**Course Project: Predicting Health Related Challenges Through Air Quality Metrics and Weather Conditions**

Christina Vosnak

Data Science, Bellevue University

DSC 630

Professor Farley

March 2nd, 2025

**Introduction**

Air quality metrics and weather conditions are globally monitored for their potential to influence various aspects of daily life. While factors like air pollution levels, temperature, humidity, and wind are often studied individually, their combined impact on public health remains less understood. One study states that “one of our era's greatest scourges is air pollution, on account not only of its impact on climate change but also its impact on public and individual health due to increasing morbidity and mortality” (Ghorani, 2020). This project aims to analyze the relationship between air quality metrics and weather variables to assess whether predictive models can accurately estimate trends in health-related challenges, such as hospital admissions and cardiovascular diseases. Prior to completing this project, I looked at some scholarly articles trying to understand the same concepts. One article in particular that looks at Air Quality Prediction states that “machine learning, as one of the most popular techniques, is able to efficiently train a model on big data by using large-scale optimization algorithms” (Zhu, 2018). This project aims to uncover similar trends using data that includes both air quality metrics and weather conditions to predict health impact on a synthetic dataset from Kaggle.

By leveraging data-driven insights, this project seeks to evaluate the feasibility of forecasting health risks based on environmental patterns. These forecasts can serve as valuable tools for proactive healthcare planning, enabling providers to allocate resources more effectively during high-risk periods. Policymakers can also use these insights to address societal challenges, such as those posed by climate change, and to develop strategies for reducing pollution levels. Ultimately, increasing global awareness of air quality and weather trends can contribute to a higher quality of life by fostering informed decision-making and preventative measures.

The findings from this project will benefit a diverse range of stakeholders. Public health agencies can use insights to implement preventive measures during high-risk periods, ensuring timely interventions for health concerns. Healthcare providers will be better equipped to prepare for patient influx during adverse environmental conditions, optimizing resources and staffing. Policymakers will gain access to data-driven insights necessary for crafting regulations to mitigate pollution and address climate-related health risks. Insurance companies can enhance their risk assessments by incorporating environmental health impact data into their decision-making processes. Lastly, the general public will benefit from these findings by staying informed and taking precautions to safeguard their health during periods of heightened environmental risk.

The data for this project was sourced from Kaggle's Air Quality and Health Impact Dataset, which contains over 5,000 records detailing various air quality and weather-related variables, along with information on health impacts such as respiratory and cardiovascular cases and hospital admissions. While the dataset is synthetic and designed for analytical purposes, its structure provides a robust foundation for exploring the relationships between environmental factors and health outcomes. In future work, more accurate, real-world data can be obtained from sources like Weather APIs and Air Quality Index platforms to enhance the reliability and precision of predictions regarding the impact of these factors on public health.

This dataset is instrumental in addressing the project’s goals as it provides a comprehensive view of the variables necessary for understanding the relationship between environmental conditions and health outcomes. It includes air quality metrics such as AQI, PM2.5, and PM10, alongside weather data like temperature, humidity, and wind speed, which are critical for identifying patterns and trends. Additionally, the dataset incorporates health metrics, including respiratory and cardiovascular cases and hospital admissions, allowing for a direct exploration of potential links between environmental factors and health challenges.

For this project, I plan to begin with a regression model to explore and capture the linear relationships between air quality metrics, weather conditions, and health outcomes. The simplicity and interpretability of a regression model make it a strong starting point, as it can reveal straightforward patterns and provide insights into the contributions of individual variables. If the regression model does not yield sufficient accuracy or fails to capture complex patterns, I will extend the analysis to include a Random Forest model. Random Forest is particularly effective at handling non-linear relationships and interactions between variables, offering greater flexibility and robustness in capturing more intricate dependencies within the data. This model will serve as an alternative approach to improve prediction accuracy and extract additional insights, such as feature importance.

To evaluate the performance of the models, I will employ a combination of metrics tailored to regression analysis. Mean Absolute Error (MAE) will provide a straightforward evaluation of average prediction errors, offering an easily interpretable measure of performance. Root Mean Squared Error (RMSE) will penalize larger errors more heavily, ensuring that outlier predictions are adequately accounted for in the evaluation. R² Score will measure the proportion of variance in the health outcomes explained by the model, providing insight into its explanatory power. This multi-metric approach will ensure a comprehensive assessment of the model's effectiveness and help guide further improvements if necessary.

Overall, I aim to uncover meaningful patterns and relationships between air quality, weather conditions, and health outcomes. Specifically, I plan to identify which environmental factors show the strongest correlations with specific health challenges. Additionally, I seek to determine the feasibility of using predictive models to forecast health risks, thereby providing actionable insights for proactive healthcare planning, policymaking, and general public awareness. I also aim to explore the potential of machine learning techniques in addressing these real-world challenges and gain a deeper understanding of model performance, strengths, and limitations when applied to this dataset. Ultimately, this project aspires to contribute valuable insights that can inform public health interventions and help mitigate the impact of environmental factors on health.

It is also important to consider the risks and ethical implications associated with this proposal. One potential risk is that the data may not represent all regions and demographics equally, which could result in biased predictions. Furthermore, machine learning models, while powerful, may oversimplify complex interactions between variables, potentially leading to inaccurate conclusions. Although this dataset uses synthetic data, future applications of this project could involve real-world health data, which raises privacy concerns. To ensure compliance with data protection regulations in such cases, I will prioritize the anonymization of any sensitive information and adhere to ethical standards in data handling.

Finally, if the original project plan encounters challenges or fails to produce meaningful results, I will implement a contingency plan to refine the scope of the project. This may involve narrowing the focus to either air quality metrics or weather data rather than analyzing both sets of factors simultaneously. Additionally, I will explore alternative datasets that include real-world information from public APIs or government health statistics to enhance the accuracy and scope of the analysis. On the modeling side, I will adjust and test different models as needed to optimize performance and ensure meaningful insights, all while avoiding over analysis of the data. This flexible approach ensures that the project can still produce valuable outcomes even if adjustments are necessary.

**Methods/Results**

To explore the data, I loaded the dataset for Air Quality and Health Impact that I found on Kaggle into a Jupyter notebook to conduct a preliminary data analysis. To understand the dataset, I first checked the information within each column and found all the features are numerical. Additionally, I checked for missing values and duplicate values and found that the dataset has neither. This means that the data consists of complete information for all 15 columns and 5,811 entries. The 15 columns are broken down as follows: 1 ID column, 6 air quality metric columns, 3 weather metric columns, 2 columns counting cases, 1 column counting hospital admissions, and 2 columns evaluating health impact.

To further look at the information contained within the dataset, I first prepared a correlation heatmap to better understand how the variables relate to one another.

A screenshot of a graph

Description automatically generated

As shown above, the heatmap displays each variable, and their degree of correlation. It is important to note the correlation between AQI and Health Impact Score being the most significant as well as AQI and Health Impact Class. Additionally, while not as strongly correlated, other air quality metrics such as PM10, PM2\_5, NO2, and 03 also correlate to Health Impact Score whereas weather metrics do not show a correlation. From the image on the next page

To further visualize the data, I created boxplots of all air quality metric features, weather features, cases, and hospital admissions. While the most of the air quality metrics and weather metrics are evenly distributed, it is important to note that both Respiratory and Cardiovascular cases as well as Hospital Admissions show several outliers indicating some records of unusually high health incidents.

A chart of blue rectangular shapes

Description automatically generated with medium confidence

Finally, to further explore the spread of each variable I plotted distribution charts for all air and weather metrics as well as cases and hospitalizations. I found that Respiratory Cases, Cardiovascular Cases, and Hospitalizations are all right skewed, further confirming the outliers within the dataset.

A group of graphs showing different types of data

Description automatically generated

As stated previously, the data contains no missing values or duplicates. After completing the heatmap, boxplots, and distribution visualizations, I confirmed that there are outliers in the data and went back into the dataset to understand those outlier cases to understand whether those cases are due to a data entry error or are significant to the dataset.

Overall, based on this preliminary analysis, I will be using a regression model to further evaluate the data. I will also use a Random Forest model to compare results. I will evaluate both models using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R² Score as each metric brings a different value for evaluation as stated in the introduction.

I chose these metrics because each provides a unique perspective on model performance. MAE measures the average magnitude of errors in a straightforward manner, making it easy to interpret. RMSE, by squaring the errors before averaging, gives more weight to larger errors, which is useful when I want to penalize significant deviations. Lastly, R² Score evaluates how well the model explains the variance in the data, helping assess overall goodness-of-fit. By using these three metrics, I can ensure a comprehensive evaluation of both regression and Random Forest models, balancing accuracy, interpretability, and reliability.

To prepare the data for this model, I luckily didn’t have to create any dummy variables to one-hot encode categorical variables as all the data in the air quality health impact dataset is numerical. I created my features and target variable in python, and I then standardized the numerical features in the dataset using a standard scaler. I also split the dataset into a training and test set to be able to evaluate my model. I built a linear regression model and a random forest model. From the two models we see the following results.

A screenshot of a computer code

AI-generated content may be incorrect.

The results indicate that the Random Forest Regressor significantly outperforms Linear Regression in predicting the target variable. The Random Forest model has a much lower Mean Absolute Error (1.58 vs. 7.36) and Root Mean Squared Error (3.21 vs. 9.63), demonstrating greater accuracy in its predictions. Additionally, its R² score of 0.944 suggests that it explains 94.4% of the variance in the data, whereas Linear Regression only accounts for 50.5%. The higher errors and lower explanatory power of Linear Regression suggest that the dataset may have non-linear relationships, which Random Forest, being a non-linear model, can better capture. Overall, Random Forest is the superior model for this dataset, providing more precise and reliable predictions.

A graph with blue squares

AI-generated content may be incorrect.

**Conclusion**

After determining that the Random Forest Regressor model is the optimal model to use. I wanted to trace back the accuracy of the model to the features that contributed to it. The feature importance plot highlights the top three factors influencing the Random Forest regression model’s predictions. Among them, AQI (Air Quality Index) has the highest importance score, indicating that it plays the most significant role in determining the target variable. This suggests that variations in AQI heavily impact the model’s predictions. The second most important feature is PM2.5 (Particulate Matter ≤ 2.5 micrometers), which also contributes substantially to the model's accuracy, likely due to its strong correlation with air quality and health effects. Lastly, PM10 (Particulate Matter ≤ 10 micrometers) ranks third in importance, playing a meaningful but relatively smaller role in comparison to AQI and PM2.5. The visualization clearly shows that AQI dominates in influence, while PM2.5 and PM10 remain significant predictors. These results emphasize the importance of air pollution metrics in the model, reinforcing their critical role in forecasting air quality-related outcomes.

To improve predictive accuracy and support informed decision-making, several recommendations can be made. First, AQI should be prioritized as a primary predictor, as it has the greatest influence on the model’s performance. Additionally, enhanced monitoring and regulation of PM2.5 and PM10 levels are essential, as these pollutants also significantly impact air quality. Since non-linear models have proven to be more effective, future studies should explore alternative machine learning models such as Gradient Boosting or Neural Networks for further improvements in prediction accuracy. Lastly, deploying this predictive model in a real-time monitoring system could provide continuous insights into air quality trends, assisting policymakers and public health officials in taking proactive measures to reduce respiratory cases and hospitalizations.

While the Random Forest model demonstrated strong predictive performance with a high R^2 score and significantly lower errors compared to the linear Regression, I don’t believe that it is fully ready for deployment. The dataset used in this analysis is limited to only 5000 entries. It is also important to note that this is fictional data, and real-world data should be evaluated prior to deployment. I would like this model to undergo further validation prior to implementation. I’d like to also get more information in terms of how seasonality may affect this data as that also might be a significant factor in determining spikes in respiratory conditions and hospitalizations.

To look at further information, I’d like to look at data from verified sources such as government health agencies, NOAA, and the EPA to validate model accuracy. I’d like to incorporate additional features to refine the impact of individual variables such as weather metrics which showed weak correlations in the above analysis. Additionally, I’d like to turn this data into accessible information prior to deployment to build an interactive dashboard that can be easily read by policymakers, healthcare providers, and the general public to ensure the information is easily accessible once the model is deployed.

Ethical considerations for this project revolve around data integrity, bias, and transparency in communicating results. The synthetic dataset used in this study may not fully represent real-world distributions, which can introduce biases when applied to actual populations. Furthermore, health-related predictions carry significant implications, and misinterpretation of the model’s results could lead to undue panic or misinformed public health decisions. When presenting findings, it is essential to clearly communicate model limitations to avoid misinterpretation from the general public.

If deployed in a real-world setting, this model could influence public health interventions, healthcare resource allocation, and policy decisions. Ethical concerns arise if the model disproportionately benefits or harms certain populations due to data biases. For example, certain communities which may experience higher exposure to pollutants, could be underrepresented in the dataset, leading to inaccurate predictions. An overall approach may not be best to deploy to the general public and location information should be included in a dashboard to further breakdown the results. Additionally, if insurance companies or employers misuse these predictions, individuals in high-risk areas could face discrimination in coverage or workplace accommodations. There is also a risk of public misinterpretation—if forecasts predict high hospitalization rates due to air quality, it could cause unnecessary alarm or lead to healthcare resource misallocation.

Overall, to mitigate these ethical concerns, it is important to utilize real-world data that captures information across diverse graphical regions to generate more accurate results. I’d like to create features to include regional information, which could be pulled from a Weather API to design a more accurate model that spans across a variety of climates. One example of an API to use for this expansion is the OpenWeather API which “collects and process weather data from different sources such as global and local weather models, satellites, radars and a vast network of weather stations.” (Open Weather). Regular audits of this model should be done to ensure its accuracy over time as air quality metrics and weather conditions change due over time as well. If the accuracy starts to decrease, future analysts should build alternative models and features that will improve accuracy over time. Finally, it is important to prevent misuse of this model by policymakers, heath officials and the public by creating a fully transparent dashboard to explain the models’ limitations.

Overall, this project explored the relations between air quality, weather conditions and health related challenges using a data-driven approach to create a predictive model of information. Through an exploratory data analysis, and slight data manipulation, a Random Forest model was created to capture non-linear relationships within the data achieving a high level of accuracy. While the results provide valuable insights for proactive healthcare planning and policymaking, there is still so much to explore regarding this information in the future. I hope to take the information learned from this project to implement a model on real-world data with additional features to explore. Ultimately, this project highlights the potential of leveraging environmental data for public health forecasting, paving the way for informed decision-making and preventative healthcare strategies in the future.

**References**

Ghorani, V., Gharaei, M., & Mahmodianfard, N. (2020). Environmental and health impacts of air

pollution: A review. *Frontiers in Public Health, 8*, 14. [**https://www.frontiersin.org/articles/10.3389/fpubh.2020.00014/full**](https://www.frontiersin.org/articles/10.3389/fpubh.2020.00014/full)

OpenWeather. OpenWeatherMap API. Retrieved March 1, 2025, from

<https://openweathermap.org/api>

Zhu, D., Cai, C., Yang, T., & Zhou, X. (2018). A machine learning approach for air quality prediction:

Model regularization and optimization. *Big Data and Cognitive Computing, 2*(1), 5. https://doi.org/10.3390/bdcc2010005