Air Pollutant Concentration Prediction over Ahmedabad Using Machine Learning

CSE523 - Machine Learning
Winter Semester 2023
Weekly Report - 5/3/2023

Yash Dahima - AU2129002

The tasks of *literature review* and *dataset compilation* have been performed this week.

Paper:

Kumar, K., Pande, B.P. Air pollution prediction with machine learning: a case study of Indian cities. Int. J. Environ. Sci. Technol. (2022). https://doi.org/10.1007/s13762-022-04241-5

Review:

The research paper investigates the air quality analysis and prediction in 23 Indian cities for a period of six years. The study is motivated by the adverse health effects of air pollution caused by industrial, transport, and domestic activities. The authors used machine learning techniques for predicting air quality, which are more efficient than traditional methods. The authors cleaned and preprocessed the dataset and used correlation-based feature selection to filter AQI affecting pollutants. The exploratory data analysis methods were exercised to find hidden patterns present in the dataset, and it was found that almost all pollutants exhibited a significant fall in 2020. The data imbalance problem was addressed using the SMOTE analysis. The authors employed five machine learning models for predicting air quality, and the XGBoost model performed the best among them. The authors compared the results of these models using standard performance parameters, and the Gaussian Naive Bayes model achieved the highest accuracy, while the Support Vector Machine model exhibited the lowest accuracy.

The authors conclude that air quality prediction is a challenging task due to the dynamic environment, unpredictability, and variability in space and time of pollutants. Consistent air quality monitoring and analysis are necessary, especially in developing countries like India, where air pollution has severe consequences on humans, animals, plants, monuments, climate, and the environment. The authors suggest that the present research contributes to the literature by addressing air quality analysis and prediction for India, which might have not been properly studied. The study can be extended by employing deep learning techniques for AQI prediction.

In summary, the research paper highlights the importance of air quality monitoring and analysis, especially in developing countries like India, and proposes machine learning techniques for predicting air quality. The authors analyze the air pollution data of 23 Indian cities for six years and apply various techniques to preprocess and analyze the dataset. The authors compare the performance of five machine learning models and find that the XGBoost model performs the best among them. The authors suggest that

the present research contributes to the literature by addressing air quality analysis and prediction for India, which might have not been properly studied, and can be extended by employing deep learning techniques.

Dataset Compilation:

The time series data of various meteorological and aerosol parameters such as temperature, humidity, wind speed, aerosol optical depth, particulate matter concentration, water vapor, etc., has been downloaded from the CAMS and ERA5 reanalysis data center.

(https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-reanalysis-eac4?tab=overview, https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview)

The data was in a *netCDF* format. The data files corresponding to different parameter were read, combined and converted into an excel file using a python program shown below:

```python
# -*- coding: utf-8 -*-
"""
Created on Fri Feb 24 21:56:09 2023

@author: Yash Dahima
"""

import xarray as xr, pandas as pd, os, numpy as np

ds1 = xr.open_mfdataset('C:/Users/Yash Dahima/PhD/Course Work/ML/Project/AQI/Datasets/CAMS Dataset/1/*.nc').mean(dim=['latitude',
'longitude']).sel(time=slice('2011-01-01T00:00:00', '2019-12-31T21:00:00'))
ds2 = xr.open_mfdataset('C:/Users/Yash Dahima/PhD/Course Work/ML/Project/AQI/Datasets/CAMS Dataset/2/*.nc').mean(dim=['latitude', 'longitude'])

ds = xr.merge([ds1, ds2], compat='override')

wind_speed = np.sqrt(np.square(ds.u10)+np.square(ds.v10)) # m/s
wind_direction = (np.rad2deg(np.arctan2(ds.v10, ds.u10))) % 360 # in degrees , Northward = 0, Eastward = 90
dew_temp = ds.d2m - 273.15 # celcius
temp = ds.t2m - 273.15 # celcius
pm2p5 = ds.pm2p5*1e9 # ug/m3
pressure = ds.sp/100 #hPa
water_vapour = ds.tcwv # Total column vertically-integrated water vapour (kg m**-2)


blh = xr.open_mfdataset('C:/Users/Yash Dahima/PhD/Course Work/ML/Project/AQI/Datasets/ERA5 Dataset/*.nc')
blh = blh.sel(expver=1).combine_first(blh.sel(expver=5))
blh = blh.mean(dim=['latitude', 'longitude']).blh
blh = blh.sel(time=slice('2011-01-01T00:00:00', '2019-12-31T23:00:00'))
blh = blh.resample(time='3H').mean()



ml_dataset = pd.DataFrame()

ml_dataset.index, ml_dataset['ws'], ml_dataset['wd'], ml_dataset['temp'], ml_dataset['dew_temp'], ml_dataset['pressure'], ml_dataset['z'],
ml_dataset['wv'], ml_dataset['blh'], ml_dataset['bcaod550'], ml_dataset['duaod550'], ml_dataset['omaod550'], ml_dataset['ssaod550'],
ml_dataset['suaod550'], ml_dataset['aod469'], ml_dataset['aod550'], ml_dataset['aod670'], ml_dataset['aod865'], ml_dataset['aod1240'],
ml_dataset['pm2p5'] = temp.time + pd.Timedelta(hours=5, minutes=30), wind_speed.data, wind_direction.data, temp.data, dew_temp.data,
pressure.data, ds.z.data, water_vapour.data, blh.data, ds.bcaod550.data, ds.duaod550.data, ds.omaod550.data, ds.ssaod550.data, ds.suaod550.data,
ds.aod469.data, ds.aod550.data, ds.aod670.data, ds.aod865.data, ds.aod1240.data, pm2p5.data # time in IST

ml_dataset.to_excel("C:/Users/Yash Dahima/PhD/Course Work/ML/Project/AQI/Datasets/data.xlsx")
```

The tasks of *EDA* and *feature analysis* will be performed next week.