★Project 2 : Air Quality Analysis in TN Analysis(DAC_Phase5)

PHASE 5 : Project Documentation & Submission

Title: Air Quality Analysis in TN

Abstract:

An index for reporting air quality is called the air quality index (AQI). It measures the impact of air pollution on a person's health over a short period of time. The purpose of the AQI is to educate the public on the negative health effects of local air pollution. The amount of air pollution in Indian cities has significantly increased. There are several ways to create a mathematical formula to determine the air quality index. Numerous studies have found a link between air pollution exposure and adverse health impacts in the population. Data mining techniques are one of the most interesting approaches to forecast AQI and analyze it. The aim of this paper is to find the most effective way for AQI prediction to assist in climate control. The most effective method can be improved upon to find the most optimal solution. Hence, the work in this paper involves intensive research and the addition of novel techniques such as SMOTE to make sure that the best possible solution to the air quality problem is obtained. Another important goal is to demonstrate and display the exact metrics involved in our work in such a way that it is educational and insightful and hence provides proper comparisons and assists future researchers.

1. Introduction

Air quality is a measure of how clean or polluted the air is. Monitoring air quality is important because polluted air can be bad for our health—and the health of the environment.Air quality is measured with the Air Quality Index, or AQI. The AQI works sort of like a thermometer that runs from 0 to 500 degrees. However, instead of showing changes in the temperature, the AQI is a way of showing changes in the amount of pollution in the air

- To plan a comprehensive programme for the prevention, control and abatement of air pollution.
- To advise the State Government on any matter concerning the prevention, control or abatement of air pollution.
- To collect and disseminate information relating to air pollution and the prevention, control or abatement thereof.
- To inspect sewage and trade effluent treatment plants for their effectiveness and review plans, specifications for corrective measures.
- To inspect industrial plants or manufacturing process, any control equipment and to give directions to take steps for the prevention, control or abatement of air pollution.

2. Design Thinking Process

Process Air Technology

Cleaning Systems

Cleaning systems are used to remove contaminants, clean the resulting fluid flows, and collect materials before discharge of exhaust air.

Pneumatic Conveying Systems

Conveying systems are used to transport captured pollutants from processes to a collection point.

Drying System

Drying systems are used to remove moisture, gases, and vapors from a product.

3. Development Phases

3.1 Data Collection

- Data are collected at a few locations taken to represent transport activity, travel movement and traffic flow across the study area or a sample of individual travellers.
- The dataset may include information on various transport

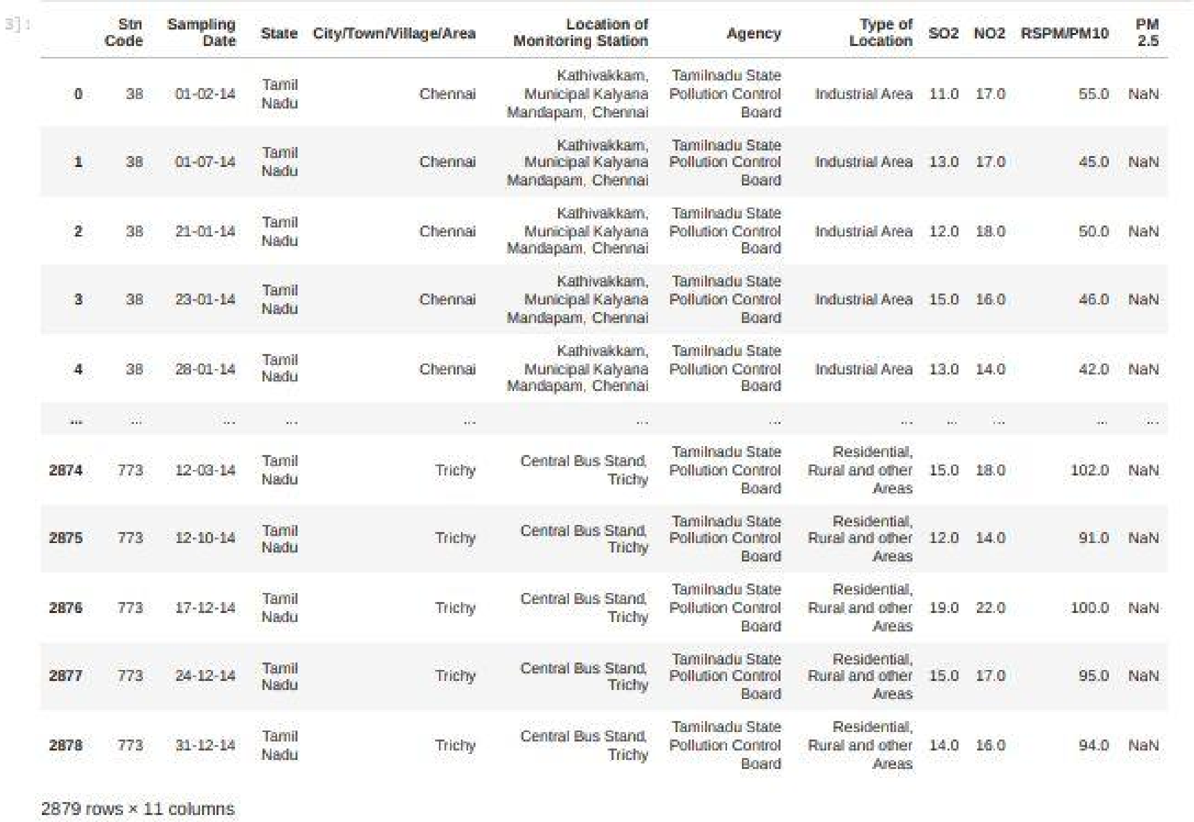
  DATASET LINK:
  https://tn.data.gov.in/resource/location-wise-daily-ambient-air-quality-tamil-nadu-year-2014

3.2 Data Preprocessing

❖ Import necessary packages: Here I have imported packages needed for preprocessing .

The dataset upon loading and observation, it’s important and easy to process the data with smaller size.

- The data types are changed while loading.
- String data is converted into categorical values, float64 to float32 etc. * *
- This process place vital role while handling dataset with larger size.



	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10	PM 2.5
0	38	01-02-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0	NaN
1	38	01-07-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0	NaN
2	38	21-01-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0	NaN
3	38	23-01-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0	NaN
4	38	28-01-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0	NaN
...	...	...	...	...	...	...	...	...	...	...	...
2874	773	12-03-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	18.0	102.0	NaN
2875	773	12-10-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	12.0	14.0	91.0	NaN
2876	773	17-12-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	19.0	22.0	100.0	NaN
2877	773	24-12-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	17.0	95.0	NaN
2878	773	31-12-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	14.0	16.0	94.0	NaN

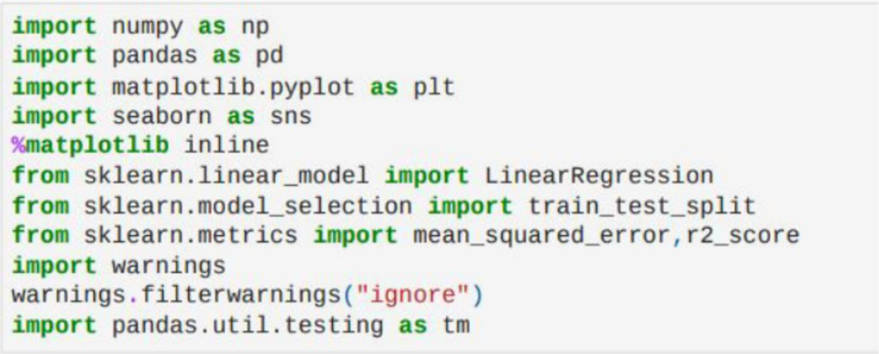
2879 rows × 11 columns

3.3 Exploratory Data Analysis (EDA)

- Conduct summary statistics and visualizations to understand the dataset's characteristics.
- Identify correlations between different air quality parameters.
- Explore potential factors influencing potability.

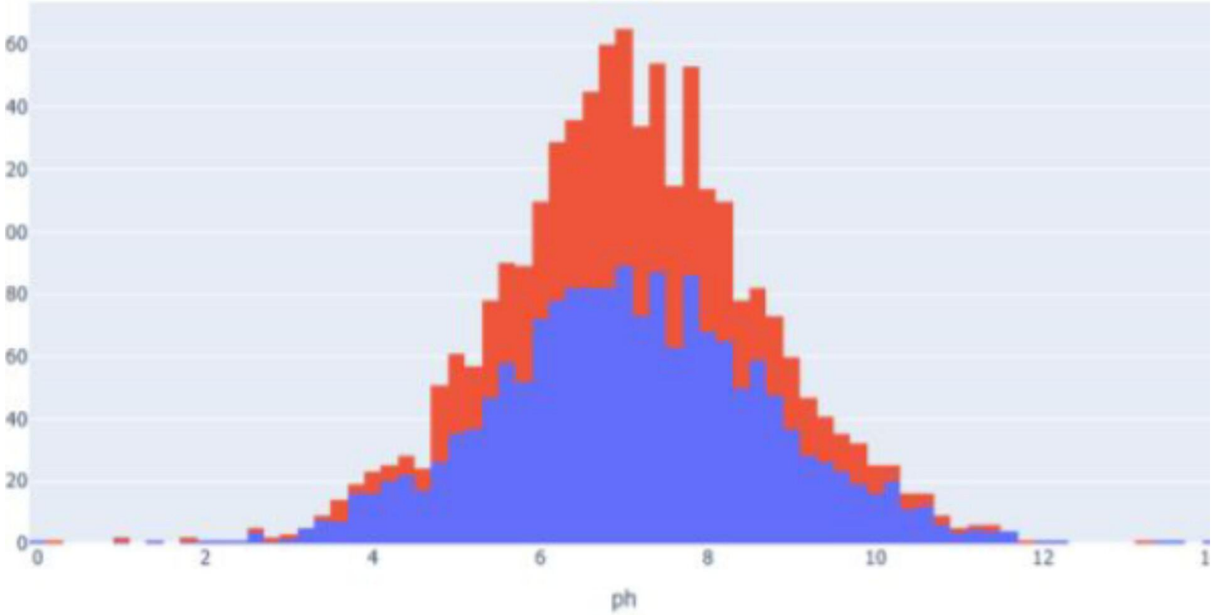
3.4 Data Visualization

- Utilize various data visualization techniques, such as scatter plots, histograms, box plots, and heatmaps, to visually represent the data.
- Visualizations will help in understanding the distribution of air quality parameters and identifying patterns.



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error,r2_score
import warnings
warnings.filterwarnings("ignore")
import pandas.util.testing as tm
```

factors affecting air quality



3.5 Predictive Modelling

- Split the dataset into training and testing sets.
- Choose appropriate machine learning algorithms (e.g., logistic regression, decision trees, random forests, or neural networks) for potability prediction.
- Train and evaluate the models using appropriate performance metrics.
- Fine-tune the models to achieve the best predictive accuracy.

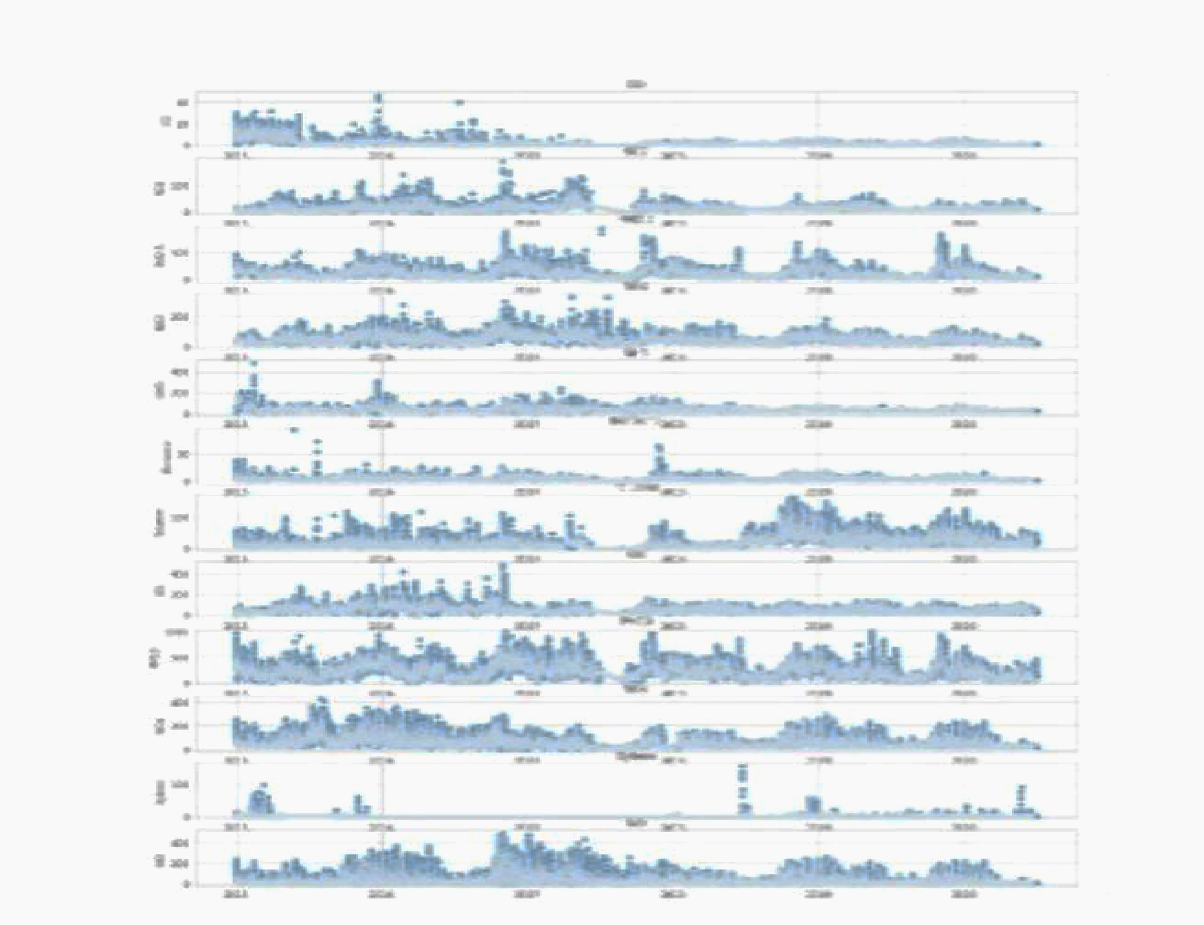
4. Analysis Objectives

4.1 Air Quality Assessment

- ❖ Determine the overall quality of air by analysing various parameters, including ozone(O3),ammonia(NH3),etc...

4.2 Factors Influencing air quality

- ❖ Identify the most influential factors affecting air , providing actionable insights for air quality improvement.

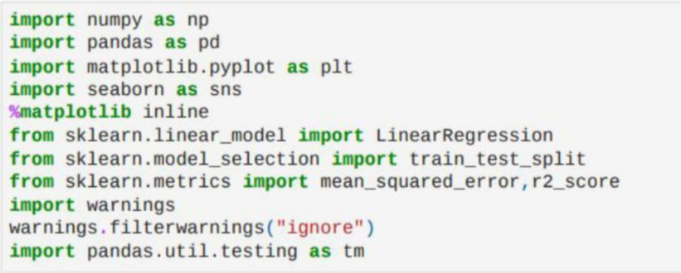


5. Data Preprocessing

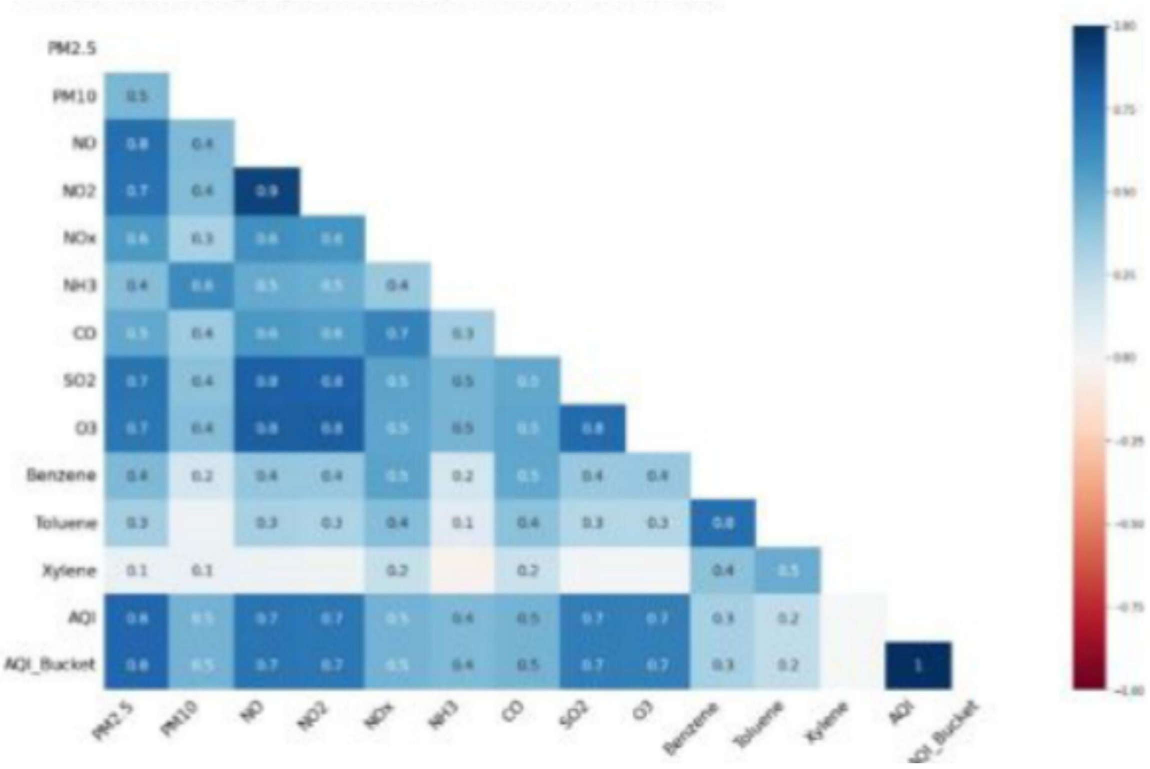
Data preprocessing is a crucial phase to ensure the reliability and accuracy of the analysis. This phase includes the following steps:

5.1 Handling Missing Values

❖ Identify and address missing data points by either imputing values or removing incomplete records.



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error,r2_score
import warnings
warnings.filterwarnings("ignore")
import pandas.util.testing as tm
```



5.2 Outlier Detection and Treatment

❖ Detect and handle outliers that may skew the analysis.

- ❖ Determine whether to remove or transform outlier .

5.3 Data Normalization

- ❖ Normalize the data if necessary to bring all variables to the same scale for accurate modelling.

6.Exploratory Data Analysis (EDA)

EDA is a critical step to understand the dataset and uncover insights. The following EDA tasks will be performed:

6.1 Summary Statistics

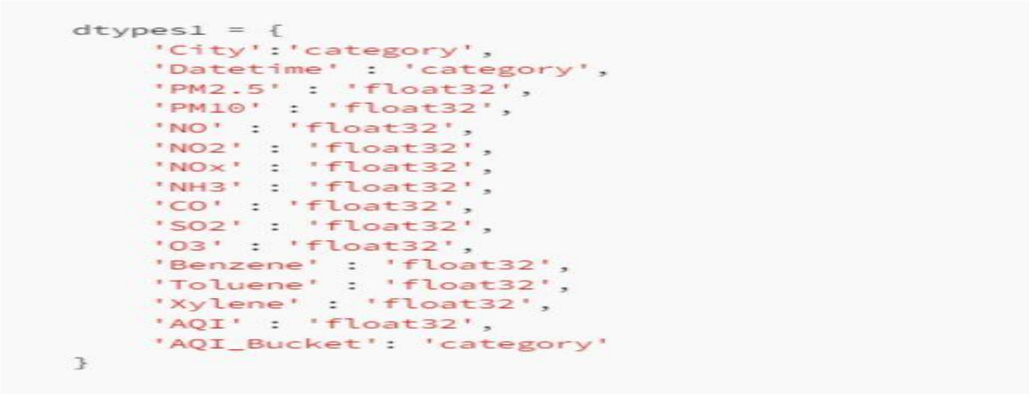
- ❖ Calculate descriptive statistics for all air quality parameters, including mean, median, standard deviation, and percentiles.

6.2 Data Distribution

- ❖ Create histograms, density plots, and box plots to visualize the distribution of each parameter.

6.3 Correlation Analysis

- ❖ Explore the relationships between different parameters by calculating correlation coefficients and creating correlation matrices.



```
dtypes1 = {
        'City':'category',
        'Datetime' : 'category',
        'PM2.5' : 'float32',
        'PM10' : 'float32',
        'NO' : 'float32',
        'NO2' : 'float32',
        'NOx' : 'float32',
        'NH3' : 'float32',
        'CO' : 'float32',
        'SO2' : 'float32',
        'O3' : 'float32',
        'Benzene' : 'float32',
        'Toluene' : 'float32',
        'Xylene' : 'float32',
        'AQI' : 'float32',
        'AQI_Bucket': 'category'
}
```

7. Data Visualization

Data visualization is essential for conveying information and patterns in the data. The following visualizations will be used:

7.1 Scatter Plots

❖ Visualize the relationships between two continuous variables to identify patterns and trends.

7.2 Box Plots

❖ Use box plots to compare the distribution of various parameters for potable and non-potable water samples.

7.3 Time Series Plots (if applicable)

❖ If the dataset includes time-related data, create time series plots to visualize trends over time.

DATA ANALYTICS WITH IBM COGNOS

I. IBM Cognos Introduction

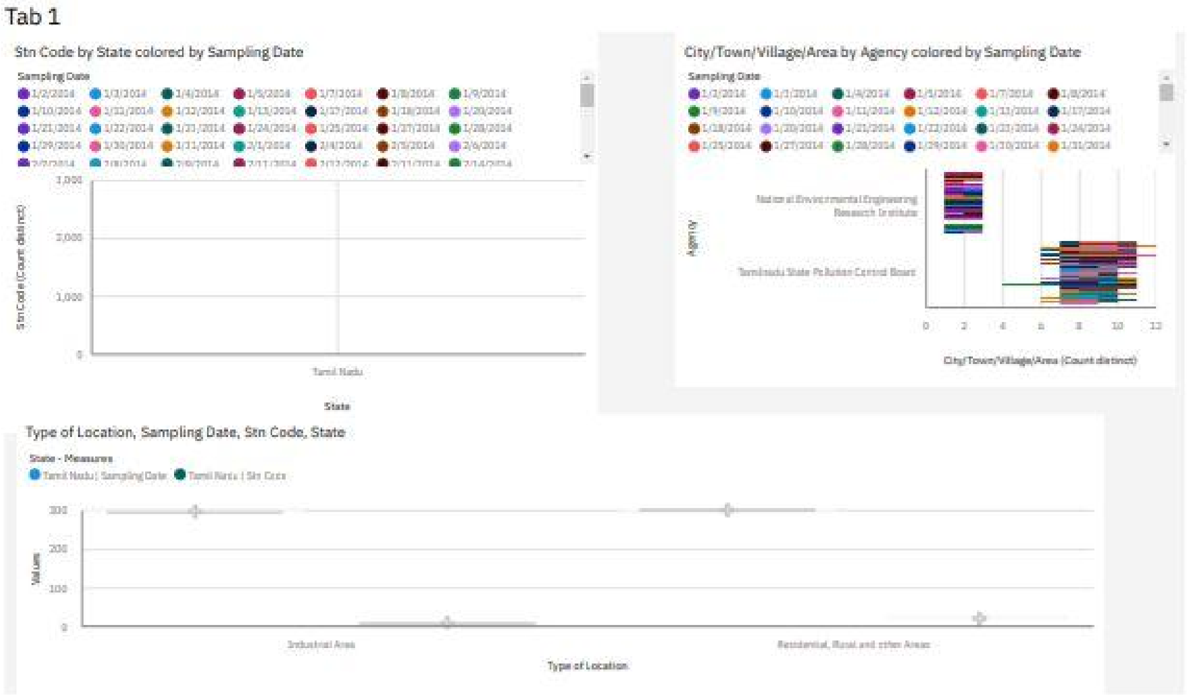
Introduce IBM Cognos as a tool for data analytics.

II. Data Exploration

Showcase how IBM Cognos aids in exploring and understanding the dataset.

III. Visualization

Demonstrate the creation of visualizations in IBM Cognos.


Tab 1
Stn Code by State colored by Sampling Date
Sampling Date
3,000
2,000
1,000
0
Stn Code (Count distinct)
Tamil Nadu
State
City/Town/Village/Area by Agency colored by Sampling Date
Sampling Date
National Environmental Engineering Research Institute
Tamilnadu State Pollution Control Board
Agency
0
2
4
6
8
10
12
City/Town/Village/Area (Count distinct)
Type of Location, Sampling Date, Stn Code, State
State - Measures
Tamil Nadu | Sampling Date
Tamil Nadu | Stn Code
300
200
100
0
Values
Industrial Area
Residential, Rural and other Areas
Type of Location

DATA VISUALIZATION WITH JUPYTER NOTEBOOK

i. Jupyter Notebook Introduction

Present Jupyter Notebook as a tool for data analysis and visualization.

ii. Visualizing Air Quality Parameters

Use Jupyter Notebook to create visualizations of air quality parameters.

iii. Geographic Mapping

Visualize air quality by location using Jupyter Notebook.

iv. Time Series Analysis

Analyze temporal changes in water quality using Jupyter Notebook.

8. Predictive Modelling

The predictive modelling phase aims to develop a model that can classify air samples as based on the analysed parameters. This phase includes the following steps:

8.1 Data Splitting

- ❖ Divide the dataset into a training set and a testing set for model training and evaluation.

8.2 Model Selection

- ❖ Choose appropriate machine learning algorithms for binary classification.
- ❖ Evaluate multiple models to select the most suitable one.

8.3 Model Training

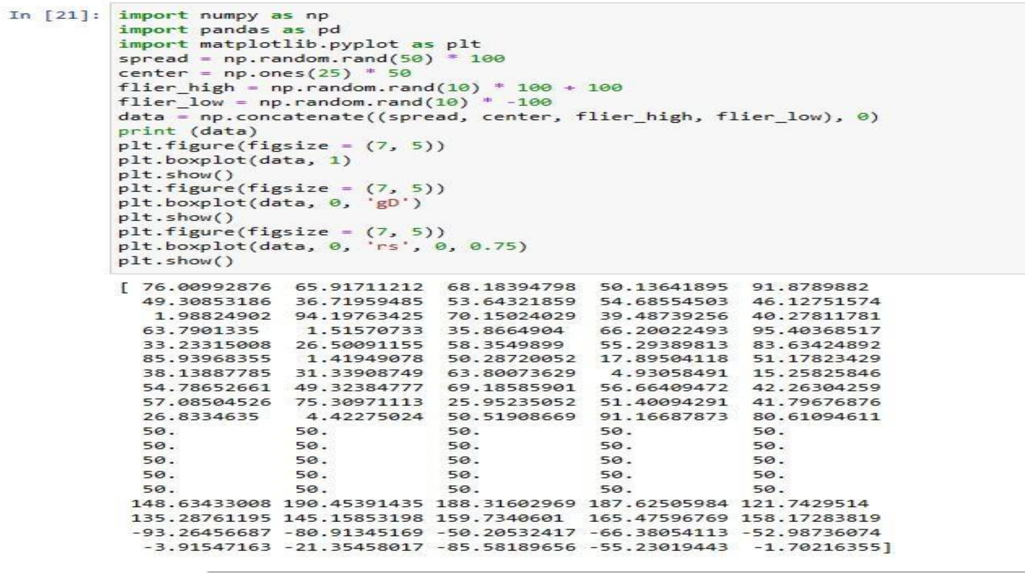
- ❖ Train the selected model on the training data using the water quality parameters as features as the target variable.

8.4 Model Evaluation

- ❖ Assess the model's performance on the testing set using relevant metrics such as accuracy, precision, recall, F1-score, and the ROC curve.

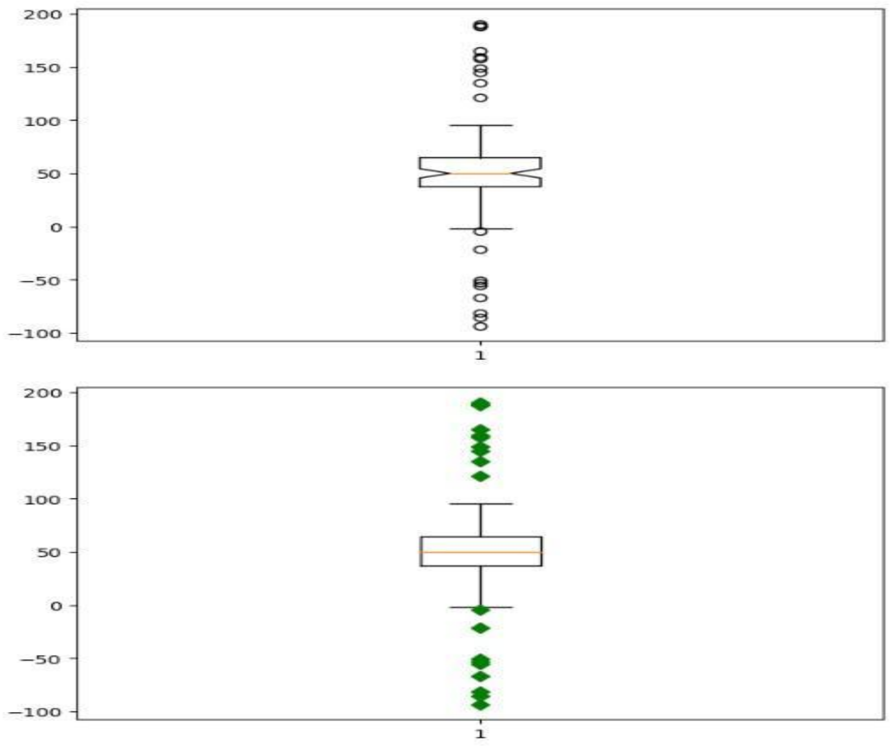
8.5 Model Optimization

- ❖ Fine-tune the model parameters, perform feature selection, and optimize hyperparameters to improve model accuracy.



```
In [21]: import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         spread = np.random.rand(50) * 100
         center = np.ones(25) * 50
         flier_high = np.random.rand(10) * 100 + 100
         flier_low = np.random.rand(10) * -100
         data = np.concatenate((spread, center, flier_high, flier_low), 0)
         print (data)
         plt.figure(figsize = (7, 5))
         plt.boxplot(data, 1)
         plt.show()
         plt.figure(figsize = (7, 5))
         plt.boxplot(data, 0, 'gD')
         plt.show()
         plt.figure(figsize = (7, 5))
         plt.boxplot(data, 0, 'rs', 0, 0.75)
         plt.show()
```

```
[ 76.00992876  65.91711212  68.18394798  50.13641895  91.8789882
  49.30853186  36.71959485  53.64321859  54.68554503  46.12751574
   1.98824902  94.19763425  70.15024029  39.48739256  40.27811781
  63.7901335    1.51570733  35.8664904   66.20022493  95.40368517
  33.23315008  26.50091155  58.3549899   55.29389813  83.63424892
  85.93968355   1.41949078  50.28720052  17.89504118  51.17823429
  38.13887785  31.33908749  63.80073629   4.93058491  15.25825846
  54.78652661  49.32384777  69.18585901  56.66409472  42.26304259
  57.08504526  75.30971113  25.95235052  51.40094291  41.79676876
  26.8334635    4.42275024  50.51908669  91.16687873  80.61094611
  50.          50.          50.          50.          50.
  50.          50.          50.          50.          50.
  50.          50.          50.          50.          50.
  50.          50.          50.          50.          50.
  50.          50.          50.          50.          50.
 148.63433008 190.45391435 188.31602969 187.62505984 121.7429514
 135.28761195 145.15853198 159.7340601  165.47596769 158.17283819
 -93.26456687 -80.91345169 -50.20532417 -66.38054113 -52.98736074
  -3.91547163 -21.35458017 -85.58189656 -55.23019443  -1.70216355]
```



9. Insights from Analysis

The insights derived from the analysis will provide valuable information for assessing air quality . Here are some of the key insights that can be obtained:

9.1 Identification of Critical Parameters

❖ Determine which air quality parameters have the most significant impact on pollutants.

9.2 Pattern Recognition

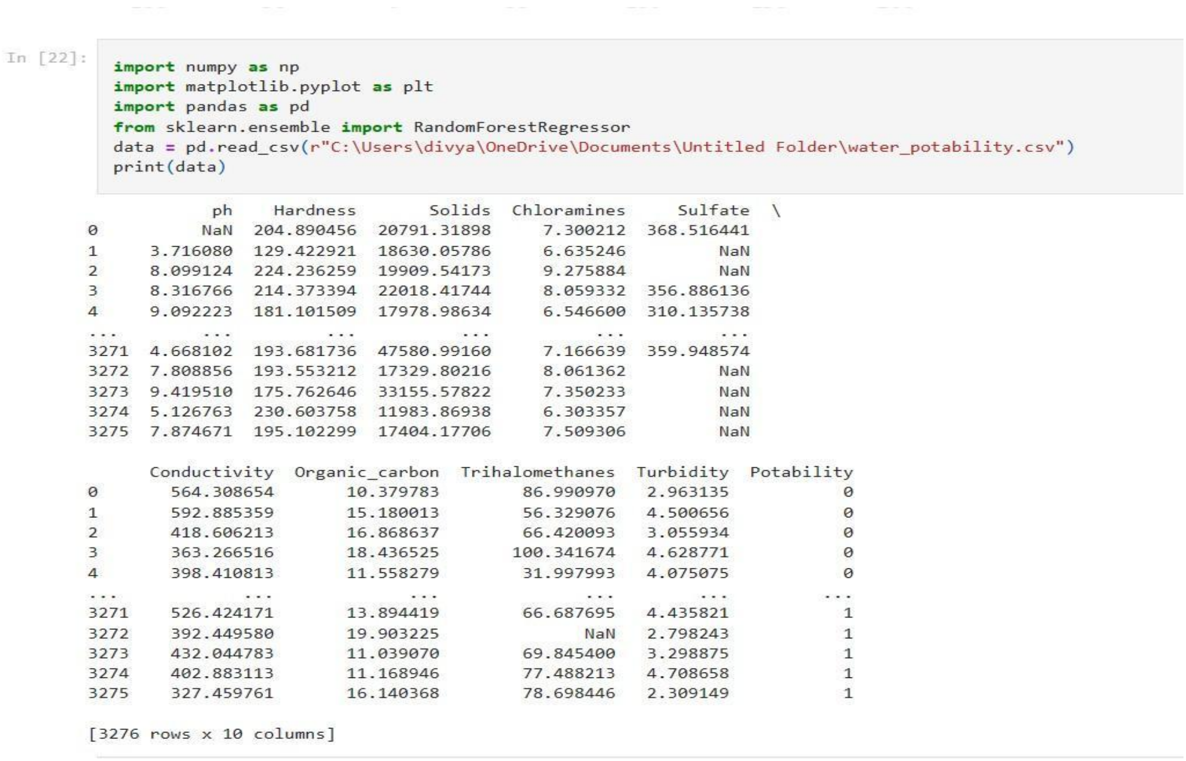
❖ Discover patterns and trends in the data that may indicate specific factors affecting air quality.

9.3 Data-Driven Recommendations

❖ Provide data-driven recommendations for improving air quality based on the identified factors.

9.4 Decision Support

❖ Offer decision support tools for stakeholders, such as air treatment plants or regulatory authorities, to make informed decisions about air quality management.



```
In [22]:
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
data = pd.read_csv(r"C:\Users\divya\OneDrive\Documents\Untitled Folder\water_potability.csv")
print(data)
```

```
            ph    Hardness        Solids  Chloramines     Sulfate  \
0          NaN  204.890456  20791.31898     7.300212  368.516441
1     3.716080  129.422921  18630.05786     6.635246         NaN
2     8.099124  224.236259  19909.54173     9.275884         NaN
3     8.316766  214.373394  22018.41744     8.059332  356.886136
4     9.092223  181.101509  17978.98634     6.546600  310.135738
...        ...         ...          ...          ...         ...
3271  4.668102  193.681736  47580.99160     7.166639  359.948574
3272  7.808856  193.553212  17329.80216     8.061362         NaN
3273  9.419510  175.762646  33155.57822     7.350233         NaN
3274  5.126763  230.603758  11983.86938     6.303357         NaN
3275  7.874671  195.102299  17404.17706     7.509306         NaN

      Conductivity  Organic_carbon  Trihalomethanes  Turbidity  Potability
0       564.308654       10.379783        86.990970   2.963135           0
1       592.885359       15.180013        56.329076   4.500656           0
2       418.606213       16.868637        66.420093   3.055934           0
3       363.266516       18.436525       100.341674   4.628771           0
4       398.410813       11.558279        31.997993   4.075075           0
...            ...             ...              ...        ...         ...
3271    526.424171       13.894419        66.687695   4.435821           1
3272    392.449580       19.903225              NaN   2.798243           1
3273    432.044783       11.039070        69.845400   3.298875           1
3274    402.883113       11.168946        77.488213   4.708658           1
3275    327.459761       16.140368        78.698446   2.309149           1

[3276 rows x 10 columns]
```

Conclusion

This project's objective is to analyze air quality data and predict pollutants using a dataset, following the design thinking process and various development phases. By conducting data preprocessing, exploratory data analysis, data visualization, and predictive modelling, we aim to provide valuable insights for assessing air quality . The insights obtained from this analysis can have a significant impact on air management, public health, and environmental conservation.

In summary, this comprehensive analysis will not only assess the quality of air but also empower decision-makers with the tools and knowledge needed to safeguard and enhance the safety of air sources. air quality analysis and pollutants prediction play a vital role in ensuring access to clean and pure air, a fundamental human right.

LINK FOR JUPYTER NOTEBOOK (ipynb) :
https://github.com/HarshiniSivakumar30/AirQassesment.git

LINK FOR JUPYTER NOTEBOOK (pdf) :
https://github.com/HarshiniSivakumar30/AirQassesment.git

LINK FOR IBM COGNOS VISUALIZATION (pdf) :
https://github.com/HarshiniSivakumar30/AirQassesment.git