

**A PROJECT REPORT ON**  
**AIR QUALITY ANALYSIS IN TAMIL NADU**  
**DOMAIN: Data analytics with Cognos**  
**IBM – DOCUMENTATION UNDER THE GUIDANCE**  
**OF**  
**Faculty Mentor(s) Name : Er. V. Sudha**

**SUBMITTED BY:**

P. Paul Raj

421321106033



**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING**  
**KRISHNASAMY COLLEGE OF ENGINEERING & TECHNOLOGY**  
**ANNA UNIVERSITY: 2021-2025**

## **BONAFIDE CERTIFICATE**

Certified this project report “**AIR QUALITY ANALYSIS IN TAMILNADU**” is the bonfide work of **P. Paul Raj**(421321106033) who carried out the project work under my supervision.

### **SIGNATURE**

#### **HEAD OF THE DEPARTMENT**

Er. S. Senthazhai M.E.,

#### **Associate Professor**

Electronics & Communication Engineering

Krishnasamy College of Engineering &

Technology,

Cuddalore – 607109.

### **SIGNATURE**

#### **MENTOR & EVALUATOR**

Er. V. Sudha, M.E.,

#### **Assistant Professor**

Electronics & Communication Engineering

Krishnasamy College of Engineering &

Technology,

Cuddalore – 607109.

<b>S.NO</b>	<b>TABLE OF CONTENTS</b>
<b>1</b>	<b>INTRODUCTION</b>
1.1	Project Overview
1.2	Purpose
<b>2</b>	<b>PROBLEM STATEMENT AND PROPOSED SOLUTION</b>
2.1	Problem Statement Definition
2.2	Proposed Solution
<b>3</b>	<b>REQUIREMENTS</b>
<b>4</b>	<b>DEVELOPMENT PART-1</b>
4.1	Loading the dataset
4.2	Exploratory Data Analysis (EDA)
4.3	Data Cleaning and Transformation
<b>5</b>	<b>DEVELOPMENT PART-2</b>
5.1	Feature Engineering
5.2	Model Selection and Training
5.3	Model Comparison

---

<b>6</b>	<b>DATA VISUALIZATION</b>
<b>7</b>	<b>ADVANTAGES AND DISADVANTAGES</b>
<b>8</b>	<b>FUTURE SCOPE</b>
<b>9</b>	<b>CONCLUSION</b>
<b>10</b>	<b>APPENDIX</b>

---



# AIR QUALITY ANALYSIS IN TAMILNADU

## 1. INTRODUCTION

### 1.1 Project overview

The “Air Quality Analysis in Tamil Nadu” project is a comprehensive initiative aimed at analyzing and Visualizing air quality data from diverse monitoring stations across Tamil Nadu. This endeavor is driven by the critical need to address air pollution issues, which directly impact the health and well-being of the State’s residents. The project revolves around assessing the levels of Sulfur Dioxide (SO<sub>2</sub>), Nitrogen Dioxide (NO<sub>2</sub>), and Respirable Suspended Particulate Matter/Particulate Matter 10 (RSPM/PM<sub>10</sub>) in Various urban, suburban, and rural regions of Tamil Nadu.

### 1.2. Purpose:

Clean air is a fundamental human right, and it has a direct impact on public health and the environment. This project is crucial for several reasons:

- **Health:** Poor air quality can lead to a range of health issues, including respiratory problems, cardiovascular diseases, and even premature death. Understanding pollution patterns can lead to healthier lives.
- **Environmental Impact:** Air pollution can harm ecosystems, damage buildings and infrastructure, and contribute to climate change. This project contributes to a greener and more sustainable Tamil Nadu.
- **Policy and Decision-Making:** The project results will assist policymakers in making informed decisions regarding pollution control measures and urban planning.

## 2. PROBLEM STATEMENT& PROPOSED SOLUTION

### 2.1. Problem Statements Definition:

The problem at hand is the pressing issue of air pollution in Tamil Nadu, which poses a substantial threat to public health and the environment. The region is experiencing elevated levels of Sulfur Dioxide (SO<sub>2</sub>), Nitrogen Dioxide (NO<sub>2</sub>), and RSPM/PM 10 matter in the Various cities, towns, villages, and areas. This pollution problem necessitates a comprehensive analysis and proactive management approach to mitigate its adverse effects. The primary problem statement can be defined as follows:

High levels of SO<sub>2</sub>, NO<sub>2</sub>, and RSPM/PM10 pollutants in Tamil Nadu have led to compromised air quality, endangering the health of its residents and causing environmental degradation. The challenge is to systematically analyze air quality data, identify pollution trends, pinpoint pollution hotspots, and develop a predictive model to support evidence-based decision-making and safeguard the well-being of the population and the ecosystem. This problem statement underscores the urgency of addressing air pollution in Tamil Nadu and sets the stage for the objectives and actions required to tackle this critical issue.

### 2.2Proposed Solution

S.NO	PARAMETERS	DESCRIPTION
01)	Problem Statement ( Problem to be solved)	This problem statement underscores the urgency of addressing air pollution in Tamil Nadu and sets the stage for the objectives and actions required to tackle this critical issue.

---

02)	Idea / Solution Description	This project aims to analyze and predict air quality in Tamil Nadu by assessing SO <sub>2</sub> , NO <sub>2</sub> , and RSPM/PM <sub>10</sub> levels. It utilizes data analysis, trend identification, and predictive modeling.
03)	Novelty / Uniqueness	The project's uniqueness lies in its multi-parameter approach, predictive modeling, and localized focus on Tamil Nadu.
04)	Social Impact / Customer Satisfaction	It improves public health, quality of life, and environmental sustainability while aiding decision-making.
05)	Business Model ( Revenue Model)	Collaboration with government, data services, consultancy, and grant-seeking form the business model.
06)	Scalability	The project is scalable to other regions, additional pollutants, public awareness campaigns, and Technology integration.

---



### 3. REQUIREMENTS

#### 1. Software Requirement:

The software requirement involves a capable computer or server to execute the data analysis and modeling tasks. Additionally, it encompasses the need for Python (version 3.x) to be installed for data analysis, along with Jupyter Notebook for code development. This software environment serves as the foundation for the entire project.

#### 2. Jupyter Notebook:

Jupyter Notebook is an integral part of the project, serving as the development environment for coding, data exploration, and documentation. It provides an interactive platform for creating and sharing code, making it an essential tool for the project team.



#### 3. Libraries for Python:

The project relies on essential Python libraries, including Pandas for data manipulation, Numpy for numerical operations, and Matplotlib for data visualization. These libraries are the building blocks of data analysis and enable the team to process, analyze, and visualize air quality data effectively.



#### **4.Scikit-learn for Model Building:**

Scikit-learn is utilized for developing machine learning models. It offers a wide range of tools for data preprocessing, model training, and evaluation. This requirement is fundamental for building predictive models to forecast air quality.



#### **5.IBM Cognos for Visualization:**

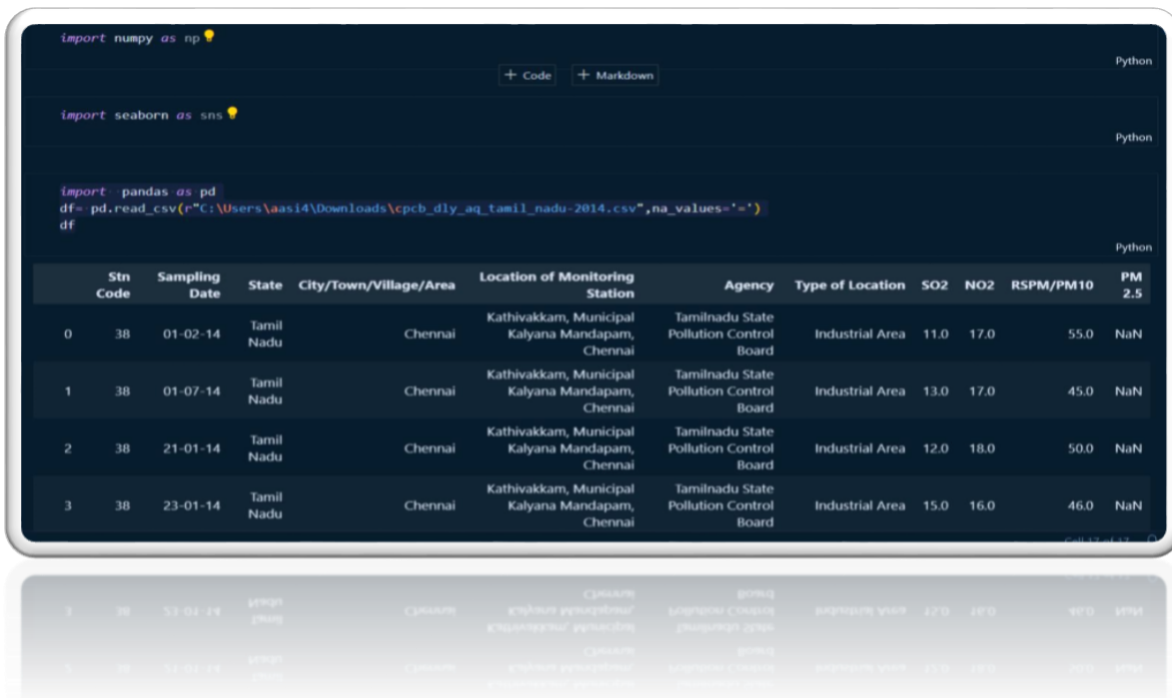
IBM Cognos serves as the platform for data visualization. It enables the creation of informative dashboards and reports to present air quality insights. This requirement is critical for communicating the project's findings in a clear and accessible manner to stakeholders and the public.



## 4.DEVELOPMENT PART-1

### 4.1 Loading the dataset:

We import the necessary Python libraries, including pandas for data Manipulation, pandas is a common library used in data analysis and Jupyter Notebook Environments. If you have ‘pandas’ installed and are using it in your Jupyter Notebook, upgrading ‘nbformat’ is an independent step to ensure that you can render content properly, such as plots or Visualizations, which might be related to other libraries like ‘matplotlib’ or ‘plotly.’ Once pandas is imported, you can load your dataset. You typically do this by providing the path to the dataset file (usually a CSV file) .in a CSV file, into a pandas DataFrame. Replace ‘‘your\_dataset.csv’’ with the actual file path of your dataset is the name of the pandas DataFrame that will hold your dataset. `pd.read_csv()` is a pandas function designed to read CSV Files and load them into a DataFrame. ‘‘my\_dataset.csv’’ should be replaced with the actual file Path or URL of your dataset.



```
import numpy as np

import seaborn as sns

import pandas as pd
df= pd.read_csv(r"C:\Users\Aasi4\Downloads\cpcb_dly_aq_tamil_nadu-2014.csv",na_values='')
df
```

	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

## 4.2 Exploratory Data Analysis (EDA):

**Exploring the Loaded Dataset:**After loading the dataset, it's a good practice to explore it and get a better understanding of its structure. You can use various pandas functions to achieve this

**Display the First Few Rows:**You can use `df.head()` to display the first few rows of your dataset. This helps you get an initial sense of the data's content.

**Check Column Names and Data Types:**Use `df.info()` to check the column names, data types, and non-null counts for each column. This is useful for understanding the dataset's structure.

**Check for Missing Values:**To identify missing values in your dataset, use `df.isnull().sum()`. This will show the count of missing values in each column.By loading and exploring your dataset, you set the foundation for data analysis, cleaning, and manipulation. Understanding the structure and content of your data is essential for making informed decisions and preparing it for further analysis.

```
df.info()

[4]

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2879 entries, 0 to 2878
Data columns (total 11 columns):
 #   Column                                Non-Null Count  Dtype  
---  -
 0   Stn Code                             2879 non-null  int64  
 1   Sampling Date                        2879 non-null  object  
 2   State                                2879 non-null  object  
 3   City/Town/Village/Area              2879 non-null  object  
 4   Location of Monitoring Station       2879 non-null  object  
 5   Agency                              2879 non-null  object  
 6   Type of Location                    2879 non-null  object  
 7   SO2                                 2868 non-null  float64 
 8   NO2                                 2866 non-null  float64 
 9   RSPM/PM10                          2875 non-null  float64 
10  PM 2.5                             0 non-null     float64 
dtypes: float64(4), int64(1), object(6)
memory usage: 247.5+ KB
```

```
df.isnull().sum()
```


```
[5]
... Stn Code      0
    Sampling Date  0
    State         0
    City/Town/Village/Area  0
    Location of Monitoring Station  0
    Agency         0
    Type of Location  0
    SO2           11
    NO2           13
    RSPM/PM10     4
    PM 2.5       2879
    dtype: int64
```

```
df.head(6)
```

```
[6]
...
   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      01-03-14  Tamil Nadu  Chennai  Kathivakkam, Municipal Kalyana Mandapam, Chennai  Tamilnadu State Pollution Control Board  Industrial Area  11.0  17.0      42.0  NaN
3      38      01-05-14  Tamil Nadu  Chennai  Kathivakkam, Municipal Kalyana Mandapam, Chennai  Tamilnadu State Pollution Control Board  Industrial Area  11.0  17.0      22.0  NaN
4      38      01-05-14  Tamil Nadu  Chennai  Kathivakkam, Municipal Kalyana Mandapam, Chennai  Tamilnadu State Pollution Control Board  Industrial Area  11.0  17.0      22.0  NaN
5      38      01-05-14  Tamil Nadu  Chennai  Kathivakkam, Municipal Kalyana Mandapam, Chennai  Tamilnadu State Pollution Control Board  Industrial Area  11.0  17.0      22.0  NaN
```

### 4.3 Data Cleaning and Transformation:

If there are missing values in your dataset, you'll need to decide how to handle them. Common Strategies include removing rows with missing values, filling them with mean or median values, or using more advanced imputation techniques. Here's an example of how to fill missing values which is shown below .Depending on your dataset, you may need to perform additional data cleaning and Transformation. For example, converting date and time columns to date time objects, dropping Irrelevant columns, or encoding categorical variables. Depending on your dataset, you may need to perform additional data cleaning and Transformation. For example, converting date and time columns to date time objects, dropping Irrelevant columns, or encoding categorical variables.



```
df['SO2'].fillna(0, inplace=True)
df['NO2'].fillna(0, inplace=True)
df['RSPM/PM10'].fillna(0, inplace=True)
missing_values = df.isna().sum()
print(missing_values)

[21]
```

```
City/Town/Village/Area    0
SO2                      0
NO2                      0
RSPM/PM10                0
dtype: int64
```

```
dist = (df['RSPM/PM10'])
distset = set(dist)
dd = list(distset)
dictofwords = {dd[i]:i for i in range (0,len(dd))}
df["RSPM/PM10"] = df["RSPM/PM10"].map(dictofwords)
```

```
columns_to_drop = ['Stn Code', 'Sampling Date', 'State', 'Location of Monitoring Station', 'Agency', 'Type of Location']
df = df.drop(columns_to_drop, axis=1)
df
```

[3] ✓ 0.0s Python

	City/Town/Village/Area	SO2	NO2	RSPM/PM10	PM 2.5
0	Chennai	11.0	17.0	55.0	NaN
1	Chennai	13.0	17.0	45.0	NaN
2	Chennai	12.0	18.0	50.0	NaN
3	Chennai	15.0	16.0	46.0	NaN
4	Chennai	13.0	14.0	42.0	NaN
...	...	...	...	...	...
2874	Trichy	15.0	18.0	102.0	NaN
2875	Trichy	12.0	14.0	91.0	NaN
2876	Trichy	19.0	22.0	100.0	NaN
2877	Trichy	15.0	17.0	95.0	NaN
2878	Trichy	14.0	16.0	94.0	NaN

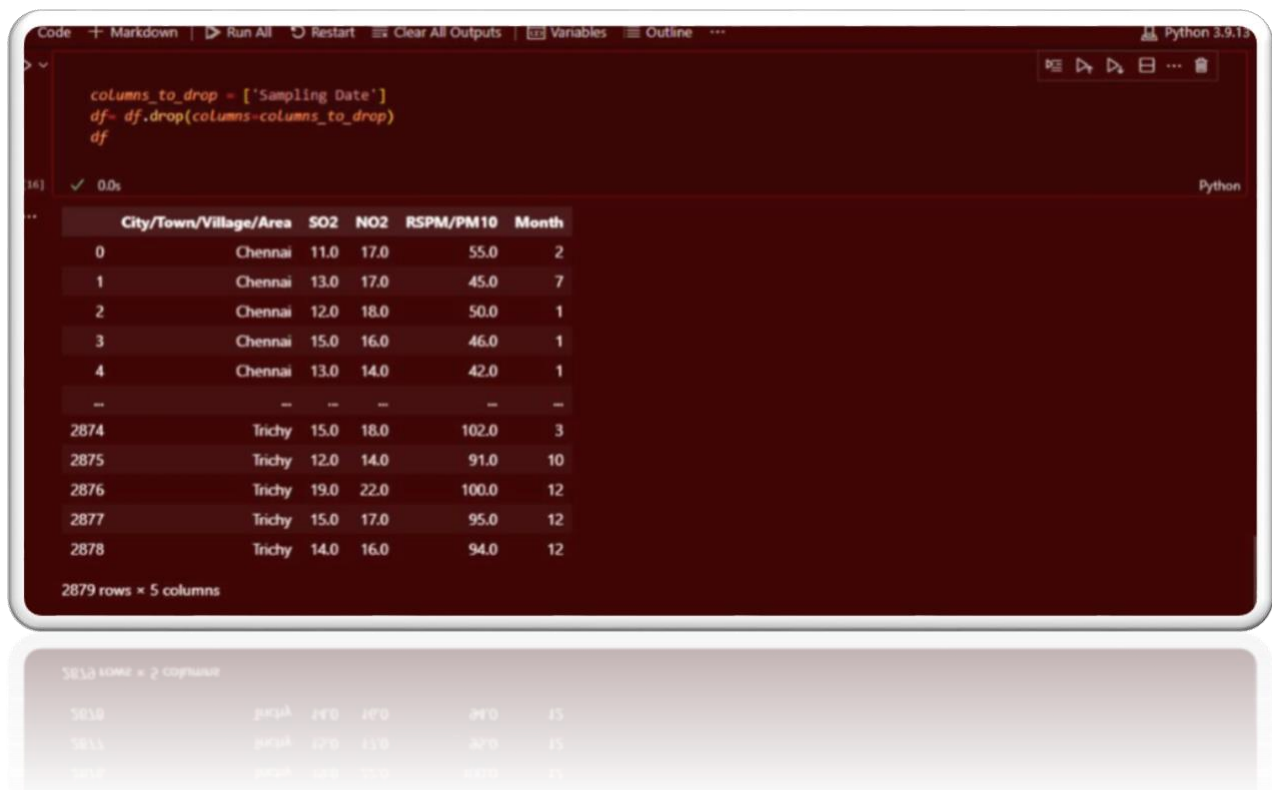
2879 rows x 5 columns

## 5. DEVELOPMENT PART-2

### 5.1 Feature engineering:

- Feature Selection: Decide which features (columns) of your dataset to include in your analysis. You may need to consider feature importance and domain knowledge to make informed choices.
- Feature Extraction: Create new features from existing ones if they can provide valuable information for your analysis.
- Handling Categorical Data: Convert categorical data into a numerical format, often using techniques like one-hot encoding or label encoding.

-



```
columns_to_drop = ['Sampling Date']
df = df.drop(columns=columns_to_drop)
df
```

	City/Town/Village/Area	SO2	NO2	RSPM/PM10	Month
0	Chennai	11.0	17.0	55.0	2
1	Chennai	13.0	17.0	45.0	7
2	Chennai	12.0	18.0	50.0	1
3	Chennai	15.0	16.0	46.0	1
4	Chennai	13.0	14.0	42.0	1
...	...	...	...	...	...
2874	Trichy	15.0	18.0	102.0	3
2875	Trichy	12.0	14.0	91.0	10
2876	Trichy	19.0	22.0	100.0	12
2877	Trichy	15.0	17.0	95.0	12
2878	Trichy	14.0	16.0	94.0	12

2879 rows x 5 columns

## 5.2 Model Selection and Training

Choose the appropriate machine learning models for your specific problem. This could include regression, classification, clustering, or other types of models. Split your data into training and testing sets to assess the performance of your models. Common methods include train-test splitting and cross-validation. Train your models on the training data. This involves fitting the model to the data to learn the underlying patterns. Hyperparameter Tuning: Optimize the model's hyper parameters to achieve the best performance.

```
[12] df['SO2'].fillna(0,inplace=True)
     df['NO2'].fillna(0,inplace=True)
     df['RSPM/PM10'].fillna(0,inplace=True)
Python

> (module) model_selection
#splitting x
y= df['RSPM/PM10']
x=df.drop('RSPM/PM10', axis=1)
#linear regression model
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=100)
from sklearn.linear_model import LinearRegression
lr= LinearRegression()
lr.fit(x_train,y_train)
Python

* LinearRegression
LinearRegression()

x_test
Python

Stn Code  City/Town/Village/Area  Location of Monitoring Station  SO2  NO2
98         71                    5                    5  13.0  19.0
```

```
#Random Forest
from sklearn
rf= RandomForestRegressor()
rf.fit(x_train,y_train)
y_rf_train= rf.predict(x_train)
y_rf_test= rf.predict(x_test)
R^2 (coefficient of determination) regression score function.
Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). In the general case when the true y is
non-constant, a constant model that always predicts the average y
disregarding the input features would get a R^2 score of 0.0.
rf_test_r2= r2_score(y_test,y_rf_test_pred)
#table for rfr
rf_results=pd.DataFrame(['RandomForestRegressor',rf_train_mae,rf_train_r2,rf_test_mae,rf_test_r2]).transpose()
rf_results
rf_results.columns=['ALG','TRAINED MSE','TESTED R2','TRAINED MSE','TESTED R2']
rf_results
```

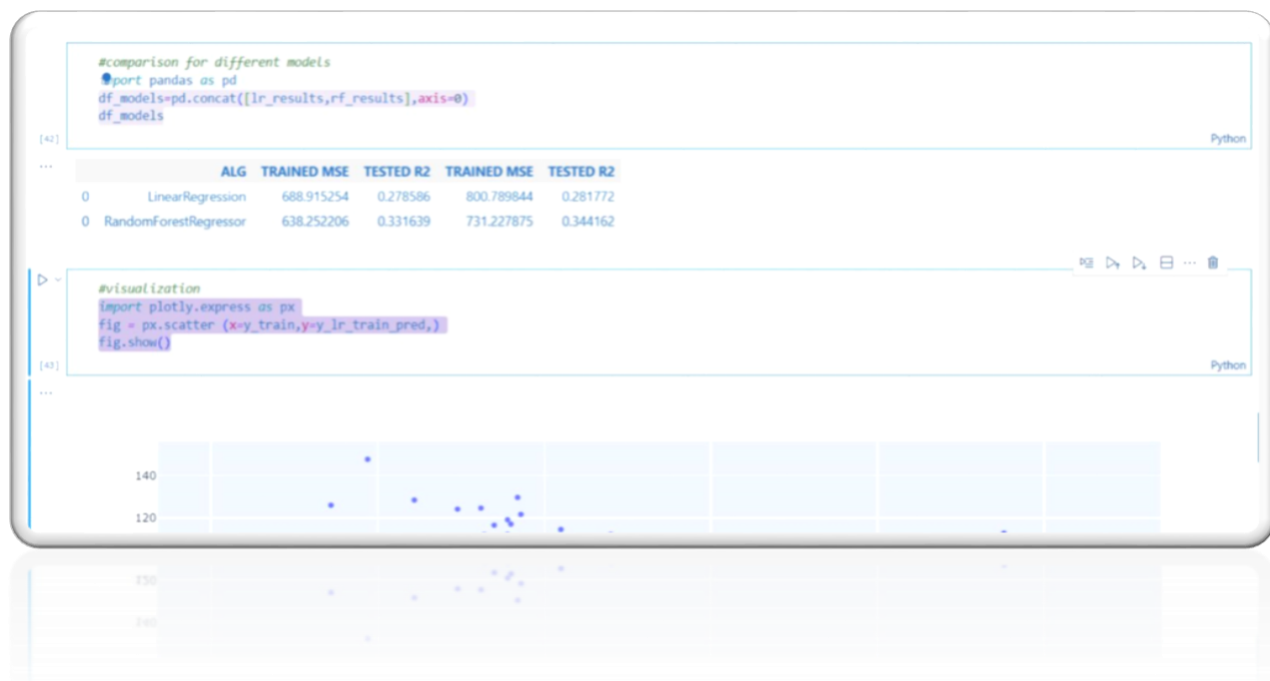
	ALG	TRAINED MSE	TESTED R2	TRAINED MSE	TESTED R2
0	RandomForestRegressor	638.252206	0.331639	731.227875	0.344162

0	gmpgus.org/gdsc	238725306	0711/PM	13/35162	07PM105
	UTC	TRAINED MSE	TESTED R2	TRAINED MSE	TESTED R2



## 5.3 MODEL COMPARISON

Evaluate the performance of your models using appropriate metrics. The choice of metrics depends on the type of problem (classification, regression, etc.). Common evaluation metrics include accuracy, precision, recall, F1-score, mean squared error (MSE), and others. Compare the performance of different models to determine which one works best for Your problem. Make use of model selection techniques to choose the best-performing model.



## 6. DATA VISUALIZATION

Data visualization is a pivotal aspect of the “Air Quality Analysis in Tamil Nadu” project, as it encompasses the art of transforming complex air quality data into clear, accessible, and meaningful visual representations. It plays a crucial role in communicating the project’s findings, trends, and predictive insights to a diverse audience, including policymakers, environmental agencies, and the general public. Through IBM Cognos, the project team creates informative dashboards, reports, and interactive visual elements that provide a visual narrative of air quality patterns and pollution hotspots. These visualizations not only facilitate data-driven decision-making but also engage and empower stakeholders and the community by fostering a deeper understanding of air quality issues, ultimately contributing to the project’s overarching goal of improving public health and environmental sustainability in Tamil Nadu.

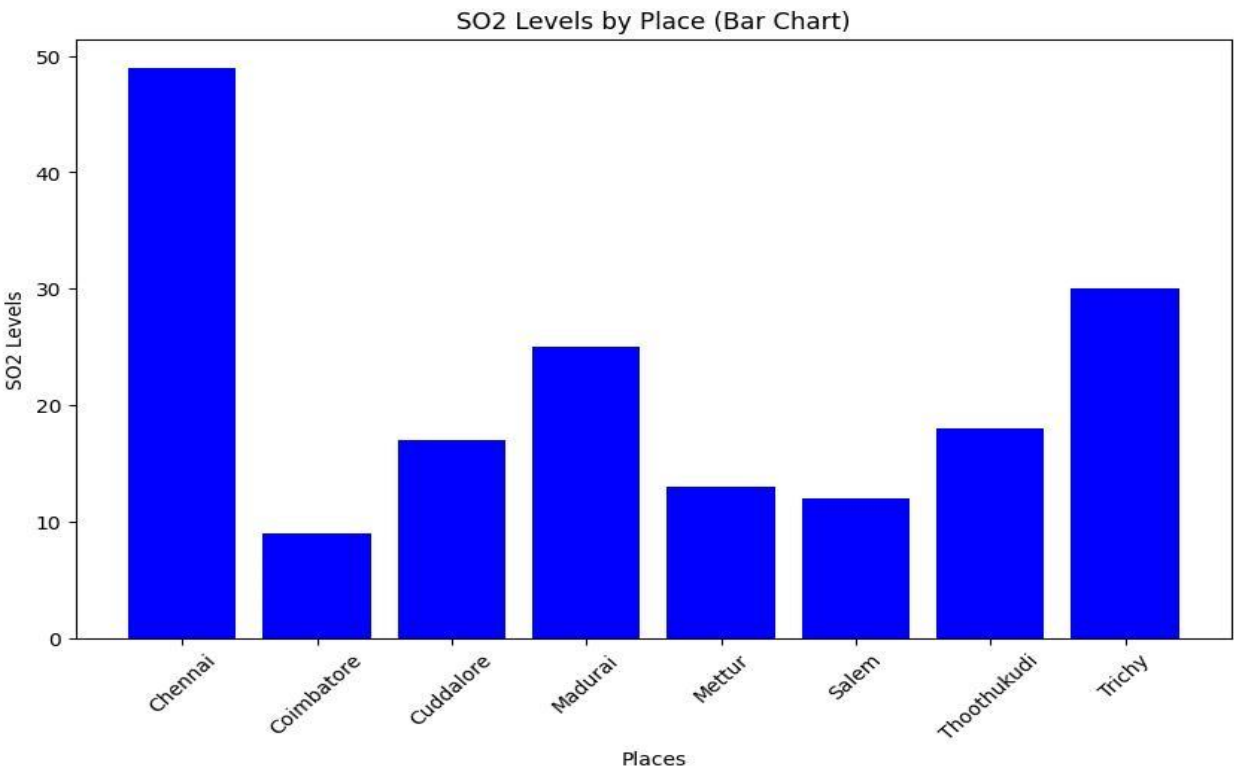


Figure 1: SO2 levels in different cities and areas

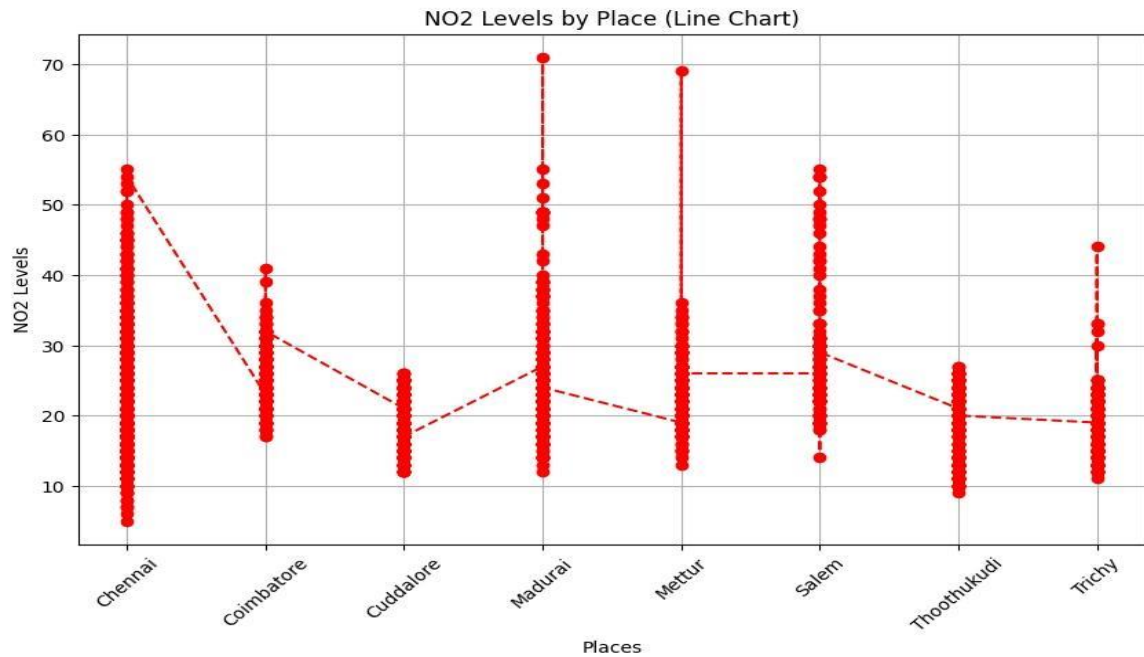


Figure 2: NO2 levels in various places illustrated using line chart

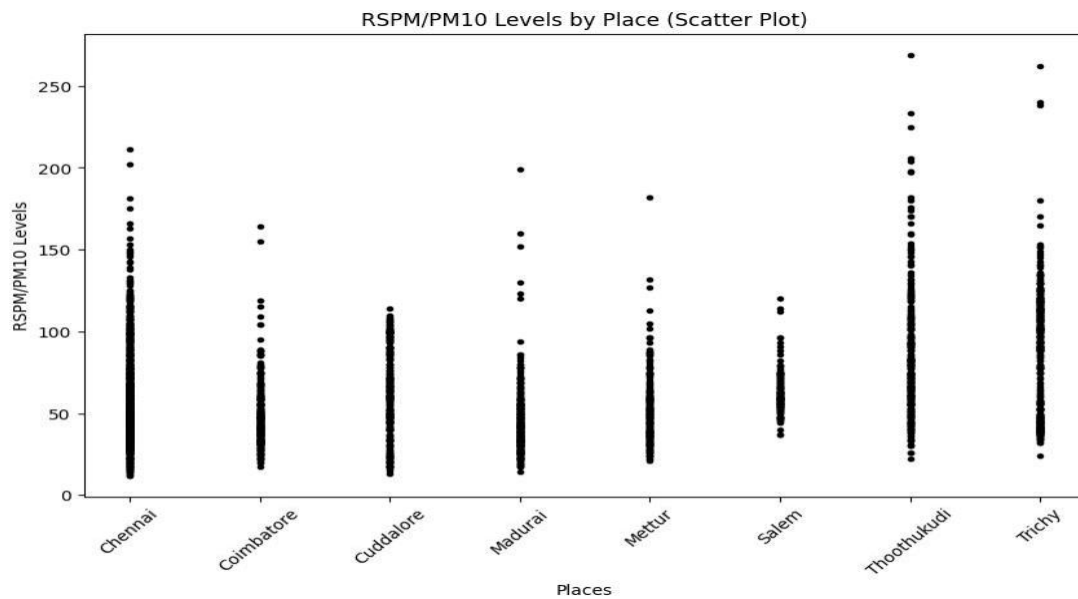
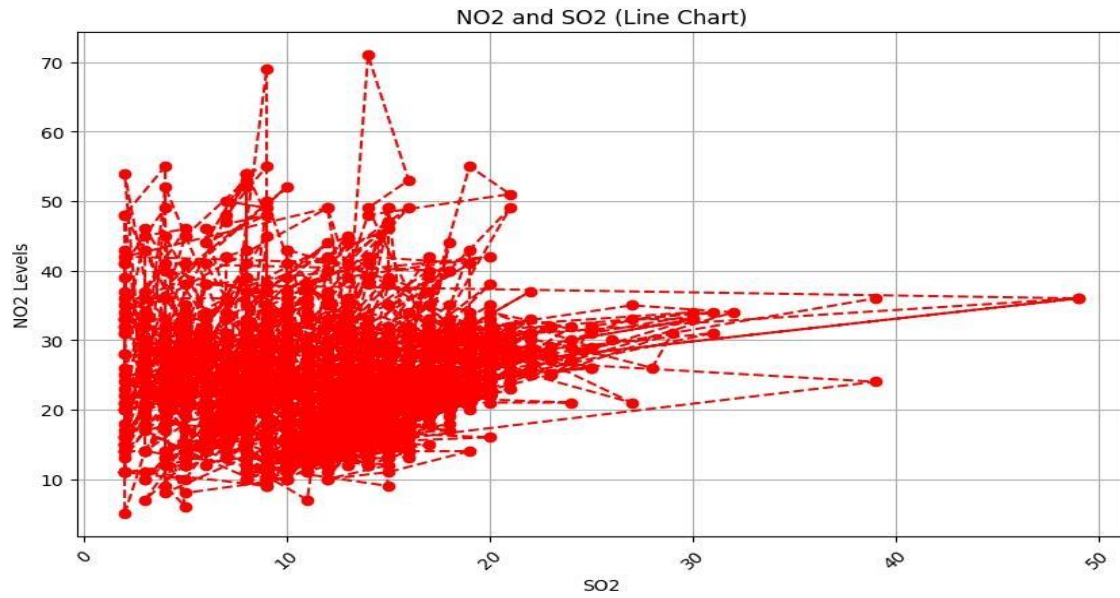
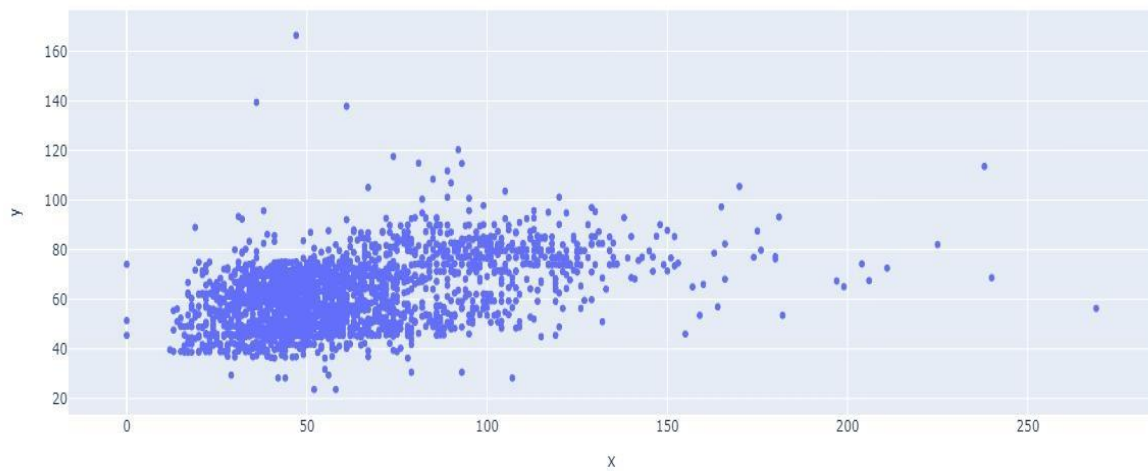


Figure 3: RSPM/PM10 levels in various places illustrated using Scatter plot



**Figure 4: Comparison of SO2 and NO 2 level using line chart**



**Figure 5: Visualization of prediction model using scatter plot**

---

## 7. ADVANTAGES AND DISADVANTAGES

### Advantages:

- **Improved Public Health:** By identifying and addressing pollution hotspots, the project directly contributes to improved public health by reducing the risk of respiratory illnesses and related healthcare costs.
- **Environmental Conservation:** The project's efforts to reduce air pollution have a positive impact on ecosystems, preventing damage to flora and fauna and mitigating the effects of climate change.
- **Informed Decision-Making:** The data-driven insights from the project empower policymakers to make informed decisions regarding pollution control measures, urban planning, and public health interventions.
- **Predictive Modeling:** The inclusion of predictive modeling enables proactive measures by forecasting air quality, allowing authorities to take preventive actions before pollution reaches critical levels.

### Disadvantages:

- **Data Accuracy:** The accuracy of the project relies heavily on the quality of the input data. Inaccurate or unreliable data can lead to flawed conclusions and predictions.
  - **Resource Intensive:** Effective data analysis and modeling require significant computational resources, which may pose challenges for organizations with limited hardware capabilities.
  - **Data Privacy:** Handling sensitive air quality data demands strict data privacy and security measures to prevent unauthorized access or misuse, which can be complex and costly to implement.
  - **Regulatory Challenges:** Adherence to and compliance with various environmental regulations can be challenging and may require close collaboration with governmental agencies.
-

## **8. Future Scope:**

The “Air Quality Analysis in Tamil Nadu” project presents a promising trajectory for future advancements and impact. The predictive modeling capabilities developed can be extended to encompass more air pollutants, offering a comprehensive air quality forecasting system. Moreover, the localized approach can serve as a blueprint for similar projects in other regions, fostering a network of environmental initiatives. As technology advances, integration with real-time IoT sensors can enhance data accuracy and timeliness. Collaborations with private sector partners, industries, and research institutions can further expand the project’s reach and capabilities. Additionally, the project’s data-driven insights may spark policy changes, emphasizing the potential for long-lasting improvements in public health and environmental sustainability. Overall, the project has the potential to be a catalyst for broader environmental and public health initiatives, driving positive change beyond Tamil Nadu’s borders.

-

## **9.Conclusion:**

The “Air Quality Analysis in Tamil Nadu” project underscores the power of data-driven solutions in the pursuit of cleaner air, improved public health, and a sustainable environment. It stands as a testament to the commitment to safeguarding the well-being of the people and ecosystems of Tamil Nadu. Through comprehensive analysis, predictive modeling, and informative data visualization, the project not only addresses current air quality challenges but also paves the way for a future where environmental stewardship and informed decision-making lead to a brighter and healthier tomorrow. It is a testament to the potential of collaborative efforts, technology, and innovation in creating a world where clean air is a fundamental right for all.

## **10. APPENDIX :**

**Github link :** <https://github.com/shyam7028/air-quality-analysis.git>

---

---

---

---