JOURNAL OF
EMERGING INVESTIGATORS

Article

Predicting asthma-related emergency department visits and hospitalizations with machine learning techniques

Kushal Chatterjee[1], Richard Williams[2]

[1] Saint Francis High School, Mountain View, California

[2] Schmahl Science Workshops, San Jose, California

SUMMARY

With the World Health Organization's declaration that 90% of the world breathes hazardous air, there is an increasing need to investigate the effects of ambient pollutants on human respiratory health. This research uses machine learning (ML) to examine asthma in Los Angeles County, an area with substantial pollution, and determine the success of classifiers in predicting future asthma-related hospitalizations. The objectives of applying ML-based solutions into healthcare are trifold. Firstly, it identifies the ML models that most accurately predict asthma hospitalizations. Secondly, it evaluates the significance of the correlation among ambient pollution, weather, and asthma. Ultimately, the model serves as a clinical support system to forewarn health care providers of asthma exacerbations. The hypothesis was that ML classification techniques would be able to predict asthmatic census based on ambient pollution and weather, displaying a positive trend between pollutant levels and asthma hospitalizations. The models revealed that nitrogen dioxide (NO_2) and ozone (O_3) levels were significantly correlated with asthma hospitalizations. A simple decision tree was the least accurate but useful in selecting features for other models. K-nearest neighbors, random forest, gradient boosting, and support vector machine classifiers predicted the asthma quartile with accuracies of 68%, 65%, 64%, and 62%, respectively. Overall, four ML classifiers were promising predictors, all showing consistency in k-fold crossover testing. The seasonal surge in asthma hospitalizations suggests that further research should explore other seasonal variables and ML classifiers to improve the models.

INTRODUCTION

Asthma has developed into a major concern in the United States; the disease costs the US $50.3 billion annually and leads to 2.1 million Emergency Department (ED) visits (1, 2). Ambient pollution has been historically understood to affect the respiratory system of individuals, exacerbating diseases such as asthma, chronic obstructive pulmonary disease (COPD), and lung cancer, among other chronic respiratory issues (3). Ambient pollution renders itself extremely problematic since it accounts for 4.2 million deaths every year worldwide (4). Studies performed by researchers have discovered that air pollution in England exhibited a significant association with coronary heart disease (CHD) mortality rates (5).

According to the American Lung Association's (ALA) annual State of the Air Report, Los Angeles (LA) County has the highest O_3 pollution in the United States. LA County, as well as neighboring counties in Southern California, have consistently received F grades from the ALA with respect to both O_3 and 24-hour particulate pollution (6). Despite legislation to improve emissions, the fact that the vast majority of LA's large population is heavily reliant on automobiles, coupled with the large number of imports arriving through Los Angeles ports, primarily explain why residents of Los Angeles continue to breathe hazardous air daily.

With the vast amounts of data now available at our fingertips, we decided that ML would be an extremely valuable tool for this project to conduct analysis and derive conclusions from clinical data. With the availability of electronic health datasets, using ML to predict asthma exacerbations has the potential to revolutionize the prevention and treatment of this respiratory disease. ML techniques are currently applied in the realm of precision medicine, which focuses on the customization of treatment based on the context of each individual patient (7). As the healthcare industry is increasingly storing patient data to enable research into disease prevention, using computerized data analysis has become attractive for handling large amounts of data.

ML is extremely useful for automation of decision-making processes. Named "supervised learning", the programmer provides the ML model with a series of known inputs (features) and outputs (target). Then, the algorithm can "predict" an output given an input it has not seen before. A common example of this used in the clinical world is for detecting the severity of a tumor. Several ML models have been created to detect whether a tumor is benign or malignant based on an image. In this scenario, the input is the image of the tumor, and the output is the binary prediction of whether the tumor is benign or malignant.

The primary goal for this project was to create a ML model to predict asthma-related ED visits and hospitalizations. Pediatric asthma has become a "growing American epidemic," as it is now the leading chronic disease in American children. From the 1980s through the following 20 years, asthma rates for children under five years of age have skyrocketed by 160% (8), indicating asthma is extremely troublesome and ever increasing.

In ML, input variables, known as features, go into the

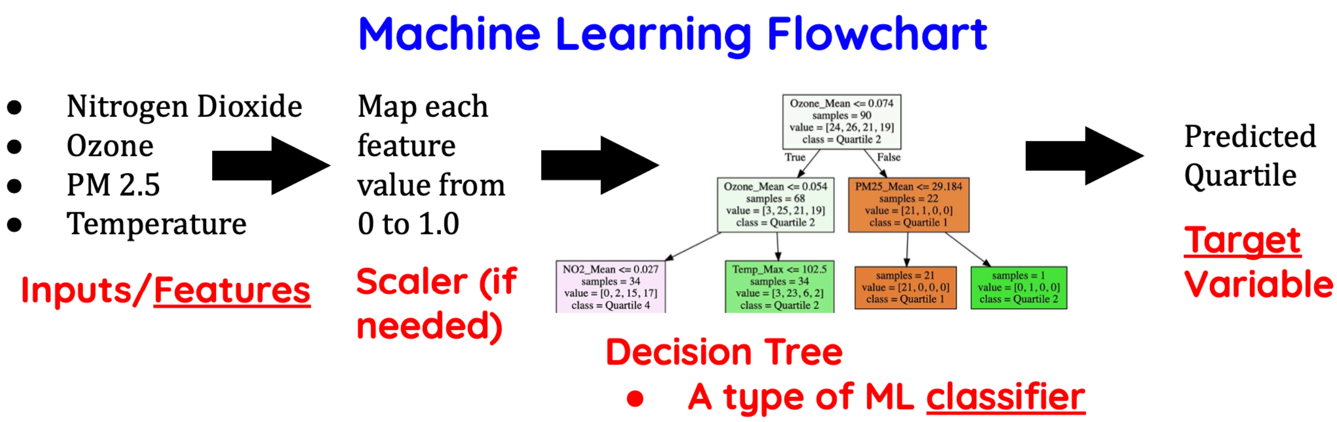

Figure 1: Flow chart illustrating the steps for machine learning classification. Shows the features passed in as inputs to the classifier, which then predicts the ED visits quartile (target).

classifier, which then outputs the target variable. For this experiment, the feature variables included ambient pollutant levels of O_3, NO_2, and $PM_{2.5}$, as well as temperature. PM2.5 is a measure of airborne fine particulate matter that is less than or equal to 2.5 microns in diameter (9). The target variable, or final prediction, was the quartile of asthma hospitalizations. Within the scope of ML, the features are entered into a classifier, which can then provide desired information regarding the target variable **(Figure 1)**.

There were various required tasks to achieve our project goal of creating a quartile predictor. Important steps along the way were identifying features and correlations, as well as comparing classifiers for accuracy. We hypothesized that ML classification techniques would be able to predict the measure of ED visits and hospitalizations in Los Angeles County based on ambient pollution and weather. Since the literature review had heavily emphasized the harms of pollution on the respiratory system, we believed that the classifiers would display a generally positive trend between pollutant levels and asthma hospitalizations. Furthermore, we posited that weather would help the classifiers in prediction, as it is intricately connected with ambient pollution.

To understand the effects of environmental features more specifically on asthma, we created various ML models that were able to provide key insight into the treatment of the burdensome disease. The models displayed that NO_2 exhibited a significant positive correlation with asthma hospitalizations, while O_3 exhibited a negative correlation. Additionally, the k-nearest neighbors, random forest, gradient boosting classifier, and support vector machine classification techniques all served as a reliable predictor for hospital asthmatic census. The capability of a prognostic tool like the one created in this study is a promising idea for the future of clinical medicine, as it would greatly aid with effective patient care.

RESULTS

Decision Tree

The ML work started with a decision tree, which was used for the feature importance information it provides, to select the features to use as inputs for the other ML models. Although the decision tree served as the least accurate predictor itself (58% accuracy), it may have performed better with more samples or features. Regardless, the decision tree was extremely useful because of its capability to evaluate feature importance, which was applied to best create the other models. In a decision tree, each node finds a feature and value, regardless of positive or negative correlation, that best separates the data into two groups, making each group as pure as possible. The first three levels of the decision tree show that O_3 mean, $PM_{2.5}$ mean, NO_2 mean, and maximum temperature provided the greatest separation of classes **(Figure 2)**. This realization helped with feature selection, as these four features were then chosen to train/test the four other ML classifiers.

Feature Importance

In this experiment, selection helped to decide the specific pollution and weather-related variables that were helpful for predicting asthma census. The feature importances from the decision tree summarizes how much each feature was used to predict the quartiles in the previously discussed decision tree **(Figure 3)**. Mean $PM_{2.5}$, mean O_3, minimum NO_2, and maximum temperature were most important. Furthermore, mean O_3 was by far the most important feature, which confirmed our prior findings from the separation of classes in the decision tree, which also had mean O_3 as the feature in the root node. However, the information provided by feature importance needs to be considered carefully. Since many of the features were highly correlated with each other, it is important to note that this chart does not say O_3 was 6 times more important than NO_2. Additionally, the graph does not make an argument regarding the effect of mean O_3 on asthma hospitalizations; it simply states that mean O_3 was important

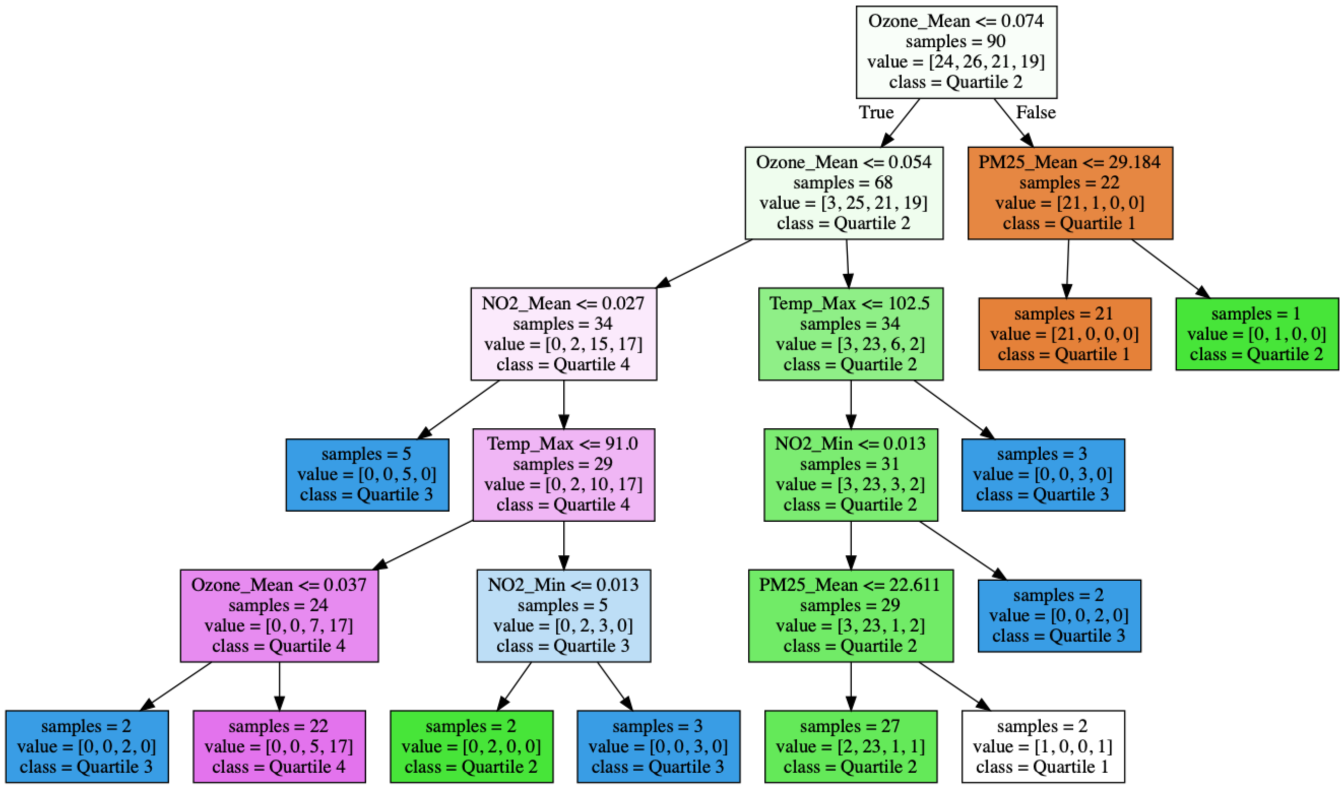

Figure 2: Mean O_3 yielded the greatest separation of classes in the decision tree model. A graphical tree (.dot file) is shown from the decision tree classifier. The separation of classes for various features are represented. The nodes higher in the tree provided greater separation.

for the tree. The next step, therefore, becomes finding what specifically made mean O_3 important (i.e., how it affected our target variable, asthma ED visits and hospitalizations).

This correlation matrix corroborates the feature importance graph from the decision tree that indicates mean O_3 and NO_2 having the strongest correlation with the target of asthma-related ED visits and hospitalizations

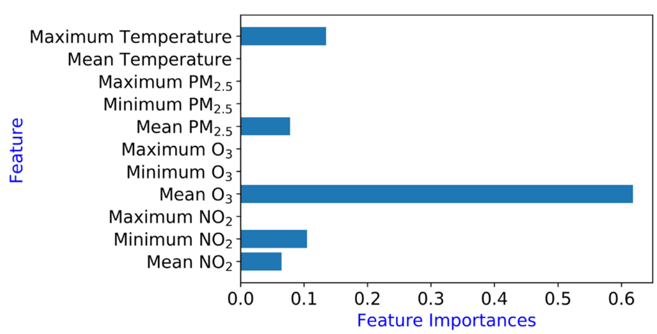

Figure 3: Mean O_3 had the highest feature importance, still being the most important feature when considering all the decisions in the tree. Chart that shows the feature importances from a decision tree. Feature importances are always positive and add up to 1.0. They are a measure of how much the tree used each feature to make decisions. The importance number for each feature is a value from 0 to 1; 0 would mean that the feature was not used at all, while 1 would mean that the feature can independently make a perfect prediction of the target.

(Table 1). Unlike feature importance, a correlation has directionality: it can either be positive or negative, depending on the slope of the relationship between variables. Mean O_3 displayed a correlation of approximately -0.83 with asthma hospitalizations. Furthermore, mean NO_2 displayed a correlation of approximately 0.73 with asthma ED visits and hospitalizations, which, in conjunction with the line graph of NO_2 and asthma **(Figure 4)**, led us to believe that the strong positive correlation between the two variables may be indicative of a causal relationship, which points to further research.

Four Prediction Models

Using the information regarding which features were most crucial to include in training a ML model, four prediction models were created, including models using Euclidean measurements (k-nearest neighbors and support vector machine), as well as ensemble models using multiple decision trees as base learners (random forest and gradient boosting classifier). Stratified five-fold crossover testing of k-nearest neighbors, random forest, gradient boosting classifier, and support vector machine predicted the asthma ED visits and hospitalization quartile of the testing data with accuracies of 68%, 65%, 64%, and 62%, respectively. New ML classifiers were made and tested with each of the five shuffled test sets, and the results were averaged to compute the accuracy

	Mean $PM_{2.5}$	Mean O_3	Mean NO_2	Maximum Temperature	Asthma Hospitalizations	Heatmap Scale
Mean $PM_{2.5}$	1.000	0.148	0.254	0.194	-0.020	1.000
Mean O_3	0.148	1.000	-0.702	0.569	-0.828	0.500
Mean NO_2	0.254	-0.702	1.000	-0.172	0.734	0.000
Maximum Temperature	0.194	0.569	-0.172	1.000	-0.390	-0.500
Asthma Hospitalizations	-0.020	-0.828	0.734	-0.390	1.000	-1.000

Table 1: The correlation heatmap of the target and features, showing the strength of the relationship between various pairs of variables used in the study. Unlike feature importances, correlations have directionality, so they can be both positive and negative. This table demonstrates that mean O_3 and mean NO_2 have the highest correlations, which agrees with the feature importances chart. Strong negative correlations are colored in red, and strong positive correlations are blue in color. Correlations are rounded to 3 decimal places.

percentages. Additionally, the Leave-one-out validation technique (also known as one-fold crossover testing) helped to validate the results as it produced similar findings. When predicting the test set with k-nearest neighbors, more than half of the predictions made perfectly aligned with the actual quartile **(Figure 5)**.

DISCUSSION

There are many types of ML classifiers, each with their own strengths and weaknesses. This study used five different ones to test, including decision tree, k-nearest neighbors, random forest, gradient boosting tree, and support vector machine. A basic decision tree is a tree of yes/no "decisions." Unfortunately, the drawback of a single decision tree is that it fits the training data perfectly but may be less successful when given previously unseen data. Therefore, both random forests and gradient boosting trees were created to overcome that issue by using an ensemble of decision trees as base learners. Multiple trees are created by limiting the depth or using subsets of the features or training samples. Each decision tree may be less accurate by itself. However, when they are combined in a majority vote, they can provide us with a more generalizable predictor that performs better with unseen data, as compared to the sole decision tree model. On the contrary, both k-nearest neighbors and support vector machines are not based on decision trees. K-nearest neighbors makes a prediction about an unknown sample based on the quartile of the nearest neighboring samples. The number of neighbors can be chosen by the programmer. This technique is analogous to a service like Netflix, which recommends movies based on those watched by "similar" people. Support vector machines map features into a higher dimensional space, and that can be separated into classes using hyperplanes. Additionally, we decided to use quartiles for the target variable because it provided the best results. Quantiles of higher specificity, like deciles, were not feasible in this project due to the limited size of the available datasets. In the future, the quantizing specificity would be determined based upon the desires of hospital ER staffers, who will eventually use such a prognostic tool. Similarly, in order to decide how accurate the ML model needs to be, we would need information from the hospitals as to what they require.

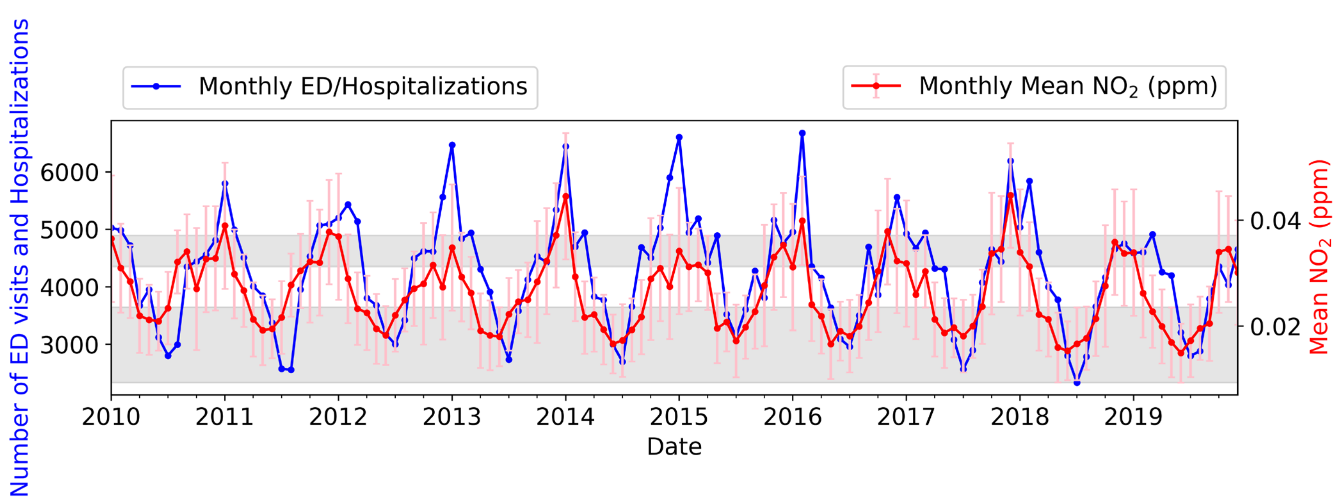

Figure 4: NO_2 mean and asthma hospitalizations each displayed a distinct seasonal relationship, with crests and troughs in different parts of the year. Line graph showing monthly hospitalizations and NO_2 mean from 2010 to 2019 in Los Angeles. The unit for NO_2 mean is parts per million (ppm), which is the number of particles of a pollutant for every 1 million particles of air. The quartiles for hospitalizations are shown by bands at the binning boundaries. The vertical lines in pink are +/- 1 standard deviation from the mean, showing the variability of NO_2 during the month.

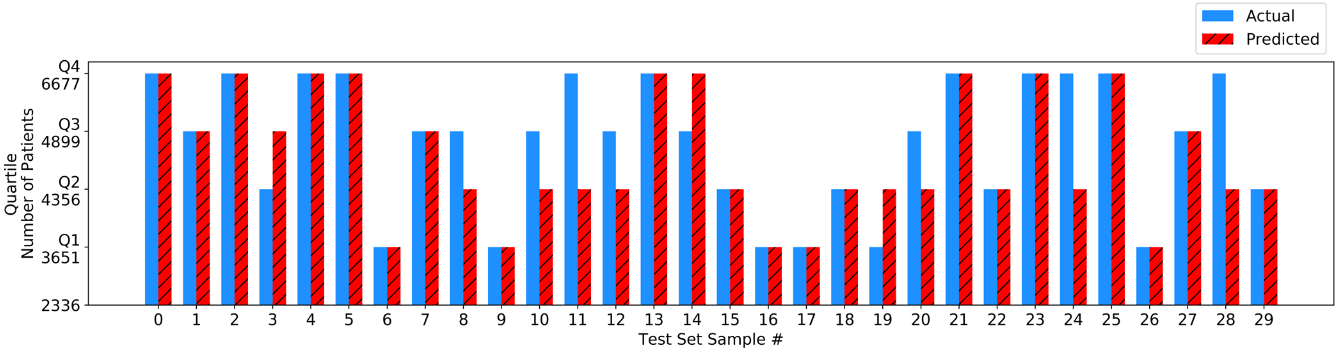

Figure 5: Chart comparing the test set predictions generated by k-nearest neighbors to the actual values. The blue bars represent the actual quartile value of each sample, while the red bars show the predicted quartile for that respective sample. Looking at the difference between the blue and red bar for each sample shows how the ML predictions compared with reality. The number of ED/hospitalized patients associated with each quartile boundary is added to the tick labels on the y-axis.

For now, this research project serves to demonstrate the feasibility of constructing ML models for hospital staffing; we claim that using ML in the healthcare industry is a promising idea for the future.

Our hypothesis that ML classifiers would be able to serve as a predictor for ED visits was supported by the results of this experiment, as the predictors were able to speculate the quartile of asthma hospitalizations with an average accuracy of 65%. Although 65% is likely not an ideal accuracy value, it points to the feasibility of using ML as a tool in the clinical world. However, the positive trend between pollutant levels and hospitalizations that we had hypothesized was not completely correct; some pollutants showed the opposite trend. Since this relationship between asthma and ambient pollution was more complicated than we expected, simple ML models that rely on linearity did not work well at all. The 2020 predictions made with the k-nearest neighbors show a pattern like the previous known years that were used to train and test the model, which is encouraging **(Figure 6)**. Generally,

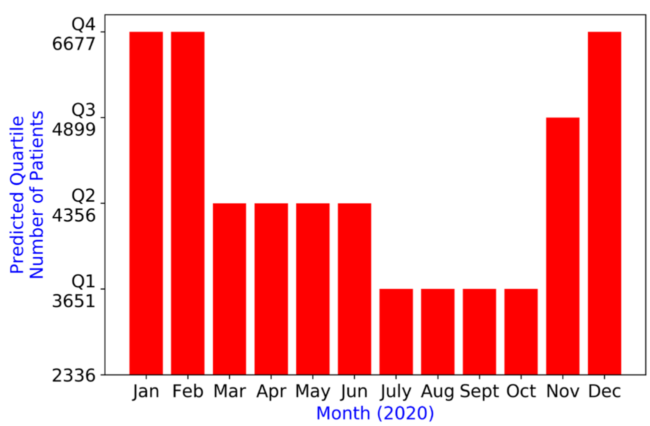

Figure 6: Chart displaying the 2020 asthma quartile predictions made by the k-nearest neighbors classifier. The number of ED/hospitalized patients associated with each quartile boundary is added to the tick labels on the y-axis. The seasonal asthma dip in the summer months and surge in the winter observed in earlier years is prevalent in the graph. Actual 2020 asthma hospitalization data was not yet available at the time of the study.

asthma quartiles are high in the winter and low during the summer months. While 2020 ED data was not available at the time this paper was written, it would be interesting to calculate the 2020 prediction accuracy of the k-nearest neighbors once the medical records are accessible. Additionally, in 2020, NASA computer models observed a worldwide NO_2 decrease by 20% on a city level due to the effects of Shelter-in-Place (SIP) orders (10). SIP is an extremely interesting test scenario, as a worldwide shutdown is a rare occurrence. This NASA theory raises the fascinating possibility that ML asthma predictors like the ones in this project can be combined with climate forecasting models to predict both the health-related and economic effects of climate change and human activity.

A possible limitation in this experiment involved the possibility that many individuals who experience asthma do not visit the hospital; ED visits and hospitalizations usually only occur in severe cases. Additionally, there are many other causes of asthma attacks that were not included as features for this project. This may have also hindered model accuracy.

Future experimentation could further improve accuracy by comparing results from other counties and states, acquiring data over a longer span of time, or acquiring daily asthma data. Monthly data may have smoothened some features. The correlation matrix showed that the various pollutants displayed differing relationships with asthma hospitalizations **(Table 1)**. NO_2 displayed a positive correlation with asthma hospitalizations, which suggests that NO_2 could play a causal role, which is an area for future investigation. However, O_3 displayed a negative correlation with the asthma dataset, which was not hypothesized and is an important finding. Although this negative correlation between O_3 and NO_2 is worthy of further investigation, the chemical interactions between NO_2 and O_3 may account for the inverse relationship. NASA's Earth Observatory explains that the ratio of NO_x to volatile organic compounds (VOC) dictates O_3 formation (11). The NASA study explains how higher concentrations of nitrous oxides can inhibit the production of O_3 in the troposphere. Combined with the reduced hours of sunlight during the winter, which is also required for O_3 production, these factors may explain the inverse correlation between O_3 and NO_2, and the inverse relationship that we observed between O_3 and asthma. The chemical interactions occurring in the atmosphere between the ambient pollutants studied is an interesting confounding variable that can explain the statistical correlations used in data analysis. As $PM_{2.5}$'s correlation was extremely low, the correlation can effectively be declared as nonexistent. Maximum temperature showed a weak negative correlation with asthma and was relatively 'important' compared to the other features tested, as suggested in **Figure 1**.

A ML model that predicts asthma census will be extremely useful for hospital staffers. Optimizing the staffing to match the predicted demand will help to relieve the economic burden and ensure the highest quality of care for all patients in healthcare (12). This research suggests that ML can transform vast and currently existing datasets into clinically actionable data. Moreover, the findings from this project can inspire societal change. Understanding the link between NO_2 and respiratory disease can incentivize individuals to pursue environment-friendly lifestyle interventions, such as renewable energy and public transport. Moreover, this information may encourage policymakers to pursue the regulation of pollution sources, in an effort to partially reduce the troublesome burdens of respiratory disease.

Future research should investigate whether the correlation involving NO_2 is indicative of a causal relationship. Furthermore, it would also be interesting to look at other environmental variables, like pollen count and type, to strengthen the accuracy of the models.

METHODS

General Preprocessing of the Datasets

Before the ML aspect of the project could begin, all the datasets necessary for training and testing the models were acquired. This study required ambient pollution data, monthly asthma hospitalization data, and temperature data. The asthma-related hospitalization data was acquired from the California Office of Statewide Health Planning and Development. This CSV dataset contained monthly asthma-related ED visits and hospitalizations counts (summed up) for most counties in California from the years 2010 to 2019. The ambient pollution data was acquired from the California Air Resources Board, which provided pollutant data for $PM_{2.5}$, O_3, and NO_2, among many other ambient pollutants, updated daily. Finally, this project used temperature data from the Los Angeles Almanac, which contained the monthly maximum temperature, average temperature, and minimum temperature for the years 1877 to 2020.

Overall, from the datasets obtained, the sample data that we used for the ML models consisted of 10 years of monthly environmental and asthma hospital data. Therefore, there were 120 observations, one for each month in the 10-year span, used for each input variable and the hospital output.

All the ML processes were conducted using many in-built ML packages in Python, in conjunction with the *JupyterLab* environment. The *Jupyter* Notebook is very similar to a lab notebook, and each cell can hold Python code or text documentation.

Once the datasets were obtained, they were "cleaned" from their raw form into a more usable form for the model. Data cleaning is the process that checks for invalid and missing data and removes or adjusts values accordingly.

Once all the datasets were cleaned, the cleaned datasets were resampled to provide monthly data. Although the asthma hospitalization and temperature data were already provided at the monthly level, the pollution dataset provided daily data. Therefore, we calculated the mean, standard deviation, maximum, and minimum of daily pollution for each month and converted it into a monthly scale.

The datasets were then processed for use, as they had been cleaned and resampled to match the monthly scale. Next, we statistically inspected the data for feature relevance, to understand which features are most important. Feature relevance can also be determined graphically using the *Matplotlib* package, which allows us to plot together a specific pollutant with asthma ED visits and hospitalization counts.

A complete *Pandas* dataset was compiled with the month and year as the index value, and columns for the mean O_3, mean $PM_{2.5}$, mean NO_2, maximum temperature, and asthma ED visits and hospitalizations during that month, using insight gained from feature importance.

We then binned the hospitalization counts into four quartiles, which were used as the final prediction output in the ML classifier. Rather than predicting the exact number of hospitalizations given the feature pollutants and temperature conditions, our models predicted the quartile that the hospitalizations were expected to be in.

Next, the compiled dataset was separated using the standard ML split, which allocates training data as 75% of the dataset and testing data as the remaining 25% of the dataset.

Creating and Testing the ML Classifiers

At this point, we completed all the general pre-processing steps, so we started creating the ML models using the *Scikit-Learn* package. For each model, there were some model-specific preprocessing steps that needed to be programmed. For example, algorithms like k-nearest neighbors and support vector machine required scaling of the data, for the various pollutants and temperature features to all be on a consistent scale. Other models that were used in this project, such as decision trees and random forests, did not require scaling of the data, since scaling did not affect the model's performance.

Eventually, an 80% training set and 20% testing set were created and shuffled within five folds of 24 samples each. This k-fold crossover testing is done after the initial testing with the 75/25 split, to avoid a single 'lucky' or 'unlucky' random split giving an unusually high or low accuracy. The K value that we used was five, representing the number of folds of the dataset, which were combined to make training and test sets. The folds were stratified, making the proportions of classes (quartiles) in each fold the same as in the entire dataset, which guaranteed that there was a sample of every quartile in each split and kept variance low by making the folds similar to the original dataset. Each of the five folds were tested, and the accuracy was then averaged to yield an overall score. With an ML classifier, accuracy is calculated using the percent of correct predictions for the testing set. All these measures ensured more consistent results.

Once the accuracy of the model was tested using the testing data, the various parameters of the ML algorithm were tweaked to provide best results, in a process known as "parameter tuning".

Once an accurate model was achieved, current ambient pollution and weather data was passed into the model, which then predicted the quantile of asthma hospitalizations.

ACKNOWLEDGMENTS

I would like to thank Ms. Schmahl, who provided guidance with my inquiries regarding proofreading of project materials and the paper. I would like to thank my Computer Science teacher, Mrs. Vollucci, as well as medical professionals, Dr. Leon Castaneda and Dr. Jasmine Dhaliwal, for providing their feedback and revisions on my draft. I would like to thank my parents, Ananya and Krishanu Chattopadhyay, for funding the project and always providing invigorating encouragement. Finally, I am grateful to Schmahl Science Workshops for providing me with the platform to complete my project.

Received: April 29, 2021
Accepted: October 9, 2021
Published: October 25, 2021

REFERENCES

1. "Considerable Economic Burden for Asthma in United States." PracticeUpdate, Primary Care, 19 Jan. 2018, www.practiceupdate.com/content/considerable-economic-burden-for-asthma-in-united-states/63021. Accessed 17 July 2020.
2. "Reducing 2.1 Million Emergency Room Visits, One Count At A Time." ACAAI Public Website, 2 Oct. 2015, acaai.org/news/reducing-21-million-emergency-room-visits-one-count-time. Accessed 17 July 2020.
3. "Impact of Air Pollution on Asthma Outcomes." PubMed Central (PMC), www.ncbi.nlm.nih.gov/pmc/articles/PMC7503605. Accessed 17 July 2020.
4. "Ambient Air Pollution." *World Health Organization,* World Health Organization, www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/ambient-air-pollution.
5. "Contribution of Climate and Air Pollution to Variation in Coronary Heart Disease Mortality Rates in England." PubMed Central (PMC), 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC3299689. Accessed 17 July 2020.
6. *California: Los Angeles.* Los Angeles | American Lung Association. (n.d.). https://www.lung.org/research/sota/city-rankings/states/california/los-angeles.
7. Davenport, Thomas, and Ravi Kalakota. "The potential for artificial intelligence in healthcare." *Future Healthcare Journal* vol. 6,2 (2019): 94-98. doi:10.7861/futurehosp.6-2-94. Accessed 07 Apr. 2020.
8. POPULATION REFERENCE BUREAU. "Childhood Asthma: A Growing American Epidemic – Population Reference Bureau." PRB, 1 Aug. 2001, www.prb.org/childhoodasthmaagrowingamericanepidemic. Accessed 17 July 2020.
9. "California Air Resources Board." *Inhalable Particulate Matter and Health (PM2.5 and PM10)*, ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health.
10. Gray, Ellen. "Model Shows Extent COVID-Related Pollution Levels Deviated from Norm." NASA, NASA, 16 Nov. 2020, www.nasa.gov/feature/goddard/2020/nasa-model-reveals-how-much-covid-related-pollution-levels-deviated-from-the-norm.
11. "Chemistry in the Sunlight." NASA Earth Observatory, earthobservatory.nasa.gov/features/ChemistrySunlight/chemistry_sunlight3.php. Accessed 23 Apr. 2021.
12. Everhart, Damian et al. "The effects of nurse staffing on hospital financial performance: competitive versus less competitive markets." Health care management review vol. 38,2 (2013): 146-55. doi:10.1097/HMR.0b013e318257292b. Accessed 7 Apr. 2021.
13. Gass, Katherine. "Classification and Regression Trees for Epidemiologic Research: An Air Pollution Example." Environmental Health, 13 Mar. 2014, ehjournal.biomedcentral.com/articles/10.1186/1476-069X-13-17. Accessed 08 Sept. 2020.
14. Müller, Andreas. *Introduction to Machine Learning with Python.* O'Reilly Media, 2016.