AI-EnviroScan: Pollution Source Identification and Visualization

Presented by : Riya Verma, B-Tech CSE (3rd Year), NIST University, Berhampur, Odisha

Mentor : Rahul

Internship: Infosys SpringBoard Virtual Internship 6.0

Project Overview: Unveiling Pollution Sources

The challenge of air pollution monitoring often lacks precise source attribution, making targeted interventions difficult. AI-EnviroScan addresses this critical gap by automatically identifying pollution sources.

Our project integrates advanced Machine Learning techniques, spatial data analysis, and intuitive visualization tools. The primary focus is on Indian cities, where rapid urbanization demands innovative environmental monitoring solutions.

Automated Identification

Pinpointing sources with AI

ML & Spatial Data

Combining diverse datasets

Enhanced Visualization

Interactive and insightful

Data Acquisition: Fueling the AI-EnviroScan Model

Spatial Context

Coordinates, boundaries, land use

OpenAQ API

Pollutants: PM2.5, PM10, NO2

Central Data Repository

Air Quality Metadata

Timestamped sensor and station info

OpenStreetMap

Geographic coverage: roads, zones, landmarks

Key Data Sources:

- **OpenAQ API:** Provides real-time and historical pollutant and meteorological data.
- **OpenStreetMap (OSM):** Offers crucial spatial information including roads, industrial zones, waste disposal sites, and agricultural areas.

Critical Pollutants Tracked:

- PM2.5, PM10
- NO2, SO2, CO, O3

Geographic Coverage: Comprehensive data collection across Indian cities at state and district levels.

Data Preprocessing & Cleaning: Ensuring Data Integrity

Missing Value Handling

Systematic removal of incomplete records and imputation using median-based strategies to maintain data density.

Outlier Detection

Identification and management of extreme data points through statistical limits to prevent model skew.

Datetime Standardization

Conversion of all temporal data to a consistent local timezone from UTC, ensuring accurate time-series analysis.

Data Validation

Implementation of checks for invalid entries, ensuring only credible data feeds the model.

Feature Engineering: Crafting Predictive Power

Feature engineering is pivotal in transforming raw data into meaningful variables that enhance model performance.

1 Pollutant Features

Concentration levels of PM2.5, NO2, SO2, etc., providing direct pollution indicators.

2 Meteorological Features

Temperature, humidity, wind speed/direction, and precipitation influencing pollutant dispersion.

3 Spatial Proximity Features

- Distance to major roads and highways
- Proximity to industrial complexes
- Distance to waste dump sites
- Vicinity to agricultural areas and their activities

4 Temporal Features

Hour of the day, day of the week, and seasonal variations capturing cyclical patterns in pollution.

Machine Learning Model: Random Forest for Source Identification

01

Model Selection

Evaluated Decision Tree, XGBoost, and Random Forest for suitability.

02

Final Choice

Random Forest emerged as the superior model due to its robustness and accuracy.

03

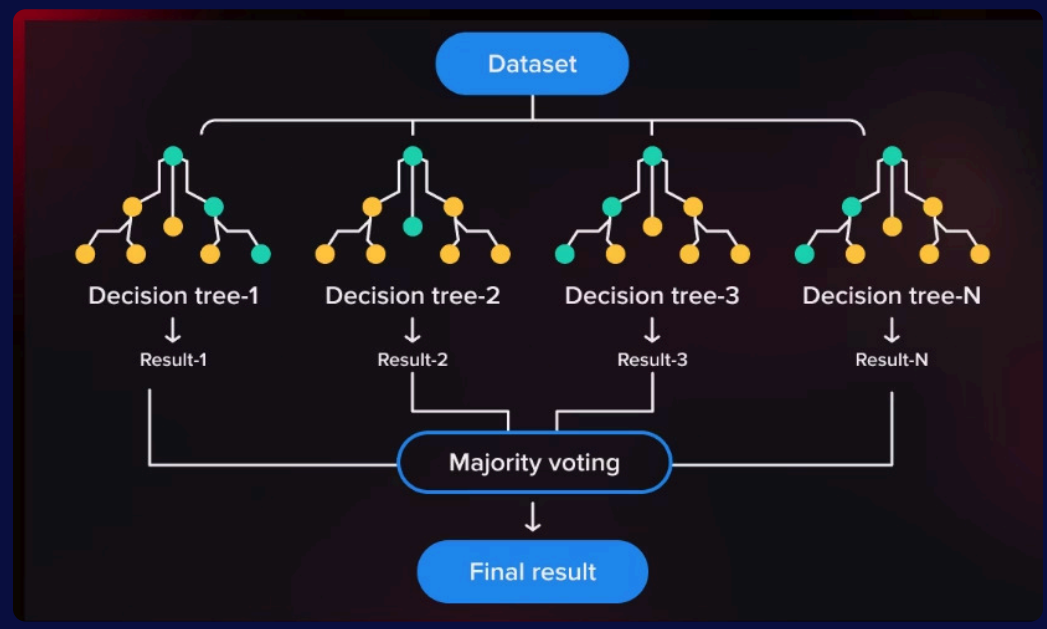
Feature Selection

Utilized importance scores to identify and prioritize the most impactful features, reducing dimensionality and improving interpretability.

04

Hyperparameter Tuning

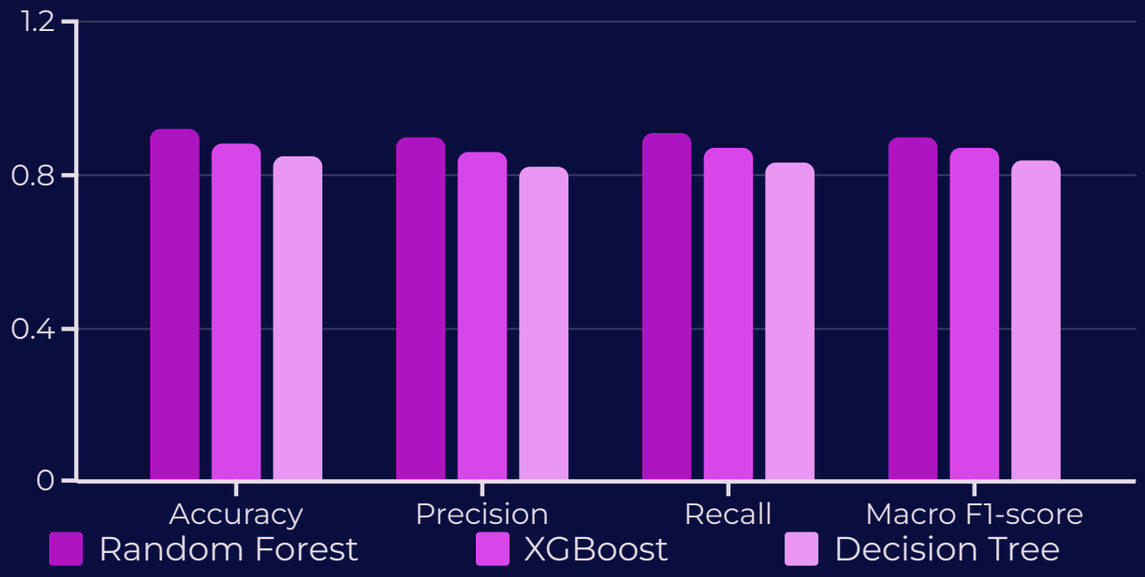
Optimized model parameters using k-fold cross-validation, enhancing generalization and preventing overfitting.



The Random Forest model aggregates predictions from multiple decision trees, resulting in improved accuracy and stability for complex datasets like pollution sources.

Model Evaluation: Performance and Validation

Rigorous evaluation using multiple metrics confirmed the Random Forest model's efficacy.



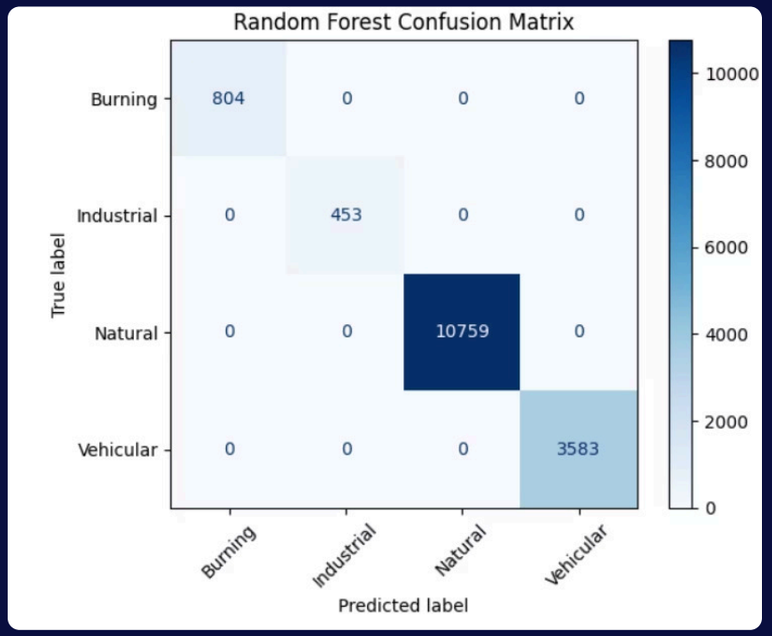

```
================ Random Forest TEST RESULTS ================

Classification Report:
              precision    recall  f1-score   support

     Burning       1.00      1.00      1.00       804
  Industrial       1.00      1.00      1.00       453
     Natural       1.00      1.00      1.00     10759
   Vehicular       1.00      1.00      1.00      3583

    accuracy                           1.00     15599
   macro avg       1.00      1.00      1.00     15599
weighted avg       1.00      1.00      1.00     15599

Macro F1-score: 1.0
```



System Architecture & Data Flow: From Raw Data to Insight

Data Collection

Gathering raw environmental and spatial data.

Preprocessing

Cleaning, standardizing, and imputing missing values.

Feature Engineering

Creating predictive features from processed data.

ML Model

Random Forest performs pollution source identification.

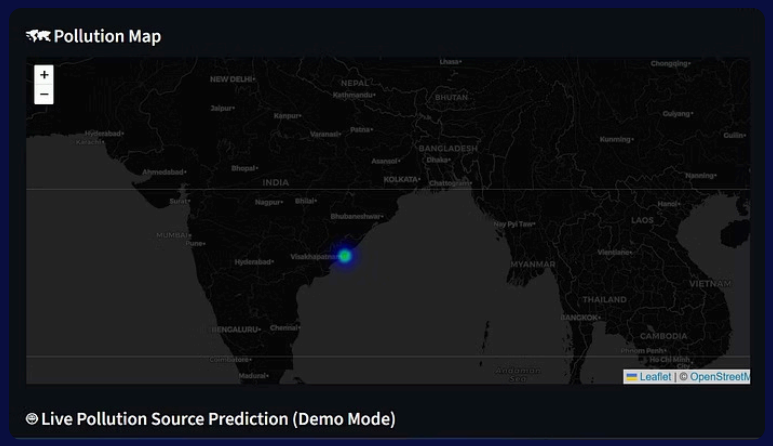
Visualization

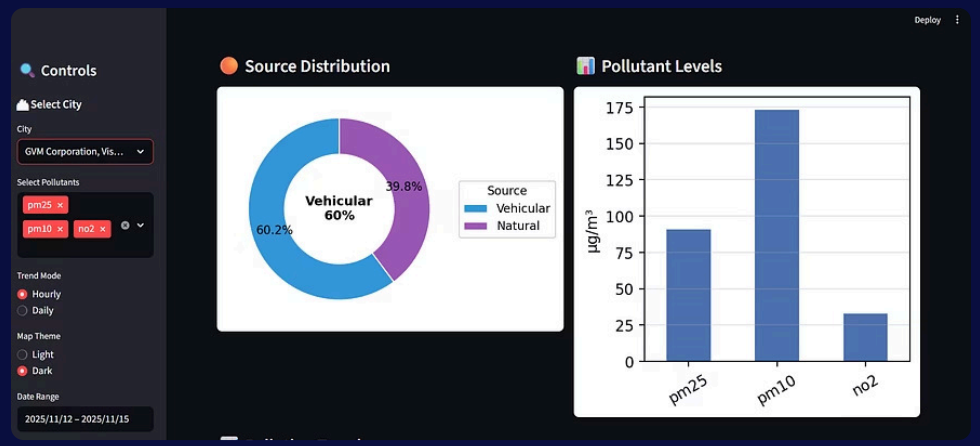
Presenting insights via interactive dashboards.

Streamlit Dashboard

User-friendly interface for monitoring.

Interactive Visualization & Dashboard: Empowering Decision-Makers





The Streamlit dashboard provides a comprehensive suite of visualization tools for granular insights into pollution patterns.

Pollution Heatmaps

Geospatial representation of pollutant concentrations.

Source Attribution Charts

Donut and bar charts detailing contributions from various sources.

Temporal Trends

Analysis of pollution variations over hours, days, and seasons.

City-Specific Analysis

Customizable views for detailed urban area assessments.

Conclusion & Future Directions

Successful Identification

Achieved precise pollution source identification.

Geospatial Integration

Effectively combined ML with location data.

User-Friendly Dashboard

Developed an intuitive visualization tool.

Environmental Impact

Enhanced monitoring for better decision-making.

AI-EnviroScan has demonstrated a robust framework for identifying and visualizing pollution sources. This project provides a valuable tool for environmental agencies and policymakers to implement targeted interventions, fostering healthier urban environments.

Future Scope:

- Integration of real-time sensor data.
- Expansion to more pollutants and regions.
- Development of predictive modeling for pollution forecasts.