Water Quality Prediction

The consumption of water constitutes the physical health of most of the living species and hence management of its purity and quality is extremely essential as contaminated water has to potential to create adverse health and environmental consequences. This creates the dire necessity to measure, control and monitor the quality of water. The primary contaminant present in water is Total Dissolved Solids (TDS), which is hard to filter out. There are various substances apart from mere solids such as potassium, sodium, chlorides, lead, nitrate, cadmium, arsenic and other pollutants. The proposed work aims to provide the automation of water quality estimation through Artificial Intelligence and uses Explainable Artificial Intelligence (XAI) for the explanation of the most significant parameters contributing towards the potability of water and the estimation of the impurities. XAI has the transparency and justifiability as a white-box model since the Machine Learning (ML) model is black- box and unable to describe the reasoning behind the ML classification. The proposed work uses various ML models such as Logistic Regression, Support Vector Machine (SVM), Gaussian Naive Bayes, Decision Tree (DT) and Random Forest (RF) to classify whether the water is drinkable. The various representations of XAI such as force plot, test patch, summary plot, dependency plot and decision plot generated in SHAPELY explainer explain the significant features, prediction score, feature importance and justification behind the water quality estimation. The RF classifier is selected for the explanation and yields optimum Accuracy and F1-Score of 0.9999, with Precision and Re-call of 0.9997 and 0.998 respectively. Thus, the work is an exploratory analysis of the estimation and management of water quality with indicators associated with their significance. This work is an emerging research at present with a vision of addressing the water quality for the future as well.

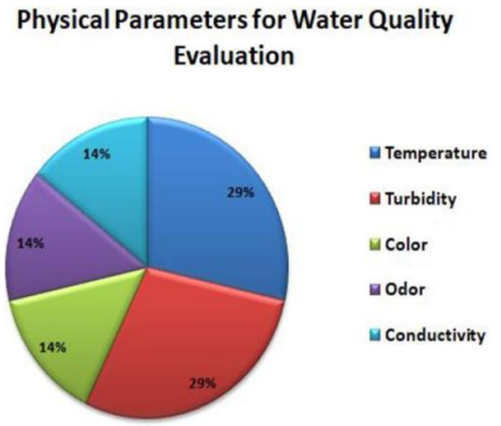
The major part of our earth comprises water and it is extremely important for the survival of all humans and animal species. Water makes up over 326 cubic metres of the planet's surface, which is almost 71% of its total area out of which 97% is seawater. Only 0.5 percentage of the drinkable water on earth is accessible, while the remaining 2.5 percentage is either trapped in glaciers, polar ice caps, the atmosphere, on soil, is polluted, or lies beneath the earth's surface far beyond human reach. If the global water supply is 100 L, consequently the amount of drinking water would be only 0.003 L, which is just a teaspoon. Therefore, the management and preservation of drinking water is regarded as a top priority. It is the most critical issue for mankind to address given the extremely limited amount of water that is accessible for use. The quantum of water around the world is represented in Table 1.

Water is a common and crucial resource shared among all humans, animals, and plants and is a necessity for all species. Each one of these species has its own respective needs for water quality. Total Dissolvable Solids (TDS) of soft water for human consumption range from the best quality stated, which is between 50 mg/dL and 150 mg/dL. Between 150 mg/dL and 300 mg/dL is the next level that can be applied to humans. The plants need water that is between 700mg/dL and 800mg/dL. The animals, especially cattle consume water around the quality of 1000 mg/dL. It is thus evident from all these observations that water quality management is essential to ensure.

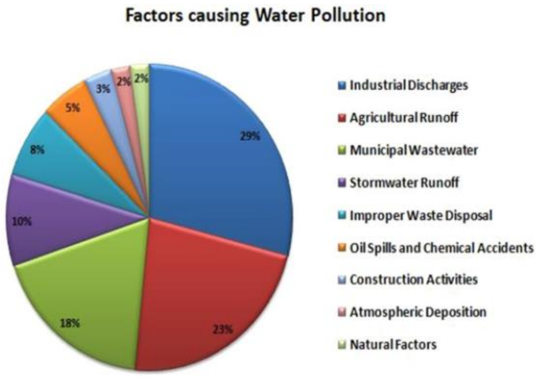
Location	Quantity (%)
Oceans	97.2
Ice Caps/Glaciers	2.0
Groundwater	0.62
Freshwater Lakes	0.009
Inland seas/salt lakes	0.008
Atmosphere	0.001
Rivers	0.0001

Table 1. Water availability around the globe.

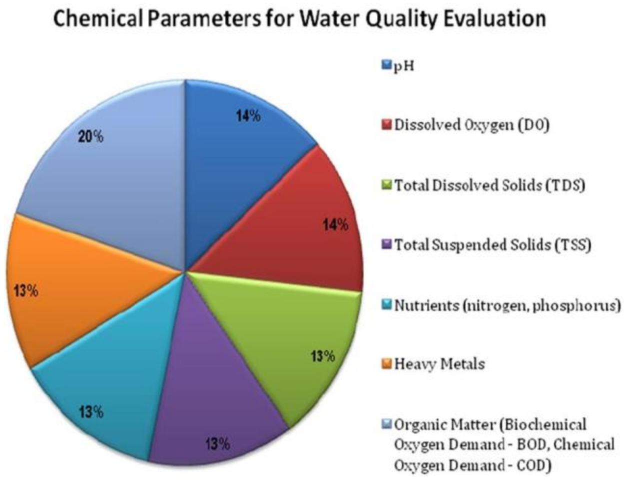
sustainability and a healthy life on Earth. The impact of water quality prediction is crucial at a global level for many reasons. First of all, to get clean and safe water is a basic human necessity and water quality prediction aids to guarantee the availability of potable water for societies worldwide. Water quality is related to public health as polluted water may cause waterborne diseases which could affect millions of humans globally. A sustainable envi- ronment is an important aspect of human well-being by preserving ecosystems and biodiversity. The significance of water quality assessment is profound and intricate by various organizations globally. The WHO (World Health Organization) , UNEP (United Nations Environment Programme), EPA (United States Environmental Protec- tion Agency), EEA (European Environment Agency), IWA (International Water Association) and WEF (Water Environment Federation) are fanatical for water quality assessment and addressing the mitigation strategies for water quality challenges. Water quality creates impact on public health globally and resulting in dissemination of waterborne diseases like typhoid, dysentery, cholera, dengue and malaria and cause substantial risks worldwide. The advancement in computing technologies and artificial intelligence have elevated the standards of water quality assessments[1]. Measurements and estimations about the quality of the water have become easier to calculate and accurate, especially with the development of Industry 4.0 standards and Internet of Things (IoT) sensors. With the integration of IoT sensors, AI solely serves as a supporting tool to automate water quality checks. Classification and Regression models based on machine learning help in determining the water quality. Depending on the outcomes, classification results tend to be binary or multi-classified. Real-time sensor data are collected, given feature labels, and then classified based on the importance of the feature labels. Earlier, these measurements used to be carried out with fuzzy-based decision support systems with subjective decision-making models.



Physical Parameters for Water Quality Evaluation



Factors causing Water Pollution



Chemical Parameters for Water Quality Evaluation

Statement of objectives

The proposed work offers a comprehensive analysis and white-box description of the classification problem for water quality. The framework incorporates extensive pre-processing of the dataset to ensure it fits into the XAI model. The proposed approach employs both model-based and model-agnostic interpretations, using model-based ML. Donnelly et al. implementations and model-agnostic XAI implementations. The quality of water is greatly challenged by innumerable influencing factors. These factors vary from condition to condition and place to place. For example, Microplastics (MP) are emerging pollutants in the marine environment with potential toxic effects on littoral and coastal ecosystems and as well as identifying the mixing of pollutants in water sources. The laboratory evaluations show the presence of polyethene (PE) particles in the waves of the ocean with wave steepness Sop of 2–5%. The transportation of which could cause severe water pollution on the seashores. These measurements require quantification and feature analysis when it is evaluated with AI. This is where the XAI plays a vital role in measuring the order and degree of the pollutants causing the quantifiable pollution in the water.

System model and architecture

Worldwide, numerous water bodies are contaminated by a variety of anthropogenic and natural processes, resulting in a variety of health problems for human life. Thus water quality requires rigorous monitoring and management to prevent pollution. In accordance with WHO guidelines, the polluted water must be treated using the proper water treatment techniques before consumption. The quality of water is contaminated by the incessant addition of toxic chemicals and microbes and also by the relentless addition of local and industrial sewage sludge, trash, and extra hazardous waste that are toxic to humans and society. Many uncertainties are required to be quantified for all machine learning models. The uncertainties such as selecting and gathering the training data, absolute and accurate training data, understanding the machine learning models with performance bounds and drawbacks and finally the uncertainties which are based on the operational data. In current years, the validation of water quality has taken active momentum because of ever-increasing water pollutants which spoil water that is dedicated for domestic use and irrigation. Machine Learning techniques play a substantial role in identifying the quality of water using explainable AI.XAI model is implemented in the framework wherein LIME and Shapely are used to provide explainability and interpretability to the results generated by the machine learning model .

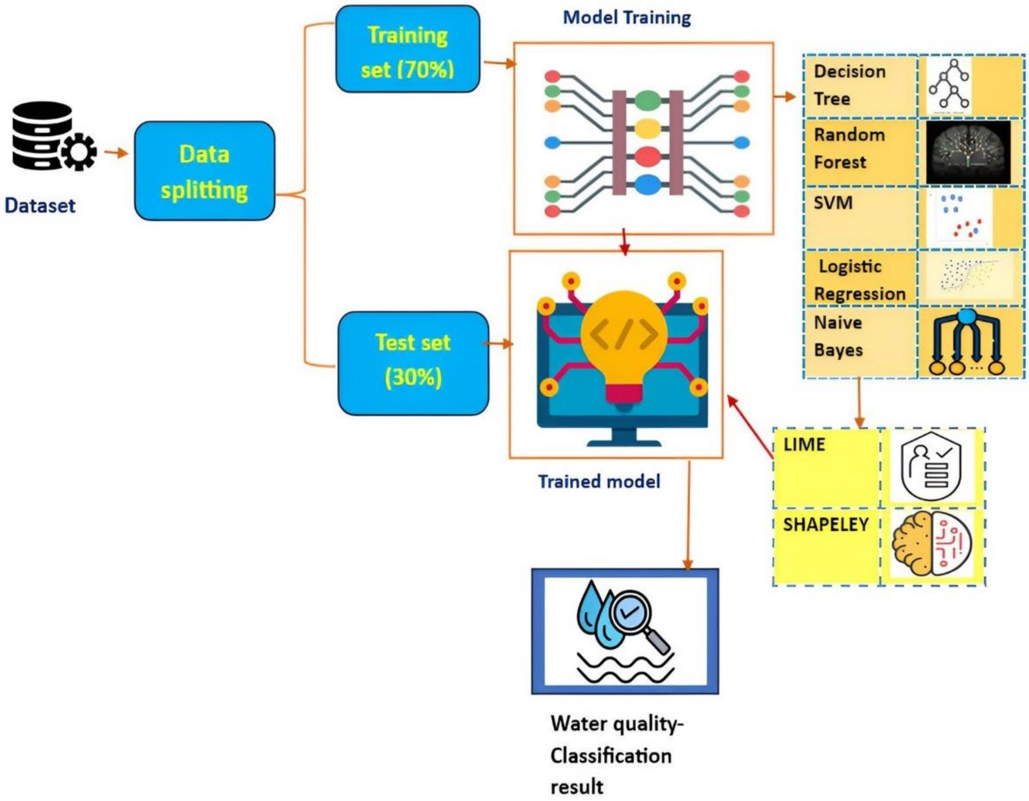
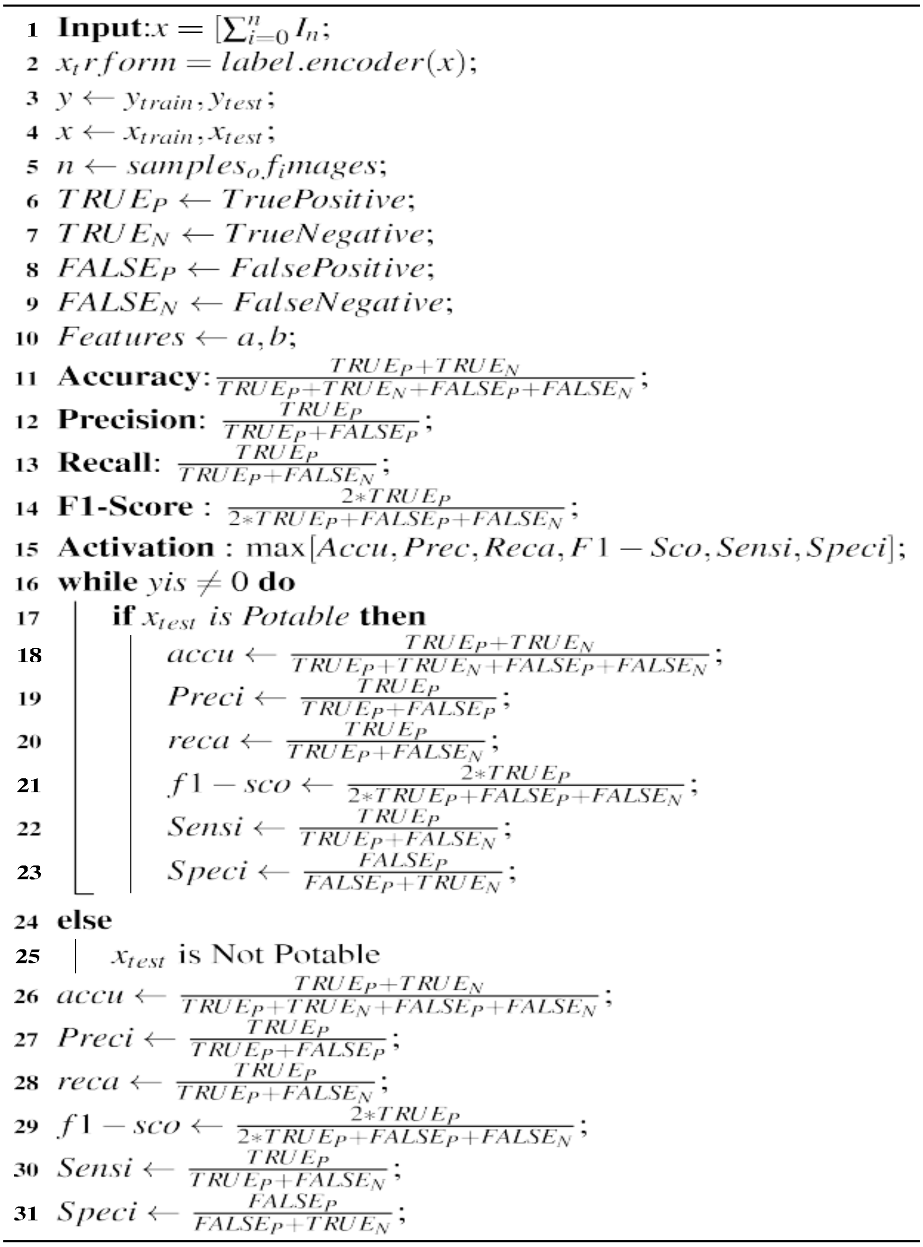


Figure : Interfacing ML algorithms with XAI.

Algorithm

In this section two algorithms are discussed: one for the algorithm-based evaluation of water quality 1 and another for the algorithm-based explanation of water quality 2. These two algorithms provide a holistic analysis and explanation of water quality management.



```
Input: $x = [\sum_{i=0}^{n} I_n$;
$x_t rform = label.encoder(x)$;
$y \leftarrow y_{train}, y_{test}$;
$x \leftarrow x_{train}, x_{test}$;
$n \leftarrow samples_o f_i mages$;
$TRUE_P \leftarrow TruePositive$;
$TRUE_N \leftarrow TrueNegative$;
$FALSE_P \leftarrow FalsePositive$;
$FALSE_N \leftarrow FalseNegative$;
$Features \leftarrow a, b$;
Accuracy: $\frac{TRUE_P + TRUE_N}{TRUE_P + TRUE_N + FALSE_P + FALSE_N}$;
Precision: $\frac{TRUE_P}{TRUE_P + FALSE_P}$;
Recall: $\frac{TRUE_P}{TRUE_P + FALSE_N}$;
F1-Score : $\frac{2 * TRUE_P}{2 * TRUE_P + FALSE_P + FALSE_N}$;
Activation : $\max[Accu, Prec, Reca, F1 - Sco, Sensi, Speci]$;
while $yis \neq 0$ do
    if $x_{test}$ is $Potable$ then
        $accu \leftarrow \frac{TRUE_P + TRUE_N}{TRUE_P + TRUE_N + FALSE_P + FALSE_N}$;
        $Preci \leftarrow \frac{TRUE_P}{TRUE_P + FALSE_P}$;
        $reca \leftarrow \frac{TRUE_P}{TRUE_P + FALSE_N}$;
        $f1 - sco \leftarrow \frac{2 * TRUE_P}{2 * TRUE_P + FALSE_P + FALSE_N}$;
        $Sensi \leftarrow \frac{TRUE_P}{TRUE_P + FALSE_N}$;
        $Speci \leftarrow \frac{FALSE_P}{FALSE_P + TRUE_N}$;
else
    $x_{test}$ is Not Potable
$accu \leftarrow \frac{TRUE_P + TRUE_N}{TRUE_P + TRUE_N + FALSE_P + FALSE_N}$;
$Preci \leftarrow \frac{TRUE_P}{TRUE_P + FALSE_P}$;
$reca \leftarrow \frac{TRUE_P}{TRUE_P + FALSE_N}$;
$f1 - sco \leftarrow \frac{2 * TRUE_P}{2 * TRUE_P + FALSE_P + FALSE_N}$;
$Sensi \leftarrow \frac{TRUE_P}{TRUE_P + FALSE_N}$;
$Speci \leftarrow \frac{FALSE_P}{FALSE_P + TRUE_N}$;
```

Algorithm 1. Algorithm for water quality classification

Challenges

The proposed work may be influenced by the following challenges which are described in detail as follows,

Global unity

For the successful implementation of the system, a unanimously accepted implementation is essential. Unfortunately, water quality estimation and related research are limited to consideration of specific datasets acquired for a particular region, wherein the generated results may differ with the changes in geographic location. Thus the generated results can never be considered suitable on a global scale. The parameters that influence the water quality may also vary across the world, and hence the proposed work can never be considered as a universal solution.

Training and re-training

The qualifying attributes that determine the quality of water vary across the globe and hence the proposed model needs to be re-trained[69] when applied to a new environment of study. This would allow the model to unlearn and re-learn new environments. On the contrary, the complexity of the model would also increase. The accuracy and other performance metrics which are measured in the proposed work may drastically decrease as well in a different environment of study. Thus applying this model to versatile environments is complex and would be a challenging task.

Subjective or quantitative

The trade-off from subjective analysis (which was done through fuzzy-based methods in the form of the Analytical Hierarchy Process (AHP) and The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)) has improved the performance and ability to classify the models with better accuracy. However, the involvement of a subject matter expert is a missing point in the current research. Despite all the implementation and analysis from an engineering perspective, the involvement of an environmental scientist in any aspect of water research would contribute towards the enhancement of research quality.

Confusing solids

The proposed work identifies Solids as the primary influencing factor that affects potability. In real-world applications, solids can be of any form. For example, in sewage water treatment plants it can be either mud, Fat-Oil-Grease(FOG), or any other substances. The attributes of research are too complex to handle in real-life scenarios, which acts as an inevitable yet detrimental impact.

Environmental challenges

Water resources are under serious threat due to water scarcity, water contamination, water conflicts and climate changes. Chemical and the municipal wastewater contaminates the water and endangering the life of the aquatic organisms and affect their ability to reproduce. This also makes them an easier prey to their predators. The food cycle and livelihood of the human is also greatly affected by the water contamination. Chemical substances make the water hard to recycle and consume by reducing the regeneration ratios.

Water quality and industrial sustainability

The era of Industry 5.0 focuses on the consumer centric industrial evolution with the idea of environmental sustainability. The futuristic technologies evolve with the improvement of technical viability, with the mission

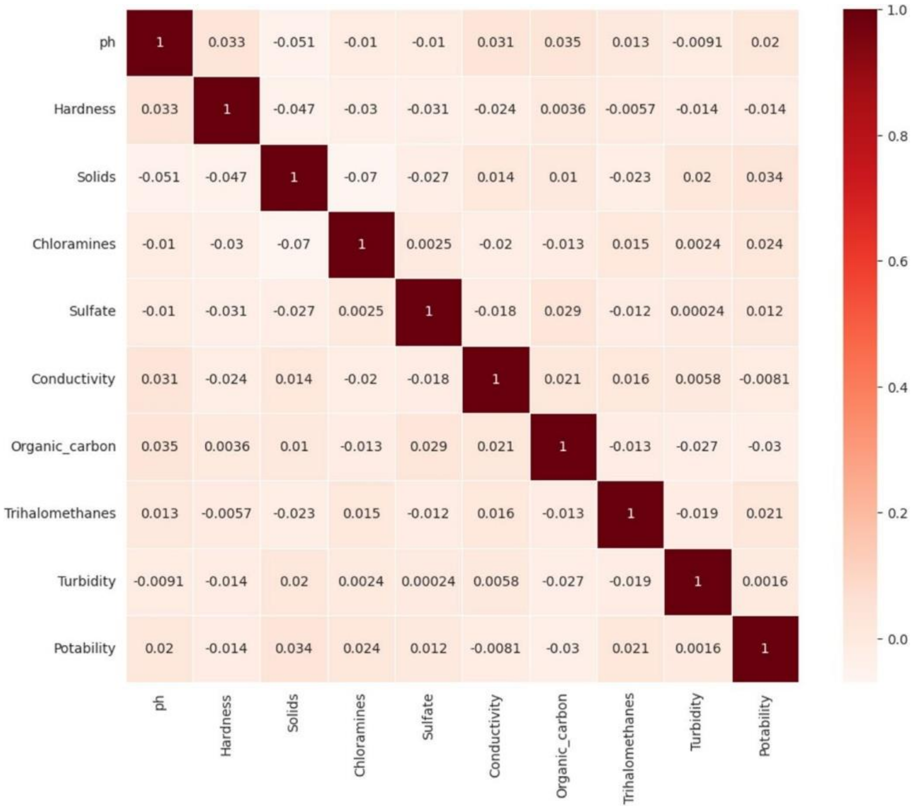


Figure: Correlation analysis for water quality attributes.

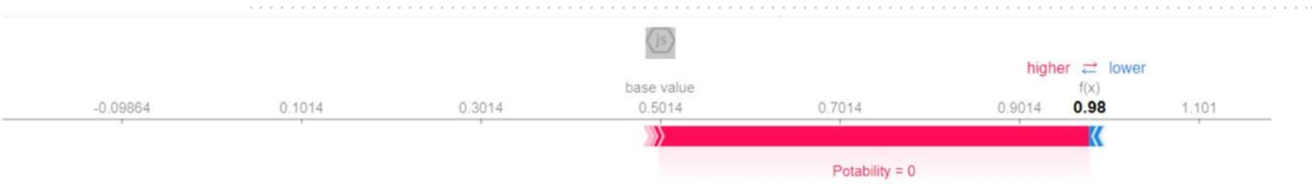


Figure : Force plot for water quality

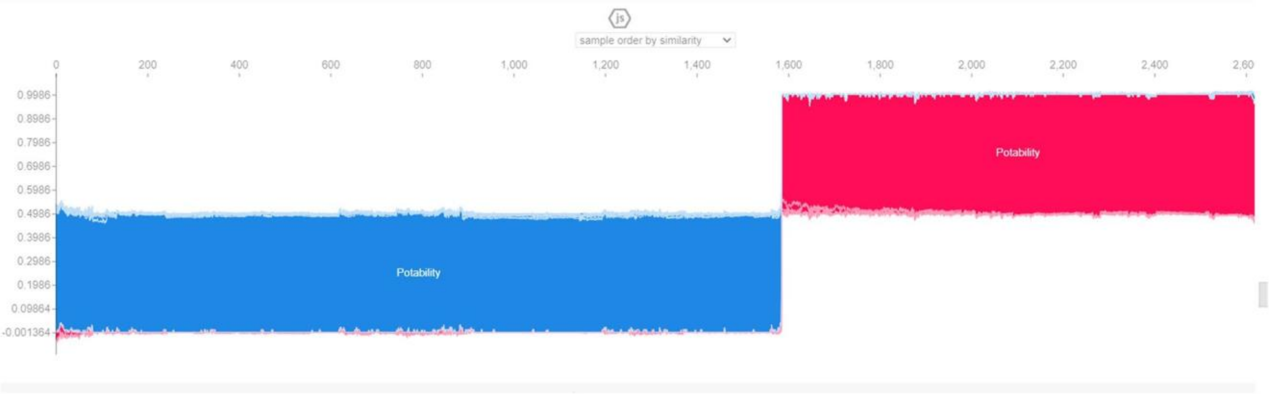


Figure : Test patch for potability

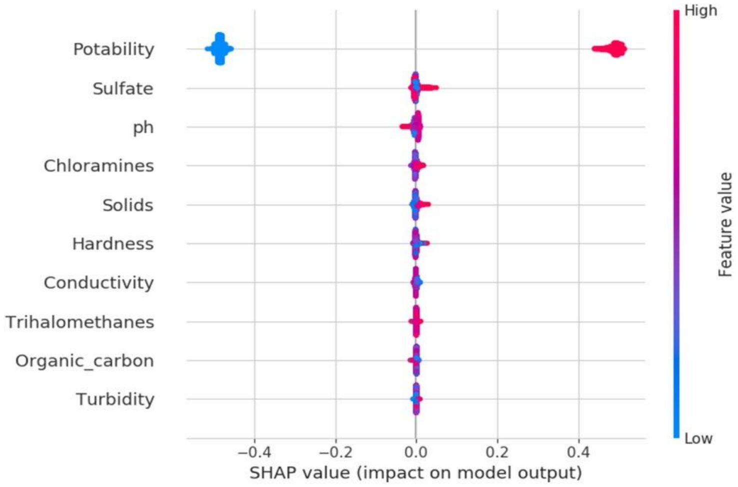


Figure : Summary plot for potability.

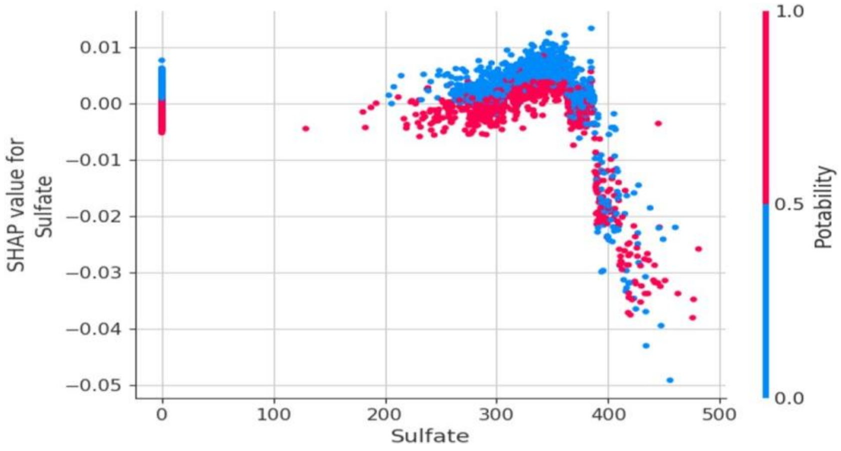


Figure : Dependency plot for potability.

Research finding of the proposed work

The following items are presented as the findings are outcomes of the proposed work

->The proposed work performs an exploratory analysis with XAI implementation providing an ability to improve the reliability of machine learning models providing explanation and transparency to the classification process.

->The proposed work acquires data from a single dataset, where the performance of classification yields optimized results. This result may vary if the model is subjected to a different dataset constituting different features and instances.

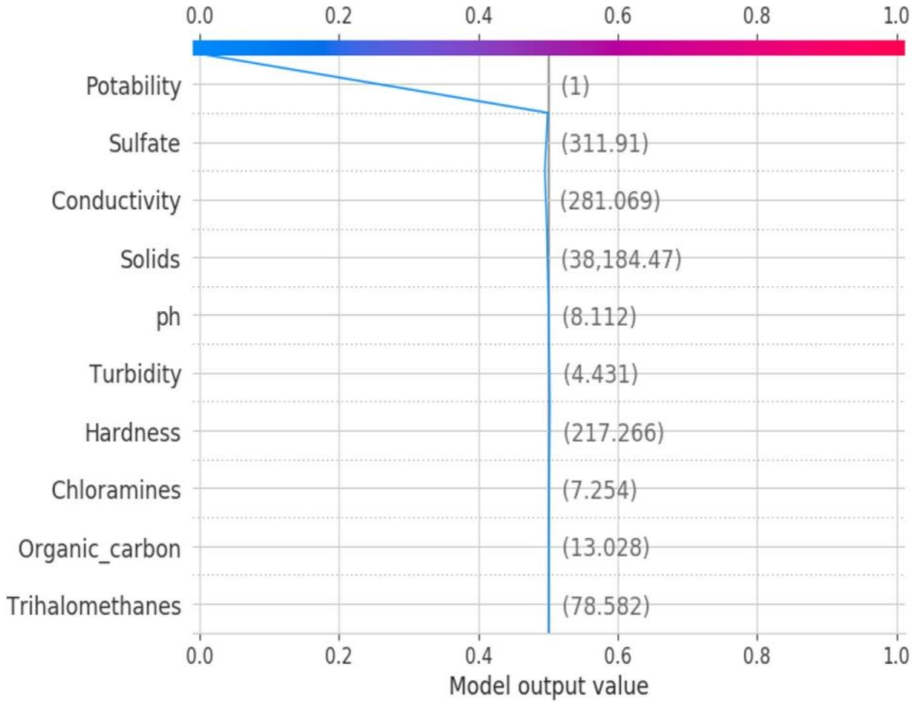


Figure : Decision plot for potability.

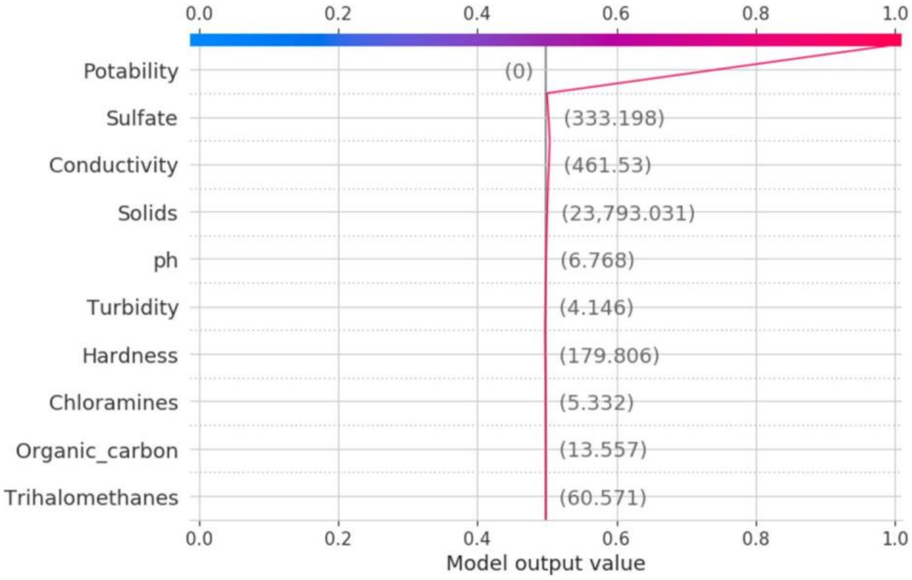


Figure: Decision plot for potability

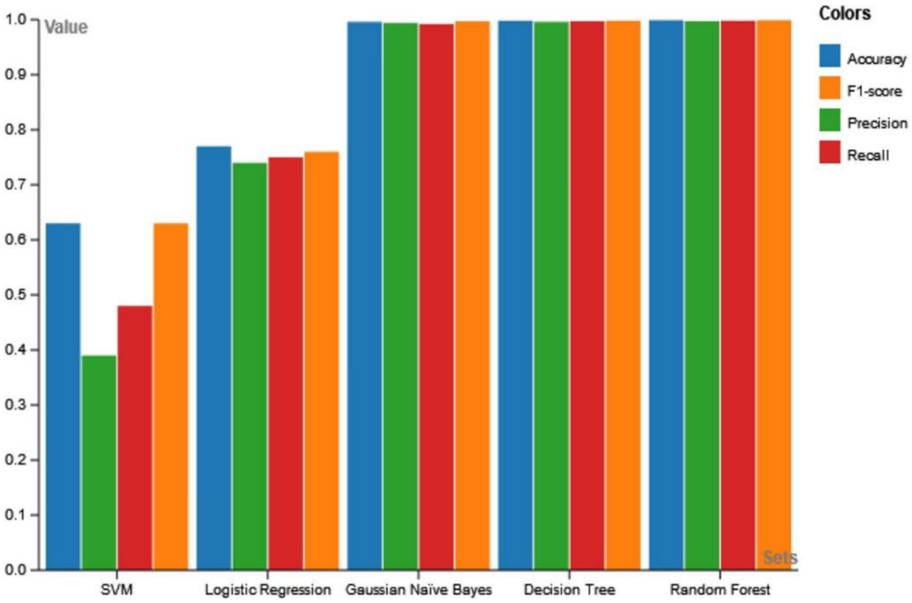


Figure : Comparative analysis of machine learning models used.

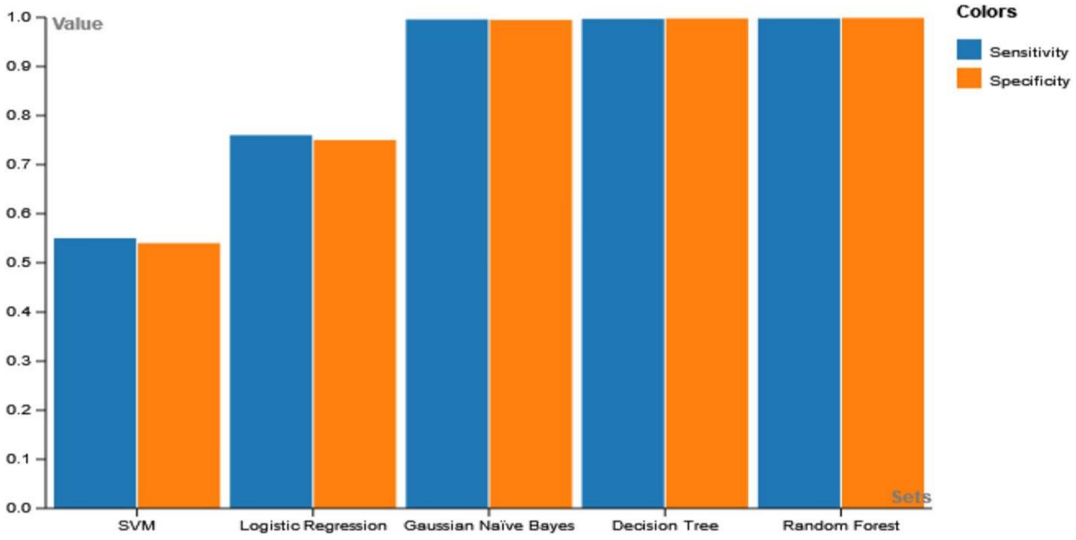


Figure 17. Comparative analysis of sensitivity and specificity.

- The XAI reveals the most significant features contributing towards classification results and also explains the same.
- The best fitting machine learning model is chosen for the explanation through an exhaustive analysis and evaluation of all the models considering the essential performance metrics. Thus the results produced by SHAPELY can be considered as the most reliable and acceptable.
- The proposed work also suggests the importance of the subject matter expert, which can extend the usability of the proposed model at the universal level.
- The predictions of the proposed work with the support of an explainer, helps end users and consumers to understand the quality of the water they use.
- The features related to the classification and explanation, can be further controlled to diminish the levels of chemicals and pollutants in water recycling.
- Total dissolvable solids quantification and the feature weights for the same determine the levels of filtration and carbon purification required in the recycling plants.
- The proposed work brings insights of pollutants on the seashore and how the explainability can support the impurity estimations for such conditions also.

Conclusion

Water quality management impacts almost all aspects of life on earth and clean water is a basic necessity. The proposed work is extremely relevant in this regard wherein an exploratory analysis conducted to analyze and control the factors that deteriorate the quality of the water. The impact of these factors is explained using XAI models. The contribution of the XAI model lies in its ability to explain the role of the underlying parameters towards the classification of water being potable or not, based on their relative importance and unique properties. The XAI model uses SHAPELY considering the probabilistic prediction generated from the Random Forest classifier. This RF model in this regard is chosen as it yields the highest accuracy of 0.999 with sensitivity and specificity of 0.999 and 0.998, which is found to be superior in comparison to the other state-of-the-art models considered in the study. This justifies the reason for the RF to be selected for XAI implementation. The proposed model identifies the parameter “solid” as the most significant in terms of its impact on the potability of water. The proposed model yields optimized and explainable results considering the dataset used in the study. Future work may involve more complex and heterogeneous datasets to generate predictions. In such scenarios, the metric evaluations may differ. The usage of deep learning algorithms could further enhance the examination the solid sediments and generate classification results based on their mass, dimensions, and shape. The use of XAI in such a model would ensure a better explanation of factors relevant to the solid sedimentation in water.