Traffic-Based Air Pollution Prediction in Urban Areas

UMBC

ABSTRACT

This project aims to use machine learning techniques to predict levels of air pollution in urban areas using traffic and weather conditions. Traffic congestion and weather changes influence critical pollutants like PM2.5 and NO_2. By using regression techniques and clustering techniques, we analyze real-time and past data to predict pollution and identify hotspots. Random Forest and K-Means Clustering tools help us understand environmental trends, enabling anticipatory city planning and public health interventions.

BACKGROUND

Urban air quality has been severely impacted by rising traffic volumes. Fine particulate matter (PM2.5) and nitrogen dioxide (NO_2), primarily from vehicle emissions, are health threats. Traditional monitoring methods are location-bound and responsive. Our approach integrates data from OpenAQ, Google Maps, and OpenWeatherMap to build a predictive, scalable, and interactive pollution prediction model. The study highlights the significance of data-driven solutions in mitigating the urban pollution issue.

EXPERIMENTAL METHODS

Data Collection and Integration

- **Pollution:** OpenAQ
- **Traffic:** Google Maps API
- **Weather:** OpenWeatherMap
- Data is synchronized by **timestamp** and **location.**

Modeling Techniques

- **Random Forest Regression:** Robust for noisy and multi-feature datasets.
- **Gradient Boosting Regression:** High accuracy in trend detection.
- **K-Means Clustering:** Identifies **pollution hotspots.**

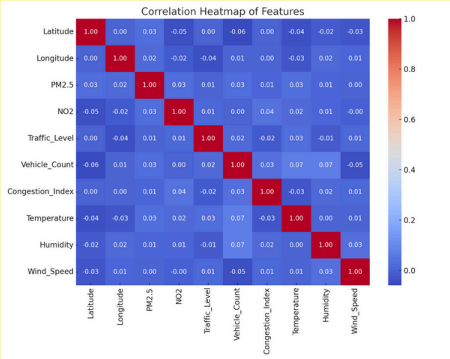
Preprocessing

- Handled missing/inconsistent values.
- Standardized **units** and **time formats.**

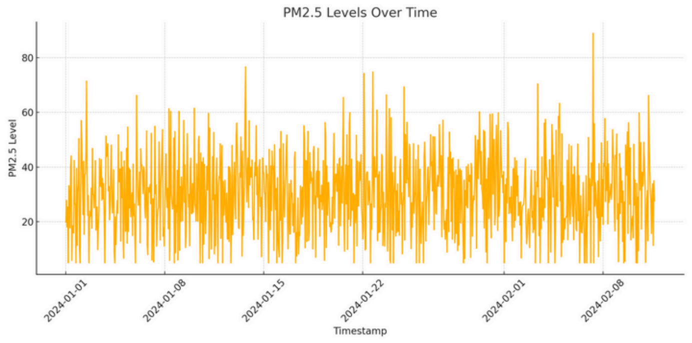
Evaluation

- **Metrics:** RMSE, MAE, R^2 Score
- **Visualizations:** Heatmaps, Line Graphs, Cluster Maps

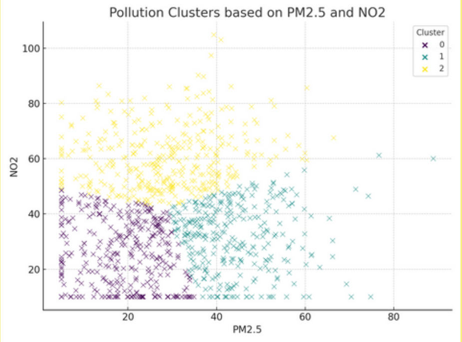
RESULTS & DISCUSSIONS



Correlation Heatmap – Displays how all numerical variables relate to one another.



PM2.5 Trend Over Time – Shows how PM2.5 levels vary across timestamps.



Pollution Clustering – Identifies hotspot clusters using PM2.5 and NO_2 with K-Means.

- Random Forest exhibited great predictive performance with good generalizability.
- Gradient Boosting achieved outstanding precision when identifying trend patterns.
- Through K-Means clustering the algorithm achieved success in grouping regions based on their pollution severity levels.
- Visualizations displayed pollution spikes which occurred during peak hours while showing weather-related patterns.
- A conclusional analysis enabled researchers to create evidence-based traffic regulation and zoning policies.

FUTURE WORK

- Integrate real-time API data feeds for live dashboard updates.
- Explore LSTM and other deep learning models for time-series forecasting.
- Conduct seasonal trend analysis and anomaly detection.
- Expand pollutant scope and apply in more urban areas.
- Partner with civic authorities for real-world application.

CONCLUSION

A dependable prediction system emerged because of the integration process for traffic data and pollution data and weather data. Multiple machine learning algorithms including Random Forest and Gradient Boosting have already shown their effectiveness in modeling urban air quality levels. Stakeholders obtain practical information through the system's hot-spot detection capabilities. As a final outcome this project shows how data-oriented methods can scale up to manage environmental risks.

Team Members:

1. Tushara Sai Maddikara
2. Nandini Doma
3. Sai Abhilash Mandava
4. Saiteja Vallabudas
5. Naimisha Nallapuri

UMBC
DEPARTMENT OF INFORMATION SYSTEMS