Prediction of Water Quality Index using Machine Learning

Janvi Behal
Computer Science and Engineering
IGDTUW
Delhi, India
janvi076btcse23@igdtuw.ac.in

Jia Dholpuria
Computer Science and Engineering
IGDTUW
Delhi, India
jia078btcse23@igdtuw.ac.in

Dikshita Yadav
Computer Science and Engineering
IGDTUW
Delhi, India
dikshita052btcse23@idgtuw.ac.in

***Abstract*— With the advancements of industries and needs, industrial scraps and wastes are hugely dumped in the water bodies. Hence, we have proposed a machine learning predictive model to analyse the Water Quality Index (WQI), or weighted arithmetic water quality index (WAWQI) to be specific. It is a useful tool to understand the levels of pollution in water bodies, groundwater, and surface waters in the spotlight. There are several parameters, such as pH, conductivity, temperature, biological oxygen demand (BOD), dissolved oxygen (DO), etc., which are used to determine the WAWQI and determine whether the water is safe for drinking and civil use or not. According to the standards set by the National Library of Medicines (NIH), the ranges of WAWQI established(figure 10) are Excellent (Grade 'A' or 1, range 0-25), Good ('B' or 2, range 26-50), Poor ('C' or 3, range 51-75), Very Poor ('D' or 4, range 76-100) and Unsuitable ('E' or 5, range>100). For the given dataset, multiple linear regression has been employed to predict the WQI, whereas the potability is determined through the random forest algorithms learning from the WQI.**

In conclusion, for multiple linear regression, it requires eight inputs and one output, which is WQI. In the random forest algorithm, WQI is taken as input and the output will be a number between 1 and 5, corresponding to the potability of water. After training and testing of the model, it has exhibited precision.

***Keywords*— WQI, WAWQI, water potability, prediction, prediction model, machine learning algorithms, random forests, multiple linear regression, parameters, features**

I. Introduction

Water is the backbone of human existence. It is an essential resource which is abundantly distributed all over the earth and covers about 71% of the earth's surface. Freshwater constitutes about 0.3% of the total water on earth. Without access to clean water, survival cannot be imagined for more than 3-4 days.

With the advancement of human needs and technology, water faces some major threats. The industrial waste and scraps are dumped into the lakes and rivers, which severely pollutes the water. Every day, pollutants like chemicals, plastics, and waste are poured into our oceans, rivers, and lakes, harming aquatic life and even human health. This poses a threat to aquatic life, promotes water-borne diseases, hence making the freshwater unfit for consumption.

More than 80 % of the world's wastewater streams back into the environment without being treated or reused, according to the United Nations; in a few least-developed nations, the figure tops 95%. Hence, it is essential to check the contamination and pollution levels in water.

The Water Quality Index is a vital tool for evaluating the suitability of water resources for human consumption and other uses. It uses a numerical scale to assess the level of pollution in water bodies, considering factors such as pH, temperature, conductivity, and biological oxygen demand.

In our model, we have made use of the Weighted Arithmetic Water Quality Index (WAWQI), which takes into account various physical, chemical, and biological parameters like pH, conductivity, nitrate levels, and bacterial presence. According to resources, the permissible limits for potable water is 6.5 to 8.5 for pH, 200 to 800 umho/cm for conductivity, 5 to 10 mg/ L for BOD (Biological Oxygen Demand), 6.5 to 8 mg/L for DO (Dissolved Oxygen), 25 to 30 degrees centigrade for temperature, 50-100 mg/ L for nitrate, <2500 MPN/100 mL for faecal coliform and 11 to 100 MPN/100 mL for total coliform. A lower WAWQI score indicates better water quality, while a higher score suggests poor water quality, potentially dangerous to human health and the environment.

This paper points to create a machine learning- based model for determining Water Quality (WQ) by foreseeing the WQI score and potability of water. Data was examined on different parameters to understand feature distribution and relationship between multiple parameters. Machine learning algorithms are utilised to figure the water quality and to classify the water quality index score to determine the potability. Mean squared error (MSE) and R squared precision is used for assessing the precision of prediction algorithms. Water quality index classification models are assessed using the metrics such as R2 score.

II. Literature Review

(2021) Md. Mehedi Hassan et al. [1] developed a reliable method for determining water quality by joining numerous machine learning calculations and procedures such as SVM (Support Vector Machine), NN (Neural Networks), MLR

(Multilayer Perceptron), BTM (Biterm Topic Model), and RF (Random Forest) to optimize show execution, wherein the maximum r2 score was evaluated to be 1.00 through multiple linear regression and 0.9698 was the least exactness obtained from support vector machine. The findings uncovered that the applied models performed well in predicting water quality parameters; though, the highest performance was linked with the MLR.

(2021) Saber Kouadri et al. [2] developed a machine learning model to improve the reliability of water quality appraisal. ANN (Artificial Neural Networks), MLR (Multiple Linear Regression), SVM (Support Vector Machine), M5P (M5 Model) Trees, RF (random forest), LWLR (Logical Weighted Linear Regression), RS (Remote Sensing), and AR (Augmented Reality) models were made in use for the estimation of WQI. Out of these models' minimum accuracy was 0.9412 proposed by the support vector machine and maximum precision was 0.1 through MLR and 0.9926 through RF model. It was hence noted that the ensemble tree-based model such as RF outperformed all the other models with considerable accuracy which can perform predictions without requiring regular large datasets and it appears that the MLR model is the most suitable for predicting the values of the water quality.

(2022) Jitha P Nair et al. [3] developed a machine learning-based model for forecasting Bhavani River water quality by integrating various machine learning techniques like the Random Forest, MLP regressor, Linear Regression, and Support Vector Regressor and Decision Tree, MLP classifier (Multilayer Perceptron), Naive Bayes, and support vector machine algorithms were enforced to make WQI classifiers. The R-squared value of support vector regressor based prediction model is-2.7132, whereas prediction models based on linear regression, random forest and MLP regressor labours 0.6375, 0.6923, and 0.7342 correspondingly. Thus, R- squared value is negatively produced by the support vector regressor which means it has lower precision and has high precision for MLP regressor. The accuracy of MLP classifier-based classification model is 0.8132, whereas classification models grounded on Naïve Bayes, Decision Tree and Support Vector Machine yield 0.7738, 0.74, and 0.61 correspondingly. Therefore, accuracy of MLP classifier is advanced as compared to other classifiers whereas support vector machine has low precision.

(2022) Mir Talas Mahammad Diganta et al. [4] developed a weighted quadratic mean for assessing coastal water quality by integrating various machine learning techniques like the random forest, decision trees, K nearest neighbours, linear regression, and the gaussian naive Bayes. The RMSE, MSE, MAE, R2, and PREI metrics were used in this study. The tree-based DT (RMSE = 0.0, MSE = 0.0, MAE = 0.0, R2 = 1.0 and PERI = 0.0) and the ExT (RMSE = 0.0, MSE = 0.0, MAE = 0.0, R2 = 1.0 and PERI = 0.0) and ensemble tree-based XGB (RMSE = 0.0, MSE = 0.0, MAE = 0.0, R2 = 1.0 and PERI = +0.16 to −0.17) and RF (RMSE = 2.0, MSE = 3.80, MAE = 1.10, R2 = 0.98, PERI = +3.52 to −25.38) models outperformed other models. The performance metrics of this model came out to be tree-based DT 100. One of its limitations is its inadequacy to assess the water quality in terms of temporal resolution.

(2022) Nida Nasir et al.[5] assessed various AI algorithms to develop a machine learning model for predicting water quality as precisely as possible. Various machine learning classifiers like as Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), CATBoost, XGBoost, and Multilayer Perceptron (MLP) and their stacking ensemble models were used to classify the water quality data via the Water Quality Index (WQI). Precision recall curves and Receiver Operating Characteristic curves (ROC) were also utilized to evaluate the performance of the different classifiers. The findings uncovered that the CATBoost model advertised the most precise classifier with an odd of 94.51. Moreover, after applying stacking ensemble models with all classifiers, precision reached 100 in varied Meta- classifiers. Likewise, the CATBoost achieved the topmost precision as a primary gradient boosting algorithm and a meta classifier. In terms of perfection, the order of the classifiers was as follows CATBoost (94.51%), random forest (94%), followed by MLP (88.6%), XGBoost (88.1%), decision tree (81.6%), SVM (80.7%), and logistic regression (72.9%).

(2023) Fei Ding et al. [6] used machine learning and game theory to optimize the combined weights by integrating various machine learning techniques like coefficient of variation (CV), Kaiser-Meyer-Olkin (KMO), Sinusoidal Weighted Mean (SWM) , Log-weighted Quadratic Mean (LQM) , Analytic Hierarchy Process and Entropy Weight Method (EWM).The three water quality assessment models (WQIS, WQIL and WQIW) were established based on the optimal weights besides. All three models had good reliability. Both WQIS and WQIW models had low eclipsing problems (25.49% and 18.63%). The accuracy of the models was ranked as WQIS > WQIW > WQIL. The weights and aggregation functions are important structures of the WQI model and determine the model accurately. Machine learning was used to calculate the weights in this study.

III. Materials and Methods

We collected water quality related data from Kaggle and more concretely, this dataset contained features such as Temperature, pH, dissolved oxygen, nitrate etc that can affect the water quality making it good or bad. We used 795 samples as data and each class contains 8 features. We visualised our data in a histogram to see how the data distributes. Next, we looked at how changes in one variable correspond with changes in another for correlation analysis using heatmap. We then separated the data into two sets — one for training our models and another to test them. Random Forest and Multiple Regression are the two machine learning algorithms we have used to train our models. We have used the trained models to make predictions on our testing set and evaluated accuracy using the R squared method as a performance metric. We have also conducted thorough research on the factors in our dataset like temperature, pH, conductivity, dissolved oxygen as shown below.

A. Data Preprocessing

Preprocessing data helps to increment data quality and efficiency. The quality of unprocessed data is hampered by its inconsistency.

- BOD:- Biological Oxygen Demand is a measure of how much oxygen is needed to break down organic matter in water
- DO:- Dissolved Oxygen is the amount of oxygen dissolved in water
- pH:- It is a measure of the acidity or basicity level in solution that ranges from 0 to 14:

  ➢ A solution with pH less than 7 is termed as an acidic solution.

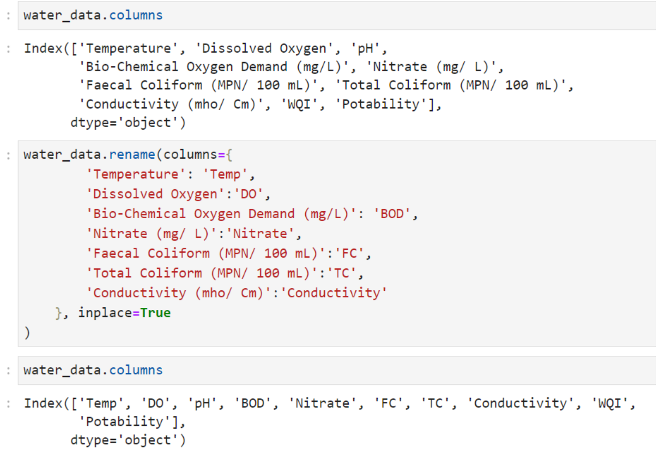
  ➢ Neutral solution has a pH of 7, this indicates the solution is neither acidic nor basic.

  ➢ A pH of more than 7 signifies a basic (or also an alkaline) solution.

  ➢ In the case of water quality, pH is important because it has an impact on solubility and bioavailability of nutrients/pollutants but also aquatic life.
- TC:- Total coliform is a large group of bacteria that are found in soil, water, and human or animal waste
- FC:- Faecal coliforms are types of total coliform that mostly exist in faeces

B. *Exploratory Data Analysis*

For exploratory data analysis, we first checked the shape of our dataset and got a view of the same to get an introductory idea of our data and what we are dealing with. We then ran some methods to get information of data types which make up our data. To make the data easier to handle, we renamed the columns as per our convenience. (fig 1)



```
water_data.columns
```

```
Index(['Temperature', 'Dissolved Oxygen', 'pH',
       'Bio-Chemical Oxygen Demand (mg/L)', 'Nitrate (mg/ L)',
       'Faecal Coliform (MPN/ 100 mL)', 'Total Coliform (MPN/ 100 mL)',
       'Conductivity (mho/ Cm)', 'WQI', 'Potability'],
      dtype='object')
```

```
water_data.rename(columns={
        'Temperature': 'Temp',
        'Dissolved Oxygen':'DO',
        'Bio-Chemical Oxygen Demand (mg/L)': 'BOD',
        'Nitrate (mg/ L)':'Nitrate',
        'Faecal Coliform (MPN/ 100 mL)':'FC',
        'Total Coliform (MPN/ 100 mL)':'TC',
        'Conductivity (mho/ Cm)':'Conductivity'
    }, inplace=True
)
```

```
water_data.columns
```

```
Index(['Temp', 'DO', 'pH', 'BOD', 'Nitrate', 'FC', 'TC', 'Conductivity', 'WQI',
       'Potability'],
      dtype='object')
```

Fig 1

C. *Handling Missing Values*

Every data is not entirely perfect. There are always some missing values which can cause discrepancies. Hence, we need to handle these null values.

Firstly, we viewed the sum of the number of null values for each column. Next, we calculated the mean of the data for columns which had null values. These null values were replaced with the mean we calculated earlier. Just to make sure there are no null values, we again viewed the sum of the number of null values present.

Figure 2 represents a bar graph of sum of null values of all the columns of the dataset.

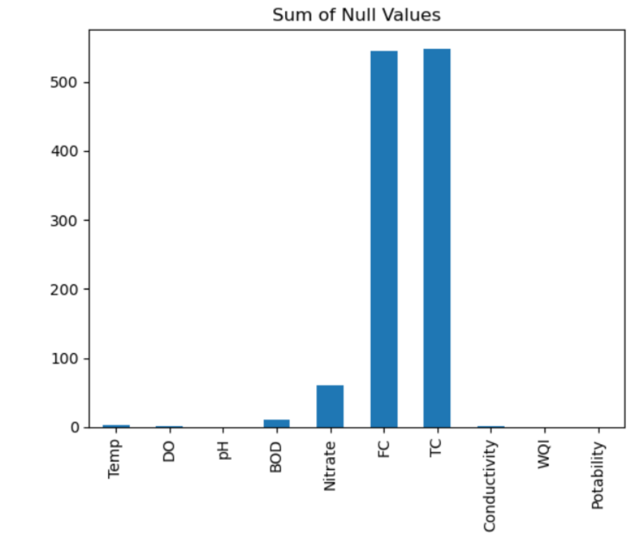


Fig 2

D. *Data Distribution*

Data distribution in machine learning is how data is arranged across a range of values, and how it may cluster around a specific value, be scattered evenly, or skew in one direction. Data distribution refers to the way data values are spread or scattered in a dataset.

It points at giving important insights, informs decision-making, and ensures that suitable methods are utilized for statistical examination and modelling.

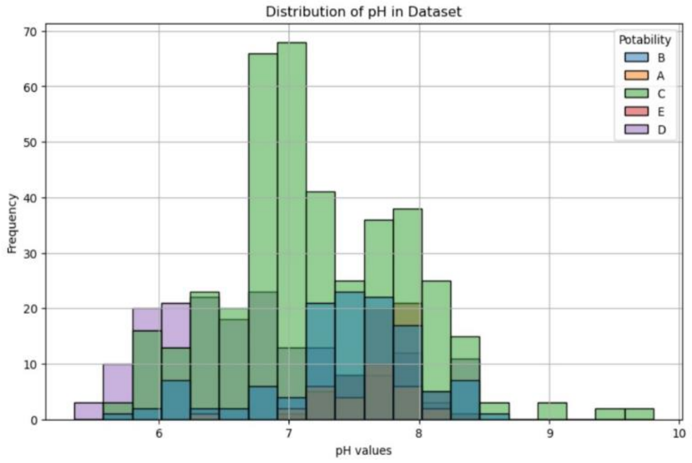


Fig 3

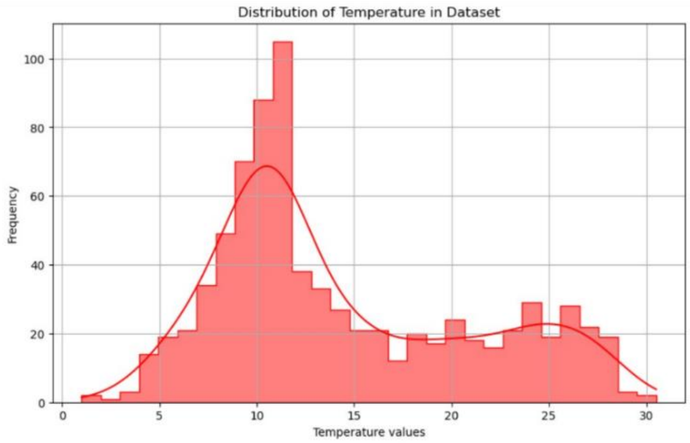


Fig 4

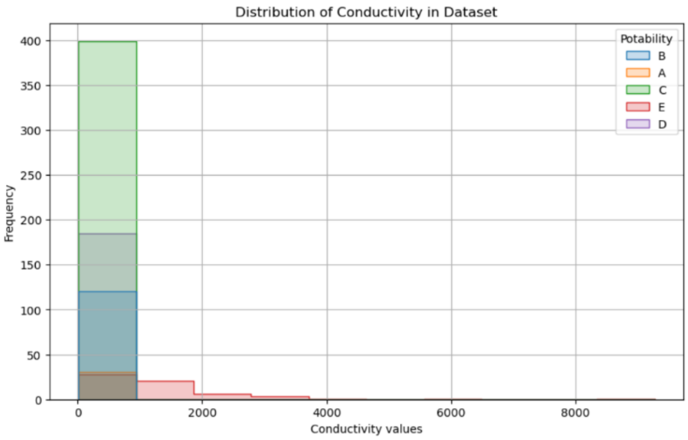


Fig 5

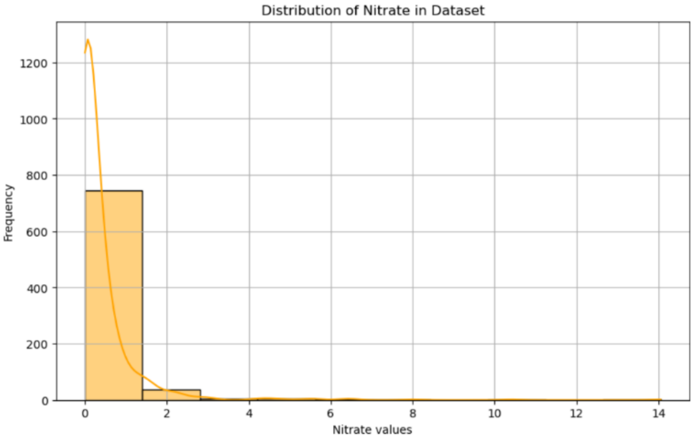


Fig 6

In this work, the data distribution is shown in Figure 3, 4, 5, and 6 for the features pH, Temperature, conductivity and nitrate respectively. All the features of data have been distributed differently with missing value and without missing value.

E. Data Splitting

- The dataset was split into two parts for multiple linear regression, X and y. Later this X was further split into X_train and X_test and y was further split into y_train, y_test. In X, values of all the rows of 1 to 8 columns were extracted, while in y only all rows of the 9th column were extracted. 80% of the data was split into X_train and y_train while 20% of data was stored in X_test and y_test.
- The dataset was again split into two parts for random forest classification, A and B. These were further split into A_train, A_test, B_train, B_test respectively. In A, values of all the rows of the 9th column were extracted and for B, values of all the rows of the 10th column were stored. Again, the training and testing data was split in a ratio of 4:1 respectively.
- The ML develops a relationship between independent and dependent features to predict on the basis of training data, later it is tested on the basis of testing data to assess the accuracy.

Dataset
Training Set
Test Set

[7] Fig 7

F. Feature Correlation

Correlation speaks to the relationship between two or more parameters. One of the parameters with correlation of more than a given value is dropped as the data is not much dependent on that parameter.

To produce a relation between the various parameters of a dataset, we use correlation matrix. The matrix contains all the parameters with potential values signifying the intensity of correlation between all parameters.

This matrix helps us in analysing the trends followed by the parameters. To plot this correlation matrix, we used the heatmap visualisation from seaborn library. As shown in figure 8, no two parameters are highly related to each other, or have value greater than 1.0.

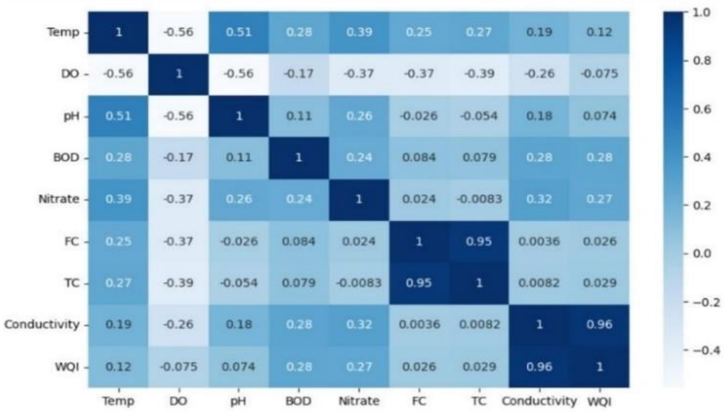


Fig 8

G. *Water Quality Index*

Water quality index (WQI) is calculated to deliver an estimate about the quality or potability of water from various resources. Water Quality Index is calculated after considering various factors which influence the water quality. We have used the Weighted Arithmetic Water Quality Index (WAWQI) method for estimating the water quality. In this method various physical and chemical parameters such as temperature, pH, conductivity etc are used to work out the quality of water.

The WAWQI is assessed from Equation (1).

$$WQI = \frac{\Sigma wiqi}{\Sigma wi} \qquad (1)$$

[8] Fig 9

The dataset which we came through already had the calculation of WAWQI done. Hence, we didn't need to calculate WQI again which narrowed down our task

0-25	A	Excellent
25-50	B	Good
50-75	C	Poor
75-100	D	Very Poor
>100	E	Unsuitable

Fig 10

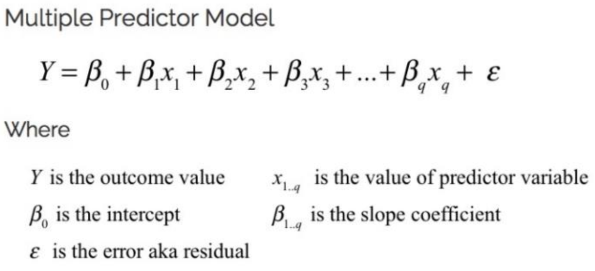
IV. Applying Machine Learning Algorithms

For our research, after examining numerous above-mentioned case studies related to predicting water quality, we came to the conclusion that machine learning algorithms like Multiple Regression and Random Forest Classifier give the best output by producing maximum accuracy in predicting water quality. Hence, to accomplish this aim, we have made use of the respective algorithms for our model.

A. *Multiple Linear Regression*

In Machine Learning, when we want to set up a connection between multiple independent variables and one dependent variable, we use Multiple Linear Regression.

Multiple linear regression, also known as Multiple Regression works on the model given below (fig 11):



Multiple Predictor Model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_q x_q + \varepsilon$$

Where

Y is the outcome value

β_0 is the intercept

ε is the error aka residual

$x_{1..q}$ is the value of predictor variable

$\beta_{1..q}$ is the slope coefficient

[9] Fig 11

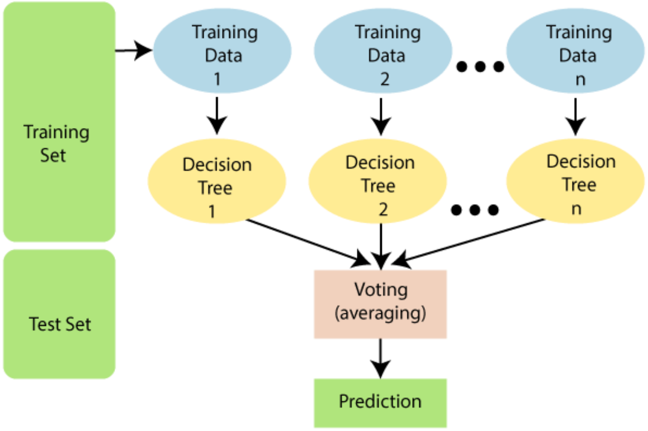
In our model, the WQI is set to be the dependent variable while all the parameters before WQI i.e., 1 to 8 columns are set to be the independent variables. The value of dependent variable WQI is forecasted on certain values of independent variables like pH, DO, BOD, temperature, conductivity, faecal coliform and total coliform. Multiple linear regression works on numeric information for predicting the numeric output.

Ordinary Linear Squares (OLS) regression makes use of the changes in the independent variable for predicting the dependent variable. Since precision cannot be achieved using a single independent variable in predicting the behaviour of dependent variable, we have applied Multiple Regression in our model by considering various independent variables like temperature, pH, BOD, conductivity etc. to accurately forecast the trends in dependent variable i.e. WQI.

B. *Random Forest Classifier*

Random forest is a commonly-used machine learning calculation that combines the yield of different decision trees to reach a single result. Its ease of use and adaptability has made it a popular algorithm, as it handles both classification and regression problems. Random forest is faster than decision trees as it works with subsets of data. Fig 12 shows the basic working of a Random Forest algorithm.

The out-of-bag (OOB) score in a random forest is a way to evaluate the model's accuracy and how well it can generalize to new data In a Random Forest model, each tree within the ensemble calculates the Out-of-Bag (OOB) error utilizing the information samples it did not select for training amid the bootstrap sampling process.



[10] Fig 12

In our model we have implemented ensemble learning method random forest on the 9th and 10th column, i.e., WQI and Potability. Before the actual implementation, we have converted the object data type of Potability to numeric data type.

Grade conversion is as follows:

- 'A' is 1 (This represents excellent water quality)
- 'B' is 2
- 'C' is 3

- ‘D’ is 4
- ‘E’ is 5 (This represents unsuitable quality of water that shouldn't be consumed)

The data was divided into two groups as priorly discussed; training was done on 80% of the data and testing was done on the remaining 20% of the data.

V. Results and Discussion

The model is made using the Jupyter notebook version (7.0.8). Python scripts run very easily on the Jupyter notebook. It is very convenient to write on it. It is an open-source model implementation and execution tool for AI and ML. The proposed models' precision is compared to that of various existing models. There are many libraries and modules that come handy. Some python environments can be shared, modified, saved, reused and, shared and can be created. For our model we have made use of libraries such as NumPy, Pandas, Matplotlib and Seaborn.

	Accuracy	Error
Multiple Linear Regression	97.2896574544092%	596.5803046833935(MSE)
Random Forest	100.0%	0.9984276729559748(OOB)

Multiple linear regression was used to predict the values of WQI (Water Quality Index).

It was achieved that the r2 score came out to be 97.28965274544092% over the testing data and MSE (Mean Squared Error) to be 596.5803046833935 over the testing data. While checking the results of accuracy after the random forest algorithm to classify the WQI for potability, we got a meticulous precision of 100.0% and an OOB (Out of Bag) score to be 0.9984276729559748. As priorly discussed about the usage of algorithms, we have utilised Multiple Linear Regression and Random Forest Classifier for our model and applied them on the dataset.

VII. Conclusion

To safeguard human health, ensuring the safety and purity of drinking water is of utmost importance.

Accurate prediction of water potability plays a crucial role in achieving this objective. Access to clean drinking water is a fundamental right for every individual, as it is vital for maintaining overall well-being and preventing waterborne diseases. However, the escalating global population and increasing pollution levels have raised significant concerns about the quality of water sources. Using the power of ML techniques can incredibly contribute to predicting the potability of water thus guaranteeing safe water for everybody.

This research depicts the potential of all implemented algorithms as valuable tools for monitoring and managing water quality,which has profound implications for both the water sector and public health However,the limitations of this study should be acknowledged. The dataset utilised in the research is relatively small, consisting of only 795 observations. Consequently, it might be challenging to generalise the findings to larger populations. Additionally, the research focused on a limited set of water quality parameters, and it is advisable for future investigation to consider other relevant factors that could influence the potability of water.

References

1. https://discovery.researcher.life/article/efficient-prediction-of-water-quality-index-wqi-using-machine-learning-algorithms/9831d44023cb354c90e9a40652026966

2. Kouadri, S., Elbeltagi, A., Islam, A.R.M.T. et al. Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). Appl Water Sci 11, 190 (2021). https://link.springer.com/article/10.1007/s13201-021-01528-9

3. Jitha P Nair and M S Vijaya 2022 J. Phys.: Conf. Ser. 2325 012011. https://www.researchgate.net/publication/363199529_River_Water_Quality_Prediction_and_index_classification_using_Machine_Learning#pf18

4. https://www.sciencedirect.com/science/article/abs/pii/S004313542300773X

5. https://www.sciencedirect.com/science/article/pii/S0301479722014967

6. https://www.sciencedirect.com/science/article/abs/pii/S004313542300773X

7. https://www.javatpoint.com/train-and-test-datasets-in-machine-learning

8. https://www.researchgate.net/publication/363199529_River_Water_Quality_Prediction_and_index_classification_using_Machine_Learning#pf18

9. https://app.myeducator.com/reader/web/1421a/6/ye1ay/

10. https://www.javatpoint.com/machine-learning-random-forest-algorithm