JOURNAL OF
EMERGING INVESTIGATORS

Article

Environmental contributors of asthma via explainable AI: Green spaces, climate, traffic & air quality

Andrew Chen[1], Shreya Parchure[2]
[1] Urbana High School, Ijamsville, Maryland
[2] University of Pennsylvania, Philadelphia, Pennsylvania

SUMMARY

Asthma, a chronic respiratory disease affecting 28 million Americans, imposes a substantial burden on public health, with 986,453 emergency department visits reported for asthma in 2020. While prior research has examined the impact of green spaces, climate, traffic, and outdoor air quality (GCTA) on asthma prevalence, these factors have largely been studied separately. Despite their importance individually, their combined and interactive effects remain less understood. We hypothesized that the number of asthma-related emergency department visits can be predicted by multimodal factors such as GCTA. This study applied three machine learning algorithms—k-nearest neighbors (KNN), Extreme Gradient Boosting (XGBoost), and random forest regression (RFR)—to analyze environmental, health, and traffic data, optimizing model performance to achieve high prediction accuracy in understanding asthma prevalence. Using explainable AI, we demonstrated that GCTA interact in intricate ways, collectively shaping asthma prevalence. Our findings underscored the significance of GCTA as key contributors to asthma risk, with green space effects being the most significant and complex, as green spaces appeared to mitigate air pollution but also potentially increased allergen exposure. This study greatly advanced our understanding of the interplay of green cover, regional climate, and air pollution as it relates to urban health. These insights provided actionable guidance for urban planning by highlighting the need to increase green space coverage strategically, manage traffic-related emissions, and consider localized climate conditions—ultimately supporting more targeted interventions to mitigate respiratory disease burden in urban populations.

INTRODUCTION

Asthma, a chronic respiratory disease, affects 28 million people in the United States, representing about 1 in 12 individuals (1,2). In 2020, asthma led to 94,560 hospital inpatient discharges and 986,453 emergency department visits, highlighting the severe impact of this chronic condition on daily life (3). Given the severe impact of asthma, there is a strong need to predict the number of asthma-related emergency department visits and prevent asthma cases through public health initiatives. One such impactful method is urban planning, which focuses on environmental factors that could reduce asthma rates (4).

Previous studies have primarily focused on the impact of poor outdoor air quality on increasing asthma severity, and some studies have examined the complex relationship between urban green spaces and asthma prevalence (5,6). Climate factors such as extreme heat and cold can increase the risk of asthma (7). Additionally, infrastructure planning contributes to differences in vehicular density, which has a complex interplay with other urban planning factors that can then influence asthma (8,9). However, no studies to date have combined all these factors to explore their collective and interactive impact on asthma rates. Considering these environmental factors in isolation may overlook important interactions and confounding effects. Assessing their combined effects therefore allows for a more holistic analysis of urban health risks.

A powerful methodology to study the impact of environmental factors on asthma is using machine learning (ML) on large publicly available datasets that contain this information. We employed ML algorithms—including k-nearest neighbors (KNN), Extreme Gradient Boosting (XGBoost) and random forest regression (RFR)—to conduct the analysis and ensure robust, consistent insights across different modeling approaches (10–12). We selected these three algorithms for their complementary strengths in handling nonlinear relationships, capturing complex feature interactions, and providing interpretable measures of feature importance. KNN, a non-parametric, instance-based algorithm, interprets local feature relationships using proximity and similarity (10). XGBoost captures complex patterns with high accuracy through ensemble-based gradient boosting, minimizing overfitting and enabling feature importance analysis (11). RFR reduces variance and generalizes well with multiple decision trees, providing robust feature interpretation (12). We applied these three ML algorithms to datasets encompassing environmental, health, and traffic data (13–18). New York City (NYC) serves as an excellent case study for analyzing asthma prevalence due to its diverse urban landscape, variation in outdoor air quality, and extensive green spaces amid dense traffic and residential areas (19–21). In this study, the analysis of feature importance used SHapley Additive exPlanations (SHAP) values—a key explainable AI technique (22). Explainable AI is a broader field focused on developing tools and techniques to help humans make sense of decisions made by AI models (22). SHAP is a method for interpreting the predictions of machine learning models by assigning a score to each feature for each data point, representing how much that feature contributed to the final prediction (23). We chose Explainable AI because it enhances transparency in

https://doi.org/10.59720/25-031

Feature(s) Removed from Analysis	R^2	MSE
None	0.9010	0.0750
Green Space: vegetative cover percentage (Veg_Cover), tree canopy cover (Tree_Cover)	0.8677	0.1002
Climate: surface temperature (Surf_Temp)	0.8889	0.0841
Traffic: vehicle miles traveled (Vehicle_Miles), truck miles traveled (Truck_Miles)	0.8954	0.0792
Air Quality: PM, NO_2, O_3	0.8728	0.0964

Table 1: Prediction performance of Extreme Gradient Boosting (XGBoost) after removing features from green space, climate, traffic and air quality category. A decline in performance was observed with the removal of each category of data.

model decisions, and SHAP because it provides instance-level feature importance explanations.

By applying machine learning models to integrated environmental and health datasets, we sought to understand how they interact to influence asthma outcomes. We hypothesized that the prevalence of asthma-related emergency department visits (EDVs) could be predicted by the combined effects of outdoor air quality, availability of green spaces, climate conditions (e.g., surface temperature), and traffic data. We aimed to uncover nuanced, interpretable patterns by analyzing these environmental variables collectively using explainable AI techniques such as SHAP—patterns that may be overlooked by traditional statistical methods or analyses of individual factors. We recommended that reducing urban asthma risks requires a multifaceted approach that addresses air quality improvements, carefully designed green space interventions, and traffic considerations. These findings provided actionable insights for policymakers to develop targeted strategies that mitigate asthma risks through integrated urban planning and tailored health initiatives.

RESULTS

We investigated whether the prevalence of asthma-related EDVs in New York City can be predicted by the combined effects of green spaces, climate, traffic, and outdoor air quality (GCTA). We applied three machine learning algorithms—KNN, XGBoost, and RFR—to an integrated dataset encompassing environmental, health, and traffic data sourced from NYC.gov public repositories (13–18). We included eight features in this investigation: vegetative cover percentage, tree canopy cover, surface temperature, particulate matter (PM), nitrogen dioxide (NO_2), Ozone (O_3), vehicle miles traveled, and truck miles traveled (13–17). The target variable was asthma-related EDVs (18).

After data preprocessing, we used the XGBoost model to evaluate feature contributions and assess the effectiveness of our feature selection, leveraging its ability to generate interpretable importance scores. We set all hyperparameters to default values and grouped eight features into four categories—green space, climate, traffic, and air quality (**Table 1**). We calculated the feature importance scores, where higher scores indicate a greater influence on prediction (**Figure 1**). Grouping features by category revealed that green space was the most influential data category with an average importance score of 0.2447, followed by traffic (0.1872), climate (0.0935), and air quality contributing the least with an average importance score of 0.0142 (**Figure 1**). These results showed that green space was the most important contributor to predicting asthma-related EDVs.

To further investigate the role of each data category in

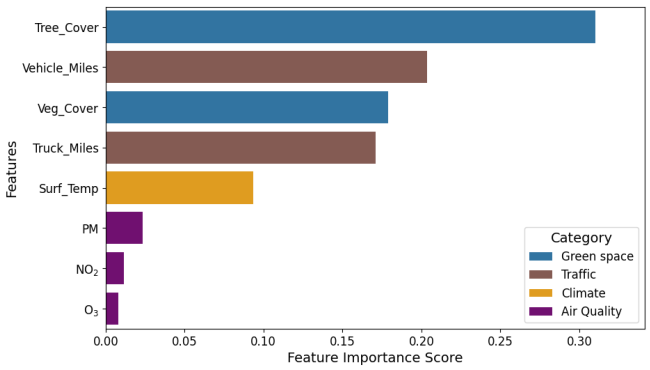

Figure 1: Extreme Gradient Boosting (XGBoost) feature importance before hyperparameter tuning. XGBoost model before hyperparameter tuning achieved a Mean Squared Error (MSE) of 0.0750 and an R^2 of 0.9010 on the training dataset. Features were ranked from the most important feature (Tree_Cover at the top) to least important (O3 at the bottom), and the importance score of each feature is given by the x-axis. Features were categorized as either green space (blue), traffic (brown), climate (yellow), or air quality (purple). The datasets were sourced from NYC.gov public repositories (13-18).

prediction, we conducted experiments by systematically removing features from specific categories and evaluating whether the model's performance declined compared to the prediction results using all features. This analysis provided deeper insights into the relative contributions of each category to the overall prediction results. We observed a decline in performance with the removal of each category of data (**Table 1**). The largest decline occurred when the green space category was removed (R^2 of 0.8677 without green space vs. R^2 of 0.9010 with green space; **Table 1**). This observation aligned with the feature importance scores, indicating that features from all four categories contributed to the prediction, with green space being the most significant contributor (**Figure 1**).

To ensure optimal model performance and validate the robustness of our results across different algorithms, we performed hyperparameter tuning for all three algorithms. Hyperparameter tuning involves selecting the best configuration of model settings that guide how it learns from data. The hyperparameter tuning process included a pre-search process followed by the Grid Search and cross-validation. The pre-search process revealed that only one hyperparameter significantly impacted the prediction accuracy of the KNN model: the number of neighbors (n_neighbors) (**Table 2**). The results of the pre-search process revealed that five hyperparameters had the most impact on the prediction accuracy of the XGBoost model (**Table 2**). These key hyperparameters were the number of estimators (n_estimators), the maximum depth of a tree (max_depth), the minimum sum of weights of observations required in a child node (min_child_weight), lambda (L2 regularization), and alpha (L1 regularization) (**Table 2**). The pre-search analysis identified four hyperparameters as the most influential on the prediction accuracy of the RFR model (**Table 2**). These critical hyperparameters included the number of estimators (n_estimators), the maximum depth of splits allowed in each tree (max_depth), the number of features considered for splitting at each node (max_features), and the minimum number of

https://doi.org/10.59720/25-031

Model	Hyperparameters Value	Training dataset		Test dataset	
		R^2	MSE	R^2	MSE
KNN	n_neighbors = 3	0.8870	0.0700	0.8401	0.1211
RFR	n_estimators = 30 max_depth = 10 max_features = 8 min_samples_split = 5	0.9637	0.0225	0.8803	0.0906
XGBoost	n_estimators = 25 max_depth = 5 min_child_weight = 12 lambda = 5 alpha = 0	0.9599	0.0249	0.9104	0.0679

Table 2: Hyperparameters value and prediction performance on training dataset and test dataset across K-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGBoost) and Random Forest Regression (RFR) models. The KNN model's final predictions results for EDVs showed lowest predictive accuracy with an MSE of 0.1211 and an R^2 of 0.8401. XGBoost achieved the best performance with an MSE of 0.0679 and an R^2 of 0.9104, followed closely by RFR with an MSE of 0.0906 and an R^2 of 0.8803. XGBoost model demonstrated strong and comparable results on both the training dataset and the testing dataset, both datasets achieved R2 values above 0.9.

samples required in a node for a split to occur (min_samples_split) (**Table 2**). We conducted the Grid Search and cross-validation on these selected hyperparameters, and identified the optimal hyperparameter values (**Table 2**).

The KNN model's final predictions results for EDVs showed lowest predictive accuracy with an MSE of 0.1211 and an R^2 of 0.8401 (**Table 2**). XGBoost achieved the best performance with an MSE of 0.0679 and an R^2 of 0.9104, followed closely by RFR with an MSE of 0.0906 and an R^2 of 0.8803 (**Table 2**). Based on its superior predictive capability, we selected the XGBoost model for further feature importance analysis.

The feature importance rankings changed after hyperparameter tuning, as adjustments to the model parameters influenced how features were evaluated and interacted with one another. Post-tuning rankings are more accurate and reliable, reflecting the optimized model's ability to generalize and capture relationships in the data, while pre-tuning rankings offer initial insights for feature selection (24). We calculated the feature importance scores again using the optimized XGBoost model (**Figure 2**). Compared to the scores before hyperparameter tuning, the rankings shifted and were more evenly distributed among the eight features, confirming that all features contributed meaningfully to the prediction (**Figure 1** and **Figure 2**). Interestingly, the most important contributors before and after tuning both belonged to the green space category—tree canopy cover before tuning, and vegetative cover percentage after tuning (**Figure 1** and **Figure 2**). Similarly, the least important contributor remained the same—O_3 (**Figure 1** and **Figure 2**). Based on these results, we concluded that green space was consistently the most important factor in EDVs prediction, while air quality had the least impact, demonstrating the robustness of these findings regardless of hyperparameter tuning.

To further understand each data features' contributions to these predictions, we applied explainable AI techniques, focusing on the XGBoost and RFR models. We used SHAP values to interpret the predictions, enabling a detailed understanding of how each feature contributed to the predicted asthma-related EDVs values for individual data points (**Figure 3** and **Figure 4**). Red data points indicated that a feature's value was relatively high for a specific instance, while blue points represented relatively low values. The x-axis shows the SHAP value, which reflects the magnitude and direction of a feature's impact on the prediction—positive values indicate increased predicted EDVs, while negative values indicate a reduction. The SHAP values allowed us to understand not just which features matter most, but how they influence predictions. The features were ranked on the y-axis based on their importance. For the best-performing model, XGBoost, the SHAP value analysis of feature importance aligned closely with the feature importance scores.

The color scheme in SHAP beeswarm plots revealed that three features had a relatively direct association with predicted EDVs values (**Figure 3** and **Figure 4**). Lower PM and NO_2 levels were strongly associated with lower predicted EDVs values, while lower Vehicle_Miles was also associated with higher predicted EDVs values. The distribution of points in the SHAP beeswarm plot provided additional statistical insights. For the best-performing model, XGBoost, lower Veg_Cover instances (blue points) extended far to the left, while lower Tree_Cover instances extended far to the right (**Figure 3**). This indicated that less vegetation strongly reduced predicted EDVs values, whereas less tree coverage increased predicted EDVs values. For PM, a dense cluster of high PM level instances (red points) appeared with small but positive SHAP values (**Figure 3**). In contrast, lower PM instances extended further to the left, suggesting that low PM had a stronger negative impact on predicted EDVs values than high PM had a positive impact (**Figure 3**). Surf_Temp and Truck_Miles exhibited a wide distribution of SHAP values, with red and blue points appearing on both sides of the SHAP axis, indicating complex interactions with other features and suggesting that these features did not consistently increase or decrease predicted EDVs values (**Figure 3**). O_3 had the least impact on predicted EDVs values, with a tightly clustered distribution near the center (**Figure 3**). Based on these results, we determined that the presence of green spaces was the most

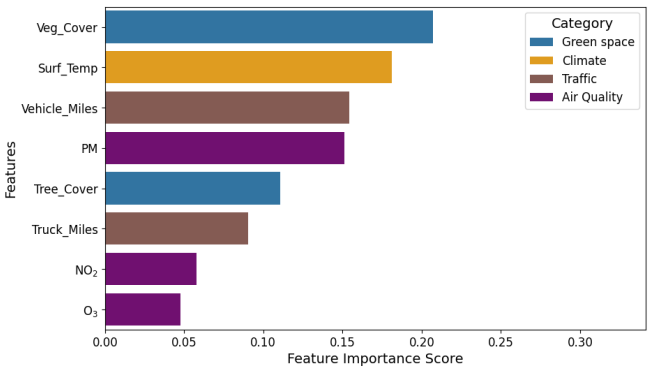

Figure 2: XGBoost feature importance after hyperparameter tuning. XGBoost model after hyperparameter tuning achieved an MSE of 0.0249 and an R^2 of 0.9599 on the training dataset. The feature importance rankings changed after hyperparameter tuning, as adjustments to the model parameters influenced how features were evaluated and interacted with one another. The graph is showing the most important feature (Veg_Cover) to least important (O3) and the importance score of each feature. Features were categorized as either green space (blue), traffic (brown), climate (yellow), or air quality (purple). The datasets were sourced from NYC.gov public repositories (13-18).

https://doi.org/10.59720/25-031

important contributor to EDVs prediction results, exhibiting a complex and context-dependent influence. In contrast, air quality showed a relatively direct association with EDVs but had the least impact on the prediction outcomes.

From the SHAP beeswarm plot of the XGBoost model, the top four most important features represented the four categories of factors examined in this study concerning asthma prevalence: GCTA (**Figure 3**). This observation confirmed these four factors' critical roles in achieving the predictive accuracy, as their prominence among the top-ranked features suggested that each contributes uniquely and significantly to the model's ability to predict EDVs values.

DISCUSSION

In this study, our findings highlighted that GCTA were all meaningful factors influencing asthma prevalence. The prevalence of asthma-related EDVs could be predicted by the combined effects of GCTA. The robustness of the prediction results stemmed from the use of a well-optimized machine learning model, where hyperparameter tuning played a critical role in enhancing performance (25). With a robust model in place, feature importance analysis provided fundamental insights into the relative influence of various factors on prediction outcomes. Beyond overall feature importance, a deeper examination using an explainable AI technique revealed intriguing patterns in individual feature contributions, highlighting their unique and sometimes context-dependent effects on EDVs prediction.

For the XGBoost model, lambda and alpha settings suggested that the model prioritizes reducing overfitting through L2 regularization without enforcing sparsity or feature elimination via L1 regularization. The absence of strong L1 regularization indicated that the model did not benefit from penalizing or eliminating individual features, further supporting the conclusion that the predictive signal was distributed across multiple environmental features rather than being concentrated in a few dominant ones. This aligned with our earlier SHAP analysis, which showed meaningful contributions from each of the GCTA factors, reinforcing the importance of a multifactorial approach in predicting asthma-related EDVs.

When comparing the XGBoost prediction accuracy of the training dataset with the performance on the testing dataset, both achieved R^2 values above 0.9. This indicated that the trained XGBoost model successfully overcame overfitting—a common issue in machine learning where a model performs well on training data but fails to generalize to new, unseen data.

Notably, the maximum features per tree matched the total number of features in the training dataset for the RFR model, which implied that all input features contributed meaningfully to the model's predictive capabilities. This finding was consistent with the findings from XGBoost regularization parameters setting, highlighting that the results were not reliant on any single feature but instead stemmed from the combined influence of all selected features. When comparing the performance of the RFR model on the training dataset and the testing dataset, the testing dataset showed a noticeable decline in performance, with an R^2 of 0.9637 on training dataset and an R^2 of 0.8803 on the testing dataset. This decline in performance on the testing dataset indicated that the RFR model experienced a certain degree of overfitting. In our experiment, XGBoost therefore outperformed RFR

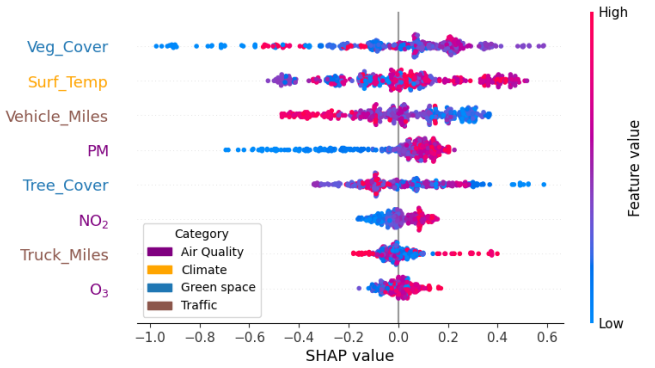

Figure 3: XGBoost SHapley Additive exPlanations (SHAP) beeswarm plot. The graph is showing the SHAP values derived from XGBoost model after hyperparameter tuning. The features were ranked on the y-axis according to their importance. Red data points indicated that a feature's value was relatively high for a specific instance, while blue points represented relatively low values. The x-axis (SHAP value) indicated the magnitude of a feature's impact on the final prediction, with higher values reflecting a greater influence. The top four most important features identified were Veg_Cover, Surf_Temp, Vehicle_Miles, and PM. These features represented the four categories of factors examined in this study concerning asthma prevalence: green space, climate, traffic, and air quality.

in mitigating overfitting, and this is likely due to XGBoost's built-in regularization techniques, such as L1 (alpha) and L2 (lambda) regularization, which penalize complex models and large feature weights (11). Our hyperparameter tuning process further optimized the regularization parameters setting for the XGBoost model. In contrast, RFR relies solely on averaging predictions from multiple decision trees, which can reduce variance but lacks explicit regularization mechanisms (12). This makes RFR more prone to overfitting, especially when the model attempts to fit noise present in the training data.

Beyond hyperparameter tuning, we analyzed which features most influenced model predictions. Green space consistently emerged as the top contributors, with Tree_Cover leading before tuning and Veg_Cover taking precedence after tuning. This shift resulted from the tuned model better balancing contributions from different green space variables. Notably, and contrary to expectations, O_3 was consistently ranked lowest in feature importance, despite prior literature suggesting that O_3 was a significant factor in asthma prevalence (5). This might be due to its relatively low spatial variability across NYC regions, especially compared to other air quality features like PM, which could vary near traffic corridors or industrial zones (19). When a feature showed little variation across regions, models such as XGBoost might find it less useful for distinguishing differences in EDVs rates. Importantly, this did not suggest that O_3 lacks relevance to asthma risk in general. Rather, it indicated that within the scope of this region-based prediction task, O_3 contributed less to model performance than the other examined features. This interpretation was supported by the SHAP analysis, where the SHAP values for O_3 were tightly clustered around the center, indicating they did not contribute to higher or lower predicted EDVs outcomes. It should also be noted that low variance does not always imply low predictive importance; features with limited variability can still be important if they capture subtle but meaningful patterns (26). In this case, however, the low

https://doi.org/10.59720/25-031

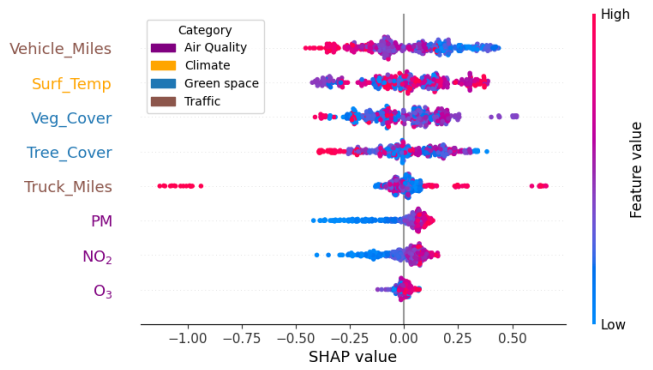

Figure 4: Random forest regression (RFR) model's SHAP beeswarm plot. The graph is showing the SHAP values derived from RFR model after hyperparameter tuning. The features were ranked on the y-axis according to their importance. Red data points indicated that a feature's value was relatively high for a specific instance, while blue points represented relatively low values. The x-axis (SHAP value) indicated the magnitude of a feature's impact on the final prediction, with higher values reflecting a greater influence. The top six most important features identified were Vehicle_Miles, Surf_Temp, Veg_Cover, Tree_Cover, Truck_Miles and PM. These features represented the four categories of factors examined in this study concerning asthma prevalence: traffic, climate, green space, and air quality.

importance of O_3 likely reflected its limited influence in our specific study parameters rather than its overall influence on asthma outcomes.

The comparison between traffic-related features (Vehicle_Miles and Truck_Miles) and climate-related features (Surf_Temp) revealed an intriguing dynamic before and after hyperparameter tuning. Before tuning, traffic-related features collectively ranked higher in importance than climate-related factors. However, after tuning, Surf_Temp rose to become the second most important feature, surpassing Vehicle_Miles while Truck_Miles dropped even further in importance. This change suggested that the optimized model placed greater emphasis on climate-related factors, recognizing their stronger impact on prediction performance compared to traffic metrics (24). The sudden increase in the importance of PM after hyperparameter tuning reflected how tuning adjustments could alter the model's interpretation of feature contributions. Regularization might have suppressed the dominance of certain features, such as Tree_Cover, allowing previously less dominant features like PM to gain prominence. Additionally, hyperparameter tuning likely reduced overfitting, helping the model better generalize and more accurately assess PM's contribution to prediction outcomes.

We observed that the feature importance score difference was narrower, and feature contributions were more evenly distributed after hyperparameter tuning. This reflected a refined understanding of feature contributions by the XGBoost model, suggesting better generalization as it relied on a broader range of variables rather than overemphasizing a few dominant ones. The even distribution also indicated improved recognition of interactions among features that were less apparent before tuning.

The SHAP values revealed that air quality exerted a more direct influence on asthma prevalence; whereas, green space, climate, and traffic had more complex and nuanced effects. For example, while green spaces may reduce air pollution and encourage physical activity, they may also introduce allergens, such as pollen, which can trigger asthma symptoms in sensitive individuals (27). For XGBoost, the best-preforming model, an intriguing observation was that areas with low asthma prevalence were associated with low vegetative cover, while areas with high asthma prevalence were associated with low tree canopy cover. A possible explanation was that general vegetative cover might increase exposure to pollen, a known asthma trigger, whereas tree canopy cover is more effective than general vegetation at improving air filtration and reducing pollutants, which are critical factors in mitigating asthma symptoms. Another observation was the high EDVs associated with low vehicle miles traveled. Possible explanations for this included economically disadvantaged areas having limited access to healthcare and lower insulation from pollution, as well as reduced urban infrastructure leading to direct exposure to pollutants (8,9).

Finally, the similarity between the SHAP beeswarm plots for the XGBoost and RFR models further validated the robustness of the findings, demonstrating that the results were not dependent on the specific machine learning algorithm employed. We also observed discrepancies in the feature importance rankings shown in the SHAP beeswarm plots. While air quality consistently ranked lowest in importance, the RFR model ranked Vehicle_Miles and Surf_Temp higher than green space features. However, due to the overfitting tendencies of RFR, we placed greater emphasis on the XGBoost results and maintained our conclusion that green space was the most important contributor to EDVs prediction (28).

While this study focused specifically on environmental predictors of asthma-related EDVs rates, we acknowledge that many other relevant factors influence asthma outcomes. Variables such as housing conditions, indoor air quality, access to healthcare, socioeconomic status, and individual behavioral factors were well-documented contributors to asthma prevalence and severity (9). These features were not included in this analysis because our primary objective was to investigate the role of environmental variables. We recommended that future research explore these additional dimensions to build a more comprehensive understanding of asthma risk, particularly in studies aiming to model health outcomes across different population groups or geographic contexts.

Our findings highlighted that GCTA were all meaningful factors influencing asthma prevalence, thereby providing critical insights for urban planning initiatives. These insights could inform strategies such as optimizing the design of parks and green spaces and enhancing public transportation systems to improve environmental health. Additionally, this study uncovered the intricate and interdependent relationships between green space, climate, and traffic, revealing their collective influence on asthma prevalence. Future research should delve deeper into these interactions to uncover more nuanced relationships among the four studied categories.

The findings from this study, based on environmental and health data from New York City, are likely applicable to other large urban centers with comparable infrastructural complexity, population density, and environmental stressors. Cities such as Toronto may exhibit similar relationships between green space, traffic-related air pollution, and asthma-

https://doi.org/10.59720/25-031

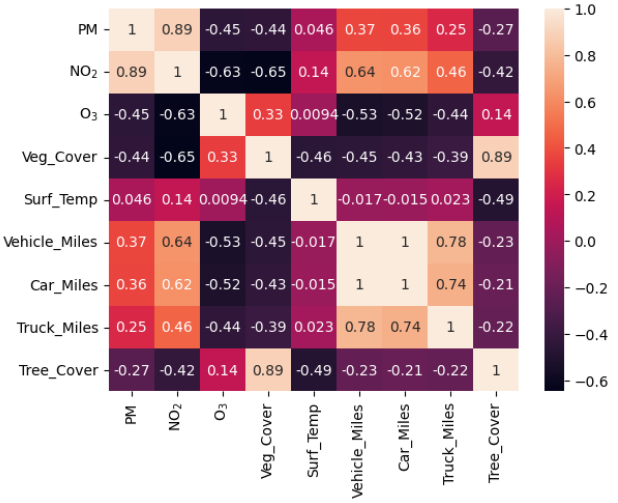

Figure 5: Correlation matrix data from the nine originally selected features. The graph is showing the correlation coefficients between nine features sourced from datasets of NYC.gov public repositories (13-18). The correlation coefficients were calculated to show how strongly two features are related. A value close to 1 means they increase together, close to -1 means one increases while the other decreases, and around 0 means there is little or no relationship. The x-axis and y-axis represented the nine features originally selected for this study, and each cell at the intersection of two features showed their correlation coefficient. The color bar on the right indicated the gradient scale of correlation strength, with tan representing positive correlation, black representing negative correlation, and purple indicating little or no correlation.

related EDVs (6). However, generalizing these findings to rural or less densely populated areas requires caution. The structure and distribution of environmental exposures differ in these contexts, potentially altering the relative importance of predictors.

MATERIALS AND METHODS

Data Sources

The datasets used in this study were sourced from publicly available repositories on NYC.gov (13–18). These datasets were organized geographically according to the United Hospital Fund (UHF) neighborhoods, which were defined by ZIP code-based boundaries. There are 42 UHF regions across New York City. Data were collected over a 10-year period from 2014 to 2023, resulting in a total of 420 data entries. The environmental variables included vegetative cover, defined as the percentage of land covered by plants, and tree canopy cover, which measured the percentage of a neighborhood shaded by trees (13,14). Climate data consisted of daytime summer surface temperatures in degrees Fahrenheit (15). Air quality variables included levels of PM measured in micrograms per cubic meter, and concentrations of NO_2 and O_3, both measured in parts per billion (16). Traffic data included vehicle miles traveled, car miles traveled, and truck miles traveled, each measured in millions of miles per square mile (17). The outcome variable, asthma-related EDVs per 10,000 people, was log-transformed to normalize the distribution (18). In cases where data for a specific year was unavailable, values from the closest available year were used to complete the dataset.

Preprocessing

Initially, nine features were selected including car miles traveled (Car_Miles) (17). A correlation matrix was computed for these nine features (**Figure 5**). From the correlation matrix, a strong correlation between Vehicle_Miles and Car_Miles was observed, leading to the exclusion of Car_Miles from further analysis. To analyze each feature's contribution and validate the feature selection, feature importance scores were calculated and experiments were conducted by removing features from the prediction model.

Hyperparameter tuning

After preprocessing, the next step was to select hyperparameters and define their ranges for tuning. Considering the computational intensity of Grid Search, it was essential to narrow down the hyperparameter space before performing a comprehensive search. To accomplish this, a preliminary search process was conducted by isolating the effect of each hyperparameter. This was done by holding all other parameters at their default values while varying one parameter at a time. For example, for the RFR model, all other hyperparameters were set to their default values and only the number of estimators was varied. The default values for other hyperparameters were: max_depth was unlimited, max_features was 8, and min_samples_split was 2. Metrics such as MSE and R^2 were used to evaluate predictive accuracy. MSE was calculated as the average of the squared differences between the predicted and actual values, providing a measure of the model's error magnitude. R^2 was computed to quantify the proportion of variance in the observed data explained by the model, The MSE results showed that performance peaked at 30 and then declined slightly (**Figure 6**). Based on these results, the range for the Grid Search was selected as [10, 15, 20, 25, 30, 35, 40, 45].

In this study, hyperparameter tuning involved Grid Search and cross-validation. The dataset was randomly partitioned into training and test datasets using an 80/20 split ratio to ensure

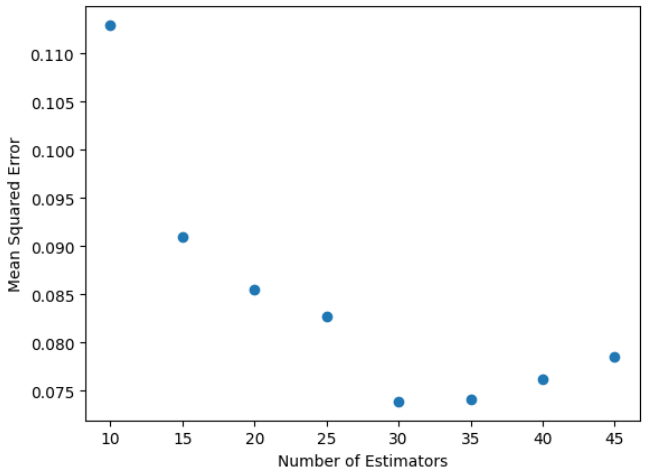

Figure 6: Effect of changing number of estimators in the RFR model. The graph is showing the MSE of predicted asthma-related emergency department visits (EDVs) using the RFR model when the number of estimators was changing and all other hyperparameters were set to default value. The default values were: max_depth was unlimited, max_features was 8, and min_samples_split was 2.

https://doi.org/10.59720/25-031

a balanced and representative sample. The training dataset was further divided into five equal subsets for cross-validation. During the tuning process, Grid Search was performed over the hyperparameter space on each fold of the training data, and the performance metrics obtained from each fold were averaged to determine the optimal set of hyperparameters. This method provided a robust approach to assessing model performance while reducing the risk of overfitting or bias from a single train-test split.

Model training

Next, each model was trained using the optimal set of hyperparameters determined in the previous step. Once the models were trained, their performance was evaluated on the test dataset to show generalizability and accuracy. The feature importance scores were calculated again using the highest-performing model, XGBoost. These scores served as the foundation for analyzing which factors were the most significant contributors to the prediction results.

Explainable AI

Finally, based on the performance results, XGBoost and RFR were selected for further analysis of feature contributions using explainable AI techniques. SHAP beeswarm plots were generated to provide insights into the contribution of each feature.

The python packages used for this study were sklearn (scikit-learn), extreme gradient boosting python package, and shap python package.

Received: January 20, 2025
Accepted: June 07, 2025
Published: August 12, 2025

REFERENCES

1. "NHIS Adult Summary Health Statistics." *Centers for Disease Control and Prevention*. https://data.cdc.gov/d/25m4-6qqq. Accessed 19 Sept. 2024.
2. "NHIS Child Summary Health Statistics." *Centers for Disease Control and Prevention*. https://data.cdc.gov/d/wxz7-ekz9. Accessed 19 Sept. 2024.
3. "Healthcare Cost and Utilization Project 2020 Healthcare Use Data." *Centers for Disease Control and Prevention*. https://www.cdc.gov/asthma/healthcare-use/2020/data.htm. Accessed 19 Sept. 2024.
4. Nieuwenhuijsen, Mark J. "New urban models for more sustainable, liveable and healthier cities post covid19; reducing air pollution, noise and heat island effects and increasing green space and physical activity." *Environment International*, vol. 157, Dec. 2021, https://doi.org/10.1016/j.envint.2021.106850.
5. Tiotiu, Angelica I., et al. "Impact of Air Pollution on Asthma Outcomes." *International Journal of Environmental Research and Public Health*, vol. 17, no. 17, 27 Aug. 2020, p. 6212, https://doi.org/10.3390/ijerph17176212.
6. Dong, Yuping, et al. "Association Between Green Space Structure and the Prevalence of Asthma: A Case Study of Toronto." *International Journal of Environmental Research and Public Health*, vol. 18, no. 11, 29 May 2021, p. 5852, https://doi.org/10.3390/ijerph18115852.
7. Han, Azhu, et al. "Asthma Triggered by Extreme Temperatures: From Epidemiological Evidence to Biological Plausibility." *Environmental Research*, vol. 216, pt. 2, 1 Jan. 2023, https://doi.org/10.1016/j.envres.2022.114489.
8. Cook, Angus G., et al. "Use of a Total Traffic Count Metric to Investigate the Impact of Roadways on Asthma Severity: A Case-Control Study." *Environmental Health*, vol. 10, 2 June 2011, https://doi.org/10.1186/1476-069X-10-52.
9. Hancox, Robert J., et al. "Relationship Between Socioeconomic Status and Asthma: A Longitudinal Cohort Study." *Thorax*, vol. 59, no. 5, May 2004, pp. 376-80, https://doi.org/10.1136/thx.2003.010363.
10. Taunk, Kashvi, et al. "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification." *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 1255-1260, https://ieeexplore.ieee.org/document/9065747.
11. Chen, Tianqi and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 13 Aug. 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.
12. Liaw, Andy and Matthew Wiener. "Classification and Regression by randomForest." *R News,* vol. 2, no. 3, Dec. 2002, pp. 18-22.
13. "Vegetative Cover." *NYC Department of Health and Mental Hygiene*. https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/climate/?id=2143. Accessed 10 Aug. 2024.
14. "Tree Canopy Cover." *NYC Department of Health and Mental Hygiene*. https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/active-design/?id=2157. Accessed 10 Aug. 2024.
15. "Daytime Summer Surface Temperature." *NYC Department of Health and Mental Hygiene*. https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/climate/?id=2141. Accessed 10 Aug. 2024.
16. "Air Quality." *NYC Department of Health and Mental Hygiene*. https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/air-quality/?id=2023. Accessed 10 Aug. 2024.
17. "Walking, Driving, and Cycling." *NYC Department of Health and Mental Hygiene*. https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/walking-driving-and-cycling/?id=2113. Accessed 10 Aug. 2024.
18. "Asthma." *NYC Department of Health and Mental Hygiene*. https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/asthma/?id=2380. Accessed 10 Aug. 2024.
19. Ren, Chongyang, et al. "Estimating of the causal effect of land use mixed on adult asthma prevalence in New York State." *Sustainable Cities and Society*, vol. 119, Feb. 2025, https://doi.org/10.1016/j.scs.2025.106125.
20. Adeyeye, Temilayo, et al. "A census tract-level assessment of social determinants of health, traffic exposure, and asthma exacerbations in New York State's Medicaid Population (2005–2015)." *Eco-Environment & Health*, vol. 3, no. 3, Sept. 2024, pp. 300-307, https://doi.org/10.1016/j.eehl.2024.04.005.
21. McPhearson, Timon, et al. "Local Assessment of New York City: Biodiversity, Green Space, and Ecosystem Services." *Urbanization, Biodiversity and Ecosystem Services: Challenges and Opportunities*, 1 Jan. 2013, pp 355–383, https://doi.org/10.1007/978-94-007-7088-1_19.
22. Dwivedi, Rudresh, et al. "Explainable AI (XAI): Core

https://doi.org/10.59720/25-031

Ideas, Techniques, and Solutions." *ACM Computing Surveys*, vol. 55, no. 9, 16 Jan. 2023, pp. 1-33, https://doi.org/10.1145/3561048.

23. Lundberg, Scott M. and Su-In Lee, "A unified approach to interpreting model predictions." *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4 Dec. 2017, pp. 4768–4777, https://dl.acm.org/doi/10.5555/3295222.3295230.
24. Tantithamthavorn, Chakkrit, et al. "The Impact of Automated Parameter Optimization on Defect Prediction Models." *IEEE Transactions on Software Engineering*, vol. 45, no. 7, 1 July 2019, pp. 683-711, https://doi.org/10.1109/TSE.2018.2794977.
25. González-Castro, Lorena, et al. "Impact of Hyperparameter Optimization to Enhance Machine Learning Performance: A Case Study on Breast Cancer Recurrence Prediction." *Applied Sciences*, vol. 14, no. 13, 6 July 2024, https://doi.org/10.3390/app14135909.
26. Chen, Jianzhong, et al. "Relationship between prediction accuracy and feature importance reliability: An empirical and theoretical study." *Neuroimage*, vol. 274, 1 July 2023, https://doi.org/10.1016/j.neuroimage.2023.120115.
27. DellaValle, Curt T., et al. "Effects of Ambient Pollen Concentrations on Frequency and Severity of Asthma Symptoms Among Asthmatic Children." *Epidemiology,* vol. 23, no. 1, Jan. 2012, pp. 55-63, https://doi.org/10.1097/EDE.0b013e31823b66b8.
28. Barreñada, Lasai, et al. "Understanding overfitting in random forest for probability estimation: a visualization and simulation study." *Diagnostic and Prognostic Research,* vol. 8, no. 14, Sept. 2024, https://doi.org/10.1186/s41512-024-00177-1.

Copyright: © 2025 Chen and Parchure. All JEI articles are distributed under the attribution non-commercial, no derivative license (http://creativecommons.org/licenses/by-nc-nd/4.0/). This means that anyone is free to share, copy and distribute an unaltered article for non-commercial purposes provided the original author and source is credited.