**Predicting And Analysing Urban Water Quality Using Machine Learning**

**Department Of Computer Science And Engineering**

**Chalapathi Institute Of Engineering & Technology**

### Team Members:

### 1.Sowmya Takkellapati

### 2.Sai Priya Maddukuri

### 3.Likhitha Mandadi

### 4. Leela Bhavani Kollisetty

### Introductoin:

### Water is considered as a vital resource that affects various aspects of human health and lives. The quality of water is a major concern for people living in urban areas. Quality of water serves as a powerful environmental determinant and a foundation for the prevention and control of waterborne diseases. However predicting the urban water quality is a challenging task since the water quality varies in urban spaces non-linearly and depends on multiple factors, such as meteorology, water usage patterns, and land uses, so this project aims at building a Machine Learning (ML) model by considering all water quality standard indicators.

### Architecture:

### 

**Python Packages:**

We need to install the following packages:

**1.Numpy:**This package is used to perform numerical computations. This package is pre-installed in anaconda.

**2.Pandas:** Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures, and data analysis tools. This package is pre-installed in anaconda.

**3.Matplotlib:** Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. This package is pre-installed in anaconda.

**4.Scikit-learn:** This is a machine learning library for the Python programming language. This package is pre-installed in anaconda.

**5.Flask:** Flask is a lightweight WSGI web

### Project Flow:

* User interacts with the UI (User Interface) to enter Data
* The entered data is analyzed by the model which is integrated
* Once model analyses the input the prediction is showcased on the UI

To accomplish this, we have to complete all the activities and tasks listed below

* Data Collection:
  + Collect the dataset or Create the dataset
* Data Preprocessing:
  + Import the Libraries.
  + Importing the dataset.
  + Checking for Null Values.
  + Data Visualization.
  + Taking care of Missing Data.
  + Label encoding.
  + One Hot Encoding.
  + Feature Scaling.
  + Splitting Data into Train and Test.
* Model Building:
  + Training and testing the model
  + Evaluation of Model
* Application Building:
  + Create an HTML file
  + Build a Python Code

**Project Objectives:**

By the end of this project:

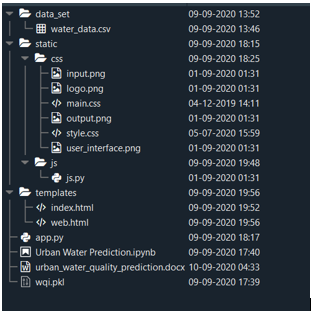
You’ll be able to understand

* Regression and Classification Problems
* To grab insights from data through visualization.
* Applying different algorithms according
* Evaluation metrics

 how to build a web application using the Flask framework

### Project Structure:

Create a Project folder which contains files as shown below

****

1. A python file called app.py for server-side scripting.
2. We need the model which is saved and the saved model in this content is (wqi.pkl).
3. Templates folder which contains web.HTML file.
4. Static folder which contains a CSS folder which contains styles.css and images for web page

### Data Collection:

ML depends heavily on data, without data, it is impossible for an “AI” to learn. It is the most crucial aspect that makes algorithm training possible. In Machine Learning projects, we need a training data set. It is the actual data set used to train the model for performing various actions.

### Download Dataset:

You can collect datasets from different open sources like kaggle.com, data.gov, UCI machine learning repository etc.

 The dataset used for this project was obtained from Kaggle.  Please refer to the link given below to download the data set and to know about the dataset.

Description of the data is also provided in the

**Data Preprocessing:**

Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn’t feasible for the analysis.

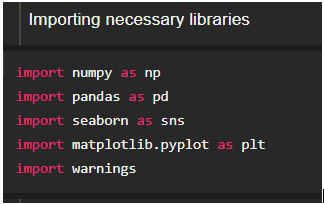
Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data pre-processing Follow the following steps to process your Data

* Import the Libraries
* Importing the dataset
* Taking care of Missing Data
* Label encoding
* One Hot Encoding
* Feature Scaling
* Splitting Data into Train and Test

**Importing The Libraries:**

It is important to import all the necessary libraries such as pandas, numpy, matplotlib.

* **Numpy**- It is an open-source numerical Python library. It contains a multi-dimensional array and matrix data structures. It can be used to perform mathematical operations on arrays such as trigonometric, statistical, and algebraic routines.
* **Pandas**- It is a fast, powerful, flexible, and easy to use open-source data analysis and manipulation tool, built on top of the Python programming language.
* **Seaborn**- Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
* **Matplotlib**- Visualisation with python. It is a comprehensive library for creating static, animated, and interactive visualizations in Python

****

**Reading The Dataset:**

You might have your data in .csv files, .excel files

Let’s load a .csv data file into pandas using read\_csv() function.We will need to locate the directory of the CSV file at first (it’s more efficient to keep the dataset in the same directory as your program).

****

If your dataset is in some other location, Then  see below command

**Data=pd.read\_csv(r”File\_location/filename.csv”)**

**Note:r stands for "raw" and will cause backslashes in the string to be interpreted as actual backslashes rather than special characters.**

* Our Dataset water\_data contains the following Columns:
* 'D.O. (mg/l)' :' dissolved\_oxygen'
* 'CONDUCTIVITY (µmhos/cm)'
* 'B.O.D. (mg/l)' : 'biological\_oxygen\_demand'
* 'NITRATENAN N+ NITRITENANN (mg/l)'
* 'TOTAL COLIFORM (MPN/100ml)Mean'
* 'STATION CODE'
* 'LOCATIONS'
* STATE
* PH

We have to predict the water quality index(wqi) by using the above input variables. By using water quality index we can predict whether the water is contaminated or not. Hence wqi is the output column to be predicted. The index ranges from 1 to 100; a higher number indicates **better water quality.**

**Analyse The Data:**

head() method is used to return top n (5 by default) rows of a DataFrame or series.

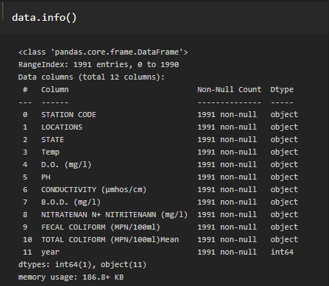


describe() method computes a summary of statistics like count, mean, standard deviation, min, max, and quartile values.

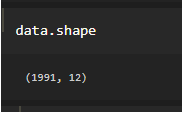


From the data we infer that there are only decimal values and no categorical values

* info() gives information about the data



the shape gives the (number of rows, number of columns)



The dataset contains 1991 rows and 12 columns.

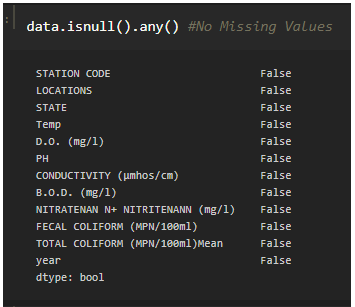
**Handling Missing Values:**

loading it is important to check the complete information of data as it can indicate After many of the hidden information such as null values in a column or a row.

Check for the null values. if it is present then the following steps can be performed

* data using the Imputation method in sklearn.
* Filling NaN values with mean, median, and mode using Imputing fillna() method.

You can check the null values with the function i**snull().any()**

****

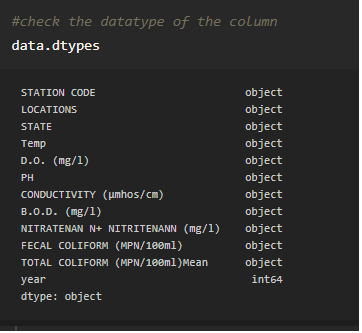
If the dataset contains null values then the above functions return as true. But if you look at the dataset you can observe that the dataset does not have any null values.

So we can skip this step.

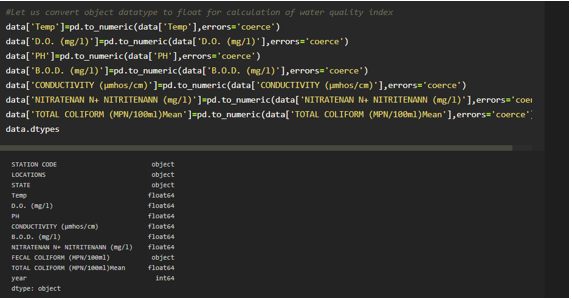
You can also check the number of null values present in the columns by the using**isnull().sum()**function.

### Handling Missing Values -2

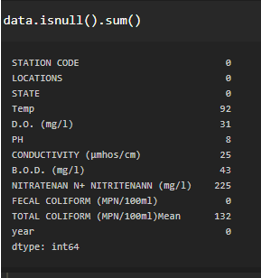
Use data.datatypes to check the type of column

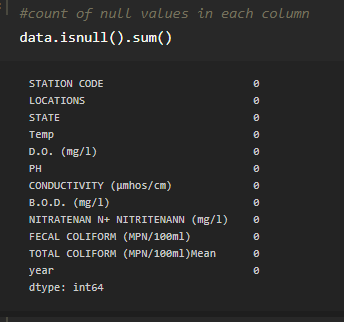


dataset contains numerical columns but if you check the above image our interpreter interpreted the datatypes of columns as objects (string values). so lets covert them to integers or floating values Our as shown below



After converting datatype from object to float, let us check for null values again

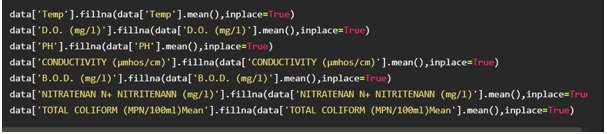


****

**You can refer to the link given at the beginning of this milestone for more info**

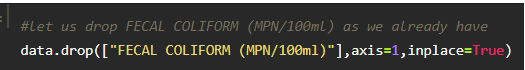
### Handling Missing Values -3

Now we can see that there are null values in the dataset. Let us handle this null data by replacing it with mean or mode



Let us check if NAN is replaced by mean(). by using isnull().any() Function. you should find all zeros with respect to columns

if you look at the data set we have two columns with the same name which is "FECAL COLIFORM".So let's delete one of the same columns using the drop function.

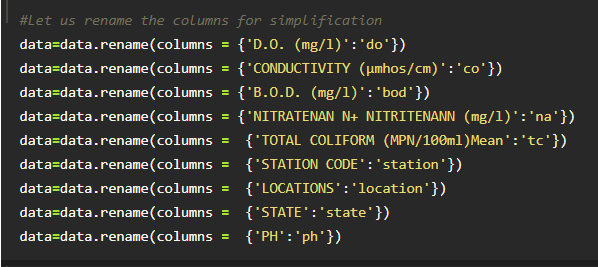


Let us rename the columns for simplification

### Water Quality Index(Wqi) Calculation -2

for the other two columns





### Data Visualization:

Data visualization is where a given dataset is presented in a graphical format. It helps the detection of patterns, trends and correlations that might go undetected in text-based data. Understanding your data and the relationship present within it is just as important as any algorithm used to train your machine learning model. Machine learning models will perform poorly on data that wasn’t visualized and understood properly.

To visualize the dataset we need libraries called Matplotlib and Seaborn. The Matplotlib library is a Python 2D plotting library that allows you to generate plots, scatter plots, histograms, bar charts etc.

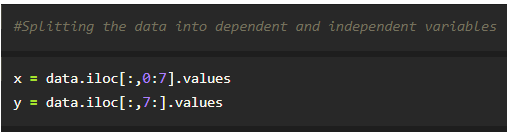
Please refer to the link given to know  about the data visualization

### Splitting Dependent And Independent Columns:

need to split our dataset into the matrix of independent variables and the vector or dependent we variable. Mathematically, Vector is defined as a matrix that has just one column.

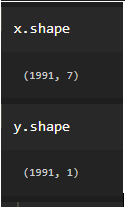
To read the columns, we will use iloc of pandas (used to fix the indexes for selection) which takes two parameters — [row selection, column selection].

Let’s split our dataset into independent and dependent variables.



From the above code “:” indicates that you are considering all the rows in the dataset and “0:7” indicates that you are considering columns 0 to 7 as input values and assigning them to variable x. In the same way in the second line “:” indicates you are considering all the rows and “7:” indicates that you are considering only the last column as output value and assigning them to variable y.

Let's check the shape of x and Y



You can see in x we have 1991 rows with  7  columns

and y has 1 column with the same number of rows

### Model Buiding:

* There are several Machine learning algorithms to be used depending on the data you are going to process such as images, sound, text, and numerical values. The algorithms that you can choose according to the objective that you might have it may be Classification algorithms are Regression algorithms.
* Example:     1. Linear Regression.
  + - 2. Logistic Regression.
    - 3. Random Forest Regression / Classification.
    - 4. Decision Tree Regression / Classification.
* You will need to train the datasets to run smoothly and see an incremental improvement in the prediction rate.
* Now we apply the RandomForestRegression algorithm on our dataset.

**Application Building:**

In this section, we will be building a web application that is integrated to the model we built. A UI is provided for the uses where he has to enter the values for predictions. The enter values are given to the saved model and prediction is showcased on the UI.

This section has the following tasks

* Building HTML Pages
* Building server-side script

### Conclusion: The quality of water is a major concern for people living in urban areas. Quality of water serves as a powerful environmental determinant and a foundation for the prevention and control of waterborne diseases. However predicting the urban water quality is a challenging task since the water quality varies in urban spaces non-linearly and depends on multiple factors, such as meteorology, water usage patterns, and land uses, so this project aims at building a Machine Learning (ML) model by considering all water quality standard indicators.

