**Water Potability Prediction Model**

**Abstract.**

One of the most important things on earth is water. People need water to live, and that includes water to drink. It's important to find out if there will be enough drinking water for everyone now and in the future. But not every place on Earth has the same amount of water supplies. Some countries and areas have a lot of water, while other places don't have enough. The water resources of each area should be looked at on their own. In this work, the authors use an Indian water potability dataset from Kaggle to study how safe water is to drink. More specifically, this study uses the binomial distribution and the k-nearest neighbour algorithm to talk about each factor of water that affects how safe it is to drink. Also, the writers make a model that lets people figure out how potable a water source is based on the data for each of its factors. The study shows that the properties of water have nothing to do with each other. To get water that can be drunk, all of the parts must meet a certain standard.

# **Introduction**

Safe and accessible water is essential for public health, whether it is used for drinking, home consumption, food production, or recreation. It is possible that countries may prosper and experience a reduction in poverty if water supplies were increased and water resources were better managed. The main reason for the deterioration of water quality is the release of pollutants into rivers; this is especially true in India, where there are many industrial regions. Water quality declines for a variety of reasons, including human waste (plastics) and the dumping of undesired items into water sources like rivers, ponds, lakes, and the ocean. Therefore, these are the causes of the current state of water. Typhoid, dysentery, polio, cholera, hepatitis, and diarrhoea are only some of the illnesses linked to dirty water and a lack of sanitation. The lack of, inadequacy of, or improper management of water and sanitation facilities exposes people to avoidable health risks. This is especially true in hospitals, where viruses and germs can spread due to inadequate water, hygiene, and cleaning services. Fifteen percent of hospital patients worldwide contract a virus during their stay, with that number rising dramatically in poorer regions. Careful consideration must be given to the selection of potable water. To overcome this obstacle, numerous domain-acknowledgements are needed. This system is designed to learn as much as it can about its data source while yet being as generic as possible. However, 70% of India's drinkable water has been tainted by industrial and household contaminants.

Eighty percent of the rural populace and twenty percent of city dwellers lack access to safe drinking water. Infectious diseases and environmental factors, particularly water supply and sanitation, are responsible for 75% of the nation's children's health problems. In children younger than five, diarrhoea is the leading cause of death (46% of all deaths). Many cases of diarrhoea are caused by water-related illnesses. The Ethiopian Ministry of Health estimates that every day 6,000 children in the country die as a result of water loss due to diarrhoea.

The contribution of this work is as follows:

1.A first review was done on the available data to filter, normalise, and run classification algorithms to improve water quality and find the smallest part of interest that allows for a high level of accuracy at a low cost. So, similar studies in the future won't have to use expensive and time-consuming lab tests with specific sensors.

2.Several supervised prediction (classification and regression) methods are used to show how they work on the dataset. The whole method is put forward in the context of numerical water quality analysis.

3.In our code, we run some models and we get the 5 best models for our dataset. So, these are Random Forest, Decision Tree, Adaptive Boosting, and Naive Bayes.

# **Literature Review**

This study investigates the approaches utilised to address water quality issues [1]. In the majority of studies, traditional laboratory analyses and data analysis are two types of analyses used to determine the quality of water. However, other studies employ machine learning techniques to discover the optimal solution to the water quality issue.

Poor potable water quality has a detrimental effect on the health of consumers. According to reports, at least 2 billion people consumed feces-contaminated potable water worldwide. Awareness of the factors influencing the purity of potable water is required for the development of accurate decisions regarding its regulation and protection. The purity of potable water is typically determined by the source water, how it is handled prior to delivery, how it is distributed, how it is maintained, and how well it is filtered at the residence. In addition, in rural areas and small towns, potable water is frequently extracted directly from wells or obtained unfiltered from rivers, lakes, and reservoirs. Therefore, the purity of the supply water has a substantial impact on the quality of the drinking water. In recent years, many developing nations have made the reduction of waterborne diseases and the development of secure water supplies a significant public health objective, and the situation has slightly improved. However, the situation is far from optimal, particularly in rural areas, and even marginally improved conditions may be jeopardised by population growth and economic development-induced increases in water consumption and decreases in water availability. It is essential to use a practicable and efficient evaluation method for drinking water quality in order to obtain reliable results and make informed decisions.

Since Horton created the first Water Quality Index (WQI) in the 1960s [2], numerous water quality evaluation methods have been proposed. The two indices for determining the general state of potable source water quality are simple, adaptable, and stable, with minimal input data sensitivity. Likewise, we utilised the weighted arithmetic WQI method to provide information on water quality. Despite significant flaws, these WQIs are the most widely used water quality assessment instrument due to their ability to convert a large number of variables into a single number that aids in understanding water quality. Recent water quality assessments in a wastewater irrigation area and a swiftly urbanising area [2] used matter element extension analysis (MEEA) and entropy TOPSIS, respectively. Both mathematical approaches are precise in estimating the overall water quality. The classification of these water quality evaluation methodologies is based on water quality standards. Therefore, it is essential to establish water quality guidelines.

All water utilities must provide consumers with an adequate, dependable source of greater potable water at a price proportional to the water system's needs. To achieve this objective, its freshwater must be purified and supplied from the largest feasible source in sufficient quantities to meet sector regulations and moisture levels standards. When determining the purity of drinking water, it is important to consider consumer acceptance, proven treatment procedures, and effective utility management. High-quality water is distinguished by the absence of hazardous organisms and biological forms that may be aesthetically unappealing. It is transparent and colourless, with no foul odour or taste. It lacks chemical concentrations that could be harmful to the body, unattractive to the eye, or financially damaging. It is also noncorrosive and does not leave behind excessive or undesirable deposits on water-transporting structures such as pipelines, tanks, and plumbing fixtures.

In their paper, Yafra Khan and Chai Soo See [3] use Artificial Neural Network and time series analysis to construct a water quality prediction model. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Regression Analysis were used to evaluate the performance of the model. In their paper, Dao Nguyen Khoi et al. [4] estimate water quality using 12 machine learning models. Two statistics, R2 and RMSE, were utilised for model evaluation. Umair Ahmed et al. [5] have estimated the Water Quality Index (WQI) with supervised machine learning algorithms. Sabre Kouadri et al. [6] utilised eight artificial intelligence algorithms to predict Water quality Index. Several statistical metrics, including correlation coefficient (R), mean absolute error (MAE), root mean square error (RMSE), relative absolute error (RAE), and root relative square error (RRSE), were used to evaluate models. Using sensor networks, Jitha Nair and Vijaya M S [7] utilised a variety of prediction models created with machine learning and big data techniques.

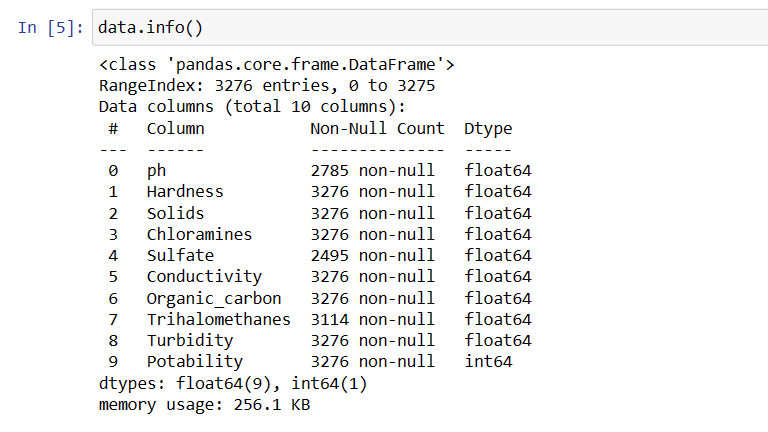
XGB (XGBoost), RF (Random Forest), DTC (Decision Tree), Adaptive Boosting (AdaBoost), and SVC were used to evaluate water quality, with XGB having the highest accuracy of 83% (XGBoost) [8]. All of the variables in the dataset, including hardness, sulphate, solid, trihalomethanes, pH, turbidity, solids, organic carbon, and conductivity, are measured in accordance with World Health Organisation (WHO) guidelines as part of their work, which is focused on water quality [5]. Using these indicators and comparing them to established values is a crucial limitation when estimating water quality. A detailed perspective of the system we have described is provided in Figure 1.

# **Methodology**

In order to trained the model, we used the water\_potability.csv dataset present on the Kaggle. This dataset size of dataset is 3276(rows)x10(columns). Each row present the some chemical property value of water like ph, Hardness, Sulfate, solids, as show in below fig.

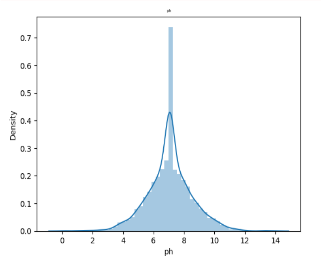
Last column which is of “Potability” value act as target column. Column contains value as 0 or 1.

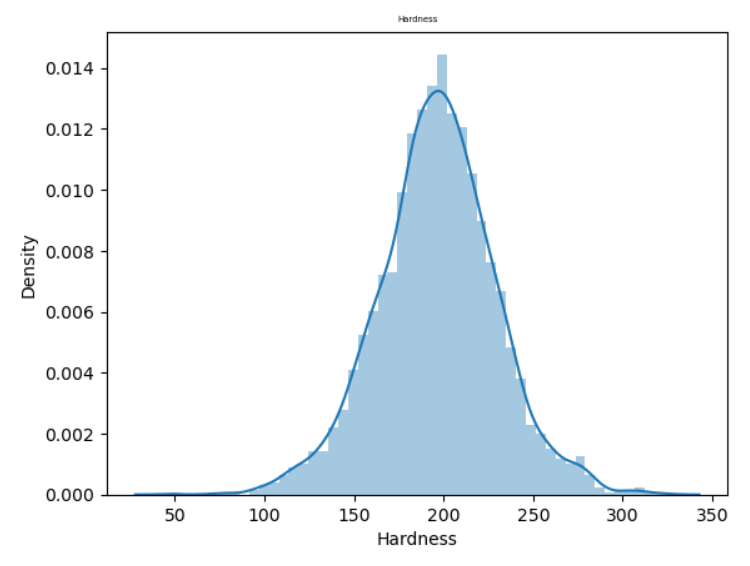
1 means water is potable and 0 is not-potable.

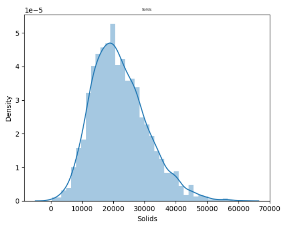


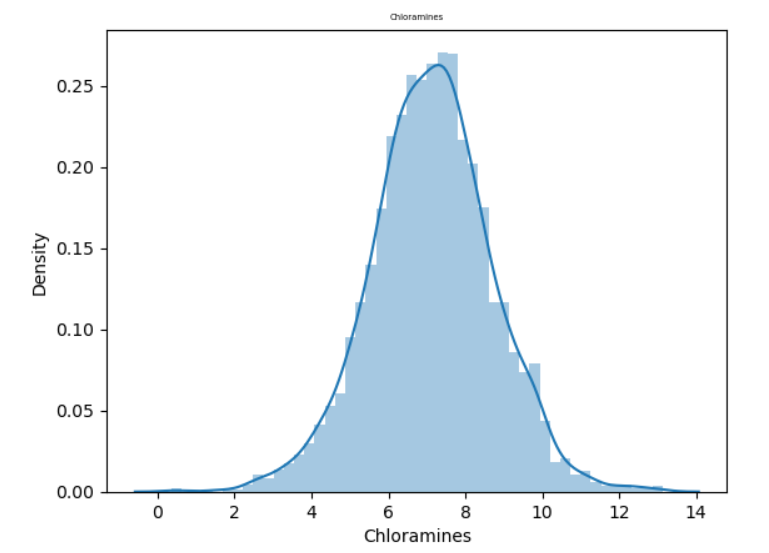
Data field had many empty values. We filled that value with mean value of whole column containing that empty value as part of pre-processing. One advantage we get is no duplicate value present in the whole data.

Every column data is almost normally distributed.

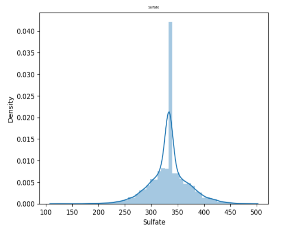
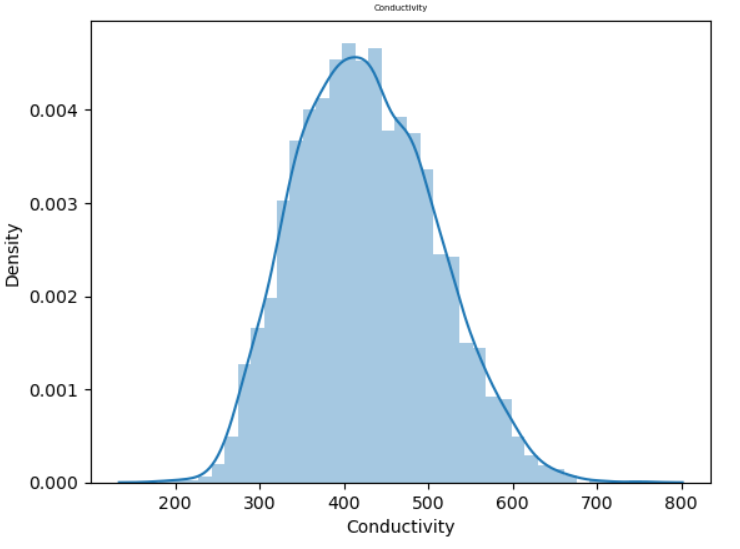




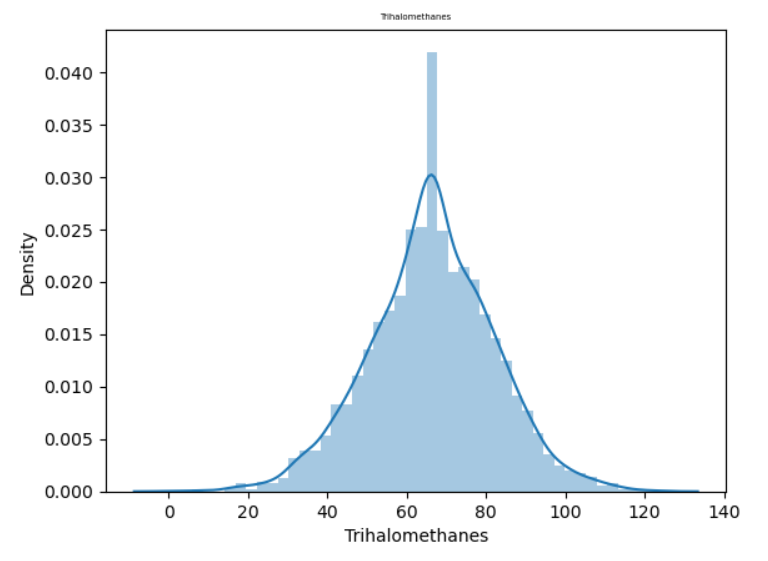
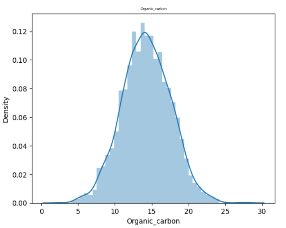
 Ph Hardness



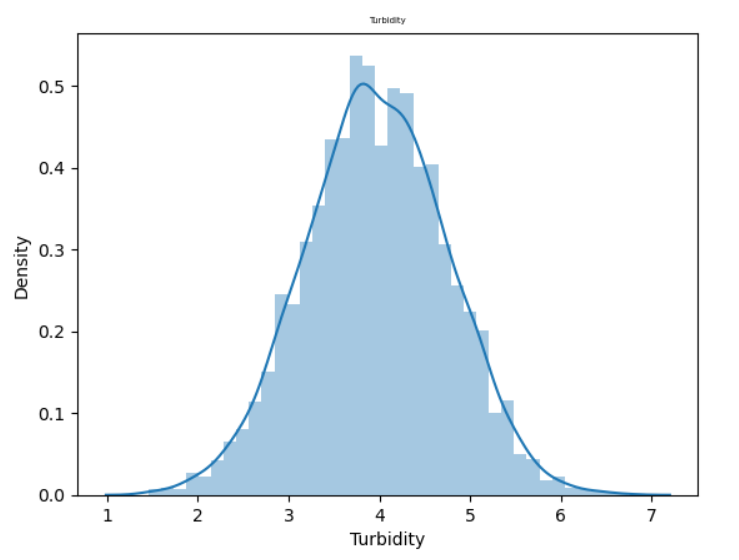
Solid Chloramines



Sulfate Conductivity

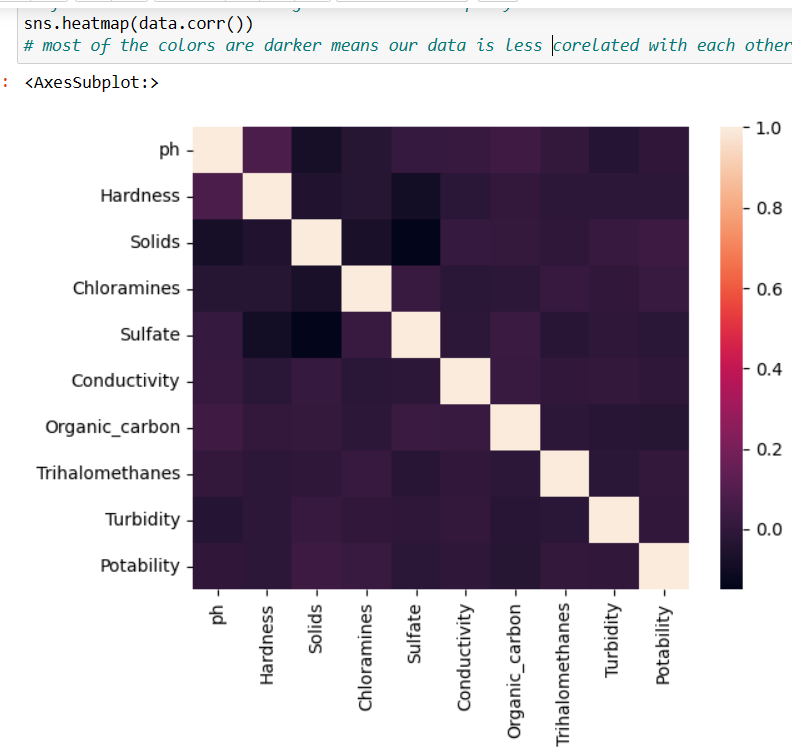


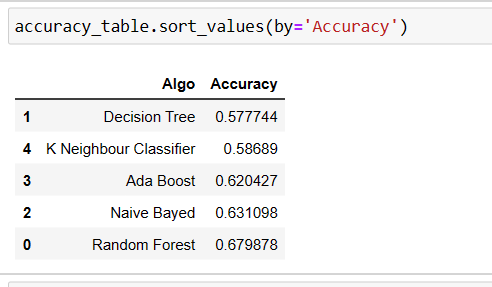
Organic\_carbon Trihalomethanes



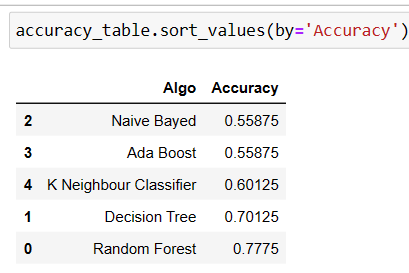
Turbidity

Even columns have less correlation with other. So, need to remove any column from dataset.



After pre-processing we train the model by splitting the data with test\_size=0.2. we used five ml algorithms to train model. The algorithms are RandomForestClassifier, DecisionTreeClassifier, GaussianNB, AdaBoostClassifier and KNeighborsClassifier. 

As given in figure we got maximum accuracy with Random Forest Classifier (67.98%). Then we observed in dataset that out of 3276 rows, “potability” columns contains 1998 rows for non-potable(0) and 1278 for potable(1) as result. Due to more values of 0(non-potable) our models also giving bias prediction. To resolve this issue we used the *oversampling technique* to make our target column data in equal proportion. We used RandomOverSampler() which make our data in proportion. After over\_sampling our efficiency got increased by 10%.



This is the result after over\_sampling and again Randomforest comes with highest accuracy of 77.75%.

Further, to improve the efficiency we used Hyper Parameter Tuning but it didn’t affect the accuracy much. We got 78.75% accuracy using GridSearchCV.

# **Conclusion and Future Work**

This study examined the machine learning performance of Random Forest, Decision Tree, Adaptive Boosting, and Naive Bayes in predicting the components of a dataset on water quality. This objective was achieved by acquiring variables from the most well-known datasets, such as pH, hardness, particulates, EC, and turbidity. The outcomes demonstrated that the employed models were effective at predicting water quality metrics. To make the selection procedure more efficient, additional research will be conducted. To develop systems that incorporate mentioned deep learning strategies and methods, as well as others.

In order to determine the characteristics that would be most helpful for forecasting models, the first step is to obtain data for our models, and then we may move on to the next step. It is necessary for us to perform data preprocessing so that we may correct any errors that may be present in our dataset. These errors may take the form of missing numbers or data that has been poorly adjusted. After then, in order to evaluate the accuracy of the model, our dataset will be divided into two parts: the Train part and the Test part. After that, we will implement a machine learning model by using our dataset as a starting point. In order to obtain the level of accuracy that we desire, we will first need to acquire accuracy, and then we will need to refine our model by modifying its hyperparameters.

In future research, we suggest incorporating the research's findings into a substantial Internet of Things system that relies on only the relevant parameter sensors. Based on the IoT system's real-time data, the investigated algorithms would generate an immediate forecast of the water quality. Before water is released for public ingestion, it would detect toxic water and notify the appropriate authorities. Eventually, fewer people will drink water of poor quality, thereby reducing the prevalence of dreadful diseases such as typhoid and diarrhoea. In this regard, the adoption of a predictive evaluation of projected values would result in the development of future decision- and policy-supporting instruments.

# **References**

1. A. N. Prasad, K. Al Mamun, F. R. Islam, and H. Haqva, “Smart water quality monitoring system,” in *Proceedings of the 2nd IEEE Asia Pacific World Congress on Computer Science and Engineering*, December 2015.
2. P. Li and J. Wu, “Drinking water quality and public health,” *Exposure and Health*, vol. 11, no. 2, pp. 73–79, 2019.
3. Y. Khan and C. S. See, “Predicting and analyzing water quality using machine learning: a comprehensive model,” in *Proceedings of the 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, April 2016.
4. D. N. Khoi, N. T. Quan, D. Q. Linh, P. T. T. Nhi, and N. T. D. Thuy, “Using machine learning models for predicting the water quality index in the La buong river, Vietnam,” *Water*, vol. 14, no. 10, p. 1552, 2022.
5. U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. Garc´ıa-Nieto, “Efficient water quality prediction using supervised machine learning,” *Water*, vol. 11, p. 2210, 2019.
6. S. Kouadri, A. Elbeltagi, A. R. M. T. Islam, and S Kateb, “Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast),” *Applied Water Science*, vol. 11, no. 12, p. 190, 2021.
7. J. P. Nair and M. S. Vijaya, “Predictive models for river water quality using machine learning and big data techniques - a Survey,” in *Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, IEEE, Coimbatore, India, March 2021.
8. M. M. Hassan, M. M. Hassan, L. Akter et al., “Efficient prediction of water quality index (WQI) using machine learning algorithms,” *Human-Centric Intelligent Systems*, vol. 1, no. 3-4, pp. 86–97, 2021.
9. B. Charbuty and A. M. Abdulazeez, “Classification based on decision tree algorithm for machine learning,” *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.
10. M. I. K. Haq, F. D. Ramadhan, F. Az-Zahra, L. Kurniaw, and A. Helen, “Classification of water potability using machine learning algorithms,” in *Proceedings of the 2021 International Conference on Artificial Intelligence and Big Data Analytics*, October 2021.
11. J. Wu, X.-Y. Chen, H. Zhang, Li-D. Xiong, H. Lei, and Si-H. Deng, “Hyperparameter Optimization for machine learning models based on bayesian Optimization,” *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, 2019.
12. C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, “An introduction to logistic regression analysis and reporting,” *The Journal of Educational Research*, vol. 96, no. 1, pp. 3–14, 2002.
13. M. Mahajan and K. Bhardwaj, “Potability analysis of drinking water in various regions of Ludhiana District, Punjab, India,” *International Research Journal of Pharmacy*, vol. 8, no. 6, pp. 87–90, 2017.
14. Y. Meride and B. Ayenew, “Drinking water quality assessment and its effects on residents health in Wondo genet campus, Ethiopia,” *Environmental Systems Research*, vol. 5, no. 1, pp. 1–7, 2016.
15. Y. Xie, B. Xie, Z. Wang et al., “Geological resource Planning and environmental Impact assessments based on GIS,” *Sustainability*, vol. 14, no. 2, p. 906, 2022.
16. C. A. Moreno-Camacho, J. R. Montoya-Torres, A. Jaegler, and N Gondran, “Sustainability metrics for real case applications of the supply chain network design problem: a systematic literature review,” *Journal of Cleaner Production*, vol. 231, no. 10, pp. 600–618, 2019.
17. G. Douzas, F. Bacao, and F. Last, “Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE,” *Information Sciences*, vol. 465, pp. 1–20, 2018.
18. A. S. Abobakr Yahya, A. N. Ahmed, F. Binti Othman et al., “Water quality prediction model based support Vector machine model for Ungauged river Catchment under dual Scenarios,” *Water*, vol. 11, no. 6, p. 1231, 2019.