PERFORMANCE ANALYSIS OF THE PORTABILITY OF WATER RESOURCES THROUGH SMOTE

Dr. Aniket Siddhaling Kothawale, Assistant Professor, Electronics Department, Tuljaram Chaturchand College Baramati 413102. kothawaleaniket71@gmail.com
Dr. Ashok Eknath Kalange, HOD & Professor, Physics Department, Tuljaram Chaturchand College Baramati 413102.
Dr. Jagdish D. Deshpande, HOD & Professor, Electronics Department, Tuljaram Chaturchand College Baramati 413102.

Abstract.

Water, a crucial resource on our planet, is essential for sustaining human life, particularly in the form of drinking water. Assessing the adequacy of drinking water, a fundamental resource for human existence, is imperative for both current and future generations. However, the distribution of water resources across the globe is uneven, with some countries and regions enjoying abundance while others face scarcity. To address this issue, it is essential to conduct individual analyses of water resources in different regions. This research endeavors to predict water potability using various algorithms, leveraging physicochemical properties extracted from a dataset of drinking water which is available on Kaggle. There are nine different parameters in the dataset including hardness, pH, solids, sulfates, chloramines, organic carbon, trihalomethanes, turbidity and conductivity. Various Machine learning techniques such as Logistic Regression, Random Forest, KNN and SVM are used for this analysis. The handling of imbalanced datasets using SMOTE significantly impacted the F1-score of the "1" potability, increasing it from 0.46 to 0.74.

Keywords: Machine Learning, SMOTE, Accuracy.

1 Introduction

Water is crucial for all living things, including humans, animals, and plants. It's used not only for drinking but also in industry, agriculture, and trade [1]. Our bodies rely on water to function properly—it provides energy, helps regulate temperature, supports kidney function, and cleanses the body. Certain elements and compounds in water are essential for our health, but too much can be harmful [2], [3]. Water also acts as a transporter for vitamins and minerals in the body. Access to clean drinking water is vital to prevent diseases. Despite Earth's abundant water, only a small fraction is suitable for drinking [4]. Unsafe drinking water causes millions of deaths each year, especially among infants and young children. Water is a major topic of concern due to its importance for life, economy, and strategic value [5]. Lack of access to clean water and sanitation facilities exposes people to preventable health risks, especially in healthcare settings where water shortages and poor hygiene can spread viruses and bacteria [6]. In India, 70% of available water is contaminated by industrial and domestic pollutants. Around 80% of rural and 20% of urban populations lack access to clean drinking water. Poor drinking water quality negatively affects consumers' health, with reports indicating that at least 2 billion people worldwide use faces-contaminated water. Making informed decisions about controlling and safeguarding drinking water quality requires understanding the factors affecting its purity. Water quality can be impacted by various factors such as the source's quality, handling procedures, distribution, and filtration methods.

2. Related Work

Khan & See in [7] created a water quality model using an Artificial Neural Network (ANN) incorporating turbidity, chlorophyll, conductivity, DO values. Ali [8] tested the effectiveness of the model using three evaluation methods. The first method involved analyzing the neural network connection weights to determine the importance of each input parameter. The second and third methods aimed to identify the most effective inputs for the model. In [9], authors proposed a deep learning ELM framework as a promising option to investigate further for anomaly detection in WQ data. In [10], authors introduced an application for real-time detection of drinking water quality using cerebellar model articulation controller (CMAC) Artificial Neural Networks (ANNs), which learned faster than multilevel perceptron (MLP) backpropagation (BP). In [11] researcher presented water potability forecasting. They utilized Random Forests, DT, and SVM.

3. Proposed Method

The proposed model for the portability of water resources is shown in Fig. 1:

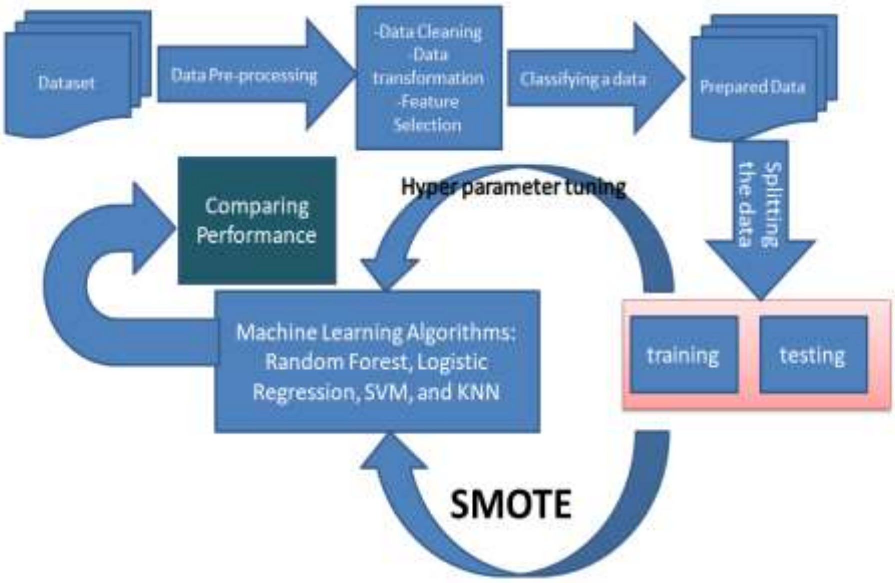


Fig. 1: Proposed Methodology

3.1 Dataset Collection and Distribution: A dataset of drinking water is taken from Kaggle for this study, provides data on more than 3000 water samples. The data contains information about chemical components and whether the water is drinkable. The dataset contains 10 different water quality parameters.

pH: The balance of acid as well as base present in water is measured using the pH value of the water. A pH-value of 6.5 - 8.5 is recommended by the WHO.

Hardness: The more calcium and magnesium the water contains, the harder the water is. Though these minerals are not harmful to consume, they could have an impact on the potability of the water.

Solids: Measurement of how many organic and inorganic materials are contained in the water.

Chloramines: Chlorine and chloramine are a common disinfectant. Chlorine of 4 mg/L or 4 parts per million(ppm) are considered save.

Sulfate: Naturally occurring mineral, that is much higher in seawater than in freshwater.

Conductivity: Measurement of how conductive the water is, meaning how well energy flows through it. The electric conductivity (EC) must not be higher than 400 μs/cm as per WHO standards.

Organic carbon: It is the result of decaying organic matter in water. According to US EPA < 2 mg/L of TOC is considered drinkable water.

Trihalomethanes: A chemical that occurs in water treated with chlorine. Levels up to 80 ppm are considered safe.

Turbidity: The WHO recommends a value of 5.00 NTU, Depending on the amount of solid matter in the water.

Potability: States whether water is safe. 0 = not safe to drink, 1 = safe to drink.

3.2 Data Preprocessing: It is a transforming technique of a dataset in a way that the content of information can be exposed to the mining tool.

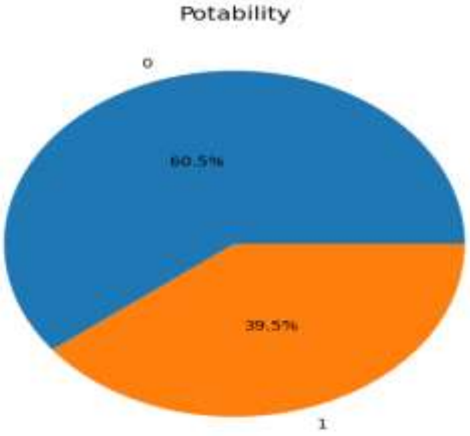
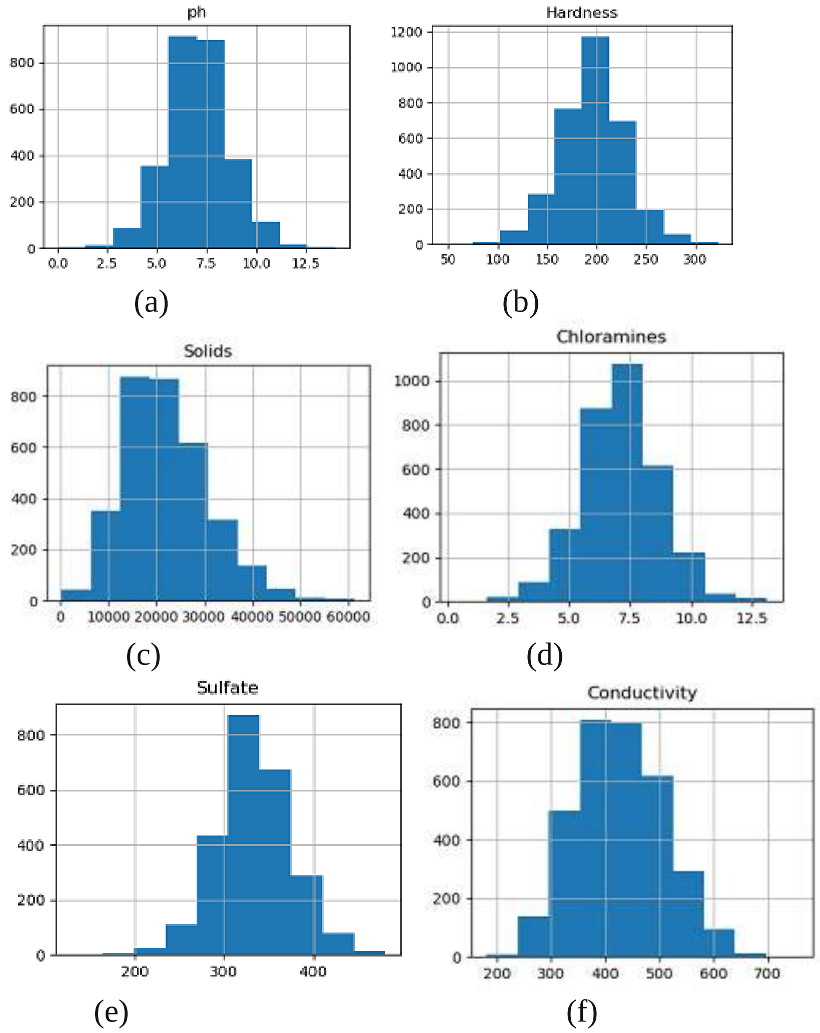


Fig: 2 Overall data in the dataset

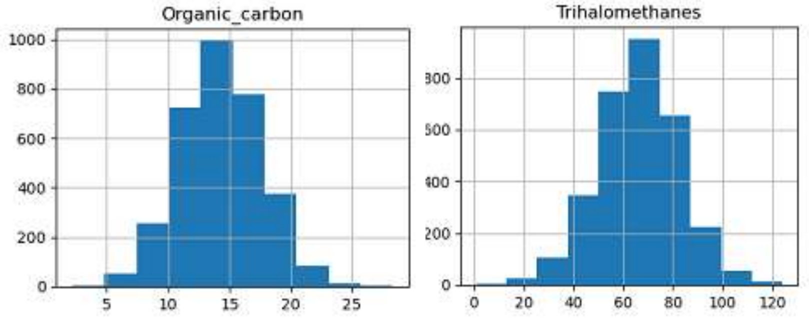
Fig. 2 shows the proportion of data collected in the dataset whether water is portable or not. From Fig. 2, it is clear that 60.5 % samples are in the category of not potable indicating that unsuitable for human consumption. Natural phenomena such as climate change and erosion are the main causes of poor water quality.



(a) (b)

(c) (d)

(e) (f)



(h) (i)

Fig.3 (a)-(k) histogram of attribute used in the dataset

In fig. : 3(a)-(k) shows the histogram distribution of each attribute used in the dataset.

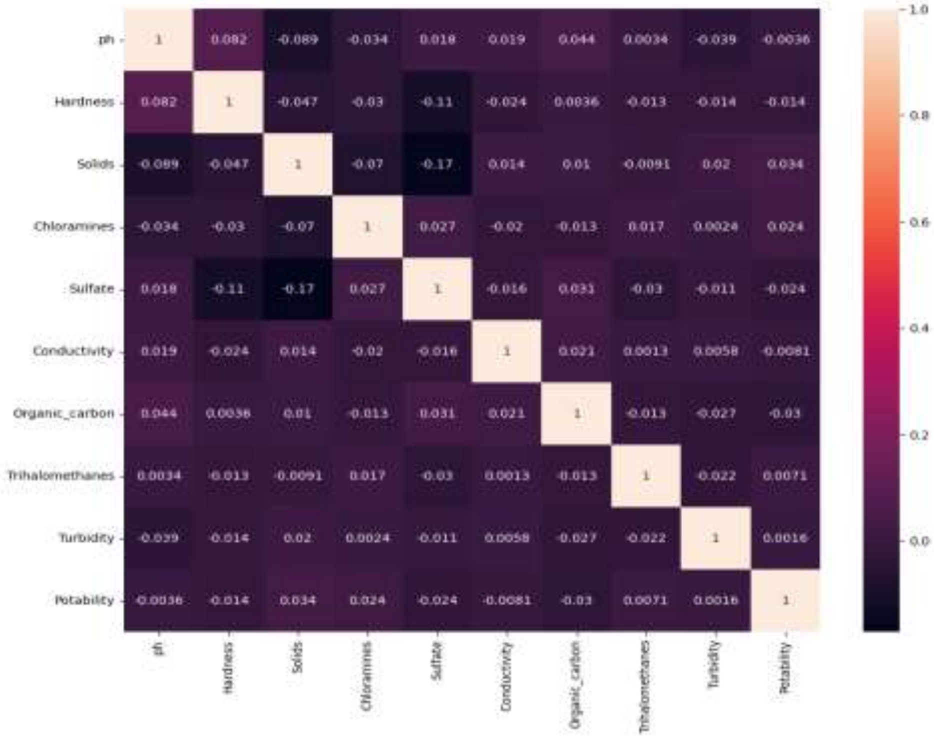


Fig. : 4 Heatmap of attributes in the dataset

In Fig.4 Shows that correlation between the data in the dataset. The relationship between various features in the dataset is depicted by correlation heatmap as shown in Figure 4.

3.3 Model Selection: SVM, Random Forest, Logistic Regression and KNN are used for predicting potability of water.

4. Result and Discussion

The simulation outcomes are derived from Python. Initially, the data undergoes several preprocessing steps to ensure cleanliness. Subsequently, the dataset is split into an 80:20 ratio, with 80% randomly allocated for training the model and 20% for testing its performance. Following this, cross-validation and hyperparameter tuning are conducted to select the optimal features from the data. Table 1 below displays the model performance achieved through cross-validation and hyperparameter tuning, along with the parameters utilized.

Table 1. Model performance achieved through cross-validation and hyperparameter tuning

Sr. No.	**Model**	**Best score**	**Best _params**
1	SVM	0.621	{'C': 50, 'kernel': 'rbf'}
2	Random Forest	0.631	{'n_estimators': 100}
3	Logistic Regression	0.610	{'C': 1}

4	KNN	0.613	{'n_neighbors': 13}

The performance of models following the handling of imbalanced data using SMOTE is displayed in Table 2.

Table 2. Performance of models following the handling of imbalanced data using SMOTE

Sr. No.	Model	Best score	Best _params
1	SVM	0.547	{'C': 50, 'kernel': 'rbf'}
2	Random Forest	0.696	{'n_estimators': 100}
3	Logistic Regression	0.516	{'C': 1}
4	KNN	0.679	{'n_neighbors': 13}

The table clearly indicates that the best score is achieved by Random Forest when utilizing cross-validation and hyperparameter tuning, as shown in Table 1. Furthermore, when employing the SMOTE technique to enhance the Random Forest model's performance, this improvement is evident in Table 2. Specifically, the accuracy for predicting portable water by Random Forest increases from 68.25% to 69.62%.

5. Conclusion

Various machine learning techniques such as SVM, Random Forest, Logistic Regression and KNN have been used to predict the portability of water resources. First, the cross validation has been performed. Then SMOTE technique has been utilized to improve the performance of various machine learning models. Using SMOTE to address the imbalanced dataset significantly improved the F1-score of "1" Potability from 0.46 to 0.74. To effectively manage water resources, it's crucial to prioritize ecological restoration and build systems to allocate water use based on industry, agricultural, and drinking demands. This framework requires an evaluation of both existing and prospective initiatives.

References:

[1] M. YURTSEVER and M. EMEÇ, “Potable Water Quality Prediction Using Artificial Intelligence and Machine Learning Algorithms for Better Sustainability,” Ege Akad. Bakis (Ege Acad. Rev., no. 2022, 2023, doi: 10.21121/eab.1252167.

[2] K. L. Christ and R. L. Burritt, “Supply chain-oriented corporate water accounting: a research agenda,” Sustain. Account. Manag. POLICY J., vol. 8, no. 2, pp. 216–242, 2017, doi: 10.1108/SAMPJ-05-2016-0029.

[3] L. Jiang, Y. Gao, H. Xu, Q. Yao, and L. Wu, “System Model of Green Building Supply Chain,” in 2ND INTERNATIONAL CONFERENCE ON EDUCATION, MANAGEMENT AND SYSTEMS ENGINEERING (EMSE 2017), 2017, pp. 397–402.

[4] J. Patel et al., “A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI,” Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/9283293.

[5] F. Azmi, M. K. Gibran, and A. Ridwan, “Enhancing Water Potability Assessment Using Hybrid Fuzzy-Naïve Bayes,” Indones. J. Comput. Sci., vol. 12, no. 3, pp. 1032–1043, 2023, doi: 10.33022/ijcs.v12i3.3232.

[6] B. Ainapure, N. Baheti, J. Buch, B. Appasani, A. V. Jha, and A. Srinivasulu, “Drinking water potability prediction using machine learning approaches: a case study of Indian rivers,” Water Pract. Technol., vol. 18, no. 12, pp. 3004–3020, 2023, doi: 10.2166/wpt.2023.202.

[7] Y. Khan and C. S. See, “Predicting and analyzing water quality using Machine Learning: A comprehensive model,” 2016 IEEE Long Isl. Syst. Appl. Technol. Conf. LISAT 2016, pp. 1–6, 2016, doi: 10.1109/LISAT.2016.7494106.

[8] A. Najah Ahmed et al., “Machine learning methods for better water quality prediction,” J. Hydrol., vol. 578, 2019, doi: 10.1016/j.jhydrol.2019.124084.

[9] E. M. Dogo, N. I. Nwulu, B. Twala, and C. Aigbavboa, “A survey of machine learning methods applied to anomaly detection on drinking-water quality data,” Urban Water J., vol. 16, no. 3, pp. 235–248, 2019, doi: 10.1080/1573062X.2019.1637002.

[10] I. O. Bucak and B. Karlik, “CMAC tabanli yapay sinir ağlari kullanilarak I⨀çme suyu kalitesinin tespiti,” Ekoloji, vol. 81, no. 78, pp. 75–81, 2011, doi: 10.5053/ekoloji.2011.7812.

[11] M. B. Addisie, “Evaluating Drinking Water Quality Using Water Quality Parameters and Esthetic Attributes,” Air, Soil Water Res., vol. 15, no. 1, 2022, doi: 0.1177/11786221221075005.