Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Water Quality Analysis and Prediction Using Machine Learning

Snehal Vijay Patil [1], Prof. Dr. Nilesh R. Wankhade [2], Dr. S. B. Bagal [3], Mohan Tukaram Patel [4]
[1] *Department of Computer Engineering, K.C.T Late G.N.Sapkal College of Engineering, Nashik, India. Email: patilsnehal9199@gmail.com*
[2] *HOD of Department of Computer Engineering, K.C.T Late G.N.Sapkal College of Engineering, Nashik, India. Email: Nileshrw_2000@yahoo.com*
[3] *Principal, Late G.N.Sapkal College of Engineering, Nashik, India*
[4] *Assistance Professor Department of Electrical Engineering K.C.T Late G.N.Sapkal College of Engineering, Nashik, India. Email: mohantp1988@gmail.com*

ARTICLE INFO

Received:18 Dec 2024

Revised: 19 Feb 2025

Accepted:28 Feb 2025

ABSTRACT

Introduction: Access to clean and safe drinking water is a fundamental human necessity and a growing global concern. With increasing industrialization and urbanization, water sources are becoming more susceptible to contamination, making it essential to monitor and assess water quality efficiently. This project focuses on the development of a predictive system using machine learning algorithms to determine the potability of water based on various physical and chemical parameters such as pH, hardness, chloramines, sulfate, and more. By leveraging advanced data science techniques, the model classifies water as either potable or non-potable, providing a data-driven approach to support public health and environmental safety. The system not only enhances decision-making for water management authorities but also empowers communities with insights into the quality of their water supply.

Objectives: To develop a machine learning-based system capable of accurately predicting the potability of water using various physicochemical parameters. To analyze key water quality indicators such as pH, hardness, chloramines, sulfate, and trihalomethanes for identifying their influence on potability.To evaluate and compare the performance of different classification algorithms, including SVM, Random Forest, KNN, and Logistic Regression, for water quality prediction. To design a user-friendly web application that allows users to input water sample values or regional names and receive real-time potability analysis.

Methods: The methodology adopted in this project follows a structured data science workflow aimed at developing an accurate and efficient water potability prediction model. Initially, the dataset was collected from a publicly available source containing essential water quality parameters such as pH, hardness, chloramines, sulfate, and trihalomethanes. Preprocessing steps were performed to address missing values using statistical imputation techniques and to handle outliers that could skew the model's performance. Normalization was applied to bring all feature values within a consistent scale, ensuring improved algorithm convergence. Following this, an exploratory data analysis (EDA) was conducted to gain deeper insights into the dataset through statistical summaries, distribution histograms, correlation heatmaps, and skewness assessments. Multiple machine learning algorithms—including Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Naive Bayes—were implemented to evaluate classification performance. The models were assessed using key evaluation metrics such as accuracy, precision, recall, and F1-score. The best-performing model was serialized and integrated into a web-based application using the Flask framework. Additionally, an AI-powered module was developed to allow users to enter a city or region name and receive detailed water quality analysis based on current or simulated environmental data.

Results: The implementation and evaluation of multiple machine learning models revealed varying levels of accuracy in predicting water potability. Among all the algorithms tested, the Support Vector Machine (SVM) classifier demonstrated the most promising performance,

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

achieving a balanced trade-off between precision and recall. With an overall accuracy of 64%, the SVM model effectively classified both potable and non-potable water samples. The classification report highlighted a precision of 0.71 for non-potable water and 0.56 for potable water, indicating a reasonably good differentiation between classes despite class imbalance in the dataset. These results underscore the potential of SVM in real-world applications, providing reliable predictions that can aid in water quality assessment and public health safety.

Conclusions: The Water Potability Prediction project demonstrates the practical application of machine learning techniques in addressing a critical public health issue—ensuring access to safe drinking water. By analyzing key water quality parameters and evaluating multiple classification algorithms, the project successfully identifies patterns and indicators that influence water potability. The Support Vector Machine model emerged as the most effective in terms of predictive accuracy and consistency, highlighting its suitability for real-world deployment. Furthermore, the integration of a user-friendly web interface and an AI-based regional analysis module enhances accessibility and usability for a wider audience. This project not only contributes to smarter environmental monitoring but also serves as a stepping stone toward data-driven solutions for sustainable water management and community well-being.

Keywords: Water Potability, Machine Learning, Support Vector Machine (SVM), Water Quality Prediction, Data Science, Classification Algorithms, Flask Web Application, Regional Water Analysis, Public Health, Environmental Monitoring.

INTRODUCTION

Access to clean and safe drinking water is one of the most critical necessities for sustaining life and promoting public health. However, due to rapid industrialization, urban development, and environmental degradation, water pollution has become a pressing global issue, affecting millions of people and posing serious health hazards. Contaminated water is responsible for numerous diseases such as cholera, dysentery, and typhoid, particularly in underdeveloped and developing regions where water quality monitoring infrastructure is limited or outdated. Thus, there is an urgent need for intelligent systems that can assess and predict water potability efficiently and accurately.

This project, titled **Water Potability Prediction**, aims to address this challenge by leveraging the power of machine learning and data science to develop a system capable of classifying water samples as potable (safe for drinking) or non-potable. The predictive model is based on key physicochemical features such as pH, hardness, chloramines, sulfate, and trihalomethanes. These parameters are critical indicators of water quality and are used by environmental and health agencies worldwide to determine whether water is suitable for human consumption. By feeding this data into supervised learning algorithms, the model is trained to learn complex patterns and correlations that indicate potability.

The methodology adopted includes extensive data preprocessing, handling missing values, outlier detection, and normalization to ensure high-quality input for model training. Various classification algorithms including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naive Bayes, and XGBoost were implemented and compared based on performance metrics such as accuracy, precision, recall, and F1-score. Among these, SVM was found to be the most effective and consistent, offering reliable predictions even in the presence of imbalanced datasets.

Furthermore, the project extends beyond traditional predictive modeling by integrating a web-based user interface developed using Flask. This interface allows users to input water sample data and instantly receive a prediction regarding its potability. Additionally, an innovative module has been introduced where users can input a **city or region name**, and the system utilizes AI to provide a detailed water quality analysis specific to that location. This feature enhances real-world applicability by making the system accessible and informative to both individuals and institutions.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

By bridging the gap between environmental data and intelligent decision-making, this project not only contributes to the field of water quality monitoring but also demonstrates the broader potential of machine learning in addressing global health and sustainability challenges. As climate change and human activities continue to impact water sources, predictive systems like the one developed here can play a pivotal role in ensuring water safety, raising awareness, and informing policy decisions.

OBJECTIVES

The primary objective of this project is to develop a machine learning-based predictive system capable of accurately determining the potability of water based on key chemical and physical parameters. The aim is to support public health initiatives and environmental safety through intelligent data-driven decision-making. By applying supervised learning algorithms, the project intends to identify patterns and correlations that signify whether a water sample is safe for human consumption.

Additionally, the project seeks to compare the performance of various machine learning classifiers—such as Support Vector Machines (SVM), Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Naive Bayes—in terms of their accuracy, precision, recall, and robustness in handling imbalanced data. Another important objective is to build a user-friendly web application using Flask, enabling users to input water quality parameters and instantly receive potability predictions.

To further enhance real-world applicability, the system also integrates an AI-powered module that allows users to input the name of a city or region to receive an automated, intelligent analysis of water quality for that specific location. This feature is designed to extend the functionality of the application from individual sample testing to broader geographical water quality insights, supporting both users and policymakers.

In essence, this project aspires to contribute a scalable, accessible, and effective solution to the ongoing global challenge of water safety, combining technological innovation with practical usability.

The methodology adopted in this project follows a comprehensive data science pipeline to build an accurate water potability prediction system. The process begins with **data collection**, where the dataset used was sourced from publicly available repositories containing water quality attributes such as pH, hardness, chloramines, sulfate, solids, and trihalomethanes. These parameters are recognized globally for their relevance in determining the potability of water.

In the **data preprocessing** stage, missing values were identified—particularly in parameters like pH, sulfate, and trihalomethanes—and handled using median imputation to maintain data consistency. Outlier detection techniques were applied to improve the reliability of the training data. Features were then normalized to bring all values into a similar scale, which is crucial for improving the performance of distance-based algorithms such as K-Nearest Neighbors.

Following this, **exploratory data analysis (EDA)** was conducted using statistical summaries, histograms, correlation matrices, and heatmaps to gain insights into the distribution and relationships between variables. This analysis helped guide feature selection and model design.

Multiple **machine learning algorithms** were implemented to classify water samples into potable and non-potable categories. These include Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and XGBoost. Each model was trained and tested using a stratified split, and their performance was evaluated based on metrics such as accuracy, precision, recall, and F1-score.

Among the models, the SVM algorithm yielded the highest predictive accuracy, making it the most suitable for deployment. The best-performing model was **serialized** using the joblib library for deployment purposes.

The final step involved **developing a Flask-based web application**. This user-friendly interface allows users to input water quality parameters and receive real-time predictions about water potability. Moreover, an **AI-based module** was incorporated, enabling users to enter the name of a city or region to receive an intelligent summary of water quality indicators for that location, based on AI-generated insights or region-specific datasets.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This multi-step methodology ensures a robust and scalable solution for real-world water quality assessment and public awareness.

RESULTS

The experimental phase involved the training and evaluation of several machine learning models to determine the most effective approach for predicting water potability. Each model—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and XGBoost—was assessed based on performance metrics such as accuracy, precision, recall, and F1-score. Among these, the Support Vector Machine (SVM) model demonstrated the most balanced and consistent performance.

The SVM classifier achieved an overall **accuracy of 64%**, with a **precision of 0.71** for non-potable water and **0.56** for potable water. Despite the presence of class imbalance in the dataset, SVM effectively identified both classes and maintained a reliable balance between false positives and false negatives. The classification report indicated that SVM provided a good trade-off between sensitivity and specificity, which is crucial in public health-related applications.

The results also highlighted that while other models like Random Forest and KNN performed reasonably well, they showed minor variations in recall or precision, making them slightly less suitable for deployment in a real-world setting where consistent accuracy is vital. The evaluation confirmed that data preprocessing steps, such as handling missing values and normalizing inputs, significantly improved model performance.

Overall, the results validate the applicability of machine learning in water quality assessment and demonstrate that with further optimization, such models can contribute to accessible and reliable water safety monitoring solutions.

DISCUSSION

The results obtained from this study underscore the significant potential of machine learning in environmental monitoring, particularly for water quality assessment. Among the various algorithms tested, the Support Vector Machine (SVM) classifier demonstrated superior performance in terms of precision, recall, and overall accuracy, making it a strong candidate for real-world deployment. Its ability to manage complex patterns in the dataset, despite inherent class imbalance, reflects the model's robustness and adaptability.

One of the critical aspects that contributed to model performance was data preprocessing. Handling missing values and outliers, along with normalization of feature scales, played a vital role in improving prediction accuracy. These steps ensured that the model was trained on clean, standardized data, leading to more reliable results. Additionally, the inclusion of various algorithms in the comparative study allowed for a comprehensive evaluation, highlighting the trade-offs between model simplicity, interpretability, and performance.

While the achieved accuracy of 64% may not appear exceptionally high in an ideal setting, it is important to recognize the limitations posed by the dataset, such as imbalanced class distribution and limited sample size. These factors inevitably impact the model's ability to generalize. Future improvements, such as expanding the dataset, incorporating real-time water quality data from diverse regions, and employing advanced ensemble techniques, could significantly enhance performance.

The integration of an AI-based regional water analysis module further broadens the application's impact, providing localized insights based on city or region input. This feature bridges the gap between predictive modeling and public usability, offering not just binary classifications but also informative context, which is vital for awareness and preventive action in water safety.

In conclusion, this project demonstrates that machine learning, when combined with thoughtful data preparation and user-centric design, can serve as an effective tool in addressing pressing public health issues like water potability. The discussion highlights both the strengths and areas for improvement, laying the foundation for future research and practical applications.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

REFRENCES

[1] U.S. Environmental Protection Agency (EPA), "Drinking Water Requirements for States and Public Water Systems," [Online]. Available: https://www.epa.gov/dwreginfo
[2] World Health Organization (WHO), "Guidelines for Drinking-Water Quality," 4th Edition, 2017.
[3] A. K. Jha, S. Yadav, and P. Tripathi, "Water Quality Prediction Using Machine Learning Algorithms," *Procedia Computer Science*, vol. 167, pp. 2221–2228, 2020.
[4] G. S. Malathi and A. Dhivya, "Prediction of Water Potability Using Classification Techniques," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 4, 2019.
[5] A. T. Ahmed et al., "Water Quality Assessment Using Machine Learning Models: A Review," *IEEE Access*, vol. 9, pp. 16253–16277, 2021.
[6] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1137–1143, 1995.
[7] Pedregosa, F., et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[8] Y. Chen and M. P. Siontas, "Comparative Analysis of Machine Learning Techniques for Water Quality Prediction," *IEEE Global Humanitarian Technology Conference (GHTC)*, 2020.
[9] S. Raschka and V. Mirjalili, *Python Machine Learning*, 3rd ed., Packt Publishing, 2019.
[10] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.