Data Mining (2024) Submitted ; Published

Predicting Air Pollution Levels in Five Major Indian Cities

Aleksandr Jan Smoliakov ALEKSANDR.SMOLIAKOV@MIF.STUD.VU.LT
Danial Yntykbay DANIAL.YNTYKBAY@MIF.STUD.VU.LT
Davide Giuseppe Griffon DAVIDE.GRIFFON@MIF.STUD.VU.LT
Data Science study programme
Faculty of Mathematics and Informatics

Advisor: Jurgita Markevičiūtė

Abstract

Air pollution ranks among the most pressing global health threats, causing an estimated 3 to 9 million deaths annually. India, with its rapid urbanization and economic growth, faces some of the highest pollution levels globally. The country's air pollution stems from a combination of natural and anthropogenic sources, emitting harmful pollutants such as carbon monoxide (CO) and particulate matter (PM_{10} and $PM_{2.5}$).

This project aims to develop a predictive model for air pollution levels using a multivariate regression approach. The model will incorporate a wide array of predictors, including geographic location, seasonal variations, meteorological data and temperature. By evaluating various modeling techniques—starting with linear regression and potentially expanding to more complex methods such as Random Forest—and analyzing data correlations, the study seeks to identify the most effective methods for accurately forecasting urban air pollution levels.

This topic was chosen due to its relevance in addressing the escalating air quality issues in India and the availability of extensive historical weather and air quality data. Previous studies on air quality prediction often focus on pollutants like $PM_{2.5}$ and PM_{10}. Here, we broaden the scope by analyzing seven different pollutants across five major Indian cities—Bengaluru (Bangalore), Delhi, Hyderabad, Jaipur, and Mumbai—assessing their seasonal trends and interdependencies, and examining a wider set of meteorological and temporal variables. The study also enhances interpretability by comparing feature importance across cities, offering a clearer understanding of regional differences.

This study will demonstrate the practical application of data mining techniques using real-world environmental data, serving as a foundation for further exploration in this field.

1. Literature Review

Air pollution has been extensively studied worldwide due to its significant impacts on public health, ecosystems, and the economy. Numerous studies have demonstrated the adverse health effects of air pollutants, including respiratory and cardiovascular diseases, leading to increased morbidity and mortality rates. Particularly, countries like India and China have received considerable attention because of their severe air quality issues, which are exacerbated by rapid industrialization, urbanization, and population growth.

During our literature review, we encountered a vast number of articles related to our research topic. The widespread interest among scientists and data analysts facilitated the collection of numerous sources. As our research progressed, we found that many studies have attempted to develop statistical and artificial intelligence models to predict air pollution levels, utilizing both global datasets and data specific to particular regions. Researchers have employed a variety of techniques, including multivariate regression, neural networks, and machine learning algorithms, to forecast pollutant concentrations based on diverse predictors such as meteorological conditions, emission sources, and socio-economic factors. These models have been applied at various regional scales, providing valuable insights for environmental management and policy-making. Despite this extensive body of research providing a solid foundation for our study, we have not found any existing studies that have examined the five major Indian cities using the same datasets we are using in this project.

To effectively manage and synthesize the extensive body of literature, we have chosen not to include general studies on air pollution in India, as they do not directly contribute to the development of our predictive model. Instead, we have focused our review on two specific categories of research that are more pertinent to our objectives: "Causal and Correlational Studies on Urban Air Pollution" and "Predictive Models". The first category delves into the factors influencing air pollution levels in urban areas, providing valuable insights that inform the selection of variables and the structural framework of our model. The second category encompasses studies that have developed predictive models for air pollution, offering methodologies and approaches that we can build upon to enhance the accuracy and reliability of our own model. By concentrating on these two groups, we aim to leverage existing knowledge effectively and advance our research in a meaningful way.

1.1 Causal and Correlational Studies on Urban Air Pollution

Understanding the dynamics of urban air pollution is essential for developing effective strategies to enhance air quality and protect public health. Several studies have focused on analyzing the correlations and underlying causes of air pollution in metropolitan environments, rather than constructing predictive models.

For instance, [Khedekar, Sneha and Thakare, Sunil, 2023] conducted a six-year analysis in Pune, India, assessing the correlations between pollutants and meteorological factors. The study revealed that most pollutants were positively correlated with each other and with temperature, except for O_3, which had a negative correlation. Wind speed showed a strong negative correlation with pollutant levels, emphasizing its role in pollutant dispersion.

Building on similar themes, [Diya et al., 2024] investigated air pollution across various urban hotspots in Chennai, India. This research assessed hourly concentrations of pollutants such as PM_{10}, $PM_{2.5}$, SO_2, NO_2, and CO across key areas—industrial, traffic, commercial,

and residential zones—over the course of 2022. A key methodological approach employed in this study is the Coefficient of Divergence (COD), which quantifies spatial variations in pollutant concentrations among the different hotspots. One of the significant findings of the Chennai study is the impact of wind on pollution dispersion. When wind speeds are low (0–3 m/s), CO levels tend to be higher, indicating that pollutants are not dispersing effectively and are accumulating near their sources. Conversely, when the wind blows from the south and southeast at moderate speeds (2–6 m/s), the concentrations of $PM_{2.5}$ and PM_{10} increase. This suggests that pollutants from nearby industries are being transported toward the monitoring stations, highlighting the crucial role of meteorological conditions in air quality.

In another study, [Suthar et al., 2024] aimed to identify seasonal patterns and understand how meteorological factors influence pollutant levels in a different Indian city. The research included a correlation analysis between air pollutants and meteorological parameters—wind speed (WS), wind direction (WD), relative humidity (RH), and solar radiation (SR). Over three consecutive years, the analysis revealed that WD, WS, and RH generally had a negative correlation with all measured air pollutants. Calm wind conditions inhibit the dispersion of pollutants, resulting in higher concentrations near the ground, underscoring the importance of WS and WD in the dispersion and transport of air pollutants.

Expanding this line of research to European cities, [Rowland, 2024] examined the relationship between meteorological parameters and the concentrations of NO_2, O_3, PM_{10}, and $PM_{2.5}$ in Krakow, Paris, and Milan during 2021. The study found that NO_2, PM_{10}, and $PM_{2.5}$ concentrations were higher during winter and lower during summer, exhibiting negative correlations with temperature, while O_3 showed the opposite trend. Wind speed was inversely related to particulate matter and NO_2 levels but positively correlated with O_3 concentrations. These findings highlight the influence of meteorological conditions on pollutant levels and the occurrence of the "Ozone weekend effect" in these cities.

Speaking of temporal trends, [Bozhkova et al., 2020] conducted research in urban areas of Belarus. Seasonal patterns revealed higher pollution in autumn and winter, with increased dispersion of pollutants and ozone formation in spring and summer. The study observed daily pollution peaks occurring in the morning and evening, driven by human activities and affected by wind and atmospheric stability. The reduced dispersion efficiency during these periods, combined with higher emission intensities, contributes to these peaks.

These studies collectively underscore the significant impact of meteorological and temporal factors on urban air pollution across diverse geographic regions. The consistent observations of pollutant behavior in relation to temperature, wind speed, and other meteorological parameters highlight the necessity of incorporating environmental conditions into air quality management and policy-making.

1.2 Predictive Models

Predictive modeling plays a crucial role in understanding and forecasting air pollution levels, which is essential for public health planning and environmental management. Various studies have employed different statistical and machine learning approaches to predict concentrations of air pollutants and the Air Quality Index (AQI), a standardized measure that indicates the overall air quality and its potential impact on human health.

Singh et al. [Singh et al., 2012] investigated both linear and nonlinear methods for forecasting urban air quality, aiming to improve prediction accuracy in complex urban environments. The study examined the effectiveness of different modeling approaches for predicting concentrations of common urban pollutants such as PM_{10}, NO, CO, and O_3. Specifically, they applied linear models like multiple linear regression and nonlinear models including Artificial Neural Networks (ANNs) to compare their performance in capturing pollution patterns. The findings indicated that nonlinear models, particularly ANNs, provided better prediction accuracy than linear models, highlighting the importance of nonlinear approaches in modeling air pollution in urban settings.

Sanjeev et al. [Sanjeev, 2021] developed predictive models for air quality using machine learning algorithms, focusing on Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forests (RF). Their study aimed to identify the most efficient algorithm for air quality prediction. The models were evaluated based on accuracy scores, with the RF-based model achieving the highest accuracy.

Kothandaraman et al. [Kothandaraman et al., 2022] focused on predicting PM2.5 pollutant levels by employing a variety of machine learning algorithms, including linear regression, Random Forest, K-Nearest Neighbors, Ridge and Lasso regression, XGBoost, and AdaBoost. Their study utilized historical PM2.5 data and relevant meteorological features such as temperature, humidity, wind speed, and precipitation collected from monitoring stations in Anand Vihar, Delhi, over the period from January 2014 to December 2019. By evaluating the performance of these models through statistical metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2), they found that ensemble methods such as XGBoost and Random Forest outperformed other algorithms in terms of predictive accuracy. These results highlight the effectiveness of advanced machine learning techniques in modeling air pollution and the critical role of incorporating meteorological data.

Kumar et al. [Kumar and Pande, 2023] addressed the challenge of predicting the AQI by analyzing air pollution data from 23 Indian cities over six years. They carried out extensive data preprocessing, which involved handling missing values, correcting outliers, normalizing data, selecting features, and applying logarithmic transformations to fix skewed data. Their exploratory data analysis showed a significant decrease in pollution levels in 2020, likely due to COVID-19 lockdowns. To fix data imbalance, they used the Synthetic Minority Over-sampling Technique (SMOTE). They performed machine learning-based AQI predictions using various models, both with and without SMOTE resampling, and compared the results. The models were assessed using standard metrics like accuracy, precision, recall, F1-score, and error metrics (MAE, RMSE, RMSLE, R^2). The XGBoost model performed the best, achieving the highest accuracy in both training and testing phases, while the SVM model had the lowest accuracy. The Random Forest model also did well, especially when SMOTE was applied. The study emphasizes the effectiveness of ensemble learning methods in AQI prediction and suggests that future research could explore deep learning techniques to improve accuracy further.

Roy et al. [Roy et al., 2024] conducted a study in the densely populated northern Indian states of Delhi, Haryana, and Uttar Pradesh, analyzing PM2.5 concentrations in relation to meteorological factors such as temperature, precipitation, surface pressure, and wind. They employed Ordinary Least Squares (OLS) regression and Geographically Weighted

Regression (GWR) to explore the relationships between PM2.5 levels and environmental parameters across different seasons and locations. The OLS model identified significant predictors with R^2 values of 0.93 for summer and 0.94 for winter, while GWR accounted for spatial variability, enhancing model performance and highlighting the importance of geographical factors in air pollution modeling. However, our study does not utilize geographical data and cannot replicate the GWR analysis by Roy et al. Instead, we focus on the overall relationships between PM2.5 concentrations and meteorological factors without considering spatial variability.

Building on the significance of feature engineering and advanced modeling techniques, [Naz et al., 2024] emphasized the crucial role of feature engineering in time series prediction of air pollutants. They introduced a two-stage feature engineering and selection process that combines correlation-based selection with Variational Mode Decomposition (VMD). By developing and categorizing 22 new features into meteorological, temporal, statistical, and air pollutant types, their approach customizes optimal feature sets for each of the five major air pollutants. This customization enhances model performance by 1–5% compared to traditional lag-based methods and further improves accuracy by 3–13% when integrating VMD features. The optimized feature selection allows for simpler forecasting models with significant improvements in RMSE, MAE, and R^2 scores.

These studies demonstrate the effectiveness of various machine learning and statistical methods in predicting air pollution levels and AQI. Nonlinear models and ensemble learning techniques like Random Forest and XGBoost have shown high accuracy in forecasting pollutant concentrations and AQI. The incorporation of meteorological and temporal data significantly enhances model performance. For our study, which does not employ deep learning methods, these findings suggest that ensemble methods and regression techniques—especially those accounting for spatial variability—can serve as effective alternatives for accurate air quality prediction. Incorporating meteorological factors and addressing data imbalances may further improve prediction accuracy without the need for deep learning models.

2. Exploratory Data Analysis

2.1 Data Source

This project used two primary sources of data to analyze air pollution levels across five major Indian cities:

- **Air Quality Data in India**: Available at Kaggle. This dataset provides hourly measurements of various air pollutants and particulate matter. The data is collected from multiple weather stations located in 27 major Indian cities. The date range spans from January 2015 to July 2020, with over 700,000 records.
- **Historical Weather Data for Indian Cities**: Available at Kaggle. This dataset includes hourly weather-related features across 8 Indian cities, encompassing over 20 variables such as precipitation (mm), wind speed, temperature, humidity, and other meteorological parameters. The date range of the dataset spans from 2006 to 2019, with over 700,000 records. [Soneji, 2020]

To extract meaningful insights, it was necessary to merge these datasets, linking the meteorological variables from the second dataset with the air quality measurements from the first. This integration allowed us to analyze how environmental factors such as temperature, wind speed, and precipitation interact with pollutant concentrations over time. The merging process focused on aligning data by city names and hourly timestamps, ensuring that weather features corresponded accurately to pollution levels. This alignment was essential for preserving the temporal and spatial consistency of our analysis. While the original datasets contained information for additional cities and years, we focused our analysis on the five metropolitan areas and the five-year period (January 2015 to January 2019), where the two datasets overlapped.

Looking ahead to our modeling approach, we envisioned using the rich meteorological data, combined with temporal features (such as hour of day and month of year) and the city names as predictors for air quality data.

Note that we will not be forecasting time series data in this project, as we are limited in the choice of models and may not use advanced techniques like LSTM or other deep learning models. Instead, our primary goal is to predict air pollution levels using a multivariate regression approach. This will allow us to explore how various environmental and temporal variables contribute to air pollution patterns and potentially forecast air quality from meteorological data.

In the following subsections, we first examine our response variables, analyzing the various pollutants and their distributions in our dataset. We then explore the weather features, detailing how we identified the most relevant predictors from our merged dataset. Finally, we present our feature engineering methodology, which helped us capture complex temporal and spatial patterns in our data, setting the stage for our subsequent modeling efforts.

2.2 Response Variable Selection

The pollution dataset contained twelve potential response variables measuring different air quality parameters: $PM_{2.5}$, PM_{10}, NO, NO_2, NOx, NH_3, CO, SO_2, O_3, Benzene, Toluene,

and Xylene. While we initially considered calculating and using the Air Quality Index (AQI) as a comprehensive response variable, this approach proved impractical due to the dataset's structure. The AQI calculation requires complete data across multiple pollutants, and the presence of missing values in our dataset would have significantly limited our ability to compute this index accurately.

After careful consideration of data completeness and relevance, we narrowed our focus to seven key pollutants that demonstrated the most complete records: $PM_{2.5}$, PM_{10}, NOx, NH_3, CO, SO_2, and O_3. The analysis of missing data revealed varying degrees of incompleteness across these pollutants, with CO showing the lowest proportion at 4% missing values, followed by NOx at 9%, SO_2 at 19%, $PM_{2.5}$ and O_3 both at approximately 20%, PM_{10} at 26%, and NH_3 showing the highest proportion at 30% missing values. Further analysis can be viewed in section 2.3.1.

To maintain data integrity and avoid potential biases, we chose to handle missing values through complete case deletion rather than imputation. This decision was supported by the large volume of available data in our merged dataset, which allowed us to maintain a robust sample size even after removing incomplete records. We applied this deletion approach consistently across both air quality parameters and meteorological features to ensure data consistency.

2.2.1 Distribution Analysis and Transformation

Initial exploratory analysis revealed that the distributions of these pollutants exhibited strongly right-skewed patterns across all five cities. Figure 1 illustrates this pattern for three representative pollutants, but the same asymmetric distribution was observed across all pollutants, characterized by a high frequency of lower values and a long tail extending toward higher concentrations.

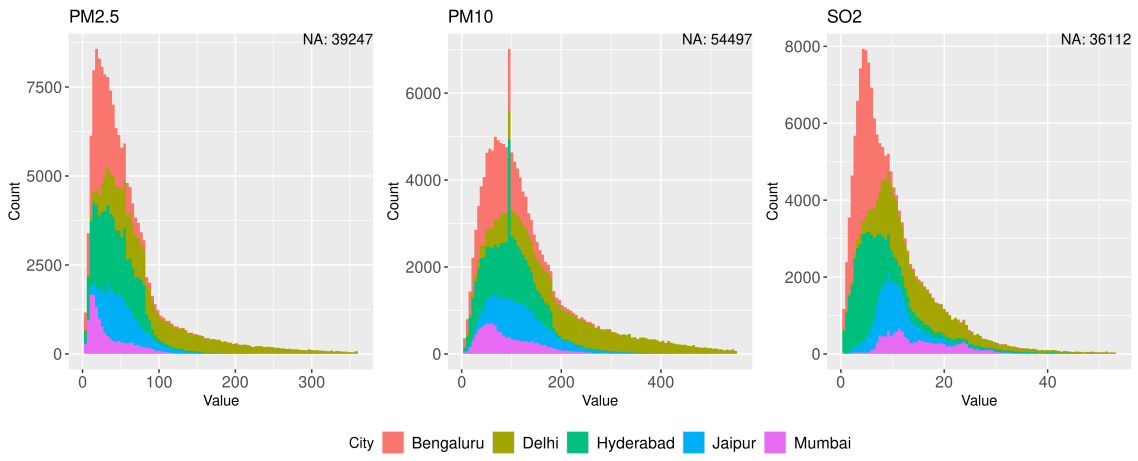

Figure 1: Pollutants levels by city

While the middle plot for PM10 shows a spike at 100 ppm, it represents a small fraction of the data and could still be correct (but rounded) data. We didn't pursue a further investigation and decided to leave these datapoints in the dataset.

To address this skewness and make the data more suitable for statistical analysis, we applied a logarithmic transformation to all pollutant measurements. Since logarithmic transformation requires strictly positive values, we added a constant of 1 to each measurement before applying the transformation. The effectiveness of this transformation is demonstrated in Figure 2, where the transformed distributions more closely approximate normal distributions, providing a more suitable basis for our subsequent analyses.

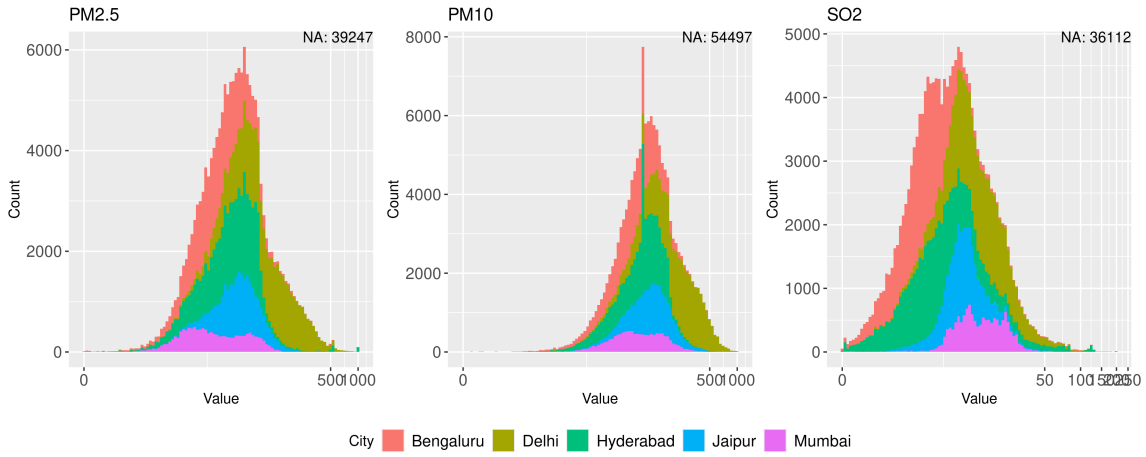

Figure 2: Pollutants levels by city on a logarithmic scale

2.2.2 Temporal Pattern Analysis

Further investigation of temporal patterns revealed distinct seasonal trends in pollutant concentrations, particularly for $PM_{2.5}$ and PM_{10}, as shown in Figure 3. Notable patterns can be observed for $PM_{2.5}$ and PM_{10}, where consistently lower concentrations were observed across all cities. Among the cities studied, Delhi consistently exhibited higher pollutant concentrations throughout the year. These temporal patterns suggest the importance of incorporating seasonal factors into our modeling approach. While there are clear seasonal trends in the data, they do not always follow a simple sinusoidal pattern, and the minimums are observed at different times of the year for different pollutants.

All this complexity suggests that different cities should be modeled separately, and the seasonal trends may require more sophisticated modeling techniques to capture accurately.

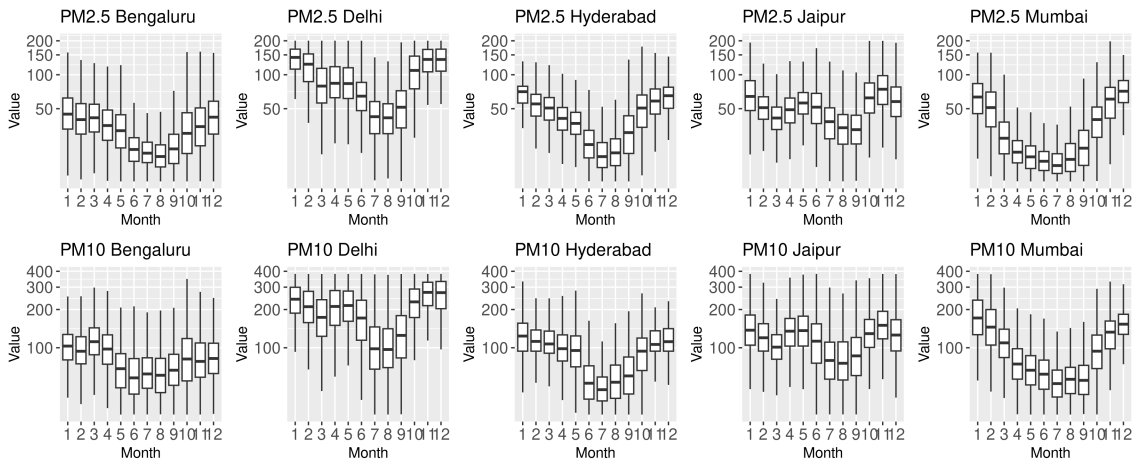

Figure 3: Monthly $PM_{2.5}$ and PM_{10} distributions by city

2.3 Predictor Selection

Having completed our analysis of the pollutant response variables and their distributions, we turned our attention to examining the meteorological predictors in our dataset. Our weather data comprises 25 distinct variables that collectively capture a comprehensive range of atmospheric conditions and celestial phenomena.

The dataset includes several categories of measurements. The temporal information category encompasses timestamps and celestial events, including *date_time* for precise temporal tracking, along with daily solar cycles (*sunrise*, *sunset*) and lunar phenomena (*moonrise*, *moonset*, *moon_illumination*). For temperature characterization there are several metrics: the basic temperature readings (*tempC*, *maxtempC*, *mintempC*), perceived temperature indicators (*FeelsLikeC*, *HeatIndexC*, *WindChillC*), and the *DewPointC* measurement that indicates atmospheric moisture saturation points.

The atmospheric conditions are documented through several key parameters: *humidity* measuring atmospheric moisture content, *pressure* indicating atmospheric force per unit area, *cloudcover* quantifying sky obstruction, precipitation measurements (*precipMM*, *totalSnow_cm*), and *visibility* recording atmospheric clarity. Wind conditions are characterized by three primary measurements: *windspeedKmph* for current wind velocity, *WindGustKmph* for peak wind speeds, and *winddirDegree* indicating wind direction. Additionally, solar radiation intensity is captured through *sunHour* measurements and *uvIndex* readings, providing information about solar exposure levels.

Before applying feature selection, we conducted a correlation analysis of the meteorological data to understand the relationships between weather variables. The correlation matrix visualization (Figure 4) revealed several significant relationships between weather parameters. Some of the key findings were as follows:

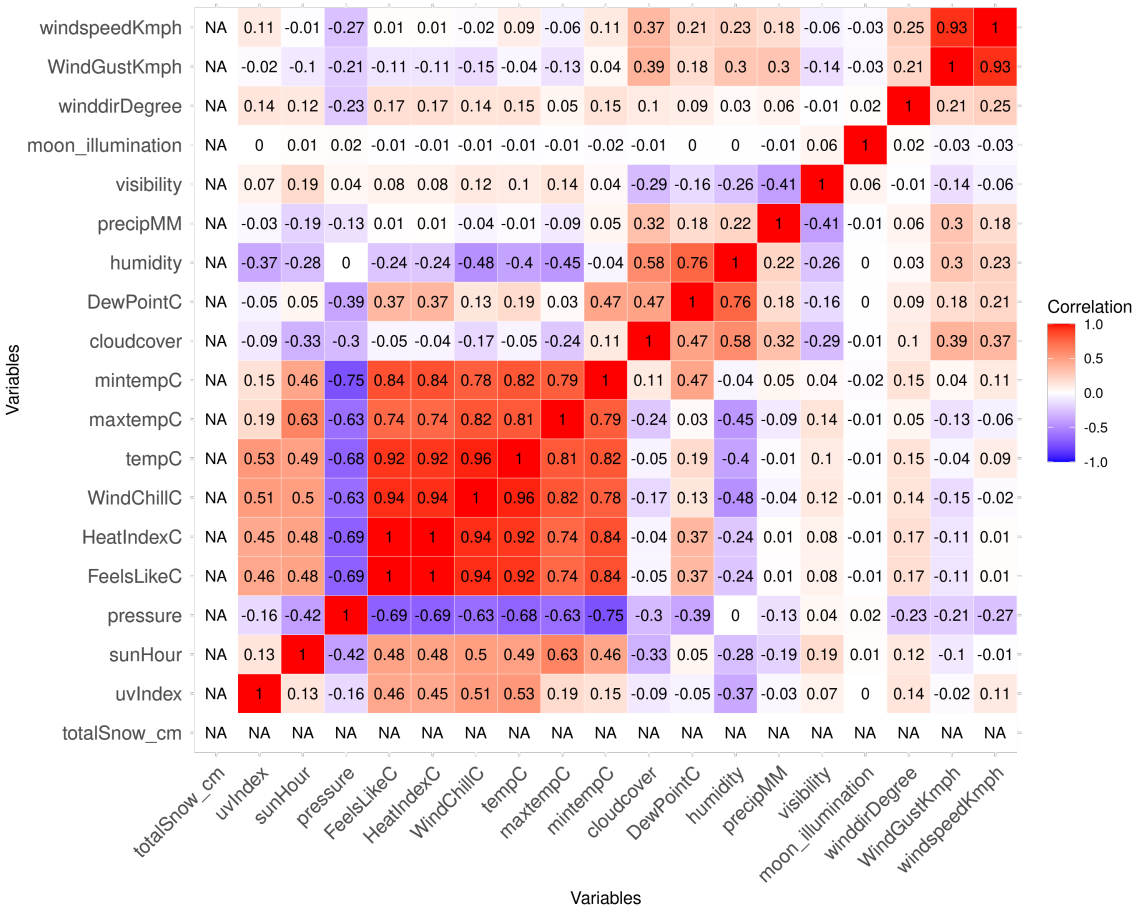

Figure 4: Correlation matrix of weather variables.

- *FeelsLikeC*, *HeatIndexC*, *WindChillC*, *minTempC*, and *maxTempC* were very strongly correlated with *tempC*, as expected, given that they are derived from temperature, humidity, and wind speed.
- Features like *sunHour* and *uvIndex* were also correlated with temperature-related variables.
- *Humidity* was positively correlated with *DewPointC* and *cloudcover*, while inversely correlated with temperature-related variables.
- *Pressure* was inversely correlated with temperature-related variables and showed slight correlations with several other features, suggesting that pressure is influenced by other meteorological factors.
- *Visibility* exhibited slight positive correlations with temperature-related variables and inverse correlations with *humidity*, *cloudcover*, *precipitation*, and *DewPointC*.
- *Windspeed* and *windgust* were strongly correlated with each other.

- *Moon_illumination* was not correlated with any other variable, as expected.

Based on these findings, we identified several highly correlated group of features, such as *tempC* and *FeelsLikeC*, *HeatIndexC*, and *WindChillC*, which could potentially introduce collinearity issues in our models. To address this, we have removed the highly correlated features, retaining only one representative variable from each correlated group. This feature selection process aimed to reduce redundancy and improve model interpretability by focusing on the most relevant predictors. The following predictors were removed from our dataset:

- *totalSnow_cm* was removed due to zero variance.
- *FeelsLikeC*, *HeatIndexC*, and *WindChillC* were removed in favor of *tempC*.
- *WindGustKmph* was removed in favor of *windspeedKmph*.
- *Visibility* was removed due to it being a possible leakage feature.

We did not remove any other correlated features at this stage, as they may still provide valuable information for our models. In the subsection 3.1 we will perform a multicollinearity test and decide to remove three additional features.

Following this, we analyzed the correlation between the selected features (those obtained after removing the highly correlated ones identified in the previous analysis) and the response variables to assess their potential predictive power. The resulting correlation analysis is visualized in Figure 5.

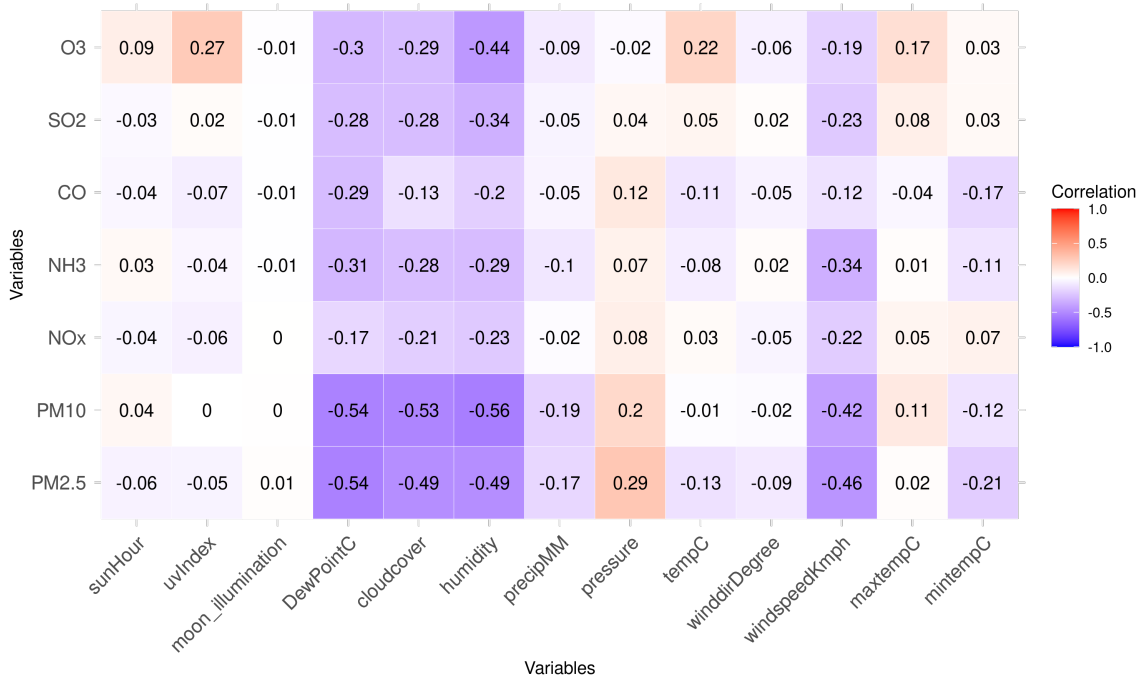

Figure 5: Correlation between weather features and response variables.

The analysis revealed the following:

- *Humidity*, *windspeed*, *cloudcover*, and *DewPointC* were negatively correlated with all but the CO pollutant, with the strongest correlations observed with particulate matter (*PM10* and *PM2.5*).
- *O3* was also slightly positively correlated with *uvIndex* and *tempC*.
- *CO* showed no notable correlation with any of the selected weather features, except a negative correlation with *DewPointC*.
- *MoonIllumination* was not correlated with any variable, as expected.

However, since this analysis relied solely on linear correlation measures, it is possible that important non-linear relationships remain undetected. To capture such relationships, several feature transformations were performed in subsequent steps.

2.3.1 Missing Value Analysis

In our analysis of the missing values, we found significant proportions of missing values for several pollutants: $PM_{2.5}$ (29.3%), O_3 (20.1%), PM_{10} (27.6%) and NH_3 (29.3%). The missing data for these variables is not uniformly distributed but appears to be geographically concentrated, with Mumbai accounting for a significant portion of the missing values, suggesting that the missingness is not random but rather geographically determined.

Table 1 shows the detailed breakdown of missing values for Mumbai.

	Pollutant	Missing values	Missing (%)
1	NH_3	38,750	88.4
2	PM_{10}	30,797	70.3
3	SO_2	30,024	68.5
4	$PM_{2.5}$	30,008	68.5
5	O_3	29,927	68.3
6	NO_x	13,139	30.0
7	CO	1,888	4.3

Table 1: Missing values analysis for each pollutant in Mumbai.

A significant proportion of missing data for several variables is concentrated in Mumbai. Specifically:

These percentages indicate that a large portion of the missing data for these key environmental variables is associated with Mumbai. This suggests that the missing data are not missing at random and may be influenced by factors specific to the Mumbai region, such as data collection or reporting challenges. Given the concentration of missing data in this specific city, we can conclude that ignoring these missing values will not significantly affect the general analysis of the dataset.

Pollutant	Missing Values from Mumbai
NH_3	67.0
PM_{10}	56.5
O_3	76.0
$PM_{2.5}$	77.0
SO_2	82.9
NO_x	71.0
CO	24.1

Table 2: Percentage of total missing values from Mumbai by pollutant

2.4 Feature Engineering

Our feature engineering process focused on creating meaningful derived variables that could capture complex temporal patterns and weather-related phenomena not directly represented in the raw data. These engineered features proved to be effective in improving our model's predictive performance.

2.4.1 Temporal Features

To capture the cyclical nature of air pollution patterns, we developed several temporal features from the datetime information. First, we created a column for the hour of the day (*hour_of_day*, ranging from 0 to 23). To properly represent the circular nature of time, we then transformed this hour variable into two continuous features using trigonometric functions: $hour\_cos = \cos(2\pi \times \text{hour}/24)$ and $hour\_sin = \sin(2\pi \times \text{hour}/24)$. This cyclic encoding ensures that hour 23 is correctly represented as being close to hour 0. We applied analogous transformations with a 12-hour period to create *hour_cos_12h* and *hour_sin_12h* features to capture any possible bi-daily patterns. Finally, we created categorical variables for the day of the week (*day_of_week*, ranging from 1 to 7) and month of the year (*month_of_year*, ranging from 1 to 12) to capture weekly and seasonal patterns in pollution levels. These predictors were made categorical to account for the complex seasonal patterns observed in the data.

2.4.2 Weather-Related Features

For weather data, we implemented several transformations. Wind speed and direction, originally measured in *Kmph* and degrees (0-359), respectively, were transformed into two new wind speed predictors, namely the wind speed along the X and Y axes. This transformation was achieved using the following trigonometric functions:

$$windspeed\_x = windspeedKmph \times \cos(2\pi \times \text{winddir}/360)$$
$$windspeed\_y = windspeedKmph \times \sin(2\pi \times \text{winddir}/360)$$

This encoding ensured that directions like 359° and 0° were appropriately represented as being nearly identical.

A key feature that emerged from our literature review was the significant impact of precipitation on pollution levels during rainy seasons. To capture this effect, we developed a cumulative precipitation feature (*precipMM_24h*) that calculates the total rainfall over the previous 24 hours for each city. This feature proved to be effective in our final model, helping to capture how sustained rainfall patterns affect pollutant concentrations. Similarly, we created a 24-hour cumulative wind speed feature (*windspeedKmph_24h*) to account for the sustained impact of wind on pollutant dispersion.

Additionally, we normalized sunrise times (*sunrise_normalized*) to account for seasonal variations in daylight patterns, scaling the values between 0 and 1 to allow for better comparability across different seasons and cities. This normalization helped capture the influence of varying daylight patterns on pollution accumulation and dispersion.

These engineered features, particularly the temporal cyclic encodings and the 24-hour cumulative precipitation metric, significantly enhanced our model's ability to capture complex patterns in air pollution levels. The effectiveness of these features suggests that both the cyclic nature of pollution patterns and the sustained impact of weather conditions play crucial roles in determining air quality levels.

Once feature engineering was complete, we scaled all continuous features to ensure that they were on a similar scale, preventing any single feature from dominating the model. This step was essential for maintaining the stability and convergence of our regression models.

Finally, we performed the feature correlation analysis once again to make sure that we have not introduced any new collinearity issues. The results of this analysis are visualized in Figure 6.

The *hour_cos* looks highly negatively correlated with *uvIndex*, which is expected, as the UV index is higher during the day. We decided to keep both features as they represent different types of data - one is temporal, the other is meteorological.

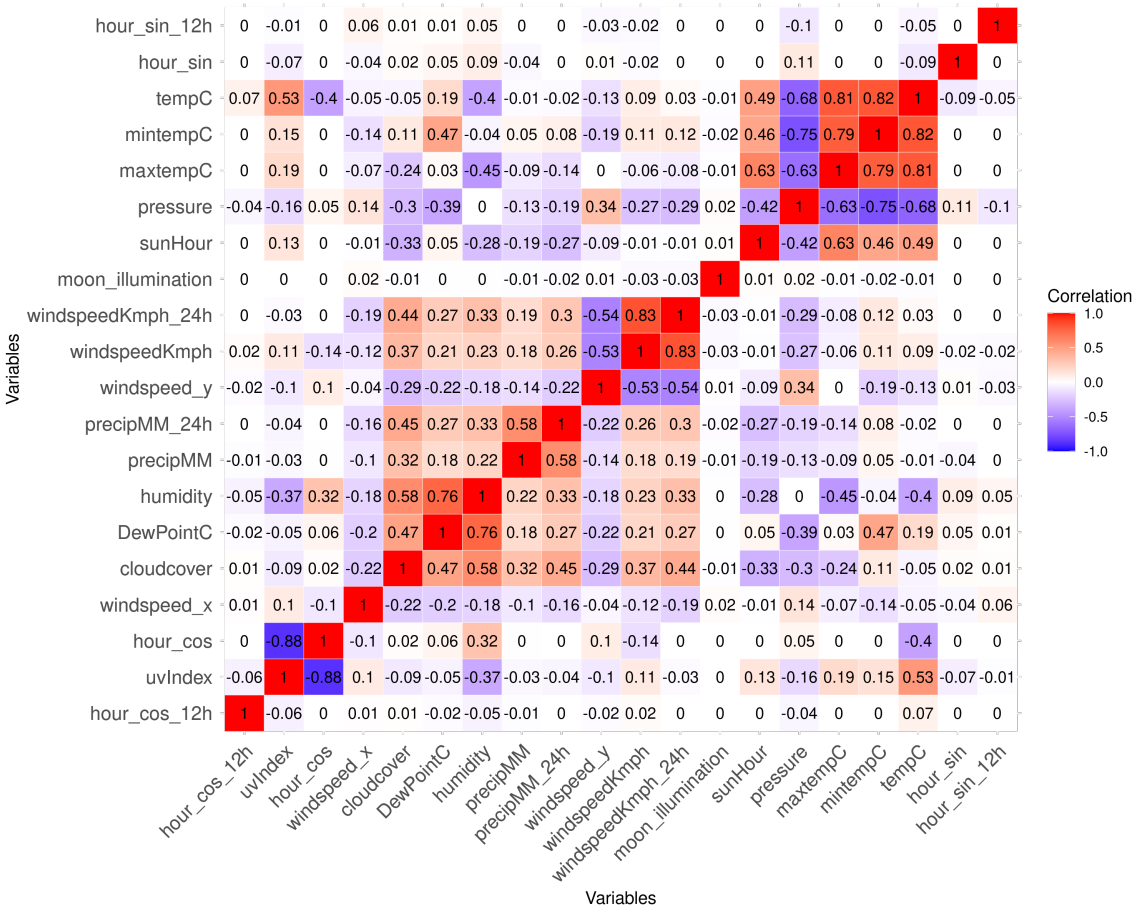

Figure 6: Correlation between final features.

3. Methodology

The above exploratory data analysis showed that different cities exhibit unique temporal and environmental patterns that influence air pollution levels. To capture these dynamics effectively, we chose to fit independent linear regression models for each city and response variable.

3.1 Feature Selection

Initially, the analysis incorporated a set of 7 response variables and 22 features, chosen for their potential relevance to air pollution levels. The set of predictors included:

- **Meteorological Features**: *tempC*, *minTempC*, *maxTempC*, *DewPointC*, *humidity*, *cloudcover*, *precipMM*, *pressure*, *windspeedKmph*, *uvIndex*, *moon_illumination*, *sunHour*

- **Derived Meteorological Features**: *windspeed_x*, *windspeed_y*, *precipMM_24h*, *windspeedKmph_24h*
- **Temporal Features**: *day_of_week*, *month_of_year*, *hour_cos*, *hour_sin*, *hour_cos_12h*, *hour_sin_12h*

During the modeling process, we performed Variance Inflation Factor (VIF) analysis to identify and remove any predictors that introduced multicollinearity issues. There were three features that consistently exhibited VIF > 4 across all models: *minTempC*, *maxTempC*, and *DewPointC*. To address this, we removed these features from our dataset, ensuring that our models were not affected by multicollinearity. Following the removal of these variables, multicollinearity analysis no longer identified any variables with high VIF values.

3.2 Data Splitting and Filtering

For each city in the dataset, the data were split into training (2015-2018) and testing (2019) sets to ensure temporal independence in model evaluation. Records with missing or incomplete data were omitted, and cities or response variables were excluded if:

- The response variable had zero values in the training period.
- The training data lacked representation for all months of the year, potentially skewing temporal patterns.

4. Model Fitting

Linear regression models were fitted independently for each of the seven response variables in each city. The general model specification was:

$$Y_C = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \epsilon$$

where Y_C is the response variable for city C, X_i are the predictor variables, β_i are the estimated coefficients, and ϵ is the residual error. This approach allowed for individualized models that could account for city-specific environmental dynamics.

4.1 Feature Importance Analysis

For each fitted model, feature importance was quantified using the estimated regression coefficients. Due to scaling done during feature engineering, the coefficients were normalized and assessing the relative contribution of each feature to the model was possible.

4.2 Performance Evaluation

Model performance was assessed using three primary metrics on the test dataset:

- Pearson Correlation (r) between actual and predicted values.

- Coefficient of Determination (R^2) as a measure of variance explained by the model.
- Root Mean Square Error (RMSE) to quantify the average prediction error.

Performance metrics were aggregated by response variable and city to assess general trends.

5. Results

5.1 Data Characteristics and Filtering Outcomes

Out of the initial dataset, several city-response combinations were excluded due to missing data or incomplete temporal coverage:

- Mumbai: $PM_{2.5}$, PM_{10}, NH_3, SO_2, O_3.

Otherwise, the retained datasets included robust representations across all months and had sufficient training records for model fitting.

5.2 Model Fit and Performance Metrics

The test R^2 across all 30 models ranged from 0.012 to 0.670 (median: 0.289), indicating varying levels of explanatory power. The test r values ranged from 0.108 to 0.819 (median: 0.538), suggesting varying (low to moderate) predictive accuracy across cities and response variables.

The aggregated performance metrics for each response variable are presented in Table 3.

	Response variable	r	R^2	RMSE
1	CO	0.311	0.132	0.307
2	NH3	0.338	0.161	0.420
3	NOx	0.430	0.214	0.604
4	O3	0.660	0.443	0.576
5	PM10	0.694	0.492	0.435
6	PM2.5	0.728	0.538	0.437
7	SO2	0.273	0.102	0.388

Table 3: Aggregated performance metrics by response variable.

The metrics suggest that the linear model's performance varied significantly across pollutants, with $PM_{2.5}$, PM_{10} and O_3 showing the highest predictive accuracy and average R^2 values of 0.538, 0.492 and 0.443, respectively, thus explaining roughly half of the variance in the test data. The models for CO and SO_2 had the lowest performance, with average R^2 values of 0.132 and 0.102, respectively.

The aggregated performance metrics for each city are presented in Table 4.

The city-specific results indicate that the models performed best in Delhi and Hyderabad, with average R^2 values of 0.403 and 0.390, respectively. Mumbai had the lowest performance, with an average R^2 of 0.007, although only two out of seven response variables were modelled for this city. All in all, the models demonstrated varying levels of predictive

	City	r	R^2	RMSE
1	Bengaluru	0.427	0.246	0.481
2	Delhi	0.615	0.403	0.412
3	Hyderabad	0.598	0.390	0.390
4	Jaipur	0.411	0.198	0.472
5	Mumbai	0.061	0.007	0.649

Table 4: Aggregated performance metrics by city.

accuracy across cities, reflecting the diverse environmental dynamics and pollutant patterns in each region.

The full raw results for each individual model are presented in Appendix A.

5.3 Key Predictors

The full feature importance heatmap is presented in Appendix B. The analysis revealed consistent predictors across multiple cities:

- **Temporal variables**: Month of the year commonly emerged as the most significant predictor, with the months of March to September or October showing the highest negative coefficients across many response variables. The hour of the day features also showed some importance.
- **Meteorological variables**: Humidity was consistently a negative predictor across multiple pollutants, with negative coefficients indicating an inverse relationship with pollutant levels. Wind speed and UV index also showed importance for some pollutants.
- **Derived features**: *precipMM_24* demonstrated negative associations with PM_{10} and $PM_{2.5}$, while showing positive associations with NH_3 and NO_x. The *hour_cos* feature exhibited positive correlations with PM_{10}, $PM_{2.5}$, and NO_x, while showing a negative correlation with O_3.

6. Conclusion

This study developed and evaluated independent linear regression models to predict air pollution levels across five major Indian cities, analyzing seven key pollutants using meteorological and temporal predictors. Our research yielded several important findings:

First, the predictive performance varied significantly across both pollutants and cities. The models showed strongest performance in predicting $PM_{2.5}$ ($R^2 = 0.538$), PM_{10} ($R^2 = 0.492$), and O_3 ($R^2 = 0.443$) levels, while struggling with CO ($R^2 = 0.132$) and SO_2 ($R^2 = 0.102$). Geographically, the models performed best in Delhi ($R^2 = 0.403$) and Hyderabad ($R^2 = 0.390$), suggesting that pollution patterns in these cities may be more strongly influenced by the meteorological and temporal factors included in our analysis. The notably poor performance in Mumbai ($R^2 = 0.007$) could potentially be attributed to the extensive missing data in this city's dataset, which led to the exclusion of five out of seven pollutants from the analysis.

Second, our feature importance analysis revealed that temporal patterns, particularly monthly variations, were consistently among the strongest predictors of pollution levels. This finding highlights the significance of seasonal effects on air quality in Indian cities. Naturally, the *month_of_year* variable also captures a good part of the yearly climate trends. Among meteorological variables, humidity emerged as a particularly important predictor, showing consistent negative correlations with multiple pollutants across different cities.

Several limitations of our study should be noted. The models' performance varied considerably across different pollutants and cities, with some combinations showing relatively weak predictive power. Additionally, our analysis was limited to a five-year period and five cities, which may not capture the full range of pollution patterns across India's diverse urban landscapes.

Additionally, the linear regression models used in this study may not fully capture the complex non-linear relationships between air pollution and meteorological factors. More advanced modeling techniques, such as generalized additive models or machine learning algorithms, could potentially provide more accurate predictions by accounting for non-linear interactions and complex temporal patterns.

Despite these limitations, our findings demonstrate that air pollution levels in Indian cities can be predicted with moderate accuracy using readily available meteorological and temporal data. The reliable prediction of $PM_{2.5}$ and PM_{10} levels is particularly significant given their well-documented health impacts.

7. Future Work

Our study suggests several promising directions for future research. First, while our linear models showed reasonable performance, exploring more sophisticated approaches such as generalized additive models (GAMs) could better capture non-linear relationships in the data. These models could be particularly valuable for pollutants like CO and SO_2 where our current approach showed limited predictive power.

Finally, the incorporation of additional predictors could enhance model performance. Satellite data could provide insights into regional air masses and pollution transport, while traffic data and industrial activity metrics could help account for anthropogenic emission sources. Additionally, data about special events (such as festivals or policy changes) could improve predictions during unusual pollution episodes.

8. Source code

The EDA and model training code can be found in our Github repository github.com/Griffosx/data-mining-project. The training code is in an R notebook located at notebooks/EDA_air_quality.Rmd.

Appendix A: Model Performance Metrics

	City	Response variable	r	R^2	RMSE
1	Bengaluru	PM2.5	0.709	0.503	0.443
2	Bengaluru	PM10	0.662	0.438	0.400
3	Bengaluru	NOx	0.416	0.173	0.911
4	Bengaluru	NH3	0.075	0.006	0.392
5	Bengaluru	CO	0.309	0.095	0.192
6	Bengaluru	SO2	0.117	0.014	0.308
7	Bengaluru	O3	0.702	0.493	0.718
8	Delhi	PM2.5	0.817	0.667	0.484
9	Delhi	PM10	0.750	0.562	0.451
10	Delhi	NOx	0.614	0.377	0.472
11	Delhi	NH3	0.671	0.451	0.258
12	Delhi	CO	0.319	0.102	0.337
13	Delhi	SO2	0.465	0.216	0.344
14	Delhi	O3	0.665	0.443	0.535
15	Hyderabad	PM2.5	0.796	0.634	0.391
16	Hyderabad	PM10	0.820	0.672	0.421
17	Hyderabad	NOx	0.472	0.223	0.497
18	Hyderabad	NH3	0.352	0.124	0.381
19	Hyderabad	CO	0.598	0.358	0.132
20	Hyderabad	SO2	0.406	0.165	0.508
21	Hyderabad	O3	0.744	0.554	0.399
22	Jaipur	PM2.5	0.590	0.348	0.429
23	Jaipur	PM10	0.545	0.298	0.470
24	Jaipur	NOx	0.532	0.284	0.440
25	Jaipur	NH3	0.255	0.065	0.648
26	Jaipur	CO	0.321	0.103	0.273
27	Jaipur	SO2	0.105	0.011	0.394
28	Jaipur	O3	0.529	0.280	0.650
29	Mumbai	NOx	0.116	0.014	0.698
30	Mumbai	CO	0.006	0.000	0.599

Appendix B: Feature Importance Heatmap

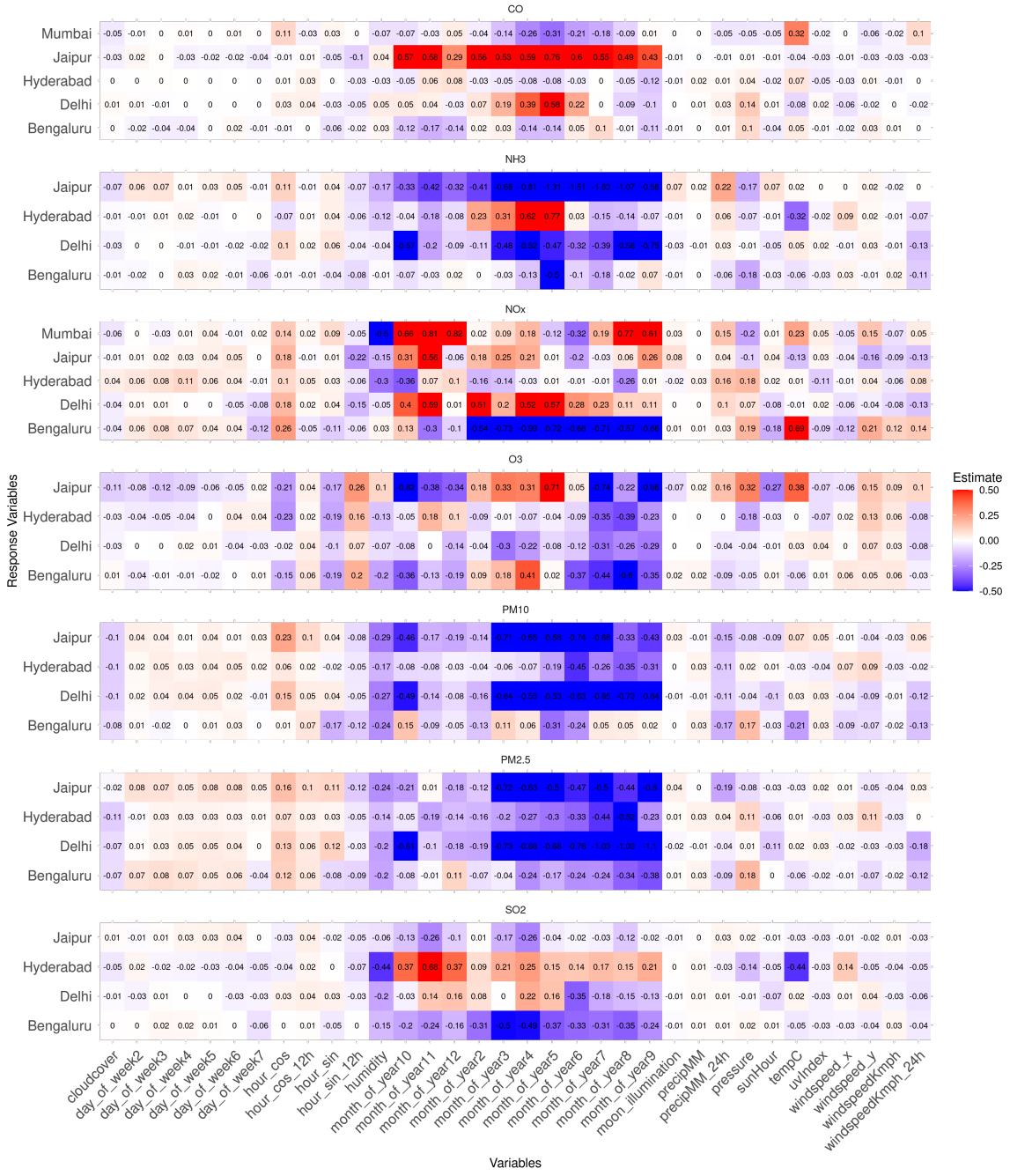

Figure 7: Feature importance heatmap.

References

Victoria V. Bozhkova, Alexander M. Liudchik, and Siarhey D. Umreika. Influence of meteorological conditions on urban air pollution. *Acta Geographica Silesiana*, 14(4):5–21, 2020. ISSN 1897-5100. URL https://ags.wnp.us.edu.pl/download/wydawnictwa/ags/tom_40_2.pdf.

M. Diya, Sudheer Kumar Kuppili, and S. M. Shiva Nagendra. Air quality in different urban hotspots in a metropolitan city in india and the environmental implication. *Environmental Monitoring and Assessment*, 196:1102, 2024. doi: 10.1007/s10661-024-13272-z. URL https://doi.org/10.1007/s10661-024-13272-z. Received: 28 May 2024 / Accepted: 16 October 2024.

Khedekar, Sneha and Thakare, Sunil. Correlation analysis of atmospheric pollutants and meteorological factors using statistical tools in pune, maharashtra. *E3S Web Conf.*, 391: 01190, 2023. doi: 10.1051/e3sconf/202339101190. URL https://doi.org/10.1051/e3sconf/202339101190.

D. Kothandaraman, N. Praveena, K. Varadarajkumar, B. Madhav Rao, Dharmesh Dhabliya, Shivaprasad Satla, and Worku Abera. Intelligent forecasting of air quality and pollution prediction using machine learning. *Adsorption Science & Technology*, 2022:5086622, 2022. doi: 10.1155/2022/5086622. URL https://doi.org/10.1155/2022/5086622.

K. Kumar and B. P. Pande. Air pollution prediction with machine learning: a case study of indian cities. *International Journal of Environmental Science and Technology*, 20 (5):5333–5348, 2023. ISSN 1735-2630. doi: 10.1007/s13762-022-04241-5. URL https://doi.org/10.1007/s13762-022-04241-5. Received: 2023/05/01.

F. Naz, M. Fahim, A. A. Cheema, N. T. Viet, T.-V. Cao, R. Hunter, and T. Q. Duong. Two-stage feature engineering to predict air pollutants in urban areas. *IEEE Access*, 12: 114073–114085, 2024. doi: 10.1109/ACCESS.2024.3443810.

Olawale Emmanuel Rowland. Comparative analysis of meteorological parameters and their relationship with no_2, pm_{10}, $pm_{2.5}$ and o_3 concentrations at selected urban air quality monitoring stations in krakow, paris, and milan. *Discover Environment*, 2(1):75, 2024. ISSN 2731-9431. doi: 10.1007/s44274-024-00060-2. URL https://doi.org/10.1007/s44274-024-00060-2.

S. Roy, C.M. Rao, and M. Abioui. Evaluation of non-stationary spatial relationship between meteorological-environmental parameters and pm2.5. *Advances in Space Research*, 73(8): 4106–4124, 2024. ISSN 0273-1177. doi: https://doi.org/10.1016/j.asr.2024.01.009. URL https://www.sciencedirect.com/science/article/pii/S0273117724000279.

D. Sanjeev. Implementation of machine learning algorithms for analysis and prediction of air quality. *International Journal of Engineering Research & Technology*, 10(3):533–538, 2021. doi: 10.17577/IJERTV10IS030323.

Kunwar P. Singh, Shikha Gupta, Atulesh Kumar, and Sheo Prasad Shukla. Linear and nonlinear modeling approaches for urban air quality prediction. *Science of The Total Environment*, 426:244–255, 2012. ISSN 0048-9697. doi: https://doi.org/10.1016/j.

scitotenv.2012.03.076. URL https://www.sciencedirect.com/science/article/pii/S0048969712004809.

Hitesh Soneji. Historical weather data for indian cities, 2020. URL https://www.kaggle.com/dsv/1129180.

Gourav Suthar, Rajat Prakash Singhal, Sumit Khandelwal, Nivedita Kaul, Vinod Parmar, and Abhay Pratap Singh. Annual and seasonal assessment of spatiotemporal variation in $pm_{2.5}$ and gaseous air pollutants in bengaluru, india. *Environment, Development and Sustainability*, 26:20629–20652, 2024. doi: 10.1007/s10668-023-03495-4. URL https://doi.org/10.1007/s10668-023-03495-4. Received: 17 August 2022 / Accepted: 9 June 2023 / Published online: 14 June 2023.