Ching-Chun Lin, Qiuxuan Li, Ziyi Zhang, Ryan Limpasurat, Heather Sizlo
Professor Paul Grigas
INDENG 142
December 15, 2023

Asthma and Air Quality - Final Project Report

Motivation: Problem and Objective

It is estimated that over 27 million people in the U.S. suffer from asthma, a chronic respiratory condition characterized by inflammation of the airways. It has long been known that genetics play a huge role in the development of asthma; if you have a parent with asthma, you are up to six times more likely to develop asthma than someone who does not have a parent with asthma. Additionally, there are severely higher asthma rates, deaths, and hospitalizations among Black, Hispanic, and American Indian individuals: a consequence of structural racism in America, as these communities are often exposed to greater levels of air pollution. Several recent studies linked here have attributed ambient outdoor air pollutants to be strongly associated with asthma-related emergency department visits, especially on the East Coast of the United States. We thus aim to explore if these attributions can be observed across other regions in the U.S., and if we can find a more generalizable pattern amongst the data. The goal of our project is to better understand how levels of particulate matter impact the severity of asthma across various geographic regions. Due to the range of particulate matter across the states and the consequent severity of asthma, we ultimately aim to predict whether or not a state will suffer any asthma-related mortalities each year.

Data: Sources and Preparation

The data used in our study is derived from two main sources: the US Chronic Disease Indicator (Asthma) dataset from the CDC website, spanning from 2010 to 2020, and the particle concentration datasets from the Environmental Protection Agency (EPA) for the same period. The asthma dataset comprised 80,342 rows with 33 columns with details about demographic stratification, causation (i.e. Hospitalization rate, death rate), and other variables. The EPA dataset has over 3 million aggregate rows across 55 columns, including state, year, pollutant, and units of measure. Initially, the two datasets were combined by grouping them by year and state. Next, categorical data (demographics) was managed through dummy encoding. We also omitted variables that are irrelevant to the study's goals and filled in the missing values by calculating the average values of the respective geographical regions. The final cleaned dataset was condensed to 4400 rows and 47 columns which included essential variables such as demographics (categorical), particle concentrations (continuous), geographical regions (categorical), and asthma death rates per 10,000 individuals (continuous) which is the outcome variable we will focus on in our models. We also did a simple EDA of pairplot to visually catch the relationships between asthma rate and the air quality indicators (Figure 0).

Analytics models

OLS (Linear Regression)

Starting off, we employed OLS models to explore the relationship between the mortality rate and the independent variables. The first model, which included all available features, yielded an R-squared value of 0.463 and an Adjusted R-squared value of 0.456 (Figure 1). These metrics suggest that approximately 46.3% of the variability in the mortality rate can be explained by the model. Upon calculating the Variance Inflation Factor (VIF) for the features (Figure 2), we identified multicollinearity with high VIF scores, which can distort the true relationship between variables and inflate the variance of the coefficient estimates. To address this, we refined the model by removing features with low

correlations, high VIFs, and high p-values as they were not statistically significant predictors. This iterative process of model adjustment led to a pared-down model that, while more statistically sound with selected features, had a slightly lower R-squared value of 0.427 and an Adjusted R-squared of 0.424 (Figure 3). This indicates a small decrease in the model's explanatory power although all the VIF scores are now low (Figure 4).

When evaluating the out-of-sample predictive performance using the OSR^2 metric, the first model achieved a value of 0.391, while the refined model scored slightly lower at 0.358 (Figure 5). This reduction in OSR^2 suggests that, despite our efforts to mitigate multicollinearity and overfitting, the refined model's generalizability to new data has diminished. The regression results indicate that while we can explain a portion of the mortality rate variability, a significant amount of variance remains unaccounted for. This could be due to non-linear relationships, unmeasured confounding variables, or inherent variability in the data that cannot be captured by linear regression. This shows a trade-off between a model's complexity and its performance. Given these considerations, we have decided to extend our approach to classification models.

Baseline Model

For the next models, we then applied classification for the binary mortality, where 0 (the negative class) indicates a mortality rate being 0 per 10,000 people and 1 (the positive class) indicates a high mortality rate of a nonzero value per 10,000 people. As for our baseline model, since there are more zeros in the training set, the baseline model predicts all the outcomes to be 0.

Logistic Regression

The performance of the logistic regression was quantitatively strong, with an accuracy rate of 0.9233. This suggests that the model correctly predicted the presence or absence of asthma-related mortalities with high reliability. Precision, the proportion of correct identifications, was 94.80%. This indicates that when the model predicted that a state would have asthma-related deaths, it was correct approximately 85.14% of the time. We primarily report results from Scikit-Learn instead of Statsmodels for logistic regression because Scikit-Learn is optimized for model performance and predictions, whereas statsmodel focuses on inference and model complexity. Since we decided to include all the features in OLS, the smf.logit in statmodel will keep warning for multicollinearity and it does generate lower accuracy than Scikit-Learn even if we did feature selections using p-values and VIF (Figure 6).

CART (With Cross-Validation)

Exploring other predictive models, our group then transitioned to CART (Classification and Regression Tree), implementing cross-validation for our ccp_alpha value to optimize our model accuracy. Primarily, we used this model to predict the mortality crude rate was non-zero. Using the full list of features, we iterated through a thousand fits for ccp_alpha, selecting the fit that provided the highest accuracy value. From cross-validation, we selected a ccp_alpha value of 0.015, producing a training set accuracy of 0.8750. Keeping all other parameters default, we produced a tree with 7 splits (Figure 8). Calculating the relative feature importance values, we found that Copper PM2.5 LC and several race-based variables had the highest significance (Figure 9). As seen in our ROC curve, our AUC was 0.898, a slight improvement from previous models (Figure 10).

Linear Discriminant Analysis

Next, our group chose to run Linear Discriminant Analysis (LDA), primarily finding one LDA component that mapped the BoolMort class to produce a prediction for mortality from our feature variables. Our training set accuracy for this model was 0.8783, and our AUC was 0.95 (Figure 12). These indicate better prediction performance, but comparing the TPR (0.8217) and FPR (0.0850) will reveal a better selection. To visualize our LDA model, we plotted the following mapping of our one LDA component, showing a sizable separation between high mortality rate and minuscule mortality rate (Figure 11).

Random Forest (with cross validation)

We used k-fold cross-validation to choose the hyperparameter 'max-features.' The cross-validated highest accuracy is 0.9025 and the corresponding max-features is 38. Thus, we use max-features = 38 in our random forest model. We found that the top 11 most important features in our random forest model are Asian or Pacific Islander, American Indian or Alaska Native, Copper PM2.5 LC, Hispanic, Arsenic PM2.5 LC, Black (non-Hispanic), White (non-Hispanic), Male, Female, Chlorine PM2.5 LC, Nitrogen dioxide (NO2), and Nickel PM2.5 LC (Figure 13).

Gradient Boosting (with cross validation)

Finally, we created a Gradient Boosting model using GradientBoostingClassifier() by Sklearn. The test set accuracy was 0.7558 and the AUC was 0.9 (Figure 15). The subsequent TPR, FPR, TNR, and FNR are 0.8408, 0.1289, 0.8711, 0.1592, respectively. When remodeled with cross-validation the highest accuracy we got was 0.8525 accuracy.

Results:

We calculated performance metrics to assess performance of these models, including the accuracy, TPR, FPR, TNR, and FNR. The metrics are shown in the table below:

	Baseline Model	Logistic Regression	CART (CV)	LDA	Gradient Boosting (CV)	Random Forest (CV)
Accuracy	**61.38%**	**92.33%**	**85.92%**	**87.83%**	**75.58%**	**90.25%**
TPR	0.00%	80.47%	84.08%	82.17%	84.08%	83.23%
FPR	0.00%	8.5%	15.8%	7.74%	12.89%	5.21%
TNR	1.00%	91.5%	84.2%	92.26%	87.11%	94.79%
FNR	1.00%	19.53%	15.92%	17.83%	15.92%	16.77%

Based on the results, we recommend that agencies use Random Forest to best predict whether their state will experience asthma related deaths. Compared to other classification models, it has a relatively low false negative rate i.e. it doesn't often predict that a state will not predict deaths when they actually will.

Extending Analysis

- Advanced modeling techniques: Other machine learning methods could improve the accuracy, such as neutral network and cluster-then-predict methodology. Moreover, we could experiment with more hyperparameters in the cross-validation.
- Incorporate more comprehensive data sources: We could add more dependent variables in the dataset, such as education level, household income, lifestyle factors, and healthcare access. This enables us to study the potential impact of other factors on the asthma mortality rate.
- Method to deal with missing values: Since the original dataset from the United States Environmental Protection Agency (EPA) contains a lot of missing values for some of the airborne pollutants, we filled the remaining missing values by the average of features in their regions. We could find better methods to handle these missing values to improve the authenticity of the data.
- Enhance interpretability: Our model of the highest accuracy, random forest with cross-validation, is hard to interpret. Implementing tools and choosing models that are explainable are important steps in the decision-making process in the context of healthcare.

Impact: Application, Significance, and Scope

Broadly, our predictive models for asthma mortality have an impact in the environmental domain, primarily for hazard identification, public health policy, and potential integration into industrial markets. Through several of our models, we have deduced certain key environmental and demographic factors that elevate the risk for asthma, and likely other lung diseases as well. Such information is not only beneficial as public knowledge, but for policies, activism, and economic activity. Particularly, our models have identified certain compounds that correlate with increased mortality rate, primarily particles of the PM 2.5 class and several greenhouse gasses. With deeper analysis that isolates these pollutants, we can concretely define and raise awareness for such harmful chemicals. Alongside these environmental factors, knowledge of the correlations of various demographic factors can also benefit the quality of healthcare treatments; if we are able to identify exact pollutants that are harmful, healthcare professionals can tailor treatments and interventions to meet the unique needs of their patients. This would also allow for more effective resource allocation in areas where areas of predicted risk (death, hospitalization rates, etc) are higher. There are also several aspects of public policy that could integrate our model into their decision-making and regulation-planning processes. In regards to the environment, policymakers can use the insights from our model to develop more efficient regulations that reduce air pollution, especially in areas known to have high rates of industrial production and thus particle emission . The implementation of more specific pollutant ceilings can lead to long-term improvements in air quality as well as human health, as the rate of severe asthma cases might decrease.

During our modeling, we included several demographic features like gender and ethnicity that are highly correlated with asthma mortality rates. As our prior research indicated, people of color tended to suffer from higher rates of asthma mortality, which may be linked to race-based biases in infrastructure. Such factors are also interconnected with environmental factors, primarily whether an area is defined as urban or rural; higher levels of pollutants were associated with states considered more urban, and thus our model would provide more information regarding hazards for such areas.

Nonetheless, our model might not have accounted for potential impacts in rural areas. Given that our data primarily originates from air quality monitors, rural regions were less likely to be represented, introducing a potential gap in our data. In retrospect, our asthma mortality predictive models offer

crucial insights for public health and public policy, emphasizing the need to address disparities and potential gaps in our collective approach to both human and environmental health.

Appendix

Project Folder which includes cited studies and all of our project data (raw and cleaned) and code. See readme file for project replication instructions.

Figure 0: Pairplot for Quantitative Feature Variables (EDA phase)

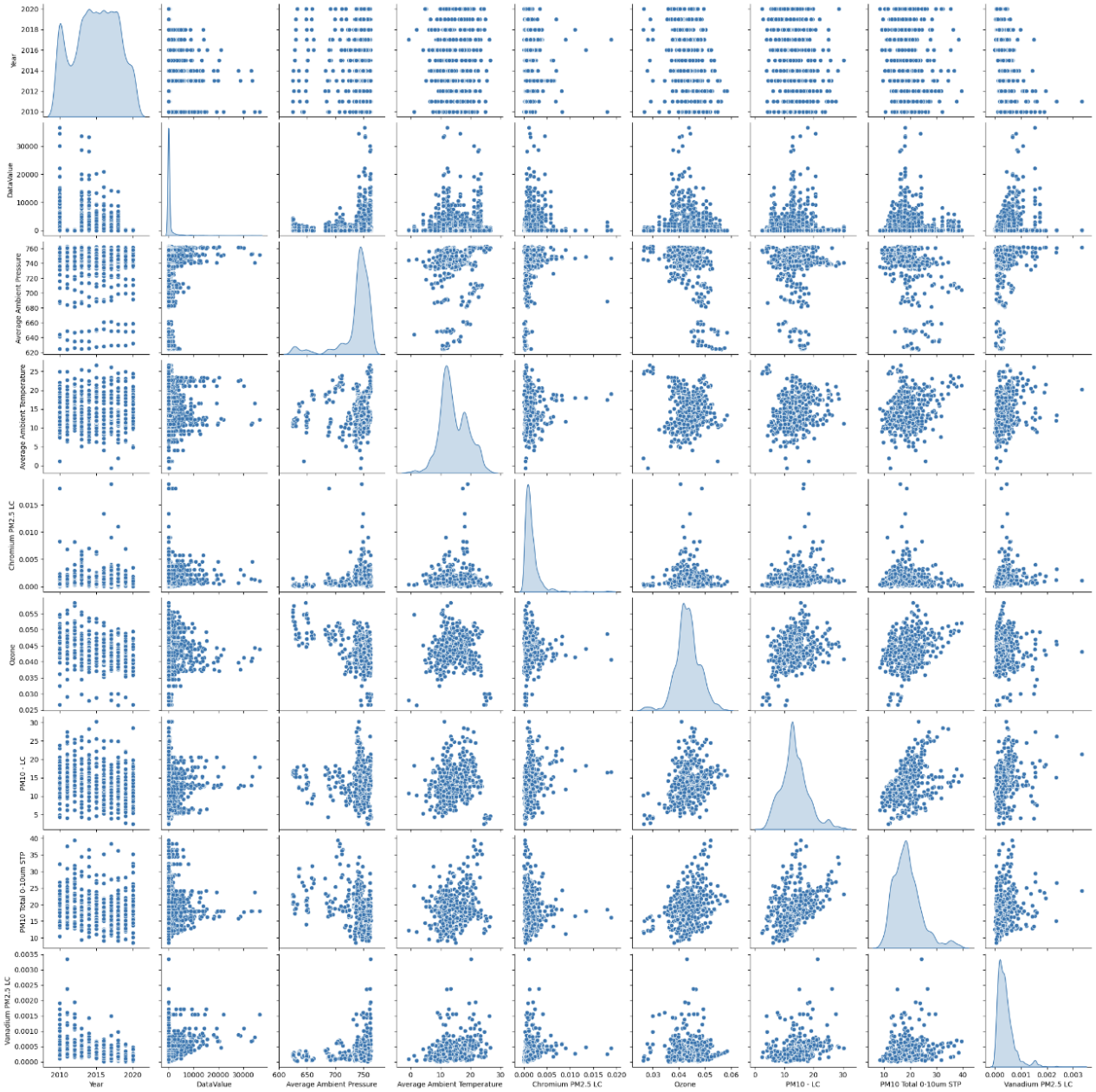
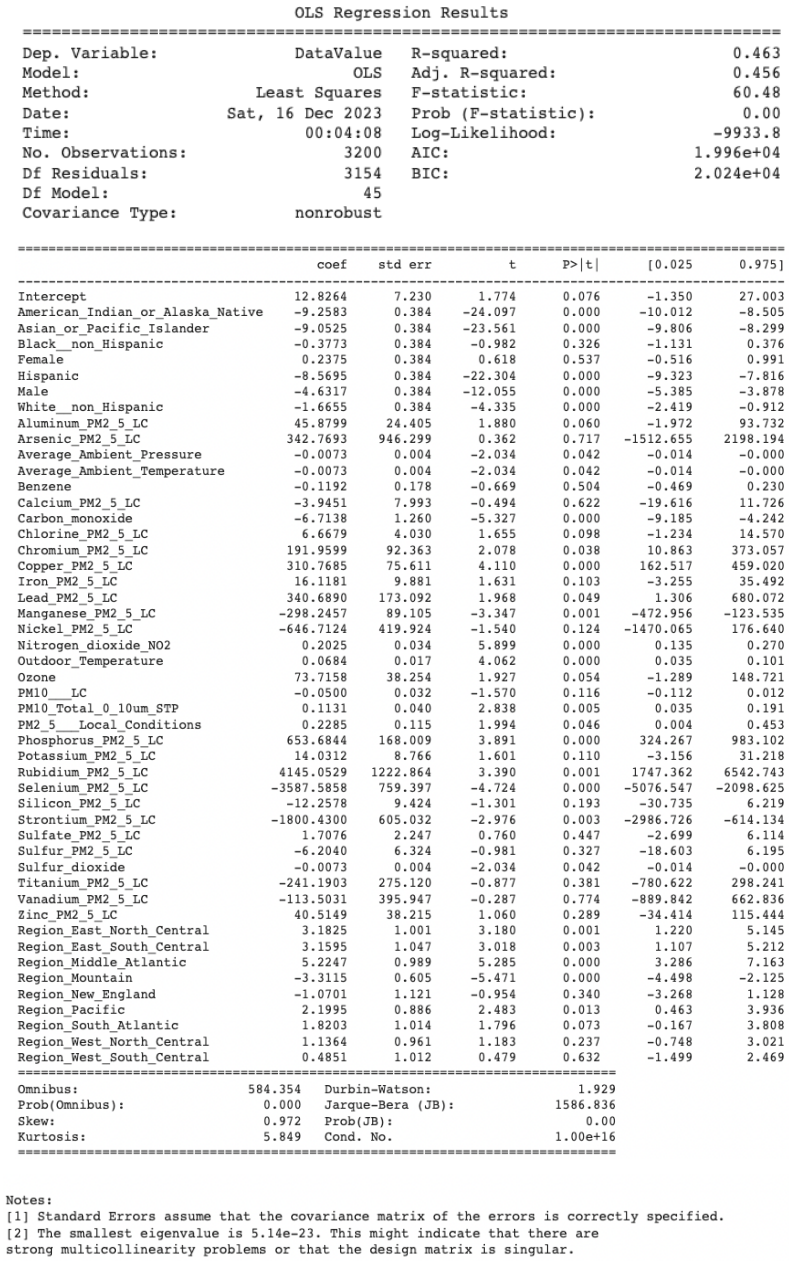


Figure 1: OLS Result Summary of Model with All Features



OLS Regression Results

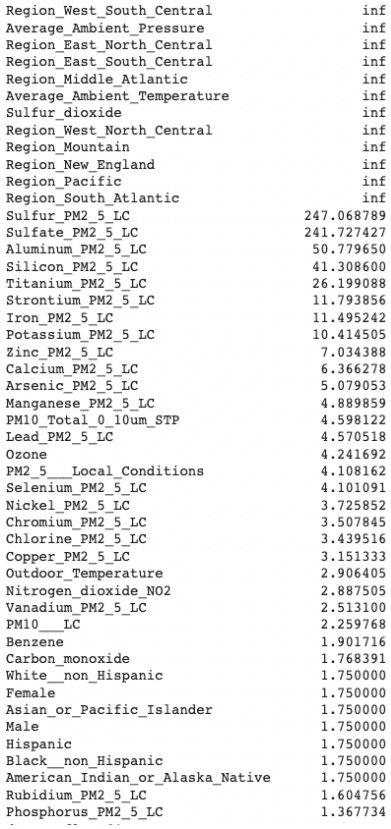
Dep. Variable:	DataValue	R-squared:	0.463
Model:	OLS	Adj. R-squared:	0.456
Method:	Least Squares	F-statistic:	60.48
Date:	Sat, 16 Dec 2023	Prob (F-statistic):	0.00
Time:	00:04:08	Log-Likelihood:	-9933.8
No. Observations:	3200	AIC:	1.996e+04
Df Residuals:	3154	BIC:	2.024e+04
Df Model:	45		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.8264	7.230	1.774	0.076	-1.350	27.003
American_Indian_or_Alaska_Native	-9.2583	0.384	-24.097	0.000	-10.012	-8.505
Asian_or_Pacific_Islander	-9.0525	0.384	-23.561	0.000	-9.806	-8.299
Black__non_Hispanic	-0.3773	0.384	-0.982	0.326	-1.131	0.376
Female	0.2375	0.384	0.618	0.537	-0.516	0.991
Hispanic	-8.5695	0.384	-22.304	0.000	-9.323	-7.816
Male	-4.6317	0.384	-12.055	0.000	-5.385	-3.878
White__non_Hispanic	-1.6655	0.384	-4.335	0.000	-2.419	-0.912
Aluminum_PM2_5_LC	45.8799	24.405	1.880	0.060	-1.972	93.732
Arsenic_PM2_5_LC	342.7693	946.299	0.362	0.717	-1512.655	2198.194
Average_Ambient_Pressure	-0.0073	0.004	-2.034	0.042	-0.014	-0.000
Average_Ambient_Temperature	-0.0073	0.004	-2.034	0.042	-0.014	-0.000
Benzene	-0.1192	0.178	-0.669	0.504	-0.469	0.230
Calcium_PM2_5_LC	-3.9451	7.993	-0.494	0.622	-19.616	11.726
Carbon_monoxide	-6.7138	1.260	-5.327	0.000	-9.185	-4.242
Chlorine_PM2_5_LC	6.6679	4.030	1.655	0.098	-1.234	14.570
Chromium_PM2_5_LC	191.9599	92.363	2.078	0.038	10.863	373.057
Copper_PM2_5_LC	310.7685	75.611	4.110	0.000	162.517	459.020
Iron_PM2_5_LC	16.1181	9.881	1.631	0.103	-3.255	35.492
Lead_PM2_5_LC	340.6890	173.092	1.968	0.049	1.306	680.072
Manganese_PM2_5_LC	-298.2457	89.105	-3.347	0.001	-472.956	-123.535
Nickel_PM2_5_LC	-646.7124	419.924	-1.540	0.124	-1470.065	176.640
Nitrogen_dioxide_NO2	0.2025	0.034	5.899	0.000	0.135	0.270
Outdoor_Temperature	0.0684	0.017	4.062	0.000	0.035	0.101
Ozone	73.7158	38.254	1.927	0.054	-1.289	148.721
PM10___LC	-0.0500	0.032	-1.570	0.116	-0.112	0.012
PM10_Total_0_10um_STP	0.1131	0.040	2.838	0.005	0.035	0.191
PM2_5___Local_Conditions	0.2285	0.115	1.994	0.046	0.004	0.453
Phosphorus_PM2_5_LC	653.6844	168.009	3.891	0.000	324.267	983.102
Potassium_PM2_5_LC	14.0312	8.766	1.601	0.110	-3.156	31.218
Rubidium_PM2_5_LC	4145.0529	1222.864	3.390	0.001	1747.362	6542.743
Selenium_PM2_5_LC	-3587.5858	759.397	-4.724	0.000	-5076.547	-2098.625
Silicon_PM2_5_LC	-12.2578	9.424	-1.301	0.193	-30.735	6.219
Strontium_PM2_5_LC	-1800.4300	605.032	-2.976	0.003	-2986.726	-614.134
Sulfate_PM2_5_LC	1.7076	2.247	0.760	0.447	-2.699	6.114
Sulfur_PM2_5_LC	-6.2040	6.324	-0.981	0.327	-18.603	6.195
Sulfur_dioxide	-0.0073	0.004	-2.034	0.042	-0.014	-0.000
Titanium_PM2_5_LC	-241.1903	275.120	-0.877	0.381	-780.622	298.241
Vanadium_PM2_5_LC	-113.5031	395.947	-0.287	0.774	-889.842	662.836
Zinc_PM2_5_LC	40.5149	38.215	1.060	0.289	-34.414	115.444
Region_East_North_Central	3.1825	1.001	3.180	0.001	1.220	5.145
Region_East_South_Central	3.1595	1.047	3.018	0.003	1.107	5.212
Region_Middle_Atlantic	5.2247	0.989	5.285	0.000	3.286	7.163
Region_Mountain	-3.3115	0.605	-5.471	0.000	-4.498	-2.125
Region_New_England	-1.0701	1.121	-0.954	0.340	-3.268	1.128
Region_Pacific	2.1995	0.886	2.483	0.013	0.463	3.936
Region_South_Atlantic	1.8203	1.014	1.796	0.073	-0.167	3.808
Region_West_North_Central	1.1364	0.961	1.183	0.237	-0.748	3.021
Region_West_South_Central	0.4851	1.012	0.479	0.632	-1.499	2.469

Omnibus:	584.354	Durbin-Watson:	1.929
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1586.836
Skew:	0.972	Prob(JB):	0.00
Kurtosis:	5.849	Cond. No.	1.00e+16

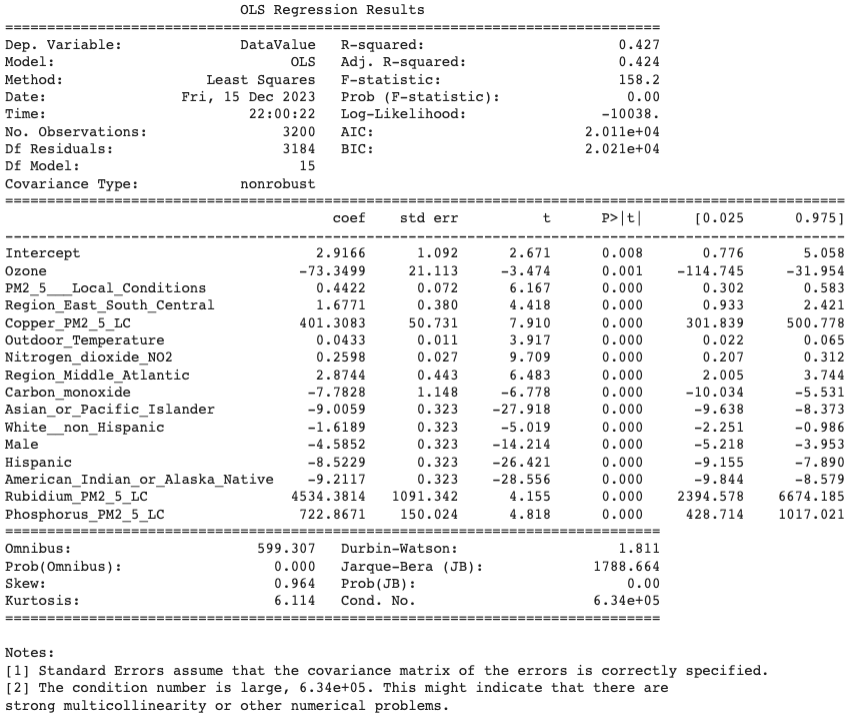
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 5.14e-23. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Figure 2: VIF Scores of All Features Together



Region_West_South_Central	inf
Average_Ambient_Pressure	inf
Region_East_North_Central	inf
Region_East_South_Central	inf
Region_Middle_Atlantic	inf
Average_Ambient_Temperature	inf
Sulfur_dioxide	inf
Region_West_North_Central	inf
Region_Mountain	inf
Region_New_England	inf
Region_Pacific	inf
Region_South_Atlantic	inf
Sulfur_PM2_5_LC	247.068789
Sulfate_PM2_5_LC	241.727427
Aluminum_PM2_5_LC	50.779650
Silicon_PM2_5_LC	41.308600
Titanium_PM2_5_LC	26.199088
Strontium_PM2_5_LC	11.793856
Iron_PM2_5_LC	11.495242
Potassium_PM2_5_LC	10.414505
Zinc_PM2_5_LC	7.034388
Calcium_PM2_5_LC	6.366278
Arsenic_PM2_5_LC	5.079053
Manganese_PM2_5_LC	4.889859
PM10_Total_0_10um_STP	4.598122
Lead_PM2_5_LC	4.570518
Ozone	4.241692
PM2_5___Local_Conditions	4.108162
Selenium_PM2_5_LC	4.101091
Nickel_PM2_5_LC	3.725852
Chromium_PM2_5_LC	3.507845
Chlorine_PM2_5_LC	3.439516
Copper_PM2_5_LC	3.151333
Outdoor_Temperature	2.906405
Nitrogen_dioxide_NO2	2.887505
Vanadium_PM2_5_LC	2.513100
PM10___LC	2.259768
Benzene	1.901716
Carbon_monoxide	1.768391
White__non_Hispanic	1.750000
Female	1.750000
Asian_or_Pacific_Islander	1.750000
Male	1.750000
Hispanic	1.750000
Black__non_Hispanic	1.750000
American_Indian_or_Alaska_Native	1.750000
Rubidium_PM2_5_LC	1.604756
Phosphorus_PM2_5_LC	1.367734

Figure 3: OLS Result Summary of Model with Selected Features



OLS Regression Results

Dep. Variable:	DataValue	R-squared:	0.427
Model:	OLS	Adj. R-squared:	0.424
Method:	Least Squares	F-statistic:	158.2
Date:	Fri, 15 Dec 2023	Prob (F-statistic):	0.00
Time:	22:00:22	Log-Likelihood:	-10038.
No. Observations:	3200	AIC:	2.011e+04
Df Residuals:	3184	BIC:	2.021e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.9166	1.092	2.671	0.008	0.776	5.058
Ozone	-73.3499	21.113	-3.474	0.001	-114.745	-31.954
PM2_5___Local_Conditions	0.4422	0.072	6.167	0.000	0.302	0.583
Region_East_South_Central	1.6771	0.380	4.418	0.000	0.933	2.421
Copper_PM2_5_LC	401.3083	50.731	7.910	0.000	301.839	500.778
Outdoor_Temperature	0.0433	0.011	3.917	0.000	0.022	0.065
Nitrogen_dioxide_NO2	0.2598	0.027	9.709	0.000	0.207	0.312
Region_Middle_Atlantic	2.8744	0.443	6.483	0.000	2.005	3.744
Carbon_monoxide	-7.7828	1.148	-6.778	0.000	-10.034	-5.531
Asian_or_Pacific_Islander	-9.0059	0.323	-27.918	0.000	-9.638	-8.373
White__non_Hispanic	-1.6189	0.323	-5.019	0.000	-2.251	-0.986
Male	-4.5852	0.323	-14.214	0.000	-5.218	-3.953
Hispanic	-8.5229	0.323	-26.421	0.000	-9.155	-7.890
American_Indian_or_Alaska_Native	-9.2117	0.323	-28.556	0.000	-9.844	-8.579
Rubidium_PM2_5_LC	4534.3814	1091.342	4.155	0.000	2394.578	6674.185
Phosphorus_PM2_5_LC	722.8671	150.024	4.818	0.000	428.714	1017.021

Omnibus:	599.307	Durbin-Watson:	1.811
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1788.664
Skew:	0.964	Prob(JB):	0.00
Kurtosis:	6.114	Cond. No.	6.34e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.34e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 4: VIF Scores of Selected Features

Nitrogen_dioxide_NO2	1.659339
PM2_5___Local_Conditions	1.520876
Carbon_monoxide	1.387944
Copper_PM2_5_LC	1.341667
Ozone	1.221907
Rubidium_PM2_5_LC	1.208758
Outdoor_Temperature	1.185239
Male	1.166667
Hispanic	1.166667
Asian_or_Pacific_Islander	1.166667
White__non_Hispanic	1.166667
American_Indian_or_Alaska_Native	1.166667
Region_Middle_Atlantic	1.136665
Region_East_South_Central	1.087239
Phosphorus_PM2_5_LC	1.031393

Figure 5: OLS Feature Selection Comparison

	Model with All Features	Model with Selected Features
OSR^2	0.391098	0.357743
RMSE	6.139530	6.305445

Figure 6: Logistic Regression Result Summary of Model with Selected Features

Logit Regression Results

Dep. Variable:	DataValue	No. Observations:	3200
Model:	Logit	Df Residuals:	3186
Method:	MLE	Df Model:	13
Date:	Sat, 16 Dec 2023	Pseudo R-squ.:	0.3249
Time:	04:22:35	Log-Likelihood:	-1441.0
converged:	True	LL-Null:	-2134.5
Covariance Type:	nonrobust	LLR p-value:	9.410e-289

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.1591	0.527	-7.898	0.000	-5.191	-3.127
Ozone	-21.0565	10.470	-2.011	0.044	-41.578	-0.535
PM2_5___Local_Conditions	0.1903	0.034	5.597	0.000	0.124	0.257
Region_East_South_Central	0.4685	0.167	2.811	0.005	0.142	0.795
Copper_PM2_5_LC	252.1961	27.964	9.019	0.000	197.388	307.005
Outdoor_Temperature	0.0274	0.005	5.301	0.000	0.017	0.038
Nitrogen_dioxide_NO2	0.1023	0.013	7.991	0.000	0.077	0.127
Region_Middle_Atlantic	0.8061	0.222	3.628	0.000	0.371	1.242
Carbon_monoxide	-2.5742	0.576	-4.468	0.000	-3.703	-1.445
Asian_or_Pacific_Islander	-4.5217	0.397	-11.377	0.000	-5.301	-3.743
White__non_Hispanic	1.2670	0.135	9.379	0.000	1.002	1.532
Hispanic	-2.7252	0.203	-13.393	0.000	-3.124	-2.326
Rubidium_PM2_5_LC	1672.6757	496.403	3.370	0.001	699.743	2645.608
Phosphorus_PM2_5_LC	467.6950	85.263	5.485	0.000	300.582	634.808

Accuracy: 0.7492
Precision: 0.6968
Recall (TPR): 0.6391
F1 Score: 0.6667
ROC AUC: 0.8358
TPR (Recall): 0.6391
FNR: 0.3609
FOR: 0.2214
TNR: 0.8203

Figure 7: ROC Curve for Logistic Regression

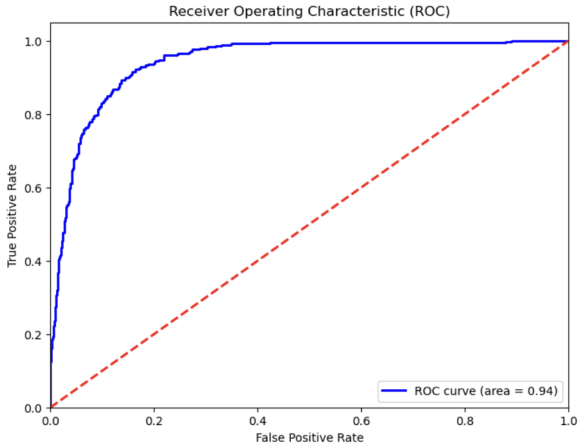


Figure 8: CART with Cross-validation Visualization

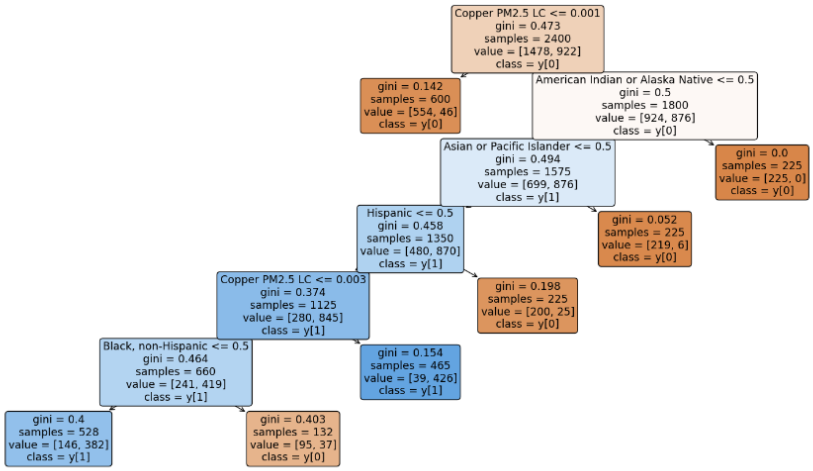


Figure 9: CART Feature Importances

	Feature	Importance
17	Copper PM2.5 LC	0.295280
5	Hispanic	0.233239
2	Asian or Pacific Islander	0.223532
1	American Indian or Alaska Native	0.184960
3	Black, non-Hispanic	0.062989
0	Year	0.000000
6	Male	0.000000
4	Female	0.000000
8	Aluminum PM2.5 LC	0.000000
9	Arsenic PM2.5 LC	0.000000

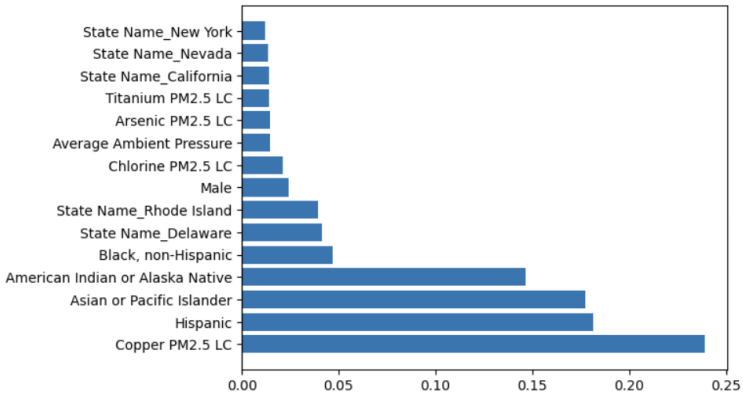


Figure 10: ROC Curve for CART with Cross Validation

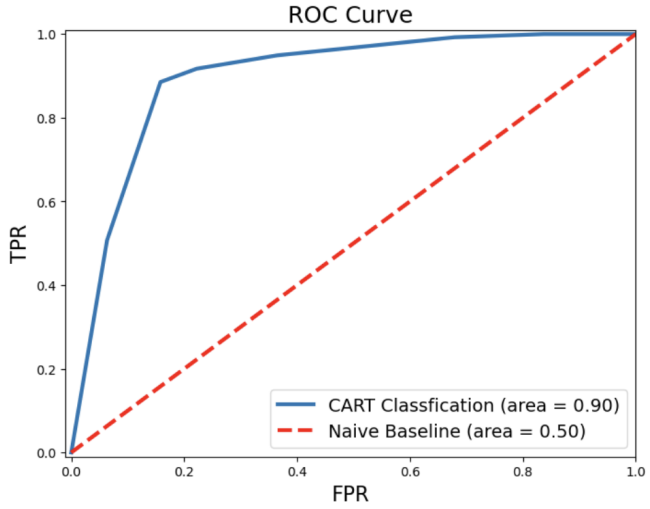


Figure 11: LDA 1-Dimensional Mapping Visualization

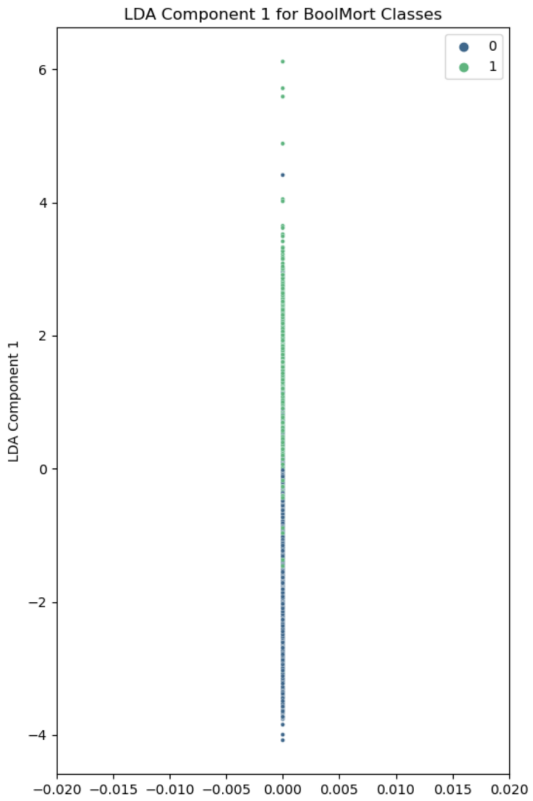


Figure 12: ROC Curve for LDA

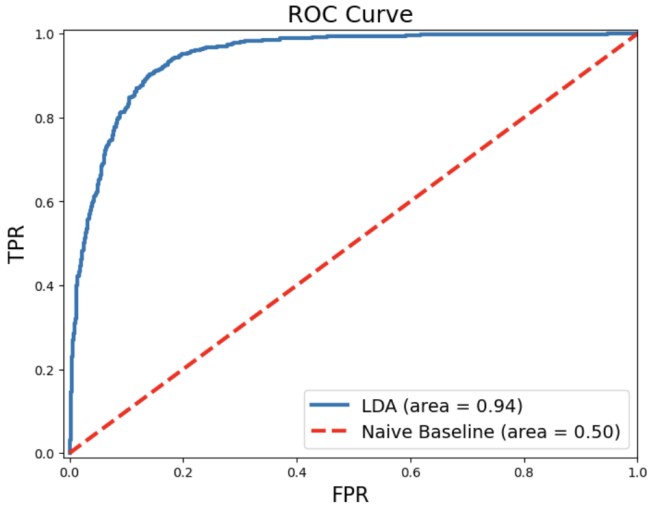


Figure 13: Random Forest Feature Importances

Feature	Importance score
Asian or Pacific Islander	14.2
American Indian or Alaska Native	13.9
Copper PM2.5 LC	13.3
Hispanic	12.4
Arsenic PM2.5 LC	5.1
Black, non-Hispanic	4.3
White, non-Hispanic	2.4
Male	2.3
Female	1.9
Nitrogen dioxide (NO2)	1.8

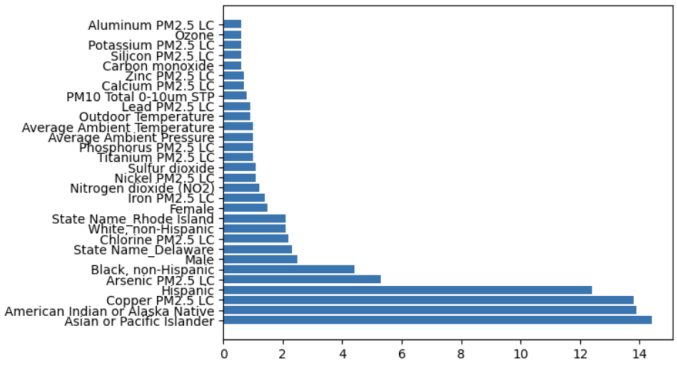


Figure 14: ROC Curve for Random Forest Classification

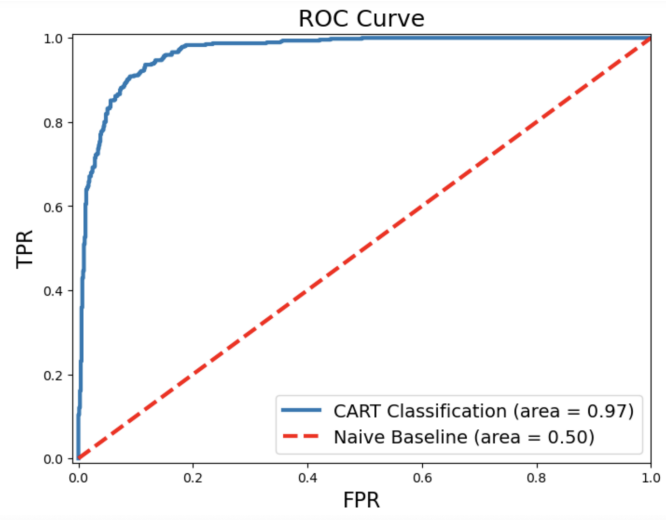
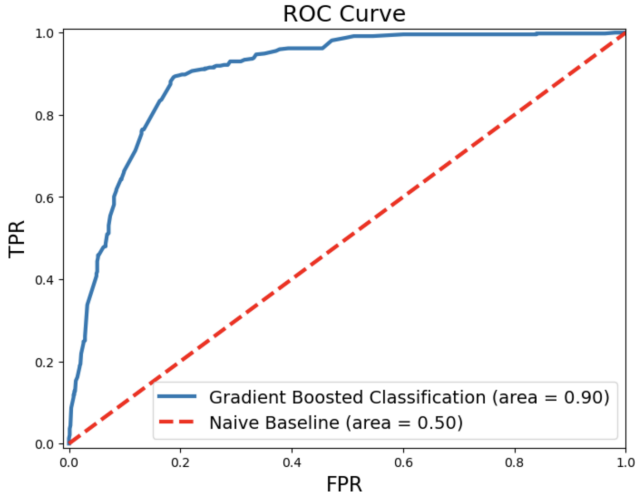


Figure 15: ROC Curve for Gradient Boosting Classification



Works Cited

Association, American Lung. What Causes Asthma? https://www.lung.org/lung-health-diseases/lung-disease-lookup/asthma/learn-about-asthma/what-causes-asthma.

Byrwa-Hill, Brandy M., et al. "Lagged Association of Ambient Outdoor Air Pollutants with Asthma-Related Emergency Department Visits within the Pittsburgh Region." International Journal of Environmental Research and Public Health, vol. 17, no. 22, Nov. 2020, p. 8619. PubMed Central, https://doi.org/10.3390/ijerph17228619.

Altman, Matthew C., et al. "Associations between Outdoor Air Pollutants and Non-Viral Asthma Exacerbations and Airway Inflammatory Responses in Children and Adolescents Living in Urban Areas in the USA: A Retrospective Secondary Analysis." The Lancet Planetary Health, vol. 7, no. 1, Jan. 2023, pp. e33–44. DOI.org (Crossref), https://doi.org/10.1016/S2542-5196(22)00302-3.

U.S. Chronic Disease Indicators: Asthma | Data | Centers for Disease Control and Prevention. https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-Asthma/us8e-ubyj/about_data.

AirData Website File Download Page. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual.