# 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, $NO_2$, PM10, and $O_3$ 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?

1

**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

2

**Pollutants**
OpenAQ API (Target: PM2.5 / $NO_2$ / $O_3$ / 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

3

**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***

### Gradient Boosting


When the loss (error) is **low enough**, the cycle stops and the **Model is Ready**.

### Random Forest Regression


The final averaged prediction reduces variance and overfitting, making the **Model Ready**.

### K-Nearest Neighbor


The mean of the K nearest neighbors becomes the final prediction, and the **Model is Ready**.

### Experimental Design

#### Data Splits*

80% Training / 20% Testing

65% Training / 35% Testing

50% Training / 50% Testing

*allows test of overfitting vs underfitting

#### Evaluation Metrics

**$R^2$**
How much variance the model explains.

**RMSE**
Average error in pollutant units.

**MAE**
Average absolute error (robust to outliers).

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

### 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**
- $y_i$: The **Actual Value**
- $\hat{y}_i$: The **Model's Guess**
- $\bar{y}$: The **Average**

### 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.

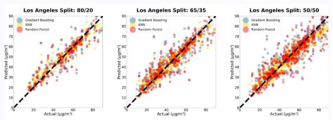
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

### $R^2$

How much your input features determine your output.

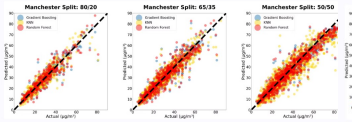
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

### LA, United States 🇺🇸



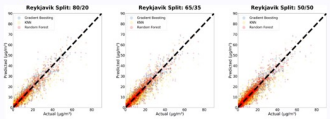

Used $NO_2$ - Made by vehicle exhaust.

### Manchester, England 🇬🇧



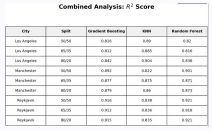

Used $NO_2$ - Made by vehicle exhaust.

### Reykjavik, Iceland 🇮🇸




Used $NO_2$ - Made by vehicle exhaust.

### Evaluation Metrics
$R^2$ results



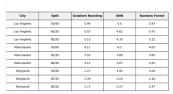

- Shows how well the model captured patterns in the pollution data
- Low or negative $R^2$ means the data is hard to predict or very noisy
- Higher $R^2$ = model fits the data better

### Evaluation Metrics
RAE results: Relative Absolute Error (RAE)



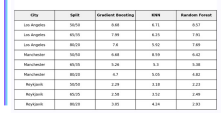

- Compares the models to a baseline
- RAE < 1 = model is better than baseline
- RAE > 1 = model performs worse than baseline
- Helps decide if the model is actually useful

### Evaluation Metrics
MAE results: Mean Absolute Error (MAE)




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

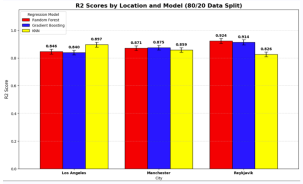
### Evaluation Metrics
RMSE results: Root Mean Squared Error (RMSE)




- 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

### Analysis: Cross-City Generalizability

A grouped bar chart comparing $R^2$ 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. NO2 - higher than 25 ug/m3 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.
- "Common Air Pollutants." *United States Environmental Protection Agency*, 2024, www.epa.gov/environmental-topics/air-common-air-pollutants.
- "National Oceanic and Atmospheric Administration." *U.S. Department of Commerce*, 2026, www.noaa.gov/.
- "OpenAQ Explorer." *OpenAQ*, 2026, explorer.openaq.org/.
- Rohith. "Traffic Volume Dataset." *Kaggle*, 2023, 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.