

# PREDICTING URBAN AIR POLLUTANT LEVELS USING TRAFFIC & WEATHER DATA: A MACHINE LEARNING APPROACH

Samarth Tomar, West Lafayette Jr Sr High, IN

## Rationale & Problem Statement



### The Problem

Vehicle exhaust and industrial emissions create hazardous PM2.5, NO<sub>x</sub>, PM10, and O<sub>3</sub> levels, impacting public health.



### Current Limitation

Traditional monitoring stations are expensive and sparse; AQI updates can be delayed.



### The Gap

Can we accurately predict local pollution levels using only widely available proxy data (traffic volume + weather) without expensive sensors?

## Research Question & Hypothesis



### Research Question

Is a real time air pollution level predictor based on traffic and weather data accurate when using different machine learning models?



### Hypothesis

If traffic density and weather conditions (wind, precipitation) significantly drive pollutant accumulation, then predictions based on such data are model independent.



### Engineering Goal

Develop an effective, accessible predictive tool.

## Background: What do these factors do?



### Rain

Washes out pollutants. Lowers the concentration.



### Traffic

Releases pollutants. Increases the concentration.



### Wind

Disperses the pollutants. Lowers the concentration.



### Temperature

Traps pollutants. Increases the concentration.

## Methodology: Data Acquisition

### Pollutants

OpenAQ API (Target: PM2.5 / NO<sub>2</sub> / O<sub>3</sub> / PM10).

### Weather

Meteosat API (Features: Temperature, Humidity, Wind speed, Precipitation).

### Traffic

Kaggle Traffic Volume Dataset (Features: Vehicle Count, Congestion Level).

### Dataset Size

- 5,000 hourly data points for pollution.
- 365 daily data points for weather and traffic.

### Locations

- Reykjavik, Iceland
- Manchester, England
- Los Angeles, United States

### Time Resolution

- Hourly data points for precise analysis.

## Methodology: Machine Learning Models

### Gradient Boosting (GB)

Builds models sequentially; each corrects the errors of the previous one. Best for minimizing bias.

Continuous Math

### Random Forest (RF)

Averages many decision trees. Best for reducing variance and overfitting.

Tree Based

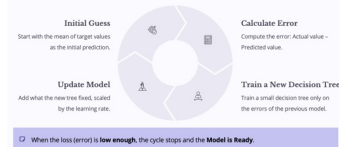
### K-Nearest Neighbors (KNN)

Simple instance-based learning. Baseline for comparison. Pattern identification

Averaging classification

## Experimental Design

### Gradient Boosting



### Random Forest Regression



### K-Nearest Neighbor



## Data Splits\*

80% Training / 20% Testing

65% Training / 35% Testing

50% Training / 50% Testing

\*allows test of overfitting vs underfitting

## Evaluation Metrics

### R<sup>2</sup>

How much variance the model explains.

### MAE

Average absolute error (robust to outliers).

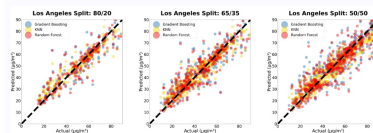
### RMSE

Average error in pollutant units.

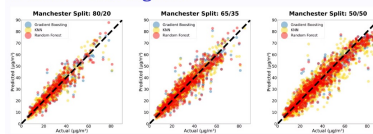
### RAE

Compares model to a baseline model to make sure the model is not getting lucky.

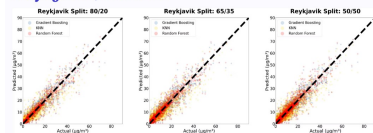
### LA, United States



### Manchester, England



### Reykjavik, Iceland



Used NO<sub>2</sub>. Made by vehicle exhaust.

### R<sup>2</sup>

How much your input features determine your output.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

### Relative Absolute Error (RAE)

Compares model to a baseline model to make sure the model is not just getting lucky.

$$RAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{\sum_{i=1}^n |y_i - \bar{y}|}$$

### Root Mean Squared Error (RMSE)

Similar to MAE but the penalty for large errors make the values bigger.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

### Mean Absolute Error (MAE)

Average distance between the prediction of the model and the actual value.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- n: The number of data points
- y<sub>i</sub>: The Actual Value
- ŷ<sub>i</sub>: The Model's Guess
- ȳ: The Average

## Analysis: Cross-City Generalizability

### Evaluation Metrics

R<sup>2</sup> results

City	Split	Random Forest	MAE	Random Forest
Los Angeles	80/20	0.82	0.18	0.18
Los Angeles	65/35	0.82	0.18	0.18
Los Angeles	50/50	0.82	0.18	0.18
Manchester	80/20	0.80	0.20	0.20
Manchester	65/35	0.80	0.20	0.20
Manchester	50/50	0.80	0.20	0.20
Reykjavik	80/20	0.80	0.20	0.20
Reykjavik	65/35	0.80	0.20	0.20
Reykjavik	50/50	0.80	0.20	0.20

- Shows how well the model captured patterns in the pollution data
- Low or negative R<sup>2</sup> means the data is hard to predict or very noisy
- Higher R<sup>2</sup> = model fits the data better

### Evaluation Metrics

RAE results: Relative Absolute Error (RAE)

City	Split	Random Forest	MAE	Random Forest
Los Angeles	80/20	0.82	0.18	0.18
Los Angeles	65/35	0.82	0.18	0.18
Los Angeles	50/50	0.82	0.18	0.18
Manchester	80/20	0.80	0.20	0.20
Manchester	65/35	0.80	0.20	0.20
Manchester	50/50	0.80	0.20	0.20
Reykjavik	80/20	0.80	0.20	0.20
Reykjavik	65/35	0.80	0.20	0.20
Reykjavik	50/50	0.80	0.20	0.20

- Compares the models to a baseline
- RAE < 1 = model is better than baseline
- RAE = 1 = model performs same as baseline
- RAE > 1 = model is actually awful

### Evaluation Metrics

MAE results: Mean Absolute Error (MAE)

City	Split	Random Forest	MAE	Random Forest
Los Angeles	80/20	0.82	0.18	0.18
Los Angeles	65/35	0.82	0.18	0.18
Los Angeles	50/50	0.82	0.18	0.18
Manchester	80/20	0.80	0.20	0.20
Manchester	65/35	0.80	0.20	0.20
Manchester	50/50	0.80	0.20	0.20
Reykjavik	80/20	0.80	0.20	0.20
Reykjavik	65/35	0.80	0.20	0.20
Reykjavik	50/50	0.80	0.20	0.20

- Shows the average size of the model's mistakes
- Lower MAE = more accurate predictions
- Helps compare which model makes the smallest everyday errors

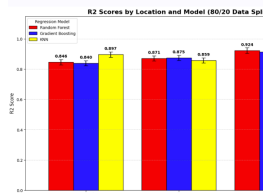
### Evaluation Metrics

RMSE results: Root Mean Squared Error (RMSE)

City	Split	Random Forest	RMSE	Random Forest
Los Angeles	80/20	0.82	0.18	0.18
Los Angeles	65/35	0.82	0.18	0.18
Los Angeles	50/50	0.82	0.18	0.18
Manchester	80/20	0.80	0.20	0.20
Manchester	65/35	0.80	0.20	0.20
Manchester	50/50	0.80	0.20	0.20
Reykjavik	80/20	0.80	0.20	0.20
Reykjavik	65/35	0.80	0.20	0.20
Reykjavik	50/50	0.80	0.20	0.20

- Shows how big the larger mistakes are
- If RMSE is much higher than MAE, the model sometimes makes big errors
- Lower RMSE = more stable and reliable predictions

A grouped bar chart comparing R<sup>2</sup> scores for KNN, Random Forest, and Gradient Boosting across LA, Manchester, and Reykjavik.



- All models statistically in same range
- Reykjavik scores highest due to lower baseline pollution variance
- The model generalizes across geographically and climatically diverse cities without city-specific retraining
- Supports the hypothesis of model independence

## Extension: Threshold Alert System

### Application

Convert the regression output to a binary "Unhealthy Air Alert" (Yes/No) based on WHO thresholds. NO<sub>2</sub> - higher than 25 µg/m<sup>3</sup> is dangerous.

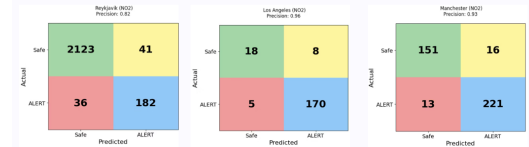
### Metric

Calculated Precision and Recall for these alerts.

### Visual

A Confusion Matrix showing how many "Unhealthy Days" were correctly flagged.

### \*Random Forest 80/20 split



## Conclusion & Future Work

### Hypothesis Supported:

Yes, traffic and weather data do drive the pollutants, and all the models were statistically equal and results are not dependent on model.

### Key Takeaway

We can build a pollution tracker using machine learning and existing data, while implying weather and traffic data.

### Limitations

Does not account for sudden pollution changes (e.g. sudden temperature change, unexpected heavy rain/wind).

### Future Improvement

Test model in different cities to check generalizability. Add temporal predictions.

## Key Sources & Acknowledgements

- "Ambient (Outdoor) Air Quality and Health." *World Health Organization*, 19 Dec. 2023, [www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health).
- "Common Air Pollutants." *United States Environmental Protection Agency*, 2024, [www.epa.gov/environmental-topics/air-common-air-pollutants](https://www.epa.gov/environmental-topics/air-common-air-pollutants).
- "National Oceanic and Atmospheric Administration." *U.S. Department of Commerce*, 2026, [www.noaa.gov/](https://www.noaa.gov/).
- "OpenAQ Explorer." *OpenAQ*, 2026, [explorer.openaq.org/](https://explorer.openaq.org/).
- Rohith. "Traffic Volume Dataset." *Kaggle*, 2023, [www.kaggle.com/datasets/rohith203/traffic-volume-dataset](https://www.kaggle.com/datasets/rohith203/traffic-volume-dataset).
- "WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide." *World Health Organization*, 2021, [iris.who.int/server/api/core/bitstreams/551b515e-2a32-4e1a-a58c-cdaecd395b19/content](https://iris.who.int/server/api/core/bitstreams/551b515e-2a32-4e1a-a58c-cdaecd395b19/content).