Water Quality Prediction Using Machine Learning

Sarthak Kapaliya, Kaxit Pandya, Dhruvil Patel

Abstract:

India has 4% of the world's water resources, making it a water-rich nation. But most of India's rivers, lakes, and surface waters are polluted by industry, untreated sewage, and solid waste. The purpose of the proposed effort is to assess the Water Quality Index (WQI) for a few of India's well-known rivers the Ganga, to determine the water quality and whether or not it is potable. The available dataset contains the latest water quality parameters of Rivers in India collected in recent years. Using several water quality parameters like pH, Total coliforms, biochemical oxygen demand, electrical conductivity, nitrate, fecal coliforms, fecal streptococci, and temperature to calculate wqi for the river water. The dataset used here is real-time data taken from Central Pollution Control Board (CPCB). This work shows a comparative analysis of different machine learning approaches like Random Forest, XgBoost, Logistic Regression Decision Tree (DT), and K-NN for classification, Linear Regressor, Elastic Net, and Random Forest for Regression. Synthetic Minority Oversampling Technique (SMOTE) is used to balance the given dataset since it is unbalanced. The results of the experiments show that the maximum accuracy of 0.98, provided by Logistic Regression and Elastic Net Regressor is the most efficient algorithm for Regression with an R2 value of 0.99 and an RMSE value is 0.032. This research is expected to provide a baseline for future studies on the water quality of Indian rivers and provide information to decision-makers on how to establish appropriate sampling and analysis techniques for managing pollution's effects on river surface water quality.

Introduction:

For all of the earth's living things, water is kept in second place behind air in terms of essential natural resources. Usable water has also been seen as a priceless gift from nature for all people, and for human applications, it should be clear, sanitary, and fragrance-free. Water is the most crucial natural resource for the entire human race, not just a state or a nation. For a country to succeed, this resource must be used carefully. Water, which flows in rivers and streams, can be considered a country's fundamental resource. This demonstrates the importance of rivers, and further justification is not required to underline this. This demonstrates the importance of rivers, and further justification is not required to underline this. Since water has no political boundaries, river basins have been recognized as a domain for planning and management all across the world. One of India's most distinctive features is its rivers, which are revered by its citizens. The growth of India's rural areas has been greatly aided by the 329 million hectares of land covered by its rivers. In terms of the country's development on the cultural, economic, geographical, and religious fronts, its numerous rivers play a crucial role. They offer tourists a wonderful look into India's traditional, cultural, and historical aspects. Among the many different kinds of inland freshwater basins, the riverine system is a special kind of ecology.

The country's water issue has started to have an impact on people's lives and the environment they live in. Water covers the planet's surface to a depth of around 75%. The oceans, which hold 97% of the water on earth, are unsuited for human use due to their high salt content. Only 1% of the remaining 2% is available as freshwater that is suitable for human consumption in rivers, lakes, streams, reservoirs, and groundwater. Polar ice caps hold the remaining 2% in place. **(Singh G, Kamal R. K., et al)** Natural resource extraction is crucial for industrialization and economic growth. Additionally, it brings in money and creates job opportunities for the neighborhood. Natural resources, notably water resources, have been degraded and exhausted as a result of mining in numerous locations.

Usable water has also been seen as a priceless gift from nature for all people, and for human applications, it should be clear, sanitary, and fragrance-free. Water is the most crucial resource and is necessary for all forms of life, but it is constantly in danger of being contaminated by life. One of the most effective communication tools with a wide range is water. As a result of rapid industrialization, the quality of the water is rapidly declining. One of the main contributors to the spread of terrible diseases is recognized to be poor water quality. Surface and groundwater resources are both heavily utilized natural resources, and as a result, there are severe pollution and scarcity issues with them at the moment.**(Ahmad,et al)**

Therefore, providing the preservation and enhancement of their quality and quantity substantial consideration is crucial. Thus, for sustainable development and the protection of human health, it is necessary to create efficient procedures for the evaluation of groundwater and surface water resources. Contrary to surface water, groundwater is typically not metered for use, which has resulted in extreme overuse. Surface water, on the other hand, is more vulnerable to pollution from a variety of sources and typically has a metered supply. However, these types of water supplies are frequently contaminated for a variety of reasons, including home and industrial pollution, agricultural runoff, and others. Surface water has historically been the most accessible source for broad usage and is hence more vulnerable to domestic and other types of pollution. The ecosystems that thrive there are seriously threatened by their ongoing decline. A careful and watchful approach is therefore required to the monitoring and evaluation of surface water, as water-borne diseases rank among the top 10 global killers. In addition to that, the rising quantities of nitrate, iron, chloride, and TDS in groundwater pose a serious problem for a long-term drinking water program. These problems should all be adequately addressed. As a result of excessive groundwater extraction, the concentration of dissolved components and ionic concentrations are continually increasing.

Water-borne infections, which account for 80% of sickness in poor countries, have reportedly sickened 2.5 billion people and killed 5 million. Currently, expensive and time-consuming lab and statistical studies that require sample collection, transportation to laboratories, a substantial amount of time, and computation are used to determine the quality of water. If the water is polluted with disease-causing trash, speed is important because it is a highly infectious medium. Given the serious consequences of water contamination, a quicker and more affordable solution is required. As a result, the major goal of this work is to suggest and assess a novel approach based on supervised machine learning for the real-time prediction of water quality.

The WQI study aims to:

(i) provide an overview of the basin's water quality;

(ii) identify the spatial distribution so that the trend of the water quality can be assessed for future development plans;

(iii) map changes in surface and groundwater quality in the study area using GIS and Geo-statistical techniques; and

(iv) Find potential equivalences between different regression and classification models to determine the best modeling approach for the independent variable.

By combining complex data and providing a score that finally defines the water quality state, WQI provides a better way to comprehend problems with water quality. The major goal of this study is to gather the necessary data or trends regarding water quality to develop certain water pollution control initiatives. To improve decision-making for future facilities like water treatment plants, this model can aid in prescriptive analysis using expected values

Literature Review:

Debnath Palit et al. 2019 focuses on the analysis of several factors that affect water quality. This study used a variety of physicochemical characteristics to determine the water quality index, including pH, total hardness, total conductivity, alkalinity, dissolved oxygen, biological oxygen demand, and chloride. Calculations are made to determine the parameter's lowest and maximum values as well as its mean, standard deviation, and connection to the water quality index of a few chosen pit lakes. The ICMR and BIS requirements for the quality of drinking water are compared to the mean values of the researched parameters. In all five pit lakes, the WQI ratings indicate poor to extremely poor water quality samples. For the majority of aquatic animals and plants, pH is an important characteristic because it impacts their metabolic activities.

This work by **Aladejana J. A., et** al.,2013 evaluated the Abeokuta groundwater qualityfort drinking and irrigation purposes. A multiparameter portable meter was used to measure the in-situ parameters (pH, EC, temperature, and TDS). A Nutrient Agar medium was used for the bacterial analyses. Ion levels in the groundwater were within acceptable ranges according to WHO and NAFDAC regulations. According to the estimated water quality index, 22% of the water samples came into the category of good water quality, while 72.2% and 5.5% of the samples fell into the categories of medium and bad water quality, respectively. This study has demonstrated the value of hydro chemical and bacteriological analyses in determining the quality of groundwater. Although not potable, the groundwater in the study area was of good irrigation quality.

Vinod Kumar Chaudhary et. Al,2022 The Yamuna River's water quality improved during the shutdown since all businesses and enterprises were shut down. Based on information found on the CPCB, India website, this report shows how the Yamuna river's water quality improved during the first phase of shutdown in terms of DO, BOD, COD, pH, conductivity, and suspended solids. The pH readings that were measured both before and after the lockdown declined. Post-lockdown, it was observed that both conductivity in river water and TSS in drain water decreased. In barely three weeks of the lockdown period, Yamuna has decreased by an average of 62% BOD and 60% COD load of the river.

Umair Ahmed, et al., 2019 Poor water quality necessitates a faster, less expensive alternative approach, according to the report. The water quality index (WQI), a distinct index used to reflect the overall quality of water, and the water quality class (WQC), a new class formed using the WQI, are estimated in this work using supervised machine learning approaches. The four input parameters for the proposed approach are temperature, turbidity, pH, and total dissolved solids. The best techniques for forecasting the WQI are gradient boosting with a learning rate of 0.1 and polynomial regression with a degree of 2. By obtaining respectable accuracy with a constrained set of parameters, the suggested technique verifies the viability of its application in real-time water quality detecting systems.

In M Ramchandra Mohan, 2022 's research work, Thorapalli lake water samples were collected and examined for their physicochemical parameters from January 2018 to December 2018. This study aimed to characterize the physicochemical parameters pH, Biochemical oxygen demand (BOD), Electrical Conductivity (EC), phosphates, Total Hardness (TH), Dissolved oxygen (DO), Turbidity (TY), Total Alkalinity (TA), Total Dissolved Solids (TDS), Nitrates, Temperature, and Chemical Oxygen Demand (COD). Each measure was compared to its World Health Organization-prescribed standard acceptable limit (WHO). The investigation finds that few of the metrics exceed the WHO-recommended maximum value.

Iqbal Ahmad and Sadhana Chaurasia, 2019 investigated the physicochemical properties of the Ganga River in Kanpur. Five sample points were chosen for the investigation of river water. In this work, many physicochemical parameters, including pH, turbidity, and electrical conductivity (EC), were measured. The collected findings were compared to the specified benchmark. Water Quality Index (WQI) was also computed to determine the river's overall water quality in the research region. Observed data indicate that the water quality of all test stations was over 100 on a scale from 0 to 100, indicating that all sampling stations were unfit for consumption.

In **Prerna Sengar, et al., 2022 '**s work, a WQI-based marking system for the Chambal River has been established. The concentrations of several parameters at 10 distinct Chambal River locations were imported from the database of the Central pollution control board to do this. In addition, a prediction model was created utilizing an Artificial Neural Network and an artificial dataset. Using random sampling on the original dataset, a fictitious dataset was created. In this approach, Levenberg Marquardt (LM), Bayesian Regularization (BR), and Scaled Conjugate Gradient (SCG) methods with various hyper-parameter values were trained and evaluated. The Bayesian Regularization technique yields the best results (RMSE = 0.00, R2 = 0.99), followed by the Levenberg Marquardt and Scaled Conjugate Gradient algorithms (RMSE = 1.89, R2 = 0.99 and 1.94, respectively).

Proposed Work:

A fundamental human right and a component of any strategy to safeguard one's health, access to clean water to drink is necessary for good health. As a matter of health and development, this is significant on a global, regional, and local level. Since the decreases in negative health effects and healthcare costs outweigh the cost of implementing the interventions, it has been shown in some places that investments in water supply and sanitation can have a net economic benefit. As was already said, a method for predicting pure, high-quality water is urgently needed.

The data on the water quality of Indian rivers were collected from Central Pollution Control Board (CPCB) database. CPCB is a statutory organization that promotes the purity of streams and wells in various States by preventing, controlling, and abating water pollution. It collects, collates, and disseminates technical and statistical data relating to water pollution and proposes guidelines devised for their effective prevention, control, or abatement (CPCB). This data is collected from several stations that are spread across the country at the river banks for several years. The dataset used in the Model was from data collection in the year 2021.

Now as we have seen different datasets on water quality. Let us go through different pre-processing tasks we performed on the data. Data preparation is a process of preparing raw data that can be useful for data preprocessing and analysis. Also, Data Preparation is the main task for Machine learning. Firstly we used Central Pollution Control Board, Data. The Central Pollution Control Board has stored the water quality data in different ways. The site contains year-wise data about different rivers. The Website also contains data for Groundwater, canals, Lakes, and Drains. For this Research Purpose, we have taken the River data of 2020.

The data was in a structured format but it was not accessible for data analysis. The CPCB has stored its data in form of a pdf. The pdf contains different tables related to different rivers of India. The Initial task we did was to convert this pdf data to an accessible manner. For that, we thought to convert it to excel format.

We converted the whole Water Quality Data of rivers under the national water quality Monitoring program to an excel file. We did this using a pdf to excel converter provided by Adobe open-source software. The software converts the pdf which contains table-like data to excel spreadsheet data. Our next step was to separate the data of different rivers into different data files which can be used for data pre-processing. So we took the three most famous river data and stored them in different excel sheets. After getting accessible data we still cannot perform data analysis as the table contains large messages from CPCB and unnecessary data features.

A. Data Description:

The Excel file contains different columns. The table was not good for processing as the first row contains different columns and there were maximum and minimum values associated with most of the columns, so we cannot train the machine learning algorithms on these data types. So we edited and updated the excel data so each column has only one value and try to make data in a structured format for machine learning to perform on it. We renamed the columns and made significant changes that can be shown in the below figures.

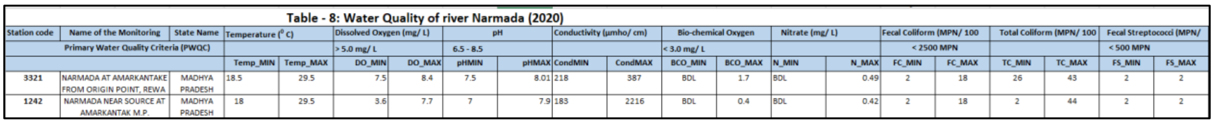
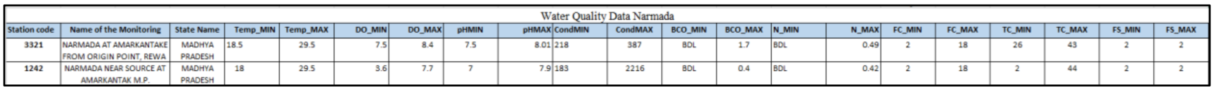


Table - 8: Water Quality of river Narmada (2020)

Station code	Name of the Monitoring	State Name	Temperature (^{0} C)		Dissolved Oxygen (mg/ L)		pH		Conductivity (µmho/ cm)		Bio-chemical Oxygen		Nitrate (mg/ L)		Fecal Coliform (MPN/ 100		Total Coliform (MPN/ 100		Fecal Streptococci (MPN/	
	Primary Water Quality Criteria (PWQC)				> 5.0 mg/ L		6.5 - 8.5				< 3.0 mg/ L				< 2500 MPN				< 500 MPN	
			Temp_MIN	Temp_MAX	DO_MIN	DO_MAX	pHMIN	pHMAX	CondMIN	CondMAX	BCO_MIN	BCO_MAX	N_MIN	N_MAX	FC_MIN	FC_MAX	TC_MIN	TC_MAX	FS_MIN	FS_MAX
3321	NARMADA AT AMARKANTAKE FROM ORIGIN POINT, REWA	MADHYA PRADESH	18.5	29.5	7.5	8.4	7.5	8.01	218	387	BDL	1.7	BDL	0.49	2	18	26	43	2	2
1242	NARMADA NEAR SOURCE AT AMARKANTAK M.P.	MADHYA PRADESH	18	29.5	3.6	7.7	7	7.9	183	2216	BDL	0.4	BDL	0.42	2	18	2	44	2	2

Fig1. Original Excel data



Water Quality Data Narmada

Station code	Name of the Monitoring	State Name	Temp_MIN	Temp_MAX	DO_MIN	DO_MAX	pHMIN	pHMAX	CondMIN	CondMAX	BCO_MIN	BCO_MAX	N_MIN	N_MAX	FC_MIN	FC_MAX	TC_MIN	TC_MAX	FS_MIN	FS_MAX
3321	NARMADA AT AMARKANTAKE FROM ORIGIN POINT, REWA	MADHYA PRADESH	18.5	29.5	7.5	8.4	7.5	8.01	218	387	BDL	1.7	BDL	0.49	2	18	26	43	2	2
1242	NARMADA NEAR SOURCE AT AMARKANTAK M.P.	MADHYA PRADESH	18	29.5	3.6	7.7	7	7.9	183	2216	BDL	0.4	BDL	0.42	2	18	2	44	2	2

Fig2. Processed Excel data

So Fig2 displays Data Prepared to apply data pre-processing techniques.

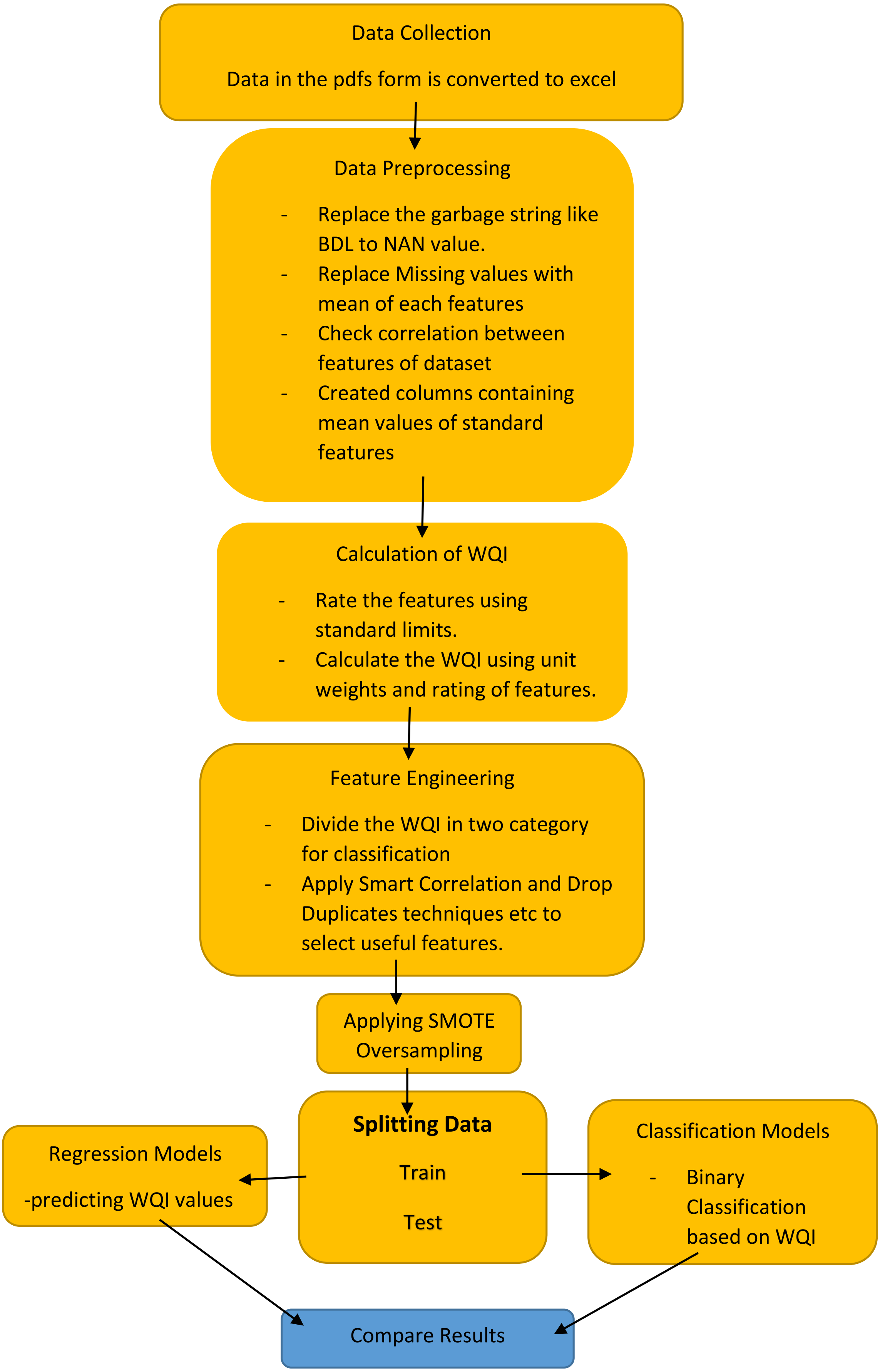

Data Collection
Data in the pdfs form is converted to excel
Data Preprocessing
- Replace the garbage string like BDL to NAN value.
- Replace Missing values with mean of each features
- Check correlation between features of dataset
- Created columns containing mean values of standard features
Calculation of WQI
- Rate the features using standard limits.
- Calculate the WQI using unit weights and rating of features.
Feature Engineering
- Divide the WQI in two category for classification
- Apply Smart Correlation and Drop Duplicates techniques etc to select useful features.
Applying SMOTE Oversampling
Splitting Data
Train
Test
Regression Models
-predicting WQI values
Classification Models
- Binary Classification based on WQI
Compare Results

Fig4. Flowchart Diagram

```
1. Data collection – CPCB data repositories
2. Data Pre − processing −
        • identifying missing values
        • replacing string values with NAN
        • filling missing values with mean of each column
        • averaging 'min' and 'max' columns into new single parameter column
        • fitting the data values into weighted rating cycle
3. Calculation of WQI using rating scale values and weights
   with respective parameters
4. feature selection pipeline
5. applying smote
6. splitting dataset into train and test
7. Applying both Classification and Regression models on preprocessed data
     a. Classification
          i. classifying the dataset into potable and non − potable water
     b. Regression
          i. predicting WQI value using regression model
8. Model evaluation
9. Compare Results
```

Fig.3 Proposed Work Algorithm

B. Data Preprocessing Techniques

Data pre-processing is the process of converting raw data into a usable, intelligible format. Real-world or raw data often contain irregular formatting, and human mistakes, and is incomplete. Data preparation resolves such difficulties and makes datasets more comprehensive and efficient for data analysis. It is a critical step that can impact the performance of data mining and machine learning initiatives.

1. Dealing with Missing values and unwanted values in features:

We imported python pre-processing libraries like NumPy and Pandas. We used the Pandas IsNull () function to detect missing values. We got no missing values in the dataset. After going through the dataset, we observed that certain features contain string values in them. For example, the N_min (Nitrates minimum) columns contain too many strings named BDL, and also FS_MIN (Fecal Streptococci minimum) contains strings like "-". So basically all these columns should contain numerical values but string values were unnecessary things. So we replace this string "BDL" and "- "with Nan value. So now we got total null values in the dataset. We fill those missing values with a mean value of that data column. So now all feature does not have null values.

2. Simplifying dataset:

Now some features have max and min values so we cannot give models this many features and it will not be able to extract features efficiently. So we created new features for the columns like Temp_min, Temp_max to Temperature (0 C). The new feature is the mean of the max and minimum values of the min and max columns. So our final dataset contains 12 Features.

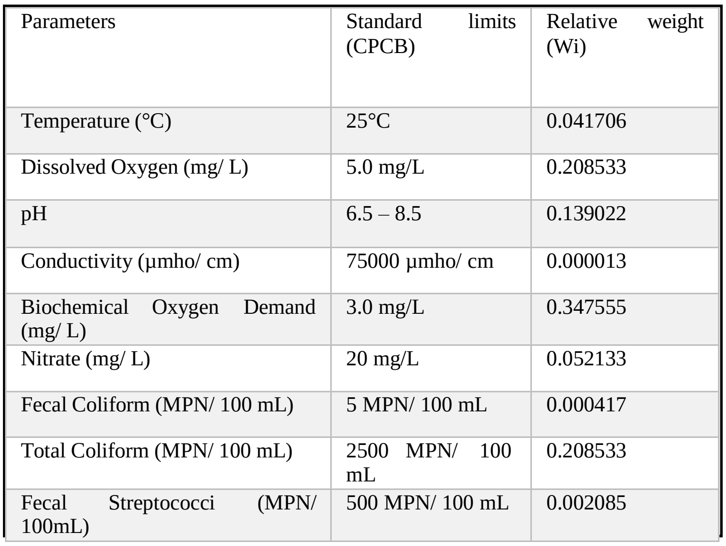
Methodology:

The water quality index (wqi) is a numerical measure that indicates the overall quality of water based on various parameters, such as dissolved oxygen, pH, turbidity, temperature, and others. It helps to compare and evaluate different water sources and identify the main problems affecting water quality. A higher value means better water quality, while a lower value means worse water quality.

Our Criteria for Calculating Water Quality Index (WQI) are illustrated below –

a) WQI is a single defining criterion as Satisfactory or Unsatisfactory.

b) Taken into account 4 (four) water quality indicators for the Water Quality Index (WQI), for which Water Quality Criteria are given, namely Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), and Faecal Coliform & Total Coliform counts.



Parameters	Standard limits (CPCB)	Relative weight (Wi)
Temperature (°C)	25°C	0.041706
Dissolved Oxygen (mg/ L)	5.0 mg/L	0.208533
pH	6.5 – 8.5	0.139022
Conductivity (µmho/ cm)	75000 µmho/ cm	0.000013
Biochemical Oxygen Demand (mg/ L)	3.0 mg/L	0.347555
Nitrate (mg/ L)	20 mg/L	0.052133
Fecal Coliform (MPN/ 100 mL)	5 MPN/ 100 mL	0.000417
Total Coliform (MPN/ 100 mL)	2500 MPN/ 100 mL	0.208533
Fecal Streptococci (MPN/ 100mL)	500 MPN/ 100 mL	0.002085

Fig5. Parameters Table

c) The water quality will be classified as good or unsatisfactory based on the observed ambient concentrations and the associated standards.

d) Each parameter's requirements are listed in the table.

e) Even one single parameter from these four parameters exceeding the criteria values will consider Unsatisfactory. Now our data does not have any Water Quality Index so we have to make one.

The best approach we got after going through different research work and thesis was to make our water quality index based on feature parameters. We select the parameters for the measurement of water quality. We selected all the features namely- Temperature(0 C)', 'Dissolved Oxygen (mg/ L)', 'pH',' Conductivity (µmhos/cm)', 'BCO (mg/ L)', 'Nitrates(mg/l)', 'Total Coliform(mg/l)', 'Fecal Coliform (MPN/ 100 mL)', 'Fecal Streptococci(MPN/100 ml). The next step is to develop a rating scale to obtain the rating **Vr**. The unit weight of each parameter (**Wi**) was calculated using the weightage criteria of each feature. Sub index value was determined with the formula $(\boldsymbol{Wi\ X\ Vr})$.

The final Value of the Water Quality Index was calculated using the Summation of all the sub-index values of each feature. Weights in units for each parameter's weightage (**Wi**) and unit weight (**Si**) are inversely related. **Wi** is the unit weight of the parameter, and n is the total number of water quality parameters.

$$\boldsymbol{Wi = K/Si}$$

Where

$$\boldsymbol{K = 1\,/\,(\,1/Si)}$$

And **K** is the proportionality constant. The table displays the computed unit weight for each parameter. The sub-index value is calculated by multiplying the rating received by the sub-unit index's weight.

Calculating the overall water quality index by adding the sub-indices together (WQI).

$$\boldsymbol{WQI} = \sum(\boldsymbol{Wi\,X\,Vr}).$$

So, After performing all the above-mentioned pre-processing tasks our data is ready for model training.

Model Training

There is no accepted definition for the phrase "machine learning" (ML). However, machine learning is frequently referred to be a subset of artificial intelligence that emphasizes using data and algorithms to replicate how people learn, simulate predicted patterns, and steadily improve the system's accuracy (Vieira et, al.,2020). Machine learning (ML) may be a key viewpoint for finding a practical and workable solution to the water quality problem of Indian rivers. One of the methods in ML is Supervised Learning in which machines train on “labeled data” i.e., input data and corresponding output data are given. Classification and Regression are examples of Supervised learning.

The data mining process called classification divides data points into several groups. This method's primary objective is to correctly anticipate the target class for each dataset's data case. The Classification models used in the proposed work are as follows –

1. Decision Tree classifier
2. K – Nearest Neighbors (K-NN) classifier
3. Random Forest Classifier
4. Xgboost classifier
5. Logistic Regression

Another data mining technique is Regression which is used to predict numeric values in a given dataset. This technique is only used to forecast continuous – values. The regression models used are as follows –

1. Random Forest Regressor
2. Linear regressor
3. Elastic Net regressor

We have used Synthetic Minority Oversampling Technique (SMOTE) Oversampling technique. A machine learning method called SMOTE addresses issues that arise from employing an unbalanced data set. It is vital to learn the skills required to work with this type of data because imbalanced data sets frequently appear in practice. It generates synthetic data points that resemble the original ones rather than duplicating the minority class. With SMOTE, your model will begin identifying more instances of the minority class, increasing recall but decreasing precision.

Classification Models

Decision Tree classifier – Using a tree-structured classifier, the decision tree classifier uses internal nodes to represent characteristics of a dataset, branches to represent decision rules, and each leaf node to represent the classification outcome. This method takes some assumptions on the data. The identification of the attribute for the root node is done using measures like Information Gain and Gini Index. After training, the decision tree makes choices based on the values of all essential input parameters **(J. R. Quinlan, 1990)**.

KNN- A non-parametric classifier called the K-NN classifier employs closeness to categorize or forecast how a single data point will be categorized. Regression and classification may both be handled using this strategy. When a fresh dataset is presented, it merely sorts the data into a category that closely resembles the training dataset. Since all processing takes place during testing and closest neighbors are repeatedly calculated for the whole training set, K nearest neighbor is not advised for large datasets. **(Kevin Beyer, et al., 1997).**

Random Forest classifier – This classifier aggregates the results from different decision trees applied to different dataset subsets to improve the projected accuracy of the dataset. More trees in the forest prevent the problem of overfitting and lead to greater accuracy. Scalable and distributed, the Extreme Gradient Boosting (XGBoost) gradient-boosted decision tree (GBDT) machine learning system. The greatest machine learning tool for regression, classification, and ranking tasks, it enables parallel tree boosting. Decision trees serve as the foundational model of random forests, which combine their benefits with the effectiveness of mixing several models. **(Liaw Andy & Wiener Matthew, 2001).**

Logistic Regression:

One of the Machine Learning algorithms that are most commonly used in the Supervised Learning category is logistic regression. Using a predetermined set of independent factors is used to forecast the categorical dependent variable. A categorical dependent variable's output can be predicted using logistic regression. Consequently, the result must be a discrete or categorical value. Instead of the precise numbers between 0 and 1, it delivers the probabilistic values that fall between 0 and 1. Either True or False, 0 or 1, or Yes or No, are possible outcomes.

Regression Models:

Random Forest Regressor –

It is an ensemble learning algorithm. In ensemble learning, you combine several methods or the same approach used several times to create a model that is stronger than the original. Because it considers numerous predictions, prediction based on trees is more precise. The utilized average value is the reason behind this. These techniques are more reliable because alterations to the dataset only affect individual trees, not the entire forest. **(Nida Nasir et. al.)**

Elastic Net Regressor - Regularization and variable selection are both used simultaneously by the regression method known as the elastic net. L1 and L2 penalties, or lasso and ridge regression, are combined in this model **(Umair Ahmed, et al., 2019)**. Lattice regression has the flaw of being unable to choose the number of predictors. The elastic net, which becomes the ridge regression when used alone, incorporates the lasso regression penalty. In the regularisation with an elastic net method, the ridge regression coefficient is first calculated. The ridge regression coefficient is then reduced using a lasso method.

Model Evaluation:

As mentioned above, both the types of supervised learning algorithms i.e., classification and Regression were implied on the dataset. Both types of algorithms' outputs were assessed using various metrics. Measures used for regression are as follows:

Here $\boldsymbol{Y_i}$ denotes the predicted value while $\boldsymbol{y_1}$ denotes the actual value

1. Mean Absolute Error - MAE calculates the accuracy of the regression. The mistakes' absolute values are tallied, then divided by the total number of values. Because MAE measures average error size without ambiguity (unlike RMSE), it is the most logical way to assess average error magnitude **(Willmott, et al., 2005**). Each inaccurate value is given the same weight.

$$MAE = \frac{\sum_{i=1}^{n} | y_{i-} \overset{\wedge}{y_i} |}{n}$$

2. Mean Square Error - The sum of squared errors divided by the total number of predicted values is known as the mean square error (MSE). This gives larger errors more significance. This is especially helpful in situations when a heavier weight for larger faults is required.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overset{\wedge}{y_i})^2$$

3. Root Mean Square Error - Simply taking the square root of mean squared error (MSE), or RMSE, scales the values of MSE near the ranges of measurements.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overset{\wedge}{y_i})^2}$$

4. R Squared Error - The coefficient of determination, often known as R squared error (RSE), and frequently abbreviated as R^2, assesses how well the model fits the data. It specifically explains the percentage of the dependent variable's volatility that the independent variable may account for.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \overset{\wedge}{y_i})^2}{\sum_{i=1}^{n}(y_i - y)^2}$$

For Classification, the measures used are as follows:

A true positive is an outcome where prediction and actual positive class is the same while in true negative prediction equals negative class. False negative means the model has made a wrong prediction about the negative class while false Positive means an incorrect prediction of the positive class.

1. Accuracy - The model's accuracy is measured by how many of the observed values it correctly predicted. The precision-recall curve provides an ostensibly comprehensive perspective of a system's performance, which is often summarised by a single indication using the average accuracy across several standard recall levels **(Cyril Goutte and Eric Gaussier, 2005).**

$$Accuracy = \frac{True\ Negative + True\ Positive}{TruePositive + False\ Positive + True\ Negative + False\ Negative}$$

2. Precision - Precision measures the proportion of positive class occurrences properly categorized out of all positive class instances classified.

$$Precision = \frac{True\ Positive}{TruePositive + FalsePositive}$$

3. Recall - Recall is the proportion of affirmative class instances that were classified with accuracy.

$$Recall = \frac{True\ Positive}{TruePositive + FalseNegative}$$

4. F1 Score - Since recall and precision alone cannot account for all facets of accuracy, we used their harmonic mean to depict the F1 score. F-score is a compound metric that rewards algorithms with more sensitivity and penalizes those with greater specificity **(Sokolova, et al., 2006).**

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Result:

Sensors for measuring water quality parameters are expensive, this study aimed to forecast water quality using a limited set of characteristics and cheap sensors. In the table below, regression algorithm results are displayed. The regression algorithms we used revealed Random Forest Regressor having an MAE of 0.131, MSE of 0.023, RMSE of 0.152, and R^2 of 0.904, to be the most efficient algorithm.

Model Name	R2	RMSE	MAE	MSE
Linear Regressor	0.710	0.252	0.149	0.067
ElasticNet	0.717	0.262	0.152	0.068
Random Forest Regressor	0.904	0.152	0.131	0.023

Fig6. Evaluation metrics Table for Regression Model

After applying the regression algorithms, we divided the datasets into two classes using the water quality index(wqi). These two classes indicate whether the water quality is 'satisfactory' or 'non-satisfactory'. Using classification algorithms on the dataset, accuracy, precision, recall, and the F- score were predicted. All Classification results were evaluated on 10-fold Cross Validation. Out of all the classification algorithms, Logistic regression outperformed all by having an accuracy of 0.98, precision of 1.00, recall of 0.96, and F-score of 0.98.

Model Name	Accuracy	Precision	Recall	F1 Score
Random Forest Classifier	0.90	0.90	0.95	0.832
XgBoost	0.82	0.808	0.875	0.804
Logistic Regression	0.98	1.00	0.96	0.98
Decision Tree	0.96	1.0	0.966	0.96
KNN	0.55	0.583	0.625	0.539

Fig7. Evaluation metrics Table for Classification Model

Conclusion and Future Scope:

Water is one of the most important resources for survival, and the WQI rates the water's quality. The cleanliness of the water is often checked by a costly and time-consuming lab study. This study investigated a different machine learning approach to forecast water quality by employing the fewest possible and most accessible water quality indicators. The WQI calculated ranges from 58.321 to 99.124, where 65% of the samples were found in excellent water quality (90 - 100).

In future studies, we suggest incorporating the results of this study into a giant Internet of Thing-based online surveillance using just sensors for the necessary parameters. Based on the IoT system's real-time data feed, the evaluated algorithms could estimate the water quality in real-time. To harness the potential of the suggested system, it will be developed into a commercial product that can be used as a decision support system in the industrial sector, at water quality monitoring stations, and in residential settings. The commercial value of the product is in its ability to measure water quality and manage water flow in real-time. Issues with water quality in the agriculture industry and other sectors can be effectively resolved with this technique. a reduction in the frequency of hazardous diseases like typhoid and diarrhea (**Uferah Shafi, et al.,2018).** To put it another way, the use of a prescriptive analysis based on anticipated values would lead to the creation of future resources to help decision- and policy-makers.

Acknowledgment:

This Research work has been done under Assistant Professor Dr. Rajeev Gupta, Pandit Deendayal Energy University.

References:

1. Singh G, Kamal R. K. Application of Water Quality Index for Assessment of Surface Water Quality Status in Goa. Curr World Environ 2014;9 (3) DOI:http://dx.doi.org/10.12944/CWE.9.3.54
2. Ahmad, Iqbal & Chaurasia, Sadhana. (2019). Water Quality Index Of Ganga River at Kanpur (U.P.). 8. 66-77. 10.26643/tjg.v8i11.
3. Palit, Debnath & Mondal, Saikat & Chattopadhyay, Pinaki. (2019). Analysing Water Quality Index of Selected Pit-Lakes of Raniganj Coal Field Area, India. 1167-1175.
4. Aladejana, Jamiu & Talabi, Abel. (2013). Assessment of Groundwater Quality in Abeokuta Southwestern, Nigeria. 21-31.
5. Chaudhary, V. Kumar, Shrivastav, A. Lav, Rinku, & Patel, N. (2021). *Water quality index of Yamuna River in Delhi region due to community lockdown amid COVID-19 pandemic*. http://doi.org/10.53550/EEC.2022.v28i01.072
6. Ahmed, U.; Mumtaz, R.; Anwar, H.; Shah, A.A.; Irfan, R.; García-Nieto, J. Efficient Water Quality Prediction Using Supervised Machine Learning. *Water* **2019**, *11*, 2210. https://doi.org/10.3390/w11112210
7. Mohan, M Ramachandra. (2022). Assessment and Evaluation of Water Quality of Thorapalli lake. Ecology, Environment,t, and Conservation. 28. 28. 10.53550/EEC.2022.v28i01s.041.
8. Ahmad Iqbal & Chaurasia Sadhana. (2019). Water Quality Index Of Ganga River at Kanpur (U.P.).8.6677.10.26643/tjg.v8i11.https://www.researchgate.net/publication/338345424_Water_Quality_Index_Of_Ganga_River_at_Kanpur_UP
9. Sengar, Prerna & Rajput, Jay & Saxena, A. (2022). Development of Marking System based Water Quality Index for Chambal River and their Prediction Model Using Artificial Neural Network. International Journal of Innovative Research in Medical Science. 7. 896-903.
10. CPCB, ADSORBS/3 1978–1979) Scheme for zoning and classification of Indian Rivers: estuaries and coastal waters. CPCB website: www.CPCB.nic.in.
11. Jinal Patel, Charmi Amipara, Tariq Ahamed Ahanger, Komal Ladhva, Rajeev Kumar Gupta, Hashem O. Alsaab, Yusuf S. Althobaiti, Rajnish Ratna, "A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI", *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 9283293, 15 pages, 2022. https://doi.org/10.1155/2022/9283293
12. Sandra Vieira, Walter Hugo Lopez Pinaya, Andrea Mechelli, Chapter 1 - Introduction to machine learning, Academic Press, 2020. https://doi.org/10.1016/B978-0-12-815739-8.00001-8.
13. Beyer, Kevin & Goldstein, Jonathan & Ramakrishnan, Raghu & Shaft, Uri. (1997). When Is "Nearest Neighbor" Meaningful? ICDT 1999. LNCS. 1540.

14. Liaw, Andy & Wiener, Matthew. (2001). Classification and Regression by RandomForest. Forest. 23.
15. J. R. Quinlan, "Decision trees and decision-making," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 20, no. 2, pp. 339-346, March-April 1990, doi: 10.1109/21.52545.
16. Ahmed, U.; Mumtaz, R.; Anwar, H.; Shah, A.A.; Irfan, R.; García-Nieto, J. Efficient Water Quality Prediction Using Supervised Machine Learning. *Water* **2019**, *11*, 2210. https://doi.org/10.3390/w11112210
17. Willmott, Cort J. and Kenji Matsuura. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance." *Climate Research* 30 (2005): 79-82.
18. Sokolova, M., Japkowicz, N., Szpakowicz, S. (2006). Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In: Sattar, A., Kang, Bh. (eds) AI 2006: Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science(), vol 4304. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11941439_114
19. Goutte, Cyril and Éric Gaussier. "A Probabilistic Interpretation of Precision, Recal,l, and F-Score, with Implication for Evaluation." *European Conference on Information Retrieval* (2005).
20. Shafi, R. Mumtaz, H. Anwar, A. M. Qama,r, and H. Khurshid, "Surface Water Pollution Detection using Internet of Things," 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), Islamabad, Pakistan, 2018, pp. 92-96, doi: 10.1109/HONET.2018.8551341.
21. Control Pollution Control Board https://cpcb.nic.in/nwmp-data-2020/
22. Nida Nasir, Afreen Kansal, Omar Alshaltone, Feras Barneih, Mustafa Sameer, Abdallah Shanableh, Ahmed Al-Shamma'a, Water quality classification using machine learning algorithms, Journal of Water Process Engineering, Volume 48, 2022, 102920, ISSN 2214-7144,https://doi.org/10.1016/j.jwpe.2022.102920.