Using Machine Learning Models to Predict Asthma Hospitalizations Based on Air Pollutants

Nikhil Tole

10/24/24

Abstract

This paper explores the use of a Decision Tree regression model to predict asthma hospitalizations across 42 regions of New York City based on air pollution data. The study focuses on three key air pollutants: PM2.5, NO2, and O3. The model's performance is evaluated using Mean Absolute Error, Mean Absolute Percentage Error, R-Squared Score, and Root Mean Squared Error. While the study does not aim to produce definitive forecasts, it assesses the viability of this machine learning approach for identifying health vulnerabilities related to air pollution. Forecasting the asthma hospitalizations by region can help identify and rectify inequities in health vulnerability. The model achieved a Mean Absolute Error of 50.01 and a Mean Absolute Percentage Error of 84.3%, indicating significant prediction challenges. Results show that high levels of NO2 and O3 are associated with increased asthma-related ER visits. These findings highlight potential areas for targeted healthcare interventions.

1. Introduction

Exposure to outdoor air pollution can reduce lung function as well as exacerbate asthma symptoms [1]. According to the American Lung Association's standards, 39% of Americans live in places with failing grades on levels of air pollution, a rise of 11.7 million people from the previous year [2]. Sixteen million Americans live in counties with failing air pollution grades in at least one pollutant, and 5.4 million Americans live in counties where at least three pollutants are above acceptable levels [2]. The motivation for this project was to discover the regions of New York City that are most vulnerable to asthma due to air pollution.

We customized and trained a model to predict asthma-related emergency room (ER) visits in all New York City regions with the goal of determining which regions are most at risk and which factors increase those risks. We use New York City's United Hospital Fund or UHF-42 mapping system, which divides New York City into 42 regions designated by NYC Community Planners [4]. We then built a supervised learning decision tree regression model to predict the age-adjusted rate of asthma related ER visits in a given region per 10,000 people using three numerical features, namely the incidence of PM2.5, NO2, and O3. By predicting which neighborhoods will see the highest number of asthma-related emergency room visits due to air pollution, one can identify regions that are vulnerable to health issues due to disproportionately high air pollution.

The predictions of our model can have important planning and policy implications in areas such as environmental regulation and healthcare initiatives. For example, additional health resources could be moved to regions that are likely to see an increase in asthma related incidents and people living in those regions could be advised on the precautions they can take based on the pollutant that is impacting them on a given day, such as wearing the proper type of mask or limiting outdoor activity.

2. Background

Prior research has used regression models to predict asthma outcomes based on environmental factors in cities such as Seoul [5]. In the paper studying Seoul, the authors trained a Random Forest model that could study the association between air pollutants and the frequency of hospital visits. For their paper, the authors focused on the city of Seoul as a whole, whereas I focused on New York City, and divided the city into 42 different regions, allowing for the study of inequities within regions.

3. Dataset

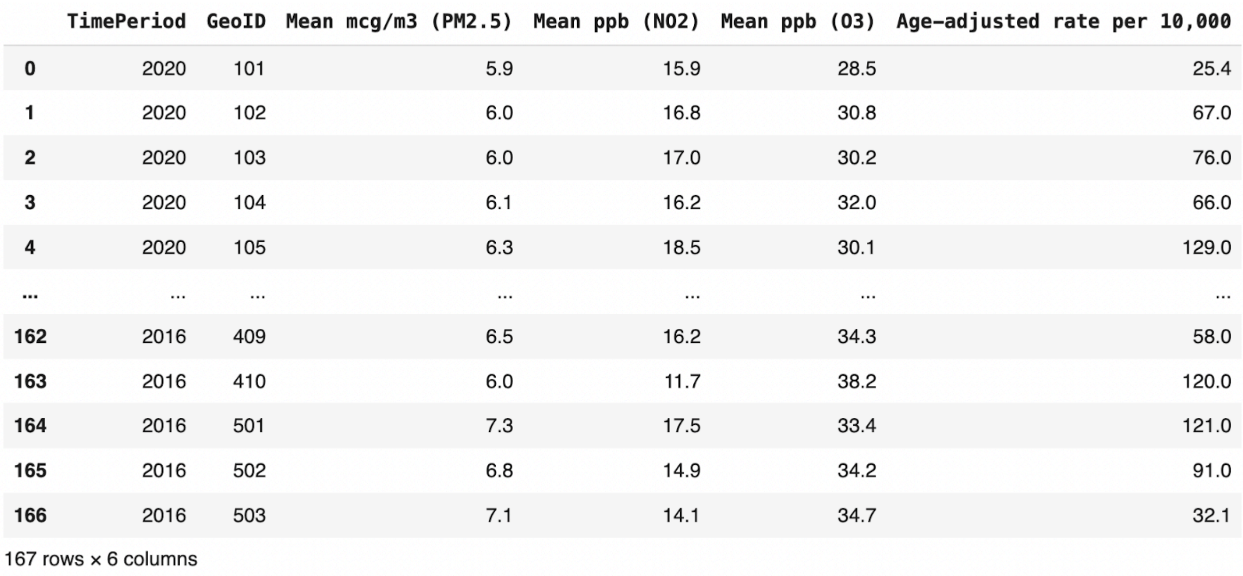
The data used in this project consists of the following four datasets, all sourced from the NYC Department of Health Environment and Health portal: asthma hospitalizations, PM2.5, NO2, and O3 [6], [7], [8], [9].

The initial asthma dataset focused on asthma related ER visits in New York City's UHF-42 regions from 2010-2020 and includes total number of ER visits, annual rate of ER visits per 10,000, and age-adjusted annual rate of ER visits per 10,000 as key features. The age-adjusted annual rate per 10,000 was the most viable option for the target variable due to the elimination of age bias, while also normalizing the data into a common proportion, allowing for clearer comparisons. Normalization ensures comparability across different rows in the datasets, aiding model fitting. The asthma hospitalizations dataset includes columns displaying neighborhood name, total number of hospitalizations, and annual rate per 10,000. Descriptive features such as neighborhood names were dropped.

While this dataset ranges from 2010 to 2020, data from 2010-2015 was excluded from this study due to data quality issues such as missing values and because the other three datasets only date back to 2016.

Air pollution data was obtained from three different datasets: PM2.5 (Mean mcg/m3), NO2 (Mean ppb), and O3 (Mean ppb), covering the years 2016 to 2020. These three datasets contain columns representing the time of year that the data was collected, the region name, the unit of measure, and the 10th and 90th percentiles. To maintain simplicity, the values used in this study for all three pollutants represent the mean amount, in parts per billion.

The final dataset includes 167 rows and 6 columns, with age-adjusted rate per 10,000 as the target.



	TimePeriod	GeoID	Mean mcg/m3 (PM2.5)	Mean ppb (NO2)	Mean ppb (O3)	Age-adjusted rate per 10,000
0	2020	101	5.9	15.9	28.5	25.4
1	2020	102	6.0	16.8	30.8	67.0
2	2020	103	6.0	17.0	30.2	76.0
3	2020	104	6.1	16.2	32.0	66.0
4	2020	105	6.3	18.5	30.1	129.0
...	...	...	...	...	...	...
162	2016	409	6.5	16.2	34.3	58.0
163	2016	410	6.0	11.7	38.2	120.0
164	2016	501	7.3	17.5	33.4	121.0
165	2016	502	6.8	14.9	34.2	91.0
166	2016	503	7.1	14.1	34.7	32.1

167 rows × 6 columns

Fig. 1: Top and bottom rows of dataset

4. Methods

To explore the prediction of asthma-related emergency room visits, we tested multiple regression models using the scikit-learn library in Python, including linear regression, random forest regression, ridge regression, and decision tree regression. We considered mean absolute error, mean squared error, r2 score, root mean squared error, and mean absolute percentage error as evaluation metrics and reached the conclusion that a decision tree model, a supervised regression model, would be most suitable for this project, due to the limited amount of data at our disposal and possibility of non-linear relationships.

The decision tree regression model works by splitting the data into two groups based on one feature at a time. A decision tree model starts at the root node, representing the entire dataset, and splits the data into two subsets based on whether the value of the feature chosen for that node is larger than a certain split value. The fitting algorithm chooses the feature to split on as well as the split value in such a way that the target variable has the largest difference in behavior in the two subsets. This process is repeated recursively to further refine the prediction in each branch of the tree. Ultimately, when we reach a leaf node, the average value of the target variables of the subset of observations that fall in that node is used as the final prediction.

This project uses the following Python libraries: numpy, pandas, seaborn, matplotlib, and scikit-learn. The model was implemented using SKLearn's DecisionTreeRegressor class. The features used were the three pollutants being used to look for correlation with asthma: PM2.5, NO2, and O3. The column Age-Adjusted Rate per 10,000 was used as the y vector for the target. The scikit-learn function train-test-split was used to do the test and train split, and the data was split using 20% testing data and 80% training data. Using 80% of the data as training data, the amount of rows in the training data was 34. After splitting the data into training and testing sets, a decision tree regression model was trained.

Hyperparameters were determined based on which yielded the lowest mean absolute error. The mean absolute error (MAE) takes the absolute error, summing it over all samples.

MAE is beneficial in minimizing the adverse impact of outliers. The result is an absolute value, so there is no directionality in the value.

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

The hyperparameters used are: max_depth, max_leaf_nodes, min_samples_split, and ccp_alpha. For the decision tree, a combination of max_depth = 5, max_leaf_nodes = 10, min_samples_split = 17, and ccp_alpha = 0.01 found to be an optimal combination to yield the lowest possible mean absolute error for the Decision Tree model. This is limited by the min_samples_split and max_leaf_nodes parameters, which inhibit the tree from going to its full depth. Max_depth refers to the depth of the decision tree, max_leaf_nodes grows the tree by taking the best nodes first based on impure nodes, min_samples_split affects the minimum amount of samples needed to split a node, and ccp_alpha uses cost complexity pruning to control the size of the tree [3]. The optimal hyperparameter values were determined through a grid search using scikit-learn's GridSearchCV function and are shown on the highlighted row below.

Hyperparameter 1: ccp_alpha	**Hyperparameter 2:** max_depth	**Hyperparameter 3:** max_leaf_nodes	**Hyperparameter 4:** min_samples_split	**Result** (MAE)
0.01	5	6	9	59.79
0.01	6	8	10	56.44
0.01	5	10	17	50.01

Fig 2: Table showing results of three combinations of hyperparameters tested during hyperparameter tuning; the hyperparameter combination used is highlighted.

5. Results and Discussion

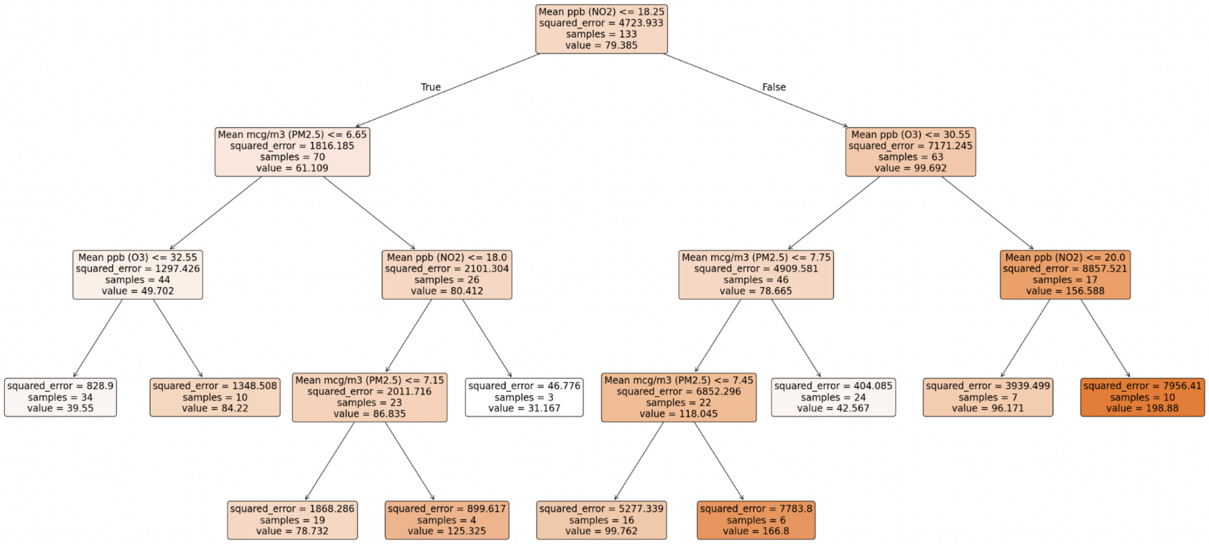


Fig 3: The visualization of the decision tree with the hyperparameters chosen by GridSearchCV
We see that the rightmost leaf node has the highest asthma hospitalization rate, almost 2.5 times the average rate. This leaf node corresponds to high NO2 and high O3.

The Mean Absolute Error (MAE) using the decision tree model is shown to be 50.01. The mean absolute percentage error (MAPE) and MAE indicate a fairly ineffective forecast of the expected asthma hospitalizations. The MAPE of the test data shows that the model's predictions deviate from the actual values by approximately 84.3%.

$$R^2 = 1 - \frac{\text{Variance of Line}}{\text{Variance of Y}}$$

Mean absolute error	50.01
R^2 score	0.07
Mean squared error	4889.48
Root mean square error	69.92
Mean absolute percentage error	84.3%

Fig 4: Resulting error metrics for the decision tree model

Using asthma hospitalization data from 2016-2020 (excluding 2017) and PM2.5, NO2, and O3, pollutants in the UHF-42 boundaries of New York City, we built a model that aims to forecast asthma hospitalizations rate in the 42 regions of New York City.

The resulting MAPE score from our Decision Tree regression model shows that the model forecasts an asthma diagnosis with an error of approximately 84%, showing a high error in the prediction. A major limitation of our study is the small amount of data used and the narrow focus on the New York City area. Using a larger, more diverse dataset and additional features that capture other known asthma causing factors may also help in improving the model's performance.

Another limitation of this study is that the data used in this paper does not explicitly account for socioeconomic or other possible factors beyond air pollution, a potential area to address in future research. Our model could be enhanced in future research by incorporating additional socioeconomic factors, such as percentage of the population under the poverty level, unemployment rate, average size of household, and population density; as well as by incorporating naturally occurring irritants such as pollen.

6. Conclusions

With data from 2016-2020 excluding 2017 focused on asthma hospitalizations and NO2, O3, and PM2.5 pollutants in the UHF-42 boundaries of New York City, this study aimed to correctly forecast the asthma hospitalizations in the 42 regions of New York City. The resulting MAPE score from the Decision Tree regression model shows that the model forecasts asthma diagnosis with approximately 84.3% error, showing that the predictions were not very successful. We found that high NO2 and O3 levels were predictive of higher asthma hospitalization rate.

References

[1] Thurston, G. D., & Rice, M. B. (2019). Air pollution exposure and asthma incidence in children. JAMA, 321(19), 1875. https://doi.org/10.1001/jama.2019.5343

[2] State of the air 2024. (n.d.). https://www.lung.org/getmedia/dabac59e-963b-4e9b-bf0f-73615b07bfd8/State-of-the-Air-2024.pdf

[3] Decision Trees. scikit-learn. (n.d.). https://scikit-learn.org/1.5/modules/tree.html

[4] New York City Department of Health and Mental Hygiene. (2020). Neighborhood boundaries on the EH Data Portal. NYC Environment & Health Data Portal. https://a816-dohbesp.nyc.gov/IndicatorPublic/data-stories/geographies/

[5] Lee, S., Ku, H., Hyun, C., & Lee, M. (2022). Machine learning-based analyses of the effects of various types of air pollutants on hospital visits by asthma patients. Toxics, 10(11), 644. https://doi.org/10.3390/toxics10110644

[6] Asthma data for NYC: Fine Particles. NYC Environment & Health Data Portal. (n.d.). https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/asthma/?id=2382#display=summary

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

[8] Air Quality Data for NYC: Ozone. Environment & Health Data Portal. (n.d.). https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/air-quality/?id=2025#display=summary

[9] Air Quality Data for NYC. Environment & Health Data Portal. (n.d.). https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/air-quality/?id=2027#display=summary