Department of Computer Science and Engineering (Data Science)

Name: Ayush Jain SAP-ID: 60004200132 Branch: Computer Engineering

Machine Learning Minors – Mini Project Task 1

Problem Statement: Air Quality Prediction for Indian Cities

India Air Quality Data

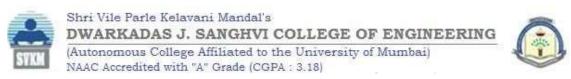
Context : Since industrialization, there has been an increasing concern about environmental pollution. As mentioned in the WHO report 7 million premature deaths annually linked to air <u>pollution</u>, air pollution is the world's largest single environmental risk. Moreover as reported in the NY Times article, <u>India's Air Pollution Rivals China's as World's Deadliest</u> it has been found that India's air pollution is deadlier than even China's.

Using this dataset, one can explore India's air pollution levels at a more granular scale.

Data Content: This data is combined(across the years and states) and largely clean version of the <u>Historical Daily Ambient Air Quality Data</u> released by the Ministry of Environment and Forests and Central Pollution Control Board of India under the National Data Sharing and Accessibility Policy (NDSAP).

The data attributes are as followed,

- 1. 'stn code': The station code,
- 2. 'sampling date': The date when the entry was made,
- 3. 'state': Name of the State of the entry made,
- 4. 'location': City name,
- 5. 'agency': Name of State Pollution Control Board from which the entry was taken,
- 6. 'type': The type of area region for which the entry was calculated,
- 7. 'so2': The SO2 % in air,
- 8. 'no2': The NO2 % in air.
- 9. 'rspm': The Respirable Suspended Particulate Matter % in air,
- 10. 'spm': The Suspended Particulate Matter % in air,
- 11. 'location monitoring station': Location of the monitoring station,
- 12. 'pm2 5' : PSI 2.5 and
- 13. 'date': Date of recording



Department of Computer Science and Engineering (Data Science)

```
df.info()
# Checking the over all information on the dataset.
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435742 entries, 0 to 435741
Data columns (total 13 columns):
    Column
                                 Non-Null Count Dtype
                                 _____
___
                                                 ____
    stn_code
                                 291665 non-null object
 0
                                 435739 non-null object
    sampling_date
 1
 2
    state
                                 435742 non-null object
                                 435739 non-null object
 3
    location
                                 286261 non-null object
 4
    agency
                                 430349 non-null object
 5
    type
 6
    so2
                                 401096 non-null float64
                                 419509 non-null float64
 7
    no2
                                 395520 non-null float64
 8
    rspm
                                 198355 non-null float64
 9
     spm
 10 location_monitoring_station 408251 non-null object
                                 9314 non-null
 11 pm2_5
                                                 float64
 12 date
                                 435735 non-null object
dtypes: float64(5), object(8)
memory usage: 43.2+ MB
```

Dataset Link: https://www.kaggle.com/datasets/shrutibhargava94/indi a-air-quality-data