A Comprehensive Analysis of Delhi's Air Quality Using Predictive Modeling, Clustering Techniques, and Apriori-Based Knowledge Discovery (2021–2024)

Himanshu Mishra, Subhan Kumar Rai, Krish Amit Modi
Team Name: NAME
B.Tech CSE (AI & ML), Newton School of Technology
https://github.com/Himanshu197200/Delhi_Air_Quality_Analysis

***Abstract*—This study presents a complete multi-stage data mining pipeline for analyzing Delhi's air quality from 2021 to 2024. The project integrates exploratory data analysis, predictive modeling (KNN, Random Forest, Decision Tree), dimensionality reduction (PCA), clustering (K-Means), anomaly detection (DBSCAN), and association rule mining (Apriori). The objective is both predictive and explanatory: to accurately classify AQI levels and to uncover hidden pollutant interactions. Random Forest achieved the strongest predictive accuracy (98.9%), while Decision Trees provided clear interpretability. K-Means revealed seasonal pollution clusters, DBSCAN detected extreme pollution events, and Apriori highlighted strong rules such as {High PM2.5, High PM10} → {Severe AQI}. This research demonstrates an end-to-end analytical framework for environmental monitoring and air quality decision-making.**

***Index Terms*—Air Quality Index, Machine Learning, Random Forest, PCA, KNN, Decision Tree, DBSCAN, K-Means, Apriori.**

I. Introduction

Delhi consistently ranks among the world's most polluted cities due to traffic emissions, industrial activity, construction dust, biomass burning, and meteorological conditions such as winter inversion. AQI fluctuates sharply across seasons, and traditional statistical models often fail to capture its nonlinear patterns. Machine learning enables a deeper examination of pollutant interactions, while clustering and rule mining reveal hidden structures within pollution data.

This project follows a progressive five-notebook pipeline: (1) data preprocessing and EDA, (2) PCA-enhanced KNN modeling, (3) improved tree-based models, (4) clustering and anomaly detection, and (5) Apriori-based knowledge discovery. Each notebook builds on the previous one to form a complete environmental analytics ecosystem.

II. Related Work

Prior work in AQI forecasting mainly relies on regression or ARIMA models. While useful for trend analysis, these methods capture limited nonlinear behavior. Ensemble-learning models such as Random Forests have shown strong performance by modeling complex pollutant interactions. Unsupervised methods like K-Means and DBSCAN have been applied to pollution episode detection, and association rule mining has occasionally been used to analyze pollutant relationships. However, few studies combine all these techniques into a unified and interpretable framework, which is a key contribution of this project.

III. Methodology

A. Dataset Description

The dataset contains daily pollutant measurements: PM2.5, PM10, NO_2, SO_2, CO, Ozone, and AQI. Date fields were converted into a Datetime index, enabling extraction of year, month, season, and day-based attributes. The dataset spans January 2021–December 2024.

B. Preprocessing

The preprocessing pipeline includes:

- Handling missing values and inconsistent entries.
- Creating a chronological Datetime index.
- Standardizing pollutant values for ML models.
- Adding seasonal categories (Winter, Summer, Monsoon, Post-Monsoon).
- Engineering lag features (1-day, 2-day, 7-day) for temporal dependence.
- Creating rolling features (3-day, 7-day averages) to capture short-term behaviors.
- Categorizing AQI into six levels: Good, Satisfactory, Moderate, Poor, Very Poor, Severe.

C. Feature Engineering

Feature engineering focuses on enhancing model interpretability and performance:

- Lag and rolling averages capture AQI momentum and pollution cycles.
- Pollutant bins (Low/Medium/High) support Apriori rule generation.
- PCA reduces dimensionality for KNN, improving efficiency and accuracy.

D. Models Used

- **KNN**: Baseline classifier with PCA optimization.
- **Random Forest**: Strong predictive model for nonlinear relationships.
- **Decision Tree**: Simple, interpretable pollutant threshold visualization.
- **K-Means**: Identifies natural pollution patterns.
- **DBSCAN**: Detects outlier pollution events (e.g., Diwali spikes).
- **Apriori**: Extracts high-confidence pollutant–AQI rules.

E. Training Strategy

The dataset was chronologically split to avoid data leakage: 2021–2023 for training and 2024 for model evaluation. Metrics include accuracy, precision, recall, F1-score, rule lift, cluster inertia, and anomaly counts.

IV. Results

A. Model Performance

TABLE I
PERFORMANCE OF MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1
KNN(Classifier)	0.68	0.68	0.68	0.68
Random Forest	**0.989**	**0.99**	**0.99**	**0.99**
Decision Tree	1.00*	1.00	1.00	1.00

TABLE II
REGRESSION PERFORMANCE METRICS FOR KNN (REGRESSOR)

Model	Accuracy	MAE	MSE	RMSE
KNN (Regressor)	0.88	30.1443	1449.43	38.0715

**Decision Tree displays overfitting due to seasonal patterns and lag-driven predictability.*

B. Clustering Insights

K-Means revealed 3 dominant pollution clusters corresponding to monsoon clean days, transition periods, and severe winter smog. DBSCAN isolated extreme anomalies closely aligned with festival fireworks and agricultural burning episodes.

C. Apriori Rule Mining

High-lift rules provide interpretable pollutant combinations:

- $\{\text{High PM2.5, High PM10}\} \rightarrow \{\text{Severe AQI}\}$
- $\{\text{Low PM2.5, Low PM10}\} \rightarrow \{\text{Good AQI}\}$

These interactions reinforce known environmental patterns and validate model outputs.

V. Discussion

The project exhibits clear progression. PCA-enhanced KNN established a strong baseline, and Random Forest significantly improved predictive strength. Decision Trees, despite overfitting, contributed valuable interpretability. Clustering highlighted seasonal shifts and pollution regimes, while DBSCAN effectively identified unusual pollution spikes. Apriori rules served as a bridge between machine learning predictions and environmental reasoning, enabling transparent decision-making.

VI. Conclusion

This work presents an integrated pipeline that unifies pollution forecasting, clustering, anomaly detection, and knowledge extraction. The dominance of particulate matter (PM2.5 and PM10) was consistently observed across predictive modeling, clustering behavior, and rule mining. The pipeline provides a strong foundation for future environmental analytics, including LSTM time-series forecasting, real-time dashboards, and integration of satellite-based inputs.

Appendix

GitHub Repository: https://github.com/Himanshu197200/Delhi_Air_Quality_Analysis

- **data/**: Dataset files.
- **notebooks/**: Complete workflow (Notebook 1–5).
- **models/**: Exported trained models.
- **results/**: Plots, clustering outputs, Apriori rules.
- **README.md**: Steps for reproducing results.