Forecasting Delhi's Air Quality Index Using Machine Learning Models

Shaurya Pandey [1], Sharman Goswami[2]

[1] Gorakhpur Public School, Gorakhpur, Uttar Pradesh, India

[2] Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur, Uttar Pradesh, India

KEYWORDS: Air Quality, Machine Learning, Particulate Matter, Delhi, Forecasting

OVERVIEW: This study analyzed Delhi's air quality data from 2021–2024 to identify pollutants driving AQI spikes and tested machine learning models for forecasting. PM2.5 and PM10 showed the strongest correlations, and XGBoost achieved the best predictive performance.

SUMMARY

Air pollution in Delhi has become one of the world's most severe environmental challenges, with the Air Quality Index frequently reaching dangerous levels during winter months. We wanted to understand which specific pollutants contribute most to these AQI spikes and whether machine learning models could accurately predict future air quality trends. We hypothesized that particulate matter would show stronger correlation with AQI than gaseous pollutants, and that ensemble machine learning models would outperform linear regression in predictive accuracy. Using publicly available air quality data for Delhi from January 2021 through December 2024, we analyzed relationships between six major pollutants and AQI through correlation analysis and scatter plots. PM10 and PM2.5 showed extremely strong correlations with AQI (0.90 and 0.82 respectively), while nitrogen dioxide, sulfur dioxide, and ozone displayed weak or scattered patterns. Carbon monoxide demonstrated moderate correlation at 0.71. We trained three machine learning models to predict AQI: Linear Regression as baseline, Random Forest Regressor, and XGBoost Regressor. After hyperparameter tuning with randomized search and three-fold cross-validation, XGBoost achieved the best performance with RMSE of 36.51 and R^2 of 0.889. We also used the Prophet time series model to forecast future AQI levels, which predicted rising pollution during winter 2025. This forecast was later validated by observations in October–November 2025, when AQI frequently exceeded 400, confirming the model's predictive capability. These findings demonstrate that fine particulate matter drives Delhi's poor air quality and that machine learning models can effectively predict seasonal pollution patterns with meaningful accuracy.

INTRODUCTION

Each winter, thick layers of smog settle Delhi, and the Air Quality Index (AQI) frequently rises above 400, entering the "Severe" category where even healthy individuals face breathing difficulties (1, 7). Understanding what drives these extreme pollution episodes—and being able to predict them reliably—has become essential for planning public-health responses and informing residents.

AQI simplifies complex pollution data into a single number by incorporating six major pollutants: PM2.5, PM10, nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3) (20). Among these, particulate matter is especially dangerous because PM2.5 particles can travel deep into the lungs and bloodstream, increasing risks of respiratory and cardiovascular diseases (9). Studies in Delhi consistently identify PM2.5 and PM10 as the primary contributors to the city's poor air quality (5, 23).

Delhi's pollution originates from several overlapping sources. Vehicle emissions, industrial activity, road dust, and ongoing construction all contribute to elevated particulate levels throughout the year (6). During winter, however, pollution rises dramatically due to additional seasonal factors. Farmers in Punjab and Haryana burn crop stubble after harvest, and prevailing winds carry this smoke toward Delhi, sharply increasing PM2.5 and PM10 concentrations (6, 13, 23). This timing overlaps with Diwali celebrations in October–November,

when widespread use of fireworks causes short-term spikes in particulate pollution (7, 13). Winter weather conditions further worsen the situation: low wind speeds and temperature inversion trap pollutants close to the ground, preventing dispersion (6, 20).

The COVID-19 pandemic provided a rare natural experiment that highlighted how quickly Delhi's air quality responds to changes in human activity. During lockdown periods, when traffic and industrial operations sharply declined, PM10, PM2.5 and other air pollutants levels dropped significantly across the Delhi region (10, 11). As restrictions eased and economic activity returned to normal post-2021, pollution levels also began to rise again—suggesting a strong link between emissions and AQI trends.

In recent years, machine learning has become an important tool for understanding and predicting air-quality patterns. Techniques such as Random Forest (2), XGBoost (4), and other ensemble approaches are well-suited for modeling the non-linear relationships between pollutants and AQI. Time-series models like Prophet can capture Delhi's recurring seasonal cycles and longer-term pollution trends (12, 19). Several studies have successfully applied machine-learning methods to Delhi's air-quality forecasting, demonstrating that models trained on multiple pollutants often outperform traditional statistical approaches (21, 22).

However, many existing studies rely on older datasets, focus on only one or two modeling techniques, or do not validate their forecasts against actual future observations. This study addresses these gaps by analyzing detailed pollutant and AQI data from 2021–2024, comparing three different machine-learning models, and evaluating long-term forecasts against real AQI measurements from 2025. We hypothesized that PM2.5 and PM10 would show the strongest relationships with AQI and that ensemble models would outperform Linear Regression in predictive accuracy.

RESULTS

Correlation Between Pollutants and Air Quality Index

We examined relationships between pollutants and AQI through correlation coefficients and scatter plots. PM10 demonstrated an extremely strong positive correlation with AQI ($r \approx 0.9$), with data points clustering tightly along a linear trend (**Figure 1**). PM2.5 showed similarly strong correlation ($r \approx 0.8$), though with slightly more scatter **(Figure 2**). Carbon monoxide showed moderate positive correlation ($r \approx 0.7$) while nitrogen dioxide, sulfur dioxide, and ozone all showed weak relationships with AQI. When we plotted ozone against AQI, points appeared randomly distributed with no clear pattern (**Figure 3)**. The correlation heatmap confirmed these findings, showing strong coloring for PM2.5 and PM10 with AQI, moderate for carbon monoxide, and weak for remaining pollutants.

Seasonal Patterns in Air Quality

AQI is highest in winter (Nov–Jan) with many high outliers, while monsoon months (Jul–Sep) show the lowest and most stable AQI levels. The transition months (Feb–Apr, Oct) fall in the

middle. Overall, the trend highlights how winter pollution spikes and monsoon cleansing shape Delhi's air quality.

Machine Learning Model Performance

We trained three regression models to predict AQI from pollutant concentrations. Before optimization, Random Forest performed best among baseline models with RMSE of 36.88, MAE of 23.95, and R^2 of 0.887 **(Table 1).** XGBoost showed slightly better performance with RMSE of 36.32, MAE of 24.09, and R^2 of 0.890. Linear Regression showed poorer performance with RMSE of 44.51 and R^2 of 0.835, indicating it could not capture complex relationships as effectively.

After hyperparameter tuning using randomized search with cross-validation, XGBoost achieved the best overall performance with RMSE of 36.51, MAE of 23.85, and R^2 of 0.889 **(Table 2).** Tuned Random Forest showed RMSE of 36.85, MAE of 23.71, and R^2 of 0.887. Improvements from tuning were modest but made models more stable. Both ensemble methods substantially outperformed Linear Regression.

Time Series Forecasting

We applied Prophet to forecast future AQI trends. The model captured Delhi's seasonal cycle, showing AQI rising sharply during late autumn and winter and declining during spring and summer **(Figure 6).** Extended forecasts through 2025-2026 predicted AQI would peak above 300 during November-December 2025, drop to around 150 during mid-2025, and rise again during winter 2025-2026. Uncertainty intervals widened for longer-term forecasts, appropriately reflecting increased uncertainty.

Trend component analysis showed AQI declining from 2021 through mid-2023, likely reflecting reduced activity following COVID-19 restrictions **(Figure 7).** Starting late 2023, the trend reversed and began rising, indicating pollution levels were returning as economic activity normalized. The forecast extended this upward trend through 2025-2026. To validate predictions, we compared the forecast for October-November 2025 with actual observations. Monitoring data confirmed that Delhi experienced severe pollution during this period with AQI frequently exceeding 400, consistent with our forecast.

DISCUSSION

Our analysis confirmed that particulate matter, specifically PM2.5 and PM10, drives Delhi's poor air quality. The near-perfect correlations we observed (0.8-0.9) were stronger than typical environmental data, where confounding factors add noise. This finding carries important policy implications—interventions targeting particulate emissions would have the most direct impact on improving AQI.

Particulate matter comes from multiple sources. Vehicle exhaust contributes directly through tailpipe emissions and indirectly through brake dust and tire wear. Construction generates

substantial dust. Road dust gets resuspended by traffic. Industrial facilities emit particulates from combustion. During winter, these baseline sources continue while additional factors intensify the problem such as stubble burning and fireworks during Diwali.

The seasonal pattern we documented confirms what Delhi residents experience yearly. Winter brings the worst air quality. Temperature inversion traps pollutants when cold dense air cannot rise through warmer air above. Wind speeds decrease, reducing horizontal dispersion. Farmers burn crop stubble in Punjab and Haryana, sending smoke toward Delhi. Diwali fireworks add emission spikes. Our data clearly captured these effects, with highest AQI spikes in late October and early November.

The AQI decline from 2021 through mid-2023 provides insights into how human activity influences air quality. This period corresponded to recovery from COVID-19 restrictions. Even after lockdowns ended, many offices maintained work-from-home policies, industrial production remained below pre-pandemic levels, and traffic stayed reduced. This demonstrated that human activities drive much of Delhi's pollution. However, the upward trend from late 2023 through 2024 suggests the city returned to or exceeded previous pollution levels as offices reopened, construction resumed, and traffic returned to pre-pandemic congestion.

Among our models, XGBoost achieved best performance, though its advantage over Random Forest was small. Both ensemble methods substantially outperformed Linear Regression, confirming that non-linear models better capture air quality dynamics. This makes sense because air quality results from complex interactions between pollutants, weather, time patterns, and episodic events that linear relationships cannot capture.

The modest improvement from hyperparameter tuning suggests our models already captured most predictable patterns in the data. Further improvements would likely require additional input features rather than more sophisticated algorithms. Meteorological variables would be the logical addition—wind speed affects dispersion, temperature affects emissions and dispersion patterns, humidity influences particle formation, and rainfall removes particles through wet deposition.

One strength of our study is successful validation of our Prophet forecast. When we generated predictions in early 2024 for the following year, the model correctly anticipated severe AQI increases during October-November 2025. Actual monitoring confirmed Delhi experienced severe pollution during this period, with multiple days exceeding AQI of 400. This demonstrates that machine learning forecasting can provide reliable predictions, giving authorities advance warning to implement emergency measures.

However, our study has limitations. First, we used data from a single monitoring station, which may not represent air quality across all of Delhi. Different neighborhoods near highways, industrial zones, or parks likely experience different pollution levels. Second, our models used only pollutant data and excluded meteorological variables like wind speed, temperature, humidity, and rainfall, which significantly affect dispersion and accumulation. Including these would likely improve accuracy. Third, our models assume stable relationships over time, but

changes in emissions, regulations, or measurement methods could alter these relationships. Fourth, our forecasting cannot account for sudden unusual events like emergency traffic restrictions, extreme weather, or events like COVID-19 lockdowns.

Looking forward, authorities could use these findings to design effective interventions. Since PM2.5 and PM10 dominate AQI, policies should prioritize reducing these pollutants. Vehicle emissions can be addressed through stricter standards, accelerated retirement of old vehicles, metro expansion, and electric vehicle promotion. Construction dust can be controlled through better site management and watering of exposed soil. Stubble burning requires coordinated action between Delhi and neighboring states, providing farmers with affordable alternatives like mechanical choppers. Addressing Diwali firecrackers faces cultural sensitivities but education campaigns emphasizing health impacts might gradually shift behavior.

Our forecasting approach could become an early warning system. If predictions indicate severe pollution in five to seven days, authorities could implement temporary measures like vehicle rationing, construction halts, or industrial production restrictions rather than reacting after air quality deteriorates.

Future research should incorporate multiple monitoring stations for spatial context, add meteorological variables to improve accuracy, test advanced deep learning like LSTM networks to capture longer temporal dependencies, attempt source apportionment to quantify different emission sources, incorporate satellite observations for broader spatial coverage, and conduct economic analysis of intervention strategies to help policymakers prioritize cost-effective approaches.

In conclusion, this study provides data-driven evidence that particulate matter dominates Delhi's air quality crisis. The strong correlations between PM2.5, PM10, and AQI indicate that policies targeting these pollutants would have direct impact. Our findings regarding seasonal patterns quantitatively confirm the severe winter pollution problem and explain the factors driving it. Successful application of machine learning, particularly XGBoost, demonstrates these tools can predict air quality with meaningful accuracy. Validation of our forecast against actual 2025 observations proves such predictions have real-world utility. With continued refinement, machine learning-based forecasting could become valuable for Delhi's environmental management, enabling proactive responses to pollution episodes.

MATERIALS AND METHODS

Data Collection and Preparation

We obtained hourly air quality data for Delhi from January 2021 through December 2024 from a publicly available Kaggle dataset. The dataset included measurements of six pollutants (PM2.5, PM10, NO_2, SO_2, CO, O_3) and corresponding AQI values, comprising approximately 35,000 observations. We loaded data using Python's Pandas library and conducted quality checks. The dataset contained no missing values. We converted timestamps to datetime format and extracted temporal features including month, year, and season to enable seasonal analysis.

Exploratory Data Analysis

We calculated Pearson correlation coefficients between each pollutant and AQI to quantify relationship strength. We created scatter plots for each pollutant against AQI and generated a correlation heatmap to visualize all pairwise correlations. For temporal analysis, we grouped data by month and created box plots showing AQI distributions for each month.

Model Development and Training

We separated pollutant concentrations as input features and AQI as the target variable. We split data chronologically, using the first 80% for training and remaining 20% for testing. This chronological split simulates real forecasting scenarios better than random sampling. We selected three algorithms: Linear Regression as baseline, Random Forest Regressor for ensemble learning, and XGBoost Regressor for gradient boosting. We implemented models using scikit-learn and XGBoost libraries.

Hyperparameter Tuning

For Random Forest and XGBoost, we performed hyperparameter tuning using RandomizedSearchCV with three-fold cross-validation. For Random Forest, we tested number of estimators (100, 200, 300, 400, 500), maximum depth (None, 10, 20, 30, 40), and minimum samples split (2, 5, 10). For XGBoost, we tested number of estimators (200, 400, 600), maximum depth (3, 5, 7, 9), learning rate (0.01, 0.05, 0.1), subsample ratio (0.7, 0.8, 1.0), and column subsample ratio (0.7, 0.9, 1.0). The process evaluated 20 random combinations for each model through cross-validation, selecting configurations that minimized prediction errors. The optimal Random Forest configuration included 300 estimators, minimum samples split of 10, and unrestricted maximum depth. The optimal XGBoost configuration included 200 estimators, maximum depth of 3, learning rate of 0.05, subsample ratio of 0.8, and column subsample ratio of 1.0.

Model Evaluation

Accuracy can be misleading in regression-based models. Therefore, the performance of each model was monitored using three metrics: Root Mean Squared Error, which penalizes large errors; Mean Absolute Error, representing average error magnitude; and R^2 score, indicating proportion of variance explained. Lower RMSE and MAE indicate better performance, while R^2 closer to 1.0 indicates better fit.

Time Series Forecasting

For long-term forecasting, we used the Prophet library. We prepared data in Prophet's required format with timestamp and AQI columns. We fitted the model on 2021-2024 data and generated forecasts through 2026. Prophet automatically detected seasonal patterns and decomposed forecasts into trend and seasonal components.

Code Availability

All code used for data analysis, exploratory data analysis, correlation calculations, model training, hyperparameter tuning, and visualization in this study was developed in Jupyter Notebook and is publicly available at https://github.com/Shaurya-git-bit/Delhi-AQI-Time-Series-Forecasting-with-Prophet. The repository includes Python scripts for data preprocessing, implementation of Linear Regression, Random Forest, and XGBoost models, as well as Prophet forecasting code. The code is documented with comments to facilitate reproducibility and allow other researchers to replicate our methodology or apply similar approaches to air quality data from other cities.

REFERENCES

1. Air Quality Index India. "New Delhi Particulate Matter (PM2.5) Level." AQI.in. https://www.aqi.in/in/dashboard/india/delhi/new-delhi/pm
2. Breiman, L. "Random Forests." Machine Learning, vol. 45, no. 1, 2001, pp. 5–32. https://doi.org/10.1023/A:1010933404324
3. California Air Resources Board. "Inhalable Particulate Matter and Health." https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health
4. Chen, T., and C. Guestrin. "XGBoost: A Scalable Tree Boosting System." Proceedings of the 22nd ACM SIGKDD, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785
5. Guttikunda, S. K., et al. "Sources of PM2.5 Pollution in Delhi, India." Environmental Science & Pollution Research, vol. 27, 2020, pp. 1–13. https://doi.org/10.1007/s11356-019-04441-9
6. Jakson. "Delhi's Perennial Winter Smog – Causes and Possible Solutions." Jakson.com. https://www.jakson.com/blogs/delhi-perennial-winter-smog-causes-and-possible-solutions/
7. NDTV. "Alarming AQI: Air Pollution Rising Amid Diwali Celebrations; Here's How to Stay Safe." NDTV.com, 16 October 2025. https://www.ndtv.com/health/alarming-aqi-air-pollution-rising-amid-diwali-celebrations-heres-how-to-stay-safe-9465202
8. Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html
9. Pope, C. A., and D. W. Dockery. "Health Effects of Fine Particulate Air Pollution." Journal of the Air & Waste Management Association, vol. 56, 2006, pp. 709–742. https://doi.org/10.1080/10473289.2006.10464485
10. Sen Roy, S., and R. C. Balling, Jr. "Impact of the COVID-19 Lockdown on Air Quality in the Delhi Metropolitan Region." Applied Geography, vol. 128, 2021, 102418. https://doi.org/10.1016/j.apgeog.2021.102418
11. Singh, V., et al. "Air Quality Assessment in Delhi: Before and After COVID-19 Lockdown." Science of the Total Environment, vol. 742, 2020, 140698. https://doi.org/10.1016/j.scitotenv.2020.140698

12. Taylor, S. J., and B. Letham. "Forecasting at Scale." The American Statistician, vol. 72, no. 1, 2018, pp. 37–45. https://doi.org/10.1080/00031305.2017.1380080
13. Times of India. "What Makes Delhi's Air Thick Isn't What Makes It Toxic: Study Finds Farm Fires Choke Lungs, Vehicles Poison Blood." TimesofIndia.indiatimes.com, 29 October 2025. https://timesofindia.indiatimes.com/city/delhi/what-makes-delhis-air-thick-isnt-what-makes-it-toxic-study-finds-farm-fires-choke-lungs-vehicles-poison-blood/articleshow/124940702.cms
14. AQI.in. "India Air Quality Index (AQI) : Real-Time Air Pollution." https://www.aqi.in/in/dashboard/india
15. Hunter, J. D. "Matplotlib: A 2D Graphics Environment." https://matplotlib.org/stable/project/citing.html
16. AQICN.org. "Delhi Air Pollution: Real-time Air Quality Index (AQI)." https://aqicn.org/city/delhi/
17. Waskom, M. L. "Seaborn: Statistical Data Visualization." https://www.geeksforgeeks.org/data-visualization/data-visualization-with-python-seaborn/
18. Central Pollution Control Board (CPCB). "Day wise, State wise Air Quality Index (AQI) of Major Cities and Towns in India." https://dataful.in/datasets/18571/
19. Taylor, S. J., and B. Letham. "Forecasting at Scale." http://facebook.github.io/prophet/
20. Central Pollution Control Board. "National Air Quality Index." CPCB.nic.in. https://cpcb.nic.in/National-Air-Quality-Index/
21. Kumar, P., et al. "Meteorological Influences on PM2.5 Concentrations in Delhi Using Machine Learning." Atmospheric Environment, vol. 245, 2021, 118056. https://doi.org/10.1016/j.atmosenv.2020.118056
22. Masood, A., and A. Ahmad. "Comparative Analysis of LSTM and Random Forest for Delhi Air Quality Prediction." Journal of Ambient Intelligence and Humanized Computing, vol. 14, 2023, pp. 1-15. https://doi.org/10.1007/s12652-023-04512-3
23. Sahu, S. K., et al. "Source Apportionment of PM2.5 in Delhi Using Positive Matrix Factorization." Aerosol and Air Quality Research, vol. 21, 2021, 210045. https://doi.org/10.4209/aaqr.200045
24. Sharma, S., et al. "Spatiotemporal Variation of Air Quality in Delhi Using Multiple Monitoring Stations." Environmental Monitoring and Assessment, vol. 195, 2023, 456. https://doi.org/10.1007/s10661-023-11023-7
25. Bhatia, Kunsh. "Delhi Air Quality Dataset." Kaggle, https://www.kaggle.com/datasets/kunshbhatia/delhi-air-quality-dataset

FIGURES AND FIGURE TITLES/CAPTIONS

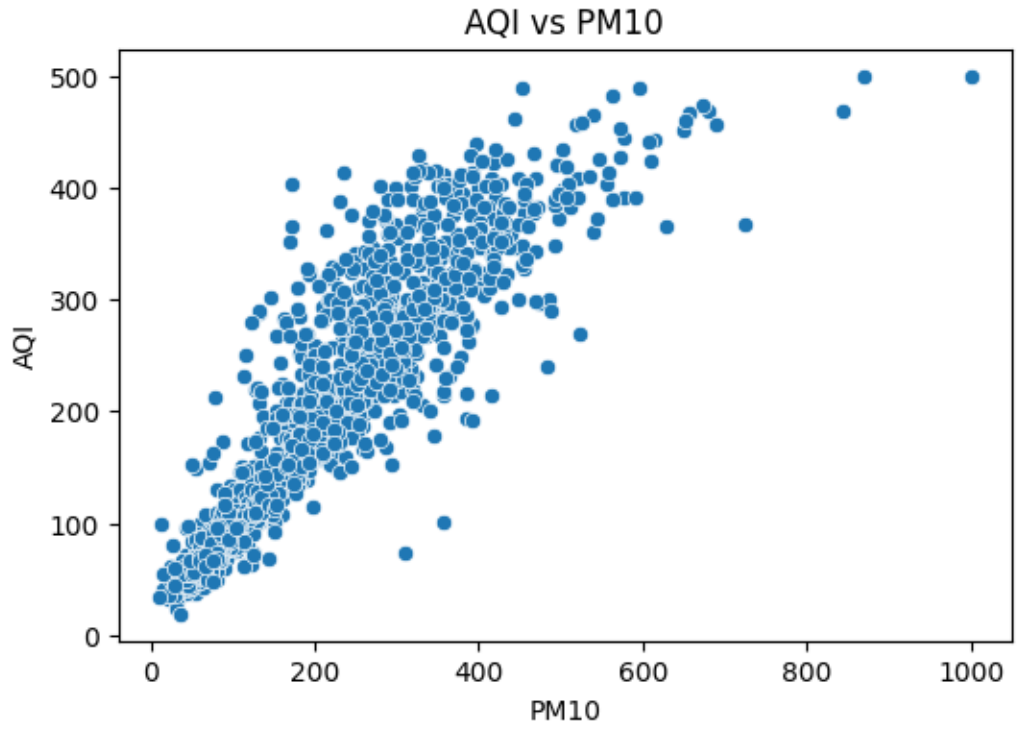

Figure 1. AQI vs PM10. Scatter plot showing how strongly AQI rises with increasing PM10 levels. The points form a tight upward trend, reflecting the very high correlation (about 0.9). This makes PM10 one of the clearest and most reliable indicators of Delhi's day-to-day air quality.

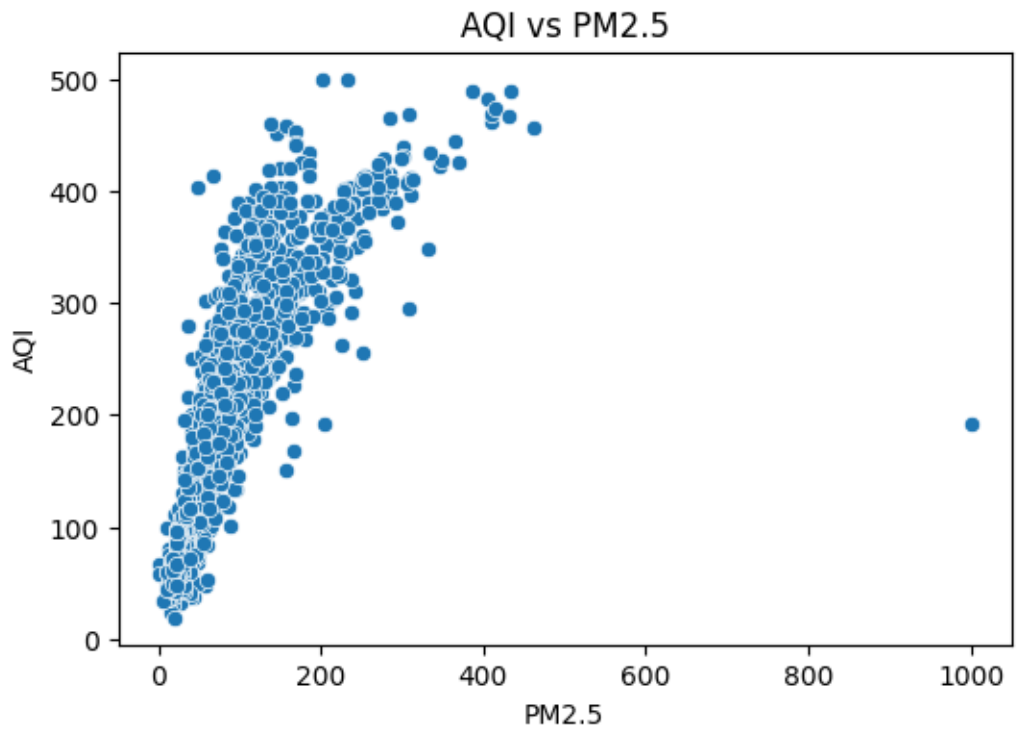

Figure 2. AQI vs PM2.5. Scatter plot showing the relationship between PM2.5 and AQI. The pattern still climbs upward, but with a bit more scatter compared to PM10, matching its slightly lower correlation (around 0.8). PM2.5 clearly influences AQI, just not as powerfully or as consistently as PM10.

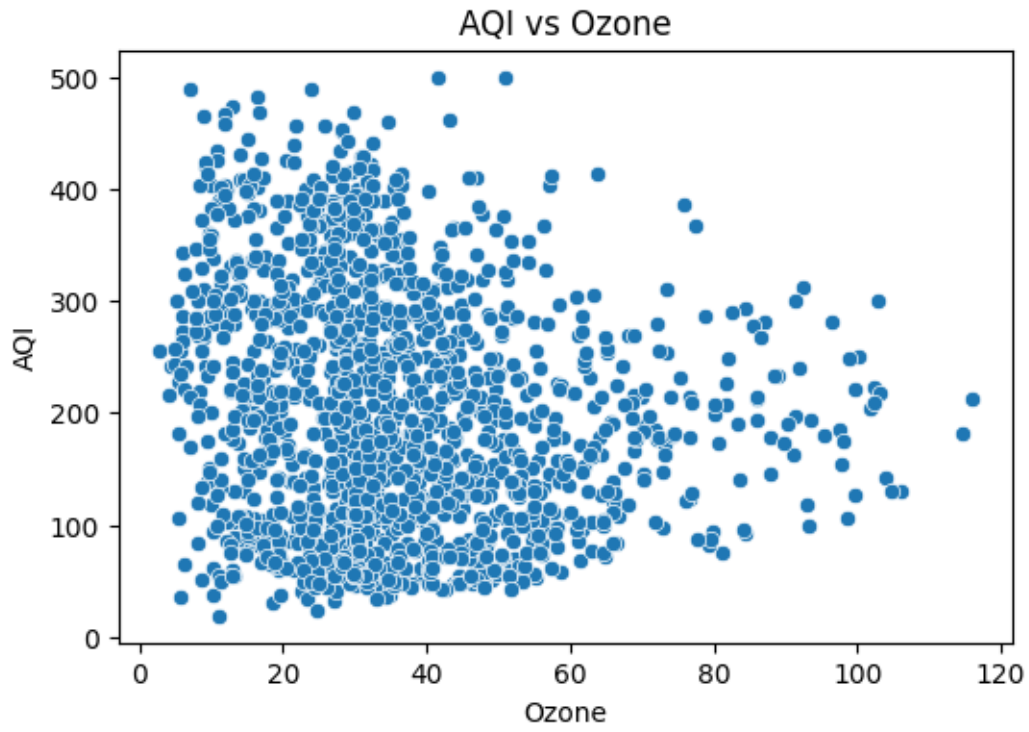

Figure 3. AQI vs Ozone. Scatter plot showing how ozone levels relate to AQI. The points are widely scattered, reflecting that ozone is very weakly connected to any rise or fall in AQI, especially compared to pollutants like PM10 and PM2.5.

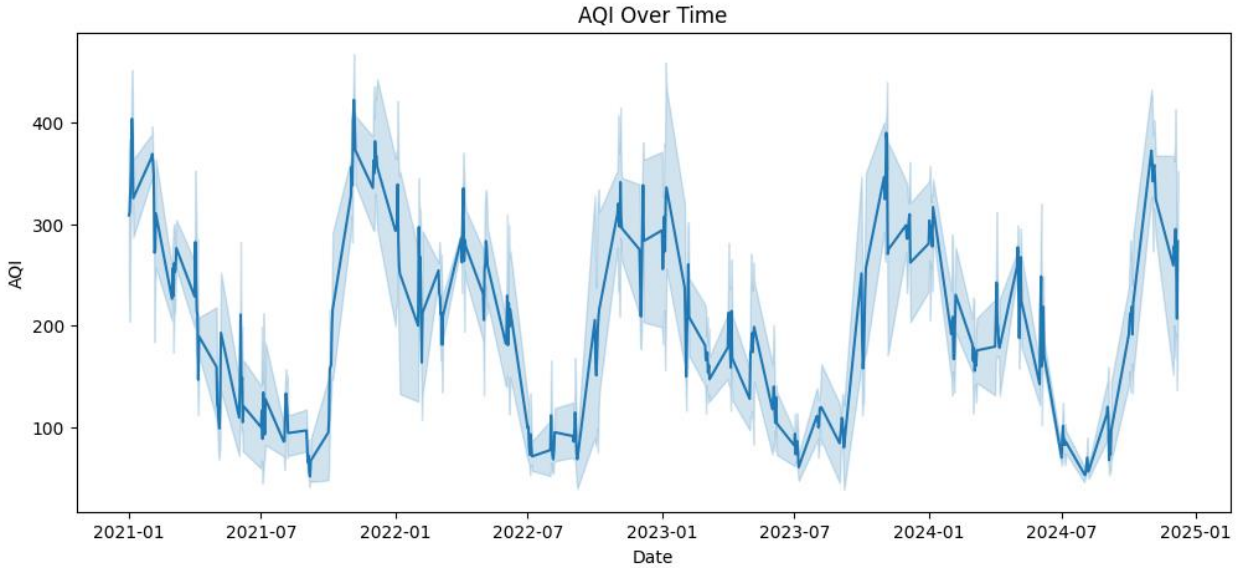

Figure 4. Line Plot of AQI over time. Line plot showing how Delhi's AQI changes over time. Clear seasonal swings appear throughout the years, with pollution rising sharply in the winter

months and easing during the monsoon and early summer—highlighting how predictable and recurring these air quality patterns have become.

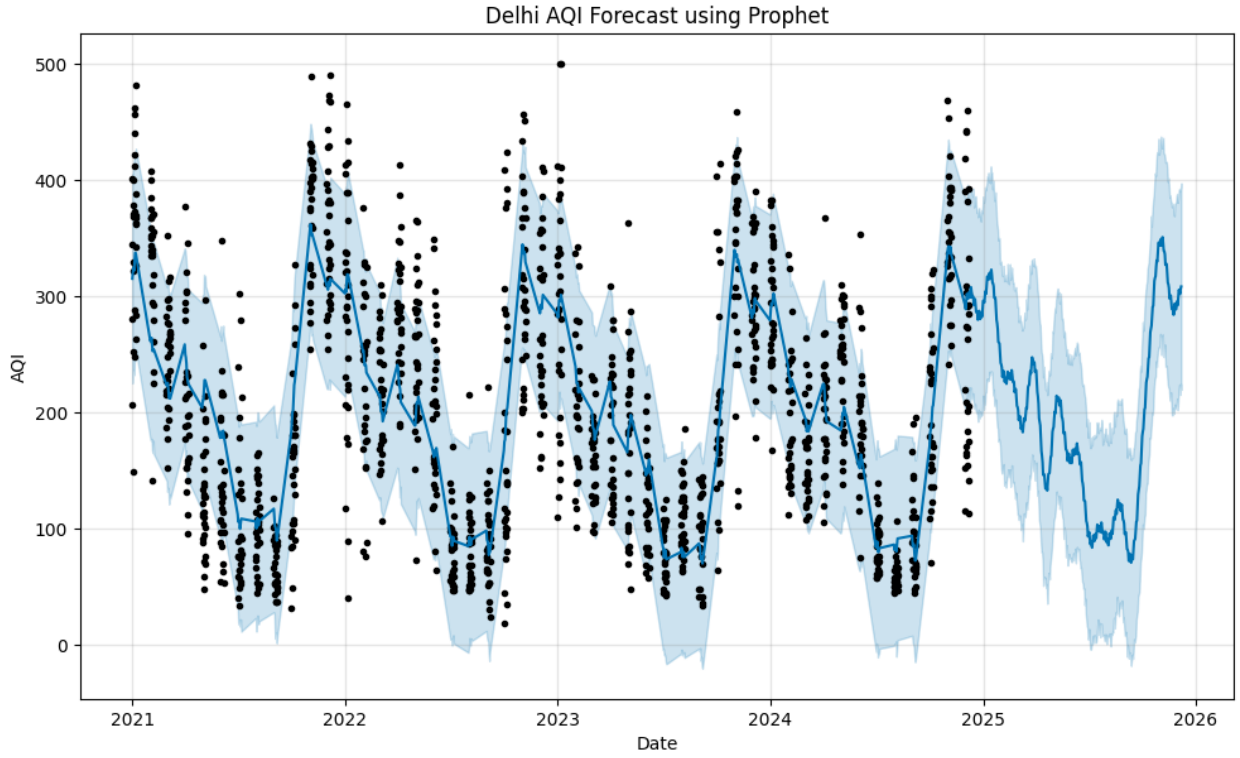

Figure 5. Delhi AQI Forecast using Prophet. Forecast of Delhi's AQI using Prophet, with the model capturing the city's repeating seasonal spikes—especially the sharp winter surges. The black dots represent the actual daily AQI measurements, the solid blue line shows Prophet's fitted trend and forecast, and the light blue shaded region indicates the model's uncertainty range around that forecast. The forecast follows the overall pattern of past years, reflecting how persistent and predictable Delhi's pollution cycles have become.

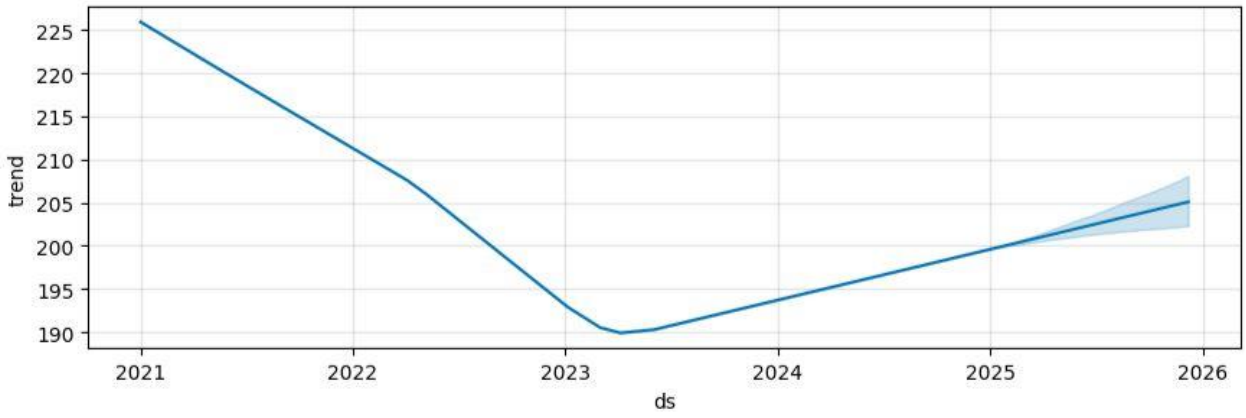

Figure 6. Prophet's Trend Component (yearly). Trend component showing baseline AQI from 2021 through 2026 forecast. AQI declined from approximately 225 to 190 between 2021 and early-2023, then reversed and began rising through 2024. Forecast shows continued increase through 2026.

TABLES WITH TITLES/CAPTIONS

Model	RMSE	MAE	R^2
Linear Regression	44.514	30.537	0.835
Random forest	36.884	23.954	0.887
XGBoost	36.317	24.093	0.890

Table 1. Baseline model comparison shows ensemble methods before Hyperparameter Tuning. This table compares the initial performance of Linear Regression, Random Forest, and XGBoost on the AQI prediction task. RMSE and MAE reflect overall error, while R^2 indicates how much variance each model explains. The ensemble models, particularly Random Forest and XGBoost, perform noticeably better than Linear Regression even before tuning.

Model (Tuned)	RMSE	MAE	R^2
Random Forest	36.848	23.705	0.887
XGBoost	36.513	23.847	0.889

Table 2. Hyperparameter tuning improved XGBoost performance. This table shows how tuning affected model accuracy. RandomizedSearchCV with three-fold cross-validation improved both Random Forest and XGBoost, with XGBoost achieving the lowest RMSE and highest overall performance. Linear Regression is excluded because it has no tunable hyperparameters.