Delhi Air Quality Pollution Trends and Next-Day AQI Category Prediction

Priyank Gaur, Dhruv Sareen, Shreyas Sarkar
Team Name - Teen

***Abstract*—This project presents a comprehensive analysis of air pollution trends in Delhi along with a machine learning–based next-day AQI category prediction system. A four-year daily dataset (2021–2024) containing major pollutant concentrations and temporal attributes was analyzed to study long-term trends, seasonal variations, pollutant correlations, and WHO guideline exceedances. Exploratory data analysis revealed strong seasonal patterns and dominant influence of particulate matter. A baseline PCA+KNN classifier was first implemented, followed by a leakage-free Random Forest forecasting model using only historical pollutant trends. The final model achieved a true forecasting accuracy of approximately 79% on unseen 2024 data, demonstrating the feasibility of practical AQI prediction without meteorological inputs.**

***Index Terms*—Air Quality Index, exploratory data analysis, machine learning, random forest, time-series forecasting.**

I. Introduction

Delhi is consistently ranked among the most polluted cities in the world, with air pollution posing severe risks to public health and environmental sustainability. Long-term exposure to polluted air contributes to respiratory and cardiovascular diseases and reduced quality of life. The Air Quality Index (AQI) is widely used to communicate air quality conditions to the public. This project combines detailed exploratory data analysis of pollution trends in Delhi with the development of a next-day AQI category prediction model using machine learning. The goal is to extract meaningful environmental insights and build a practical forecasting system for early warning and policy planning.

II. Related Work

Prior studies on air quality analysis and prediction employ a range of statistical and machine learning approaches including linear regression, ARIMA models, neural networks, and ensemble classifiers. Deep learning techniques such as LSTM have been used for time-series forecasting, while tree-based models such as Random Forest have shown strong performance for nonlinear pollutant interactions. Many existing approaches depend heavily on meteorological parameters. In contrast, this project focuses on extracting maximum predictive value using only historical pollutant and temporal information.

III. Methodology

A. Dataset Description

The dataset consists of 1460 daily observations collected from January 1, 2021 to December 31, 2024. It includes pollutant concentrations (PM2.5, PM10, NO_2, SO_2, CO, Ozone), date attributes (Date, Month, Year, Day), a derived AQI value, and Holidays_Count. The target variable is the AQI category with six classes: Good, Satisfactory, Moderate, Poor, Very Poor, and Severe.

B. Preprocessing

The data were sorted chronologically and verified for missing values. A proper datetime column was constructed from calendar attributes. AQI categories were generated using standard regulatory threshold rules. For predictive modeling, the dataset was split into training (2021–2022), validation (2023), and testing (2024) sets to ensure time-consistent evaluation.

C. Feature Engineering

Temporal memory was introduced using lag features at 1, 2, and 7 days for each pollutant. Rolling mean statistics were computed over 3-day and 7-day windows to capture short-term and weekly trends. Seasonal effects were encoded using categorical seasons, sine and cosine transforms of month, and a weekend indicator. All AQI-derived features were excluded from model inputs to prevent data leakage.

D. Models Used

Two classification models were evaluated. PCA coupled with KNN was used as a baseline approach to handle high-dimensional lag features. PCA reduced the feature space while KNN classified AQI categories based on similarity between past pollution profiles. A Random Forest classifier was then implemented as the final forecasting model to capture nonlinear pollutant interactions and to better handle class imbalance.

E. Training and Validation

Model performance was evaluated using Accuracy, Precision, Recall, and F1-score. Random Forest hyperparameters were tuned using the 2023 validation set. The final model was retrained on combined training and validation data and evaluated on the unseen 2024 test set.

IV. Results

A. Quantitative Results

B. Figures

V. Discussion

Exploratory analysis revealed that AQI improved from 2021 to 2023 but showed a slight increase again in 2024. PM2.5 declined significantly, while PM10 remained persistently high, indicating continuous dust and construction activity. SO_2

TABLE I
MODEL PERFORMANCE COMPARISON

Model	Acc	Prec	Rec	F1
PCA+KNN	0.68	0.66	0.65	0.63
Random Forest (Final)	**0.79**	**0.81**	**0.79**	**0.79**

```
CONFUSION MATRIX:
[[ 8  0  0  1  0  0]
 [ 0 99 18  7  0  1]
 [ 0  7 96  0  0  2]
 [11  3  0 55  0  0]
 [ 0  0  0  0  3  5]
 [ 0  0 20  0  2 28]]

CLASSIFICATION REPORT:
              precision    recall  f1-score   support

        Good       0.42      0.89      0.57         9
    Moderate       0.91      0.79      0.85       125
        Poor       0.72      0.91      0.80       105
Satisfactory       0.87      0.80      0.83        69
      Severe       0.60      0.38      0.46         8
   Very Poor       0.78      0.56      0.65        50

    accuracy                           0.79       366
   macro avg       0.72      0.72      0.69       366
weighted avg       0.81      0.79      0.79       366
```

Fig. 1. Confusion matrix visualization of Random Forest model.

showed an increasing trend, suggesting possible industrial contributions, while Ozone exhibited a gradual rise due to photochemical formation. Seasonal analysis confirmed that pollution peaks during winter due to temperature inversion, low wind speeds, and biomass burning, while monsoon seasons showed the lowest pollution levels due to rainfall washout. Correlation analysis demonstrated that PM2.5 and PM10 have the strongest positive relationship with AQI, while gaseous pollutants showed moderate influence. WHO exceedance analysis revealed that PM2.5 and PM10 exceeded safe limits on more than 95% of days. The Random Forest forecasting model significantly outperformed PCA+KNN and provided reliable next-day AQI predictions under realistic data constraints. However, prediction of extreme events such as Severe AQI remains difficult without meteorological inputs.

VI. CONCLUSION

This project successfully combined exploratory data analysis of Delhi's air pollution with a machine learning–based next-day AQI category prediction framework. Strong long-term trends, seasonal pollution cycles, dominant particulate influence, and frequent WHO limit exceedances were identified. A Random Forest forecasting model achieved approximately 79% true predictive accuracy on unseen test data without information leakage. The system demonstrates strong potential for real-world air quality monitoring and early warning systems. Future improvements will depend on integrating meteorological parameters and extending predictions beyond a one-day horizon.

REFERENCES

APPENDIX

This appendix provides details regarding the project's GitHub repository and associated file structure for reproducibility and open-source collaboration.

A. Repository Link

The complete source code, processed datasets, trained models, and documentation are available at:

https://github.com/Shreyas-Sarkar/Delhi-Air-Quality-Pollution-Trends

B. Repository Structure

The project repository follows the structure shown below:

C. Description of Key Components

- **data/**: Contains raw and processed pollutant datasets.
- **notebooks/**: Jupyter notebooks for EDA, feature engineering, and model training and evaluation.
- **reports/**: Comprehensive Project Report containing brief of analysis and model results.
- **README.md**: Project documentation and execution guidelines.