World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

Assessing recreational water quality with machine learning and SHAP-based interpretability

Vedanto Bhowmik *

Dougherty Valley High School, California, US.

World Journal of Advanced Research and Reviews, 2025, 27(03), 922–930

Publication history: Received on 05 August 2025; revised on 11 September 2025; accepted on 13 September 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.3.3166

Abstract

This paper aims to explore how machine learning (ML) can be used to forecast water quality tailored for recreational use by identifying contamination levels in water. The project investigates the physical and chemical properties of water that can be used to forecast bacterial contamination, with the goal of identifying the predominant water features influencing bacterial contamination forecast. Multiple classification models were developed and evaluated including Random Forest, Logistic Regression, k-Nearest Neighbors (KNN), and XGBoost. I additionally utilized SHAP (SHapley Additive ex Planations) and SHAP analysis to provide a detailed explanation of the model's performance. Based on the Fecal Coliform levels, the impact on the model's performance was the strongest, with both the Conductivity and Biodegradable Oxygen Demand (BOD) coming after and being of the same level. This aligns the findings with the general notions of environmental science. The limitations of the research include regional basis and applicability of the model to recreational water use. The developments of the research can help incorporate larger datasets with coverage of larger regions and try to predict drinking water safety to help improve the model's accessibility.

Keywords: Water Quality; Recreational Water; Coliform Bacteria; Chemical and Physical Properties; Machine Learning; Feature Importance Analysis; Shap Analysis

1. Introduction

Water quality remains a major concern regarding public health, environmental stability, and the economies worldwide [1, 3]. In the hierarchy of human needs, clean water comes first, whether for drinking, irrigation, industrial processes, or for recreation. Fecal bacteria have remained majorly problematic to recreational water safety; they include *Escherichia coli* (E. coli) and enterococci, which can cause gastrointestinal illnesses, skin infections, and respiratory ailments [9]. According to the CDC in the United States, millions of cases of recreational water illnesses occur globally every year on account of contact with polluted water, which constitutes a major public health concern [6]. Drinking water contaminated with pollutants or pathogens may cause diseases that threaten human life or disability from these diseases; for example, cholera, typhoid, and dysentery [8]. On the contrary, such water acts as a sinister polluter, damaging the environment by way of interactions between aquatic species and impairing biodiversity within the ecosystems [4]. Well-being water is not just a health issue; it is also an environmental issue for sustainability [5].

Water quality is a critical matter for India because it has a large population which is still growing [2]. While many works attempt to determine the general aspect of water-quality analysis, another aspect comprises the levels of fecal bacteria and whether that water is safe for recreational purposes [7]. The significance of this research lies in the contamination of Indian rivers and groundwater not only by industrial wastes, agricultural runoffs, and sewage wastes but also by the discharge of human wastes in the environment where sanitation facilities are not available [1]. In India, rivers and groundwater sources are contaminated by industrial discharges, agricultural inputs, and sewage wastes [10]. Among

[*] Corresponding author: Vedanto Bhowmik

them, the Ganges being the most culturally and spiritually significant river in India is maybe the most polluted and hence posing a great risk to communities and ecosystems [3]. Rural water users mainly depend on groundwater sources that may be contaminated with bacteria, besides arsenic, fluoride, and nitrates—this only amplifies the water quality crisis [9]. This is wherein begins the water quality crisis [8].

If ever one hopes to counteract all the more pressing challenges on water quality faced by the people in India, it has to be a quickest achievement of active community water management, good infrastructure, and pollution control mechanisms [4]. Initiatives by the governments are in process, wherein water will be piped to every house (e.g., many large-scale projects currently on construction such as the Jal Jeevan Mission), but the bigger solution of safe water is yet to be implemented, more so for recreational purposes [2]. There is a need for public awareness, continuous monitoring, and the use of predictive tools that could warn of unsafe fecal bacteria levels before people are exposed to much risk [6]. In this regard, urbanization and climate change will escalate water stress in India and will need more attention [5].

A study by Rong Xiao et al. (2013) investigated the presence and ecological risk of six heavy metals-cadmium, chromium, copper, nickel, lead, and zinc-in sediment samples from two river systems in the Pearl River Delta-the Shiqiao River (urban) and the Shawan River (rural). Researchers collected both surface and core sediment samples to analyze horizontal and vertical pollution distribution. They also analyzed factors like organic matter, grain size, and salinity to figure out how these factors influenced the accumulation of these six metals. The study aimed to find out pollution sources, evaluate toxicity levels, and derive high-risk areas requiring remediation with the data. The results from this study displayed that all of the studied metals exceeded the background values, with cadmium being the most common pollutant in both of the rivers. Urban river sediments exhibited a widespread non-point source pollution from industrial activity, whilst the rural rivers displayed localized high-risk zombies primarily from agricultural and little industrial runoff. Vertical profiles pointed to more stable deposition in the rural river and disturbance in the urban one, most likely from dredging. Toxicity assessments proved moderate to high ecological risks, with chromium and nickel adding heavily to toxicity, and the ecological risk index values showed “hot areas” needing attention in both river systems.

Another study by Zhang et al. (2023), pinpointed spatiotemporal trends in water quality over 117 monitoring sites in Zhenjiang China, using five years of surface water data. The researchers analyzed ten water quality indicators, some including ammonia nitrogen, biological oxygen demand, and dissolved oxygen. They used statistical and trend models to identify pollution drivers. Findings displayed clear spatial and seasonal patterns, urban areas experienced worse pollution during dry seasons due to lower amounts of dilution and increased amounts of discharge. Important influencing factors consisted of land use types, topographic slope, and economic activity. The limitations in the research included the simplicity of the statistical and trend models used, which restricted the depth of insight into more complex pollutant interactions. The authors of the paper made the importance of needing tailored management based on regional land use and water flow conditions clear.

Reza et al. (2021) discussed AutoML use to predict water quality according to a number of physicochemical parameters. To reduce manual model tuning, they aimed to compare AutoML-generated models versus expert-designed ones. Working with a refined data set containing water quality indicator variables, they applied both approaches to model classification levels of water quality. Cleaning and structuring the data was one of the challenges arising out of inconsistencies and missing values. Despite these challenges, AutoML models had an edge in some respect over the expert-tuned ones in terms of accuracy and speed, achieving an average classification accuracy of 91.3% and an F1-score of 0.89. However, another limitation cited by the study was the lack of generalizability of the present study because of the relatively small size and geographic scope of the analyzed data set. Their conclusion indicates that potentially, AutoML offers a workable solution with scalability in promoting and supporting water quality monitoring and decision-making systems in areas with low technical expertise or scarcity of resources.

Ahmed et al (2019) worked on supervised machine learning models to predict the Water Quality Index and to classify water quality classes using few features. The dataset was based on water samples from Rawal watershed. It contained four parameters that were easy to measure and were temperature, turbidity, pH, and total dissolved solids. They experimented with a collection of regression and classification algorithms, including gradient boosting, polynomial regression, and multilayer perceptron (MLP). The gradient-boosting algorithm provided the best WQI prediction with the lowest value of mean absolute error (roughly 1.96), while its counterpart, the MLP classifier, gave the maximum level of classification accuracy, that is, nearly 85%, for WQC. A severe problem was in handling noisy/incomplete data, addressed through preprocessing and feature selection to ensure model reliability. The study found that with or without limited input data, water assessment can be done by machine learning in an accurate and efficient manner, thereby rendering traditional means of lab-base methods obsolete.

Building on these prior studies about water quality, the primary objective of this research is to evaluate whether the physical and chemical properties of water can be used to predict water quality through supervised learning techniques. The goal is to estimate how accurately ML models can replicate or even enhance traditional water quality test methods by using environmental data. While much of the previous literature has focused on general water quality assessment, this study is more specifically concerned with recreational water safety, distinguishing whether water is suitable for body-contact activities such as swimming or only for non-body-contact uses like boating or irrigation. This investigation also tries to understand which elements of a physical or chemical nature contribute to model predictions and thus the most important factors in classifying water quality. The research's use of mixed methods and combining supervised and exploratory learning, introduces a data-driven approach to understanding water quality, with applications for better decision-making, improved monitoring, and early warning signals in both urban and more remote rural locales, particularly in regions like India that continue to struggle with water quality assessment.

1.1. Dataset

The dataset used here provides a collection of 619 water quality observations from various sources of water bodies in India: lakes, ponds, wetlands, and tanks (Baskar, 2022). It contains observations on important physical and chemical parameters needed to determine the quality of water. While the specific place of extraction is not given on the dataset for larger bodies of water like wetlands and lakes, it is meant to depict measurement in different regions of India which makes it convenient for comparison with rural and urban areas. The sources of data are environmental condition monitoring organizations or field studies in the different areas.

The data has many columns where each column represents a minimum and maximum amount of a water quality indicator. For instance, water pH (acid/base), temperature, dissolved oxygen (oxygen gas in liquid), turbidity (clarity of water), biological oxygen demand (the measure of the amount of dissolved oxygen used by micro-organisms), conductivity, nutrients such as nitrate, nitrite, bacteria such as fecal coliforms and total coliforms (each attribute with minimum and maximum values). Each of these characteristics of the water bodies will measure the sustainability of water for ecosystem and human use and therefore are all valuable inputs to predictive modelling.

2. Methodology

2.1. Input Feature and Output Target

The input features considered for this study encompass a number of chemical and biological water-quality parameters. These parameters include Temperature (Min and Max in °C), Dissolved Oxygen (Min and Max in mg/L), pH (Min and Max), Conductivity (Min and Max in mhos/cm), BOD (Min and Max in mg/L), Nitrate + Nitrite Nitrogen (Min and Max in mg/L) The output target features are Fecal Coliform and Total Coliform counts (Min and Max in MPN/100ml). These were chosen because of their relevance to assessing the water bodies' health and their potential influence on the water quality classification. As for the output, it is a categorical variable of water quality classes for body contact that were induced from the California State Water Resources Control Board and classified as discrete categories, such as "safe for body contact" for less than 200 coliform bacteria colonies per 100 mL water and "unsafe for body contact" for more than 200 coliform bacteria colonies per 100 mL water [11]. The model is aimed at predicting this water quality class from the input features so that it can be used to classify and monitor water conditions across several sites.

2.2. Data Pre Processing

First, missing/null values were detected and treated, usually by removal, depending on the nature and severity of the missing data. 0.484% of values were missing/null from the min and max Temperature C columns, 0.323% of values were missing/null from the min and max Conductivity (mhos/cm) columns, and 0.161% of values were missing/null from the min and max Fecal Coliform (MPN/100ml) and the min and maxTotal Coliform (MPN/100ml) columns.

Afterward, 55 anomalous data points (extreme) were detected by the Isolation Forest [14] and removed to lessen their possible influence on the training process. Occasional outliers found in the (all have minimum and maximum value columns) BOD, Fecal Coliform, Total Coliform, Conductivity, and Nitrate and Nitrite columns were present due to extreme readings caused by environmental noise, sensor errors, or rare local events. The Isolation Forest was trained in all the features relevant to the data set to get an underpinning distribution of normal observations. One can find the data records that are easier to isolate by recursively splitting the datasets. It assigns every instance an anomaly score and removes those that were quite distant from the observed standard behavior. This filtering thus guaranteed that the remaining data mostly represented common trends in water quality. So, common trends will probably prevent any misleading patterns from coming up in training and enable a much more reliable prediction stage.

For the target variable of bacterial counts or water quality, the values were binned into discrete categories (e.g., “safe”, “not safe”) through the use of certain threshold guidelines by the California State Water Resources Control Board. They state that the fecal coliform concentration should not exceed a log mean of 200 per 100 ml for body to water contact activities. With these steps, the dataset came out clean, consistent, and robust enough for both supervised and unsupervised learning.

2.3. Models and Evaluation Metrics

To evaluate the predictive performance of multiple supervised learning algorithms for water quality classification using physicochemical parameters such as temperature, pH, turbidity, and total dissolved solids, five models were considered: Logistic Regression [15], k-Nearest Neighbors (kNN) [16], Random Forest [17], and XGBoost [18]. Logistic Regression was selected for its simplicity and interpretability, kNN for its non-parametric, neighborhood-based decision process, Decision Tree for its explicit rule-based structure, Random Forest as an ensemble of decorrelated decision trees to improve generalization, and XGBoost as a gradient boosting method leveraging sequential error correction for high predictive accuracy.

For model optimization, hyperparameter tuning was conducted using grid search in conjunction with 5-fold stratified cross-validation repeated 50 times to ensure robust and low-variance performance estimates. For Random Forest, the parameter grid included n_estimators ∈ {100, 500}, min_samples_split ∈ {10, 20, …, 100}, and max_features ranging from 1 to (total_features − 1), with class_weight fixed to ‘balanced’. For XGBoost, the search space included eta ∈ {0.001, 0.01, 0.05, 0.1, 0.2}, subsample and colsample_bynode ∈ {0.6, 0.7, 0.8, 0.9, 1.0}, and n_estimators ∈ {50, 100, 200}, with eval_metric fixed to logloss and use_label_encoder=False. Logistic Regression, kNN, and Decision Tree were also tuned using grid search over their respective key hyperparameters. All models were evaluated during tuning using f1_weighted to account for minor label imbalance while capturing per-class performance.

The repeated stratified cross-validation process produced per-label and overall (weighted) precision, recall, and F1-scores, reported as mean ± 95% confidence interval (CI) across all folds and repeats, providing a statistical measure of performance stability. After selecting the best hyperparameters for each model, final performance on unseen test data was assessed using bootstrap resampling of the test set with 1000 iterations. For each bootstrap sample, predictions were generated, and per-class and weighted precision, recall, and F1-scores were computed to derive distributions and 95% CIs, enabling uncertainty quantification and bias detection across labels. Comparing CV-derived metrics with bootstrap-derived test metrics allowed for an assessment of generalization capability and overfitting risk.

2.4. Feature Importance Analysis

In this research, the feature importance analysis holds major significance because it enables one to get an idea about which physical and chemical parameters of water are more significant in making an overall prediction about water quality. While a machine learning model can make very accurate predictions, there is always an imprint of operating as “black boxes," unable to provide insight into the contribution of each input variable to the final decision. By associating feature importance, the analysis brings out the respective parameters-the pH, turbidity, or dissolved oxygen-that influence the model predictions the most. Hence, the analysis brings more clarity and also helps direct policymakers, environmental agencies, and communities to monitor and intervene mainly on those water quality indicators that bear the most significance. In other words, the analysis helps ensure that our results are meaningful and actionable along with being predictive.

To reach this aim, we are analyzing the results using the method most often used to interpret machine learning models. SHAP (SHapley Additive ex Planations) assigns a value of importance to a feature concerning the model output for a single prediction under consideration with a game-theoretic approach: how much has each feature "contributed" to either pushing a prediction higher or lower than the expected value? For instance, SHAP would tell whether high conductivity strongly contributes to a negative classification of water quality, or if dissolved oxygen contributes to the positive classification of water quality. Traditional feature importance metrics provide an explanation valid globally for the dataset, whereas they lack local explanations that would explain why a model made a certain prediction for an individual data point. That is what makes SHAP so powerful for environmental studies, because it not only tells which features have the biggest say but also sheds light on more subtle relationships between water quality parameters and prediction outcomes.

3. Results and Discussion

This section of the study will draw conclusions from the obtained results. Accordingly, baseline models are presented with main procedures that serve as contrasting bases for other models. Subsequently, hyperparameter tuning was implemented, during which gains in performance and accuracy made by the respective algorithms were discussed. In the next step, the models were again ranked to determine which among them are dominating the others in classification accuracy, precision, recall, and F1-Score. Interpreting using SHAP analysis the best-performing model involved the contribution of the physicochemical features to the final prediction. Once a conclusion is drawn on the strength of the modeling technique, an actual cause to some extent can be made for water quality outcomes.

Table 1 shows the results of the basic kNN and Logistic Regression on both the training and testing data sets. The kNN reports an F1-score of 74-77%, with a slight improvement in the training part of the classification and a strong validation setting. The test performance includes a more significant decline, especially for Class 0, where the recall is 56%. This under-performance points to overfitting of the model, as is shown by the training viewed data. As for the logistic regression, it records a slightly stronger overall test performance with an F1-score of 70-73%, with a strong recall that outweighs defensive improvements. The decrease in the testing region as compared to as viewed on the training data is lesser because the Logistic regression now provides an acceptable fall. Consequently, based on the provided baseline, the Logistic Regression outperforms kNN as it continues to cover the classes better as reflected in the F1-score and recall performance.

Table 1 Evaluation metrics of baseline models

Model	Train/Test	Label	Precision	Recall	F1	Samples
KNN	Train	0	74%	75%	74%	182
		1	78%	76%	77%	207
	Test	0	72%	56%	63%	55
		1	56%	72%	63%	43
Logistic Regression	Train	0	78%	80%	79%	182
		1	82%	81%	81%	207
	Test	0	79%	67%	73%	55
		1	65%	77%	70%	43

The Random Forest (RF) model after adjustments achieved strong performance across both training and testing when utilizing its optimized parameters; nevertheless, minimal overfitting was observed. With regards to the recall, Label 0 (safe for recreational purposes) reached a value of 82.88% and its precision 83.37%, while Label 1 (unsafe for recreational purposes) had a superior F1 score of 84.97% and its precision and recall as 85.18% and 85.08% respectively. In contrast, the values presented in the testing dataset were of F1 score and therefore slightly lower. Upon closer inspection, it seems that the performance values for RF are of good quality. With minor refinement, the model could achieve both the desired precision and the recall. is introduced, and both precision and recall seem to be clearly defined, the model would work best with the precision and recall to be set, it would be best if tests were run with the given parameters in terms of what shape the data should test on.

The random forest model was outperformed by the XGBoost in the majority of tests. The XGBoost model has a better generalization and test performance. For training, it has an F1 score of 83.07% ± 0.52 for Label 0 and 85.07% ± 0.45 for Label 1 values that are near the random forest models but with more confidence. However, the Label 0 and 85.07% ± 0.45 for Label 1 and the significant region around them is more the result of the random forest. For the testing set, the XGBoost model has Label 0 and Label 1 having an F1 score and both labels having one with 87.42% ±0.20 and 83.09%, respectively. This shows that XGBoost was more successful in using the data to its full potential and correctly flagging bad water conditions, which is vital for public health and recreational activities. XGBoost actually outperforms both models, especially when the F1 score for Label 1 is considered.

Table 2 Evaluation metrics of tuned models

Model	**Train/Test**	**Label**	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Samples**
Random Forest *'max_features': 2,* *'min_samples_split': 10,* *'n_estimators': 500*	CV Train	0	83.37 +/- 0.72	82.88 +/- 0.76	82.92 +/- 0.54	182
		1	85.18 +/- 0.57	85.08 +/- 0.75	84.97 +/- 0.49	207
	Test	0	85.14 +/- 0.30	83.79 +/- 0.32	84.33 +/- 0.23	55
		1	79.61 +/- 0.39	81.19 +/- 0.38	80.19 +/- 0.3	43
Xgboost *'colsample_bynode': 0.8,* *'eta': 0.2,* *'n_estimators':50,* *'subsample': 0.7*	CV Train	0	83.33 +/- 0.67	83.24 +/- 0.80	83.07 +/-0.52	182
		1	85.50 +/- 0.59	84.99 +/- 0.72	85.07 +/- 0.45	207
	Test	0	85.93 +/- 0.29	89.19 +/- 0.27	87.42 +/- 0.2	55
		1	85.42 +/- 0.35	81.25 +/- 0.38	83.09 +/- 0.27	43

According to table 2, the model that performed best in this research was XGBoost, and it was superior to Random Forest, as demonstrated in the training and testing datasets. Through SHAP (SHapley Additive exPlanations) the feature importance analysis was leveraged not only to interpret the model's decision-making process, but also to pinpoint out which water quality parameters were mostly responsible for correct predictions. SHAP further enables the quantification of each feature's effect on the output, thereby providing practical insight on the influence of the water's physical and chemical classifications for distinguishing water safety and is therefore essential in ensuring that machine learning outcomes are applied correctly to environmental science.

In order to expound on how each variable obtained from the visual models for the SHAP's visualization, SHAP's analysis in Figure 1 and 2 reveal to us that the predictive module has a very strong relation with the level of oxygen dissolved and BOD. These variables have a more important positive meaning than others. This means that the output from the model is aided with higher levels of the DO, whereas the higher levels of BOD have a positive harm. With high levels of BOD, it explains that the environment is rather unhealthy. On the other hand, higher levels of DO are really great for the aquatic world. As for the model, it pays no attention to things like Conductivity and ph, which means that the SHAP values are very close to 0, which do not assist the model in any way. Temperature and Nitrates are showing as moderately significant, but at the same time they will have a positive effect on the output. The models can be seen to have a gloomier relationship with respect to the relationships of the variables, which means they could differ in the functional relationships and depend on the context. The analysis shows that the model seems to be connecting the basic biological major keys so that its output is more refined than the singular variables.

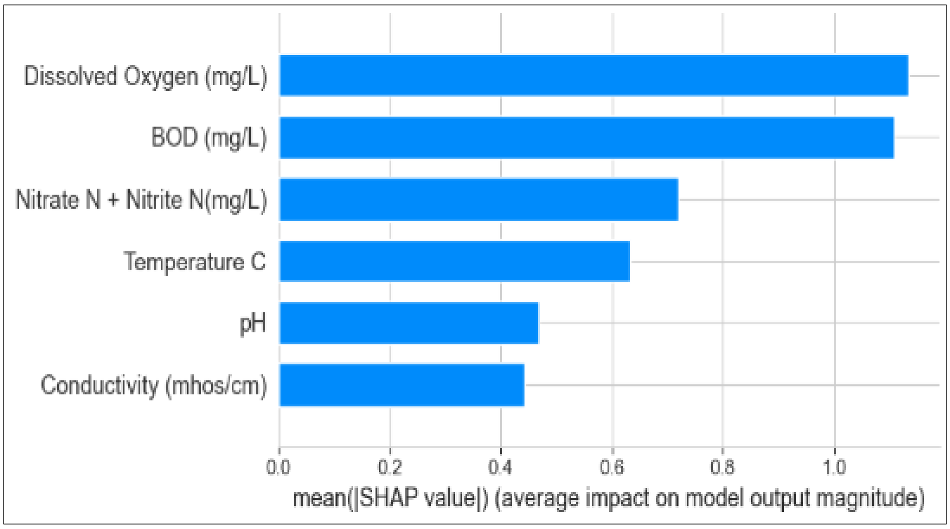

Figure 1 Mean absolute SHAP values for the water quality model

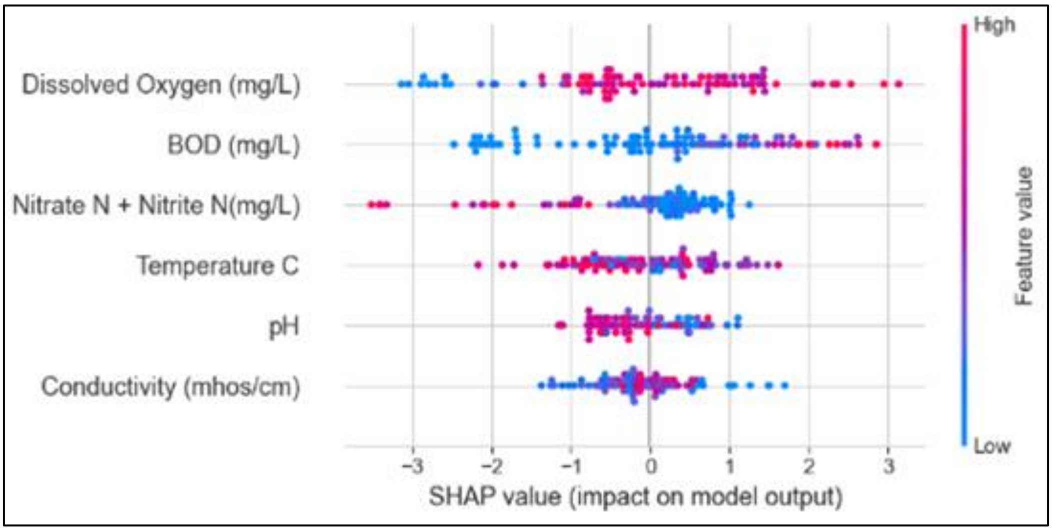

Figure 2 SHAP beeswarm plot illustrating the impact of feature values on the model's output

As stated in various studies, machine learning models offer commendable accuracy in the interpretation of water quality. For example, Reza et al. (2021) and Ahmed et al. (2019) have a machine learning model with an average accuracy level of about 85%, which is also clearly visible in this particular research's XGBoost model, which has an F1 score of 87.42% and 83.09% for safe-to-drink and unsafe-to-drink waters, respectively. Utilizing the SHAP analysis, once again, offers an in-depth explanation of the impact of each feature, which also aids identification of Nitrate, DO, and BOD, which is known as the recent in-depth analysis of the features. Since the assessment of Nitrate, DO, and BOD along with the other parameters directly indicate the severity of pollution in the ever changing and multifaceted ecosystems, all of the methods Reza et al. (2021) and Ahmed et al. (2019) were no less beneficial. The neglect of the recreational water/ the contact of the water with the body that was discussed and explained is the same from the finding and the presented research.

In respect to the research provided, the dataset provided was helpful, but would be more helpful if expanded. It would have provided more insight and accuracy had it not only been limited to a specific geographical region. Water testing from different geographic regions, climates, pollution sources, different land uses, and socioeconomic activity can be helpful in further enhancing the results. The analysis from my research is helpful, but needs seasonal statistics to show the effectiveness of the model.

Rather than concentrating on recreational water quality classification as a whole, another limitation of this study is that they classify whether the water is safe or not for recreational water activities like swimming. Looking wider, I believe it can also be used to determine if water is safe for humans to drink based on the rigorously tested results. Secondly, the study looks at physical and chemical parameters alone and makes an assumption that they are completely responsible for the characteristics of water, thus overlooking the biological and hydrological parameters that in reality help in making water quality determination much more precise.

In the future, the study should be looking at acquiring much larger and more comprehensive data sets for their study qualification, expanding their parameters to include microbial counts and hydrodynamic data. Using deep learning frameworks could further their cause in terms of accuracy. The study needs to go beyond recreational water study to capture both agricultural and potable water, and their models will be much more useful. Real-time data gathering will also help by making the model more accurate.

4. Conclusion

The main goal of this study was to predict the conditions of water safety for recreational activities such as body-contact swimming through the application of physical and chemical properties of water. With the use of supervised learning models, the primary aim of the research was to enhance the pre-existing water quality evaluations with data-driven methods. The study further sought to trace the contributions of each independent feature to the prediction, thereby using feature selection to outline the bare minimum properties for water quality and enhanced methods. Moreover, supervised and exploratory learning techniques were employed to understand better the regions and the expected pollution sources.

The XGBoost model showcased the highest predictive accuracy and strongest balance across training and test sets when compared to the other models. For both labels, the model also demonstrated robust precision, recall, and F1 scores, thus indicating that the model's predictions are well generalized to the world, even while the world is changing. On the other hand, the Random Forest model (RF) also performed well, but it showed relatively larger performance gaps across sets. On balance, the Random Forest model remains reasonable when the XGBoost model's superior metrics and consistency are taken into account for predicting recreational water safety.

An evaluation of the XGBoost best performing model SHAP results was also conducted to test and gain better understanding regarding the particulars that have a say in the model's prediction. The analysis revealed that Fecal Coliform (MPN/100ml) (Max), key to a very critical and neurological disease that may be associated with bacteria, stands out as the most important condition while also correlating to the recreational water safety standards and guidelines provided. Moreover, Conductivity and BOD (Biological Oxygen Demand) also stood significant, though less, and some other conditions with slightly less impact. These results show that the model is good at identifying water safety using biological and chemical markers.

While these findings are promising, the study has limitations in the amount of information available in the dataset used. Going forward, resources should be allocated to expanding the datasets to cover other geographic areas, in addition to the physical and biological parameters. In addition to drinking water regulations being improved, increasing public safety with water systems should be achieved, and efforts should be made to increase recreational water safety as well. It is highly likely that public health and environmental sustainability will benefit from the improvements made to the new monitoring and enforcement systems, as the systems could be made in real time.

References

[1] Prasad, D. V. V., Raju, S. B., Sajja, G. S., Reddy, P. V. G. D., Suresh, A., and Reddy, P. V. G. (2021). Automating water quality analysis using ML and AutoML Environmental Research, 204, 112945. https://www.sciencedirect.com/science/article/pii/S0013935121010148

[2] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., and García-Nieto, J. (2019). Efficient Prediction of Water Quality Index Using Machine Learning. Water, 11(11), 2210. https://www.mdpi.com/2073-4441/11/11/2210

[3] Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., Ye, L. (2022). A review of the application of machine learning in water quality evaluation. Eco-Environment and Health, 1, 107–116. https://doi.org/10.1016/j.eehl.2022.06.001.

[4] Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Chow, M. F., Hossain, M. S., Ehteram, M., Elshafie, A. (2019). Machine learning methods for better water quality prediction. Journal of Hydrology, 578, 124084. https://doi.org/10.1016/j.jhydrol.2019.124084.

[5] Abuzir, Y. S., and Abuzir, S. Y. (2022). Machine learning for water quality classification. Water Quality Research Journal, 57(3), 152–164. https://iwaponline.com/wqrj/article/57/3/152/88998/Machine-learning-for-water-quality-classification

[6] Dogo, E. M., Nwulu, N. I., Twala, B., and Aigbavboa, C. (2019). A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water Journal, 16(3), 235–248. https://www.tandfonline.com/doi/abs/10.1080/1573062X.2019.1637002

[7] Alghamdi, M. (2018). Classification model for water quality using data mining techniques. International Journal of Advanced Computer Science and Applications, 9(10), 42–49. https://d1wqtxts1xzle7.cloudfront.net/99641969/479b6cfcd3fefe47dd1b20050489ce7bf368-libre.pdf

[8] Chen, W., Xu, D., Pan, B., Zhao, Y., and Song, Y. (2024). Machine Learning-Based Water Quality Classification Assessment. Water, 16(20), 2951. https://www.mdpi.com/2073-4441/16/20/2951

[9] Abdelmoula, A. M., El-Masry, M. F., Tawfik, M. A., and El-Bendary, A. A. (2022). Simple Prediction of an Ecosystem-Specific Water Quality Index and the Water Quality Classification of a Highly Polluted River through Supervised Machine Learning. Water, 14(8), 1235. https://www.mdpi.com/2073-4441/14/8/1235

[10] Xin, L., and Mou, T. (2022). Research on the Application of Multimodal-Based Machine Learning Algorithms to Water Quality Classification. Wireless Communications and Mobile Computing, 2022, 9555790. https://doi.org/10.1155/2022/9555790

[11] Xiao, R., Bai, J., Huang, L., Zhang, H., Cui, B., and Liu, X. (2013). Distribution and pollution, toxicity and risk assessment of heavy metals in sediments from urban and rural rivers of the Pearl River Delta in southern China. Environmental Monitoring and Assessment, 185(9), 7747–7758.

[12] Reza, M., et al. (2021). "Assessment of water quality in groundwater resources of Iran using a modified drinking water quality index (DWQI)." Ecological Indicators, 30, 28–34. https://doi.org/10.1016/j.ecolind.2013.02.008

[13] Baskar, B. (2022). Water bodies quality data – India [Data set]. Kaggle. https://www.kaggle.com/datasets/balabaskar/water-quality-data-india/data

[14] GeeksforGeeks. (2025, July 23). What is Isolation Forest? GeeksforGeeks. Retrieved August 18, 2025, from https://www.geeksforgeeks.org/machine-learning/what-is-isolation-forest/

[15] IBM. (2025, May 14). What is logistic regression? IBM Think. Retrieved August 18, 2025, from https://www.ibm.com/think/topics/logistic-regression

[16] IBM. (n.d.). What is the k-nearest neighbors (KNN) algorithm? IBM Think. Retrieved August 18, 2025, from https://www.ibm.com/think/topics/knn

[17] IBM. (n.d.). What is random forest? IBM Think. Retrieved August 18, 2025, from https://www.ibm.com/think/topics/random-forest

[18] IBM. (2024, May 9). What is XGBoost? IBM Think. Retrieved August 18, 2025, from https://www.ibm.com/think/topics/xgboost