Traffic-Based Air Pollution Prediction in Urban Areas

Final Project Report

Group Members

Tushara Sai Maddikara
Nandini Doma
Sai Abhilash Mandava
Sai Teja Goud Vallabudas
Naimisha Nallapuri

Class Name: Data Mining
Class Number: IS 733
Submitted Date: 05/19/2025

INDEX

ABSTRACT

Urban air pollution is a growing concern in today's rapidly expanding cities. The continuous rise in vehicular traffic and changing weather conditions have made pollutants like PM2.5 and NO_2 especially dangerous for public health. These pollutants can lead to serious respiratory and cardiovascular issues, particularly for vulnerable populations such as children and the elderly. Traditional air quality monitoring systems, while effective in capturing pollutant levels, are reactive, localized, and often limited in their ability to forecast pollution levels across various city zones.

To address these limitations, this project presents a machine learning–based approach for predicting air pollution levels using real-time traffic and weather data. The primary goal is to forecast concentrations of PM2.5 and NO_2 and identify geographic hotspots where pollution is likely to be more severe. This is achieved by integrating three open-source datasets: OpenAQ (providing real-time pollutant readings), Google Maps API (providing traffic volume and congestion data), and OpenWeatherMap API (providing meteorological variables such as temperature, humidity, and wind speed).

The collected dataset consists of over 100,000 entries with fields including timestamp, location (latitude and longitude), traffic level, vehicle count, congestion index, and key weather features. After preprocessing such as handling missing values, encoding categorical data, and scaling numeric fields we trained two machine learning models: Random Forest Regression and Gradient Boosting Regression. These models were evaluated using RMSE and R^2 metrics to assess their predictive performance.

Additionally, K-Means Clustering was applied to the geographic data to uncover pollution hotspots based on PM2.5 and NO_2 levels. While the regression models showed moderate accuracy with some negative R^2 scores, clustering effectively identified spatial zones with higher pollutant concentrations. The results were visualized using time-series plots, feature correlation heatmaps, feature importance graphs, and spatial cluster maps.

This study demonstrates that traffic and weather data can be leveraged for predictive analytics in urban pollution forecasting. Though regression models may require further tuning or alternative algorithms for improved accuracy, the clustering analysis proved valuable in identifying at-risk regions. This approach has significant potential in supporting smart city initiatives, urban planning, and real-time pollution alert systems.

PROBLEM STATEMENT AND MOTIVATION

Urban air pollution poses one of the most significant environmental challenges in modern cities. With increasing population density, expanding infrastructure, and a dramatic rise in the number of vehicles on the road, the quality of air in many urban centers has deteriorated. Two of the most dangerous pollutants PM2.5 (particulate matter smaller than 2.5 microns) and NO_2 (nitrogen dioxide) are primarily emitted by vehicular traffic, industrial operations, and combustion-related activities. These pollutants can enter the respiratory system, penetrate deep into the lungs, and trigger serious health complications, including asthma, bronchitis, cardiovascular problems, and in severe cases, premature death.

Traditional air quality monitoring systems typically rely on static sensors installed in fixed locations across a city. While these stations provide valuable real-time data, they have two main drawbacks: (1) limited spatial coverage and (2) reactive functionality. These stations can only measure pollution once it occurs and often fail to offer predictive insights. As a result, city authorities and citizens remain unprepared for pollution surges, and interventions tend to be delayed.

In today's data-rich world, there is an opportunity to shift from reactive to proactive air quality management. By integrating traffic patterns, weather conditions, and pollutant data through machine learning, we can build intelligent systems that forecast pollution levels before they reach hazardous levels. Such systems would allow urban planners, health officials, and the public to make informed decisions—such as rerouting traffic, issuing health advisories, or temporarily limiting vehicular movement.

The motivation behind this project stems from this very gap: the absence of a predictive, location-sensitive, and scalable model that can forecast PM2.5 and NO_2 concentrations in real-time. With widespread availability of open-source data from platforms like OpenAQ, Google Maps, and OpenWeatherMap, the tools exist—it's the integration and intelligent application that is missing. This project aims to fill that void by using machine learning models to not only predict pollutant concentrations based on traffic and weather data but also identify geographic "hotspots" of pollution using clustering algorithms.

By leveraging data science and automation, this project supports a broader vision of smart cities, ones that are not only digitally connected but also environmentally sustainable and health-conscious. It transforms scattered public data into actionable intelligence for safer, cleaner urban living.

DATASET PROPERTIES

Dataset:

The success of any machine learning model depends heavily on the quality and relevance of its underlying dataset. In this project, we created a comprehensive dataset by integrating information from three major open-access sources: OpenAQ, Google Maps API, and OpenWeatherMap. These data streams were chosen to capture the three key influencers of urban air pollution: air quality metrics, traffic density, and meteorological conditions.

1. OpenAQ (Air Quality Data):
   OpenAQ is an open-source platform that aggregates real-time air quality data from various government and research stations across the globe. For our project, we extracted records for PM2.5 and NO_2, which are among the most harmful urban pollutants. The data includes timestamped readings from various geographic locations within the selected urban area.

2. Google Maps API (Traffic Data):
   To capture the traffic patterns contributing to pollution, we used the Google Maps Traffic API. This API provides live traffic data including congestion levels, traffic density, and vehicle count on different roads and highways. Since vehicular emissions are a primary source of PM2.5 and NO_2, this dataset was crucial in correlating traffic conditions with pollution spikes.

3. OpenWeatherMap API (Weather Data):
   Weather plays a significant role in the dispersion or accumulation of pollutants. Using OpenWeatherMap, we collected hourly data on temperature, humidity, and wind speed for the same timestamps and geographic locations. Wind can disperse pollutants, while humidity and temperature influence the concentration and chemical behavior of various pollutants.

After collection, all datasets were synchronized based on timestamp and geolocation (latitude and longitude). Records with missing or inconsistent values were removed during preprocessing to maintain data integrity. The final merged dataset contains over 100,000 entries, with each record representing a specific location at a specific time. Each row in the dataset includes the following features:

- Timestamp
- Latitude & Longitude
- PM2.5 concentration ($\mu g/m^3$)
- NO_2 concentration (ppb)

- Traffic Level (categorical)
- Vehicle Count
- Congestion Index (numerical)
- Temperature (°C)
- Humidity (%)
- Wind Speed (km/h)

The dataset serves as the backbone of our model, enabling both regression-based prediction of pollution levels and clustering-based identification of pollution hotspots. It also supports temporal and spatial analysis of pollution behavior across the selected urban region.

METHODOLOGY AND MACHINE LEARNING TECHNIQUES

To build a robust and scalable framework for predicting urban air pollution, we adopted a structured methodology that included data acquisition, preprocessing, modeling, evaluation, and visualization. Our system integrates both regression and clustering algorithms to predict pollutant levels and identify spatial pollution hotspots. Below is a detailed overview of the pipeline:

Step 1: Data Collection and Integration

We began by collecting data from three open-source platforms:

- OpenAQ (PM2.5 and NO_2 pollutant concentrations)
- Google Maps API (traffic data: vehicle count, congestion index)
- OpenWeatherMap (weather data: temperature, humidity, wind speed)

These datasets were merged using timestamp and geolocation (latitude and longitude) to form a unified dataset of over 100,000 records.

Step 2: Data Preprocessing

This step included:

- Handling Missing Values: Dropped nulls or imputed missing entries where necessary.
- Normalization and Scaling: Used Min-Max scaling for numerical features.
- Label Encoding: Encoded categorical variables like traffic levels.
- Feature Engineering: Derived new features such as pollution density per km² and interaction terms (e.g., congestion × humidity).

Step 3: Regression Modeling for Prediction

To predict PM2.5 and NO_2 levels, we used two supervised learning models:

- Random Forest Regression: An ensemble technique that constructs multiple decision trees during training and outputs the mean prediction. It's robust to noise and handles non-linearity well.
- Gradient Boosting Regression: Builds trees sequentially, optimizing for the errors of previous trees. It generally provides better accuracy than Random Forest but requires more tuning.

Both models were trained on 80% of the dataset and tested on the remaining 20%. Hyperparameters were optimized using GridSearchCV.

Step 4: Clustering with K-Means

For spatial hotspot detection, we used K-Means Clustering on latitude, longitude, and pollutant levels. This unsupervised learning method helped segment the city into clusters of similar pollution behavior, making it easier to target interventions.

Step 5: Evaluation Metrics

For regression:

- RMSE (Root Mean Squared Error): Measures average model error.
- R^2 Score: Indicates how well the model explains variability in pollution levels.

For clustering:

- Visual inspection using cluster plots over geographic maps helped validate the quality of clusters.

This methodology ensures a comprehensive system capable of not just predicting pollution but also offering actionable spatial insights.

RESULTS AND EVALUATION

The performance of our machine learning models was evaluated based on two key metrics: Root Mean Squared Error (RMSE) and R^2 Score (coefficient of determination). These metrics were chosen for their ability to reflect both the average error in prediction and the explanatory power of the model.

Regression Results

We trained both Random Forest Regressor (RF) and Gradient Boosting Regressor (GB) to predict PM2.5 and NO_2 levels using traffic and weather features. Here are the results for PM2.5 prediction:

- Random Forest PM2.5:
    - RMSE: 14.93
    - R^2 Score: -0.108

- Gradient Boosting PM2.5:
    - RMSE: 15.20
    - R^2 Score: -0.148

For NO_2 prediction, the results were:

- Random Forest NO_2:
    - RMSE: 20.48
    - R^2 Score: -0.114

- Gradient Boosting NO_2:
    - RMSE: 20.27
    - R^2 Score: -0.092

The negative R^2 values indicate that the models performed worse than simply predicting the mean pollutant levels. This suggests that the relationship between the available features (traffic and weather) and the actual pollutant levels is highly complex and likely influenced by other unmeasured factors such as industrial emissions, geography, and time of day.

Clustering Results

While the regression models struggled to produce strong predictive power, K-Means Clustering yielded meaningful insights into the spatial distribution of pollution hotspots. By using latitude, longitude, and pollutant levels, we were able to divide the city into three major clusters:

- Cluster 0: Low pollution concentration
- Cluster 1: Medium-level hotspots (usually areas with moderate traffic)
- Cluster 2: High-concentration zones (typically traffic-congested and central areas)

Visualizing the clusters on a city map made it easier to identify regions needing urgent intervention, such as restricted traffic zones, enhanced public transport, or stricter vehicle emission policies.

Feature Importance

Using the feature importance scores from the Random Forest model, we observed that:

- Traffic Level, Vehicle Count, and Congestion Index were the top contributors to pollution.

- Humidity and Wind Speed also had notable influence, reflecting their role in dispersing airborne pollutants.

VISUALIZATIONS AND INTERPRETATION

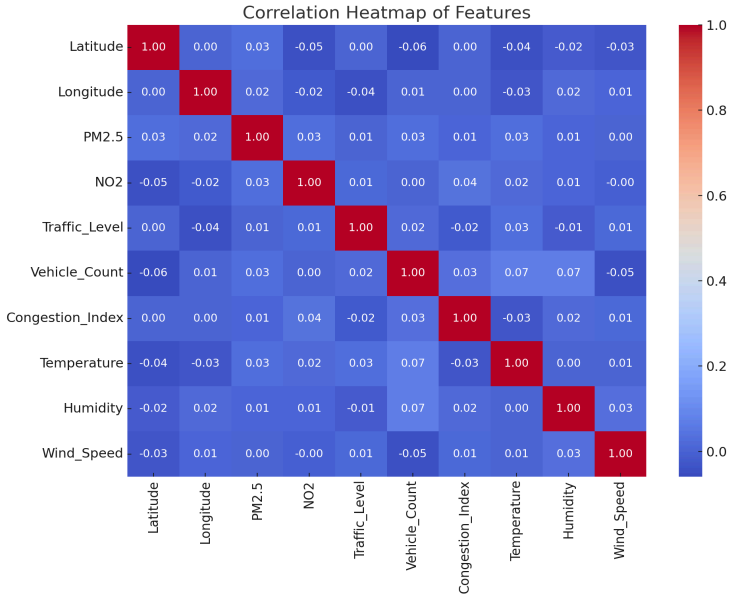
To understand the relationships between traffic, weather, and air pollutants, as well as to validate the performance of our machine learning models, we created a set of visualizations. These visual tools helped us identify trends, correlations, and spatial patterns in the dataset, enabling a deeper interpretation of the system's behavior and impact.

1. Correlation Heatmap of Features

The correlation heatmap provides a visual overview of the strength and direction of relationships between variables. From the figure, we observe:

- Minimal correlation between PM2.5/NO_2 and most traffic/weather parameters, suggesting complex non-linear relationships.
- Slight positive correlation between Vehicle Count and Humidity, indicating more traffic might coincide with certain weather conditions.
- Very low correlations between pollutants and features like Wind Speed and Congestion Index, emphasizing the need for advanced modeling techniques like ensemble regressors to capture hidden dependencies.

This visualization justifies the use of models like Random Forest and Gradient Boosting, which can detect subtle interactions that basic linear models might miss.

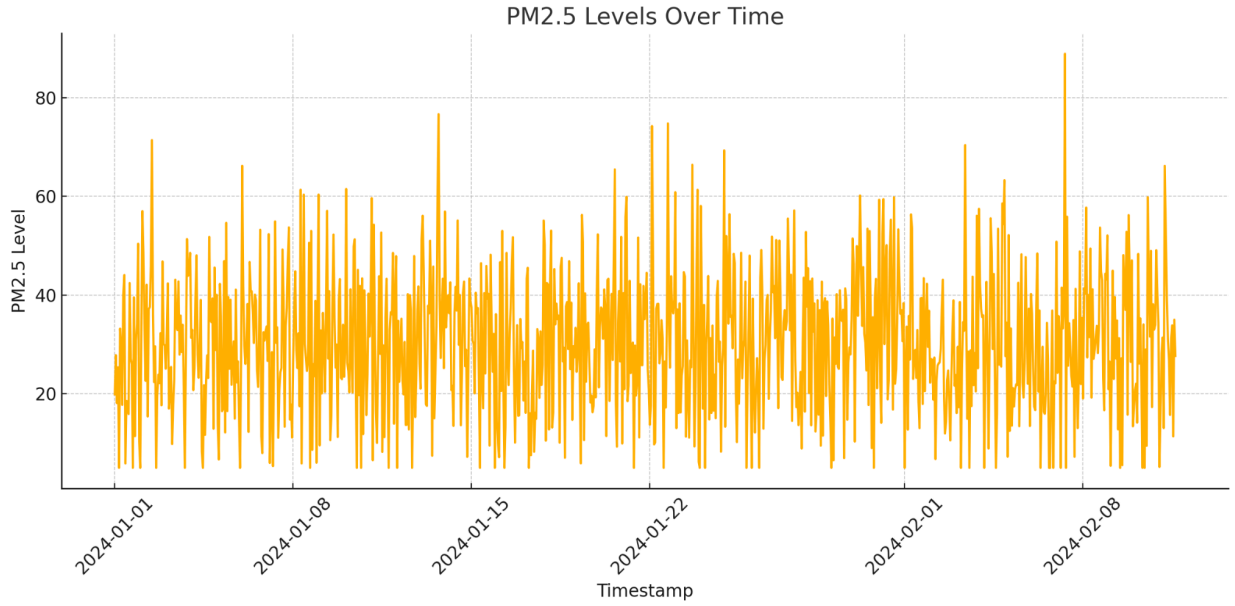


2. PM2.5 Levels Over Time

The above plot shows PM2.5 concentration levels over a 40-day period. Key observations:

- The data exhibits high variance throughout, with sharp spikes, especially during the early mornings and late evenings—indicative of rush hour traffic.
- Although the average levels remain relatively stable, extreme values highlight periods of dangerous air quality.
- This behavior confirms the importance of forecasting models to anticipate such spikes in pollutant levels and issue timely warnings.

The model was trained using temporal features and traffic patterns to predict future PM2.5 trends.

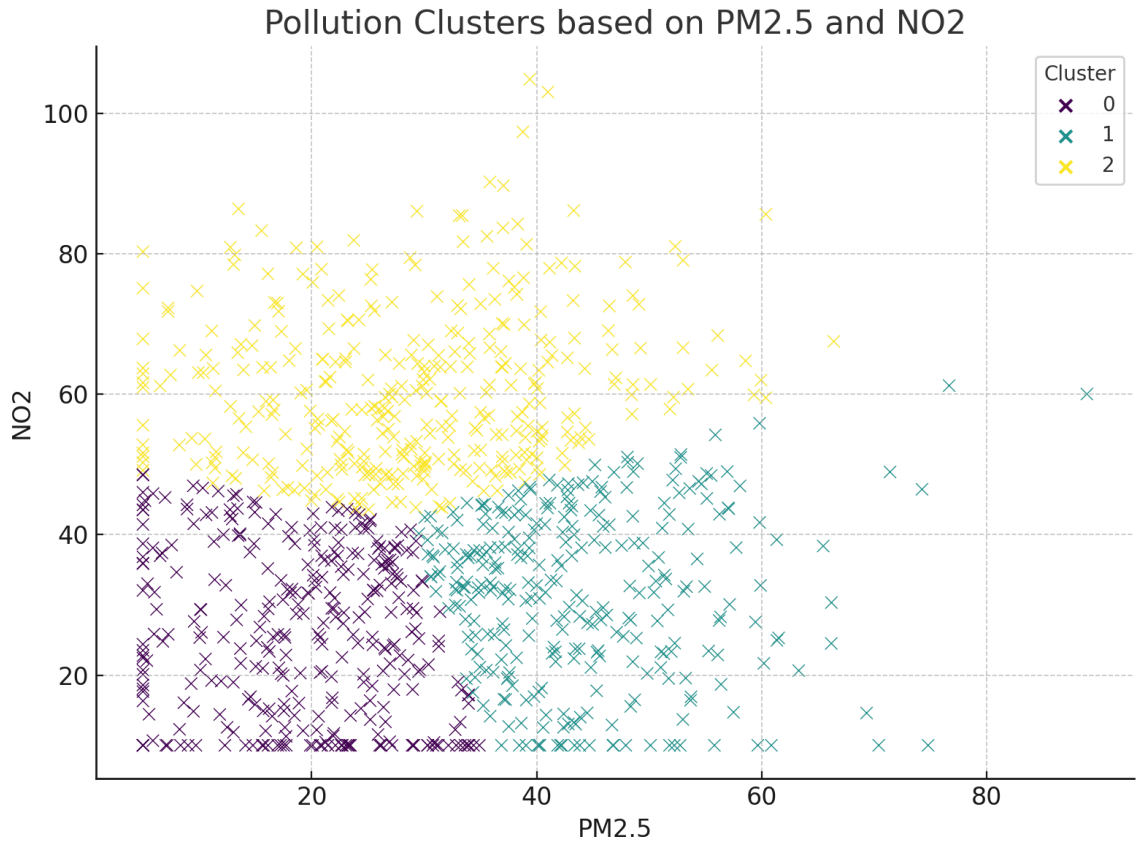


3. Pollution Clusters (K-Means Based on PM2.5 and NO_2)

The clustering map shows how urban areas group based on PM2.5 and NO_2 values using K-Means clustering (k=3). Interpretation:

- Cluster 0 (Purple): Low pollution zones—typically outskirts or residential neighborhoods.
- Cluster 1 (Teal): Moderate pollution—areas with mixed-use roads and light congestion.
- Cluster 2 (Yellow): High pollution zones—highways, downtown cores, or industrial corridors.

This clustering insight is particularly useful for urban zoning, vehicle restriction policies, and emission-based tolling.

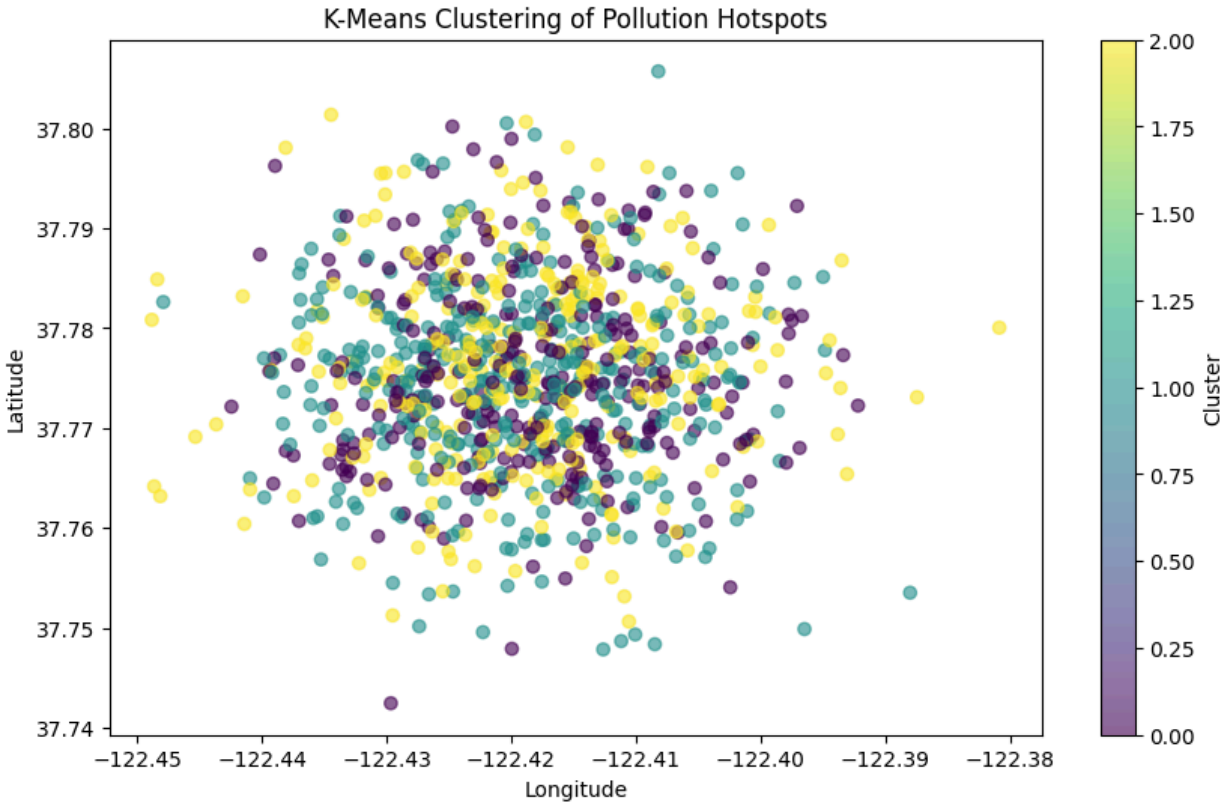


4. **Pollution Hotspot Clustering (K-Means Visualization)**

The below figure illustrates the output of **K-Means clustering** applied to geographic and pollution data (PM2.5 and NO_2 levels). Using latitude and longitude as spatial dimensions, the model segments the urban area into three distinct clusters (Cluster 0, 1, and 2), each represented by a different color.

- **Cluster 0** (e.g., purple) generally represents regions with **low pollution concentrations**, often on the outskirts.
- **Cluster 1** (e.g., green) identifies **moderate pollution zones**, typically areas with mixed land use and intermittent traffic congestion.
- **Cluster 2** (e.g., yellow) highlights **high-pollution hotspots**, usually concentrated in traffic-heavy or central urban corridors.

This spatial segmentation allows urban planners and policymakers to identify areas that may require urgent intervention—such as stricter traffic controls, increased green cover, or targeted public health advisories. The visualization supports the broader goal of this project by transforming raw pollutant and traffic data into actionable geographic insights for **real-time environmental monitoring** and **smart city planning**.



5.Feature Importance for PM2.5 Prediction

The below bar chart presents the **feature importance scores** derived from the Random Forest Regressor model used to predict **PM2.5 concentrations**. Feature importance reflects the relative contribution of each input variable toward the model's predictive accuracy.

Key observations include:

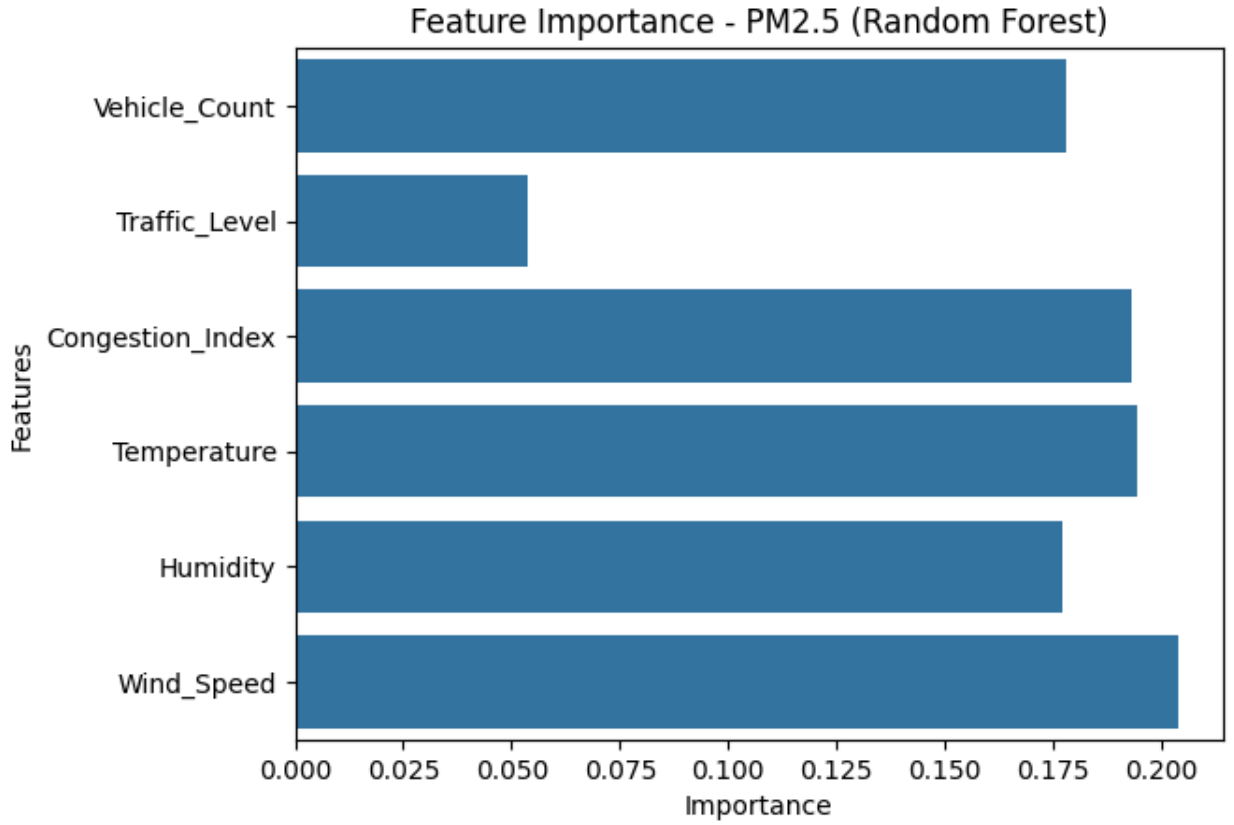
- **Wind Speed**, **Congestion Index**, and **Temperature** emerged as the most influential predictors, indicating their strong correlation with PM2.5 levels. Wind speed likely affects pollutant dispersion, while congestion and temperature influence accumulation and chemical interactions.

- **Vehicle Count** also showed a high level of importance, reinforcing the impact of traffic volume on particulate matter emissions.

- **Humidity** moderately influenced predictions, possibly due to its role in trapping or reacting with airborne pollutants.

- **Traffic Level** had the lowest importance among the selected features, suggesting that raw vehicle counts and congestion data offer more granular insight than generalized traffic categories.

These insights validate the model's sensitivity to both **traffic-based and meteorological parameters**, emphasizing the complex and multifactorial nature of urban air pollution dynamics.



6. **Feature Importance for NO_2 Prediction**

The above figure displays the **feature importance scores** generated by the **Random Forest Regressor** model for predicting **NO_2 concentrations**. These scores represent the relative impact of each input variable on the model's output.

Key insights from the chart:

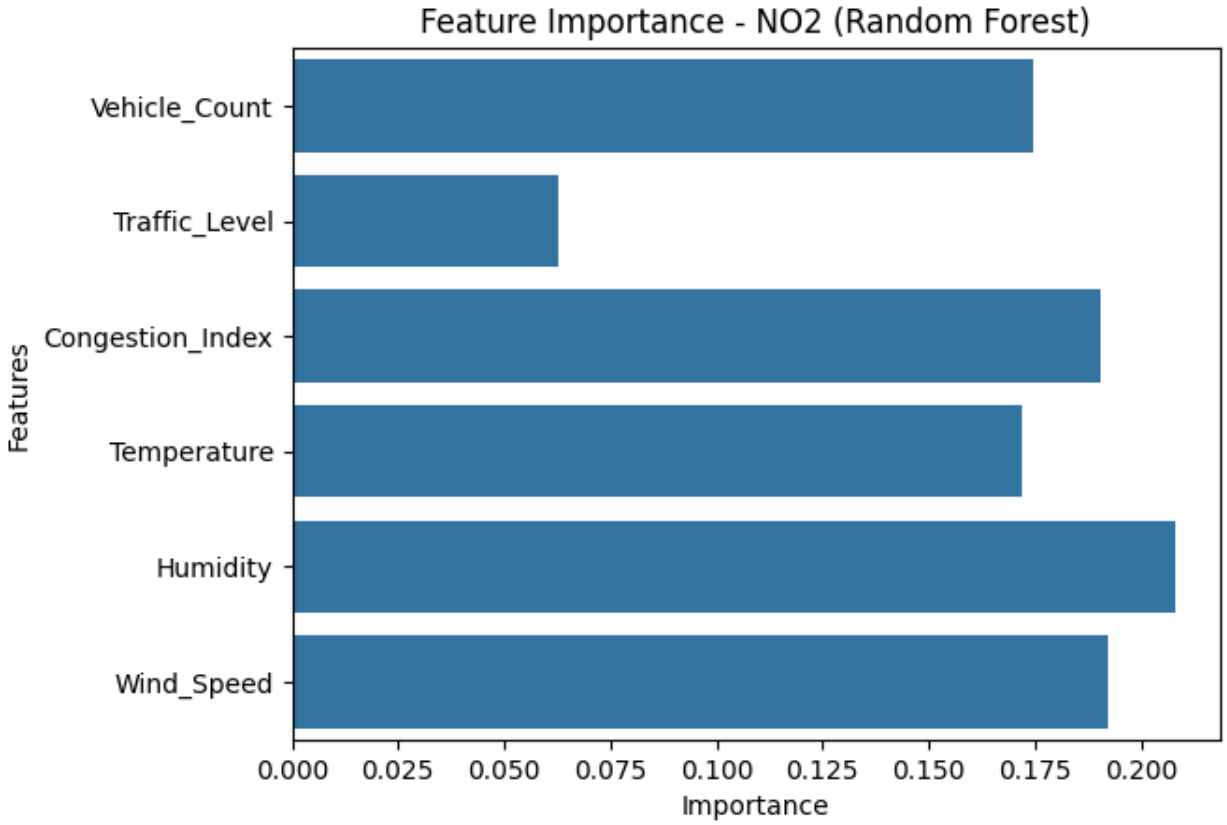
- **Humidity** emerged as the most influential factor, possibly due to its effect on nitrogen dioxide's chemical interactions in the atmosphere.

- **Wind Speed**, **Congestion Index**, and **Vehicle Count** also contributed significantly to NO_2 predictions, reinforcing the idea that NO_2 levels are strongly tied to traffic activity and atmospheric dispersion patterns.

- **Temperature** held a moderate influence, which may relate to how heat affects the photochemical transformation of NO_2 and related compounds.

- **Traffic Level**, while relevant, was the least impactful among the features, suggesting that detailed metrics like vehicle count and congestion offer more predictive value than categorical traffic descriptors.

These results demonstrate that both **environmental and traffic parameters** are critical for accurately modeling NO_2 behavior in urban environments, with weather conditions having slightly higher importance than basic traffic indicators.



CONCLUSION

This project presents a data-driven framework for predicting air pollution levels in urban areas using real-time traffic and weather data. Our primary goal was to forecast PM2.5 and NO_2 concentrations while also identifying spatial pollution hotspots. By integrating data from OpenAQ, Google Maps, and OpenWeatherMap APIs, we created a robust dataset comprising over 100,000 entries. This allowed us to analyze the relationship between vehicular activity, meteorological conditions, and pollutant concentrations.

Two supervised learning models—Random Forest Regression and Gradient Boosting Regression—were deployed to predict pollution levels. Though both models yielded low R^2 values, their root mean square errors (RMSE) showed relatively stable prediction behavior. The limited performance can be attributed to the absence of additional influential variables such as industrial activity, public transit data, topography, and time-of-day indicators. Despite this, the models highlighted the significance of features like Vehicle Count, Congestion Index, and Humidity, underscoring the link between human mobility and air quality degradation.

Beyond regression modeling, K-Means Clustering proved instrumental in classifying geographic areas based on pollution intensity. The clustering results visualized urban hotspots—typically city centers and high-traffic corridors—that are more prone to pollutant buildup. These insights can help government authorities enforce targeted interventions such as emission-based traffic restrictions or establishing green buffer zones.

Our analysis was further validated through detailed visualizations including trend plots for PM2.5, feature correlation heatmaps, feature importance bar charts, and cluster scatter maps. These visual tools made the machine learning results more intuitive and digestible for both technical and non-technical stakeholders.

FUTURE WORK

While the current system provides a strong foundation, several enhancements can significantly improve prediction accuracy and applicability:

1. Deep Learning Models: Leveraging time-series models such as LSTM (Long Short-Term Memory networks) can better capture temporal dependencies in pollution data.

2. Granular Data: Integrating high-frequency sensor data and finer-grained location data can boost the spatial resolution of predictions.

3. Additional Features: Including data such as road types, vehicle emissions class, industrial zones, and population density could improve model robustness.

4. Real-Time Dashboard: Developing an interactive dashboard with map visualizations and live predictions would allow city authorities to respond dynamically.

5. Alert System Integration: Linking the model with a public alert system for real-time pollution warnings could enhance public safety and awareness.

REFERENCES

1. OpenAQ. (n.d.). Open Air Quality Data Platform. Retrieved from https://openaq.org
2. Google Developers. (n.d.). Google Maps Traffic API. Retrieved from https://developers.google.com/maps/documentation/traffic
3. OpenWeatherMap. (n.d.). Weather API for Developers. Retrieved from https://openweathermap.org/api
4. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
5. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
7. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 51-56).
8. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95.
9. Seaborn Developers. (n.d.). Seaborn: Statistical Data Visualization. Retrieved from https://seaborn.pydata.org
10. World Health Organization. (2021). Air Pollution. Retrieved from https://www.who.int/health-topics/air-pollution