-Ritika Mathur

**Title:**

Prediction of drinking water’s potability using Machine Learning models

**Abstract:**

Clean water for drinking is one of the most essential requirements for humans and in general, life to exist on Earth. This research paper investigates the prediction of the potability of drinking water using machine learning models by classifying it into the categories of potable or non potable. The results show the accuracy of different machine learning models and that the machine learning model achieved around 73% classification accuracy The dataset employed in this study was sourced from the Kaggle repository. This study highlights the potential of machine learning techniques in areas as big as water quality assessment and aims to contribute to the development of automated systems for monitoring of safety of drinking water.

**Introduction:**

“Potable water” term is defined as the water that is purified as well as feasted perfectly and is eventually free from all the toxins and hazardous microorganisms. This filtered water is appropriate to consume, or it may be dubbed as “the drinking water” following the cleansing operations and is secure for both cooking and drinking. [1]  [Potable water](https://www.sciencedirect.com/topics/chemical-engineering/potable-water) can be compromised due to a myriad of physical, operational, and environmental factors, such as contaminants intrusion into water pipelines, leaching, [disinfection](https://www.sciencedirect.com/topics/engineering/disinfection) byproducts, chemical or microbial permeation, and pollution. The prediction of potable water quality has seldom been researched, while a novel automated model that offers a [proactive approach](https://www.sciencedirect.com/topics/engineering/proactive-approach) can be developed to promote sustainability based strategies. [2]

The access to clean drinking water is a pressing global issue. Water scarcity, pollution, contamination, and the effects of climate change are some of the factors which exacerbate the problem. Inadequate water quality impacts industrial processes hence making it important to monitor water quality and also identify which source of water is potable. Access to clean drinking water is essential for improving people's lives, whether they live in small villages or large cities. One of the key benefits of this is the reduction of waterborne diseases, In rural areas, diseases like diarrhea, cholera, and typhoid are common and providing safe drinking water can help to prevent them. Clean water also plays a big role in the health and development of children which eventually leads to a better school attendance and more opportunities for education. From an economic perspective, having reliable access to water helps agriculture and industry thrive, boosting local economies. Additionally, easier access to clean water encourages more sustainable and eco friendly practices, helping to protect natural resources. Therefore, ensuring access to safe water has a positive impact on public health, the environment, and the well being of all communities on Earth.

The World Health Organization also emphasizes the need for innovative solutions to predict and monitor water potability. The traditional water testing methods require intensive labour, high budget and it often takes time to deliver the results. It is then machine learning comes in the picture for a real time water quality prediction which is essential for taking prompt actions in order to protect public health, prevent waterborne diseases and ensure safe water access to every individual.

The 17 United Nations' Sustainable Development Goals lay out a roadmap in order to address critical global issues. In particular, SDG 6: Clean Water and Sanitation aims to ensure access to safe drinking water and sanitation for all by 2030. To achieve this goal, the development of sustainable water management practices and machine learning can contribute by enabling predictive analytics and real time monitoring. This research aligns with SDG 6 by using machine learning to predict water potability. The unique feature of machine learning is that it can both identify a definite pattern in the water quality and also adjust to environmental changes. This capacity can be employed to find the possible sources of pollution and constantly improve the water quality. The blending of machine learning and remote sensing techniques provides a huge benefit for water quality remote monitoring. [3] Moreover, integrating Explainable AI enhances transparency in AI driven water quality monitoring systems. Techniques like SHAP and LIME provide insight into how individual features influence the model’s decisions, making them valuable for sensitive applications such as water quality monitoring.

**2. Literature Review**

**Machine Learning in Water Quality Prediction**

Several studies have applied machine learning to predict water quality and potability such as:

This is an experimental analysis of the different machine learning techniques that was carried out to create a generic water quality classifier.

Dalal, S., Onyema, E. M., Tavera Romero, C. A., Ndufeiya-Kumasi, L. C., Maryann, D. C., Nnedimkpa, A. J., & Bhatia, T. K. (2022). Machine learning-based forecasting of potability of drinking water through adaptive boosting model. *Chemistry & Biodiversity*, *19*(7), e20220187.

This uses deep learning methods to predict water quality.

Dalal, S., Onyema, E., Romero, C., Ndufeiya-Kumasi, L., Maryann, D., Nnedimkpa, A. & Bhatia, T. (2022). Machine learning-based forecasting of potability of drinking water through adaptive boosting model. *Open Chemistry*, *20*(1), 816-828.

In this paper, various models, including support vector machines and k Nearest Neighbours were employed to classify water quality, with performance evaluation based on accuracy and other metrics​.

These studies indicate that machine learning models can be used effectively for water potability prediction, especially when the datasets are large having the availability of multiple features.

**3. Methodology**

**3.1 Dataset**

The dataset, sourced from Kaggle, contains records of various water parameters such as pH, hardness, solids, chloramines, sulphates, conductivity, organic carbon, trihalomethane, turbidity, and the target variable Potability (where 1 indicates potable water and 0 indicates non potable water).

**3.2 Data Preprocessing**

Data preprocessing includes several steps to clean and prepare the data for modelling:

* **Missing Value Imputation**: Missing values in the dataset were handled using the mean imputation technique. When working with real world datasets, they often contain gaps or incomplete information. In many cases, some data points may be missing due to errors during data collection or other factors. To ensure that these missing values don't negatively affect the accuracy of a machine learning model, imputation techniques are used to fill in the missing data.
* **Feature Scaling**: The features were scaled using Standard Scaler to ensure that each feature has zero mean and unit variance. When dealing with algorithms that are sensitive to the scale of input features like k nearest neighbours and support vector machines, It involves transforming the features of a dataset so that they all have a similar scale, making it easier for the algorithm to learn patterns effectively.
* **Class Imbalance Handling**: Given that the dataset may have an imbalanced distribution of potable vs. non-potable water, we applied SMOTE  which stands for Synthetic Minority Oversampling Technique, to balance the classes. SMOTE creates synthetic instances by randomly selecting one of the nearest neighbours and creating a new data point along the line segment that joins the original data point and the selected neighbour. The new data point is formed by adding a weighted difference between the original point and its neighbour to the original point.
* **Principal Component Analysis (PCA)**: To visualize the data and reduce dimensionality, PCA was performed which reduced the dataset to two dimensions.

**3.3 Model Selection**

The following are the machine learning models which were applied to predict the potability of water:

* **Naive Bayes**: A probabilistic classifier based on Bayes' theorem. Naïve Bayes uses the features’ probabilities to find out the class to which a data point belongs. Using these probabilities, it predicts for each class and the class with the highest probability is selected.
* **Decision Tree**: A tree based model that splits data based on feature values. Each non-leaf node in the tree is linked to a feature from the dataset, and each terminal node is linked to a class. The tree partitions the dataset into similar subsets and uses these partitions to make predictions.
* **Random Forest**: A model that combines multiple decision trees for improvement in performance. This algorithm is built by aggregating many decision trees. Every tree is trained on a randomly chosen subset of the data. The model considers the majority vote from all the trees. By training each tree on different subsets, this algorithm helps minimize overfitting and enhances the model’s ability to generalize.
* **Support Vector Machine (SVM)**: A linear classifier that finds the hyperplane to classify best the data points and place this hyperplane to maximize the margin between classes.
* **K-Nearest Neighbours (KNN)**: A classifier that assigns a class based on the majority class of its neighbours. The KNN algorithm relies on distance metrics like Euclidean or Manhattan distance to calculate the proximity between data points. The value of k is set by the user to see the influence of those k nearest neighbours to determine the class or value of a data point.

**3.4 Model Evaluation**

The performance of each model was evaluated using the K Fold Cross Validation with 20 splits. The accuracy score was used as the primary evaluation metric. As seen in Figure 1., the model that performed the best, Random Forest was further tested on a separate test set.

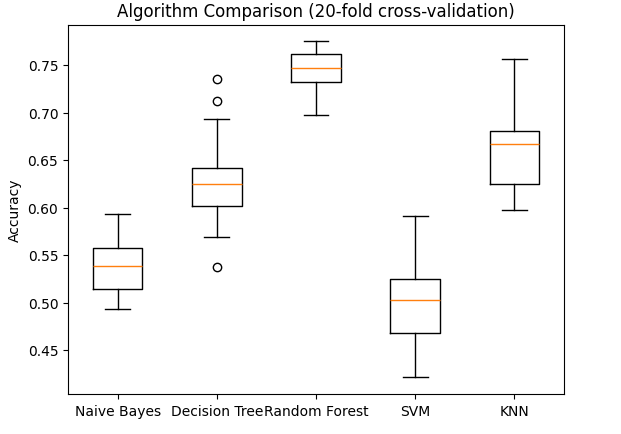


Figure 1.

**3.5 XAI Techniques**

To explain the predictions of the best-performing model (Random Forest), we applied **SHAP** and **LIME**:

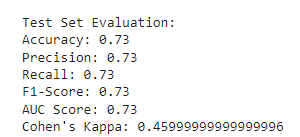
* **SHAP**: We used SHAP values to determine the feature importance and understand how each feature contributed to the model's decision.
* **LIME**: We used LIME to generate local explanations for specific predictions and visualize the impact of individual features.

**4. Results**

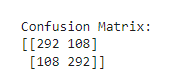
**4.1 Model Performance**

The performance of different models was evaluated using cross validation. The results showed that the Random Forest model achieved the highest accuracy at 73%, followed by KNN and Decision Tree models.

* **AUC:** The area under the curve The Area Under the Curve can be considered as the measurement of the area under the ROC curve. AUC measures the model performance for various thresholds by looking at the size of the area under the ROC curve. AUC can take any value from 0 to 1. A high AUC value indicates that the model has a good trade off between high true positives and low false negatives. This means that the model is skilled in classifying the data.
* **Precision:** Precision is a measure of how many of the cases that a classification model predicts as positive are actually true positive. Precision is the ratio of false positive predictions to the total number of positive predictions. A high precision value indicates that the model's positive predictions are reliable, while a low precision value indicates a high rate of false positive predictions.
* **Recall:** Recall is a quality index that is used to find out how well the positive samples have been recognized by the model. Recall is the proportion of true positive predictions and total positive samples. A high recall value tells us that the model is very good at detecting positive instances.
* **F1 Score:** The F-measure or F1 score is a gauge utilized to determine the success of a classification model by fusing two measures of precision and recall. The F-measure, which combines these two measures at a balanced point, evaluates how well the model can both minimize the number of false positive predictions and identify the true positives accurately.

****

* **Confusion Matrix:** The confusion matrix for the Random Forest model is as follows:



This matrix indicates a reasonable balance between true positives and true negatives, confirming that the model effectively differentiates between potable and non-potable water.

**4.2 Explainable AI**

The SHAP plot as in Figure2., visualizes how the interaction between two features here - Hardness and pH is affecting the whole model's predictions. Hence, this plot shows how the SHAP values of a feature change with respect to another feature.

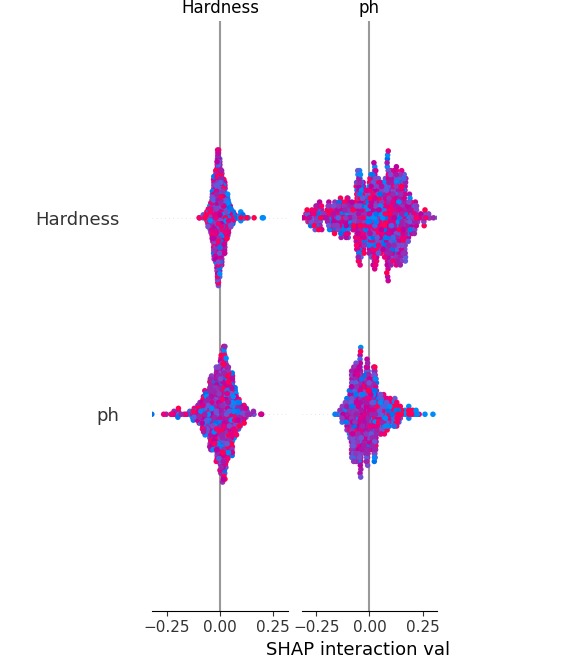


Figure 2.

**5. Conclusions:**

This paper demonstrates one of the applications of machine learning models for predicting the potability of drinking water and applies XAI techniques to interpret the model's decisions. The Random Forest model achieved the maximum accuracy of 73%, while SHAP and LIME were used to explain the feature contributions to the model's predictions. The combination of accurate predictions and model interpretability offers a solution which is promising for the monitoring of water quality and can be extended to other environmental applications too. Addressing class imbalance and handling missing values improved the accuracy of analysis and modelling. To address class imbalance, two sets of samples with different values that are - 0 and 1 were used for the 'Potability' column and were handled separately. Oversampling techniques like SMOTE were applied to increase the number of instances in the minority class, and the expanded minority group was combined with the majority class to create a balanced data frame. Future research can explore the inclusion of additional features, such as seasonal variations or geographical data, to further enhance the model's accuracy.

**6. Refrences:**

[1] Dalal, S., Onyema, E., Romero, C., Ndufeiya-Kumasi, L., Maryann, D., Nnedimkpa, A. & Bhatia, T. (2022). Machine learning-based forecasting of potability of drinking water through adaptive boosting model. *Open Chemistry*, *20*(1), 816-828.

<https://doi.org/10.1515/chem-2022-0187>

[2] Dawood, T., Elwakil, E., Mayol Novoa, H., & Gárate Delgado, J. F. (Year). Toward urban sustainability and clean potable water: Prediction of water quality via artificial neural networks.

<https://doi.org/10.1016/j.jclepro.2020.125266>

[3] Arias-Rodriguez, L. F., Author2, A. B., Author3, C. D., & Author4, E. F. (2021). Integration of remote sensing and Mexican water quality monitoring system using an extreme learning machine. *Sensors, 21*(12), 4118.

[https://doi.org/10.3390/s21124118](https://d.docs.live.net/c868662f7ec7dc8c/Documents/ https:/doi.org/10.3390/s21124118)