The DSE Capstone Project

Cohort 2024

The City College of New York

Analyzing Air Pollution Exposure in Affordable Housing Programs: A Data-Driven Approach to Environmental Risk

by

Sajida Malik

Master's in Data Science and Engineering

The Grove School of Engineering

December 10, 2024

Mentor: Yana Kucheva

Table of Contents

Abstract:

Federally assisted affordable housing programs, such as the Low-Income Housing Tax Credit (LIHTC) and Public Housing, are critical in addressing housing insecurity for low-income populations across the United States. However, many of these properties are situated in areas with elevated air pollution levels, particularly Particulate Matter 2.5 (PM2.5) and Ozone, exposing residents to significant health risks. This study evaluates pollution exposure levels across housing programs using socioeconomic and environmental data, employing machine learning and statistical analysis. Key findings of hypothesis testing confirmed significant differences in pollution exposure, with LIHTC properties experiencing higher PM2.5 and Ozone levels compared to Public Housing. For machine learning Stacking Classifier model achieved high accuracy for PM2.5 classification up to 85.89% for LIHTC and 86.46% for public housing properties. However, Ozone classification exhibited lower accuracy for LIHTC 82.65% and for public housing 70.75%, highlighting challenges in modeling its variability. SHAP analysis and feature importances values identified percentage of people of color as the most influential predictors of pollution exposure. Regression results demonstrated that Random Forest outperformed linear regression, achieving higher R-squared values up to 0.54 for PM2.5 in LIHTC properties. These results underscore the need for targeted policy interventions to reduce pollution exposure for federally assisted housing residents. Future work will integrate health outcomes, explore temporal trends, and develop interactive visualization tools to support evidence-based policy decisions. This research provides critical insights into the intersection of affordable housing, environmental justice, and public health.

1. Problem Statement:

1.1. Project Goals:

Federally assisted affordable housing, such as Low-Income Housing Tax Credit (LIHTC) and Public Housing programs, plays a crucial role in providing housing for low-income

populations across the United States. However, these properties are often situated in areas with elevated levels of air pollution, which poses significant health risks to vulnerable residents. Current housing policies may lack sufficient measures to mitigate environmental risks related to air quality. This project seeks to assess the exposure of federally assisted affordable housing residents to air pollutants, specifically focusing on Particulate Matter (PM2.5) and Ozone levels. By analyzing the relationship between affordable housing locations, pollution exposure, and the socioeconomic demographics of residents, this study aims to evaluate the effectiveness of current housing policies in protecting residents from environmental hazards. The results will provide data-driven insights to support policy improvements for environmental health equity.

1.2. Motivation:

Federally assisted tenants face heightened health risks due to housing policies that inadequately address environmental hazards like air pollution. This research aims to provide data-driven insights into environmental injustices impacting vulnerable communities. By analyzing air pollution exposure levels in federally assisted housing programs, this project will underscore the need for housing policies that prioritize environmental health equity, ultimately advocating for safer, more sustainable living conditions for underserved populations.

2. Introduction:

Public housing and the Low-Income Housing Tax Credit (LIHTC) program are essential components of the affordable housing landscape in the United States, serving low-income families, seniors, and individuals with disabilities. The U.S. Department of Housing and Urban

Development (HUD) oversees the Public Housing Program, which provides safe rental housing to eligible households through local housing agencies. Currently, approximately 970,000 households benefit from public housing units nationwide, highlighting its critical role in addressing housing insecurity. Similarly, the LIHTC program incentivizes private developers to build and maintain affordable housing by providing tax credits. Through LIHTC, over three million affordable housing units have been developed across the country, significantly increasing the supply of affordable housing. Both programs aim to ensure that vulnerable populations have access to stable and safe living conditions. (HUD 2024).

Air pollution is a complex mixture of harmful substances originating from both human-made and natural sources. Ground-level ozone, often referred to as smog, forms when pollutants from vehicles and industrial activities chemically react in sunlight. Another significant pollutant, particulate matter (PM), is composed of various chemicals, including sulfates, nitrates, and carbon, with fine particulate matter (PM2.5) being especially harmful due to its ability to penetrate deep into lung tissue, causing serious health issues. The intertwined relationship between air pollution and climate change stems from shared sources like fossil fuel combustion, posing global threats to human health and the environment. (NIEHS 2024).

Despite significant progress since the 1970s in reducing visible air pollution and achieving cleaner air through regulations like the Clean Air Act, air pollution remains a pressing issue in the United States. Harmful pollutants such as particulate matter (PM2.5) and ozone continue to exceed national air quality standards in many areas, posing risks to human health and the environment. Long-term exposure to PM2.5 is linked to cardiovascular and respiratory diseases, while elevated ozone levels exacerbate conditions like asthma and lung disease. Efforts by the Environmental Protection Agency (EPA) to update and enforce air quality standards, coupled

with state and local initiatives, highlight the ongoing need to address both outdoor and indoor air quality challenges to safeguard public health and ecosystems. (EPA 2024).

However, despite their importance, many properties within Public Housing and LIHTC programs are often located in areas with poor environmental conditions, including elevated levels of air pollution. This study seeks to evaluate the exposure of these federally assisted affordable housing programs to air pollution, particularly focusing on pollutants such as Particulate Matter (PM2.5) and Ozone, to assess the effectiveness of current housing policies in mitigating environmental risks for their residents.

2.1. Research Questions:

What are the levels of exposure to PM2.5 and Ozone in LIHTC and Public Housing properties?

Is there a significant difference in average air pollution exposure between LIHTC and Public housing properties?

Is there any relationship between socioeconomic, demographic factors and air pollution levels exposure in LIHTC and Public Housing Properties?

Are vulnerable communities disproportionately affected by air pollution in LIHTC and Public Housing properties?

3. Background/Related Work:

A variety of studies have investigated various aspects of affordable housing, environmental justice, and air pollution, each contributing unique perspectives to these complex issues. To

inform our project and ensure comprehensive analysis, we draw upon key insights from the following works:

The Spatial Relationship Between the Low-Income Housing Tax Credit Program and Industrial Air Pollution (2022) conducted an important investigation into whether LIHTC properties are more likely to be in neighborhoods with higher levels of industrial pollution. The study finds that, within cities, neighborhoods with LIHTC properties tend to have more industrial pollution, though this relationship changes when considering factors like racial and economic composition. These findings offer critical context for understanding how affordable housing programs might unintentionally expose vulnerable populations to environmental hazards, aligning with our project's focus on air pollution exposure to federally assisted housing. This relates closely with my research, which examines the exposure of federally assisted housing, including LIHTC and Public Housing properties, to air pollution levels like PM2.5 and Ozone.

Another study, An Environmental Justice Analysis of Air Pollution Emissions in the United States from 1970 to 2010 (2024), highlighted how racial and socioeconomic disparities persist despite overall reductions in air pollution levels following the Clean Air Act. This study's findings that low-income and minority communities continue to bear a disproportionate pollution burden reinforce the need for our project's analysis of housing policy effectiveness in mitigating environmental risks. My research relates to this by considering socioeconomic and demographic factors to assess whether current housing policies inadvertently place vulnerable populations in environmentally hazardous areas, providing critical insights into the intersection of affordable housing and environmental justice.

The study Intersections Among Housing, Environmental Conditions, and Health Equity: A Conceptual Model for Environmental Justice Policy (2024) explored how housing and

environmental conditions are deeply intertwined and play a significant role in shaping health outcomes. This review highlighted the links between housing disparities and environmental injustices, focusing on how low-income and minority communities are disproportionately affected by environmental risks due to historical practices such as redlining and unfair zoning. It proposed a conceptual model that links poor housing quality with environmental hazards, advocating for better housing code enforcement, more affordable housing, and stricter air quality standards to address inequities and promote health equity. This study aligns with my research by exploring how housing and environmental conditions are deeply intertwined, disproportionately affecting low-income and minority communities through environmental risks.

Another study on Reducing Mortality from Air Pollution in the United States by Targeting Specific Emission Sources (2020) examined that in the United States, despite significant improvements in air quality, air pollution is still linked to a considerable number of deaths each year, ranging from 100,000 to 200,000. To effectively and fairly reduce these deaths, it's important to pinpoint specific sources of pollution. However, most existing studies focus on broad categories or only a few sources. This research aims to fill that gap by estimating how many deaths are caused by human-made emissions of primary and secondary PM2.5. The findings break down the sources into four distinct groups, helping to identify the most effective strategies for reducing pollution-related deaths. Remarkably, five key activities across various sectors account for almost half of the deaths, with a significant portion resulting from fossil fuel combustion. Other causes include non-fossil fuel combustion, agriculture, and non-combustion activities. Both types of PM2.5, including those from pollutants like ammonia that are not currently regulated, are critical contributors. The study suggests that continuing to target traditional pollution sources, along with developing new methods to address emerging ones, can improve both public health and environmental sustainability. This research relates to my study on analyzing air pollution exposure in affordable housing programs by providing insights into the

sources of air pollution and their disproportionate impact, which may help identify how housing policies intersect environmental risks.

4. Methodology/Design:

4.1. Data Sources and Variables:

This study utilizes data from the National Housing Preservation Database (NHPD) and the U.S. Environmental Protection Agency's (EPA) Environmental Justice Screen (EJScreen) dataset. The NHPD provides geographic information on federally assisted housing properties, including Census Tract, longitude, latitude, city, state, county and housing program (Low Income Housing Tax Credit (LIHTC) and Public Housing). The EJScreen dataset includes air pollution metrics such as PM2.5 and Ozone levels, as well as socioeconomic and demographic variables like percentages of low-income populations and people of color. The detailed features of the dataset are below:

4.1.1. **Geographic Variables:** Census Tract, longitude, latitude, city, state, and county.

4.1.2. **Target Variables (Air Pollution Metrics):** PM2.5 and Ozone levels, classified into health categories based on EPA standards as follows:

Classification of PM2.5 levels

- Good: 0-9
- Moderate: 9.1- 35.4
- Unhealthy for Sensitive Group: 55.5-125.4
- Unhealthy: 125.5-225.4

- Very unhealthy: 125.4- 225.4
- Hazardous: 225.5+

Classification of Ozone Levels

- Good: 0- 54
- Moderate: 55- 70
- Unhealthy for Sensitive Group: 71-85
- Unhealthy: 86-105
- Very unhealthy: 106-200
- Hazardous: 201

4.1.3. **Socioeconomic and Demographic Variables:** Percentage of people of color (PEOPCOLORPCT), low-income percentage (LOWINCPCT), unemployment percentage (UNEMPPCT), Under 5 years of age percentage (UNDER5PCT), Over 64 years of age percentage (OVER64PCT), Life expectancy percentage (LIFEEXPPCT)

4.2. Tools and Software:

Data preprocessing and analysis were conducted using Python, which provided a versatile platform for exploratory data analysis, statistical analysis and machine learning. ArcGIS Pro was utilized for spatial analysis, allowing for the creation of detailed geospatial visualizations to map air pollution exposure across Census Tracts.

4.3. Analytical Framework:

The study followed a structured analytical framework comprising six key stages:

4.3.1. Data Collection and Preparation:

Datasets from the NHPD and EPA were loaded, cleaned, and merged using Census Tract. The analysis focused on active LIHTC and Public Housing properties, ensuring that only relevant entries were included. Data cleaning involved handling missing values and formatting variables to ensure consistency across datasets.

4.3.2. Exploratory Data Analysis (EDA):

Descriptive statistics were generated to summarize the data. Histograms were created to check the distribution of the quantitative features of the datasets. Visualizations, including heatmaps, were created to identify patterns in the data.

4.3.3. Statistical Analysis:

Inferential statistical tests, including T-tests and Mann-Whitney U-tests, were applied to compare pollution exposure across housing types. Correlation analysis was conducted to examine initial relationships between air pollution levels and socioeconomic characteristics, such as the percentage of People of Color, Low-Income residents, Unemployed residents, residents Under Age 5, Over Age 64, and Life Expectancy. Regression modeling, with linear regression used as a baseline and Random Forest regression applied subsequently, was employed to quantify the impact of socioeconomic characteristics on air pollution levels. This

approach provided deeper insights into the relationships identified through correlation analysis.

4.3.4. Feature Visualization:

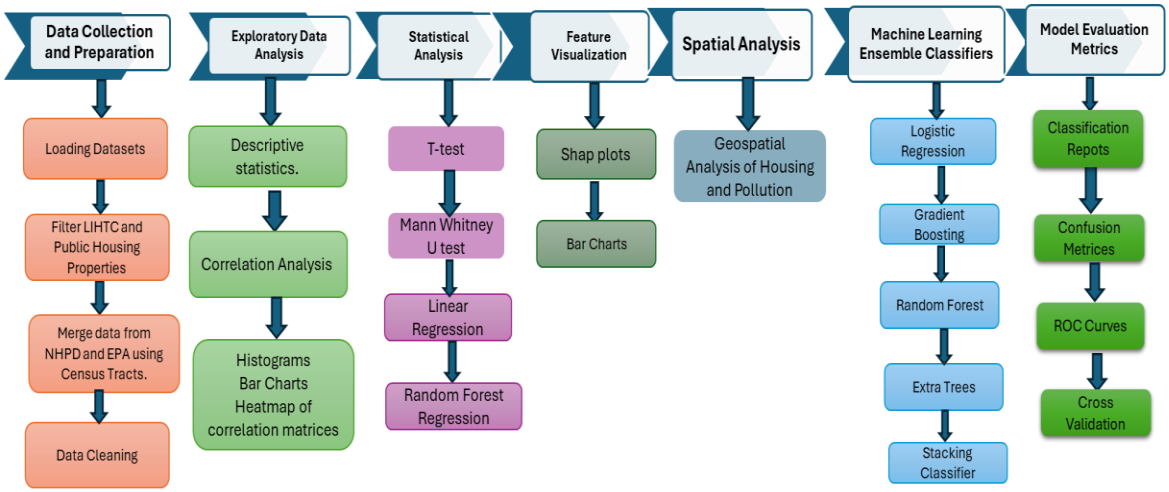
SHAP (Shapley Additive Explanations) plots were used to identify and visualize the importance of various features in predictive models.

4.3.5. Machine Learning:

A range of ensemble classifiers, including Logistic Regression, Gradient Boosting, Random Forest, and Extra Trees, were implemented to predict air pollution exposure levels PM2.5 and Ozone. A stacking classifier combining multiple models was used to improve prediction performance. SMOTE over sampling technique was used to balance the classes. Model evaluation metrics, including classification reports, confusion matrices, Receiver operating curves (ROC), and cross-validation, were employed to assess the accuracy and robustness of the models.

4.3.6. Spatial Analysis:

Geospatial analysis was performed using ArcGIS Pro to visualize the geographic distribution of pollution levels relative to housing locations. Interactive maps were generated to highlight areas with elevated pollution exposure, enabling a deeper understanding of spatial disparities.



5. Results:

The dataset comprises a total of 30,321 LIHTC buildings and 5,601 Public Housing buildings, providing a substantial basis for analyzing air pollution exposure across these federally assisted housing types.

5.1. Exploratory Data Analysis:

Table 1: Descriptive Statistics of all quantitative features of the dataset for LIHTC and Public Housing Properties:

	LIHTC	Public Housing

Variables	Mean	Standard Deviation	Mean	Standard Deviation
PM2.5	8.46	2.20	7.98	1.64
Ozone	60.84	8.68	59.14	6.23
% of People of Color	0.51	0.31	0.50	0.32
% Low Income residents	0.41	0.18	0.47	0.18
% Unemployed Residents	0.07	0.05	0.08	0.06
% Underage 5	0.06	0.03	0.06	0.03
% Over Age 64	0.16	0.07	0.17	0.07
% of Life Expectancy	**0.21**	**0.04**	**0.23**	**0.04**

Table 1 shows the mean and standard deviation of all quantitative features in LIHTC and public housing properties. The average pollution level is higher in LIHTC.

Figure 1: Data Distribution of LIHTC Properties

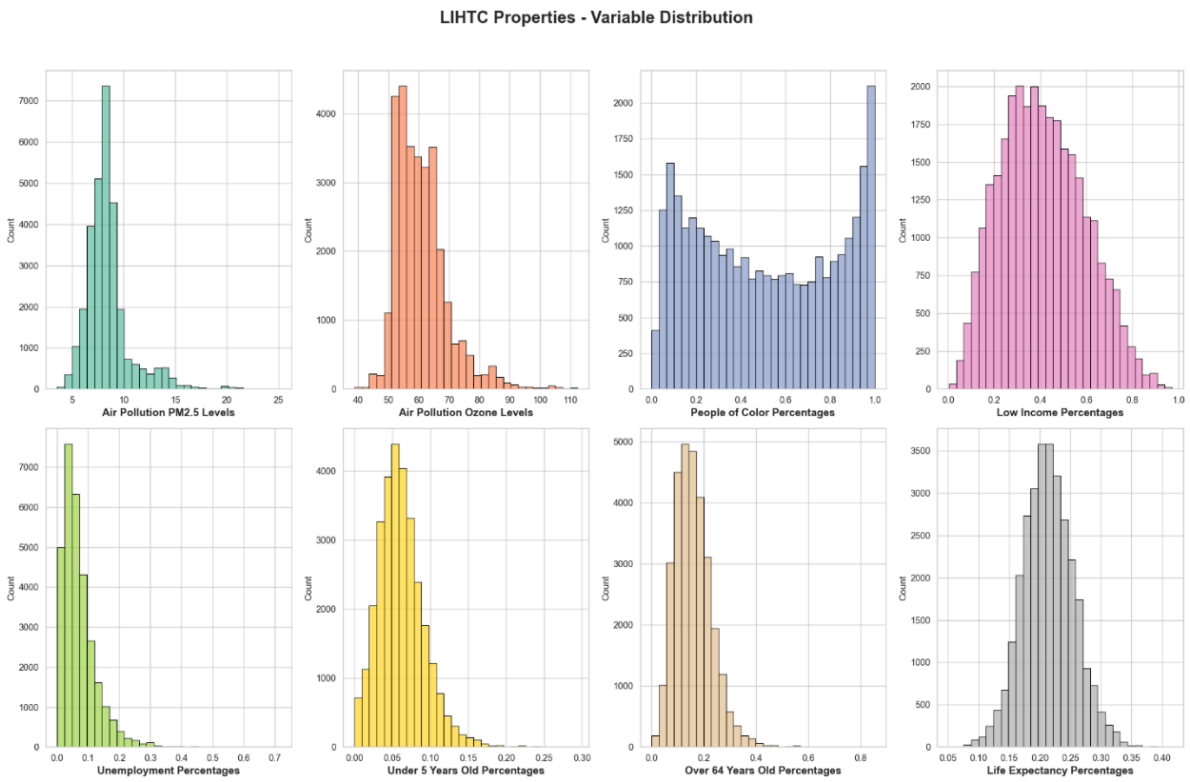
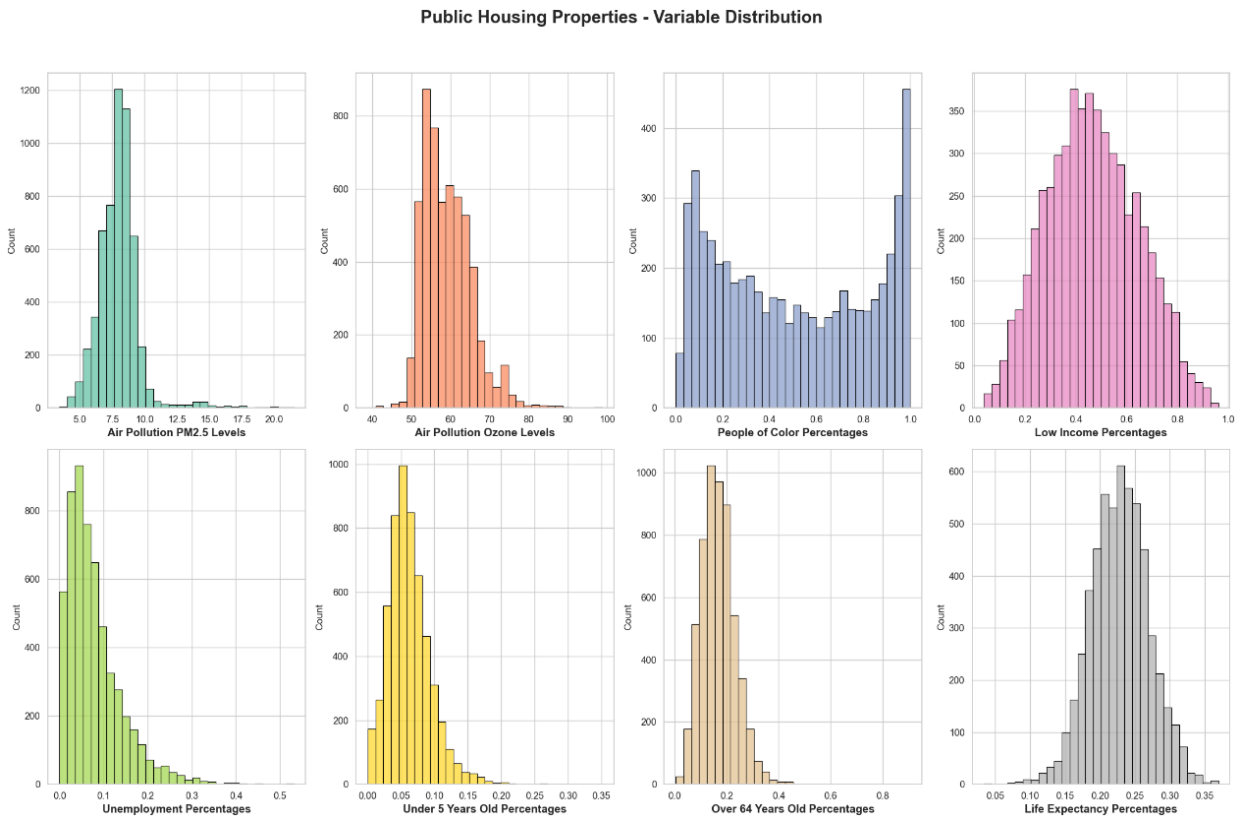


Figure 2: Data Distribution of Public Housing Properties



5.2. Statistical Analysis:

5.2.1. Hypothesis Testing:

T- test:

Null Hypothesis (H0): The mean air pollution exposure (PM2.5 and Ozone levels) for LIHTC properties is the same as for Public Housing properties.

Alternative Hypothesis (H1): The mean air pollution exposure (PM2.5 and Ozone levels) is higher in LIHTC properties as compared to Public Housing properties.

Table 2: Results of T-Test for PM2.5 and Ozone between LIHTC and Public Housing:

	PM2.5	Ozone

t-statistic	18.812	17.53
p-value	1.4063e-77	8.807e-68

The T-test resulted in a p-value of less than 0.05, confirming that the average (mean) exposure to air pollution is significantly higher for LIHTC properties compared to Public Housing properties.

Mann-Whitney U test:

Null Hypothesis (H0): The distribution of air pollution exposure (PM2.5 and Ozone levels) for LIHTC properties is the same as for Public Housing properties.

Alternative Hypothesis (H1): The distribution of air pollution exposure (PM2.5 and Ozone levels) differs between LIHTC properties and Public Housing properties.

Table 3: Results of Mann Whitney U test for PM2.5 and Ozone between LIHTC and Public Housing:

	PM2.5	**Ozone**
p-value	5.1951e-30	1.003e-26

The Mann-Whitney U test yielded p-values less than 0.05 for both PM2.5 and Ozone levels, indicating a significant difference in the overall distribution of exposure levels between LIHTC and Public Housing properties. The results suggest that LIHTC properties are generally associated with higher pollution levels.

5.2.2. Correlation Analysis:

Figure 3: Correlation Matrix of all Quantitative Variables in LIHTC Properties:

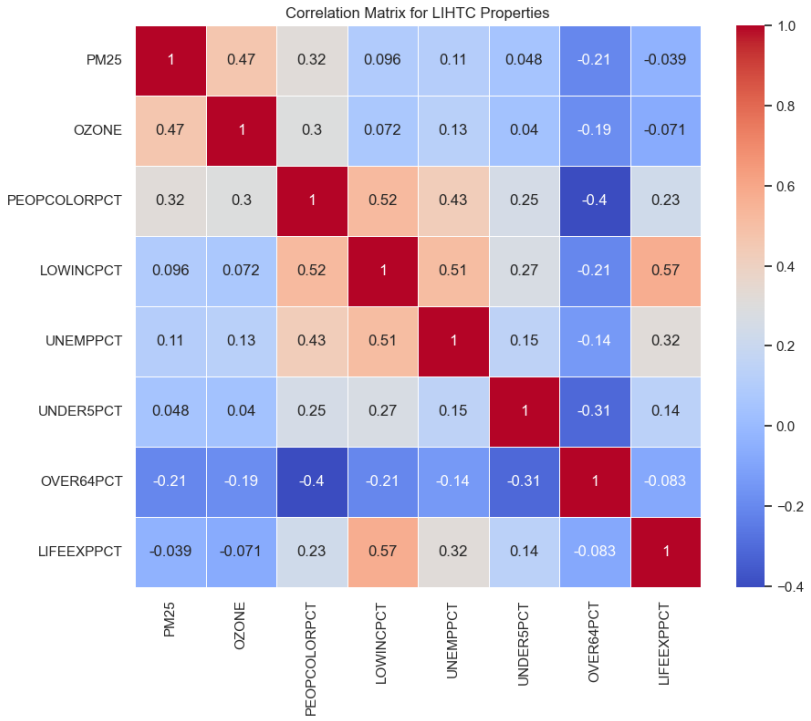
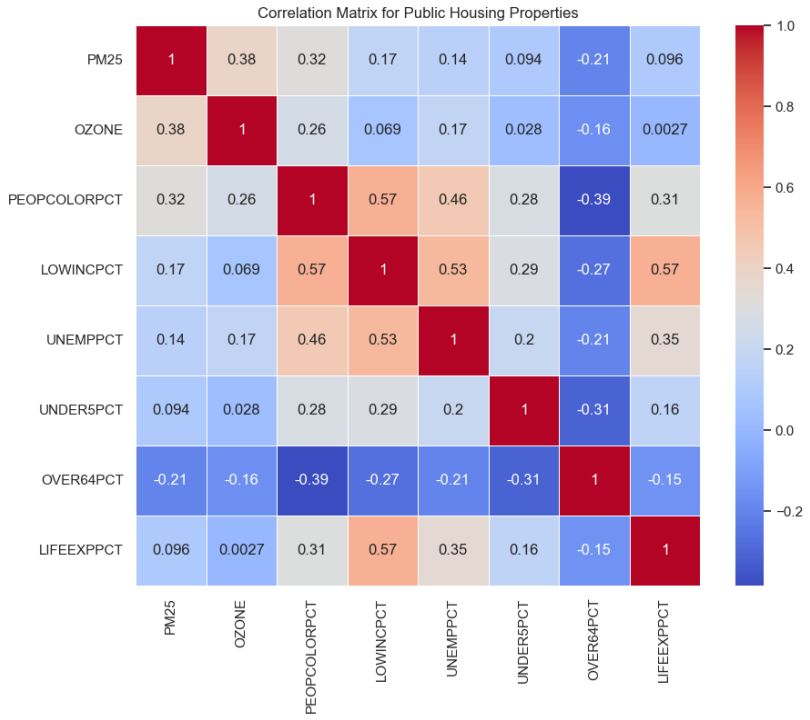


Figure 4: Correlation Matrix of all quantitative variables in Public Housing Properties:



People of color is positively correlated with PM2.5 and Ozone in both LIHTC and Public Housing Properties.

5.2.3. Regression Analysis:

The performance of regression models for predicting air pollution levels, specifically PM2.5 and Ozone, was evaluated for LIHTC and Public Housing properties. The results are summarized below:

Table 4: Results of Regression Analysis

	LIHTC		Public Housing	
	PM2.5	**Ozone**	**PM2.5**	**Ozone**

Air Pollution Levels								
Regression	**Linear Regression**	**Random Forest**	**Linear Regression**	**Random Forest**	**Linear Regression**	**Random Forest**	**Linear Regression**	**Random Forest**
Testing R-squared	0.12	0.54	0.11	0.52	0.10	0.19	0.08	0.19
Testing Mean Square Error (MSE)	4.26	2.24	64.90	34.96	2.07	1.87	35.86	31.50

Figure 5: Comparison of Linear and Random Forest Regression for PM2.5 in LIHTC

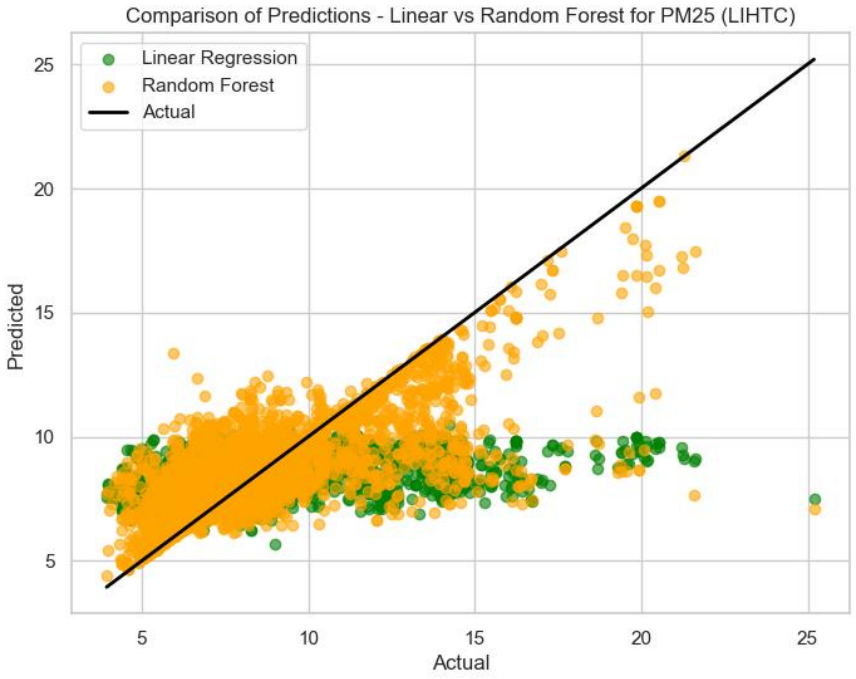


Figure 6: Comparison of Linear and Random Forest Regression for Ozone in LIHTC

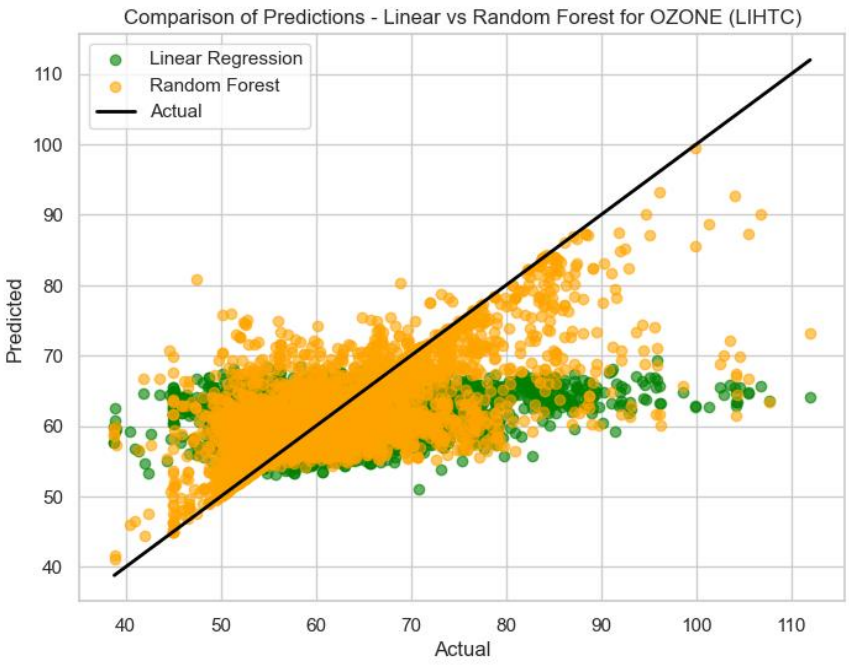


Figure 7: Comparison of Linear and Random Forest Regression for PM2.5 in Public Housing

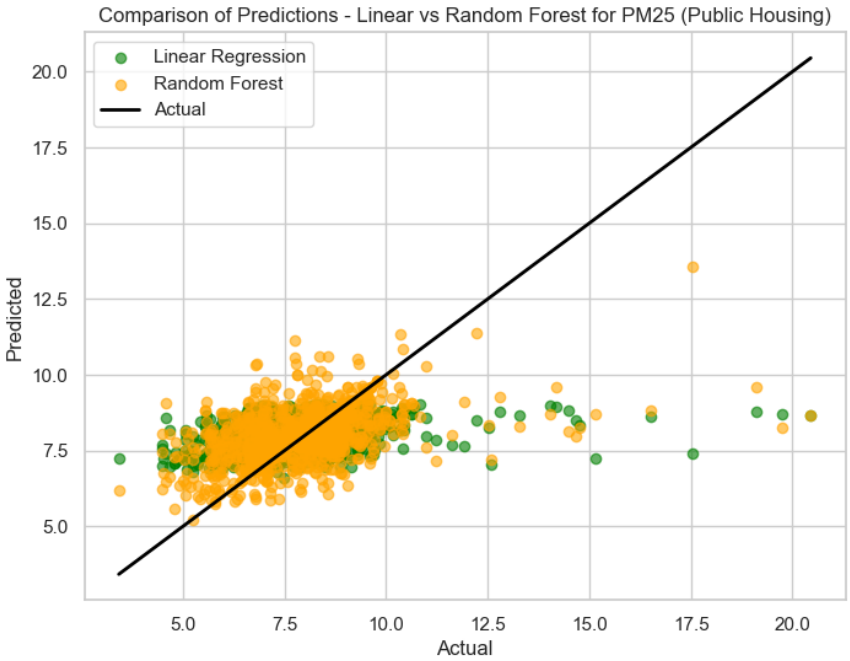
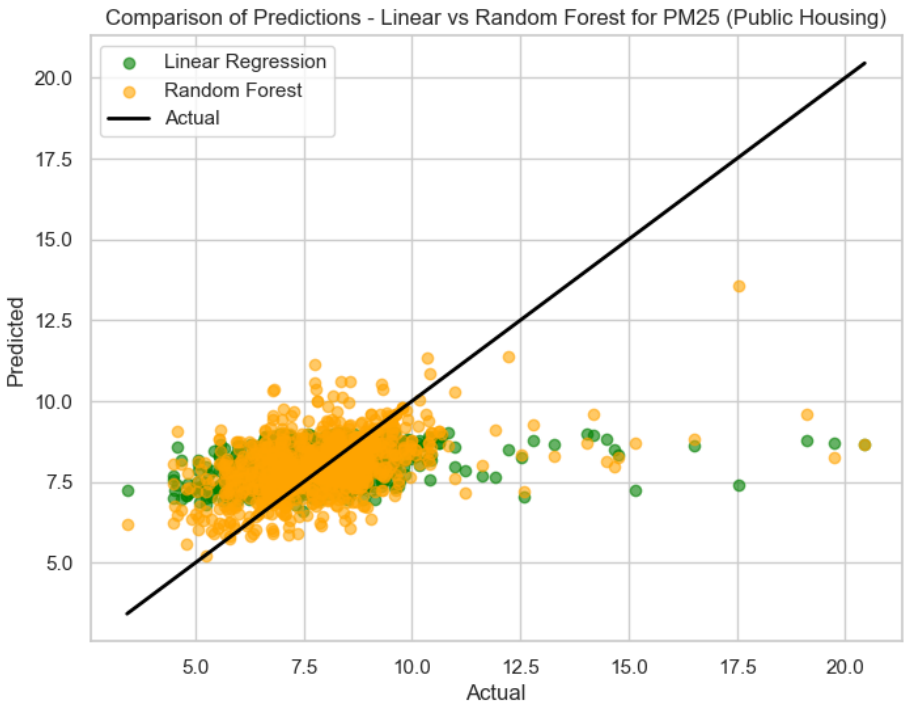


Figure 8: Comparison of Linear and Random Forest Regression for Ozone in Public Housing



The lower MSE value and higher R squared values indicate that Random Forest's predictions are closer to the actual values than those of linear regression, showing it to be a more accurate model in this context. The scatter plots compare the actual PM2.5 and Ozone values to the predicted values for both models. The closer the points are to the diagonal line, the better the predictions. Random Forest (yellow points) is generally closer to the diagonal line, whereas Linear Regression (green points) is more spread out, indicating less accuracy.

5.3. Feature Importances and Visualization

Table 5: Feature Importances ranking for Air Pollution Predictors

	LIHTC	Public Housing

Variables	PM2.5	Ozone	PM2.5	Ozone
% of People of Color	0.26	0.23	0.26	0.22
% Low Income residents	0.15	0.15	0.15	0.17
% Unemployed Residents	0.15	0.14	0.14	0.16
% Underage 5	0.14	0.15	0.13	0.15
% Over Age 64	0.15	0.17	0.17	0.16
% of Life Expectancy	0.14	0.15	0.15	0.15

Figure 9: Shap Analysis for PM2.5 Feature Importance for LIHTC

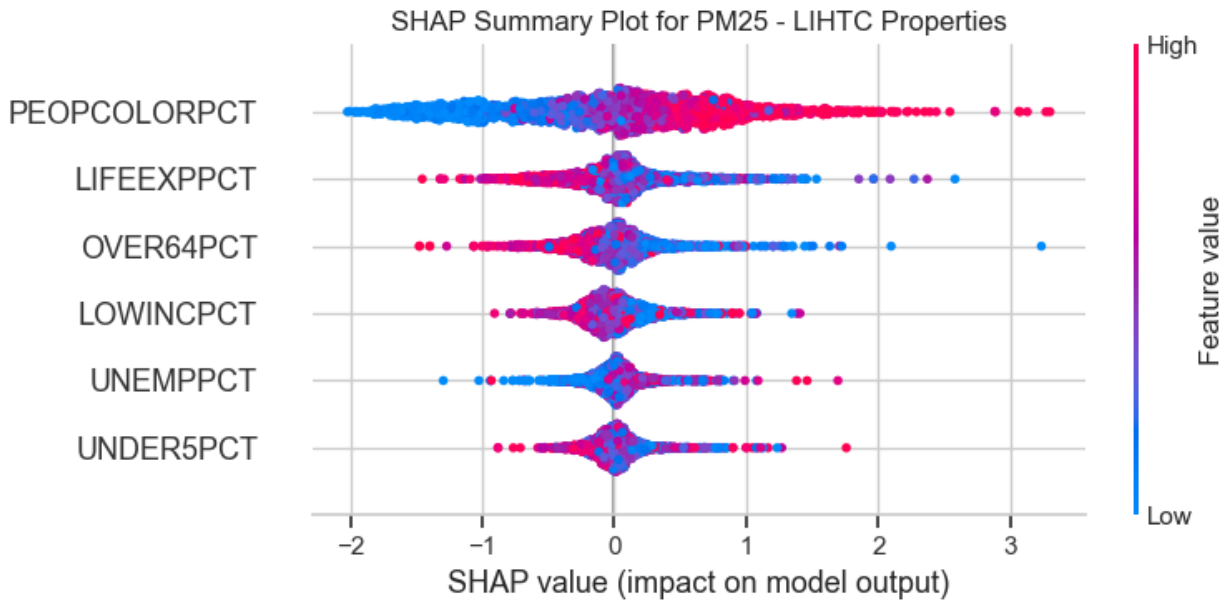


Figure 10: Shap Analysis for Ozone Feature Importance for LIHTC

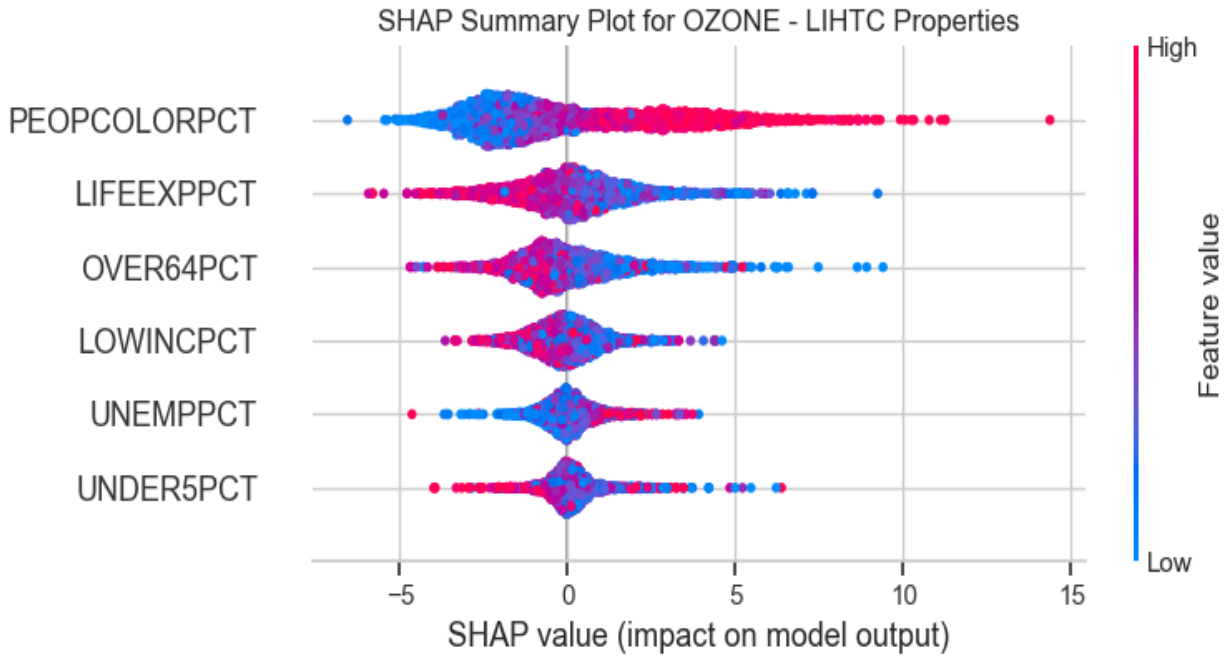


Figure 11: Shap Analysis for PM2.5 Feature Importance for Public Housing

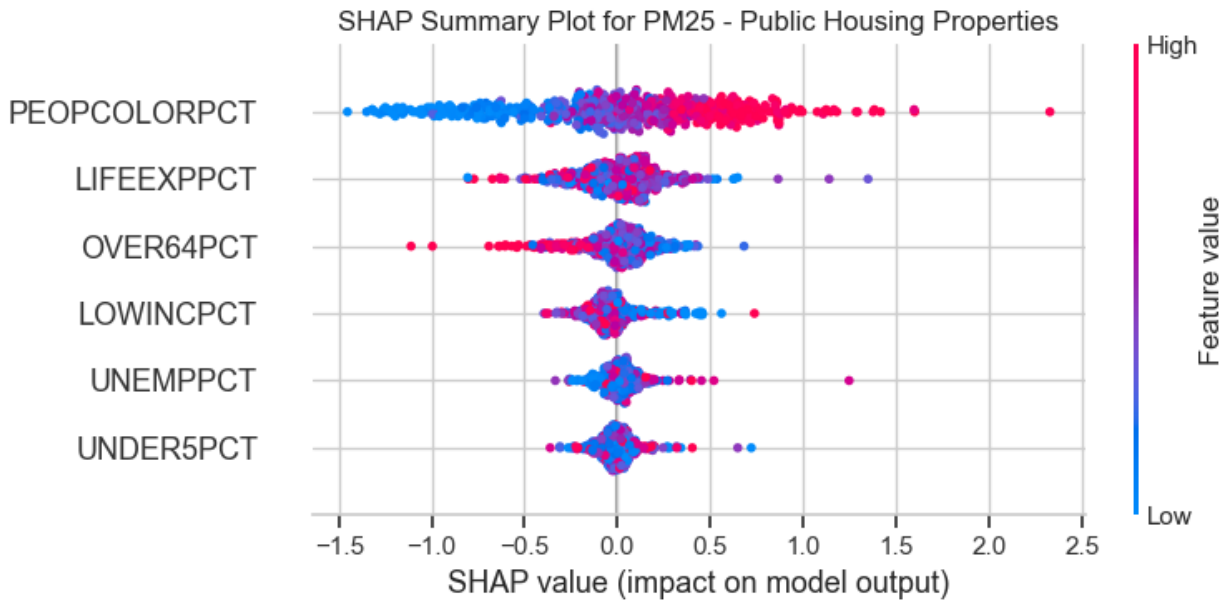
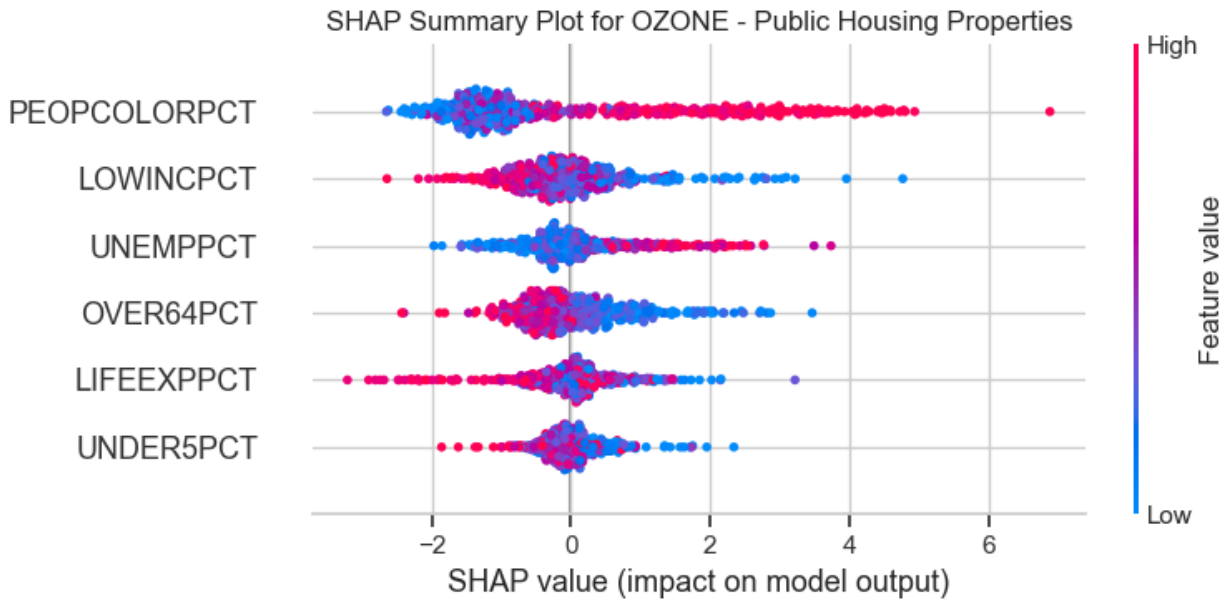


Figure 12: Shap Analysis for Ozone Feature Importance for Public Housing

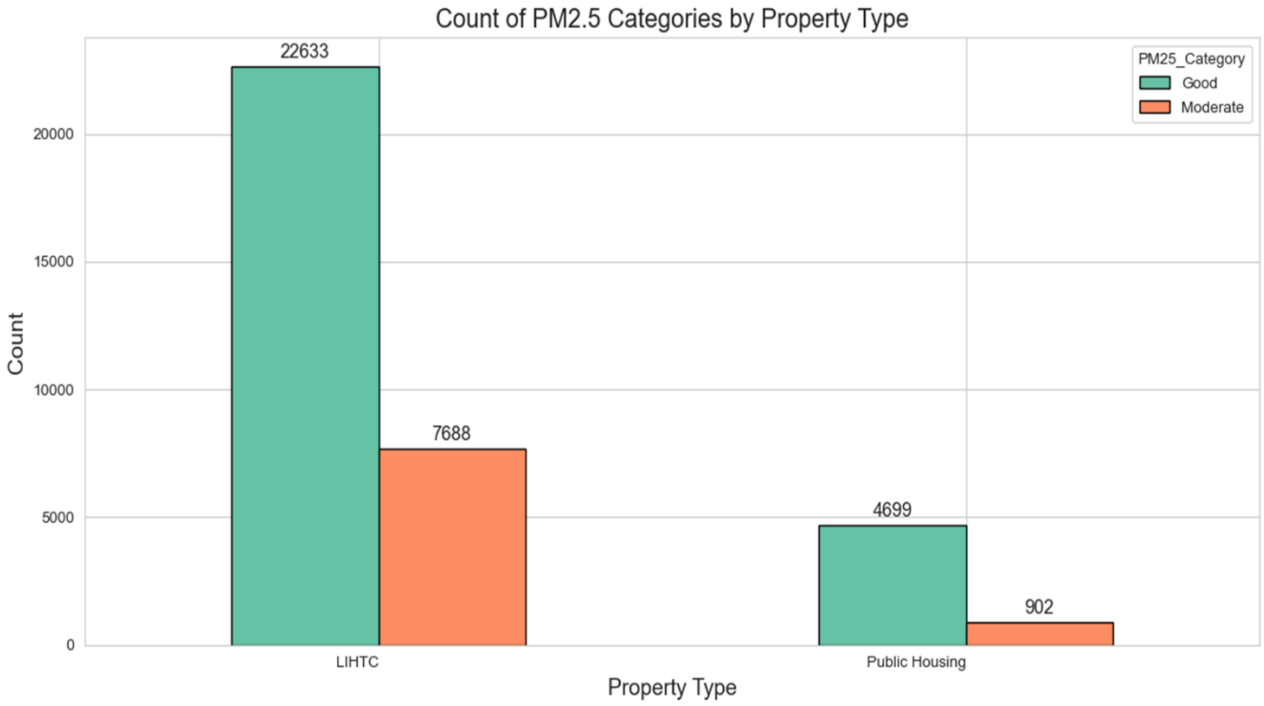


Percentages of people of color consistently shows the highest impact on model output, indicating it is a significant predictor for pollution levels. High values contribute positively to the output, suggesting areas with higher percentages of people of color are associated with higher predicted pollution levels.

5.4. Machine Learning Analysis:

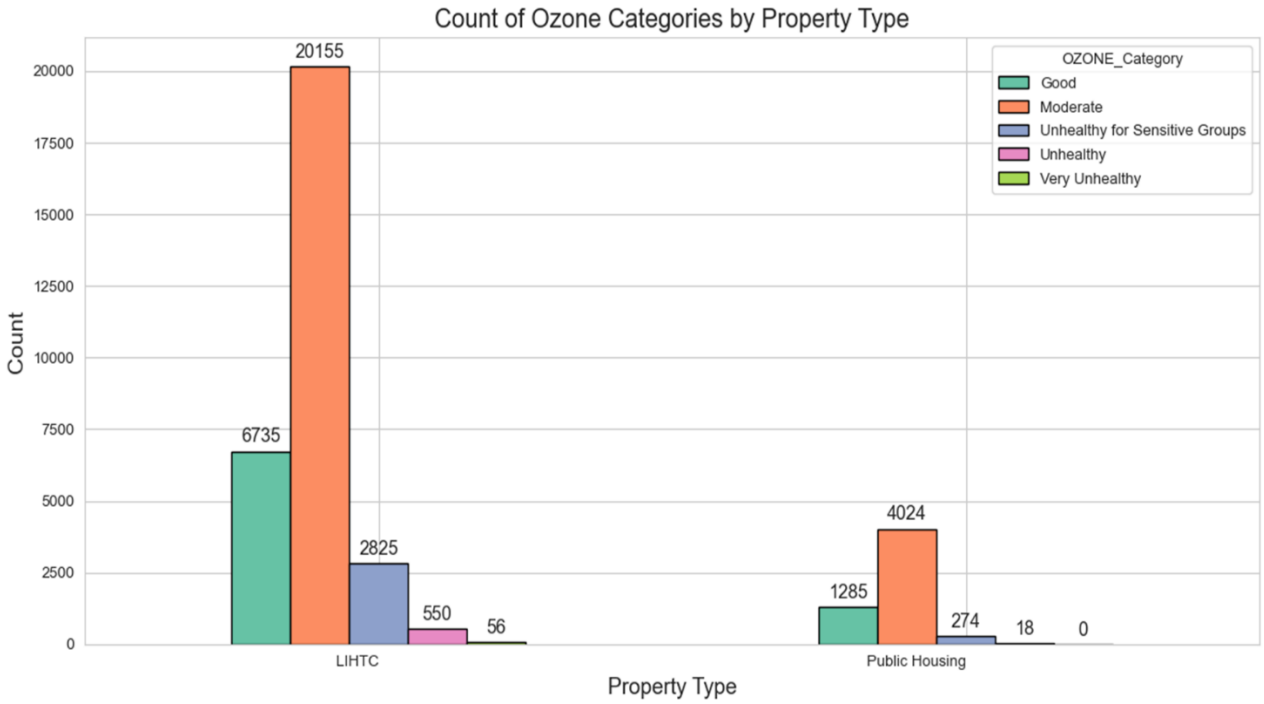
The data split is 70%/30%/10% for training, testing and validation set.

Figure 13: Distribution of PM2.5 levels among affordable housing Programs



Total 7688 LIHTC and 902 public housing properties have moderate exposure to PM2.5 levels.

Figure 14: Distribution of Ozone levels among affordable housing Programs



Total 20155 LIHTC and 4024 Public Housing are exposed to moderate ozone levels.

Figure 15: Test Accuracies of Machine Learning Ensemble Classifiers for PM2.5 in LIHTC Properties

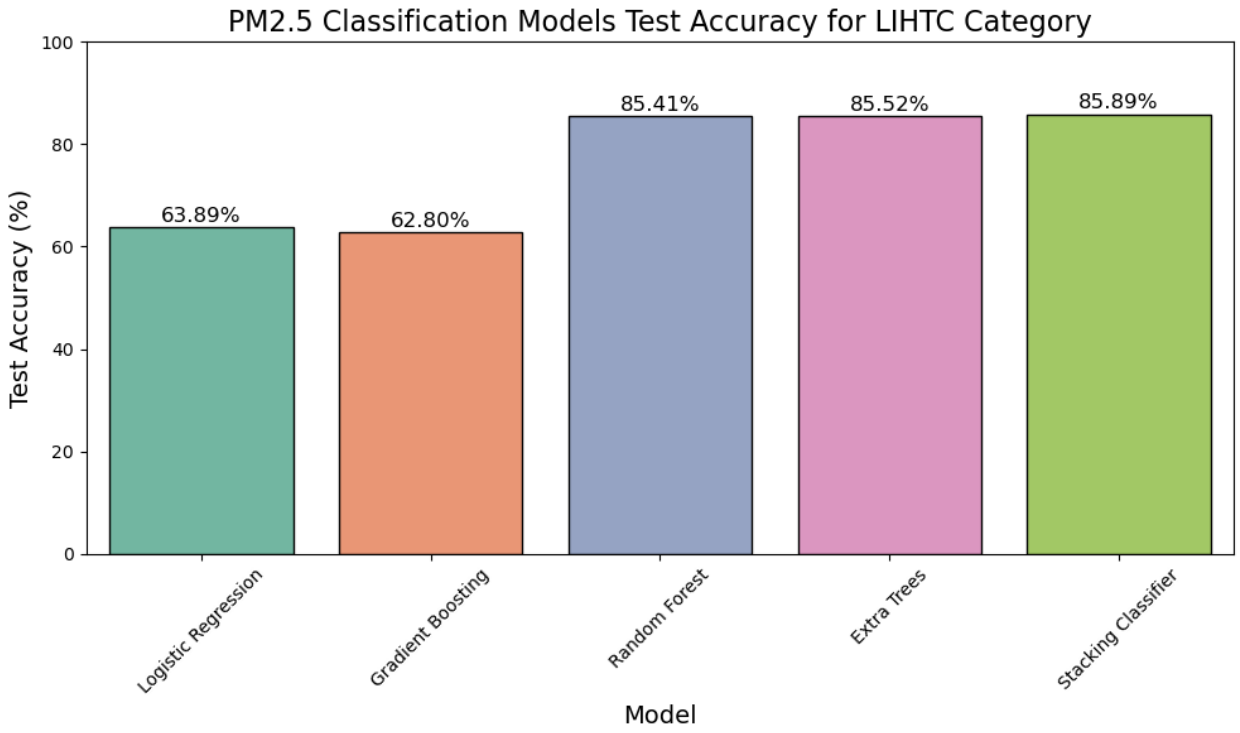
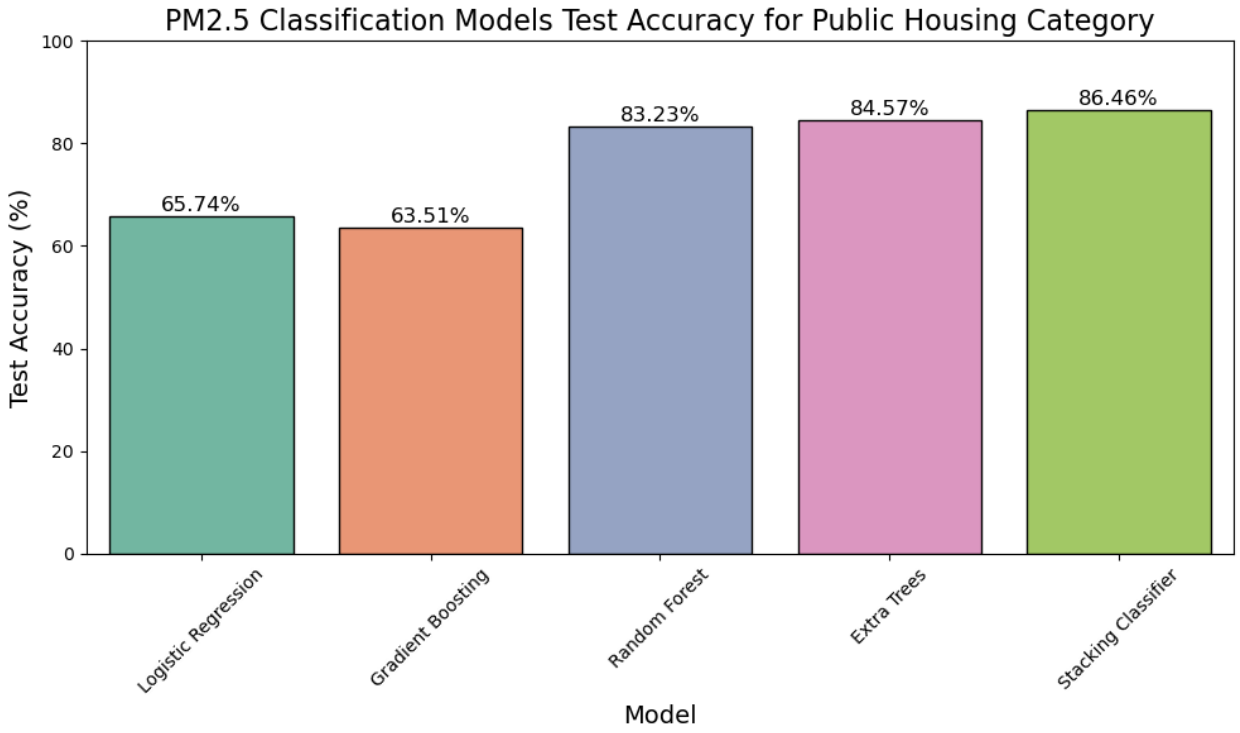
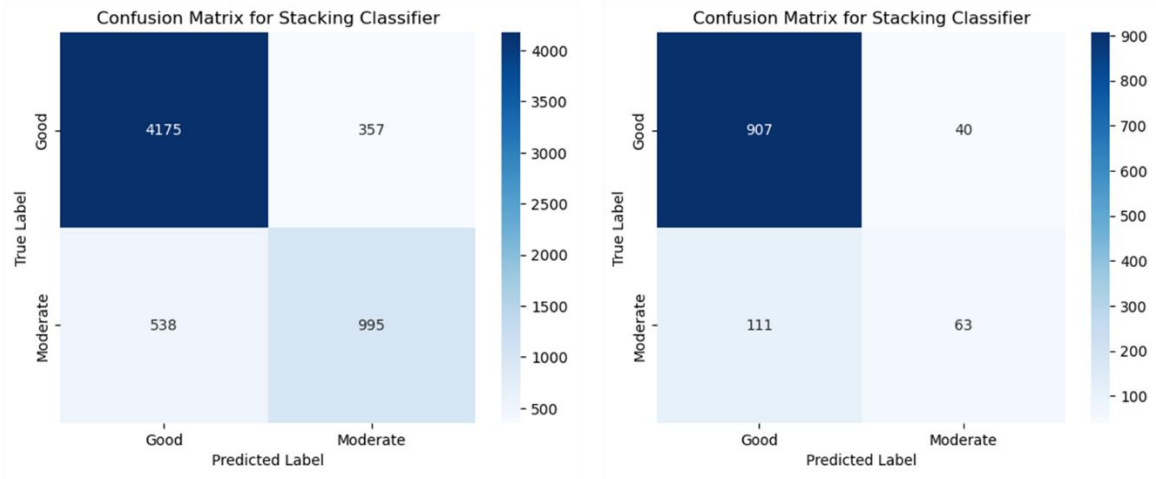


Figure 16: Test Accuracies of Machine Learning Ensemble Classifiers for PM2.5 in Public Housing Properties



Stacking Classifier performed better in predicting PM2.5 Levels as compared to other classifiers.

Figure 17: Confusion Matrices of PM2.5 for LIHTC and Public Housing Properties for Stacking Classifier



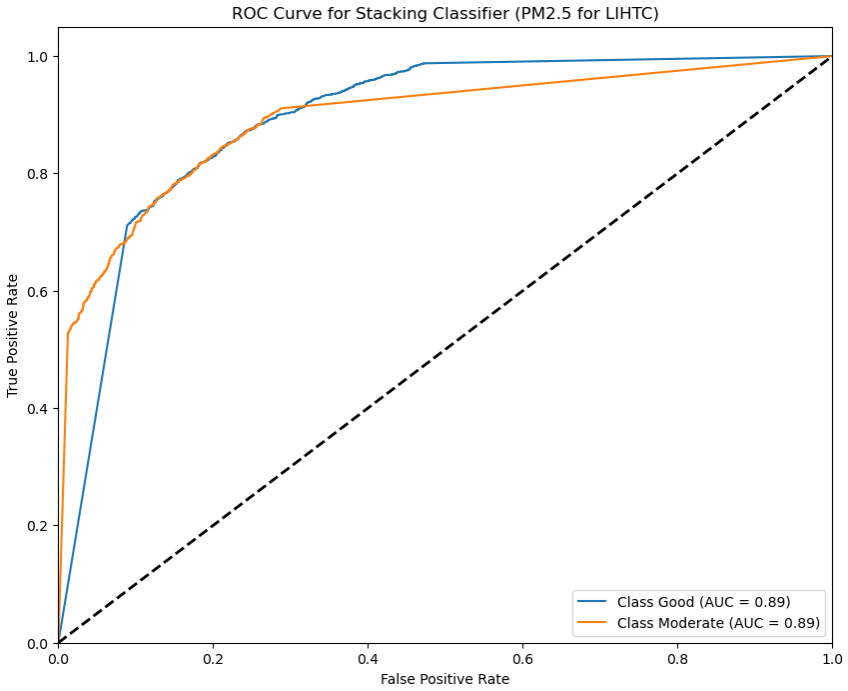
The model performed better in predicting Good as compared to moderate classes for both properties.

Table 6: Classification Report for Stacking Classifier in Predicting PM2.5 for LIHTC and Public Housing

Classification Report	LIHTC		Public Housing	
	PM2.5		PM2.5	
	Good	Moderate	Good	Moderate
Precision	0.89	0.74	0.90	0.54
Recall	0.92	0.68	0.93	0.43
F1-Score	0.91	0.71	0.92	0.48

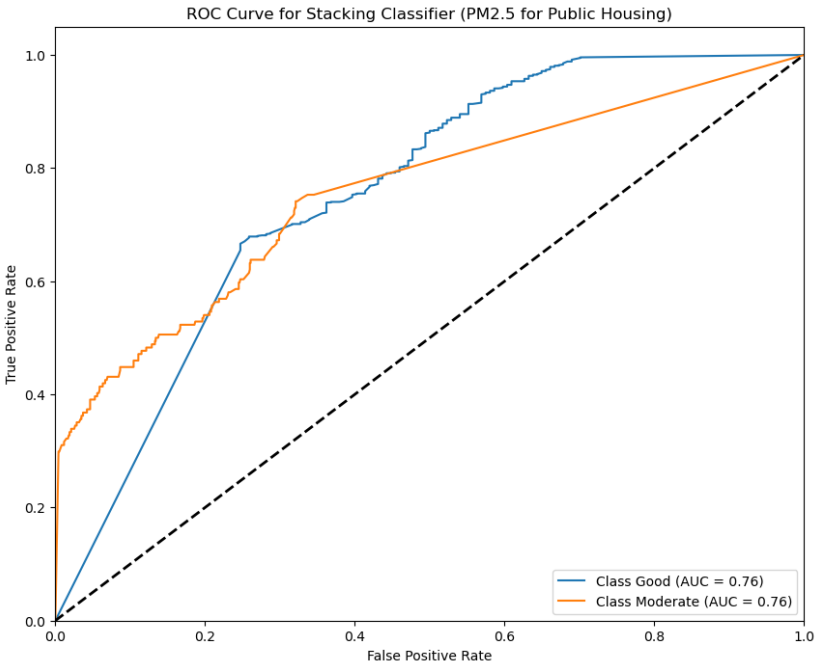
The classification report indicates that the model achieved high performance in predicting "Good" PM2.5 levels for both LIHTC and Public Housing properties, with precision, recall, and F1-scores exceeding 0.89, while performance for "Moderate" levels was comparatively lower, especially for Public Housing.

Figure 18: Receiver Operating Curve (ROC) for PM 2.5 for LIHTC



The stacking classifier is performing consistently well for both classes ("Good" and "Moderate") in distinguishing between different PM2.5 pollution levels, as indicated by the high AUC (Area under the curve) score (0.89)

Figure 19: Receiver Operating Curve (ROC) for PM 2.5 for Public Housing



The stacking classifier is performing consistently for both classes ("Good" and "Moderate") in distinguishing between different PM2.5 pollution levels, as indicated by the high AUC score (0.76).

Figure 20: Test Accuracies of Machine Learning Ensemble Classifiers for Ozone in LIHTC Properties

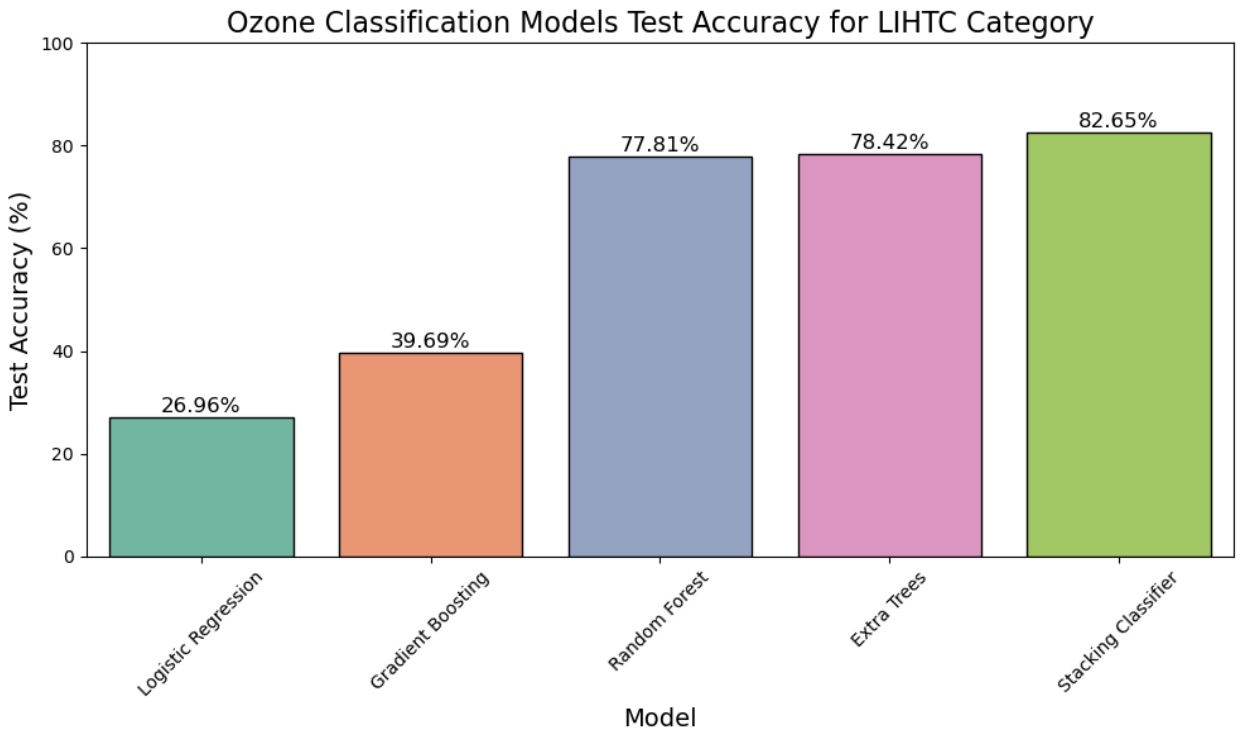
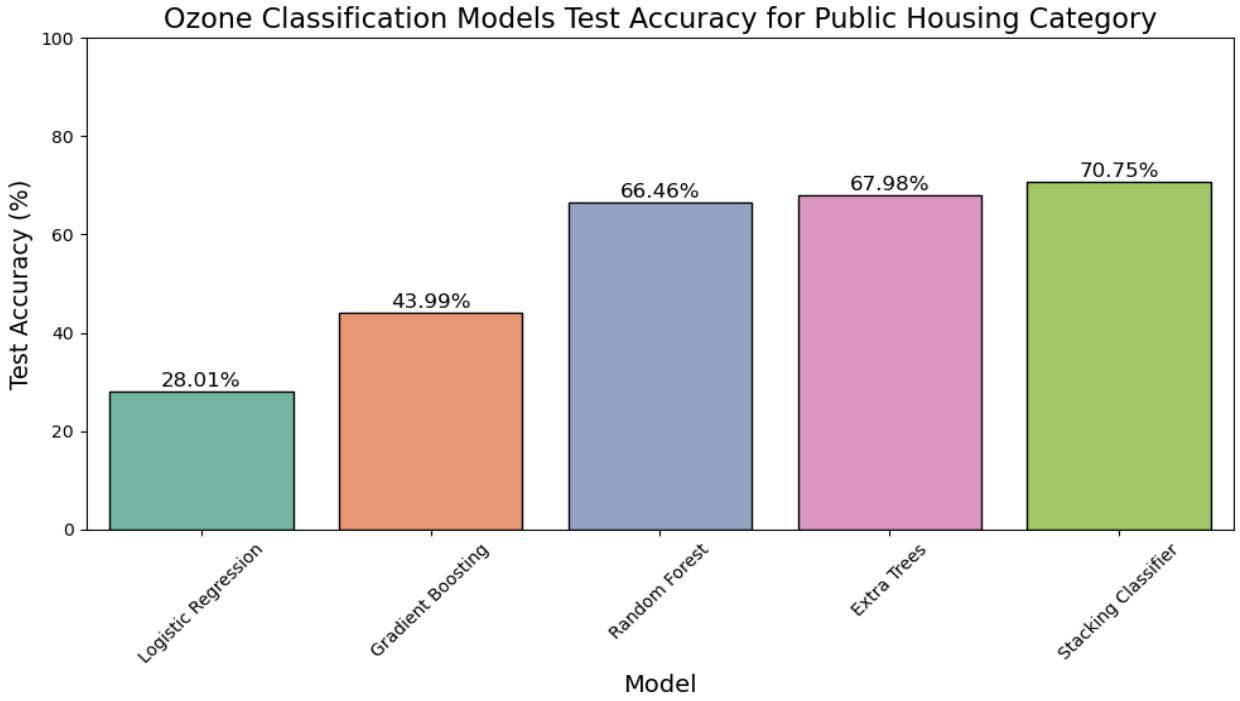
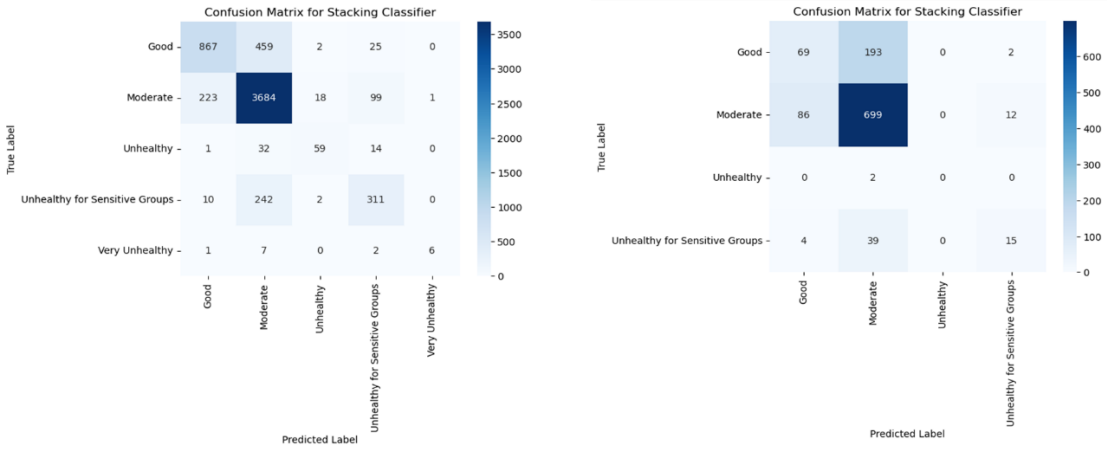


Figure 21: Test Accuracies of Machine Learning Ensemble Classifiers for Ozone in Public Housing Properties



Stacking Classifier performed better in predicting PM2.5 Levels as compared to other classifiers.

Figure 22: Confusion Matrices of Ozone for LIHTC and Public Housing Properties for Stacking Classifier



The model performed well in predicting good and moderate categories as compared to Unhealthy classes.

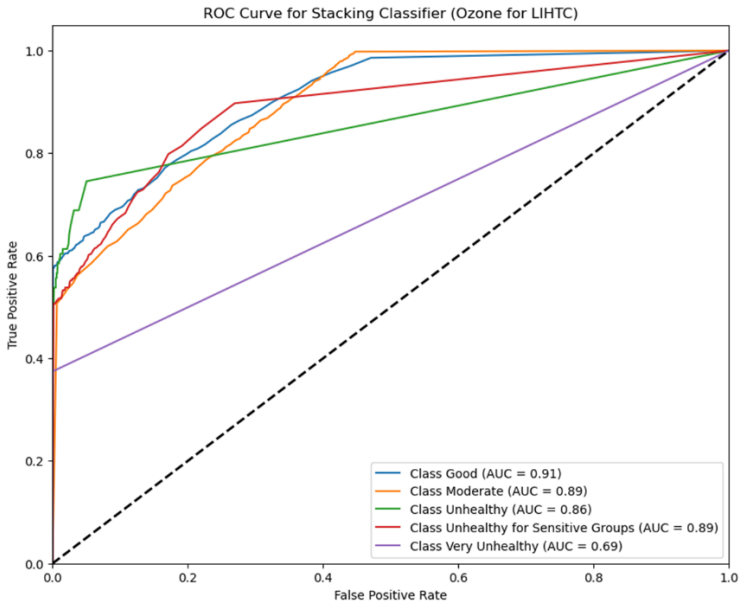
Table 7: Classification Report for Stacking Classifier in Predicting Ozone for LIHTC and Public Housing

	LIHTC	Public Housing
Classification Report	**Ozone**	**Ozone**

	Good	Moderate	Unhealthy for Sensitive Group	Unhealthy	Very Unhealthy	Good	Moderate	Unhealthy for Sensitive Group
Precision	0.78	0.83	0.62	0.67	0.86	0.48	0.75	0.47
Recall	0.64	0.91	0.58	0.55	0.38	0.25	0.89	0.29
F1-Score	0.70	0.87	0.60	0.60	0.52	0.33	0.81	0.36

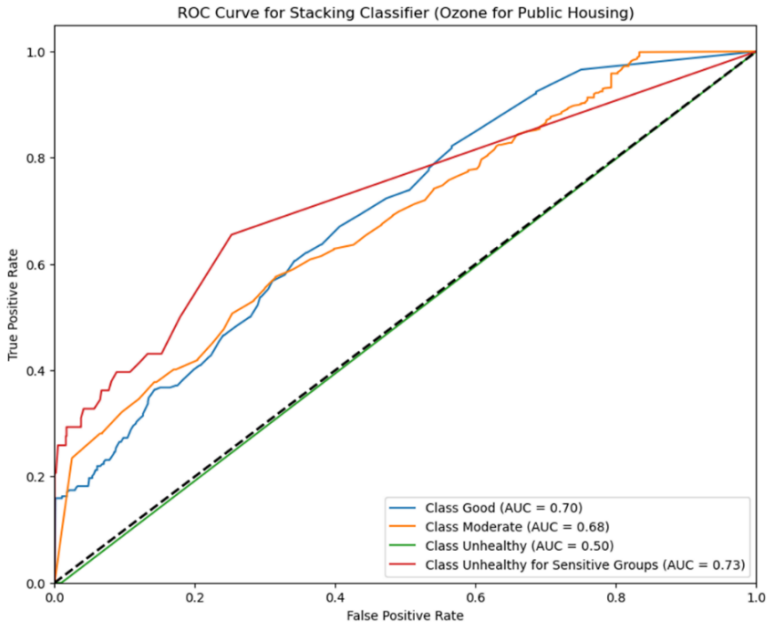
There is no very unhealthy exposure and Unhealthy has very small exposure of Ozone for Public Housing so the classification report for public housing is for just three categories. The F1-scores reveal that the model performs best for "Moderate" Ozone levels in both LIHTC and Public Housing, reflecting a balanced precision and recall. However, the performance decreases significantly for more severe categories like "Unhealthy" and "Unhealthy for Sensitive Group," particularly for Public Housing, highlighting challenges in accurately predicting these cases.

Figure 23: Receiver Operating Curve (ROC) for Ozone for LIHTC



The classifier performs very well for most classes, especially "Good" and "Moderate". Performance declines significantly for the "Very Unhealthy" category, due to limited data for this class or class imbalance in the dataset.

Figure 24: Receiver Operating Curve (ROC) for Ozone for Public Housing



Overall performance for Public Housing is weaker compared to LIHTC properties. The classifier struggles the most with the "Unhealthy" class, showing an AUC of 0.50, indicating it cannot distinguish this class from others.

Table 8: Validation Accuracies for Stacking Classifier for PM2.5 and Ozone in LIHTC and Public Housing Properties

	LIHTC	Public Housing

PM2.5	Ozone	PM2.5	Ozone
86.61%	82.66%	85.89%	71.61%

5.5.Spatial Analysis:

Figure 25: Spatial Analysis of PM2.5 Exposure in United States

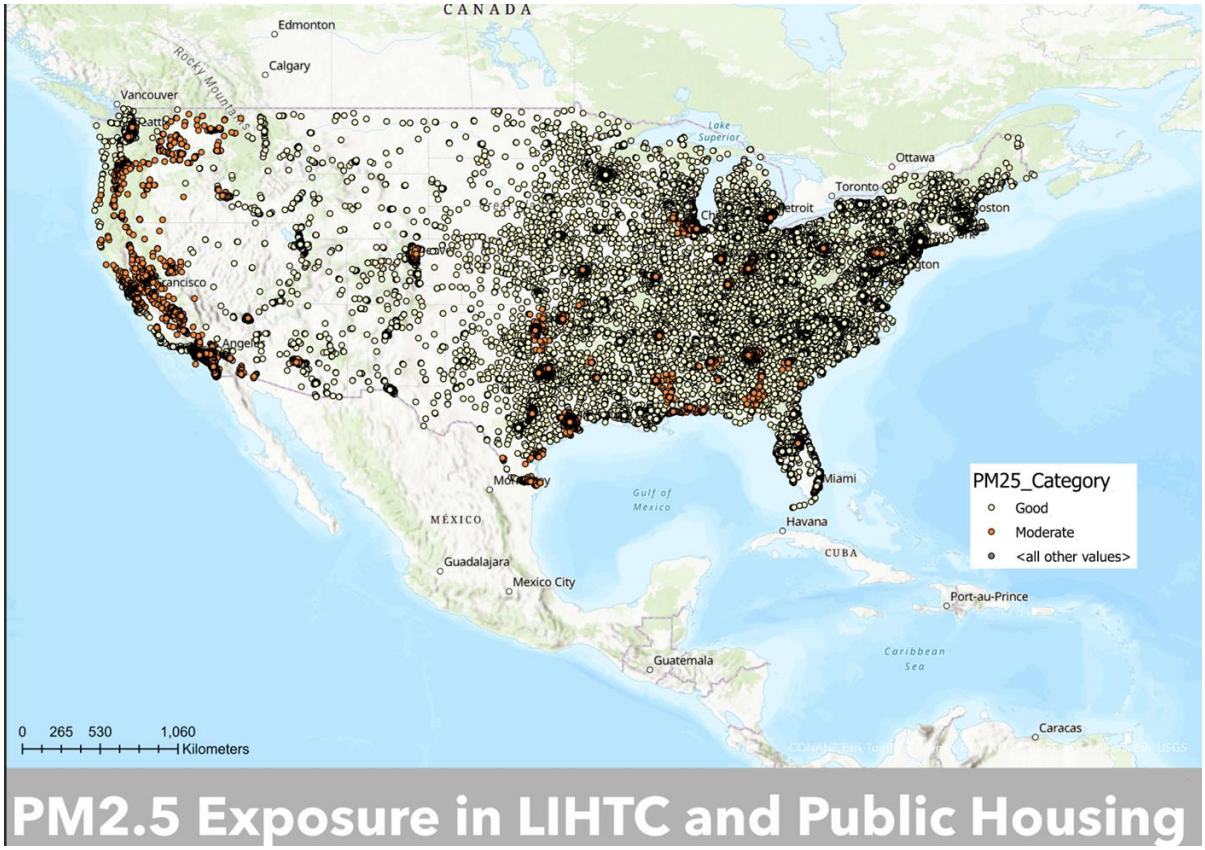


Figure 26: Spatial Analysis of Ozone Exposure in United States

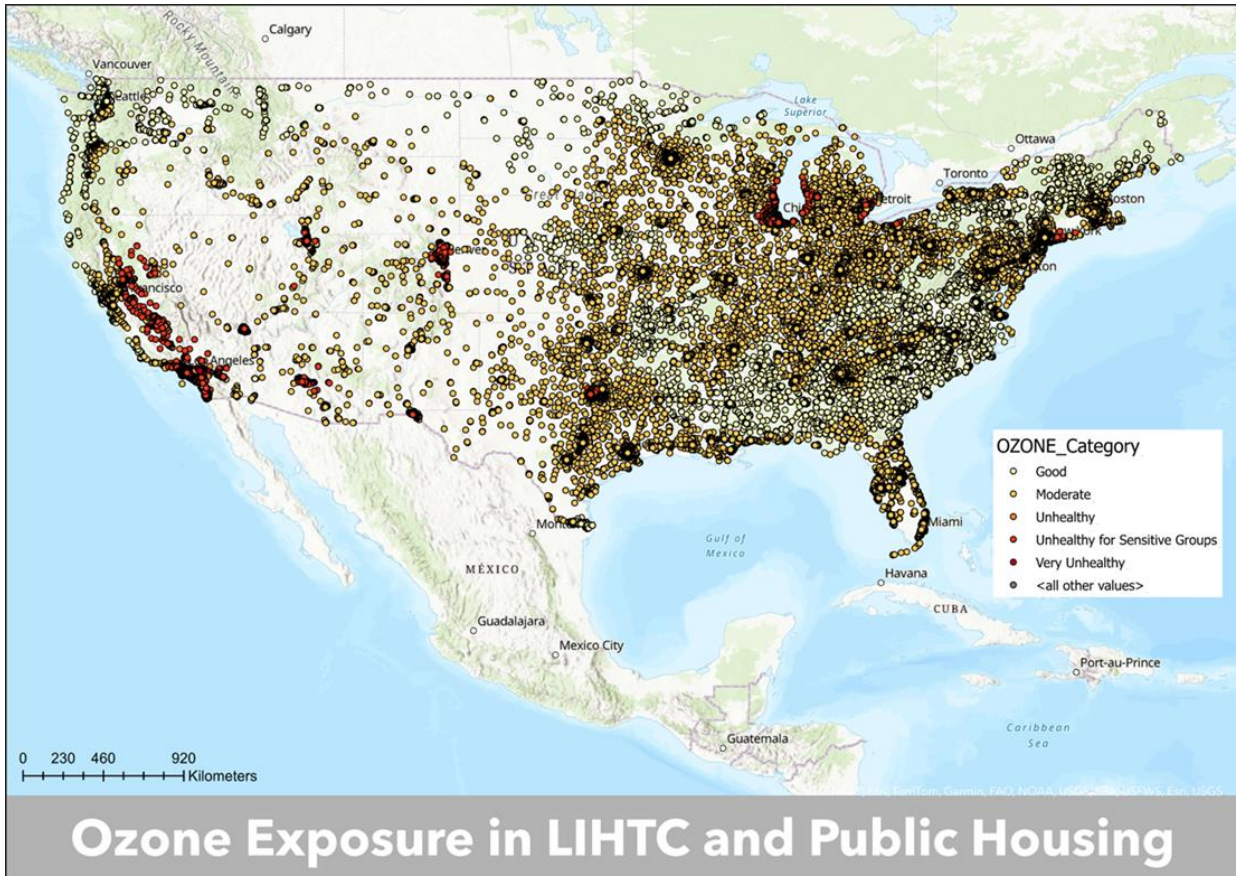


Figure 27: Air Pollution Exposure in Chicago

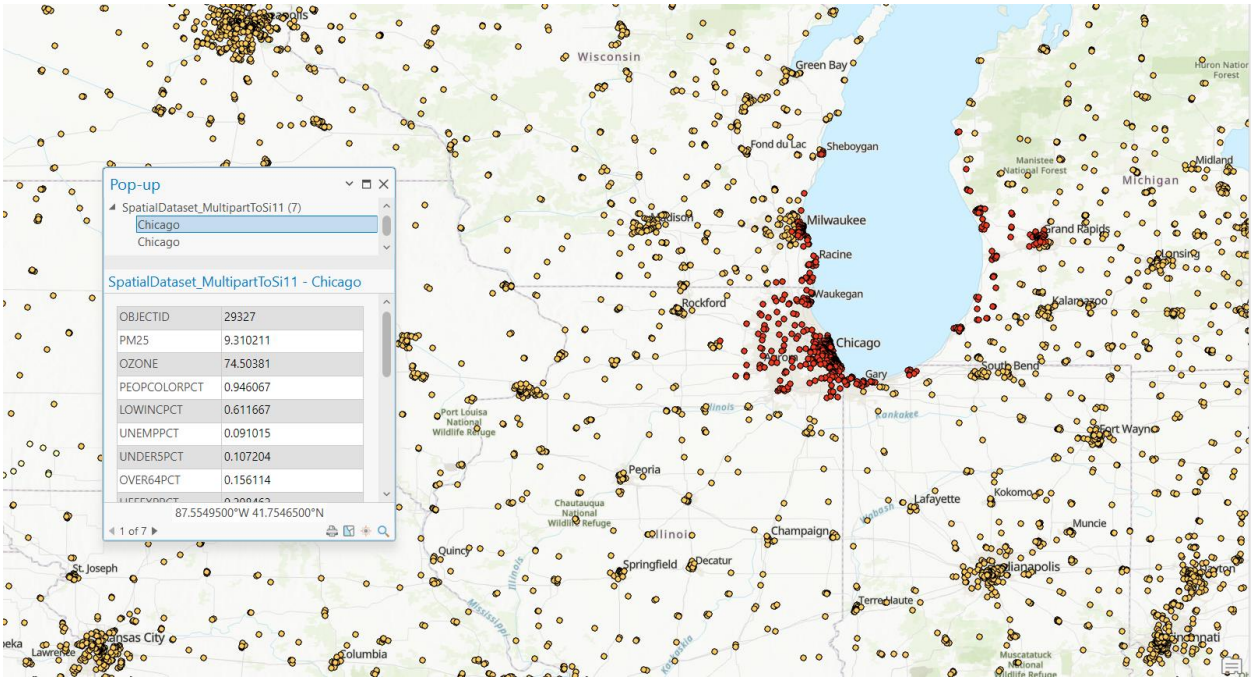
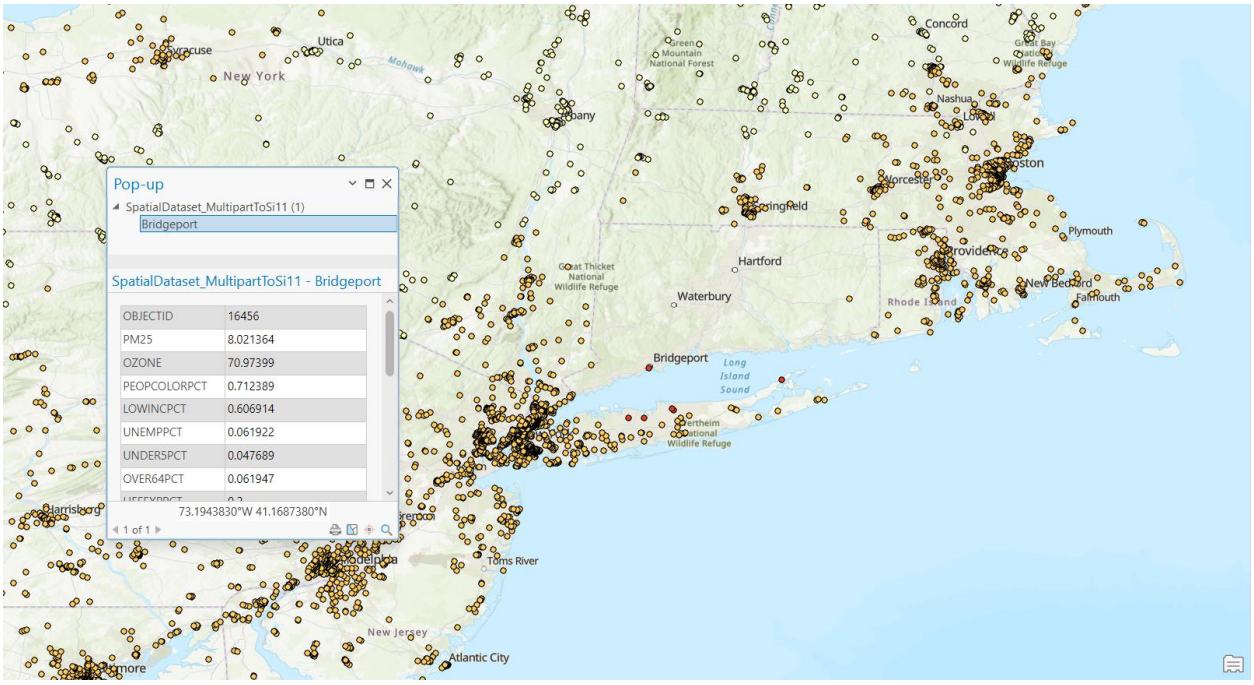


Figure 28: Air Pollution Exposure in Long Island, New York



6. Discussion:

This study found that federally assisted affordable housing properties, particularly LIHTC and Public Housing, are often located in areas with elevated PM2.5 and Ozone levels, exceeding EPA thresholds. These findings highlight potential gaps in housing policies aimed at mitigating environmental risks. Total LIHTC and public housing buildings are 30321 and 5601 respectively. The average PM2.5 exposure is 8.46 and 7.98 for LIHTC and public housing properties (Table 1). The average Ozone exposure is 60.84 and 59.14 for LIHTC and public housing properties (Table 1). The average percentage of people of color, low-income residents, unemployed residents, underage of 5 children, over the age of 64 residents, life expectancy are 51%, 41%, 7%, 6%, 16% and 21% respectively for LIHTC properties (Table 1). The average percentage of people of color,

low-income residents, unemployed residents, underage of 5 children, over the age of 64 residents, life expectancy are 50%, 47%, 8%, 6%, 17% and 23% respectively for public housing properties (Table 1). The Socio economic and demographic average percentages are a little bit but not much different in both LIHTC and Public Housing properties.

The average exposure of air pollution levels PM2.5 and Ozone are significantly higher in LIHTC properties as compared to public housing as p-values is <0.001(Table 2). These findings align with the results of the study by Goplerud et al. (2022), which observed that, within metropolitan areas, LIHTC properties are more likely to be located in tracts with poorer air quality. While our analysis quantifies these differences in exposure between LIHTC and Public Housing, the Goplerud et al. (2022) study highlights the broader implications, suggesting that even small differences in air quality between LIHTC and non-LIHTC neighborhoods could have significant health impacts when applied across millions of LIHTC residents nationwide. These findings emphasize the critical need for targeted interventions to mitigate the disproportionate environmental burden faced by residents of LIHTC properties.

The analysis shows positive correlations between pollution levels and socioeconomic indicators. Specifically, the percentage of people of color (PEOPCOLORPCT) is positively correlated with PM2.5 levels, with correlation coefficients of 0.32 for LIHTC properties and 0.32 for public housing properties. Similarly, Ozone levels also exhibited positive correlations with % people of color, with coefficients of 0.30 for LIHTC properties and 0.28 for public housing properties (Figure 1 & 2). Linear Regression is used as a baseline regression model. Random Forest regression outperformed linear regression in explaining variance (higher R-squared) and reducing prediction error (lower MSE) for both PM2.5 and Ozone, suggesting the non-linear relationships between pollution levels and socioeconomic variables (Table 4). SHAP Analysis highlights percentage of people of color is the most influential factors across both LIHTC and Public

Housing properties for predicting PM2.5 and Ozone levels and % life expectancy is the second influential factor in prediction air pollution levels except for ozone in public housing where low income is the second influential factor. (Figure 9,10,11 & 12).

Feature importance values for % people of color were 0.26 for PM2.5 and 0.23 for Ozone in LIHTC properties, and 0.26 for PM2.5 and 0.22 for Ozone in Public Housing (Table 5). Percentage of people of color is the most significant feature for predicting both PM2.5 and Ozone levels across LIHTC and Public Housing properties. Socioeconomic factor % low-income values are 0.15-0.17 for both pollutants and property types (Table 5). It shows that economic disadvantages are another key factor in pollution exposure. The importance of socioeconomic factors like % low-income and %Unemployment indicates the intersection of economic disadvantage and environmental risk (Table 5). The moderate importance of age-related variables (%Underage 5 and %Over 64 of age) suggests a need to prioritize pollution mitigation efforts in areas with higher populations of vulnerable age groups (Table 5). The feature importance aligns with previous correlation and SHAP analysis findings, supporting the robustness of the model's predictions and the relationships identified. These findings align with broader research on an environmental justice analysis of air pollution, indicating that marginalized communities, such as those with higher percentages of people of color, are disproportionately exposed to poor air quality (Nunez et al., 2024).

For the distribution of PM2.5 air pollution levels, the majority of properties in both LIHTC and Public Housing categories were located in areas with "Good" PM2.5 levels. Specifically, 22,633 LIHTC properties were categorized as having "Good" PM2.5 levels, while 7,688 were in the "Moderate" category. Similarly, 4,699 Public Housing properties were classified as "Good," with only 902 falling into the "Moderate" category. These findings suggest that both housing types generally experience lower exposure to elevated PM2.5 levels (Figure 13). For the distribution of

Ozone air pollution levels, the majority of LIHTC and public housing properties were located in areas with "Moderate" ozone levels. For LIHTC properties, 20,155 were categorized as "Moderate," while only 6,735 had "Good" ozone levels. Higher-risk categories such as "Unhealthy" had very few properties, with counts of 550 or less. Similarly, for Public Housing properties, 4,024 were classified as "Moderate," and only 1,285 were in the "Good" category. Extreme ozone levels, including "Unhealthy" and "Very Unhealthy," were even smaller, with negligible counts across both property types (Figure 14).

The machine learning analysis aimed to classify pollution levels (PM2.5 and Ozone) across LIHTC and Public Housing properties using multiple classifiers. SMOTE over sampling technique is used to balance the classes. Logistic regression is used as a baseline model. Random Forest and Extra Trees perform well but Stacking Classifier models achieved the highest accuracy for PM2.5 and Ozone classification. However, Ozone classification showed relatively lower accuracy compared to PM2.5, indicating challenges in predicting Ozone exposure due to its variability. The detailed results are discussed below:

For PM2.5 classification, the Stacking Classifier consistently outperformed other models across both LIHTC and Public Housing properties. It achieved the highest test accuracy of 85.89% for LIHTC and 86.46% for Public Housing properties, demonstrating its ability to integrate the strengths of multiple base models (Random Forest, and Extra Trees) (Figure 15 & 16). The validation accuracies of stacking classifier are 86.61% and 85.89% for LIHTC and Public Housing respectively (Table 8). The stacking classifier is performing consistently well for both classes ("Good" and "Moderate") in distinguishing between different PM2.5 pollution levels, as indicated by the high AUC (Area under the curve) score (0.89) (Figure 18). The stacking classifier is performing consistently moderately for both classes ("Good" and "Moderate") in distinguishing between different PM2.5 pollution levels, as indicated by the high AUC score

(0.76). The stacking classifier for Public Housing performs less effectively than for LIHTC properties, as shown by the lower AUC values. (Figure 19)

For Ozone classification, the Stacking Classifier also showed promising performance but faced challenges due to class imbalance, particularly in predicting "Unhealthy" and "Very Unhealthy" categories. The test accuracy is 82.65% for LIHTC and 70.75% for Public Housing properties (Figure 20 & 21). The validation accuracies of stacking classifier are 82.66% and 71.61% for LIHTC and Public Housing respectively (Table 8). The classifier performs very well for the majority of classes, especially "Good" (AUC = 0.91) and "Moderate" (AUC = 0.89) for LIHTC properties. Performance declines significantly for the "Very Unhealthy" category, due to limited data for this class or class imbalance in the dataset (Figure 23). Overall performance for Public Housing is weaker compared to LIHTC properties as for "Good" AUC = 0.70 and for "Moderate" AUC is 0.68. The classifier struggles the most with the "Unhealthy" class, showing an AUC of 0.50, indicating it cannot distinguish this class from others (Figure 24).

Spatial analysis revealed significant geographic disparities in air pollution exposure for federally assisted housing. In Chicago, neighborhoods with high percentages of people of color and low-income residents, showed elevated PM2.5 levels of 9.3 and Ozone levels of 74.5, indicating disproportionate exposure for vulnerable communities (Figure 27). Similarly, in Long Island, hotspots were identified with ozone levels exceeding thresholds, emphasizing localized pollution risks (Figure 28). These findings highlight the importance of targeted interventions to address air pollution in specific high-risk regions and provide actionable insights for policymakers to focus on areas most affected by environmental inequities.

7. Conclusion:

In conclusion, tree-based ensemble models, particularly the Stacking Classifier, outperformed other approaches in classifying PM2.5 and Ozone levels for LIHTC and Public Housing properties, effectively capturing complex, non-linear relationships in the data. Regression analysis highlighted significant associations between air pollution levels and socioeconomic factors, such as the percentage of people of color, shedding light on systemic environmental inequities. Statistical tests confirmed substantial disparities in pollution exposure between LIHTC and Public Housing properties, providing compelling evidence of differential environmental impacts based on housing type.

Spatial analysis further revealed regions with disproportionately high pollution exposure affecting affordable housing residents, emphasizing the need for geographically targeted interventions. Feature importance analysis consistently identified socioeconomic factors as key predictors of pollution exposure, reinforcing the urgent need to address structural inequities in housing and environmental policies. While the models achieved good F1-scores for major pollution categories, performance for minority classes, particularly in ozone classification, indicated the need for enhanced features and improved data collection.

These findings suggest that current housing policies must be reevaluated to integrate environmental equity considerations, prioritizing efforts to mitigate pollution exposure in underserved communities. Policymakers should focus on addressing the systemic disparities highlighted in this study to ensure healthier living environments for vulnerable populations.

8. Future Work:

For future work, several enhancements can be explored to enrich the analysis and expand its impact. Incorporating additional air pollution variables, such as nitrogen dioxide (NO_2) and sulfur dioxide (SO_2), could provide a more holistic view of environmental risks faced by affordable housing residents. Leveraging the time-series would enable tracking trends in pollution exposure over time, offering insights into the long-term effects of housing policies and environmental regulations, including the Clean Air Act. Refining ozone exposure modeling through additional features and more robust datasets could address its unique complexities and improve prediction accuracy. Further, evaluating the effectiveness of policies, such as the Clean Air Act, in reducing pollution levels near federally assisted housing would provide valuable insights to guide future policy adjustments. Engaging with local communities and stakeholders to validate findings and integrate lived experiences could ensure that the research outcomes are actionable, equitable, and responsive to the needs of impacted populations.

9. References

1. Air pollution and your health. (n.d.). National Institute of Environmental Health Sciences. Retrieved November 24, 2024, from https://www.niehs.nih.gov/health/topics/agents/air-pollution.
2. Anyanwu, C., & Beyer, K. M. M. (2024). Intersections among housing, environmental conditions, and health equity: A conceptual model for environmental justice policy. Social Sciences & Humanities Open, 9, 100845. https://doi.org/10.1016/j.ssaho.2024.100845

3. Goplerud, D.K., Gensheimer, S.G., Schneider, B.K., Eisenberg, M.D., Smith, G.S. The Spatial Relationship Between the Low-Income Housing Tax Credit Program and Industrial Air Pollution. Cityscape: A Journal of Policy Development and Research, 24(3), 2022. U.S. Department of Housing and Urban Development, Office of Policy Development and Research. https://www.huduser.gov/portal/periodicals/cityscape/vol24num3/ch7.pdf

   Nunez, Y., Benavides, J., Shearston, J.A. et al. An environmental justice analysis of air pollution emissions in the United States from 1970 to 2010. Nat Commun 15, 268 (2024). https://doi.org/10.1038/s41467-023-43492-9

4. Thakrar, S. K., Balasubramanian, S., Adams, P. J., Azevedo, I. M. L., Muller, N. Z., Pandis, S. N., Polasky, S., Pope, C. A. III, Robinson, A. L., Apte, J. S., Tessum, C. W., Marshall, J. D., & Hill, J. D. "Reducing Mortality from Air Pollution in the United States by Targeting Specific Emission Sources." Environmental Science & Technology Letters, 7(9), 639-645, 2020. DOI: 10.1021/acs.estlett.0c00424

5. U.S. Department of Housing and Urban Development (HUD), Office of Policy Development and Research. Low-Income Housing Tax Credit (LIHTC) Database.

Accessed November 18, 2024. https://www.huduser.gov/portal/datasets/lihtc.html

6. U.S. Department of Housing and Urban Development (HUD). Public Housing Program. Accessed November 18, 2024. https://www.hud.gov/topics/rental_assistance/phprog

7. U.S. Environmental Protection Agency (EPA). "Air Pollution: Current and Future Challenges." Accessed November 18, 2024. https://www.epa.gov/clean-air-act-overview/air-pollution-current-and-future-challenges