Impact of Air Pollution on Respiratory Health Across US Counties

E Krishna Koushik-23MIA1107, R Amarender Reddy-23MIA1012, T Guru Raghav Raj-23MIA1105 | Pattabiraman V | SCOPE

Motivation/Introduction

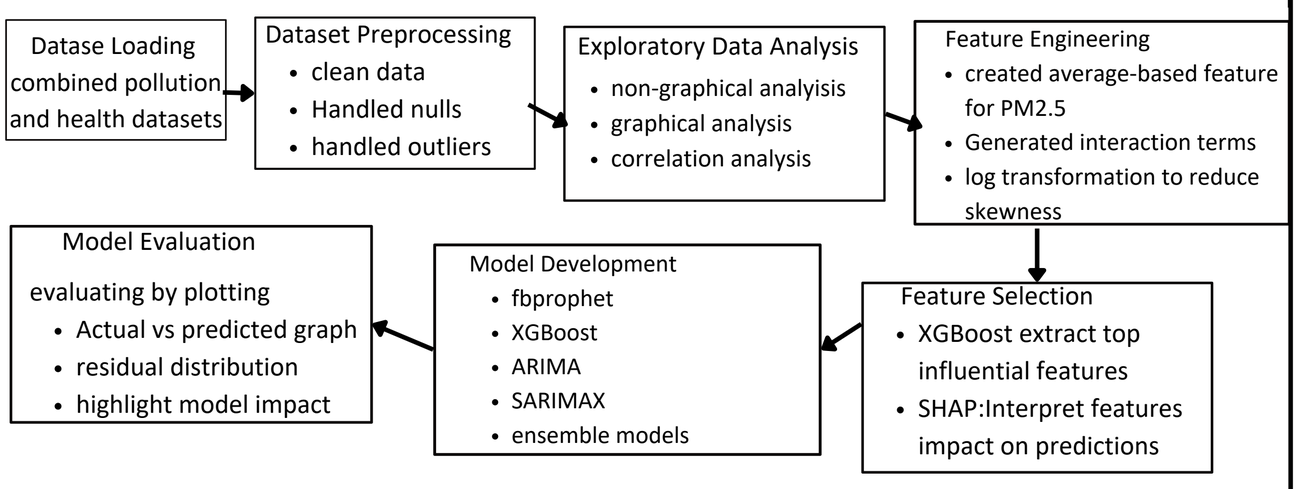
Air pollution has become a pressing concern across the US, particularly due to its impact on public health. Among the most affected are individuals with pre-existing respiratory conditions such as asthma and COPD. This study explores how pollutant levels correlate with respiratory health trends using predictive modeling and feature-based analysis.

SCOPE of the Project

This project focuses on:

- Exploring the link between pollution levels and asthma/COPD rates
- Performing data cleaning, preprocessing and feature engineering
- Applying statistical and ML-based models to identify trends
- Evaluating models using accuracy metrics and visualisations
- Forecasting asthma rates for public health insights

Methodology



Feature Engineering

- Introduced average-based PM2.5 levels[e.g:weekly/monthly averages]
- Created interaction terms [e.g.: Pollution Index . COPD_Rate, PM25 and Asthma Rate]
- Applied log transformation on shared variables for: 1. Asthma Rate 2. COPD Rate 3.Pollution Index 4. PM25

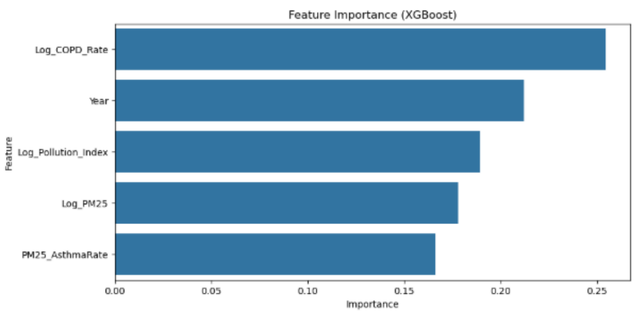


Fig01:Feature importance using XGBoost

SHAP summary plot shows the impacts of each feature on the model's prediction "PM2.5 levels & asthma rates have the strongest positive influence on the predicted health impact"

Feature Selection

- Checked collinearity & skewness of features
- Performed log transformations & created interaction terms
- Final selected features: Year, PM25_30DAY_AVG, PM25_AsthmaRate, Log_PM25, Log_Pollution_Index, Log_COPD_Rate
- Used XGBoost to validate feature importance after validating we removed PM25_30DAY_AVG

- Used XGBoost for feature importance
- Top prediction: Log_COPD_Rate,Year,Log_Pollution_Index
- Log and interaction features improved model performance
- By the insight from this we removed PM25_30DAY_AVG

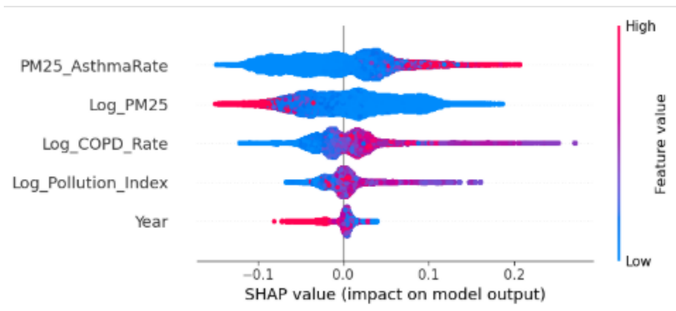


Fig02:Feature impact on target variable

Modelling Techniques

- Tried ARIMA, SARIMAX, and fbprophet.
- ARIMA and SARIMAX performed poorly ($R^2 \approx -20$).
- fbprophet also performed poorly ($R^2 \approx 10.2$).
- Tried XGBoost performed very well ($R^2 \approx 0.9802$).
- Implemented ensemble models: Bagging, Boosting, Voting, Stacking.

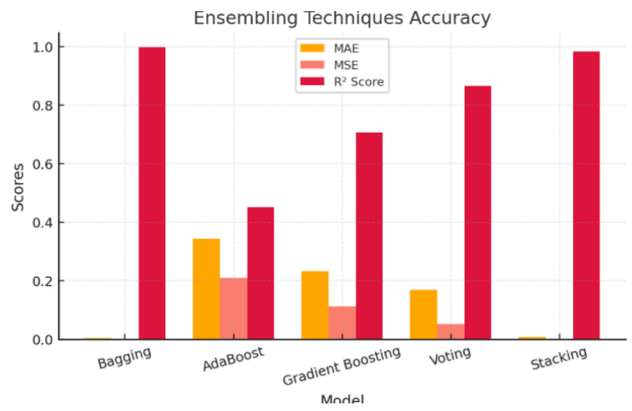
Evaluation Metrics

- R^2 Score, MAE, Residuals Analysis.
- Visualizations: Actual vs Predicted, Residual Plot, LOESS Smoothed Plot.

Result

Table-1 Ensampling Techniques

Model	MAE	MSE	R^2 Score
Bagging	[0.0037]	[0.0013]	[0.9967]
AdaBoost	[0.3440]	[0.2095]	[0.4507]
Gradient Boosting	[0.2337]	[0.1120]	[0.7064]
Voting	[0.1694]	[0.0516]	[0.8648]
Stacking	[0.0073]	[0.0001]	[0.9835]



- Bagging and Stacking outperform other methods, with very low errors and R^2 scores close to 1.
- AdaBoost performs the worst among the ensemble techniques based on both error metrics and R^2.

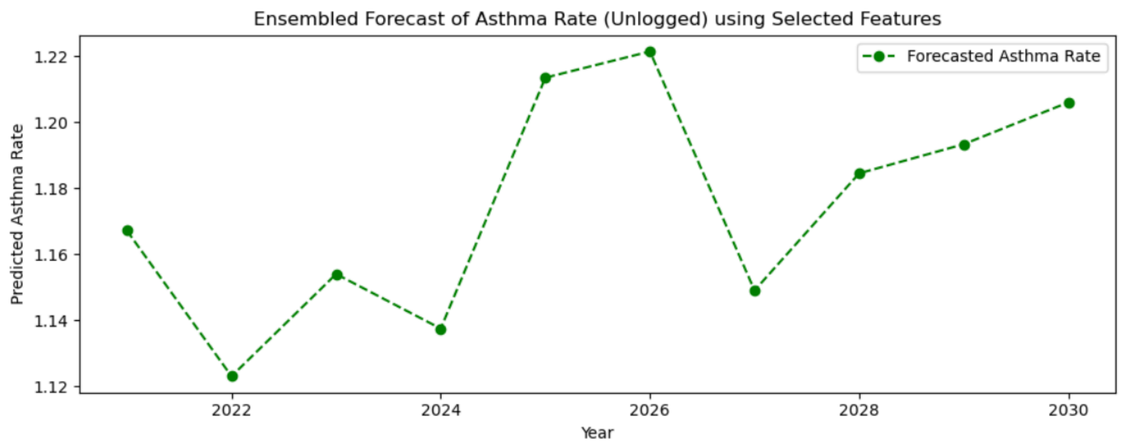


Fig03:Future Prediction of Asthama Rate

The ensemble model shows reliable performance with low prediction error (MAE ~0.007), enabling confident asthma rate forecasting. A significant spike is observed around 2025–2026, possibly indicating environmental or policy shifts. Although a dip occurs post-2026, the rate steadily rises again, suggesting a potential long-term health risk trend. This pattern emphasizes the importance of sustained intervention in pollution control and health planning.

Table -2 Accuracy for models

Model	MAE	MSE	R^2 Score
Prophet	[0.0600]	[0.0800]	[0.1063]
XGBoost	[0.0080]	[0.0002]	[0.9802]
ARIMA	[0.2530]	[0.0672]	[-20.0400]
SARIMAX	[0.2559]	[0.0684]	[-20.4100]

XGBoost is the Best – It achieves the lowest MAE (0.0080), MSE (0.00016), and highest R^2 (0.98), indicating highly accurate predictions.
Prophet is Consistent but Weak – Its MAE (0.06) and MSE (0.08) are improved, but a low R^2 (0.1063) suggests it explains little variance.
ARIMA Performs Poorly – With high error values and a negative R^2 (-20.04), ARIMA struggles to fit the data well, possibly due to overfitting.
SARIMAX is Similar to ARIMA – It shows nearly identical error rates and a negative R^2 (-20.41), offering no improvement over ARIMA.

XGBoost is the most effective with highly accurate predictions. Prophet is consistent but weak in explaining variance. ARIMA and SARIMAX both struggle with poor performance and high errors.

Conclusion/Summary

In this study, we developed a hybrid model for analyzing the relationship between air pollution and respiratory health by merging six years of pollution datasets with a dedicated respiratory health dataset. Unlike many studies that rely on single-source or Kaggle-based datasets, our approach integrated real-world, multi-year environmental data with health records to uncover deeper patterns. By applying machine learning techniques — including ensemble methods such as bagging, boosting, voting, and stacking — we were able to enhance predictive performance and provide meaningful insights. The results demonstrate the strength of combining environmental and health data for improved prediction and highlight the potential of ensemble models in supporting public health and pollution control strategies.

Contact Details

elluru.krishna2023@vitstudent.ac.in,rayini.amarender2023@vitstudent.ac.in, talari.guru2023@vitstudent.ac.in

Acknowledgements/References

Link: https://github.com/krishnakoushik2792005/Impact-of-Air-Pollution-on-Respiratory-Health-Across-US-Cities