AQI VS HEALTH (DEATH RATE) PREDICTION

Nattawaree Piyarat

Brainstation – data science bootcamp

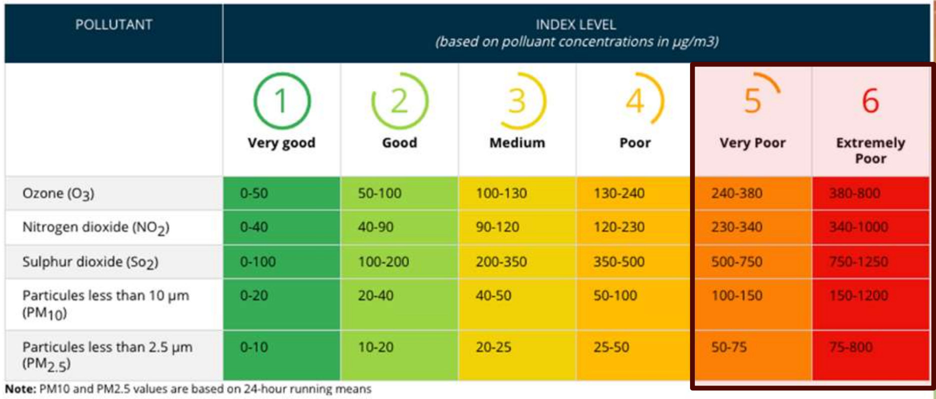
PROBLEM STATEMENT

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?

Focus group (selected Asian countries):

- China
- India
- Thailand

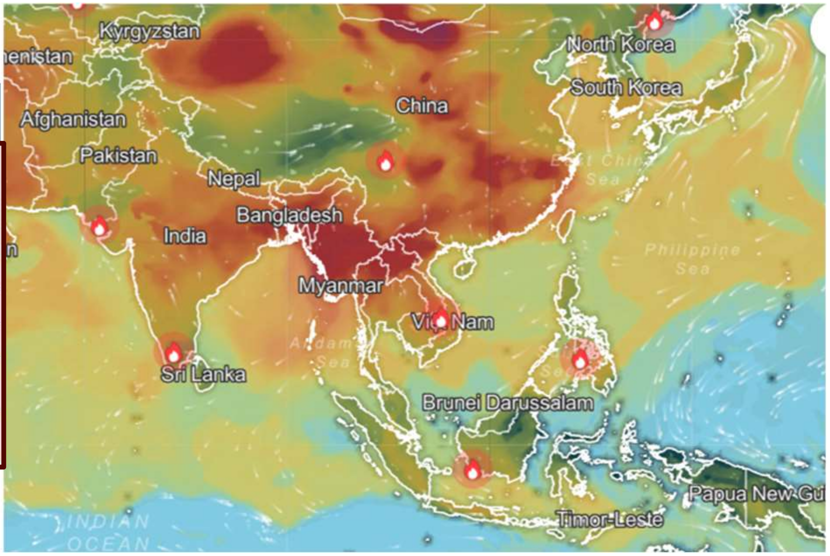
AQI (AIR QUALITY INDEX)



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 >= Level5



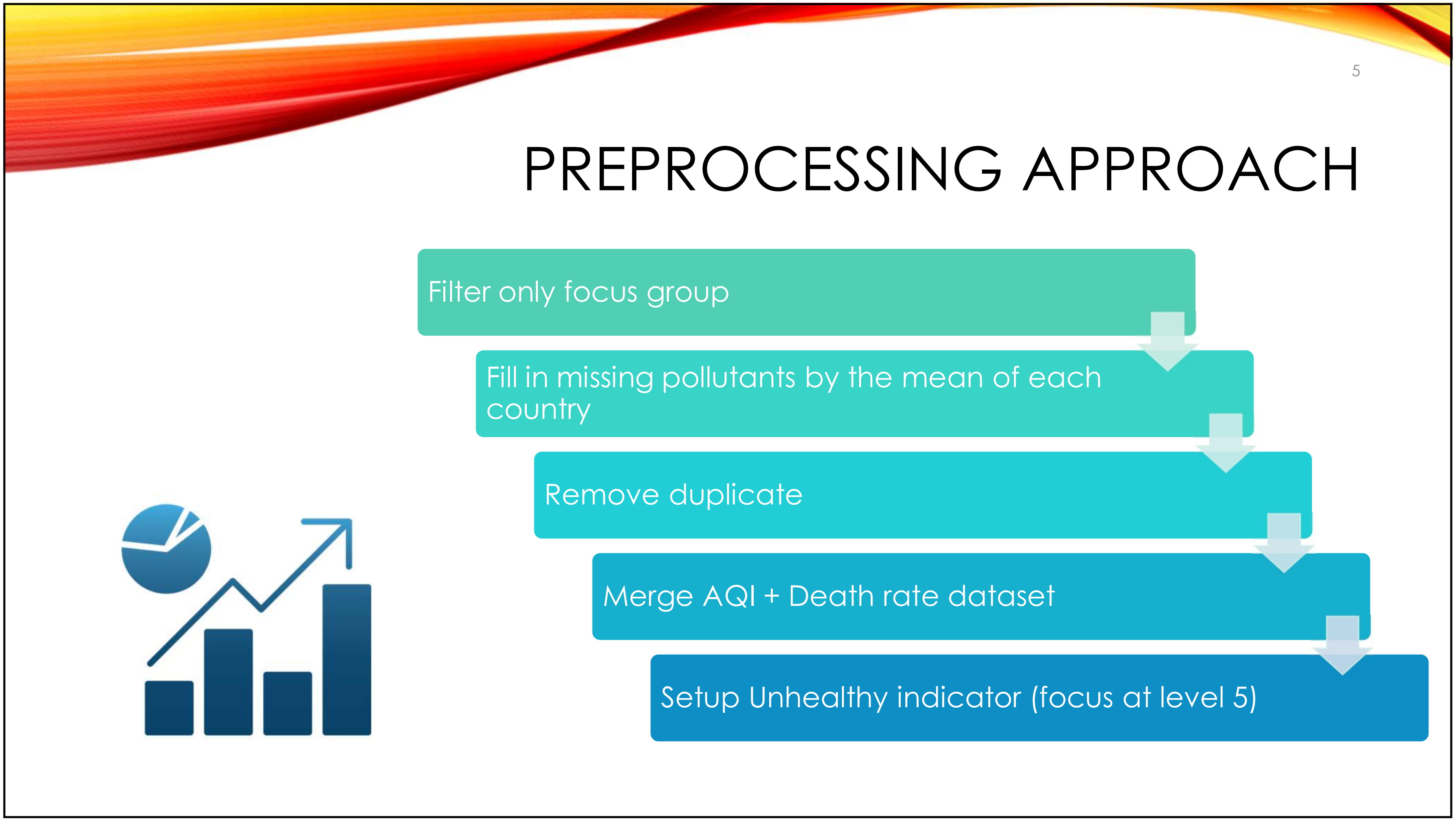
Picture from IQAir website

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

DATA SET

Currently, I reference the datasets below.

1. Air Quality Index (2009–2019) from WHO
   - This dataset has been removed and replaced with another due to the inclusion of more pollutants while retaining similar details.
2. Number of Deaths by Cause (2016–2019) from WHO.
   - Data by country and year
3. Air Quality Open Data Platform (2019–2020): https://aqicn.org/data-platform/covid19/
   - Sourced from The World Air Quality Index Project.
   - The data includes city-specific measurements, aggregated to calculate the mean air quality for each country on each date for this project.



PREPROCESSING APPROACH

Filter only focus group

Fill in missing pollutants by the mean of each country

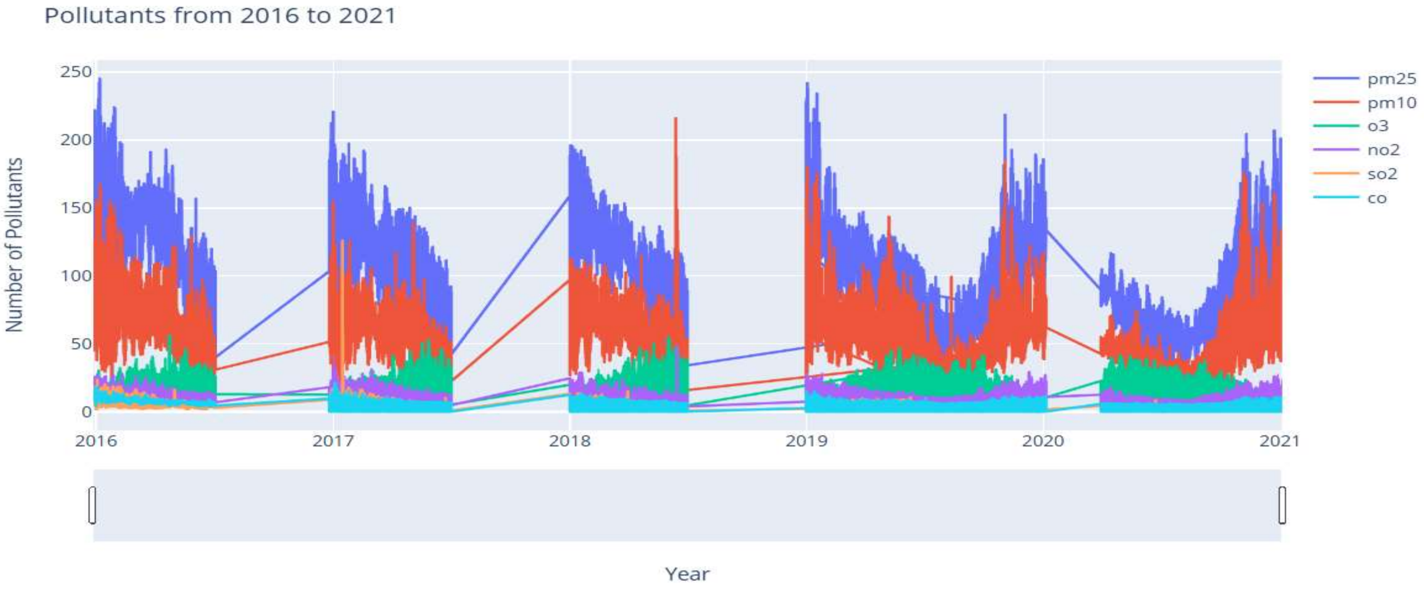
Remove duplicate

Merge AQI + Death rate dataset

Setup Unhealthy indicator (focus at level 5)

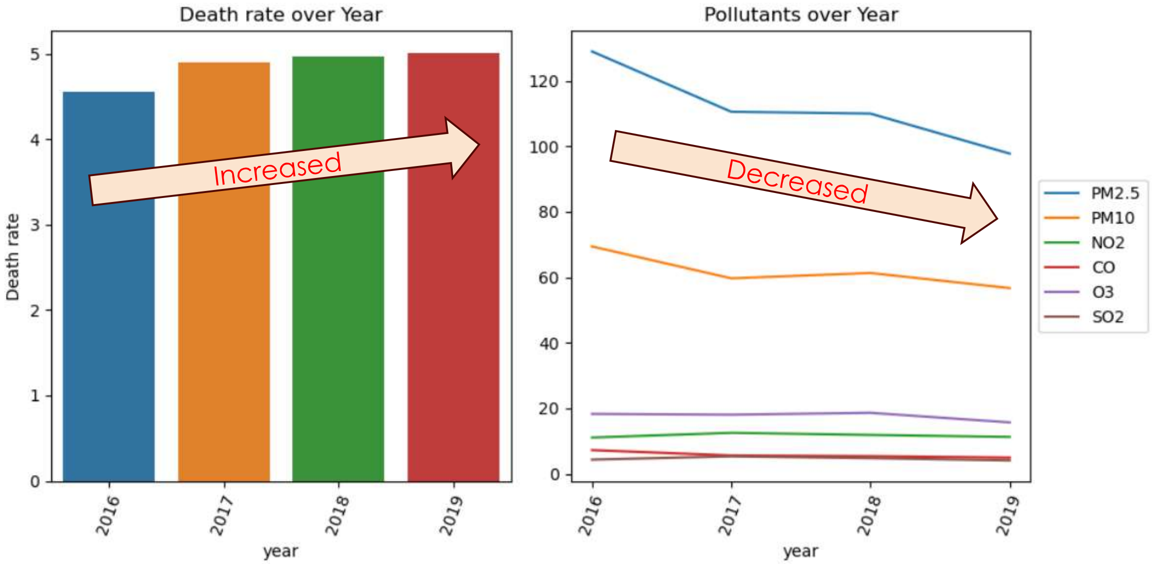
FINDING (1)

From 2016 to 2018, data is available only for the first half of each year, showing a peak in January and a subsequent increase through the mid-year. For 2019 to 2021 Q1, the full-year data follows a similar trend: a peak in January, a gradual decline until mid-year, and a climb back up until the next January peak.

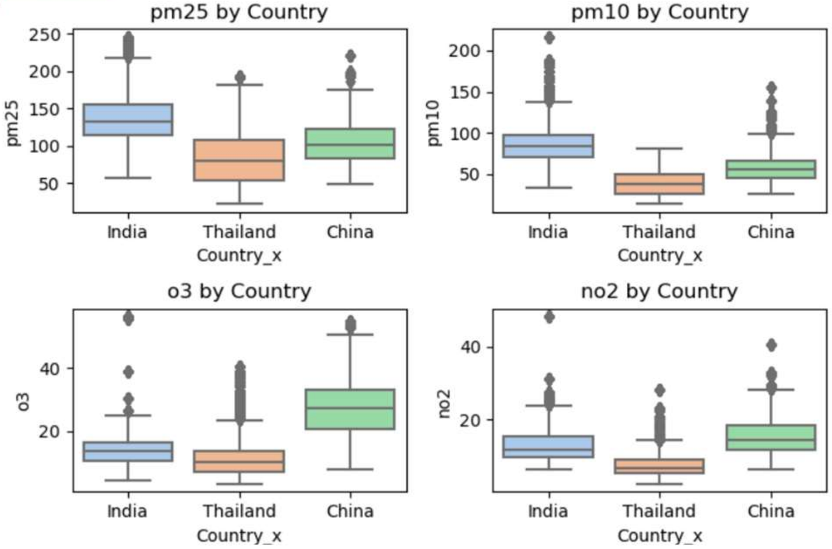


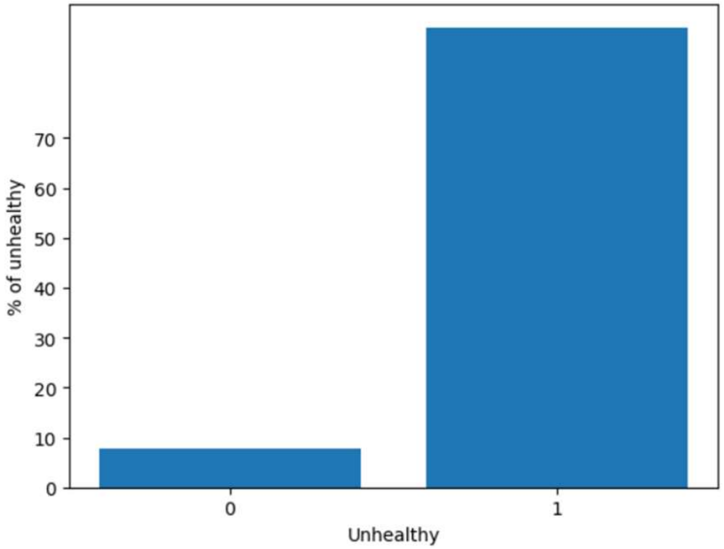
FINDING (2)

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)





Each pollutant has many outliners higher than the mean value meaning some parts of the country produced high pollutants

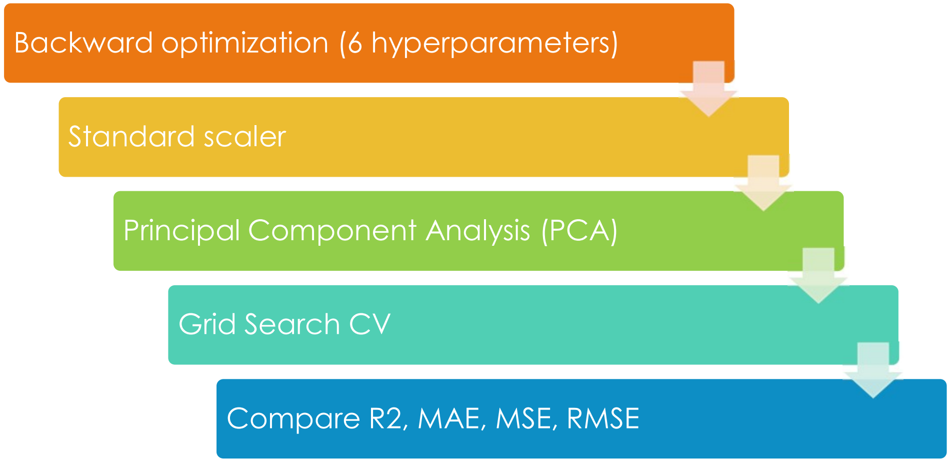
The selected countries show high unhealthy ratio (9:1)

MODEL COMPARISON

CANDIDATES

- Logistic Regression (Baseline)
- Decision Tree
  - Max depth 5, 7
- K-nearest Neighbor
  - Neighbors : 5, 9
- Xgboost
  - binary:logistic

STEPS



Backward optimization (6 hyperparameters)

Standard scaler

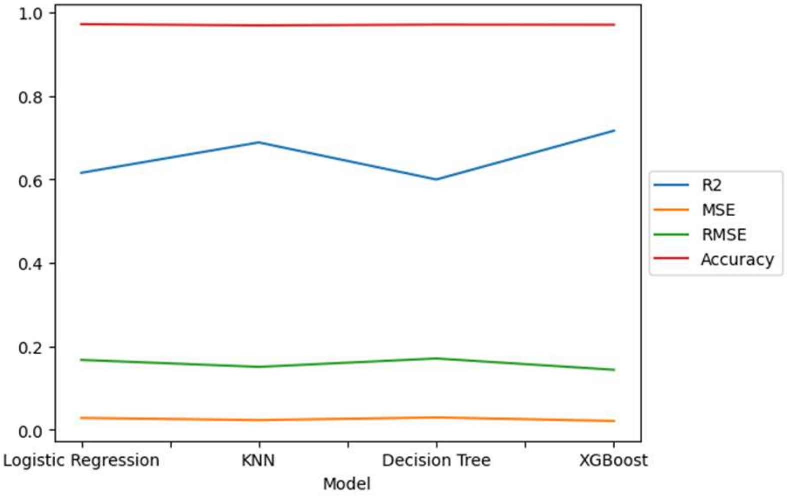
Principal Component Analysis (PCA)

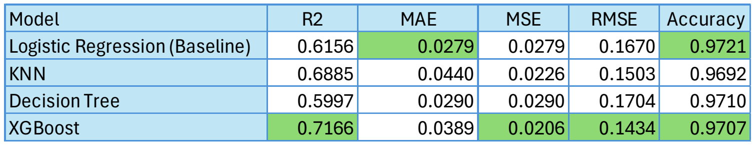
Grid Search CV

Compare R2, MAE, MSE, RMSE

MODEL COMPARISON

- Logistic Regression
- Decision Tree
- K-Nearest Neighbor
- XGBoost



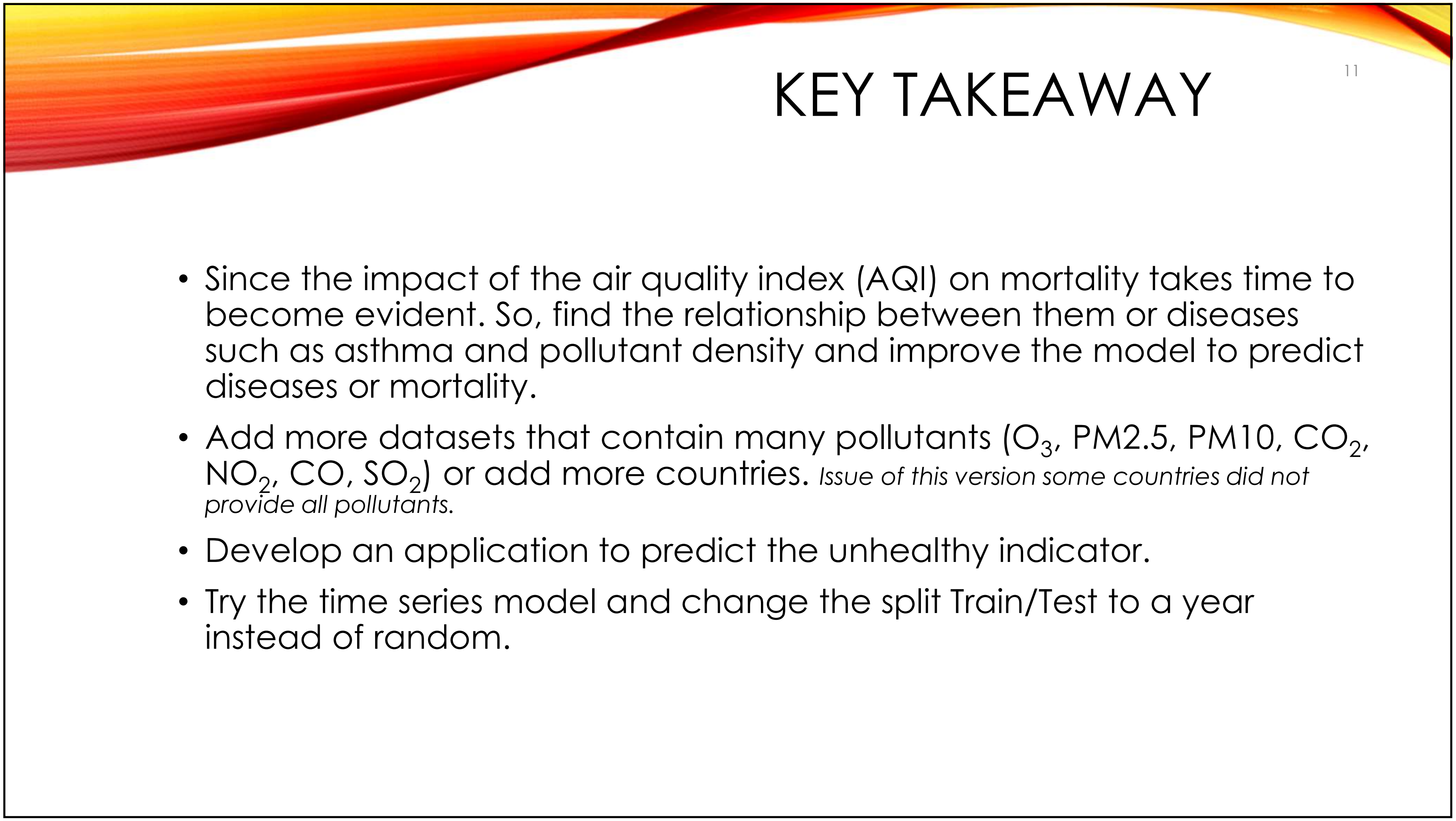


Model	R2	MAE	MSE	RMSE	Accuracy
Logistic Regression (Baseline)	0.6156	0.0279	0.0279	0.1670	0.9721
KNN	0.6885	0.0440	0.0226	0.1503	0.9692
Decision Tree	0.5997	0.0290	0.0290	0.1704	0.9710
XGBoost	0.7166	0.0389	0.0206	0.1434	0.9707

🏅 **Selected Model is XGBoost:**

Reasons:

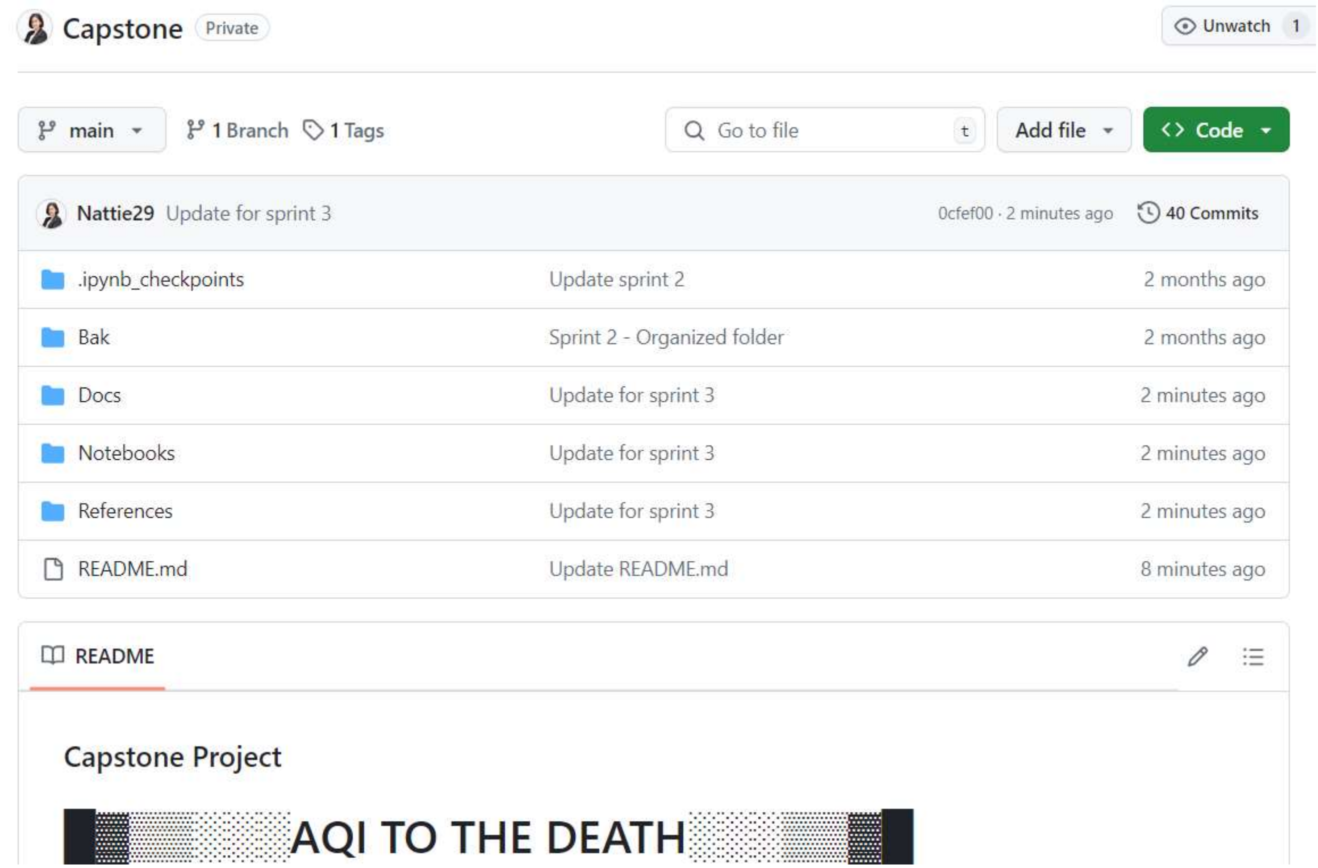
- A good score of R^2 should be close to 1
- MAE and MSE should close to 0
- RMSE, the lower is better
- The above table shows that each model’s results are almost the same but the XGBoost is the best from those 4 models.



KEY TAKEAWAY

- Since the impact of the air quality index (AQI) on mortality takes time to become evident. So, find the relationship between them or diseases such as asthma and pollutant density and improve the model to predict diseases or mortality.
- Add more datasets that contain many pollutants (O_3, PM2.5, PM10, CO_2, NO_2, CO, SO_2) or add more countries. *Issue of this version some countries did not provide all pollutants.*
- Develop an application to predict the unhealthy indicator.
- Try the time series model and change the split Train/Test to a year instead of random.

12
THANK YOU
Q&A



Nattie29 Update for sprint 3 0cfef00 · 2 minutes ago 40 Commits

.ipynb_checkpoints	Update sprint 2	2 months ago
Bak	Sprint 2 - Organized folder	2 months ago
Docs	Update for sprint 3	2 minutes ago
Notebooks	Update for sprint 3	2 minutes ago
References	Update for sprint 3	2 minutes ago
README.md	Update README.md	8 minutes ago

README

Capstone Project

AQI TO THE DEATH