Data Analytics with Cognos

Project: Air Quality Assessment Tamil Nadu

Phase 2 Project

Table of Contents

1. Introduction:

Chennai beats Delhi in pollution, records 'very poor' air quality

Velachery, Ramapuram, Manali, Kodungaiyur, Anna Nagar, Chennai Airport clocked pollution levels as high as 341, while Delhi stood at 254.

Published: 07th November 2019 11:51 AM | Last Updated: 07th November 2019 09:43 PM

A+ A A-

Many areas in the Tamil Nadu capital recorded 'very poor' air quality. | (Photo | Martin Louis/EPS)

Air quality is a critical concern for public health and environmental well-being.

Tamil Nadu have PM 2.5 of 28.2 microgram/m3 levels. In Tamil Nadu, as in many other regions, understanding and predicting ambient air quality is essential for informed decision-making and effective pollution control measures. This document outlines a design to improve the accuracy of predictive models for ambient air quality in Tamil Nadu using machine learning algorithms.

2. Problem Statement:

The objective is to incorporate a machine learning algorithm to improve the accuracy of the predictive model. The model should provide reliable forecasts for different air quality parameters such as PM10, NO2 and SO2.

3. Data Collection:

The dataset containing location-wise daily ambient air quality records for Tamil Nadu in the year 2014 has been obtained from the below datalink.

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

Data set:

	A	B	C	D	E	F	G	H	I	J	K
1	Stn Code	Sampling D	State	City/Town	Location o	Agency	Type of Lo	SO2	NO2	RSPM/PM	PM 2.5
2	38	1/2/2014	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	11	17	55	NA
3	38	1/7/2014	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	13	17	45	NA
4	38	21-01-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	12	18	50	NA
5	38	23-01-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	15	16	46	NA
6	38	28-01-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	13	14	42	NA
7	38	30-01-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	14	18	43	NA
8	38	2/4/2014	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	12	17	51	NA
9	38	2/6/2014	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	13	16	46	NA
10	38	########	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	10	19	50	NA
11	38	13-02-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	15	14	48	NA
12	38	18-02-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	14	16	32	NA
13	38	20-02-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	14	14	29	NA
14	38	25-02-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	13	17	17	NA
15	38	27-02-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	15	16	44	NA
16	38	3/4/2014	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	12	17	25	NA
17	38	3/6/2014	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	13	16	29	NA
18	38	########	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	11	18	29	NA
19	38	13-03-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	15	16	41	NA
20	38	18-03-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	14	17	43	NA
21	38	20-03-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	14	14	42	NA
22	38	25-03-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	14	17	54	NA
23	38	27-03-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	15	19	62	NA
24	38	4/1/2014	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	14	15	66	NA
25	38	4/3/2014	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	11	16	40	NA
26	38	4/8/2014	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	14	17	56	NA
27	38	########	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	15	17	50	NA
28	38	15-04-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	12	14	49	NA
29	38	17-04-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	15	16	63	NA
30	38	22-04-14	Tamil Nad	Chennai	Kathivakka	Tamilnadu	Industrial A	15	18	42	NA

cpcb-dly-aq-tamil-nadu-2014 (2)

4. Data Cleansing:

Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model.

Data pre-processing is a crucial step in preparing the dataset for machine learning.

Data Cleansing involves:

- Identifying and removing any missing, duplicate or irrelevant data.
- Handling missing data.
- Removing outliers.

5. Feature Engineering:

Feature engineering involves creating new features or modifying existing ones to improve model performance. In this context, it may involve generating lag features, aggregating data over time intervals, or incorporating weather data if available to capture external factors influencing air quality.

6. Machine Learning Model Selection:

Select appropriate machine learning algorithms for the task. Potential models include:

- Regression models (e.g., Linear Regression, Random Forest Regression).
- Time series forecasting models (e.g., ARIMA, LSTM).
- Ensemble methods for improved accuracy.

7. Model Training and Evaluation:

Split the dataset into training and testing sets. Train the selected models on the training data and evaluate their performance using suitable metrics (e.g., Mean Absolute Error, R-squared). Perform hyper parameter tuning to optimize model performance.

8. Program:

- First, import the required Python libraries. As given below.

```
# In[1]: import pandas as pd
         from sklearn.linear_model import LinearRegression
         from sklearn.model_selection import train_test_split
         from sklearn.metrics import mean_squared_error,r2_score
         import matplotlib.pyplot as plt
         from sklearn import datasets
         import numpy as np
```

- To import the database into the jupyter notebook, you can use the following Python code given below.

```
# In[2]:data =pd.read_csv(r"C:\Users\Administrator\Downloads\airquality.csv")


# In[3]:print(data.head())
```

 # OP [3]:

```
   Stn  Code Sampling Date        State City/Town/Village/Area  \
0         38      01-02-14  Tamil Nadu                 Chennai
1         38      01-07-14  Tamil Nadu                 Chennai
2         38      21-01-14  Tamil Nadu                 Chennai
3         38      23-01-14  Tamil Nadu                 Chennai
4         38      28-01-14  Tamil Nadu                 Chennai

                     Location of Monitoring Station  \
0  Kathivakkam, Municipal Kalyana Mandapam, Chennai
1  Kathivakkam, Municipal Kalyana Mandapam, Chennai
2  Kathivakkam, Municipal Kalyana Mandapam, Chennai
3  Kathivakkam, Municipal Kalyana Mandapam, Chennai
```

```
4  Kathivakkam, Municipal Kalyana Mandapam, Chennai

                                    Agency Type of Location   SO2   NO2  \
0  Tamilnadu State Pollution Control Board  Industrial Area  11.0  17.0
1  Tamilnadu State Pollution Control Board  Industrial Area  13.0  17.0
2  Tamilnadu State Pollution Control Board  Industrial Area  12.0  18.0
3  Tamilnadu State Pollution Control Board  Industrial Area  15.0  16.0
4  Tamilnadu State Pollution Control Board  Industrial Area  13.0  14.0

   RSPM/PM10  PM 2.5
0       55.0     NaN
1       45.0     NaN
2       50.0     NaN
3       46.0     NaN
4       42.0     NaN
```

- Data Cleansing and Transformation are done by following Python Code.

  Using Drop()- To drop an entire Column.

  Using dropna()- To drop a the NaN Values.

  Using drop_duplicates()- To drop the Duplicates in the Dataset.

```
# In[3]: df=data.drop(['PM 2.5'],axis=1)
# In[4]: df.head()

# OP[4]:
```

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10
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
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
2	38	21-01-14	Tamil	Chennai	Kathivakkam, Municipal	Tamilnadu State	Industrial Area	12.0	18.0	50.0

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10
			Nadu		Kalyana Mandapam, Chennai	Pollution Control Board				
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
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

```
# In[5]: newf=df.dropna()
         newd=newf.drop_duplicates()

# In[6]: newf.head()

# OP[6]:
```

Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10	
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
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

Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10	
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
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
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

- Split the data for Training and Testing and train the model for Linear Regression.

```
# In[7]: x = newf[['NO2','SO2']]


# In[8]: y = newf['RSPM/PM10']


# In[9]:x_train,x_test,y_train,y_test =
         train_test_split(x,y,test_size=0.3,random_state=0)


# In[8]: model = LinearRegression()
```

```
# OP[8]:  LinearRegression

            LinearRegression()
```

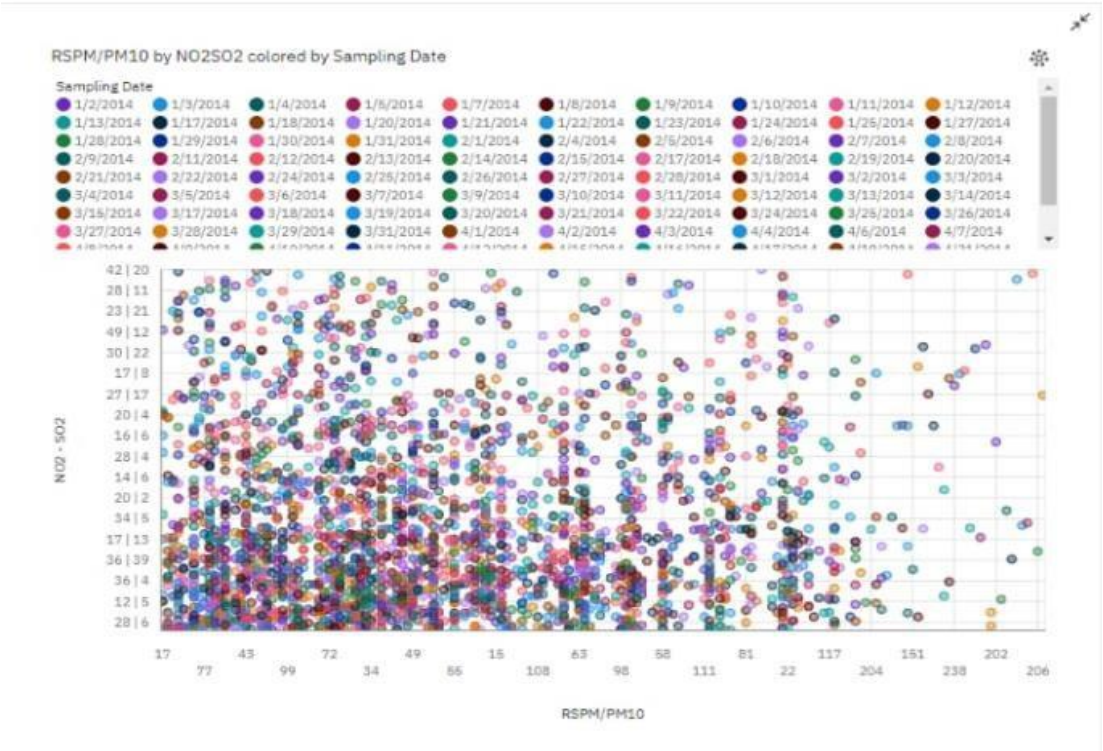
```
# In[9]: model.fit(x_train,y_train)


# In[10]: y_pred = model.predict(x_test)
```

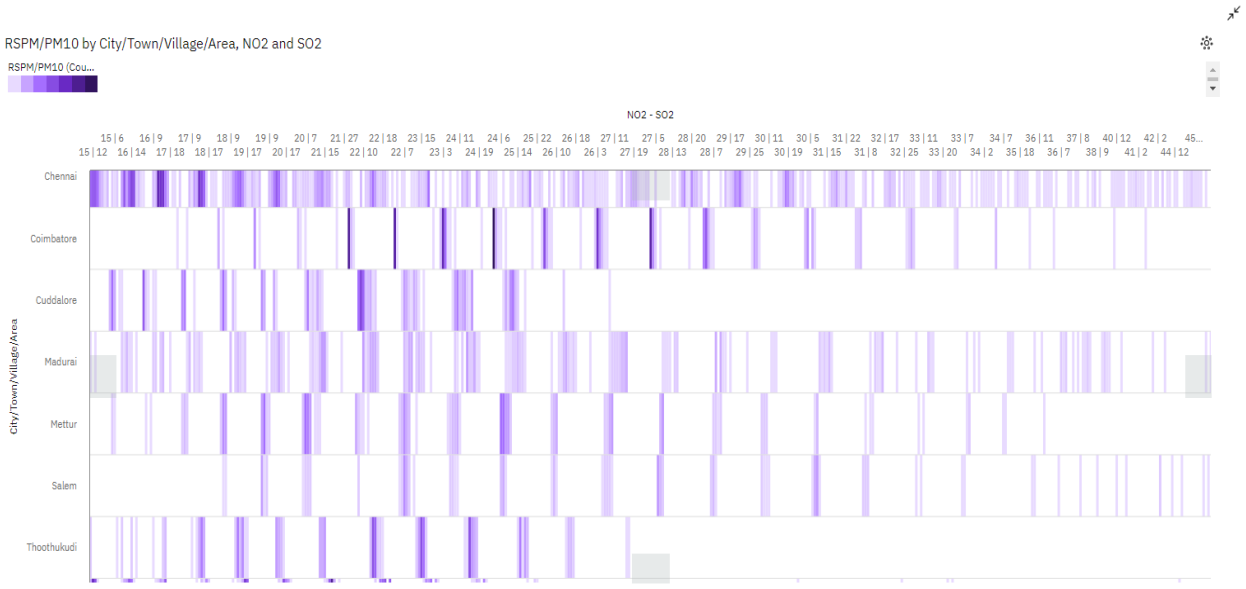
- Evaluate the Model using Mean Squared Error and R2 score.

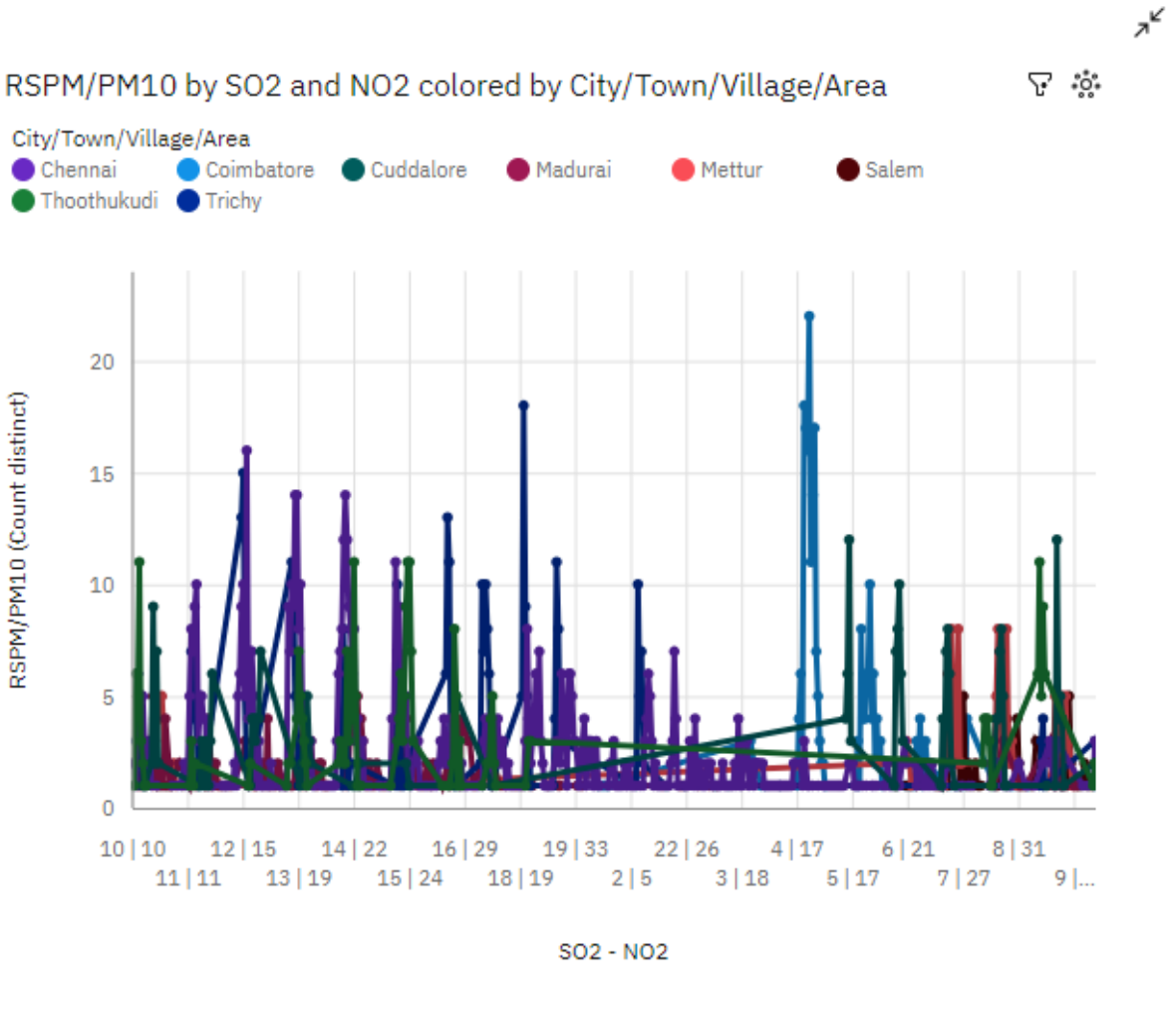
```
# In[11]: mse = mean_squared_error(y_test,y_pred)


# In[12]: r2 = r2_score(y_test,y_pred)


# In[13]: print("Mean Squared Error:",mse)
           print("R-squared:",r2)

# OP[13]: Mean Squared Error: 908.4528649741137
          R-squared: 0.19877081345863346
```

- Data Visualization can be done by IBM Cognos

The Below Data Visualization is to visualize the NO2 and SO2 using IBM Cognos.




RSPM/PM10 by City/Town/Village/Area, NO2 and SO2
RSPM/PM10 (Cou...
NO2 - SO2
City/Town/Village/Area
Chennai
Coimbatore
Cuddalore
Madurai
Mettur
Salem
Thoothukudi


RSPM/PM10 by SO2 and NO2 colored by City/Town/Village/Area
City/Town/Village/Area
Chennai
Coimbatore
Cuddalore
Madurai
Mettur
Salem
Thoothukudi
Trichy
RSPM/PM10 (Count distinct)
0
5
10
15
20
10 | 10
11 | 11
12 | 15
13 | 19
14 | 22
15 | 24
16 | 29
18 | 19
19 | 33
2 | 5
22 | 26
3 | 18
4 | 17
5 | 17
6 | 21
7 | 27
8 | 31
9 |...
SO2 - NO2

9. Conclusion:

The proposed approach aims to enhance the accuracy of predictive models for ambient air quality in Tamil Nadu through the incorporation of machine learning algorithms. The success of this project will lead to better air quality predictions, enabling more effective pollution control measures and safeguarding public health and the environment.