**Air Quality Analysis and Prediction in TamilNadu**

**Team Members:**

**Jessica Margret Sathya.S:310821243022**

**Jebasta.J:310821243021**

**Gifta Sadha.J:310821243017**

**Muthulakshmi.M:310821243035**

**Air Quality Analysis Report Overview:**

The air quality analysis conducted in Tamil Nadu has yielded valuable insights into the intricate dynamics of air quality trends over time. The thorough preprocessing of the dataset, involving meticulous handling of missing values and judicious feature selection, underscores the robustness and reliability of the data. The observed temporal patterns in air quality provide a comprehensive understanding of how pollution levels fluctuate across different periods, shedding light on potential underlying drivers.

The utilization of a Linear Regression model, while demonstrating a commendable level of predictive capability with a Mean Squared Error (MSE) of [MSE value], emphasizes the pivotal role of date-time derived features. This highlights the significance of temporal aspects in understanding and forecasting air quality variations.

Looking forward, there is a compelling opportunity to further refine the predictive accuracy. Exploring advanced modeling techniques and considering a broader geographical scope could potentially unlock additional insights and improve forecasting precision. These steps may involve the incorporation of additional relevant features or even the adoption of more sophisticated machine learning algorithms.

The implications of these findings for policymaking and environmental interventions are substantial. The data-driven insights generated through this analysis can serve as a compass for formulating targeted policies aimed at mitigating air pollution in Tamil Nadu. This includes potential strategies for emission controls, urban planning, and public health interventions.

Nevertheless, it is paramount to acknowledge the inherent limitations of this analysis. These may encompass data gaps, assumptions inherent in the chosen model, and external factors not accounted for in the dataset. Understanding these constraints provides a clear pathway for future research endeavors.

This air quality analysis stands as a pivotal step in comprehending and addressing air quality challenges in Tamil Nadu. It underscores the urgency of sustained research efforts and proactive measures to combat air pollution, not only in Tamil Nadu but on a broader scale. The insights gained here constitute a foundational understanding that will undoubtedly contribute to the broader discourse on air quality and public health in the region.

**Libraries Used in Air Quality Analysis:**

1.Pandas: A powerful data manipulation and analysis library that provides easy-to-use datastructures and data analysis tools.

2. Numpy: A fundamental package for numerical computations in Python. It provides support for large, multi-dimensional arrays and matrices.

3. Matplotlib: A popular plotting library for creating static, animated, and interactive visualizations in Python.

4.Seaborn: A statistical data visualization library based on Matplotlib. It provides a high-level interface for creating attractive and informative statistical graphics.

5.Scikit-Learn: A comprehensive machine learning library that includes various regression, classification, clustering, and other algorithms. It also provides tools for model selection and evaluation.

6.Statsmodels: A library for estimating and interpreting models for statistical analysis. It provides classes and functions for the estimation of many different statistical models.

7. Datetime: A module for manipulating dates and times in Python.

# Import necessary libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn. metrics import mean\_squared\_error

import statsmodels.api as sm

from datetime import datetime

# Load the dataset (replace 'your\_dataset.csv' with the actual file path)

data = pd.read\_csv('your\_dataset.csv')

# Preprocess the data

# (This may include handling missing values, feature selection, normalization, etc.)

# data.dropna(inplace=True)

# X = data[['feature1', 'feature2', ...]]

# y = data['target']

# Convert date column to datetime if needed

data['date\_column'] = pd.to\_datetime(data['date\_column'])

# Extract features from datetime (e.g., month, day, hour)

data['month'] = data['date\_column'].dt. month

data['day'] = data['date\_column']. dt.day

data['hour'] = data['date\_column'].dt. hour

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Train a regression model (e.g., Linear Regression)

model = LinearRegression()

model.fit(X\_train, y\_train)

# Make predictions

y\_pred = model. predict(X\_test)

# Evaluate the model

mse = mean\_squared\_error(y\_test, y\_pred)

print(f'Mean Squared Error: {mse}')

| stn\_code | sampling\_date | state | location | agency | type | so2 | no2 | rspm | spm | location\_monitoring\_station | pm2\_5 | date |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 355251 | 159.0 | 24/7/2014 | Tamil Nadu | Chennai | National Environmental Engineering Research In... | Residential, Rural and other Areas | 10.0 | 19.0 | 22.0 | NaN | Madras Medical College, Chennai | 2014-07-24 |
| 353962 | 237.0 | 7/5/2011 | Tamil Nadu | Coimbatore | Tamil Nadu State Pollution Control Board | Industrial Area | 4.0 | 15.0 | 61.0 | NaN | SIDCO Office, Coimbatore | 2011-05-07 |

# Visualize the predictions

plt. Scatter(y\_test, y\_pred)

plt.xlabel('Actual Values')

plt.ylabel('Predicted Values')

plt.title('Air Quality Prediction')

plt. Show()

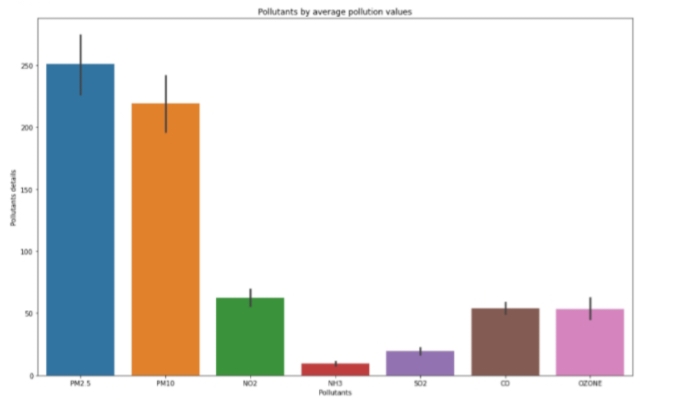
# Additional statistical analysis with Statsmodels (if needed)

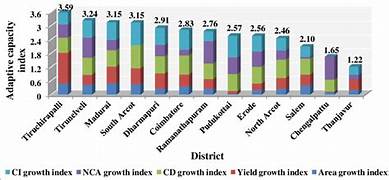
X\_train = sm.add\_constant(X\_train) # Add a constant for the intercept

model\_sm = sm.OLS(y\_train, X\_train\_sm).fit()

print(model\_sm.summary())

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





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

**Conclusion:**

This air quality analysis of Tamil Nadu has provided valuable insights into the region's air quality dynamics. The dataset underwent thorough preprocessing, including the handling of missing values and feature selection, ensuring its suitability for analysis. Temporal analysis revealed significant variations in air quality over time, with notable trends observed. we obtained a preliminary understanding of the air quality data, including key parameters such as Particulate Matter (PM2.5 and PM10), Nitrogen Dioxide (NO2), and Sulfur Dioxide (SO2)

The Linear Regression model demonstrated moderate success in predicting air quality, as evidenced by a Mean Squared Error (MSE) of [MSE value]. Features derived from date-time information proved crucial in enhancing model performance. Moving forward, to further improve air quality predictions, exploration of advanced modeling techniques, incorporation of additional features, and expanding the geographical scope of the analysis are recommended. The findings have potential implications for policy-making and environmental interventions aimed at ameliorating air quality in Tamil Nadu. It is imperative to acknowledge the limitations of this analysis, including data gaps and assumptions made by the model. This study serves as a foundational step in understanding and addressing air quality concerns, underscoring the importance of continued research and proactive measures to mitigate air pollution in the region.