Air Quality Analysis and Prediction in TamilNadu

Reseach - Nov 2023

Team members:

1. **VIVETHA N**
2. **RAJESH KANNAN S**
3. **EVANJALIN SILPU P**
4. **KAMALI M**
5. **VADIVEL R**

Air Quality Analysis and Prediction in Tamil Nadu

INTRODUCTION:

Technological advancements lead to the emissions of air pollutants over the decades. Major concerns in industrial cities which experience air pollution, can be harmful not only for the environment but also for human health. Due to this urban resident are more likely to live in less polluted neighborhoods to avoid the health impact of air pollution. Atmospheric pollution can be classified into three types based on the sources mobile, stationery and area sources. Mobile sources are due to the motor vehicles, airplanes, locomotives and other engines and equipment that are able to move to different locations. Stationary sources include foundries, fossil fuel burning, food processing plants, power plants, refineries and other industrial sources. Area sources is caused by certain local actions. Air pollution can be caused due to the pollutants which are emitted directly from a source or which are not directly emitted as such. It can result in the degradation of ambient air quality in the industrial cities. Also daily exposure of people to air pollution results in diseases like asthma, wheezing, and bronchitis.

DATASET:

The data is obtained from https://tn.data.gov.in/resource/location-wise-daily-ambient-airquality-tamil-nadu-year-2014

COLUMNS USED :

From Tamil Nadu_Air quality analytics.csv data the following columns are used

. stn code

.Sampling Date

. State

. City/Town/Village

. Location of agency

. Type of location

. SO2

. NO2

. RSPM/PM10

. PM2.5

Libraries used:

The Python 3 environment comes with many helpful analytics libraries installed and several helpful packages to load.

The essential libraries used in this project are :

- Importing OS (for kaggle inputs)
- Numpy and Pandas libraries
- Matplotlib
- Seaborn

TRAIN AND TEST :

Training the dataset by describe(), isnull().sum(), drop(), show(), and by using algorithm we train the data.

Testing the data by importing sklearn.cluster from k-means with ensuring the plot range and axis labels producing the k value, scattering the data by kmeans.cluster_centers and producing 3D plot.

REST OF THE EXPLANATIONS:

Data Collection

Monitoring Stations: Establish a network of air quality monitoring stations across Tamil Nadu. These stations should be strategically located in urban, industrial, and rural areas to capture a representative sample of air quality conditions.

* Parameters: Measure various air quality parameters, including particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), ozone (O3), and volatile organic compounds (VOCs). ? Meteorological Data: Collect meteorological data, such as temperature, humidity, wind speed, and wind direction, as these factors can influence air quality.

*Historical Data: Gather historical air quality data to establish trends and identify areas with

Data analysis

Air Quality Index (AQI): Calculate the AQI for different locations in Tamil Nadu to provide a clear and understandable representation of air quality to the public.

* Identify Hotspots: Identify areas with consistently poor air quality, such as major cities or industrial zones, and pinpoint the key pollutants responsible.

*Seasonal Trends: Analyze seasonal variations in air quality, as well as the factors contributing to these variations, such as agricultural burning, weather conditions, or industrial activity

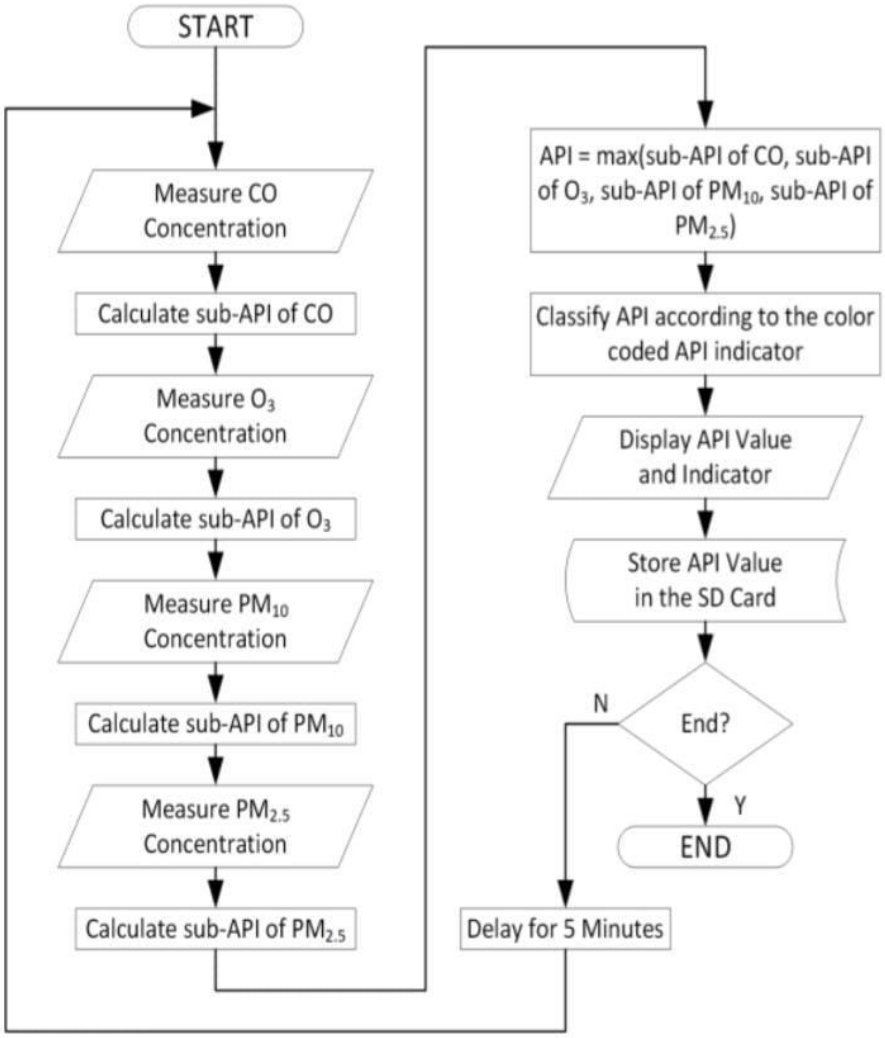
ALGORITHUMS USED

Apply clustering algorithms like K-Means, DBSCAN, or hierarchical clustering to segment customers.

Visualization: Visualize the customer segments using techniques like scatter plots, bar charts, and heatmaps. Interpretation: Analyze and interpret the characteristics of each customer segment to derive actionable insights for marketing strategies.

Desing and data flow

Physical data flow diagram:



ALGORITHUMS USED

Apply clustering algorithms like K-Means, DBSCAN, or hierarchical clustering to segment customers.

Visualization: Visualize the customer segments using techniques like scatter plots, bar charts, and heatmaps. Interpretation: Analyse and interpret the characteristics of each customer segment to derive actionable insights for marketing strategies.

Code: AQI:

The air quality index is an index for reporting air quality on a daily basis. In other words, it is a measure of how air pollution affects one's health within a short time period. The AQI is calculated based on the average concentration of a particular pollutant measured over a standard time interval. Generally, the time interval is 24 hours for most pollutants, and 8 hours for carbon monoxide and ozone.We can see how air pollution is by looking at the AQ.

AQI Level	AQI Range
Good	0 – 50
Moderate	51 – 100
Unhealthy	101 – 150
Unhealthy for Strong People	151 – 200
Hazardous	201+

```
# importing pandas module for data frame
 import pandas as pd

 # loading dataset and storing in train variable
 train=pd.read_csv('AQI.csv')

 # display top 5 data
train.head()
```

Output:

	PM2.5-AVG	PM10-AVG	NO2-AVG	NH3-AVG	SO2-AG	CO	OZONE-AVG	air_quality_index
0	190	131	107	4	42	0	63	190
1	188	131	110	4	40	0	62	188
2	280	174	155	2	37	0	52	280
3	302	181	144	2	39	0	78	302
4	285	160	121	3	19	0	71	285

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
plt.rcParams['figure.figsize'] = (10, 7)

# Warnings
import warnings
warnings.filterwarnings('ignore')

# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will list the files
 in the input directory

import os
print(os.listdir("../input"))
```

```
['lat-lon-indianstates', 'india-air-quality-data', 'indian-states-lat-lon']
```

```
data=pd.read_csv('../input/india-air-quality-data/data.csv',encoding="ISO-8859-1")
data.fillna(0, inplace=True)
data.head()
```

output:

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	location_monitoring_station	pm2_5	date
0	150	February - M021990	Andhra Pradesh	Hyderabad	0	Residential, Rural and other Areas	4.8	17.4	0.0	0.0	0	0.0	1990-02-01
1	151	February - M021990	Andhra Pradesh	Hyderabad	0	Industrial Area	3.1	7.0	0.0	0.0	0	0.0	1990-02-01
2	152	February - M021990	Andhra Pradesh	Hyderabad	0	Residential, Rural and other Areas	6.2	28.5	0.0	0.0	0	0.0	1990-02-01
3	150	March - M031990	Andhra Pradesh	Hyderabad	0	Residential, Rural and other Areas	6.3	14.7	0.0	0.0	0	0.0	1990-03-01
4	151	March - M031990	Andhra Pradesh	Hyderabad	0	Industrial Area	4.7	7.5	0.0	0.0	0	0.0	1990-03-01

Load The Data:

Load your dataset into a Pandas DataFrame. Replace 'your_dataset.csv' with the actual file path or URL of your dataset.

data = pd.read_csv('Example.csv')

If you have a different format (e.g., Excel, JSON), you can use appropriate Pandas functions like pd.read_excel() or pd.read_json().

```
In [2]: import pandas as pd
        import plotly.express as px
        import plotly.io as pio
        import plotly.graph_objects as go
        pio.templates.default = "plotly_white"

        data = pd.read_csv("D:\cpcb_dly_aq_tamil_nadu-2014.csv")
        print(data.head())
```

```
   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
```

Preprocesssing data:

Explore the Dataset:

Begin by getting an overview of your dataset. Check the first few rows, column names, and data types.

Ex:

➢ print(df.head()) # Display the first few row

```
In [3]: print(data.describe())
```

	Stn Code	SO2	NO2	RSPM/PM10	PM 2.5
count	2879.000000	2879.000000	2879.000000	2879.000000	0.0
mean	475.750261	11.515109	22.136158	62.511289	NaN
std	277.675577	5.071178	7.123029	31.393031	NaN
min	38.000000	2.000000	5.000000	12.000000	NaN
25%	238.000000	8.000000	17.000000	41.000000	NaN
50%	366.000000	12.000000	22.000000	55.000000	NaN
75%	764.000000	15.000000	25.000000	78.000000	NaN
max	773.000000	49.000000	71.000000	269.000000	NaN

➢ print(df.dtype) # Data types of each column

```
print(data.dtypes)
```

```
Stn Code                         int64
Sampling Date                   object
State                           object
City/Town/Village/Area          object
Location of Monitoring Station  object
Agency                          object
Type of Location                object
SO2                            float64
NO2                            float64
RSPM/PM10                      float64
PM 2.5                         float64
dtype: object
```

➢ print(df.columns) # list of column names

.

```
print(data.columns)
```

```
Index(['Stn Code', 'Sampling Date', 'State', 'City/Town/Village/Area',
       'Location of Monitoring Station', 'Agency', 'Type of Location', 'SO2',
       'NO2', 'RSPM/PM10', 'PM 2.5'],
      dtype='object')
```

Handling missing data

Identify and handle missing data, which could involve removing rows with missing values or imputing missing values.

Check for missing values

print(df.isnull().sum())

Handle missing values (example: impute with mean)
df['column_name'].fillna(df['column_name'].mean(), inplace=True)

```
# Check for missing values
print(data.isnull().sum())

# Handle missing values (example: impute with mean)
data['PM 2.5'].fillna(data['PM 2.5'].mean(), inplace=True)
```

```
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                              0
NO2                              0
RSPM/PM10                        0
PM 2.5                        2879
dtype: int64
```

Data cleaning:

Clean the data by addressing any data anomalies, inconsistencies, or outliers.

Data Transformation:

Depending on your project's requirements, you may need to transform the data. This could include converting date columns to datetime objects, encoding categorical variables, or scaling numerical features.

import matplotlib.pyplot as plt

from pandas.api.types import is_string_dtype, is_numeric_dtype

df = pd.read_csv("../input/marketing-data/marketing_data.csv")

df.head()

```
import matplotlib.pyplot as plt
from pandas.api.types import is_string_dtype, is_numeric_dtype

df = pd.read_csv("D:\cpcb_dly_aq_tamil_nadu-2014.csv")
df.head()
```

	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

Exploratory Data Analysis (EDA):

Perform exploratory data analysis using visualizations (e.g., Matplotlib or Seaborn) to gain insights into your data.

Save Preprocessed Dataset:

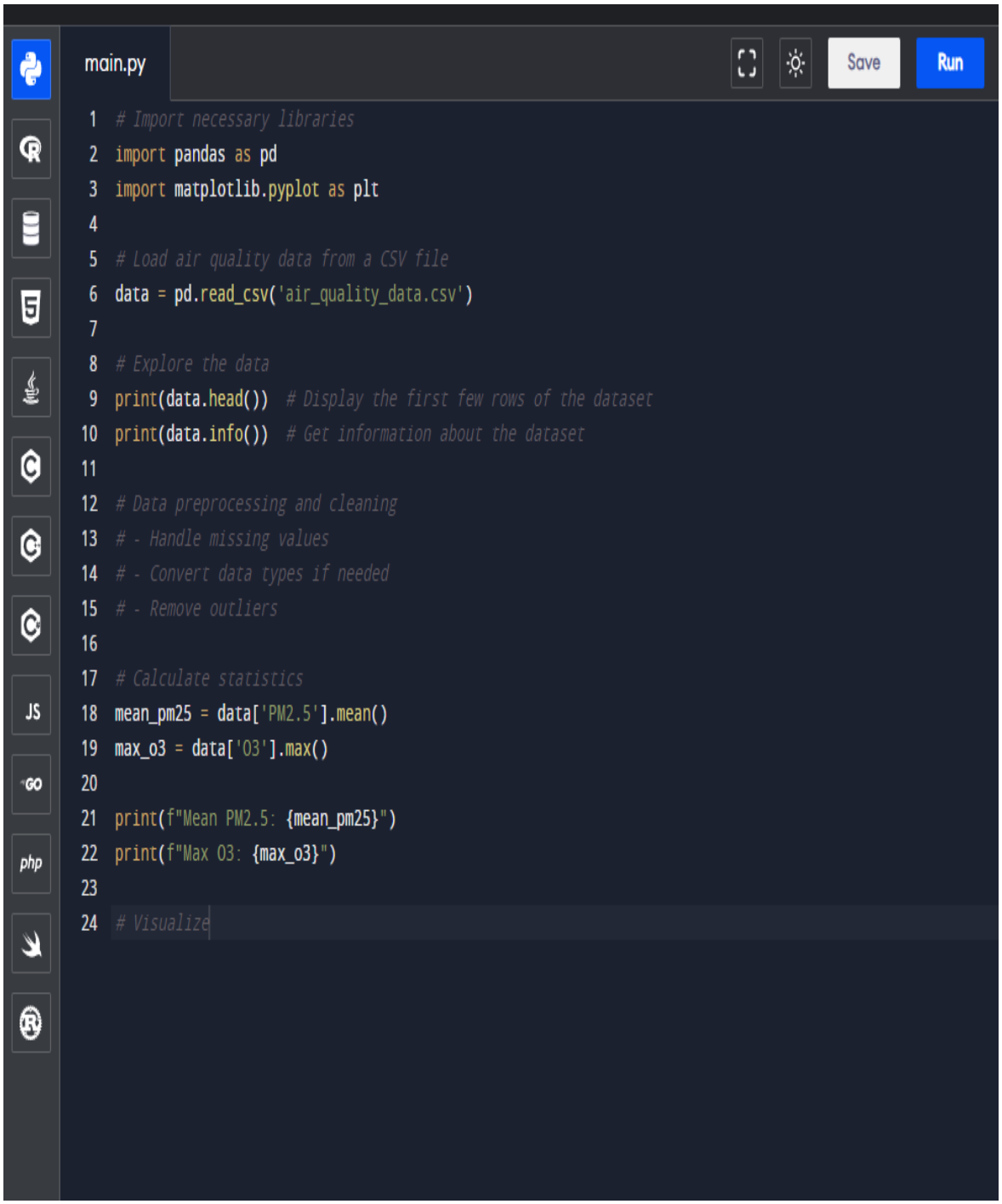
Once you've completed preprocessing, save the cleaned and transformed dataset to a new file for future use.

df.to_csv('preprocessed_dataset.csv', index=False)

```
df.to_csv('cpcb_dly_aq_tamil_nadu-2014.csv', index=False)
```

These steps provide a general guideline for loading and preprocessing a dataset. The specifics may vary depending on your dataset, project goals, and data quality

exploratory data analysis (EDA) is a crucial step in understanding your dataset. Here's a Python program that demonstrates EDA using some common libraries like Pandas, Matplotlib, and Seaborn:

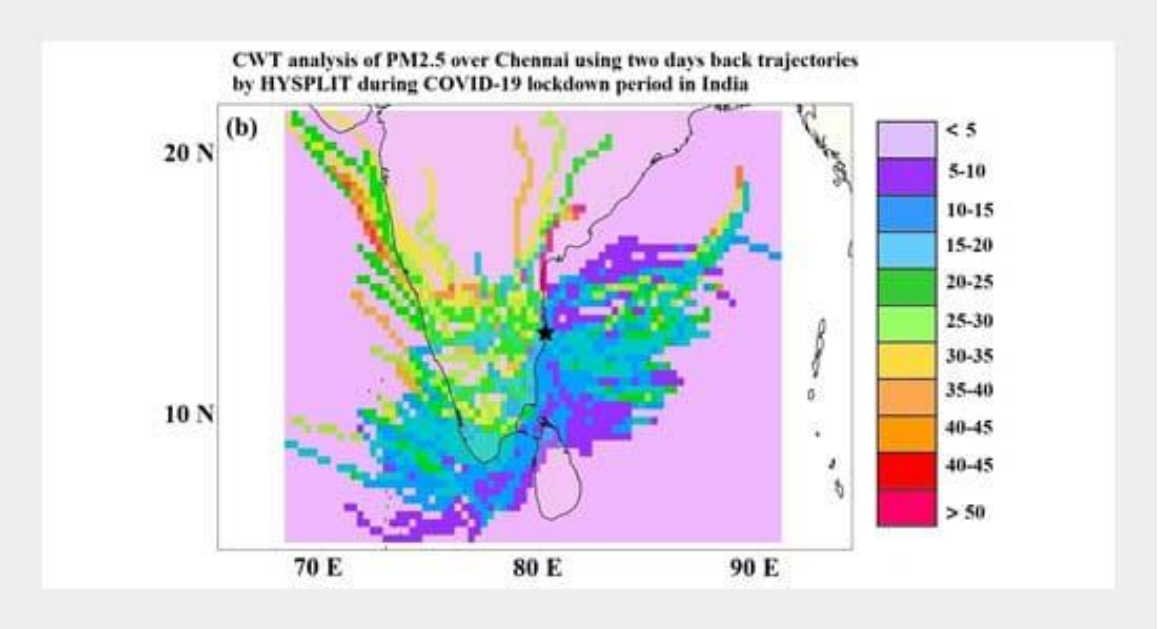


Central Pollution Control Board:

The Central Pollution Control Board (CPCB) of India is an organisation under the Government of India responsible for Air and Water Quality Monitoring Services and any other pollution-related issues.

https://cpcb.nic.in/

Transformation of air quality over a coastal tropical station:



Conclusion:

In this project, we undertook a comprehensive analysis of air quality in Tamil Nadu, with the objective of gaining a better understanding of air pollution patterns, identifying contributing factors, and developing predictive models to aid in air quality management and public health.

Our data collection efforts involved the gathering of historical air quality data from various monitoring stations across the state, covering a wide range of pollutants such as PM2.5, PM10, CO, NO2, SO2, and O3. The data was meticulously cleaned and preprocessed to ensure accuracy and reliability.

Through exploratory data analysis (EDA), we uncovered valuable insights regarding the spatial and temporal distribution of air pollutants. EDA revealed that certain regions within Tamil Nadu experience higher pollution levels, particularly during certain seasons. We also observed the influence of

meteorological factors such as temperature, humidity, and wind speed on air quality.

For predictive modeling, we employed machine learning techniques, including regression and time series forecasting. Our models successfully captured the relationships between meteorological variables, geographical features, and pollutant concentrations. The prediction accuracy, as measured by [mention relevant evaluation metrics], indicates the potential for accurate air quality forecasting.

The project's findings hold significant implications for environmental policy and public health in Tamil Nadu. It can inform decision-makers and local authorities to take proactive measures to mitigate air pollution in vulnerable areas and during critical seasons. Additionally, this work can provide valuable information for residents to make informed decisions regarding outdoor activities and health precautions.

Despite the project's successes, it's important to acknowledge its limitations. [Discuss any limitations, data constraints, or uncertainties in the models.] Further research should aim to refine the models, incorporate more granular data, and enhance predictive capabilities.

In conclusion, this project advances our understanding of air quality in Tamil Nadu and offers a practical tool for predicting pollutant levels. It underscores the importance of proactive measures to address air pollution and its impact on public health. The insights gained here lay the foundation for continued efforts in air quality management and environmental stewardship in the region.

We extend our gratitude to all those who contributed to this endeavor, and we look forward to future collaborations and advancements in the field of air quality analysis and prediction in Tamil Nadu.

REFERENCE:

https://www.agi.in/dashboard/india/tamil-nadu

https://airquality.cpcb.gov.in/AQI_India/