Behnke, Lopez

Predicting Air Quality - Final Report

1. Motivation - What is the problem you are considering, and why is this an interesting problem? Has this problem been attempted before?

Air quality is something that affects each individual person every day, but only certain people are in the proper positions to take action to improve air quality for any given area. The goal of this project is to identify which characteristics have the greatest effect on health outcomes in order to inform a plan of action for policymakers, as well as an opportunity for continued research on the topic. Additionally, of these most impactful features, we aim to identify which ones policy makers in a region can easily alter to provide better health outcomes for its people. As pollution in the air around the world rises, being able to predict health outcomes enables us to target areas that are most severely affected by air quality and help prevent adverse health outcomes.

In particular, there are industrial areas of most cities across the developed world that experience acute levels of air pollution. An example of one of these cities is Lubbock, Texas. Here, the industrial and manufacturing sector is in the east side of the city, and the people who live here may have higher risk for adverse health effects. According to First Street[1], Lubbock has worse air quality than around 80% of other cities in Texas, and air quality predictions for the future do not indicate improvement.

Different studies have explored using machine learning models to predict and determine Air Quality Indexes. Though they have been targeted toward air quality index, of the studies we have discovered, few have additionally explored measures focused on predictions of health outcomes regarding the quality of air in these regions.

In the study performed by Mauro Castelli and partners[2], they explored using machine learning modes to predict air quality in California. As a part of their study, they employed the support vector regression model, using the radial basis function kernel to forecast pollutant and particulate levels and to predict the air quality index with the highest accuracy of 94.1%. Additionally, they classified their results into 6 air quality index categories based on their index number. In this study they identified O3, CO, and SO2 as being notable pollutants as they worked to model their concentrations in regions across California.

[1] FIRST STREET TECHNOLOGY, INC. (n.d.). *Lubbock, TX Poor Air Quality Map and forecast*. firststreet.org. https://firststreet.org/city/lubbock-tx/4845000_fsid/air

[2] Castelli, Mauro, Clemente, Fabiana Martins, Popovič, Aleš, Silva, Sara, Vanneschi, Leonardo, A Machine Learning Approach to Predict Air Quality in California, *Complexity*, 2020, 8049504, 23 pages, 2020. https://doi.org/10.1155/2020/8049504

In another study, performed by T. M. Chiwewe and J. Ditsela[3], the researchers used machine learning algorithms to estimate ozone concentration. This study made use of data collected from air quality monitoring facilities that included spatio-temporal features. The team here employed Linear Regression models and Artificial Neural Networks to predict and measure the ozone levels for any particular facility. They then utilized a cross-correlation approach and a spatial correlation approach to measure and predict ozone with a data partition of 80% for training and 20% for testing. The ANN model resulted in an accuracy of 77%, while the regression model had an adjusted R-squared of .579. The most important predictors for the ANN were temperature, Nitrous Oxide, and relative humidity. The most important predictors for the regression model were temperature, relative humidity, and Nitrogen Dioxide.

A final study that we researched was performed by Sachin Bhoite, Sejal Pitale, Pooja Bhalgat. These researchers set out to predict the air quality index of different regions in India using levels of SO2 in the atmosphere as a measure. They implemented an integrated model using Artificial Neural Networks and Kriging to predict the level of air pollutants in different Indian cities. They also utilized logistic regression to detect whether a dataset was polluted or not and autoregression to predict future values of the P2.5 measure. As their focus was on SO2, they have identified this compound as particularly important, and their autoregression model achieved an MSE of 166.358.

2. Problem framing - What is the proposed solution to the problem?

Due to increasing levels of air pollution, people may begin to experience adverse health effects, especially in industrial and manufacturing areas. We are targeting solutions toward areas with industrial and manufacturing capacities—like those in the eastern region of Lubbock, Texas—because these areas are the most susceptible to poor air quality. Our data, which we derived from Kaggle, examines over 5000 observations. Though we do not have location data, we have determined that each observation corresponds to a different area or region and provides us with measurements of different air pollutants, temperature, humidity, and air quality related health issues that were reported.

Considering our data is well organized and requires minimal pre-processing, and since there are only numerical and factor type variables, we are interested in creating a regression model and XGBoost model to predict health outcomes. Based on this analysis, we can also derive feature importance to determine which pollutants or other variables are most impactful when determining health outcomes and propose policy makers implement policies that help protect those exposed to these factors. Our data has been easy to work with thus far and we believe that our models will provide us with clear results that we can use to help improve these

[3] T. M. Chiwewe and J. Ditsela, "Machine learning based estimation of Ozone using spatio-temporal data from air quality monitoring stations," 2016 IEEE 14th International Conference on Industrial Informatics (INDIN), Poitiers, France, 2016, pp. 58-63, https://ieeexplore.ieee.org/abstract/document/7819134.

health outcomes. We could run into issues of data imbalance based on our response variables, but this can be addressed using bagging, bootstrapping, or additional models that have built in functionality to deal with data imbalance.

3. Data overview - What is the dataset you are going to use to solve this problem? Describe the dataset and its characteristics.

The dataset we’re using to solve this problem is one from Kaggle: air quality and health impact. This dataset contains a comprehensive look at the characteristics of air quality, weather conditions, health impact metrics, and a health impact response class.

This dataset has 5811 observations and 15 variables, two of which can be used as outcome variables. The variables are as follows:

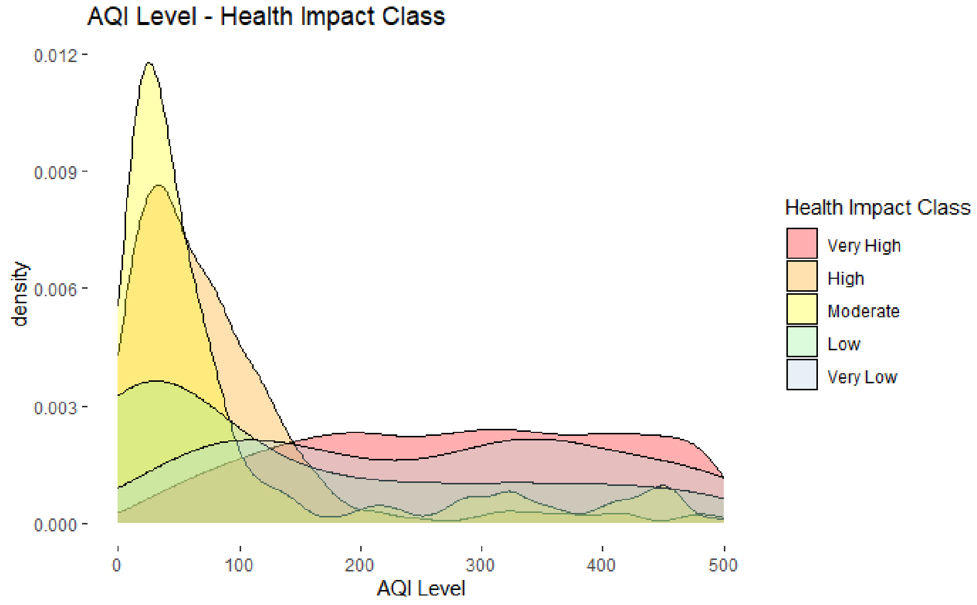
- RecordID: A unique identifier assigned to each record (1 to 2392)
- AQI: Air Quality Index, a measure of how polluted the air currently is or how polluted it is forecast to become
- PM10: Concentration of particulate matter less than 10 micrometers in diameter (µg/m^3)
- PM2_5: Concentration of particulate matter less than 2.5 micrometers in diameter (µg/m^3)
- NO2: Concentration of nitrogen dioxide (ppb)
- SO2: Concentration of sulfur dioxide (ppb)
- O3: Concentration of ozone (ppb)
- Temperature: Temperature in degrees Celsius (°C)
- Humidity: Humidity percentage (%)
- WindSpeed: Wind speed in meters per second (m/s)
- RespiratoryCases: Number of respiratory cases reported
- CardiovascularCases: Number of cardiovascular cases reported
- HospitalAdmissions: Number of hospital admissions reported
- HealthImpactScore: A score indicating the overall health impact based on air quality and other related factors, ranging from 0 to 100
- HealthImpactClass: Classification of the health impact based on the health impact score

(variable descriptions sourced from Kaggle)

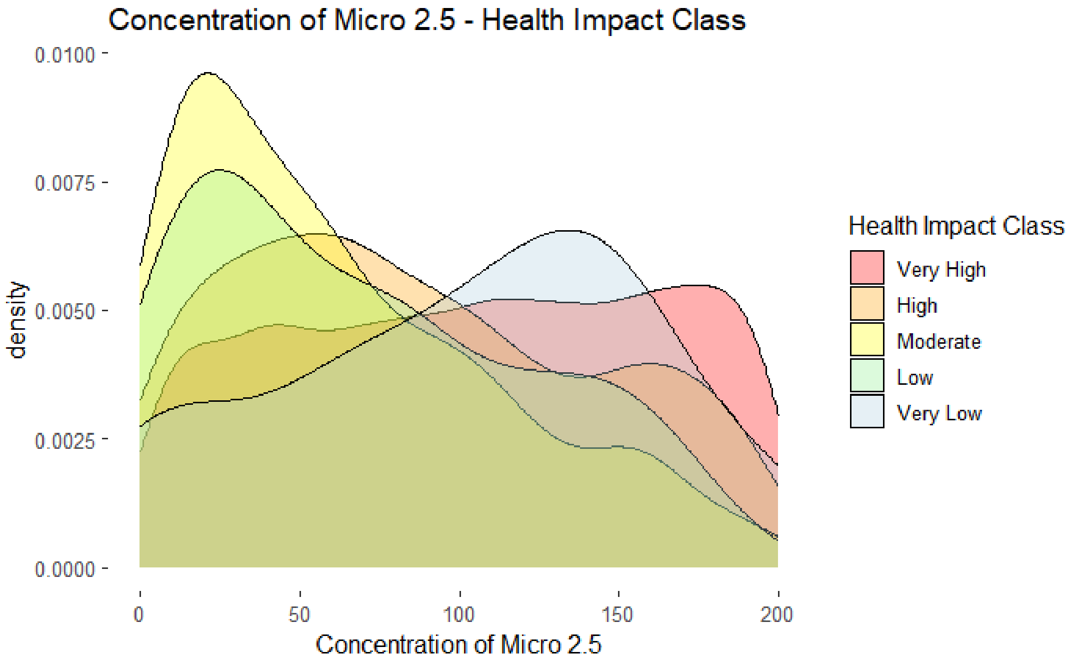
There are no missing values in this dataset.

Mary performed the initial data exploration for the project and provided useful articles for tailoring our project to gear towards areas that are affected by poor air quality. Additionally, she provided the data analysis section of the report. Allen performed the research for previous studies done in the area, created the bibliography, and wrote the explanation to the motivation section.

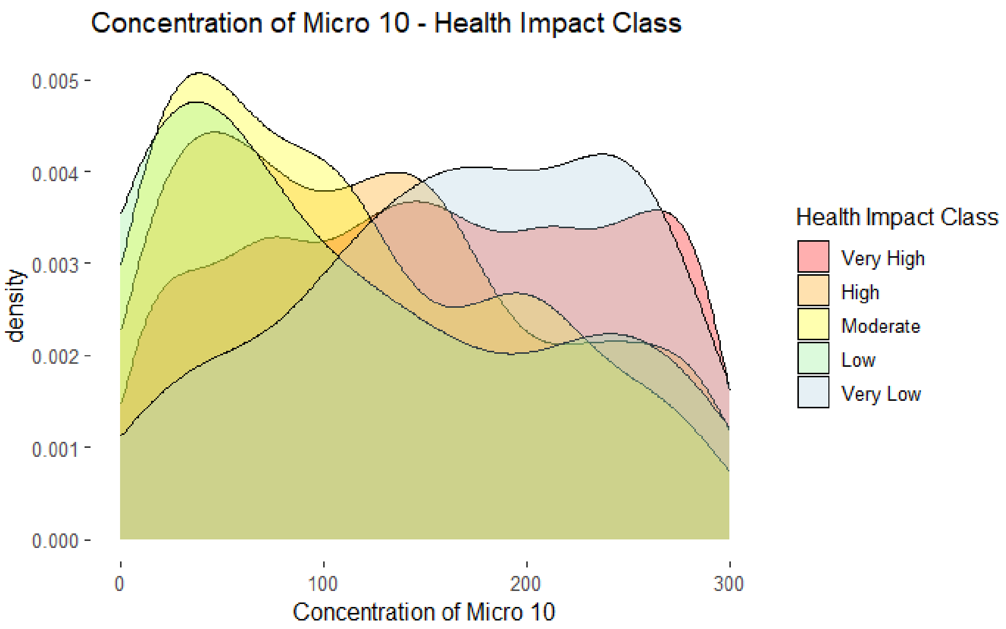
Visualizations - Exploratory Data Analysis



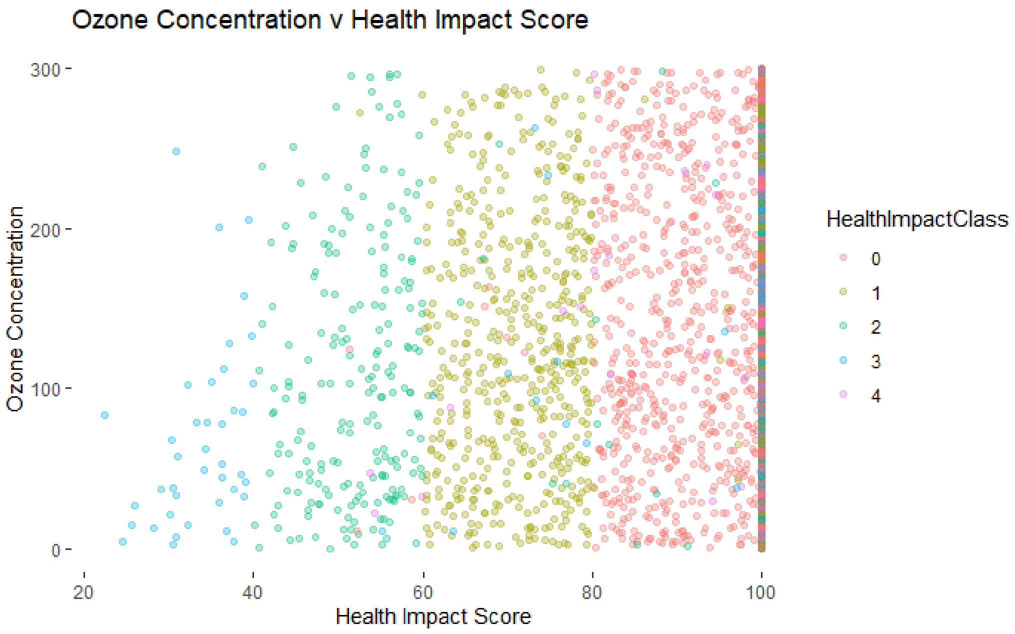
This visualization shows that there is a high concentration of High and Moderate Health Impact Class between the AQI levels of 0-100, with the Low class also primarily concentrated in the same area, and the Very High and Very Low classes picking up after the 100 AQI level.



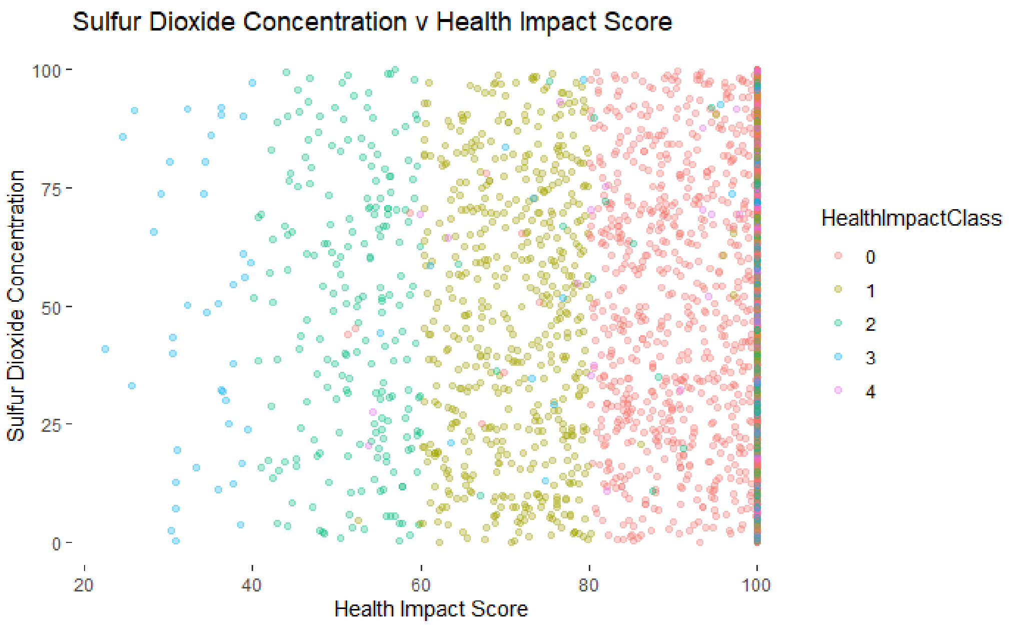
This visualization shows the Health Impact Class relating to the concentration of particulate matter less than 2.5 micrometers in diameter, with Moderate and Low classes concentrated in the lower area, and the Very Low class picking up in the 125 to 150 range.



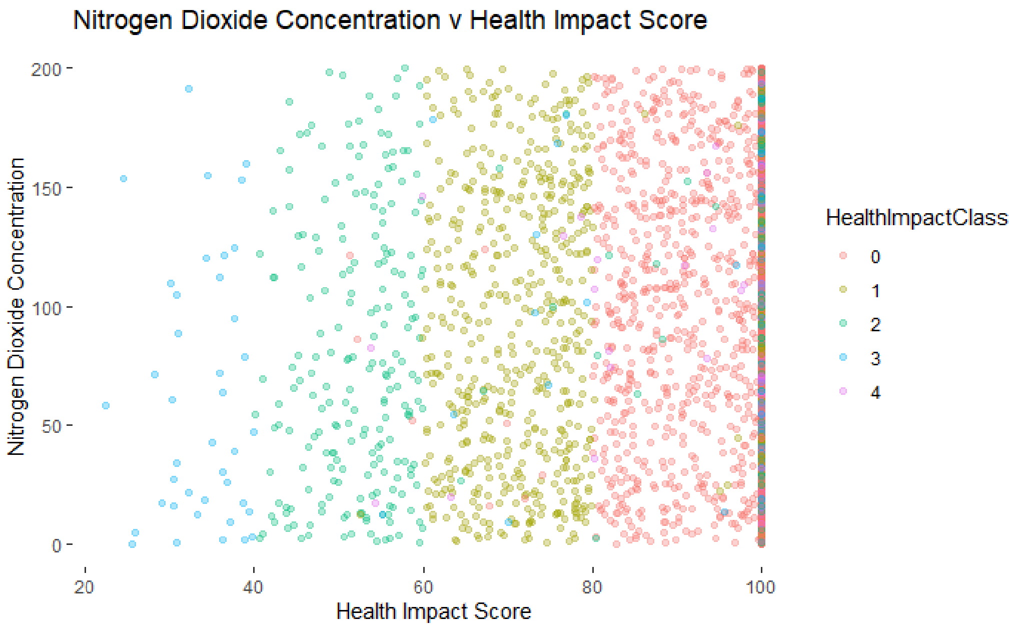
This visualization shows the Health Impact Class relating to the concentration of particulate matter less than 10 micrometers in diameter, with High, Moderate, and Low classes concentrated in the lower area, and the Very Low class picking up in the 125 to 150 range.



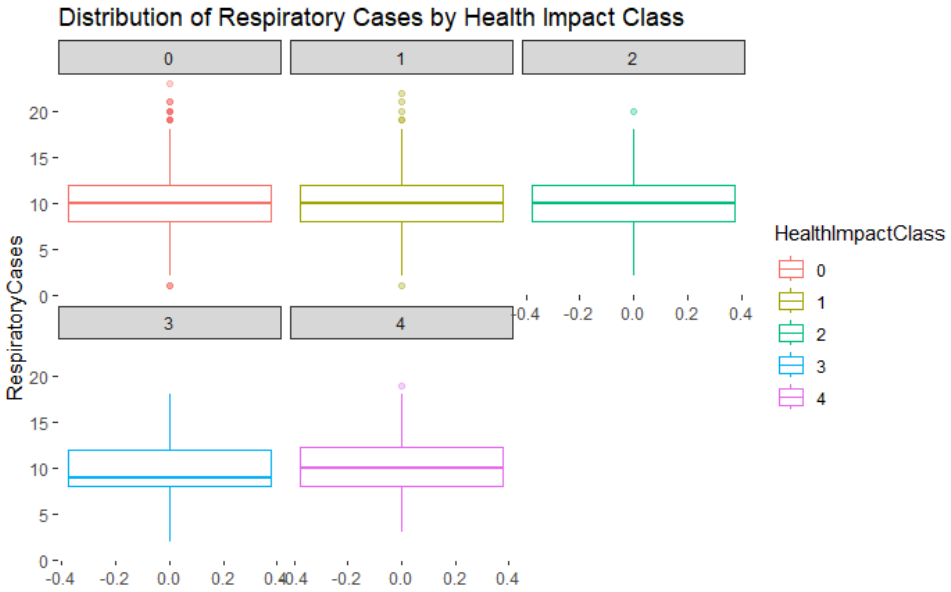
This visualization shows the Ozone concentration in relation to the Health Impact Score, colored by the Health Impact Class.



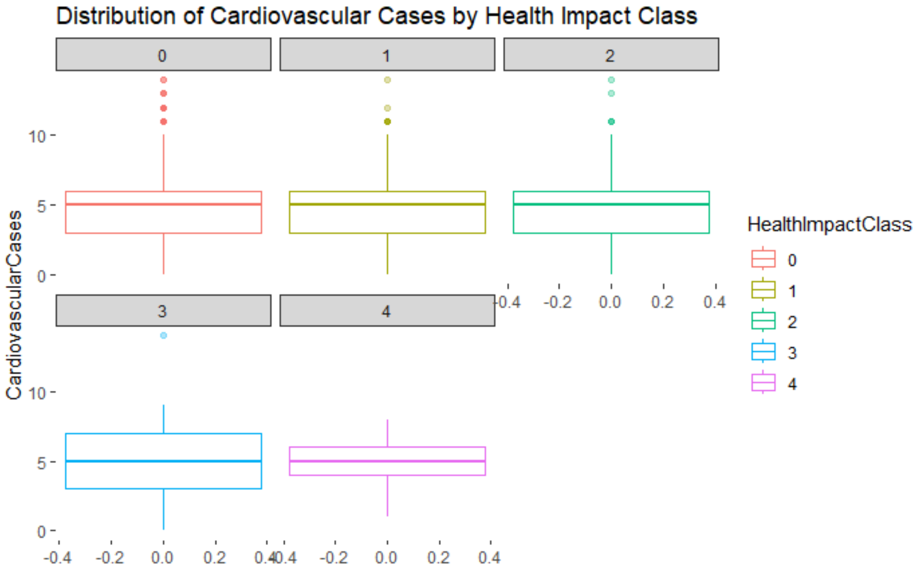
This visualization shows the Sulfur Dioxide concentration in relation to the Health Impact Score, colored by the Health Impact Class.



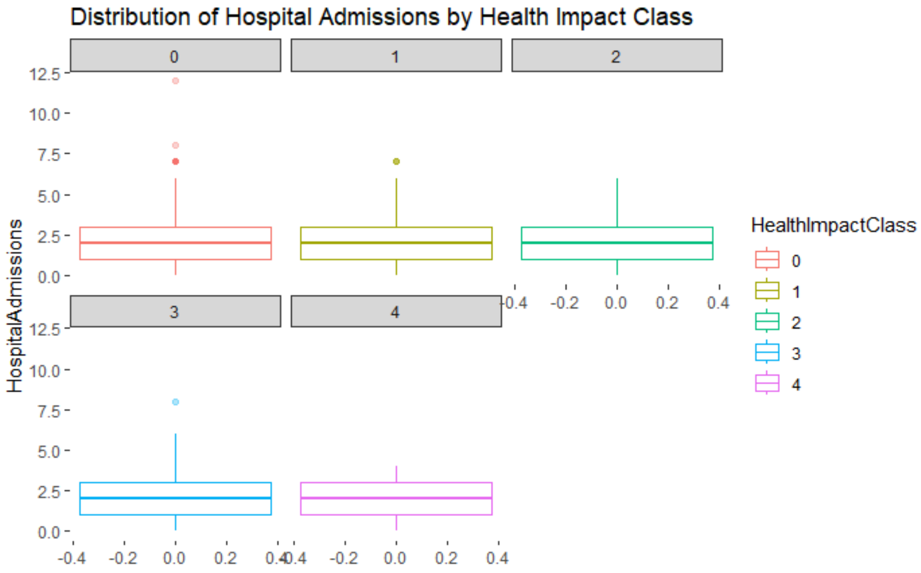
This visualization shows the Nitrogen Dioxide concentration in relation to the Health Impact Score, colored by the Health Impact Class.



This visualization shows the distribution of Respiratory cases by Health Impact Class, seeing a decrease in outliers as the severity of the Health Impact Class decreases.



This visualization shows the distribution of Cardiovascular cases by Health Impact Class, seeing a decrease in outliers as the severity of the Health Impact Class decreases.



This visualization shows the distribution of Hospital Admissions cases by Health Impact Class, seeing a decrease in outliers as the severity of the Health Impact Class decreases.

4. Modeling Process and Interpretation

In order to identify the most important factors for determining health outcomes and air quality, we made use of a few models and tuned them in order to find the most accurate results.

First, we decided to utilize a Random Forest model to discover the highest predictors and their ranking of importance. The Random Forest model was chosen for its random sampling of variables, ensuring that every significant predictor emerged in our model in the situation that there was one particularly strong predictor. The decision tree base model also provides the flexibility to extend to a regression problem, and it allows for the opportunity to display the statistics of one tree to aid interpretability. The initial model gave a Mean Squared Error (MSE) of 9.816 and a Root Mean Squared Error (RMSE) of 3.133. After tuning, the final model gave a MSE of 9.644 and a RMSE of 3.106, so an improvement from the previous model. A Random Forest model is well suited to this problem, as the ensemble technique allows for the emergence of predictors that may not be seen in a normal decision tree, and it returns overall pretty accurate predictions without overfitting (due to the variance introduced). Even so, we wanted to see if we could train a model with even greater predictive power, as the problem at hand requires care and accuracy in the way it's handled.

Secondly, we used an XGBoost model to see if there were any differences in the most important contributors for Health Impact score and air quality. We had to tailor this XGBoost to work with entirely numerical inputs which required us to turn our test and training data into matrices, which did not pose much of a problem. After creating an initial model, it resulted in an

RMSE of 2.818, and the final model ultimately ended with an RMSE of 1.825984. The XGBoost model is very applicable to the problem we have at hand since it has strong predictive power and is a slow learner. We want our insights to be as accurate as possible so that policy makers would be more willing to act to protect public health and air quality. This model is also resilient to overfitting and sensitive to outliers. Considering that air quality and health impact fall closely in a normal range, understanding outliers is important because it allows us to better address the air quality and health impact for those who live in poor air quality zones. Though the model trains slowly and cannot be run in parallel, that did not detract from our project in building a model that can adapt well to new data.

5. Results

The Random Forest model was chosen due to its predictive power, low bias, and ability to produce variable importance. The hyperparameters that were tuned were the number of trees and node size, which dictate the total number of trees in the model and the minimum node size for a split to occur in each tree. After tuning, the values for the hyperparameters were chosen:

- ntree = 1000
- nodesize = 1

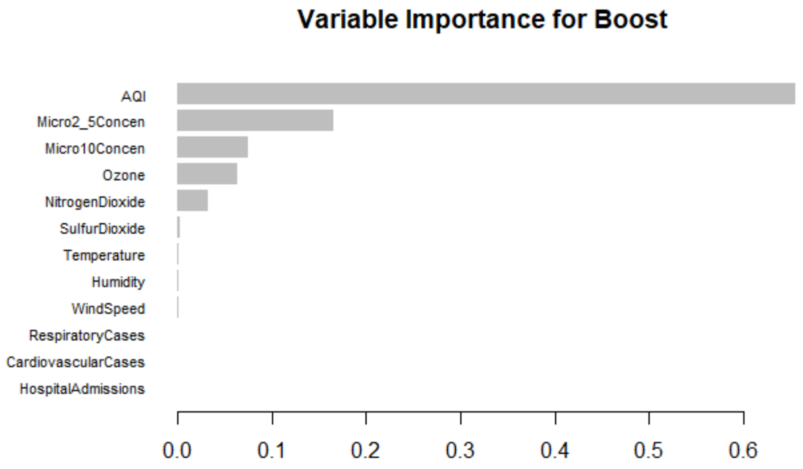
In order to evaluate this model, we looked at mean squared error between training and testing sets, then also calculated root mean squared error for better comparison to other models. Variable importance can be found in the visualizations at the bottom of the document, but are extremely similar to the following model.

With the XGBoost model, we decided to apply it to determining Health Impact Score because of its high accuracy. Due to its high accuracy, it would provide us with the best possible determination of impactful factors for determining air quality and Health Impact Score. To create the best model for our application, we performed hyperparameter tuning, resulting in the following hyperparameters:

- eta = 0.05
- max.depth = 7
- min_child_weight = 15
- gamma = 0.05
- subsample = 0.6
- colsample_bytree = 1
- nrounds = 650
- early_stopping_rounds = 20

In order to determine the performance of our model, we looked at its RMSE for our training and testing sets, as we did not have a validation set. This model had a training RMSE of 0.324107 and a test RMSE of 1.825984.

Looking at the variable importance for both of the models yielded us with the same results in determining that AQI, micro_2.5, micro_10, Ozone, and Nitrogen Dioxide were all impactful in determining the HealthImpactScore. The AQI had a much larger effect on predicting this score, and that is likely because it is a measure of these particles, thus it may be absorbing some of their effect. Additionally, based on the RMSE we can see that the XGBoost model outperformed the bagging model.



Based on our results and being able to successfully identify some meaningful contributors, we have determined so recommendations for policy makers. Given our findings, policymakers should make an effort to reduce particle pollution the most, followed by ozone, then nitrogen dioxide. Furthermore, The EPA notes that particle pollution can originate from a variety of sources, including construction sites, fires, power plants, and industry. Policymakers could strengthen regulations on the minimum distance from the nearest residential or public use area that industry/power plants can be built or operated in, as well as prioritizing efficient construction and fire prevention.

In addition to Policy makers, researchers have a responsibility to continue developing filtration systems for manufacturing, industrial, and energy plants to minimize the amount of particulate matter they release into the air. Furthermore, advanced research should be performed on helping prevent ozone deterioration and wild fires that also produce harmful levels of particulate matter.

Finally, individual citizens should monitor the Air Quality Index of their area, and on poor air quality days, they can wear masks or other protective equipment to filter particles from the air they breathe. They can also invest into air purifiers or other home filtration systems if they have the means to do so.

6. Conclusion and Future Work

Air quality is something that has an effect on the entire population, but only a few people are in the positions of power to deal with it. Predicting the health impact on a group based on the factors of air quality would create a better idea of what constitutes dangerous levels of pollutants, as well as identify the most important target factors for improvement. With predictors relating to air quality, weather conditions, health metrics, and the linked outcomes of Health Impact Class/Score (multiclass/numeric), we opted to use the numeric outcome to create a regression problem. The data was unbalanced, with a greater number of high health impact samples, and this was dealt with in both the train-test split and models chosen.

Both a Random Forest and XGBoost model were created, prioritizing low bias and high predictive power. For each one, an initial model was created then tuned, and a final model was created using the tuned hyperparameters. Both models saw an improvement in root mean square error (RMSE) from the initial to the tuned models, with the tuned XGBoost model performing the best with an RMSE of 1.825984.

The ranking of variable importance was consistent with each model. The best predictor by far was AQI, or air quality index. The application of this would be predicting the health impact on a group based on the overall air quality in the area. This also gives the individual a metric to monitor to lower their exposure to air of a dangerous air quality level. The following predictors were the concentration of particulate matter with differently sized diameters, then the concentration of Ozone, then Nitrogen Dioxide. These give a ranking of priority for policymakers and researchers to target in order to improve the air quality in their areas.

Given more time and resources, we may also incorporate a multiclass prediction model, using the other outcome variable. This would allow for a different sort of interpretation, separating the outcomes into ranges to delve deeper into. We could also make it easier for a user to predict the health impact for an area using their own input with the creation of something like a Shiny app. Lastly, we could test our model on real-life data, particularly from areas such as Lubbock, Texas, using publicly available records.

Bibliography

Castelli, Mauro, Clemente, Fabiana Martins, Popovič, Aleš, Silva, Sara, Vanneschi, Leonardo, A Machine Learning Approach to Predict Air Quality in California, *Complexity*, 2020, 8049504, 23 pages, 2020. https://doi.org/10.1155/2020/8049504

FIRST STREET TECHNOLOGY, INC. (n.d.). *Lubbock, TX Poor Air Quality Map and forecast*. firststreet.org. https://firststreet.org/city/lubbock-tx/4845000_fsid/air

Iskandaryan, Ditsuhi, Francisco Ramos, and Sergio Trilles. "Air quality prediction in smart cities using machine learning technologies based on sensor data: a review." *Applied Sciences* 10.7 (2020): 2401. https://www.researchgate.net/profile/Sachin-Bhoite/publication/335911816_Air_Quality_Prediction_using_Machine_Learning_Algorithms/links/5d836662299bf1996f77746f/Air-Quality-Prediction-using-Machine-Learning-Algorithms.pdf

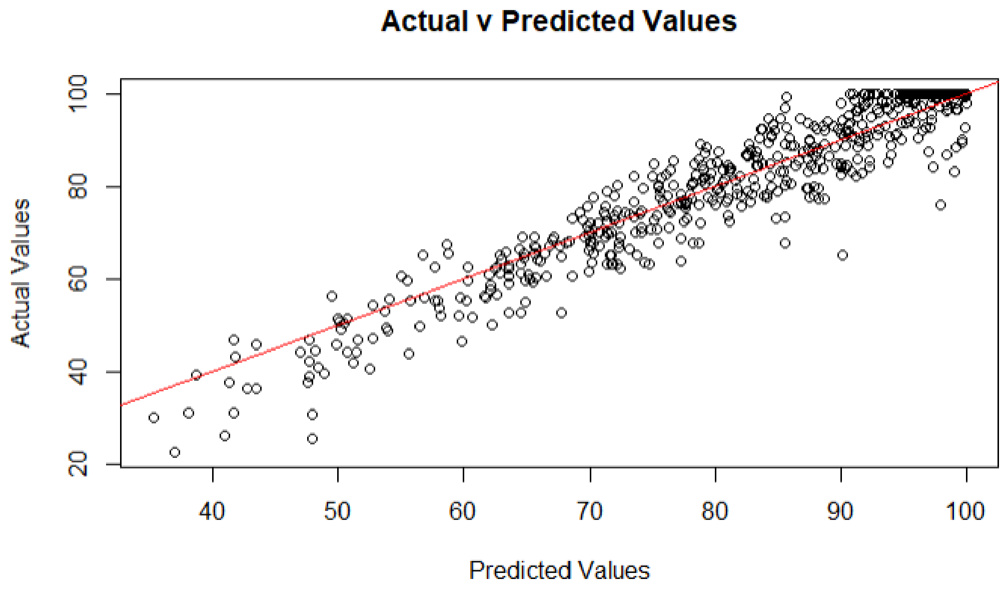
Kharoua, Rabie El. "🌍 Air Quality and Health Impact Dataset 🌍." *Kaggle*, 12 June 2024, www.kaggle.com/datasets/rabieelkharoua/air-quality-and-health-impact-dataset.

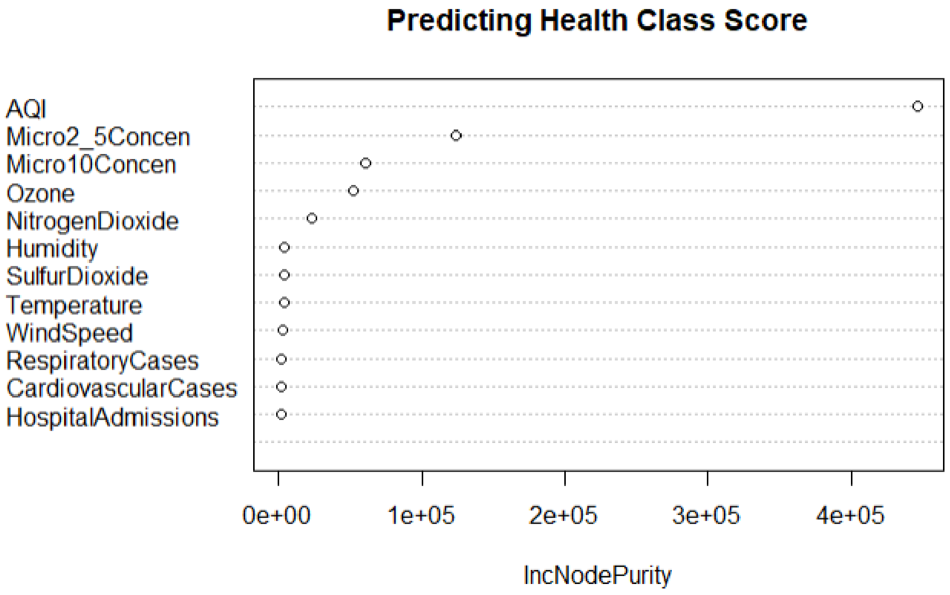
T. M. Chiwewe and J. Ditsela, "Machine learning based estimation of Ozone using spatio-temporal data from air quality monitoring stations," 2016 IEEE 14th International Conference on Industrial Informatics (INDIN), Poitiers, France, 2016, pp. 58-63, https://ieeexplore.ieee.org/abstract/document/7819134.

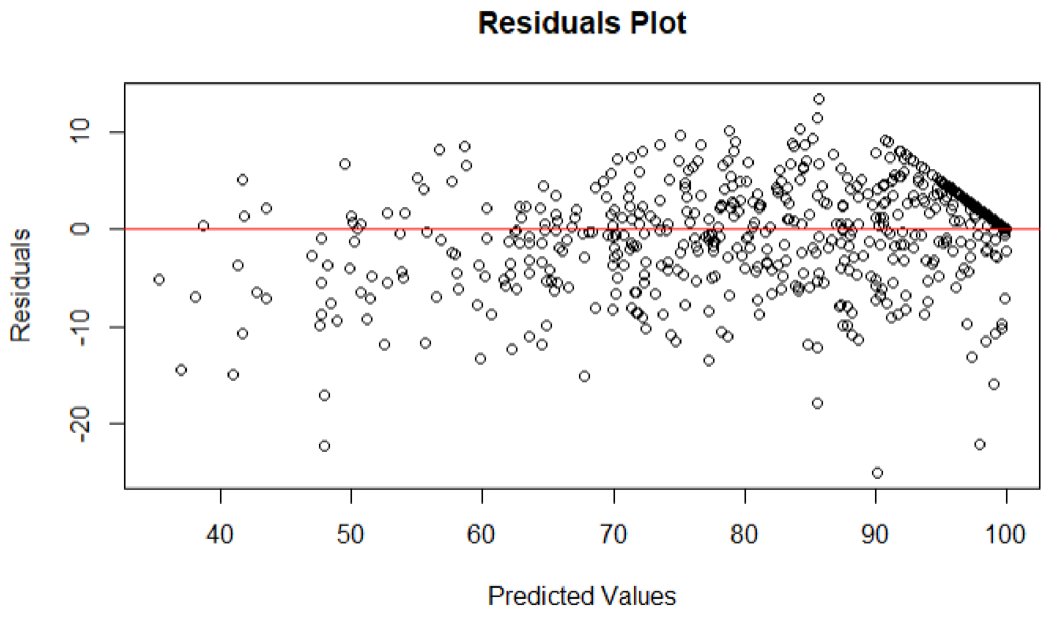
https://www.niehs.nih.gov/health/topics/agents/air-pollution

Visualizations - Model Performance Evaluation

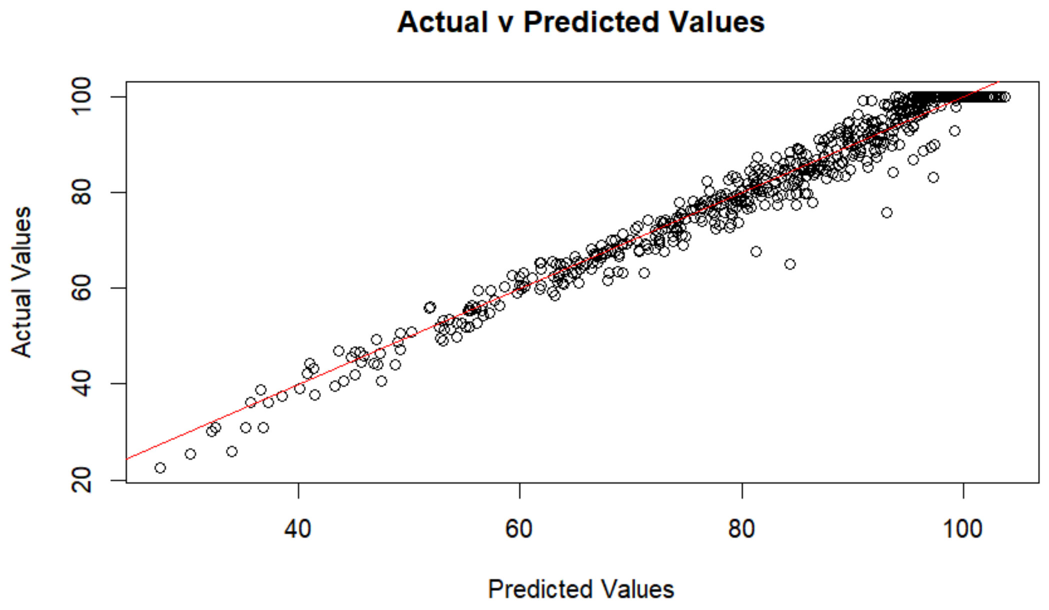
Random Forest model:

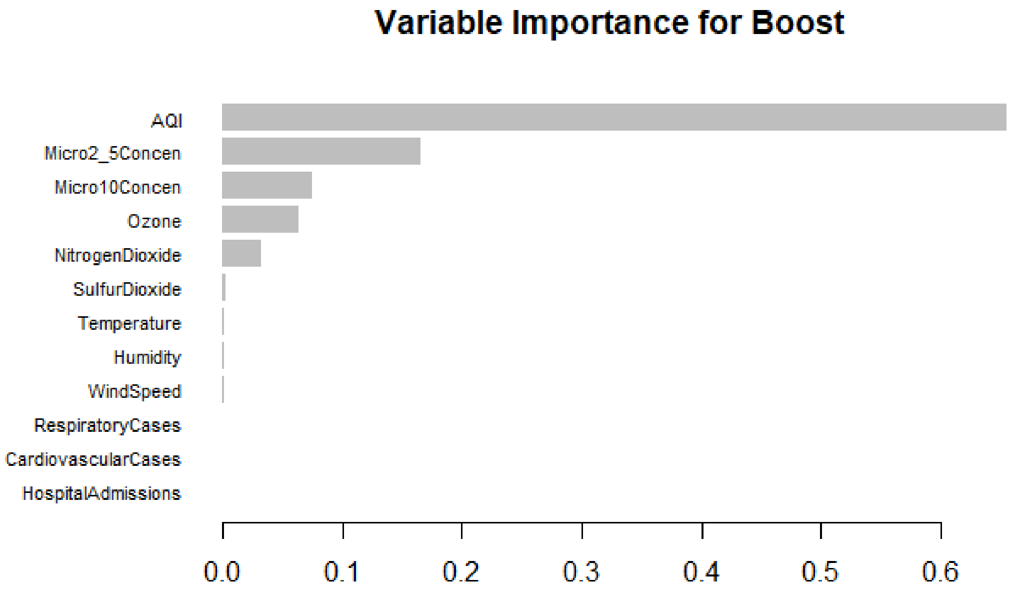


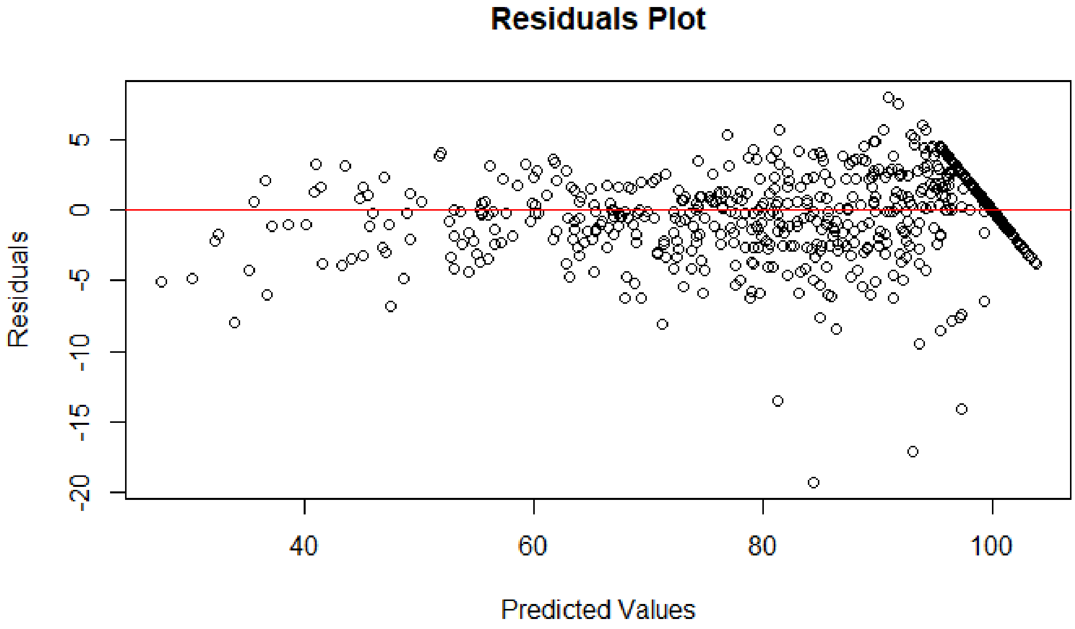




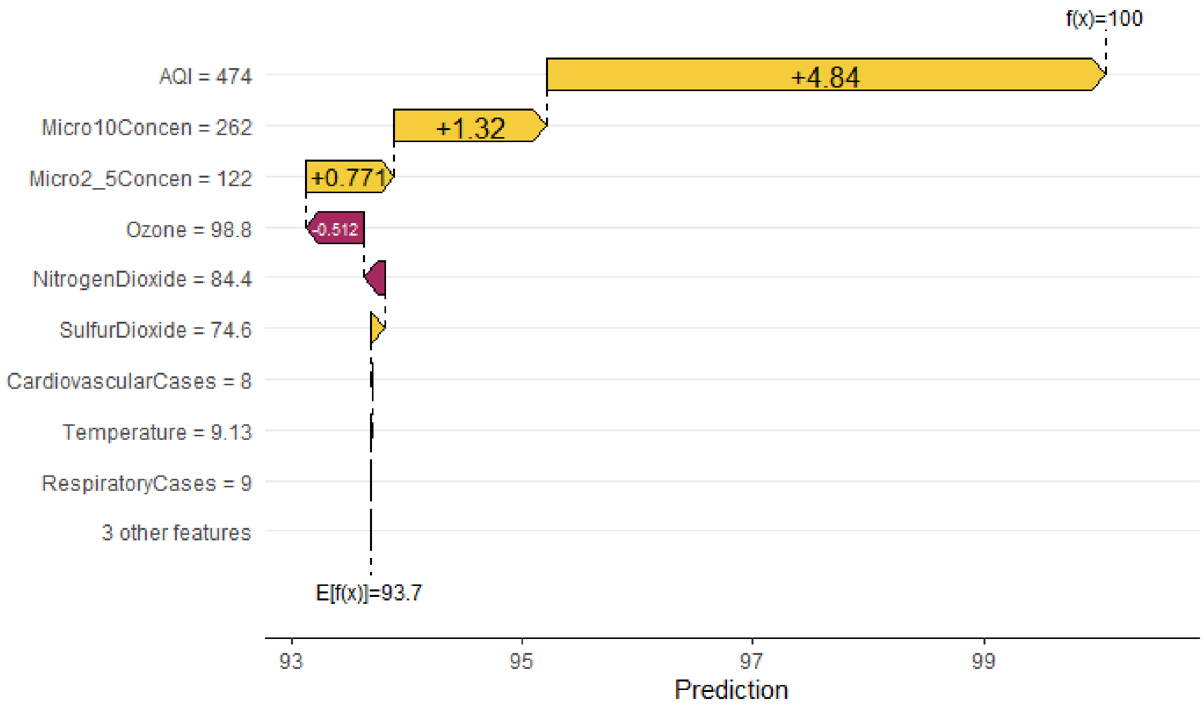
XGBoost model:




Variable Importance for Boost
AQI
Micro2_5Concen
Micro10Concen
Ozone
NitrogenDioxide
SulfurDioxide
Temperature
Humidity
WindSpeed
RespiratoryCases
CardiovascularCases
HospitalAdmissions
0.0
0.1
0.2
0.3
0.4
0.5
0.6


Residuals Plot
Residuals
5
0
-5
-10
-15
-20
40
60
80
100
Predicted Values

One sample SHAP example (XGBoost):


f(x)=100
AQI = 474
Micro10Concen = 262
Micro2_5Concen = 122
Ozone = 98.8
NitrogenDioxide = 84.4
SulfurDioxide = 74.6
CardiovascularCases = 8
Temperature = 9.13
RespiratoryCases = 9
3 other features
E[f(x)]=93.7
Prediction