Applying Machine Learning for Ensuring Sustainable Management of Water (SDG6)

Mohd Talha[1], Namrata Nagpal[1] and Meenakshi Srivastava[1]

[1] *Amity Institute of Information Technology, Amity University Uttar Pradesh, Lucknow Campus, India,*

Abstract

Clean water is one of the necessities for the sustainability of the life of living beings. But with various affluents mixed in water, the condition of getting clean water for drinking is becoming hazardous. There is a lot of water pollution in various sources of water, specifically potable water. The physical and chemical characteristics of potable water are typically assessed, and the acceptable range of every factor needs to be determined. This can be achieved by incorporating the use of machine learning algorithms that evaluate the given factors like Ph, Chloramines, Sulfate, and Turbidity found in water. This research paper investigates ten factors required to assess water quality using six different machine learning algorithms and describes the best way to ensure Sustainable Management of Water, thus enabling sustainable development goal 6. Random Forest was able to give the highest accuracy of 82% among all the machine learning algorithms used.

Keywords

Machine Learning, Sustainable Management, SDG6, sustainable development goals, water treatment, potable, Ph, water treatment, Random Forest

1. Introduction

Clean water is one of the necessities for the sustainability of the life of living beings. According to the United Nations, Sustainable Development Goal 6 (SDG6) is all about ensuring the availability of drinking water and sustainable management of water and sanitation for all the people of the world. 74% of the world's population has managed to have access to clean drinking water during 2020 which is an increase from 70% in 2015 [1]. To achieve SDG6, the United Nations has set targets for the year 2030. Special attention needs to be given to providing universal access to drinking water. By 2030, the water quality must be improved by reducing pollution, avoiding dumping, minimizing the release of hazardous chemicals in river water from factories, and increasing the recycling and reuse of water [2].

To improve the water quality, the elements of water first need to be assessed using a variety of chemical and physical factors. As numerous factors affect water quality, its analysis is a challenging task. This concept is closely related to the various ways in which water is used. Water quality prediction is a topic of extensive research. The physical and chemical characteristics of water are typically assessed about the water's intended usage [3]. An acceptable and unacceptable range for each factor should be determined. Water is deemed suitable for a given purpose when it complies with the established requirements.

AISD 2023: First International Workshop on Artificial Intelligence: Empowering Sustainable Development September 4-5, 2023, co-located with International Conference on Artificial Intelligence: Towards Sustainable Intelligence, (AI4S-2023), Pune, India

talha.mohd.shaikh1@gmail.com (M. Talha); nnagpal@lko.amity.edu (N. Nagpal); msrivastava@lko.amity.edu (M. Srivastava)

https://dblp.org/pid/64/5313-1.html (N. Nagpal); https://dblp.org/pid/340/0794.html (M. Srivastava)

If the water doesn't satisfy the given standards, it needs to be treated before being used. Water quality can be assessed using a variety of physical and chemical factors. Because of this, it is impractical to appropriately quantify water quality on a spatial or temporal basis by looking at each variable's behavior separately. The more challenging method is to combine the values of a variety of physical and chemical components into a single value. The quality value function of each variable's index, which was frequently linear, represented the correspondence between the variable and its quality level [4]. The calculations were made using physical variables taken from water samples or direct measurements of a substance's concentration. The main goal is to investigate the feasibility of using Machine Learning (ML) Algorithms to predict the quality of the water.

Water systems can improve their performance by using machine learning to learn from their prior actions. It is quite like how the human brain develops information and comprehension to comprehend things, domains, and the relationships among them, machine learning requires input, such as training data or knowledge graphs. Deep learning can start once entities are defined. Machine learning is based on observations or data in the form of narratives, anecdotes, or directions. To make decisions based on the examples that are presented, it searches for similarities in the data. The main objective of ML is to enable computers to alter their behavior based on what they have learned on their own, without any help from humans.

The water quality measures that were used in our study to assess the overall water quality in terms of potability are Solids, Hardness, Conductivity, Organic Carbon, Trihalomethanes, Ph, Chloramines, Sulfate, and Turbidity. These factors are utilized as feature vectors to represent the quality of water. The machine learning algorithms used in determining the level of water quality are Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), XGBoost, and K-Nearest Neighbour (KNN) classification methods [5]. According to the results of several types of classifiers and approaches, the K-Nearest Neighbour (KNN) classifier performs better than other classifiers. The findings demonstrate that potability may be well predicted quite precisely using machine learning approaches.

2. Literature Review

The process of improving the water quality for drinking purposes is a global process that requires attention for the sustainability of mankind on earth. The study on water quality is very old and was described by Horton who created the Water Quality Index (WQI) in the 1060s. Since then, there have been many methods to carry out the said process. The two indices used to assess the overall quality of drinking water sources are simple, flexible, and stable, and they are only somewhat sensitive to the input data. Similarly, the weighted mathematical WQI technique provides information on water quality [6]. These WQIs, despite having serious limitations, are the most used evaluation methods of water quality because they help with understanding water quality by converting many factors into a digital number.

According to studies, at least 2 billion people consume dirty water tainted with excrement throughout the world. The microbial contamination brought on by fecal contamination has become the biggest threat to the safety of drinking water. Clean water is important for avoiding many overlooked tropical and diarrheal diseases. Safe drinking water makes it easy to practice cleanliness and prevent various serious respiratory infections.

High-quality water is defined as being free of dangerous organisms and biological substances that might be unsightly. It has no flavor or smell and is clear and colorless. It is devoid of chemical concentrations that might be bad for the body, ugly to look at or cost you money. It does not cause excessive or undesirable deposits on water-conveying systems like pipes, tanks, and plumbing fixtures since it is non-corrosive.

Traditional machine learning methods such as Logistic Regression (LR), RF (Random Forest), DT (Decision Tree), Support Vector Machine (SVM), XGBoost, and K-Nearest Neighbour (KNN) Classifier were used to predict the quality of the water, with Random Forest having the highest accuracy [7].

Guangtao et al. in their work discussed deep learning methods being used for analyzing urban water and wastewater management [8]. Similarly, Mahalakshmi and Yogalakshmi tried to predict water quality for Indian rivers using similar machine learning methods [9].

Ahmad, Anwar, and Irfan in their work described the water prediction using supervised learning methods. They proposed methods that took 663 samples using 12 spring water samples and assessed WQI [10]. Gakii and Jennifer in their work depicted a classification model for water quality prediction using decision trees. The water samples were taken from Kenya that were used to indicate clean drinking water for residents [11]. Alexander and Bridget in their work discussed how big data and machine learning can be used to get benefits from water management and the environment. The paper discussed various methods and applications of big data for water management [12]. Mohamad Sakizadeh in his work predicted water quality by analyzing various parameters using artificial neural networks. The variables used by them were mainly groundwater variables collected from wells and springs in Iran [13].

Ruixing et al. discussed utilizing machine learning methods for predicting water quality using various water quality indicators. They also talked about water purifying techniques and evaluating the toxicity of natural water systems [14]. Krishnan et al. [15] in their work depicted the use of artificial intelligence (AI), deep learning, and the Internet of Things (IoT) approach to managing water resources smartly. They developed a methodology that aims to use all technologies to provide sustainability to water usage from natural resources.

Castillo et al. in their work depicted the classification of water quality in Mexico rivers by simply predicting water quality based on WQI obtained from the ecosystem. They used supervised machine learning for their study and experimentation [16]. Jinal Patel et al. used various machine learning approaches using SMOTE (Synthetic Minority Oversampling Technique) to maintain a balance of the dataset. Experiment results showed that gradient boost gave 81% results. The best features were determined by explainable AI (XAI) [17].

The focus of the work is water quality; every component of the dataset, including conductivity, turbidity, organic carbon hardness, sulfate, and trihalomethanes, needs to be examined [18]. Using these indicators and comparing them to established values is a crucial limitation when estimating water quality.

3. Methodology

The methodology used to assess the sustainability of water talks about detailing the strategy of water quality assessment (See Figure 1). The process needs to be evaluated from time to time as new factors keep emerging in the scene. With the latest technologies being used, the methodology tends to deviate and give better results than its previous versions.

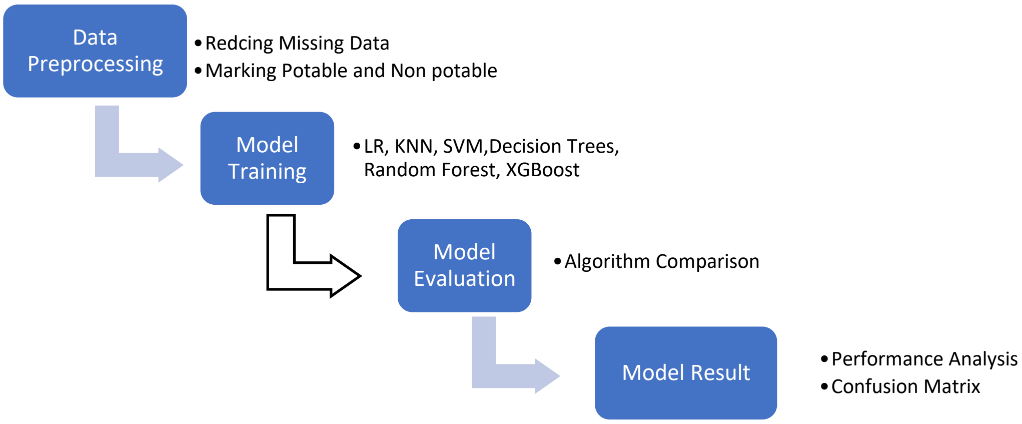
The study aims to assess the water quality in the best possible ways focusing on the prime factors like turbidity, solid, sulfate, trihalomethanes, pH, Hardness, solids, organic carbon, etc. The water quality assessment is necessary to conclude if the given water samples are potable or not.

The potability of water refers to the validity of water quality and whether it is worth drinking or not.

The primary strategy involves working in four phases, namely:

1. **Data Preprocessing:** It involves performing the initial analyses of the raw data. When data are gathered and converted into useful information, data processing takes place. It is crucial that data processing is done appropriately to prevent harming the final output of data. This procedure entails record inconsistency detection, data transformation, and source data rectification. Python 3.7 is used for data exploration analysis (EDA) which is used to import and manage the raw data. Data preprocessing entails converting unstructured data into well-formed sets of data so that algorithms for data mining can be used. Raw data frequently has irregular formatting and is unfinished. Data Preprocessing and Data modeling were done using the SKlearn package after the data was determined to be consistent, with 70% of the dataset going towards training and 30% going towards testing.

2. **Model Training:** The dataset is then scaled to ensure that the data points are within a suitable scale so that lower value ranges do not predominate while calculating data point distances. Model training can now be done with the dataset using the six algorithms that support vector modeling, decision tree classification, k-nearest neighbor classification, and the XGBoost algorithm. The modeling process went through iterations, with default settings.


Data Preprocessing

- Redcing Missing Data
- Marking Potable and Non potable

Model Training

- LR, KNN, SVM,Decision Trees, Random Forest, XGBoost

Model Evaluation

- Algorithm Comparison

Model Result

- Performance Analysis
- Confusion Matrix

Figure 1: Methodology for Water Quality Assessment (WQA)

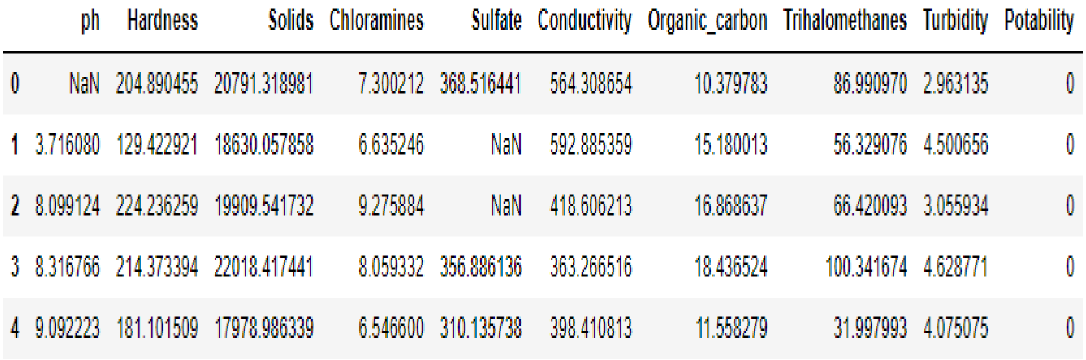
3. **Model Evaluation:** This process deals with the major analysis phase where the execution details of all six algorithms are analyzed individually. The dataset is divided into the ratio of potable and non-potable data. The results obtained from each algorithm are compared to one another leading to acquiring some meaningful results that would finally conclude the fate of water quality; if the water is good enough to be used by global people or requires some cleanliness.

4. **Model Result:** A confusion matrix is created using Python that gives output w.r.t. all 6 machine learning algorithms in detail. The matrix also determines the accuracy, recall, and F1 Score of all the methods applied in assessing the water quality. Type II error is also calculated to determine the potability or quality of water. Based on the results obtained from the confusion matrix and graphical analysis, required results are drawn that help to predict the best-suited method of all.

3.1. Factors for Water Quality Assessment

To determine whether the water is fit for drinking, ten features from the selected dataset will be used. The 10 variables given below in Figure 2 that describe the water's quality are pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, and finally Potability [19]. Below is a list of these properties, values, and associated definitions.

- **pH-**The pH measure determines the amount of hydrogen ions present in a solution and distinguishes between acidic, basic, and neutral media, such as water. Drinkable water must have a pH that falls between 6.52 and 6.83, according to the WHO's recommendations.
- **Hardness-**Hardness is caused by magnesium and calcium salts that build up in the geologic environment of flowing water. How much hardness is present in raw water depends on how long it has been in touch with materials that cause hardness.
- **Solids-** When we talk about dissolved solids in water, we're talking about salts like potassium, magnesium, calcium, bicarbonates, sodium, chlorides, etc. In addition to impacting the safety of the water, the presence of these dissolved solids in it changes the flavor. TDS concentrations in drinking water should be kept at 500 mg/l or below and should not exceed 1000 mg/l.
- **Chloramines-** When treating water and sanitizing it against bacteria and other pathogens, chloramine is frequently used in conjunction with chlorine. Drinkable water shouldn't have more than 4 mg of chloramine per liter for safety's sake.
- **Sulfate-** Sulfates are organic compounds that occur naturally and are present in soil, food, minerals, groundwater, plants, and rocks. However, the chemical industry uses them extensively. In freshwater, the sulfate concentration should range from 3 to 30 milligrams per liter.
- **Conductivity-** Water's ability to conduct electricity is determined by its electric conductivity. Since it does not conduct electricity, pure water is referred to as an insulator. Ionic chemicals present in ionic water, however, cause it to have a higher electric conductivity.
- **Organic Carbon-** Total organic carbon (TOC) is the term used to describe the total quantity of carbon derived from organic elements in water. This organic carbon can come from unnatural synthetic sources or the decomposition of real organic matter. Average organic carbon concentrations in potable water should be less than 2 mg per liter, and in treated water, they should be less than 4 mg per liter.
- **Trihalomethanes-** THMs, or Trihalomethanes, are molecules that are frequently present when water is treated with chlorine. The amount of THMs depends on several factors, including the water's organic content, the amount of needed chlorine, and the temperature of the water being treated. The THM value must be under 80 ppm for water to be fit for human consumption.



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

Figure 2: Features for Accessing the Potability of Water

- **Turbidity-** The term "turbidity" refers to the state of water, including whether particulates are suspended in it. Water's ability to emit light, which serves as the benchmark for waste disposal in terms of colloidal matter, can be used to compute the turbidity of water. The World Health Organisation recommends a turbidity level of 5.00 NTU.
- **Potability-**The term "potability" is used to indicate whether water can be consumed or drunk by people. It should be considered whether using the same water to water plants is appropriate. While a value of 0 indicates that the water is unfit or non-potable for human consumption and a value of 1 indicates that the water is potable or drinkable.

3.2. Machine Learning Algorithms for Water Quality Assessment

There are many different machine learning algorithms and technologies available currently. The machine learning algorithms that are used in our study predict the quality of water by analyzing different parameters as discussed in the above section. The six machine learning algorithms can be explained here.

- **Support Vector Machine:** One of the fundamental techniques, support vector machine is mostly employed in this study as a baseline to compare the model performances. This algorithm's fundamental concept makes use of the linear model family. The original vector is transferred into a higher-dimensional space where the model is trained by looking for the dividing hyperplane in this space with the largest gap. The approach is predicated on the idea that the average classification error will decrease the greater the difference and spacing between these parallel hyperplanes.
- **Random Forest:** Leo Breiman and Adele Cutler created the random forest (RF) model near the end of the 1990s. The method is repeated until the desired number of decision trees is constructed, selecting a random subset of variables at each split of the observed sample data. A bootstrap sample chosen with replacement from the observed data serves as the foundation for each tree, and majority voting is used to aggregate all of the trees' predictions. One of the features that this model offers is the ability to execute a feature selection, which not only makes the process simpler and lowers the processing costs of the analysis, but also makes it easier to grasp the relationships between variables and the dependencies between them. The weighted sums of the absolute regression coefficients serve as the basis for this variable importance measurement.
- **K-Nearest Neighbour:** In this study, missing data was imputed (changing missing values with the closest available value) using the k-nearest neighbors' technique, which is one of the fundamental and intuitively simple algorithms for tasks like classification. In general, any model can be used for imputation; however, in this study, the KNN algorithm is selected because it can give sufficient outcomes while maintaining computing expenses within reasonable limitations. The model's main concept is that the missing value of a measurement is attributed to the value that is most common among its predetermined group of neighbors.
- **Decision Tree:** Models for classification and regression use non-parametric supervised learning techniques called Decision Trees. The objective is to learn straightforward decision rules based on the data features to build a model that predicts the value for a target variable. Regression is a predictive modeling technique; hence these trees are used to categorize data or predict future events. Decision trees, like flowcharts, feature a root node with a particular data question that relates to branches with potential answers. Following the branching, the decision (inner) nodes increase the number of outputs and queries they pose. This continues until the data ultimately reaches a terminal node (also known as a "leaf") and comes to an end. Boolean examples, like yes or no, are typically categorized using the decision tree technique.
- **Logistic Regression:** The statistical method of logistic regression computes the probability that a binary event will occur, just like linear regression does. The model uses log odd ratios and an iterative maximum likelihood approach to forecast the likelihood that a classification will occur based on the independent variables present in a dataset. LR calculates

the target's probability based on the input features. This method is frequently applied for binary classification jobs in the field of water quality and can be extended to address multiclass classification problems.

- **XGBoost Classifier:** Gradient boosting is a method for fixing mistakes in older models by building new ones. The final prediction is created by combining the results. A highly efficient machine learning technique called XGBoost is now widely used to forecast the potability of water. Its impressive capacity to handle large and complex datasets and ability Its wide use in this field is mostly due to its ability to generate accurate results for several classification and regression tasks. XGBoost incorporates several DTs into a model as a decision tree-based ensemble learning technique. The main advantage of XGBoost for forecasting water potability is its good handling of missing values, which enables it to handle real-world water quality data without the need for time-consuming pre-processing.

4. Experimental Setup

A dataset of 3276 samples taken from Kaggle was used for the proposed investigation [20]. Nine water quality characteristics were measured for each sample by analysis of the following parameters: pH, Organic Carbon, Chloramines, Turbidity, Trihalomethanes, Sulfate, Hardness, Conductivity, and Solids. To make analysis easier, the dataset was divided into 70:30 training and test data ratios.

4.1. Data Preprocessing

While preparing data for further experimentation, data preprocessing was required to be done for better analysis and testing. Preprocessing involves cleaning of data such that better quality of data yields better results.

The data were first converted from a string to a float as part of the pre-processing of the data. In addition, redundant data were removed, leaving only the relevant information behind (See Figure 3). The metrics considered for the study are conductivity, the degree of hardness, sulfate, trihalomethanes, pH level, turbidity, and solids. To get a reliable outcome that will decide whether the water being tested is potable or not, the study takes the cumulative behavior of all the factors into account.

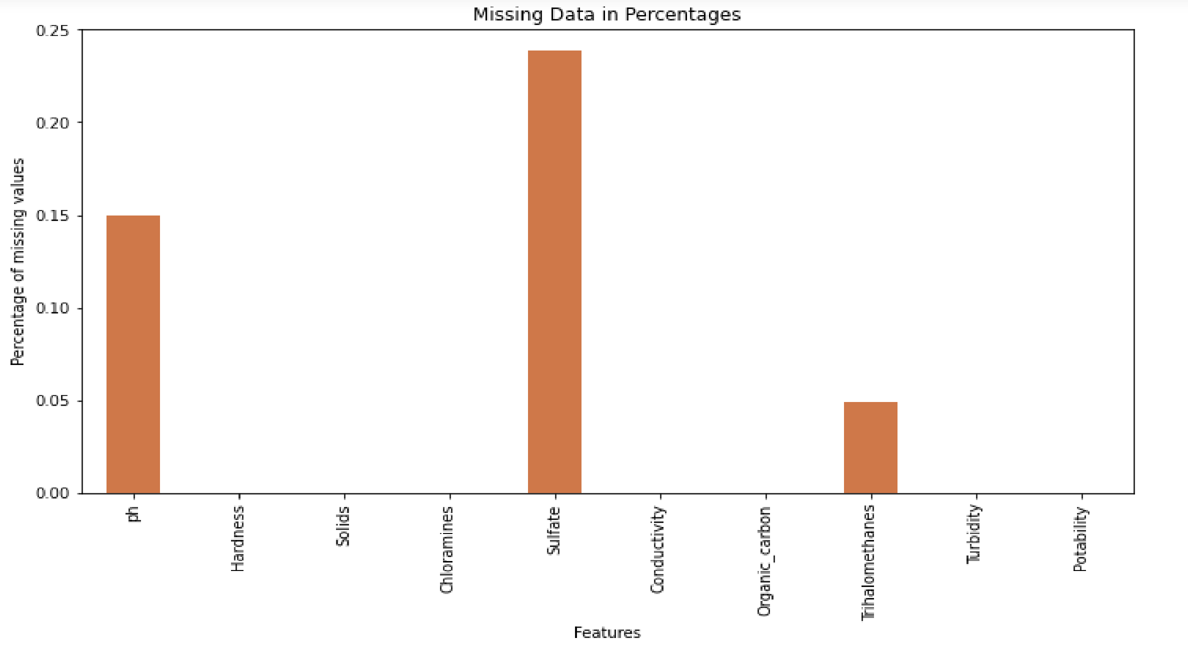


Figure 3: Preprocessing showing necessary Data Values

The null values in the data were replaced by taking the mean or average of the given categories. The Water Quality Index (WQI) is then precisely calculated to assess water quality to fulfill the goal.

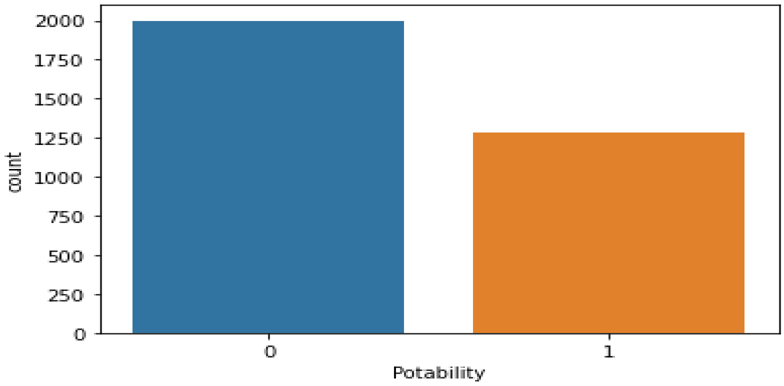


Figure 4: The dataset showing water potability

As a more accurate depiction, a histogram (See Figure 4) of the dataset is taken which enables a clear data distribution of the entire set. The dataset's whole range depicts water potability values of 1 and unpotable water with a 0 value.

Both non-potable and potable records exhibit a normal/Gaussian distribution pattern when the distribution of the data within each of the predictor variables is plotted (See Figure 5). The plot depicts a bell-shaped curve, except for the solids being somewhat right-skewed. This helps to make decisions during the data cleaning stages and informs that the data distribution is acceptable without forcing one to reject any of the predictor variables.

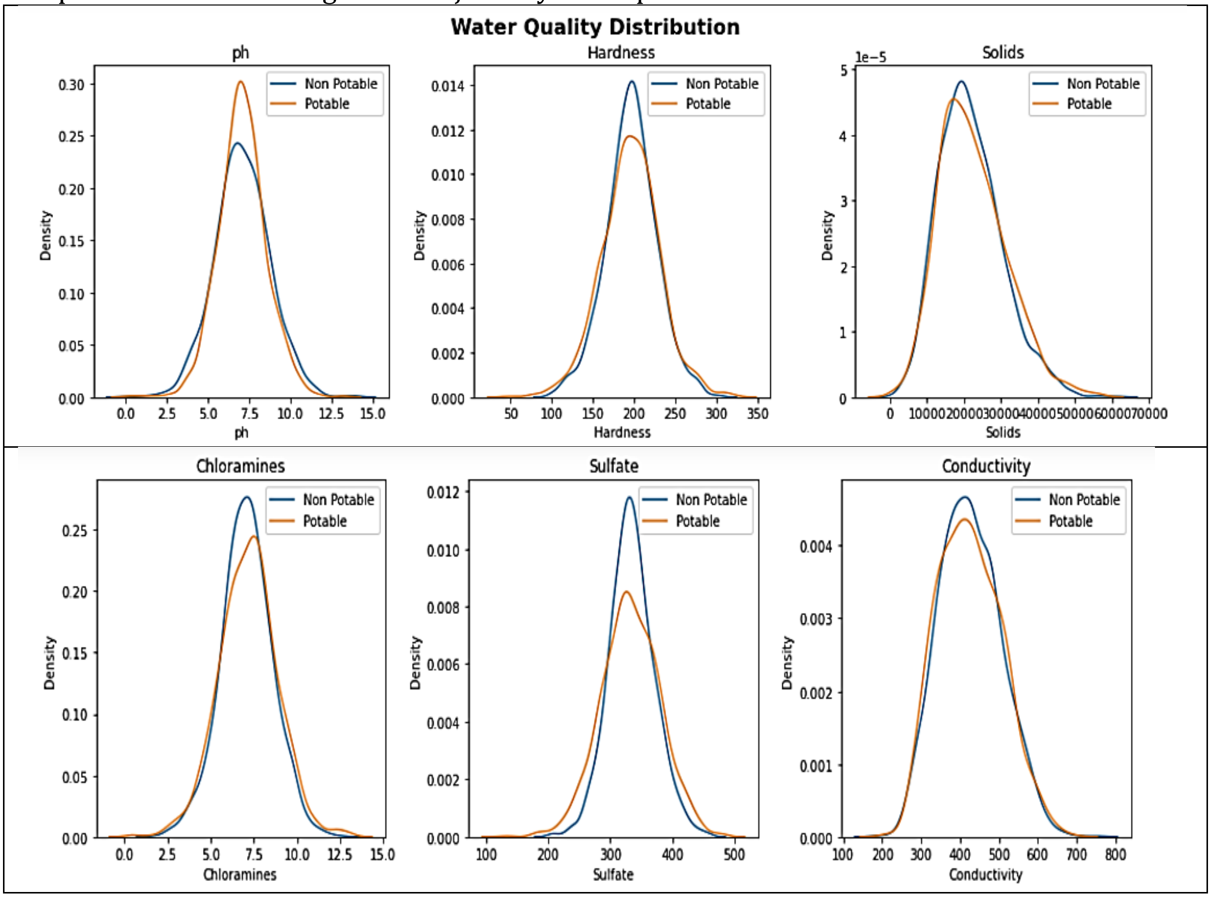


Figure 5: The distribution graph of each predictor variable separating non-potable records from records with potable water.

A box and whisker plot, often known as a box plot, is also used to visualize the data distribution of the nine predictor variables. The distribution plots' depiction of the skewness in Solids also allows us to confirm it (see Figure 6).

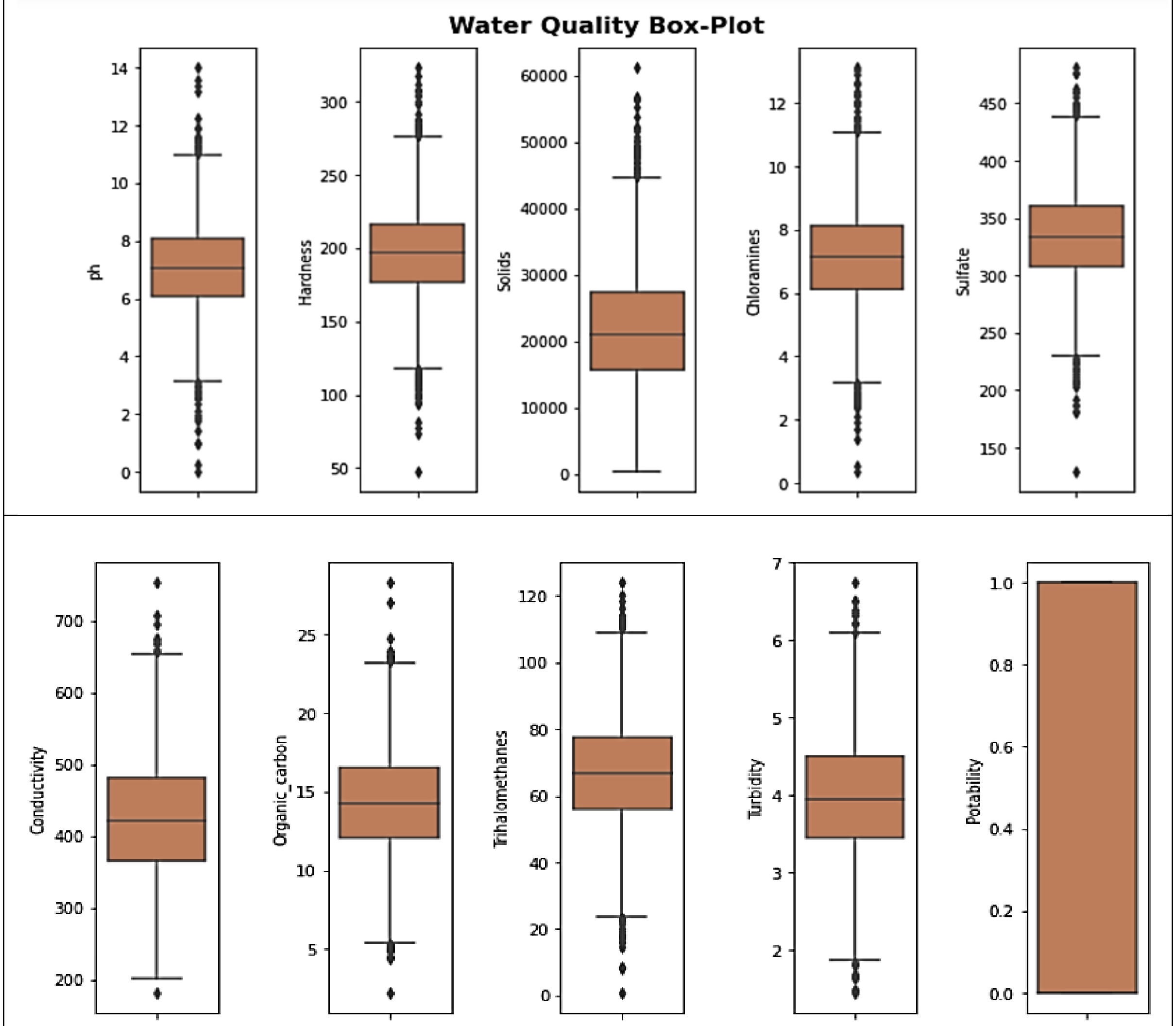


Figure 6: Box plot showing outlier data points for each predictor variable.

Since the dataset's distribution is primarily Gaussian, as mentioned above, the outlier data outside of three standard deviations are excluded without eliminating all the remaining outlier data (outside the maximum and minimum). This enabled to saving of part of the outlier data while still preserving the consistency of the dataset's variability. Again, the Solids variable showed the greatest visual change, and it also showed the greatest skewness (See Figure 7).

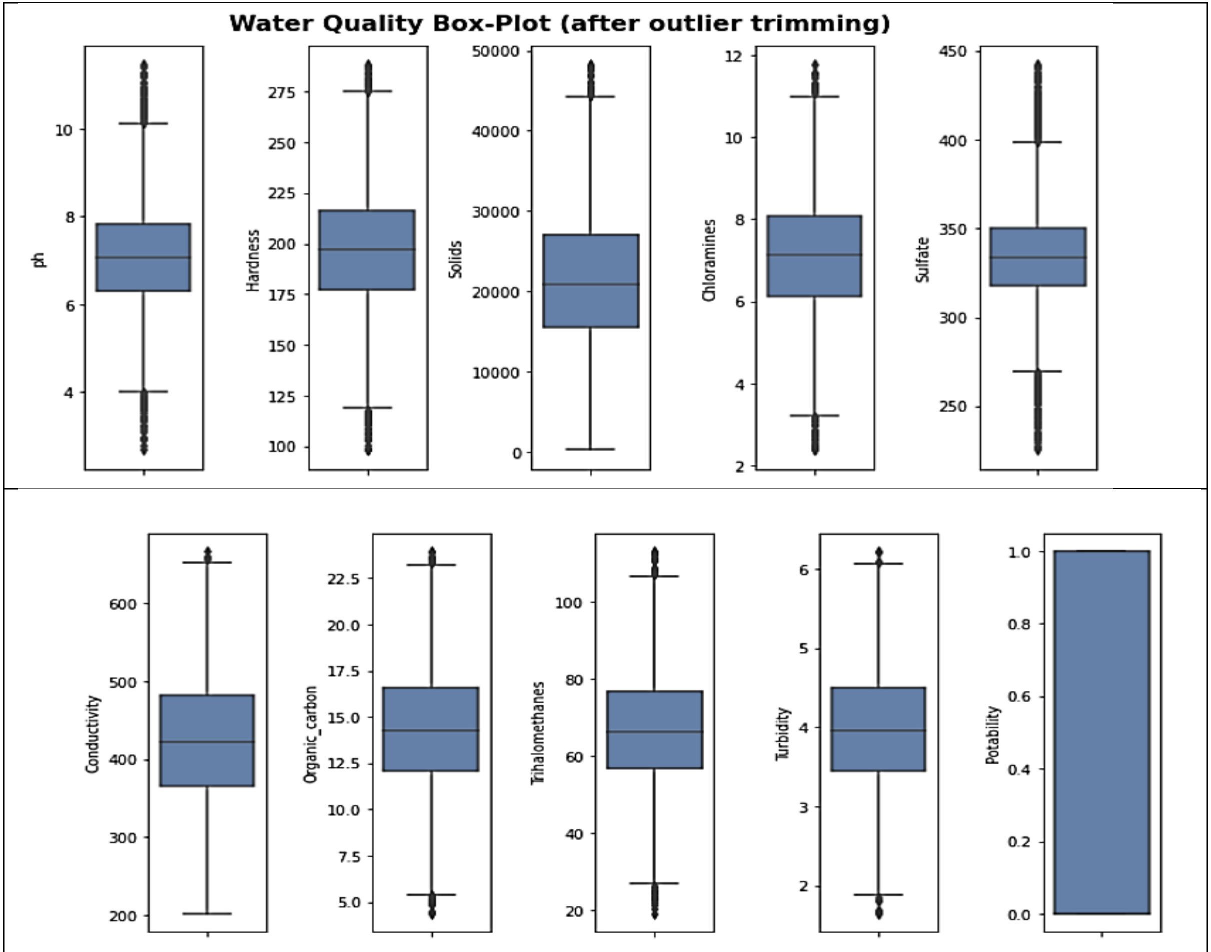


Figure 7: Box-plot showing the dataset after removing outliers with a standard deviation >3

After managing the missing data points and identifying outliers using trimming, 1198 potable records and 1930 non-potable records were found in total. After using the resampling technique, the balanced data set was split into 1930 potable and 1930 non-potable values as shown in Figure 8.

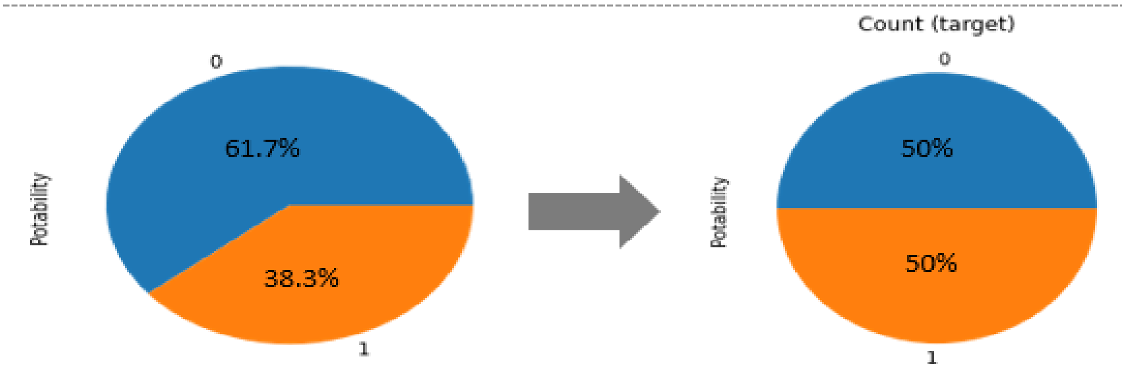


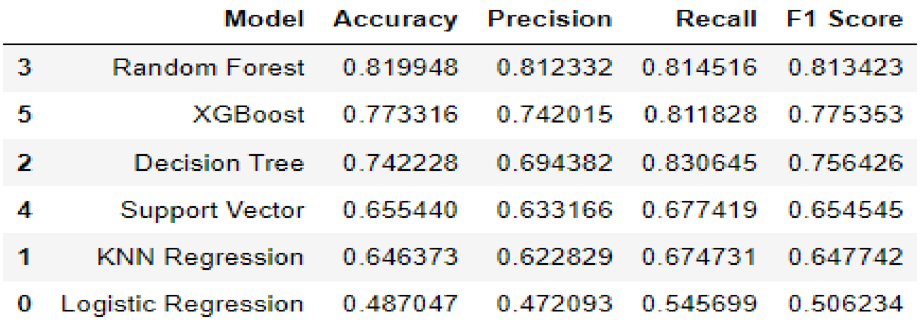
Figure 8: Pie chart comparing statistics from unbalanced and balanced classes

The data is then divided into training and test sets, thereby machine learning algorithms were used to train the dataset, and the accuracy of the models was then compared. Finally, using the accuracy rating of suggested models, results were further compared.

5. Result & Discussion

All six algorithms were run with simple default settings of their respective functions in the initial modeling iteration. In terms of statistics, type II errors are more dangerous, especially when determining the potability or quality of water. A community that uses contaminated water would suffer if a false positive result was returned.

This study used a confusion matrix to combine many performance indicators to assess the effectiveness of different algorithms, including logistic regression, KNN, RF, XGBoost, and SVM.



	Model	Accuracy	Precision	Recall	F1 Score
3	Random Forest	0.819948	0.812332	0.814516	0.813423
5	XGBoost	0.773316	0.742015	0.811828	0.775353
2	Decision Tree	0.742228	0.694382	0.830645	0.756426
4	Support Vector	0.655440	0.633166	0.677419	0.654545
1	KNN Regression	0.646373	0.622829	0.674731	0.647742
0	Logistic Regression	0.487047	0.472093	0.545699	0.506234

Figure 9: Performance analysis of the proposed algorithms

Considering the information in Figure 9 it was clear how the Random Forest algorithm performed better than the other methods, displaying outstanding outcomes. Its 81.99% accuracy rate, 0.812 precision, and 0.81 F1 score indicate perfect classification performance with fewer mistakes as compared with other algorithms.

Reviewing the confusion matrices from Figure 4 reveals that the Decision Tree has the largest type II error (40.03%), and the Logistic Regression has the lowest type II error (26.30%).

The evaluation metrics for each method during the modeling depict that the Random Forest algorithm performed the best with the highest accuracy score of 81.99%, while the Logistic Regression approach performed the worst, with an accuracy score of 48.70%.

Figure 10 shows the confusion matrices for each of the six ML methods that were employed in the study. Here 1 stands for potable water and 0 for not potable in the confusion matrix.

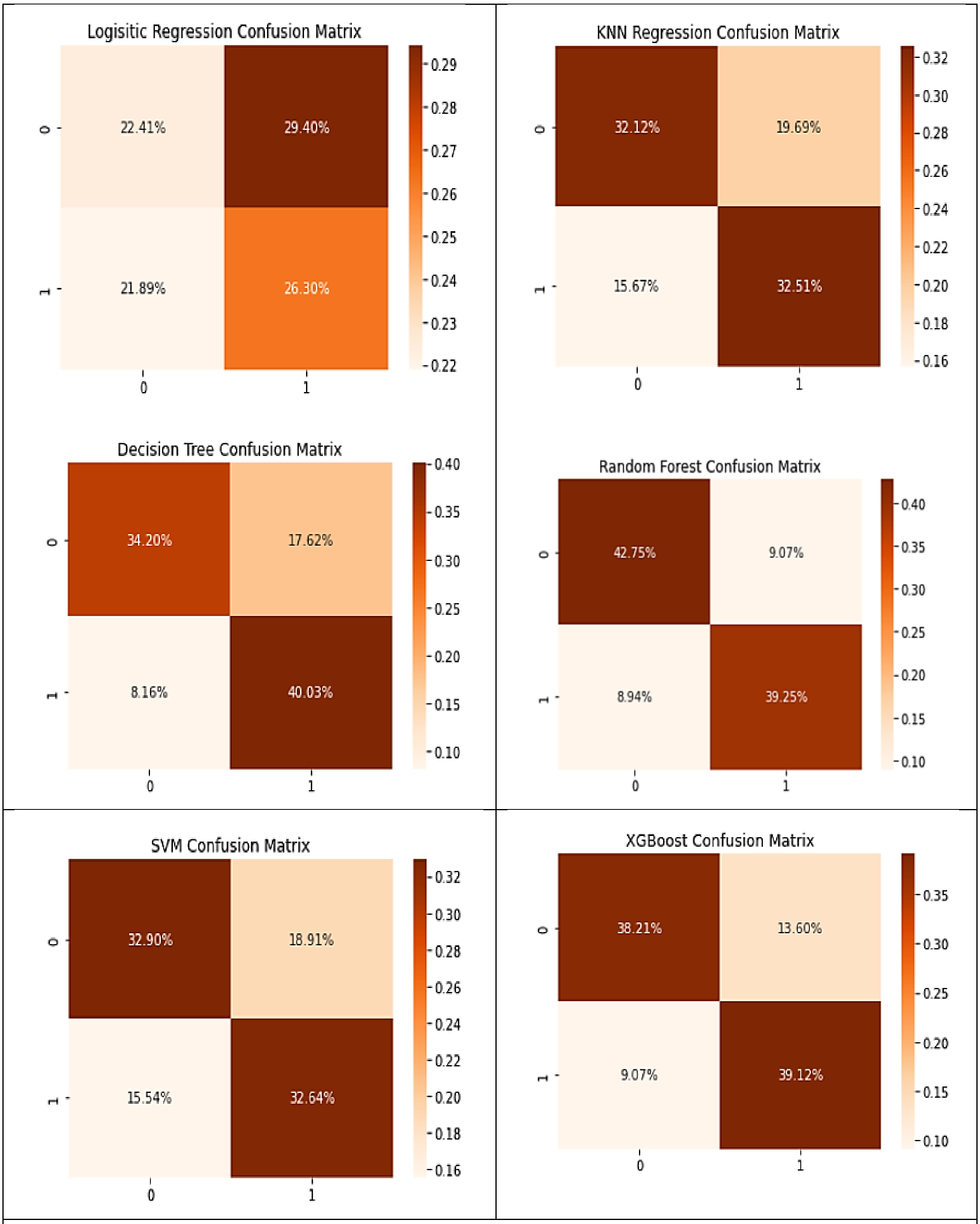


Figure 10: Confusion Matrix of All ML Algorithms

When it comes to machine learning, the F1 Score is crucial for evaluating the efficacy of binary classification models. This suggested metric gives a combined score that successfully integrates the equilibrium between these two metrics and provides a holistic predictive capacity by concurrently taking precision and recall into account.

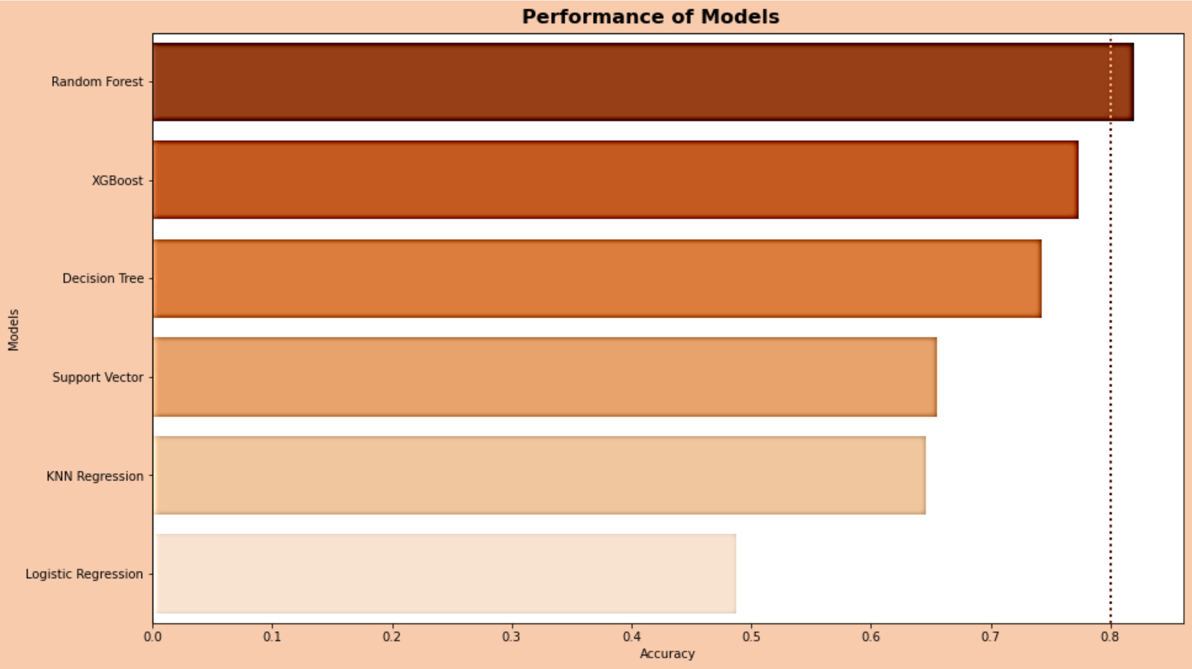


Figure 11: Performance of Algorithm by Accuracy Score

The study offered machine learning algorithms' accuracy scores, which are represented in a bar graph in Figure 11. The chart indicates that Random Forest, XGBoost, followed by Decision Tree (DT) had the best accuracy rate.

In contrast, among the algorithms, the accuracy rate for the Logistic Regression model was the lowest. After model training, Logistic Regression performed the lowest while XGBoost, Decision Tree had close accuracy values around 77.33% to 74.22%. Support Vector Machine had a 65% accuracy & KNN Regression had 64% accuracy while Random Forest was able to be accurate to 82% which is the highest accuracy among all the algorithms that are used.

6. Conclusion

To protect human health, it is crucial to ensure that drinking water is safe and pure. The whole world is working towards achieving this sustainability development goal (SDG6). So, to accomplish this goal, accurate water potability prediction is essential. The algorithm that demonstrates the water potability with the highest accuracy of 81.99% was given by random forest.

Due to the small size of the dataset, no evident processing time delays were found when running the models. As a result, there wasn't enough variation between the models in run-time evaluations to identify one as being more effective than the others. Preparing the data was crucial to the modeling procedure. Managing missing values and outliers gave modelers access to a larger dataset and improved overall accuracy. To prevent the depicted modeling from becoming skewed or biased to the dominant class of the underlying dataset, addressing class imbalance was equally crucial. This dataset, from the authors' view, lacks several crucial predictive criteria such as coliform or bacteria and toxic metals like lead or copper based on a literature analysis of prior studies for testing water as well as additional basic water testing research done.

The most notable examples of these characteristics are in at-home water testing equipment and the identification of diseases transmitted by water. These characteristics also correlate more strongly with water potability, which would have facilitated the initial exploratory investigation.

Machine learning techniques can be used to improve the classification of water, and it is possible to do so with high accuracy. With an accuracy rate of 81.99%, the Random Forest Classifier demonstrated the greatest overall performance. While water is successfully categorized using the predictor factors in the dataset, certain additional crucial features, such as coliform levels and heavy metals, could have been included in the future before deployment. An investigation into more sophisticated deep learning algorithms can also be done in the future.

Acknowledgments

This research is supported by the Sparks Foundation as a part of the Graduate Rotational Internship Program (GRIP). I would express my heartfelt gratitude towards the Sparks Foundation for their valuable guidance and support. The dataset for the experimentation was collected from Kaggle.com.

References

[1] “Goal 6: Ensure access to water and sanitation for all”, https://www.un.org/sustainabledevelopment/water-and-sanitation/, [Date Accessed: 01-Jul-2023]

[2] “Youth at UN 2023 Water Conference”, https://www.unwater.org/, [Date Accessed: 01-Jul-2023]

[3] Kleespies, M.W., Dierkes, P.W. The importance of the Sustainable Development Goals to students of environmental and sustainability studies—a global survey in 41 countries. Humanit Soc Sci Commun 9, 218 (2022). https://doi.org/10.1057/s41599-022-01242-0

[4] The Canadian government (2015), Maintaining water availability and quality. https://www.canada.ca/en/environment-climate-change/services/archive/sustainable-development/2015-progress-report/water-quality-availability.html

[5] Nida Nasir, Afreen Kansal, Omar Alshaltone, et al., “Water quality classification using machine learning algorithms”, Journal of Water Process Engineering, Volume 48, 2022, ISSN 2214-7144, https://doi.org/10.1016/j.jwpe.2022.102920.

[6] WHO (2022). recommendations for drinking water quality. The information was obtained from https://www.who.int/publications/i/item/9789240045064.

[7] Thanh Noi, P.; Kappas, M. Comparison of Random Forest, k-nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. Sensors 2018, 18, 18. https://doi.org/10.3390/s18010018

[8] Guangtao Fu, Yiwen Jin, Siao Sun, Zhiguo Yuan, David Butler, "The role of deep learning in urban water management: A critical review", Water Research, Volume 223, 2022, 118973, ISSN 0043-1354, https://doi.org/10.1016/j.watres.2022.118973.

[9] Mahalakshmi, A. and Yogalakshmi, S. Effective Machine Learning-Based Water Quality Prediction for Indian Reivers. www.ajast.net.

[10] Ahmed, U. Mumtaz, R. Anwar, H. Anwar, A. A. Shah, R. Ifran, & J. Garcia-Nieto (2019). Effective Prediction of Water Quality Using Supervised Machine Learning. www.mdpi.com/journal/water.

[11] Consolata Gakii and Jeniffer Jepkoech (2019), ‘A Classification Model for Water Quality Analysis Using Decision Tree’- European Journal of Computer Science and Information Technology Vol.7, No.3, pp.1-8

[12] Alexander Y Sun and Bridget R Scanlon 2019 Environ. Res. Lett. 14 073001, 10.1088/1748-9326/ab1b7d

[13] Sakizadeh, M. Artificial intelligence for the prediction of water quality index in groundwater systems. Model. Earth Syst. Environ. 2, 8 (2016). https://doi.org/10.1007/s40808-015-0063-9

[14] Ruixing Huang, Chengxue Ma, Jun Ma, Xiaoliu Huangfu, Qiang He, Machine learning in natural and engineered water systems, Water Research, Volume 205, 2021, 117666, ISSN 0043-1354, https://doi.org/10.1016/j.watres.2021.117666.

[15] Krishnan, S.R.; Nallakaruppan, M.K.; Chengoden, R.; Koppu, S.; Iyapparaja, M.; Sadhasivam, J.; Sethuraman, S. Smart Water Resource Management Using Artificial Intelligence—A Review. Sustainability 2022, 14, 13384. https://doi.org/10.3390/su142013384

[16] Fernández del Castillo, A.; Yebra-Montes, C.; Verduzco Garibay, M.; de Anda, J.; Garcia-Gonzalez, A.; Gradilla-Hernández, M.S. Simple Prediction of an Ecosystem-Specific Water Quality Index and the Water Quality Classification of a Highly Polluted River through Supervised Machine Learning. Water 2022, 14, 1235. https://doi.org/10.3390/w14081235

[17] Jinal Patel, Charmi Amipara, Tariq Ahamed Ahanger, Komal Ladhva, Rajeev Kumar Gupta, Hashem O. Alsaab, Yusuf S. Althobaiti, Rajnish Ratna, "A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI", Computational Intelligence and Neuroscience, vol. 2022, Article ID 9283293, 15 pages, 2022. https://doi.org/10.1155/2022/9283293

[18] Ghobadi, F.; Kang, D. Application of Machine Learning in Water Resources Management: A Systematic Literature Review. Water 2023, 15, 620. https://doi.org/10.3390/w15040620.

[19] Bhateria, R., Jain, D. Water quality assessment of lake water: a review. Sustain. Water Resour. Manag. 2, 161–173 (2016). https://doi.org/10.1007/s40899-015-0014-7

[20] A. Kadiwal (2021). Water quality. Retrieved 20-Jun-2023 from https://www.kaggle.com/datasets/adityakadiwal/water-potability/.