Examining $PM_{2.5}$ Mitigation Potential via Urban Forestry Using Machine Learning Methods in Austin, Texas

Asmit Chakraborty
Asmit.Chakraborty@utexas.edu
The University of Texas at Austin
Austin, Texas, USA

Berkeley Ho
BerkNHo@utexas.edu
The University of Texas at Austin
Austin, Texas, USA

Grace Nguyen
NguyenGrace@utexas.edu
The University of Texas at Austin
Austin, Texas, USA

Victor Nguyen
victor.bt.nguyen@utexas.edu
The University of Texas at Austin
Austin, Texas, USA

Vijetha Ramdas
VijethaRamdas@utexas.edu
The University of Texas at Austin
Austin, Texas, USA

Figure 1: City of Austin. Photograph by Roschetzky Photography, via Shutterstock

Abstract

High levels of $PM_{2.5}$, associated with poor air quality, pose a threat to health in the Austin area and, in general, have been shown to be associated with negative short- and long-term health effects. Using PurpleAir sensor data in conjunction with weather data from the National Oceanic and Atmospheric Association and tree data from the City of Austin, we constructed ensemble models aimed at understanding how particulate matter concentrations are affected by environmental conditions, weather patterns, and local forestry throughout the Austin area. We evaluated these models using mean squared error, mean absolute error, and the coefficient of determination R^2 and used interpretability methods including partial dependency plots and Shapley Vvlues to gain further insights from our selected model. Our findings show that $PM_{2.5}$ is positively correlated with temperature and humidity and negatively correlated with canopy cover percentage and tree diameter, although our insight into the latter two was limited by the availability of our data.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

We also observed that $PM_{2.5}$ tends to spike in the morning, with higher overall levels in the first half of the year. These insights seek to drive efficient city resource allocations and forestry directives to combat high levels of $PM_{2.5}$.

CCS Concepts

• **Applied computing** → **Physical sciences and engineering**; • **Computing methodologies** → *Machine learning*; • **Information systems** → Data analytics.

Keywords

XGBoost, Random Forest, Air quality, Air pollution, PurpleAir

ACM Reference Format:

Asmit Chakraborty, Berkeley Ho, Grace Nguyen, Victor Nguyen, and Vijetha Ramdas. 2018. Examining $PM_{2.5}$ Mitigation Potential via Urban Forestry Using Machine Learning Methods in Austin, Texas. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The ability to predict air quality is vital in the prevention of negative health outcomes. A significant contributor to the Air Quality Index (AQI) is particulate matter. Particles less than 2.5 micrometers in diameter ($PM_{2.5}$) are a substantial factor in this calculation. Although particulate matter levels of less than 10 micrometers (PM_{10}) are also measured, the connections between adverse health effects and smaller particles are stronger, thus being of greater interest in indicating air particle pollution.

Inhalation of or exposure to $PM_{2.5}$ presents negative health implications for populations [1]. Additionally, children and older adults who have preexisting conditions are at greater risk [2]. Due to its small size, it has the ability to disperse within the human body, travelling among the respiratory system and damaging immune response [3]. Additional avenues of research have also considered connections between pollution and economic loss [4][5].

In urban environments, there are a variety of different sources contributing to $PM_{2.5}$. Among them, traffic, fuel combustion and emissions are the most prevalent [6]. While prevention or reduction of $PM_{2.5}$ from these sources is possible, it is not entirely feasible to eliminate pollution via these methods [7]. Instead, combating the spread of particulate matter is also an important consideration for areas susceptible to these conditions. The measurement of $PM_{2.5}$ levels has evolved significantly over time. Traditional methods relied on large, stationary monitoring stations, which provided accurate but spatially limited data. Recently, low-cost sensors such as PurpleAir monitors have gained popularity due to their ability to provide dense, localized readings [8]. This increased availability of granular data has advanced research on air quality, allowing for more detailed studies on sources and mitigation strategies. One promising method for mitigating $PM_{2.5}$ pollution involves urban forestry. Trees can act as natural air filters, capturing and absorbing pollutants, including particulate matter [9]. Studies have shown that tree canopy coverage and species composition can significantly influence their effectiveness in reducing $PM_{2.5}$. Understanding these dynamics can inform urban planning and resource allocation, particularly in rapidly growing urban areas. Austin, Texas, presents a unique case study at the intersection of rapid urban development and environmental management. The city has experienced significant economic and population growth, resulting in increased construction and vehicular traffic—key contributors to $PM_{2.5}$ [10]. However, Austin's growing efforts to implement urban forestry offers an opportunity to explore how tree coverage and diversity might mitigate pollution levels [11].

Previous research has employed machine learning models to demonstrate predictive capabilities in regards to estimating $PM_{2.5}$. Neural networks, random forest, and XGBoost have been implemented, providing different advantages and insight into modeling pollution [12][13][14]. However, the opportunity remains in understanding how urban forestry can be optimized to combat air pollution at a smaller resolution within cities.

In our work we hope to contribute to a more nuanced understanding of how tree coverage and urban forestry can mitigate particulate matter pollution, providing insight for policy makers in their city planning. Thus, the objective of our study is twofold. We aim to predict $PM_{2.5}$ levels given the specific characteristics of the surrounding area, and also determine any connections between pollution levels and these features. By understanding these relationships, this research seeks to assist in producing actionable insight for urban planning, ultimately contributing to improved air quality and public health outcomes in Austin and similar urban environments.

The rest of the paper is organized as follows. Section 2 presents the dataset in detail, including its source, preprocessing steps, and training and testing details. Section 3 provides the methodology and justification for the modeling portion. Section 4 presents the key findings and Section 5 interprets these results and their significance. Section 6 concludes with a summary of the project and final thoughts.

2 Data

2.1 Data Collection

Data was sourced from several publicly accessible data portals at no cost. PurpleAir, a company that facilitates the measurement of air quality, provides historical records on particulate matter using their sensors. The PurpleAir PA-II monitors have been evaluated heavily for outdoor usage in research [15][16]. The instruments use two Plantower PMS 5003 sensors to obtain the mass concentration of airborne particles. However, density of particulate matter can vary due to the source of pollutant, introducing possible error via overestimation in measurements. Despite this, the sensors provide a more cost effective and accessible solution to monitoring air quality at a finer spatiotemporal resolution [17]. Since customers are responsible for purchasing and installing the company's outdoor air sensors, the distribution of sensors in a geographic region is reliant upon the user-base.

For the purpose of this project, measurements were obtained if they were recorded in the city of Austin between September of 2022 and November of 2024. The source pool of instruments in the city consisted of 191 total sensors. From these, observations were tracked using six-hour averages of $PM_{2.5}$. The outdoor monitors

provide multiple algorithms for calculating $PM_{2.5}$ concentrations, including the ATM and ALT-CF3.4 (ALT) methods. The latter measurement was preferred due to its improved precision and overall accuracy in predicting pollution levels [18]. As a result, ALT was designated as the main target variable when modeling.

To account for possible confounding variables, additional climate information was retrieved from the National Oceanic and Atmospheric Administration. The Climate Data Online tool was used to access elevation and historical daily precipitation totals across Travis County for the same period of September 2022 to November 2024 [19]. The region contains 233 unique stations with recorded rainfall data. However, the number of measurements observed at each location was not consistently available across the two year period. As a result, only 111 different stations had usable data logged within this time frame.

Tree canopy coverage was collected from the City of Austin open data portal [20]. To produce this, the city compiled satellite and aerial imagery from Maxar Technologies and the United States Department of Agriculture National Agricultural Imagery Program (USDA NAIP), respectively. The source files were translated into Geographic Information System (GIS) vector data via machine learning algorithms. Image pixels were then labeled based on whether or not they belonged to tree canopy, after which GIS professionals manually reviewed and annotated the canopy designations. To derive total tree canopy, acres of tree canopy were divided by acres of land. The resulting coverage was provided in GIS format on the data portal for 2022.

Additional information on individual trees is also accessible through the City of Austin open data portal. A tree inventory dataset was used to obtain characteristics of locally monitored trees in the city of Austin [21]. Several departments contributed to its collection, including the Development Services Department's Tree Division, Austin Independent School District, the Austin Parks and Recreation Department, and the Austin Public Works Department. The compiled set included 62,274 trees as of March of 2020. The features observed included the longitude and latitude for each tree, as well as its species and diameter at 4.5-feet from the ground.

2.2 Data Processing

Before using the data for modeling, several quality checks were performed to reduce the possibility of using errant measurements. The initial sensor dataset had 139,801 observations. However, rows that contained missing values were removed. Additionally, observations that had temperature readings outside of typical weather conditions for Austin were filtered out. More explicitly, rows that reported conditions below 0°F or above 115°F were filtered out.

To discard remaining outliers, the mean temperature across month-long periods within each cluster of sensors was calculated. Sensors were first clustered using Density-Based Spatial Clustering of Applications with Noise (DBSCAN). These clusters provided a means to aggregate data within clusters containing sensors with similar characteristics such as their coordinates and elevation. The average temperature within the sensor readings for each cluster was then calculated for each month. Any data that was more than two standard deviations from the mean was then tagged as an outlier and removed from the training data.

Several new features were derived from existing data in order to investigate possible temporal or spatial relationships with $PM_{2.5}$. From the `time_stamp` associated with each observation, the `season` category was encoded. Seasons were based on the equinox and solstice dates for each year.

Additionally, the total count of trees surrounding a sensor was added to the dataset using the `species` data in the tree inventory data. The five most prevalent trees logged in the inventory consisted of Oak, Elm, Pecan, Crape Myrtle and Ashe Juniper. The individual counts of these trees within a 1.5 mile radius of each sensor was accounted for, and the sum of all other species was housed within an overall `Other` designation.

While the Purple Air PA-II sensors measure relative humidity, pressure, and temperature readings, additional precipitation readings were also integrated. Across the available weather stations, the nearest location containing rainfall data for a day was used to estimate total precipitation.

2.3 Training and Testing Data

After cleaning the dataset, a total of 20 features and 121,625 observations remained. The features primarily consist of climate, spatial, temporal, and tree related characteristics. An example of the dataset in given in Table 1.

An 80-20 random split of data was used for training and testing the models. K-fold cross-validation was used to segment the training set into smaller sets when evaluating the models. While leave-one-out cross validation was considered for use, it would have been more computationally expensive given the data and computing resource constraints.

3 Methodology

3.1 Predictive Modeling

For the task of predicting $PM_{2.5}$, our team introduced and investigated 4 predictive models. We focused on the following methods: K-nearest neighbors (KNN), Random Forest, XGBoost, and Temporal Convolutional Network (TCN). We note here that a key benefit of all of these model classes is a lack of distributional assumptions.

3.1.1 KNN. In KNN, a pre-specified parameter K and measure of distance $d(\cdot)$, determine the output for the data [22]. Given data $\mathbf{X} = \{x_1, \ldots, x_n\}$ and an instance x_i of interest, KNN determines the K nearest other instances $\{x_{j_1}, x_{j_2}, \ldots, x_{j_K}\}$ to x_i. KNN finds the subset of length K that satisfies

$$\sum_{j=1}^{K} d(x_i, x_j) = \min_{S} \sum_{x \in S} d(x_i, x)$$

where S is any other subset of the data (not including the instance of interest) with $|S| = K$. The necessity to choose a measure for distance diminishes the performance of KNN in a high-dimensional space due to increased sparsity in the data space [23]. KNN also relies on the assumption that data with similar features tend to have similar outputs. In the case of $PM_{2.5}$ prediction, we hypothesize that similar conditions in the covariates (e.g. weather and time) will tend to be associated with similar levels of $PM_{2.5}$. This framework makes KNN a good potential fit for modeling. Because the data follows normality constraints, each instance $x_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,d}\}$

Table 1: Feature and Target Variables

Variable Name	Data Type
Features	
Weather	
Humidity	Numeric
Temperature	Numeric
Pressure	Numeric
Altitude	Numeric
Precipitation	Numeric
Tree Type	
Oak	Numeric
Elm	Numeric
Pecan	Numeric
Crape Myrtle	Numeric
Ashe Juniper	Numeric
Other	Numeric
Tree Characteristics	
Total Diameter	Numeric
Canopy Coverage Percentage	Numeric
Sensor Location	
Latitude	Numeric
Longitude	Numeric
Time	
Hour	Numeric
Day of Week	Categorical (encoded)
Month	Categorical (encoded)
Season	
Fall	Binary
Spring	Binary
Summer	Binary
Winter	Binary
Target	
PM2.5	Numeric

(where d is the number of dimensions) was standardized into a corresponding z score. Then, Euclidian distance was selected as the distance metric. Next, a distance-based weight function was selected to give more weight to "nearer neighbors" in prediction. Finally, a grid search with 5-fold cross-validation. was performed on choices of $K \in \{1, 2, \ldots, 20\}$, with the elbow method being used to determine the best choice of K.

3.1.2 Random Forest. Tree-based models are well-suited for the task due to their interpretable framework for prediction, with visual representation being simple and intuitive compared to other model classes. A random forest was chosen rather than a single decision tree because of the tendency of ensemble methods to output predictions with decreased variance when compared to a single tree [24]. In a random forest, each tree is independently trained on a bootstrapped sample of the data [25]. In each tree, binary recursive partitioning is used to split the data into a hierarchical series of parent and child nodes. Each step selects the split that greedily minimizes a user-chosen function of the respective child nodes to group similar instances. Once N trees have been trained, the final prediction for a given instance x_i is the average of the N predictions on x_i. Crucially, this splitting method implies a hierarchical nature of feature importance that does not change with different realizations of the features. Although this may not hold in the data, it provides a simple top-level explanation for potentially important variables. Here, we opted to minimize the sum of squared deviations from the respective means of the child nodes created from a split. A grid search using 5-fold cross-validation was then performed on the following hyper-parameters: the number of trees in the forest (`n_estimators`), the maximum depth of each tree (`max_depth`), the minimum number of samples required to split at a given node (`min_samples_split`), and the minimum number of samples required to be in a leaf node (`min_samples_leaf`).

3.1.3 XGBoost. Although a random forest does well in decreasing variance, it does nothing about the potential contribution of bias to prediction error [26]. To address this issue, we used XGBoost [27], which aims to reduce prediction bias by sequentially training trees rather than independently training them. We note that, as a tree-based algorithm, XGBoost also assumes that the features hierarchically contribute to $PM_{2.5}$ levels, regardless of the actual feature values. In the XGBoost algorithm, each tree attempts to predict the residual of the previous tree (the first tree predicts the actual value). After training K trees in this manner, the final prediction for a given instance $\boldsymbol{x}_i$ is the sum of the predictions from each tree, all scaled by the shrinkage parameter or learning rate. This might look like

$$\hat{y}_i = \eta \sum_{k=1}^{K} f_k(\boldsymbol{x}_i)$$

where $\hat{y}_i$ is the final prediction of $PM_{2.5}$, η is the learning rate, and f_k is the prediction for the k-th tree. Each tree was again constructed with the objective of greedily minimizing the sum of squared deviations. Finally, a grid search with 5-fold cross validation was performed on the following parameters: the maximum depth of each tree (`max_depth`), the learning rate for weight updates (`learning_rate`), the number of trees constructed (`n_estimators`), the proportion of data to sample during bootstrapping (`subsample`), and the number of features used to train each tree (`colsample_bytree`).

3.1.4 TCN. Although the previous three models each have their own benefits, none of them explicitly account for potential temporal dependencies of the data. To address this, we introduce a temporal convolutional network (TCN) [28]. Intuitively, TCN combines two key aspects of interest: long-term memory capacity through dilated convolutions and convolutional layers for noise reduction. However, TCN also comes with a uniquely high risk of over-fitting compared to the previously discussed models and may require a large sample of data to capture temporal dependencies accurately. TCN includes an input layer, one or more hidden layers, and an output layer. It is trained for a user-specified number of epochs by, at each epoch, passing each instance in the training set through the input layer. Then, a series of linear computations using weights and

nonlinear activation functions in the hidden layers result in a final prediction, $\hat{y}_i$, for the observed $PM_{2.5}$ value, y_i, in the output layer. Finally, weights are updated using backpropagation. We use the popular choice $\text{ReLU}(x) = \max(0, x)$ for all activation functions for its simplicity and ability to handle issues such as vanishing gradients in hidden layers [29]. We use a TCN model with three temporal blocks, each of which has two 1-D convolutional layers, and pass the output of the three blocks to an average pooling layer, which is then passed to a dense layer. We use mean-squared error as the objective function, and the Adam [30] optimizer for optimization. To mitigate overfitting, we take multiple approaches. We adjust the kernel size of the convolutional layers, utilize dropout during training, and implement early-stopping when validation loss plateaus over $n = 10$ epochs. Finally, we implement a learning rate scheduler and utilize regularization. Then, we perform a grid search on the following hyper-parameters: the kernel size of the convolutional layers (`kernel_size`), the dropout rate (`dropout`), the initial learning rate passed to the Adam optimizer (`lr`), and the regularization parameter (`weight_decay`), and use the best model for $PM_{2.5}$ prediction.

3.2 Partial Dependence Plots

The partial dependence plot (PDP) is a global method used to interpret black-box models [31]. It shows the marginal effect of a feature on the predicted outcome of a model. The function is defined as:

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C)$$

where x_s are the features whose marginal effect on the predicted $PM_{2.5}$ we would like to know and X_C are the remaining features used in the machine learning model $\hat{f}$. The function can be estimated using:

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

We chose to utilize one-way PDPs where there is only one feature in the set S.

4 Results

In this section, we introduce the results for each model, along with key metrics and evaluation criteria. Then, we focus on potential interpretations of the predictions from the best model.

4.1 Model Results

Models were scored using three primary metrics: mean-squared error (MSE), mean-absolute error (MAE), and the coefficient of determination R^2. The MSE is defined as follows: let $\mathbf{Y} = \{y_1, \ldots, y_n\}$ and $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the observed target ($PM_{2.5}$) and features for n data points, respectively. Then, let $\hat{\mathbf{Y}} = \{\hat{y}_1, \ldots, \hat{y}_n\}$ be the predicted values from a constant model, whose inputs are the corresponding features. The MSE is defined as the average residual sum of squares:

$$\text{MSE} = \frac{1}{n} \cdot \text{SS}_{\text{residual}} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Moreover, the MAE is defined as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Finally, the R^2 coefficient is given by one minus the ratio of the residual sum of squares to the total sum of squares

$$R^2 = 1 - \frac{\text{SS}_{\text{residual}}}{\text{SS}_{\text{total}}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where $\bar{y}$ is the average observed $PM_{2.5}$ value.

MSE was used as a metric in order to penalize large errors, whereas MAE was used due to its robustness in the face of noise (where large errors may be inevitable). Because $PM_{2.5}$ rates appear to exhibit a large amount of variance in the data, MAE may be uniquely well-suited for local interpretations of prediction error. Moreover, since R^2 is a measure of the amount of variance in the data a model is able to capture, we evaluate it as a general guideline for overfitting and/or underfitting. Namely, the presence of large amounts of noise inherent to weather-related data suggests that a high R^2 value may be artificial, whereas a low value for R^2 may imply a necessity to increase model complexity. By using all 3 of these metrics, we assess two measures of prediction error, while also focusing on capturing a signal in the face of uncertainty. These metrics are displayed in Table 2.

Table 2: Model Evaluation Results

Model	MSE	MAE	R^2
KNN	25.03	3.24	0.41
Random Forest	16.14	2.51	0.62
XGBoost	12.98	2.13	0.69
TCN	15.41	2.50	0.64

All three metrics indicate that XGBoost performs the best. This is a surprising result, as XGBoost does not have a built-in long-term memory capacity like TCN, though it is likely the case that the TCN over-fit to the training data. Conversely, the KNN and random forest models both appear to suffer from high bias. Table 3 displays the optimized hyperparameters (obtained via grid search) from XGBoost modeling.

When interpreting the R-squared of our optimal model, we can say that 69% of the variability in our data can be estimated by our model.

Table 3: XGBoost Optimized Parameters

Parameter	Value
`colsample_bytree`	0.9
`learning_rate`	0.04
`max_depth`	11
`n_estimators`	800
`subsample`	0.9

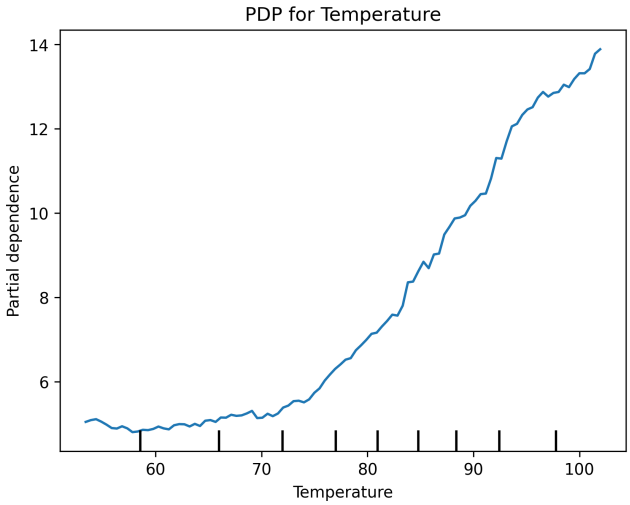


Figure 2: PDP of $PM_{2.5}$ on Temperature

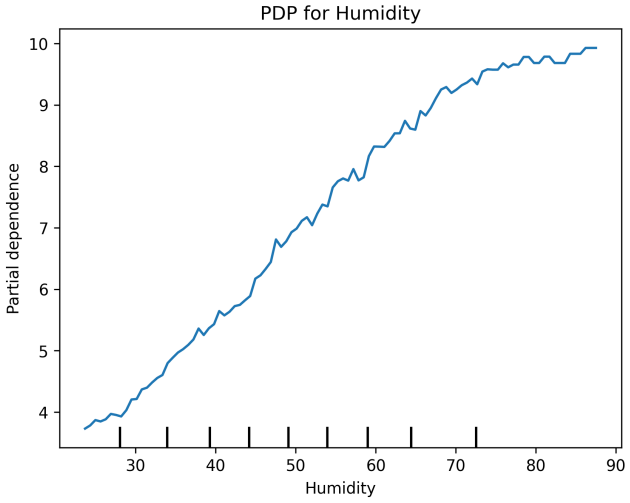


Figure 3: PDP of $PM_{2.5}$ on Humidity

4.2 Partial Dependence Plots

4.2.1 Weather Features. Figure 2 shows the PDP of temperature (in Fahrenheit) for the XGBoost model. The overall plot depicts a positive correlation between temperature and $PM_{2.5}$, beginning at around 70 °F, with $PM_{2.5}$ concentrations continually increasing in an approximately linear fashion. Below 70 °F, we observe a flatter trend line, or more minimal levels of association.

Figure 3 shows the PDP of average daily humidity level. The overall plot depicts a positive association between humidity levels and $PM_{2.5}$ concentrations, with a steadily linear trend line that begins to plateau around the 75% humidity level.

4.2.2 Tree Characteristics. Figure 4 shows the one-way PDP of canopy cover percentage using the XGBoost model. There were very few observations in the data with a canopy coverage percentage above 40%, so the model was unlikely to learn a meaningful prediction for this range. We see that there is a somewhat negative relationship between $PM_{2.5}$ and canopy coverage above this range, but there are inconsistencies in this trend.

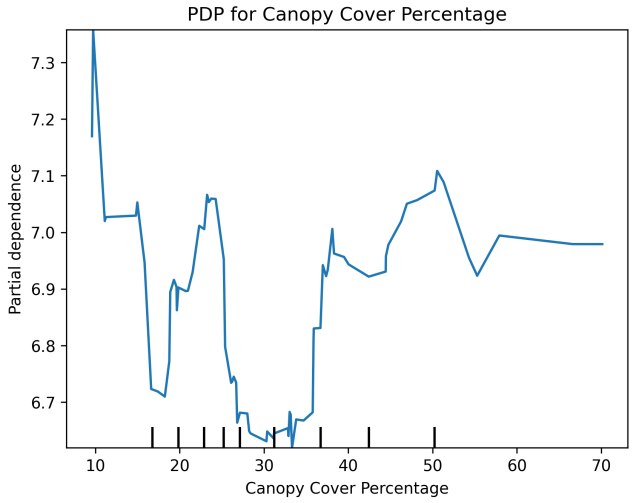


Figure 4: PDP of $PM_{2.5}$ on Canopy Cover Percentage

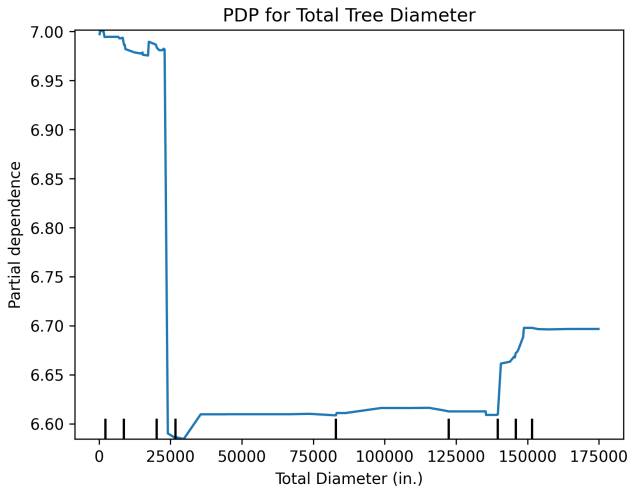


Figure 5: PDP of $PM_{2.5}$ on Total Tree Diameter

Figure 5 shows the PDP of total tree diameter for the XGBoost model. There were few observations in the data with total diameters between 25,000 and 150,000 inches. Thus, it is possible the model did not learn a meaningful prediction for this range. When analyzing the areas with total diameters outside of this range, areas with lower total diameter appear to have higher predicted $PM_{2.5}$ values compared to areas with a higher total diameter.

4.2.3 Temporal Variables. Figure 6 shows the one-way PDP plots of month and time. There appears to be a seasonal trend where winter and spring months are related to a higher predicted $PM_{2.5}$

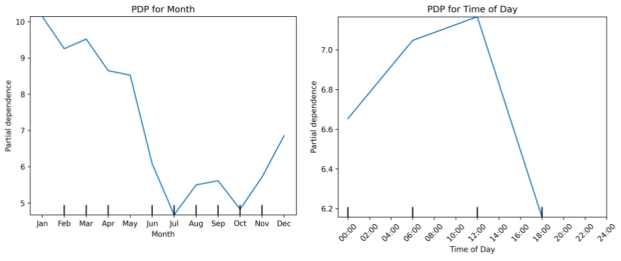


Figure 6: PDPs of $PM_{2.5}$ on Month and Time of Day

level than summer and early fall months in Austin. Additionally, we see that the late morning and early afternoon times are related to higher predicted $PM_{2.5}$ levels than the evening and night time.

4.2.4 Tree Types. Figures 7(a)-7(f) show PDPs for counts of the most common types of trees. The PDPs for counts of oak, elm, Ashe juniper, and crape myrtle trees show a slight downward trend. This could mean that there is an association between these types of trees and reduced levels of $PM_{2.5}$. Among the pecan species and remaining trees categorized under the remaining `other` category, the plots are flat. This might suggest there is no discernible trend when controlling for the other features. However, it is important to still consider the noise present in the data.

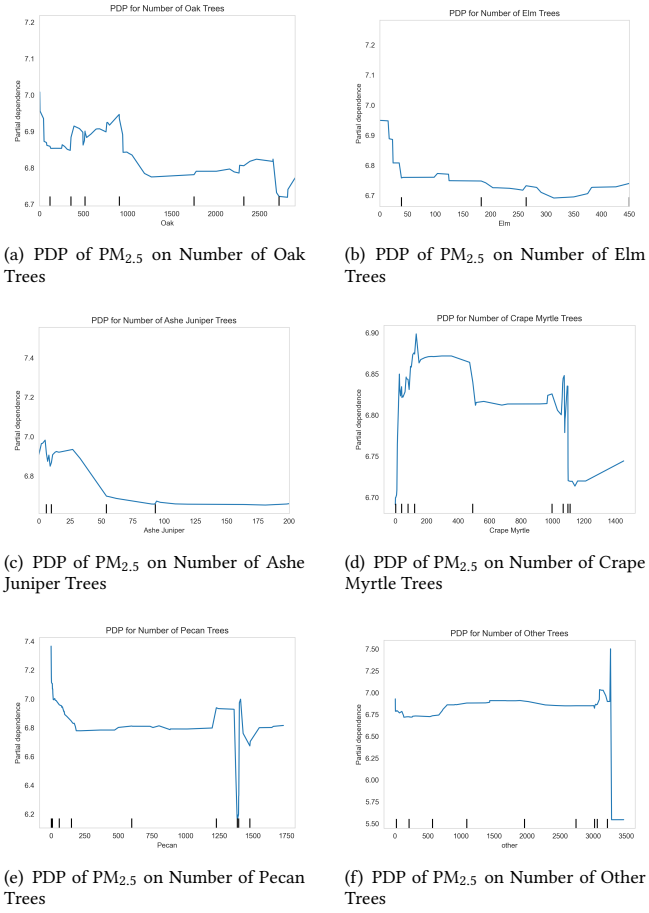
4.3 SHAP

Our exploration of the model also consisted of using SHAP (SHapley Additive exPlanations). For each observation, SHAP calculates the contribution of each feature to the model's expected prediction, relative to the mean of all of the model's predictions (base value) [32]. An advantage of this interpretability method is that we do not need to make strong assumptions to use it. Due to computational restrictions, these values were computed for a subset of the data. Figure 8 shows that temperature and humidity contribute the most to $PM_{2.5}$. Higher values of these features are correlated with higher $PM_{2.5}$. Following these two are the month and the indicator for the Spring season. Early months in the year are associated with higher $PM_{2.5}$, and the Spring indicator tends to have positive SHAP values when it is Spring and negative otherwise. Air quality varies with the weather and season, so these findings align with that.

5 Discussion

5.1 Weather

Through feature importance analysis, we discovered temperature, and humidity to be significant contributors to our model's decision making. These features were shown through partial dependence to be positively correlated with $PM_{2.5}$, where higher temperatures and humidity levels are associated with higher levels of $PM_{2.5}$. The association we found between humidity and $PM_{2.5}$ is in line with findings of Zalakeviciute et al., who observed higher readings of $PM_{2.5}$ as daily relative humidity increased within urban centers in Brazil. Notably, readings would plateau around the 70 percent humidity level [33]. The trend between $PM_{2.5}$ and temperature as displayed by our model showcases higher $PM_{2.5}$ levels associated



(a) PDP of $PM_{2.5}$ on Number of Oak Trees

(b) PDP of $PM_{2.5}$ on Number of Elm Trees

(c) PDP of $PM_{2.5}$ on Number of Ashe Juniper Trees

(d) PDP of $PM_{2.5}$ on Number of Crape Myrtle Trees

(e) PDP of $PM_{2.5}$ on Number of Pecan Trees

(f) PDP of $PM_{2.5}$ on Number of Other Trees

Figure 7: PDPs for the Most Common Tree Types

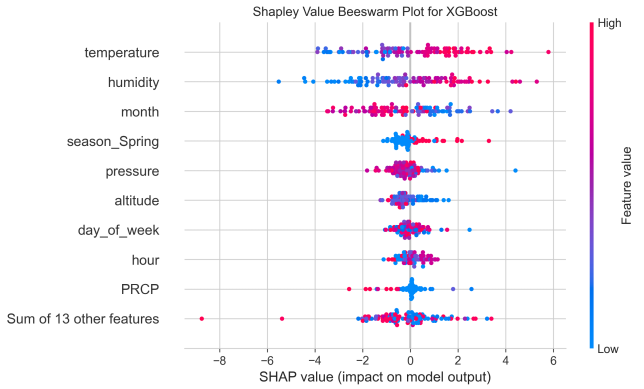


Figure 8: Beeswarm Plot of SHAP Values

with higher temperatures, similar to trends found by Kioumourtzoglou et al., who observed a stronger association between $PM_{2.5}$ and mortality in warmer cities [34].

5.2 Urban Forestry

The PDPs for 4 of the most common tree types—oak, elm, Ashe juniper, and crape myrtle—revealed slight negative trends between the counts of these types of trees and PM2.5 levels.

Additionally, the PDPs for canopy cover percentage and total tree diameter demonstrated a generally negative relationship between these variables and predicted $PM_{2.5}$. Based on these findings, we recommend the City of Austin to plant more trees—specifically oak, elm, Ashe juniper, or crape myrtle trees—and promote green spaces in places with higher $PM_{2.5}$. Additionally, since tree age is positively correlated with diameter [35], we urge the city to limit the removal of well-established trees.

5.3 Temporal Insights

Figure 6 showed the impact of month and time of day on $PM_{2.5}$ levels in Austin. The finding that predicted $PM_{2.5}$ levels are higher in winter than in other seasons is consistent with existing literature [36]. This is largely due to increased heating by combustion of wood and coal as well as motor vehicle use during the wintertime [37]. We recommend that the City of Austin promote alternate heating sources and improve its public transportation system to combat this issue. In addition, Figure 6 shows that $PM_{2.5}$ levels are high in the morning, which corresponds with rush hour traffic. This finding aligns with existing literature [38]. Our suggestion is to plant trees and minimize tree removal near these high-traffic areas in the city to potentially reduce the impact of pollution from personal vehicles.

5.4 Limitations

While we gained several valuable insights from our interpretability methods, certain PDPs such as the canopy cover percentage and Crepe Myrtle tree-type plots included large fluctuations. The noisiness of our data might be a contributing factor to these inconsistencies. Additionally, as described in Section 4.2.2, there were ranges of the features with limited observations. As a result, we must be cautious in our interpretation of these plots.

6 Conclusion

As the city of Austin continues to experience rapid population and industrial growth, air quality is an increasingly important factor in local health outcomes and lifestyles. Our work focuses on identifying and examining the key factors that affect $PM_{2.5}$ rates, which are critical in determining the AQI. We collect and use data containing several key indicators for $PM_{2.5}$, including weather conditions and details on urban forestry, with the aim of identifying any potential impact of the latter on $PM_{2.5}$ levels. We perform standard data cleaning and develop four machine learning models (KNN, Random Forest, XGBoost, and TCN) to predict $PM_{2.5}$. Then, we extract insights from the best-performing model (XGBoost) using common interpretability methods such as PDP plots and SHAP values. Through this analysis, we find that weather and seasonal changes affect $PM_{2.5}$ significantly, but the presence of trees can potentially reduce its concentration. Additionally, we find that areas of high total tree diameter are related to lower $PM_{2.5}$ levels, indicating the potential of urban forestry to combat high $PM_{2.5}$. Therefore, it could benefit the City of Austin to plant more trees and promote greenery in areas of low air quality, high motor vehicle traffic, and especially those with high temperatures and humidity levels, by specifically cultivating oak, elm, Ashe juniper, and crape myrtle trees.

7 Acknowledgment

Table 4: Contribution Table

Team Member	Contributions
Asmit	Modeling, model evaluation and drafting.
Berkeley	Pulling data, modeling and drafting.
Grace	Modeling, model evaluation and drafting.
Victor	Pulling data, modeling and drafting.
Vijetha	Modeling and drafting.

We used ChatGPT, Gemini, and Claude to generate code for our predictive modeling and interpretability method tasks [39][40][41]. We also used these large language models to refine our report's written content and generate LaTeX code.

References

[1] Shaolong Feng, Dan Gao, Fen Liao, Furong Zhou, and Xinming Wang. The health effects of ambient PM2.5 and potential mechanisms. *Ecotoxicology and Environmental Safety*, 128:67–74, 2016.

[2] California Air Resources Board. Inhalable Particulate Matter and Health (PM2.5 and PM10). https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health#:~:text=Also%2C%20children%20and%20infants%20are,9%2C300)%20each%20year%20in%20California.

[3] Zheng Duan, DU Fy, YD Yuan, YP Zhang, HS Yang, and WS Pan. Effects of PM2.5 exposure on Klebsiella pneumoniae clearance in the lungs of rats. *Chinese journal of tuberculosis and respiratory diseases*, 36(11):836–840, 2013.

[4] Siyuan Yang, Delin Fang, and Bin Chen. Human health impact and economic effect for PM2.5 exposure in typical cities. *Applied Energy*, 249:316–325, 2019.

[5] Kamal Jyoti Maji, Wei-Feng Ye, Mohit Arora, and SM Shiva Nagendra. PM2.5-related health and economic loss assessment for 338 Chinese cities. *Environment International*, 121:392–403, 2018.

[6] Federico Karagulian, Claudio A Belis, Carlos Francisco C Dora, Annette M Prüss-Ustün, Sophie Bonjour, Heather Adair-Rohani, and Markus Amann. Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. *Atmospheric environment*, 120:475–483, 2015.

[7] W Gene Tucker. An overview of PM2.5 sources and control strategies. *Fuel Processing Technology*, 65:379–392, 2000.

[8] Iasonas Stavroulas, Georgios Grivas, Panagiotis Michalopoulos, Eleni Liakakou, Aikaterini Bougiatioti, Panayiotis Kalkavouras, Kyriaki Maria Fameli, Nikolaos Hatzianastassiou, Nikolaos Mihalopoulos, and Evangelos Gerasopoulos. Field evaluation of low-cost PM sensors (Purple Air PA-II) under variable urban air quality conditions, in Greece. *Atmosphere*, 11(9):926, 2020.

[9] Thithanhthao Nguyen, Xinxiao Yu, Zhenming Zhang, Mengmeng Liu, and Xuhui Liu. Relationship between types of urban forest and PM2.5 capture at three growth stages of leaves. *Journal of Environmental Sciences*, 27:33–41, 2015.

[10] Caleb Pritchard. Inhalable Particulate Matter and Health (PM2.5 and PM10). https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health#:~:text=Also%2C%20children%20and%20infants%20are,9%2C300)%20each%20year%20in%20California., March 2024.

[11] City of Austin Urban Forestry Board. Austin's Urban Forest Plan: A Master Plan for Public Property. https://www.austintexas.gov/sites/default/files/files/Parks/Forestry/AUFP_Final_DRAFT_01-07-14_No_Appendices.pdf, January 2014.

[12] Qian Di, Itai Kloog, Petros Koutrakis, Alexei Lyapustin, Yujie Wang, and Joel Schwartz. Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. *Environmental Science & Technology*, 50(9):4712–4721, 2016.

[13] Xuefei Hu, Jessica H Belle, Xia Meng, Avani Wildani, Lance A Waller, Matthew J Strickland, and Yang Liu. Estimating PM2.5 concentrations in the conterminous United States using the random forest approach. *Environmental Science & Technology*, 51(12):6936–6944, 2017.

[14] Qingyang Xiao, Howard H Chang, Guannan Geng, and Yang Liu. An ensemble machine-learning model to predict historical PM2.5 concentrations in China from satellite data. *Environmental science & technology*, 52(22):13260–13269, 2018.

[15] Karin Ardon-Dryer, Yuval Dryer, Jake N Williams, and Nastaran Moghimi. Measurements of PM2.5 with PurpleAir under atmospheric conditions. *Atmospheric Measurement Techniques*, 13(10):5441–5458, 2020.

[16] Ethan Breinholt. How do purpleair sensors compare to regulatory particulate matter sensors? https://community.purpleair.com/t/q-how-do-purpleair-sensors-compare-to-regulatory-particulate-matter-sensors/810, June 2022.

[17] Karoline K Barkjohn, Brett Gantt, and Andrea L Clements. Development and application of a United States-wide correction for PM2.5 data collected with the PurpleAir sensor. *Atmospheric Measurement Techniques*, 14(6):4617–4637, 2021.

[18] Lance Wallace. Intercomparison of PurpleAir sensor performance over three years indoors and outdoors at a home: bias, precision, and limit of detection using an improved algorithm for calculating PM2.5. *Sensors*, 22(7):2755, 2022.

[19] National Oceanic and Atmospheric Administration. Climate data online data tools. https://www.ncei.noaa.gov/cdo-web/datatools.

[20] City of Austin Development Services Department. Tree canopy 2022, 2022.

[21] City of Austin Development Services Department. Tree inventory, March 2020.

[22] Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[23] Radovanovic, Milos and Nanopoulos, Alexandros and Ivanovic, Mirjana. Nearest neighbors in high-dimensional data: The emergence and influence of hubs. *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, 382:109, 06 2009.

[24] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[25] Breiman, Leo. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[26] Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome. *Random Forests*, pages 587–604. Springer New York, New York, NY, 2009.

[27] Chen, Tianqi and Guestrin, Carlos. XGBoost: A Scalable Tree Boosting System. pages 785–794, 08 2016.

[28] Lea, Colin and Vidal, René and Reiter, Austin and Hager, Gregory. Temporal Convolutional Networks: A Unified Approach to Action Segmentation. 08 2016.

[29] Glorot, Xavier and Bordes, Antoine and Bengio, Yoshua. Deep Sparse Rectifier Neural Networks. In Gordon, Geoffrey and Dunson, David and Dudík, Miroslav, editor, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[30] Kingma, Diederik and Ba, Jimmy. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, 12 2014.

[31] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[32] Lundberg, Scott. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[33] Zalakeviciute, Rasa and López-Villada, Jesús and Rybarczyk, Yves. Contrasted effects of relative humidity and precipitation on urban PM2.5 pollution in high elevation urban areas. *Sustainability*, 10(6):2064, 2018.

[34] Kioumourtzoglou, Marianthi-Anna and Schwartz, Joel and James, Peter and Dominici, Francesca and Zanobetti, Antonella. PM2.5 and mortality in 207 US cities: Modification by temperature and city characteristics. *Epidemiology*, 27(2):221–227, 2016.

[35] Bingham, B.B. and Jr..Sawyer, J.O. Canopy structure and tree condition of young, mature, and old-growth Douglas-fir/hardwood forests, Jan 1992.

[36] Naizhuo Zhao and Ying Liu and Jennifer K. Vanos and Guofeng Cao. Day-of-week and seasonal patterns of PM2.5 concentrations over the United States: Time-series analyses using the Prophet procedure. *Atmospheric Environment*, 192:116–127, 2018.

[37] Courtney A. Gorin, Jeffrey L. Collett Jr. and Pierre Herckes. Wood Smoke Contribution to Winter Aerosol in Fresno, CA. *Journal of the Air & Waste Management Association*, 56(11):1584–1590, 2006.

[38] J I Levy and E A Houseman and J D Spengler and P Loh and L Ryan. Fine particulate matter and polycyclic aromatic hydrocarbon concentration patterns in Roxbury, Massachusetts: a community-based GIS analysis. *Environmental Health Perspectives*, 109(4):341–347, 2001.

[39] OpenAI. ChatGPT: OpenAI Language Model. https://openai.com/chatgpt, 2024.

[40] Google DeepMind. Gemini: Advanced AI Model. https://gemini.google.com/, 2024.

[41] Anthropic. Claude: AI Assistant. https://www.anthropic.com/claude, 2024.