Predicting Urban Air Pollutant Levels using Traffic & Weather Data: A Machine Learning Approach

Samarth Tomar, West Lafayette Jr Sr High School, 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.

Methodology: Data Acquisition

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

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

Data Splits*

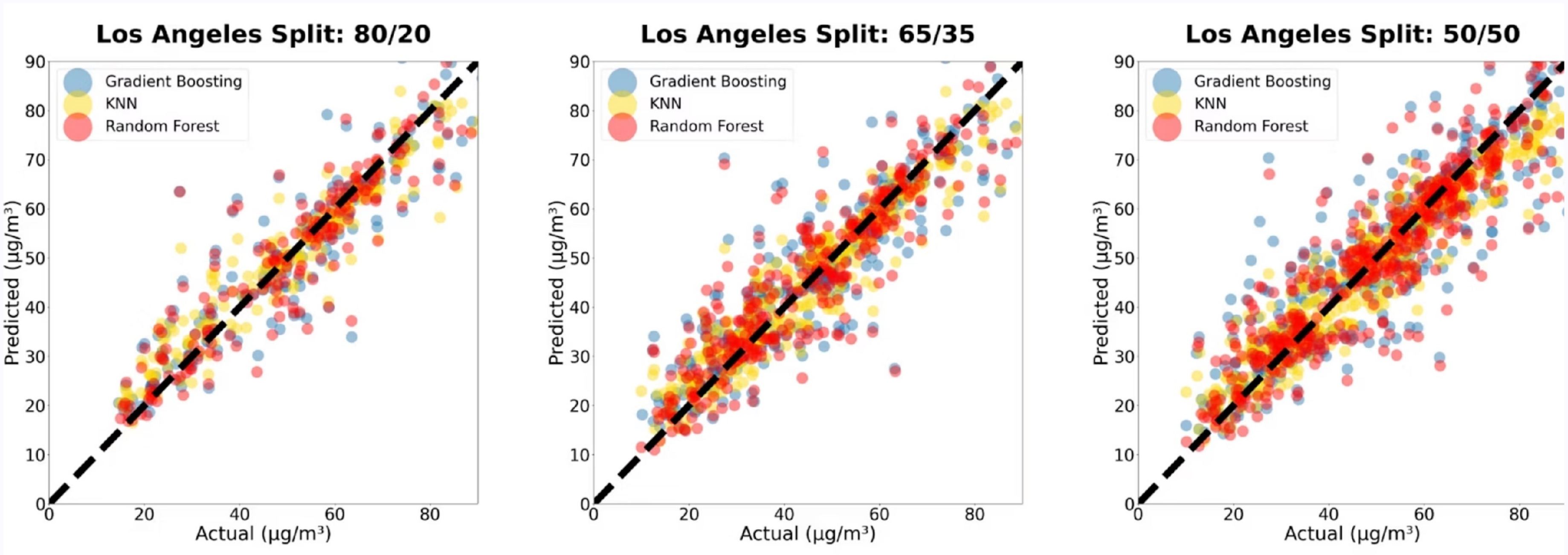
80% Training / 20% Testing

65% Training / 35% Testing

50% Training / 50% Testing

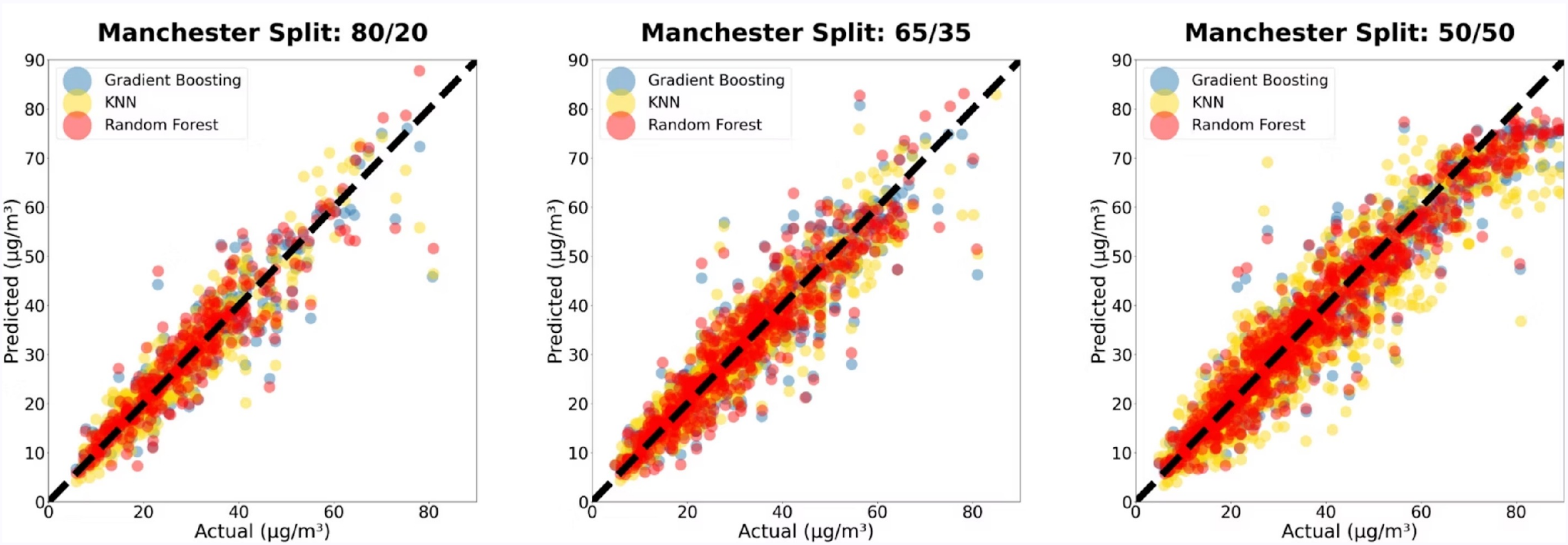
*allows test of overfitting vs underfitting

LA, United States 🇺🇸



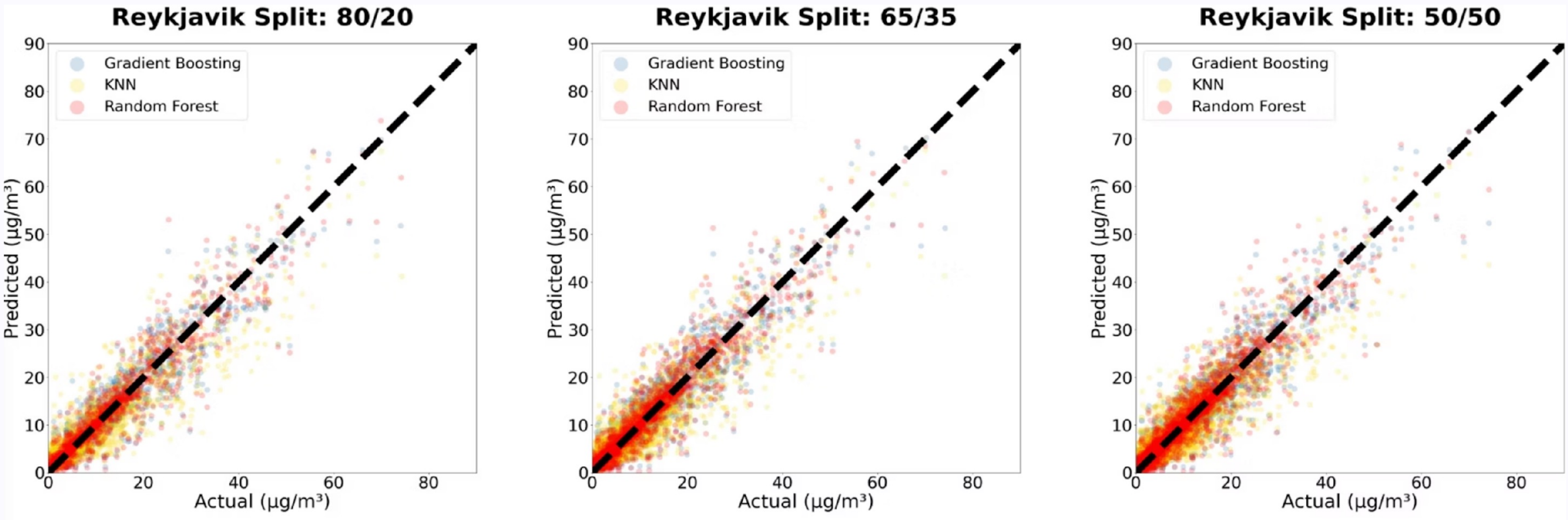
Used NO_2 - Made by vehicle exhaust.

Manchester, England 🇬🇧



> Used NO_2 - Made by vehicle exhaust.

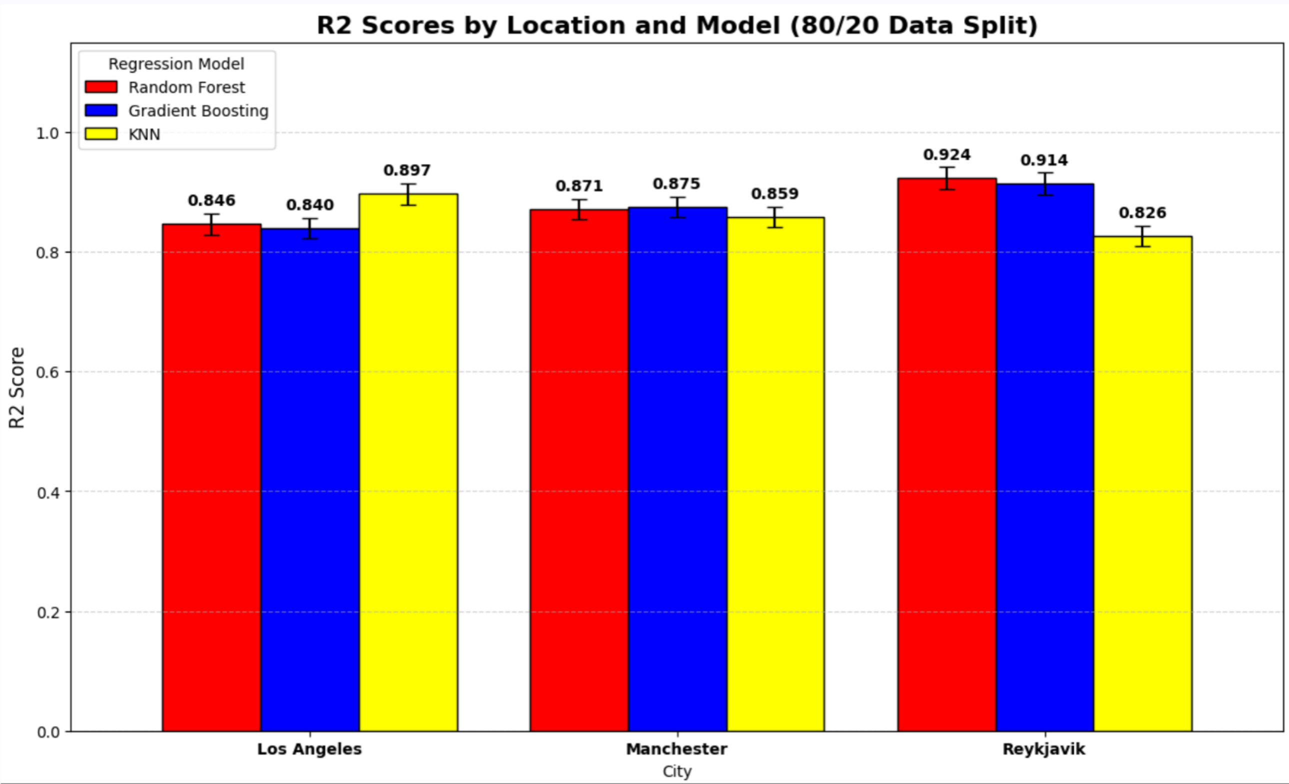
Reykjavik, Iceland 🇮🇸



Used NO_2 - Made by vehicle exhaust.

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

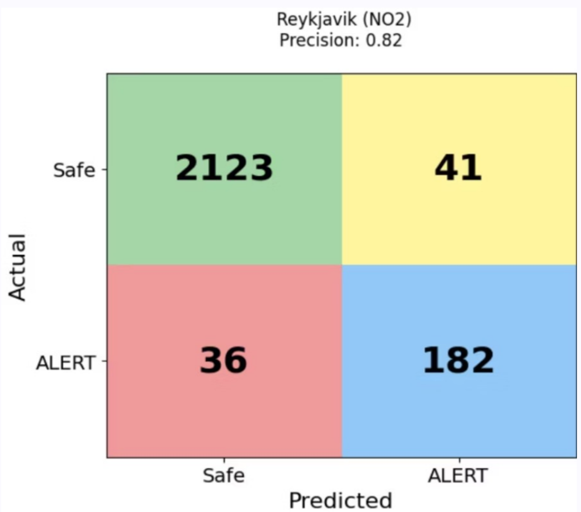
Metric

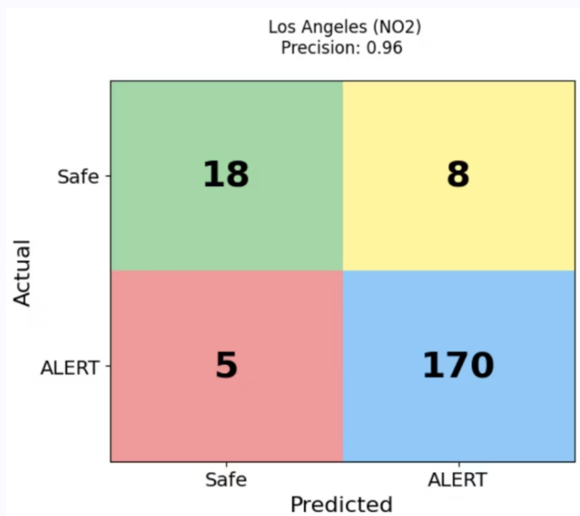
Calculated Precision and Recall for these alerts.

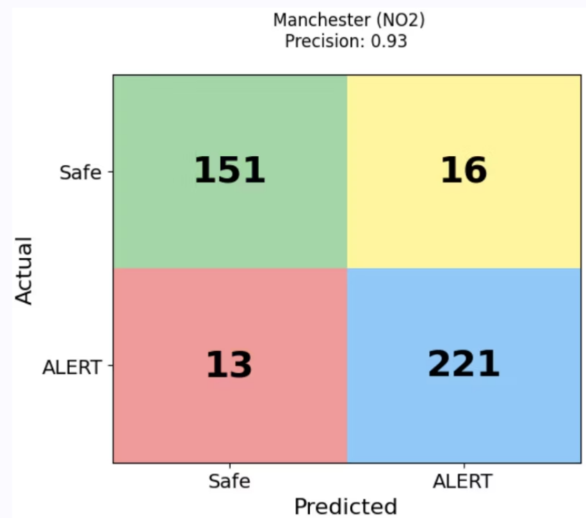
Visual

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.