CleanAir Futures International

Created by: Kari Hattabaugh, Jessica Davis, Manali Sudhir Chavaj, Daniel Nagy

TABLE OF CONTENTS

01
Introduction

Details about the business problem

02
Data processing

How we have cleaned and prepared our data

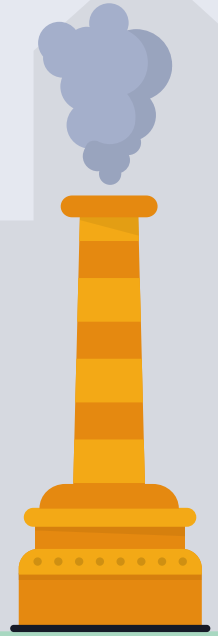
03
Our models

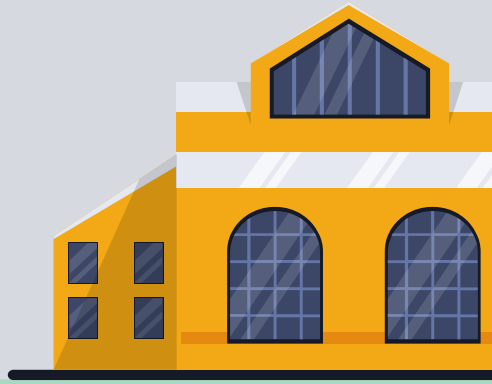
Key features and statistical outcomes

04
Insights

Business insights and policy recommendation

01
Introduction





Who We Are and What We Do

About CAFI

CleanAir Futures International is a global nonprofit focused on improving air quality and protecting public health through research, innovation, and collaboration

Why This Matter

Air pollution remains one of the world's leading health risks. It is linked to nearly 7 million premature deaths each year (WHO, 2024). Understanding where and how pollution affects communities helps us act where it matters most.

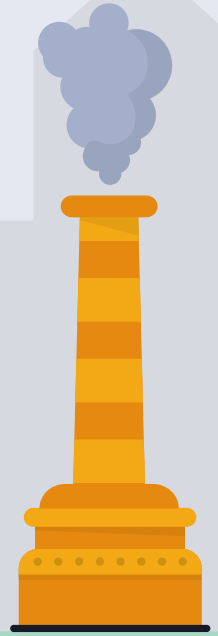
Our Work

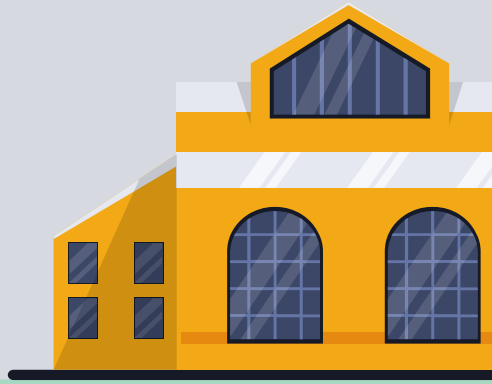
We combined IoT-based air quality data with global health indicators to identify countries facing the greatest environmental and health burdens. Our predictive analysis points to clear priorities for policy action, investment, and targeted interventions.

The Problem

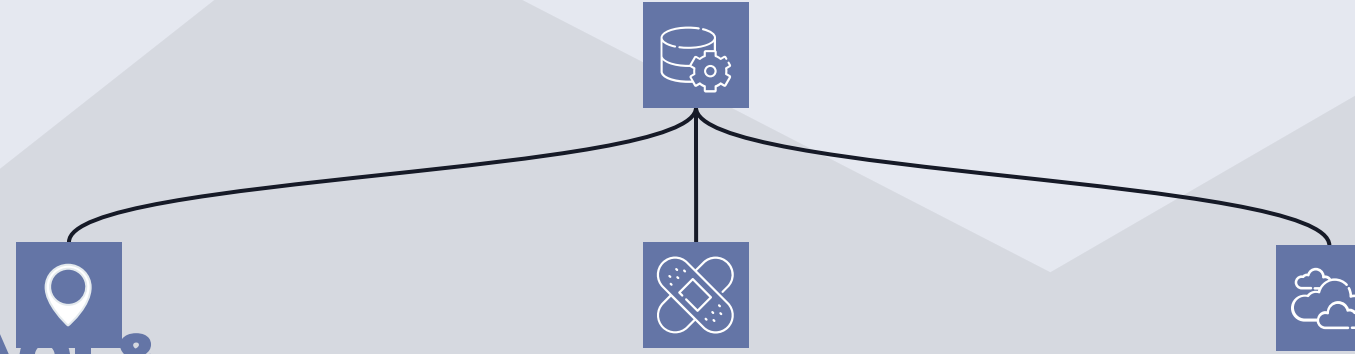
Air pollution continues to cause millions of preventable deaths each year because global air-quality, health, and economic data remain fragmented, limiting our ability to predict risks, target interventions, and drive coordinated action

02
Data processing





Integrated Datasets



AQI & Coordinates

Records- 16,695
Columns- 14
PK- Country

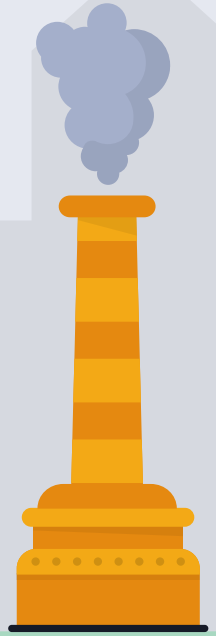
WHO Demographics

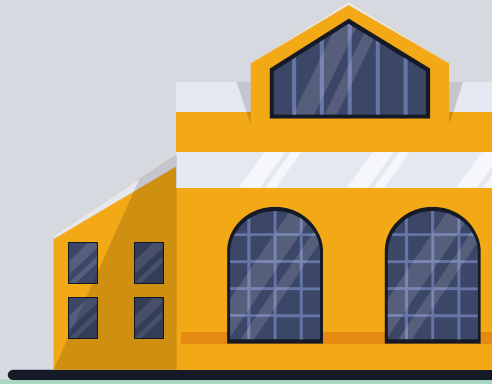
Records- 594
Columns- 54
PK- Country

Air Quality & Health

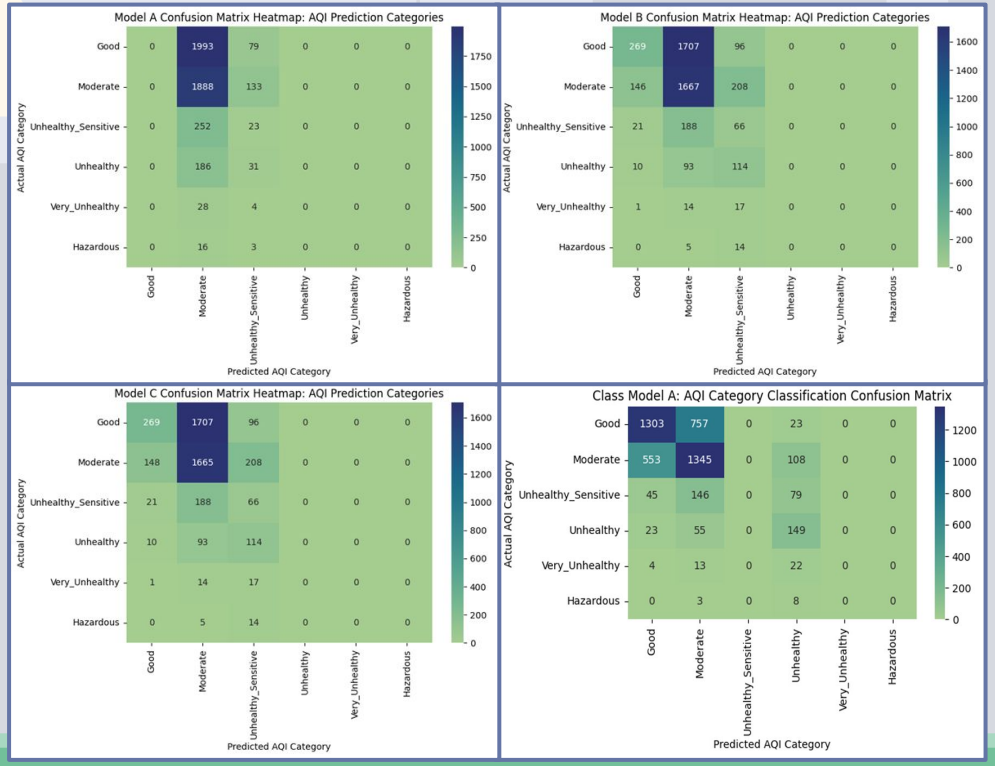
Records- 88,489
Columns- 12
PK- City

03
Our models





Model Development and Validation



Data

70/30

Model A

Linear Regression
3 Features
RMSE = 40.59

Model B

Linear Regression
4 Features
RMSE = 39.80

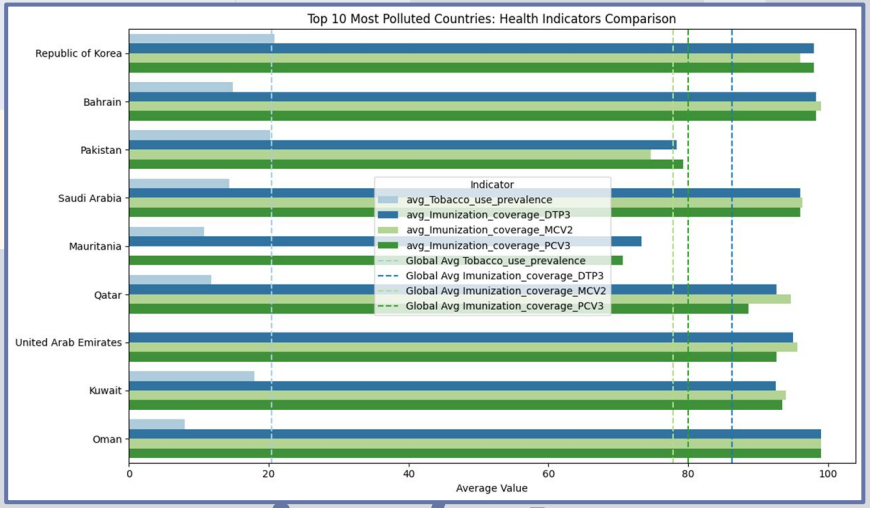
Model C

Linear Regression
5 Features
RMSE = 39.78

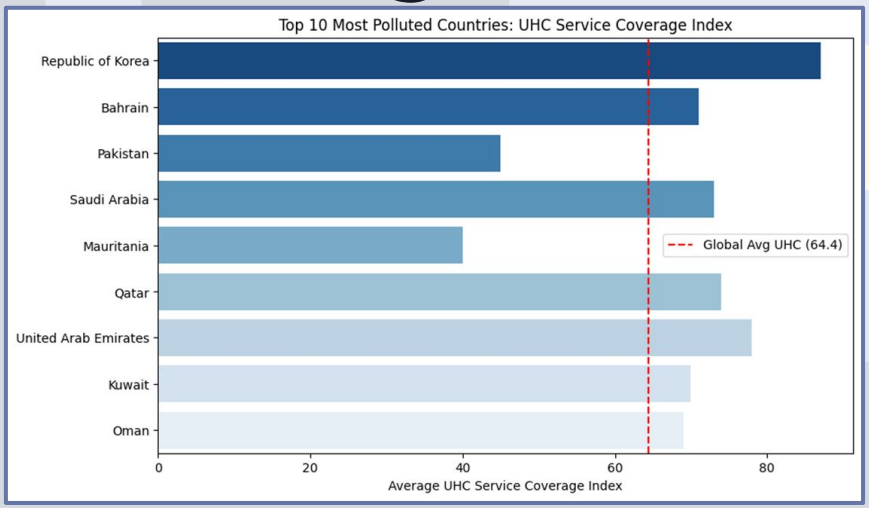
Classification Model A

Random Forest
Number of Trees = 10
Max Depth = 5
Accuracy 60.33%

Additional Data Investigation



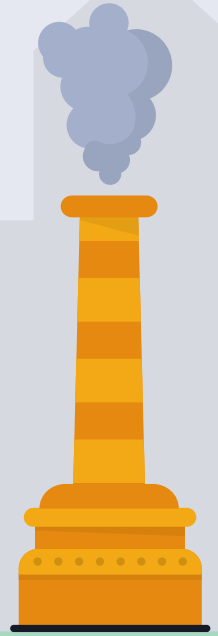
Vaccines/Tobacco Prevalence

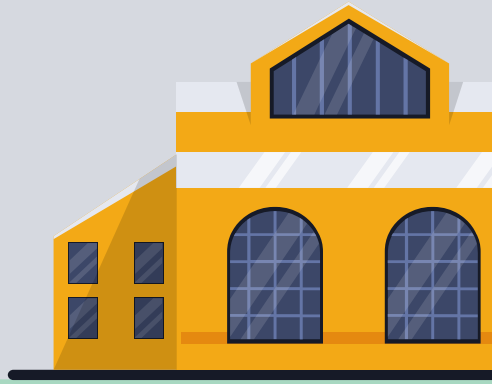


UHC Coverage Index

coverage of essential health services

04
Insights





Key Findings: AQI, Health, and Hotspots

Local factors had no impact on AQI

- Temp, humidity, admissions, capacity: $R^2 < 0.002$ each
- Combined local model: $R^2 \approx 0$

Country-level AQI–health link

- Higher AQI ment higher under-5 mortality lower life and healthy life expectancy
- One unit of increase in AQI adds 0.425 deaths / 100k, and decreases 0.028 healthy life years

Model C: limited but directional

- 5 health metrics explains some 14% of AQI variation
- Pollution and under-5 mortality: positive coefficients
- Healthy life expectancy: negative coefficient

Pollution hotspots

- Very high AQI in Republic of Korea, Bahrain, Pakistan
- Additional hotspots in Middle East, Mauritania, Guinea-Bissau

Strategy & Business Relevance for NGOs

Use AQI + mortality as screening

- Jointly flag high-AQI / high-mortality countries
- Move them into Tier 1 priority list
- Republic of Korea, Bahrain, Pakistan, Mauritania, Guinea-Bissau

Infrastructure & operations strategy

- Random Forest (10 trees, depth 5): 60.33% accuracy
- Good at Good/Moderate, weak for Unhealthy/Hazardous
- More reliable for cleaner at-risk zones than extreme cases

Practical NGO actions

- Place fixed sites in cleaner but at-risk areas
- Serve hazardous areas with mobile, short-term, high-impact programs
- Always layer model outputs with feasibility and equity considerations

Limits, Lessons & Future Work

Data & measurement limits

- AQI definitions differ across countries
- AQI = max pollutant index, non-linear category thresholds
- City-level IoT vs country-level WHO averages
- Uneven sensor coverage and possible sensor drift

Model limits

- Regressions: R^2 ~0.10–0.14
- Random Forest: 60.33% accuracy, strong class imbalance
- Worst performance in most polluted ranges

Technical & teamwork lessons

- Databricks/PySpark: cleaning and schema choices are critical
- Standardizing names, dates, types, keys before modeling
- Low R^2 can still support ranking and screening
- Divide-and-conquer team structure around strengths

Future work

- Real-time AQI and health APIs for dynamic monitoring
- Add socioeconomic and structural variables
- Improve modeling of extreme AQI and class imbalance

Questions