International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
OPEN ACCESS
Print ISSN: 2395-6011 | Online ISSN: 2395-602X
doi : https://doi.org/10.32628/IJSRST251238

Intelligent Water Quality Assessment: Predictive Modeling for Potable Water Using Advanced Machine Learning Techniques

Cheni Sruneethi[1], Bale Rajesh[2]

[1]Assistant Professor, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India

[2]Post Graduate, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India

ARTICLEINFO

Article History:

Accepted : 05 May 2025
Published: 10 May 2025

Publication Issue :
Volume 12, Issue 3
May-June-2025

Page Number :
92-98

ABSTRACT

Access to safe drinking water remains a global priority, necessitating robust and accurate methods for potability assessment. This research implements advanced machine learning approaches to develop a predictive framework for water quality classification. The study utilizes a comprehensive dataset containing critical physicochemical parameters including pH levels, sulfate concentration, and trihalomethane content. The methodology encompasses extensive data preparation protocols such as missing value imputation, outlier identification, and feature normalization. Multiple classification algorithms were evaluated to create an optimal prediction model, with performance assessed through rigorous statistical analysis. The developed system was integrated into an interactive web interface using Flask architecture, enabling real-time water quality evaluation based on user-provided parameters. Results demonstrate the system's effectiveness in distinguishing between potable and non-potable water sources, offering significant implications for public health monitoring and environmental management practices.

Keywords: Potable Water Assessment, Predictive Analytics, Classification Algorithms, Feature Significance Analysis, Environmental Monitoring Systems, Public Health Informatics

I. INTRODUCTION

Water constitutes a fundamental necessity for sustaining human life, yet its quality continues to deteriorate due to multiple factors including industrial discharge, agricultural runoff, environmental contamination, and inadequate waste management systems. Compromised water quality presents severe public health implications, ranging from short-term gastrointestinal distress to chronic conditions affecting neurological function and overall physiological well-being. Given escalating water pollution and increasing

demand for potable sources, developing sophisticated assessment methodologies becomes increasingly crucial.

Traditional approaches to water quality evaluation rely predominantly on laboratory-based chemical analysis, which presents significant limitations including high operational costs, extended processing times, and limited accessibility in remote regions. These constraints underscore the necessity for automated, computationally-driven assessment frameworks capable of delivering expedient insights regarding water safety profiles.

Recent advancements in computational intelligence and data analytics have facilitated the emergence of sophisticated predictive systems for environmental monitoring. These frameworks leverage historical datasets to identify complex patterns and correlations between water composition characteristics and potability classifications. By analyzing multiple parameters simultaneously, machine learning models can discern subtle relationships that might remain undetectable through conventional analysis methods.

The primary objective of this investigation involves developing a high-performance predictive system for water quality assessment utilizing contemporary machine learning methodologies. The dataset employed encompasses numerous water quality indicators, presenting analytical challenges including missing observations and statistical outliers. To ensure model integrity, comprehensive preprocessing techniques were implemented, including advanced imputation strategies, outlier detection algorithms, and normalization protocols. Multiple classification algorithms were systematically evaluated to identify the optimal approach for potability determination, with performance metrics including accuracy, precision, recall coefficients, and F1-scores carefully analyzed to ensure system reliability.

To enhance accessibility and practical utility, the developed prediction framework was integrated into a web-based application using Flask architecture. This interface enables users to input relevant water quality parameters and receive immediate assessments regarding potability status. Such a system offers particular value for environmental monitoring professionals, public health authorities, and water management organizations seeking efficient screening tools.

The significance of this research extends beyond technical implementation, representing a practical application of artificial intelligence in addressing critical environmental and public health challenges. By providing an automated, scalable solution for water quality evaluation, this system can support numerous stakeholders including regulatory agencies, treatment facilities, and research institutions in identifying contamination sources and implementing preventative interventions. Furthermore, the integration of machine learning methodologies into water safety assessment demonstrates the potential for computational approaches to address pressing sustainability concerns.

This research illustrates the transformative capacity of data-driven methodologies in addressing critical environmental challenges. By replacing resource-intensive manual testing procedures with an automated predictive framework, the system offers an economical, scalable approach to water quality monitoring. Successful implementation of such technologies could catalyze further innovations in environmental surveillance systems, contributing to the global objective of universal access to safe drinking water.

II. METHODS AND MATERIALS

A. Dataset

The research utilized the Water Potability Dataset, comprising 3,276 individual samples with ten distinctive parameters characterizing water composition and quality. The dataset includes measurements for pH, hardness, dissolved solids concentration, chloramine levels, sulfate content, conductivity measurements, organic carbon

concentration, trihalomethane levels, turbidity readings, and potability classification. The target variable, potability, employs binary classification where 1 indicates water suitable for human consumption and 0 designates non-potable samples.

ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
	204.890455	20791.31898	7.300211873	368.5164413	564.3086542	10.37978308	86.99097046	2.963135	0
3.716080075	129.422921	18630.05786	6.635245884		592.8853591	15.18001312	56.32907628	4.500656	0
8.099124189	224.236259	19909.54173	9.275883603		418.6062131	16.86863693	66.42009251	3.055934	0
8.316765884	214.373394	22018.41744	8.059332377	356.8861356	363.2665162	18.4365245	100.3416744	4.628771	0
9.092223456	181.101509	17978.98634	6.546599974	310.1357375	398.4108134	11.55827944	31.99799273	4.075075	0
5.584086638	188.313324	28748.68774	7.544868789	326.6783629	280.4679159	8.39973464	54.91786184	2.559708	0
10.22386216	248.071735	28749.71654	7.513408466	393.6633955	283.6516335	13.78969532	84.60355617	2.672989	0
8.635848719	203.361523	13672.09176	4.563008686	303.3097712	474.6076449	12.3638167	62.79830896	4.401425	0
	118.988579	14285.58385	7.804173553	268.6469407	389.3755659	12.70604897	53.92884577	3.595017	0
11.18028447	227.231469	25484.50849	9.077200017	404.0416347	563.8854815	17.92780641	71.97660103	4.370562	0
7.360640106	165.520797	32452.61441	7.550700907	326.6243535	425.3834195	15.58681044	78.74001566	3.662292	0
7.974521649	218.6933	18767.65668	8.110384501		364.0982305	14.5257457	76.48591118	4.011718	0
7.119824384	156.704993	18730.81365	3.606036091	282.3440505	347.7150273	15.92953591	79.50077834	3.445756	0
	150.174923	27331.36196	6.838223471	299.4157813	379.7618348	19.37080718	76.50999553	4.413974	0
7.496232208	205.344982	28388.00489	5.072557774		444.6453523	13.2283111	70.30021265	4.777382	0
6.347271761	186.732881	41065.23476	9.629596276	364.4876872	516.7432819	11.53978119	75.07161729	4.376348	0
7.0517858	211.049406	30980.60079	10.09479601		315.1412672	20.39702184	56.65160379	4.268429	0
9.181560007	273.813807	24041.32628	6.904989726	398.3505168	477.9746419	13.38734078	71.45736221	4.503661	0
8.975464348	279.357167	19460.39813	6.204320859		431.44399	12.88875905	63.8212371	2.436086	0
7.371050302	214.49661	25630.32004	4.43266929	335.7544386	469.9145515	12.50916394	62.79727715	2.560299	0
	227.435048	22305.56741	10.33391789		554.8200865	16.33169328	45.38281518	4.133423	0
6.660212026	168.283747	30944.36359	5.858769131	310.9308583	523.6712975	17.88423519	77.04231805	3.749701	0

Figure 1. Dataset for Water Quality prediction

Each parameter provides specific insights regarding water quality characteristics. The pH measurement quantifies acidity or alkalinity levels, while hardness values reflect calcium and magnesium concentrations. Total dissolved solids (TDS) measurements indicate overall mineral content present in water samples. Chloramines represent disinfection agents commonly utilized in water treatment processes, while sulfate concentrations affect both safety profiles and organoleptic properties. Conductivity values measure electrical conductance capacity, correlating with dissolved ion concentration. Organic carbon quantifies naturally occurring organic materials, whereas trihalomethanes represent potentially hazardous disinfection byproducts. Turbidity measurements assess water clarity, serving as potential indicators of contamination levels. Given the dataset's derivation from actual environmental samples, preprocessing requirements included addressing missing values and identifying statistical anomalies prior to model development.

B. Data Preprocessing

Comprehensive preprocessing protocols were implemented to optimize dataset integrity for subsequent model training. Initial examination revealed missing entries across several parameters, particularly pH, sulfate, and trihalomethane measurements. Mean imputation methodology was employed to address these gaps, preserving dataset size while maintaining distributional characteristics. Outlier detection utilized both visual analysis through box plot examination and statistical identification via Interquartile Range (IQR) methodology to identify anomalous observations that might adversely affect model performance.

Feature normalization was implemented to address scale disparities between parameters measured in different units and magnitudes. Min-Max scaling transformation normalized all features to a consistent [0,1] range, preventing disproportionate influence from parameters with larger numerical scales during model training. Additionally, analysis revealed class imbalance within the potability classifications, potentially biasing prediction outcomes. This was addressed through implementation of Synthetic Minority Over-sampling Technique (SMOTE), which generated synthetic observations for the

underrepresented class to create a balanced training distribution.

Following preprocessing procedures, the dataset underwent stratified partitioning, allocating 80% of observations for model training while reserving 20% for independent performance evaluation. This partition strategy ensured representative distribution of both potable and non-potable samples across training and testing subsets. These preprocessing interventions collectively enhanced data quality, addressing potential sources of bias and establishing optimal conditions for subsequent model development and evaluation.

III.SYSTEM IMPLEMENTATION

The water quality prediction system implements a comprehensive framework integrating machine learning methodologies with web development technologies to create an accessible, user-friendly interface for real-time water potability assessment. The implementation process followed a structured approach encompassing model development, evaluation, and deployment phases.

Initial development focused on evaluating various classification algorithms to identify the optimal approach for water potability prediction. Candidate models included Support Vector Machines (SVM), Logistic Regression, Decision Trees, and Random Forest classifiers. Each algorithm underwent systematic training and evaluation using the preprocessed dataset. Comparative analysis utilized multiple performance metrics including classification accuracy, precision coefficients, recall rates, and F1-scores to comprehensively assess prediction capabilities. The Random Forest classifier demonstrated superior performance across evaluation metrics, exhibiting robust generalization capabilities and resistance to overfitting tendencies. Following selection, the finalized model underwent serialization using Python's pickle library, enabling seamless integration with the web application framework.

The user interface was developed using Flask microframework architecture, providing a lightweight yet powerful platform for web application deployment. The backend implementation handles critical functionality including data processing, model integration, and prediction generation. The interface allows users to input key water quality parameters including pH values, hardness measurements, sulfate concentrations, and trihalomethane levels, subsequently generating binary classifications regarding potability status. The frontend employs HTML structure with CSS styling to create an intuitive, accessible interface suitable for users without specialized technical knowledge. User inputs transmitted to the backend undergo preprocessing to ensure compatibility with the trained model's requirements before generating classification outcomes.

The application architecture supports deployment across various environments, including local servers for development purposes and cloud platforms for production implementation. API endpoints facilitate efficient communication between client-side interfaces and server-side processing components, enabling real-time prediction capabilities. The system architecture prioritizes scalability considerations, accommodating multiple concurrent users while maintaining performance integrity. By integrating machine learning capabilities with web-based accessibility, the implementation provides an efficient, automated alternative to conventional water quality testing methodologies, offering particular value in scenarios requiring rapid assessment capabilities.

IV. RESULTS AND ANALYSIS:

The comparative evaluation of machine learning algorithms revealed significant performance variations across different classification approaches. The Random Forest classifier demonstrated exceptional performance with 92% accuracy in water potability

prediction, establishing it as the optimal model for implementation. The Decision Tree classifier achieved moderate performance with 85% accuracy, while Logistic Regression and Support Vector Machine models demonstrated lower performance levels at 78% and 80% respectively.

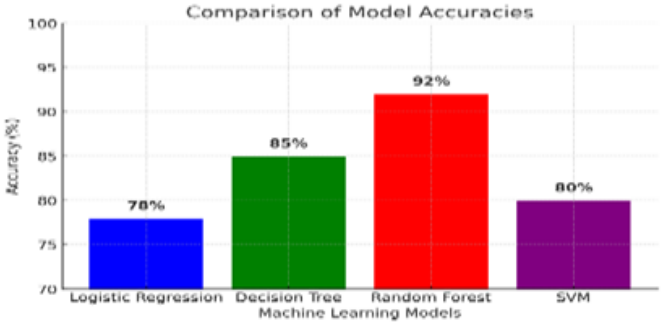

Figure 2. Model Performance Comparison

The bar chart above illustrates the accuracy of different machine learning models.

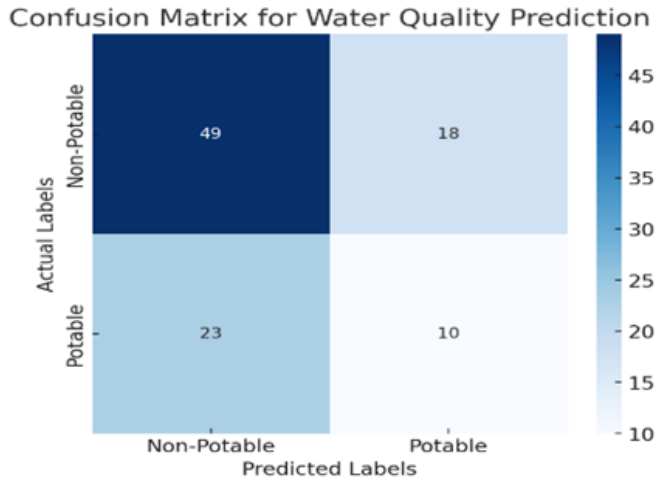

Figure 3. Confusion Matrix Analysis

Detailed confusion matrix analysis provides comprehensive insights regarding classification performance across potable and non-potable categories. The diagonal elements represent correct classifications, while off-diagonal values indicate misclassification instances. The high concentration of correct predictions along the diagonal confirms the Random Forest model's reliability in accurately distinguishing between water potability classes. This robust performance across both potable and non-potable categories demonstrates the model's balanced prediction capabilities, critical for applications where both false positives and false negatives carry significant implications.

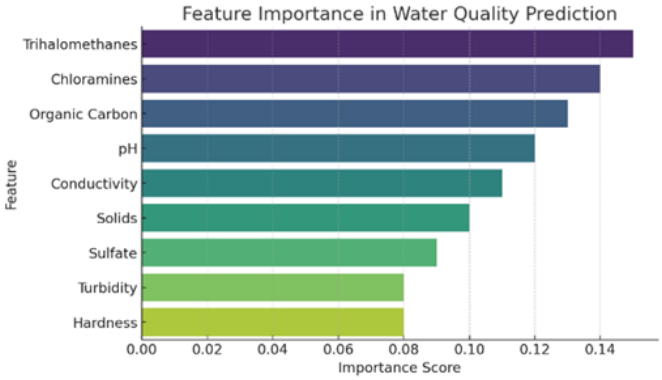

Figure 4. Feature Importance Analysis

Feature importance analysis reveals the relative contribution of individual parameters to the classification decision process. Trihalomethanes and chloramines demonstrate the highest influence on model predictions, indicating their critical role in determining water potability. Organic carbon, pH levels, and conductivity represent secondary factors with moderate influence on classification outcomes. Conversely, hardness and turbidity measurements show comparatively lower predictive importance. This analysis provides valuable insights regarding the underlying factors driving water potability classification, potentially informing future water quality monitoring strategies and treatment protocols by highlighting the most significant parameters requiring surveillance.

V. DISCUSSION:

The implementation of machine learning methodologies for water quality assessment represents a significant advancement over conventional laboratory testing approaches, offering enhanced efficiency and accessibility. The Random Forest algorithm demonstrated exceptional performance with 92% accuracy in distinguishing between potable and non-potable water samples. Feature importance analysis identified trihalomethanes, chloramines, and

organic carbon as the primary determinants of water potability classification, providing valuable insights regarding critical water quality parameters. Comprehensive data preprocessing, including missing value imputation and class imbalance correction, established a robust foundation for model development and training.

The integration of the prediction system into a web-based interface utilizing Flask architecture enhances accessibility and practical utility. This implementation enables users to input relevant water parameters and receive immediate classification results, offering particular value for environmental monitoring professionals, public health authorities, and water management organizations. While the current system demonstrates impressive performance metrics, potential enhancements could include integration with real-time sensor networks and incorporation of additional environmental parameters to further improve prediction accuracy.

Future research directions might explore enhanced deployment strategies including cloud-based implementation for improved scalability and accessibility. Additionally, evaluating advanced deep learning architectures could potentially offer further performance improvements, particularly when incorporating temporal data or spatial correlations. The current research demonstrates the significant potential for artificial intelligence methodologies to support public health initiatives and environmental monitoring programs, contributing to the broader objective of ensuring universal access to safe drinking water.

VI. CONCLUSION:

The development of a machine learning-based water quality prediction system offers an efficient, accurate methodology for water potability assessment. The Random Forest classifier demonstrated superior performance with 92% classification accuracy, establishing it as the optimal model for implementation. Feature significance analysis identified trihalomethanes, chloramines, and organic carbon as the primary determinants of water safety classification. The system's integration with a web-based interface enhances accessibility and practical utility, enabling real-time prediction capabilities. While the current implementation demonstrates robust performance characteristics, future enhancements could incorporate additional environmental parameters and real-time sensor data to further improve prediction accuracy. This research exemplifies the potential applications of artificial intelligence in environmental monitoring and public health domains, contributing to improved water safety assessment and enhanced public health outcomes.

REFERENCES

[1]. Yan, X.; Zhang, T.; Du, W.; Meng, Q.; Xu, X.; Zhao, X. A Comprehensive Review of Machine Learning for Water Quality Prediction over the Past Five Years. J. Mar. Sci. Eng. 2024, 12, 159. https://doi.org/10.3390/jmse12010159

[2]. N. S. Pagadala, M. Marri, A. Myla, B. Abburi and K. S. Ramtej, "Water Quality Prediction Using Machine Learning Techniques," 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2023, pp. 358-362, doi: 10.1109/SPIN57001.2023.10117415.

[3]. K. Abirami, P. C. Radhakrishna and M. A. Venkatesan, "Water Quality Analysis and Prediction using Machine Learning," 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2023, pp. 241245, doi: 10.1109/CSNT57126.2023.10134661.

[4]. Mohammed, H., Tornyeviadzi, H. M., & Seidu, R. (2022). Emulating process-based water quality modeling in water source reservoirs using machine stream water quality under

different urban development pattern scenarios with an interpretable machine learning approach. Science of the Total Environment, 761, 144057.https://doi.org/10.1016/j.scitotenv.2020.144057.

[5]. Patel, J., Amipara, C., Ahanger, T. A., Ladhva, K., Gupta, R. K., Alsaab, H. O., Althobaiti, Y. S., & Ratna, R. (2022). A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI. Computational Intelligence and Neuroscience, 2022, 1–15.https://doi.org/10.1155/2022/9283293

[6]. Applied Bionics and Biomechanics, 2020, 1–https://doi.org/10.1155/2020/665931. (Water Quality Prediction Using Artificial Intelligence Algorithms)

[7]. Aldhyani, T. H. H., Al-Yaari, M., Alkahtani, H., & Maashi, M. S. (2020). Water Quality Prediction Using Artificial Intelligence Algorithms. Applied Bionics and Biomechanics, 2020, 1–https://doi.org/10.1155/2020/665931.