Correlation between air quality and respiratory diseases

Enrico D'Alberton, Pietro Girotto, Arjun Jassal

Abstract

The prediction of mortality rates for respiratory diseases based on air pollutant levels is a critical task in understanding and mitigating the impacts of environmental factors on public health. In this paper, we describe a machine learning model that uses data on air pollutants to predict respiratory disease mortality rates. A dataset containing historical records of pollutant concentrations and mortality rates associated with respiratory diseases was used to create and evaluate the model. We used a multi-step method that included data preparation, implementation of machine learning algorithms and experimental analysis. The performance of the final model are pretty got and they highlight the correlation between the data. This study presents an in-depth analysis of our approach, experimental design, results and analysis, highlighting the advantages and disadvantages of our model. In order to improve the accuracy and usefulness of mortality rate prediction models for respiratory disorders, we explore the consequences of our results and provide suggestions for future research areas.

Introduction

Much of the issue on global health is caused by respiratory disorders, which affect millions of people each year and cause significant morbidity and mortality. The development and aggravation of respiratory disorders have been linked largely to air pollution, a prevalent environmental problem (Kim et al. 2018). For public health treatments and policies to reduce related risks, understanding the relationship between air pollution and respiratory disease mortality rates is essential.

Machine learning methods have gained popularity in the field of predictive modeling recently, providing interesting new ways to identify intricate patterns and make accurate predictions. We have begun work on a machine learning model to predict respiratory disease mortality rates based on levels of contaminants in the air using these findings. Our model aims to provide insights into the possible influence of air pollution on respiratory disease outcomes, helping to identify high-risk populations.

The construction and evaluation of the model created to predict respiratory disease mortality rates are the main topics of the findings in this article. We used a comprehensive dataset consisting of historical records of pollution measurements and associated mortality rates.

The development of a predictive model is just one of the goals of this project, along with a thorough examination of its effectiveness and failures. We hope to identify areas for development and provide insight into the difficulties associated with estimating respiratory disease mortality rates based on air pollution by investigating the strengths and weaknesses of the model.

Related work

The relationship between air pollution and respiratory disease outcomes has been studied extensively in the literature. According to studies, there is a direct correlation between the amount of air pollutants and respiratory morbidity and mortality. For example, (Dong et al. 2012) found a statistically significant positive connection between respiratory disease mortality and long-term exposure to PM10.

Despite these valuable additions, there are still some gaps in the existing literature. For example, few studies also examine the combined effects of various air contaminants on respiratory disease outcomes, which requires models that reflect the complexities of real-world situations.

Our research aims to fill these gaps by creating a machine learning model that can predict respiratory disease mortality rates. Our model seeks to provide a more complete understanding of the association between air pollution and respiratory disease outcomes by exploiting a diverse dataset that includes numerous respiratory disorders and multiple measures of air pollutants.

Data Collection and Preprocessing

The two main datasets used in this survey were respiratory disease mortality rates and air pollutant measurements. The IHME (Institute for Health Metrics and Evaluation 2023) provided the mortality rates dataset, while the EPA (Environmental Protection Agency 2023) provided the air pollutants dataset.

The respiratory disease mortality rates dataset covered several categories of respiratory diseases, including asthma, interstitial lung disease, pulmonary sarcoidosis, chronic respiratory diseases, chronic obstructive pulmonary disease, pneumoconiosis (such as silicosis, asbestosis, and coal workers' pneumoconiosis), and other chronic respiratory diseases. Federal Information Processing Standards codes

(FIPS 2023) and years were used to further subdivide the data, covering the years 1980 to 2014.

The dataset contained measurements of PM10, PM2.5, ozone, nitrogen dioxide, sulfur dioxide, and carbon monoxide in the air. The data were collected from numerous monitoring sites scattered throughout the research area. To ensure data accuracy, only measurements from stations with the most days of operation within a year were chosen.

The data collection process, including precise information on data collection and any obstacles encountered, was not disclosed.

The chosen air contaminants were combined into a single dataset during data pre-processing. The data set units for each pollutant were made consistent. Based on FIPS code and year matching, pollutant values and mortality rates were combined to create a final integrated dataset. In order to complete missing values with the closest accessible values in the dataset, forward and backward filling techniques were used to handle null values in the dataset.

In addition, based on the final integrated dataset, separate databases were created for each type of respiratory disease to facilitate analysis and modeling. This classification allowed more focused research on the relationship between air pollution levels and mortality rates for each respiratory disease.

The aforementioned data collection and pre-processing techniques sought to ensure the accuracy and consistency of the data sets and to establish a comprehensive and reliable framework for further analysis and model development.

Methodology

This section presents the methodology used to develop and evaluate machine learning models to predict respiratory disease mortality rates based on amounts of air pollutants. The models were implemented using the PyTorch (PyTorch 2023) and Scikit-learn (scikit learn 2023) libraries in Python, taking advantage of their robust machine learning capabilities.

To explore different modeling approaches, we experimented with various regression algorithms and explored different hyperparameters. The models were trained and evaluated on the prepared dataset consisting of air pollutant measurements and respiratory disease mortality rates.

First, we applied a linear regressor to establish a baseline performance. We then extended our analysis by employing polynomial regression models with degrees between 1 and 5. The goal was to assess whether higher regression models could provide a more accurate analysis.

Next, we explored Support Vector Machine (SVM) models with different kernels, including linear, polynomial, and radial-based (RBF). SVM models are well suited to capture complex relationships in the data and can provide flexibility in modeling nonlinear interactions between predictors and outcomes.

In addition, we used a neural network architecture, consisting of three layers with an equal number of neurons in each layer. We experimented with varying the number of neurons between 10 and 50 to evaluate the impact on model performance.

By employing these methodologies and evaluating a range of models with different hyperparameters, we aimed to identify the most suitable model(s) for predicting respiratory disease mortality rates based on air pollutant quantities. The performance evaluation results will be discussed in the subsequent section, providing insights into the strengths and limitations of each model and guiding future research in this domain.

Experimental Setup

To assess the performance of the developed machine learning models, we employed a rigorous experimental setup that involved data splitting, training, and testing stages.

The integrated dataset, consisting of air pollutant measurements and respiratory disease mortality rates, was randomly divided into two subsets: a training set and a test set. We allocated 80% of the data for training purposes and reserved the remaining 20% for model evaluation.

The training phase involved fitting the models to the training data using the selected machine learning algorithms. For each model, we utilized the training set to estimate the model parameters and optimize the performance.

During the training process, we utilized various hyperparameters, as mentioned earlier, to fine-tune the models and optimize their predictive accuracy. This involved conducting a grid search or systematic exploration of hyperparameter combinations to identify the optimal settings.

By following this experimental setup, including the data splitting, training, and testing procedures, we aimed to objectively evaluate and compare the performance of the different machine learning models. The results of the experiments will be presented and discussed in the subsequent section, shedding light on the effectiveness of the models and guiding further analysis and conclusions.

Results and Analysis

In this section we present the results we got from our models. First of all we inpsect the correlation between data:

As shown above, the correlation between pollutants and mortality rates is not so relevant. One aspect that stands out in both Figure 1 and 2 is the correlation between PM10 and PM2.5 that is the highest in the whole dataset. Another interesting aspect is the difference of correlation between pollutants and the mortality rate for the two different categories of respiratory disease. Indeed, in Figure 1 we can notice a weaker link between pollutants and mortality rate linked to Asbestosis. Instead, in Figure 2 we can highlight a better correlation of the mortality rate, in particular with PM10, PM2.5 and Ozone. This suggests that some categories of respiratory diseases are not so linked to air pollution and other instead could have a correlation with it.

Now instead, we present the obtained results with the tested models:

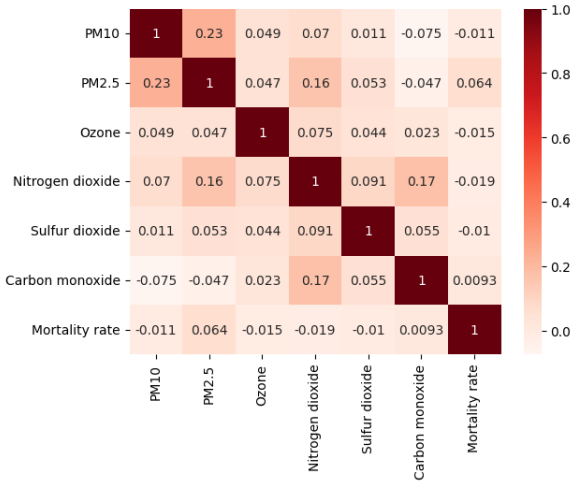

Figure 1: Correlation matrix with mortality rate relative to Asbestosis

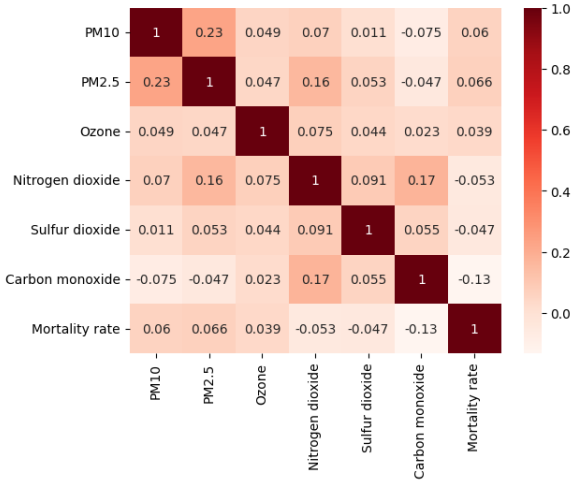

Figure 2: Correlation matrix with mortality rate relative to chronic respiratory diseases

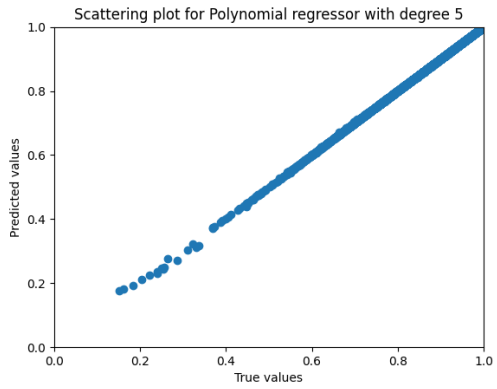

Figure 3: Executed on the test set of the Chronic respiratory disease dataset

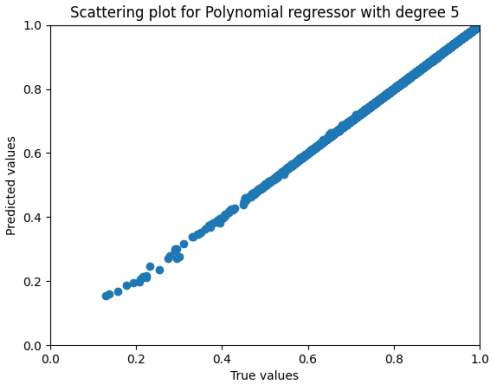

Figure 4: Executed on the test set of the Chronic obstructive pulmonary disease

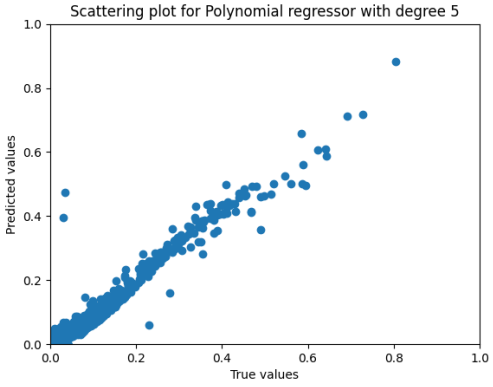

Figure 5: Executed on the test set of the Pneumoconiosis

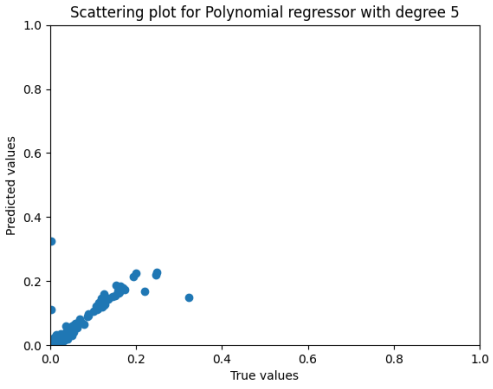

Figure 6: Executed on the test set of the Silicosis

Cause	MSE	MAE	MAPE	R2
508	4.38×10^{-7}	0.0002	0.0344	0.9999
509	7.68×10^{-7}	0.0003	0.0577	0.9999
510	1.33×10^{-4}	0.0065	32.13	0.9416
511	2.33×10^{-5}	0.0017	128.20	0.7964

Table 1: Compared performances of Polynomial regressor with degree 5 on different dataset

In Figure 3, 4, 5, 6 are reported the scatter plots of the best models we obtained for that specific dataset relative to a respiratory disease. These kind of plots provide a visual representation of the relationship between the predicted values and the actual values. Each point on the plot represents an individual data point, with the x-coordinate indicating the predicted value by the model and the y-coordinate representing the actual value. Ideally, we would like to see the points clustered closely around a diagonal line, which would indicate a strong correlation between the predicted and actual values. The closer the points are to the diagonal line, the higher the precision and accuracy of the model's predictions. As we can observe, the model work good with the chronic respiratory diseases and with the chronic obstructive pulmonary diseases. However, as already hypothesised from the correlation matrixes, the models are not able to predict the mortality rate accurately for some kind of respiratory disease, in this case Pneumoconiosis and Silicosis.

Moreover, Table 1 compares different metrics for 4 different mortality disease. In fact, the codes in the 'Cause' column refers to a category:

- **Cause 508**: Chronic respiratory diseases
- **Cause 509**: Chronic obstructive pulmonary disease
- **Cause 510**: Pneumoconiosis
- **Cause 511**: Silicosis

The exhibited data are clear and show a huge difference between the well and the bad performing models.

Discussion and Conclusion

The purpose of this study was to develop a machine learning model that could predict mortality for various respiratory diseases based on the amount of air pollutants. We explored different regression models, including linear regression, polynomial regression of orders 1-5, support vector machines (SVMs) with different kernels, and neural networks with different numbers of layers and neurons. As highlighted in the previous section, among these models, the fifth-order polynomial regressor performed the best. All the tests executed are visible at this Github page (https://github.com/gp-1108/air_pollution_ai).

In summary, our study successfully developed a machine learning model to predict mortality from respiratory diseases based on the amount of air pollutants. Polynomial regressors proved to be the most effective models, performing well for several causes of respiratory disease. This model can be a valuable tool for assessing the potential impact of air pollution on respiratory health.

Although the polynomial regression function shows promising results, it is important to recognize that there may still be room for improvement. Future research should explore further techniques such as ensemble methods and more advanced neural network architectures to further improve the prediction accuracy of all causes of respiratory disease.

Overall, the results of this study contribute to our understanding of the complex relationship between the amount of air pollutants and mortality from respiratory disease. The models developed will help public health professionals, policy makers, and researchers identify and address potential health risks associated with air pollution, ultimately leading to more targeted interventions and prevention. lead to measures.

References

Dong, G.-H.; Zhang, P.; Sun, B.; Zhang, L.; Chen, X.-F.; Ma, N.-N.; Yu, F.; Guo, H.-Z.; Huang, H.; Lee, Y. L.; and et al. 2012. Long-term exposure to ambient air pollution and respiratory disease mortality in Shenyang, China: a 12-year population-based retrospective cohort study. *Respiration*, 84(5): 360–368.

Environmental Protection Agency. 2023. Environmental Protection Agency. https://www.epa.gov/.

FIPS. 2023. Federal Information Processing System (FIPS) Codes for States and Counties.

Institute for Health Metrics and Evaluation. 2023. Institute for Health Metrics and Evaluation. https://www.healthdata.org/.

Kim, D.; Chen, Z.; Zhou, L.-F.; and Huang, S.-X. 2018. Air pollutants and early origins of respiratory diseases. *Chronic Diseases and Translational Medicine*, 4(2). Special Issue: Air Pollution and Chronic Respiratory Diseases.

PyTorch. 2023. PyTorch. https://pytorch.org/.

scikit learn. 2023. scikit-learn. https://scikit-learn.org/stable/.