Air Quality Prediction Using Machine Learning Algorithm in Maharashtra

Tarun Vinod Pandit[1], Prof. Vishnupant Potdar[2], Dr. Nagnath Biradar[3]
[1]*Department of Data Science,*
[2]*Guide, Department of Data Science,*
[3]*Co-Guide, Department of Data Science,*
Symbiosis Skills and Professional University

Abstract- **Air pollution poses a severe and multifaceted risk to public health, economic productivity, and environmental sustainability, particularly in rapidly urbanizing regions such as Maharashtra, India. This research presents a comprehensive comparison of classical and ensemble machine learning algorithms—for example, Linear Regression, Random Forest, and XGBoost—in forecasting the Air Quality Index (AQI) across three major metropolitan areas: Mumbai, Pune, and Nagpur. The dataset comprises hourly pollutant concentrations (PM2.5, PM10, O3, NO2, CO, SO2) and meteorological variables (temperature, humidity, wind speed, rainfall) collected via Mendeley Data and the CPCB API over a two-year period. Key steps include rigorous data preprocessing, advanced feature engineering—including lag and interaction terms—and systematic hyperparameter tuning with five-fold cross-validation. Model performance is evaluated using RMSE, MAE, and R^2 metrics. Results indicate that XGBoost consistently yields superior predictive accuracy (R^2 up to 0.92), while Random Forest offers robust interpretability through feature importance analysis. City-specific findings reveal that PM2.5 and NO2 are the dominant drivers of AQI variation in Mumbai, whereas meteorological factors play a larger role in Pune and Nagpur. These insights can guide targeted mitigation strategies and inform data-driven policy development.**

Index Terms- **Air Quality Index, Machine Learning, Maharashtra, XGBoost, Random Forest, Pollution Forecasting, Feature Engineering, Hyperparameter Tuning**

I. INTRODUCTION

Air pollution is one of the most pressing environmental challenges facing India today, exacerbated by rapid industrialization, increasing vehicular traffic, and variable climatic conditions. Maharashtra, home to over 100 million residents, consistently records AQI values that exceed safe thresholds defined by the World Health Organization, particularly in urban centers like Mumbai, Pune, and Nagpur. High particulate matter and gaseous pollutant concentrations contribute to respiratory and cardiovascular diseases, imposing significant health burdens and economic costs. Accurate forecasting of AQI can enable authorities to implement proactive public health advisories, optimize traffic regulation, and deploy pollution-control measures. This research aims to develop and compare multiple machine learning models tailored to regional characteristics of Maharashtra, thereby enhancing the precision and applicability of AQI predictions for local governance and community awareness.

II. LITERATURE REVIEW

Previous studies have demonstrated the efficacy of machine learning approaches in air quality forecasting. For instance, LSTM-based deep learning models applied to Beijing and Delhi datasets achieved R^2 values above 0.85, highlighting the importance of capturing temporal dependencies. Random Forest and hybrid ensemble methods have also shown robust performance in urban settings, balancing accuracy and computational efficiency. However, most literature focuses on North Indian or East Asian megacities, with limited exploration of Maharashtra's unique industrial and meteorological profiles. Additionally, existing works often omit comprehensive hyperparameter optimization and localized feature engineering. This paper addresses these gaps by integrating region-specific emissions data from the Maharashtra Pollution Control Board (MPCB) and applying systematic model tuning protocols, thereby offering a more granular understanding of pollutant dynamics in Maharashtra.

III. METHODOLOGY

3.1 Data Collection

- Primary Source: Mendeley Data repository "Air Quality Index of Major Indian Cities" (2024), providing hourly readings for PM2.5, PM10, O_3, NO_2, CO, and SO_2.
- Supplementary Source: Central Pollution Control Board (CPCB) API and Maharashtra

Pollution Control Board (MPCB) emission inventories, offering station-level pollutant data and emission factors.

- Meteorological Data: Indian Meteorological Department (IMD) archives for temperature (°C), humidity (%), wind speed (m/s), and daily rainfall (mm).
- Geographic Coverage: Data from 15 monitoring stations across Mumbai, Pune, and Nagpur between January 2022 and December 2023.

3.2 Data Preprocessing

- Time Alignment: Consolidated hourly pollutant and weather readings into daily aggregates (mean, max, and min) to reduce noise and computational load.
- Missing Values: Applied linear interpolation for short gaps (<6 hours) and KNN imputation for larger gaps, followed by manual verification against neighboring station trends.
- Outlier Detection: Removed observations exceeding three standard deviations (Z-score > 3), accounting for sensor malfunctions and extreme weather events.
- Feature Scaling: Employed Min-Max normalization to map all continuous variables to the [0,1] range, facilitating model convergence.

3.3 Feature Engineering

- Temporal Features: Lag variables up to seven days for primary pollutants (e.g., PM2.5_t−1, PM2.5_t−3) to capture autocorrelation.
- Interaction Terms: Pollutant ratio features (PM2.5/PM10, NO_2/O_3) to highlight combined pollutant effects.
- Categorical Encoding: One-hot encoded day-of-week and month to capture weekly and seasonal patterns.
- Emission Inventory Integration: Incorporated monthly emission factors from MPCB reports as static features for each station.

3.4 Model Selection and Training

- Baseline: Multiple Linear Regression to benchmark linear relationships.
- Ensemble Models: Random Forest Regressor (n_estimators=100–300, max_depth=10–30) and XGBoost (learning_rate=0.01–0.1, n_estimators=200, max_depth=6), tuned via grid search.
- Data Partitioning: 70% of data for training, 30% for testing, ensuring stratified sampling by city and season.
- Validation: Five-fold cross-validation on the training set to mitigate overfitting and ensure model generalizability.

IV. RESULTS AND EVALUATION

Ciy	**Model**	**RMSE**	**MAE**	**R^2**
Mumbai	XGBoost	12.4	8.9	0.91
Pune	Random Forest	14.1	10.3	0.88
Nagpur	XGBoost	11.7	8.1	0.92

XGBoost outperformed both Linear Regression and Random Forest in Mumbai and Nagpur, indicating its stronger capacity to capture nonlinear pollutant–weather interactions. Random Forest provided competitive results in Pune, where variable meteorological influences necessitate robust ensemble averaging. Feature importance analysis (via SHAP values) identified PM2.5, NO_2, and temperature as the top three predictors for all cities, underscoring the critical role of fine particulate matter and urban heat dynamics in AQI fluctuations.

V. ENTITY-RELATIONSHIP DIAGRAM OVERVIEW

Entities:

- data schema is designed to support scalable ingestion and querying, comprising the following entities:
- CityStation: station_id (PK), city_name, latitude, longitude
- PollutantData: record_id (PK), station_id (FK), timestamp, PM2.5, PM10, NO_2, CO, O_3, SO_2
- WeatherData: record_id (PK), station_id (FK), timestamp, temperature, humidity, wind_speed, rainfall
- AQIRecord: record_id (PK), station_id (FK), date, AQI_value, AQI_category
- MLModel: model_id (PK), algorithm, hyperparameters, train_date
- PredictionResult: result_id (PK), model_id (FK), station_id (FK), date, predicted_AQI, error_metrics

VI. DISCUSSION

The superior performance of XGBoost highlights the importance of gradient boosting in handling complex, high-dimensional datasets typical of urban air quality monitoring. The pronounced significance of PM2.5 and NO_2 aligns with epidemiological studies linking these pollutants to adverse health outcomes. Model interpretability via SHAP allowed us to pinpoint pollutant–meteorology interactions, such as the exacerbating effect of high temperatures on ozone formation. These insights can inform targeted interventions—for example, optimizing traffic restrictions during peak smog episodes or deploying mobile air-quality sensors in pollution hotspots. Moreover, the methodology demonstrates transferability to other Indian states with similar data infrastructures.

VII. FUTURE WORK

- Deep Learning Models: Incorporate recurrent neural networks (LSTM, GRU) to capture long-term temporal dependencies and compare them against ensemble approaches.
- Spatial Analysis: Extend the framework to include spatial interpolation methods (kriging, graph-based neural networks) for gap-filling and high-resolution AQI mapping.
- Mobile Application: Develop a real-time mobile platform leveraging model APIs and push notifications to alert citizens of hazardous AQI levels.
- Satellite Data Integration: Fuse satellite-derived aerosol optical depth (Sentinel-5P) and vegetation indices (NDVI) to enrich feature sets with land-cover and vegetation effects.

VIII. CONCLUSION

This study underscores the efficacy of machine learning algorithms—particularly XGBoost—in forecasting AQI in Maharashtra's major cities. Through meticulous preprocessing, feature engineering, and hyperparameter optimization, our models achieved high accuracy and interpretability. The findings emphasize pollutant-specific drivers and meteorological influences, offering actionable insights for policymakers, environmental agencies, and urban planners. By adopting such data-driven approaches, stakeholders can implement timely interventions to mitigate air pollution and safeguard public health.

REFERENCES

[1]. JTawade, J., & Kulkarni, N. (2024). Air Quality Index of Major Indian Cities and Stations [Dataset]. Mendeley Data.
[2]. Sharma, A. (2025). Fine-Grained AQI Forecasting Using Mobile Sensor Data. arXiv:2506.10332.
[3]. Maharashtra Pollution Control Board (MPCB). (2024). Mumbai Emission Inventory & Source Apportionment Report.
[4]. Times of India. (2025). Breathing in Nagpur Nearly Equivalent to Smoking Two Cigarettes a Day.
[5]. Chen, Y., Li, Y., & Zhang, X. (2019). A Deep Learning Model for Air Quality Prediction Based on LSTM. Environmental Science and Pollution Research, 26(4), 3693–3703.
[6]. Zhang, Y., et al. (2012). Real-Time Air Quality Forecasting. Atmospheric Environment, 60, 632–655.