AQI VS DEATH RATE PREDICTION

Nattawaree Piyarat

Brainstation – data science bootcamp

PROBLEM STATEMENT

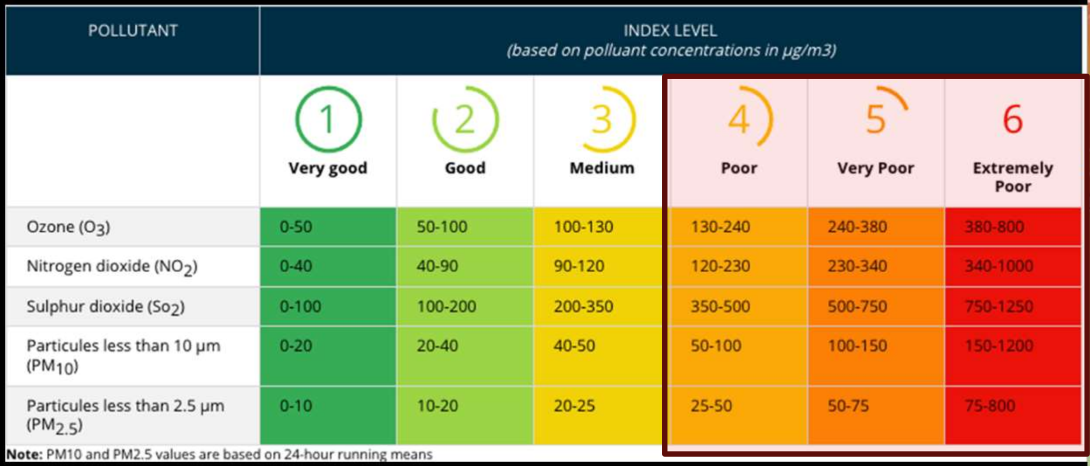
Can machine learning accurately predict mortality rates based on the Air Quality Index (AQI)?

Are there correlations between air quality and health outcomes, particularly mortality rates, and air pollution stems from various sources such as wildfires, dust, vehicular emissions, and industrial activities?

Focus group (selected Asian countries).

- Bangladesh
- Bhutan
- China
- India
- Indonesia
- Sri Lanka
- Thailand

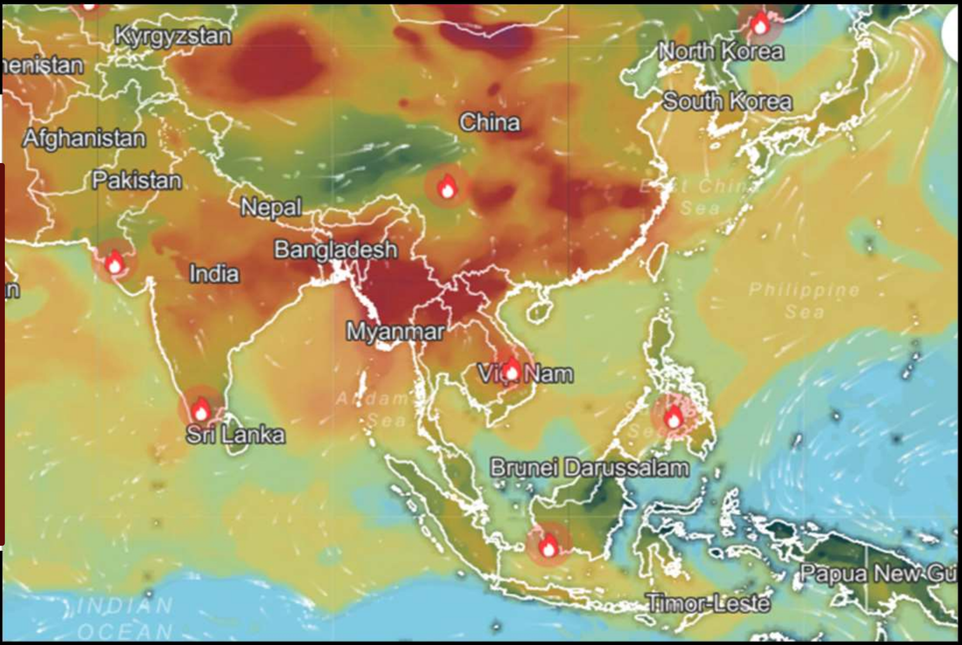
AQI (AIR QUALITY INDEX)



POLLUTANT	INDEX LEVEL *(based on polluant concentrations in µg/m3)*					
	1 Very good	2 Good	3 Medium	4 Poor	5 Very Poor	6 Extremely Poor
Ozone (O_3)	0-50	50-100	100-130	130-240	240-380	380-800
Nitrogen dioxide (NO_2)	0-40	40-90	90-120	120-230	230-340	340-1000
Sulphur dioxide (So_2)	0-100	100-200	200-350	350-500	500-750	750-1250
Particules less than 10 µm (PM_{10})	0-20	20-40	40-50	50-100	100-150	150-1200
Particules less than 2.5 µm ($PM_{2.5}$)	0-10	10-20	20-25	25-50	50-75	75-800

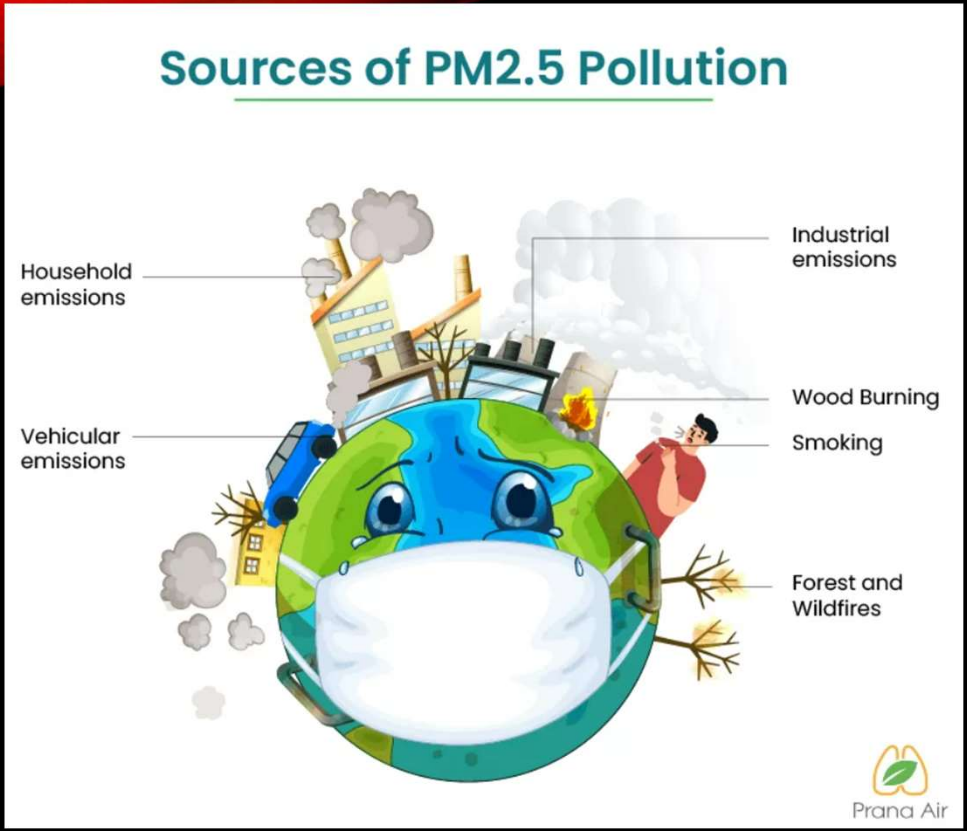
Note: PM10 and PM2.5 values are based on 24-hour running means

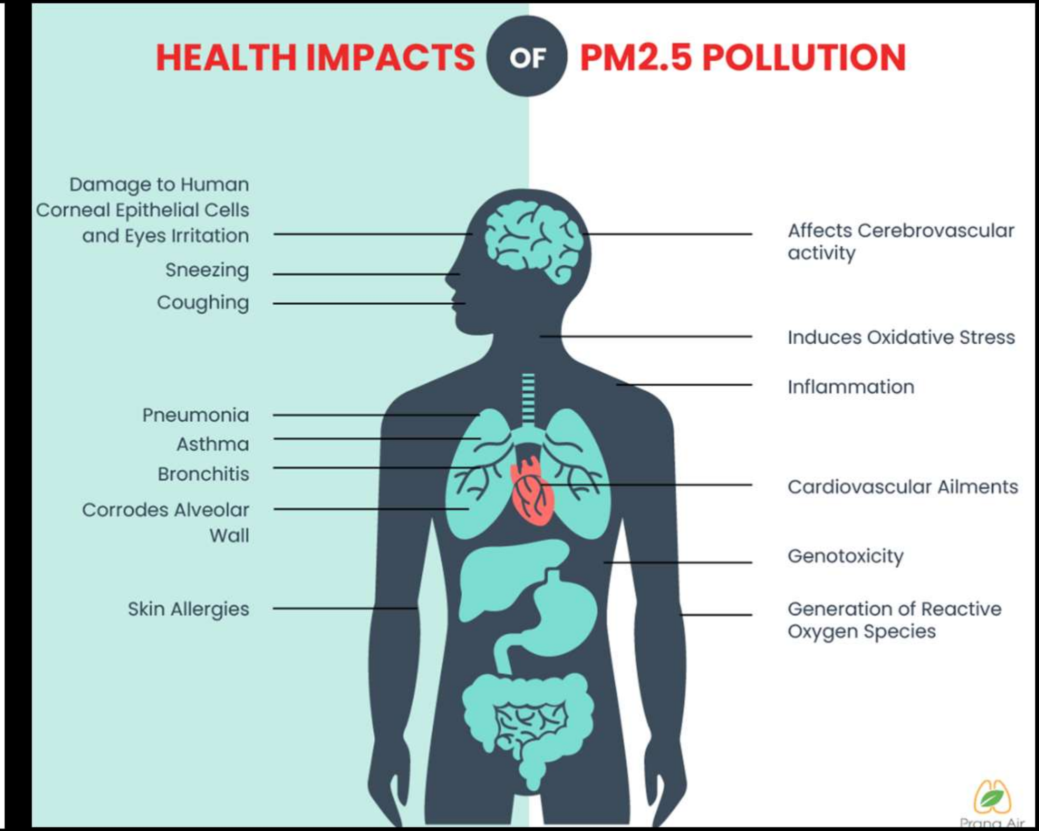
Focus >= Level4



Picture from IQAir website

PM 2.5 meaning: Fine particulate matter is defined as particles that are 2.5 microns or less in diameter


Sources of PM2.5 Pollution
Household emissions
Industrial emissions
Vehicular emissions
Wood Burning
Smoking
Forest and Wildfires
Prana Air


HEALTH IMPACTS OF PM2.5 POLLUTION
Damage to Human Corneal Epithelial Cells and Eyes Irritation
Sneezing
Coughing
Pneumonia
Asthma
Bronchitis
Corrodes Alveolar Wall
Skin Allergies
Affects Cerebrovascular activity
Induces Oxidative Stress
Inflammation
Cardiovascular Ailments
Genotoxicity
Generation of Reactive Oxygen Species
Prana Air

DATA SCIENCE

The machine learning approach will use historical AQI data, and historical death causes data. And predict the relationship between air quality and the number of deaths which is assumed caused by high AQI.

Approach

- Linear Regression
- Logistic Regression
- Time series

The results of this project should raise awareness of air pollution and reduce the number of deaths of people by at least 1%, especially in the concerned areas.

DATA SET

Dataset from WHO database (Air quality index, number of deaths by cause, pollution, dust, wildfires)

Currently, I reference the datasets below.

Air quality index year 2009 – 2019 from WHO

- Country, City
- Year
- Number of PM10, PM2.5, NO2

Number of deaths by cause year 2016 – 2019 from WHO

- Country (no city)
- Year
- Cause of death
- Number of deaths, death rate

Additional dataset (not in this sprint)

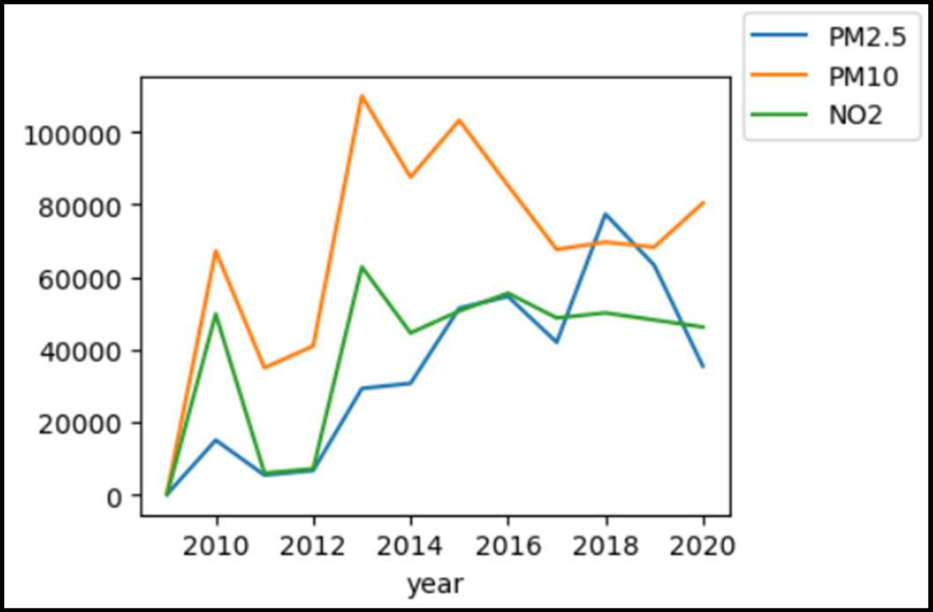
- Historical AQI separated by pollutants to predict AQI

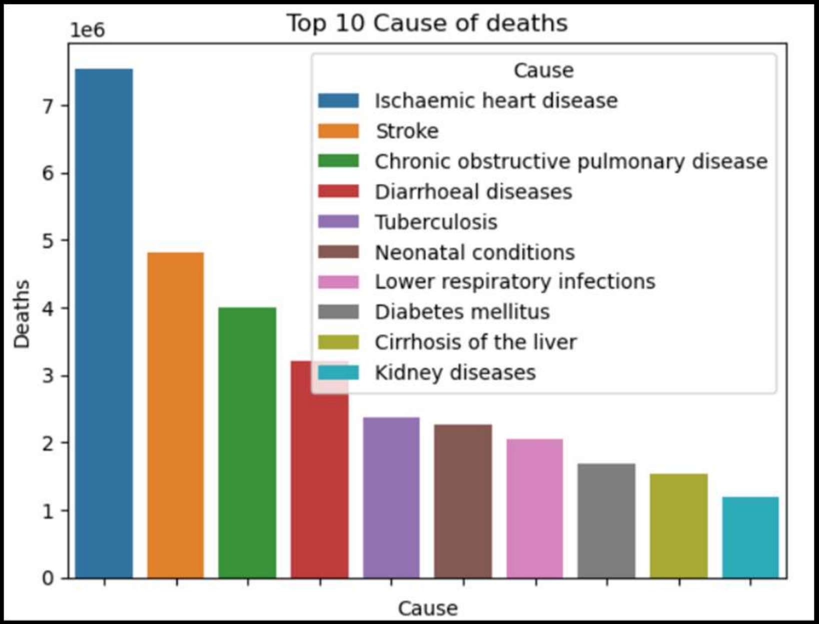
FINDING (1)

PM 10 peaked during 2013 – 2015 and PM2.5 increased and peaked in 2018.

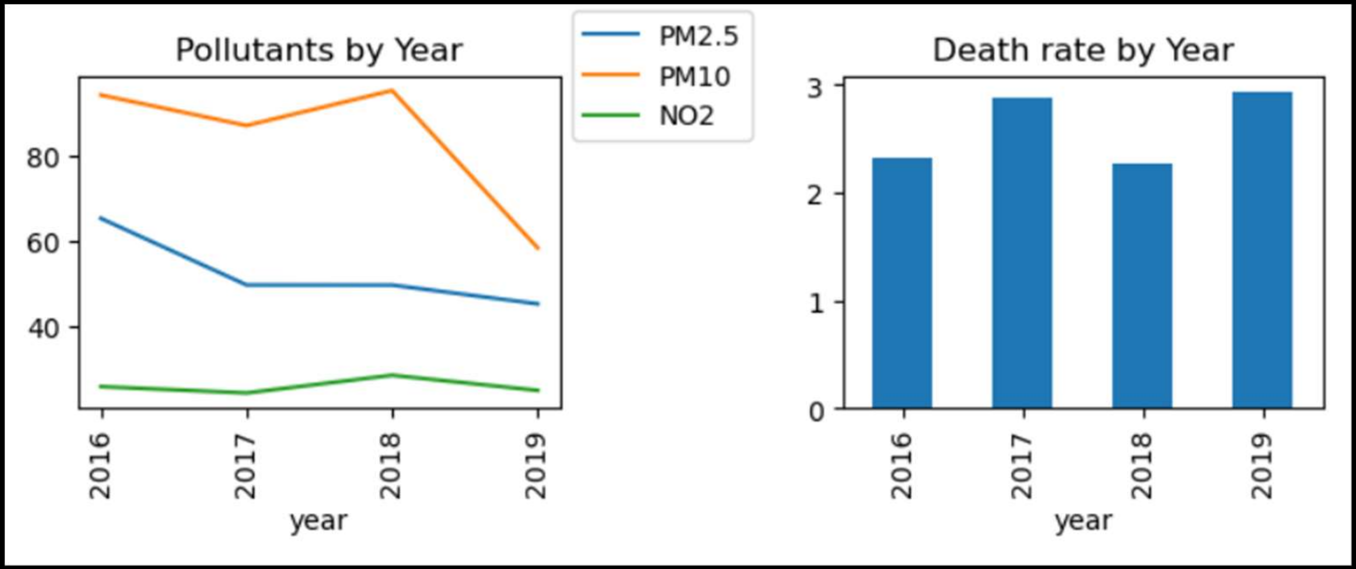
'Ischaemic heart disease' is the top 1 of the cause of death.

The WHO explains that household air pollution is identified as a primary source of outdoor air pollution.



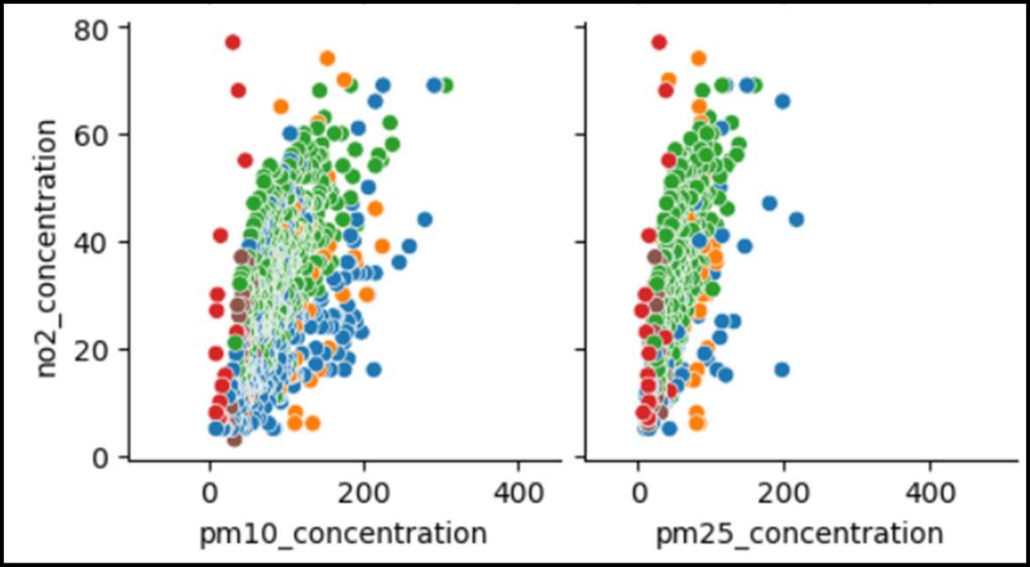


FINDING(2) –AFFECTING THE GOAL



Observation In 2019, pollutant levels tended to decrease, yet the death rate increased. This suggests that pollution may not have an immediate impact but requires several years to affect mortality rates.

FINDING (3)



NO_2, PM10, PM2.5 linear relationship

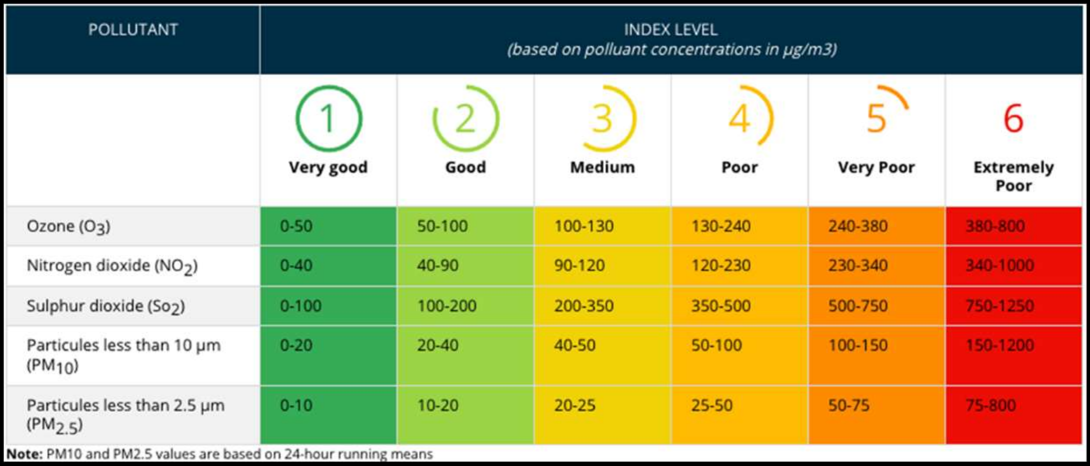
FINDING (4)

Linear Regression	X (features)	Result (R²)	(R²) add China
y = number of deaths	PM10	0.031	0.077
	PM2.5	0.032	0.087
	PM10, PM2.5	0.048	0.127 ★

Focus only on 'Ischaemic heart disease' the Death rate correlates with PM10, PM2.5.

If use all diseases R^2 shows 0.00

Logistic Regression	X	Result (R²)
y = Unhealthy	PM10, PM2.5, NO2	0.77



POLLUTANT	INDEX LEVEL *(based on polluant concentrations in µg/m3)*					
	1 Very good	2 Good	3 Medium	4 Poor	5 Very Poor	6 Extremely Poor
Ozone (O_3)	0-50	50-100	100-130	130-240	240-380	380-800
Nitrogen dioxide (NO_2)	0-40	40-90	90-120	120-230	230-340	340-1000
Sulphur dioxide (So_2)	0-100	100-200	200-350	350-500	500-750	750-1250
Particules less than 10 µm (PM_{10})	0-20	20-40	40-50	50-100	100-150	150-1200
Particules less than 2.5 µm ($PM_{2.5}$)	0-10	10-20	20-25	25-50	50-75	75-800

Note: PM10 and PM2.5 values are based on 24-hour running means

NEXT STEP

- The time delay in observing the impact of the Air Quality Index (AQI) on mortality, my approach involves predicting the AQI from each pollutant or forecasting illness instead.
- Plan to add more datasets that contain many pollutants (O_3, PM2.5, PM10, CO_2, NO_2, CO, SO_2). And illness dataset.
- Using machine learning to predict AQI and illness
    - Linear Regression
    - Time series
    - Decision Tree
    - KNN

Feedback after the presentation (I will consider this for next sprint)

- Change the target to Death rate (percentage) instead of number of deaths. (Arun)
- Change the target to air-bound disease. (Arun)
- Adding some factor that correlated with AQI (Wuyang)
- Any global warming impact AQI (Marco)

Q&A