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-air-quality-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 k-means 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

The samples are collected from NAMP stations are analysed for the Respirable Suspended Particulate matter (RSPM) and gaseous pollutants such as Sulphur dioxide(SO2) and Nitrogen dioxides(NO2)

Data analysis

ANOVA (one way), Tukey HSD, and Pearson correlation coefficient (*r*) were computed using self-coded software on Microsoft Excel 2019 to statistically analyze the collected data.

ALGORITHMS USED

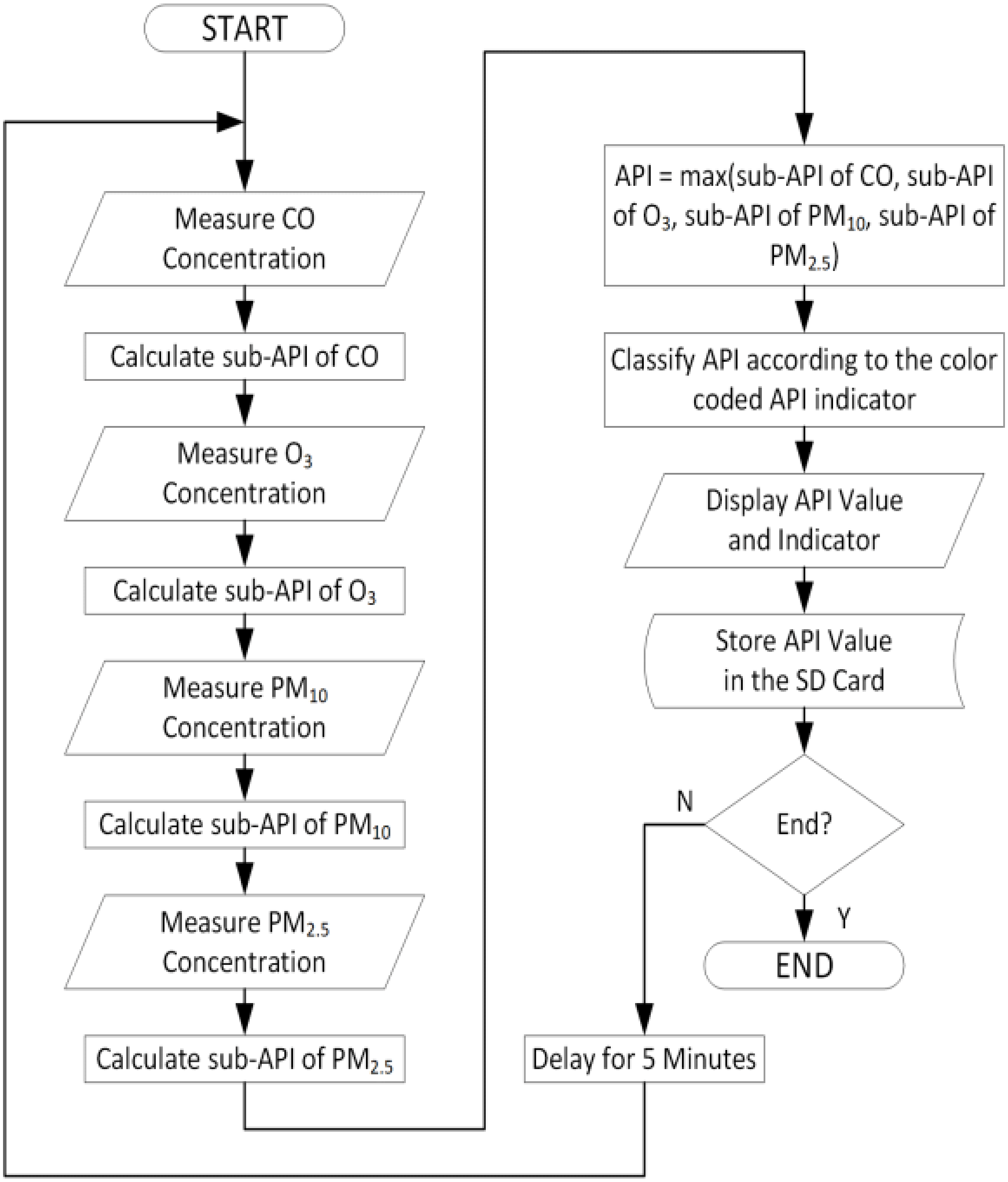
Apply clustering algorithms like K-Means, DBSCAN, or hierarchical clustering to segment customers.

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.

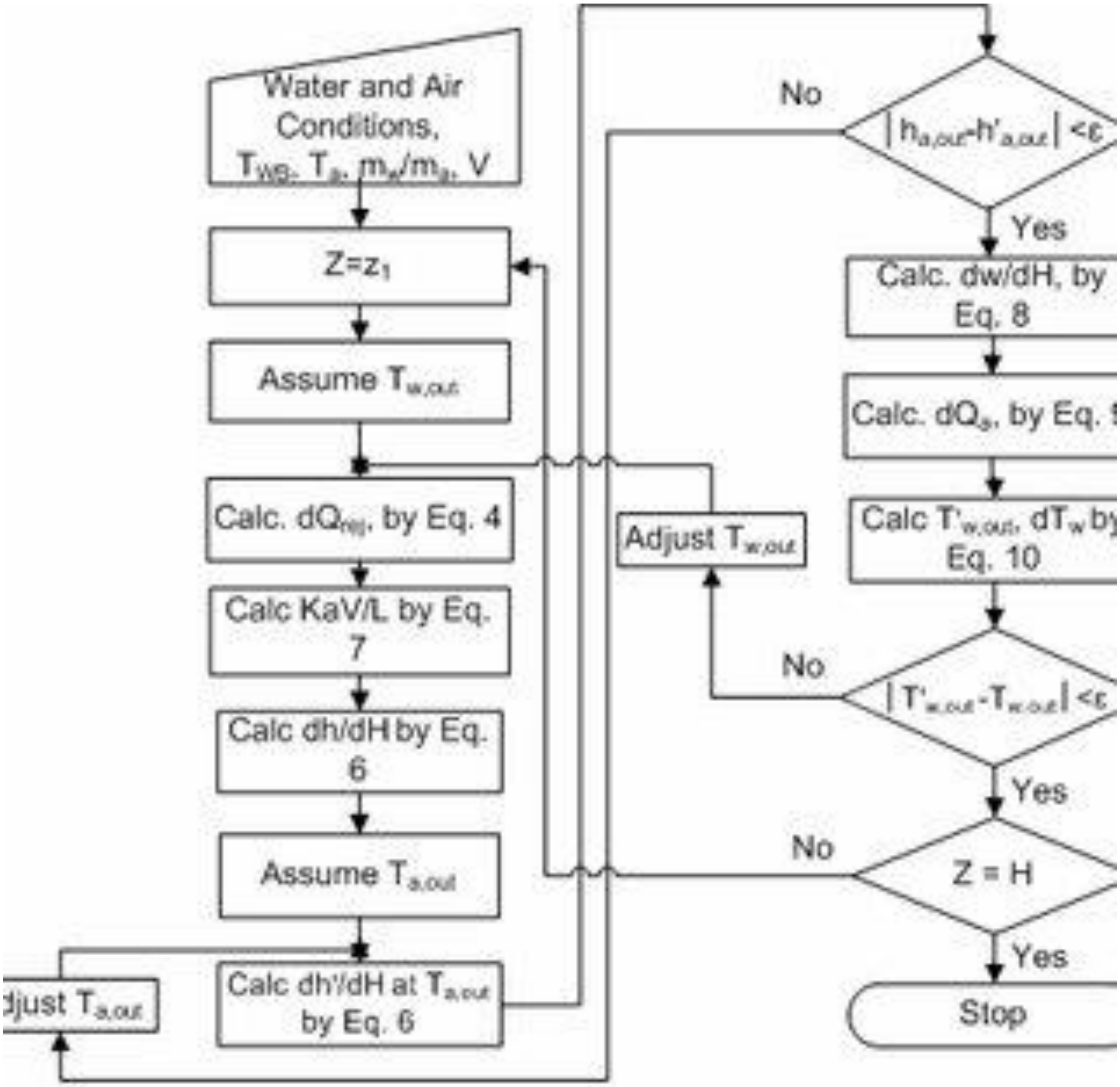
DESIGN AND DATAFLOW

1.Physical data flow diagram:

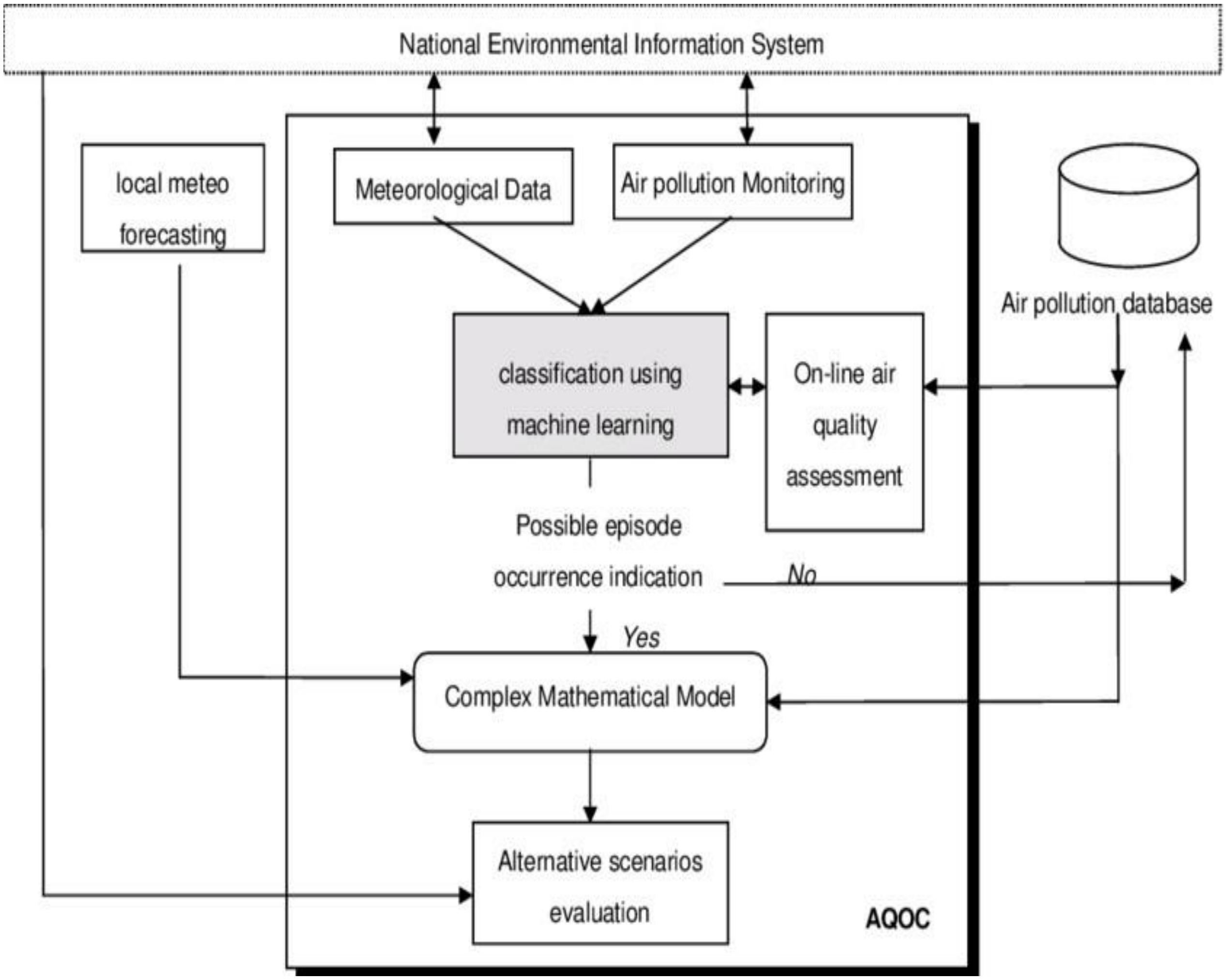
AIR QUALITY ANALYSIS AND PREDICTION
IN TAMIL NADU


START
Measure CO Concentration
Calculate sub-API of CO
Measure O_3 Concentration
Calculate sub-API of O_3
Measure PM_{10} Concentration
Calculate sub-API of PM_{10}
Measure $PM_{2.5}$ Concentration
Calculate sub-API of $PM_{2.5}$
API = max(sub-API of CO, sub-API of O_3, sub-API of PM_{10}, sub-API of $PM_{2.5}$)
Classify API according to the color coded API indicator
Display API Value and Indicator
Store API Value in the SD Card
End?
N
Y
END
Delay for 5 Minutes

2.Logical data flow diagram:


Water and Air Conditions, T_{WB}, T_a, m_w/m_a, V
$Z=z_1$
Assume $T_{w,out}$
Calc. dQ_{rej}, by Eq. 4
Calc KaV/L by Eq. 7
Calc dh/dH by Eq. 6
Assume $T_{a,out}$
Calc dh'/dH at $T_{a,out}$ by Eq. 6
djust $T_{a,out}$
Adjust $T_{w,out}$
No
$|h_{a,out}-h'_{a,out}| <\varepsilon$
Yes
Calc. dw/dH, by Eq. 8
Calc. dQ_a, by Eq. 9
Calc $T'_{w,out}$, dT_w by Eq. 10
No
$|T'_{w,out}-T_{w,out}| <\varepsilon$
Yes
No
Z = H
Yes
Stop

3. Data flow diagram


National Environmental Information System
local meteo forecasting
Meteorological Data
Air pollution Monitoring
Air pollution database
classification using machine learning
On-line air quality assessment
Possible episode occurrence indication
No
Yes
Complex Mathematical Model
Alternative scenarios evaluation
AQOC

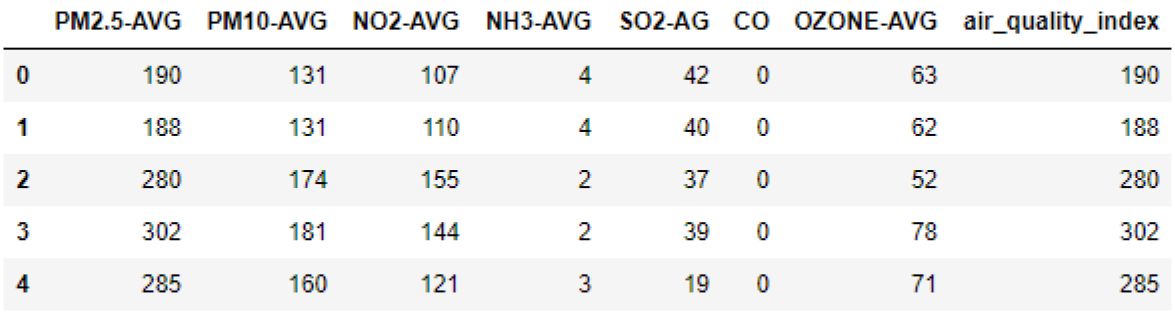
Code:

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 AQI

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:



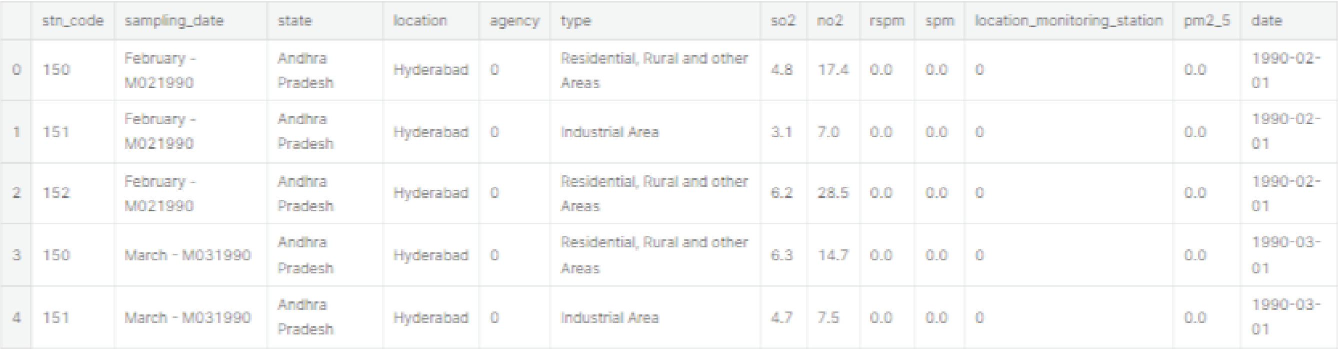
```
   PM2.5-AVG  PM10-AVG  NO2-AVG  NH3-AVG  SO2-AG  CO  OZONE-AVG  air_quality_index
0        190       131      107        4      42   0         63                190
1        188       131      110        4      40   0         62                188
2        280       174      155        2      37   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