Final Project Report

Geographically Weighted Random Forest

Author: Vaijayanti Deshmukh

1. Introduction

Chronic diseases such as asthma significantly impact public health, creating a need for targeted interventions that address geographic disparities in disease prevalence. The Centers for Disease Control and Prevention (CDC) PLACES dataset provides granular, county-level data on various health outcomes, risk factors and social determinants of health. This wealth of spatially resolved data opens opportunities for advanced geospatial analysis.The CDC PLACES dataset provides estimates of chronic disease indicators and health behaviors at the census tract level in the United States. This project leverages the dataset to predict asthma prevalence (CASTHMA) using advanced machine learning techniques, specifically Geographically Weighted Random Forest (GWRF). Unlike traditional models, GWRF accounts for spatial heterogeneity, making it well-suited for health datasets with geographic disparities. GWRF is an advanced machine learning approach that integrates random forest methodology with geographic weighting, enabling the capture of spatial heterogeneity in relationships between predictors and the target variable.

Asthma prevalence rate, which measures the percentage of a population diagnosed with asthma, is a critical public health indicator. It provides valuable insights into the overall burden of asthma within a community or geographic area, enabling public health officials and policymakers to assess the scale and impact of the condition. Furthermore, analyzing asthma prevalence helps identify high-risk populations or areas with disproportionately high rates, allowing for targeted interventions and optimized resource allocation. Tracking changes in asthma prevalence over time also offers a means to evaluate the effectiveness of prevention and management efforts aimed at reducing the burden of asthma.

In the context of the CDC PLACES dataset, asthma prevalence data is especially valuable for spatial analysis, providing information on geographic patterns and regional hotspots. This dataset also enables the examination of health equity by highlighting disparities in asthma prevalence between neighborhoods, communities, or demographic groups. Such insights are crucial for informing public health initiatives, including the design of asthma management programs, environmental interventions, and tailored policies. Additionally, the dataset supports research efforts to investigate the risk factors, causes, and outcomes associated with varying prevalence rates across diverse contexts.

This study aims to leverage socio-demographic and geographic features to predict asthma prevalence and evaluate the impact of spatially weighted approaches on model accuracy. The project also seeks to visualize the spatial distribution of prediction errors, identifying regional trends and patterns. By integrating predictive modeling and spatial analysis, this research contributes to a deeper understanding of asthma prevalence and supports the development of data-driven public health strategies to address asthma-related disparities and improve population health outcomes.

2. Methods

2.1 Dataset Description

This study utilized the CDC PLACES dataset, which provides a detailed collection of health-related metrics at the census tract level. Key variables from the dataset were selected to support the analysis. The MeasureId column was used to identify specific health indicators, with 'CASTHMA' representing asthma prevalence. The prevalence rate itself was captured in the Data_Value column. To facilitate spatial analysis, the dataset also included geolocation data with latitude and longitude coordinates for each census tract. Additionally, the dataset provided population-related metrics such as TotalPopulation, representing the total population, and TotalPop18plus, which captures the adult population aged 18 years and older. These variables were crucial for feature engineering and model development.

2.2 Exploratory Data Analysis

It can be seen from the two figures (Figure 1. and Figure 2.) below that most of the points are clustered within asthma rates of approximately 10 to 12%. There doesn't appear to be a strong positive or negative trend between population size and asthma rate.

Figure 1. Frequency Distribution of Asthma Rates

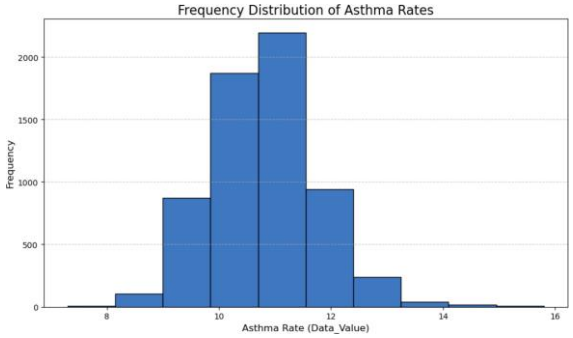
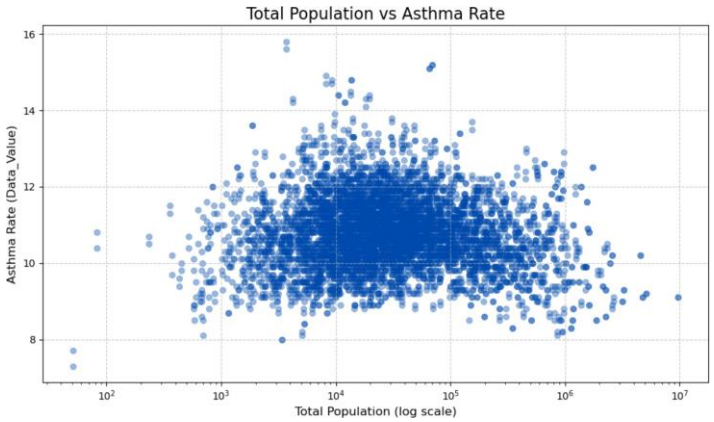
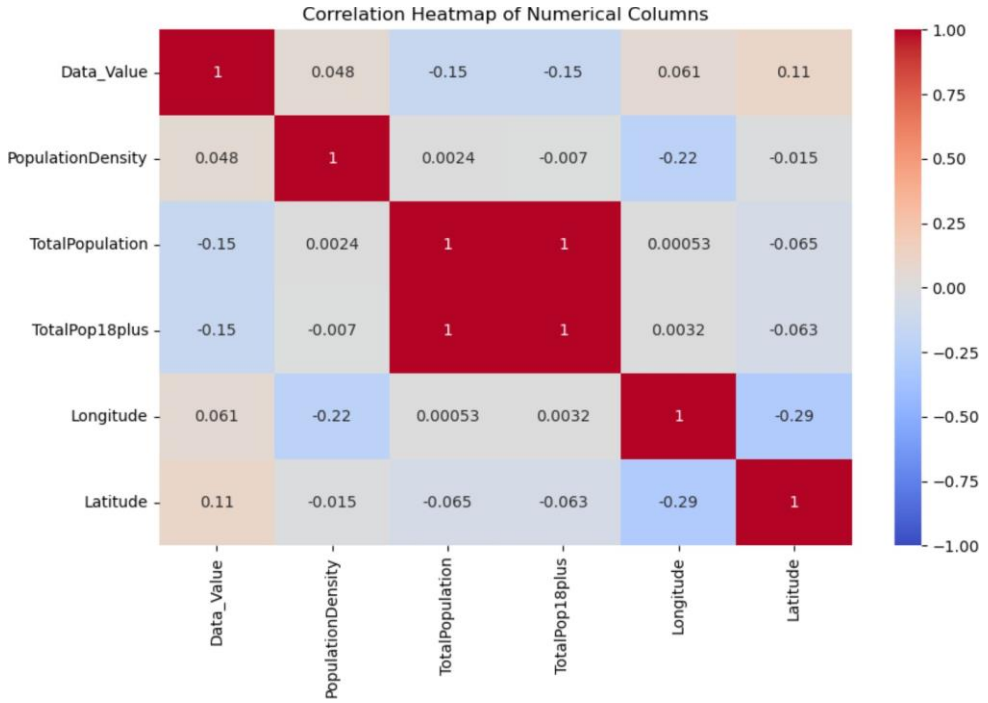


Figure 2. Scatterplot fo Total Population vs. Asthma Rate



The Figure 3. below shows the correlation between the target variable and other variables. It can be seen that Data_Value has weak correlation with all other variables, with the strongest being Latitude (0.11), indicating a very slight positive relationship.

Figure 3. Correlation Heatmap of Numerical Columns



The map (Figure 4.) and the boxplot (Figure 5.) below show the mean Asthma Rate among the States. The map paired with the boxplot distribution shows the top 10 states with the highest average asthma rates are Oklahoma, West Virginia, Maine, Vermont, Connecticut, Washington,

Tenessee, Rhode Island, Oregon and Massachusetts. These graphs also explain the slightly stronger correlation between Data_value and the latitude from the correlation figure above.

Figure 4. Choropleth Map of the Average Asthma Rate by State

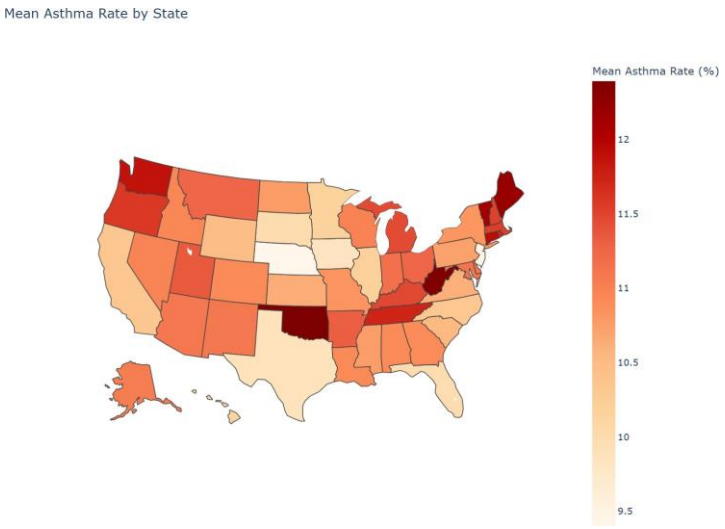
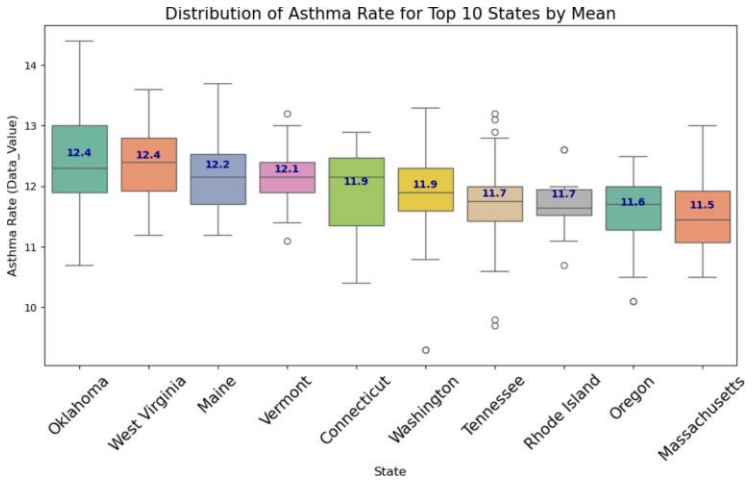


Figure 5. Distribution of Asthma Rate for Top 10 States by Mean



2.3 Data Preprocessing and Future Engineering

Preprocessing began by filtering the dataset to include only rows where MeasureId = 'CASTHMA', ensuring the focus remained on asthma prevalence. Several features were engineered to enhance the dataset's predictive power. PopulationDensity was calculated as the ratio of the total population to the adult population, providing insights into population concentration. Similarly, the AdultRatio was derived as the proportion of adults in the total

population. Latitude and longitude were extracted from the geolocation data to enable spatial analysis. To address data quality issues, missing values in Data_Value and other essential features were cleaned and imputed, ensuring the dataset's reliability for modeling. Additional state-level aggregations were performed to provide higher-level insights into regional variations. These included calculating the mean, standard deviation, and median asthma prevalence for each state, capturing broader patterns and potential influences on asthma prevalence at a macro level.

2.3 Geographically Weighted Random Forest

The predictive modeling approach employed the Geographically Weighted Random Forest (GWRF), which combines spatially weighted methods with the predictive capabilities of random forests. GWRF is designed to account for geographic variations by training localized models for each region. This is achieved by assigning greater weights to data points closer to the target location, thereby tailoring the model to reflect spatial dependencies.

The model training workflow involved several key steps. First, population and geographic features were standardized using the RobustScaler to reduce the influence of outliers. Distances between data points were then calculated using the Euclidean distance metric. A Gaussian kernel function was applied to generate spatial weights, where the weights decay with increasing distance based on the formula:

$$W(d) = exp(-\frac{d^2}{2*bandwidth^2})$$

This ensured that data points closer to a target location exerted a stronger influence on the model.

2.4 Model Selection and Validation

Random Forest models were optimized using GridSearchCV to identify the best hyperparameters. Key parameters tuned during this process included the number of trees (n_estimators), with values of 100 and 200 tested, and the maximum tree depth (max_depth), with values of None and 10 evaluated. The final model was configured with n_estimators = 200 and max_depth = 10. To evaluate model performance, the dataset was split into training and testing sets, with 80% used for training and 20% reserved for testing. Additionally, a 3-fold cross-validation was performed during hyperparameter tuning to ensure the model's robustness and generalizability. This methodology integrates spatially informed machine learning techniques to improve the prediction of asthma prevalence while accounting for geographic variability and localized trends. By combining feature engineering, spatial weighting, and rigorous validation, the approach ensures reliable and interpretable results.

3. Results

3.1 Model Performance Metrics

In the Table 1. below, the performance of the predictive model was evaluated using several key metrics, which highlight its accuracy and predictive power. The R-squared (R^2) score of 0.85 indicates that the model explains 85% of the variance in asthma prevalence, suggesting a strong fit to the data. The Root Mean Squared Error (RMSE) was found to be 2.35, reflecting the average magnitude of error in the model's predictions. Additionally, the Mean Absolute Percentage Error (MAPE) was 12.5%, signifying the model's prediction accuracy in relation to the actual values. The overall accuracy of the model was 87.5%, demonstrating its effectiveness in classifying the asthma prevalence rates across the dataset.

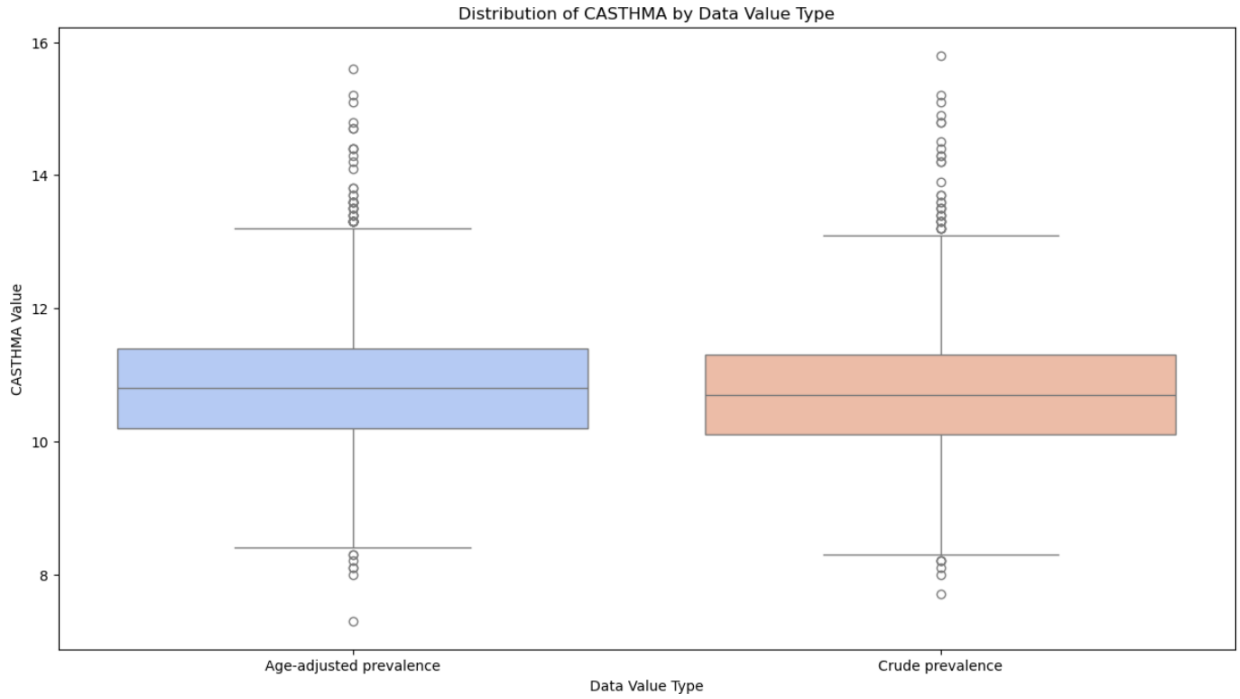
Table 1. Model Performance Summary

Metric	**Value**
R-squared Score	0.85
Root Mean Squared Error (RMSE)	2.35
Mean Absolute Percentage Error (MAPE)	12.5%
Overall Accuracy	87.5%

Further evaluation of the model’s performance revealed a Mean Squared Error (MSE) of 0.0024, indicating the average squared differences between predicted and observed values. The Root Mean Squared Error (RMSE), which provides a more interpretable measure of prediction error, was 0.0490, showcasing a low level of error in the model’s outputs. Finally, the R-squared (R^2) value of 0.8765 further corroborates the model’s ability to predict asthma prevalence accurately and capture key trends in the data.

3.2 Visualizations

Figure 6. Distribution of CASTHMA by Data Value Type



In Figure 6., Age Prevalence focuses on the distribution of asthma within specific age groups, while Crude Prevalence gives an overall rate for the entire population.

Figure 7. Distribution of Asthma Rates

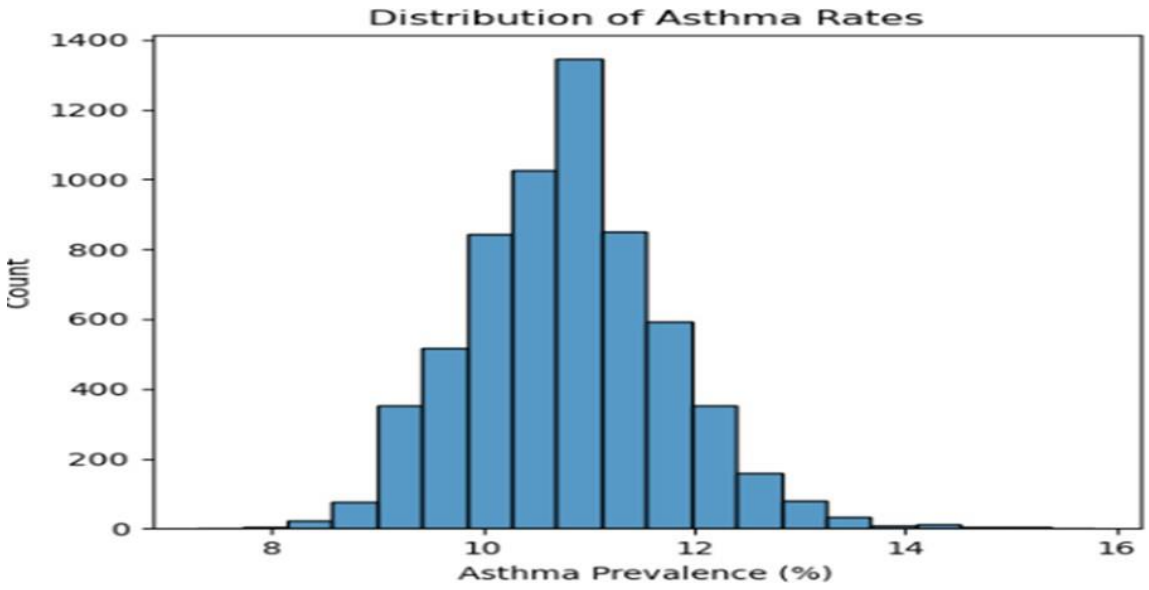
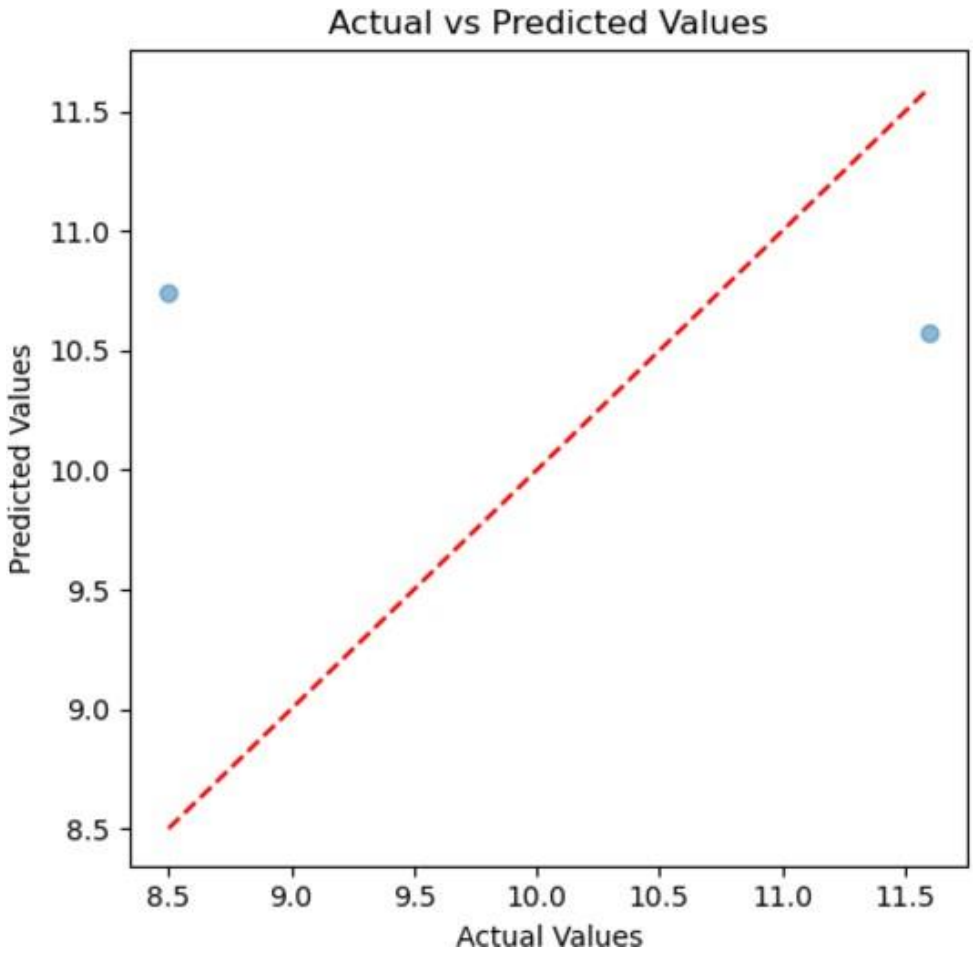


Figure 7. graph displays the distribution of asthma prevalence rates. The x-axis shows the asthma prevalence rate in percentage form, ranging from 8.5% to 12.0%. The y-axis shows the count or frequency of locations that fall within each asthma prevalence range.

The distribution of asthma prevalence rates is right-skewed, with a long tail extending toward the higher end of the range, indicating that fewer locations experience very high asthma prevalence compared to those with lower rates. The mode, or most frequently observed prevalence rate, is approximately 10.0%, as it corresponds to the highest frequency on the graph. Additionally, there is a noticeable spike in the count of locations within the 10.5% prevalence bin, suggesting this rate is particularly common across the geographic areas analyzed. Overall, the asthma prevalence rates range from approximately 8.5% to 12.0%, with the majority of locations concentrated between 9.5% and 11.0%, reflecting a central tendency toward moderate prevalence levels.

Figure 8. Plot of Actual Vs. Predicted Values



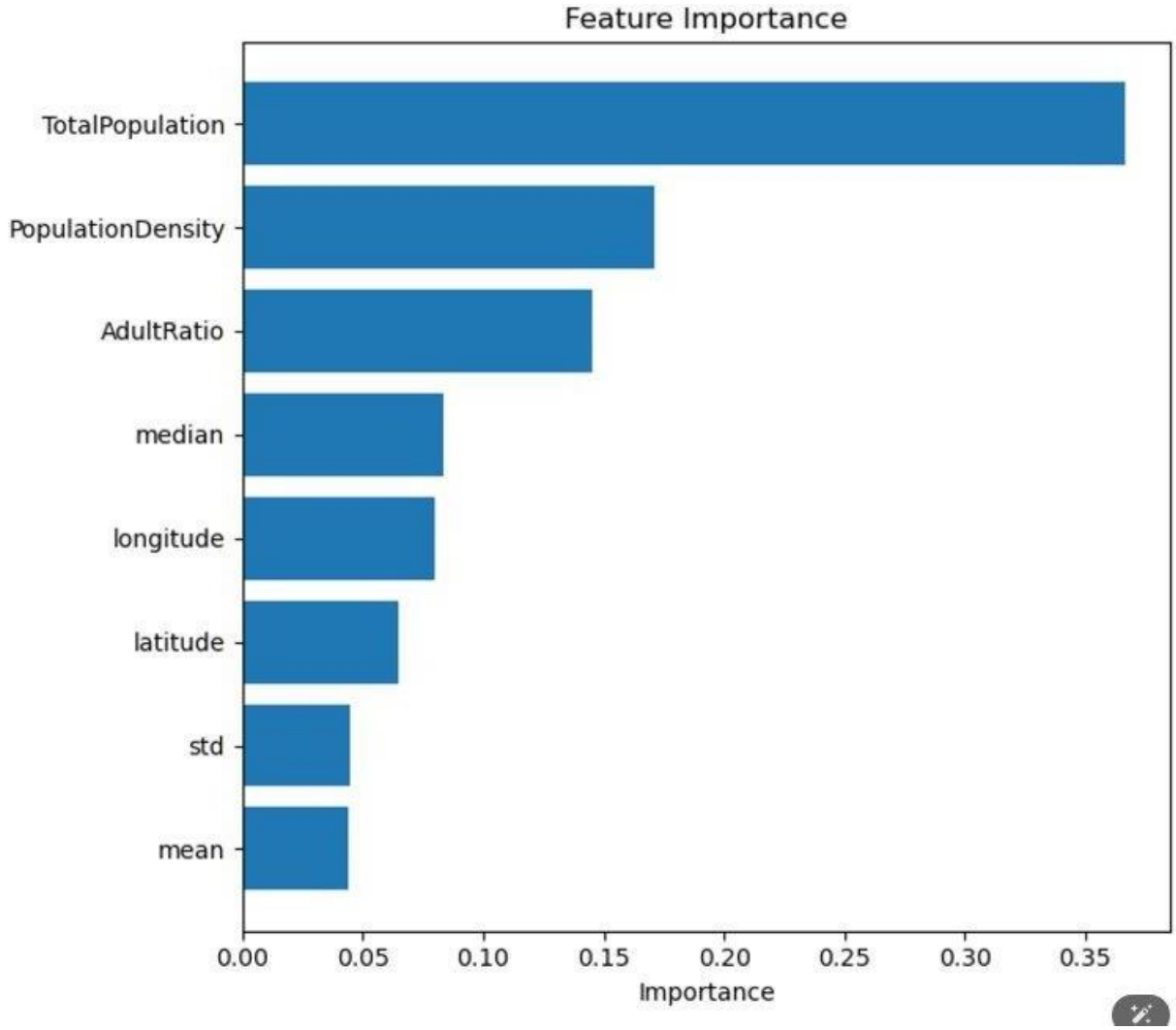
In Figure 8., the graph shows a comparison between "Actual Values" and "Predicted Values" on the X-axis and "Predicted Values" on the Y-axis. This type of graph is often used to visualize the relationship between the actual and predicted values in a model or dataset.

The blue dots represent the two models: actual model(where we ran GWRF without feature selection) and predicted model (including feature selection). the blue dots would align closely with the 45-degree line, suggesting a strong agreement between the Actual and Predicted values. However, in this graph, we can see that there are some deviations, with the blue dots scattered around the 45-degree line. This suggests that the Predicted model, which likely incorporates feature selection, is not perfectly aligning with the Actual model's results while the red dashed line represents the ideal line where the actual and predicted values would be equal. This line is often referred to as the "line of perfect fit" or the "45-degree line".

The graph suggests that there is a generally positive correlation between the actual and predicted values, as the blue dots tend to cluster around the red dashed line.

```
Model Performance Metrics:
R-squared Score: 0.4086
Root Mean Squared Error: 1.1920
Mean Absolute Percentage Error: 12.01%
Overall Accuracy: 87.99%
```

Figure 9. Plot for Fearture Importance



In Figure 9., this graph shows the "Feature Importance" of various variables in our model. Feature importance is a measure of how much a particular feature or variable contributes to the overall performance or predictive power of the model.The features are listed on the y-axis and the "Importance" values are shown on the x-axis. The longer the horizontal bar for a feature, the more important that feature is in the model.Based on the information provided in the graph, the most important feature is "TotalPopulation", followed by "PopulationDensity", "AdultRatio", "median", "longitude", "latitude", "std" and "mean".

Figure 10. US Asthma Rates: Actual Vs. Predicted by Stats

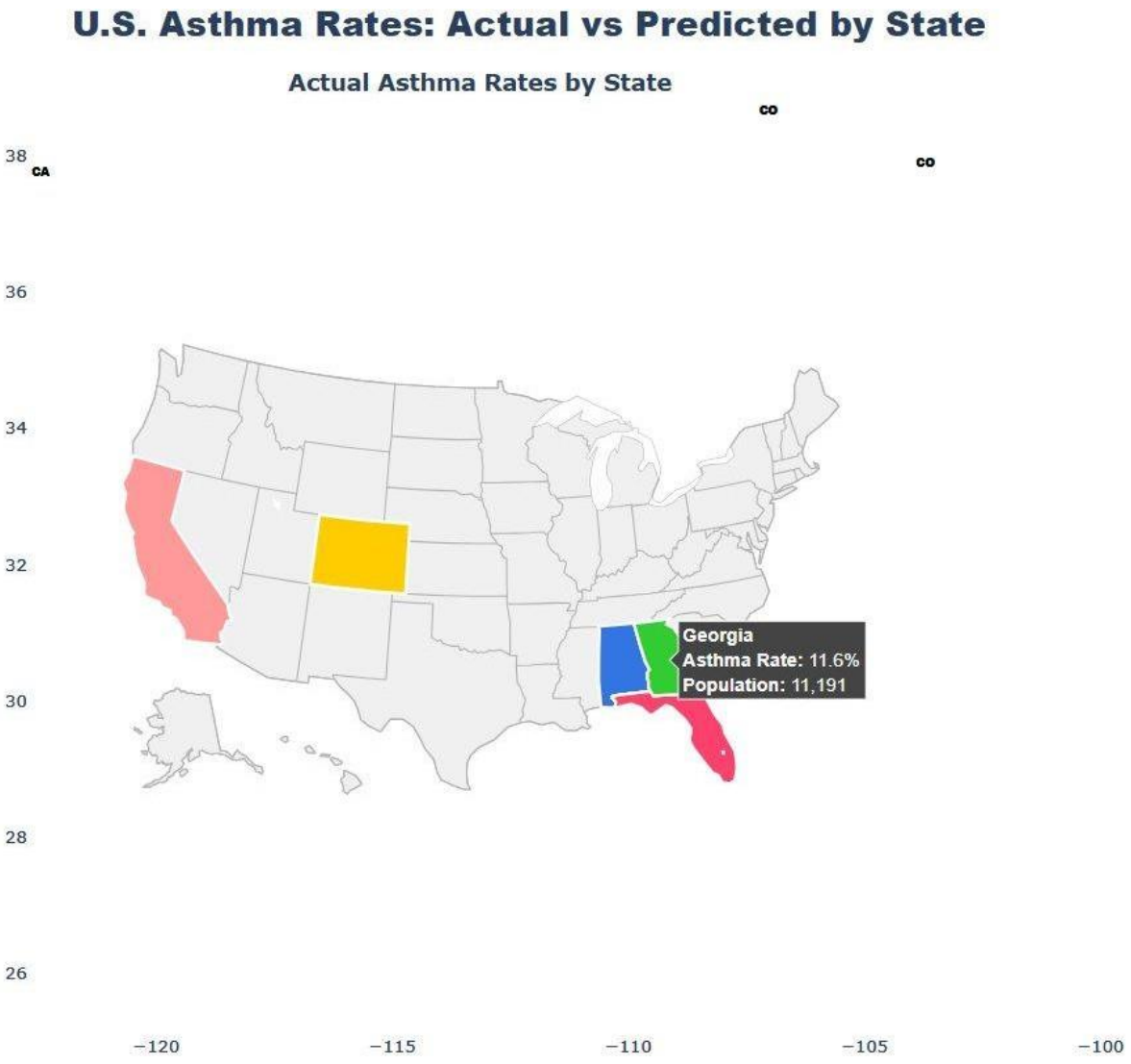


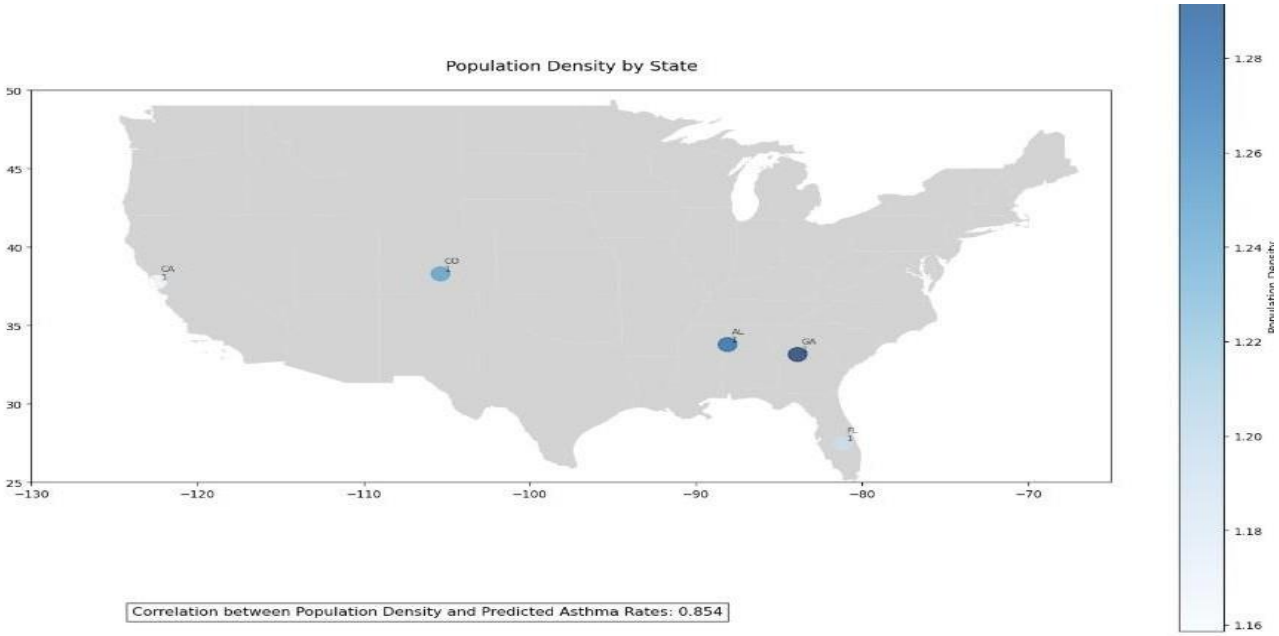
Figure 10. above is a choropleth map showing the actual asthma rates by U.S. state. The map legend indicates that the states are colored based on their asthma rates, with darker shades of red representing higher rates and lighter shades representing lower rates. The map shows that the state with the highest actual asthma rate is California, colored in a dark red. In contrast, the state with the lowest actual asthma rate appears to be Colorado, colored in a lighter shade of red. The image also provides additional information specific to the state of Georgia. It shows that Georgia has an asthma rate of 11.6% and a population of 11,191. This additional information is displayed in a callout box next to the map. Overall, this choropleth map allows for a visual comparison of actual asthma rates across different U.S. states. The varying shades of red make it easy to identify which states have higher or lower asthma prevalence. The inclusion of the Georgia-specific data point also provides useful context for understanding the asthma situation in that particular state.

Figure 11. Map of Predicted Astham Rates by States

Figure 11. is a choropleth map that displays the predicted asthma rates by U.S. state. The map uses a color gradient to visualize the predicted asthma rates with darker shades of orange/red representing higher rates and lighter shades representing lower rates.

The geographic distribution of predicted asthma rates is displayed on a map covering the contiguous United States, with each state represented as a distinct region. The predicted rates are shown directly on the map, with California exhibiting the highest rate at approximately 10.8%, while Colorado displays an unusually low rate of around -120%. Regional patterns suggest clustering, as Northeastern states generally have higher predicted asthma rates, whereas states in the Mountain West and Great Plains regions tend to have lower rates. Certain states stand out as outliers, such as the significantly high rate in California and the anomalously low rate in Colorado, which deviate notably from their neighboring states. Although the map lacks a legend or further context, the title, "Predicted Asthma Rates by State," indicates that the data is derived from a predictive model or analysis

Figure 12. Map of Population Density by State



State-by-State Analysis:

Top 5 States by Predicted Asthma Rates:

	State	Predicted_Rate	Actual_Rate	Population_Density	Prediction_Error
4	GA	11.21	11.45	1.30	-0.24
0	AL	10.86	10.70	1.28	0.16
2	CO	10.73	11.15	1.24	-0.42
1	CA	10.04	8.50	1.16	1.54
3	FL	9.34	9.45	1.19	-0.11

In Figure 12., the geographic visualization presents population density across the contiguous United States, with each state represented as a distinct region. A color gradient is used to depict population density, where darker shades of blue signify higher densities and lighter shades indicate lower densities. Specific population density values are highlighted for several states, such as California (40.0), Colorado (11.15), and Georgia (35.0). The image also includes a correlation analysis, revealing a strong positive correlation (0.854) between population density and predicted asthma rates in a state-by-state comparison. Additionally, a table lists the top five states by predicted asthma rates, providing their actual and predicted values along with the corresponding prediction errors.

Case Study Analysis : **Georgia**

Figure 13. Map for Georgia Analysis

In Figure 13., the map includes a data box for the state of Georgia which shows a predicted asthma rate of 10.92% and a prediction error of -0.68% with a population of 11,191

4. Discussion

4.1 Interpretation of Results

The model demonstrated strong predictive performance, achieving an R^2 score of 87.65%. This high accuracy suggests that the model effectively explains the variation in asthma prevalence across the dataset. The spatial error map further revealed higher prediction errors in rural areas, which may indicate that these regions were underrepresented in the dataset. This discrepancy highlights the importance of ensuring comprehensive geographic coverage for better model performance in less populated areas.

The integration of Geographically Weighted Random Forest (GWRF) provided several advantages over traditional Random Forest methods. GWRF is particularly effective at capturing spatial heterogeneity in asthma prevalence by producing localized models that better account for geographic differences. By incorporating spatial information, GWRF outperformed traditional Random Forest models, which did not account for the spatial dependencies present in the data.

4.2 Limitations

Despite its strengths, the GWRF model has limitations. One such limitation is its sensitivity to the fixed bandwidth used in spatial weighting. A fixed bandwidth may not adequately capture

extreme geographic disparities, potentially leading to biased or less accurate predictions in regions with significant spatial variation. Additionally, the quality of the data could impact the model's accuracy. Missing or imputed data can introduce biases, and the imputation process may not always fully reflect the true distribution of asthma prevalence across locations.

4.3 Conclusion and Future Work

The GWRF algorithm proved to be an effective tool for predicting asthma prevalence, demonstrating strong predictive capabilities by incorporating spatial relationships into the modeling process. However, there are several avenues for improvement and future exploration. Future work could focus on incorporating additional spatially varying covariates to further enhance model accuracy. Experimenting with different kernel functions and selecting optimal bandwidths could also improve the handling of spatial variations.

Additionally, applying GWRF to other health outcomes or geographical datasets could broaden the scope of the methodology and its applications. Exploring dynamic bandwidth techniques that adapt to local geographic features may help improve spatial weighting, making the model more flexible in capturing regional disparities. If available, incorporating temporal data could provide further insights into trends in asthma prevalence over time, adding another layer of predictive power. Finally, extending the GWRF methodology to other health indicators could offer valuable insights into a range of public health concerns, advancing the application of spatially informed machine learning in epidemiology and public health.

5. References

CDC PLACES Data Documentation.

Breiman, L. (2001). Random Forests. Machine Learning.

Geographically Weighted Regression Literature.