# AI-EnviroScan: Pollution Source Identification and Visualization System

---

## Student Details

**Name:** Riya Verma
**Degree:** B.Tech in Computer Science & Engineering (3rd Year)
**University:** NIST University, Berhampur, Odisha

**Mentor:** Rahul
**Internship Program:** Infosys SpringBoard Virtual Internship 6.0
**Organization:** Infosys SpringBoard

---

## 1. Abstract

Air pollution is a major environmental and public health concern in rapidly urbanizing regions. Identifying the dominant sources of air pollution is critical for effective mitigation strategies and policy formulation. This project, *AI-EnviroScan*, presents an end-to-end intelligent system for pollution source identification, analysis, and visualization using geospatial data, machine learning, and interactive dashboards.

The system integrates real-time and historical air quality data, meteorological variables, spatial proximity features derived from OpenStreetMap, and domain-driven heuristics to label pollution sources. Machine learning models including Random Forest, Decision Tree, and XGBoost are trained to automate pollution source classification. The results are visualized through interactive heatmaps and a real-time dashboard built using Streamlit.

The proposed system demonstrates high accuracy in pollution source identification and provides an interpretable, scalable framework suitable for smart city and environmental monitoring applications.

## 2. Introduction

Air pollution has emerged as one of the most serious environmental challenges in India, especially in urban and industrial regions. Pollutants such as PM2.5, PM10, $NO_2$, $SO_2$, CO, and $O_3$ pose severe health risks. Traditional air quality monitoring systems primarily focus on pollutant concentration reporting but lack actionable insights regarding pollution sources.

With advancements in data availability, geospatial analytics, and machine learning, it is now possible to move beyond monitoring toward intelligent pollution source identification. *AI-EnviroScan* aims to bridge this gap by integrating environmental data, spatial features, and machine learning techniques into a unified decision-support system.

---

## 3. Problem Statement

The key challenges addressed in this project are:

- Lack of automated systems to identify pollution sources
- Difficulty in correlating pollutant levels with spatial features
- Absence of interactive visualization for public and policy usage
- Limited interpretability in existing black-box pollution models

**Objective:**
To design an AI-based system that identifies pollution sources using environmental, spatial, and temporal

features and presents the insights through interactive visualizations and dashboards.

---

## 4. Literature Review

Previous research in air pollution analysis primarily focuses on pollutant prediction using statistical and deep learning models. Studies have utilized regression, time-series forecasting, and neural networks for AQI estimation. However, limited work has been done on explicit pollution source identification.

Geospatial studies indicate strong correlations between pollution levels and proximity to roads, industrial areas, and agricultural activities. Recent research highlights the importance of hybrid approaches combining domain knowledge and machine learning for interpretable environmental modeling.

This project builds upon these findings by integrating rule-based labeling with supervised learning for scalable pollution source classification.

---

## 5. Data Sources & APIs

The system utilizes the following data sources:

- **OpenAQ API:** Air quality and meteorological data
- **OpenStreetMap (OSM):** Spatial features such as roads, industries, dump sites, and agricultural land
- **Government AQ Monitoring Stations:** Station metadata and coordinates

**Pollutants Collected:** PM2.5, PM10, $NO_2$, $SO_2$, CO, $O_3$
**Weather Features:** Temperature, humidity, wind speed, wind direction

---

## 6. Data Collection Methodology

Air quality data was collected programmatically using the OpenAQ API for major districts across Indian states. For each district, a unique monitoring station was selected to avoid duplication. Data was collected for the most recent 30-day window to ensure relevance.

Spatial features were extracted using OpenStreetMap data through OSMnx and GeoPandas within a 5 km radius of each monitoring location.

---

## 7. Data Preprocessing & Cleaning

The preprocessing steps included:

- Removal of invalid coordinates and timestamps
- Handling missing pollutant and weather values using median imputation
- Outlier handling using statistical thresholds
- Standardization of timestamps to Indian Standard Time (IST)
- Coordinate precision normalization

These steps ensured a clean, consistent, and ML-ready dataset.

---

## 8. Feature Engineering

Multiple categories of features were engineered:

**Environmental Features**

- PM2.5, PM10, $NO_2$, CO, $SO_2$, $O_3$

**Weather Features**

- Temperature, humidity, wind speed, wind direction

**Spatial Proximity Features**

- Distance to nearest road
- Distance to nearest industrial area
- Distance to nearest dump site

- Distance to nearest agricultural land

**Temporal Features**

- Hour, day of week, weekend indicator
- Month and season encoding

## 9. Source Labeling Strategy

Pollution sources were labeled using domain-driven heuristics:

- **Vehicular:** High $NO_2$ near roads
- **Industrial:** High $SO_2$ near industries
- **Agricultural:** High PM2.5 during dry seasons near farmland
- **Burning:** High PM2.5 with elevated CO
- **Natural:** Low pollutant concentrations

Thresholds were defined using the 75th percentile of pollutant distributions.

## 10. Machine Learning Models

The following supervised models were implemented:

- Random Forest Classifier

- Decision Tree Classifier
- XGBoost Classifier

All models were trained using pollution source labels as the target variable.

## 11. Model Training & Hyperparameter Tuning

- Dataset split: 80% training, 20% testing
- Class imbalance handled using SMOTE
- Hyperparameter tuning performed using RandomizedSearchCV
- Cross-validation applied for stability analysis

## 12. Evaluation Metrics & Results

Models were evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score (Macro)
- Confusion Matrix

The Random Forest model achieved the best overall performance with high macro-F1 score and stable cross-validation results.

## 13. Model Comparison

| Model | Accuracy | Macro F1 |
|---|---|---|
| Random Forest | Highest | Best |
| XGBoost | High | Competitive |
| Decision Tree | Moderate | Lower |

Random Forest was selected for deployment due to interpretability and robustness.

## 14. Geospatial Visualization

Pollution intensity and sources were visualized using:

- Folium-based interactive heatmaps
- Source-wise markers with toggle controls
- PM2.5 severity gradients

These maps provide intuitive spatial understanding of pollution patterns.

## 15. Dashboard Design

A real-time dashboard was developed using Streamlit featuring:

- City-wise selection
- Trend analysis (hourly/daily)
- Source distribution charts
- Interactive pollution maps
- PDF report generation

---

## 16. System Architecture

**Data Flow:**

Data Sources → Preprocessing → Feature Engineering → Source Labeling → ML Models → Dashboard & Visualization

The architecture ensures modularity, scalability, and ease of deployment.

---

## 17. Results & Discussion

The system successfully identifies dominant pollution sources with high confidence. The hybrid rule-based and ML approach ensures interpretability and practical relevance. Visualization components enhance decision-making and public awareness.

## 18. Limitations

- Dependency on availability of monitoring stations
- Rule-based labeling may introduce bias
- Limited real-time streaming capability

## 19. Future Scope

- Integration with live sensor streams
- Deep learning-based source inference
- Mobile application deployment
- Government policy integration

## 20. Conclusion

AI-EnviroScan demonstrates an effective, interpretable, and scalable solution for pollution source identification. By combining geospatial analytics, machine learning, and interactive visualization, the system provides valuable insights for environmental monitoring and smart city planning.

## 21. References

1. **OpenAQ Platform**
   OpenAQ. *Open Air Quality Data API Documentation (v3)*.
   Available at: https://docs.openaq.org/
   *(Used for air quality and meteorological data collection across Indian cities)*
2. **OpenStreetMap Contributors**
   OpenStreetMap. *OpenStreetMap Data and Contributors*.
   Available at: https://www.openstreetmap.org/
   *(Used for extracting roads, industrial zones, dump sites, and agricultural land features)*
3. **OSMnx Library**
   Boeing, G. (2017). *OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks*.
   Computers, Environment and Urban Systems, 65, 126–139.

*(Used for spatial feature extraction from OpenStreetMap)*

4. **GeoPandas Project**
   GeoPandas Developers. *GeoPandas: Python tools for geographic data*.
   Available at: https://geopandas.org/
   *(Used for geospatial processing and distance calculations)*

5. **Scikit-learn**
   Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python*.
   Journal of Machine Learning Research, 12, 2825–2830.
   *(Used for Random Forest, Decision Tree, preprocessing, evaluation, and model pipelines)*

6. **XGBoost**
   Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*.
   Proceedings of the 22nd ACM SIGKDD Conference.
   *(Used for pollution source classification)*

7. **Imbalanced-learn (SMOTE)**
   Lemaitre et al. (2017). *Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets*.

Journal of Machine Learning Research, 18(17), 1–5.
*(Used for handling class imbalance using SMOTE)*

8. **Folium Library**
   Folium Developers. *Folium: Python data visualization on interactive maps*.
   Available at: https://python-visualization.github.io/folium/
   *(Used for heatmaps and geospatial visualization)*

9. **Streamlit**
   Streamlit Inc. *Streamlit Documentation*.
   Available at: https://docs.streamlit.io/
   *(Used for building the interactive AI-EnviroScan dashboard)*

10. **World Health Organization (WHO)**
    World Health Organization. *WHO Air Quality Guidelines*.
    *(Used for understanding pollutant health thresholds and interpretation)*