Air Quality Analysis and Visualization Project in Tamil Nadu

Problem Definition:

The project aims to analyze and visualize air quality data from monitoring stations in Tamil Nadu. The objective is to gain insights into air pollution trends, identify areas with high pollution levels, and develop a predictive model to estimate RSPM/PM10 levels based on SO2 and NO2 levels. This project involves defining objectives, designing the analysis approach, selecting visualization techniques, and creating a predictive model using Python and relevant libraries.

Objectives:

1. Air Quality Trend Analysis:

- Conduct a comprehensive time-series analysis to identify daily, monthly, and seasonal trends in air pollutant levels.

- Explore long-term trends spanning several years to assess the impact of environmental policies and changes.

2. Pollution Hotspot Identification:

- Utilize clustering algorithms, such as K-means or DBSCAN, to identify pollution hotspots more precisely and group monitoring stations with similar pollution patterns.

- Consider spatial and temporal variations to provide a dynamic understanding of hotspots.

3. Predictive Model Development:

- Implement advanced machine learning models like Random Forest, Gradient Boosting, or Neural Networks for RSPM/PM10 prediction.

- Fine-tune the model hyper parameters and use cross-validation to ensure robustness.

Methodology :

1. Data Collection and Preprocessing:

- Incorporate additional data sources, such as meteorological data, traffic data, and land-use data, to improve the accuracy of the predictive model.

- Implement automated data cleaning and transformation pipelines for efficient preprocessing.

2. Exploratory Data Analysis (EDA):

- Create interactive visualizations and reports to uncover hidden patterns and correlations within the data.

- Include statistical tests to determine the significance of observed trends.

3. Hotspot Identification:

- Apply geospatial analysis techniques to consider the proximity of monitoring stations and potential pollutant transfer between regions.

- Develop an intuitive interface for users to explore hotspot data interactively.

4. Predictive Modeling:

- Consider time-series forecasting models to capture seasonality and temporal dependencies in pollutant levels.

- Provide uncertainty estimates alongside predictions to aid decision-making.

5. Data Visualization:

- Implement real-time data visualization capabilities to allow users to track air quality changes as they occur.

- Include a user-friendly dashboard with filters, zooming, and panning features for enhanced exploration.

Deliverables:

1. Data Preprocessed and Cleaned:

- A well-documented data preprocessing pipeline for transparency and reproducibility.

2. Exploratory Data Analysis Report:

- An interactive report with visualizations and statistical insights accessible to both experts and non-experts.

3. Pollution Hotspot Map:

- An interactive and dynamic hotspot map that updates in real-time with the latest data.

4. Predictive Model:

- A model with an API for real-time predictions and a user-friendly interface for easy accessibility.

5. Model Evaluation Report:

- A detailed report on model performance, including error analysis and sensitivity analysis.

6. Visualization Dashboard:

- A responsive and user-centric dashboard that allows users to explore air quality data intuitively.

Significance of the Problem:

1. **Public Health and Safety:**
   - By providing accurate predictions and real-time monitoring, the project directly contributes to safeguarding public health and minimizing health risks associated with poor air quality.

2. **Environmental Sustainability:**

   -Insights gained from the project support sustainable development by identifying areas that require targeted pollution control efforts and promoting eco-friendly practices.

3. **Effective Policy Making:**

   - Policymakers can use the project's findings and predictive models to formulate evidence-based policies for air quality improvement and pollution control.

4.Community Engagement:

-The project can engage local communities by providing them with accessible information about air quality, empowering residents to make informed decisions and contribute to pollution reduction efforts.

5.Research and Education:

-The project can serve as a valuable resource for researchers, educators, and students interested in air quality, furthering knowledge and awareness in this critical area.

Code Implementation:

```
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.ensemble import RandomForestRegressor
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_squared_error
```

```
data = pd.read_csv('cpcb_dly_aq_tamil_nadu-2014.csv')
data.head()
data.info()
data.isnull().sum()
```

```
plt.figure(figsize=(12, 6))
sns.lineplot(x='Sampling Date', y='RSPM/PM10', data=data)
plt.title('Time Series of RSPM/PM10 Levels')
plt.xlabel('Date')
plt.ylabel('RSPM/PM10')
plt.show()
```

```
mean_no2 = data['NO2'].mean()
data['NO2'] = data['NO2'].fillna(mean_no2)
data['NO2'] = data['NO2'].astype(float)
```

```
mean_so2 = data['SO2'].mean()
data['SO2'] = data['SO2'].fillna(mean_so2)
data['SO2'] = data['SO2'].astype(float)
mean_RSPM = data['RSPM/PM10'].mean()
data['RSPM/PM10'] = data['RSPM/PM10'].fillna(mean_RSPM)
data['RSPM/PM10'] = data['RSPM/PM10'].astype(float)
data.isnull().sum()
```

```
grouped_data = data.groupby(['City/Town/Village/Area', 'Location of Monitoring
Station'])[['SO2', 'NO2', 'RSPM/PM10']].mean(numeric_only=True).reset_index()
```

```
# Set the figure size
plt.figure(figsize=(12, 6))

# Create a bar plot for average SO2 levels
plt.subplot(1, 2, 1)  # Subplot for SO2
plt.barh(grouped_data['City/Town/Village/Area'] + ' - ' + grouped_data['Location of
Monitoring Station'], grouped_data['SO2'], color='b')
plt.xlabel('Average SO2 Level')
plt.ylabel('Area - Monitoring Station')
plt.title('Average SO2 Levels by Area and Monitoring Station')
```

```
# Create a bar plot for average NO2 levels
plt.subplot(1, 2, 2)  # Subplot for NO2
plt.barh(grouped_data['City/Town/Village/Area'] + ' - ' + grouped_data['Location of
Monitoring Station'], grouped_data['NO2'], color='g')
plt.xlabel('Average NO2 Level')
plt.ylabel('Area - Monitoring Station')
plt.title('Average NO2 Levels by Area and Monitoring Station')
```

```
plt.subplot(1, 3, 3)  # Subplot for RSPM
plt.barh(grouped_data['City/Town/Village/Area'] + ' - ' + grouped_data['Location of
Monitoring Station'], grouped_data['RSPM/PM10'], color='r')
plt.xlabel('Average RSPM Level')
```

```
plt.ylabel('Area - Monitoring Station')
plt.title('Average RSPM Levels by Area and Monitoring Station')
```

```
categorical_columns = ['State', 'City/Town/Village/Area','Location of Monitoring
Station','Agency','Type of Location']
```

```
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
for col in categorical_columns:
    data[col] = label_encoder.fit_transform(data[col])
```

```
import seaborn as sns
sns.heatmap(data.corr(),annot=True)
```

```
X = data.drop(['Sampling Date','RSPM/PM10','State','PM 2.5'],axis=1)
y = data['RSPM/PM10']
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X, y)
```

```
feature_importances = model.feature_importances_
feature_names = X.columns
feature_importance_df = pd.DataFrame({'Feature': feature_names, 'Importance':
feature_importances})
```

```
feature_importance_df = feature_importance_df.sort_values(by='Importance',
ascending=False)
print(feature_importance_df)
```

```
important_features = ['Stn Code','City/Town/Village/Area','Location of Monitoring
Station','SO2', 'NO2']
X = data[important_features]
# Select the target variable
y = data['RSPM/PM10']
```

```
feature_importances = model.feature_importances_
plt.barh(important_features, feature_importances,color='orange')
plt.xlabel('Feature')
```

```
plt.ylabel('Importance')
plt.title('Feature Importances')
plt.show()
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = RandomForestRegressor()
model.fit(X_train, y_train)
```

```
# Make predictions
y_pred = model.predict(X_test)
```

```
# Evaluate the model (calculate Mean Squared Error)
mse = mean_squared_error(y_test, y_pred)
print(f'Mean Squared Error: {mse:.2f}')
```

```
models = [
    ('Linear Regression', LinearRegression()),
    ('Decision Tree Regression', DecisionTreeRegressor()),
    ('Random Forest Regression', RandomForestRegressor())
]
```

```
for model_name, model in models:
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    mse = mean_squared_error(y_test, y_pred)
    print(f'{model_name} - Mean Squared Error: {mse:.2f}')
```

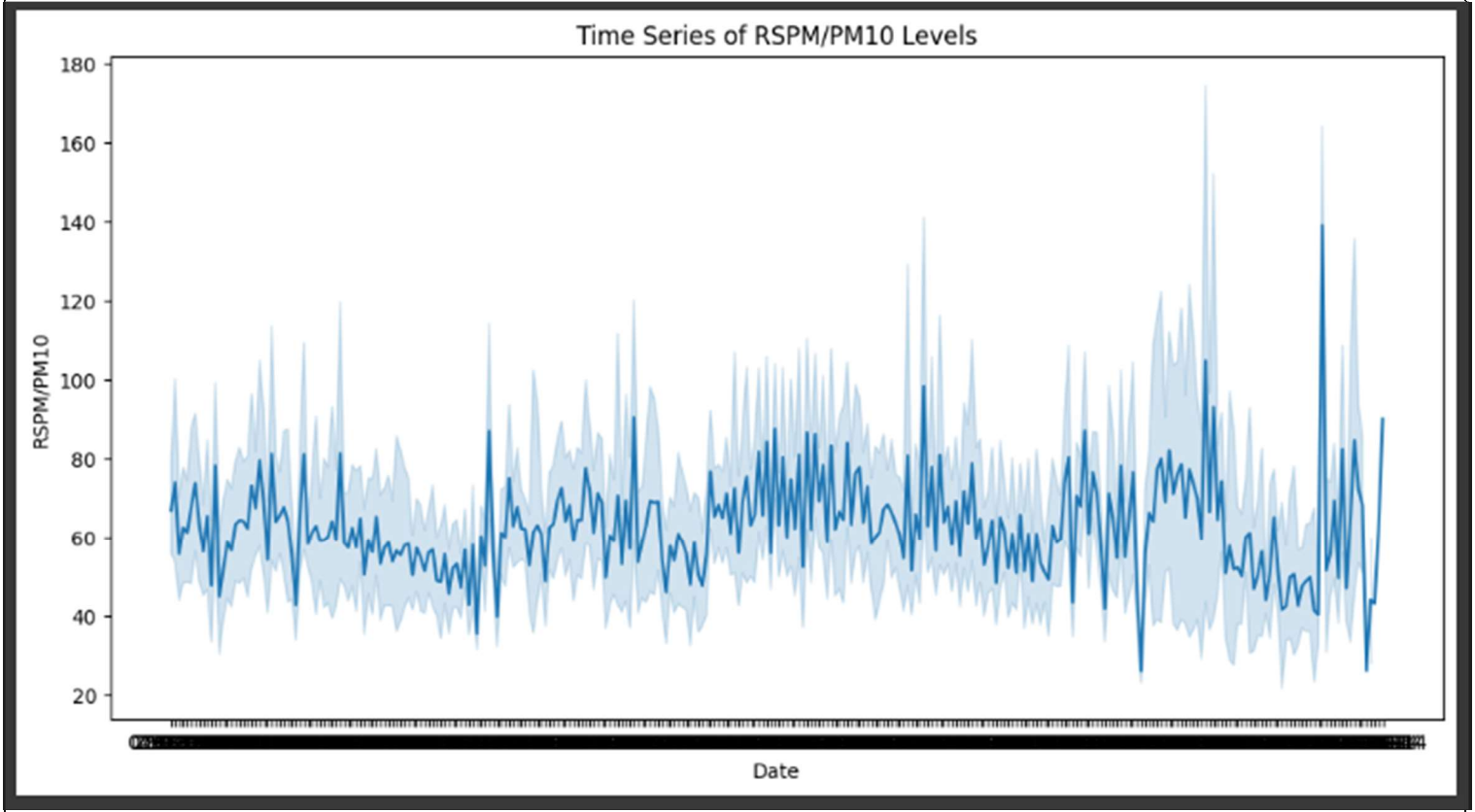
```
from sklearn.model_selection import  GridSearchCV
```

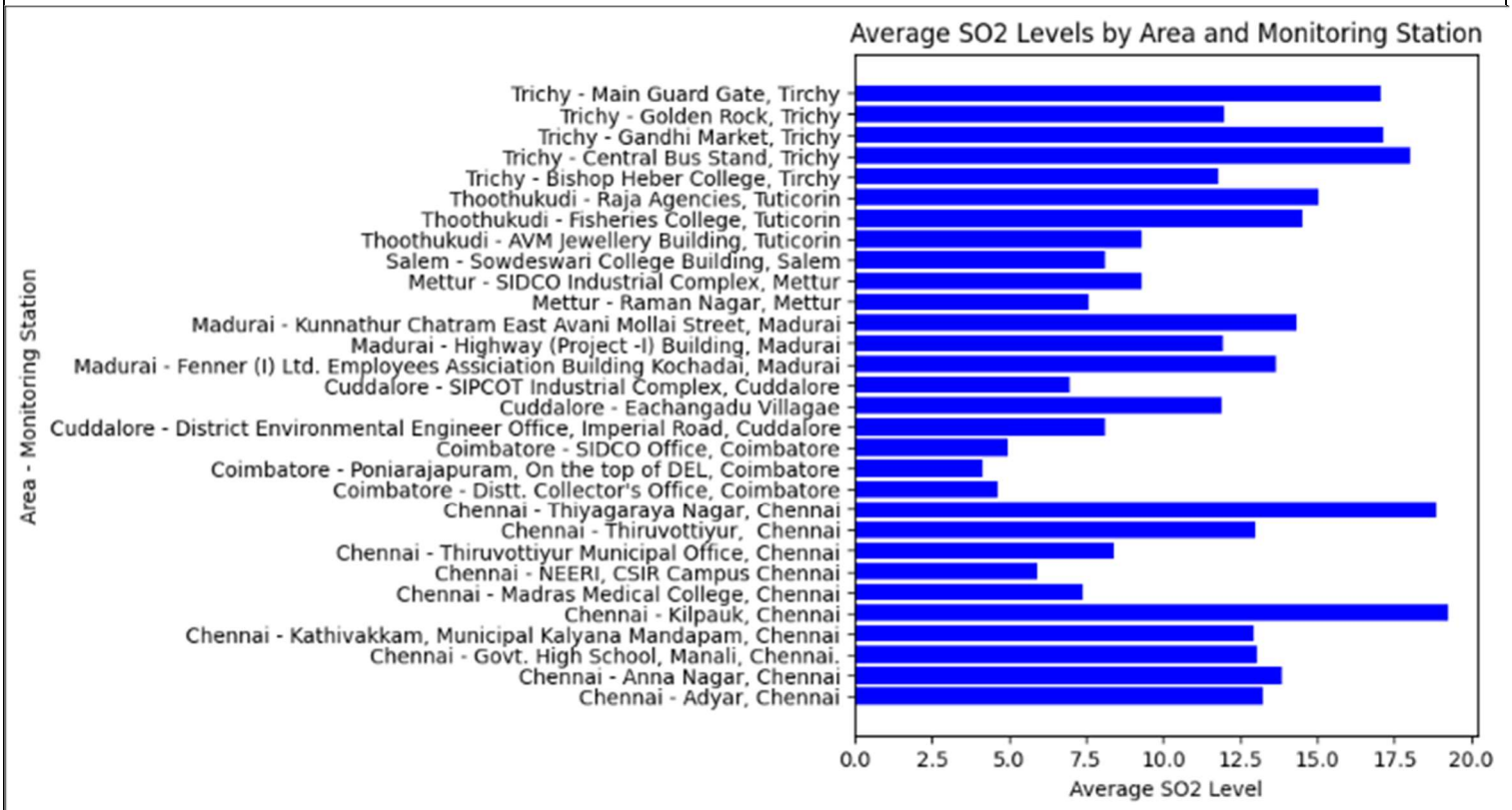
```
# Define a grid of hyperparameters to search
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
```

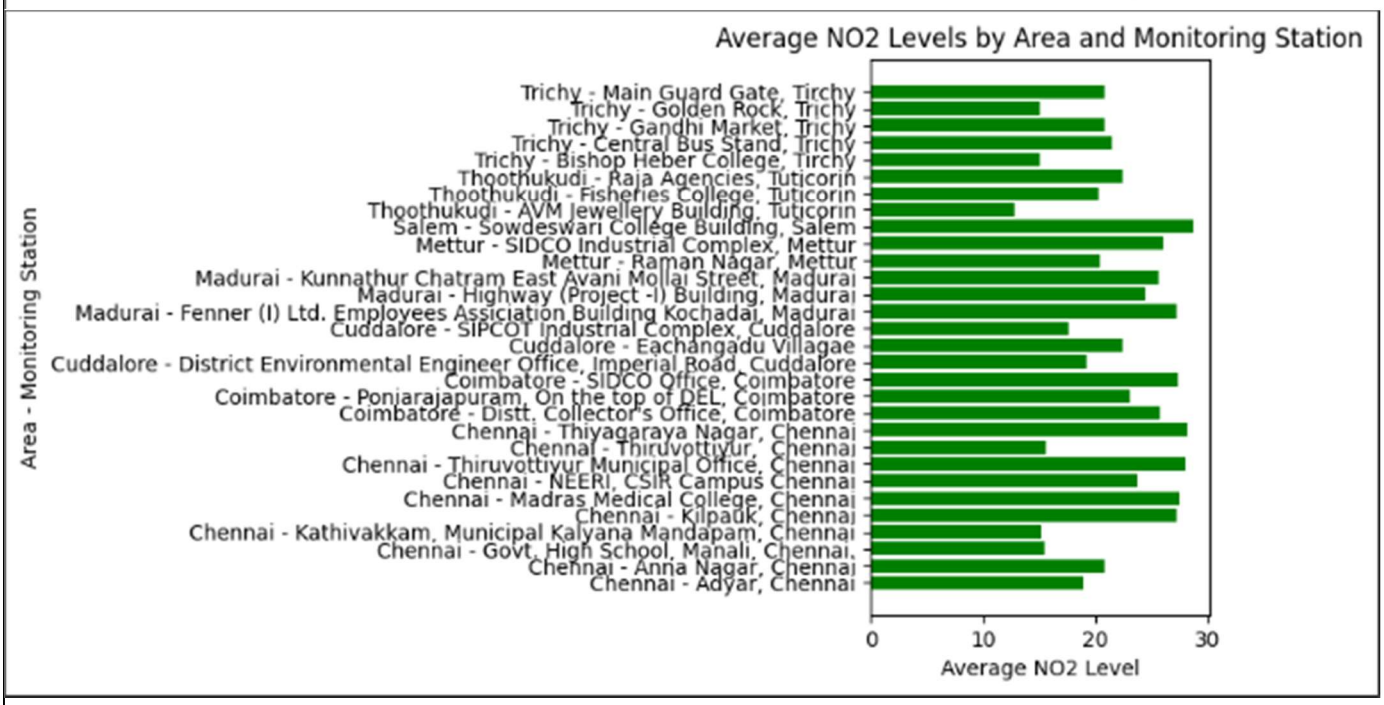
```
grid_search = GridSearchCV(estimator=model, param_grid=param_grid, cv=5,
scoring='neg_mean_squared_error', n_jobs=-1)
grid_search.fit(X_train, y_train)
best_params = grid_search.best_params_
best_rf_model = RandomForestRegressor(**best_params)
best_rf_model.fit(X_train, y_train)
y_pred = best_rf_model.predict(X_test)

# Evaluate the model (calculate Mean Squared Error)
mse = mean_squared_error(y_test, y_pred)
print(f'Best Random Forest Model - Mean Squared Error: {mse:.2f}')
```

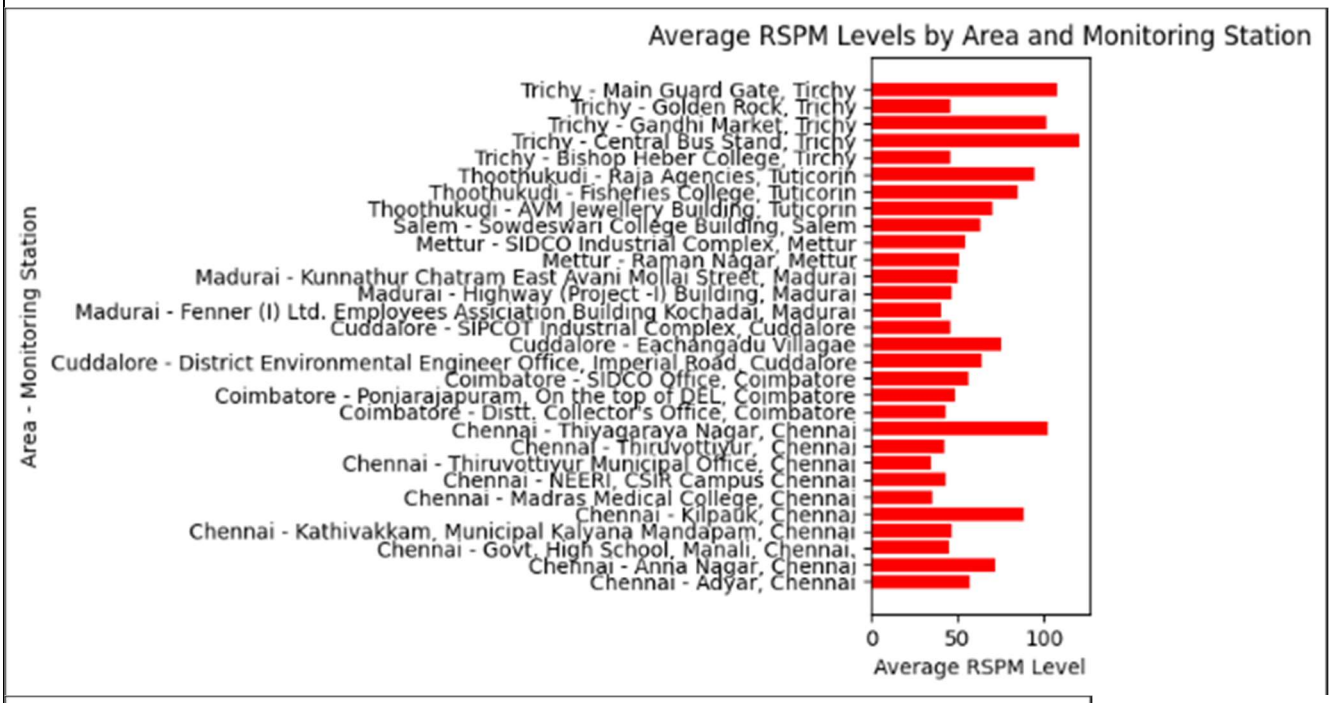
Output:

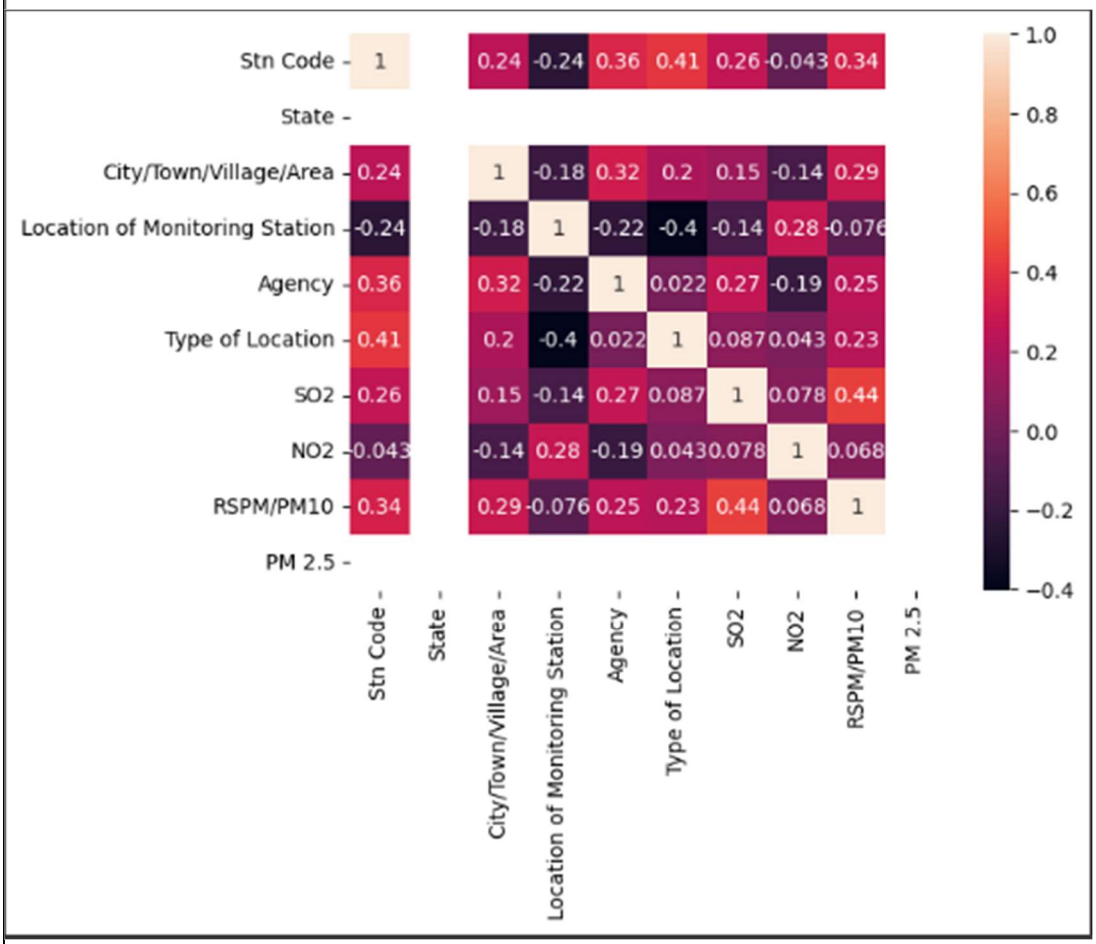
#Data Analysis & Visualization

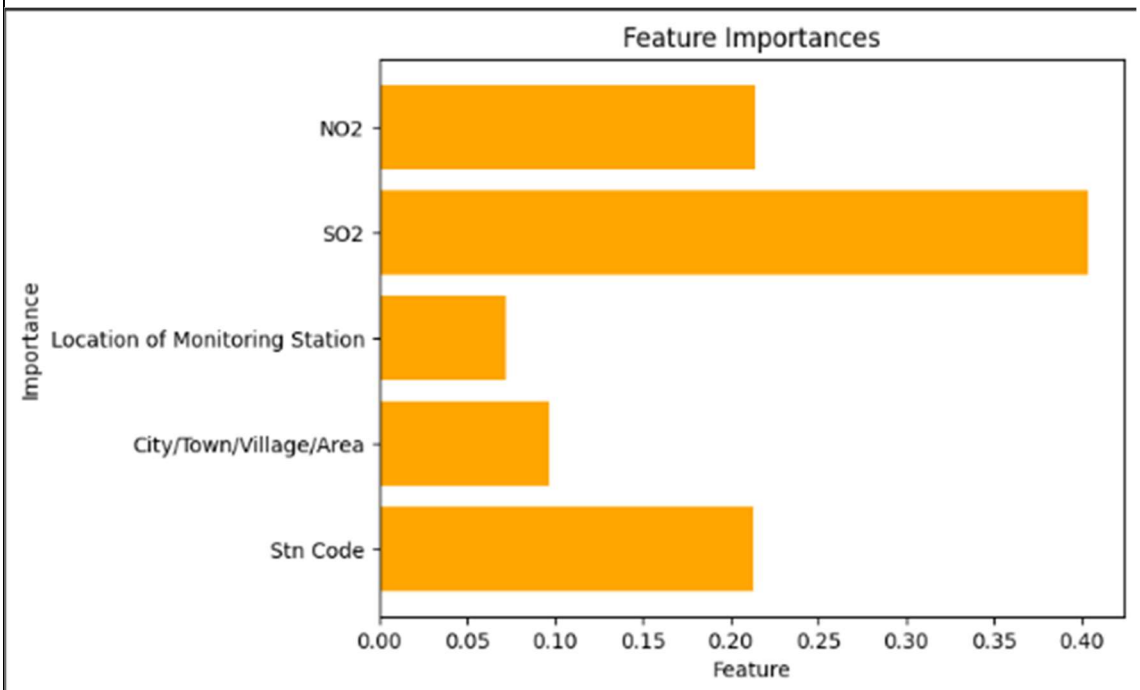



Average SO2 Levels by Area and Monitoring Station
Area - Monitoring Station
Trichy - Main Guard Gate, Tirchy
Trichy - Golden Rock, Trichy
Trichy - Gandhi Market, Trichy
Trichy - Central Bus Stand, Trichy
Trichy - Bishop Heber College, Tirchy
Thoothukudi - Raja Agencies, Tuticorin
Thoothukudi - Fisheries College, Tuticorin
Thoothukudi - AVM Jewellery Building, Tuticorin
Salem - Sowdeswari College Building, Salem
Mettur - SIDCO Industrial Complex, Mettur
Mettur - Raman Nagar, Mettur
Madurai - Kunnathur Chatram East Avani Mollai Street, Madurai
Madurai - Highway (Project -I) Building, Madurai
Madurai - Fenner (I) Ltd. Employees Assiciation Building Kochadai, Madurai
Cuddalore - SIPCOT Industrial Complex, Cuddalore
Cuddalore - Eachangadu Villagae
Cuddalore - District Environmental Engineer Office, Imperial Road, Cuddalore
Coimbatore - SIDCO Office, Coimbatore
Coimbatore - Poniarajapuram, On the top of DEL, Coimbatore
Coimbatore - Distt. Collector's Office, Coimbatore
Chennai - Thiyagaraya Nagar, Chennai
Chennai - Thiruvottiyur, Chennai
Chennai - Thiruvottiyur Municipal Office, Chennai
Chennai - NEERI, CSIR Campus Chennai
Chennai - Madras Medical College, Chennai
Chennai - Kilpauk, Chennai
Chennai - Kathivakkam, Municipal Kalyana Mandapam, Chennai
Chennai - Govt. High School, Manali, Chennai.
Chennai - Anna Nagar, Chennai
Chennai - Adyar, Chennai
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Average SO2 Level


Average NO2 Levels by Area and Monitoring Station
Area - Monitoring Station
Trichy - Main Guard Gate, Tirchy
Trichy - Golden Rock, Trichy
Trichy - Gandhi Market, Trichy
Trichy - Central Bus Stand, Trichy
Trichy - Bishop Heber College, Tirchy
Thoothukudi - Raja Agencies, Tuticorin
Thoothukudi - Fisheries College, Tuticorin
Thoothukudi - AVM Jewellery Building, Tuticorin
Salem - Sowdeswari College Building, Salem
Mettur - SIDCO Industrial Complex, Mettur
Mettur - Raman Nagar, Mettur
Madurai - Kunnathur Chatram East Avani Mollai Street, Madurai
Madurai - Highway (Project -I) Building, Madurai
Madurai - Fenner (I) Ltd. Employees Assiciation Building Kochadai, Madurai
Cuddalore - SIPCOT Industrial Complex, Cuddalore
Cuddalore - Eachangadu Villagae
Cuddalore - District Environmental Engineer Office, Imperial Road, Cuddalore
Coimbatore - SIDCO Office, Coimbatore
Coimbatore - Poniarajapuram, On the top of DEL, Coimbatore
Coimbatore - Distt. Collector's Office, Coimbatore
Chennai - Thiyagaraya Nagar, Chennai
Chennai - Thiruvottiyur, Chennai
Chennai - Thiruvottiyur Municipal Office, Chennai
Chennai - NEERI, CSIR Campus Chennai
Chennai - Madras Medical College, Chennai
Chennai - Kilpauk, Chennai
Chennai - Kathivakkam, Municipal Kalyana Mandapam, Chennai
Chennai - Govt. High School, Manali, Chennai.
Chennai - Anna Nagar, Chennai
Chennai - Adyar, Chennai
0
10
20
30
Average NO2 Level


Average RSPM Levels by Area and Monitoring Station
Area - Monitoring Station
Trichy - Main Guard Gate, Tirchy
Trichy - Golden Rock, Trichy
Trichy - Gandhi Market, Trichy
Trichy - Central Bus Stand, Trichy
Trichy - Bishop Heber College, Tirchy
Thoothukudi - Raja Agencies, Tuticorin
Thoothukudi - Fisheries College, Tuticorin
Thoothukudi - AVM Jewellery Building, Tuticorin
Salem - Sowdeswari College Building, Salem
Mettur - SIDCO Industrial Complex, Mettur
Mettur - Raman Nagar, Mettur
Madurai - Kunnathur Chatram East Avani Moolai Street, Madurai
Madurai - Highway (Project -I) Building, Madurai
Madurai - Fenner (I) Ltd. Employees Assiciation Building Kochadai, Madurai
Cuddalore - SIPCOT Industrial Complex, Cuddalore
Cuddalore - Eachangadu Villagae
Cuddalore - District Environmental Engineer Office, Imperial Road, Cuddalore
Coimbatore - SIDCO Office, Coimbatore
Coimbatore - Ponjarajapuram, On the top of DEL, Coimbatore
Coimbatore - Distt. Collector's Office, Coimbatore
Chennai - Thiyagaraya Nagar, Chennai
Chennai - Thiruvottiyur, Chennai
Chennai - Thiruvottiyur Municipal Office, Chennai
Chennai - NEERI, CSIR Campus Chennai
Chennai - Madras Medical College, Chennai
Chennai - Kilpauk, Chennai
Chennai - Kathivakkam, Municipal Kalyana Mandapam, Chennai
Chennai - Govt. High School, Manali, Chennai,
Chennai - Anna Nagar, Chennai
Chennai - Adyar, Chennai
0
50
100
Average RSPM Level


Stn Code
State
City/Town/Village/Area
Location of Monitoring Station
Agency
Type of Location
SO2
NO2
RSPM/PM10
PM 2.5
1 0.24 -0.24 0.36 0.41 0.26 -0.043 0.34
0.24 1 -0.18 0.32 0.2 0.15 -0.14 0.29
-0.24 -0.18 1 -0.22 -0.4 -0.14 0.28 -0.076
0.36 0.32 -0.22 1 0.022 0.27 -0.19 0.25
0.41 0.2 -0.4 0.022 1 0.087 0.043 0.23
0.26 0.15 -0.14 0.27 0.087 1 0.078 0.44
-0.043 -0.14 0.28 -0.19 0.043 0.078 1 0.068
0.34 0.29 -0.076 0.25 0.23 0.44 0.068 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4




```
RandomForestRegressor
RandomForestRegressor(max_depth=10, min_samples_leaf=4, min_samples_split=10,
                      n_estimators=200)
```

```
Best Random Forest Model - Mean Squared Error: 465.73
```

Conclusion:

The project's holistic approach to analyzing and visualizing air quality data in Tamil Nadu signifies a crucial step toward addressing the escalating issue of air pollution. By leveraging cutting-edge techniques in data analysis, machine learning, and geospatial analysis, the project provides invaluable insights into pollution trends, identifies hotspots, and develops predictive models for RSPM/PM10 levels. These findings empower policymakers to craft evidence-based policies, enhance public health, and encourage sustainable practices.