atmosphere

MDPI

Article

Implementing Machine Learning Algorithms to Predict Particulate Matter ($PM_{2.5}$): A Case Study in the Paso del Norte Region

Suhail Mahmud [1,*], Tasannum Binte Islam Ridi [2], Mohammad Sujan Miah [3], Farhana Sarower [4] and Sanjida Elahee [5]

[1] Karen Clark and Company, Boston, MA 02116, USA
[2] Department of Economics, East West University, Aftabnagar, Dhaka 1212, Bangladesh
[3] Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA
[4] Mathematics Department, New Mexico Sate University, Las Cruces, NM 88003, USA
[5] Department of Chemistry, The Pennsylvania State University, State College, PA 16801, USA
* Correspondence: smahmud@karenclarkandco.com

Abstract: This work focuses on the prediction of an air pollutant called particulate matter ($PM_{2.5}$) across the Paso Del Norte region. Outdoor air pollution causes millions of premature deaths every year, mostly due to anthropogenic fine $PM_{2.5}$. In addition, the prediction of ground-level $PM_{2.5}$ is challenging, as it behaves randomly over time and does not follow the interannual variability. To maintain a healthy environment, it is essential to predict the $PM_{2.5}$ value with great accuracy. We used different supervised machine learning algorithms based on regression and classification to accurately predict the daily $PM_{2.5}$ values. In this study, several meteorological and atmospheric variables were retrieved from the Texas Commission of Environmental Quality's monitoring stations corresponding to 2014–2019. These variables were analyzed by six different machine learning algorithms with various evaluation metrics. The results demonstrate that ML models effectively detect the effect of other variables on $PM_{2.5}$ and can predict the data accurately, identifying potentially risky territory. With an accuracy of 92%, random forest performs the best out of all machine learning models.

Keywords: ground PM; air quality; machine learning algorithms; Paso del Norte; classification

Citation: Mahmud, S.; Ridi, T.B.I.; Miah, M.S.; Sarowar, S.; Elahee, S. Implementing Machine Learning Algorithms to Predict Particulate Matter ($PM_{2.5}$): A Case Study in the Paso del Norte Region. *Atmosphere* **2022**, *13*, 2100. https://doi.org/10.3390/atmos13122100

Academic Editors: Kangwei Li, Huan Yu, Christian George

Received: 2 November 2022
Accepted: 9 December 2022
Published: 14 December 2022

1. Introduction

Paso del Norte (PdN) is the largest metropolitan area on the border between the United States of America and Mexico, with a population estimation of 2.4 million. This region is made up of three large cities: El Paso, Texas; Las Cruces, New Mexico; and Ciudad Juarez, Mexico, all of which share the PdN airshed. Similar to any other developing metropolis, PdN is confronted with the ever-increasing issues of poor air quality. Additionally, as an international cross-border location, there is growing concern about its air quality for both the United States and Mexico [1,2].

Fine particulate matter ($PM_{2.5}$) is an air pollutant with an aerodynamic diameter of less than or equal to 2.5 µm, which becomes hazardous to people's health when the PM concentration levels in the air are above a certain standard. These small particles can absorb a variety of chemical components, including metals, salts, poisons, organic compounds, and biological groups, such as pollens [3]. $PM_{2.5}$ levels are rising as a result of automobiles, power generation, and other anthropogenic factors. Ammonium sulfate, Ammonium nitrate, and organic and elemental carbon are all major components of $PM_{2.5}$ [4]. These $PM_{2.5}$ chemicals have a significant impact on human health and can lead to cardiovascular problems [5]. Even the tiniest airways and lungs can be invaded by these microorganisms, causing increased respiratory oxidative stress and inflammation [6].

The ambient $PM_{2.5}$ concentrations in the PdN region surpassed the U.S. Environmental Protection Agency's (EPA) National Ambient Air Quality Standards (NAAQS) on many

occasions. The majority of $PM_{2.5}$ in this area originated from geological and industrial sources as well as vehicle exhaust and household cooking and heating. The desert environment is characterized by sporadic calm winds, frequent stagnation due to high atmospheric stability, and sporadic shallow convective and nocturnal boundary layer heights, all of which contribute to the rising $PM_{2.5}$ concentrations [7].

In recent years, machine learning algorithms showed their feasibility in predicting the concentration of air pollutants. Scientists and researchers from across the globe have used different algorithms and techniques to predict air pollutants. Several studies [8–12] found that machine learning, including deep learning [13], random forest [14], and ensemble models [15], are highly capable of estimating $PM_{2.5}$ concentration on different temporal and spatial scales.

According to the literature review above, the majority of the existing forecasting models are capable of predicting daily $PM_{2.5}$ concentrations and high PM days; however, because of the complex geography and topography of the Paso del Norte region [7,16,17], as well as its exceptional meteorological conditions, these analyses cannot be applied to our study area. Therefore, this study is dedicated to conducting an in-depth analysis of historical $PM_{2.5}$ concentrations and proposes an efficient ML method for forecasting future high/low PM concentration days in the PdN region. The novelties of this study are as follows: (1) analyzing the temporal characteristics of $PM_{2.5}$ concentration patterns by month based on the historical data collected from designated locations in the PdN region; (2) analyzing $PM_{2.5}$ concentration data using several ML models with good prediction effectiveness and comprehensible results to address various inadequacies from prior studies; (3) identifying the primary variables causing the high particulate matter concentration in this area; and (4) investigating the complex link between $PM_{2.5}$ and other meteorological and air pollutant variables based on ground station data using various ML techniques.

In this study area, researchers have conducted a number of studies to better understand the chemical and physical processes responsible for causing high $PM_{2.5}$ concentrations [16,18–20]. The majority of these investigations were diagnostic, or they modeled the situation using an idealized profile [21,22] or a specific method that was limited by the technology at that time [23]. Furthermore, the topography of the study area makes it difficult for forecast and prediction accuracy of air quality models to accurately predict pollutants [24]. Hence, the approach in our study certainly overcomes those limitations and fills the research gap.

In this study, various machine learning (ML) algorithms are utilized to predict particulate matter concentration. The study uses data on air pollutants and meteorological variables collected from several locations in the Paso del Norte region during the years from 2014 to 2019. The ML models used include ridge regression, logistic regression, MARS (multivariate adaptive regression splines), SVM (support vector machine), and RF (random forest). This study has three objectives: first, to predict high/low $PM_{2.5}$ levels; second, to determine the features that contribute to high $PM_{2.5}$ concentrations; and finally, to forecast the $PM_{2.5}$ concentration values using penalized linear regression. Detailed research has been conducted to determine how $PM_{2.5}$ concentrations affect other air pollutants and meteorological variables.

We organized this article according to the following pattern: Section 2 presents a basic overview of machine learning methods with regularization techniques to suit the best model. A significant part of Section 3 is dedicated to the details of experimental data, including an explanation of the data properties and air quality standards. Section 4 discusses the association between the variables and other exploratory analyses. Then, Section 5 presents results from the application of ML models to the data sets, as well as the accuracy and parameter estimates of the models. Finally, Section 6 summarizes the fitted models used to classify the $PM_{2.5}$ levels, along with the most important variables responsible for a high or low $PM_{2.5}$ concentration in the study area.

2. Methodology

In this section, the machine learning techniques used in this study are briefly discussed. Regularization approaches are also presented in order to achieve the best prediction model. We also show how to estimate the tuning parameters to obtain the lowest misclassification rate, predictive accuracy, mean squared error (MSE), and other metrics. The flowchart of the methodology is presented in Figure 1.

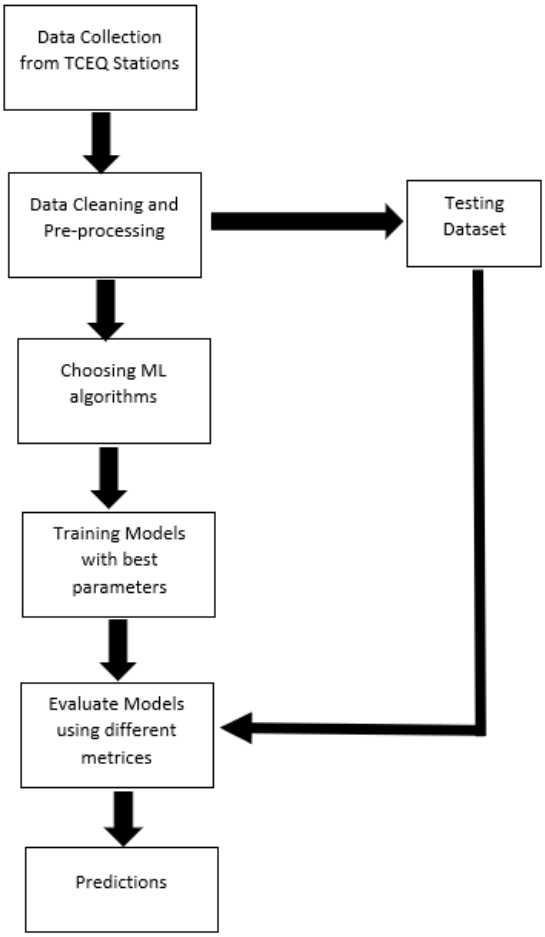


Figure 1. Flowchart of the methodology.

2.1. Penalized Linear Regression

A linear regression model can be expressed as:

$$y = \delta_0 + \sum_{i=1}^{k} \delta_i x_i$$

where δ_0 is the intercept, and $\delta_1 \cdots \delta_k$ are weights.

The penalty in the lasso is determined by the L_1 norm, i.e., the sum of the absolute values of the weights. This penalty makes the estimated weights shrink towards zero, where we can identify the significant and insignificant variables. Therefore, the lasso works for both shrinkage and variable selection [25]. In particular, the lasso is an efficient method to find the sparse/parsimonious model when a data set has a large number of features. It can be expressed as:

$$\frac{1}{2m} \sum_{i=1}^{m} \{h_\delta(x_{(i)}) - y_i\}^2 + \frac{\lambda}{2m} \sum_{j=1}^{n} |_j|$$

where λ is a tuning parameter [26]. In a ridge regression, $\frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2$ (L_2 norm) is added to the cost function to shrink the large coefficients of model as follows:

$$\frac{1}{2m}\sum_{i=1}^{m}\{h_w(x_{(i)}) - y_i\}^2 + \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2$$

However, lasso and ridge sometimes are assumed to be involved with some bias. In these models, the prediction of the target variable might be highly dependent on specific predictors. In such a case, elastic net is used as a combination of the lasso and ridge regressions as follows:

$$\frac{1}{2m}\sum_{i=1}^{m}\{h_w(x_{(i)}) - y_i\}^2 + \frac{\lambda}{2m}\left((1-\alpha)\sum_{j=1}^{m} w_j^2 + \alpha\sum_{j=1}^{m}|w_j|\right)$$

where α is any value between 0 and 1. Elastic net turns into a ridge when $\alpha = 0$ and into a lasso when $\alpha = 1$.

2.2. Logistic Regression

Logistic regression is one of the most popular models for classification. Typically, this method is used when the target variable is binary categorical. The probability of the response variable can be classified by employing the Sigmoid function [27]. Mathematically the model has the following form:

$$p = (1-p)e^{0+\sum_{i=1}^{k} ii} \quad (1)$$

here, p is the probability, and $\{_0, _1, \cdots _k\}$ are the coefficient parameters of the event. To estimate $_i$s, the maximal likelihood is used to fit the model as follows:

$$l(\,) = \sum_{i=1}^{N}\left(\,^T_{i\ i} - log(1+e^{\,^T_i})\right) \quad (2)$$

In this study, a lasso regularization technique is used with L_1 regularization presented in Equation (2). We can write the penalized versions to maximize it as follows:

$$l_\lambda(\,) = \sum_{i=1}^{N}\left(\,^T_{i\ i} - log(1+e^{\,^T_i})\right) - \lambda\sum_{j=1}^{p}|\,_j| \quad (3)$$

here, $\lambda > 0$ is the regularization parameter. We obtain the ordinary least square estimation when lambda is 0, and an under-fitting situation occurs when lambda is large due to the high weight of the data. The lasso approach can help us to reduce the coefficient of the less significant feature to zero. Moreover, in the case of feature selection, it works well if the data contains a large number of features [28].

2.3. Ridge Regression

We add Equation (2), known as a ridge regression function, to narrow the coefficients. This also helps us decrease the model's complexity and multicollinearity. The optimized model and the L_2 penalized versions are as follows:

$$l_\lambda(\,) = \sum_{i=1}^{N}\left(\,^T_{i\ i} - log(1+e^{\,^T_i})\right) - \lambda\sum_{j=1}^{p}\,^2_j \quad (4)$$

We use cross-validation to examine the findings and determine which tuning parameter has the lowest mean squared error λ [29].

2.4. Random Forest

Random forest is another frequently used classification algorithm similar to the decision tree. Many decision trees are required to build the model. Then, the majority of votes from the trees are used to perform the classification [30]. For optimum accuracy of the model, we tuned the parameters, such as node size, number of predictor samples from splitting, and number of trees. The mean decrease Gini (MDG) and mean decrease accuracy (MDA) are applied to accumulate other remarkable information. Here, the MDA helps us measure the significance of a variable, and the MDG determines the variable's contribution to the homogeneity of the nodes and leaves [31].

2.5. Multivariate Adaptive Regression Splines (MARS)

MARS is a non-parametric regression model. The model is built with different predictor values using multiple linear regression. For training, the data sets are divided into piece-wise linear segments with varying gradients that are connected using knots in a smooth manner [32]. We develop the basic function to improve the flexibility of the model as follows:

$$(-t)_+ = \begin{cases} -t, & \text{if } > t \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$\text{and } (t-)_+ = \begin{cases} t-, & \text{if } < t \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Now, we can represent $f(x)$ in the following manner:

$$f() = {}_0 + \sum_{m=1}^{M} {}_m\lambda_m(X) \tag{7}$$

where each λ_m is a BF, ${}_i$ is the constant coefficient when estimating a least-squares method. Similar to the other tree-based model, the MARS model also uses a pruning technique to fit a very large model and avoid over-fitting [33]. In addition to pruning, the backward deletion technique is used to obtain the best sub-model for eliminating less important terms. Finally, a generalized cross-validation (GCV) technique is implemented to determine the best model among the sub-models.

2.6. Support Vector Machine (SVM)

Support vector machine (SVM) is a machine learning model that examines the data used for both classification and regression analysis. This model tries to put a hyperplane between two classes in order to maximize the margin between them. In this model, we use kernel functions to handle the linearly non-separable data. A discriminant function that maximizes the geometric margin is known as the maximum margin classifier $\frac{1}{|||}$ [34], which is the same as minimizing the function $\frac{1}{2}|| \, ||^2$ with the following constraints:

$${}_i({}^t_i+) \geq 1 \quad \text{for, } i = 1, \cdots, n. \tag{8}$$

3. Data Background

The air quality index(AQI) is used for reporting the daily air quality for any specific location. It indicates the quality of air, such as whether it is clean or polluted. It also demonstrates the health risks associated with air quality. Different countries have different air quality indices. The Table 1 below shows the average $PM_{2.5}$ concentration in the United States for a 24-h period.

Table 1. EPA AQI Table for $PM_{2.5}$ concentration (24 hour avg.).

AQI Range	$PM_{2.5}$ Value ($\mu g/m^3$)	AQI Category
0–50	0–12	Good
51–100	12.1–35.4	Moderate
101–150	35.5–55.4	Unhealthy for sensitive
151–200	55.5–150.4	Unhealthy
201–300	150.5–250.4	Very Unhealthy

El Paso is a non-attainment city for carbon monoxide (CO) and $PM_{2.5}$, and it has several days of high $PM_{2.5}$ concentration during the months from May to September. Data from urban, suburban, industrial, and rural areas were used to calculate $PM_{2.5}$ precursor substances in Table 2. The data sets included both air pollutants and meteorological variables.

Table 2. Ground station locations in the Paso Del Norte region.

Sites	Latitude	Longitude	Type
UTEP/CAMS 12	31.7709 N	106.5046 W	Urban
Santa Teresa	31.8729 N	106.6978 W	Rural
Skyline Park	31.8924 N	106.4257 W	Urban
Socorro Hueco	31.61712 N	106.28822 W	Rural

The Paso del Norte (PdN) region has become a major environmental concern for both countries in recent decades. The PdN region is made up of the cities of El Paso, Texas; Ciudad Juarez, Mexico; and a few more cities from New Mexico [35]. Our study area includes deserts such as The Chihuahuan, mountain ranges, shared rivers, wetlands, state parks, and protected areas. Around 12 million people live along the border, almost equally divided between the two countries. With 0.7 million people, El Paso is the U.S.'s eighth largest city; adjacent to it, another 1.3 million people live in Cuidad Juarez, Mexico [23]. El Paso, a southwestern U.S. city, has the typical warm and arid climate, but its air quality is typically large because of the industrial activity along the U.S./Mexico border, as well as the unique meteorological conditions created by geography [18,20]. Due to a mix of high population density, industrial effects, and weather circumstances, El Paso has historically been in non-attainment for the U.S. NAAQS for O_3, CO and PM_{10}, and $PM_{2.5}$ [19,36].

In this work, data were collected from the Continuous Ambient Monitoring Station (CAMS) of TCEQ. Data from different CAMS in the Paso del Norte region were used to collect hourly average $PM_{2.5}$ concentrations at ground level. In Figure 2, the AQI days were displayed from the year 2014 to 2019 in the El Paso region. As shown, most of the moderate and unhealthy AQI days were in the summer and winter seasons. In addition, recent years have shown increases in unhealthy days throughout the year.

Table 2 and Figure 3 provide details about those locations.

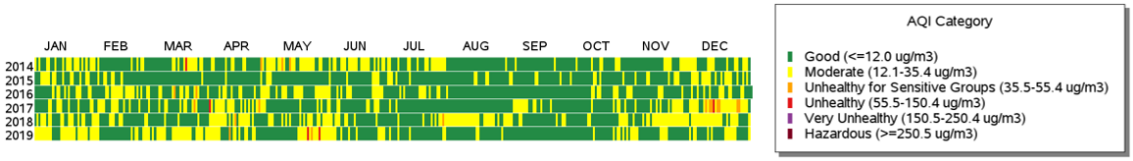


Figure 2. Different PM concentration days in El Paso during 2014–2019 [37].

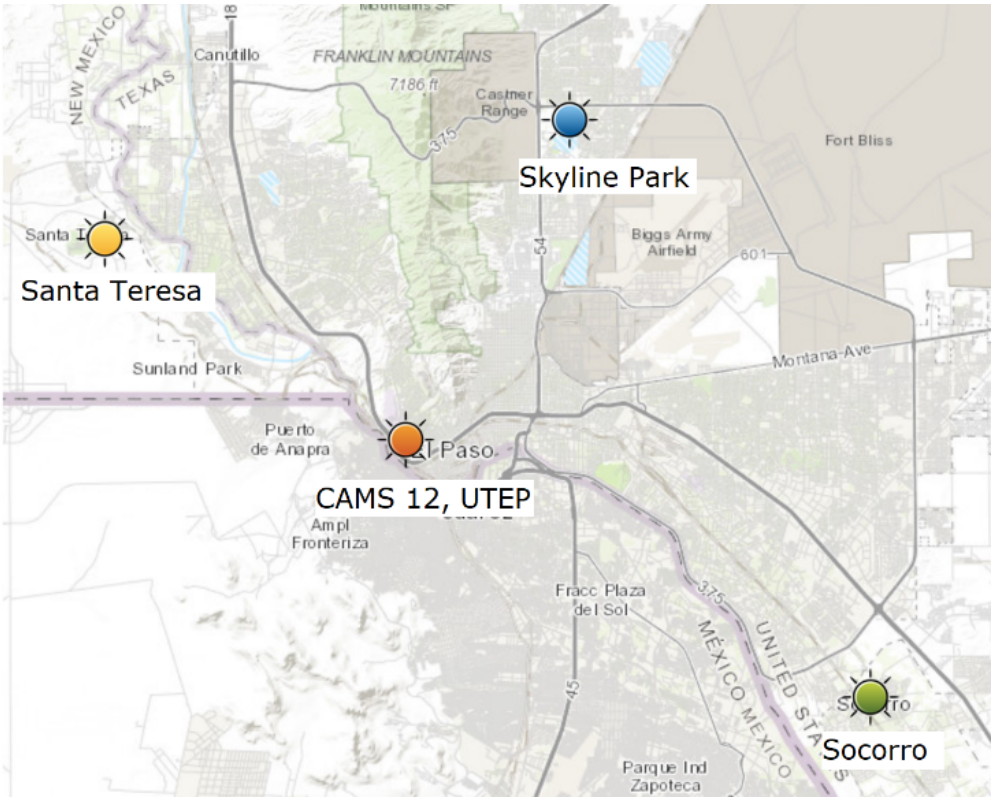


Figure 3. Data location sites in the Paso del Norte area.

4. Exploratory Data Analysis

El Paso, Texas, is considered to have the highest levels of $PM_{2.5}$ in the United States. It has a history of high $PM_{2.5}$ exceedances every year. To demonstrate the trend for high $PM_{2.5}$ days, we conducted an extensive study of the years from 2014 to 2019.

Figure 4 shows the box and whisker plot of all the meteorological and air pollutant variables used for our study. On the vertical axis, the numerical values for all of the variables are presented, and the names of the variables are presented on the horizontal axis. As most of the data were collected in the summer season, due to the high PM concentration days, the mean value of the outdoor temperature is around 80 degrees Fahrenheit. The relative humidity is around 30–40%, and the Ozone is around 50–60 parts per billion.

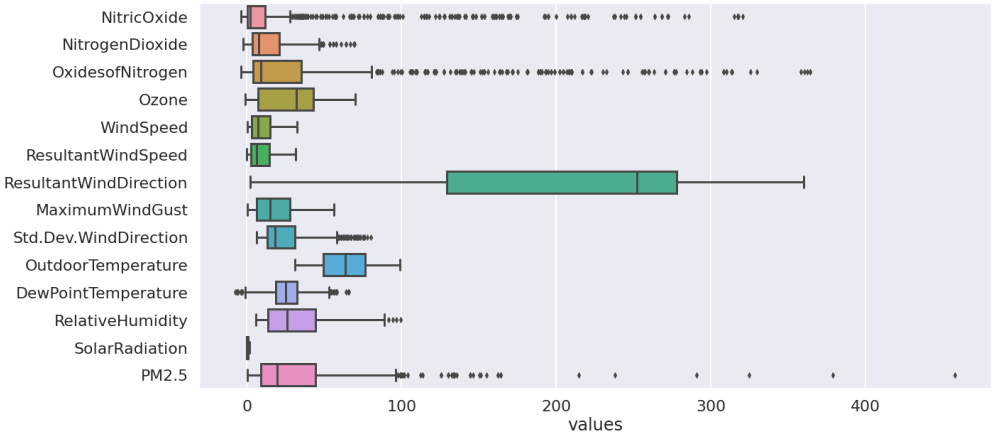


Figure 4. Box and whisker plot of all variables.

Figure 5 illustrates the correlation between all of the predictors of $PM_{2.5}$ from the data set. Several meteorological variables, such as wind speed, resultant wind speed, and maximum wind gust, have a better positive correlation with $PM_{2.5}$. On the contrary, dew point

temperature and relative humidity have a negative correlation with the target variable, i.e., $PM_{2.5}$. Figure 6 shows the scatter plot matrix with the slope values between the variables and a histogram of the diagonal element. This histogram provides a sense of the shape of the univariate distribution for each variable. Additionally, above each scatter plot, the slope of the linear fit is demonstrated with its statistical significance indicated by one asterisk (*) sign, which denotes $p < 0.05$, or two asterisk signs (**), which shows $p < 0.01$. As illustrated, with our target variable, $PM_{2.5}$, all pollutants are in positive associations, and there are statistically significant relationships between them.

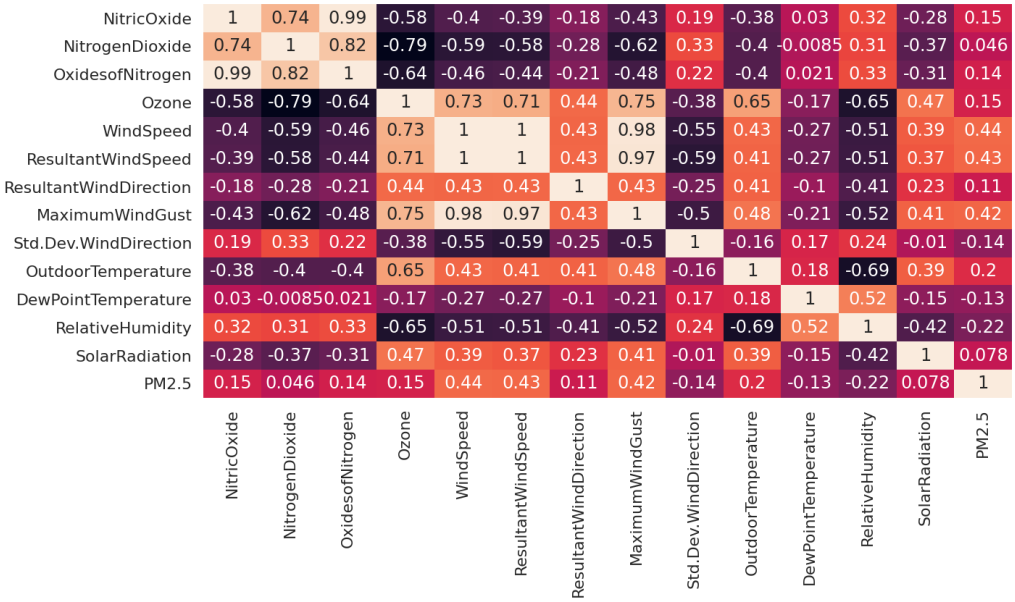


Figure 5. Correlation among the predictors with histogram of $PM_{2.5}$ data set.

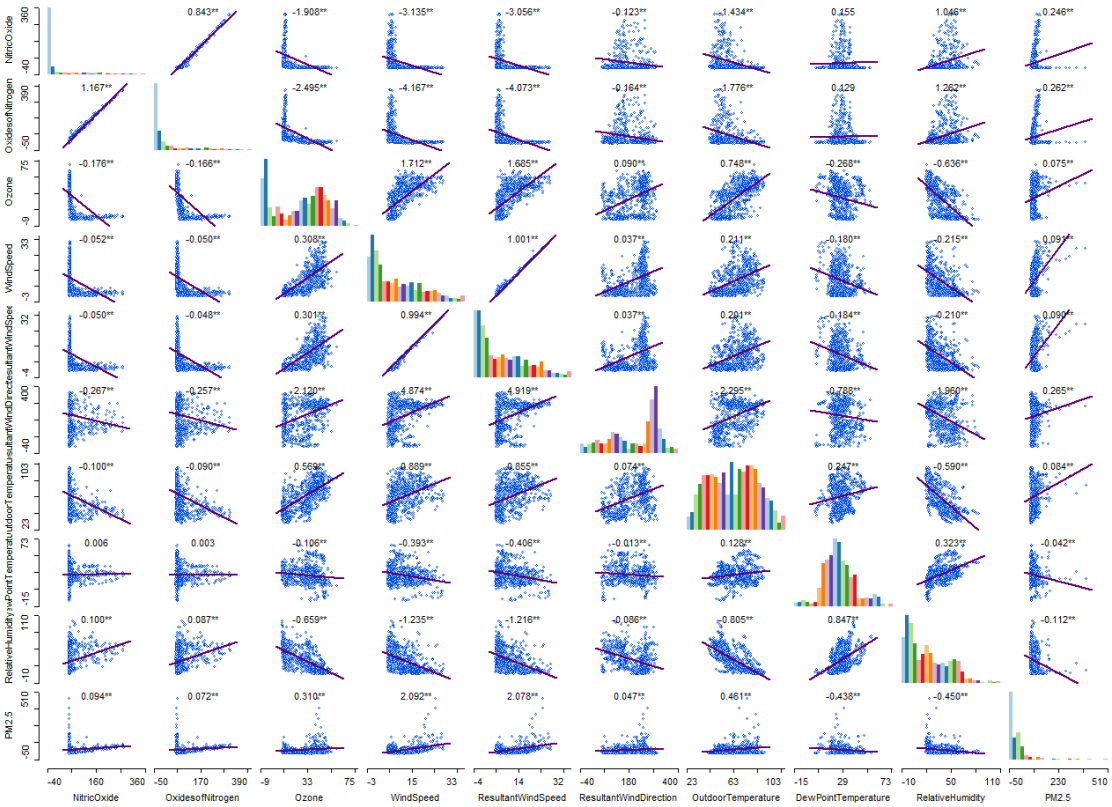


Figure 6. Scatter plots with the slope values.

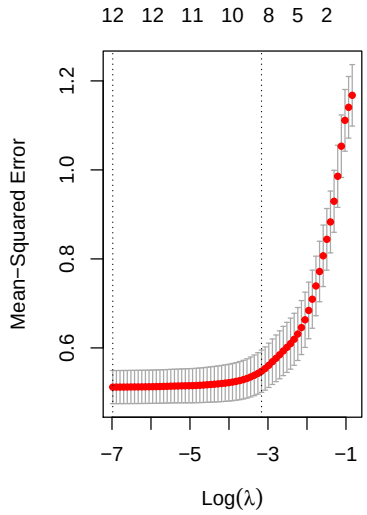
5. Results

This section will discuss the data processing approaches, the results, and the applications of machine learning models for predicting the ground-level $PM_{2.5}$ concentration in the atmosphere. In this study, 70% of the training data was used in the prediction model, and 30% of the test data was used to evaluate the model. The regularization techniques are considered to obtain the best model by using bias-variance trade-off rules. To predict $PM_{2.5}$ data, we first use penalized regression based on several meteorological variables. The lasso regression was used to predict the $PM_{2.5}$ with reduced predictors. The coefficients obtained from the lasso and its evaluation metrics are presented in Table 3.

Table 3. Coefficients of lasso model.

Variables	Coefficients
Nitric Oxide	0.0057
Nitrogen Dioxide	0.03147
Oxides of Nitrogen	0.0183
Ozone	−0.0039
Wind Speed	0.0282
Resultant Wind Speed	0.0576
Resultant Wind Direction	−0.0019
Maximum Wind Gust	0.0198
Std. Dev. Wind Direction	0.0109
Outdoor Temperature	0.0275
Dewpoint Temperature	−0.0137
Relative Humidity	0.0066
Solar Radiation	−0.3004

We also used the ridge and elastic net regression methods to overcome the multicollinearity issue and to predict the $PM_{2.5}$, including all the variables. Figure 7 shows the sample path of the tuning parameter λ for the above three models when cross-validation is applied. From these figures, we see how the tuning parameter λ was picked using cross-validation. At this point, the two dotted lines show the two lambda values. The left one gives the minimum cross-validation error, and the right one gives the most highly regularized model within 1 S.D. of the minimum error for a fixed α.



(**a**)

Figure 7. *Cont.*

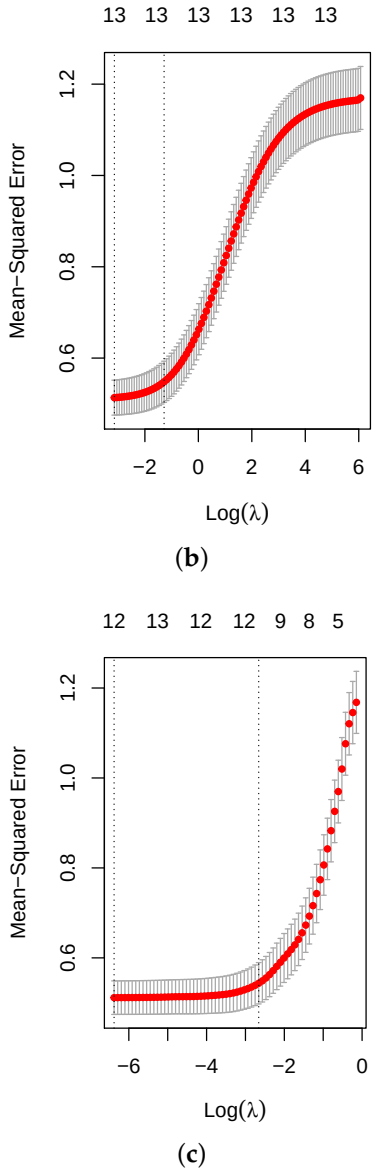


Figure 7. Sample path of tuning parameter λ for the three penalized models. (**a**) Lasso λ path. (**b**) Ridge λ path. (**c**) Net λ path.

In the logistic regression, we used the lasso regularization with the L_1 penalty and obtained the tuning parameter λ with cross-validation. The L_1 penalty is significant for variable selection and shrinkage because it forces some of the coefficients' estimates to be zero [38]. Table 4 demonstrates the coefficients of the predictors where Nitrogen Dioxide, Oxides of Nitrogen, Wind Speed, Resultant Wind Direction, Std. Dev. Wind Direction, Outdoor temperature, and Relative Humidity are the important factors for $PM_{2.5}$ classification.

In addition to L_1, we used the logistic regression model with an L_2 penalty to reduce the multicollinearity issue of the data set. The tuning parameter λ is optimized via ten-fold cross-validation until we achieve the best predictive model. For the MARS model, the cross-validation of the training data was used to choose a reliable classifier. This model regulates the training process with the residual sum of squares (RSS). In the random forest model, five-hundred trees were used. and three variables were sampled at each split to classify the levels. Using the mean decrease accuracy and mean decrease Gini indices, we ordered the predictors based on their importance. Figure 8 shows that the predictors

Nitrogen Dioxide, *Wind Speed*, *Oxides of Nitrogen*, and *Maximum Wind Gust* are important variables for high $PM_{2.5}$ levels in the atmosphere.

Table 4. Coefficients of important predictors using LGR(L_1) model.

Variables	Coefficients
Nitric Oxide	0.000
Nitrogen Dioxide	0.049
Oxides of Nitrogen	0.018
Ozone	0.000
Wind Speed	0.225
Resultant Wind Speed	0.000
Resultant Wind Direction	-0.002
Maximum Wind Gust	0.000
Std. Dev. Wind Direction	0.005
Outdoor Temperature	0.024
Dewpoint Temperature	0.000
Relative Humidity	-0.006
Solar Radiation	0.000

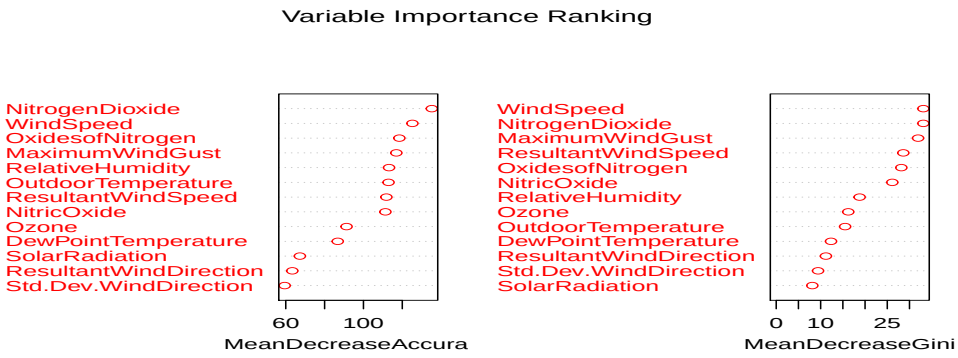


Figure 8. Variable importance plot using mean decrease accuracy and Gini indices.

Table 5 shows the prediction mean squared error and misclassification rate of the models, and we also analyze the confusion metrics to choose the best classifier for the $PM_{2.5}$ concentration.Lastly, the kernel SVM is studied to classify the $PM_{2.5}$ concentration, where ten-fold cross-validation and different cost levels were used. The optimized cost and accuracy were found for the parameter γ of 0.001.

Table 5. Model evaluation.

Models	Prediction Mean Squared Error	Misclassification Rate
LGR	0.100	0.120
RGR	0.106	0.112
RF	0.067	0.072
MARS	0.113	0.163
SVM	0.084	0.109

In Table 6, RMSE and R-Square values of tuning parameter λ for the three penalized models are presented.

Table 6. Model parameter and evaluation metrics for penalized linear regression.

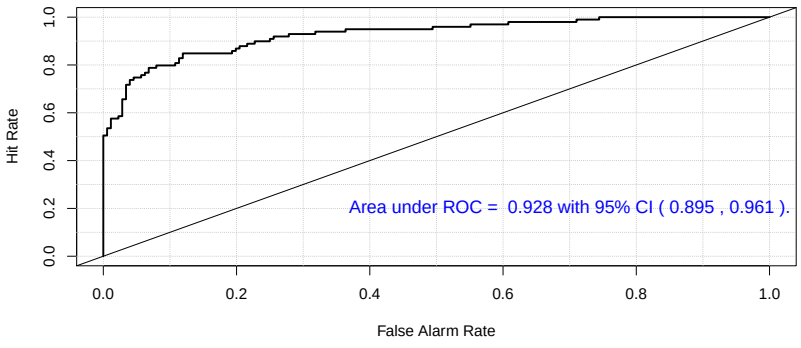
Regression Models	α	λ (min)	RMSE	R-Square
Lasso	1.0000	0.0009	0.7028	0.5761
Ridge	0.0000	0.0431	0.7056	0.5737
Elastic Net	0.5000	0.00169	0.7028	0.5760

Model Accuracy

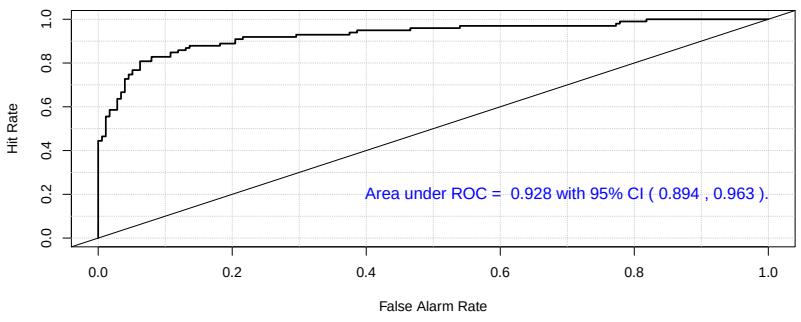
This section compares and contrasts our proposed models using a variety of evaluation metrics (see Table 7). The true positive rate, or the fraction of detected positives in the target variable, represents the sensitivity in this case. The true negative rate (TNR), or the fraction of recognized negatives, is measured by specificity. At this point, the ROC curve is also presented where the X axis shows the true positive rate, or sensitivity, and the Y axis shows the false positive rate, or 1-specificity. The confidence of interval is a range of values that is likely to include a population value with a certain degree of confidence. Accuracy is the proportion of the total number of predictions that are correct. The diagonal line of the ROC curve represents the threshold (0.5), which separates the ROC space (see Figure 9). A good classifier tends towards a value of one. From Table 7 and Figure 9, it is concluded that the random forest model performs well compared to others.

Table 7. Model evaluation using classification metrics

Models	Sensitivity (%)	Specificity (%)	Accuracy (%)	Conf. Interval (%)
LGR	88.65	86.67	88.00	(83.50–91.60)
RGR	87.77	87.36	87.64	(83.15–91.28)
MARS	88.64	74.75	83.64	(78.72–87.81)
SVM	92.44	83.50	89.09	(84.79–92.52)
RF	93.18	91.92	92.73	(88.99–95.50)

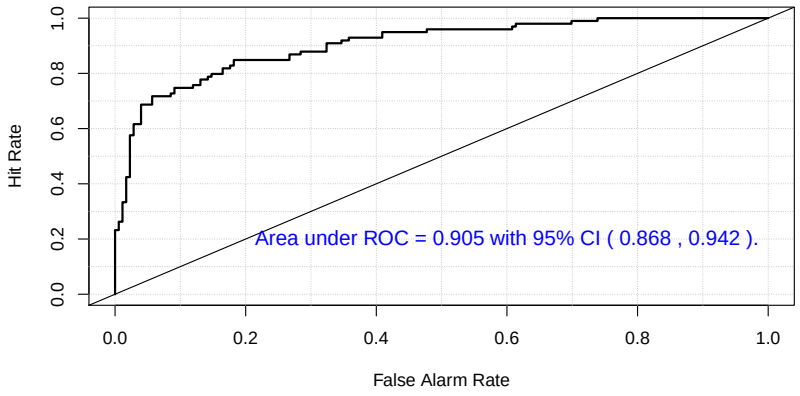


(**a**)

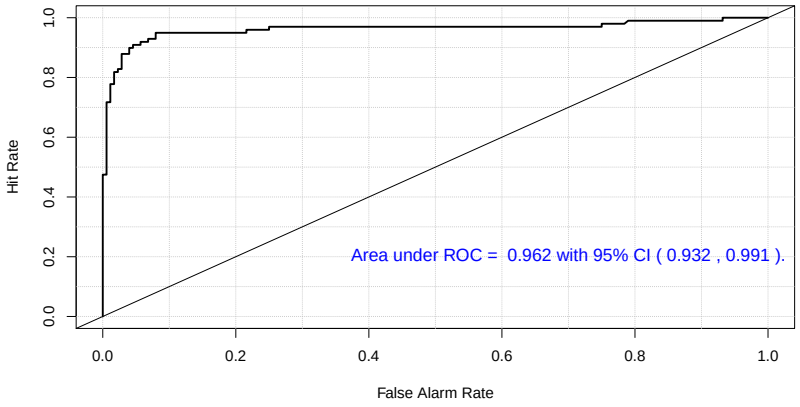


(**b**)

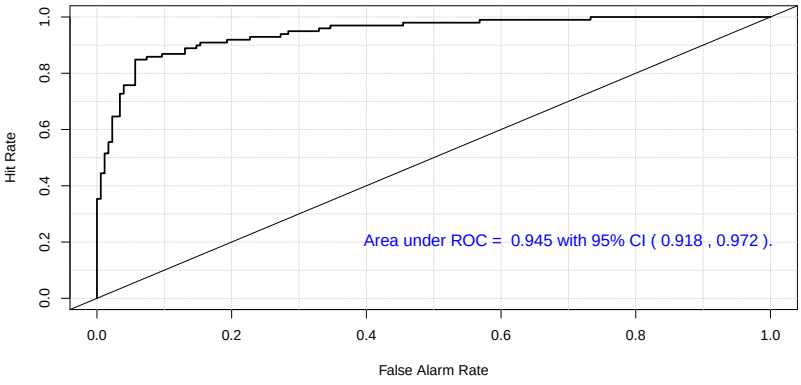
Figure 9. *Cont.*



(**c**)



(**d**)



(**e**)

Figure 9. (**a**) ROC curve for logistic regression. (**b**) ROC curve for ridge regression. (**c**) ROC curve for MARS model. (**d**) ROC curve for random forest. (**e**) ROC curve for SVM kernel.

6. Conclusions

In recent years, scientists have proposed and implemented numerous models for forecasting and predicting air pollutants across different geographical locations. The results of this study suggest that machine learning techniques are effective for predicting $PM_{2.5}$ concentrations based on meteorological and air pollution variables. The purpose of this study was to analyze methodologies for predicting ground-level fine particle concentrations. Several meteorological parameters, including temperature, wind speed, relative humidity, and different air pollutants, including CO, NOx, and Ozone are used for classification.

Our proposed penalized regression models with L1 and L2 regularization provide important features for detecting high or low $PM_{2.5}$ days. To determine the significant predictors for

high $PM_{2.5}$ concentration days, we used various ML algorithms such as random forest, MARS, logistic regression, and SVM. Cross-validations of the training data were used to examine the various cost functions, yielding the models' tuning parameters. The tuning parameter determines which model is best for predicting the test data. After fitting the test data with the optimized predictive models, several metrics were computed to assess the prediction. In addition, the accuracy, sensitivity, specificity, precision, and recall metrics have been compared (see Table 6) to obtain the best classifier. According to empirical research, the random forest model correctly classifies 92.73% of $PM_{2.5}$ data as high or low with a confidence interval of (89% to 96%). It also demonstrates that several meteorological elements, such as Nitric Oxide, Wind Speed, and Maximum Wind Gust have a significant impact on $PM_{2.5}$'s high concentration. The areas under the ROC curves for all ML approaches are shown in Figure 9, where the RF and SVM models depict high accuracy in classifying high and low $PM_{2.5}$ days. The results of this study can contribute to an evaluation of the long-term effects of $PM_{2.5}$ air pollution and the diseases caused by exposure to $PM_{2.5}$. Furthermore, the analysis provides valuable information which can be useful in the prevention and control of air pollution in the binational airshed. The future work of this research work will focus on the prediction of an unanticipated increase in $PM_{2.5}$ in the study area during the peak season of air pollution using deep learning, i.e, LSTM (long short-term memory) analysis, a feed-forward neural network using multiple neurons, stochastic approaches [39], causality discovery approaches [40], etc. Further, it can be used as the basis for many future advanced research projects involving machine learning/deep learning-based air pollution prediction, since extended historical data can be collected for training tailored to this region.

Author Contributions: S.M., T.B.I.R., and S.E. contributed to the supervision and project administration; M.S.M., F.S., S.E., and S.M. contributed to the conceptualization, methodology, and results analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This research project did not receive any additional funding.

Informed Consent Statement: All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: All data supporting this study are provided as supplementary information accompanying this paper.

Acknowledgments: Our thanks and heartful gratitude goes to the Texas Commission of Environmental Quality, the Earth and Environmental Systems Institute (Penn State), and the Computational Science Program (UTEP) for all their support.

Conflicts of Interest: It is declared that neither the authors nor their associates have any competing interests that may have affected the content of this paper.

References

1. Chow, J.C.; Watson, J.G.; Edgerton, S.A.; Vega, E. Chemical composition of $PM_{2.5}$ and PM10 in Mexico City during winter 1997. *Sci. Total Environ.* **2002**, *287*, 177–201. [CrossRef] [PubMed]
2. Quintero, M.; Meza, L.; Canales, M.; Ahumada, S. The program to improve the air quality of Mexicali, Baja California, Mexico 2010–2015. *Procedia Environ. Sci.* **2010**, *2*, 800–813. [CrossRef]
3. Seinfeld, J.; Pandis, S. *Atmospheric Chemistry and Physics. 1997*; Yale University Press: New Haven, CT, USA, 2008.
4. Khanna, I.; Khare, M.; Gargava, P.; Khan, A.A. Effect of $PM_{2.5}$ chemical constituents on atmospheric visibility impairment. *J. Air Waste Manag. Assoc.* **2018**, *68*, 430–437. [CrossRef] [PubMed]
5. Kim, K.H.; Kabir, E.; Kabir, S. A review on the human health impact of airborne particulate matter. *Environ. Int.* **2015**, *74*, 136–143. [CrossRef]
6. Zhang, H.h.; Li, Z.; Liu, Y.; Xinag, P.; Cui, X.y.; Ye, H.; Hu, B.l.; Lou, L.p. Physical and chemical characteristics of $PM_{2.5}$ and its toxicity to human bronchial cells BEAS-2B in the winter and summer. *J. Zhejiang Univ. Sci. B* **2018**, *19*, 317. [CrossRef]
7. Karle, N.N.; Mahmud, S.; Sakai, R.K.; Fitzgerald, R.M.; Morris, V.R.; Stockwell, W.R. Investigation of the Successive Ozone Episodes in the El Paso–Juarez Region in the Summer of 2017. *Atmosphere* **2020**, *11*, 532. [CrossRef]
8. Dai, H.; Huang, G.; Wang, J.; Zeng, H.; Zhou, F. Prediction of Air Pollutant Concentration Based on One-Dimensional Multi-Scale CNN-LSTM Considering Spatial-Temporal Characteristics: A Case Study of Xi'an, China. *Atmosphere* **2021**, *12*, 1626. [CrossRef]

9. Dai, H.; Huang, G.; Zeng, H.; Zhou, F. $PM_{2.5}$ volatility prediction by XGBoost-MLP based on GARCH models. *J. Clean. Prod.* **2022**, *356*, 131898. [CrossRef]
10. Dai, H.; Huang, G.; Wang, J.; Zeng, H.; Zhou, F. Spatio-Temporal Characteristics of $PM_{2.5}$ Concentrations in China Based on Multiple Sources of Data and LUR-GBM during 2016–2021. *Int. J. Environ. Res. Public Health* **2022**, *19*, 6292. [CrossRef]
11. Wong, Y.J.; Shimizu, Y.; Kamiya, A.; Maneechot, L.; Bharambe, K.P.; Fong, C.S.; Nik Sulaiman, N.M. Application of artificial intelligence methods for monsoonal river classification in Selangor river basin, Malaysia. *Environ. Monit. Assess.* **2021**, *193*, 1–22. [CrossRef]
12. Wong, Y.J.; Shiu, H.Y.; Chang, J.H.H.; Ooi, M.C.G.; Li, H.H.; Homma, R.; Shimizu, Y.; Chiueh, P.T.; Maneechot, L.; Sulaiman, N.M.N. Spatiotemporal impact of COVID-19 on Taiwan air quality in the absence of a lockdown: Influence of urban public transportation use and meteorological conditions. *J. Clean. Prod.* **2022**, *365*, 132893. [CrossRef]
13. Park, Y.; Kwon, B.; Heo, J.; Hu, X.; Liu, Y.; Moon, T. Estimating $PM_{2.5}$ concentration of the conterminous United States via interpretable convolutional neural networks. *Environ. Pollut.* **2020**, *256*, 113395. [CrossRef] [PubMed]
14. Hu, X.; Belle, J.H.; Meng, X.; Wildani, A.; Waller, L.A.; Strickland, M.J.; Liu, Y. Estimating $PM_{2.5}$ concentrations in the conterminous United States using the random forest approach. *Environ. Sci. Technol.* **2017**, *51*, 6936–6944. [CrossRef] [PubMed]
15. Xiao, Q.; Chang, H.H.; Geng, G.; Liu, Y. An ensemble machine-learning model to predict historical $PM_{2.5}$ concentrations in China from satellite data. *Environ. Sci. Technol.* **2018**, *52*, 13260–13269. [CrossRef] [PubMed]
16. Mahmud, S.; Bhuiyan, M.A.M.; Sarmin, N.; Elahee, S. Study of wind speed and relative humidity using stochastic technique in a semi-arid climate region. *AIMS Environ. Sci.* **2020**, *7*, 156–173. [CrossRef]
17. Mahmud, S.; Karle, N.N.; Fitzgerald, R.M.; Lu, D.; Nalli, N.R.; Stockwell, W.R. Intercomparison of Sonde, WRF/CAMx and Satellite Sounder Profile Data for the Paso Del Norte Region. *Aerosol Sci. Eng.* **2020**, *4*, 277–292. [CrossRef]
18. Brown, M.J.; Muller, C.; Wang, G.; Costigan, K. Meteorological simulations of boundary-layer structure during the 1996 Paso del Norte Ozone Study. *Sci. Total Environ.* **2001**, *276*, 111–133. [CrossRef]
19. Einfeld, W.; Church, H.W.; Yarbrough, J.W. *Winter Season Air Pollution in El Paso-Ciudad Juarez*; Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 1995.
20. Funk, T.; Chinkin, L.; Roberts, P.; Saeger, M.; Mulligan, S.; Figueroa, V.P.; Yarbrough, J. Compilation and evaluation of a Paso del Norte emission inventory. *Sci. Total Environ.* **2001**, *276*, 135–151. [CrossRef]
21. Hutchison, K.D.; Smith, S.; Faruqui, S.J. Correlating MODIS aerosol optical thickness data with ground-based $PM_{2.5}$ observations across Texas for use in a real-time air quality prediction system. *Atmos. Environ.* **2005**, *39*, 7190–7203. [CrossRef]
22. Mahmud, S.; Wangchuk, P.; Fitzgerald, R.; Stockwell, W. Study of Photolysis Rate Coefficients to Improve Air Quality Models. *Bull. Am. Phys. Soc.* **2016**, *61*.
23. Ordieres, J.; Vergara, E.; Capuz, R.; Salazar, R. Neural network prediction model for fine particulate matter ($PM_{2.5}$) on the US–Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). *Environ. Model. Softw.* **2005**, *20*, 547–559. [CrossRef]
24. Mahmud, S. *The Use of Remote Sensing Technologies and Models to Study Pollutants in the Paso del Norte Region*; The University of Texas at El Paso: El Paso, TX, USA, 2016; p. 15.
25. Heckman, N.E.; Ramsay, J.O. Penalized regression with model-based penalties. *Can. J. Stat.* **2000**, *28*, 241–258. [CrossRef]
26. Wu, T.T.; Lange, K. Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* **2008**, *2*, 224–244. [CrossRef]
27. Menard, S. *Applied Logistic Regression Analysis*; Sage: New York, NY, USA, 2002; Volume 106.
28. Peng, C.Y.J.; Lee, K.L.; Ingersoll, G.M. An introduction to logistic regression analysis and reporting. *J. Educ. Res.* **2002**, *96*, 3–14. [CrossRef]
29. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
30. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
31. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **2005**, *26*, 217–222. [CrossRef]
32. Zhang, W.; Goh, A.T.; Zhang, Y. Multivariate adaptive regression splines application for multivariate geotechnical problems with big data. *Geotech. Geol. Eng.* **2016**, *34*, 193–204. [CrossRef]
33. Kuter, S.; Akyurek, Z.; Weber, G.W. Retrieval of fractional snow covered area from MODIS data by multivariate adaptive regression splines. *Remote Sens. Environ.* **2018**, *205*, 236–252. [CrossRef]
34. Suykens, J.A.; Vandewalle, J. Least squares support vector machine classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300. [CrossRef]
35. Mahmud, S. Optimization Of Regional Scale Numerical Weather Prediction & Air Quality Model For The Paso Del Norte Region. Doctoral Dissertation. The University of Texas at El Paso, El Paso, TX, USA, 2020.
36. MacDonald, C.P.; Roberts, P.T.; Main, H.H.; Dye, T.S.; Coe, D.L.; Yarbrough, J. The 1996 Paso del Norte Ozone Study: analysis of meteorological and air quality data that influence local ozone concentrations. *Sci. Total Environ.* **2001**, *276*, 93–109. [CrossRef] [PubMed]
37. Environment Protection Agency. Outdoor Air Quality. Available online: https://www.epa.gov/outdoor-air-quality-data/air-data-multiyear-tile-plot (accessed on 20 November 2022).
38. Bhuiyan, M.; Mahmud, S.; Sarmin, N.; Elahee, S. A Study on Statistical Data Mining Algorithms for the Prediction of Ground-Level Ozone Concentration in the El Paso–Juarez Area. *Aerosol Sci. Eng.* **2020**, *4*, 293–305. [CrossRef]

39. Bhuiyan, M. Predicting Stochastic Volatility for Extreme Fluctuations in High Frequency Time Series. Doctoral Dissertation, The University of Texas at El Paso, El Paso, TX, USA, 2020.
40. Hussung, S ; Mahmud, S; Sampath, A ; Wu, M; Guo, P ; Wang, J. Evaluation of data-driven causality discovery approaches among dominant climate modes. *UMBC Faculty Collection* **2019**.