Machine Learning Approach To Predicting Water Quality: An Evaluation Of Potability Using Ensemble Models

By Azubuike H. Chibuike
Henryc.azubuike@gmail.com

17 September 2024

ABSTRACT

Ensuring safe drinking water is vital for public health, yet conventional water quality assessment methods can be time-consuming and resource-intensive. This study explores the application of machine learning (ML) models to predict water potability based on physicochemical parameters such as pH, hardness, and sulfate concentrations. Using a dataset containing various water quality features, models including Support Vector Classifier (SVC), Random Forest Classifier (RFC), and Gradient Boosting Classifier (GBC) were evaluated based on their precision scores. After hyperparameter tuning, SVC and GBC were combined in a voting classifier, yielding a precision score of 91%. SHAP analysis identified pH and sulfate as key determinants of water potability. This research highlights the potential of ML in water quality prediction, offering a cost-effective alternative to conventional testing methods.

Keywords: Water quality prediction, Machine learning, Ensemble methods, SHAP analysis.

1. Introduction

Ensuring the potability of water is a fundamental concern for public health and environmental safety. Traditionally, determining whether a water sample is safe for consumption involves a series of laboratory tests to measure various physicochemical, microbiological, chemical, physical, biological, and radiological parameters. Each of these parameters is analyzed against established safety standards, such as the Maximum Contaminant Levels (MCLs) set by the U.S. Environmental Protection Agency (EPA) or the World Health Organization (WHO) guidelines, to determine the overall safety of the water (EPA, 2020; WHO, 2017). However, this process can be complex and may not always provide a straightforward answer due to the interactions between different contaminants and varying thresholds for safety standards.

This study explores the potential of machine learning models to predict water potability by analyzing physicochemical parameters. By leveraging advanced data analytics, this research aims to provide a cost-effective and efficient alternative to conventional water quality assessment methods, ultimately contributing to safer drinking water management and public health protection.

2. Literature Review

2.1. *Conventional approaches to water quality assessment*

Classical water quality assessment methods involve a comprehensive range of tests to evaluate different aspects of water contamination. Physicochemical tests measure parameters such as pH, conductivity, hardness, and dissolved oxygen, which provide an indication of the water's overall quality (Gharibi et al., 2012). Microbiological testing focuses on detecting pathogenic organisms, such as Escherichia coli (E. coli), which can indicate fecal contamination (Sharma et al., 2020). Chemical tests assess the levels of various contaminants, including heavy metals, pesticides, and organic compounds, while biological and radiological tests look for harmful biological organisms and radioactive substances (Gupta et al., 2020).

Each parameter must be compared against the relevant standards to determine the potability of the water. For example, the EPA's MCL for lead is 0.015 mg/L, while the WHO's guideline value is 0.01 mg/L (EPA, 2020; WHO, 2017). The complexity arises because these standards are not always consistent across different regulatory bodies, and the interaction of multiple parameters can complicate the interpretation of results. Additionally, traditional testing methods can be time-consuming and require specialized equipment and trained personnel, which limits their applicability in resource-constrained settings (Gharibi et al., 2012).

2.2. *Some limitations to conventional methods*

The conventional approach of measuring individual parameters against set standards has several limitations. First, it assumes that each parameter independently affects water potability, which is often not the case. Interactions between different contaminants can create synergies that increase or decrease the overall risk, complicating the assessment (Kazi et al., 2009). Second, the thresholds for safety standards may not account for all possible combinations of contaminants, leading to potential misclassification of water quality. Third, traditional testing methods are often costly and time-consuming, making them unsuitable for real-time monitoring or widespread deployment in low-resource areas (Gupta et al., 2020).

2.3. *Machine learning approaches to water quality prediction*

Machine learning (ML) offers a promising alternative to traditional water quality assessment methods by leveraging large datasets to identify complex patterns and interactions among multiple parameters. ML algorithms can analyze combinations of physicochemical, microbiological, and chemical parameters to predict water potability with high accuracy (Rani & Thakur, 2021). Unlike traditional methods that rely on predefined thresholds, ML models learn from historical data to understand how different factors, in varying quantities, contribute to water quality outcomes.

Several studies have demonstrated the potential of ML for predicting water quality. For example, Alfian et al. (2018) utilized a Random Forest model to predict the potability of water samples based on multiple parameters, achieving a classification accuracy of over 90%. The Random Forest algorithm effectively handles non-linear relationships and complex interactions among variables, making it particularly suitable for this application. Similarly, Rani and Thakur (2021) employed Support Vector Machines (SVM) and Gradient Boosting algorithms to classify water samples into potable and non-potable categories, highlighting the ability of these models to improve prediction accuracy over traditional methods.

2.4. *Challenges and future directions*

While machine learning presents a promising approach to water quality prediction, several challenges remain. Data quality is a significant concern, as missing or inaccurate data can adversely affect model performance (Rani & Thakur, 2021). Additionally, the interpretability of ML models, particularly complex ones like Neural Networks, can be limited, making it difficult to understand the underlying reasons for specific predictions (Gupta et al., 2020). Ensuring model robustness and generalizability across different regions and water sources also requires careful consideration.

3. Methodology

The methodology for this project involves several steps, from data collection to model evaluation.

3.1. *Data collection*

The dataset used in this project is sourced from water quality monitoring systems that include various chemical and physical properties, such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, and trihalomethanes. The target variable is the ”Potability” of water, which is a binary classification (0 for not potable and 1 for potable). Below is a brief description of the water features contained in the dataset:

- pH: The degree of acidity or alkalinity of a solution (in this case, water). Ranges between 0 and 14, where lower values indicate acidity, higher values indicate alkalinity, and 7 is neutral.
- Hardness: Capacity of water to precipitate soap in mg/L.
- Solids: Total dissolved solids in ppm.
- Chloramines: Amount of Chloramines in ppm.
- Sulfate: Amount of Sulfates dissolved in mg/L.
- Conductivity: Electrical conductivity of water in S/cm.
- Organic Carbon: Amount of organic carbon in ppm.
- Trihalomethanes: Amount of Trihalomethanes in g/L.
- Turbidity: Measure of light emiting property of water in NTU.
- Potability: Indicates if water is safe for human consumption. Potable - 1 and Not potable - 0

3.2. *Preprocessing:*

The steps include:

- Handling missing values using imputation techniques, such as mean or median imputation. Based on the distributions, the data was not skewed so mean imputation can be used to fill the null values.
- Standardization of features due to variance within the dataset.
- Splitting the dataset into training and test sets using an 80/20 split ratio.

3.3. *Model development*

We will employ multiple ml models and choose the top performers for hyperparameter tuning. These models include: Logistic Regression, SVC, K Neighbors Classifier, Decision Tree Classifier, GaussianNB, Random Forest Classifier, Gradient Boosting Classifier, Ada Boost Classifier, Random Forest Classifier

3.4. *Model evaluation*

The choice metric for this project will be the precision score as we want to focus on correctly predicting the cases where the water is truely potable. Cross-Validation was also employed to validate the model's generalizability by splitting the data into multiple subsets and training/testing the model on different subsets.

3.5. *Hyperparameter tuning*

Hyperparameters of top performing models (which in this case include: Support Vector Classifier, Ada Boost Classifier, Gradient Boost Classifier, Random Forest Classifier) are optimized using Random Search technique to find the best combination of parameters that maximize model performance.

3.6. *Final model*

To further improve the model, we use the voting classifier. A voting classifier is an ensemble learning method in machine learning where multiple models are combined to make predictions. The idea behind a voting classifier is to combine the predictions of several models and make a final prediction by selecting the most frequent class prediction among the models. The main advantage of using a voting classifier is that it can improve the accuracy of predictions by combining the strengths of different models. For this project, we will use the top two models based on precision score i.e. Gradient Boosting Classifier and Support Vector Classifier.

4. Results

The models were evaluated based on precision, and the performances were as follows:

- **SVC:** 74%
- **RFC:** 66%
- **ABC:** 65%
- **GBC:** 61%
- **KNN:** 56%
- **GNB:** 47%
- **DTC:** 46%

After this initial evaluation, the top four (4) performing models were selected for hyperparameter tuning, and yielded a much more improved precision score as seen below:

- **SVC:** 88%
- **GBC:** 88%
- **RFC:** 78%
- **ABC:** 65%

The top two (2) performing models from the hyperparameter tuning (GBC and SVC) were combined using a voting classifier to get an improved mean precision score of **91%**.

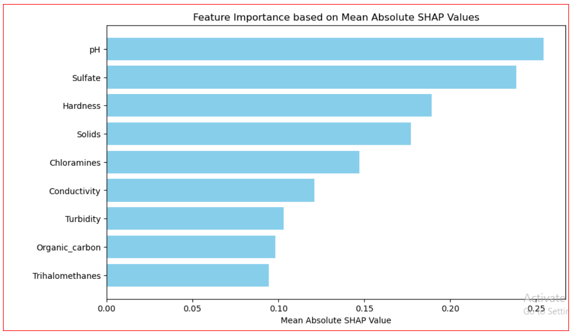

Fig. 1. Feature importance

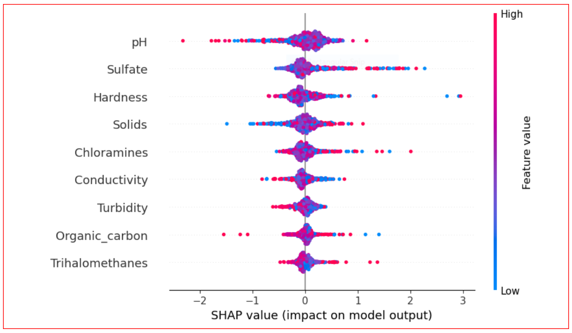

Fig. 2. SHAP plot

SHAP Analysis: SHAP analysis was finally carried out to understand the overall importance of each feature to the model, and how these features contributed (in magnitude) to each predicted outcome.

As seen in the plots above, the SHAP analysis revealed that overall, two very important features to the model were **pH** and **Sulfate**. This implies that from our data, on average, the two most important physicochemical properties that determines to a large extent if a water sample is potable or not, are **pH** and **Sulfate**. The **pH** feature which had the most impact on the model, will be briefly explained, and also how it contributed to predictions.

- **pH:** This simply represents the level of acidity or alkalinity contained in each water sample. It ranges from 0 to 14, where lower values indicate alkalinity, higher values indicate alkalinity, and 7 is neutral. From the SHAP plot, extreme pH values (tending towards 0 and 14) give rise to negative SHAP values, which in this case represents ”non portability” predictions; while values tending towards the center (7) tend to give rise to positive SHAP values, indicating ”potability” predictions. This is in line with general standards that pure water has a neutral pH of 7, which indicates that it is neither acidic or basic, and hence it is normal for a potable water to have a range of between 6.5 and 8.5 on the scale.

5. Discussion

The machine learning models employed in this study provide a modern solution to the challenges posed by traditional water quality assessments. ML models such as SVC, RFC, and GBC showed considerable accuracy in predicting the potability of water samples. For this study, the focus on precision is essential as the cost of misclassifying non-potable water as safe can have severe health implications.

The SHAP analysis deepened the understanding of the model's behavior, particularly highlighting the importance of pH levels in water quality assessments. This insight aligns with established guidelines, where pH values between 6.5 and 8.5 are considered ideal for potable water. The integration of these physicochemical properties within the model not only increases predictive accuracy but also provides interpretable results that can guide water treatment decisions.

However, machine learning approaches are not without challenges. Data quality, including missing or inaccurate values, can hinder model performance, and while methods such as imputation were employed in this study, ensuring high-quality data remains a significant concern. Additionally, the need for robust and generalizable models is crucial, as water quality varies across regions due to differences in environmental factors and pollution levels. Further research should explore larger datasets with greater geographical diversity to enhance model applicability.

6. Conclusion

This study demonstrates the potential of machine learning models in predicting water potability, offering a viable alternative to conventional water quality assessments. The application of ensemble models, particularly the voting classifier combining SVC and GBC, improved the precision of predictions. SHAP analysis identified pH and sulfate as critical determinants of water quality, reinforcing the importance of pH levels in potable water standards.

References

- Alfian, G., Rhee, J., & Yoon, B. (2018). Real-Time Water Quality Monitoring System Using IoT and Machine Learning Techniques. *IEEE Internet of Things Journal*, 5(5), 4221-4229.
- EPA. (2020). Drinking Water Regulations and Contaminants. U.S. Environmental Protection Agency.
- Gharibi, F., Mahvi, A. H., Nabizadeh, R., & Arabalibeik, H. (2012). A novel approach in water quality assessment based on fuzzy logic. *Journal of Environmental Management*, 112, 87-95.
- Gupta, A., Kumar, S., & Sharma, P. (2020). Predictive Modeling for Water Quality Analysis Using Machine Learning Techniques. *Environmental Science and Pollution Research*, 27(5), 5571-5582.
- Kazi, T. G., Arain, M. B., Jamali, M. K., Jalbani, N., Afridi, H. I., & Sarfraz, R. A. (2009). Assessment of water quality of polluted lake using multivariate statistical techniques: A case study. *Ecotoxicology and Environmental Safety*, 72(2), 301-309.
- Rani, S., & Thakur, R. (2021). Application of Random Forest and Gradient Boosting Methods for Water Quality Prediction. *Water Resources Management*, 35(2), 613-624.
- Sharma, S., Bhardwaj, N., & Kumar, V. (2020). Microbiological Quality of Drinking Water: Challenges and Technological Solutions. *Journal of Water and Health*, 18(1), 1-13.
- WHO. (2017). Guidelines for Drinking-Water Quality: Fourth Edition Incorporating the First Addendum. World Health Organization.

About The Author: Azubuike Chibuike Henry is an innovative Civil Engineer and Data Analyst, who specializes in applying Data Science to various aspects of Civil Engineering, finding valuable insights & trends in data sets, and developing useful solution-oriented predictive models where needed.

For access to the project files, visit my GitHub repository: Water Quality Prediction.