Predicting Asthma Attacks by Utilizing Machine Learning

John Pascoe

## **Introduction**

Asthma stands as a major chronic respiratory condition affecting millions globally, marked by episodes of airway inflammation and bronchoconstriction. These episodes manifest through distressing symptoms such as wheezing, shortness of breath, chest tightness, and persistent coughing, significantly impairing daily activities and overall quality of life. The ability to predict these exacerbations is crucial as it enables the implementation of preemptive management strategies that not only reduce the frequency of emergency hospital visits but also tailor treatment plans to individual needs, thereby enhancing patient outcomes. As a prevalent health issue, asthma is a significant contributor to global morbidity and extensive healthcare utilization, driving the need for innovative approaches to manage and mitigate its impact effectively.

Early and accurate prediction of asthma exacerbations plays a pivotal role in improving disease management and controlling its progression, which can substantially decrease the physical and economic burden of the disease. Machine learning emerges as a powerful tool in this context, offering the capability to uncover complex patterns and relationships within large datasets that traditional statistical methods might overlook. A report following the same vein as this project was utilized to understand the relationship of machine learning to asthma exacerbations [1]. This project was centered around developing a predictive model capable of assessing the likelihood of asthma attacks across diverse populations. By synthesizing data that mirrors real-world conditions, the project aimed to delve deeper into understanding how various environmental, demographic, and behavioral factors converge to influence the risk of asthma exacerbations. This model seeks to provide actionable insights that could lead to more personalized, effective, and proactive asthma care strategies.

## **Methodology**

In the development of the predictive model for asthma exacerbations, a robust dataset was synthesized, integrating multiple data sources to ensure comprehensive coverage of factors influencing asthma attacks. This synthesis involved combining regional Air Quality Index (AQI) data from the Environmental Protection Agency (EPA) with demographic health data from the Centers for Disease Control and Prevention (CDC) and the National Health Interview Survey (NHIS) [2,3,4]. The AQI data provided baseline environmental risk factors by supplying average air pollution levels for distinct U.S. regions—Northeast, Southeast, Northwest, and Southwest. Concurrently, CDC data contributed insights into asthma attack prevalence across diverse demographic groups, differentiated by age, gender, and race, while NHIS data offered additional context regarding the frequency of rescue and preventative inhaler usage among these populations [2,3,4].

To achieve a realistic simulation of environmental and health variables, AQI values were statistically generated to mirror actual conditions, using region-specific means and standard deviations [2]. This method ensured that the synthetic dataset reflected true-to-life variability and could effectively support the modeling of asthma attack likelihoods across different U.S. regions. Demographic characteristics assigned to individuals within the dataset matched the actual distribution reported by the CDC, allowing for the precise calibration of the baseline asthma attack probability to approximately 40.7%, aligning with the figures reported in 2020 [3]. This value would vary a negligible amount based on the size of the population created, with a population size of 200,000 and 1,000,000 being utilized as synthesized datasets.

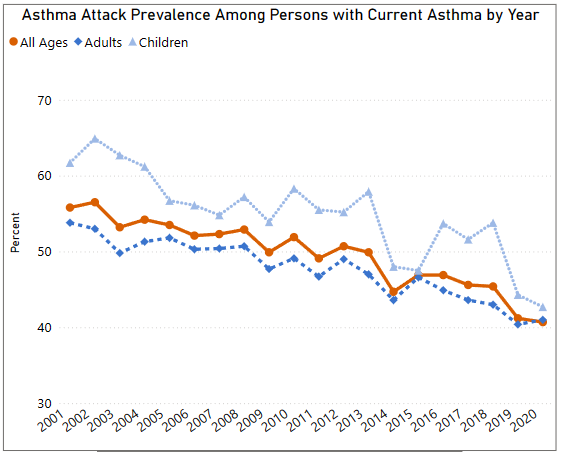


Figure 1: CDC data of recent attacks experienced by those with asthma [3].

The methodology for predictive modeling employed advanced machine learning techniques. A key component of the data preprocessing stage was the application of Synthetic Minority Oversampling Techniques (SMOTE), which addressed class imbalance within the dataset (such as with the non-even split of asthmatics who have and haven’t had a recent attack), ensuring no bias towards any particular outcome in model training. For the core analysis, an XGBoost classifier was selected for its efficacy in handling sparse data and its superior performance over traditional models like the RandomForest Classifier, which served as an initial comparison baseline. The use of XGBoost was predicated on its ability to efficiently manage diverse datasets and improve over time with parameter tuning.

The model's architecture was embedded within a cross-validated framework to bolster its robustness and generalizability across unseen data. This approach ensured that the model was not only trained but also validated across multiple subsets of data, preventing overfitting and promoting a reliable estimation of its real-world applicability. Performance evaluation metrics such as F1 scores, precision, and recall for both "Yes" (attack predicted) and "No" (no attack predicted) categories were meticulously calculated to assess the effectiveness of the model. These metrics provided a detailed view of the model’s ability to accurately classify and predict instances of asthma exacerbations, which is critical for its potential deployment in clinical settings where accurate predictions could significantly influence patient outcomes and healthcare strategies.

Table 1 shows the variable assigned to each individual person within the synthesized population.

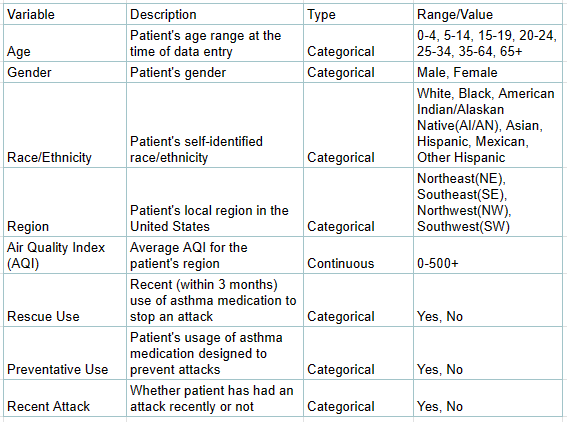


Table 1. Variables utilized to create a distinct population

## **Results**

The optimized XGBoost model showed promising results. For the “Yes” category of recent attacks, it achieved a precision of 69%, a recall of 74%, and a f1-score of 71%. For the “No” category, the model achieved a precision of 81%, a recall of 78%, and a f1-score of 79%. The overall accuracy of the model turned out to be 76%. These values can be seen in the output of testing shown in Figure 1.

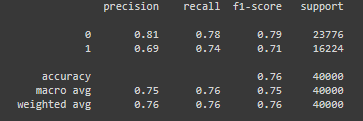


Figure 2: The test output after training with a complete population size of 200,000.

These metrics indicate that the created model is not only reliable, but also effective in distinguishing between individuals who are likely and unlikely to experience an asthma attack.

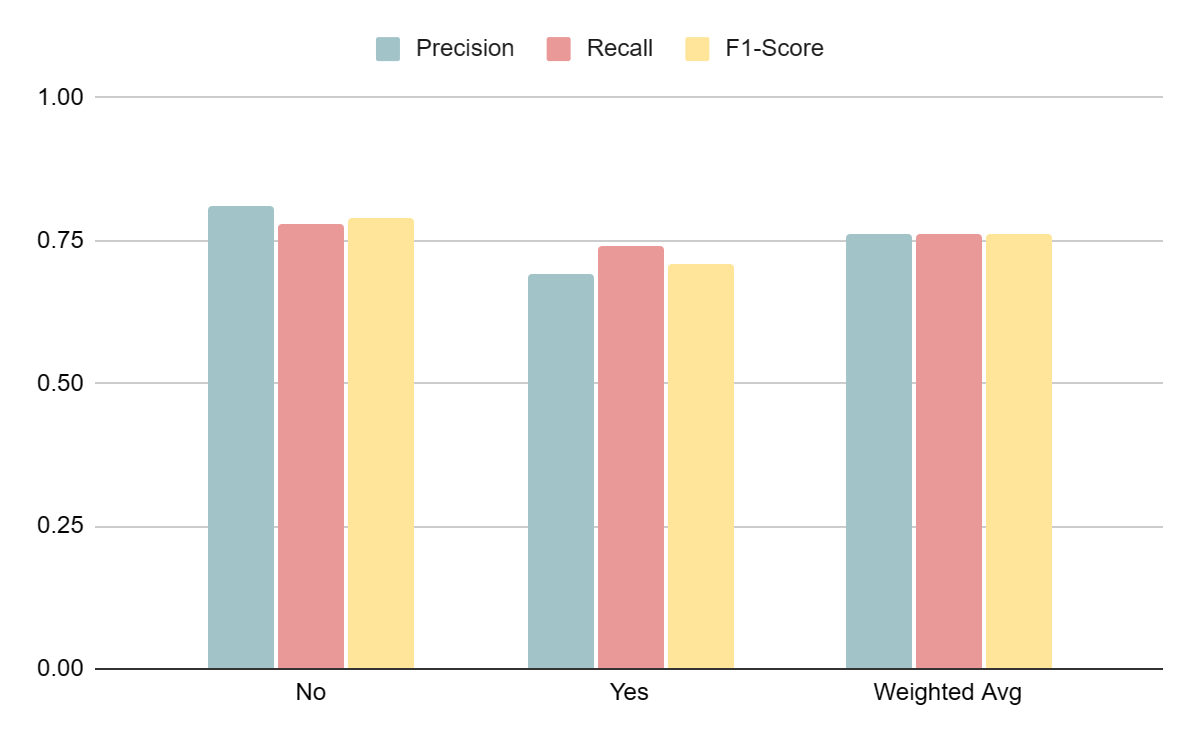


Figure 3: Bar chart of the test output’s results

## **Conclusion**

This project demonstrates the potential of using machine learning to predict health outcomes in the context of asthma. By combining environmental data with individual demographics and other health data, it is possible to offer valuable insights into asthma management. This approach could be further extended to other respiratory conditions, enhancing the ability to prevent and manage episodic health issues effectively. Further avenues to enhance the realism and applicability of my model include: Incorporating Real-Time Data, and Expanding Demographic Data. Integrating real-time environmental data such as pollen counts, weather conditions, and air quality fluctuations as they could provide more dynamic and precise predictions. This would allow the model to adapt to sudden changes in environmental conditions that significantly impact asthma symptoms. For expanding demographic data, while my synthesized dataset effectively mirrors real-world distributions, acquiring and incorporating more granular demographic data from healthcare providers or longitudinal health studies could refine our understanding of asthma attack triggers. This includes detailed lifestyle data, socio-economic factors, and geographical mobility patterns of individuals.

## **Citations**

1. Xiang, Y., Ji, H., Zhou, Y., Li, F., Du, J., Rasmy, L., Wu, S., Zheng, Wenjin. Jim., Xu, H., Zhi, D., Zhang, Y., & Tao, C. (2019). Asthma Exacerbation Prediction and Interpretation based on Time-sensitive Attentive Neural Network: A Retrospective Cohort Study. MedRxiv (Cold Spring Harbor Laboratory). <https://doi.org/10.1101/19012161>
2. US EPA, O. (2016, August 12). Air Quality Index Daily Values Report. US EPA. <https://www.epa.gov/outdoor-air-quality-data/air-quality-index-daily-values-report>
3. CDC. (2020, December 1). Asthma Data Visualizations | CDC. Www.cdc.gov. <https://www.cdc.gov/asthma/data-visualizations/default.htm>
4. NHIS - 2020 NHIS. (2021, February 4). Www.cdc.gov. [https://www.cdc.gov/nchs/nhis/2020nhis.htm‌](https://www.cdc.gov/nchs/nhis/2020nhis.htm%E2%80%8C), [NHIS Frequency Codebook (cdc.gov)](https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2020/adult-codebook.pdf)