Long-Term $PM_{2.5}$ Exposure and Chronic Disease Prevalence in the United States

Final Report

Course: DSA210 – Introduction to Data Science

Student Name: Ahmet Vedat Kurt

Student ID: 34454

Date: 09/01/2026

1 Motivation and Problem Statement

Air pollution is widely recognized as a major environmental determinant of public health, contributing to millions of premature deaths worldwide each year. Among air pollutants, fine particulate matter ($PM_{2.5}$) poses a particular threat due to its small size, which allows it to penetrate deep into the respiratory system and enter the bloodstream. While $PM_{2.5}$ has been extensively linked to cardiovascular and respiratory outcomes, its relationship with chronic metabolic diseases such as diabetes and hypertension remains less clearly quantified at fine geographic scales.

Diabetes and hypertension are among the most prevalent chronic conditions in the United States, imposing significant long-term burdens on individuals, healthcare systems, and public policy. Understanding whether environmental exposure contributes to the prevalence of these diseases—beyond behavioral and lifestyle risk factors—has important implications for prevention strategies.

This project investigates whether long-term exposure to $PM_{2.5}$ is associated with higher county-level prevalence of diabetes and hypertension across the United States. Rather than relying on coarse national or state averages, the analysis integrates high-resolution environmental data with public health indicators and applies population-weighted aggregation to ensure that exposure estimates accurately reflect where people live.

2 Data Sources and Integration Strategy

Health outcome and behavioral risk data were obtained from the *CDC PLACES* project, which provides age-adjusted prevalence estimates for chronic conditions such as diabetes and hypertension, as well as behavioral indicators including obesity, smoking, and physical inactivity [1]. In this study, PLACES variables were utilized at the census-tract level to enable fine-grained spatial alignment with environmental exposure data.

Environmental exposure data were sourced from *EPA Air Quality System (AQS)*-derived $PM_{2.5}$ estimates, also available at the census-tract level [2, 3]. A census tract is a small, relatively permanent geographic unit defined by the U.S. Census Bureau, typically containing between 2,500 and 8,000 residents. Census tracts are designed to be internally homogeneous with respect to population characteristics and the built environment, making them well suited for capturing localized environmental exposure and health patterns.

Using both health and environmental variables at the tract level ensures spatial consistency and reduces ecological bias prior to aggregation. To support population-level

analysis and align with the scale of public health decision-making, tract-level variables were subsequently aggregated to the county level. Counties were further classified using the *National Center for Health Statistics (NCHS) Urban–Rural Classification Scheme* [4] to enable stratified analysis across urbanization contexts.

Population-weighted aggregation was employed to derive county-level estimates:

$$\text{County Value} = \frac{\sum(\text{Tract Value} \times \text{Tract Population})}{\sum \text{Tract Population}}$$

This approach ensures that county-level exposure and health measures reflect the average experience of residents rather than an unweighted average of tracts.

3 Data Preparation and Quality Control

Data preparation followed a structured *tract-to-county* pipeline to ensure accuracy, spatial consistency, and population representativeness. The key steps are summarized below:

- **Tract-level processing:** Daily $PM_{2.5}$ concentration estimates were averaged by year for each census tract and subsequently combined across multiple years to compute a long-term exposure metric. PLACES health and behavioral variables were retained at the tract level during this stage.
- **Spatial merging:** Tract-level $PM_{2.5}$ and PLACES datasets were merged using the 11-digit census tract FIPS code to ensure precise geographic alignment.
- **Population-weighted aggregation:** Tract-level variables were aggregated to the county level using population-weighted averaging, ensuring that densely populated tracts contributed proportionally more to county estimates.
- **Urbanization mapping:** County-level records were assigned urbanization categories using the NCHS Urban–Rural Classification Scheme, joined via the 5-digit county FIPS code.
- **Quality control:** Records with missing or inconsistent geographic identifiers, population values, or key variables were removed. Continuous variables were examined for scale and distribution, and categorical variables were encoded appropriately.

These steps resulted in a clean, analysis-ready county-level dataset integrating long-term environmental exposure, behavioral risk factors, and chronic disease prevalence across the United States.

4 Research Hypotheses

This report evaluates three hypotheses about how long-term $PM_{2.5}$ exposure relates to county-level prevalence of diabetes and hypertension, and how these relationships behave under different modeling choices and geographic contexts:

- **Null hypothesis (H_0):** $PM_{2.5}$ exposure has no statistically significant relationship with diabetes or hypertension prevalence.
- **Main hypothesis (H_1):** Counties with higher long-term $PM_{2.5}$ exposure have higher diabetes and hypertension prevalence even after controlling for key behavioral and socioeconomic covariates. This is evaluated using correlation tests and multivariable regression/regularization analyses.
- **Secondary hypothesis (H_2):** Nonlinear ensemble models (e.g., Random Forest and Gradient Boosting/XGBoost) outperform simpler baselines (Linear Regression and KNN), implying non-linear and interacting risk relationships. This is evaluated using model performance comparison.
- **Third hypothesis (H_3):** The $PM_{2.5}$–disease relationship varies by urbanization level, with stronger association expected in more urban counties. This is evaluated using stratified modeling and interaction-based approaches.

5 Exploratory Analysis

5.1 Univariate Distribution of Key Variables

Univariate analysis was conducted to assess the distributional properties of environmental exposure, health outcomes, and behavioral risk factors prior to correlation analysis and modeling. Examining marginal distributions is essential for identifying skewness, outliers, and heterogeneity, which may influence statistical assumptions and model choice.

Figure 1 shows the distributions of all continuous variables, while Figure 2 summarizes their central tendency and dispersion.

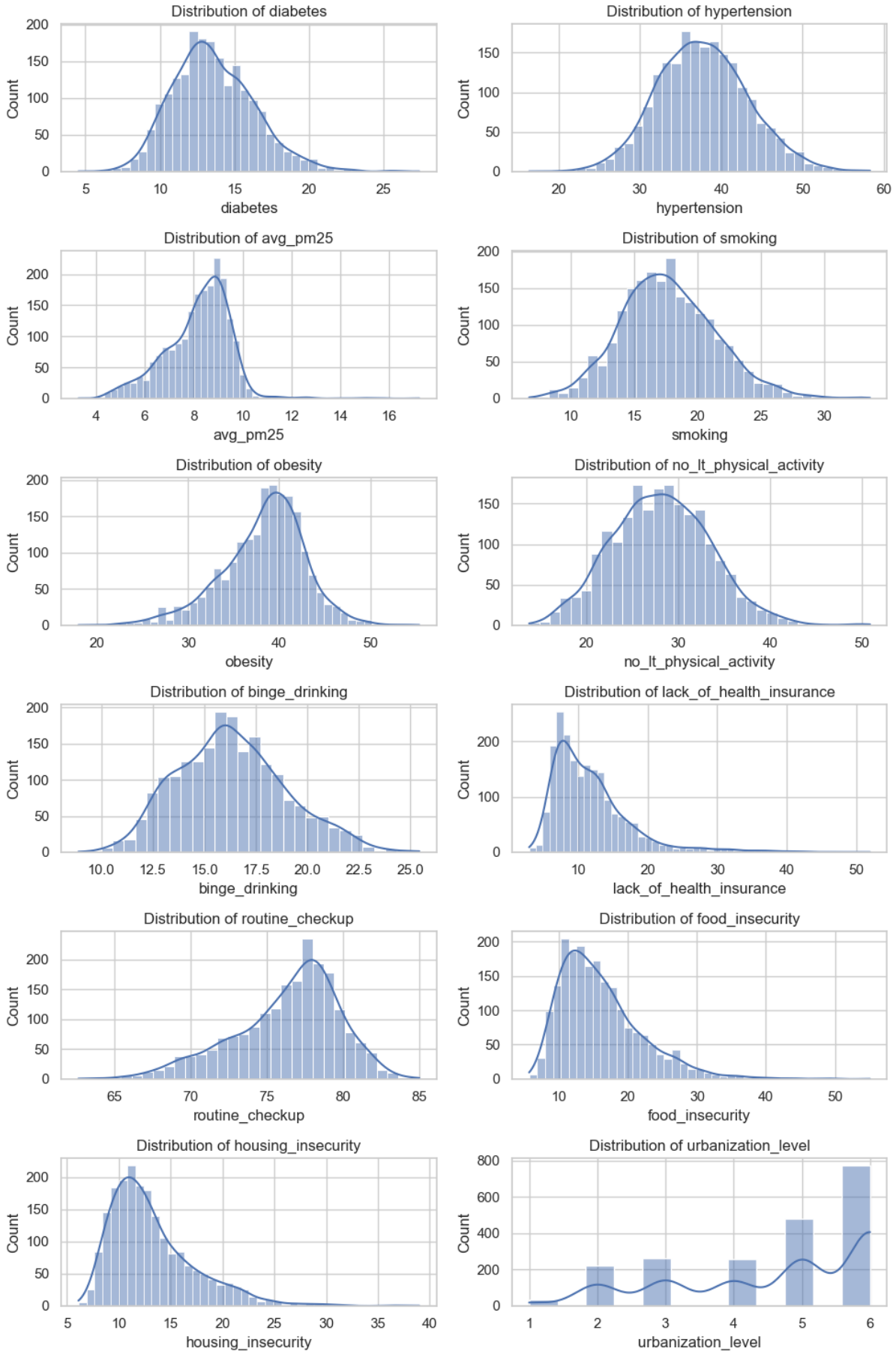

Figure 1: Univariate distributions of environmental, health, and behavioral variables across U.S. counties.

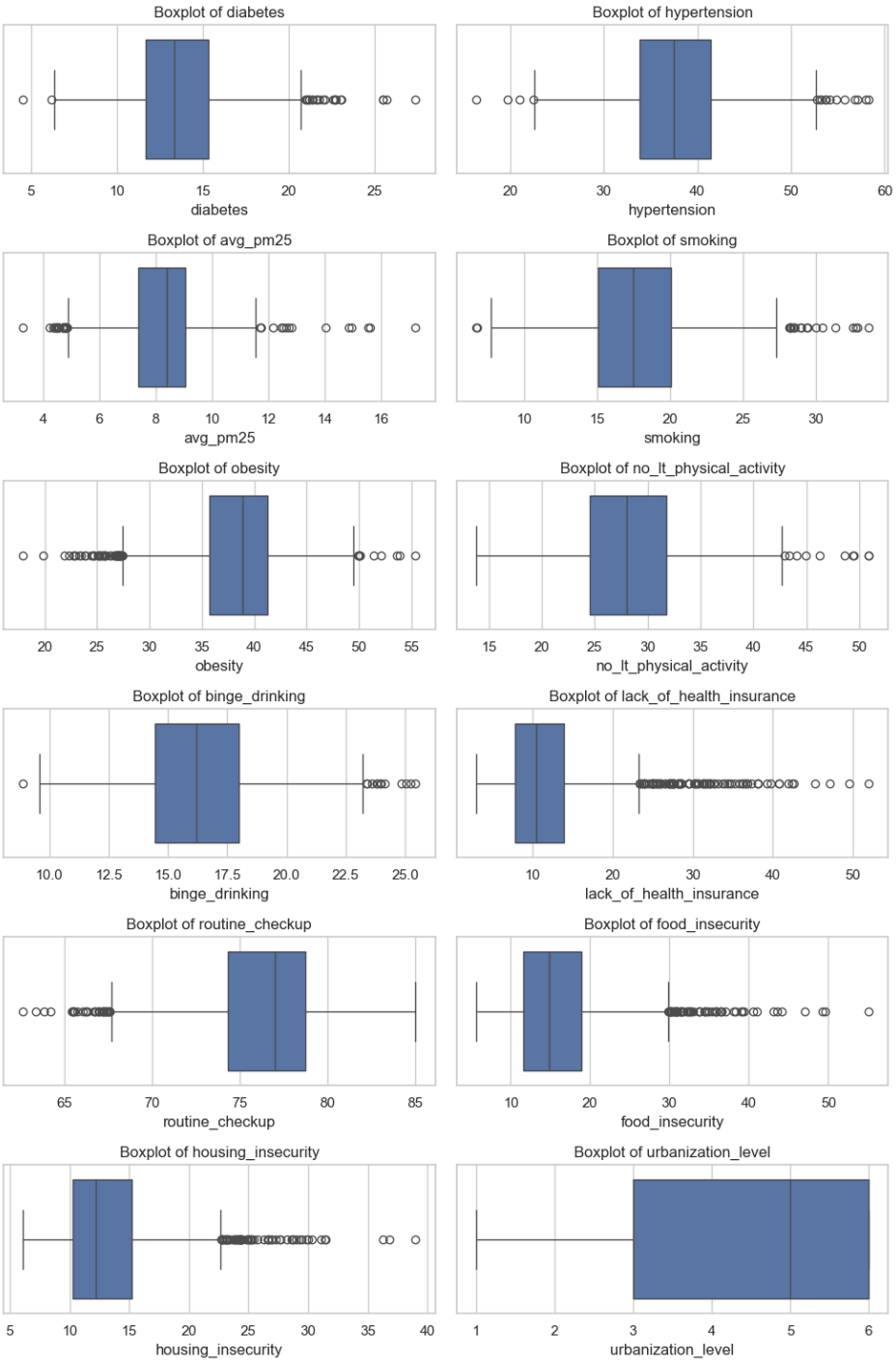

Figure 2: Boxplots of key variables illustrating medians, interquartile ranges, and outliers.

Long-term $PM_{2.5}$ exposure exhibits moderate right skewness, indicating the presence of high-exposure counties. Diabetes and hypertension prevalence show substantial inter-county variation, with hypertension displaying a broader distribution. Behavioral and socioeconomic risk factors demonstrate pronounced dispersion and right-tailed distributions, particularly for obesity, physical inactivity, and measures of insecurity. Routine checkup rates are left-skewed, while urbanization levels show uneven representation across categories.

Overall, the presence of skewed distributions and outliers motivates the use of robust correlation measures and nonlinear predictive models in subsequent analyses.

5.2 $PM_{2.5}$ Exposure and Chronic Disease Relationships

To assess the association between long-term $PM_{2.5}$ exposure and chronic disease prevalence, bivariate relationships were examined using scatter plots with fitted regression lines. These plots provide an initial visual assessment of direction, strength, and dispersion of the association without imposing strict modeling assumptions.

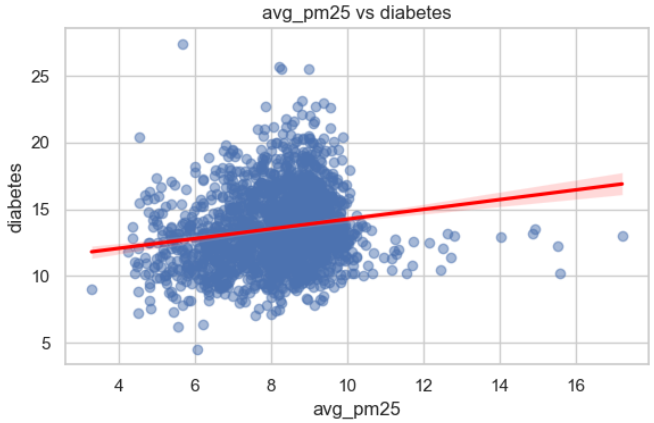

Figure 3: Relationship between $PM_{2.5}$ exposure and diabetes prevalence.

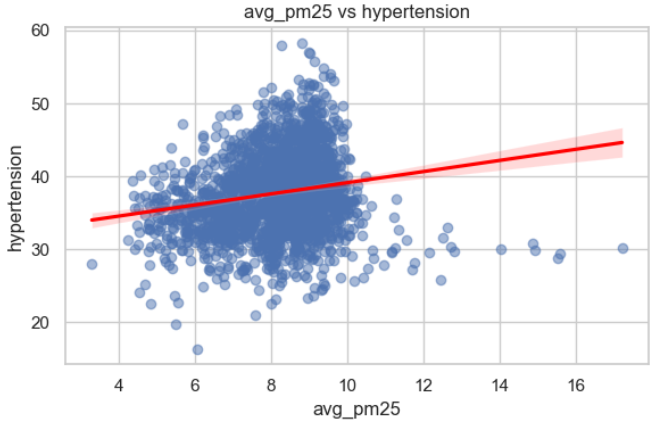

Figure 4: Relationship between $PM_{2.5}$ exposure and hypertension prevalence.

Both scatter plots reveal a positive but highly dispersed relationship between $PM_{2.5}$ exposure and disease prevalence. The upward slope of the fitted lines indicates that counties with higher long-term $PM_{2.5}$ exposure tend, on average, to exhibit higher prevalence of diabetes and hypertension. However, the substantial spread of observations suggests that $PM_{2.5}$ alone explains only a limited portion of the variation in chronic disease prevalence, motivating multivariate and nonlinear analyses.

5.3 Quantile-Based Comparison Across $PM_{2.5}$ Levels

To complement the continuous scatter plots and reduce sensitivity to outliers, counties were grouped into quintiles based on long-term $PM_{2.5}$ exposure. This approach facilitates comparison of disease prevalence across exposure strata and highlights population-level shifts rather than pointwise variation.

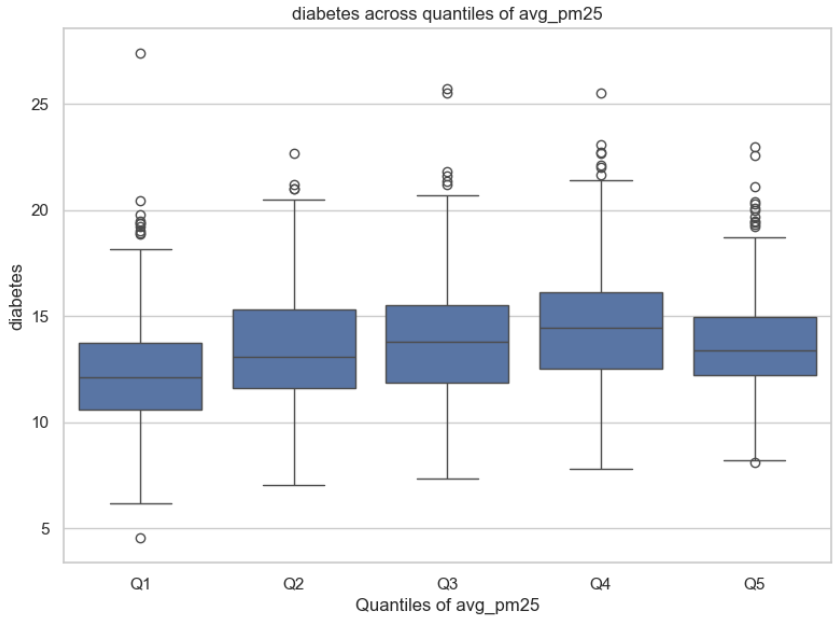

Figure 5: Diabetes prevalence across quintiles of $PM_{2.5}$ exposure.

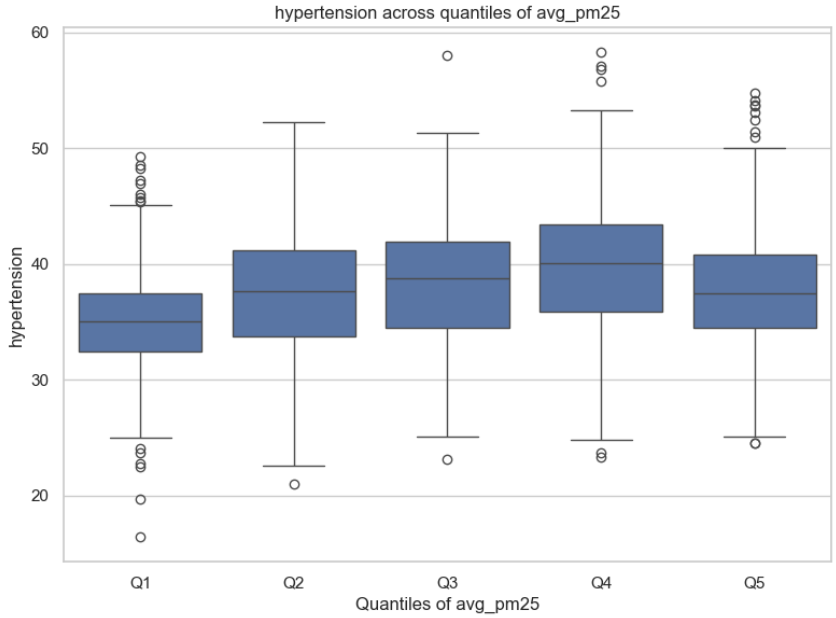

Figure 6: Hypertension prevalence across quintiles of $PM_{2.5}$ exposure.

The quantile-based boxplots show a gradual increase in median diabetes and hypertension prevalence from lower to higher $PM_{2.5}$ exposure groups, despite overlapping

distributions. This pattern provides additional evidence of a positive association at the population level and supports subsequent statistical testing by demonstrating that higher exposure is associated with a systematic upward shift in disease prevalence rather than being driven solely by extreme values.

5.4 Correlation Structure

To evaluate linear associations among environmental, health, behavioral, and socioeconomic variables, a Pearson correlation matrix was constructed (Figure 7). $PM_{2.5}$ exposure shows a positive but relatively weak correlation with both diabetes and hypertension, indicating that higher pollution levels are associated with higher disease prevalence but do not act as dominant predictors. In contrast, behavioral risk factors—particularly obesity, physical inactivity, and smoking—exhibit substantially stronger correlations with chronic disease outcomes. Socioeconomic indicators such as food and housing insecurity are also strongly correlated with disease prevalence, highlighting the multifactorial nature of chronic health conditions. These patterns suggest that while $PM_{2.5}$ contributes to disease risk, its effect is likely mediated or confounded by behavioral and socioeconomic factors.

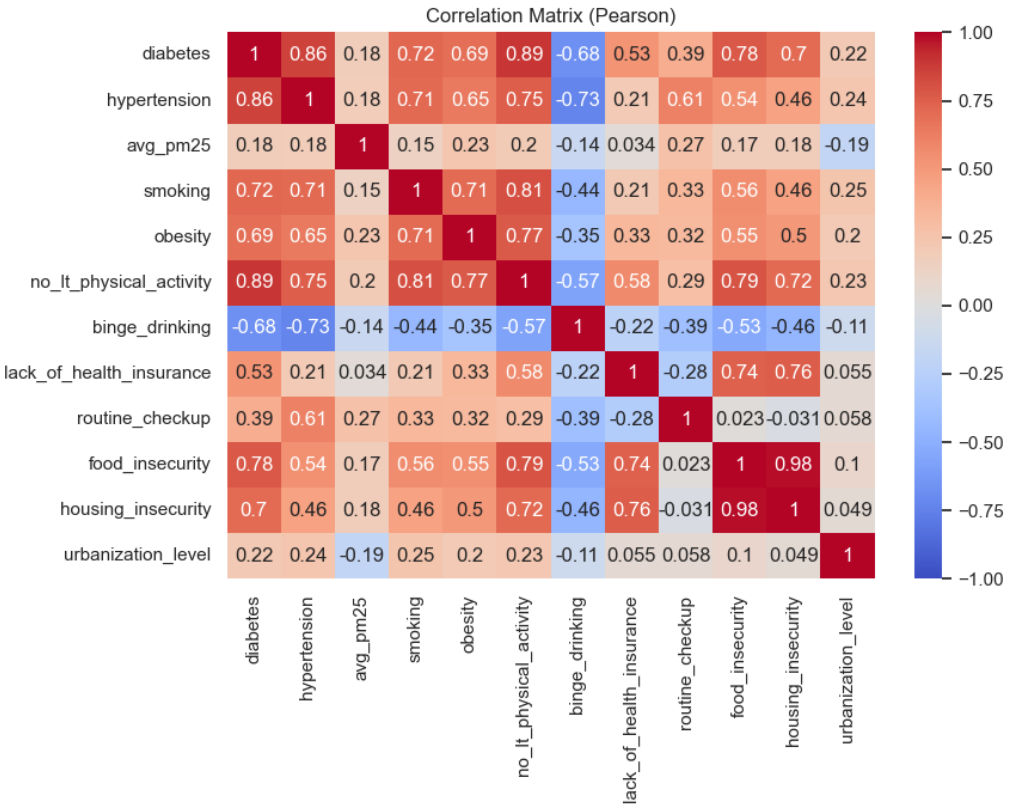

Figure 7: Pearson correlation matrix of environmental, health, behavioral, and socioeconomic variables.

5.5 Urbanization Effects

To examine whether the association between $PM_{2.5}$ exposure and chronic disease prevalence varies by structural context, fitted regression curves were compared across urban and rural counties (Figure 8). The resulting interaction plot indicates that the $PM_{2.5}$–diabetes relationship differs in slope between urban and rural settings, with rural counties exhibiting a steeper increase in diabetes prevalence as $PM_{2.5}$ levels rise. This divergence suggests that urbanization modifies the pollution–health relationship, potentially reflecting differences in population characteristics, baseline health status, or co-occurring environmental and socioeconomic factors. These findings motivate stratified analysis and interaction-aware modeling in subsequent stages.

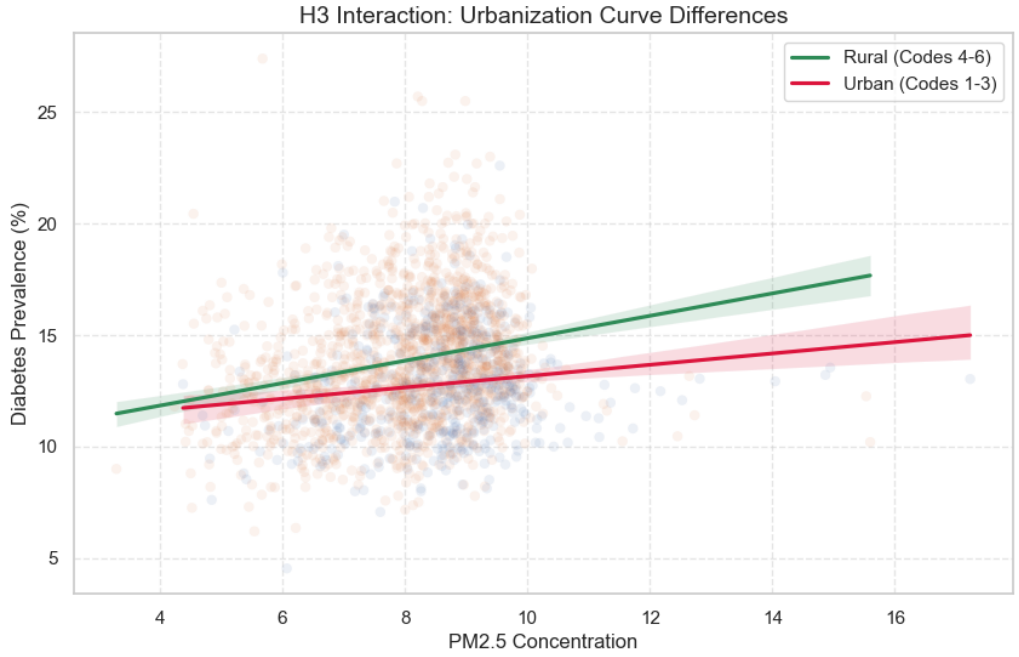

Figure 8: Interaction between $PM_{2.5}$ exposure and urbanization level in relation to diabetes prevalence.

6 Statistical Hypothesis Testing

The null hypothesis (H_0) states that long-term $PM_{2.5}$ exposure has no association with chronic disease prevalence (diabetes and hypertension) at the county level. Pearson (linear association) and Spearman (rank-based, outlier-robust) correlation analyses both indicate statistically significant positive relationships between $PM_{2.5}$ and each outcome. Therefore, H_0 is rejected in favor of the alternative hypothesis that higher $PM_{2.5}$ exposure is associated with higher chronic disease prevalence.

7 Predictive Modeling and Hypothesis Evaluation

7.1 Predictive Contribution of $PM_{2.5}$ (Hypothesis H_1)

Hypothesis H_1 states that long-term $PM_{2.5}$ exposure contributes to chronic disease prevalence after accounting for major behavioral, socioeconomic, and structural confounders. To evaluate this hypothesis, regression analyses were conducted in a structured sequence designed to assess robustness, diagnose multicollinearity, and interpret coefficient stability.

7.1.1 Fully Adjusted OLS Model

The analysis began with a fully adjusted Ordinary Least Squares (OLS) regression that included $PM_{2.5}$ exposure alongside a comprehensive set of confounders: smoking, obesity, physical inactivity, binge drinking, lack of health insurance, food insecurity, housing insecurity, routine checkup prevalence, and urbanization level. In this specification, the estimated coefficient of $PM_{2.5}$ was negative and statistically significant for both diabetes and hypertension (diabetes: $\beta = -0.1158$, $p < 0.001$; hypertension: $\beta = -0.2403$, $p < 0.001$).

The negative sign contrasts with the positive bivariate association observed earlier, suggesting that conditioning on highly correlated behavioral and socioeconomic variables substantially alters the estimated marginal contribution of $PM_{2.5}$. This pattern raised concerns regarding multicollinearity and suppression effects, motivating further diagnostic analysis.

7.1.2 Multicollinearity Diagnostics

To assess whether the negative $PM_{2.5}$ coefficient resulted from multicollinearity among covariates, Variance Inflation Factors (VIF) were computed for the fully adjusted model. $PM_{2.5}$ itself exhibited low multicollinearity (VIF = 1.21), while several covariates showed severe collinearity, particularly food insecurity (VIF = 54.43), housing insecurity (VIF = 41.64), and physical inactivity (VIF = 10.56). These results indicate that multiple predictors capture overlapping socioeconomic and healthcare-access dimensions.

To reduce redundancy, highly collinear insecurity-related variables were combined into a composite measure, and the model was re-estimated. Despite this correction, the $PM_{2.5}$ coefficient remained negative and statistically significant (diabetes: $\beta = -0.1368$, $p < 0.001$; hypertension: $\beta = -0.3269$, $p < 0.001$), indicating that the sign of the association is not solely an artifact of collinearity.

7.1.3 Stepwise OLS Regression

To further understand how the $PM_{2.5}$ coefficient evolves as confounders are introduced, a stepwise OLS procedure was conducted. In simpler specifications including $PM_{2.5}$ and a limited set of behavioral variables, $PM_{2.5}$ exhibited a positive coefficient (diabetes: $\beta = 0.1989$, $p < 0.001$; hypertension: $\beta = 0.4441$, $p < 0.001$). As additional covariates—particularly obesity and physical inactivity—were introduced, the magnitude of the $PM_{2.5}$ coefficient diminished and eventually changed sign.

This progression indicates a suppression effect: variables strongly correlated with both $PM_{2.5}$ exposure and disease prevalence absorb shared variance, altering the conditional interpretation of the $PM_{2.5}$ coefficient. The stepwise results demonstrate that the negative coefficient observed in the fully adjusted model reflects conditional association rather than numerical instability.

7.1.4 Regularized Regression via LASSO

Finally, Least Absolute Shrinkage and Selection Operator (LASSO) regression was applied to evaluate the robustness of $PM_{2.5}$ under penalized estimation. With the full covariate set, LASSO retained $PM_{2.5}$ with a non-zero negative coefficient for both outcomes (diabetes: $\beta = -0.1014$; hypertension: $\beta = -0.1967$). When the most collinear healthcare-access and insecurity variables were removed, the $PM_{2.5}$ coefficient became positive (diabetes: $\beta = 0.0494$; hypertension: $\beta = 0.0801$).

These results indicate that while $PM_{2.5}$ consistently contributes predictive information, its estimated direction is sensitive to model specification in the presence of highly correlated confounders.

7.1.5 Interpretation with Respect to Hypothesis H_1

Taken together, the regression analyses support Hypothesis H_1 in a qualified sense. $PM_{2.5}$ exposure consistently retains non-zero predictive contribution across multiple modeling frameworks, but its conditional association depends on the structure of correlated behavioral and socioeconomic variables. The results suggest that $PM_{2.5}$ is embedded within a complex risk environment rather than acting as an isolated driver of chronic disease prevalence, motivating subsequent analyses using nonlinear and interaction-aware models.

7.2 Model Complexity and Nonlinearity (Hypothesis H_2)

To evaluate Hypothesis H_2, multiple predictive models with increasing representational capacity were trained and compared, including Linear Regression, K-Nearest Neighbors (KNN), Random Forest, Gradient Boosting, and XGBoost. These models span linear,

distance-based, and ensemble tree-based architectures, allowing assessment of whether nonlinear methods better capture the relationship between environmental exposure, behavioral risk factors, and chronic disease prevalence.

Model performance was evaluated using the coefficient of determination (R^2). For diabetes prediction, ensemble-based models achieved the highest performance, with XGBoost attaining the best fit ($R^2 = 0.9105$), followed closely by Random Forest ($R^2 = 0.9094$) and Gradient Boosting ($R^2 = 0.8996$). Linear Regression and KNN yielded lower performance ($R^2 = 0.8809$ and $R^2 = 0.8836$, respectively). A similar pattern was observed for hypertension, where Random Forest performed best ($R^2 = 0.8913$), followed by XGBoost ($R^2 = 0.8853$) and Gradient Boosting ($R^2 = 0.8820$), again outperforming Linear Regression ($R^2 = 0.8410$) and KNN ($R^2 = 0.8531$). These results are summarized in Figures 9 and 10.

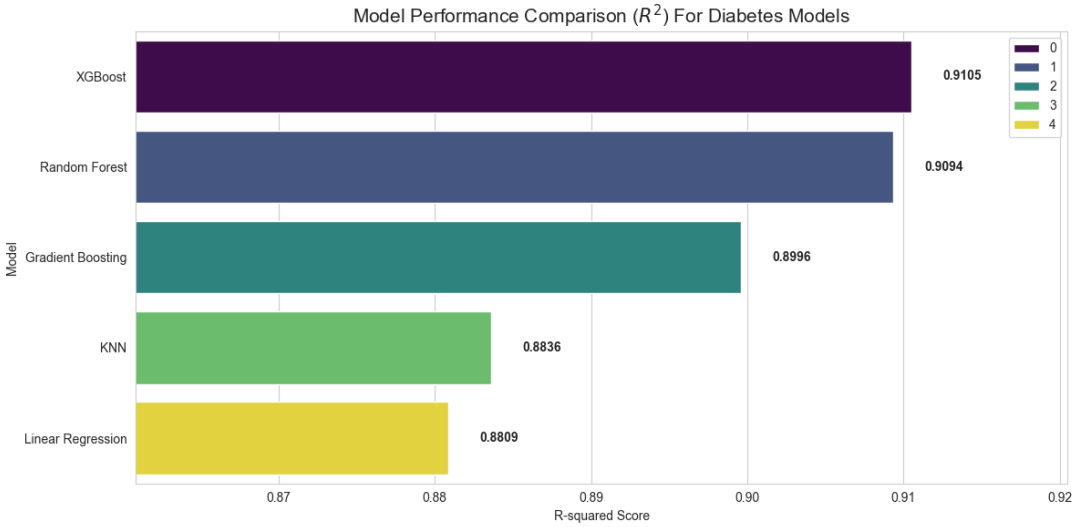

Figure 9: Model performance comparison (R^2) for diabetes prediction across Linear Regression, KNN, Gradient Boosting, Random Forest, and XGBoost.

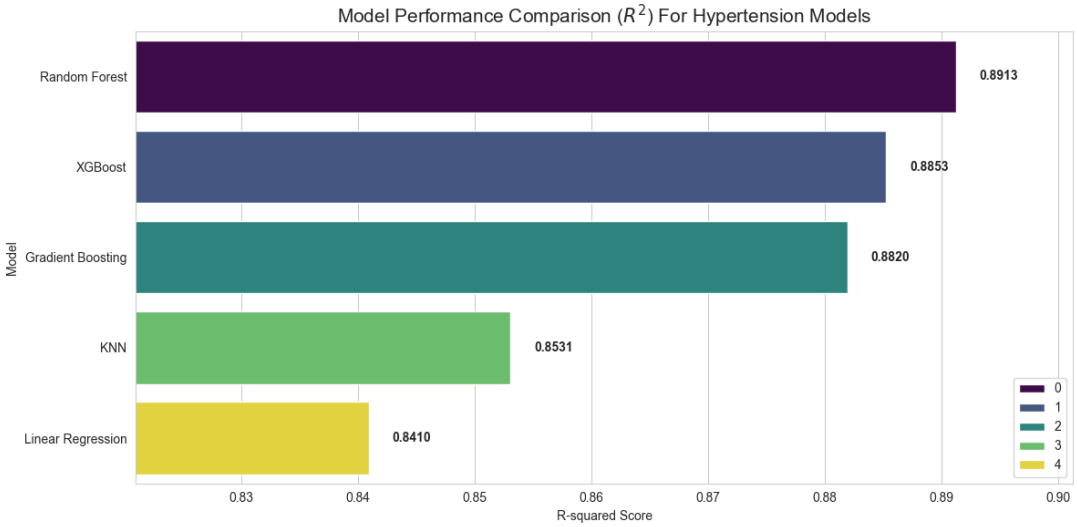

Figure 10: Model performance comparison (R^2) for hypertension prediction across Linear Regression, KNN, Gradient Boosting, Random Forest, and XGBoost.

The consistent performance advantage of ensemble tree-based models indicates that the relationships among predictors are nonlinear and involve interactions that simpler linear or distance-based models cannot fully capture. These findings support Hypothesis H_2 and motivate the use of flexible models for capturing complex pollution–health dynamics.

7.3 Feature Importance

To interpret the relative contribution of individual predictors within the best-performing nonlinear models, feature importance scores were extracted from the ensemble-based models (Figures 11 and 12). Across both diabetes and hypertension predictions, behavioral risk factors—particularly physical inactivity, binge drinking, and smoking—emerged as the dominant contributors to model performance.

Despite being less influential than behavioral variables, long-term $PM_{2.5}$ exposure consistently appeared with non-zero importance in both disease models. This indicates that $PM_{2.5}$ provides additional predictive signal beyond lifestyle and socioeconomic factors, even when complex nonlinear interactions are accounted for. Preventive care utilization and socioeconomic indicators exhibited moderate importance, while urbanization level contributed comparatively less in direct prediction once other variables were included.

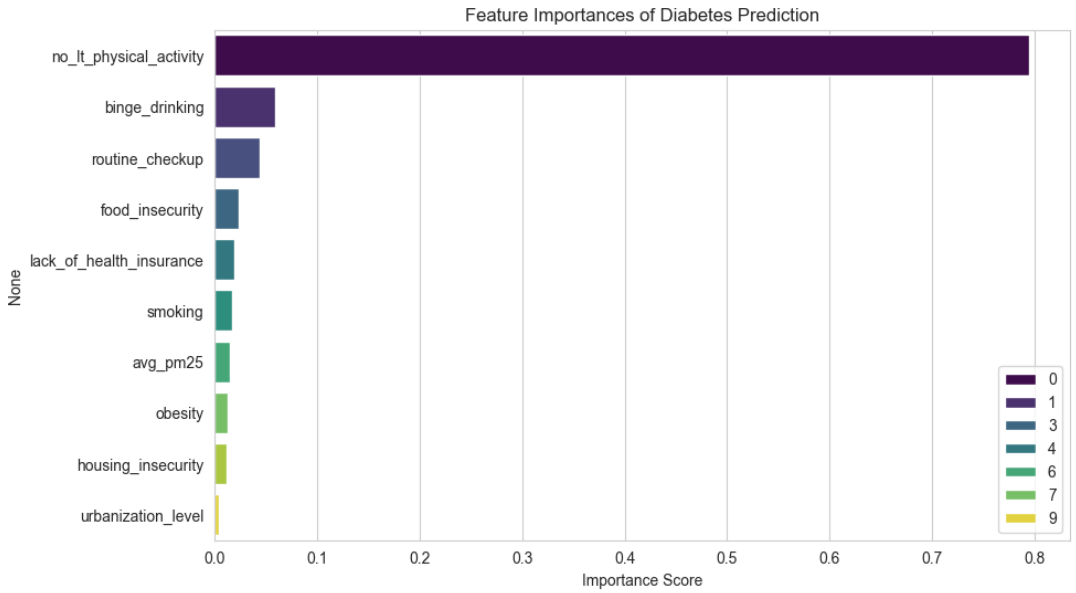

Figure 11: Feature importance scores for diabetes prediction from the best-performing ensemble model.

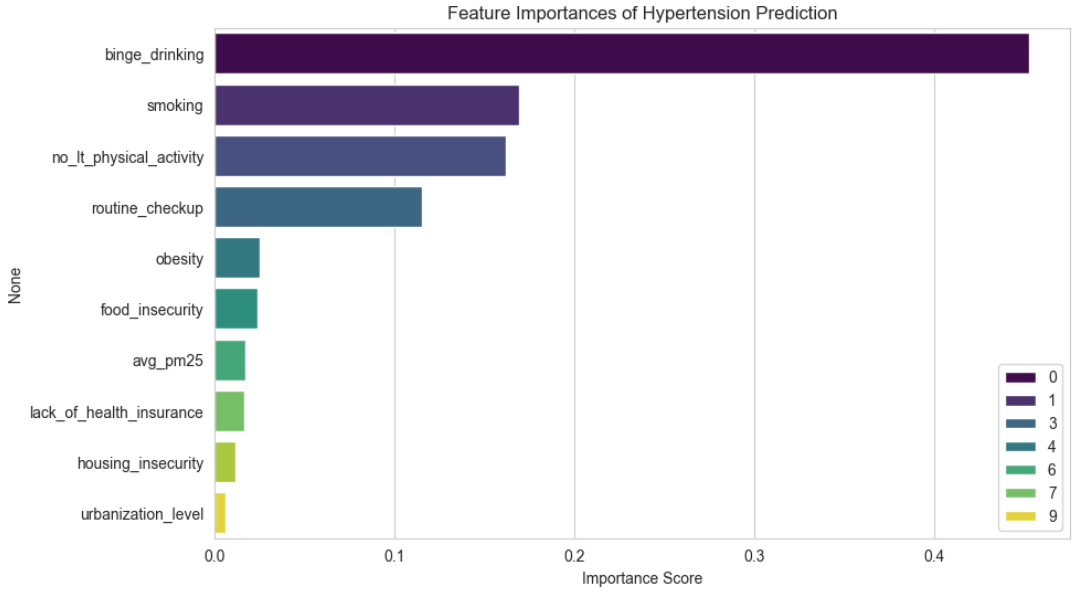

Figure 12: Feature importance scores for hypertension prediction from the best-performing ensemble model.

Overall, the feature importance analysis reinforces earlier regression findings: chronic disease prevalence is primarily driven by behavioral and socioeconomic factors, but environmental exposure remains a meaningful component of the predictive landscape. Importantly, these importance measures reflect predictive contribution rather than causal effect, and should be interpreted accordingly.

7.4 Urbanization as an Effect Modifier (Hypothesis H_3)

Hypothesis H_3 examines whether the relationship between long-term $PM_{2.5}$ exposure and chronic disease prevalence differs across urbanization contexts. Because effect modification may not be well captured by a single global linear coefficient, two complementary machine learning strategies were used: (i) stratified modeling to compare $PM_{2.5}$ importance in urban versus rural subsets, and (ii) an explicit interaction-term model to test whether a combined $PM_{2.5}$×urbanization signal provides additional predictive value.

7.4.1 Stratified Modeling (Urban vs. Rural Random Forests)

To directly compare the predictive role of $PM_{2.5}$ across geographic contexts, the dataset was stratified using NCHS urbanization codes into *urban* counties (codes 1–4) and *rural* counties (codes 5–6). Separate Random Forest regressors were trained within each subset for diabetes and hypertension, and feature importances were extracted. This approach isolates how strongly $PM_{2.5}$ contributes to prediction within each environment rather than averaging effects across heterogeneous county types.

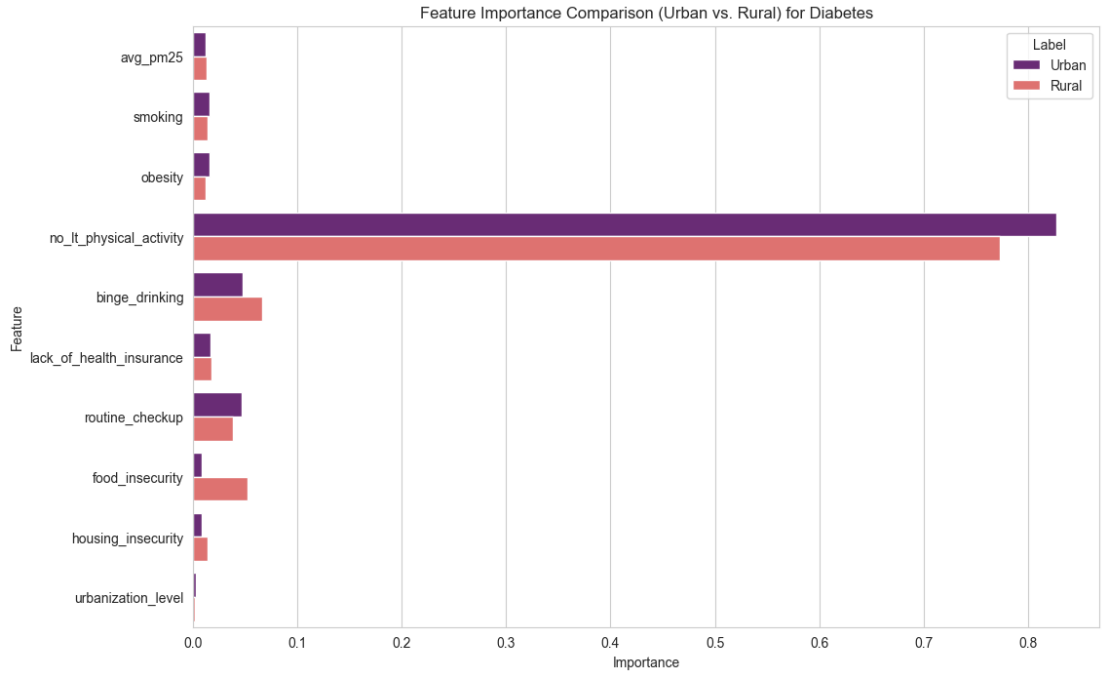

Figure 13: Feature importance comparison (Urban vs. Rural) for diabetes prediction using stratified Random Forest models.

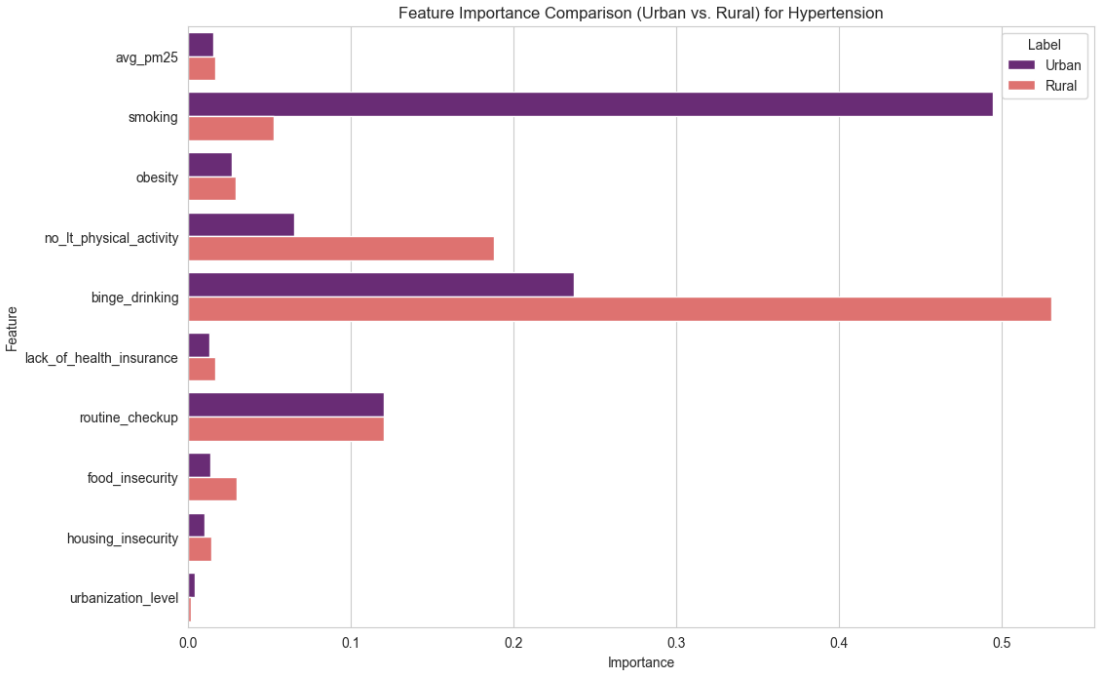

Figure 14: Feature importance comparison (Urban vs. Rural) for hypertension prediction using stratified Random Forest models.

Across both disease models, $PM_{2.5}$ exhibited a small but non-zero predictive contribution in both contexts, with slightly higher importance in rural counties. For diabetes, $PM_{2.5}$ importance was 0.0121 (urban) versus 0.0127 (rural). For hypertension, $PM_{2.5}$ importance was 0.0156 (urban) versus 0.0168 (rural). While these differences are modest, they suggest that $PM_{2.5}$ is not exclusively an urban risk signal and may retain comparable (or slightly stronger) relevance in rural settings.

7.4.2 Interaction-Term Modeling ($PM_{2.5}$×Urbanization)

To test whether urbanization amplifies or suppresses the $PM_{2.5}$ signal beyond what is captured by the main effects, an interaction feature was constructed:

$$\texttt{pm25\_urban\_interaction} = \texttt{avg\_pm25} \times \texttt{urbanization\_level}.$$

A Random Forest model was then trained using the original feature set plus this interaction term, and the relative importance rank of the interaction feature was examined.

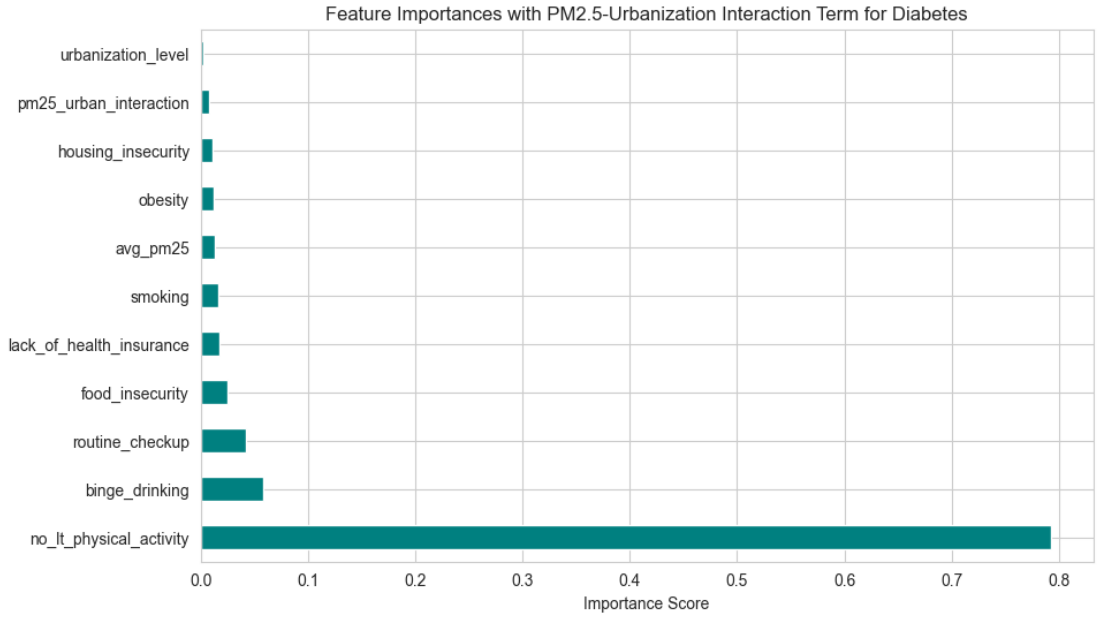

Figure 15: Feature importances for diabetes prediction including the $PM_{2.5}$×urbanization interaction term.

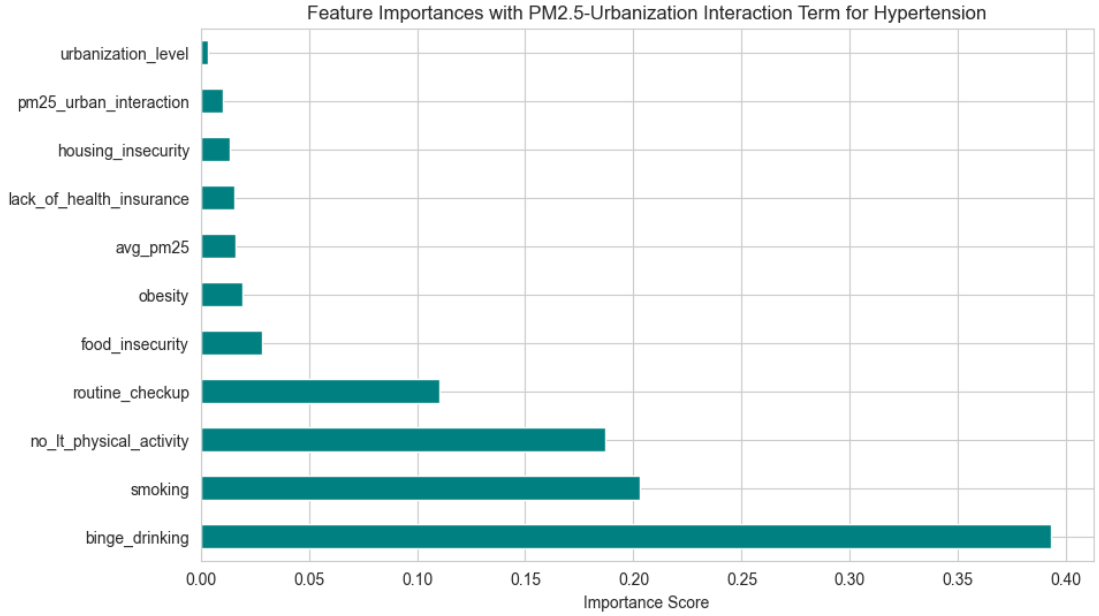

Figure 16: Feature importances for hypertension prediction including the $PM_{2.5}$×urbanization interaction term.

In both diabetes and hypertension models, the interaction term received very low importance relative to dominant behavioral predictors (e.g., physical inactivity, binge drinking, smoking). This indicates limited evidence that urbanization acts as a strong multiplicative modifier of the $PM_{2.5}$–disease relationship within the predictive framework used here.

7.4.3 Conclusion for H_3

Overall, the H_3 analyses provide weak support for strong urban amplification. Stratified Random Forest models show that $PM_{2.5}$ remains a consistent predictor in both urban and rural settings, with slightly higher importance in rural subsets, while the explicit $PM_{2.5}$×urbanization interaction term contributes minimal additional predictive value. These results suggest that $PM_{2.5}$ is best interpreted as a broadly relevant environmental risk factor whose predictive contribution persists across contexts, rather than a risk factor whose impact is substantially intensified by urbanization alone.

8 Key Findings

This study integrated tract-level health and environmental datasets into a population-weighted county-level analysis to examine the relationship between long-term $PM_{2.5}$ exposure and chronic disease prevalence. The main findings are summarized below:

- **Positive bivariate association:** Exploratory plots and correlation analysis indicate that counties with higher long-term $PM_{2.5}$ exposure tend to show higher prevalence of diabetes and hypertension, although the relationships are highly dispersed.
- **$PM_{2.5}$ contributes predictive signal but is not dominant:** Across regression and machine learning models, behavioral and socioeconomic risk factors (e.g., physical inactivity, smoking, binge drinking, obesity, and insecurity measures) explain substantially more variation in disease prevalence than $PM_{2.5}$. Nevertheless, $PM_{2.5}$ consistently retains non-zero predictive contribution.
- **Strong evidence of nonlinearity:** Ensemble tree-based models (Random Forest, Gradient Boosting, XGBoost) outperform linear and distance-based approaches for both outcomes (diabetes $R^2 \approx 0.91$; hypertension $R^2 \approx 0.89$), indicating nonlinear relationships and interactions among predictors.
- **Urbanization modifies predictive patterns only weakly:** Stratified models suggest $PM_{2.5}$ remains relevant in both urban and rural contexts with slightly higher importance in rural subsets, while explicit $PM_{2.5}$×urbanization interaction features contribute minimal additional importance relative to dominant behavioral predictors.

9 Limitations and Future Work

Several factors should be considered when interpreting these results. The analysis is observational and ecological, meaning county-level associations do not imply individual-level causal effects and may be influenced by unmeasured confounders. $PM_{2.5}$ exposure is based on modeled, aggregated estimates, which can introduce measurement error and temporal mismatch. In addition, strong correlations among behavioral and socioeconomic predictors can affect coefficient stability and interpretation in linear models. Finally, the predictor set is limited and does not include other pollutants or detailed source-related variables.

Future work should incorporate longitudinal designs and causal inference methods, expand environmental predictors (e.g., multiple pollutants and source proxies), and examine heterogeneity more directly through stratified and fairness-aware analyses. More detailed interpretation tools (e.g., SHAP dependence and interaction analyses) could further clarify how $PM_{2.5}$ contributes within complex risk environments.

10 Conclusion

This project integrated tract-level environmental and public health data into a population-weighted county-level dataset to study the relationship between long-term $PM_{2.5}$ exposure and chronic disease prevalence in the United States. Exploratory analysis indicates a positive but highly variable association between $PM_{2.5}$ and both diabetes and hypertension.

Predictive modeling shows that behavioral and socioeconomic factors explain most variation in disease prevalence, while $PM_{2.5}$ retains a consistent, non-zero predictive contribution across regression, regularization, and nonlinear ensemble models. The superior performance of tree-based methods further suggests nonlinear relationships among predictors. Urbanization analyses provide limited evidence of strong interaction effects, with $PM_{2.5}$ remaining relevant in both urban and rural contexts.

Overall, the results support a multifactorial view of chronic disease prevalence: lifestyle and structural determinants are dominant, but long-term environmental exposure adds measurable predictive signal that is relevant for public health monitoring and prevention planning.

AI Tool Usage Disclosure

ChatGPT was used to support writing, formatting, and overall presentation of this report. All analyses, computations, and primary findings are original work.

References

[1] Centers for Disease Control and Prevention (CDC). *PLACES: Local Data for Better Health*. Available at: https://www.kaggle.com/datasets/cdc/500-cities/data

[2] U.S. Environmental Protection Agency (EPA). *Daily Census Tract-Level $PM_{2.5}$ Concentrations (2011)*. Available at: https://healthdata.gov/CDC/Daily-Census-Tract-Level-PM2-5-Concentrations-2011/wwnf-fvrd/about_data

[3] U.S. Environmental Protection Agency (EPA). *Daily Census Tract-Level $PM_{2.5}$ Concentrations (2016)*. Available at: https://healthdata.gov/CDC/Daily-Census-Tract-Level-PM2-5-Concentrations-2016/k9st-jhz8/about_data

[4] National Center for Health Statistics (NCHS). *Urban–Rural Classification Scheme for Counties*. Available at: https://www.cdc.gov/nchs/data-analysis-tools/urban-rural.html

[5] Ahmet Vedat Kurt. *Long-Term $PM_{2.5}$ and Chronic Disease Prevalence (DSA210 Project)* (GitHub repository). Available at: https://github.com/ahmetvkrt/pm25-diabetes-hypertension