Water Quality Prediction Using Machine Learning

Sarthak Kapaliya and Kaxit Pandya
Department of Computer Science
Pandit Deendayal Energy University
{kapaliyasarthak & Kaxitpandya}@gmail.com

Dhruvil Patel
Department of Computer Science
Pandit Deendayal Energy University
Dhruvilpatel120102@gmail.com

***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, including the Ganga, Yamuna, and Sabarmati, in order to determine the water quality and whether or not it is potable. The water quality index (WQI) is a useful and distinctive assessment that summarises the current state of water quality in a single term, making it easier to choose the best treatment options for the challenges at hand.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 usedhere is real-time data taken from Central Pollution Control Board (CPCB). Both Regression (Elastic Net Regressor) and Classification (Random Forest Classifier) are applied in the model for the wqi analysis. The outcome of the analysis shows that the quality of river water is very poor in the river and is not fit for drinking or bathing.**

Index Terms - Water Quality Index, Portability, Random Forest, Ensemble Methods, Machine Learning

I. 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. For the entire human race, not just a state or a nation, water is the most crucial natural resource. For a country to succeed, this resource mustbe used carefully. Water, which flows in rivers and streams, can therefore be referred to as 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 justificationis not required to underline this Since water has no political boundaries, river basins have been recognised as a domain for planning and management all across the world. Oneof 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 fresh water that is suitable for human consumption in rivers, lakes, streams, reservoirs, and groundwater. Polar ice caps hold the remaining 2% in place. Natural resource extraction is crucial forindustrialization 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 aresult of rapid industrialization, the quality of the water is rapidly declining. One of the main contributors to the spreadof 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 pollutionand scarcity issues with them at the moment.

Therefore, it is crucial to provide the preservation and enhancement of their quality and quantity substantial consideration. 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 ametered 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. A major issue for a long-term drinking water programme, aside from that, is Chloride, TDS, nitrate, and iron concentrations in groundwater are increasing.

All of these issues need to be thoroughly addressed. The concentration of dissolved components/ionic concentrations is constantly rising as a result of excessive groundwater extraction.

According to reports, 2.5 billion people have fallen ill and 5 million have died as a result of water-borne diseases, which account for 80% of illnesses in underdeveloped nations. The quality of water is currently estimated through expensive and time-consuming lab and statistical analyses, which call for sample collection, transportation to labs, and a significant amount of time and calculation. Water is a highly contagious medium, and time is of the essence if it is contaminated with disease-causing waste. Given the severe effects of water pollution, there must be a speedier and less expensive remedy. In light of this, the main objective of this study is to propose and evaluate a different method based on supervised machine learning for the prediction of water quality in real-time.

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 modelling approach for the independent variable.

The current study takes into account the following goals: I locate the best places in the current research area for various uses. (ii) Researching the basin's water quality patterns. (iii) To determine, for both surface and groundwater, the statistical relationships between the biophysical and chemical water quality parameters of the basin. (iv) To determine the basin's Water Quality Index (WQI).
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 into 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.

II. Literature Review

Debnath Palit et. al has written this paper which focuses on the study of various water quality influencing parameter. In this research water quality index was established by different Physico-chemical parameters such as pH, total hardness, total conductivity, alkalinity, dissolved oxygen, biological oxygen demand, and chloride. The minimum and maximum value, mean, standard deviation, and correlation between the parameter and water quality index of selected pit lakes are calculated. The mean values of studied parameters arecompared with ICMR and BIS standards for drinking water quality. The WQI scores show poor to very poor quality water samples in all five pit lakes. Since it affects the metabolic processes of aquatic organisms, pH is a crucial parameter for the majority of aquatic animals and plants. This work by Aladejana J. A. et al. evaluated the Abeokuta groundwater quality with regard to drinking and irrigation purposes. A multiparameter portable metre 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 hydrochemical 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 During the lockdown, the water quality of river Yamuna got improved as the entire commercial premises and industries were shut down. This paper presents the improvement of Yamuna river water qualityin terms of DO, BOD, COD, pH, conductivity, and suspended solids during the first phase of lockdown on the basis of data available on the website of CPCB, India. pH values which were examined in the pre-lockdown and it was decreased post-lockdown. Similarly, conductivity reduction in river water andTSS reduction in drain water were recorded post-lockdown.On an average of 62% BOD and 60% COD load of the river, Yamuna has been reduced in just three weeks of the lockdowntime period. Umair Ahmed et. al This paper discusses
Poor water quality requires an alternative solution that is quicker and less expensive. In this study, supervised machine learning techniques are investigated to estimate the water quality index (WQI), a unique index used to represent the general quality of water, and the water quality class (WQC), a different class established using the WQI. The suggested methodology uses temperature, turbidity, pH, and total dissolved solids as its four input parameters. The most effective methods for predicting the WQI are polynomial regression with a degree of 2, and gradient boosting with a learning rate of 0.1. The proposed methodology validates the possibility of its use in real-time water quality detection systems by achieving reasonable accuracy with a limited number of parameters.

III. Proposed Work

Access to clean water to drink is essential for health, a fundamental human right, and a part of any plan to protect one's health. This is significant as a national, regional, and local health

and development issue. It has been demonstrated in some locations that investments in water supply and sanitation can result in a net economic benefit since the reductions in adverse health impacts and healthcare expenditures surpass the price of carrying out the interventions. So as stated above there is a serious need for a clean and good-quality water prediction system.

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. 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 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 to most of the columns, so we cannot train the machine learning algorithms on these data types. So we

Fig1 Original pdf Data

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

Station code	Name of the Monitoring Location	State Name	Temperature (° C)		Dissolved Oxygen (mg/ L)		pH		Conductivity (µmho/ cm)		Bio-chemical Oxygen Demand (mg/ L)		Nitrate (mg/ L)		Fecal Coliform (MPN/ 100 mL)		Total Coliform (MPN/ 100 mL)		Fecal Streptococci (MPN/ 100mL)	
Primary Water Quality Criteria (PWQC) notified by MoEF & CC under E (P) Rules, 1986					> 5.0 mg/ L		6.5 - 8.5				< 3.0 mg/ L				< 2500 MPN / 100 mL				< 500 MPN / 100 mL	
			MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	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
1241	NARMADA AT MANDLA NEAR ROAD BDG. M.P.	MADHYA PRADESH	17.2	26	6.8	8.3	7.2	8.7	160	363	BDL	2.6	BDL	0.8	2	14	17	48	-	-
1240	NARMADA AT NARSINGHPUR M.P.	MADHYA PRADESH	17	23.5	7	8.6	7	8.3	170	357	BDL	1.9	BDL	1.16	2	11	12	49	-	-
2106	NARMADA AT NEMAWAR	MADHYA PRADESH	22	28	7	7.9	7.3	8.3	279	360	BDL	2	0.31	51	2	18	28	49	2	2
45	NARMADA AT MANDLESHWAR	MADHYA PRADESH	20	24	7.3	8.6	7.4	8.5	214	340	BDL	1.3	0.51	1.33	2	2	32	47	-	-
1239	NARMADA AT BADWANI, M.P.	MADHYA PRADESH	20	24.6	7.2	8	7.7	8.3	152	594	BDL	1.2	BDL	1.24	2	2	34	48	-	-
1431	NARMADA AT MAHESHWAR, M. P.	MADHYA PRADESH	19	24.3	7.3	8	7.4	8.1	258	338	BDL	1.3	0.61	1.11	2	2	38	49	-	-
2099	NARMADA LALPUR, JABALPUR	MADHYA PRADESH	17	25	7	8.7	7.4	8.6	130	378	BDL	1.9	BDL	0.87	2	4	9	39	-	-
1234	NARMADA AT HOSHANGABAD U S	MADHYA PRADESH	17	32	8.3	9.8	7.5	7.9	246	298	1.1	1.7	0.95	1.75	2	2	26	63	2	2
2112	NARMADA NEAR MORTAKKA BRIDGE, BADWAH	MADHYA PRADESH	19	24.2	7.2	8.3	7	8.3	270	489	BDL	1.2	BDL	3.65	2	2	34	49		
44	NARMADA AT SETHANIGHAT	MADHYA PRADESH	18	31	5.8	9.1	7.6	8.39	268	367	1	1.5	1.62	7.64	2	35000	38	160000	2	8

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 figure.

Station code	Name of the Monitoring	State Name	Temp_MIN	Temp_MAX	DO_MIN	DO_MAX	pHMIN	pHMAX	CondMIN	CondMAX	BOD_MIN	BOD_MAX	N_MIN	N_MAX	FC_MIN
3321	NARMADA AT AMARKANTAKE FROM ORIGIN POINT, REWA	MADHYA PRADESH	18.5	29.5		7.5	8.4	8.01	218	387	BDL	1.7	BDL	0.49	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
1241	NARMADA AT MANDLA NEAR ROAD BDG. M.P.	MADHYA PRADESH	17.2	26		6.8	8.3	7.2	8.7	160	363	BDL	2.6	BDL	0.8	2
1240	NARMADA AT NARSINGHPUR M.P.	MADHYA PRADESH	17	23.5	7		8.6	7	8.3	170	357	BDL	1.9	BDL	1.16	2
2106	NARMADA AT NEMAWAR	MADHYA PRADESH	22	28	7		7.9	7.5	8.3	279	360	BDL	2	0.31	51	2
45	NARMADA AT MANDLESHWAR	MADHYA PRADESH	20	24		7.3	8.6	7.4	8.5	214	340	BDL	1.3	0.51	1.33	2
1239	NARMADA AT BADWANI, M.P.	MADHYA PRADESH	20	24.6		7.2	8	7.7	8.3	152	594	BDL	1.2	BDL	1.24	2
1431	NARMADA AT MAHESHWAR, M. P.	MADHYA PRADESH	19	24.3		7.3	8	7.4	8.1	258	338	BDL	1.3	0.61	1.11	2
2099	NARMADA LALPUR, JABALPUR	MADHYA PRADESH	17	25	7		8.7	7.4	8.6	130	378	BDL	1.9	BDL	0.87	2
1234	NARMADA AT HOSHANGABAD U S	MADHYA PRADESH	17	32		8.3	9.8	7.5	7.9	246	298	1.1	1.7	0.95	1.75	2
2112	NARMADA NEAR MORTAKKA BRIDGE, BADWAH	MADHYA PRADESH	19	24.2		7.2	8.3	7	8.3	270	489	BDL	1.2	BDL	3.65	2
44	NARMADA AT SETHANIGHAT	MADHYA PRADESH	18	31		5.8	9.1	7.6	8.39	268	367	1	1.5	1.62	7.64	2
1235	NARMADA AT HOSHANGABAD D S	MADHYA PRADESH	18	30		5.6	8.7	7.4	8.2	278	365	1	1.5	2.12	8.12	2
2113	NARMADA NEAR PUNASA DAM, PUNASA	MADHYA PRADESH	19	24.3		7.1	8.2	7.5	8.4	242	1676	BDL	1.1	0.59	1.02	2
1430	NARMADA AT D S OF OMKARESHWAR M.P.	MADHYA PRADESH	19	24.4		7	8.1	7.4	8.2	224	292	BDL	1.2	0.58	1.11	2

Fig2 Processed excel data

So this is the Data Prepared to apply data pre-processing techniques.

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) contain 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 there are features that 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.

IV. METHODOLOGY

Water Quality Index Model

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) Considered 4(four) water quality parameters (viz. Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), Faecal Coliform & Total Coliform counts) for Water Quality Index (WQI) for which Water Quality Criteria is prescribed.

c) Based on the measured ambient concentrations and corresponding criteria, water quality will be defined as satisfactory or unsatisfactory.

d) The criteria for each parameter are as follows –

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

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 to make our own water quality index based on features 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 (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 Each parameter's weightage (Wi) and unit weight (Wi) are inversely related. Wi is the unit weight of the parameter, and n is the total number of water quality parameters. Wi = K/Si where 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 subindices together (WQI)

WQI = $\sum$ (Wi X Vr).

So After performing all the above-mentioned preprocessing tasks our data is ready for model training

V. MODEL TRAINING

The term "machine learning" (ML) has no common definition. Nevertheless, machine learning is sometimes described as a subset of artificial intelligence that emphasizes on use of data and algorithms to simulate how humans learn, simulate predictive patterns, and gradually increase the accuracy of the system. 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.

Classification is a data mining technique which segregates datapoint into various classes. The main aim of this technique is to precisely predict the target class for each data case in the dataset. 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

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

K - NN classifier – It is a non-parametric classifier that uses proximity to classify or predict how a single data point will be grouped. This approach is applicable to both regression and classification. When a new dataset is provided, it simply classifies the data into a category that is very similar to the dataset that was used for training.

Random Forest classifier – This classifier uses a variety of decision trees on various subsets of the dataset, then averages the results to increase the dataset's predicted accuracy. As there are more trees in the forest, it avoids the issue of overfitting and results in improved accuracy. The Extreme Gradient Boosting (XGBoost) gradient-boosted decision tree (GBDT) machine learning system is scalable and distributed. It supports parallel tree boosting and is the best machine learning tool for regression, classification, and ranking problems.

Logistic Regression:
Logistic regression is one of the Machine Learning algorithms that is most frequently employed in the Supervised Learning category. It is used to forecast the categorical dependent variable using a specified set of independent variables. Logistic regression is used to predict the output for a dependent variable that is categorical. The outcome must thus be a discrete or categorical value. It offers the probabilistic values that lie between 0 and 1 rather than the precise values between 0 and 1. It can be either True or False, 0 or 1, or Yes or No.

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.

Elastic Net Regressor - Regularization and variable selection are both used simultaneously by the regression method known as elastic net. L1 and L2 penalties, or lasso and ridge regression, are combined in this model. 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.

VI. 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:
1. Mean Absolute Error (MAE) - Regression accuracy is measured by MAE. The absolute values of the errors are added up and divided by the overall number of values. It gives each incorrect value the same amount of weight.
2. Mean Square Error (MSE) - 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.
3. Root Mean Square Error (RMSE) - Simply taking the square root of mean squared error (MSE), or RMSE, scales thevalues of MSE near the ranges of measurements.
4. R Squared Error (RSE) - 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.
For Classification, the measures used are as follows:
1. Accuracy - The model's accuracy is measured by how many of the observed values it correctly predicted.
2. Precision - Precision is the percentage of instances of a particular positive class that is correctly classified out of all instances of that class that are classified.
3. Recall - The percentage of instances of a specific positive class that was really accurately categorized is known as recall.
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.

VII. Results

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 Elastic Net Regressor having an MAE of 0.981, MSE of 8.916, RMSE of 2.986, and RSE of 0.927, to be the most efficient algorithm.

Model Name	R2	RMSE	MAE	MSE
Linear Regressor	0.999	0.032	0.016	0.0010
ElasticNet	0.927	2.986	0.981	8.916
Random Forest Regressor	0.687	0.065	0.205	0.2561

After applying the regression algorithms, we divided the datasets into two classes on the basis of 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 F-score was 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

VIII. CONCLUSION

One of the most vital resources for survival is water, and the WQI measures the quality of water. Traditionally, one must undergo an expensive and time-consuming lab analysis totest the purity of the water. 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).

ACKNOWLEDGMENT

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

REFERENCES

[1] Palit, Debnath & Mondal, Saikat & Chattopadhyay, Pinaki. (2019). Analysing Water Quality Index of Selected Pit-Lakes of Raniganj Coal Field Area, India. 1167-1175.

[2] Aladejana, Jamiu & Talabi, Abel. (2013). Assessment of Groundwater Quality in Abeokuta Southwestern, Nigeria. 21-31.

[3] DOI No.: http://doi.org/10.53550/EEC.2022.v28i01.072B.

[4] J.-G. Lu, "Title of paper with only the first word capitalized," *J. Name Stand. Abbrev.*, in press.

[5] Ahmed U, Mumtaz R, Anwar H, Shah AA, Irfan R, García-Nieto J. Efficient Water Quality Prediction Using Supervised Machine Learning. Water. 2019; 11(11):2210. https://doi.org/10.3390/w11112210M.

[6] References –
CPCB, ADSORBS/3 1978–1979) Scheme for zoning and classification of Indian Rivers: estuaries and coastal waters. CPCB website: www.CPCB.nic.in.

[7] 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.

[8] Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 2005, 30, 79–82.

[9] Menard, S. Coefficients of determination for multiple logistic regression analysis. Am. Stat. 2000, 54, 17–24.

[10] Shafi, U.; Mumtaz, R.; Anwar, H.; Qamar, A.M.; Khurshid, H. Surface Water Pollution Detection using Internet of Things. In Proceedings of the 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), Islamabad, Pakistan, 8–10 October 2018; pp. 92–96.

[11] Goutte, C.; Gaussier, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In Proceedings of the European Conference on Information Retrieval, Santiago de Compostela, Spain, 21–23 March 2005; pp. 345–359.