Shodhshauryam, International Scientific Refereed Research Journal
Available online at : www.shisrrj.com
OPEN ACCESS
© 2024 SHISRRJ | Volume 7 | Issue 2
ISSN : 2581-6306
doi : https://doi.org/10.32628/SHISRRJ

Predictive Model of Water Quality Analysis

B. Rupadevi[1], B. Naveen[2]

[1]Associate Professor, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India

[2]Post Graduate, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India

Article Info

Article History
Received : 02 April 2024
Published : 13 April 2024

Publication Issue :
March-April-2024
Volume 7, Issue 2

Page Number : 560-567

ABSTRACT

In order to provide safe drinking water for people and aquatic life, the water quality analysis and prediction project describes how machine learning algorithms will be employed in this regard. It talks about how important these kinds of predictive models are to preserving water safety and emphasizes how machine learning is becoming more and more popular in this field since it can manage complicated datasets and produce precise forecasts. The research discusses the use of random forests, a machine learning method, in the analysis and prediction of water quality. It focuses on how machine learning models may be taught to predict changes in these characteristics based on historical data and environmental variables. These factors include pH, temperature, dissolved oxygen, and nutrient levels. This paper proposes a method for the Random Forest Regressor-based prediction of water quality. This study investigates various supervised machine learning techniques to estimate the water quality class (WQC), a unique class defined based on the WQI, and the water quality index (WQI), a single index to characterize the overall quality of water. Several tests are carried out with real-world datasets related to water quality to assess the model's efficacy. Performance evaluation metrics include R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The efficacy of the Random Forest Regressor approach is demonstrated by comparisons with other machine learning algorithms. The outcomes show that the system can accurately predict water quality measures.. The importance of predicting and analyzing water quality as well as the potential benefits of using machine learning techniques are highlighted in the abstract's conclusion. It suggests that these techniques may be used to develop accurate and reliable models that effectively safeguard water resources, ensuring their sustainability and security.

Keywords : Predictive modelling, Machine Learning, Random Forest

Regressor, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared, Water Quality Index (WQI), Water Quality Class (WQC), Comparative Analysis, Scalability, Real-Time Monitoring

I. INTRODUCTION

Predictive models for analyzing water quality are essential for maintaining ecosystem sustainability, promoting economic growth, and protecting public health. Conventional techniques for evaluating water quality, like laboratory analysis and manual monitoring, frequently have drawbacks such as drawn-out procedures and a deficiency of real-time information about variations in water quality. The aim of this research is to evaluate the predictive performance of RandomForestRegressor, a machine learning method that can handle non-linear correlations in high-dimensional datasets, for urban water quality forecasting. Evaluating the water's quality is essential to protecting water resources.

To address this need, our research creates a machine learning predictive model for assessing water quality. The main objective of the model will be to anticipate the water Quality Index(WQI), a through indicator of the water quality, by utilizing significant factors such as conductivity, pH, and dissolved oxygen. By accurately predicting WQI the model will provide meaningful data for informed environmental management decision making. This project has several steps. Data preprocessing will first be used to clean up and prepare the dataset for analysis.

This address encoding categorical variables, managing missing values, and fixing formatting errors. The most important variables for WQI

After that, prediction will be discovered by feature extraction. The Random Forest technique will next be applied to train the predictive model with the pre-processed data. The model's performance and accuracy will be assessed using metrics like Mean Squared Error (MSE), R-squared (R2), Root Mean Squared Error (RMSE), and Mean Average Error (MAE). This project has several steps. Data pre-processing will first be used to clean up and prepare the dataset for analysis. This includes handling missing values, correcting formatting problems, and encoding categorical variables. Feature extraction will next be used to identify the variables that are most crucial for WQI prediction. The Random Forest technique will next be applied to train the predictive model with the pre-processed data. The model's performance and accuracy will be assessed using metrics like Mean Squared Error (MSE), R-squared (R2), Root Mean Squared Error (RMSE), and Mean Average Error (MAE). The main objective of this project is to supply a reliable and efficient tool for evaluating water quality, which will support environmental management and monitoring. Through the use of machine learning, the research seeks to improve the effectiveness of water quality monitoring and support water resource conservation

II. METHODOLOGY

A. Gathering water information :-

The first phase involves obtaining vital water quality indicators, including Total Coliforms Mean, Dissolved Oxygen, Potential of Hydrogen (pH), Conductivity,

Biochemical Oxygen Demand (BOD), and levels of Nitrate and Nitrite. The input data for these parameters is this.

B. **Data Validation and Evaluation**:

Upon gathering the necessary water quality information, through validation and evaluation of the dataset are conducted. This involves checking for data consistency, completeness, and correctness.

C. **Training and Testing Models:**

The data is trained using machine learning models, like RandomForestRegressor, to forecast water quality metrics. In order to evaluate the trained models' prediction accuracy and generalization skills, a thorough assessment is conducted utilizing cross-validation techniques and R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE)

D. Data visualization

Data can be visualized through the histogram, pie chart, countplot, lineplot, data.hist() function from pandas can be used to create a histogram for multiple variables. In the context of water quality analysis, a histogram can be used to visualize the distribution of multiple water quality parameters, such as temperature, pH, and dissolved oxygen e.t.c The figsize argument is used to specify the size of the plot

E. Determining Water Quality Classification:

Once the WQI is predicted, the water quality is classified into different categories based on established standards and guidelines. Common classifications include: (Drinking Water | Recreational Use (Swimming)| Agricultural Irrigation | Industrial Aquatic Ecosystem Support | Fishery Support | No Recommended Use)

F. Integration with User-Friendly Interface :

A user-friendly web-based interface seamlessly incorporates the final prediction model, giving stakeholders instant access to projections about the quality of the water. Users can investigate historical patterns in water quality data, see prediction findings, and input pertinent parameters using interactive capabilities available in the user interface.

III. Symbols and Mathematical Expressions

A number of mathematical expressions and symbols can be included in the suggested machine learning-based water quality prediction system to represent different facets of the approach. Here are a few instances:

Water Quality Index (WQI) Calculation:

- $WQI = \sum^{n} (w_i \times I_i)\ i{=}1$
- The Water Quality Index (WQI)
  Each sub-index, such as pH, dissolved oxygen, nitrate, etc., has a weight (Wi). Each water quality parameter has an individual sub-index (Ii).
- n = number of parameters considerd

Data preprocessing equations

- **Mean Imputation for Missing Values:**
  - This is a method to handle missing data by replacing them with the mean value of the available data.
  - $\text{mean}=1/n \sum^{n}_{i=1} x^{i}$

- n = number of available data points
- xi: Value of the i^{th} data point
- fillna(x): Function to fill missing values in variable x with its mean.

- **Model evaluation metrices**

Mean absolute error (MAE) MAE = 1/n |yi – yi^| ∑ni=1
o˜yi^ = Expected value o˜yi = Actual value
o^n = Total Samples
oIt calculates the mean absolute difference between the calculated and actual values.
MEAN SQUARED ERROR (MSE)
mean squared error (MSE) is equal to 1/n ∑ni=1 (yi – yi^).2 o˜yi: Real value o˜y^i: Estimated value
o: The quantity of samples
oThe average squared difference between the actual and anticipated numbers is what it calculates.

• RMSE, or root mean squared error

• RMSE equals $\sqrt{MSE}$

• R-squared ($R2$)

• yi - yi^ = ∑ni = 1 (R-squared ($R2$) Score = 12/yi – y¬-– = ∑ni = 12.

Where yi is the actual value and y^I is the anticipated value.
• y- = actual values' mean
• n = the quantity of samplesIt calculates the percentage of the dependent variable's variance that can be predicted based on the independent variables.

The mean absolute error (MAE), mean squared error (MSE), and root mean square error (RMSE) between the actual and predicted water quality index (WQI) values for the test set are determined using the scikit-learn mean_absolute_error(), mean_squared_error(), and np.sqrt(metrics.mean_squared_error()) functions.

The average absolute difference between the actual and anticipated WQI values is measured by the MAE. Regardless of the direction of the errors, it calculates their average magnitude. A better fit between the actual and expected WQI values is indicated by a lower MAE.

The average squared difference between the actual and anticipated WQI values is measured by the MSE. It calculates the average error size, but squares the mistakes first, so big errors are given more weight than little errors.
greater errors are given more weight. A better fit between the actual and anticipated WQI values is shown by a lower MSE.

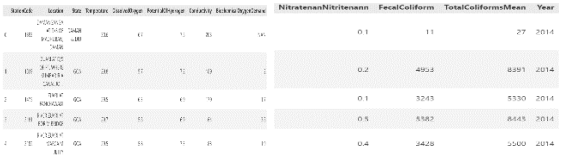
The square root of the MSE is the RMSE. On the same scale as the WQI values, it calculates the average magnitude of the mistakes. A better fit between the actual and expected WQI values is shown by a lower RMSE.

IV. Analysis

The initial process begins by importing necessary libraries such as NumPy, Pandas, Seaborn, Matplotlib, and scikit-learn. For jobs involving data manipulation, visualization, and machine learning, these libraries are indispensable. examining the dataset Pandas is used to read the dataset including data on water quality from a CSV file. pd.read_csv('/content/DataSet.csv') = data
. Basic analysis of the dataset is performed using

methods like head(), describe(), and info() to understand its structure and contents.

data.head()



NitratenanNitritenann	FecalColiform	TotalColiformsMean	Year
0.1	11	27	2014
0.2	4953	8391	2014
0.1	3243	5330	2014
0.5	5382	8443	2014
0.4	3428	5500	2014

The data.info() function is used to display the data types and missing values for the data DataFrame. The output includes the data type, number of non-null values, and memory usage for each variable in the DataFrame.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1991 entries, 0 to 1990
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   StationCode             1991 non-null   object
 1   Location                1991 non-null   object
 2   State                   1991 non-null   object
 3   Temperature             1991 non-null   object
 4   DissolvedOxygen         1991 non-null   object
 5   PotentialOfHydrogen     1991 non-null   object
 6   Conductivity            1991 non-null   object
 7   BiochemicalOxygenDemand 1991 non-null   object
 8   NitratenanNitritenann   1991 non-null   object
 9   FecalColiform           1991 non-null   object
 10  TotalColiformsMean      1991 non-null   object
 11  Year                    1991 non-null   int64
dtypes: int64(1), object(11)
memory usage: 186.8+ KB
```

The next step is to check the missing values in the dataset using the isnull().any() method. It provides information about the presence of missing values in the variables. This can help you to understand the completeness of the data and identify any issues that need to be addressed

```
StationCode              False
Location                 False
State                    False
Temperature              False
DissolvedOxygen          False
PotentialOfHydrogen      False
Conductivity             False
BiochemicalOxygenDemand  False
NitratenanNitritenann    False
FecalColiform            False
TotalColiformsMean       False
Year                     False
dtype: bool
```

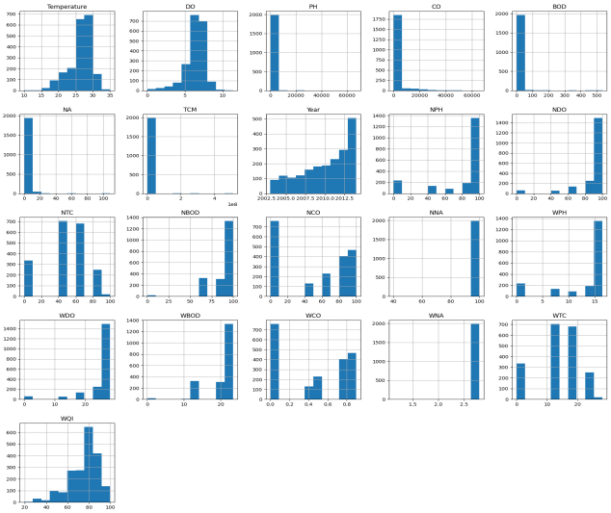
Renaming the columns name with their respective short forms using “data” variable for example data=data.rename(columns = {'DissolvedOxygen': 'DO'}) and then print the data

	SC	LC	ST	Temperature	DO	PH	CO	BOD	NA	TCM	Year
0	1393	DAMANGANGA AT D/S OF MADHUBAN, DAMAN	DAMAN & DIU	30.600000	6.7	7.5	203.0	6.940049	0.100000	27.0	2014
1	1399	ZUARI AT D/S OF PT. WHERE KUMBARJUA CANAL JOI...	GOA	29.800000	5.7	7.2	189.0	2.000000	0.200000	8391.0	2014
2	1475	ZUARI AT PANCHAWADI	GOA	29.500000	6.3	6.9	179.0	1.700000	0.100000	5330.0	2014
3	3181	RIVER ZUARI AT BORIM BRIDGE	GOA	29.700000	5.8	6.9	64.0	3.800000	0.500000	8443.0	2014
4	3182	RIVER ZUARI AT MARCAIM JETTY	GOA	29.500000	5.8	7.3	83.0	1.900000	0.400000	5500.0	2014

Calculating the Water quality index(WQI) for each data points on different parameters, The apply() function from pandas can be used to apply a function to calculate a parameter values of water quality analysis, this can be used to calculate the different parameters value based on their measurements.

Data visualization the next step is to visualize the data by univariant, bivariant and multi-variant analysis. Multivariant analysis displays the entire water parameter histogram in a single frame; subsequent analysis displays separate frames.

data.hist() function from pandas is used to create a histogram for all the numerical variables in the data DataFrame. The figsize argument is used to specify the size of the plot.



The next action is to create an instance of the class using the LabelEncoder() class. The data in the 'LC' and 'ST' columns of the DataFrame are then transformed by using the fit_transform() method to fit the encoder to those columns and replace the

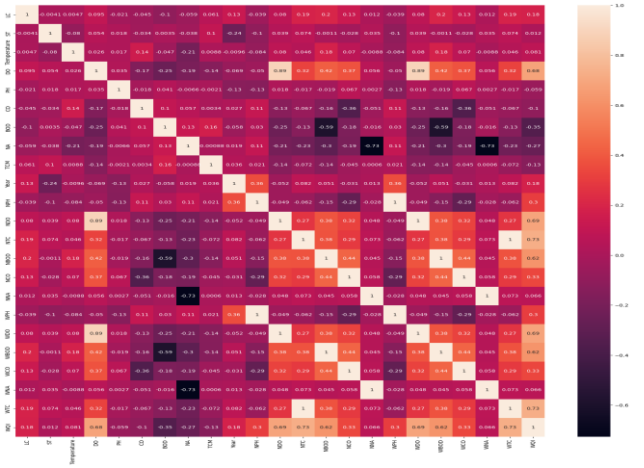
categorical values with numerical values. Machine learning algorithms that need numerical input can perform better if categorical variables are encoded into numerical variables. The encoded variables, for instance, can be fed into a machine learning algorithm.

The succeeding step is to finding correlation matrix using heat map. The correlations between water quality metrics can be seen in the context of an investigation of water quality using a correlation matrix. It offers details on the direction and strength of the correlations that exist between the variables. This can assist you in figuring out how the various water quality parameters relate to one another and in identifying any problems that require attention.The correlation matrix of the data DataFrame using the corr() function from pandas. Next, a heatmap of the correlation matrix is made using seaborn's heatmap() method. The correlation coefficients are displayed as text on the heatmap using the annot=True option.The correlations between the water quality measures can be seen by making a heatmap of the correlation matrix.

A perfect positive linear relationship, or one in which the value of one variable grows as the value of the other increases, is indicated by a correlation coefficient of +1.

• A complete negative linear relationship, where the value of one variable increases and the value of the other variable drops, is indicated by a correlation coefficient of -1

sns.heatmap (data.corr(), annot=True) : The above line uses the seaborn library to create a heatmap. The correlation matrix of the dataset, which most likely includes several characteristics or factors connected to water quality (such as pH, temperature, dissolved oxygen, turbidity, etc.), is calculated by the function data.corr().



The heatmap's color intensity indicates how strongly the two variables are correlated. In general, stronger correlations—whether positive or negative—are shown by deeper hues, whereas weaker correlations are indicated by lighter hues.

Splitting train and test data :- x typically represent the dataset's features or independent variables. When predicting the quality of water, x may consist of several factors such as conductivity (CO), pH, dissolved oxygen (DO), biochemical oxygen demand (BOD), and other pertinent indicators of water quality.

y: Stands for the variable you're trying to predict. Most likely, in this instance, the Water Quality Index (WQI) determined using the input features x.

test_size is equal to 0.2. The percentage of the dataset to be included in the test split is specified by this option. In this instance, the model will be trained using 80% of the data, with the remaining 20% being used for testing.

random_state = 10: The split's repeatability is guaranteed by this option. Using a random state guarantees that the split will always be the same.

Cross validation:- The k-fold cross-validation namely with k=5 (cv=5). This indicates that the

model is trained and assessed five times, using a split water quality dataset into equal portions. Every time, the remaining four components are used as the training set, while one of the five components serves as the validation set (testing set). Five times, this process is carried out, using a new portion each time that is chosen as the validation set.

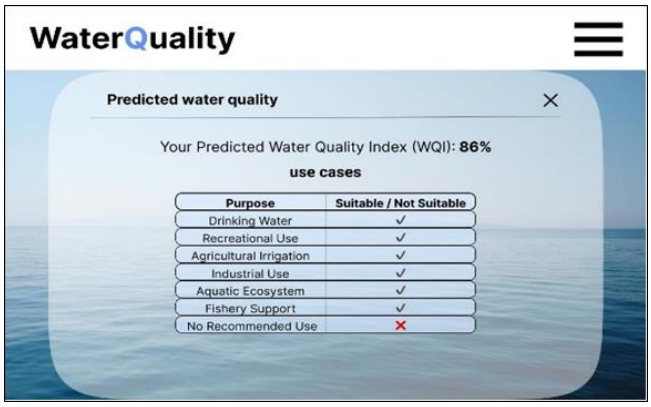
Model evaluation :- The model is evaluated by importing indicators like root mean square error (RMSE), mean absolute error (MAE), and mean squared error (MSE)

For the testing set, the root mean squared error between the actual and predicted values is computed using the np.sqrt() function. The model's prediction accuracy of water quality indicators is quantified by these evaluation metrics. Metrics like MAE, MSE, and RMSE are used to compare the model's predictions with actual values in order to evaluate the model's performance and pinpoint areas that require improvement.

An HTML template for the web application, likely providing a user interface for inputting data and displaying predictions. Finally developing a a Flask application that uses a trained model to forecast water quality. "/" route: Serves the index.html file from the build folder. This route is associated with the root URL of the Flask application. "/upload" route: Accepts POST requests with water quality information, processes the input, and returns a success response. This route seems to handle uploading water quality data to your application. "/predict" route: Accepts POST requests with water quality parameters in JSON format, processes the input, makes predictions using the loaded model, and returns the predicted Water Quality Index (WQI) as a JSON response.

V. Result

User input form



Predicted values and use cases

VI. Conclusion

In summary, by forecasting water quality, our effort has shown how machine learning may be used to address the problems associated with water management. We have demonstrated a proactive strategy for efficiently managing water supplies by developing an interactive web application and utilizing prediction models. We were able to assess complex data on water quality, pinpoint important characteristics, and develop predictive models that offer useful information for making decisions. by applying machine learning approaches. By taking a proactive stance, stakeholders can predict shifts in the quality of the water and take prompt action to reduce hazards to ecosystems and public health. In addition, the use of cutting-edge technologies like It

is now simpler to develop a user-friendly application that enables users to interact with and obtain real-time water quality predictions thanks to HTML, CSS, React.js, Flask, and MySQL. This intuitive interface enhances the utility and practicality of our prediction model, empowering stakeholders to make knowledgeable decisions regarding water management strategies.

VII. REFERENCES

[1]. Brown R. M., McClelland N. I., Deininger R. A., Tozer R. G. 1970 A water quality Index do we dare? Water and Sewage Works, October 1970, 339-343

[2]. Bureau of Indian Standards 2012 Indian Standard Drinking Water Specification (Second Revision)

[3]. Deshpande L. undated Water Quality Analysis: Laboratory Methods. National Environmental Engineering Research Institute (NEERI), Nagpur, Council of Scientific & Industrial Research, New Delhi, Govt. of India

[4]. KoriR., Parashar S., Basu, D.D. undated Guide Manual: Water and Wastewater Analysis. Central Pollution ControlBoard, Ministry of Environment and Forest, India

[5]. Metcalf E., Eddy H. 2003 Wastewater Engineering: Treatment and Reuse.Tata McGrawHill Publishing Co Ltd, India.

[6]. RoyR. 2018 An Approach to Develop an Alternative Water Quality Index

[7]. FLDM. In: Majumder M. (eds) Application of Geographical Information Systems and Soft

[8]. Computation Techniques in Water and Water Based Renewable Energy Problems.Water Resources Development and Management. Springer, Singapore