International Journal of Innovative Research and Scientific Studies, 8(4) 2025, pages: 1321-1335

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Air quality forecasting using a modified statistical approach: Combining statistical and machine learning methods

Mohamed C. Ali[1], Ehab Ebrahim Mohamed Ebrahim[2], Mohamed R. Abonazel[3*]

[1]*Faculty of Business Administration, Deraya University, Minya, Egypt.*
[2]*Department of Economics, College of Business, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia.*
[3]*Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research, Cairo University, Giza 12613, Egypt.*

Corresponding author: Mohamed R. Abonazel (*Email: mabonazel@cu.edu.eg*)

Abstract

Accurate prediction of air quality in urban areas is critical due to the increasing impact of pollution on public health and environmental sustainability. The purpose of this study is to develop an enhanced forecasting model for urban air quality using a hybrid approach. The methodology integrates statistical techniques, namely least absolute shrinkage and selection operator (LASSO), ridge regression, and elastic net, with machine learning models such as random forest (RF), k-nearest neighbor regression (KNN), and extreme gradient boosting (XGBoost). A novel ensemble model combining elastic net and RF is proposed to improve predictive accuracy. The approach was validated using a comprehensive air quality dataset collected from multiple Indian cities between 2015 and 2020. India was chosen because it is one of the largest countries in Asia. The findings indicate that the hybrid model outperforms traditional statistical and machine learning models in terms of predictive performance, as assessed by robust goodness-of-fit metrics. In conclusion, the proposed method provides a powerful and reliable tool for predicting air quality and thus achieving environmental sustainability and meeting the basic needs of humans. Practical implications of this work include its potential use by policymakers and environmental agencies for proactive pollution management and public health planning.

Keywords: Big data, Environmental sustainability and basic needs, ridge regression, elastic net, LASSO, random forest. KNN, Xgboost model.

DOI: 10.53894/ijirss.v8i4.8061
Funding: This work is supported by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU), Saudi Arabia (Grant number: IMSIU-DDRSP2502).
History: Received: 6 May 2025 / **Revised:** 10 June 2025 / **Accepted:** 12 June 2025 / **Published:** 24 June 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

Air pollution has become a significant global issue because of its serious effects on human health, environmental standards, and climate change. The rapid growth of cities, industrial activities, and vehicle emissions has resulted in increasingly dangerous air quality in numerous urban centers, especially in developing countries like India. Fine particulate matter (PM2.5) and trace gases such as NO_2, CO, and SO_2 are strongly linked to respiratory and cardiovascular illnesses, highlighting the urgent necessity for reliable forecasting systems.

Air quality forecasting plays a critical role in achieving environmental sustainability by enabling proactive pollution control measures. This study's innovative approach combines statistical and machine learning methods for air quality prediction, directly supporting sustainable development goals (SDGs), particularly Goal 11.6 (reducing cities' environmental impact) and Goal 3.9 (minimizing pollution-related health risks). By improving forecast accuracy, our research contributes to smarter urban planning, better public health interventions, and more effective climate change mitigation strategies all critical components of building sustainable communities and safeguarding environmental resources for future generations.

According to Kumar and Pande [1] the existence of air is critical for human survival. However, advancements in other aspects of modern life, such as industry, transportation, and household activities, have had a negative impact on overall air quality. As a result of the event, harmful pollutants have been released into the surrounding ecosystem. Monitoring and forecasting air quality have grown in importance, particularly in developing countries such as India. In contrast to traditional methodologies, the use of machine learning-based prediction technologies has shown usefulness in the study of current environmental threats. This study examines a dataset containing air pollution statistics from 23 cities in India over a six-year period. The study's primary goal is to create a predictive model for air quality. The dataset has been effectively preprocessed, and key features have been identified using correlation analysis. Exploratory data analysis is used to uncover hidden patterns, with a particular emphasis on identifying contaminants that have a direct influence on the air quality index.

Surprisingly, the year 2020 saw a significant decrease in pollution levels. To address the issue of data imbalance, a resampling approach is used, followed by the application of five different machine learning models for air quality prediction. These models' performances are evaluated using well-established metrics. The maximum level of accuracy is demonstrated by the Gaussian Naive Bayes model, while the lowest level of accuracy is demonstrated by the Support Vector Machine model. Based on predefined parameters, the authors conduct a comparative evaluation of the performances of various models. Their findings show that the XGBoost model outperforms the others, with the highest level of concordance between predicted and actual data.

In the year Singgih [2] contribution was of great significance, or an event of exceptional nature took place in Singgih [2] life. There is a substantial body of research that supports the association between exposure to contaminated air and the development of many diseases, particularly those that impact the cardiovascular and pulmonary systems. Numerous air pollutants, such as nitrogen dioxide (NO2), carbon monoxide (CO), ozone (O3), sulfur dioxide (SO2), and particulate matter (PM), possess the capacity to inflict damage upon the respiratory system. Several local governments have built real-time monitoring systems in order to evaluate air quality and provide the public with information regarding the suitability of engaging in outdoor activities. Several international institutions, including Peking University, Christchurch, and Los Angeles, have successfully used data collection and analysis systems to improve their understanding of local air quality conditions. These technologies facilitate the acquisition of data and distribution of information regarding air quality to the wider population.

To increase the performance of the Post-Selection Boosting Random Forest (PBRF) algorithm, Farhadi et al. [3] developed an innovative methodology known as "Reducing and Aggregating Random Forest Trees Using an Elastic Net" (RARTEN). The random forest architecture was improved by using penalized regression techniques. The RARTEN method consists of three sequential steps. To begin, a random forest model is employed to predict. Second, by lowering the number of trees, an Elastic Net approach is used to refine both the random forest and the PBRF. Finally, the selected trees are combined to give the final result.

RARTEN's effectiveness was assessed utilizing statistical performance indicators using real-world data and Monte Carlo simulations. The simulation study findings show that including RARTEN improved the accuracy of both the classic random forest and Wang's proposed technique significantly. Reductions of 7%, 5%, and 8.5% were achieved in the linear, nonlinear, and noisy models, respectively. Furthermore, when compared to other penalized regression algorithms, RARTEN displayed a significant improvement. The empirical findings, which showed a drop of around 16%, further validated the usefulness of the proposed technique.

Because of its numerous structural components and well-documented role in various human illnesses, Cai et al. [4] highlighted the prospective therapeutic potential of RNA as a target. However, progress in drug discovery and development has been stymied by a lack of knowledge of the variables that influence the interaction between RNA and small molecules. This difficulty is exacerbated by the lack of empirically validated structure-activity relationships (QSAR). To address this issue, the researchers developed a unique method that used ensemble learning approaches to predict the binding affinity and kinetic properties of small molecules targeting the HIV-1 TAR model RNA system.

The researchers created a training dataset of small chemicals screened against the HIV-1 TAR RNA construct and used surface plasmon resonance to evaluate binding kinetics and affinities. Ensemble learning approaches and structure-based chemical descriptors were used to build predictive models. External validation was employed to assess the model's accuracy, specifically its ability to accurately predict the binding properties of new molecules not included in the training set. This marks the first application of a predicted and experimentally confirmed two-dimensional quantitative structure-activity relationship (2D QSAR) to an RNA target, specifically the HIV-1 TAR RNA. The platform developed in this study is a valuable tool for guiding future synthetic efforts and has the potential to be extended to diverse RNA architectures, providing

insights into the unique properties of small molecules that promote selective binding interactions. Ultimately, this technology could significantly enhance the efficiency of ligand design and optimization by reducing reliance on high-resolution structures.

In a separate study, Singgih [2] compared various machine learning techniques, including random forest, decision tree, neural network, and naive Bayes, to assess their effectiveness in predicting city air quality. They utilized an online dataset and applied preprocessing techniques such as removing records with missing values, splitting the data into training and testing sets, and employing K-fold cross-validation. The results revealed that the top-performing methods were random forest, gradient boosting, and K-nearest neighbors, all achieving precision, recall, and F1-score values exceeding 0.63.

Singh et al. [5] successfully addressed the inverse problem of retrieving subsurface physical features information from observed data through the utilization of a machine learning approach called Random Forest Regressor (RFR). A dataset pertaining to geophysics. Random Forest Regression (RFR) operates by partitioning the dataset into several equal subsamples and constructing multiple decision trees on each subsample. The outcomes from these decision trees are subsequently aggregated to generate a final prediction. The selection of this particular technique was based on its ability to effectively address intricate inversion problems, provided that enough quantity of training examples is available. Additionally, this strategy eliminates the need for calculating forward code at each iteration of the process. The method employed by researchers involved the analysis of synthetic magnetotelluric (MT) and DC resistivity datasets pertaining to a three-layered earth. The obtained results were subsequently compared with outcomes derived from alternative methodologies, namely Particle Swarm Optimization (PSO), genetic algorithm (GA), Ridge Regression (RR) algorithm, PSO, and Grey Wolf Optimization (GWO) techniques. The methodologies encompassed in this study comprise Particle Swarm Optimization (PSO), genetic algorithms (GA), Ridge Regression (RR) algorithms, as well as other instances of PSO. The results derived with RFR demonstrated a significant level of agreement with the true model parameters and were comparable to or even surpassed the results produced from alternative approaches. Another paper proposed two innovative hybrid estimation methods, LASSOPBRF and EnetRARTEN, aimed at enhancing prediction accuracy in high-dimensional regression frameworks. These approaches merge the variable selection advantages of LASSO and elastic net with the ensemble learning potential of the random forest algorithm. In particular, LASSOPBRF combines LASSO with an additional boosting phase for RF trees after selection, while EnetRARTEN utilizes elastic net regularization to streamline and consolidate trees in the RF model. Through comprehensive Monte Carlo simulations and a practical application in air quality assessment, the authors illustrated that these hybrid models surpass traditional methods in terms of mean square error (MSE) and root mean square error (RMSE), even in the presence of significant multicollinearity and outlier contamination. Their results highlight the promise of hybrid statistical-machine learning models in addressing the challenges present in high-dimensional datasets.

In their 2017 study, De Ávila et al. [6] used crystallographic structures of protein-ligand complexes to construct a machine-learning model for predicting binding affinity. They used an ensemble of crystallographic structures with resolutions less than 1.5 Å, as well as half-maximal inhibitory concentration (IC50) data. They used energy terms from the MolDock and PLANTS scoring systems as explanatory variables to build polynomial scoring methods for binding affinity prediction. The prediction model's performance was examined, and it was discovered that the supervised machine learning models exceeded the PLANTS and MolDock scoring functions in terms of predictive capabilities. Furthermore, when compared to scores produced by AutoDock4, AutoDock Vina, MolDock, and PLANTS, the machine-learning approach outperformed them in predicting CDK2 binding affinity.

Despite the progress made, several obstacles still persist. Firstly, there is an inconsistency in the assessment of hybrid models across various urban datasets. Secondly, current models frequently neglect the temporal and spatial fluctuations in pollutant levels. Thirdly, there has been limited research on contextualizing model outcomes for decision-making related to policy. This research aims to fill these gaps by introducing an improved hybrid model that integrates elastic net and random forest techniques for predicting urban air quality. In particular, we intend to:

- Assess the predictive capabilities of hybrid models utilizing real-world air quality information from cities in India.
- Contrast their performance against traditional and machine learning models under different data quality scenarios.
- Determine the most significant pollutants influencing the Air Quality Index (AQI).

The study addresses the following research questions:

1. How do hybrid statistical-machine learning models compare with standalone models in forecasting air quality?
2. Are these models capable of effectively managing high-dimensional, noisy, or imbalanced environmental datasets?
3. What pollutants are primarily responsible for the variability of AQI in urban settings in India?

To meet these aims, we utilized a six-year dataset (2015–2020) from 23 cities in India. We executed preprocessing techniques that included feature selection, and subsequently, we trained five predictive models: LASSO, elastic net, RF, KNN, and XGBoost. Our proposed model combines elastic net with RF to enhance both accuracy and robustness. The evaluation metrics we used encompass mean square error (MSE), root mean square error (RMSE), and various goodness-of-fit statistics.

The subsequent sections of this paper are organized as follows: Section 2 outlines the materials and methods employed in this study, Section 3 presents the exploratory data analysis conducted, Section 4 introduces the proposed method, Section 5 discusses the results and analysis derived from the study, and finally, Section 6 presents the conclusion.

2. Material and Methods

Several Indian towns are among the most polluted in the world, highlighting the growing urgency of the air pollution problem. Inadequate air quality in India is widely acknowledged as a major health issue as well as a significant obstacle to the country's economic progress. According to a recent study conducted by Dalberg [7] Advisors, a UK-based non-profit

management organization, in collaboration with the Industrial Development Corporation, has highlighted that the negative effects of air pollution in India have resulted in significant annual economic losses totaling approximately Rs 7 lakh crore (equivalent to $95 billion) [7]. Energy-generating businesses, road traffic, soil and road dust, trash incineration, power plants, and open waste burning are the principal sources of pollution emissions in India. The current study concentrated on air pollution statistics collected by India's Central Pollution Control Board (CPCB). This dataset spans the period from January 2015 to July 2020, and it contains 12 separate variables and 29,531 individual data points collected from 23 different cities across India. The primary goal of this research was to examine significant atmospheric contaminants such as the Air Quality Index (AQI), nitrogen dioxide (NO2), carbon monoxide (CO), sulfur dioxide (SO2), and ozone (O3), as well as to forecast particulate matter with a diameter of 2.5 micrometers or less (PM2.5) Figure 3 depicts the approach used in this experiment.

2.1. Data Preprocessing

Ahmedabad, Aizawl, Amaravati, Amritsar, Bengaluru, Bhopal, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram, and Visakhapatnam were included in this study. The dataset includes 29,531 observations of air quality statistics and the Air Quality Index (AQI) at hourly and daily intervals. These observations were gathered from a variety of sites located throughout India. Daily data collection was carried out between January 1, 2015, and January 7, 2020. According to Bhat [8] it has been proposed. The data set included 11 independent variables and one response variable, PM2.5. Table 1 lists the variables that were used.

The major goal of this study is to assess the relative importance of independent variables in accurately predicting the value of the dependent variable, namely particulate matter PM2.5 concentration. This inquiry considers potential difficulties in the data such as multicollinearity and outliers. To achieve this goal, the research team used a variety of model selection strategies, including LASSO, Ridge, Elastic Net, Random Forest, KNN, XGBoost, and a unique methodology called Elastic Net Random Forest.

Additionally, a comprehensive analysis was conducted to gain a deep understanding of the data and identify any potential patterns or correlations. The assessment of multicollinearity and data anomalies involved the use of various techniques, such as the correlation matrix, descriptive statistics, and boxplots. The aim of this inquiry is to identify the most reliable approach for predicting PM2.5, taking into consideration both the data characteristics and the independent variables.

As depicted in Figure 2, certain associations between variables exhibit significantly higher levels of statistical significance than others. Figure 1 presents the correlation coefficients, highlighting strong associations (exceeding 0.6) among specific independent variables. The presence of such strong associations can potentially lead to multicollinearity, which can affect the statistical analysis.

To assess multicollinearity, the variance inflation factor (VIF) was calculated for each variable in the model, as indicated in Table 3. The VIF values exceeding five for several variables suggest the presence of multicollinearity. Consequently, it can be inferred that the model's reliability in predicting the impact of individual independent factors on the dependent variable is questionable due to the strong correlation among these independent variables, which can result in unstable and unreliable estimates of the regression coefficients.

Figure 2 illustrates the identification of outliers within the dataset using both histograms and box plots. The visual representations reveal data points that exceed the upper extreme or fall below the lower extreme in all variables. However, pinpointing the exact locations of these outliers posed a challenge. The "sklearn" module in Python was utilized to generate the correlation matrix, descriptive statistics, and box plots.

Table 1.
Variable description.

Variable	Description
PM2.5	PM2.5 refers to airborne particulate matter with a mass concentration of particles that are smaller than 2.5 micrometers per cubic meter.
NO	The acronym NO stands for nitrogen oxide.
NO2	In this context, the abbreviation NO2 denotes nitrogen dioxide.
NOx	NOx refers to the collective quantity of nitrogen oxide (NO) and nitrogen dioxide (NO2).
NH3	The chemical compound NH3 is commonly known as ammonia.
CO	The term "CO" is an abbreviation for carbon monoxide.
SO2	SO2 is an abbreviation for sulfur dioxide.
O3	The term O3 is commonly used to denote the chemical compound known as ozone.
Benzene	Benzene is a chemical compound that consists of six carbon atoms arranged in a hexagonal ring
Toluene	Toluene is a colorless liquid hydrocarbon compound that belongs to the aromatic hydrocarbon family.
Xylene	Xylene is a chemical compound that belongs to the group of aromatic hydrocarbons. It
AQI	The Air Quality Index (AQI) is a metric used to evaluate and measure the quality of air in a particular location.

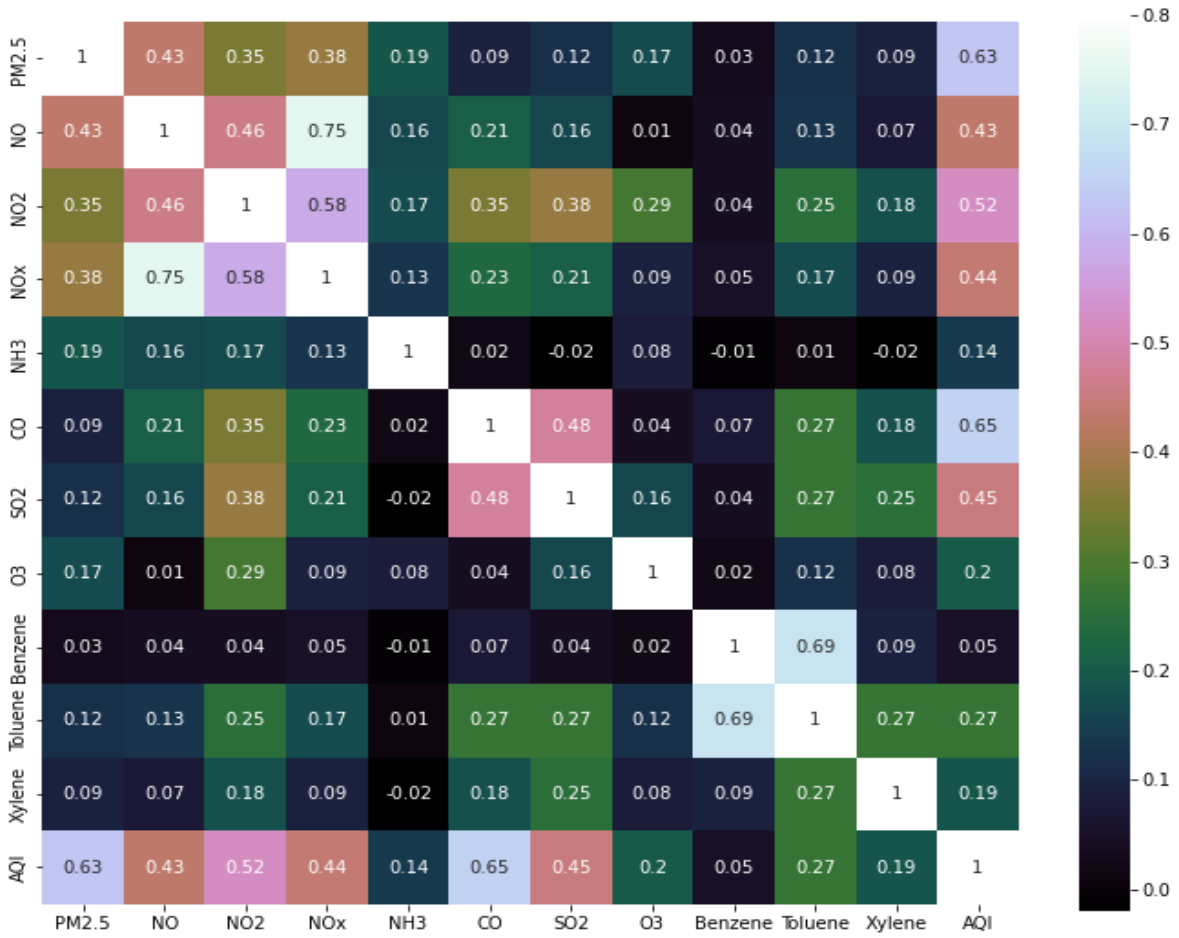
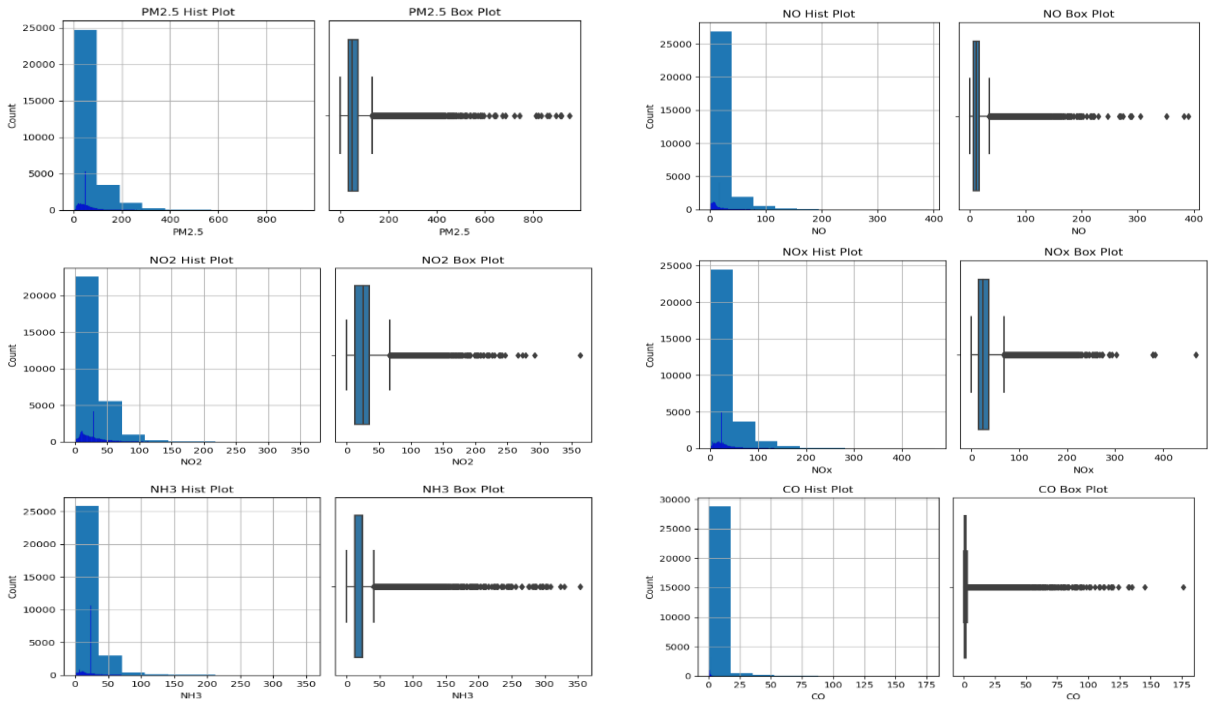


Figure 1.
Correlation Matrix.



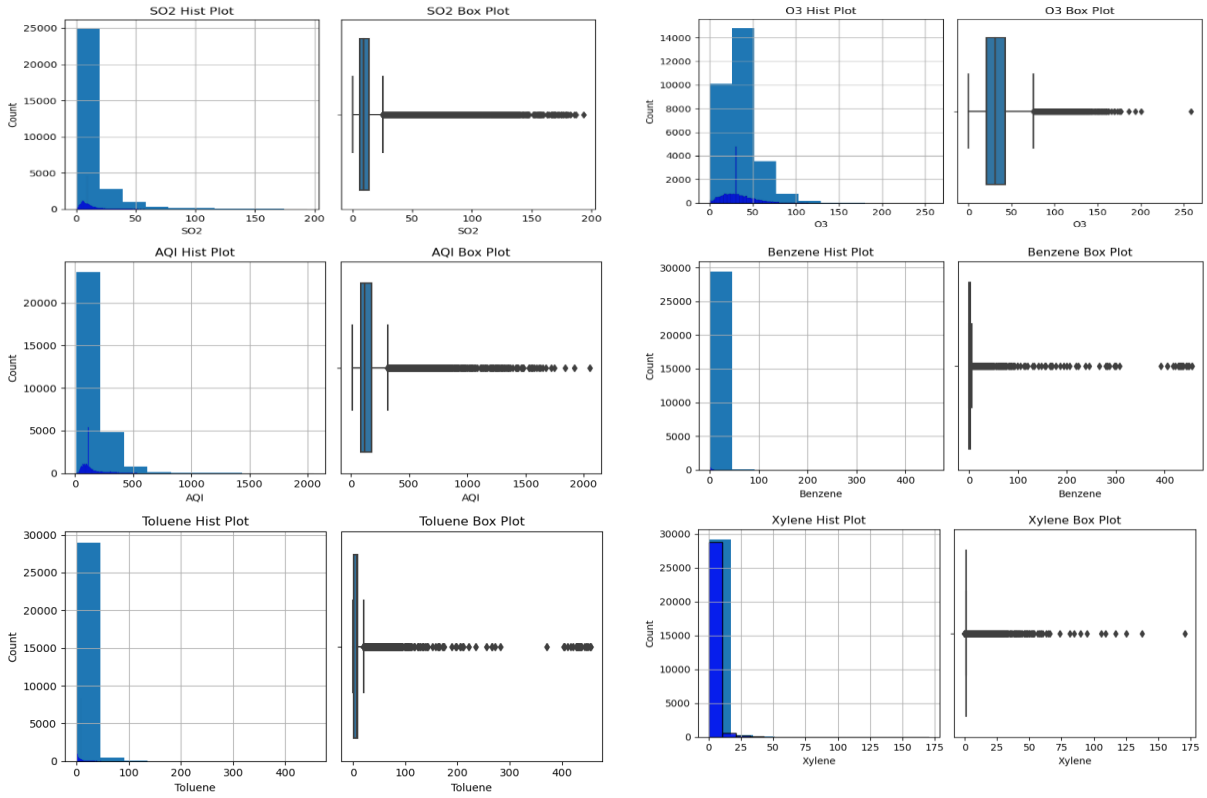


Figure 2.
Histogram and Boxplot of all variables.

Table 2.
Descriptive statistics.

Variable	Mean	Std.	Min.	25%	50%	75%	Max.
PM2.5	64.51	59.81	0.04	32.15	48.57	72.45	949.99
NO	17.5747	21.36	0.02	6.21	11.53	17.57	390.68
NO2	28.56	22.94	0.01	12.98	25.24	34.67	362.21
NOx	31.064	29.48	0	14.67	23.52	36.02	467.63
NH3	23.48	20.71	0.01	12.04	23.43	23.48	352.89
CO	2.15	6.72	0	0.54	0.89	1.38	175.81
SO2	13.83	17	0.01	6.09	9.16	13.81	193.86
O3	33.99	20.2	0.01	20.74	30.84	42.73	257.73
Benzene	2.86	14.25	0	0.24	1.07	2.42	455.03
Toluene	8.70	17.03	0	1.28	6.93	8.70	454.85
Xylene	1.79	4.06	0.	0.98	0.98	0.98	170.37
AQI	158.78	130.27	13.	88.	118.	179	2049

Table 3.
VIF.

Variable	VIF
NO	4.020
NO2	5.059
NOx	5.633
NH3	2.137
CO	2.164
SO2	2.443
O3	3.286
Benzene	2.181
Toluene	3.113
Xylene	1.355
AQI	5.803

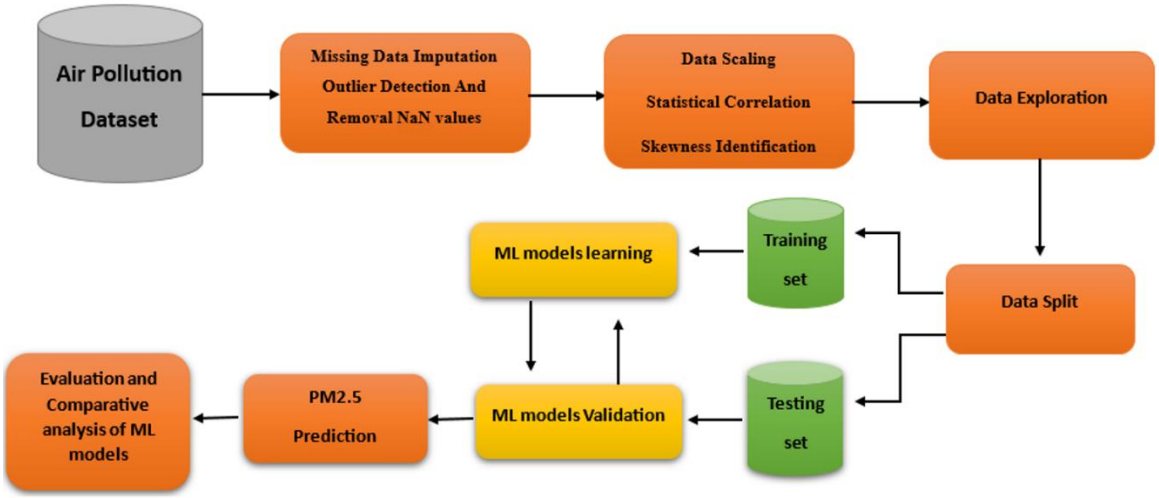


Figure 3.
Flowchart of proposed methods.

2.2. Variable Selection Methods
2.2.1. Naive Elastic Net

Zou and Hastie [9] proposed the Elastic Net as an innovative method for regularization and variable selection in linear regression, which has had a significant impact in the field. The Elastic Net is a mathematical model that integrates L1 and L2 regularization approaches in a linear fashion, thus successfully mitigating specific constraints associated with the Lasso and Ridge regression methodologies. The Elastic Net approach offers distinct advantages in scenarios where the number of predictors (p) significantly surpasses the number of observations (n), a circumstance that is not applicable to the Lasso method.

The findings derived from the conducted simulation studies revealed that the Elastic Net algorithm consistently exhibited superior performance compared to the LASSO algorithm, while still preserving a similar degree of sparsity. Furthermore, the utilization of the Elastic Net regularization technique fosters the emergence of a phenomenon commonly referred to as the "grouping effect." This impact manifests when variables that exhibit strong correlation are inclined to be either collectively included or collectively omitted from the model. The authors of this study presented an algorithm called LARS-EN to effectively calculate the regularization paths for Elastic Net.

Suppose the data set has n observations with p predictors. Let $y = (y_1, \cdots, y_n)^T$ be the response and $X = [X_1| \cdots |X_n]$ be the model matrix, where $x_j = \left(x_{1j}, \cdots, x_{nj}\right)^T, j = 1, \ldots, p$ are the predictors. After a location and scale transformation, we can assume the response is centered and the predictors are standardized,

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \text{ and } \sum_{i=1}^{n} x_{ij}^2 = 1, \text{ for }, j = 1, \ldots, p \tag{1}$$

For any fixed non-negative λ_1 and λ_2, we define the naive elastic net criterion

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1, \tag{2}$$

Where

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2, \text{ and } |\beta|_1 = \sum_{j=1}^{p} |\beta_j| \tag{3}$$

The naive elastic net estimator $\hat{\beta}$ is the minimizer of (3):

$$\hat{\beta} = \arg\min L(\lambda_1, \lambda_2, \beta) \tag{4}$$

The above procedure can be viewed as a penalized least-squares method. Let $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$, then solving $\hat{\beta}$ in (3) is equivalent to the optimization problem:

$$\hat{\beta} = \arg\min L(\lambda_1, \lambda_2, \beta), \text{ subject to } (1-\alpha)|\beta|_1 + \alpha|\beta|^2 \leq t \text{ for some t.} \tag{5}$$

The Elastic Net penalty function is defined as $(1-\alpha)|\beta|_1 + \alpha|\beta|^2$, where α is a parameter that determines the balance between the Lasso and Ridge penalties. The parameter α governs the trade-off between L1 and L2 regularization. When the value of α is set to 1, the Elastic Net method transforms into Ridge regression, while a value of α equal to 0 results in the transformation of the method into Lasso regression.

The previously mentioned phenomenon can be observed in a contour plot that represents two dimensions. The outermost contour in this plot depicts the shape of the Ridge penalty, while the diamond-shaped curve represents the Lasso penalty. Furthermore, the red solid curve represents the application of the Elastic Net penalty, where the coefficient α is set at 0.5. The contour plot demonstrates a high level of convexity along its edges, with the degree of convexity being contingent upon the parameter α. In this work, the researchers employed the "glmnet" package in R version 4.0.0 to implement the Elastic Net methodology.

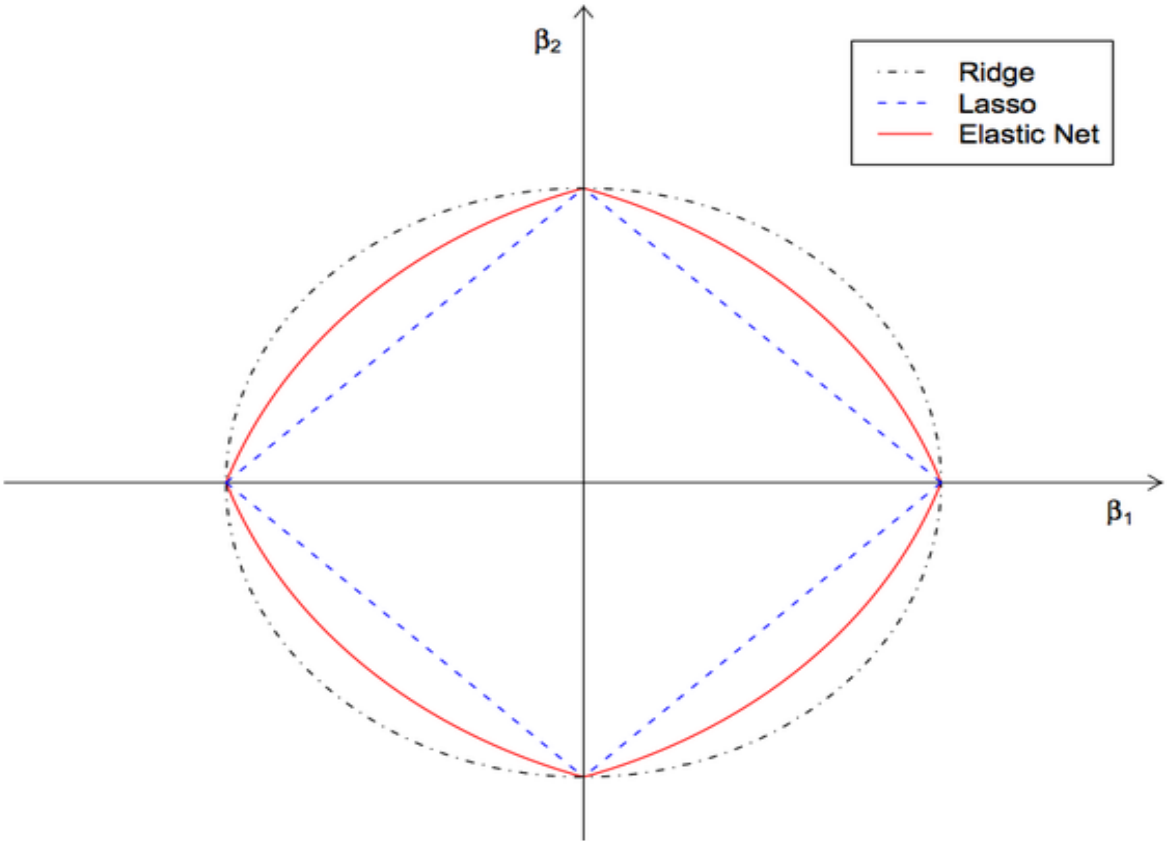


Figure 4.
The geometric properties of the elastic net penalty.

2.2.2. Breiman's Random Forest Mechanism (RF)

The Random Forest technique is well acknowledged in the field of machine learning and has demonstrated its effectiveness in tackling various practical problems. These include applications such as air quality prediction, chemoinformatics, ecology, 3D object recognition, and bioinformatics, among others. The approach, which was proposed by Breiman [10], is classified as an ensemble learning method that combines many randomized decision trees and aggregates their predictions via averaging. This phenomenon is especially beneficial in scenarios when the number of variables surpasses the number of observations.

The Random Forest algorithm is widely recognized as a valuable computational technique for effectively solving both regression and classification applications. Ensemble techniques involve the integration of multiple machine learning methods to improve the accuracy of predictions.

The Random Forest methodology is a technique that involves creating a collection of decision trees, where the entire dataset is divided into subsets to facilitate prediction. Each subgroup leads to the formation of a unique decision tree within the random forest. In the context of machine learning, it is observed that each decision tree produces a specific outcome. Subsequently, in the case of a random forest model, the final decision is determined by selecting the majority outcome from the individual decision trees. In this study, the investigators utilized the Random Forest package in R version 4.0.0 to implement the random forest technique.

Algorithm 1 outlines the steps for the Random Forest algorithm proposed by Abd Algani et al. [11]:

Step 1: Initially, construct N decision trees
Step 2: Start with the root node data
Step 3: Select an attribute and create a logical attribute check
Step 4: Direct each test result to the relevant child node by sending a subset of satisfactory examples
Step 5: Visit every node of the child
Step 6: Repeat the process until leaf nodes are 'pure'
Step 7: Take the majority decision of the Decision Trees as the final decision

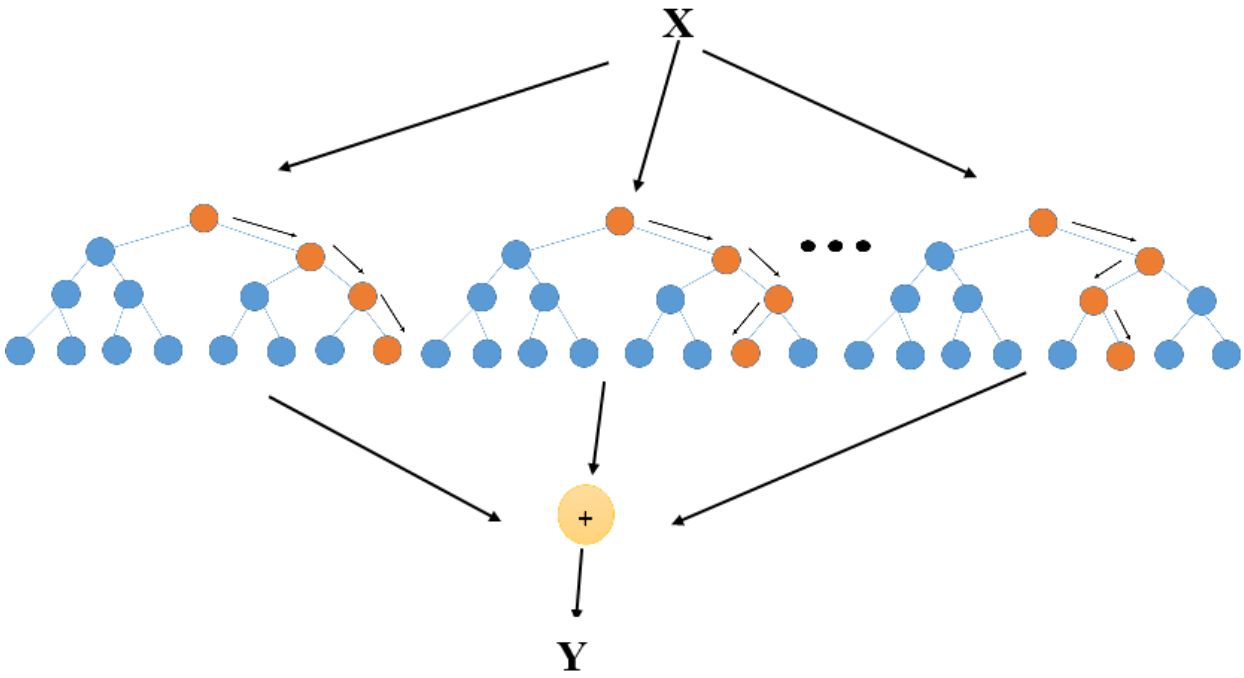


Figure 5.
RF for regression.

Figure 5 illustrates the functioning of the Random Forest Regressor (RFR), which involves the equitable and random partitioning of the training dataset into many groups. Then, every dataset is presented as input to a decision tree, which then analyzes the dataset and generates its own output. The Random Forest Regression (RFR) algorithm incorporates the predictions generated by all individual decision trees and subsequently derives an outcome by calculating the average of these predictions.

2.2.3. K-Nearest-Neighbor Regression

The KNN regression technique was employed to forecast the value of y for a specific sample, and the subsequent procedures were executed. Initially, the selection of the number of neighbors, denoted as k, was made. This implies that a collection of k samples was identified from the original dataset X, which exhibited the closest proximity to the independent variable of sample a. Next, the distance of each sample from the target sample was calculated. Ultimately, we ascertain the precise location at which the magnitude of the separation between the k samples is minimized. By employing a distance measuring technique such as Euclidean distance, it becomes possible to determine the nearest k sample points, denoted as e_1, e_2,$\cdots$, e_k, where: According to the study conducted by Jiang et al. [12].

$$d = \sqrt{\sum_{k=1}^{m}(x_{lk} - x_{ik})^2} \quad (6)$$

The y values of the k sample points are computed by taking their average, which is then used as the y value for the target sample.

$$y = \frac{y_1+y_2+y_3+\cdots+y_k}{k} \quad (7)$$

We can use the above process to obtain the predicted result for the given sample.

2.2.4. eXtreme Gradient Boosting (XGBoost) Regression

According to the study conducted by Shehadeh et al. [13] the XGBoost algorithm has gained significant popularity across various domains due to its organizational structure, adaptability, and ease of implementation. The technique integrates a cause-based decision tree (CBDT) and Gradient Boosting Machine (GBM), enabling efficient and precise processing of diverse input types. The aforementioned algorithm demonstrates significant utility in the development of forecasting models due to its capability to effectively manage both regression and classification procedures for specific datasets. XGBoost is very suitable for the analysis of extensive datasets containing several independent variables and classifications. It has the potential to provide efficacious resolutions for novel optimization challenges, particularly in cases where striking a balance between efficiency and accuracy has significance.

3. Exploratory Data Analysis

In the present investigation, an examination was conducted on the data in order to identify any latent patterns that might exist within the dataset. To achieve this objective, it is important to conduct exploratory data analysis (EDA), which conventionally serves as the initial phase in the data analytics process and precedes the implementation of any machine learning model. In this section of the analysis, we will examine two highly consequential factors: (a) an examination of the patterns and trends pertaining to air pollutants between 2015 and 2020, encompassing a span of six years (see Figure 6), and

(b) an investigation into the dispersion of pollutants in the atmosphere, along with an identification of the six cities exhibiting the highest average PM2.5 levels. Both of these components encompass a duration of six years. (c) Determine an approximation for the four most significant pollutants that are predominantly accountable for PM2.5.

3.1. This Study Aims to Examine the Temporal Patterns of Air Pollution Over a Six-Year Period

In recent years, India has experienced significant instances of air pollution due to the rapid growth of its industrial and urban sectors. The occurrence of several challenges impacting public health and the environment has emerged as a direct outcome. Notably, air pollution has risen to prominence as one of the foremost global risk factors for mortality, ranking among the top five. Based on a study conducted by the Health Effects Institute in 2017, it was found that particulate matter (PM) emissions rank as the third leading cause of mortality on a global scale. India has the highest documented death rate. India has been acknowledged by the World Health Organization (WHO) as one of the nations with the worst levels of pollution globally, mostly attributed to its significant emissions of PM2.5 particles. India was assigned this ranking based on its current standing within the global context. The presented graphic depicts the temporal trends of pollutants spanning the period from 2015 to 2020. Figure 7 is depicted in the following manner. Nevertheless, by the year 2020, a notable reduction in the concentrations of all pollutants was observed, except for O_3 and Benzene. The decline coincided with the implementation of the most rigorous nationwide lockdown measures in the country's history, prompted by the COVID-19 pandemic. The environment and air quality had a discernible enhancement due to the decrease in industrial, automotive, and aviation operations. Figure 8 depicts a visual representation of the mean PM2.5 concentrations observed in the six most heavily polluted regions in India for the duration of the monitoring period.

4. Proposed Method

In this section, the authors introduce a hybrid approach they've named EentRF, which combines Elastic Net and Random Forest methods. The approach employed in this study is analogous to the methodologies presented by El-Sheikh et al. [14], wherein they integrated LASSO with Neural Network (NN) and Random Forest (RF) with NN, respectively. One of the main advantages of employing Elastic Net is its ability to induce a grouping effect, wherein predictors that exhibit strong correlation are more likely to be simultaneously included or excluded from the model. This is particularly advantageous in scenarios where the number of indicators (p) exceeds the total number of observations significantly. The user's text is insufficient to be rewritten in an academic manner. The Random Forest (RF) algorithm exhibits superior accuracy compared to the decision tree algorithm and moreover provides an effective approach for addressing data scarcity. Furthermore, at the specific point of node division, stochastic selection is performed from a subset of the features within the Random Forest (RF) model. The authors believe that by utilizing this combination strategy, as opposed to conventional statistical methods such as OLS, LASSO, Ridge, and Elastic Net, and machine learning algorithms such as RF, they will be able to attain considerably more resilient goodness of fit.

Algorithm 2 outlines the steps for the proposed Elastic Net RF (EN-RF) method:
Step 1: The analysis commences with the utilization of an Elastic Net model.
Step 2: The task at hand involves the identification and selection of the most pertinent variables from the Elastic Net model.
Step 3: The chosen variables should be inputted into the Random Forest algorithm.

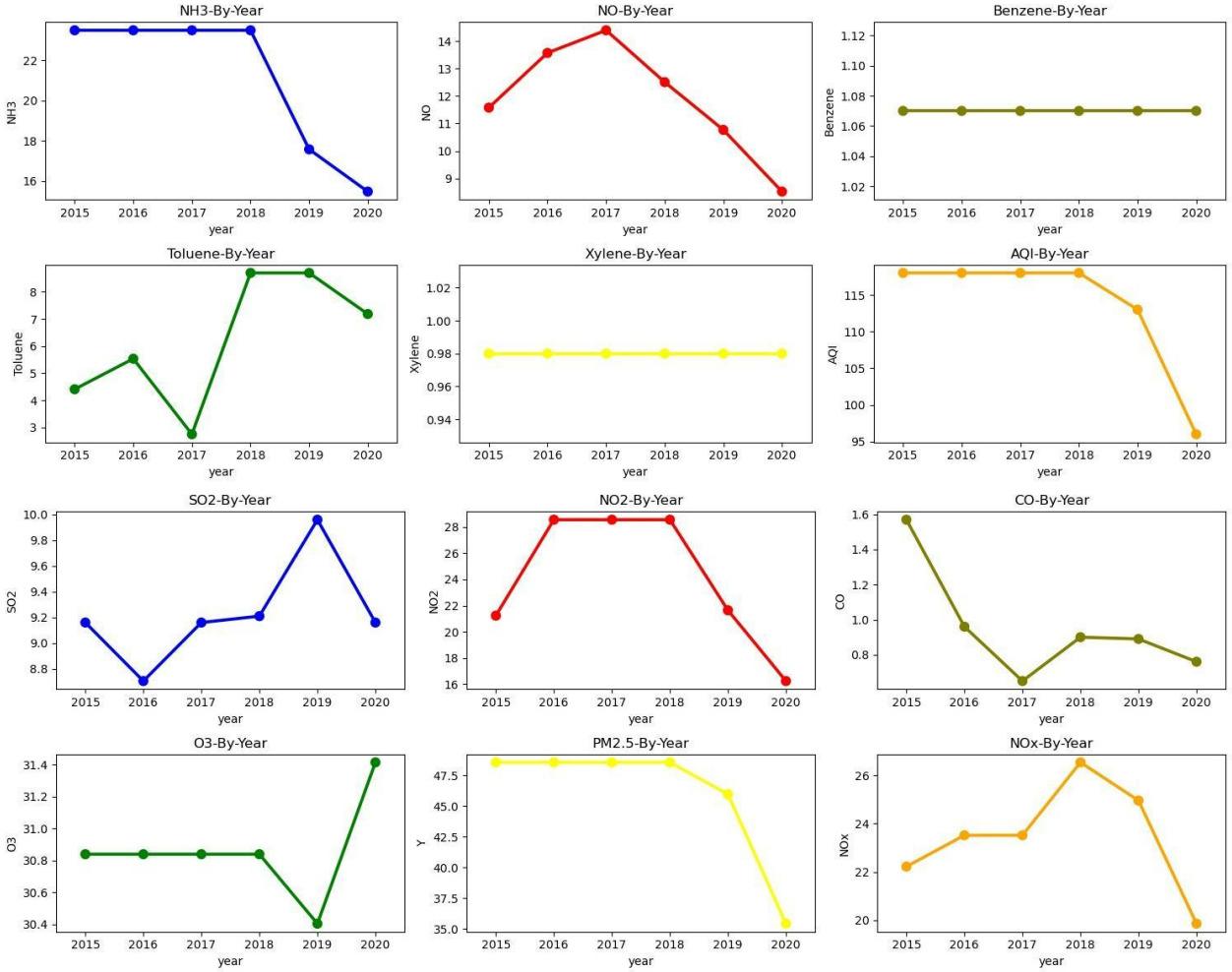
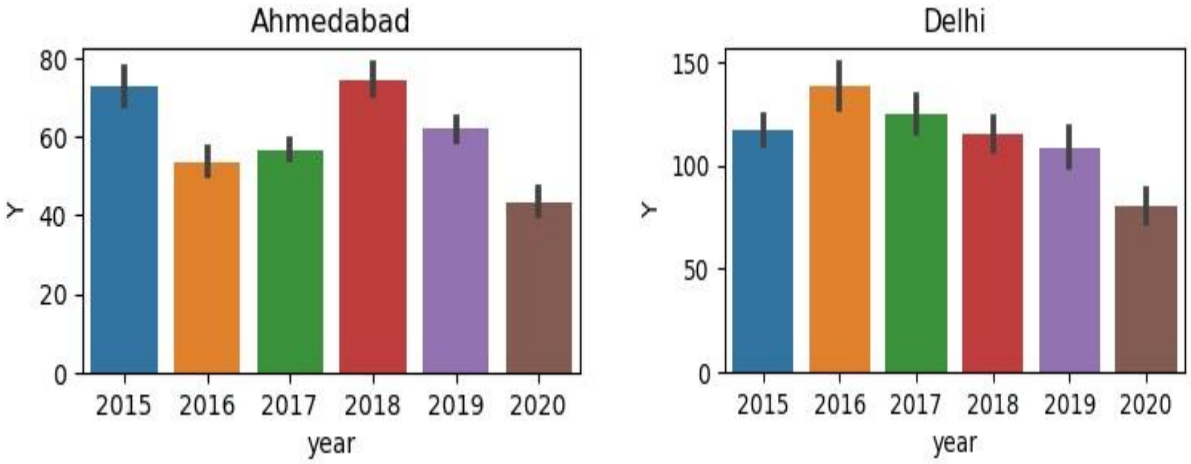


Figure 6.
Variations in pollution levels from 2015 to 2020.



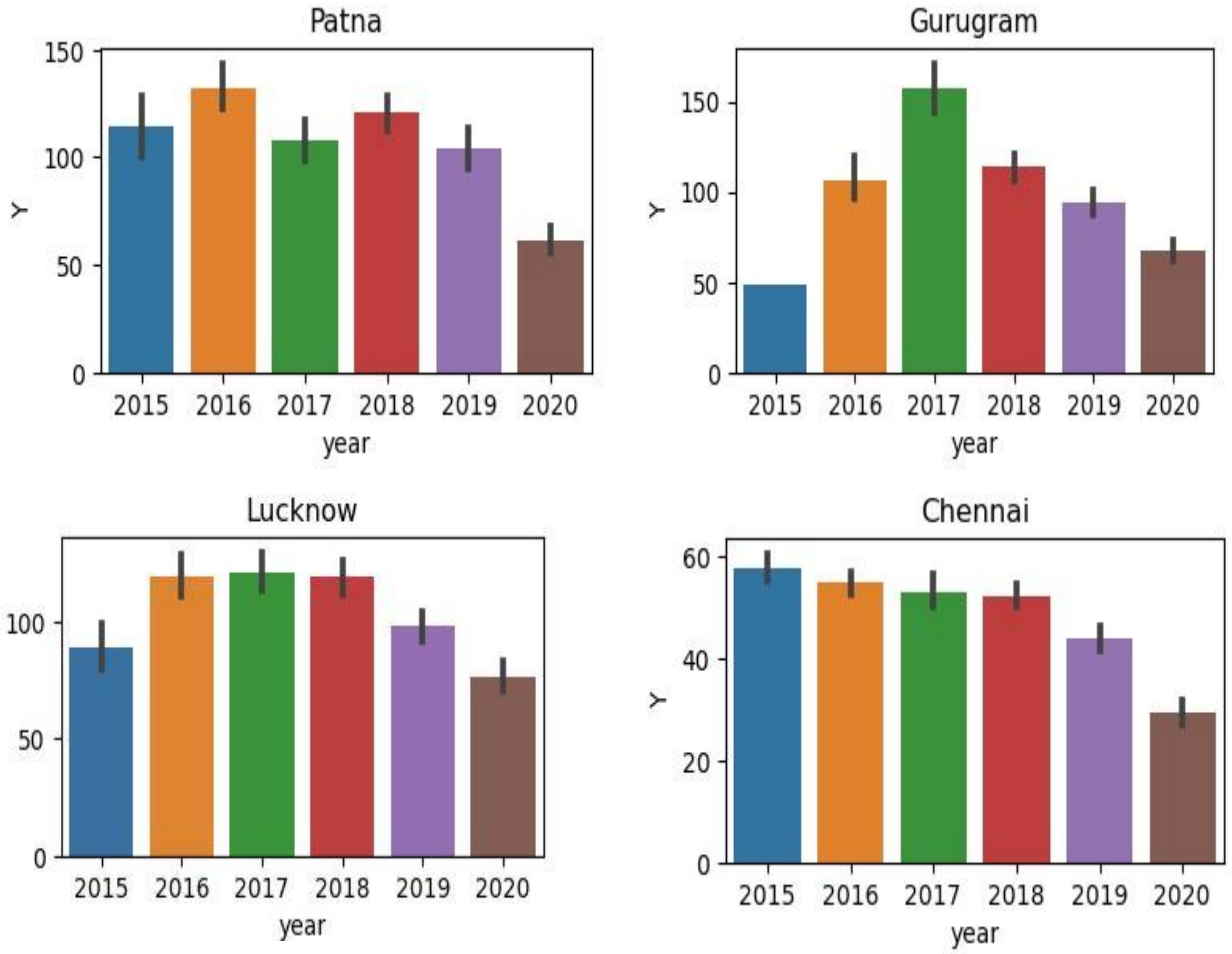
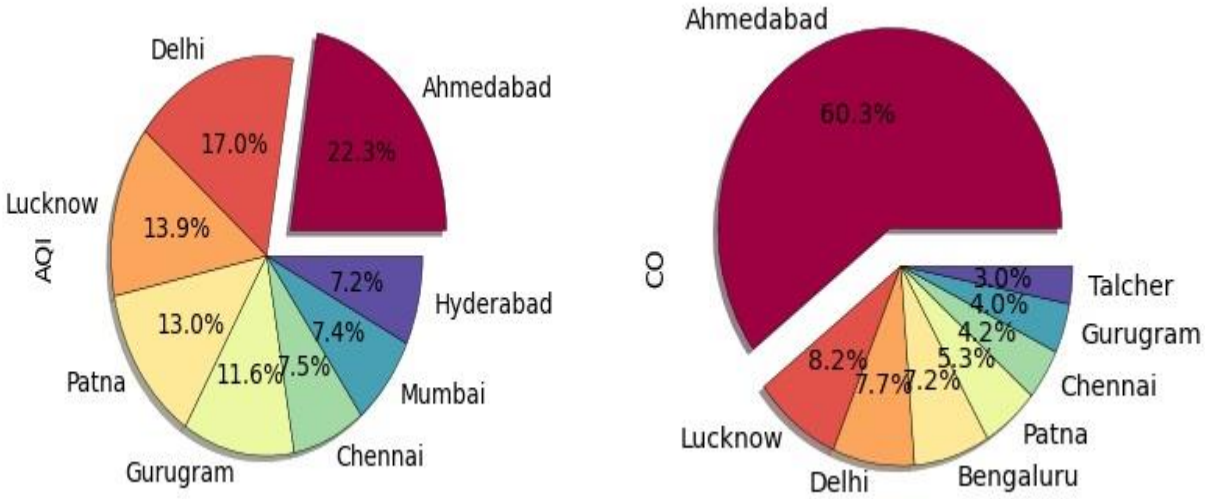


Figure 7.
The six Indian cities with the highest PM2.5 values between 2015 and 2020.



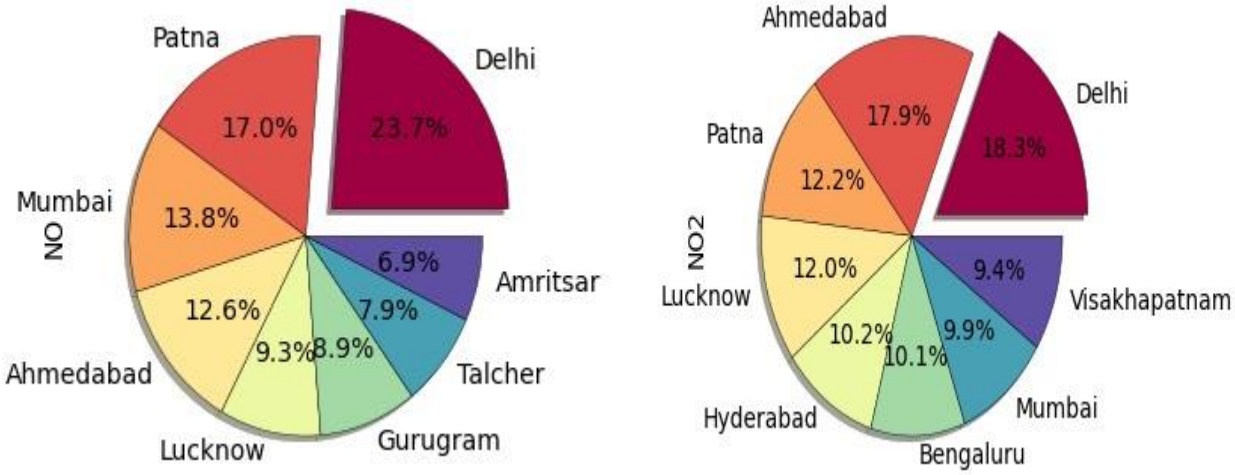


Figure 8.
Pollutants that have a direct influence on PM 2.5.

5. Results and Discussion

The analysis steps performed in this study are as follows:

1. Detection and Treatment of Outliers In our dataset, a notable presence of outliers has been observed. The qualifier variable for all pollutants, particles, and meteorological conditions exhibits a substantial proportion of missing data. Consequently, a decision was made to impute the missing values using the mean.

2. The dataset was partitioned into separate training and testing datasets. In the present study, the proportion of the testing data was designated as 50%. The data underwent a shuffling process prior to being divided.

3. The precision of each strategy was assessed by employing mean square error (MSE), root mean square error (RMSE), and mean absolute percentage error (MAPE) as evaluation metrics.

4. The procedure of model selection entails the utilization of cross-validation methodologies to assess and compare the efficacy of different models. The models encompassed in this study include Lasso, Ridge, Elastic Net, Random Forest, K-Nearest Neighbors (KNN), XGBoost, and the recently proposed Elastic Net Random Forest.

5. Comparative Analysis and Determination of the Optimal Method: The selection of the ideal technique was based on a thorough review of several metrics and its ability to successfully handle multicollinearity and outliers seen in the dataset. In light of the offered findings and analysis, it is apparent that a thorough assessment of the data substantiates the subsequent advice. The investigation produced results that informed the development of conclusions and recommendations for the key independent variables and the most effective approach for predicting PM2.5 levels in air quality. Table 4 displays the variables that have been selected as components of the Elastic Net Model, denoted as the chosen variables (SV). Figure 9 depicts the importance of the independent variable, as evidenced by the EnetRF model. Based on the results presented in Table 5, it can be deduced that the ordinary least squares (OLS), least absolute shrinkage and selection operator (LASSO), ridge regression, k-nearest neighbors (KNN), XGBoost, and random forest (RF) techniques include all independent variables in their respective analyses. In contrast, the Enet model explicitly identifies and integrates a limited set of 10 independent variables. The methodology described in this study involves employing EnetRF algorithms to reduce the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The elastic net and random forest (RF) methods demonstrated superior performance.

Table 4.
Elastic Net Model.

SV	β_0	NO	NO2	NOx	NH3	CO	SO2	Benzene	Toluene	Xylene	AQI
Coefficient	(2.97)	0.31	0.02	(0.037)	0.12	(5.57)	(0.25)	0.098	(0.024)	0.19	0.47

Table 5.
Goodness of fit measures for methods.

Criteria	OLS	LASSO	Ridge	Enet	KNN	Xgboost	RF	EnetRF
MSE	1558.371	1556.812	1556.525	1553.112	1013.827	858..291	733.665	728.132
RMSE	39.476	39.456	39.452	39.409	31.84	29.296	27.086	26.983
MAPE	0.663	0.35	0.347	0.336	0.237	0.229	0.194	0.191
#SV	11	11	11	10	11	11	11	10

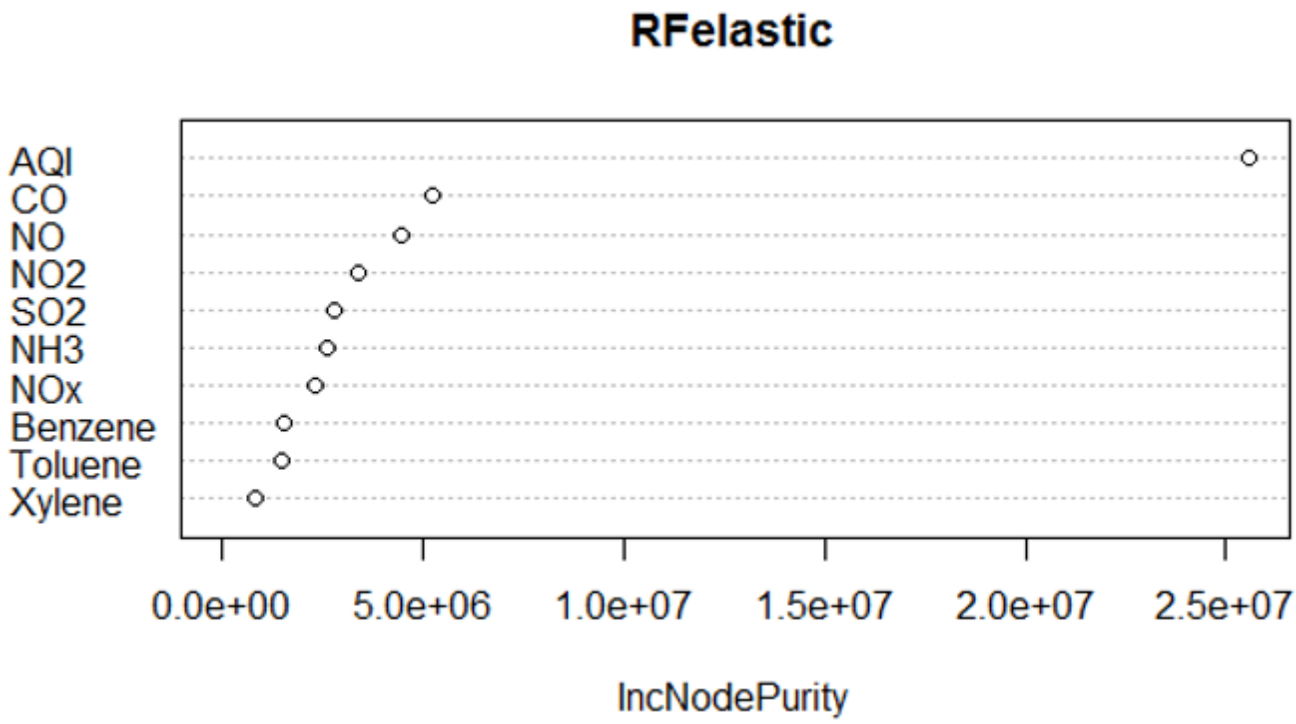


Figure 9.
Variable importance of the Elastic Net RF model.

6. Conclusions

Evaluating and predicting air quality are essential yet intricate tasks due to the ever-changing nature of environmental factors and the uneven spread of pollutants over time and geography. This research concentrated on assessing pollution levels in 23 cities across India over a span of six years by employing a hybrid modeling framework that merged statistical and machine learning approaches. After performing data cleaning, imputation, and exploratory data analysis, a range of models such as OLS, LASSO, ridge regression, elastic net, random forest, KNN, and XGBoost were utilized to forecast air quality indicators. The findings revealed the existence of multicollinearity and outliers, which were efficiently addressed using the hybrid EnetRF approach. This method showcased enhanced predictive performance regarding MSE, MAE, and RMSE, surpassing conventional models. The results highlight the efficacy of hybrid models for handling high-dimensional, noisy, and correlated environmental data. The combination of elastic net with random forest (EnetRF) facilitated effective variable selection and increased predictive accuracy. This modeling strategy can be utilized by environmental authorities and policymakers for real-time air quality assessment, particularly within smart city frameworks, aiding in the implementation of proactive measures to mitigate health hazards. Although the results are promising, the study has certain limitations. Firstly, it is confined to Indian cities and may not be applicable to other regions with distinct pollution sources or climatic conditions. Secondly, while the hybrid model tackles multicollinearity and outliers, it may still be vulnerable to extreme anomalies or patterns of missing data. Lastly, real-time streaming data was not included in this research, which could have further increased its practical applicability. Future research could explore various paths. Including satellite-based remote sensing data and meteorological factors could improve spatial-temporal accuracy. Additionally, applying the proposed methodology in other areas, like the Middle East or Southeast Asia, would assist in validating its broader applicability. Integrating deep learning techniques and real-time data feeds within IoT systems could further enhance forecasting precision and responsiveness in urban smart environments.

References

[1] 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*, vol. 20, no. 5, pp. 5333–5348, 2023.

[2] I. K. Singgih, "Air quality prediction in smart city's information system " *International Journal of Informatics, Information System and Computer Engineering*, vol. 1, no. 1, pp. 35–46, 2020.

[3] Z. Farhadi, H. Bevrani, and M. R. Feizi-Derakhshi, "Improving random forest algorithm by selecting appropriate penalized method," *Communications in Statistics - Simulation and Computation*, vol. 51, no. 10, pp. 3566–3581, 2022.

[4] Z. Cai, M. Zafferani, and A. Hargrove, "Ensemble learning-based quantitative structure-activity relationship platform predicts binding behavior of RNA-targeted small molecules," *Journal of Medicinal Chemistry*, vol. 64, no. 9, pp. 5204–5217, 2021.

[5] A. P. Singh, D. Vashisth, and S. Srivastava, "Random forest regressor for layered earth data inversion," presented at the 2019 Fall Meeting, AGU, San Francisco, CA, USA. AGU S53D-0483, 2019.

[6] M. B. De Ávila, M. M. Xavier, V. O. Pintro, and W. F. de Azevedo Jr, "Supervised machine learning techniques to predict binding affinity. A study for cyclin-dependent kinase 2," *Biochemical and Biophysical Research Communications*, vol. 494, no. 1-2, pp. 305-310, 2017.

[7] Dalberg, *Air pollution and its impact on business: The silent pandemic*. London, UK: Clean Air Fund, 2019.

[8] N. Bhat, *Air quality level of different cities in India (2015–2020)*. San Francisco, CA: Kaggle, 2020.

[9] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301-320, 2005.
[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
[11] Y. M. Abd Algani, M. Ritonga, B. K. Bala, M. S. Al Ansari, M. Badr, and A. I. Taloba, "Machine learning in health condition check-up: An approach using Breiman's random forest algorithm," *Measurement: Sensors*, vol. 23, p. 100406, 2022. https://doi.org/10.1016/j.measen.2022.100406
[12] D. Jiang, J. Zhang, Z. Wang, C. Feng, K. Jiao, and R. Xu, "A prediction model of blast furnace slag viscosity based on principal component analysis and K-nearest neighbor regression," *JOM*, vol. 72, pp. 3908–3916, 2020.
[13] A. Shehadeh, O. Alshboul, R. E. Al Mamlook, and O. Hamedat, "Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression," *Automation in Construction*, vol. 129, p. 103827, 2021.
[14] A. A. El-Sheikh, M. R. Abonazel, and M. C. Ali, "Proposed two variable selection methods for big data: Simulation and application to air quality data in Italy," *Communications in Mathematical Biology and Neuroscience*, vol. 2022, pp. 1–20, 2022.