Data Preprocessing System

*Did you know that the United Nations has declared 2021-2030 as the Decade of Ocean Science for Sustainable Development? The main goal of this initiative is to improve the management and conservation of the ocean and its resources, which are essential for life on Earth. One of the key challenges is to monitor and improve the water quality of the ocean, which is affected by pollution, climate change, overfishing, and other factors.*

Data Collection

Data preprocessing

Original Data

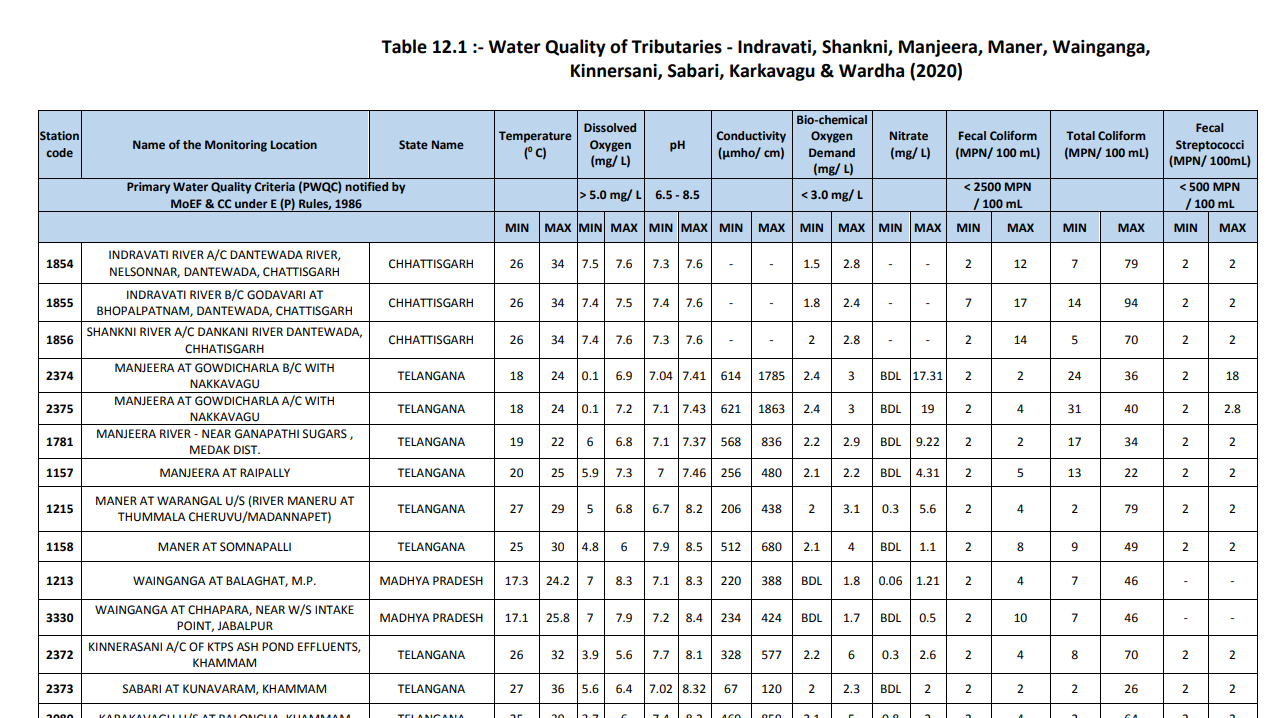
Access to safe drinking water is critical to health, a basic human right, and a component of any health-protection strategy. This is significant as a national, regional, and local health and development issue. It has been demonstrated in some locations that investments in water supply and sanitation 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 of water quality of Indian rivers was collected from Central Pollution Control Board (CPCB) database. CPCB is a statutory organisation that promotes the purity of streams and wells in various States by preventing, controlling, and abating water pollution. It collects, collate and disseminate technical and statistical data relating to water pollution and propose 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 since several years. The dataset used in the Model was from data collection of year 2021

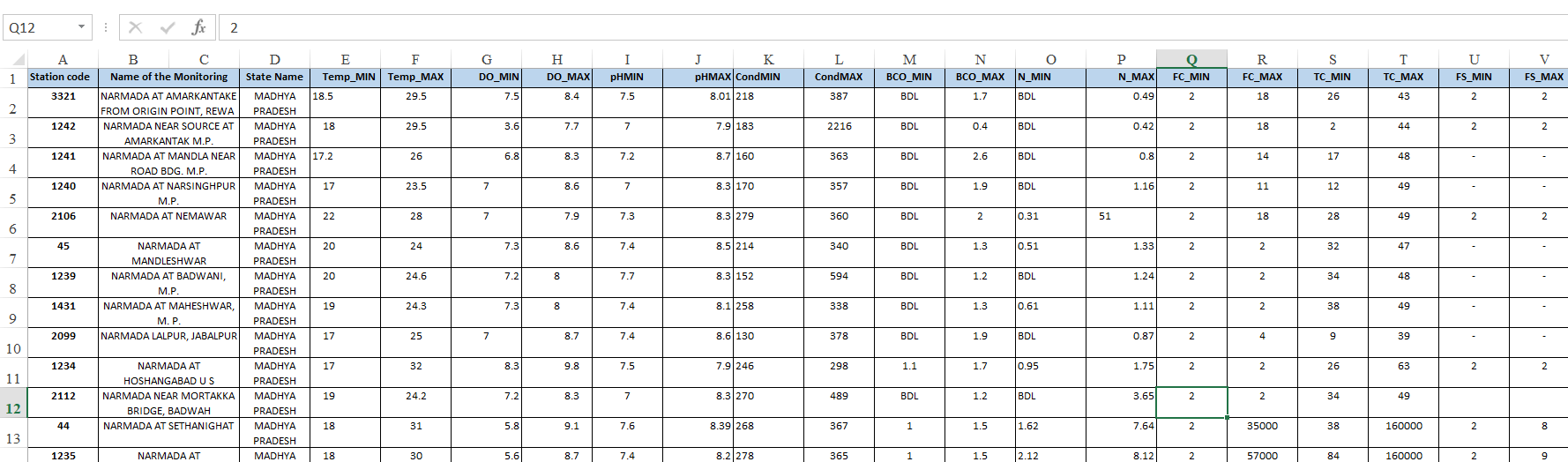
Now as we have seen different datasets on water quality. Let us go through different pre-processing tasks we performed on the data. Data preparation is a process of preparing raw data that can be useful for data preprocessing and analysis. Also Data Preparation is the main task for Machine learning.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

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

The Data preparation: We converted the whole Water Quality Data of rivers under national water quality Monitoring programme to excel file. We did this using pdf to excel converter provided by Adobe open source software. The software convert the pdf which contain table like data to excel spreadsheet data. Our next step was to separate the data of different river into different data files which can be used 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.

Data description: The Excel file contain different columns namely as shown in figure below

The table was not good for processing as the first row contains different columns and there were maximum and minimum value associated to most of the column so we cannot train the machine learning algorithms on this data types. So we edited and updated the excel data so each colums have only one value and try to make data in structured format for machine learning to perform on it.

We renamed the columns and made significant changes that can be shown in below figure.

(Table for explaining changes done to the excel file)

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

Data Pre Processing Techniques:

Data pre-processing is the process of converting raw data into a usable, intelligible format. Real-world or raw data often contains irregular formatting, 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 Pandas isnull () function to detect missing values. We got no missing values in the dataset. After going through the dataset, we observed that certain feature contain string values in them. For example, the N\_min (Nitrates minimum) colums contain too many string named BDL and alse FS\_MIN (Fecal Streptococci minimum) contain string like “-“. So basically all this 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 mean value of that data column. So now all feature does not have null values.

1. Simplifying dataset:

Now there are feature have max and min values so we cannot give models this much 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 max and minimum value of the min and max columns. So our final dataset contains 12 Features.

Methodology

Water Quality Index Model

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 wqi value means better water quality, while a lower value means worse water quality.

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

a) WQI is a single defining criteria 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 is as follows -

Dissolved Oxygen: < 4.0 mg/l

Biological Oxygen Demand: > 3.0 mg/l

Faecal Coliform: > 2500 MPN/ 100 ml

Total Coliform: > 5000 MPN/ 100 ml

- ---

e) Even one single parameter from these four parameters exceeding the criteria values will considered as 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- Tempearture(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). Next step is to develop rating scale to obtain the rating Vr. The unit weight of each parameter (Wi) was calculated using weightage criteria of eachfeatures. Sub index value was determined with formula (Wi X Vr). Final Value of Water Quality Index was calculated using Summation of all the subindex value of each features.

Unit weight of each parameter the unit weight (Wi) of each parameter is proportional to the weightage of each parameter. Wi = K/Si where K = 1 / ( ∑ 1/Si ) where K is the constant of proportionality; Wi is the unit weight of the parameter; n is the number of water quality parameters. The unit weight of each parameter calculated are shown in the table. The sub index value is determined by multiplying its unit weight with its rating obtained. Totalling the sub indices to get the general water quality index (WQI)

WQI = ∑ (Wi X Vr).

So After performing all the above mentioned pre processing task our data is ready for model training

Model Training

Dhruvil

Model Evaluation:

Metrics table 1 Regression Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

5238095238095),

('MSE', 62),

('MAE', 99999999999996),

('Root Mean Squared Error', 0.25616400996236766)])

Table 2 Classification Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

Result

Conclusion

SMOTE

SMOTE is a machine learning technique that solves problem that occur when using an imbalance dataset.SMOTE stand for Synthetic Minority Oversampling Technique. Oversampling means making duplicates of thee data that is the least present in your dataset. SMOTE is an algorithm that performs data augmentation by creating synthetic data points based on the original data points. Here we try not to generate duplicates rather we create points slightly different from the original data points.

Algorithm SMOTE(T, N, k)

Input: Number of minority class samples T; Amount of SMOTE N%; Number of nearest

neighbors k

Output: (N/100) \* T synthetic minority class samples

1. (∗ If N is less than 100%, randomize the minority class samples as only a random

percent of them will be SMOTEd. ∗)

2. if N < 100

3. then Randomize the T minority class samples

4. T = (N/100) ∗ T

5. N = 100

6. endif

7. N = (int)(N/100) (∗ The amount of SMOTE is assumed to be in integral multiples of

100. ∗)

8. k = Number of nearest neighbors

9. numattrs = Number of attributes

10. Sample[ ][ ]: array for original minority class samples

11. newindex: keeps a count of number of synthetic samples generated, initialized to 0

12. Synthetic[ ][ ]: array for synthetic samples

(∗ Compute k nearest neighbors for each minority class sample only. ∗)

13. for i ← 1 to T

14. Compute k nearest neighbors for i, and save the indices in the nnarray

15. Populate(N, i, nnarray)

16. endfor

Populate(N, i, nnarray) (∗ Function to generate the synthetic samples. ∗)

17. while N 6= 0

18. Choose a random number between 1 and k, call it nn. This step chooses one of

the k nearest neighbors of i.

19. for attr ← 1 to numattrs

20. Compute: dif = Sample[nnarray[nn]][attr] − Sample[i][attr]

21. Compute: gap = random number between 0 and 1

22. Synthetic[newindex][attr] = Sample[i][attr] + gap ∗ dif

23. endfor

24. newindex++

25. N = N − 1

26. endwhile

27. return (∗ End of Populate. ∗)

End of Pseudo-Code.