Research Question: Asthma and Pollution

Ainsley Chapman-Atherton, Nikki La, Jordan Ledbetter, Palaniappan Sithambaram

DAT 490 - Data Science Capstone Project

Arizona State University

1 Broad Question / Background

Asthma is a chronic respiratory condition that affects a large portion of the United States Population [4]. Air quality, in previous research, has been shown to correlate with increased Asthma cases, especially in larger cities [4]. This project aims to expand on previous research and use machine learning methods to discover and substantiate evidence that certain contaminants may have greater effects on Asthma rates than others. To do this, we will utilize public environmental and epidemiological data to determine if there are any correlations and add a deeper understanding through machine learning concepts.

1.1 Main Question: What contaminants are the most significant in causing Asthma?

Asthma is a complex respiratory condition influenced by various environmental factors. Previous studies suggest a correlation between air quality and increased asthma cases, particularly in urban areas. This question aims to delve deeper into the specific contaminants that play a significant role in the development and exacerbation of asthma.

1.2 Sub-Question: Do different contaminants deferentially correlate with different diseases?

Environmental contaminants are diverse, and their impact on health can vary. This sub-question explores the possibility that distinct contaminants may have specific associations with different respiratory diseases.

1.3 Sub-Question: How can machine learning algorithms leverage environmental data, including pollution levels, to identify patterns and predict asthma incidence and outcomes, ultimately contributing to a better understanding of the environmental factors influencing asthma prevalence?

In the context of asthma and environmental data, this question addresses the potential of machine learning algorithms to analyze patterns, identify trends, and predict asthma incidence and outcomes. The goal is to enhance our understanding of the intricate relationships between environmental factors and asthma prevalence through advanced computational techniques.

1.4 Sub-Question: To what extent can pollution data from 2020 to 2021 be used to predict future asthma trends, and what are the limitations and challenges associated with forecasting asthma prevalence based on past environmental conditions?

By exploring the limitations, opportunities, and reliability of such predictions, we aim to discern the factors influencing the accuracy of forecasts and contribute to a comprehensive understanding of the dynamics between pollution and future asthma trends.

2 Preliminary Data

2.1 Centers for Disease Control and Prevention (CDC) Data

Prelimanry Asthma data we plan to use for this project comes from this CDC dataset: `https://data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-County-Data-20/swc5-untb/about_data` [3]. In addition to Asthma rates for the US by county for 2020-2021, this dataset also records various other health conditions and rates like stroke and heart disease.

2.2 United States Environmental Protection Agency (EPA) Data

The EPA has a wealth of air quality and environmental data of which we will primarily be using state and county data measurements of pollutants suchs as carbon monoxide, lead, nitrogen dioxide, ozone and more. we plan to use the files from this page `https://aqs.epa.gov/aqsweb/airdata/download_files.html` [2] which has county specific data through 2023. We also plan to utilize their github files for additional datasets that may be useful `https://github.com/USEPA/Air-Trends-Report` [1].

3 Preliminary Methods

For exploratory analysis, we will map pollution data against asthma rates to confirm the possibility of correlation. We will also run an ANOVA test with Asthma rates as the independent variable and the different pollutant will compose the dependent variables. This should give us some insight into answering our main question and provide direction for how we may answer our sub-questions.

For machine learning, we plan to create a convolutional neural network(s) (CNN), an ARIMA model, multiple regression with regularization, and Random Forests. According to Mendez, Merayo, and Nunez, the afore mentioned machine learning techniques are common and reasonable to use to forecast air quality data, which may help us answer our subquestions [5].

References

[1] Environmental Protection Agency. *AirData website File Download page*. en. Data & Tools. 2023. URL: https://aqs.epa.gov/aqsweb/airdata/download_files.html (visited on 01/20/2024).

[2] United States Environmental Protection Agency. *Air Trends Report*. May 2023. URL: https://github.com/USEPA/Air-Trends-Report.

[3] Division of Population Health Centers for Disease Control and Prevention National Center for Chronic Disease Prevention and Health Promotion. *PLACES: Local Data for Better Health, County Data 2023 release — Data — Centers for Disease Control and Prevention*. Aug. 2023. URL: https://data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-County-Data-20/swc5-untb/about_data (visited on 01/20/2024).

[4] Michael Guarnieri and John R. Balmes. "Outdoor air pollution and asthma". In: *Lancet* 383.9928 (May 2014), pp. 1581–1592. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(14)60617-6. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4465283/ (visited on 01/20/2024).

[5] Manuel Méndez, Mercedes G. Merayo, and Manuel Núñez. "Machine learning algorithms to forecast air quality: a survey". en. In: *Artificial Intelligence Review* 56.9 (Sept. 2023), pp. 10031–10066. ISSN: 1573-7462. DOI: 10.1007/s10462-023-10424-4. URL: https://doi.org/10.1007/s10462-023-10424-4 (visited on 01/20/2024).