atmosphere

MDPI

Article

Assessment of Fine Particulate Matter for Port City of Eastern Peninsular India Using Gradient Boosting Machine Learning Model

Manoj Sharma [1], Naresh Kumar [2,*], Shallu Sharma [3], Vikas Jangra [4], Seema Mehandia [5,*], Sumit Kumar [6,*] and Pawan Kumar [7,*]

1 Department of Electronics and Communication Engineering, Giani Zail Singh Campus College of Engineering & Technology, Maharaja Ranjit Singh Punjab Technical University, Bathinda 151001, India; neelmanoj@gmail.com
2 Department of Electronics and Communication Engineering, University Institute of Engineering and Technology (UIET), Panjab University, Chandigarh 160014, India
3 Neuroimaging & Neuro Spectroscopy Lab, National Brain Research Centre, Manesar, Gurugram 122051, Haryana, India; shallu.hari16@gmail.com
4 Department of Statistics, Panjab University, Chandigarh 160014, India; vjangra49@gmail.com
5 Department of Biotechnology, University Institute of Engineering and Technology (UIET), Panjab University, Chandigarh 160014, India
6 Division of Research and Development, Centre for Space Research, School of Electronics and Electrical Engineering (SEEE), Lovely Professional University, Phagwara 144411, Punjab, India
7 Materials Research Application Lab (MARL), Department of Nano Sciences & Materials, Central University of Jammu, Jammu 181143, India
* Correspondence: naresh_uiet@pu.ac.in (N.K.); seema_uiet@yahoo.co.in (S.M.); kumarsumit8@gmail.com (S.K.); pawannano10@gmail.com (P.K.)

check for updates

Citation: Sharma, M.; Kumar, N.; Sharma, S.; Jangra, V.; Mehandia, S.; Kumar, S.; Kumar, P. Assessment of Fine Particulate Matter for Port City of Eastern Peninsular India Using Gradient Boosting Machine Learning Model. *Atmosphere* **2022**, *13*, 743. https://doi.org/10.3390/atmos13050743

Academic Editor: Jianmin Chen

Received: 11 April 2022
Accepted: 4 May 2022
Published: 6 May 2022

Abstract: An assessment and prediction of $PM_{2.5}$ for a port city of eastern peninsular India is presented. Fifteen machine learning (ML) regression models were trained, tested and implemented to predict the $PM_{2.5}$ concentration. The predicting ability of regression models was validated using air pollutants and meteorological parameters as input variables collected from sites located at Visakhapatnam, a port city on the eastern side of peninsular India, for the assessment period 2018–2019. Highly correlated air pollutants and meteorological parameters with $PM_{2.5}$ concentration were evaluated and presented during the period under study. It was found that the CatBoost regression model outperformed all other employed regression models in predicting $PM_{2.5}$ concentration with an R^2 score (coefficient of determination) of 0.81, median absolute error (MedAE) of 6.95 $\mu g/m^3$, mean absolute percentage error (MAPE) of 0.29, root mean square error (RMSE) of 11.42 $\mu g/m^3$ and mean absolute error (MAE) of 9.07 $\mu g/m^3$. High $PM_{2.5}$ concentration prediction results in contrast to Indian standards were also presented. In depth seasonal assessments of $PM_{2.5}$ concentration were presented, to show variance in $PM_{2.5}$ concentration during dominant seasons.

Keywords: air pollution; $PM_{2.5}$; gradient boosting; machine learning; CatBoost regression model

1. Introduction

The World Health Organization (WHO) has reported that 9 out of 10 people breathe air containing high levels of pollutants, and it is also estimated that air pollution is responsible for approximately 7 million deaths every year [1]. The burden of ill-health is not equally distributed, as approximately two-thirds of deaths occur in the developing countries of Asia. Air pollution in Asian countries is mainly due to increasing trends in economic and social development. In India, rapidly increasing industrialization, urbanization, and demand for transportation influence air pollution in many Indian cities [2]. The major health impacts of air pollutant $PM_{2.5}$ are shown in Figure 1, which indicates that $PM_{2.5}$ is the main air pollutant responsible for the most significant health problems viz. impacts on the central

nervous system, breathing problems, chronic obstructive pulmonary disease, lung cancer, cardiovascular diseases, impacts on the reproductive system, etc., [3,4]. Therefore, it is of utmost importance to determine and forecast the $PM_{2.5}$ concentration with more reliable and better forecasting models.

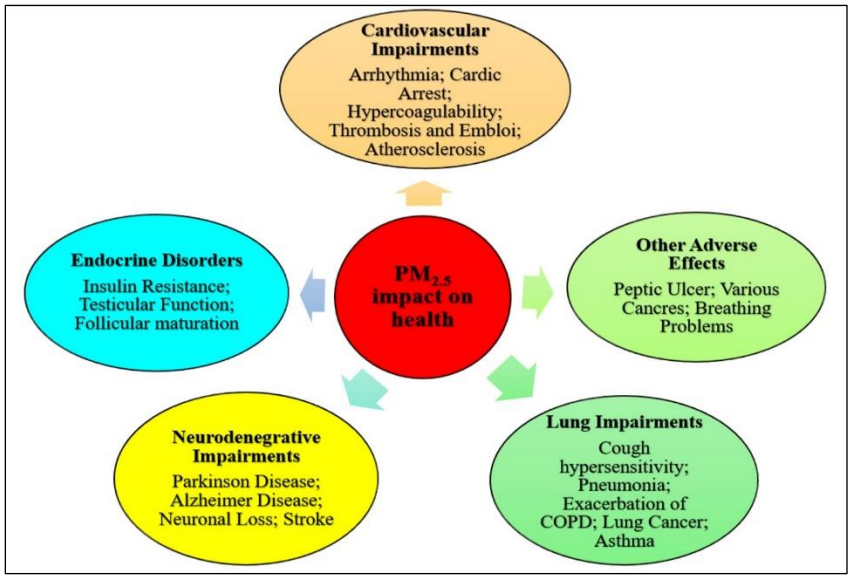


Figure 1. Health impacts of air pollution.

Many studies have reported on pollutant forecasting models using ML approaches because it is very difficult to estimate originator pollutants with varying environmental conditions by traditional methods. ML approaches are capable of determining the causes of high pollutants, along with better forecasting of the increase and decrease in pollutants in the environment. Forecasting can help regulatory agencies or the government to control emission levels of pollutants such as nitric oxide (NO), nitrogen dioxide (NO_2), nitrogen oxides (NO_x), ammonia (NH_3), sulphur dioxide (SO_2), carbon monoxide (CO), benzene (C_6H_6), toluene (C_7H_8), etc., and alert residents to avoid outdoor activity, especially patients with respiratory problems.

ML approaches are based on data collected through various sensors located in different parts of the city. ML algorithms have advanced over the past few years, and their prediction is based on the quality of the data collection, i.e., data required for training the models. In our study, we considered the data collected at Visakhapatnam, a port city on the eastern side of peninsular India, for the assessment period 2018–2019.

This paper is organized as follows: Section 2 provides information about the literature review; Section 3 introduces the study area, methods of data collection, and descriptions of data and the machine learning regression model. Section 4 describes performance measurement indicators. Section 5 describes the results related to Visakhapatnam's three predominant seasons and a comparison of prediction results of fifteen ML regression models. Section 6 presents the conclusion of the proposed work.

2. Literature Review

A number of studies have reported the implementation of ML approaches for urban air pollution, including $PM_{2.5}$ concentration in recent years. Masood and Ahmad [5] presented a comparison of support vector machine (SVM) and artificial neural network (ANN) for the prediction of anthropogenic fine particulate matter (or $PM_{2.5}$) for Delhi, based on various meteorological and pollutant parameters corresponding to a 2 year period from 2016 to 2018. ANN provided faster prediction and better accuracy, compared with SVM. Deters et al. [6] carried out research on the prediction of $PM_{2.5}$ with the help of ML regression models considering selected meteorological parameters, and reported a high correlation between

estimated data and real data. Moisan et al. [7] compared the performances of a dynamic multiple equation (DME) model in forecasting $PM_{2.5}$ concentrations, with an ANN model and an ARIMAX model. They concluded that, although ANN in very few instances showed more significant and accurate results than the DME model, the overall performance of the DME model was slightly better than the ANN. Jiang et al. [8] reported the prediction of $PM_{2.5}$ with the help of a long short-term memory (LSTM) model, with better accuracy. Suleiman et al. [9] compared three air quality control strategies, including SVM, ANN and boosted regression trees (BRT) for forecasting $PM_{2.5}$ and PM_{10} concentrations in the roadside environment. They found that regression models with neural networks had better prediction performance, compared with SVM. A regression model which used extra trees regression and AdaBoost, for further boosting, was proposed for estimating $PM_{2.5}$ for Delhi by Kumar S. et al. [10]. Regression and statistical methods were proposed by Chandu and Dasari [11] to study the causal relationship between $PM_{2.5}$ and gaseous pollutants for the city of Visakhapatnam, but no forecasting model was proposed.

We chose Visakhapatnam, a port city of east peninsular India, which is surrounded on three sides by mountains and the Bay of Bengal on the fourth. This city is studded with major industries, including Hindustan Petroleum Corporation Limited (HPCL), Hindustan Zinc Limited (HZL), Bharat Heavy Plates and Vessels (BHPV), Hindustan Polymers Limited (HPL), Visakhapatnam Steel Plant (SP), Coastal Chemicals (CC), Andhra Cement Company (ACC) and Simhadri Thermal Power Corporation (STPC) [11]. Approximately 200 ancillary industries operate to supplement these main industries thereby causing air pollution, hence air quality research becomes a priority.

There are many ML regression-based models available in the literature; of these, fifteen ML regression models (presented in Section 3.3) were tested and implemented to predict the $PM_{2.5}$ concentrations for Visakhapatnam. Among them, gradient boosting models (CatBoost, XGBoost and LightGBM) and voting regression models provided significant prediction results. Overall, the CatBoost model (that utilizes oblivious decision trees as the base learner) as proposed by Yandex Company [12] resulted in excellent prediction performance. The CatBoost model was able to archive comparable results, as obtained by other state of the art regression models such as RNN–LSTM [13], multivariate linear regression model [14] and LSTM [15]. Prediction accuracy of these regression models was validated using air pollutants (NO, NO_2, NO_x, NH_3, SO_2, CO, C_6H_6, C_7H_8) and meteorological parameters (relative humidity (RH), wind speed (WS), wind direction (WD), solar radiation (SR) and barometric pressure (BP), and temperature) as input variables.

This research work is different from previously reported studies on $PM_{2.5}$, in multiple ways. In this study, in-depth seasonal and yearly variations in $PM_{2.5}$ concentration were analyzed for two years. Fifteen machine learning methods were analyzed for optimal performance, compared with only two to three methods in most other studies, offering a more robust analysis for comparing models for real world implementation [6–10].

The main highlights of this manuscript are:

(1) Analysis of the concentration of $PM_{2.5}$ and air pollutants in the air of the eastern peninsular port city of India, on yearly and seasonal bases;
(2) Evaluation of the correlation between $PM_{2.5}$ concentration, air pollutants and meteorological parameters for the port city of Visakhapatnam;
(3) Observation of the CatBoost prediction model as the most efficient prediction model for assessing and predicting the concentration of $PM_{2.5}$ in the air;
(4) Analysis of high $PM_{2.5}$ concentration prediction results for the period under observation.

3. Materials and Methods

3.1. Study Area

Figure 2 shows the area under observation and wind rose diagram of the observation period. The study area, Visakhapatnam, a port city on the eastern side of peninsular India, with latitude 17°42′15″ N and longitude 83°17′52″ E, is located in the Indian state of Andhra Pradesh. The monitoring station for data collection was set up by the Andhra

Pradesh Pollution Control Board (APPCB) [16]. Visakhapatnam is the largest city of Andhra Pradesh and the second largest east coast city in India, lying between the coast of the Bay of Bengal and the Eastern Ghats. According to the 2011 Indian census, Visakhapatnam had a population of 1,728,128 with a population density of 18,480/km^2 [17].

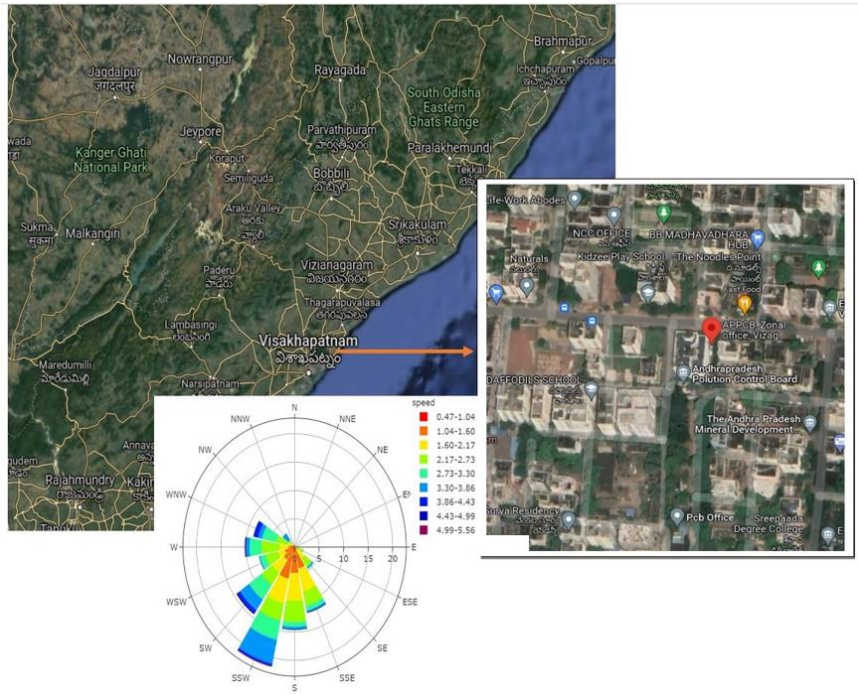


Figure 2. Location of port city Visakhapatnam (courtesy: google maps), and wind rose diagram (2018–2019) of study area.

Visakhapatnam observes tropical wet and dry climates during the year, with three dominant seasons: a summer season from March to June, a monsoon season from July to September, and a winter season from October to February. Though the summer extends from March to June, the maximum temperature is observed mainly in the month of May. The monsoon season extends from July to September, where an annual average rainfall of 44.05 inches was witnessed [18]. During winter, the minimum temperature is observed mainly in the month of January. The annual mean temperature of the city varies from 24.7 to 30.6 °C and observes an RH in the range of 68–80% [18].

During the period of observation, wind in the city blew with a mean speed of 2.32 m/sec in the direction of southwest (213.80 degrees, mean direction). The wind rose diagram in Figure 2 (based on a 24 h mean value) represents the direction of the wind, speed of wind and wind frequency of the location under observation for 2018–2019. The extent of the spoke determines the frequency of wind blowing in a specific direction. It is noted that the current $PM_{2.5}$ concentration in Visakhapatnam air is 8 times above the WHO annual air quality guideline value [2]. Due to the extreme variation in climatic conditions and a very high concentration of $PM_{2.5}$, there is a need to analyze ambient air quality, in addition to the impact of climatic factors, on the air quality of the city.

3.2. Data Description

In the present work, the data for air pollutants and $PM_{2.5}$ concentration in the air, along with meteorological parameters, were collected for two consecutive years 2018–2019. Visakhapatnam has nine monitoring stations located at Industrial Estate Marripalem, Parawada, GVMC, Raitu Bajar, Police Barracks, Pedagantyada Gajuwada, Naval Area, Seethammadhara and Ganapuram Area, to measure the air quality index of the city. The original readings of air pollutant concentration and meteorological parameters were provided by

the Central Pollution Control Board (CPCB) website [2] and APPCB [16] on 30 min, 1 h, 4 h, 8 h, 24 h and annual bases. For the present study, 24 h mean values of concentration of air pollutants and meteorological parameters were noted and utilized as prime variables for our prediction models. The raw dataset contained the record of air pollutants and meteorological parameters for 730 days. Of these days, the records for 32 days (19 days in 2018, and 13 days in 2019) were completely missed (either due to power failure, device failure or other reasons). After removing the 32 days, approximately 5% to 6% of values were further missed due to various parameters. The missing values were imputed using K-Nearest Neighbour (KNN) imputation method, and the values were imputed using the mean value from the n nearest neighbors. The value with k = 1, using the heterogeneous euclidean overlap metric (HEOM) distance, was chosen for the missing imputation for the presented dataset. After carefully processing, and missing value imputation, the observational record of air pollutants and meteorological parameters for 698 days was considered for our $PM_{2.5}$ concentration analysis. The statistical descriptions of primary variables considered for analysis are presented in Table 1. For the period under observation, the city observed mean $PM_{2.5}$ concentration of 48.63 $\mu g/m^3$ with standard deviation of 30.05 $\mu g/m^3$. Similarly, Table 1 shows the mean, standard deviation, minimum and maximum values of air pollutants and meteorological parameters.

Table 1. Statistical description of air pollutants and meteorological parameters for the assessment period 2018–2019.

Variable	Unit	Mean	Std. Deviation	Min. Value	Max. Value
$PM_{2.5}$	$\mu g/m^3$	48.63	30.05	3.06	202.52
NO	$\mu g/m^3$	13.4	11.47	0.45	99.32
NO_2	$\mu g/m^3$	38.37	16.26	0.56	131.53
NO_x	ppb	31.82	15.97	0.31	129.9
NH_3	$\mu g/m^3$	10.96	6.31	0.16	63.62
SO_2	$\mu g/m^3$	12.05	7.42	1.32	60.83
CO	mg/m^3	0.78	0.27	0.15	2.18
C_6H_6	$\mu g/m^3$	4.46	1.75	0.01	11.69
C_7H_8	$\mu g/m^3$	10.15	5.33	0.01	54.29
Temperature	°C	29.26	1.95	23.36	42.68
RH	%	71.81	5.57	43.77	86.56
WS	m/s	2.32	0.83	0.47	5.56
WD	degree	213.80	50.44	79.45	351.46
SR	W/m^2	137.39	53.48	6.00	520.41
BP	mmHg	744.19	13.82	701.00	766.58

3.3. Machine Learning Model Description

For the prediction of $PM_{2.5}$ concentration, 15 regression models, namely, voting regression (VR) [19], CatBoost regression (CB) [20], LightGBM regression (LGBM) [21], XGBoost regression (XGB) [22], LASSO least angle regression (LR-LA) [23], random forest regression (RF) [24], multi-layer perceptron regression (MLP) [25], LASSO regression (LAR) [26], partial least square regression (PLS) [27], quantile regression (QR) [28], multi-task ElasticNet (MTE) [29], ridge regression (RR) [30], Bayesian ridge regression (BRR) [31], KNN regression (KNN) [32] and linear regression (LR) [33] were trained and tested using data collected for the years 2018–19. The notion behind the consideration of 15 regression models was that the models broadly cover different regression approaches, thus offering more robust results. The broad regression approaches for models under consideration were:

(a) Gradient boosting decision tree regression models: LGBM, XGB and CB;
(b) Ensemble regression models: VR and RF (tree-based bagging ensemble technique);
(c) Penalized regression models: LR–LA (lasso model with least-angle regression algorithm), LAR (lasso regularization technique to derive the coefficients exactly to zero), MTE (regularize multi-tasking using lasso and ridge norms) and RR (regularization technique to derive weights nearer to the origin);
(d) Linear regression models: LR (using linear equation, i.e., linear relation of inputs and single target), BRR (linear regression using probability distribution) and;
(e) Miscellaneous regression models: MLP (supervised neural network model), PLS (covariance-based statistical approach), QR (evaluates the median or other quantiles of target variable conditional on feature variables) and KNN (k-nearest neighbors algorithm).

However, despite predicting efficient results, regression models have their advantages and limitations. Table 2 presents the pros and cons of the proposed regression models for prediction of $PM_{2.5}$ concentration.

Table 2. Pros and cons of presented regression models.

Models	Pros	Cons
LGBM	• Less training time is required; • Requires low memory as the model replaces continuous values to discrete bins; • The model supports parallel learning.	• Suffers with overfitting problem in small size datasets; • More tuning parameters.
XGB	• No need for scaling, normalisation and other feature engineering; • Effect of outliers are minimal; • Feature importance can be evaluated.	• Large number of tuning parameters; • The models are hard to interpret.
CB	• Handles categorical features; • Reduces the need for extensive hyper-parameter tuning and hence reduces the chances of overfitting.	• Good for heterogeneous data, but may not be the optimal learner for homogeneous data.
VR	• The performance of voting regression combines the performance of many models, so poor performance of one model can be offset by strong performance of other models; • The performance of voting regression is not largely affected by one strong/weak model.	• Due to use of multiple models, the voting models are computationally intensive.
RF	• Efficiently handles missing values in a dataset; • Gives estimation about feature importance.	• Suffers with overfitting for noisy data; • Biased in the case of categorical data with different levels.
LR-LA	• Efficient for datasets having a number of features greater than the number of samples; • Easily adapted for other estimators.	• Sensitive to noisy data.
MTE	• Multiple features share similar spareness patterns; • Advantage of lasso and ridge norms.	• Difficult to solve due to non-smoothness of lasso and ridge regression.
RR	• Reduces the variance; • Avoids overfitting of the model; • Well suited for datasets with large numbers of independent variables.	• Not able to select important features; • Shrinks the regularization coefficient towards zero and hence the model is difficult to interpret.

Table 2. *Cont.*

Models	Pros	Cons
LR	• Simple and computationally efficient; • Able to derive the influence of input variables on the target variable; • Works well irrespective of size of dataset.	• Strong assumption that input variables and target are linearly related, which may not be the case; • The linear regression models assume homoscedasticity.
BRR	• Effective for a large or small dataset; • Well suited for real time learning.	• Computationally less efficient as it requires more execution time; • Least suited for large datasets.
MLP	• Suited for real time learning; • Can be applied to non-linear models; • Handles large datasets.	• Not able to define feature importance; • High execution time.
PLS	• Can handle multiple variables (nominal, continuous and ordinal); • Deals with multicollinearity; • Handles multi-dependent and multi-independent variables.	• Difficult to interpret the model; • No information regarding distributional properties of features.
QR	• Models are more robust to outliers and non-normal errors; • Changes the quantile variations with change in covariates.	• Coefficient estimates can be noisy; • Coefficients are non-monotonic across the quantile.
KNN	• Non parametric algorithm, and has no assumptions; • Based on instance-based learning and hence does not require learning during the training phase; • Addition of new data during training phase does not impact the prediction accuracy.	• Sensitive to outliers, noisy and missing data; • Finding the optimal value of k is challenging.

3.4. CatBoost (Based on Gradient Boosting Algorithm) Model Description

Gradient boosting is a significant and effective machine learning technology implemented to deal with noisy, diverse features and complex correlated information. Using iteration, the technique amalgamates weak machine learning models with the aid of gradient descent in function space [34]. A gradient boosting based CB model was proposed by Yandex Company and the model utilizes oblivious decision trees as the base learner [12]. The decision trees are implemented for regression. Each tree indicates division of feature space and output value. Decision rule/splitting criteria are used during division of trees. Individual splitting criteria resembles a pair $p = (q, m)$ having a feature indicator $q = 1, 2, \ldots . n$, and threshold value $m \, \epsilon \, D$. On implementing the decision rule/splitting criteria, a set of feature vectors X can be disjointed into two subsets of X^C and X^D, so that for every $x = \left(x^1, x^2, x^3, \ldots . x^n\right) \epsilon \, X$, we have:

$$x \, \epsilon \begin{cases} X^C & if \; x^q \leq m \\ X^D & if \; x^q > m \end{cases} \tag{1}$$

After implementing decision rule/splitting criteria to e disjoint sets $X_1, X_2 \ldots . . . X_e \, \epsilon \, D^n$, we obtain $2e$ disjoint sets $X_1^C, \; X_1^D, \; X_2^C, \; X_2^D \ldots . . . . . X_e^C, X_e^D$. For a specified collection

of sets $N = \{X_1, X_2, X_3, \ldots\ldots, X_e\}$ and the target variable $Y : D^n \rightarrow D$, the decision rule/splitting criteria can be given as:

$$arg\ min_p\{G(p, Y, N)\} \tag{2}$$

where N functions to estimate the optimality of the splitting criteria/decision rule p and the collection N with respect to the target variable Y. For an oblivious decision tree, G can be defined as:

$$G(p, Y, N) = \frac{1}{\sum_{a=1}^{e} |X_a|} \left[\sum_{a=1}^{e} \left| X_a^C \right| Var \left\{ Y\left(X_a^C \right) \right\} + \left| X_a^D \right| Var\left\{ Y\left(X_a^C \right) \right\} \right] \tag{3}$$

where $Y(X_a)$ is the target variable score with respect to the sample X_a. In contrast to other regression models, the CB model has following advantages:

(a) Categorical features: The model is capable of handling categorical features. In conventional gradient boosting decision tree-based algorithms, categorial features are replaced by their mean label value. If mean values are used to characterize features, then it will give rise to an effect of conditional shift [35]. However, in CB, an approach known as greedy target statistics is employed, and the model inculcates prior values to greedy target statistics. The employed technique reduces overfitting with minimum information loss;

(b) Combining features: CB implements a greedy way to amalgamate all of the multiple categorical features and their combinations by the current tree during the formation of the new split. All the splits in the decision tree are considered as categories with two disjoint values and are employed during amalgamation;

(c) The CB models are fast scorers. They are based on oblivious decision trees which are balanced and less inclined to overfitting.

4. Performance Measurement Indicators

Evaluation matrices for verification of high $PM_{2.5}$ concentration: The National Ambient Air Quality Standards (NAAQS) [36] were developed by the Central Pollution Control Board, Ministry of Environment and Forests (Government of India), to regulate pollutant emissions into the air. According to the standards, a mean 24 h $PM_{2.5}$ concentration of 60 $\mu g/m^3$ was classified as higher concentration. For evaluating the high $PM_{2.5}$ concentration, the following evaluation parameters were used: hit rate (HR), false alarm rate (FAR), threat score or critical success index (CSI), and true skill statistics (TSS). These parameters were evaluated using the contingency table shown in Table 3, and are presented in Table 4. The parameters were defined in terms of "Hit", "Miss", "False Alarm" and "Correct Rejection". The terms "Hit" and "Correct Rejection" were possible cases when the prediction was accurate, and "False Alarm" and "Miss" were possible cases when the prediction was not accurate. The classification accuracy of the forecasting model was assumed to be good if "Hit" and "Correct Rejection" cases were predominant, with very low cases of "False Alarm" and "Miss".

Table 3. Contingency table.

		Observed/Actual	
		Yes	**No**
Forecast/ Predicted	**Yes**	Hit (a)	False Alarm (b)
	No	Miss (c)	Correct Rejection (d)

Evaluation matrices for ML regression models: The prediction performance of regression models has been evaluated in terms of R^2 Score, also known as "coefficient of determination", MedAE, MAPE, RMSE, and MAE. Table 4 provides the performance measurement indicators utilized to validate the proposed prediction model.

Table 4. Performance measurement indicators for regression prediction model validation and verification of high concentration.

Parameter	Significance	Mathematical Representation
HR	Used to measure accurate forecasts of events. Its value ranges between 0 and 1. A value close to 1 indicates excellent prediction performance.	$\mathrm{HR} = \frac{a}{a+c}$
FAR	Measures the ratio of false alarms and gives an indication of the occurence of an event when there is no event. Its value ranges between 0 and 1; a value close to 0 indicates better prediction.	$FAR = \frac{b}{b+d}$
CSI	Together takes into account hits, misses and false alarms. Its value ranges between 0 and 1. A value close to 1 indicates excellent prediction performance	$\mathrm{CSI} = \frac{a}{a+b+c}$
TSS	Determines the ability of the model to distinguish between "Yes" and "No" cases. Its value ranges between −1 and 1, with 1 indicating a perfect forecast, 0 defining a standard forecast, and a negative value indicating a below-standard forecast.	$\mathrm{TSS} = \frac{a}{a+c} - \frac{b}{b-d} = \mathrm{HR} - \mathrm{FAR}$
R^2 Score/Coefficient of determination	Provides a degree of discrepancy in dependent variables.	$R^2(A, \hat{A}) = 1 - \frac{\sum_{j=1}^{s}(A_j - \hat{A}_j)^2}{\sum_{j=1}^{s}(A_j - \overline{A_j})^2}$
MedAE	Provides the median value of the absolute difference between forecasted and true values. The MedAE is least influenced by outliers.	$\mathrm{MedAE}(A, \hat{A}) = \mathrm{median}(\lvert A_1 - \hat{A}_1\rvert, \ldots, \lvert A_s - \hat{A}_s\rvert)$
MAPE	Predicts the accuracy of a regression model. It defines the relative percentage error of the predicted value against the true value.	$\mathrm{MAPE}(A, \hat{A}) = \frac{1}{s}\sum_{j=1}^{s}\frac{\lvert A_j - \hat{A}_j\rvert}{\max(\varnothing, \lvert A_j\rvert)}$
RMSE	Used to evaluate the standard deviation of predicted errors.	$\mathrm{RMSE}(A, \hat{A}) = \sqrt{\frac{\sum_{j=1}^{s}(A_j - \hat{A}_j)^2}{s}}$
MAE	Evaluates the mean of absolute values of the difference between the predicted value and true value.	$\mathrm{MAE}(A, \hat{A}) = \frac{1}{s}\sum_{j=1}^{s}\left\lvert A_j - \hat{A}_j\right\rvert$

Where A_j and $\hat{A}_j$ are the true and predicted values of the dependent variable for jth sample, respectively. s is the total number of samples, $\varnothing$ is a very small positive number to define the result if $A_j = 0$ and $\overline{A_j}$ can be given as $\frac{1}{s}\sum_{j=1}^{s} A_j$.

5. Analysis of Results

In this section, we report a comprehensive analysis of the results, which comprises Sections 5.1–5.4. Section 5.1 presents the correlation of $PM_{2.5}$ concentration with air pollutants and meteorological parameters. Seasonal variation in $PM_{2.5}$ concentration, along with other parameters, is presented in Section 5.2. $PM_{2.5}$ concentration prediction using machine learning based regression models is reported in Section 5.3, and Section 5.4 presents an evaluation of the results for verification of higher concentration.

5.1. Correlation of $PM_{2.5}$ Concentration with Air Pollutants and Meteorological Parameters

To statistically explore the relation between the concentration of $PM_{2.5}$ with concentration of air pollutants and meteorological parameters, the correlation coefficients were calculated using Pearson's correlation method, and are presented in Table 5. The Pearson's correlation coefficient for two variables x and y can be evaluated as:

$$\text{Pearson}'\text{s Correlation Coefficient} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} \quad (4)$$

where x_i and y_i are the values of ith sample of variables x and y, respectively. $\overline{x}$ and $\overline{y}$ are the mean values of the variables x and y, respectively.

Table 5. Correlation of $PM_{2.5}$ with air pollutants and meteorological parameters.

Parameter	Pearson's Correlation Coefficient	Parameter	Pearson's Correlation Coefficient
NO	0.26	C_7H_8	−0.01
NO_2	0.54	Temperature	0.05
NO_x	0.44	RH	−0.25
NH_3	0.33	WS	−0.49
SO_2	0.16	SR	−0.16
CO	0.60	BP	0.42
C_6H_6	0.25		

Positive and negative correlations between the concentration of $PM_{2.5}$ with the concentration of air pollutants and meteorological parameters were observed.

The air pollutants CO, NO_2 and NO_x formed a strong correlation with $PM_{2.5}$ concentration. Among the meteorological parameters, WS, BP, and RH attained significant correlation with $PM_{2.5}$ concentration. The air pollutant C_7H_8 showed a very low correlation value with $PM_{2.5}$ concentration, which may have occurred due to fewer C_7H_8 emission sources in the city. An increment or decrement in the concentration of air pollutant stipulated a direct impact on the increment or decrement of $PM_{2.5}$ concentration. All the meteorological parameters except BP and temperature exhibited a negative correlation with $PM_{2.5}$ concentration. The meteorological parameter temperature marked a very low value of Pearson's coefficient, which was probably due to insignificant variation in mean temperature values during the major seasons. The city observed an approximate mean temperature of 29 °C during the major seasons. It was noticed that WS reported a strong negative correlation with $PM_{2.5}$ concentration.

5.2. Seasonal and Annual Behaviour of $PM_{2.5}$ Concentration with Air Pollutants and Meteorological Parameters

An in-depth analysis of seasonal and annual behavior of $PM_{2.5}$ concentration is presented in Sections 5.2.1 and 5.2.2.

5.2.1. Seasonal Behavior of $PM_{2.5}$ Concentration with Air Pollutants and Meteorological Parameters

For assessing the seasonal impact of air pollutants and meteorological parameters on $PM_{2.5}$ concentration, a detailed statistical analysis is presented in Table 6. It was comprehended that the air pollutant concentration reported significant seasonal variations. A rise in $PM_{2.5}$ concentration during December–February, and decrease from March–August, was observed for the study period. The seasonal variations in $PM_{2.5}$ and air pollutant concentration were primarily due to the varying speed and direction of the wind, seasonal variation in SR, and lack of precipitation in winter which reduces surface vertical mixing and can lead to limited dilution and dispersion [37].

It was found that the air pollutant and $PM_{2.5}$ concentration exhibit maximum variation in the winter season. The maximum value of $PM_{2.5}$ concentration recorded during the winter season was 202.52 $\mu g/m^3$, which was 106.62 $\mu g/m^3$ and 130.68 $\mu g/m^3$ higher than the maximum observed $PM_{2.5}$ concentration during summer and monsoon season, respectively. The monsoon season observed marginally greater mean and median values of $PM_{2.5}$ concentration (34.39 $\mu g/m^3$ and 33.61 $\mu g/m^3$, respectively) in contrast to the summer season. However, the summer season observed significant $PM_{2.5}$ concentration variation (24 h) in contrast with the monsoon season, and observed higher minimum and maximum $PM_{2.5}$ concentration. The summer season observed a mean $PM_{2.5}$ concentration of 33.64 $\mu g/m^3$ (standard deviation = 14.21 $\mu g/m^3$) with minimum and maximum $PM_{2.5}$

concentration of 9.09 μg/m^3 and 95.90 μg/m^3, respectively. Due to precipitation in the monsoon season, the air pollutant level dipped and led to a low concentration of air pollutants and $PM_{2.5}$ concentration in the air.

Table 6. Seasonal behavior of $PM_{2.5}$ concentration, air pollutants and meteorological parameters during summer season (sum.), winter season (win.) and monsoon season (mon.).

Parameter	**$PM_{2.5}$ (μg/m^3)**			**CO (mg/m^3)**			**NH_3 (μg/m^3)**			**NO_2 (μg/m^3)**		
	sum.	**win.**	**mon.**	**sum.**	**win.**	**mon.**	**sum.**	**win.**	**mon.**	**sum.**	**win.**	**mon.**
Mean Value	33.64	**69.84**	34.49	0.67	0.90	0.75	10.37	12.83	8.75	35.08	**42.85**	35.54
Std. Deviation	14.21	34.50	12.36	0.19	0.29	0.25	6.35	6.70	4.48	11.55	20.87	10.51
Median Value	31.24	**67.15**	33.61	0.67	0.84	0.72	9.44	12.54	8.33	33.83	38.59	37.65
Min. Value	9.09	3.06	6.19	0.27	0.41	0.15	1.83	0.16	0.70	5.39	0.56	9.23
Max. Value	95.90	202.52	71.84	1.59	2.18	1.66	42.93	63.62	40.08	72.02	131.53	58.19
Parameter	**C_6H_6 (μg/m^3)**			**NO_x (ppb)**			**SO_2 (μg/m^3)**			**NO (μg/m^3)**		
	sum.	**win.**	**mon.**	**sum.**	**win.**	**mon.**	**sum.**	**win.**	**mon.**	**sum.**	**win.**	**mon.**
Mean Value	3.92	4.49	5.16	26.62	35.17	30.90	11.77	12.90	11.04	9.92	15.43	14.83
Std. Deviation	1.65	1.77	1.58	8.00	21.51	11.21	6.84	8.14	6.86	4.67	15.04	10.15
Median Value	3.87	4.38	5.06	25.71	30.17	30.55	10.69	10.32	10.66	9.37	11.18	12.69
Min. Value	0.86	0.01	0.51	6.44	0.31	7.56	1.97	1.84	1.32	1.22	0.45	1.48
Max. Value	10.41	11.69	9.90	52.37	129.9	92.99	40.47	60.83	46.36	27.86	99.32	93.45
Parameter	**C_7H_8 (μg/m^3)**			**WS (m/s)**			**WD (degree)**			**RH (%)**		
	sum.	**win.**	**mon.**	**sum.**	**win.**	**mon.**	**sum.**	**win.**	**mon.**	**sum.**	**win.**	**mon.**
Mean Value	8.38	9.33	13.91	2.79	1.85	2.45	239.96	180.80	231.95	71.89	70.00	74.68
Std. Deviation	4.40	4.66	5.71	0.73	0.66	0.82	46.43	39.28	41.00	4.55	6.18	4.48
Median Value	7.78	8.53	13.03	2.82	1.78	2.41	223.95	177.03	226.91	72.85	70.81	74.51
Min. Value	1.28	0.01	2.84	0.72	0.47	0.55	84.74	79.45	116.00	48.18	43.77	63.06
Max. Value	26.93	34.76	54.29	5.56	4.97	5.28	351.46	313.86	315.59	82.62	86.56	85.23
Parameter	**SR (W/m^2)**			**Temperature (°C)**			**BP (mmHg)**		
	sum.	**win.**	**mon.**	**sum.**	**win.**	**mon.**	**sum.**	**win.**	**mon.**
Mean Value	167.18	120.51	124.12	29.16	29.75	28.59	745.58	751.36	730.54
Std. Deviation	56.02	38.33	54.04	1.08	2.61	1.31	6.51	12.44	13.52
Median Value	172.88	122.86	129.13	29.07	29.44	28.44	746.82	754.14	734.16
Min. Value	6.50	6.00	6.00	24.50	23.36	24.21	730.29	701.00	701.33
Max. Value	520.41	409.73	257.24	34.18	42.68	36.93	757.18	766.58	753.38

During the winter season, a higher air pollutant concentration was noticed. The maximum values observed for air pollutants CO, NO_2 and C_6H_6 were 2.18 mg/m^3, 131.53 μg/m^3 and 11.69 μg/m^3, respectively. Due to an increase in the concentration of CO and oxides of nitrogen, a rise in $PM_{2.5}$ concentration was observed. The rise in air pollutant concentration was probably due to the slow WS and temperature inversion effect in the winter season. As a result, the air pollutants and $PM_{2.5}$ particles were trapped near the earth's surface which, in turn, increased the $PM_{2.5}$ concentration [38,39].

A substantial variation in $PM_{2.5}$ concentration (24 h) was observed during the summer season, and a variation from 9.09 μg/m^3 to 95.90 μg/m^3 in $PM_{2.5}$ concentration was reported during the period of observation. The season reported high SR (mean value = 167.18 W/m^2; standard deviation = 56.02 W/m^2) and high WS (mean value = 2.79 m/s; standard deviation = 0.73 m/s). Because of the high SR acquired by the earth's near-surface

atmosphere, the near-surface temperature increased, promoting upward movement, eventually diffusing $PM_{2.5}$ concentration. In addition to higher SR, the high WS and high precipitation diluted the air pollutant concentration at the surface and caused a significant decrease in air pollutants and $PM_{2.5}$ concentration during the summer season [38,39].

It was noted that the air pollutant SO_2 concentration showed inadequate seasonal variability and remained approximately constant throughout the year. This was possibly due to SO_2 emission sources (sulphur-containing fuels such as oil, coal and diesel) which constantly emit SO_2 pollutants in the city. The primary sources of CO, SO_2 and nitrate aerosols for the city are presumed to be power generation plants, engines in vehicles and ships, ship-yard industries and steel plants. The key source of C_6H_6 emission in the city is probably due to the presence of heavy chemical and petroleum industries, as C_6H_6 is a natural element of petrol and crude oil and is produced as a by-product during the oil refining process. In addition to major industries, the city is surrounded by many small and medium-size industries which add to the concentration of air pollutants. Though heavy, medium and small industries contribute to increasing air pollution, the city has the advantage of sea breezes, by which most of the air pollutant emissions are disseminated to sea and the impact of air pollutants on air quality is reduced. However, a relatively constant NH_3 concentration was observed during the assessment period, which was probably due to the continuous emission of ammonia gases from industrial processes and vehicular emissions.

It was found that the city experiences greater humidity during the monsoon season, i.e., from 63.06% to 85.23%, with a mean humidity of 74.68%, whereas a mean humidity of 70% and 71.89% was observed by the city in the winter and summer seasons, respectively. As noted from Table 4, the metrological parameter 'humidity' exhibits a negative correlation with $PM_{2.5}$ concentration. In the highly humid season, raindrops influence gaseous air pollutants by the phenomenon of absorption and collision. The phenomenon leads to wet decomposition and reduces the $PM_{2.5}$ concentration [40,41].

As observed from the wind rose diagram (Figure 3) and from Table 6, slow and the infrequent wind blows during the winter season. To present a detailed analysis of WD, the wind rose diagram is plotted in sixteen directions from N to NNE (counterclockwise). The concentric circles in the wind rose diagram represent the probability percentage of wind blow, and are labeled with percentages increasing outward. As shown in Figure 3, the probability percentage concentric rings are placed at 5% intervals. For analysis, the WS is divided into nine bins and the bins are differentiated, with colors ranging from red to brown. The length of spoke around the circle is related to the frequency of time that the wind blows from a particular direction. The dominant wind directions during the winter season were found to be in the SSE and S directions, with a small secondary lobe in the SSW direction, and with minor lobes in the SW and SE directions, indicating less frequent wind blow in these minor lobe directions. The winter season observed a mean WS of 1.85 m/s, and approximately 65% of the time the wind blew in the direction of SSW to SSE (230–305 degrees). During the season, only 1–2% of infrequent high-speed winds (greater than 4.47 m/s) were observed. Approximately 20–25% of the time the wind blew at a speed of 2.47 m/s to 2.97 m/s. During the remaining time, the wind blew at a speed of less than 1.47 m/s. The summer and monsoon seasons observed frequent and high-speed winds during the period of observation, and the wind blew mainly in a southwest direction (primarily in direction of 180–250 degrees). During the monsoon season, approximately 40–45% of the time the wind blew with a speed greater than 3.18 m/s, and approximately 8–10% of the time the wind blew with a speed of 4.23 m/s to 4.75 m/s. However, infrequent WS greater than 4.75 m/s blew for approximately 1–2% of the entire season. Similar high-speed winds were observed for the summer season, with a SSW prominent wind direction. For the summer and monsoon seasons, the high-speed winds swept the air pollutants away from the city, hence, a low concentration of $PM_{2.5}$ was observed during these months.

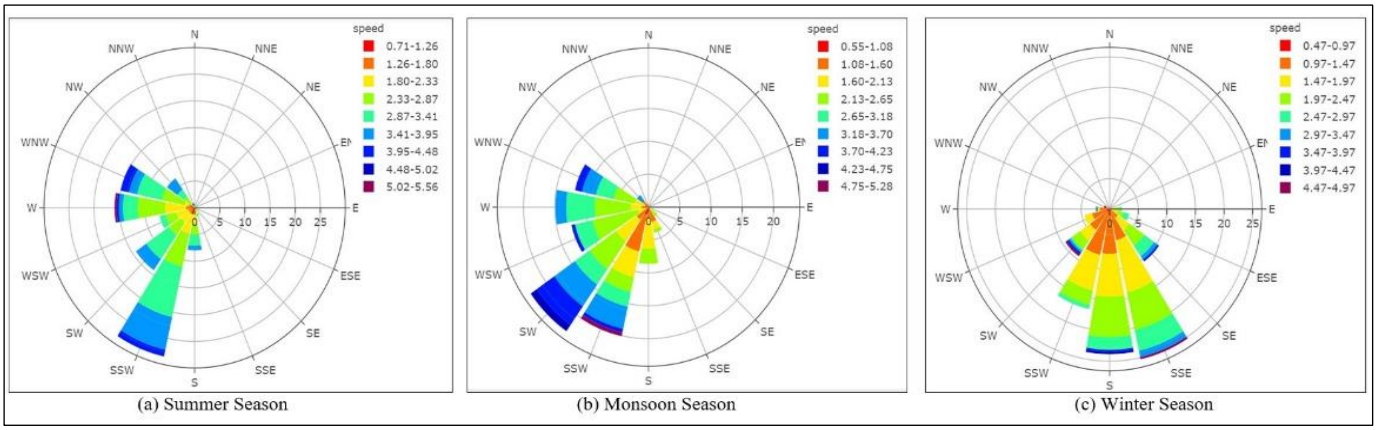


Figure 3. Wind rose diagram for summer, monsoon and winter seasons.

5.2.2. Annual Behavior of $PM_{2.5}$ Concentration, Air Pollutants and Meteorological Parameters

Annual variation in $PM_{2.5}$ concentration, air pollutant and meteorological parameters for 2018 and 2019 are presented in Table 7. As observed, a minor increase in $PM_{2.5}$ concentration was observed for 2018, and mean $PM_{2.5}$ concentration values of 49.97 $\mu g/m^3$ and 47.32 $\mu g/m^3$ were noted for 2018 and 2019, respectively. A substantial variation in CO and NO concentration was observed for 2018 and 2019. A rise of 0.15 mg/m^3 in mean value of CO emission was observed during 2019. However, the maximum value of CO emission (2.18 mg/m^3) was noted in 2018. An inconsiderable variation in the mean concentrations of NO_2, C_6H_6, and NO_x was observed during the entire period.

Table 7. Statistical description of $PM_{2.5}$ concentrations, air pollutants and meteorological parameters for the assessment year 2018–2019.

Parameter	**$PM_{2.5}$ ($\mu g/m^3$)**		**CO (mg/m^3)**		**NH_3 ($\mu g/m^3$)**		**NO_2 ($\mu g/m^3$)**	
Year	**2018**	**2019**	**2018**	**2019**	**2018**	**2019**	**2018**	**2019**
Mean	49.97	47.32	0.71	0.86	11.94	10.01	38.88	37.86
Std.	28.53	31.45	0.30	0.22	5.55	6.86	15.53	16.94
Median	43.88	37.15	0.62	0.83	11.07	8.50	37.44	36.04
Min.	5.04	3.06	0.15	0.38	0.16	0.70	0.56	5.39
Max.	202.52	201.85	2.18	1.75	63.63	42.93	93.54	131.53
Parameter	**C_6H_6 ($\mu g/m^3$)**		**NO_x (ppb)**		**SO_2 ($\mu g/m^3$)**		**NO ($\mu g/m^3$)**	
Year	**2018**	**2019**	**2018**	**2019**	**2018**	**2019**	**2018**	**2019**
Mean	4.81	4.12	30.72	31.63	11.13	12.96	12.60	14.18
Std.	1.41	1.97	16.04	15.92	6.68	7.99	11.42	11.48
Median	4.63	4.08	27.84	29.22	10.58	10.74	10.16	11.57
Min.	0.01	0.51	0.31	6.44	1.32	1.97	0.45	1.95
Max.	11.69	10.41	129.9	105.43	40.47	60.83	99.32	93.45

Parameter	**WS (m/s)**		**WD (degree)**		**RH (%)**	
Year	**2018**	**2019**	**2018**	**2019**	**2018**	**2019**
Mean	2.24	2.40	234.25	193.71	70.70	72.91
Std.	0.83	0.82	58.95	28.63	6.01	4.87
Median	2.19	2.34	244.56	199.72	71.99	73.07
Min.	0.47	0.55	79.45	84.74	43.77	57.72
Max.	4.97	5.56	351.46	263.23	85.23	86.56

Table 7. *Cont.*

Parameter	SR (W/m^2)		Temperature (°C)		BP (mmHg)	
Year	**2018**	**2019**	**2018**	**2019**	**2018**	**2019**
Mean	143.01	131.87	29.12	29.39	745.87	742.54
Std.	57.60	48.54	1.52	2.30	9.77	16.73
Median	141.97	131.34	29.14	28.95	748.18	747.07
Min.	6	6	23.36	24.50	701	701.33
Max.	520.41	261.92	38.33	42.68	766.58	765.57

A minor variation of 1.58 μg/m^3 to 1.83 μg/m^3 in the mean concentration of SO_2 and NO was noted between 2018 and2019. As observed, the period under study observed approximately continuous mean temperature. During the period of observation, the city observed a mean temperature of approximately 29°C with approximately the same median temperature in 2018–2019. In contrast to 2018, slightly stronger winds (mean speed of 2.40 m/s) blew during 2019, and the city observed maximal WS during the months of April–July. During the period of observation, the wind blew in the mean direction of 234.25 degrees to 193.71 degrees, and a change in wind direction was noticed in the months of November–January. A small variation of 3.33 mmHg in mean BP was observed for 2018–2019, and approximately the same values of minimum and maximum BP were noted for the period under study. However, variations in BP were noted for 2019. A high SR was observed in 2018 compared with 2019, and a mean SR of 143.01 W/m^2 in 2018 was noted in contrast to 131.87 W/m^2 for 2019. Higher peak values of SR were noted for 2018 in contrast to 2019. Due to the cumulative effect of variation in air pollutant concentration and meteorological parameters, continuous variable $PM_{2.5}$ concentration was observed for the period 2018–2019.

5.3. Machine Learning-Based $PM_{2.5}$ Concentration Estimation

In the present study, the performance of the machine learning based regression models, employed to estimate $PM_{2.5}$ concentration in the air, were validated using data related to eight air pollutants and six meteorological parameters collected for the year 2018–2019 for Visakhapatnam. The air pollutants and meteorological parameters were utilized as independent input variables to train 15 distinct machine learning regression models to predict $PM_{2.5}$ concentration. The model parameters were tuned using a grid search optimization technique. The dataset was divided into training and test datasets with a ratio of 80–20%, namely, 80% of observations were used to train the model and 20% of observations were sed to test the model. In our proposed methodology, the dataset was randomly categorized into a training dataset and a test dataset. The experiments were simulated using Python 3.8 open-source software on an IBM PC with Intel Core i-7–6700 CPU @ 3.40 GHz processor supported with 8 GB RAM. Table 8 presents the performance matrices of 15 regression models for the prediction of $PM_{2.5}$ concentration. It was observed that the VR and gradient boosting regression models (CB, LGBM and XGB) showed notable performance, in contrast to other presented regression models.

Table 8. $PM_{2.5}$ concentration prediction performance analysis using various regression models.

Sr. No	Regression Model	R^2 Score	MedAE (μg/m^3)	MAPE	RMSE (μg/m^3)	MAE (μg/m^3)
1	VR (GB + RF + LR)	0.73	7.98	0.31	13.74	10.23
2	CB	0.81	6.95	0.29	11.42	9.07
3	LGBM	0.76	8.05	0.29	12.94	9.85
4	XGB	0.71	7.8	0.30	14.03	10.34

Table 8. *Cont.*

Sr. No	Regression Model	R^2 Score	MedAE ($\mu g/m^3$)	MAPE	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)
5	LR-LA	0.57	8.75	0.43	17.26	13.10
6	RF	0.52	9.90	0.49	18.22	14.19
7	MLP	0.69	8.96	0.37	14.65	11.12
8	LAR	0.57	8.71	0.42	17.24	13.09
9	PLS	0.57	10.25	0.43	17.31	13.07
10	QR	0.47	10.33	0.46	19.18	14.21
11	MTE	0.58	8.35	0.43	17.09	12.86
12	RR	0.56	8.61	0.43	17.36	13.18
13	BRR	0.57	8.53	0.43	17.26	13.02
14	KNN	0.50	9.75	0.43	18.58	13.37
15	LR	0.56	8.65	0.43	17.38	13.20

5.3.1. Performance of Regression Models

From Table 8, it was noted that the gradient boosting and VR models achieved a notable R^2 score (0.71 to 0.81). The higher R^2 score signified that the dataset values were well fitted in the model. Furthermore, the gradient boosting and VR prediction models achieved low error scores in terms of RMSE (11.42 $\mu g/m^3$ to 14.03 $\mu g/m^3$) and MAE (9.09 $\mu g/m^3$ to 10.34 $\mu g/m^3$). The CB prediction model outperformed RMSE, MAE, MAPE and MedAE, in terms of R^2 score. The model yielded an R^2 score of 0.81, RMSE of 11.42 $\mu g/m^3$, MAPE of 0.29, and an MAE of 9.07 $\mu g/m^3$. The prediction performance of the LGBM model was found to have deteriorated in contrast to the CB which predicted the results with an R^2 score of 0.76, RMSE of 12.94 $\mu g/m^3$, MAPE of 0.29 and MAE of 9.85 $\mu g/m^3$. Comparable accuracies were observed for VR and XGB models. However, the models showed lower prediction performance against CB and LGBM regression models.

As observed, in addition to high R^2, the CB regression predicted the $PM_{2.5}$ concentration with minimum error amongst all the presented models. It was found that the CB model attained minimum MedAE (6.95 $\mu g/m^3$) and MAPE (0.29) errors, revealing that the model was robust to the outliers presented in the dataset. Very low performance was observed for QR and KNN regression models. Low-performance scores in terms of R^2 and high errors attained by these models indicated their inefficacy to fit the dataset values for predicting $PM_{2.5}$ concentration. However, the RF model showed slightly improved performance in comparison to the KNN model. Least prediction performance was observed for the QR model, with low R^2 scores (0.47) and unfavorable high error in terms of RMSE (19.18 $\mu g/m^3$), MAE (14.21 $\mu g/m^3$), MedAE (10.33 $\mu g/m^3$) and MAPE (0.46).

The MLP regression model showed marked performance degradation for predicting the $PM_{2.5}$ concentration. MLP predicted the results with an R^2 score of 0.69, RMSE of 14.65 $\mu g/m^3$, MAE of 11.12 $\mu g/m^3$ and median absolute error of 8.96 $\mu g/m^3$. All the penalized regularization regression models (LA-LSR, LAR, MTE, RR and BRR) and the LR model attained comparable prediction performance, with an approximate R^2 score of 0.57, and RMSE and MAE in the proximity of 17 $\mu g/m^3$ and 13 $\mu g/m^3$, respectively. The models attained a median absolute error of approximately 8.75 $\mu g/m^3$.

It was observed that the penalized–regularization model showed enhanced prediction performance in contrast to MLP, and diminished performance in comparison to voting and gradient boosting models. Comparative prediction performance with slightly increased MedAE was noted for the PLS prediction model.

Figures 4 and 5 present the regression plots and residual error plots for the test dataset. Figure 4 shows the regression plots mapped between observed and predicted $PM_{2.5}$ concentrations. As observed from Figure 4, compared with other models, the data

points for CB model were highly concentrated on the 'fitting line' in the regression curve, indicating that the values were well fitted in the model. A low concentration of data points was observed on the 'fitting line' for LGBM, XGB and VR regression models, indicating that the values were not well fitted to the models.

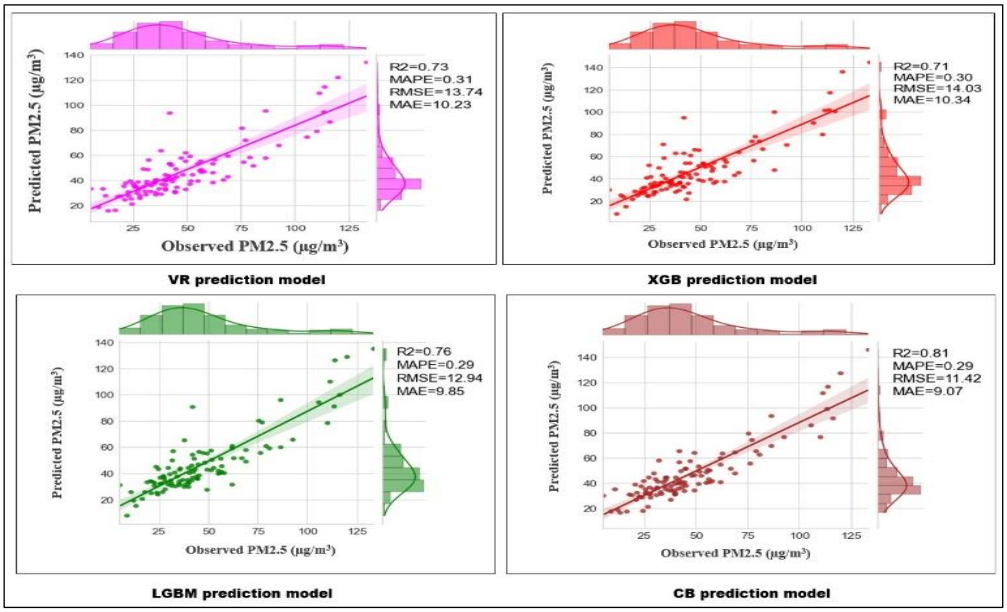


Figure 4. Regression plots of test dataset using VR, XGB, LGBM and CB prediction models.

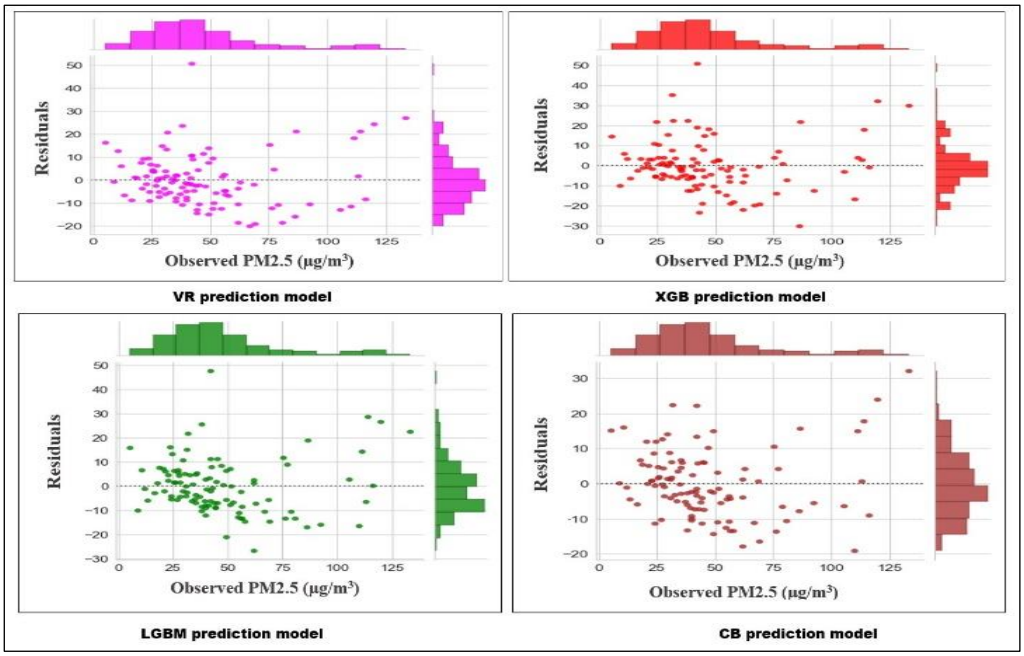
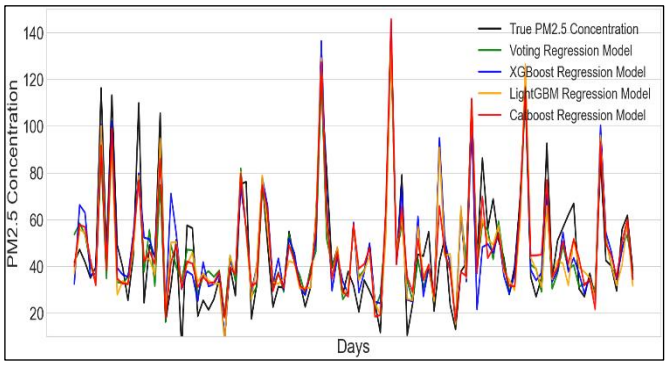


Figure 5. Residual plots of test dataset using VR, XGB, LGBM and CB prediction models.

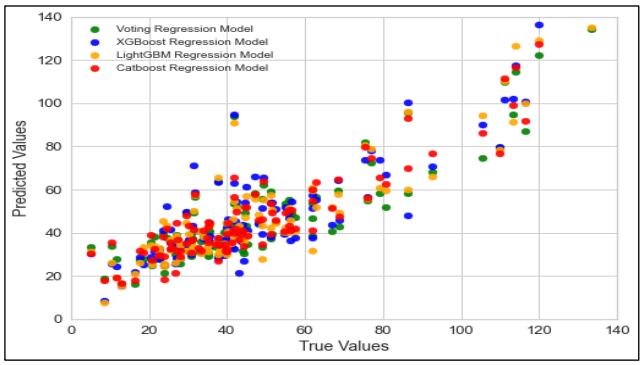
Figure 5 presents the residual error plots with their histograms showing residuals of the regression models evaluated for the test dataset. The results report that for XGB and VR regression models, the residuals and histogram peaks reside at around 0 to −20, yielding negatively biased results. It indicates that the model's prediction was too high, and that the models probably predict higher $PM_{2.5}$ concentration than compared with observed $PM_{2.5}$ concentration. The LGBM model shows satisfactory improvement having a random and dispersed distribution of residual. However, the model observes low

negatively biased residual error, and the results turn out to be slightly negative, biased with a moderate difference between predicted and observed $PM_{2.5}$ concentrations. Amongst all the fifteen implemented models, the CB model shows significant prediction performance with minimum residual error. As observed from the residual error plot Figure 5, the model is least biased to positive and negative residuals and shows random residual distribution. The residual error in CB lies in the range of −20 to 30, with maximum residual present in the range of −10 to 10.

Time series and scatter plot between true and predicted $PM_{2.5}$ concentrations for the test datasets are presented in Figure 6. As marked in Figure 6a, the models show adequate prediction performance and, as compared with other models, the CB model relatively follows the true $PM_{2.5}$ concentration. The scatter plot presented in Figure 6b shows that in contrast to VR, XGB and LGBM models, the CB model predicts $PM_{2.5}$ concentration in the proximity of observed/true $PM_{2.5}$ concentration at maximum instants of time. For the CB regression model, Figures 5 and 6 show fine agreement between the observed/true and predicted $PM_{2.5}$ concentrations.



(**a**)



(**b**)

Figure 6. (**a**) Time series plot of true and predicted $PM_{2.5}$ concentration for the test dataset using different regression models. (**b**) Scatter plot between true and predicted $PM_{2.5}$ concentration on the test dataset using VR, XGB, LGBM and CB models.

5.3.2. Impact of Input Variables (Air Pollutant Concentration and Meteorological Parameters) on the CB Model

To interpret the CB model, the impact of feature variables (air pollutants and meteorological parameters) on $PM_{2.5}$ concentrations were evaluated, and are presented in Figure 7. The influence of feature variables was measured using Shapley additive explanations (SHAP) values [42].

The SHAP framework defines the prediction in terms of a linear combination of binary variables that are used to describe whether an input feature is present in the model or not. The framework defines results in terms of Shapley values. Shapley (SHAP) values define the feature importance and impact of features on the prediction model by considering three required properties: (a) local accuracy, (b) missingness, and (c) consistency [42]. The

y-axis points to the input variables indicating their impact on the model. The input variables on the *y*-axis are arranged according to their importance. The values on the *x*-axis indicate SHAP values, and points on the plot indicate Shapley values of input variables for the instances. The color gradient (blue to red) indicates variable importance from low to high. The higher the SHAP value, the higher is the variable's impact on the model. As shown in Figure 7, variables NH_3, BP, CO and NO_2 significantly influenced the predicted $PM_{2.5}$ concentration with positive correlation, i.e., the predicted value increased with the high feature values of NH_3, BP, CO and NO_2, and conversely, predicted value decreased with the lower feature values of NH_3, BP, CO and NO_2.

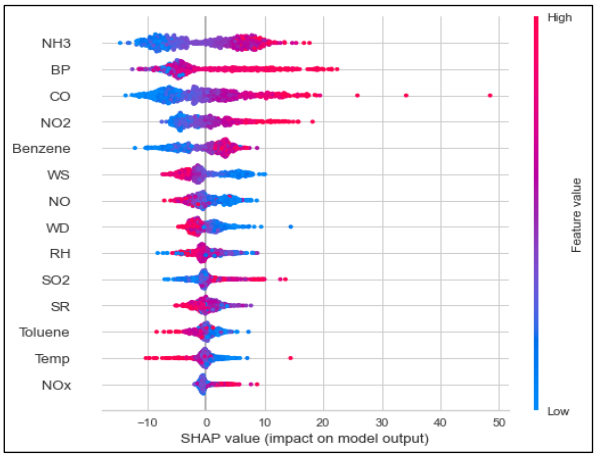


Figure 7. Impact of air pollutants and meteorological parameters on the CB regression model.

However, the variables SO_2, NO_x, and C_6H_6 also positively influenced the predicted results but had less impact on the prediction result. The variables WS, WD, SR, temperature, C_7H_8 and RH influenced the predicted $PM_{2.5}$ concentration with negative correlation, i.e., higher values of WS, WD, SR, temperature, C_7H_8 and RH tended to decrease the predicted value, and vice versa.

5.4. Evaluation of High $PM_{2.5}$ Concentration

During the period of observation, the city was exposed in 2019 to a higher value of $PM_{2.5}$ concentration compared with 2018. Among the three dominant seasons of the city, the winter season observed a maximum number of days with higher $PM_{2.5}$ concentration than NAAQS standards. The NAAQS standards classify a mean of 24 h $PM_{2.5}$ concentration of 60 $\mu g/m^3$ as higher concentration. However, the summer and monsoon seasons observed very few days of $PM_{2.5}$ concentration greater than NAAQS standards.

Table 9 presents the prediction results for high $PM_{2.5}$ concentration using the CB regression model for the period of observation. The high concentration prediction performance was evaluated using HR, FAR, CSI, TSS and OR measurement indices. The results for high concentration were evaluated on the test dataset. The results showed that the model achieved a high Hit rate of 0.85 for measuring accurate predictions and achieved a low score of 0.02 for false alarms.

Table 9. Evaluation of predictions for high PM2.5 concentration using CB model on the test dataset.

Variable	HR	FAR	CSI	TSS
$PM_{2.5}$	0.85	0.02	0.80	0.83

The model also achieved high CSI and TSS scores, indicating the model's excellent performance and its ability to correctly classify between "Yes" and "No" cases. Moreover, the CSI represented the model's sensitivity to correct forecasts of high concentration, and

the high value obtained by the CSI indicated that the high concentration cases of $PM_{2.5}$ were generally predicted correctly.

It was found that, for the period under observation, the winter season encountered 166 days with $PM_{2.5}$ concentrations greater than the standards set by NAAQS. Approximately 55–60% of days during the winter season witnessed higher $PM_{2.5}$ concentration than the prescribed NAAQS standards. In contrast, approximately 4–5% of days witnessed higher $PM_{2.5}$ concentration than the prescribed NAAQS standard in the summer and winter season. For approximately more than 90 days of the winter and monsoon seasons, the city was under the impact of higher C_6H_6 concentration than the prescribed standards.

6. Conclusions

In the present study, fifteen machine learning models were presented for analysis and prediction of $PM_{2.5}$ concentration based on time series data. This study provided detailed insight into the air pollutants and meteorological parameters contributing to $PM_{2.5}$. We targeted the statistical behavior of 24 h average air pollutant and $PM_{2.5}$ concentrations along with meteorological parameters observed at the eastern coastal city of Visakhapatnam, India. Using Pearson's correlation coefficient, the correlation of $PM_{2.5}$ with air pollutants and meteorological parameters was determined. Seasonal behavior of air pollutants, $PM_{2.5}$ concentration and metrological parameters were studied by extracting significant information from raw data collected from the Central Pollution Control Board and APPCB. It was deduced that the summer and monsoon seasons showed lower $PM_{2.5}$ and air pollutant concentrations, compared with the winter season. The results revealed that the CB machine learning model is an efficient predictive model. For comparative analysis with other prediction models, we used R^2 score, RMSE, MAE, and MedAE as MAPE performance parameters. The performance of the CB model was not only better than traditional models, such as the linear regression model and MLP, but also better than voting and other boosting models such as XGBoost and LightGBM. The prediction of $PM_{2.5}$ concentration is a challenging task, owing to changing metrological and pollutant concentration, yet an amount of credibility in prediction results can be attained using copious data. The present study was carried out for a small period and notable results were achieved, however, significant improvement in forecasting results can be expected by examining the information over a greater period and, in addition, considering nearby geographical locations.

Author Contributions: Conceptualization, N.K.; Data curation, M.S., S.S., V.J. and S.K.; Formal analysis, M.S., N.K., S.S., V.J., S.M., S.K. and P.K.; Investigation, N.K., S.S., V.J., S.M. and S.K.; Methodology, M.S. and S.M.; Resources, V.J.; Supervision, P.K., Visualization, P.K.; Writing—original draft, M.S. and N.K.; Writing—review & editing, M.S., N.K., S.S., S.K. and P.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent Statement: Not Applicable.

Data Availability Statement: The dataset is publicly available at https://cpcb.nic.in (accessed on 5 October 2021) and https://pcb.ap.gov.in/UI/Home.aspx (accessed on 5 October 2021).

Acknowledgments: The authors acknowledge Ashok K. Goel, GZSCCET, MRSPTU, Bathinda, for his valuable suggestions in conducting the research and performance analysis. Additionally, Shallu Sharma and Sumit Kumar are immensely thankful to Pravat Mandal, Manesar, and Ashok Mittal of Lovely Professional University, Punjab, for their support during the execution of this work.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. WHO. 2 May 2018. Available online: https://www.who.int/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action (accessed on 25 November 2021).
2. Central Pollution Control Board, India. Available online: https://www.cpcb.nic.in/ (accessed on 5 October 2021).
3. EEA. Healthy Environment, Healthy Lives. 2019. Available online: https://www.eea.europa.eu/ (accessed on 26 November 2021).
4. EEA. Air Pollution: How It Affects Our Health. 2021. Available online: https://www.eea.europa.eu/ (accessed on 26 November 2021).
5. Masood, A.; Ahmad, K. A model for particulate matter ($PM_{2.5}$) prediction for Delhi based on machine learning approaches. *Procedia Comput. Sci.* **2020**, *167*, 2101–2110. [CrossRef]
6. Deters, J.K.; Zalakeviciute, R.; Gonzalez, M.; Rybarczyk, Y. Modeling $PM_{2.5}$ Urban Pollution Using Machine Learning and Selected Meteorological Parameters. *J. Electr. Comput. Eng.* **2017**, *2017*, 5106045. [CrossRef]
7. Moisan, S.; Herrera, R.; Clements, A. A dynamic multiple equation approach for forecasting $PM_{2.5}$ pollution in Santiago, Chile. *Int. J. Forecast.* **2018**, *34*, 566–581. [CrossRef]
8. Jiang, X.; Luo, Y.; Zhang, B. Prediction of $PM_{2.5}$ Concentration Based on the LSTM-TSLightGBM Variable Weight Combination Model. *Atmosphere* **2021**, *12*, 1211. [CrossRef]
9. Suleiman, A.; Tight, M.; Quinn, A. Applying machine learning methods in managing urban concentrations of traffic related particulate matter (PM_{10} and $PM_{2.5}$). *Atmos. Pollut. Res.* **2019**, *10*, 134–144. [CrossRef]
10. Kumar, S.; Mishra, S.; Singh, S.K. A machine learning-based model to estimate $PM_{2.5}$ concentration levels in Delhi's atmosphere. *Heliyon* **2020**, *6*, e05618. [CrossRef]
11. Chandu, K.; Dasari, M. Variation in Concentrations of $PM_{2.5}$ and PM_{10} during the Four Seasons at the Port City of Visakhapatnam, Andhra Pradesh, India. *Nat. Environ. Pollut. Technol.* **2020**, *19*, 1187–1193. [CrossRef]
12. Ferov, M.; Modry, M. Enhancing lambdaMART using oblivious trees. *arXiv* **2016**, arXiv:1609.05610.
13. Rao, K.S.; Devi, G.L.; Ramesh, N. Air Quality Prediction in Visakhapatnam with LSTM based Recurrent Neural Networks. *Int. J. Intell. Syst. Appl.* **2019**, *11*, 18–24. [CrossRef]
14. Prasad, N.K.; Sarma, M.; Sasikala, P.; Raju, N.M.; Madhavi, N. Regression Model to Analyse Air Pollutants Over a Coastal Industrial Station Visakhapatnam (India). *Int. J. Data Sci.* **2020**, *1*, 107–113. [CrossRef]
15. Devi Golagani, L.; Rao Kurapati, S. Modelling and Predicting Air Quality in Visakhapatnam using Amplified Recurrent Neural Networks. In Proceedings of the International Conference on Time Series and Forecasting, Universidad de Granada, Granada, Spain, 25–27 September 2019; Volume 1, pp. 472–482.
16. Andhra Pradesh Pollution Control Board. Available online: https://pcb.ap.gov.in (accessed on 5 October 2021).
17. Census-India (2012) Census of India. The Government of India, New Delhi. 2011. Available online: https://censusindia.gov.in/census.website/ (accessed on 28 November 2021).
18. Indian Meteorological Department. Government of India. Available online: https://mausam.imd.gov.in/ (accessed on 23 October 2021).
19. Dietterich, T.G. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 1–15.
20. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. Catboost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31, pp. 6638–6648.
21. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
22. Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
23. Avila, J.; Hauck, T. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499. [CrossRef]
24. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
25. Haykin, S. *Neural Networks and Learning Machines*; Pearson: Upper Saddle River, NJ, USA, 2009; Volume 3.
26. Tibshirani, R. Regression Analysis and Selection via the Lasso. *R. Stat. Soc. Ser.* **1996**, *58*, 267288.
27. Abdi, H. Partial least square regression (PLS regression). In *Encyclopedia of Measurement and Statistics*; Salkind, N.J., Ed.; Sage: Thousand Oaks, CA, USA, 2007.
28. Koenker, R.; Hallock, K.F. Quantile regression. *J. Econ. Perspect.* **2001**, *15*, 143–156. [CrossRef]
29. Liu, J.; Ji, S.; Ye, J. Multi-task feature learning via efficient l 2, 1-norm minimization. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence; AUAI Press: Arlington, VA, USA, 2009; pp. 339–348.
30. Kidwell, J.S.; Brown, L.H. Ridge Regression as a Technique for Analyzing Models with Multicollinearity. *J. Marriage Fam.* **1982**, *44*, 287–299. [CrossRef]
31. Bayesian Ridge Regression. Available online: https://scikit-learn.org/stable/modules/linear_model.html#bayesian-ridge-regression (accessed on 13 December 2021).
32. Fehrmann, L.; Lehtonen, A.; Kleinn, C.; Tomppo, E. Comparison of linear and mixed-effect regression models and k-nearest neighbour approach for estimation of single-tree biomass. *Can. J. For. Res.* **2008**, *38*, 1–9. [CrossRef]
33. Seber, G.A.; Lee, A.J. *Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2012; Volume 329.

34. Prokhorenkova, L.; Gusev, G.; Vorobev, A. CatBoost: Unbiased boosting with categorical features. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; pp. 6638–6648.
35. Zhang, K.; Schölkopf, B.; Muandet, K.; Wang, Z. Domain adaptation under target and conditional shift. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 819–827.
36. The Gazette of India, Part III–Section 4, NAAQS CPCB Notification. 2009. Available online: https://cpcb.nic.in/ (accessed on 17 December 2021).
37. Tai, A.P.K.; Mickley, L.J.; Jacob, D.J. Correlations between fine particulate matter ($PM_{2.5}$) and meteorological variables in the United States: Implications for the sensitivity of $PM_{2.5}$ to climate change. *Atmos. Environ.* **2010**, *44*, 3976–3984. [CrossRef]
38. Khillare, P.S.; Sarkar, S. Airborne inhalable metals in residential areas of Delhi, India: Distribution, source apportionment and health risks. *Atmos. Pollut. Res.* **2012**, *3*, 46–54. [CrossRef]
39. Xu, R.; Tang, G.; Wang, Y.; Tie, X. Analysis of a long-term measurement of air pollutants (2007–2011) in North China Plain (NCP); Impact of emission reduction during the Beijing Olympic Games. *Chemosphere* **2016**, *159*, 647–658. [CrossRef]
40. Jian, L.; Zhao, Y.; Zhu, Y.-P.; Zhang, M.-B.; Bertolatti, D. An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. *Sci. Total Environ.* **2012**, *426*, 336–345. [CrossRef]
41. Wang, J.; Ogawa, S. Effects of Meteorological Conditions on $PM_{2.5}$ Concentrations in Nagasaki, Japan. *Int. J. Environ. Res. Public Health* **2015**, *8*, 9089–9101. [CrossRef] [PubMed]
42. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774. Available online: https://arxiv.org/abs/1705.07874 (accessed on 20 December 2021).