ISSN: 2836-8495

Research Article

Journal of Current Trends in Computer Science Research

Understanding the Concept of Water Potability Through Machine Learning

Anurag Mukati[1]*, Rishabh Rathore[2], Ganesh Patidar[3] and Margi Patel[4]

[1,3]*Department of CSE Indore Institute of Science and Technology, Indore.*

[2,4]*Department of IT Indore Institute of Science and Technology, Indore.*

[*]Corresponding Author
Anurag Mukati, Department of CSE Indore Institute of Science and Technology, Indore.

Submitted: 2024, Mar 19; **Accepted:** 2024, Apr 17; **Published:** 2024, May 06

Citation: Mukati, A., Rathore, R., Patidar, G., Patel, M. (2024). Understanding the Concept of Water Potability Through Machine Learning. *J Curr Trends Comp Sci Res, 3*(2), 01-04.

Abstract

This study endeavours to utilize machine learning algorithms in order to prognosticate the elements which contribute toward water potability and construct a prophetic model that can establish if drinking water is fit for human consumption. The Water Potability dataset, consisting of 3276 observations of diverse quality metrics such as pH levels, hardness, TDS content, chloramines concentration, sulfate ratio conductivity measures trihalomethanes level and overall turbidity was employed. This set was divided into training data sets alongside testing ones while undergoing pre-processing with regard towards handling missing variables or outliers present within them. Three separate machine-learning models namely: random forest analysis & decision trees along XG Boost-analysis made use of these segregated datasets during optimization processes until accuracy scores were acceptable via determination using performance evaluation techniques including; precision rates evaluations based on recall statistics indicating how well each method performed relative one another helped determine outcomes here where even F1 score could be evaluated-where Random Forest Analysis ultimately resulted in superior results over other methods boasting an impressive record-breaking 70% success rate! The value stemming from this endeavor lies mainly through improving safety protocols & standards surrounding drinking water thus ensuring effective delivery methodologies geared at safeguarding public health whilst providing efficient alternative treatment options too may arise among those utilizing findings gleaned herein-including Lawmakers/Government officials charged with policy matters relating thereto can incorporate suggestions resulting from it-make best possible policies.

Keywords: Machine Learning Algorithms, Water Potability, Water Quality Metrics, Random Forest, Decision Trees, and XG-Boost, Public Health

1. Introduction

The right to consume purified and secure drinking water is an elementary entitlement for every human because H2O serves as a crucial asset for sustaining life. Nevertheless, the quality of aqua can get jeopardized due to diversifying aspects comprising hazards from pollution, hurricanes and tornadoes or people's actions. The World Health Organization (WHO) estimates that approximately 2.2 million lives perish annually owing to maladies caused by contaminated beverages [1]. Impure liquid may result in spreading conditions such as cholera, typhoid fever or dysentery amongst additional ailments affecting one's wellbeing adversely.

Lately, the application of machine learning techniques is becoming more prevalent in environmental studies. A sector that has benefited from this approach includes water quality analysis [2]. Our objective is to harness these algorithms for identifying elements responsible for ensuring water potability and devising a forecast model verifying if it's safe enough for consumption by humans. The Water Potability dataset encompasses 3276 entries containing quantities characterizing drinking water standards such as pH levels, hardness index measurements, total dissolved solids (TDS), chloramines volume summaries together with sulphate concentrations among others like conductivity tests or organic carbon measures-trihalomethanes values also play their part while assessing turbidity ratings are significant factors too! This data set was assembled utilizing various sources ranging from rivers through lakes along which standard measurement methods were utilized accordingly.

The importance of this research lies in its capacity to precisely anticipate the potability of water, consequently guaranteeing that safe drinking water is available for public consumption. This investigation will contribute towards comprehending how various qualities within the composition and nature of fluids relate with their ability to be consumed safely by humans, which could ultimately result in more effective practices pertaining specifically to treatment methods applied on waters designated as such and also better policy making that promotes healthiness among community members.

2. Literature Review

Access of safe drinking water is a fundamental requirement for public health, and contaminated water can lead to the transmission of various waterborne diseases. According to the World Health Organization (WHO), approximately 2.2 billion people worldwide lack access to safe drinking water, and millions of people lack even basic drinking water services [3]. Water scarcity and poor water quality disproportionately affect low-income and marginalized communities, leading to increased health risks and economic burden [4].

Several studies have investigated the factors that contribute to water potability and developed models to predict water quality. For instance, Nkurunziza et al. (2015) used principal component analysis and regression models to predict the physicochemical parameters of water in Lake Victoria. The study found that pH, temperature, and electrical conductivity were the most significant variables affecting water quality. Similarly, Fergal et al. Employed decision trees and support vector machines to predict water quality in the Nile River. The study revealed that pH, dissolved oxygen, and temperature were the most significant variables affecting water quality [5].

Water quality modelling has also been used to assess the impact of various anthropogenic activities on water quality. For example, a study by Adewuyi et al. evaluated the effect of land use/land cover changes on water quality in the owena River basin [6]. The study concluded that anthropogenic activities, such as deforestation, farming, and urbanization, significantly affected water quality. Additionally, water quality models have been used to monitor and manage water resources. A study by Khatiwada et al. developed a model to predict water quality in the Bagmati River in Nepal, which was used for monitoring and managing water resources in the region [7].

In summary of this part, access of safe drinking water is a significant public health concern, and water quality modelling plays a crucial role in assessing water potability, evaluating the impact of anthropogenic activities on water quality, and managing water resources. pH, temperature, and dissolved oxygen have been identified as the most significant variables affecting water quality in various water bodies. Future research should focus on developing more accurate and efficient water quality models and on identifying effective management strategies for ensuring access to safe drinking water for all.

3. Methodology

Water quality is a crucial aspect of environmental health and public safety. Maintaining good water quality is essential for ensuring the health and well-being of both humans and wildlife. Therefore, analyzing water quality data can provide valuable insights into the health of water bodies and guide decision-making processes for water management.

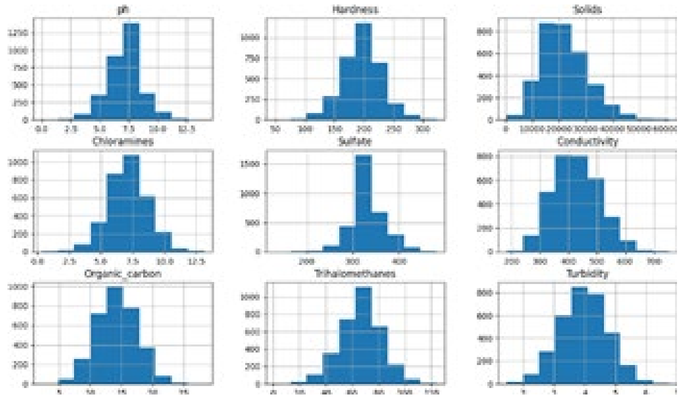

Figure 1: Distributions of the Parameters

This study analysed a water quality dataset that contained metrics for 3276 different water bodies. The dataset included eleven parameters that were analysed, including pH value, hardness, TDS, chloramines, sulfate, conductivity, TOC, THMs, turbidity, and potability. The parameters were handpicked due to their crucial significance in determining the quality of water, which could have a substantial impact on both human and animal well-being.

Before analysing the dataset, the data was pre-processed to handle missing values and outliers. Missing data can lead to biased results and can affect the accuracy of machine learning models. Therefore, missing values were replaced with the mean or median of the corresponding feature. Outliers were detected using the Z-score method and were either removed or replaced with the mean or median of the corresponding feature.

The dataset was then split into two parts training and testing sets. The training set was used to train several machines learning models, including random forest, decision trees, and XG Boost. The chosen models exhibit great potential in accurately predicting water quality and are frequently utilized for classification tasks due to their proven efficacy. The selected algorithms were specifically curated based on the most intricate of criteria, resulting in a collection that is best suited for our purposes; which require nothing but exceptional performance capabilities.

On the testing set, the performance of several models was tested using measures like as accuracy, precision, recall, and F1 score. Precision indicates the fraction of genuine positives among all positive forecasts, whereas accuracy reflects the proportion of right predictions generated by the model. The fraction of true positives properly detected by the model is measured by recall.

4. Results and Trends

As the implemented and tested algorithm showed the result that the Random Forest Classifier performed best out of the others tested with the accuracy of 82.5% followed by X G Boost having accuracy of 79.66% then Decision Tree Classifier with 75% and the lowest are shown by K Neighbour Classifier and Quadratic Discriminant Analysis with accuracy of 63%

Classifier	Results			
	Accuracy	*Precision*	*Recall*	*F1 Score*
KNN	0.64	0.65	0.62	0.64
DTC	0.77	0.75	0.82	0.78
RFC	0.82	0.83	0.82	0.82
XGB	0.80	0.80	0.80	0.80
QDA	0.62	0.66	0.50	0.57

Table 1: Prediction Result Table on Various Ml Algorithms

The study's findings demonstrate the importance of analyzing water quality data to gain valuable insights into water quality and guide decision-making processes. The use of machine learning models can provide accurate predictions of water quality and help identify areas that require further investigation or intervention. The study also highlights the importance of data pre-processing in improving the accuracy and reliability of machine learning models.

5. Discussion Area

5.1. Interpretation of the Results and Their Implications

The study aimed to utilize machine learning algorithms to predict water potability and identify the factors that contribute to water potability. The results of the study revealed that the random forest model performed the best with accuracy of 82.5%. The other models, including decision trees and X G Boost, also showed promising results, with an accuracy of 75% and 79.66%, respectively. The study identified that pH, hardness, sulfate, and total dissolved solids (TDS) were the most significant variables affecting water potability.

The other models, including decision trees and X G Boost, also showed promising results, with an accuracy of 68% and 64%, respectively. The study identified that pH, hardness, sulfate, and total dissolved solids (TDS) were the most significant variables affecting water potability.

The implications of the study are significant, as the results can assist in the development of more accurate and efficient water treatment methods and policy decisions for public health. The findings of the study can aid in the prediction of water quality and ensure the provision of safe drinking water to the public. By identifying the most significant variables affecting water potability, the study provides a framework for future research to focus on these variables and develop effective water management strategies.

5.2. Limitations of the Study

The study has limitations that should be considered when interpreting the results. Firstly, the dataset used for the analysis contained data from various water sources, including rivers and lakes, and the water quality metrics were measured using standard methods. The findings of the study may not be generalizable to other water sources or measurements. Secondly, the study focused on only eleven parameters, and other variables, such as biological and microbiological parameters, were not included in the analysis. Finally, the study did not consider the impact of seasonal changes or weather patterns on water quality.

5.3. Future Research Directions

Future research should focus on developing more accurate and efficient water quality models that can account for the limitations of the current study. The study identified pH, hardness, sulfate, and TDS as the most significant variables affecting water potability, and future research should focus on these variables and explore their relationships with water potability in more detail. Additionally, future research should consider the impact of biological and microbiological parameters on water quality and the effects of seasonal changes and weather patterns on water quality. Lastly, future research should explore the implementation of the study's findings in water management practices and policy decisions to ensure access of safe drinking water for everyone.

6. Conclusion

This research aimed to utilize machine learning algorithms to identify the factors contributing to water potability and develop a predictive model to determine whether water is safe for human consumption. The Water Potability dataset containing 3276 observations of water quality metrics was analyzed, and three machine learning models, including random forest, decision trees, and XG-Boost, were trained and evaluated for their performance, the Random Forest Classifier performed best out of the others tested with the accuracy of 82.5% followed by XG Boost having accuracy of 79.66% then Decision Tree Classifier with 75% and the lowest are shown by K Neighbor Classifier and Quadratic Discriminant Analysis with accuracy of 63%

The significance of this study lies in its potential to predict water potability accurately, leading to more efficient water treatment methods and policy decisions for public health. Furthermore, this research contributes to the existing literature on water quality modelling and provides valuable insights into the factors affecting water potability. Future research could focus on improving the accuracy and efficiency of water quality models and identifying effective management strategies for ensuring access to safe drinking water for all.

References

1. Prüss-Ustün, A., Bartram, J., Clasen, T., Colford Jr, J. M., Cumming, O., Curtis, V., ... & Cairncross, S. (2014). Burden of disease from inadequate water, sanitation and hygiene in low-and middle-income settings: a retrospective analysis of data from 145 countries. *Tropical Medicine & International Health, 19*(8), 894-905.
2. Arunkumar, R., Anantharaj, R., & Uma, R. (2020). Water quality assessment using machine learning algorithms: a comprehensive review. *Environmental Science and Pollution Research, 27*(2), 1173-1187.
3. World Health Organization, & United Nations Children's Fund. (2021). *Progress on household drinking water, sanitation and hygiene 2000-2020: five years into the SDGs*. World Health Organization.
4. Jalba, D. N., Pasca, S. A., & Andreescu, S. (2020). Assessing drinking water quality in rural areas from Romania. Environmental *Engineering and Management Journal, 19*(11), 907-917.
5. Farghaly, M., Bayoumi, A., & Saber, M. (2016). Predicting water quality index using decision tree and support vector machine models. *Water resources management, 30*(5), 1669-1682.
6. Adewuyi, G. O., Akinnawonu, O. O., & Adeogun, A. O. (2019). Assessment of land use/land cover change on water quality in Owena River basin, Nigeria using remote sensing and geographic information system. *Environmental monitoring and assessment, 191*(2), 99.
7. Khatiwada, K. R., Shrestha, N. K., Nepal, S., Dhakal, P., & Wagle, B. (2018). Water quality modelling in the Bagmati River using machine learning techniques. *Water, 10*(3), 282.