Air Aware: AQI Prediction and Analysis

Implementing data-driven techniques to predict air quality index and inform public health decisions

Introduction:

Analysing the Air Quality Index (AQI) is a vital part of environmental data science that focuses on tracking and assessing air quality in an area. The AQI is essential as it provides a clear and consistent metric for assessing and communicating air pollution levels (Medium 2023). Poor air quality is linked to various health problems, including respiratory and heart diseases, and poses greater risks to vulnerable groups such as children, the elderly, and those with existing health conditions. The AQI offers a simple numerical value that helps individuals decide when to limit outdoor activities and enables health authorities to respond promptly to pollution episodes. It supports policymakers in creating and evaluating regulations aimed at reducing pollution, contributing to a healthier environment for everyone. The goal of this project is to generate a numerical indicator using predictive modelling, that reflects the general state of the air of a city. The project will be made using Python in either a Jupyter or a Colab notebook.

Imagine the AQI as a scale ranging from 0 to 500. The higher the AQI number, the more polluted the air.

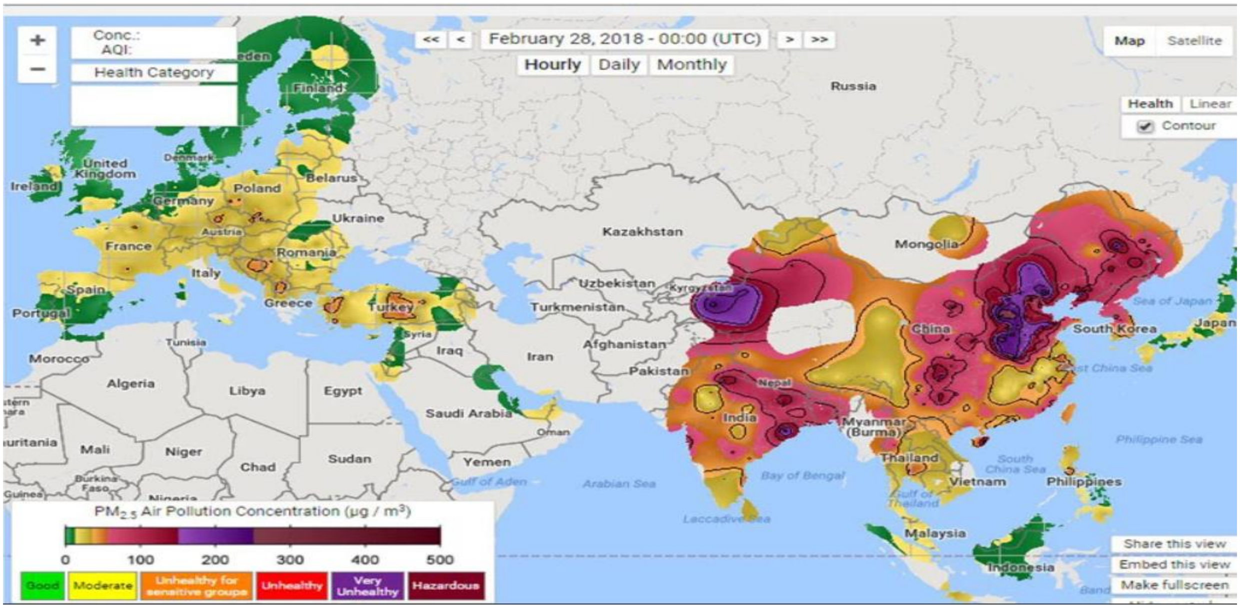

Fig 1. Real-time map of air pollution in world

Data and project description:

Q 1: Is the collected data satisfying the 4 Vs?

A: The 4Vs of any Big Data analysis project are Volume, Velocity, Veracity, and Variety. The data collected from this project comes from various sources. I will be using a total of six datasets for predictive modelling. Each dataset is in the csv form and consists of 4 columns, which are as follows:

1. **Date**: The date on which the data was recorded, with the format MM/DD/YYYY.
2. **Time**: The time at which the data was recorded, with an hourly resolution.
3. **PM2.5**: The concentration of PM2.5 particles in the air, measured in micrograms per cubic meter (µg/m^3).
4. **PM2.5 AQI**: The Air Quality Index value associated with the PM2.5 concentration.

The 4 Vs for this dataset (Tutiempo) can be described as given below:-

1. Volume: The dataset consists of a total of 43, 757 observations
2. Velocity: The dataset consists of time-series data, at an hourly rate over a period of 6 years (from 2013 to 2018, both included).
3. Veracity: There are missing and erroneous values in the dataset. There might be wrong entries, repetitions and other abnormalities too. The project aims to achieve veracity through data cleaning and feature engineering.
4. Variety: While the current dataset is temporally detailed, additional variety could be added by including spatial data, such as different locations or regions, to analyze geographic variations in air quality.

Q 2: How has the annual average AQI changed over the past 6 years?

A: To answer this question, I need to aggregate the data after the pre-processing step. This can be done by calculating the average AQI of each year. Then, I can analyze the average AQI of each year and create visual presentations such as line graphs to illustrate the changes in annual average AQI over a period of time. I can use these graphs to recognize patterns and trends and consecutively summarize my findings.

Q 3: Which machine learning algorithm will work the best?

A: Data will be split into train and test data. The project aims to predict AQI using different regression models like Linear Regression, Lasso and Ridge regression, Decision Tree Regressor, KNN Regressor, Random Forest Regressor, XG Boost Regressor, Hyperparameter tuning, and ANN. Then, I can compare using evaluation metrics like F1 score, precision or recall to understand which algorithm will work the best for our task.

Backup Questions

Q B1: What is my backup data source?

A: My backup dataset (World Health organization 2022) comprises air quality monitoring information collected by the World Health Organization (WHO) across various regions and countries. It includes several columns detailing the WHO region, country name, city or locality, and the measurement year. The dataset provides specific air quality measurements, such as PM2.5 and PM10 concentrations (in µg/m^3), along with their temporal coverage percentages, indicating how consistently data was recorded over the period. Additionally, it includes NO2 concentrations and their temporal coverage percentages. The data is referenced from different sources, primarily the European Environmental Agency. The version of the database is 2022. The status column indicates whether the data is complete or has certain qualifiers. This dataset is essential for analyzing air quality trends and aiding in environmental policy-making on a global scale.

Q B2: Can machine learning algorithms effectively predict AQI and help in a better policymaking?

A: Machine learning models are able to predict future AQI levels with high accuracy by utilizing a multitude of environmental and socio-economic parameters in addition to a massive amount of previous air quality data. With the help of these forecasts, decision-makers can take preventive steps to reduce pollution, allocate resources as efficiently as possible, and safeguard public health.

References:

Tutiempo 2024, World Weather, Tutiempo 2024, viewed 7 June 2024, <https://en.tutiempo.net/>

Aman Kharwal 2023, *Air Quality Index Analysis Using Python*, The Clever Programmer, viewed 7 June 2024, <https://thecleverprogrammer.com/2023/09/18/air-quality-index-analysis-using-python/>

World Health Organization 2022, WHO Ambient Air quality database 2022 v5, viewed 7 June 2024, <https://data.who.int/countries/036>

Air Now n.d., Air Quality Index (AQI) basics, viewed 7 June 2024, <https://www.airnow.gov/aqi/aqi-basics/#:~:text=Think%20of%20the%20AQI%20as,300%20represents%20hazardous%20air%20quality.>