Title: Predictive Asthma Prevalence Modeling
Authors: Daida, K., Petrov, A., Ryuzaki, Y.

Introduction

Objective
The study objective is to understand and predict asthma prevalence at the census tract level using social determinants of health (SDOH) and environmental factors.

Problem, Impact, and Motivation
Asthma is a chronic respiratory condition that has a high economic burden (Bhattacharya et al., 2024) and an increasing prevalence rate in the United States (Pate & Zahran, 2024). Assessing the socioeconomic and environmental factors related to asthma prevalence allows us to identify communities that are vulnerable to health disparities, such as decreased access to medical care and increased rates of poor air quality indices. One of the study authors works for a managed care organization (MCO) that serves Medicaid and dual Medicare/Medicaid recipients, and is interested in providing additional asthma prevention funding for communities with high asthma prevalence rates.

Methodology
The study methodology was set up as follows:

1. Data collection and cleaning of three datasets, joined at the census-tract level.
2. Data exploration analysis.
3. Unsupervised algorithms for:
    a. Dimensionality reduction using principal component analysis (PCA).
    b. Feature engineering using clustering and community detection.
4. Supervised algorithms for prediction.
    a. Two baseline models were evaluated: linear regression and random forest models. Each baseline model had three approaches: raw features only, PCA transformed features, and raw features combined with outputs from K-Means clustering and community detection.
    b. An additional two models were evaluated: gradient boosting and XGBoost.
5. Failure analysis and advanced model evaluation on the best performing supervised learning model.

Most related studies use one supervised or unsupervised learning method, whereas our analysis uses the unsupervised learning section to inform the supervised learning section. Additional novel contribution highlights are included in the Related Work section.

Findings
The main findings for the unsupervised learning section were focused on the interrelationship among variables.

- **Principal Components Analysis (PCA):** The first component mostly had features related to social vulnerability, such as high poverty levels, minority status, and limited English proficiency. The second component had features that related to social isolation or age-related vulnerability, with higher values driven by higher disability rates, lack of internet access, and elderly populations.
- **K-Means Clustering:** The ideal cluster size was three clusters. The top two differentiating features in the clusters were the estimated percentage minority population and the estimated percentage of Hispanic population. This suggests that the demographic and socioeconomic composition are key factors that separate the clusters.
- **Louvain Community Detection:** The optimal community detection algorithm was with 10 neighbors and 0.9 resolution produced 14 communities. The top two differentiating features across communities were the estimated percentage minority population and the estimated percentage of housing in structures with 10 or more units. This suggests the demographic composition and community density are the key factors that separate the clusters.

The main findings for the supervised learning section were focused on asthma prevalence prediction. Of the evaluated models, the random forest model on raw data features performed the best. High poverty levels contributed approximately 40% to the performance of the model, followed by racial composition factors, such as the estimated percentage of Asians, Hispanics, and African-Americans/Blacks. The air quality was a moderate contributor.

Related Work

The literature review includes the following studies:

1. Lotfata et al. (2023) did a study titled, "**Socioeconomic and environmental determinants of asthma prevalence: a cross-sectional study at the U.S. County level using geographically weighted random forests**", which used data at the U.S. county-level to examine associations between asthma prevalence and different socioeconomic and environmental factors. The researchers found that, "...poverty, minority, depression prevalence, obesity prevalence, smoking prevalence, and green space were positively and non-linearly associated with asthma prevalence, while limited language, uninsured, mean temperature, $PM_{2.5}$ and O_3 were inversely associated" (Lotfata et al., 2023).

   This study used geographically weighted random forests, whereas our study uses an unsupervised learning technique to create features that feed into a supervised learning model to predict asthma prevalence. The study also used Behavioral Risk Factor Surveillance System (BRFSS) data, whereas we use three different data sources, one of which is a census-tract imputed version of the BRFSS data to calculate the asthma prevalence (PLACE).

2. In a study called, **"Geospatial Analysis of Social Determinants of Health Identifies Neighborhood Hot Spots Associated With Pediatric Intensive Care Use for Life-Threatening Asthma"**, Grunwell et al. (2021) used social determinants of health and readmission outcomes to evaluate hotspots for school-aged children who were admitted to a pediatric intensive care unit (PICU) for asthma. Researchers found that PICU admission rates per 1000 in the 90th percentile "were associated with a higher (ie, poorer) composite Social Vulnerability Index ranking, reflecting differences in socioeconomic status, household composition and disability, and housing type and transportation" (Grunwell et al., 2021).

   Unlike our study, this study focused on how socioeconomic factors impact the asthma exacerbation severity in the pediatric population of Atlanta, Georgia, and used a geospatial analysis on composite Social Vulnerability Index (SVI) factors to identify hotspots.

3. Tiotiu et al. (2020) performed a narrative review titled, "**Impact of Air Pollution on Asthma Outcomes**", which reviewed studies up to 2020 and found that both outdoor pollutants (such as traffic-related pollution, NO_2, ozone, and particulate matter) and indoor exposures (like secondhand smoke, heating sources, and mold) worsen asthma outcomes, triggering exacerbations, reducing lung function, and increasing healthcare use in adults and children, with childhood exposure also linked to asthma development. While a clear causal link between air pollution and adult-onset asthma isn't firmly established, evidence strongly supports that reducing exposure, through public health measures and personal precautions, can significantly improve asthma control and reduce its burden. This article differs from our study, which extends beyond a literature review and focuses on a research project.

This project is not an extension of a Milestone 1 or other previous course project.

Data Sources

Our study used the following data sources:

1. **Centers for Disease Control and Prevention (CDC) PLACES:** Used for asthma prevalence at the census-tract level (*PLACES: Local Data for Better Health*, 2024). This dataset is located and accessed through the CDC API URL (https://data.cdc.gov/resource/cwsq-ngmh.json). We converted and filtered the JSON file, which has 3.18M records and 24 columns, into a dataframe with 83,522 rows and 5 columns, which represent the following variables: state, state name, county name, FIPS, and the asthma prevalence. The asthma prevalence represents the 2022 current asthma prevalence among adults.

2. **Social Vulnerability Index (SVI):** Used for socioeconomic status, household characteristics, racial or ethnic minority status, housing type and transportation, and other social determinants of health (*Social Vulnerability Index*, 2024). This dataset is located and downloaded at the SVI landing page

(https://www.atsdr.cdc.gov/place-health/php/svi/svi-data-documentation-download.html) filtered on the year 2022. We converted the CSV file, which has 84,120 rows and 158 columns[1], into a dataframe with 84,120 rows and 25 columns.

3. **Air Quality System (AQS):** Used for air quality and other environmental factors (*Air Quality System (AQS) | US EPA*, 2025). This dataset is located and accessed through the CDC API URL (https://data.cdc.gov/resource/96sd-hxdt.json). We converted the JSON file, which has 136M records and 9 columns, into a dataframe with 83,776 rows and 2 columns, which have the FIPS and EPA modeled predictions of PM2.5 levels. The PM2.5 predictions were as of December 31, 2020.

4. **Census Regions:** Used for mapping the FIPS to US regions for visualizations (U.S. Census Bureau, n.d.) and considered an auxiliary dataset. This file is a PDF that maps the state FIPS to a US region. We manually converted the mapping to a CSV file and imported it into a 51 row, 2 column dataframe.

Feature Engineering

The following steps were taken to process the data:

1. The raw input datasets were in JSON or CSV file formats and loaded into dataframes. Redundant columns, such as state abbreviations, margin of error estimates or totals for predicted percentage estimates, or flags derived from existing columns were dropped.
2. The three main datasets (i.e., PLACES, SVI, and AQS) were joined on the FIPS at the census tract level. This resulted in a dataframe with 67,546 rows and 31 columns.
3. Rows with null SVI values (designated with -999) were dropped, because the null values impacted all SVI values in a given row. This had a 0.26% impact on the total dataset.
4. The final dataframe has 67,368 rows and 31 columns. A list of the final features is included in the Appendix (Figure 0).

For the unsupervised learning methods, we dropped non-numeric variables and used a standard scaler before running the PCA. The principal components were then fed into two separate unsupervised learning algorithms: a k-means clustering algorithm and Louvain community detection algorithm. Next, we evaluated the raw variables, clusters, and communities for multicollinearity using the variable inflation factor (VIF) - variables with a high VIF were removed ahead of the supervised learning section.

For the supervised learning methods, we created two models, linear regression and random forests, using three different approaches: raw features, the PCA transformed features, and raw features combined with outputs from the k-means clustering and community detection algorithms. To run a supervised learning model on the clustering or community detection algorithms, we one-hot encoded the clusters/communities into binarized variables (e.g., cluster_1 with values 0 and 1). We also ran gradient boosting and XGBoost on the raw variables with the binarized clusters and communities.

Unsupervised Learning

Methods Description

The unsupervised learning methods used were principal component analysis (PCA), k-means clustering, and Louvain community detection.

PCA

The workflow for the PCA was:

1. Drop non-numeric columns and the asthma prevalence.

[1]We used the data dictionary to extract the estimated percentage variables and exclude redundant columns. In an auxiliary analysis (*Auxiliary SVI Data Analysis*), we confirmed that the redundant columns had high pairwise Pearson correlation coefficients and infinite VIF scores. The infinite VIF scores signaled perfect multivariate collinearity.

2. Scale the remaining columns with a standard scaler. The standard scaler[2] was selected because it centers the data on mean = 0 and scales it to the unit variance, ensuring all features contribute equally to the principal components. It was selected over the min-max scaler, which doesn't center or standardize variance and can skew PCA results, and the robust scaler, which handles outliers but doesn't standardize variance either.
3. Select the top n components that represent at least 95% of the variance.

PCA works by computing the eigenvectors and eigenvalues of the data's covariance matrix, projecting the data onto the eigenvectors (principal components) corresponding to the largest eigenvalues, which reduces dimensionality while preserving maximal variance. This method was selected because it provides an optimal linear orthogonal projection of the data that minimizes reconstruction error under an L2 loss, making it the most efficient dimensionality reduction method when linearity, global variance preservation, and interpretability are important over methods such as UMAP or t-SNE.

K-Means Clustering

The workflow for the K-Means Clustering was:

1. Use the first 19 components from the PCA analysis.
2. Create eight models with the parameter *k*, the number of clusters, ranging from two to nine.
3. Evaluate and rank the models with the Silhouette score[3].

K-means clustering partitions data into *k* clusters by minimizing the within-cluster sum of squared Euclidean distances, iteratively assigning points to the nearest centroid and updating centroids as the mean of assigned points until convergence. This model was selected because it can be used as a feature engineering model by assigning cluster labels that capture latent group structure in the data, providing a compact, informative categorical feature for downstream supervised modeling.

Louvain Community Detection

The workflow for the Louvain Community Detection was:

1. Use the first 19 components from the PCA analysis.
2. Find the ideal parameters with the following combinations:
    a. Nearest neighbors: [10, 11, 12, 15, 20]
    b. Resolutions: [0.8, 0.9, 1.0]
3. Evaluate the best model based on the modularity score[4].

The Louvain community detection algorithm is a hierarchical, modularity-optimization method that partitions a graph by iteratively aggregating nodes into communities to maximize modularity gain, then constructing a new meta-graph where communities become nodes and repeating the process until convergence. This algorithm was selected because it is well-suited for feature engineering without requiring the number of clusters as input, enabling the extraction of meaningful, data-driven groupings for downstream tasks.

Unsupervised Model Evaluation

The best models from each family were as follows:

Model Family	**Hyperparameters**	**Evaluation**
Principal Component Analysis	Components = 19	95% of variance captured

[2] Standard scaler = $(X-\mu) / \sigma$ where X is the original value, μ is the mean of the feature, and σ is the standard deviation.

[3] Silhouette score for a single point = (b - a) / max(a, b) where a is the intra-cluster distance and b is the nearest-cluster distance.

[4] Modularity score = $Q = 1/(2m) * \Sigma[(A_{ij} - (k_i * k_j) / (2m)) * \delta(c_i, c_j)]$ where A_{ij} is the element of the adjacency matrix, indicating the presence or absence of an edge between nodes i and j, k_i is the degree of node i (number of connections), m is the total number of edges in the network, and $\delta(c_i, c_j)$ is the Kronecker delta, which is 1 if nodes i and j belong to the same community and 0 otherwise.

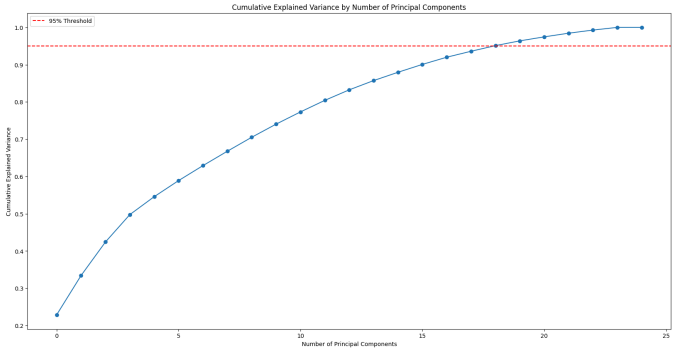
K-Means Clustering	Clusters = 3	0.266 Silhouette score
Louvain Community Detection	Resolution = 0.9 Neighbors = 10	0.720 modularity score

The justification for the evaluation metrics is described under each model family subsection.

PCA

The evaluation for the principal components is the cumulative explained variance. The results show that the top 19 components represent at least 95% of the variance.

Figure 1: Cumulative Explained Variance by Number of Principal Components

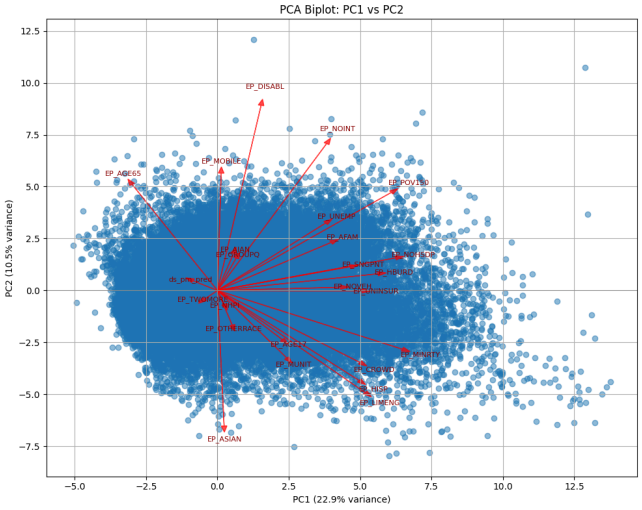


A more aggressive threshold could reduce the risk of model overfitting, but the 95% variance threshold was selected to prioritize minimal information loss while still reducing feature dimensionality.

A table of values corresponding to Figure 1 is included in the Appendix (Figure 1a).

From the PCA, we found that the first two loadings corresponded with socioeconomic vulnerabilities and social isolation or age-related vulnerabilities, respectively.

Figure 2: PCA Biplot



PC1 accounted for 22.9% of the variance and appears to represent a socioeconomic vulnerability axis, with higher values associated with poverty (EP_POV150), minority status (EP_MINRTY), and limited English proficiency (EP_LIMENG).

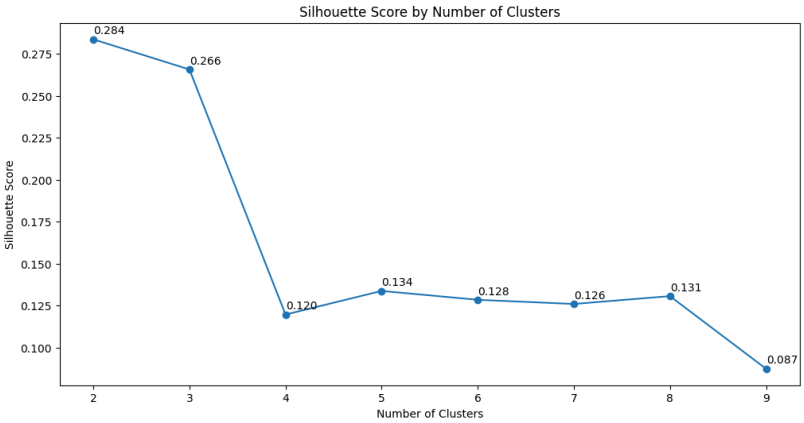
PC2 accounted for 10.5% of the variance and seems to capture a dimension of social isolation or age-related vulnerability, with higher values driven by disability (EP_DISABL), lack of internet access (EP_NOINT), and elderly populations (EP_AGE65).

K-Means Clustering

The evaluation method for the k-means clustering is the Silhouette score. The Silhouette score quantitatively evaluates clustering quality by measuring the cohesion of points within clusters and their separation from other clusters, making it a useful, label-independent metric for selecting the optimal number of clusters in unsupervised learning.

Based on the Silhouette scores, the ideal number of clusters is either two with a score of 0.284 or three with a score of 0.266.

Figure 3: Silhouette Score by Number of K-Means Clusters

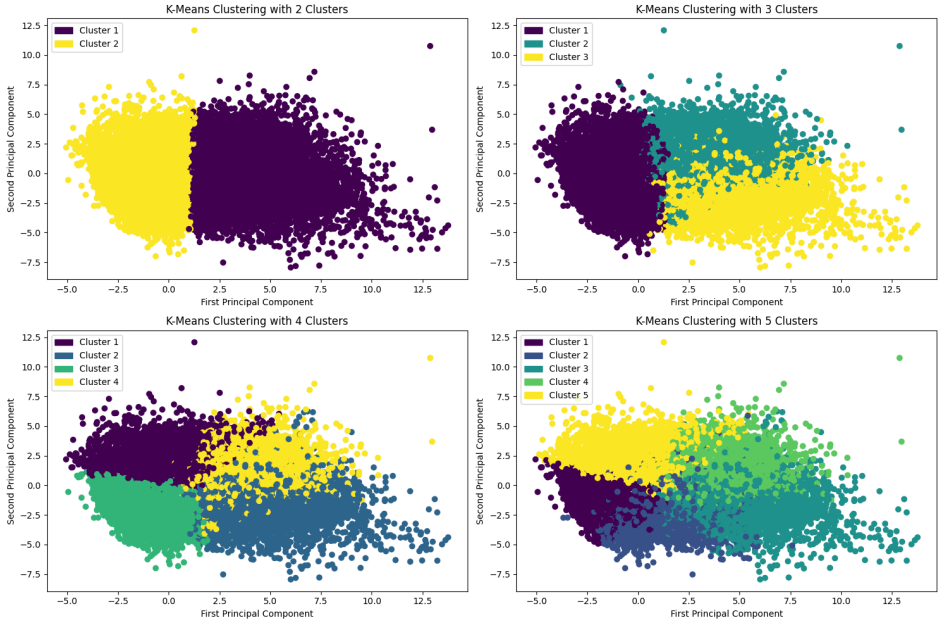


The Silhouette score drastically decreases at four or more clusters.

A table of values corresponding to Figure 3 is included in the Appendix (Figure 3a).

The following image visualizes the separation between clusters.

Figure 4: Plotting Clusters on 2D Space

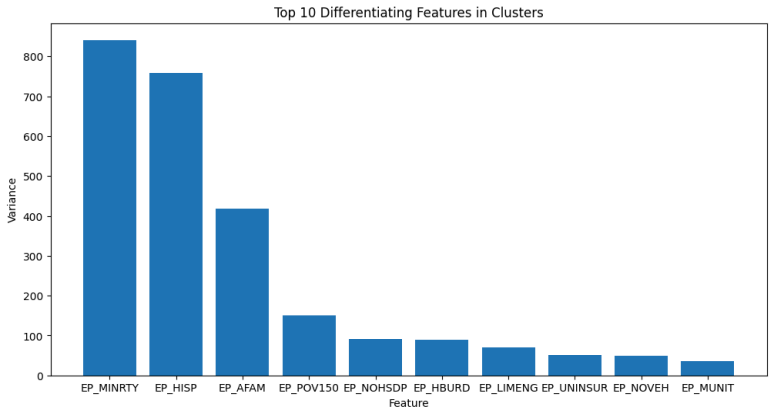


Although the two-cluster model provides the highest score, the three-cluster model provides the best balance between capturing complexity and having adequate decision boundaries for downstream modeling.

The clusters were added to the original dataset to highlight the highest and lowest asthma prevalence by cluster. The highest average asthma prevalence is in cluster 1 (12.17%) and the lowest asthma prevalence is in cluster 0 (10.33%). A table on the asthma prevalence distribution by cluster is included in the Appendix (Figure 4a).

To better understand how the three clusters differ, we visualized which features had the highest variance between clusters.

Figure 5: Top 10 Differentiating Features in Clusters



The top two differentiating features in the clusters were the estimated percentage of the minority population (EP_MINRTY) and estimated percentage of Hispanic population (EP_HISP). This suggests that the demographic and socioeconomic composition are key factors that separate the clusters.

A table of values corresponding to Figure 5 is included in the Appendix (Figure 5a).

Louvain Community Detection

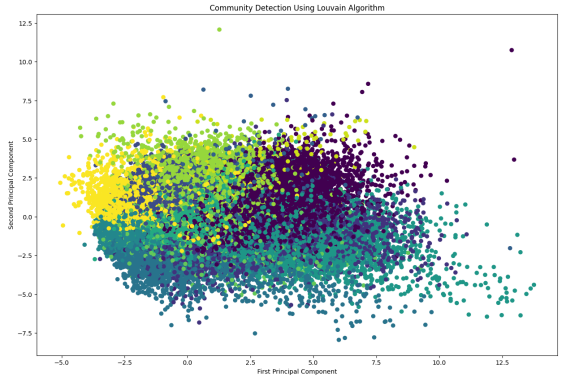
The modularity score is the evaluation method for the Louvain Community Detection algorithm. The modularity score evaluates the quality of a community partition in a graph by quantifying the difference between the observed intra-community edge density and the expected edge density under a configuration null model. Modularity is preferred in community detection because it directly leverages the graph's topology, accounting for both edge density and network structure, unlike general clustering metrics which may ignore connectivity patterns critical to community formation.

We ran 15 versions of the model with the following hyperparameters:

Modularity Score	**Resolution**		
# of Neighbors	**0.8**	**0.9**	**1.0**
10	0.7116	0.7201	0.7134
11	0.7078	0.7156	0.7124
12	0.7150	0.7175	0.7184
15	0.6948	0.6971	0.7061
20	0.7034	0.7034	0.7049

The optimal model based on the modularity score has a 0.9 resolution and 10 neighbors. Visualizing the optimal community detection model yields the following:

Figure 6: Community Detection using Louvain Algorithm on 2D Space

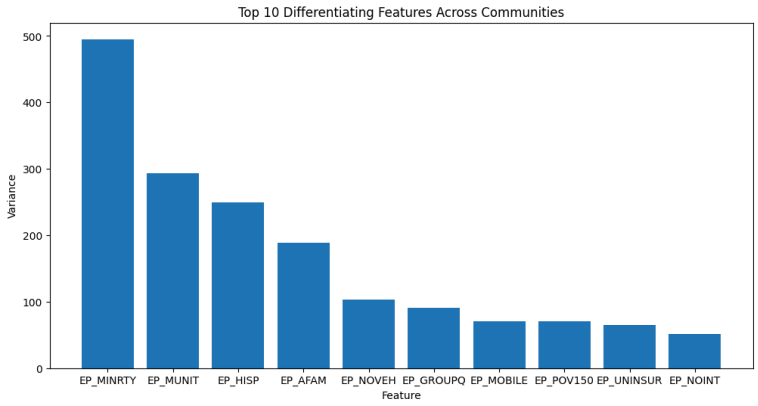


This visual shows the communities within a 2D space using results from PCA. Unlike k-means clustering which relies on spatial distance, community detection algorithms create communities based on graph connectivity. This is why some communities seem to be more intermixed when graphed on a 2D space than clusters, despite there being substantial differences between them.

The optimal community detection algorithm was with 10 neighbors and 0.9 resolution produced 14 communities. The 14 communities were added to the original dataset to highlight the highest and lowest asthma prevalence by community. Community 12 had the highest average asthma prevalence (12.68%), while community 5 had the lowest asthma prevalence (9.02%). Asthma prevalence distribution by community is included in the appendix (Figure 6a).

To better understand how the communities differ, we visualized which features had the highest variance between clusters.

Figure 7: Top 10 Differentiating Features Across Communities



The top two differentiating features across communities were the estimated percentage of the minority population (EP_MINRTY) and the estimated percentage of housing in structures with 10 or more units (EP_MUNIT). This suggests the demographic composition and community density are the key factors that separate the clusters.

A table of values corresponding to Figure 7 is included in the Appendix (Figure 7a).

Multicollinearity

Before moving to supervised learning, we evaluated the raw variables, clusters, and communities for multicollinearity. A view on the pairwise correlations is included in the Appendix (Figure 8). We used the variable inflation factor (VIF) to review multivariate correlations and removed EP_MINRTY, which likely represents mutually exclusive race-related variables. When EP_MINRTY was removed, all but one variable had a VIF score under 5.0.

Supervised Learning

Methods Description

We used four supervised learning models to create the asthma prevalence predictions: linear regression, random forests, gradient boosting, and XGboost. For the linear regression and random forest, we ran the models using three different data approaches: raw data, PCA-transformed features, and raw data with cluster and community features.

Linear Regression

The workflow for the linear regression was:

1. Drop non-numeric columns and the asthma prevalence.
2. Scale the independent columns in the training and test sets with a standard scaler. The standard scaler was selected because it preserves the underlying distribution shape, ensures unit variance for stable optimization, and allows for clearer coefficient interpretation, especially when using regularization.
3. Predict the linear regression model and evaluate the performance with the mean absolute error (MAE)[5] and R^2 score[6] as the mean metrics from a five-folds cross validation.
4. Go through steps 1 through 4 for the three different approaches: raw data, PCA-transformed features, and raw data with cluster and community features.

Linear regression estimates the coefficients of a linear equation by minimizing the residual sum of squares between the observed responses and the responses predicted by a linear combination of the input features. We used linear

[5] MSE = $(1/n) * \Sigma|Y_i - \hat{Y}_i|$ where n is the number of observations, Y_i is actual (true) value, and $\hat{Y}_i$ is the predicted value.

[6] $R^2 = 1 - (SS_{res}) / (SS_{tot})$ where SS_{res} is the residual sum of squares and SS_{tot} is the total sum of squares.

regression because it is an interpretable, computationally efficient model that quantifies the relationship between input features and a continuous target variable, making it suitable for both prediction and inference.

Random Forest

The workflow for the random forest was:

1. Drop non-numeric columns and the asthma prevalence.
2. Scale the independent columns in the training and test sets with a standard scaler.
3. Predict the random forest and evaluate the performance with the MAE and R^2 as the mean metrics from a five-folds cross validation.
4. Go through steps 1 through 4 for the three different approaches: raw data, PCA-transformed features, and raw data with cluster and community features.

A random forest is an ensemble learning algorithm that constructs multiple decision trees using bootstrapped samples and random feature selection at each split, and aggregates their predictions through averaging (for regression) or majority voting (for classification) to improve generalization and reduce overfitting. We chose this model because it captures complex, nonlinear relationships and interactions between environmental, demographic, and clinical variables without requiring prior feature transformation, while offering robustness to overfitting and strong predictive accuracy on high-dimensional or noisy datasets.

Gradient Boosting

The workflow for the gradient boosting algorithm was:

1. Drop non-numeric columns and the asthma prevalence.
2. Scale the independent columns in the training and test sets with a standard scaler.
3. Evaluate the performance with the mean absolute error (MAE)[7] and R^2 score[8] as the mean metrics from a five-folds cross validation.
4. Highlight feature importance.

Gradient boosting is an ensemble learning method that builds a sequence of decision trees, where each new tree fits the residuals (errors) of the previous ones to minimize a specified loss function through gradient descent. It is useful for predicting asthma prevalence because it models complex, nonlinear relationships with high predictive accuracy, handles mixed data types, and can be fine-tuned to reduce overfitting.

XGBoost

The workflow for the XGBoost algorithm was:

1. Drop non-numeric columns and the asthma prevalence.
2. Scale the independent columns in the training and test sets with a standard scaler.
3. Use random search to optimize hyperparameters evaluated with the root mean squared error (RMSE)[9].
4. Use L2 regularization[10] and feature subsampling to avoid overfitting.
5. Run the final XGBoost algorithm using the optimized hyperparameters. evaluate the performance with the mean absolute error (MAE)[11] and R^2 score[12] as the mean metrics from a ten-folds cross validation.

XGBoost is a scalable, gradient boosting framework that builds an ensemble of decision trees by sequentially minimizing a specified loss function using second-order gradient descent and regularization to improve predictive accuracy and prevent overfitting. We used XGBoost because it delivers high accuracy, handles nonlinear relationships and feature interactions effectively, supports regularization to reduce overfitting, and is computationally efficient even on large, noisy, or structured datasets.

---

[7] $MSE = (1/n) * \Sigma|Y_i - \hat{Y}_i|$ where n is the number of observations, Y_i is actual (true) value, and $\hat{Y}_i$ is the predicted value.

[8] $R^2 = 1 - (SS_{res}) / (SS_{tot})$ where SS_{res} is the residual sum of squares and SS_{tot} is the total sum of squares.

[9] $RMSE = \sqrt{[(\Sigma(P_i - O_i)^2) / n]}$ where P_i is the predicted value and O_i is the observed value.

[10] L2 regularization is the ridge penalty.

[11] $MSE = (1/n) * \Sigma|Y_i - \hat{Y}_i|$ where n is the number of observations, Y_i is actual (true) value, and $\hat{Y}_i$ is the predicted value.

[12] $R^2 = 1 - (SS_{res}) / (SS_{tot})$ where SS_{res} is the residual sum of squares and SS_{tot} is the total sum of squares.

Supervised Learning Evaluation

The best models from each family and data approach were as follows:

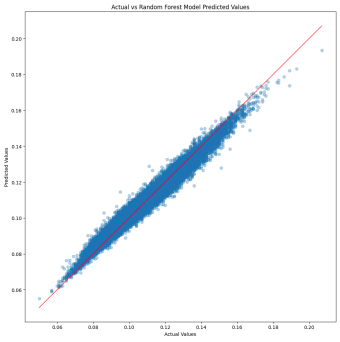
Model Family	Data	Hyperparameters	MAE	R^2
Linear Regression	Raw	-	0.0065±0.00002	0.681±0.00525
Linear Regression	PCA-Transformed	-	0.0070±0.00003	0.633±0.00642
Linear Regression	Raw + Clusters + Communities	-	0.0064±0.00003	0.688±0.00522
Random Forest	Raw	Estimators = 100	0.0056±0.00003	0.753±0.00271
Random Forest	PCA-Transformed	Estimators = 100	0.0063±0.00003	0.691±0.00229
Random Forest	Raw + Clusters + Communities	Estimators = 100	0.0056±0.00003	0.753±0.00244
Gradient Boosting	Raw + Clusters + Communities	Estimators = 200 Max depth = 5 Max features = 50%	0.0062±0.00009	0.704±0.00746
XGBoost	Raw + Clusters + Communities	Estimators = 100 Learning rate = 0.05 L2 lambda = 0.5 Col sample = 0.5	0.0073±0.0001	0.598±0.0181

The reported MAE and R^2 are the mean metrics from five-fold cross validation. MAE measures the predictive error, and R^2 is used to interpret how well the model explains the variation in the target variable. MAE was selected over the mean squared error (MSE) because the MSE does not handle scaled variables well. XGBoost hyperparameters were selected via random search using MAE, and the final model was evaluated using the MAE and R^2.

Best Model Analytics

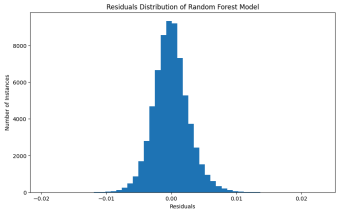
The best model is a random forest with raw data and 100 estimators.

Figure 9a: Actual versus Random Forest Predicted Values



This scatter plot compares actual values to predicted values from a random forest (RF) regression model. The predicted values align closely with the actual values along the diagonal y = x line, indicating strong model performance and high predictive accuracy. The concentration of points around the diagonal suggests low residual error across most of the range. There is minor deviation at the distribution's extremes, with slight underprediction of higher values, consistent with RF's tendency to regress predictions toward the mean. The error variance appears relatively constant across the range, suggesting homoscedasticity. Overall, the model demonstrates strong fit with limited systematic bias.

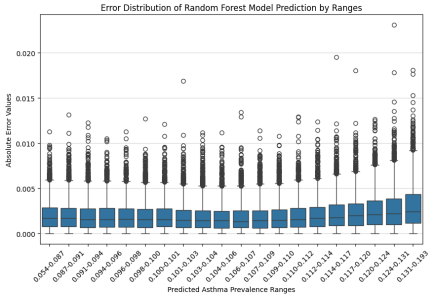
Figure 9b: Random Forest Residual Distribution



This histogram shows the distribution of residuals (prediction errors) from the random forest model. The distribution is approximately symmetric and centered around zero, indicating that the model is unbiased overall. The residuals are tightly clustered near zero, with a steep peak and rapidly decreasing frequency as the magnitude increases, suggesting low variance and high model precision.

The near-normal shape, with no heavy tails or skewness, implies that the model does not systematically over- or underpredict across the range of values. These characteristics support the conclusion that the model performs well with well-behaved, homoscedastic residuals.

Figure 9c: Error Distribution of Random Forest Model Prediction by Range

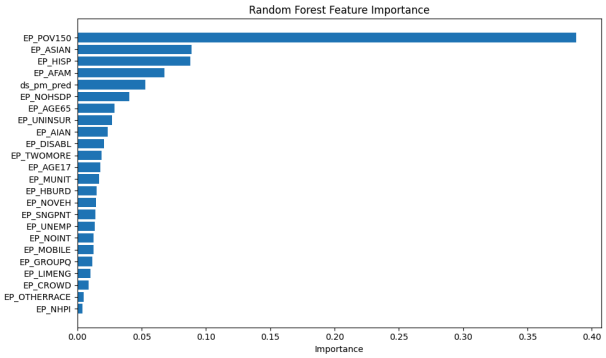


This boxplot illustrates the distribution of absolute prediction errors across binned predicted asthma prevalence values from a random forest model. Each box represents the spread of absolute errors within a specific range of predicted values. The plot shows that while the median absolute error remains low across all bins, the interquartile range (IQR) and the number of outliers increase slightly as predicted prevalence values rise. This trend suggests that the model becomes marginally less precise for higher predicted values, likely due to fewer training samples or higher variability in those regions.

The overall error remains relatively contained, indicating that the model maintains good generalization performance across the prediction spectrum, with only a modest increase in uncertainty for higher prevalence predictions.

A feature importance analysis yielded the following:

Figure 10: Random Forest Feature Importance



Poverty levels (EP_POV15) contributed approximately 40% of the model performance, followed by racial composition factors, such as the estimated percentage of Asians (EP_ASIAN), Hispanics (EP_HISP), and African-Americans/Blacks (EP_AFAM). The air quality (ds_pm_pred) was a moderate contributor.

A table of values corresponding to Figure 10 is included in the Appendix (Figure 10a).

This analysis can be interpreted as:

- Socioeconomic factors dominate. Poverty alone contributes dramatically more than other variables.
- Racial and ethnic demographics matter, but are fragmented across multiple indicators, none of which individually rival poverty.
- Environmental exposure plays a moderate role, but not as central as social vulnerability metrics.

We ran an ablation analysis on EP_POV in one batch and racial composition factors in another batch.

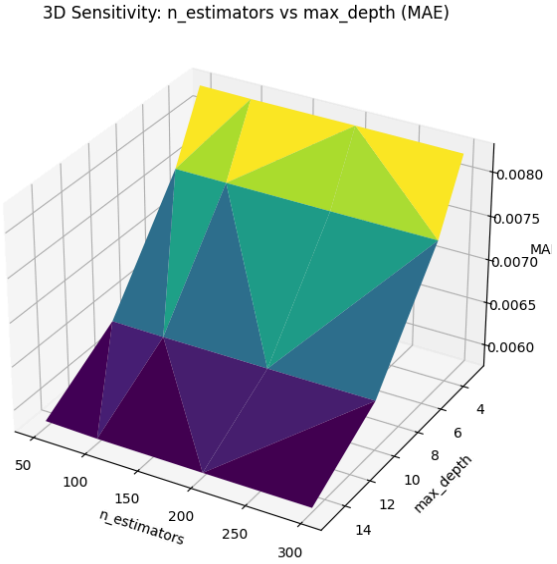
Ablation	**MAE**	**R2**	**MAE_Change**	**R2_Change**
Baseline (All Features)	0.0056	0.753	0	0
No EP_POV150	0.0061	0.708	0.0005	-0.0445
No Racial Features	0.0066	0.665	0.0009	-0.0888

Poverty (EP_POV150) is moderately important. Removing it slightly worsens predictive accuracy and reduces explanatory power. It's likely partially correlated with other socioeconomic variables, so some signal gets absorbed

elsewhere. Racial features provide distinct and meaningful signals in the model. Removing them causes the biggest performance drop, suggesting they capture variance not explained by other variables and they're not redundant in the same way poverty is.

We also ran a sensitivity analysis on two hyperparameters: the number of estimators and the max depth.

Figure 11: 3D MAE Sensitivity on # Estimators and Max Depth

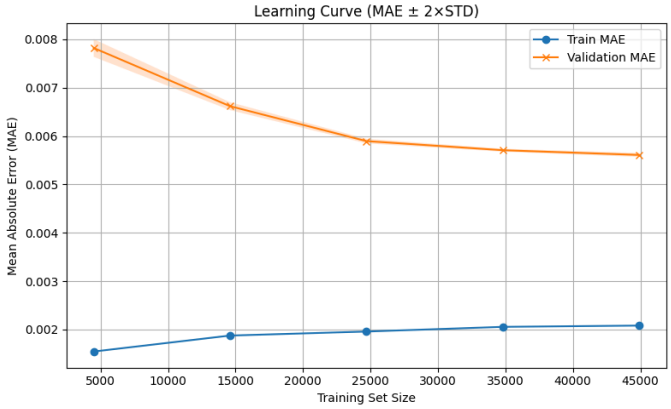


This 3D sensitivity analysis illustrates how model performance, measured by Mean Absolute Error (MAE), responds to changes in the number of estimators and tree depth in a random forest. The results show a sharp decline in MAE as max_depth increases, indicating that shallow trees (depths of 3 or 5) significantly underfit the data. In contrast, models with deeper trees (10 or 15) consistently yield lower errors, especially when paired with a sufficient number of estimators. While increasing n_estimators improves stability, its marginal impact diminishes beyond 100–200 trees. Overall, this plot confirms that tree depth is a critical driver of accuracy, and overly simplistic models severely compromise performance.

MAE	**Max Depth**			
# Estimators	**3**	**5**	**10**	**15**
50	0.002653	0.001261	0.000062	0.000042
100	0.002654	0.001255	0.000058	0.000039
200	0.002654	0.001255	0.000056	0.000037
300	0.002656	0.001255	0.000055	0.000037

We ran learning curves on this tree to determine the tradeoff on how much data is needed for optimal performance.

Figure 12: Learning Curve for Optimal Performance



The learning curve shows that validation MAE decreases steadily with more training data, leveling off around 35,000–45,000 observations. Train MAE rises slightly but remains low, reflecting reduced memorization as data increases. The small, stable gap between train and validation errors suggests good generalization and minimal overfitting. Overall, the model performs well and may benefit from modest additional data. A table of values corresponding to Figure 12 is included in the Appendix (Figure 12a).

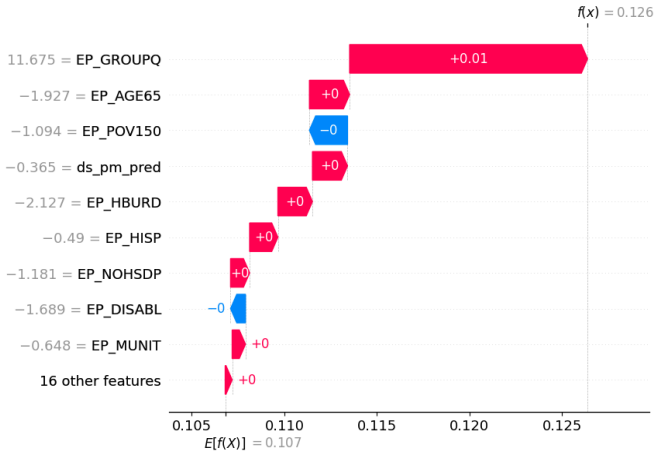
Additional tradeoffs include:

- **Bias versus Variance**: In Figure 9b, the residual histogram is centered and symmetric, and actual vs. predicted values show tight clustering around the diagonal, which implies low bias. However, slight over/underprediction at the extremes (from the scatter and boxplot) implies the model smooths predictions.The model minimizes variance with consistent performance across most predictions but introduces mild bias by underpredicting at higher prevalence levels.
- **Model Complexity versus Speed:** The random forest model achieves strong accuracy and robustness but requires more computational time and resources compared to simpler alternatives.

Failure Analysis

We used SHAP to provide insights on why the model made large prediction errors.

Figure 13: Failure Case #1 - Large Prediction Error

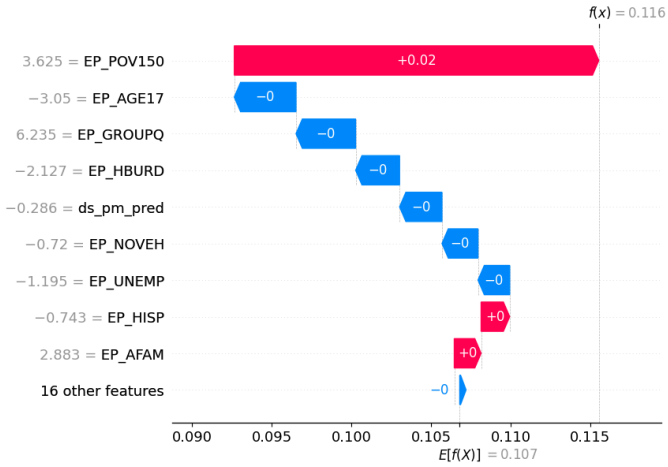


This SHAP waterfall plot explains the prediction for a single instance with the largest absolute error in the model. The base value (mean model prediction) is 0.107, and the final predicted value for this instance is 0.126. The largest positive contributor to the prediction is the percentage of persons in group quarters (EP_GROUPQ) at +0.01, followed by smaller contributions from the elderly, air quality, and other features. Negative contributions include poverty and disability rates, though their magnitudes are small. The overall error suggests the model may be overemphasizing EP_GROUPQ in this context, leading to overprediction. A table of values corresponding to Figure 13 is included in the Appendix (Figure 13a).

To address this, future improvements could disaggregate EP_GROUPQ into more specific categories, explore interaction terms (e.g., between age and housing status), and consider custom sampling or loss functions that prioritize cases that are difficult to predict.

We used SHAP to provide insights on the worst performing record in the third quintile of poverty predictions to highlight potential systemic bias.

Figure 14: Failure Case #2 - Systemic Bias

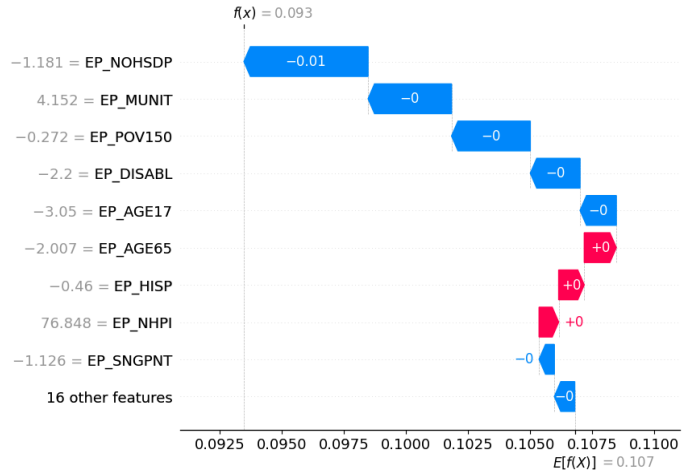


Poverty (EP_POV150) exerts an outsized influence on the predicted asthma prevalence, even when other indicators (e.g., lower household burden, reduced air pollution) suggest a lower risk. While poverty is a known correlate of health disparities, the model's overreliance on it in this case may reflect or reinforce structural assumptions that treat poverty as a deterministic proxy for poor health. This could lead to overprediction in high-poverty areas that have mitigating community or environmental factors. A table of values corresponding to Figure 14 is included in the Appendix (Figure 14a).

To reduce potential systemic bias, future improvements could consider conducting subgroup error analysis to verify if high-poverty areas are consistently overpredicted. Incorporating fairness-aware modeling techniques or regularization constraints could help ensure that poverty alone doesn't disproportionately drive predictions without sufficient contextual nuance.

We used SHAP to provide insights on an outlier case identified by total z-score across features.

Figure 15: Failure Case #3 - Outlier/Edge Case



This outlier case was identified based on extreme feature z-scores, with the percentage of Native Hawaiian/Pacific Island (EP_NHPI) standing out as highly unusual. The model predicted a lower-than-average asthma prevalence, primarily driven by negative contributions from low educational attainment (EP_NOHSDP) and low poverty (EP_POV150). Although the high NHPI value nudged the prediction upward, its small SHAP value suggests the model may underweight patterns specific to smaller or less-represented populations. This could indicate underfitting for edge communities or systemic data imbalance, raising questions about predictive reliability in high-variance demographic context. A table of values corresponding to Figure 15 is included in the Appendix (Figure 15a).

To better handle edge cases, future improvements could consider data augmentation for underrepresented subgroups (e.g., Native Hawaiian and Pacific Islander populations), and assess whether model calibration or more granular subgroup modeling can improve sensitivity to valid high-risk signals without inflating prediction errors.

Discussion

For unsupervised learning, we were surprised by how clearly socioeconomic and racial demographics consistently were found to be top factors for asthma prevalence. We expected certain environmental factors such as air quality (ds_pm_pred) to have a more significant effect on asthma prevalence, but in comparison to socioeconomic features, the effects of air quality are much more minimal. A challenge we encountered at this step was selecting the optimal amount of clusters for K-Means. Based on the Silhouette score alone, a model using two clusters performed better than a model with three clusters. To respond to this challenge, we decided to visualize the different K-Means models and compare how well each group was separated within our data. This revealed that the use of three clusters provided the best balance between capturing complexity and having adequate separation between the clusters, and ultimately we decided that despite the silhouette score being a bit lower, the three cluster solution would allow us to discover deeper patterns that would be missed with a two way split. Given more time and resources, our group would check for sources of bias within the data, such as determining whether certain socioeconomic groups are overrepresented in clusters with high asthma prevalence. The goal of this would be to ensure that our unsupervised learning results reflect true underlying patterns within the data rather than misclassifying communities based on biases in the input data.

For supervised learning, we learned how to conduct robust failure analytics, specifically using SHAP - this is not a method we have been introduced to through MADS yet. We were surprised that the most parsimonious supervised learning models had the highest predictive power and could describe more of the variance than models that depended on unsupervised learning generated features or more complex models. Despite extensive hyperparameter optimization and validation, the XGBoost model underperformed relative to Random Forest. This result appears to stem from structural characteristics of the dataset: the underlying relationships between predictors and the outcome are largely additive with limited interaction effects, and the signal-to-noise ratio is moderate. In such cases, Random Forest's averaging-based approach tends to perform better by capturing broad marginal trends while remaining

robust to noise and multicollinearity. Conversely, XGBoost's boosting framework, which relies on sequential error correction, may over-regularize or amplify residual noise when complex hierarchical patterns are absent. These findings suggest that the dataset's statistical structure is better suited to bagging methods like Random Forest than to gradient boosting. Challenges included deciding between keeping or omitting SVI redundant variables e.g., estimated percentage of minorities versus total number of minorities. We ultimately decided to remove the total values to retain population-normalized metrics and avoid underrepresenting risk in rural areas. If we had more time, we would evaluate the impacts of both populations.

Ethical Concerns

An ethical concern surrounding unsupervised learning is the overrepresentation of minority groups in clusters with high asthma prevalence. This could result in the reinforcement of previously existing disparities within the healthcare sector and within the decision-making process for public health in the United States. To address this, we would need to monitor for biases within the created clusters and monitor the interpretation of our results to prevent any misuse of our findings. Another ethical concern surrounding unsupervised learning arises from the use of data sourced from previous years (the most recent data being from 2022). This means that our results may not fully exhibit the current-day features within communities, especially in areas that have had recent population shifts or environmental changes. To address this, future work on this project would involve rerunning our clustering models whenever new data from our sources is made available, ensuring that predictions remain relevant and accurate.

The supervised learning process carries similar ethical concerns to those discussed above. Because race-related features were used directly or indirectly (features generated through unsupervised learning) for training the supervised learning models, there is a risk that the results could overrepresent minority groups or reinforce the existing systematic biases, especially if the outcomes are interpreted without care. Although racial data is commonly used in healthcare research for examining health disparities, its use must be handled carefully. In our case, ensuring the results are being used for uncovering inequity and not getting misused for discriminatory generalizations is important. To address this concern, we need to emphasize the significance of interpreting results within broader social, environmental, and economic contexts. Also, rather than making decisions or drawing conclusions solely based on racial distribution, we strongly encourage future research to explore the possible underlying factors that form the observed patterns. For example, these can include differences in diet, environmental exposures, healthcare access, socioeconomic status, or lifestyle habits between different groups. We believe understanding the root causes like these is the key and can lead to more equitable, effective, and ethically sound healthcare decision-making.

Statement of Work

The lead took on the initial work for a given task. The reviewers provided support to the task, including reviewing the work, adding code/visuals/insights, and tying the task to the project write-up. Everyone in the team was involved in the project from start to finish, including the problem formulation, data collection and cleaning, modeling, and project write-up. The task assignments are outlined in Figure 16 in the Appendix.

References

About PLACES: Local data for better health. (2024, October 29). PLACES: Local Data for Better Health. https://www.cdc.gov/places/about/index.html

Air Quality System (AQS) | US EPA. (2025, March 27). US EPA. https://www.epa.gov/aqs

Bhattacharya, A., Syamlal, G., & Dodd, K. E. (2024). Medical costs and incremental medical costs of asthma among workers in the United States. *American Journal of Industrial Medicine*, *67*(9), 834–843. https://doi.org/10.1002/ajim.23633

Grunwell, J. R., Opolka, C., Mason, C., & Fitzpatrick, A. M. (2021). Geospatial analysis of social determinants of health identifies neighborhood hot spots associated with pediatric intensive care use for Life-Threatening asthma. *The Journal of Allergy and Clinical Immunology in Practice*, *10*(4), 981-991.e1. https://doi.org/10.1016/j.jaip.2021.10.065

Lotfata, A., Moosazadeh, M., Helbich, M., & Hoseini, B. (2023). Socioeconomic and environmental determinants of asthma prevalence: a cross-sectional study at the U.S. County level using geographically weighted random forests. *International Journal of Health Geographics*, *22*(1). https://doi.org/10.1186/s12942-023-00343-6

Pate, C. A., & Zahran, H. S. (2024). The status of asthma in the United States. *Preventing Chronic Disease*, *21*. https://doi.org/10.5888/pcd21.240005

Social Vulnerability Index. (2024, July 22). Place and Health - Geospatial Research, Analysis, and Services Program (GRASP). https://www.atsdr.cdc.gov/place-health/php/svi/?CDC_AAref_Val=https://www.atsdr.cdc.gov/placeandhealth/svi/index.html

Tiotiu, A. I., Novakova, P., Nedeva, D., Chong-Neto, H. J., Novakova, S., Steiropoulos, P., & Kowal, K. (2020). Impact of air pollution on asthma outcomes. *International Journal of Environmental Research and Public Health*, *17*(17), 6212. https://doi.org/10.3390/ijerph17176212

U.S. Census Bureau. (n.d.). Census regions and divisions of the United States. In *U.S. Census Bureau*. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

Wang, N., & Nurmagambetov, T. (2024). Sociodemographic factors of asthma prevalence and costs among children and adolescents in the United States, 2016–2021. *Preventing Chronic Disease*, *21*. https://doi.org/10.5888/pcd21.230449

Project Submission

Github: The project's code and data files can be found at https://github.com/yoryuzaki/ss25_milestone2_team3

Google Doc: The Google Doc version of the written report with comments enabled can be found at https://docs.google.com/document/d/1aJ0pUjcNUKwmW_jSucgJl4vMvTBzc0cpgCcfheuOFeE/edit?usp=sharing

Appendix

Figure 0: Raw Data Features

Variable Name	Source	Description
asthma_prevalence	PLACES	Asthma prevalence
countyname	PLACES	County name
stateabbr	PLACES	Two letter state abbreviation
statedesc	PLACES	State name
Region	Census	United States region (e.g., Northeast, Midwest, South)
FIPS	SVI	11-digit FIPS number at the census tract level
EP_AFAM	SVI	Adjunct variable - Percentage of Black/African American, not Hispanic or Latino persons estimate, 2018-2022 ACS
EP_AGE17	SVI	Percentage of persons aged 17 and younger estimate, 2018-2022 ACS
EP_AGE65	SVI	Percentage of persons aged 65 and older estimate, 2018-2022 ACS
EP_AIAN	SVI	Adjunct variable - Percentage of American Indian or Alaska Native, not Hispanic or Latino persons estimate, 2018-2022 ACS
EP_ASIAN	SVI	Adjunct variable - Percentage of Asian, not Hispanic or Latino persons estimate, 2018-2022 ACS
EP_CROWD	SVI	Percentage of occupied housing units with more people than rooms estimate
EP_DISABL	SVI	Percentage of civilian noninstitutionalized population with a disability estimate, 2018-2022 ACS
EP_GROUPQ	SVI	Percentage of persons in group quarters estimate, 2018-2022 ACS
EP_HBURD	SVI	Percentage of housing cost-burdened occupied housing units with annual income less than $75,000 (30%+ of income spent on housing costs) estimate, 2018-2022 ACS estimate, 2018-2022 ACS
EP_HISP	SVI	Adjunct variable - Percentage of Hispanic or Latino persons estimate, 2018-2022 ACS
EP_LIMENG	SVI	Percentage of persons (age 5+) who speak English "less than well" estimate, 2018-2022

		ACS
EP_MINRTY	SVI	Percentage minority (Hispanic or Latino (of any race); Black and African American, Not Hispanic or Latino; American Indian and Alaska Native, Not Hispanic or Latino; Asian, Not Hispanic or Latino; Native Hawaiian and Other Pacific Islander, Not Hispanic or Latino; Two or More Races, Not Hispanic or Latino; Other Races, Not Hispanic or Latino) estimate, 2018-2022 ACS*
EP_MOBILE	SVI	Percentage of mobile homes estimate
EP_MUNIT	SVI	Percentage of housing in structures with 10 or more units estimate
EP_NHPI	SVI	Adjunct variable - Percentage of Native Hawaiian or Other Pacific Islander, not Hispanic or Latino persons estimate, 2018-2022 ACS
EP_NOHSDP	SVI	Percentage of persons with no high school diploma (age 25+) estimate
EP_NOINT	SVI	Adjunct variable - Percentage of households without an internet subscription estimate, 2018-2022 ACS
EP_NOVEH	SVI	Percentage of households with no vehicle available estimate
EP_OTHERRACE	SVI	Adjunct variable - Percentage of some other race, not Hispanic or Latino persons estimate, 2018-2022 ACS
EP_POV150	SVI	Percentage of persons below 150% poverty estimate
EP_SNGPNT	SVI	Percentage of single-parent households with children under 18 estimate, 2018-2022 ACS
EP_TWOMORE	SVI	Adjunct variable - Percentage of two or more races, not Hispanic or Latino persons estimate, 2018-2022 ACS
EP_UNEMP	SVI	Unemployment Rate estimate
EP_UNINSUR	SVI	Percentage uninsured in the total civilian noninstitutionalized population estimate, 2018-2022 ACS
ds_pm_pred	AQS	EPA modeled predictions of PM2.5 levels

Figure 1a: Cumulative Explained Variance by Number of Principal Components Table

# of Principal Components	Cumulative Explained Variance
0	0.23
1	0.33
2	0.42
3	0.50
4	0.55
5	0.59
6	0.63
7	0.67
8	0.71
9	0.74
10	0.77
11	0.80
12	0.83
13	0.86
14	0.88
15	0.90
16	0.92
17	0.94
18	0.95
19	0.96
20	0.97
21	0.98
22	0.99
23	1.00
24	1.00

Figure 3a: Silhouette Score by Cluster Number

# of Clusters	Silhouette Score
2	0.284
3	0.266
4	0.120
5	0.134
6	0.128
7	0.126
8	0.131
9	0.087

Figure 4a: Three-Means Clustering on Asthma Prevalence

Cluster	Count	Mean	STD	Min	25%	50%	75%	Max
0	47,404	0.103324	0.011741	0.050	0.095	0.103	0.111	0.156
1	12,643	0.121743	0.015067	0.076	0.111	0.121	0.131	0.207
2	7,321	0.103479	0.014190	0.061	0.094	0.102	0.112	0.169

Figure 5a: Top 10 Most Differing Features in Clusters

Feature	Variance
EP_MINRTY	841.4
EP_HISP	757.8
EP_AFAM	418.4
EP_POV150	151.2
EP_NOHSDP	90.5
EP_HBURD	89.0

EP_LIMENG	69.5
EP_UNINSUR	50.6
EP_NOVEH	49.1
EP_MUNIT	36.6

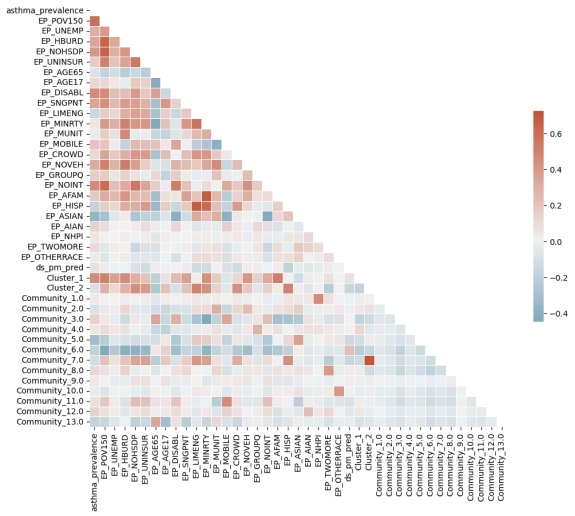
Figure 6a: Louvain Community Detection on Asthma Prevalence

Community	Count	Mean	STD	Min	25%	50%	75%	Max
0.0	8752.0	0.120407	0.015951	0.077	0.109	0.119	0.1310	0.207
1.0	1132.0	0.110575	0.011843	0.070	0.103	0.110	0.1180	0.150
2.0	2312.0	0.100660	0.016799	0.061	0.088	0.096	0.1110	0.159
3.0	9287.0	0.109944	0.009127	0.081	0.104	0.110	0.1160	0.150
4.0	2015.0	0.111907	0.016021	0.067	0.101	0.110	0.1210	0.171
5.0	4549.0	0.090202	0.010869	0.050	0.083	0.090	0.0970	0.134
6.0	11538.0	0.100663	0.008982	0.073	0.094	0.101	0.1070	0.132
7.0	4989.0	0.101159	0.012506	0.072	0.093	0.099	0.1080	0.153
8.0	9749.0	0.110229	0.012391	0.073	0.101	0.110	0.1190	0.163
9.0	135.0	0.117304	0.006968	0.103	0.113	0.117	0.1205	0.140
10.0	3055.0	0.105046	0.013206	0.073	0.096	0.104	0.1130	0.164
11.0	5303.0	0.109566	0.009827	0.078	0.103	0.109	0.1160	0.153
12.0	931.0	0.126750	0.014899	0.092	0.118	0.125	0.1340	0.189
13.0	3621.0	0.096555	0.010248	0.064	0.089	0.095	0.1030	0.132

Figure 7a: Top Most Differing Features Across Communities

Feature	Variance
EP_MINRTY	495.2
EP_MUNIT	293.0
EP_HISP	250.0
EP_AFAM	189.2
EP_NOVEH	103.5
EP_GROUPQ	91.0
EP_MOBILE	70.6
EP_POV150	70.3
EP_UNINSUR	65.2
EP_NOINT	52.0

Figure 8: Pairwise Correlation



Spearman correlation picked because it shows non-parametric correlations and handles ties well.

Figure 10a: Random Forest Feature Importance

Feature	Importance
EP_POV150	0.388
EP_ASIAN	0.089
EP_HISP	0.088
EP_AFAM	0.068
ds_pm_pred	0.053
EP_NOHSDP	0.040
EP_AGE65	0.029
EP_UNINSUR	0.027
EP_AIAN	0.024
EP_DISABL	0.021
EP_TWOMORE	0.019
EP_AGE17	0.018
EP_MUNIT	0.017
EP_HBURD	0.015
EP_NOVEH	0.015
EP_SNGPNT	0.014
EP_UNEMP	0.014
EP_NOINT	0.013
EP_MOBILE	0.012
EP_GROUPQ	0.011
EP_LIMENG	0.010
EP_CROWD	0.009
EP_OTHERRACE	0.005
EP_NHPI	0.004

Figure 12a: Learning Curve Values

Train Size	Train MAE	Val MAE
4,491	0.0017	0.0082
14,596	0.0020	0.0070
24,701	0.0020	0.0060
34,806	0.0021	0.0058
44,912	0.0021	0.0057

Figure 13a: SHAP on Largest Error

Feature	Value	SHAP Value
EP_GROUPQ	11.6751	0.0129
EP_AGE65	-1.9268	0.0022
EP_POV150	-1.0937	-0.0021
ds_pm_pred	-0.3652	0.0019
EP_HBURD	-2.1272	0.0019
EP_HISP	-0.4904	0.0015
EP_NOHSDP	-1.1809	0.0010
EP_DISABL	-1.6888	-0.0008
EP_MUNIT	-0.6479	0.0007
EP_NOINT	0.4938	0.0007
EP_AFAM	-0.3799	-0.0006
EP_UNINSUR	-0.8884	0.0004
EP_SNGPNT	-1.1258	-0.0003
EP_AGE17	-2.7377	0.0002
EP_TWOMORE	0.5481	0.0002
EP_AIAN	-0.1466	-0.0002

EP_ASIAN	2.8867	-0.0001
EP_NOVEH	0.6759	0.0001
EP_NHPI	-0.1491	-0.0001
EP_CROWD	-0.6810	0.0000
EP_UNEMP	-0.2415	0.0000
EP_MOBILE	-0.5602	0.0000
EP_OTHERRACE	-0.3558	0.0000
poverty_quartile	-1.3357	0.0000
EP_LIMENG	-0.3190	0.0000

Figure 14a: SHAP on Systemic Bias

Feature	Value	SHAP Value
EP_POV150	3.6250	0.0229
EP_AGE17	-3.0500	-0.0039
EP_GROUPQ	6.2349	-0.0037
EP_HBURD	-2.1272	-0.0028
ds_pm_pred	-0.2860	-0.0027
EP_NOVEH	-0.7199	-0.0023
EP_UNEMP	-1.1946	-0.0020
EP_HISP	-0.7432	0.0018
EP_AFAM	2.8826	0.0017
EP_AGE65	-2.0072	0.0015
EP_NOHSDP	4.5624	0.0014
EP_SNGPNT	-1.1258	-0.0013
EP_DISABL	-2.1996	-0.0011
EP_ASIAN	-0.5308	0.0011
EP_NOINT	-1.3645	-0.0009
EP_TWOMORE	-1.1158	-0.0007
EP_AIAN	-0.1466	-0.0003
EP_UNINSUR	-1.1744	-0.0002
EP_MUNIT	-0.6479	0.0001
EP_MOBILE	-0.5602	0.0001
EP_LIMENG	-0.5468	0.0000
EP_OTHERRACE	-0.3558	0.0000
EP_CROWD	-0.6810	0.0000
poverty_quartile	1.3450	0.0000
EP_NHPI	-0.1491	0.0000

Figure 15a: SHAP on Outlier/Edge Case

Feature	Value	SHAP Value
EP_NOHSDP	-1.1809	-0.0050
EP_MUNIT	4.1519	-0.0034
EP_POV150	-0.2716	-0.0032
EP_DISABL	-2.1996	-0.0020
EP_AGE17	-3.0500	-0.0015
EP_AGE65	-2.0072	0.0013
EP_HISP	-0.4601	0.0010
EP_NHPI	76.8484	0.0008
EP_SNGPNT	-1.1258	-0.0006
EP_UNINSUR	-1.1744	0.0006
EP_TWOMORE	-1.1158	-0.0005
EP_UNEMP	9.6365	0.0005

EP_AFAM	-0.6533	-0.0004
EP_NOINT	-1.3645	-0.0004
EP_ASIAN	0.0607	0.0004
EP_AIAN	-0.1466	-0.0003
ds_pm_pred	-0.2091	-0.0003
EP_HBURD	-2.1272	-0.0002
EP_GROUPQ	-0.2983	-0.0002
EP_LIMENG	-0.5468	0.0001
EP_NOVEH	6.4113	-0.0001
EP_OTHERRACE	-0.3558	0.0000
EP_CROWD	-0.6810	0.0000
EP_MOBILE	-0.5602	0.0000
poverty_quartile	-0.4421	0.0000

Figure 16: Task Distribution

Task	**Lead**	**Reviewer(s)**
Formation & Title	All	
Draft Proposal	Daida	Petrov, Ryuzaki
Data Collection & Cleaning	Daida	Petrov, Ryuzaki
Data Exploration	Petrov	Ryuzaki, Daida
Unsupervised Learning	Petrov	Ryuzaki, Daida
Supervised Learning	Ryuzaki	Daida, Petrov
Visualizations	Ryuzaki	Daida, Petrov
Failure Analysis & Advanced Model Evaluation	Daida	Petrov, Ryuzaki
Project Write-Up	Daida	Petrov, Ryuzaki
Github	Ryuzaki	Daida, Petrov