Predicting Asthma Hospitalization Visits

Table Of Contents

Predicting Asthma Hospitalization Visits

1. Motivation & Background

WHO has identified climate and pollutants as humanity's biggest health threat. Identification of the most crucial links between pollutants and health can help mitigate the health impacts by issuing warnings and helping in early adaptation for any kind of adversity arising out of the changes in the climate. In the longer term, a transformational action can be taken to reduce emissions in order to reduce the presence of harmful chemicals in the air or to find preventive ways to reduce the breaching of dangerous temperature thresholds.

We have gone through various publications [1,2,3] and government websites [4,5] that spearheaded interdisciplinary efforts to study combinations of multiple pollutants more extensively. Research has linked regulated air pollutants such as ozone and particulate matter to lung, heart disease, and other health problems.
Issues discussed in this article:

- To understand the collective impacts of multiple air pollutants, how they interact in the atmosphere, and whether the interactions modify health effects.
- Characterization of point source air pollution and climate forcers
- Characterization of fugitive and area source air pollution and climate forcers
- Characterization of mobile source emissions of air pollution and climate forcers.

2. Problem

Health of asthma patients can be affected by the presence of air pollutants through many causal pathways. A useful approach to understanding how pollutants present in air affects asthma patients considering specific exposure pathways and how they can lead to asthma or cause severity in the asthma patients. Exposure to various changes in climate, like the presence of Ozone (O3), Nitrogen dioxide (NO2), and Nitric Oxide(NOx) can affect human health in various different ways.

For our problem, we are focusing on age groups, and for disease, we are working with Asthma. We will find the correlation between asthma hospitalizations visits and air pollutants (PM_2.5, NO2, O3, and SO2).

For this dataset we will try to find out how pollutants(like NO2, SO2, etc.) affect the number of department visits for asthma in NYC?

3. Dataset Description

We are using **daily data of New York City** from two open data sources:

Air quality: epa.gov: Outdoor-air-quality-data
Asthma Hospitalization: - health.nyc.gov Syndromic diseases

The original Asthma Hospitalization dataset has 86835 rows and 9 columns and the Air quality dataset has 7308 rows and 17 columns.

4. Dataset Preparation and Feature Selection

The air quality dataset had the concentration of pollutants present in the air and also the pollutant aqi. The pollutant aqi was a redundant feature here hence we have removed all the pollutants aqi. Rest all other irrelevant features were removed from both the dataset. After removing the irrelevant features of the two dataset we **merged them based on the Date and County** and considered age group as features. Once the data was merged and cleaned we performed encoding on the county and the age group columns. The encoding for the county is as follows: Bronx:0, Brooklyn:1, Queens:2 and Manhattan: 3.

The final **predictors** are the 10 features which include pollutants like co_con, no2_con, o3_con, pm25_con, so2_con, various age groups and demographic features i.e. county. More details about the predictors are mentioned in the below table.

Feature	Data Type	Range	Scale	Description
0 - 4 years	integer	0 / 1	boolean	Age group from 0 - 4 years
5 - 17 years	integer	0 / 1	boolean	Age group from 5 - 17 years
18 - 64 years	integer	0 / 1	boolean	Age group from 18 - 64 years
64+ years	integer	0 / 1	boolean	Age group from 64+ years
county	integer	0-3	numerical	Encoded county names
co_con	float	0-1.7	numerical	CO Concentration

Feature	Data Type	Range	Scale	Description
no2_con	float	1.0-80	numerical	NO2 Concentration
o3_con	float	0.001-0.089	numerical	O3 Concentration
pm25_con	float	-3.4-30.0	numerical	PM2.5 Concentration
so2_con	float	-0.1-29.9	numerical	SO2 Concentration

Table: Predictors or Feature Set

We will be **predicting** the department visit for asthma patients across all age groups. The **asthma department visit count** is an integer that **ranges from 0 to 350**.

Feature	Data Type	Range	Scale	Description
count	integer	0 - 350	numerical	Count of asthma patients department visit

Table: Target Variable

5. Data Analysis

The below graph shows distribution of the median pollutant concentrations based on the counties.

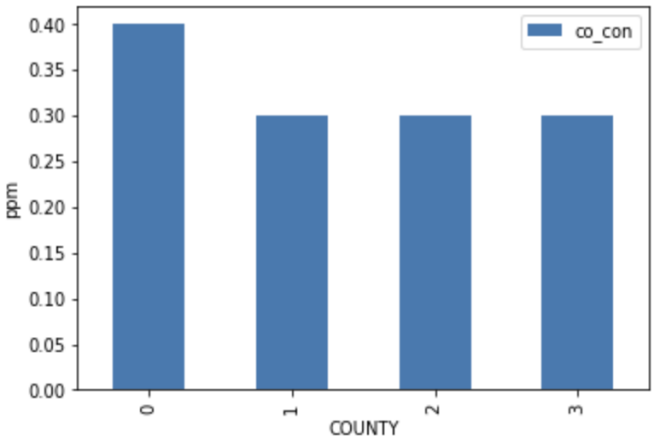

Figure: Distribution of County vs median co_conc in ppm

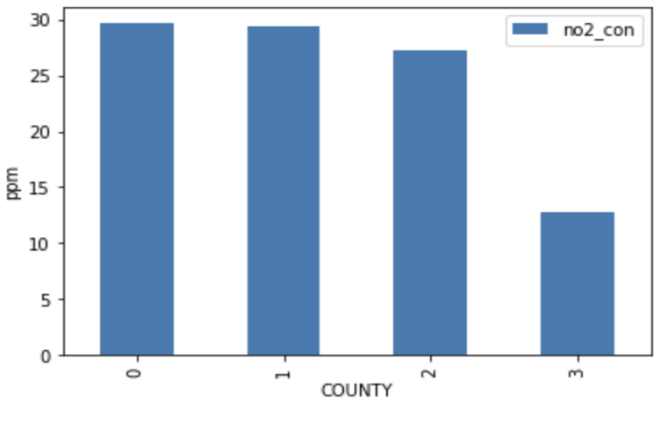

Figure: Distribution of County vs median no2_con in ppm

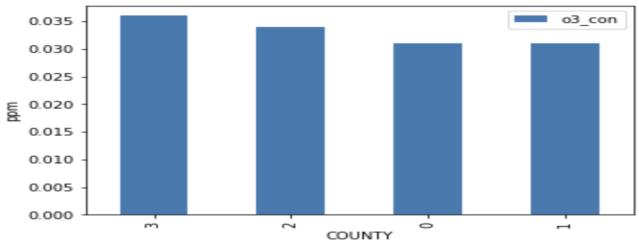

Figure: Distribution of County vs median no2_con in ppm

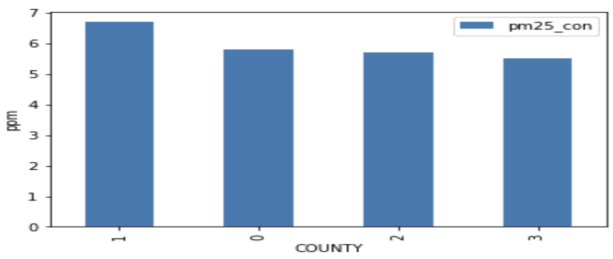

Figure: Distribution of County vs median pm25_con in ppm

For finding out the feature correlation between the pollutants and the department visit we are plotting the graph for feature correlation. As we can see from the graph the count is highly correlated with co_con, o3_con, no2_con and pm25_con.

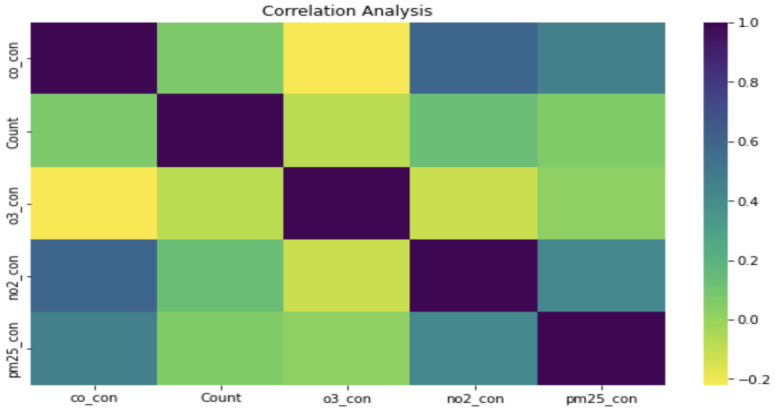

Figure: Feature correlation of the pollutants

6. Methodology

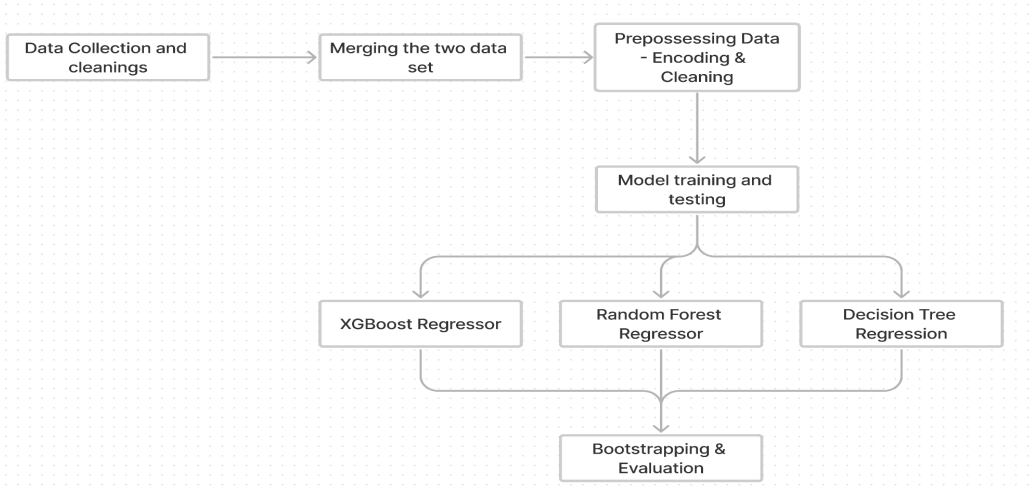

Figure: Regression Pipeline for Predicting Hospitalization Visit

6.1 Machine learning models

a. Decision Tree Regression

Decision tree breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. **We picked decision trees** for predicting asthma visits as they have an advantage that it is easy to interpret data, and less data cleaning is required. Moreover, they are not affected by non linearity.

To perform hyperparameter tuning we used GridSearchCV with appropriate range of values, and got the best values to train the model.

Parameters selected:

- **splitter:** To decide which feature and which threshold to be used.
- **max depth:** Setting max depth of tree to preclude it from overfitting
- **min sample split:** To get best trade-off between complex or generalized model
- **min sample leaf:** To retain some amount of information till tree reaches leaf node.
- **max leaf nodes:** allow the branches of a tree to have varying depths, another way to control the model's complexity.

Best parameter grid:
{'max_depth': 5, 'max_leaf_nodes': 10, 'min_samples_leaf': 4, 'min_samples_split': 8, 'splitter': 'random'}

b. Random Forest Regression

Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. **We picked random forest regression** as it provides better tuning parameters which will help us to give accurate predictions.

To perform hyperparameter tuning we used GridSearchCV with appropriate range of values, and got the best values to train the model.

Parameters selected:

- **no. of estimators:** Setting no. of decision trees, (trying to achieve higher no. of trees)
- **max depth:** Setting max depth of tree to preclude it from overfitting
- **min sample split:** To get best trade-off between complex or generalized model
- **min sample leaf:** To retain some amount of information till tree reaches leaf node.

- **bootstrap:** check effect of bootstrapping before actually performing it later.

Best parameter grid:
{'bootstrap': True, 'max_depth': 5, 'min_samples_leaf': 3, 'min_samples_split': 6, 'n_estimators': 10}

c. XGBoost Regression

XGBoost is an efficient implementation of gradient boosting that can be used for regression predictive modeling. **We picked xgboost regression model** as it provides execution speed and model performance. It dominates structured or tabular datasets on regression predictive modeling problems.

To perform hyperparameter tuning we used GridSearchCV with appropriate range of values, and got the best values to train the model.

Parameters selected:

- **no. of estimators:** No. of tries model would learn for given learning rate.
- **max depth:** Setting max depth of tree to preclude it from overfitting
- **colsample by tree :** Setting subsampling parameters for xgboost model
- **gamma:** Make the algorithm somewhat conservative for the minimum loss reduction required to make a split.
- **learning rate:** Important to get a good learning rate in correspondence to no. of estimators.

Best parameter grid:
{'clf__colsample_bytree': 0.1, 'clf__gamma': 0.0, 'clf__learning_rate': 0.01, 'clf__max_depth': 2, 'clf__n_estimators': 5, 'fs__k': 10}

7. Evaluation Metric

Our problem is based on regression as we are predicating the department visits for asthma in NYC. Therefore, we have selected the following evaluation metric -

- **Mean Absolute Error**: It represents the average magnitude of the errors in a set of predictions, without considering their direction. In other words, MAE is the average over the test sample of the absolute differences between the predicted values and the actual values.
- **R2 Score**: R2 Score, or the coefficient of determination, is a measure of how well a model fits a given dataset. It is a statistic that provides a measure of how well the regression line approximates the real data points. An R2 score of 1 indicates that the model perfectly fits the data, while a score of 0 indicates that the model does not fit the data at all.

7.1 Bootstrapping

We also performed bootstrapping on all the three model type. Bootstrapping is a technique for building a machine learning model from a relatively small amount of data. It involves repeatedly drawing samples from the original data and fitting a model to each sample. One of the main reasons for using bootstrapping is that it can be done without making any assumptions about the underlying distribution of the data, making it a useful tool for estimating statistics in cases where the distribution of the data is unknown or difficult to model. Additionally, because bootstrapping uses the data itself to estimate the distribution of the statistic, it can be used in all cases.

Logic for Bootstrapping:

```
# Defining number of iterations for bootstrap resample
# Initializing estimator
# Initializing DataFrame, to hold bootstrapped statistics
for i in range(n_iterations):
  # Sampling n_samples from data, with replacement, as train
  # Defining test to be all observations not in train
  # Fitting regression model
  # Storing stats in DataFrame, and concatenating with stats
```

For exact logic, please refer to the notebook

7.2 Results

Model Type	Mean MAE Score from Bootstrapping on 500 iterations
Decision Tree	12.392
XGBoost	12.391
Random Forest	12.42

Here for our case XGBoost regression worked the best over the decision tree regression and random forest regression models for several reasons mentioned below:

- XGBoost Regression works better as the gradient of the data is considered for each tree, the calculation is faster and the precision is more accurate than Random Forest regression.

Predicting Asthma Hospitalization Visits

- Comparing random forest regression and XGBoost regression which are opposite in nature. The former tries to reduce error by reducing the bias whereas random forest regression reduces the error by reducing variance.
- Decision tree regression is a more generalized model which is outperformed by complex ensemble models trained by hyperparameter tuning and boosting.

Decision Tree Regressor

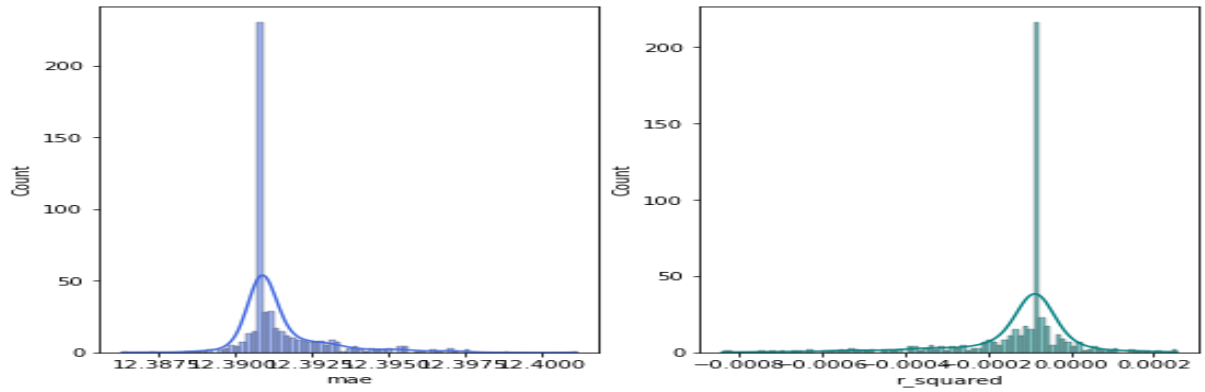

Figure: Mae & R_squared as opposed to count for Decision Tree

Random Forest Regressor

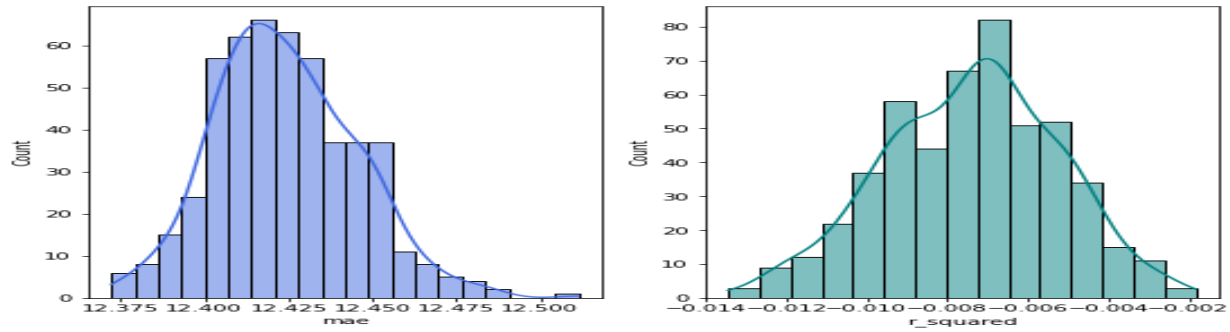

Figure: Mae & R_squared as opposed to count for Random Forest

XGBoost Regressor

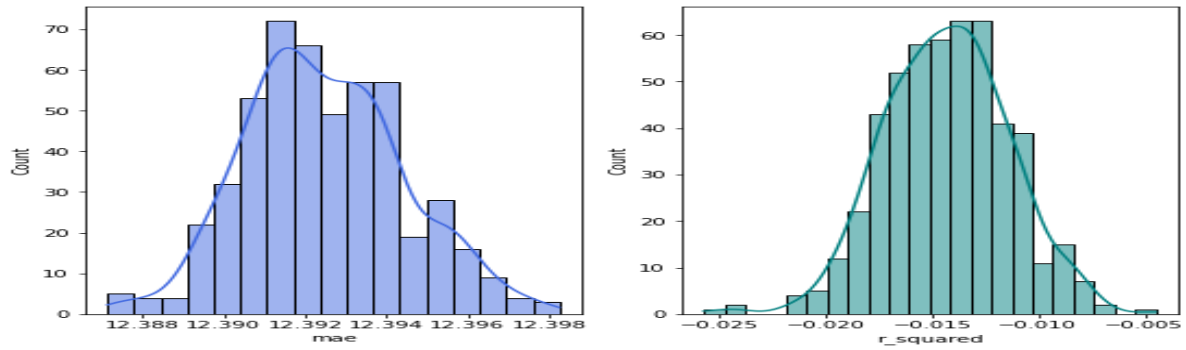

Figure: Mae & R_squared as opposed to count for Random Forest

7.3 Feature Importance

The below graph shows feature importance based on Random Forest Classifier

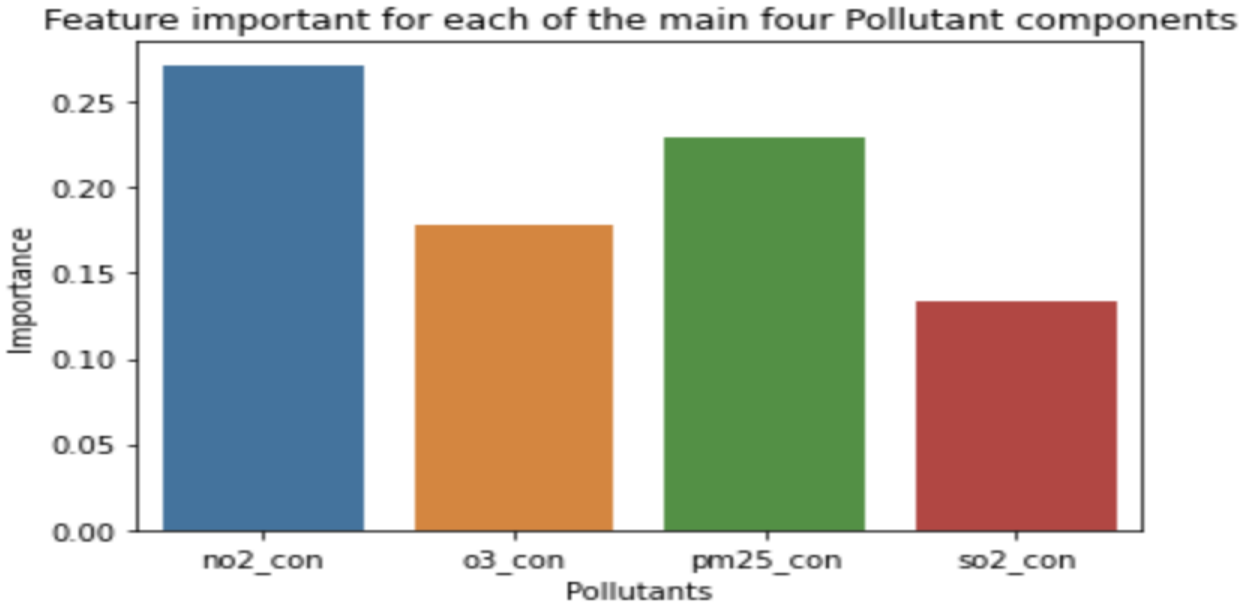

Figure: Feature Importance v/s Pollutants - Random Forest

The below graph shows feature importance based on XGBoost Classifier.

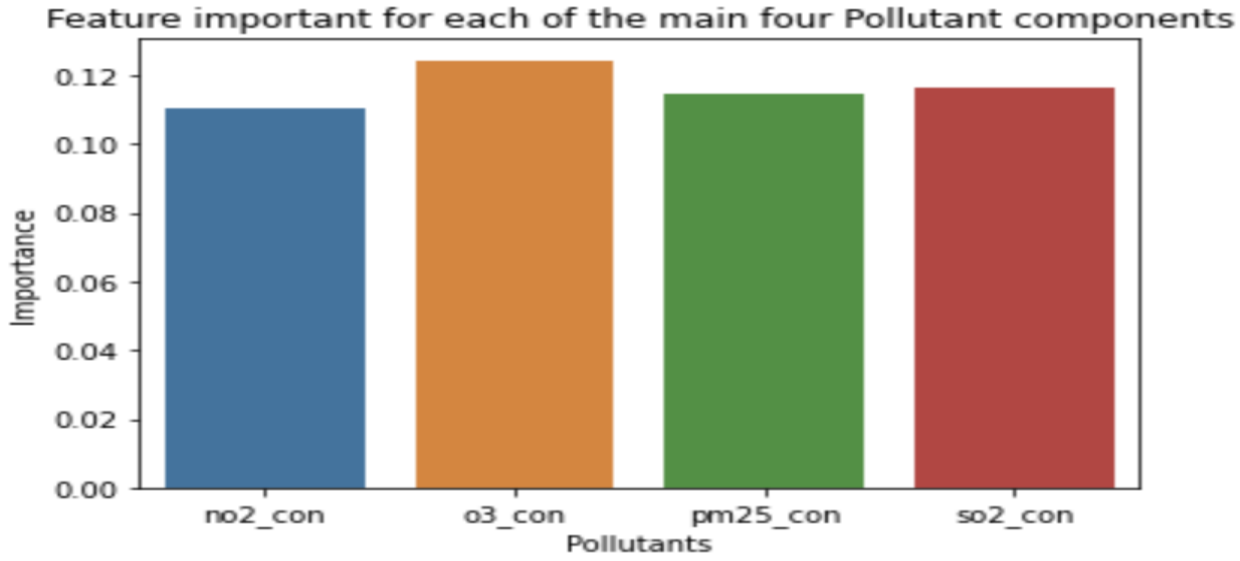

Figure: Feature Importance v/s Pollutants - XGBoost

8. Assumptions

- The number of time observations taken for air quality will be ignored.
- To build the dataset, we have merged two publicly available datasets based on the district and year. The two datasets were air quality data - pollutants and Asthma Hospitalizations.
- We are taking mean concentration for pollutants irrespective of the number of observations made daily.
- The label column has been correctly marked in the datasets we are using.

9. Limitations

- Our model is limited to only NYC data and hence may not give appropriate results for other states or places.
- Our dataset has only four pollutants, whereas there can be other pollutants that are affecting the health of asthma patients.
- No control over data quality and collection as we used pre-collected datasets from online.

10. Future Scope

- We can improve mae and r2_score with hyperparameter tuning and optimization.
- To find better correlation between target and independent variables we can add more features like demographic, economic, occupation, environment and historical data.
- Try more advanced models

11. References

1. https://doi.org/10.3389/fpubh.2020.00014
2. https://www1.nyc.gov/site/doh/data/data-sets/community-health-survey.page
3. https://rdcu.be/cYbJc
4. https://www.epa.gov/air-research/history-air-pollution
5. https://www.cdc.gov/climateandhealth/effects/default.htm