International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.029

Machine Learning Methods for Balanced and Imbalanced Datasets to Predict Consumable Water

S. VarshaaSaiSripriya
Department of Electronics and Communication Engineering
Amrita School of Engineering,
Chennai, Amrita VishwaVidyapeetham, India
ch.en.u4ece19029@ch.students.amrita.edu

Ganesh Kumar Chellamani
Department of Electronics and Communication Engineering
Amrita School of Engineering,
Chennai, Amrita VishwaVidyapeetham, India
ganeshkumar@ch.amrita.edu

G. Premalatha
Department of Data Science and Business System[3]
SRM Institute of Science and Technology,
Kattankulathur, India
premalag@srmist.edu.in

***Abstract* – The paper details various machine learning techniquesto identify the best technique to predict consumable water. Being the most essential natural element, identifying drinkable water amidst the deteriorating qualities of drinkable water is yet a worrisome issue. The versatility of employing techniques and algorithms of machine learning in solving real-world problems proves to bring efficient results. This paper details a comparative study using algorithms like – Logistic Regression, Decision Trees (DT), Random Forest (RF), XGBoost, Gradient Descent, Support Vector Machine (SVM), AdaBoost, and k-Nearest Neighbours. The used dataset is self-developed with reference to water potability parameters. Originally 5000 data entries were provided as input. XGBoost outperforms other models in terms of identifying consumable water. It is worthwhile to note that algorithm's performance is validated as best when it provides equal importance to both majority and minority classes. Thus, Synthetic Minority Oversampling technique (SMOTE) and Adaptive synthetic sampling approach (ADASYN) were employed to conclude and accurately identify for the best technique. XGBoost pertaining to SMOTE outperformed other techniques and was the best model to predict water potability.**

Keywords* – *Machine learning algorithms, ADASYN, SMOTE, XGBoost

I. Introduction

Water is indispensable for life to exist. One of the most jeopardizing issues that needs to address is identifying consumable water. According to UNICEF and WHO, 2.1 billion people worldwide lack access to safe drinking water. Adverse effects impacting every sector and aspect of life, including the vulnerability to water-borne diseases, are certain by consuming unclean or unsafe water.

Consumable water is expected to be free from pathogens that are harmful as well as from toxic chemicals. Contaminated water could be both natural and man-made. Naturally contaminated refers to those water bodies in which certain elements like fluoride, chloride, etc., have been found slightly more than the accepted level, whereas man-made contamination due primarily due to water pollution that is either by dumping waste into water bodies or by dumping chemicals from factories. Water quality is defined by various factors:

(1) Physical parameters: color, taste and odor, temperatures:

  1.1. Color of water results from form the dissolved organic substances

  1.2. Taste and odor are due to inorganic salts and dissolved organic components & gases

  1.3. Temperatures are usually typical for safe drinking water. Fluctuations in temperatures may be due to the presence of harmful chemicals and substances

(2) Turbidity:Suspended materials in water determine the turbidity levels of water. It is the amount of solid matter present in suspended form.

(3) Chemical parameters include BOD (biological oxygen demand), COD (chemical oxygen demand), etc. Amount of arsenic level (As), chloride (Cl^-), fluoride (F^-), zinc (Zn), iron (Fe), manganese (Mn), and other toxic substances also play a significant role in determining whether or not water is safe.

(4) Biological parameters like disease-causing organisms

(5) pH levels:The alkalinity of water is determined using pH levels. It is a logarithmic measure from 0 to 14 pH scale. The scale is divided into two sections – acidic and basic. Any value from 0 to 7 represents acidic nature, and any value above 7 represents basic nature. As per WHO standards, the permissible limit of pH for pure water is from 6.5 to 8.5. Any other value is considered impure in the case of water.

(6) Hardness:Calcium and magnesium salts cause water hardness.

(7) Total dissolved solids:Mineralized water consists of a high TDS value. However, the maximum limit for TDS ranges from 500 mg/l to 1000 mg/l for drinking purposes.

(8) Total Organic Carbon

(9) Conductivity:As per WHO standards, water conductivity must not exceed 400µS/cm.

(10) Other components like Trihalomethanes can be present only up to 80ppm.

The collected dataset considers pH, hardness, concentration of molecules, chloride level, sulphate levels, conductivity, TOC levels, amount of trihalomethanes, turbidity, and potability as its attributes. Machine learning serves to provide powerful tools and algorithms that can be used to bring efficient results to tackle real-world problems. For this problem statement, algorithms, namely - Logistic Regression, Decision Trees, Random Forest, XGBoost, Gradient Descent, Support Vector Machine

(SVM), AdaBoost, and k- Nearest Neighbours, were used.I. LITERATURE SURVEY

In [1], the authors employed a SVM, Group Method of Data Handling (GMDH), and ANN in order to determine the quality of water of the river Tireh. It was found that the performance of GMDH was not satisfactory, and thus ANN and SVM were more suitable for predicting the quality. It was also noted that the Tansig transfer function and RBF kernel functions gave the best performance. Considering DDR index values, SVM was evaluated to have a lower value and, therefore, was found to be the most accurate of the three models [1]. Authors of [2] used ANN with a nonlinear autoregressive time series model to develop a complete framework for efficient prediction and analysis. Scaled conjugate gradient and log sigmoid was used for the training algorithm. Chlorophyll, specific conductance, dissolved oxygen, and turbidity were primarily considered as the determining factors by them. Results concluded that ANN- NAR proves to be a reliable method to identify the potability of water [2].

The primary aim was to employ Breiman's random forests and validate the results. It was also found that random forest techniques were also effective in solving this problem [3]. Random forest, deep neural network, gaussian naïve bayes, artificial neural network, and data distribution analysis were used to tackle the issue. The algorithms performed well wither maximum testing accuracy corresponding to ANN with 98.12%. The author of [4] was able to arrive at a conclusion that a total of 89.71% of water was found to be safe for drinking purposes [4]. In this paper [5], a decision tree and k- Nearest neighbour were used to estimate the water quality class. The models were optimized, and hyperparameter tuning was performed. The model's accuracy was once again validated. The k-nearest neighbour was identified to be better than the decision tree, with accuracy scores of 61.7% and 58.5%, respectively [5].

The author of paper [6], employed used (1) stratified sampling and wavelet de-noising ANFIS Model, (2) Fuzzy models and time series analysis, and (3) integrating ANFIS model with intelligence algorithms like – genetic algorithm and particle swarm algorithm. In the first algorithm, TDS, sulphate, chloride, and fluoride were the primary inputparameters since they have higher correlation values with electric conductivity. This algorithm was used to predict EC. WT-ANFIS model trained with a stratified sampling strategy was found to perform better than MLR, ANNs, ANFIS, and EANFIS. In the second algorithm - fuzzy and time series analysis also stratified sampling strategy was employed. With parameter EC in BB, FTS was able to predict well. For the third case, a comparison was made between ANFIS, ANFIS- GA, and ANFIS-PSO, that is, ANFIS integrated with intelligent algorithms. ANFIS-PSO was identified to outperform the other two cases [6].

Supervised machine learning approaches- support vector regression and extreme gradient boost (XGBoost). The algorithms were provided with big data. Both algorithms predicted well for temperature parameters. SVR was able to perform well even for dissolved oxygen. In the case of turbidity, the prediction by the two algorithms had more than 5% variation. However, cyanobacteria and fDOM had large variations. Other parameters didn't show much variation. Thus, considering these three factors to determine the better of the two, the authors identified SVR to be better than XGBoost for this model [7]. Authors of the paper [12] considered temperature, turbidity, pH, and TDS (total dissolved oxygen) as input, to apply to supervised ML approaches. It was found that gradient descent with a learning rate of 0.01 and polynomial regression with degree 2 performed better than other models [12]. Thus, it was also noted that PCA with SVR was performing better [15].

In this paper [8], the authors integrated IoT and performed real-time water quality checks using machine learning. The idea was to analyse the data using sensor inputs from lakes in rural locations and use k means for the process. Arduino UNO and Raspberry Pi embedded devices were used to validate the same. The sensor inputs were given to the pi4 edge-level processor, where k-means were used to predict the quality. The predicted values were then stored in a cloud server for future access. Thus, water could be monitored using IoT techniques without any human interference [8]. A similar approach was used by the authors of the paper [14], where Arduino was interfaced with the ZigBee handset, which detected low- quality water [14]. Synthetic minority oversampling technique and explainable AI were used in the paper [9] to predict the water potability prediction model. Oversampling was performed using Synthetic Minority Oversampling Technique (SMOTE). Therefore, replication of minority classes is performed in oversampling. As a result, the authors were able to replicate the data as 1998 for both not portable and portable after oversampling, from 1998 not portable and 1278 portable data, respectively. Later machine learning approach is applied to validate the results. Radom forest was found to outperform other models [9]. DT, RF, and MLP methods were employed for air quality with the same approach as that of the water quality case [13].

WQI was evaluated for Ebinur Lake Watershed using ML and remote spectral indices via fractional derivatives methods. The model pertaining to a spectral index of 1.60 was found to perform better than other models. The proposed models were GA-SVR and band difference algorithms [10]. In paper [11], the authors identified that the most commonly used machine learning approaches were ANN, RF, SVM, regression cubist, genetic programming, and DT. It was found that Chlorophyll- a, temperature, suspended solids, colored dissolved organic matter, salinity, and turbidity were the commonly used determining factors for the problem statement. 138 samples of water from Agastheeswaram, Tamil Nadu, were collected pre- and post-monsoon by the authors of the paper [17] to predict groundwater quality. Off DT, KNN, and SVM used, SVM was found to achieve better results [17]. Apart from these models, a fuzzy system was used to classify water into five different groups. Water from three lakes from Hosur, Tamil Nadu was collected, and the classification was computed [18].

III. PROPOSED SOLUTION WORKFLOW

The primary objective was to perform a comparative study that validates the potability by analysing various algorithms.First, the dataset is thoroughly studied by performing data visualisation and exploratory data analysis (EDA). Then thefollowing three stages are computed to identify the best algorithm.

Stage 1: Analysing models without oversampling - At this stage, the original dataset is fed to the models, and the performance is analysed. Initially, a total of 5000 inputs were given to the model with the attribute values - pH, hardness, the concentration of molecules, chloride level, sulphate levels, conductivity, TOC levels, amount of trihalomethanes, turbidity, and potability. The values were randomly generated in Excel considering WHO standards. Logistic Regression, DT, RF, XGBoost, Gradient Descent, SVM, AdaBoost, and k- Nearest Neighbours were used, and the accurate algorithm was identified.

Further investigation was done by performing oversampling. Oversampling is done to reduce the possibility of ignoring minority classes by machine learning when unbalanced data is fed to the model requirement [9]. Thus, ML algorithms tend to be biased toward the majority class. Since potable data was found to be a minority class and is the criterion of interest. It was necessary to balance the data to get accurate and unbalanced results. Two techniques – the Synthetic Minority Oversampling technique (SMOTE) and the Adaptive synthetic sampling approach (ADASYN) were used with the same algorithms to conclude for the best technique.

Stage 2: ADASYN approach is used to balance datasets by adaptively generating minority data samples that are based on their distributions [16]. The advantage of using ADASYN is that it can shift the classifier's decision boundary to focus on- difficult to learn aspects, which allows the algorithm to improve its learning performance.

Stage 3: SMOTE works by analysing and identifying adjacent instances in feature space. A line is drawn to link them and to generate a new sample positioned at that line. By doing so, new data is replicated, and oversampling is successfully achieved.

Finally, a comparative study was computed to determine the best algorithm.

IV. Algorithms Used

A. *Logistic Regression*

Logistic Regression used in both classification and regression cases, is a statistical model that estimates the probability of the occurrence of an event. Logit transformation is applied on the odds whose function is given by:

$$log(\pi) = \frac{1}{1+e-\pi} \qquad (1)$$

It is a supervised learning technique for predicting categorically dependent variables using the provided set of independent variables.

B. *Decision Tree*

The decision tree - a supervised learning non-parametric algorithm used for both classification and regression problems is based on trees that consist of features like root node, leaf nodes, branches, and internal nodes. In order to identify optimal split points, decision tree employs a divide and conquers strategy by greedy search.

C. *Random Forest*

In this learning approach, ensemble learning is used that solves complicated problems by several classifiers. This technique is also employed for classification and regression purposes. Random forest is made up of many decision trees. The outcome is determined based on the prediction made by the decision trees. As the number of trees grows, the precision of the result also increases. It is primarily used to address the shortcomings of the decision tree method.

D. *XGBoost*

XGBoost is a machine-learning toolkit for distributed gradient boosting. It has been optimised for: (1) efficiency, (2) portability, and (3) flexibility. It is a parallel tree boosting algorithm that solves problems quickly and accurately. It is also a supervised learning technique to accurately predicttarget variables by combining weaker models. A depth of 8 was considered.

E. *K-Neighbors*

K-Neighbors or KNN is a supervised learning non-parametric algorithm that uses similarities between new and available cases and places a new case into a category that is closest to one of the available cases. KNN is also used to solve both classification and regression cases. It uses Euclidian distance (equation 2) as the determining distance metric criterion.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (yi - xi)^2} \qquad (2)$$

Nine nearest neighbors were considered, with leaf size as 20 for the comparative study.

F. *SVM*

SVM is a supervised learning algorithm that uses a decision plane and boundaries. A decision plane divides groups of data points into several classes. Therefore, in an n-dimensional space, the input data is viewed as two sets of vectors. SVM has several kernels like (1) Linear, (2) Polynomial, (3) Gaussian, (4) ANOVA, (5) RBF – Gaussian Radial Basis Function, and (6) Sigmoid. To compute the comparative study, RBF kernel was used since the results pertaining to RBF had better accuracy than other kernels.

G. *AdaBoost*

AdaBoost or adaptive boosting is a statistical meta-algorithm used for classification purposes. It is an ensemble technique used to solve complicated problems by combining weaker classifiers and building a stronger one. In AdaBoost, weaker ones are termed as decision stumps that denote decision trees with a single split. AdaBoost classifier with a learning rate of 0.002 was used.

H. *Gradient Descent*

Gradient boosting is an optimization technique. It is used to find the global minimum of a given function.

However, it applies mainly in cases with few local minima. Gradient boosting with a learning rate of 0.05 and a maximum depth of 5 was used.

V. Methodology

Data visualization is presenting data in graphical or pictorial form. Heat maps are widely used in order to represent correlation matrix graphically. Correlation is a statistical metric to represent the relationship between two variables and ranges from -1 to +1. Figure 1 indicates a heat map of the given attributes. Considering values as $xi$and $yi$and mean values as X and Y, the correlation value is calculated as:

$$correlation = \frac{\sum(xi-X)(yi-Y)}{\sqrt{\sum(xi-X)2(yi-Y)2}} \qquad (3)$$

EDA is a sophisticated method for examining datasets to highlight key features often associated with visual methods. It provides valuable information that could be used in scientific research and visualizations. Box plots are plotted to identify outliers. Figure 2 represents a distplot that is used to plot univariate data distribution against density distribution.

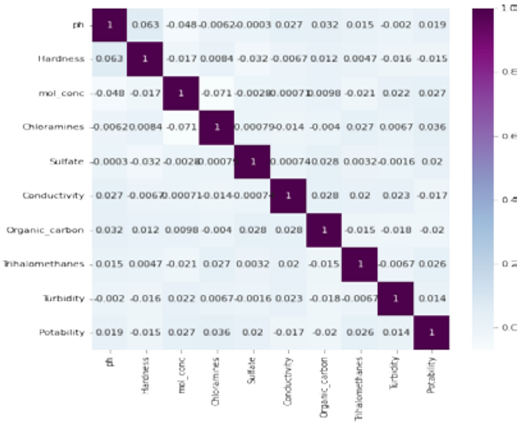

Fig.1. Heatmap representing correlation between attributes

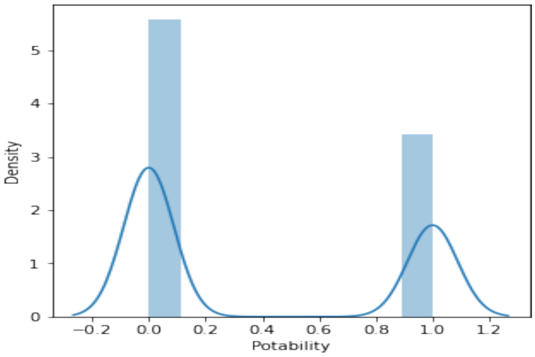

Fig.2. Distplot of potabilityvs density distribution

Observations from the dataset fed to the system is analysed as follows: in Figure 3, the blue box plot corresponds to non-potable. The minimum and maximum values found in the dataset were 1.2 and 14. However, the lower and upper fences were considered as 2.175578 and 13.10774. Thus, any value below and above the lower and upper ranges were treated as outliers. The median for the non-potable case was 7.848005, whereas q1 and q3 were 6.317914 and 9.093132, respectively. Similarly, the red box plot corresponds to potable. Here, the lower and upper ranges were 3.738116 and 12.10508. The median for the potable case was 8.004796, whereas q1 and q3 values were 6.677912 and 8.944125, respectively. Figure 4 is a box plot with respect to the hardness of water and potability. The blue plot box in figure 6 corresponds to the non-potable. The minimum and maximum entries received were 50.0006 and 354.2365, respectively. Any value other than this range was treated as an outlier. The lower and upper fence values were 168.9892 and 325.8871. The median for this case was 247.5177, whereas q1 and q3 values were 227.5106 and 267.0189. The red box plot corresponds to potable. Here, the values of the lower and upper fence are 161.4792 and 328.62, respectively. The median for the potable case was 247.3163, whereas q1 and q3 values were 224.6147 and 266.7835. It is worthwhile to note that as per WHO standards, the permissible pH value for pure water was between 6.5 and 9, while hardness limitations are in the range of 200-300mg/L.

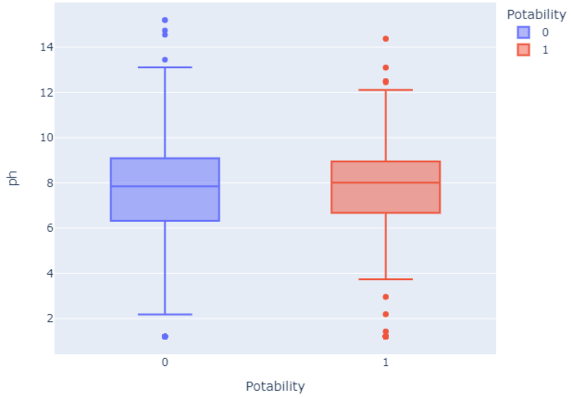

Fig. 3. pH box plot (Blue Box plot – non potable, Red Box plot – potable)

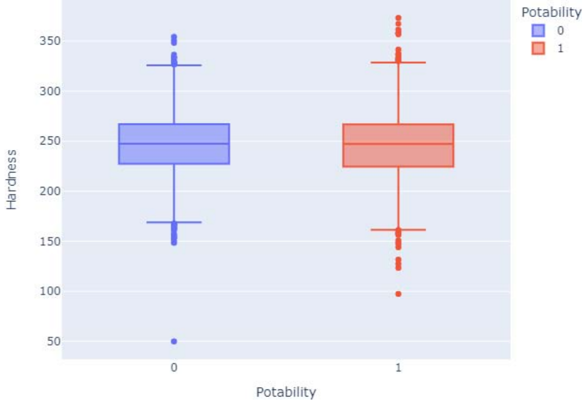

Fig.4. Hardness box plot (Blue Box plot – non potable, Red Box plot – potable)

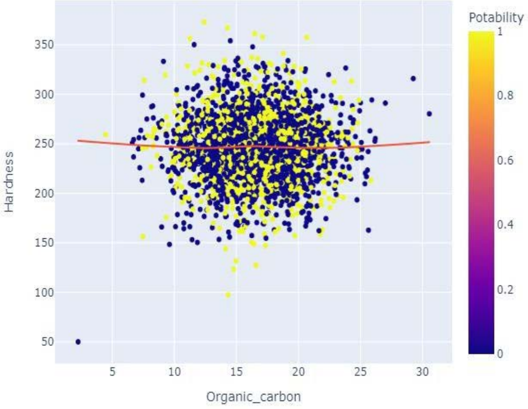

Fig.5. Hardness and Organic Carbon as determining factors to identify potability

Water contamination or determining factors could be due to two attributes, for such cases, a scatter plot corresponding to two determining factors (Figure 5) was implemented.

Considering input parameters such as pH, hardness, the concentration of molecules, chloride level, sulphate levels, conductivity, TOC levels, amount of trihalomethanes, and turbidity. Then, a feature scaling technique called Standard Scaler is applied to x. This standardizes data into a standard format that is that mean of the data to zero and standard deviation to one. Standard scaler performs an operation on the dataset as mentioned in equation 4 where xi represents values and X is mean. Later, a train test split is performed, and the models are trained and validated.

$$standardscaler = \frac{(xi - X)}{standarddeviation} \quad (4)$$

Potability represents whether or not water is drinkable. The value of potability is binary – 0 and 1. For the given conditions, if the value corresponds to 0, that indicates that water does not fall under the drinkable category. A value of 1 indicates that the water is consumable and safe. Initially, out of the 5000 entries, 62%-38% was the distribution of potability in the dataset. After ADASYN and SMOTE, a total of 6196 and 6206 entries, respectively, were used for model training. Table 1 summarises the distribution.

TABLE 1: SUMMARY OF DATA ENTRIES

Name	Total	X_train	X_test	Y_train	Y_test
Before Oversampling	5000	3350	1650	3350	1650
ADASYN	6196	4192	2004	4192	2004
SMOTE	6206	4156	2050	4156	2050

VI. RESULTS AND DISCUSSION

To validate accuracy precision, recall, and f1-scores are analysed. The accuracy of an algorithm is the score that corresponds to the number of correct predictions to all predictions. It is given by:

$$accuracyscore = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

The precision determines how many predictions (positive) are correctly made. This is a measure of true positive given by:

$$precision = \frac{TP}{TP+FP} \quad (6)$$

The recall is also referred to as sensitivity checks for correctly predicted cases over the entire positive cases in the data. It is given by:

$$recall = \frac{TP}{TP+FN} \quad (7)$$

F1-score is given by: $f1 - score = 2 * \frac{precision*recall}{precision+recall}$ (8)

TABLE 2: COMPARISON OF PRECISION RECALL F1-SCORE AND ACCURACY VALUES BEFORE OVERSAMPLING

ALGORITHM	BEFORE OVERSAMPLING 0- non-potable, 1 - potable						
	Precision		Recall		F1-Score		Accuracy
	0	1	0	1	0	1	
Logistic Regression	0.62	1.00	1.00	0.00	0.76	0.00	61.69%
Decision Tree	0.67	0.53	0.79	0.39	0.73	0.45	63.33%
Random Forest	0.62	0.00	1.00	0.00	0.76	0.00	61.63%
XGBoost	**0.76**	**0.84**	**0.94**	**0.51**	**0.84**	**0.64**	**77.51%**
k-Neighbors	0.69	0.58	0.81	0.42	0.74	0.48	65.87%
SVM	0.66	0.75	0.96	0.21	0.78	0.33	67.03%
AdaBoost	0.62	0.00	1.00	0.00	0.76	0.00	61.63%
Gradient Descent	0.66	0.84	0.98	0.20	0.79	0.33	67.93%

TABLE 3: COMPARISON OF PRECISION RECALL F1-SCORE AND ACCURACY VALUES AFTER OVERSAMPLING

ALGORITHM	AFTER OVERSAMPLING 0- non-potable, 1 - potable													
	ADASYN							SMOTE						
	Precision		Recall		F1-Score		Accuracy	Precision		Recall		F1-Score		Accuracy
	0	1	0	1	0	1		0	1	0	1	0	1	
Logistic Regression	0.52	0.52	0.51	0.52	0.51	0.52	51.59%	0.52	0.52	0.50	0.53	0.51	0.52	51.60%
Decision Tree	0.66	0.54	0.27	0.86	0.38	0.66	56.53%	0.55	0.65	0.84	0.31	0.66	0.42	57.31%
Random Forest	0.54	0.55	0.58	0.51	0.56	0.53	54.74%	0.50	0.00	1.00	0.00	0.67	0.00	50.00%
XGBoost	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**	**77.29%**	**0.80**	**0.81**	**0.81**	**0.80**	**0.80**	**0.80**	**80.34%**
k-Neighbors	0.00	0.50	0.00	1.00	0.00	0.67	50.00%	0.63	0.62	0.61	0.64	0.62	0.63	62.14%
SVM	0.50	0.50	0.65	0.36	0.57	0.42	50.29%	0.50	0.51	0.77	0.25	0.61	0.33	50.58%
AdaBoost	0.60	0.53	0.32	0.79	0.41	0.64	55.08%	0.58	0.53	0.31	0.77	0.40	0.63	54.09%
Gradient Descent	0.66	0.65	0.64	0.67	0.65	0.66	65.61%	0.66	0.66	0.65	0.67	0.66	0.66	65.90%

Table 2 details the performances of models without oversampling. It is noted that XGBoost is able to provide an accuracy of 77.51%. The idea behind oversampling is to avoid algorithms tending to be biased towards one category while predicting the output. It can also be noted that other models were able to perform well, giving approximately 70% results. Thus, it was not sufficient to conclude that XGBoost was the best method among all the other algorithms. To rectify this issue and to conclude on the best algorithm, ADASYN and SMOTE results are compared with the before oversampling case. Table 3 details the performance of models after oversampling. It is worthwhile to note that Logistic regression, decision tree, random forest, k-nearest neighbors, SVM, AdaBoost, and gradient descent have a significant decrease in the performance. This concludes that these models did not give equal importance to the minority class, which was to be required in this problem statement. However, XGBoost's performance is seen to improve from before sampling case of 77.51% to 77.29% in ADASYN (approximately same but better than other models) and to 80.34% in SMOTE with improvements in precision, recall and f1-scores, indicating that it is the best accurate model to determine consumable water. Confusion matrix are plotted to validate the results. Figures 6, 7, 8 correspond to XGBoost algorithm. Bar plot in figures 9, 10 and 11 represent accuracy scores. These are the accuracy plots for individual cases. It can be observed that accuracies of other models are gradually decreasing when unbiased data is fed. Finally, a combined performance analysis plot is plotted (Figure 12). Green line corresponds to SMOTE, blue for unbalanced dataset and yellow for ADASYN. Overall performance of XGBoost outperforms other models.

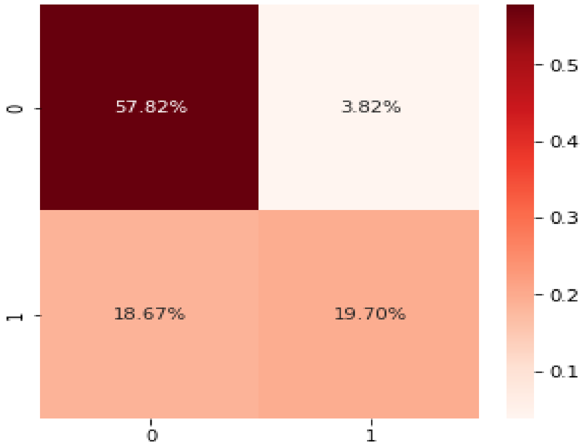

Fig.6. XGBoost Confusion Matrix – Before oversampling

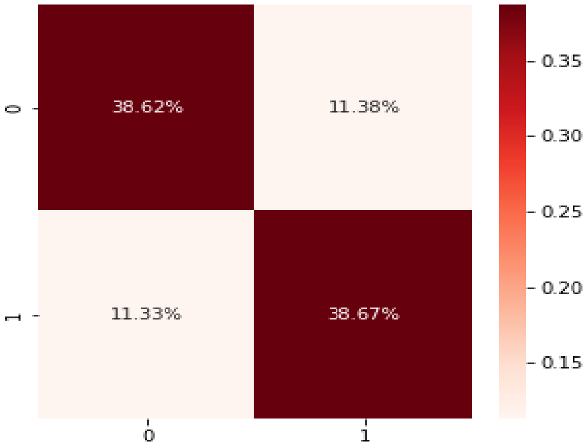

Fig.7. XGBoost Confusion Matrix – ADASYN

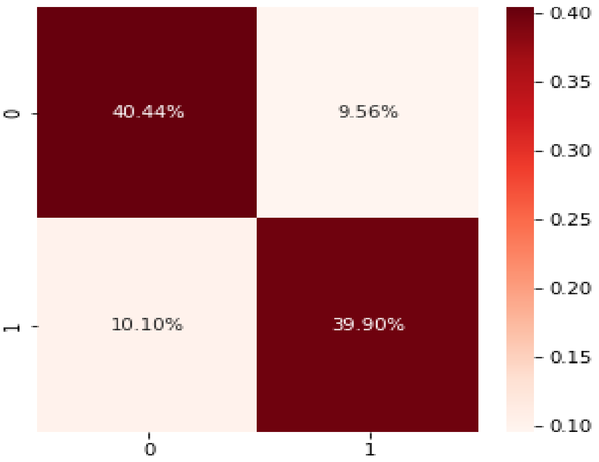

Fig.8. XGBoost Confusion Matrix – SMOTE

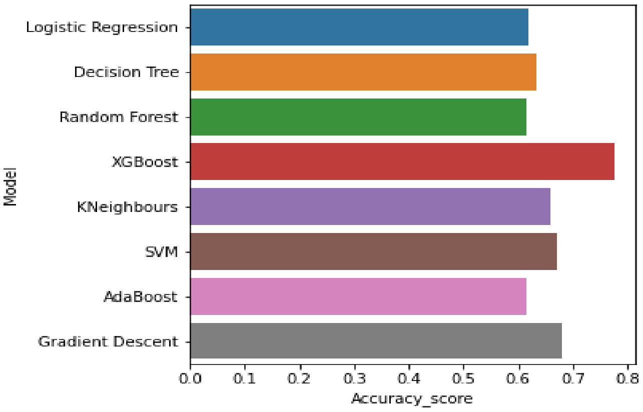

Fig.9. Accuracy scores before oversampling

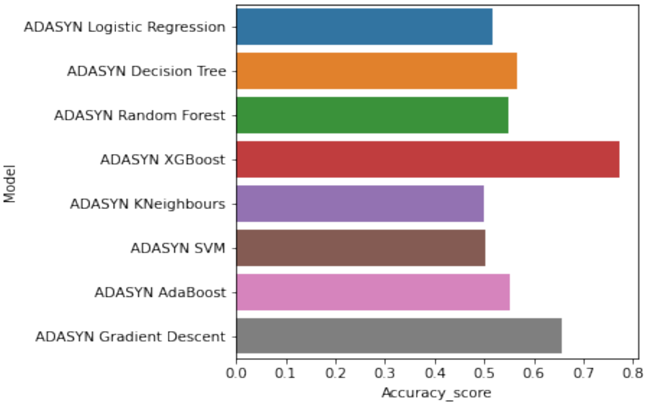

Fig.10. Accuracy scores ADASYN

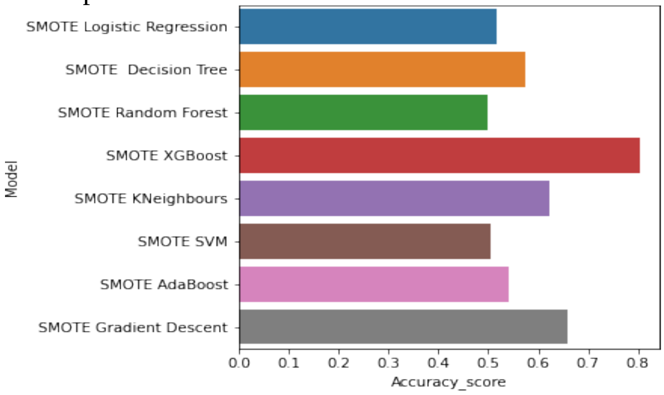

Fig.11. Accuracy scores SMOTE

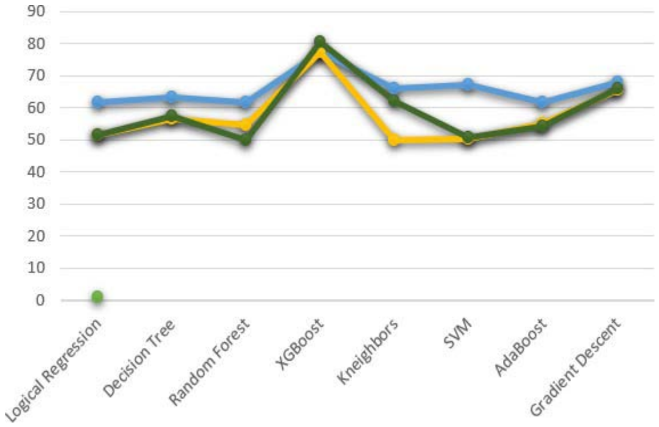

Fig.12. Performance Analysis

VII. Conclusion

Water is one of the most important and essential resources that determines life on Earth. The availability of safe consumable water is a jeopardizing situation that is being faced in many countries around the world. Apart from natural contaminants, water pollution has become a major concern for many environmentalists. Water potability is affected by a number of factors. In this paper, factors such as pH, hardness, the concentration of molecules, chloride level, sulphate levels, conductivity, TOC levels, amount of trihalomethanes, and turbidity were considered. The used dataset is self-developed. However, the models will work with the same efficiency as other river or lake datasets with the same attributes. Exploratory data analysis is widely used by researchers to study and investigate various data. The same concept is applied here to understand and analysis the distribution of data when considering different attributes. WHO standards were considered, and analysis was performed on the dataset.

Machine learning algorithms are powerful and sophisticated tools that can tackle problems such as water quality index determination or water potability. From the previous research, it was identified that various algorithms were used to analyse the same. But it was noted that the algorithms were performing differently depending on the dataset. On further investigation, it was found that the models tended to be biased toward the majority class.

Therefore, an imbalanced dataset was used to have a comparative analysis of various models. The imbalanced dataset consisted of more data entries with non-potable conditions. The goal was to predict potability, thus maintaining potable cases as a minority class. To balance the dataset, ADASYN and SMOTE were used. In both ADASYN and SMOTE cases, models like logistic regression, decision tree, Support vector machine, AdaBoost, and Gradient descent had the same performance/accuracy level with negligible percentage difference. However, it was also observed that these models had a drastic decrease in their performance from their original performance, as in the case of the unbalanced dataset.

XgBoost, on the other hand, performed well and had a slight improvement in its accuracy too. The precision, recall, and f1-scores of XgBoost were significantly higher with values - 0.80, 0.81, 0.80 for non-potable and 0.81, 0.80, 0.80 for potable cases. This model can effectively produce results when interfaced with IoT or TinyML real- time projects. Therefore, the comparative analysis concluded that XGBoost was the best technique since it outperformed other models in all three stages of the proposed workflow.

References

[1] Amir HamzehHaghiabi, Ali HeidarNasrolahi, and Abbas Parsaie; “Water quality prediction using machine learning methods”, Water Quality Research Journal 1 February 2018.

[2] Y. Khan and C. S. See, "Predicting and analyzing water quality using Machine Learning: A comprehensive model," 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), April 2016.

[3] NishantRawat, ManganiDaudiKazembe, andPradeep Kumar Mishra, “Water Quality Prediction using Machine Learning” International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538,vol. 10,issue VI June 2022

[4] Osim Kumar Pal, “The Quality of Drinkable Water using Machine Learning Techniques”, International Journal of Advanced Engineering Research and Science, vol. 9, June 2022

[5] SaiSreejaKurra, SambangiGeethika Naidu, and SravaniChowdala, “Water Quality Prediction using Machine Learning”, International Research Journal of Modernization in Engineering Technology and Science,vol. 04, issue05, May2022

[6] Fu, Zhao, "Water Quality Prediction Based on Machine Learning Techniques", UNLV Theses, Dissertations, Professional Papers, and Capstones, vol. 3994, January 2020.

[7] Joslyn, Kathleen, "Water Quality Factor Prediction Using Supervised Machine Learning", REU Final Reports, vol. 6, 2018.

[8] R.M. Bhavadharini, Kalpana Devi. S, S. Angel Vergina, and S. Kayalvizhi, “A Real Time Water Quality Monitoring Using Machine Learning Algorithm”, European Journal of Molecular & Clinical Medicine, vol. 7, 2020/12/17

[9] Jinal Patel, CharmiAmipara, Tariq AhamedAhanger, KomalLadhva, Ranjeev Kumar Gupta, Hashem O. Alsaab, Yusuf S. Althobaiti, andRajnishRatna, “A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI”, Computational Intelligence and Neuroscience,vol. 2022.

[10] Dhanabalan, S. S., Sitharthan, R., Madurakavi, K., Thirumurugan, A., Rajesh, M., Avaninathan, S. R., & Carrasco, M. F. (2022). Flexible compact system for wearable health monitoring applications. Computers and Electrical Engineering, 102, 108130.

[11] N Hassan and C S Woo, “Machine Learning Application in Water Quality Using Satellite Data”, IOP Conference Series: Earth and Environment Science, Volume 842, 3rd International Conference on Tropical Resources and Sustainable Sciences, July 2021.

[12] U.Ahmed, R.Mumtaz, H.Anwar, A.A.Shah, R.Irfan,and J.García-Nieto, “Efficient Water Quality Prediction Using Supervised Machine Learning” Water 2019.

[13] Gomathy, V., Janarthanan, K., Al-Turjman, F., Sitharthan, R., Rajesh, M., Vengatesan, K., &Reshma, T. P. (2021). Investigating the spread of coronavirus disease via edge-AI and air pollution correlation. ACM Transactions on Internet Technology, 21(4), 1-10.

[14] S., Yogalakshmi and A., Mahalakshmi, “Efficient Water Quality Prediction for Indian Rivers Using Machine Learning”, Asian Journal of Applied Science and Technology (AJAST), vol. 5, issue 1, April 18, 2021.

[15] Md. Saikat Islam Khan, Nazrul Islam, JiaUddin, Sifatul Islam, Mostofa Kamal Nasir, “Water quality prediction and classification based on principal component regression and gradient boosting classifier approach”, Journal of King Saud University - Computer and Information Sciences, vol. 34, issue 8, September 2022.

[16] Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008.

[17] A. Vijayakumar, andA.S. Mahesh, “Quality Assessment of Ground Water in Pre and Post-Monsoon using Various Classification Technique”, International Journal of Recent Technology and Engineering (IJRTE), vol. 8, issue 2, 2019.

[18] P. Kalamani and N. Rakesh, "Analysis of Lake Water Quality classification using Fuzzy inference Systems,", 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), 2018.