Understanding the Health Consequences of Air Quality using Machine Learning

Williams, Kevin
Computer Science
California State University, Los Angeles
Los Angeles
williamskevin1968@gmail.com

Pourhomayoun, Mohammad
Computer Science
California State University, Los Angeles
Los Angeles
mpourho@calstatela.edu

***Abstract*—Elevated levels of pollutants are known to have a detrimental effect on one's respiratory health and depending on the season, weather can inflate pollution levels causing more harm to those with respiratory ailments. This study involves using averaged weather data and air pollutant levels to predict the amount of daily emergency department visits for asthmatic patients in Los Angeles County. We employed various machine learning algorithms such as Random Forest (RF), Multi-Layer Perceptron (MLP), XGBoost and Support Vector Regressor (SVR) with the best results coming from Random Forest, RMSE of 20.30, and the worst results from MLP, RMSE of 33.09.**

***Index Terms*—asthma, machine learning, regression, meteorological factors, air pollution, weather**

I. Introduction

Air pollution has always been preeminent concern when it comes to the health and safety of the population, especially for those suffering from respiratory ailments. Air pollution is caused by vehicle emissions, the burning of fossil fuels, pesticides, toxic fumes, forest fires, and more [1]. These pollutants, micrometers in size, once inhaled, travel through the respiratory tract and into the lungs causing irritation and inflammation, resulting in respiratory infections, hospitalization, and in some cases, death. [2-3]

Air quality has been a topic of research for decades in the US starting with the Clean Air Act in 1963, resulting in the study and regulation of air pollution.[4] Since then, The Environmental Protection Agency (EPA) has started recording six specific pollutants known to be detrimental to one's health: Particulate Matter 2.5 (PM2.5) and 10 (PM10), Sulfur Dioxide (SO2), Carbon Monoxide (CO), Ground-Level Ozone (O3), and Nitrogen Dioxide (NO2). [5] These six pollutants are known to be harmful, increasing the level of pollutants in the air and can trigger onset asthma attacks as well as boost the rates of morbidity and mortality related to respiratory disease.[6]

One of the most common, yet dangerous respiratory diseases is Asthma. Asthma is a chronic disease that affects the lungs by narrowing and swelling the airways, resulting in restricted breathing, wheezing, shortness of breath, and more. Asthma attacks can be life-threatening and are usually triggered by increased air pollution, pollen, or weather.[7-8] This disease, paired with poor air quality, can lead to hospitalization with life-threatening symptoms. Currently, there is no cure for asthma, and it is estimated that 25 million Americans are affected by this disease including 10 million people in Los Angeles County, as well as 339 million people worldwide. [9-11]

With the robustness and effectiveness of Artificial Intelligence and Machine Learning (ML), we can build models to aid medical caregivers with additional decision-making tools. In this study, we propose a data-driven model that predicts the amount of daily asthmatic emergency department visits in Los Angeles County given forecasted weather data and daily pollutant concentration. These models can be very helpful for medical facilities, hospitals, and other health institutions to prepare in advance for patient visits as well as provide actionable advice to patients.

The proposed model includes algorithms for handling missing values, eliminating useless information, and selecting the most impactful features. After the preprocessing stage, we applied various regressive ML models to develop a predictive model to forecast the amount of asthma emergency department visits in LA county.

II. Related Work

Asthma is a common chronic disease that can reduce the quality of life and poses an economic burden. Air pollution is considered a main factor in asthma and many studies evaluate their relationship, showing a positive correlation [12] but with the growing desire to stray away from fossil fuels and finding newer energy alternatives, the association is believed to be growing weaker [13], thus inviting the idea of studying other factors that may be crucial in asthma exacerbation [14]. In light of this, recent studies have been looking at the correlation between air pollution, socioeconomic factors, race, age, and other environmental factors with asthma [15-19]. These studies look at the causation of asthma attacks or hospitalizations, mainly in adolescents with a binary output, predicting the likelihood that one would develop asthma or be hospitalized. These models result in high accuracy yielding significant promise. However, this does not benefit the population and others experiencing the same medical issues. We propose a slightly different idea. Rather than predicting the likelihood of how one's health might be affected, we want to focus and gain an understanding on how many people will be affected on a given day. In the wake of COVID-19, we have seen a shortage of medical staff and resources due to various reasons and they need to be prepared for everything. This paper focuses on studying the association of the most common environmental factors everyone in Los Angeles experiences and building a model to predict the count of ED visits based on those factors.

III. Data Collection

A. Data Set

In this paper, we constructed our data set from multiple well-known sources. The asthma emergency department visit count was given to us by the California Healthcare Access and Information and denotes the daily sum of all emergency department visitations in Los Angeles County from Los Angeles County residents.

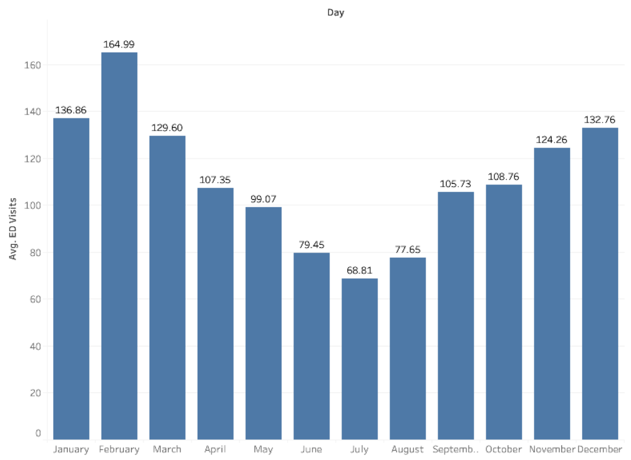

Fig. 1. Average Asthma ED Visits per Month

Our Meteorological factors or weather data was collected by Automated Surface Observation Systems (ASOS), ground-based sensors, located at all airports throughout Los Angeles County and maintained by the Federal Aviation Administration. The airport's ASOS data utilized in this study are Van Nuys, Hollywood/Burbank, Los Angeles International, Ontario International, Long Beach, and Palmdale Regional. Regarding the air pollutants (CO, PM10, PM2.5, SO2, NO2, O3), our data was collected from the EPA's online tool that queries daily air quality statistics by city, county, and state.

The following step after data collection was to handle missing values. Since there were at most two missing values within the ASOS data for each specific weather feature, the missing values were computed by averaging the values of the previous and succeeding days. With regard to the air pollutant data from the EPA, there were a handful of missing values from each different sensor for each different pollutant. If a pollutant sensor had more than 5% of values missing within our time frame, the sensor was dropped and not taken into account for this study. The remainder of the data was then passed into KNN Imputer to predict the missing values based on the mean of the surrounding data point values. We checked all odd neighbors from 1- 20 and validated their goodness of fit with the RMSE of a simple random forest model. This was repeated for each pollutant and the pollutant data with lowest RMSE from the random forest model was kept for the data set used in the ML models. After the computation of missing values, each pollutant data set was averaged to represent the value for that pollutant over Los Angeles County. This was also repeated for the weather data. In the end our data set consisted of 13 features and 1919 rows ranging from the dates of October 1, 2015 to December 31, 2020.

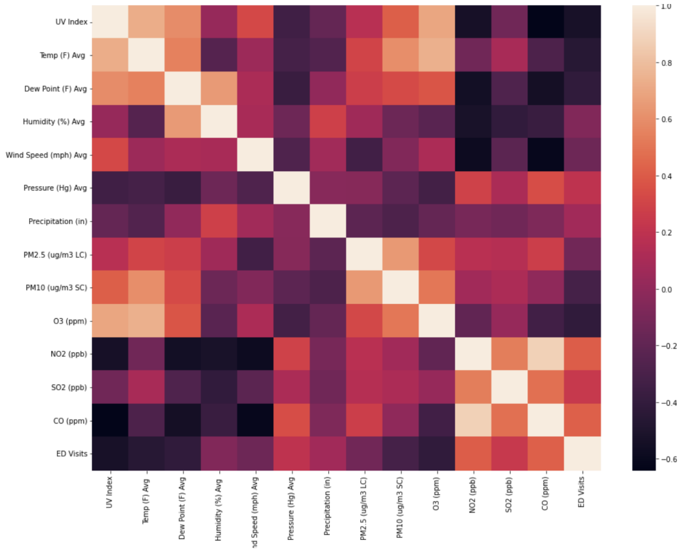

Fig. 2. Pearson Correlation Matrix

TABLE I
No. of Sensors for each pollutant

Pollutant	No. of Sensors
C0	12
NO2	11
O3	10
PM2.5	6
PM10	2
SO2	2

TABLE II
Data set Features

Meteorological Factors	Air Pollutants
UV Index	O3
Temperature	NO2
Dew Point	C0
Humidity	PM2.5
Wind Speed	PM10
Pressure	SO2
Precipitation	

B. Feature Selection

Feature selection methods are used to select the most useful or informative features for our Machine Learning models by exploring their relationship with the label, filter method, or by searching for the best subset of features, wrapper method, to remove redundant data and reduce the models complexity. After applying both the filter and wrapper method, we chose 6 features out of the original 13 to continue with this work. 3 features came from both the meteorological factors(UV Index, Temperature, and Dew Point) and air pollution (O3, NO2, CO).

TABLE III
FEATURE SELECTION

Air Pollutants	Meteorological Factors
UV Index	O3
Temperature	NO2
Dew Point	C0

C. Machine Learning Models

After the best features selection, we used 4 different machine learning algorithms for predictive analysis: Random Forest regressor, Multi-layer Perceptron (MLP), Support Vector regressor, and XGBoost (XGB).

Within MLP we normalized the data set between 0 and 1, using min max normalization.

$$Xn = \frac{Xi - X.min()}{X.max() - X.min()} \tag{1}$$

Furthermore, we employed a grid search on each algorithm specifically looking for the best parameters in each model. Within RF, we searched min_samples_leaf, min_samples_split, and n_estimators. In MLP, we searched the number of hidden layers and how many neurons in each layer, the learning rate, activation and solver. SVR was searched for the optimal C, gamma, and kernel with a fixed epsilon at 0.1, and finally XGBoost was searched for n_estimators, max_depth, gamma, and reg_lambda.

D. Evaluation

To evaluate these models, a few different techniques were used. We used both train-test-split at 2 different sizes in addition to 10-fold cross-validation. With MLP and XGBoost, a 70-30 train test split was utilized, and with SVR, we only utilized a 10-fold cross validation. Random forest used both a 90-10 training and testing size as well as 10-fold cross-validation. Our goal was to find the absolute best model with our data as well as how these algorithms would work on average. Since we utilized regression algorithms, we calculated the RMSE of all models as well as the percent error of the mean to represent how well our model prediction capabilities are in relation to the mean of the label.

$$PercentErrorofMean = RMSE/y.mean() * 100 \tag{2}$$

IV. RESULTS

Within the algorithms used, Random Forest with a 90-10 training/testing split performed the best while every other algorithm, SVR, XGBoost, and MLP all performed around the same with an average RMSE of 32. The table below describes their performance and their percent error of mean.

TABLE IV
RESULTS

Models	RMSE	Percent Error of Mean
RF 90-10	20.30	18.2
XGBoost 70-30	31.22	28.0
SVR CV	32.15	28.8
RF CV	32.89	29.0
MLP 70-30	33.09	29.6

V. CONCLUSION AND DISCUSSION

Using environmental factors such as air pollution and weather conditions, we are able to predict the amount of asthma emergency department visits in Los Angeles County. The models utilized in this study are Random Forest with a 90-10 training/testing split with an RMSE of 20.30, and RF 10-fold cross-validation having an RMSE of 32.89, XGBoost with an RMSE of 31.22, SVR 10-fold cross validation having an RMSE of 32.15, and the worst model being MLP with an RMSE of 33.09. We believe this information can be beneficial for the County's ability to prepare and set aside resources in preparation for asthmatic patients. Furthermore, this information can also be used give actionable advice such as issue patients to stay at home or give proper medication instructions. Following this work, we believe that this can be implemented on a per zip code basis in Los Angeles County to fully understand specifically when and where asthmatic patients are more common and why. Although the 90-10 split RF model did perform the best, it is believed that the other models have a more accurate representation of actual performance based on the larger testing size utilized. We think it is worthwhile to note that the number of emergency department visits started trending downward in March 2020 due to the issuance of the stay at home order for the COVID-19 pandemic. This event dropped the mean of the label by about 14%. As of now, it is unknown as to whether this was only for 2020 or the trend is on a downward decrease as data for 2021 is currently unavailable.

REFERENCES

[1] "Air Pollution," World Health Organization. [Online]. Available: https://www.who.int/health-topics/air-pollutiontab=tab_1. [Accessed: 28-Mar-2022].

[2] A. I. Tiotiu et al., "Impact of Air Pollution on Asthma Outcomes," IJERPH, vol. 17, no. 17, p. 6212, Aug. 2020, doi: 10.3390/ijerph17176212.

[3] D.-H. Tsai et al., "Effects of particulate matter on inflammatory markers in the general adult population," Part Fibre Toxicol, vol. 9, no. 1, p. 24, Dec. 2012, doi: 10.1186/1743-8977-9-24.

[4] F. Caiazzo, A. Ashok, I. A. Waitz, S. H. L. Yim, and S. R. H. Barrett, “Air pollution and early deaths in the United States. Part I: Quantifying the impact of major sectors in 2005,” Atmospheric Environment, vol. 79, pp. 198–208, Nov. 2013, doi: 10.1016/j.atmosenv.2013.05.081.

[5] Y.-C. Liang, Y. Maimury, A. H.-L. Chen, and J. R. C. Juarez, “Machine Learning-Based Prediction of Air Quality,” Applied Sciences, vol. 10, no. 24, p. 9151, Dec. 2020, doi: 10.3390/app10249151.

[6] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, “A Machine Learning Approach to Predict Air Quality in California,” Complexity, vol. 2020, pp. 1–23, Aug. 2020, doi: 10.1155/2020/8049504.

[7] T. Janssens and T. Ritz, “Perceived triggers of asthma: key to symptom perception and management,” Clin Exp Allergy, vol. 43, no. 9, pp. 1000–1008, Sep. 2013, doi: 10.1111/cea.12138.

[8] I. Rosas et al., “Analysis of the relationships between environmental factors (aeroallergens, air pollution, and weather) and asthma emergency admissions to a hospital in Mexico City,” Allergy, vol. 53, no. 4, pp. 394–401, Apr. 1998, doi: 10.1111/j.1398-9995.1998.tb03911.x.

[9] Z. He, J. Feng, J. Xia, Q. Wu, H. Yang, and Q. Ma, “Frequency of Signs and Symptoms in Persons With Asthma,” Respir Care, vol. 65, no. 2, pp. 252–264, Feb. 2020, doi: 10.4187/respcare.06714.

[10] “Los Angeles County asthma profile - cleanair.org.” [Online]. Available: https://cleanair.org/wp-content/uploads/LosAngeles2016profile.pdf. [Accessed: 30-Mar-2022].

[11] “2019 National Health Interview Survey (NHIS) data,” Centers for Disease Control and Prevention, 14-Dec-2020. [Online]. Available: https://www.cdc.gov/asthma/nhis/2019/data.htm. [Accessed: 28-Mar-2022].

[12] Y. Liu et al., “Short-Term Exposure to Ambient Air Pollution and Asthma Mortality,” Am J Respir Crit Care Med, vol. 200, no. 1, pp. 24–32, Jul. 2019, doi: 10.1164/rccm.201810-1823OC.

[13] D. P. Strachan, “The role of environmental factors in asthma,” British Medical Bulletin, vol. 56, no. 4, pp. 865–882, Jan. 2000, doi: 10.1258/0007142001903562.

[14] L. Wijesekara and L. Liyanage, “Modelling Environmental Impact on Public Health using Machine Learning: Case Study on Asthma,” in 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA), Sydney, Australia, Nov. 2020, pp. 1–7. doi: 10.1109/CITISIA50690.2020.9397488.

[15] E. T. Alharbi, F. Nadeem, and A. Cherif, “Predictive models for personalized asthma attacks based on patient’s biosignals and environmental factors: a systematic review,” BMC Med Inform Decis Mak, vol. 21, no. 1, p. 345, Dec. 2021, doi: 10.1186/s12911-021-01704-6.

[16] Q. Do, S. Tran, and A. Doig, “Reinforcement Learning Framework to Identify Cause of Diseases - Predicting Asthma Attack Case,” in 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, Dec. 2019, pp. 4829–4838. doi: 10.1109/BigData47090.2019.9006407.

[17] S. Bose, C. C. Kenyon, and A. J. Masino, “Personalized prediction of early childhood asthma persistence: A machine learning approach,” PLoS ONE, vol. 16, no. 3, p. e0247784, Mar. 2021, doi: 10.1371/journal.pone.0247784.

[18] M. R. Sills, M. Ozkaynak, and H. Jang, “Predicting hospitalization of pediatric asthma patients in emergency departments using machine learning,” International Journal of Medical Informatics, vol. 151, p. 104468, Jul. 2021, doi: 10.1016/j.ijmedinf.2021.104468.

[19] J. L. Harvey and S. A. P. Kumar, “Machine Learning for Predicting Development of Asthma in Children,” in 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, Dec. 2019, pp. 596–603. doi: 10.1109/SSCI44817.2019.9002692.