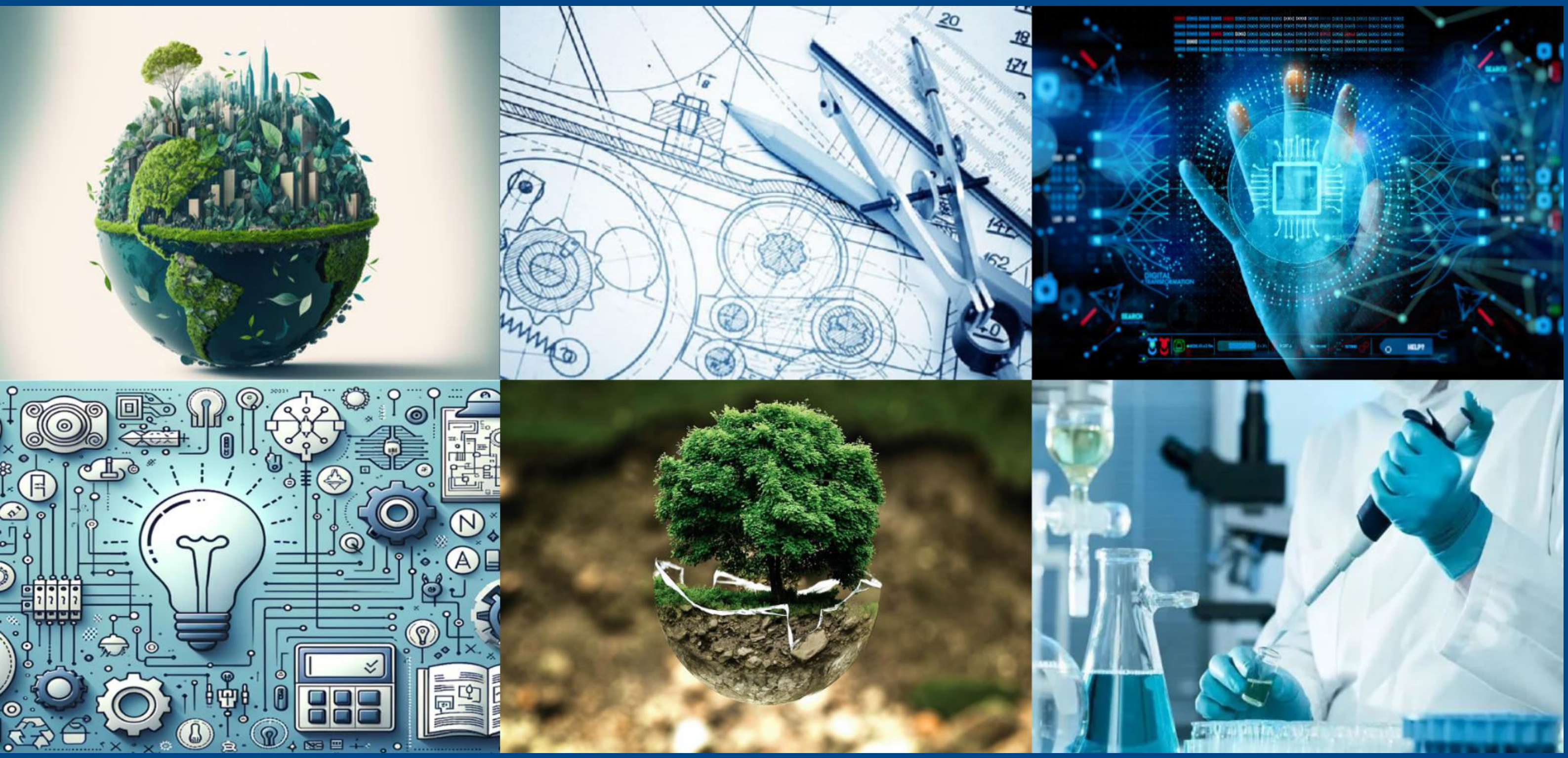
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206

Volume 8, Issue 4, April 2025

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Water Quality Analysis and Prediction Using Machine Learning Algorithms

Ajmal Sharukhan S, Prem Kumar S

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

Assistant Professor, Department of Computer Science with Data Analytics, Dr.N.G.P Arts and Science College (Autonomous), Coimbatore, India

ABSTRACT: Water quality is crucial for human health, agriculture, and industry. Traditional assessment methods are time-consuming and costly. This project leverages machine learning to automate water quality classification and prediction using parameters such as pH, Hardness, Solids, Chloramines, Sulphate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity. Supervised learning models classify water into potable (drinking), agricultural, and industrial categories. Predictive models estimate future trends, enabling proactive decision-making. Machine learning models, including Random Forest and XGBoost, optimize classification accuracy. Data preprocessing involves handling missing values, feature scaling, and data normalization, ensuring high-performance classification and prediction models.

I. INTRODUCTION

Water quality affects public health, agriculture, and industries. Traditional assessment relies on laboratory tests, which are expensive and time-consuming. This project automates classification using machine learning. The system analyzes key indicators to classify water samples and predict future contamination risks. A dashboard provides real-time classification, integrating data visualization tools like Plotly and Seaborn. Built using Python and Streamlit, the system ensures efficient water quality assessment.

OBJECTIVES

- Classify water quality into drinking, agricultural, or industrial categories using XGBoost.
- Predict future trends based on historical data for proactive resource management.
- Develop an interactive Streamlit dashboard for real-time classification and trend analysis.
- Implement pre-processing techniques to enhance accuracy.
- Provide insights to policymakers, researchers, and industries for decision-making.

II. DATASET

Kaggle: Water Quality Dataset
Includes parameters such as pH, Hardness, Solids, Chloramines, Sulphate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity. The dataset aids in water classification and risk prediction.

III. FEATURES

- **pH**: Measures acidity or alkalinity.
- **Hardness**: Indicates dissolved minerals.
- **TDS (Total Dissolved Solids)**: Represents organic/inorganic substances.
- **Chloramines**: Disinfectant preventing bacterial growth.
- **Sulfate**: Influences taste and hardness.
- **Conductivity**: Measures ion concentration.
- **Organic Carbon**: Indicates contamination presence.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- **Trihalomethanes (THMs)**: Byproduct of disinfection.
- **Turbidity**: Measures water clarity.

IV. PREPROCESSING

1. **Handling Missing Values**: Mean imputation for missing data using SimpleImputer.
2. **Feature Scaling**: StandardScaler normalizes values to improve model training.
3. **Train-Test Split**: Data is split (80% training, 20% testing) to prevent overfitting.

V. ALGORITHMS

1. **Random Forest Classifier**
   - Uses multiple decision trees for classification.
   - Reduces overfitting and improves accuracy.
   - Identifies key factors affecting water quality.

2. **XGBoost Classifier**
   - Uses gradient boosting to enhance accuracy.
   - Handles missing values efficiently.
   - Outperforms traditional models in structured data classification.

VI. FINDINGS

1. **Model Performance**: XGBoost outperformed Random Forest in accuracy.
2. **Preprocessing Impact**: Mean imputation and scaling improved model performance.
3. **Feature Importance**: pH, Hardness, and Sulfate are key classification parameters.
4. **Applications**:

   - Real-time water monitoring.
   - Government and industrial decision-making.
   - Predicting contamination risks.

VII. CHALLENGES

1. **Data Quality**: Missing/incomplete records impact prediction accuracy.
2. **Real-time Application**: High-latency models require optimization for real-time monitoring.

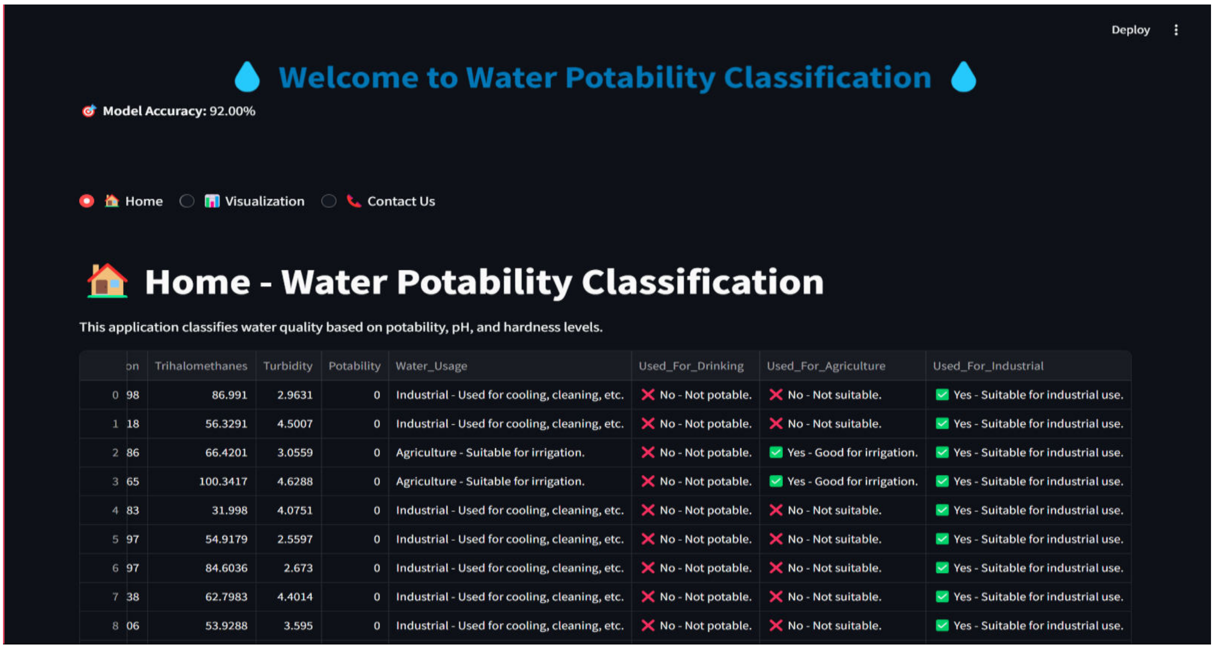
VIII. FUTURE SCOPE

1. **IoT Integration**: Real-time sensors for automated monitoring.
2. **Hybrid Models**: Combining machine learning with deep learning (ANNs, LSTMs) for improved accuracy.

This project demonstrates machine learning’s potential in water quality analysis, offering an efficient and accessible solution for public health and environmental monitoring.

SAMPLE RESULT

Fig – 1



💧 Welcome to Water Potability Classification 💧

🎯 Model Accuracy: 92.00%

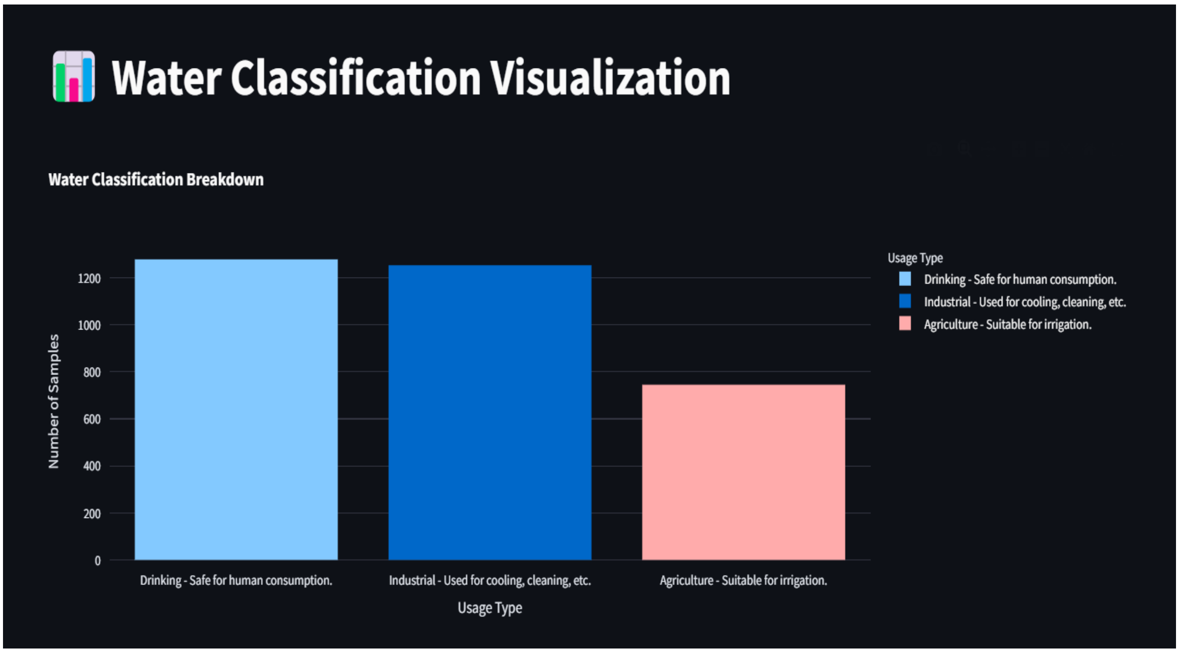
● 🏠 Home ○ 📊 Visualization ○ 📞 Contact Us

🏠 Home - Water Potability Classification

This application classifies water quality based on potability, pH, and hardness levels.

	on	Trihalomethanes	Turbidity	Potability	Water_Usage	Used_For_Drinking	Used_For_Agriculture	Used_For_Industrial
0	98	86.991	2.9631	0	Industrial - Used for cooling, cleaning, etc.	✗ No - Not potable.	✗ No - Not suitable.	✓ Yes - Suitable for industrial use.
1	18	56.3291	4.5007	0	Industrial - Used for cooling, cleaning, etc.	✗ No - Not potable.	✗ No - Not suitable.	✓ Yes - Suitable for industrial use.
2	86	66.4201	3.0559	0	Agriculture - Suitable for irrigation.	✗ No - Not potable.	✓ Yes - Good for irrigation.	✓ Yes - Suitable for industrial use.
3	65	100.3417	4.6288	0	Agriculture - Suitable for irrigation.	✗ No - Not potable.	✓ Yes - Good for irrigation.	✓ Yes - Suitable for industrial use.
4	83	31.998	4.0751	0	Industrial - Used for cooling, cleaning, etc.	✗ No - Not potable.	✗ No - Not suitable.	✓ Yes - Suitable for industrial use.
5	97	54.9179	2.5597	0	Industrial - Used for cooling, cleaning, etc.	✗ No - Not potable.	✗ No - Not suitable.	✓ Yes - Suitable for industrial use.
6	97	84.6036	2.673	0	Industrial - Used for cooling, cleaning, etc.	✗ No - Not potable.	✗ No - Not suitable.	✓ Yes - Suitable for industrial use.
7	38	62.7983	4.4014	0	Industrial - Used for cooling, cleaning, etc.	✗ No - Not potable.	✗ No - Not suitable.	✓ Yes - Suitable for industrial use.
8	06	53.9288	3.595	0	Industrial - Used for cooling, cleaning, etc.	✗ No - Not potable.	✗ No - Not suitable.	✓ Yes - Suitable for industrial use.

Fig - 2

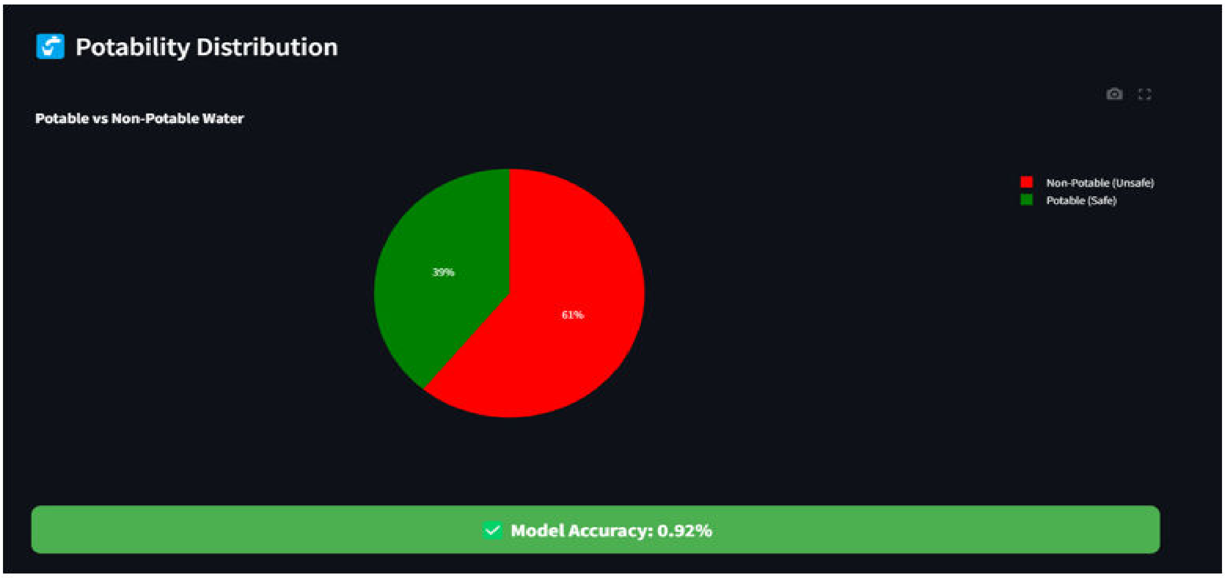


ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Fig – 3

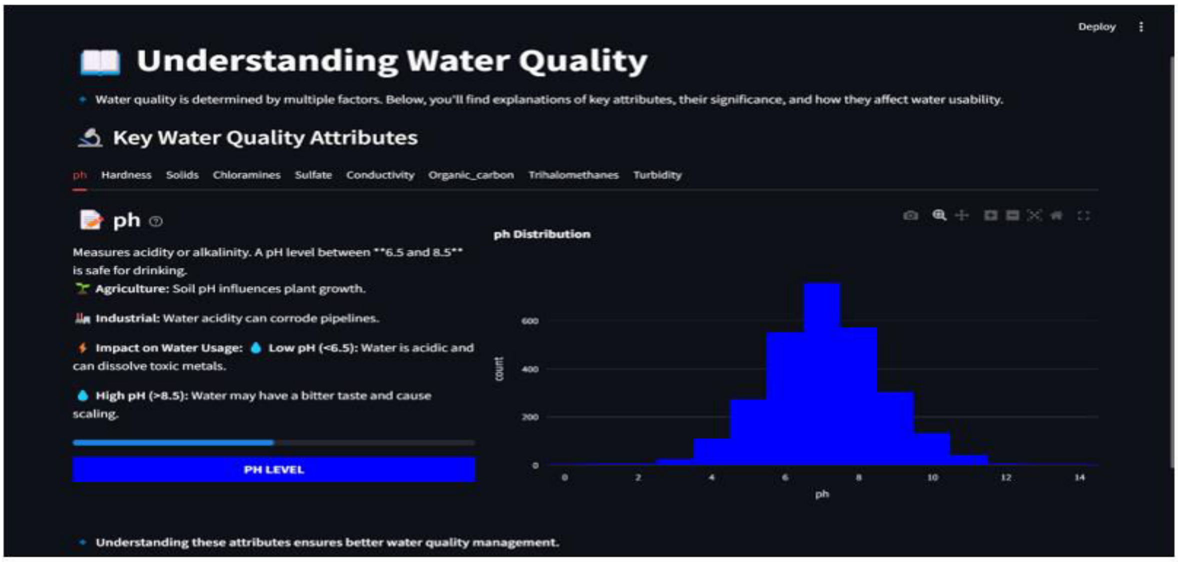


Potability Distribution

Potable vs Non-Potable Water

✓ Model Accuracy: 0.92%

Fig – 4



Understanding Water Quality

- Water quality is determined by multiple factors. Below, you'll find explanations of key attributes, their significance, and how they affect water usability.

Key Water Quality Attributes

ph

Measures acidity or alkalinity. A pH level between **6.5 and 8.5** is safe for drinking.

Agriculture: Soil pH influences plant growth.

Industrial: Water acidity can corrode pipelines.

Impact on Water Usage: 💧 **Low pH (<6.5):** Water is acidic and can dissolve toxic metals.

💧 **High pH (>8.5):** Water may have a bitter taste and cause scaling.

- Understanding these attributes ensures better water quality management.

IX. CONCLUSION

The Water Quality Analysis and Prediction System successfully classifies water samples based on key chemical and physical properties, providing insights into their suitability for drinking, agricultural, and industrial use. By leveraging machine learning models such as Random Forest and XGBoost, the system achieves high accuracy in water quality classification. The implementation of feature scaling, missing value imputation, and feature selection ensures that the models perform optimally on real-world datasets.

The findings indicate that pH, Hardness, Conductivity, and Sulfate levels are the most influential parameters affecting water potability. Random Forest and XGBoost outperform traditional models due to their ability to handle complex, non-linear relationships within the dataset. Feature importance analysis further validates that chemical parameters play a significant role in determining water quality.

The Water Quality Analysis and Prediction System provides an effective machine learning-based approach for assessing water suitability for drinking, agricultural, and industrial use. By utilizing advanced algorithms like Random Forest and XGBoost, the system ensures accurate classification of water quality based on critical chemical parameters. The integration of data preprocessing techniques, including feature scaling and missing value imputation, enhances model performance, ensuring reliable predictions.

This project contributes to public health and environmental sustainability by offering a data-driven approach to water quality assessment. The system can assist municipalities, industries, and agricultural sectors in monitoring and managing water resources efficiently. Additionally, it can serve as a decision-support tool for water treatment plants by providing predictive insights into contamination risks.

Despite its effectiveness, the system has limitations in real-time application and data availability. The accuracy of predictions heavily depends on the quality and completeness of input data. Future enhancements could include real-time data integration using IoT sensors, cloud-based monitoring, and hybrid AI models combining deep learning with machine learning techniques.

Overall, the Water Quality Analysis and Prediction System bridges the gap between data science and environmental management, ensuring that water resources are safe, sustainable, and suitable for diverse applications.

REFERENCES

1.Kaggle – Water Potability Dataset, https://www.kaggle.com/datasets/adityakadiwal/water-potability
2.WHO Guidelines for Drinking-Water Quality, Journal: World Health Organization, https://www.who.int/publications/i/item/9789241549950
3. Li, Y., et al. (2020), **Environmental Modelling & Software – Water Quality Prediction Models**
4. Kumar, R., et al. (2021), **Real-Time Water Quality Monitoring Using IoT and Machine Learning**

INNO SPACE
SJIF Scientific Journal Impact Factor

ISSN
INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com