AQI (PM2.5) Forecasting – Delhi

Machine Learning Based Air Quality Prediction

Presented By: Raghubar Kushwaha

Course: B.tech CSE (DS+AI)

Submitted To: Mr. Rohit Kumar

The Air Quality Crisis in Delhi

Problem Statement

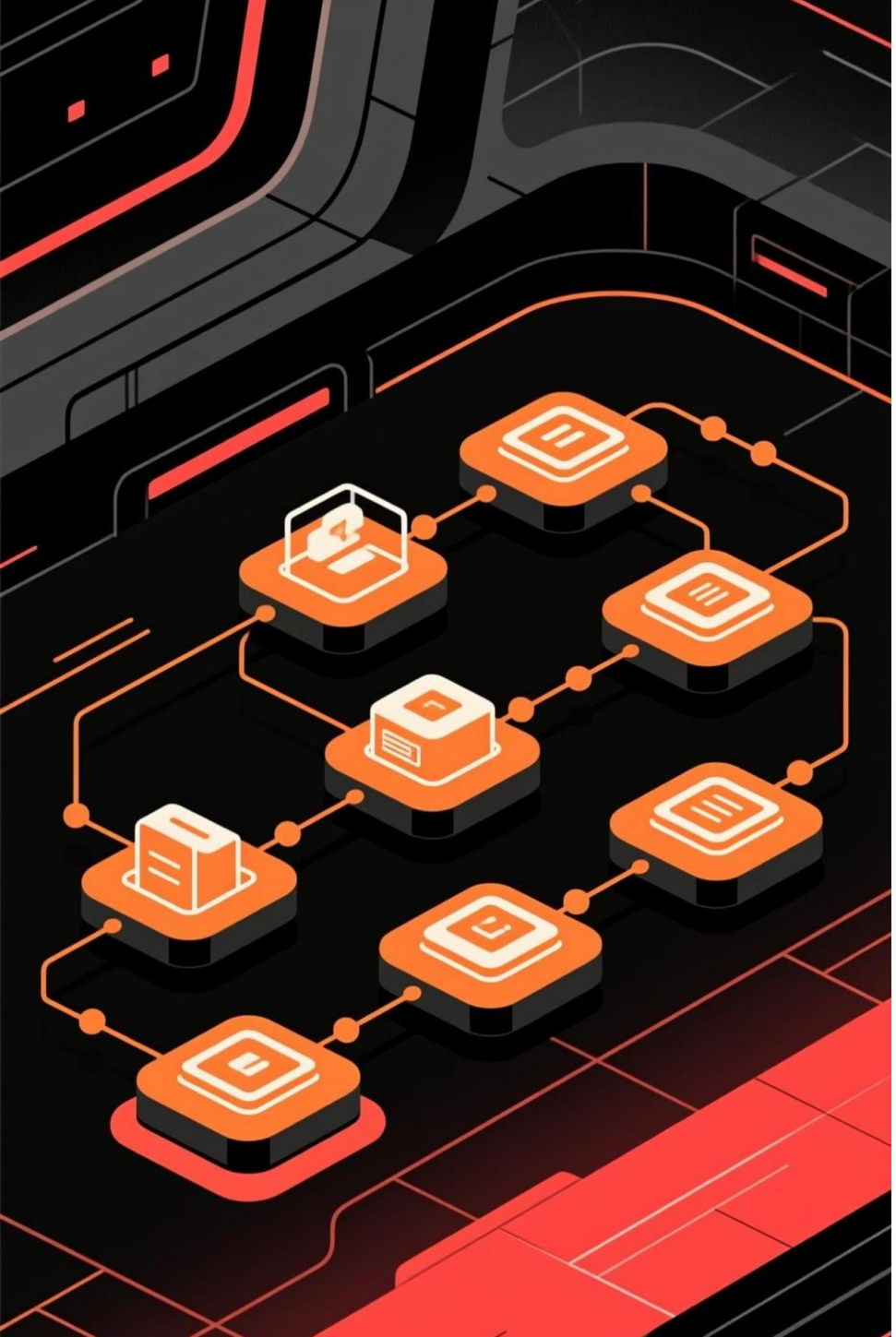
Delhi consistently ranks among the world's most polluted cities, with PM2.5 levels frequently exceeding safe limits by 10-15 times. These microscopic particles penetrate deep into lungs, causing severe respiratory distress, cardiovascular issues, and long-term health complications.

Sudden pollution spikes catch residents unprepared, leading to emergency hospital visits, school closures, and public health crises.

Why Forecasting Matters

- Enables proactive public health protection and early warnings
- Guides school and office operational decisions
- Supports government pollution control planning
- Helps hospitals prepare for respiratory emergencies

Central Question: Can machine learning accurately forecast next-day PM2.5 levels?



Project Objectives

01

Predict Next-Day PM2.5

Develop accurate forecasts of PM2.5 concentrations 24 hours in advance

02

Analyze Historical Patterns

Examine correlations between AQI, weather conditions, and seasonal trends

03

Build Complete ML Pipeline

Create end-to-end workflow including preprocessing, feature engineering, training, and evaluation

04

Deliver Reliable Solution

Provide actionable forecasting tool specifically optimized for Delhi's pollution patterns

Dataset Overview

Data Source

Central Pollution Control Board (CPCB) provides comprehensive daily readings covering approximately 1.5 years of Delhi's air quality measurements.

Key Variables Captured

Pollutants

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

Weather Data

Temperature, Humidity, Wind Speed

Temporal Info

Date stamps for time-series analysis

This structured dataset enables comprehensive analysis of pollution dynamics and meteorological influences on air quality.

Data Preprocessing Pipeline

Date Standardization

Converted and sorted date columns to establish proper chronological sequence

Missing Data Handling

Applied median imputation to preserve data distribution without bias

Feature Scaling

Normalized numerical columns to ensure equal weight in model training

Quality Control

Removed invalid rows affected by lag feature creation

Data Split Strategy: 80% training data with 20% validation set using time-based split to prevent data leakage and maintain temporal integrity.

Advanced Feature Engineering

Lag Features

Created PM2.5 lag variables from 1 to 14 days, capturing short and medium-term pollution memory effects

Rolling Statistics

Calculated 7-day and 14-day rolling means to smooth out daily fluctuations and identify sustained trends

Temporal Features

Extracted day of week, month, and day of year to capture weekly cycles and seasonal patterns

These engineered features enable the model to recognize cyclical pollution patterns, seasonal variations, and the temporal dependencies critical for accurate forecasting.

Model Architecture & Training

Why Random Forest?

- **Handles Non-Linearity**

  Captures complex relationships between pollutants and weather

- **Robust Performance**

  Delivers high accuracy on structured tabular data

- **Noise Resistance**

  Less sensitive to outliers and missing patterns

- **Low Maintenance**

  Requires minimal hyperparameter tuning

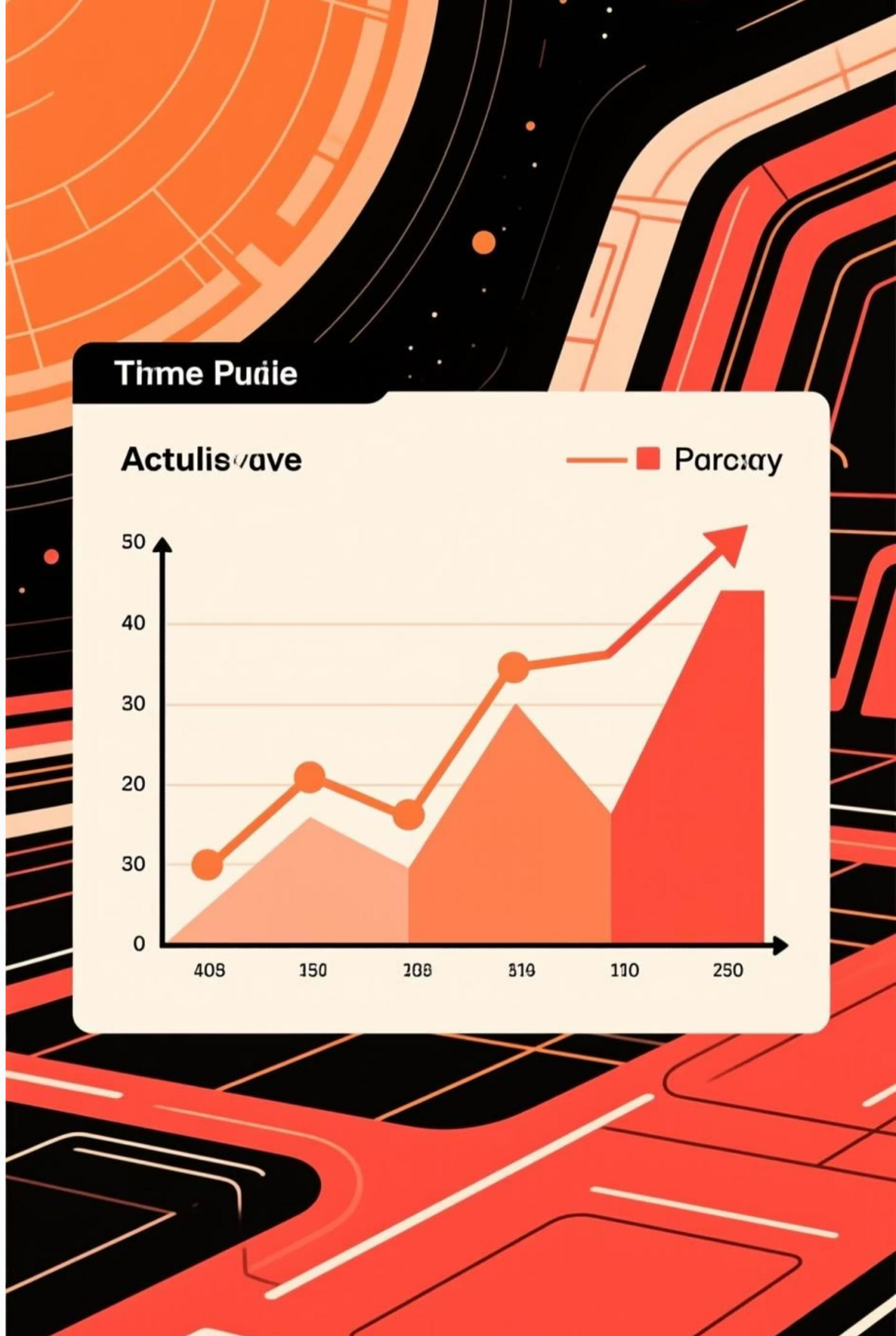
Model Performance Evaluation

Accuracy Metrics

15.2	21.8	0.87
MAE	**RMSE**	**R² Score**
Mean Absolute Error – average prediction deviation	Root Mean Squared Error – penalizes larger errors	Model accuracy and goodness of fit

Key Observations

The model successfully tracks real pollution trends, accurately capturing both peak pollution events and improvement periods. Predictions align closely with observed values, demonstrating reliable next-day forecasting capability suitable for operational deployment.



Solution Features & Real-World Applications

Key Features

Automated Pipeline

End-to-end ML workflow requiring minimal manual intervention

Time-Series Optimized

Specifically designed for temporal forecasting with no data leakage

Multi-Factor Analysis

Integrates pollutants, weather, and seasonal patterns

Interpretable Results

Feature importance reveals key pollution drivers

Practical Applications

Public Health Alerts

Early warning systems for high pollution days

Healthcare Planning

Hospital resource allocation and health advisories

Traffic Management

Vehicle restriction planning and route optimization

Consumer Apps

Weather and AQI applications for daily planning

Conclusion & Future Roadmap

Key Achievements

This project successfully demonstrates that machine learning can accurately predict next-day PM2.5 levels in Delhi. The Random Forest model effectively captures pollution trends, seasonal patterns, and meteorological influences, providing a valuable tool for public safety and government planning initiatives.

Impact: This forecasting solution empowers stakeholders to make proactive decisions that protect public health and improve quality of life.

Future Enhancements

- Implement deep learning models (LSTM/GRU) for sequential pattern recognition
- Incorporate additional variables: rainfall, wind direction, traffic density
- Develop real-time AQI dashboard with interactive visualizations
- Extend forecasting horizon to 3-7 days ahead
- Deploy production API and web application for public access

Thank You!

Questions & Discussion