Smog Prediction

Abdul Saboor (20L-1113)
Minal Urooj (20L-1388)

Introduction

This project is about predicting the time and intensity in which photochemical smog, otherwise known as Smog, might occur in particular areas or regions worldwide. This will be further helpful in developing an application that will work on real-time data. This application enables real-time analytics by combining air pollution and weather data. Once this issue of accurate and acceptable predictions is solved, users can visualize past data trends and future forecasts. However, remember that this might take some time as the project is still in its early stages, and the prime focus right now is an accurate prediction based on the collected dataset.

Dataset Availability

The dataset is being collected using the APIs from the following sources

- https://openweathermap.org/ is used to get information about
  `aqi, co, no, no2, o3, so2, pm2_5, pm10 and nh3`
- https://open-meteo.com/ is used to get information about
  `Temperature and dewpoint`

This data is mainly in the most refined form, as the APIs provide a clean and continuous data stream. The data from both sources is merged after a small pre-processing, including synchronization of date and location coordinates between both data sources.

Proposed Solution

Currently, the main task is to accurately predict the upcoming pollution patterns and attain an acceptable accuracy score. This might be evaluated among these different models:

- SARIMAX
- Neural Prophet
- Multivariable Adversarial Learning
- Other LSTM variants based on permissible time

The entire code of this dataset is being built from scratch, as no previous implementations have been found for the particular use case. However, the portion of code containing the prediction algorithm's performance will be from libraries and, therefore, might be already available in other forms on the internet. The links for pieces of code referenced from the internet will be included along with the code snippets.

Proposed Timeline

The dataset will continuously increase, adding the data as it becomes available. Work on implementing and testing the models will begin as soon as sufficient data is collected. The proposed timeline for a concrete implementation of the model after evaluation, among others, is around six weeks and shall be completed by April 20, 2023. A week's work will be dedicated to initial visualization, and this project's pilot phase will be completed by April 30, 2023.