Urban Green Space and Air Quality Impact on Asthma Analysis Using Machine Learning Algorithms

Andrew Chen
Urbana High School

Abstract - Asthma, a chronic respiratory condition, poses a significant public health concern in urban areas, prompting investigation into how factors like air quality, green space, climate, and socioeconomic conditions contribute to its prevalence. This study utilizes machine learning algorithms with feature explainability to analyze the relationships between these factors and asthma-related emergency department visits (EDV) across different neighborhoods in the New York City (NYC). The models demonstrated high accuracy in predicting EDV rates, providing reliable insights into the factors influencing asthma prevalence. SHAP values offered further understanding of feature importance, confirming the significant roles of air quality, green space, climate, and socioeconomic factors. These findings pave the way for decisive, integrated urban planning approaches that prioritize air quality improvements, targeted green space management, and socioeconomic considerations to effectively mitigate asthma risks in urban environments.

Keywords**:** Asthma, environment, machine learning, air quality, urban green space, environmental medicine

INTRODUCTION

Asthma, a chronic respiratory disorder characterized by airway inflammation and hyperresponsiveness, remains a significant public health issue in urban areas, impacting daily life through reduced physical activity, frequent medical visits, and limitations on work and school attendance. The prevalence of asthma is often worsened by various environmental factors, with air pollution and limited access to green spaces being major contributors. Poor air quality, resulting from pollutants such as particulate matter (PM), nitrogen dioxide (NO_2), and ozone (O_3), can trigger or exacerbate asthma symptoms. Conversely, the presence of green spaces can help mitigate these effects by improving air quality through pollutant absorption and providing areas that encourage physical activity, which can improve respiratory health. Given the substantial effect of asthma on individuals and communities, investigating how both green space availability and air quality influence asthma prevalence is critical for developing targeted public health strategies and urban planning policies.

Previous literature offers varied perspectives on the relationship between urban green spaces and asthma, reflecting both benefits and potential risks. For instance, Dadvand et al. (2014) explored the dual nature of green spaces, finding that while proximity to forests was associated with lower obesity rates and reduced sedentary behavior among children, living near parks was linked to an increased prevalence of asthma. This suggests that not all types of green spaces confer the same health benefits, pointing to the need for careful consideration in urban planning. Similarly, Oosterbroek et al. (2023) demonstrated the complexity of urban green space impacts by identifying both health benefits, such as reduced heat stress and improved air quality, and risks, including perceived unsafety and exposure to tick-borne diseases. This dual-edged nature of green spaces emphasizes the importance of strategically designing urban landscapes to maximize health benefits while minimizing adverse effects.

In addition, Mueller et al. (2022) provided a systematic review that generally found positive associations between exposure to greenspaces and respiratory health outcomes, though the findings for asthma specifically were inconsistent, suggesting the influence of other underlying factors. Dong et al. (2021) further contributed by examining the structural composition of green spaces, revealing that areas with higher tree diversity significantly reduced asthma risk in children and adolescents, although the expected link through air quality improvement was not strongly supported. This finding indicates that other mechanisms, such as increased physical activity or microbial exposure, might play a role in the protective effects observed. Hartley et al. (2020) underscored the complex and often indirect relationship between greenness and childhood asthma. Their systematic review indicated mixed results, with most studies finding no significant direct link between greenness and asthma, although greenness was suggested to mitigate other asthma risk factors, such as traffic-related air pollution and tobacco smoke exposure. These studies collectively highlight the intricate and sometimes contradictory effects of green spaces on respiratory health, illustrating the need for multifaceted research and strategic urban planning to effectively harness the health-promoting potential of green spaces while addressing possible risks.

This study aims to examine the relationship of urban green space availability, air quality, climate, and socioeconomic factors with asthma prevalence using machine learning algorithms. NYC serves as an ideal case study due to its dense population, diverse environmental conditions, and significant asthma burden, making it a representative case for urban health challenges. We used tree-based machine learning algorithms, including random forest regression and extreme gradient boosting models, for their proven effectiveness in handling complex, high-dimensional datasets and capturing non-linear relationships. By leveraging extensive datasets encompassing environmental, climate, health, and socioeconomic data, this research employs advanced statistical and machine learning techniques to identify patterns and associations. These algorithms enable a nuanced understanding of how variations in green space distribution, air quality, climate, and socioeconomic factors across different neighborhoods correlate with asthma rates, accounting for the multifaceted nature of urban environmental health.

METHODS

This study employs a machine learning approach to explore the relationship between urban green space availability, air quality and asthma prevalence. The analysis utilizes two advanced algorithms—Extreme Gradient Boosting (XGBoost) and Random Forest Regression (RFR) - to identify patterns and associations within the data. These models leverage extensive datasets, including environmental, climate, health, and socioeconomic factors, to gain insights into how green space and air quality may impact asthma rates across different neighborhoods.

1. Feature Selection and Preprocessing

The datasets used in this study are obtained from NYC.gov public datasets. Each data entry is combined across all feature data corresponding to the same region in NYC.

Initially, eight features were selected from four data categories based on their potential influence on asthma prevalence. These features include:

Environmental Data: Vegetative cover percentage (NYC.gov Environment & Health Data Portal Vegetative Cover, n.d.)

Climate Data: Surface temperature (NYC.gov Environment & Health Data Portal Daytime Summer Surface Temperature, n.d.)

Air Quality Data: PM, NO_2, and O_3 (NYC.gov Environment & Health Data Portal Air Quality, n.d.)

Socioeconomic Data: Vehicle miles traveled, car miles traveled, and truck miles traveled (NYC.gov Environment & Health Data Portal Walking, Driving, and Cycling, n.d.)

Data preprocessing involves handling missing values and normalizing features for consistency. Correlation analysis is performed to detect and remove highly collinear features.

To assess multicollinearity among the features, a correlation matrix was computed (Figure 1). The analysis indicated a high correlation between vehicle miles traveled and car miles traveled, suggesting strong collinearity. Therefore, the feature of car miles traveled was excluded from the feature list, leaving seven features for subsequent analysis.

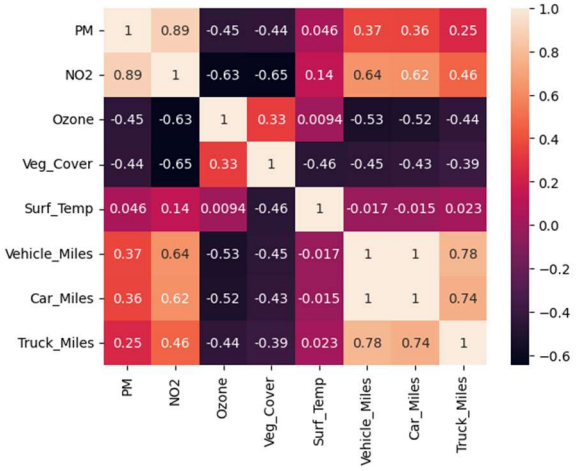

Figure 1. Collinearity Index

2. Model Implementation

The main algorithms used are XGBoost and RFR. Both models are well-suited for handling complex relationships among variables.

2.1 XGBoost

XGBoost is a gradient-boosting algorithm that minimizes a loss function by adding new models that correct the errors made by previous models. The algorithm optimizes the following objective function:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, z_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (1)$$

$\sum_{i=1}^{n} l(y_i, z_i)$ is the loss function, which measures the difference between the true labels y_i and the predicted label z_i (e.g., Mean Squared Error for regression or Logarithmic Loss for classification). The goal is to minimize this term so that the predictions become more accurate.

$\sum_{k=1}^{K} \Omega(f_k)$ is the regularization term, which penalizes the complexity of the model. The term $\Omega(f_k)$ ensures that the trees do not become too deep or too complex, which can lead to overfitting. Regularization makes the model more generalizable to unseen data.

XGBoost's objective is to strike a balance between accuracy (minimizing the loss function) and model simplicity (minimizing the regularization term). This helps create a model that performs well on both training and test datasets.

2.2 Random Forest Regression (RFR)

RFR builds an ensemble of decision trees and combines their outputs to improve predictive performance. The prediction for a given input is obtained by averaging the individual predictions from all the trees:

$$\hat{y}(x) = \frac{1}{M} \sum_{m=1}^{M} \widehat{y_m}(x) \qquad (2)$$

$\hat{y}(x)$ is the final prediction for the input x

M is the number of trees in the forest.

$\widehat{y_m}(x)$ is the prediction from the m^{th} decision tree.

Each tree is built using a different subset of the data (bootstrap sampling), and feature selection is randomized at each split to reduce overfitting. Hyperparameters such as the number of trees and maximum depth are optimized to enhance model accuracy.

3. Model Training and Evaluation

There are a total of 421 data entries. Data is randomly split into training (90%) and testing (10%) sets. The target variable, asthma related EDV per 10,000 people (NYC.gov Environment & Health Data Portal Asthma, n.d.), was log-transformed to achieve a normalized distribution. Models are evaluated using metrics such as Mean Squared Error (MSE) and R^2, to measure predictive accuracy.

.

RESULTS

1. XGBoost Model Optimization and Performance Evaluation

1.1 XGBoost Model Optimization

For the XGBoost model, the primary focus was on optimizing the number of estimators (trees). MSE and R^2 values were calculated for different numbers of estimators, as shown in Figure 2 and Figure 3, respectively. The results demonstrated that increasing the number of estimators improved model performance. However, the gains diminished beyond 25 estimators, indicating an optimal balance between model complexity and accuracy. Thus, the number of estimators was set to 25.

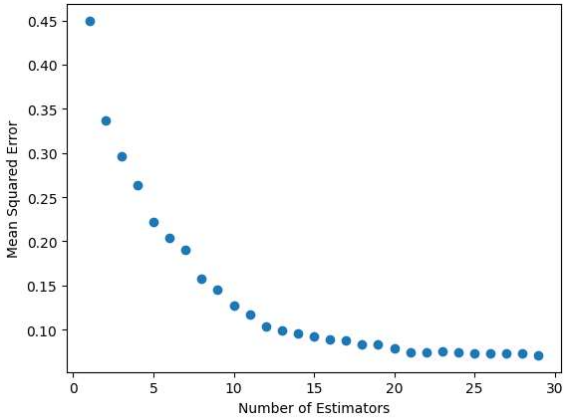

Figure 2. XGBoost optimization MSE.

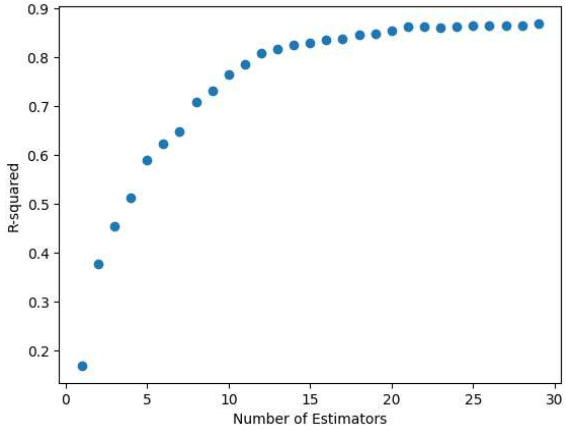

Figure 3. XGBoost optimization R^2

The final model predictions for EDV are depicted in Figure 4, where a strong correlation between the predicted and actual values is evident. The performance metrics are summarized in Table 1, showing satisfactory predictive accuracy, with MSE of 0.0339 and R^2 of 0.9375.

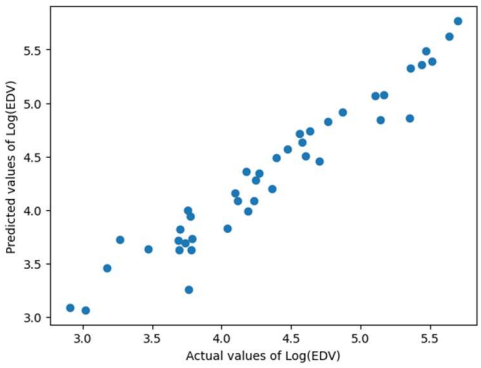

Figure 4. XGBoost Results.

1.2 Interpretation of XGBoost Model Using SHAP Values

SHAP values were employed to interpret the XGBoost model's predictions, providing insights into the contribution of each feature to the model's output. The SHAP beeswarm plot (Figure 5) visualizes the impact of each feature on the predictions. All features showed significant influence, indicating their crucial roles in determining asthma-related EDV.

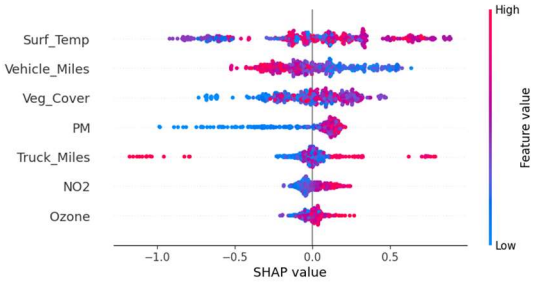

Figure 5. XGBoost SHAP beeswarm plot

2. RFR Model Optimization and Performance Evaluation

2.1 RFR Model Optimization

Similarly, the RFR model was optimized by tuning the number of estimators. The results, depicted in Figure 6 (MSE) and Figure 7 (R^2), suggest that the optimal number of estimators is 20, beyond which no significant performance improvement was observed.

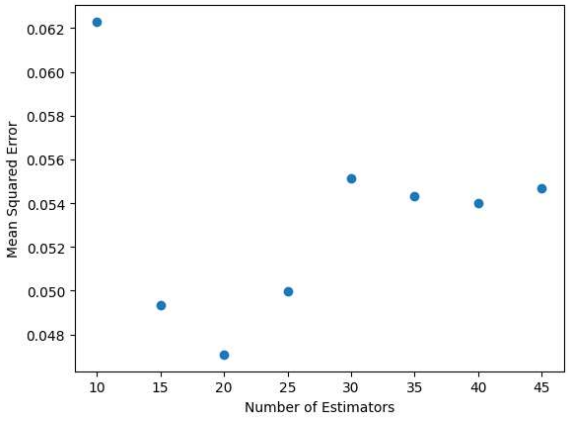

Figure 6. RFR optimization MSE

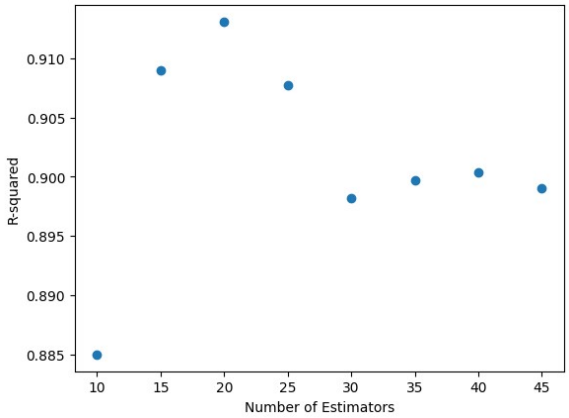

Figure 7. RFR optimization R^2

The final prediction results are shown in Figure 8. There was a strong correlation between the predicted and actual values, indicating that the model accurately captured the underlying patterns in the data. The MSE and R^2 for the RFR model are also provided in Table 1, confirming the model's predictive effectiveness, with MSE of 0.0437 and R^2 of 0.9193.

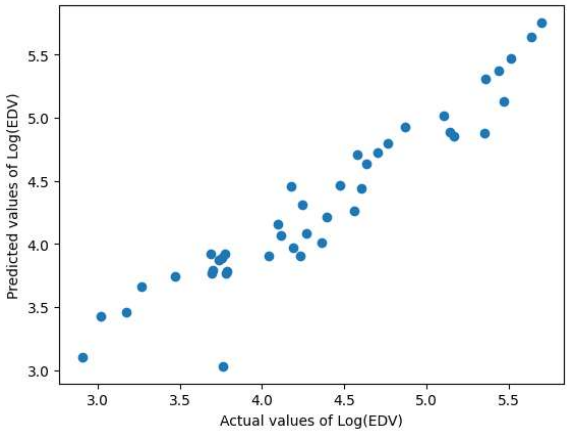

Figure 8. RFR results

2.2 Interpretation of RFR Model Using SHAP Values

SHAP values were also used to interpret the predictions made by the RFR model. The SHAP beeswarm plot (Figure 9) illustrates the feature importance. This aligns with the XGBoost model's findings, reinforcing the significance of these features in predicting asthma-related EDV.

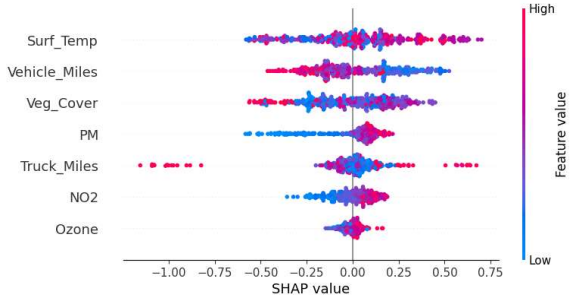

Figure 9. RFR SHAP beeswarm plot

Table 1. Model Performance Comparison

Model	R^2	MSE
RFR	0.9193486703186	0.0436836551032
XGBoost	0.9374959071398	0.0338544602529

DISCUSSION

The analysis using both XGBoost and RFR models demonstrated robust predictive capabilities, with high correlation coefficients between the predicted and actual values of EDV related to asthma. The findings indicate a significant association between air quality factors—O_3, NO_2, and PM—and EDV. As shown in Figures 5 and 9, poor air quality (higher pollutant levels) is associated with increased EDV, while better air quality is linked to lower asthma-related emergency visits. This relationship highlights the crucial role of air pollution in exacerbating asthma conditions, aligning with established research in the field. This finding is consistent with existing literature, such as Murrison et al. (2019), which emphasizes the significant impact of air pollutants on respiratory health. The positive association between higher pollutant levels and increased asthma incidence supports the need for stricter air quality regulations and interventions aimed at reducing pollutant exposure in urban areas.

Interestingly, the environmental feature representing vegetative cover percentage did not exhibit a strong, straightforward correlation with EDV, as indicated by SHAP analysis (Figures 5 and 9). This observation suggests a more complex relationship between urban green space and asthma, where the presence of vegetation alone does not directly translate to improved respiratory health outcomes. Prior studies, such as Ferrante et al. (2020), have also reported inconsistent findings regarding the protective effects of urban green space on asthma, suggesting that other factors, such as the type of vegetation, allergen levels, or proximity to pollution sources, might mediate the relationship.

This complexity indicates that the benefits of urban green space are not uniformly experienced across all communities. While green spaces may reduce air pollution or provide areas for physical activity, they can also introduce allergens, such as pollen, which may trigger

asthma symptoms in sensitive individuals. The varying impacts observed in this study underscore the need for tailored urban planning strategies that consider not only the quantity of green space but also its quality, management, and integration with other factors.

In addition to green space and air quality, this study finds that climate data (e.g., surface temperature) and socioeconomic factors (e.g., truck miles traveled) significantly influence asthma-related EDV, though some expected associations, like high truck miles correlating with increased asthma rates, were not strongly observed. Possible explanations include wealthier areas having better healthcare access and insulation from pollution, urban infrastructure limiting residents' direct exposure despite pollution sources, and well-developed public transportation reducing reliance on personal vehicles. These findings highlight the complex interplay between socioeconomic factors and environmental health, with infrastructural elements shaping exposure levels and health outcomes.

The comparison between XGBoost and RFR shows that while XGBoost achieved slightly better predictive accuracy, as reflected by lower MSE and higher R^2, there are notable differences in model stability and feature interactions. RFR exhibited more stable predictions, with tighter SHAP value clusters, indicating less variability and stronger robustness. In contrast, XGBoost's wider range of SHAP values suggests more pronounced feature interactions and greater sensitivity to feature inputs, which can capture complex patterns but also risks overfitting. Thus, while XGBoost may be more effective for modeling intricate relationships, RFR's stability makes it a suitable choice when robustness is a priority.

While this study provides valuable insights, it has some limitations. The analysis is based on the available data, which may not capture all relevant factors influencing asthma, such as indoor air quality, specific sources of pollution, or individual health behaviors. Moreover, the type and management of green spaces were not distinguished, which could affect the observed associations. Future research could focus on differentiating the quality and characteristics of green spaces, and proximity to pollution sources, to gain a deeper understanding of their impact on respiratory health.

CONCLUSION

This study effectively demonstrates the application of machine learning algorithms in analyzing the complex relationships between urban green space, air quality, climate, socioeconomic factors, and asthma-related health outcomes. The findings emphasize the critical role of air quality factors, such as O_3, NO_2, and PM, in driving asthma prevalence, while also revealing the intricate and context-dependent effects of urban green space. The results indicate that the impact of green space on asthma is influenced by various factors, including climate conditions, and socioeconomic characteristics, underscoring the importance of a more nuanced approach to urban planning rooted in environmental medicine.

ACKNOWLEGEMENTS

I would like to express my sincere gratitude to Shreya Parchure, MD-PhD candidate at the University of Pennsylvania, for her invaluable mentorship and guidance throughout this study.

REFERENCES

[1] Dadvand, P., Villanueva, C. M., Font-Ribera, L., Martinez, D., Basagaña, X., Belmonte, J., Vrijheid, M., Gražulevičienė, R., Kogevinas, M., & Nieuwenhuijsen, M. J. (2014). Risks and benefits of green spaces for children: A cross-sectional study of associations with sedentary behavior, obesity, asthma, and allergy. Environmental Health Perspectives, 122(12), 1329–1335. https://doi.org/10.1289/ehp.1308038

[2] Oosterbroek, B., de Kraker, J., Huynen, M. M. T. E., Martens, P., & Verhoeven, K. (2023). Assessment of green space benefits and burdens for urban health with spatial modeling. Urban Forestry & Urban Greening, 86, 128023. https://doi.org/10.1016/j.ufug.2023.128023

[3] Mueller, W., Milner, J., Loh, M., Vardoulakis, S., & Wilkinson, P. (2022). Exposure to urban greenspace and pathways to respiratory health: An exploratory systematic review. Science of The Total Environment, 829, 154447. https://doi.org/10.1016/j.scitotenv.2022.154447

[4] Dong, Y., Liu, H., & Zheng, T. (2021). Association between green space structure and the prevalence of asthma: A case study of Toronto. International Journal of Environmental Research and Public Health, 18(11), 5852. https://doi.org/10.3390/ijerph18115852

[5] Hartley, K., Ryan, P., Brokamp, C., & Gillespie, G. L. (2020). Effect of greenness on asthma in children: A systematic review. Public Health Nursing, 37(4), 453–460. https://doi.org/10.1111/phn.12701

[6] NYC.gov Environment & Health Data Portal Vegetative Cover. (n.d.). https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/climate/?id=2143#display=map

[7] NYC.gov Environment & Health Data Portal Daytime Summer Surface Temperature. (n.d.). https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/climate/?id=2141#display=map

[8] NYC.gov Environment & Health Data Portal Air Quality. (n.d.). https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/air-quality/?id=2023#display=summary

[9] NYC.gov Environment & Health Data Portal Walking, Driving, and Cycling. (n.d.). https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/walking-driving-and-cycling/?id=2113#display=summary

[10] NYC.gov Environment & Health Data Portal Asthma. (n.d.). https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/asthma/?id=2380#display=summary

[11] Murrison, L. B., Brandt, E. B., Myers, J. B., & Hershey, G. K. K. (2019). Environmental exposures and mechanisms in allergy and asthma development. Journal of

Clinical Investigation, 129(4), 1504–1515. https://doi.org/10.1172/JCI124612
[12] Ferrante, G., Asta, F., Cilluffo, G., De Sario, M., Michelozzi, P., & La Grutta, S. (2020). The effect of residential urban greenness on allergic respiratory diseases in youth: A narrative review. World Allergy Organization Journal, 13(1), 100096. https://doi.org/10.1016/j.waojou.2019.100096