Can Air Quality Parameters be Used to Predict Respiratory Disease Incidence?

Adriana Ramírez Flores
Alan Andrews
Trang Khong

Problem Statement

- To determine if the incidence of respiratory diseases can be accurately predicted based on various air quality parameter measurements using machine learning estimator methods.

- The 'Incidence' rate is the number of new cases divided by the population at the middle of the year for that age group and state.

- Predict the incidence rate across 25 states with six different age groups:
    - 65-69
    - 70-74
    - 75-79
    - 80-84
    - 85-90
    - 90-94

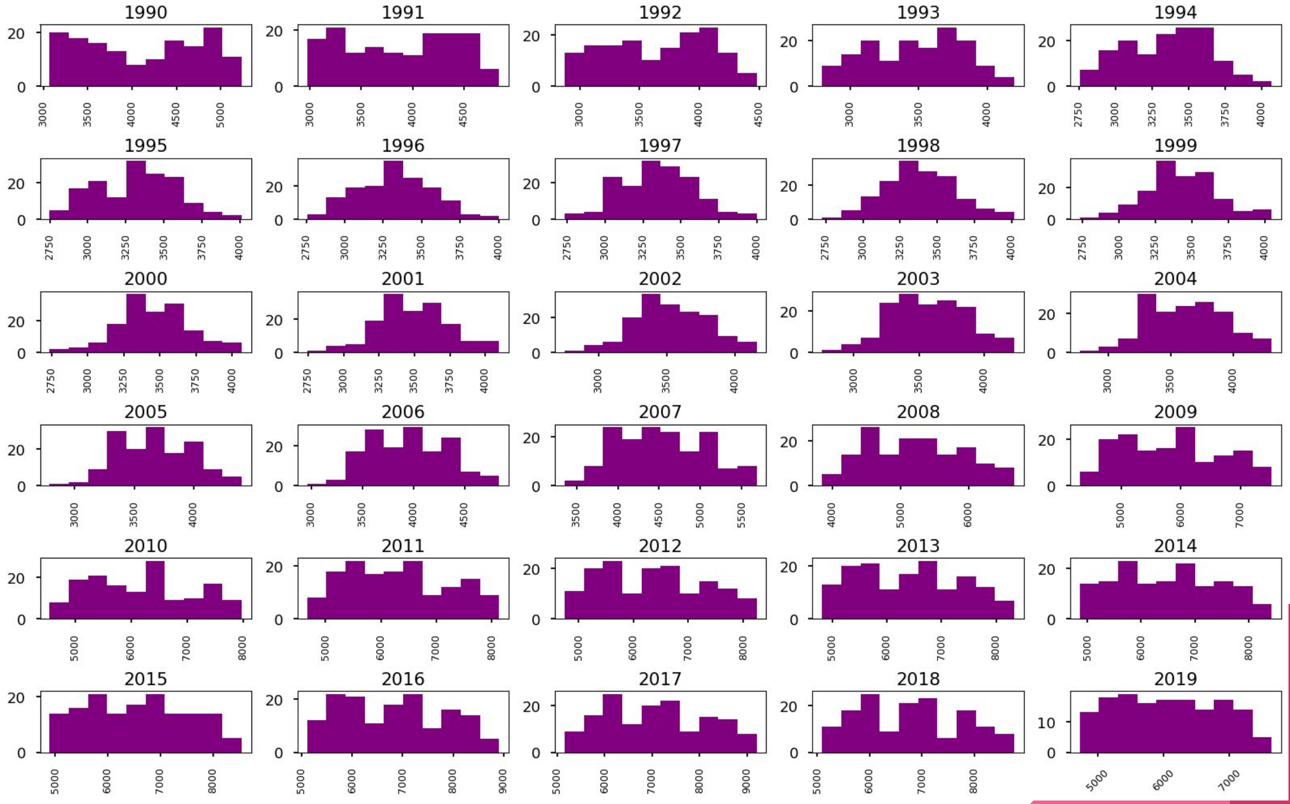
Analyzed data

- **train.csv** - consisting of incidence rates of respiratory disease in each state, by year, by age bracket.
- **test.cs**v - same as train without the incidence rates.
- 4 supplemental **parquet files** containing 167 parameters pertaining to the following air pollutants:
  - Lead (Pb)
  - Hazardous Air Pollutants (HAPs)
  - Various Nitrogen Oxide compounds (NOs)
  - Volatile organic compounds. (VOCs)

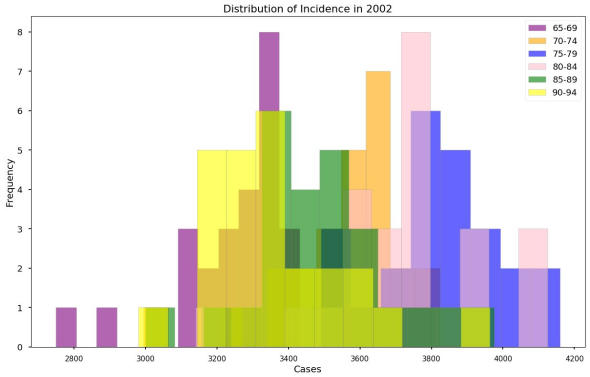
Processing The Parquet Files

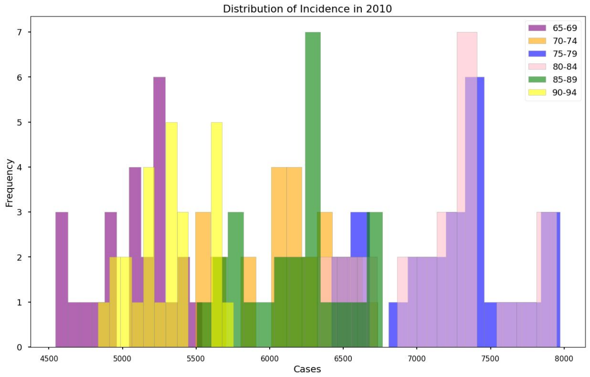
- The training data spans from 1990 to 2019, while the data in the parquet files goes back to 1980.
- Columns with more than 20% missing data were eliminated from the parquet files
- The data was aggregated by year and state.
- Lag columns were created for 2, 5, and 8 years before the date in each row to take into consideration latent effects of pollutants on respiratory disease incidence rates.
- Data from the parquet files was combined with the training and testing datasets by state/year.

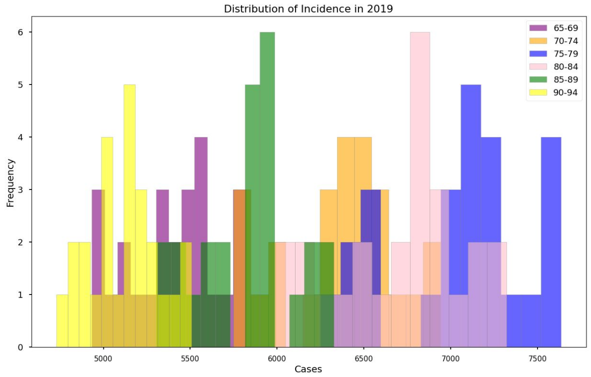
Yearly incidence of respiratory diseases



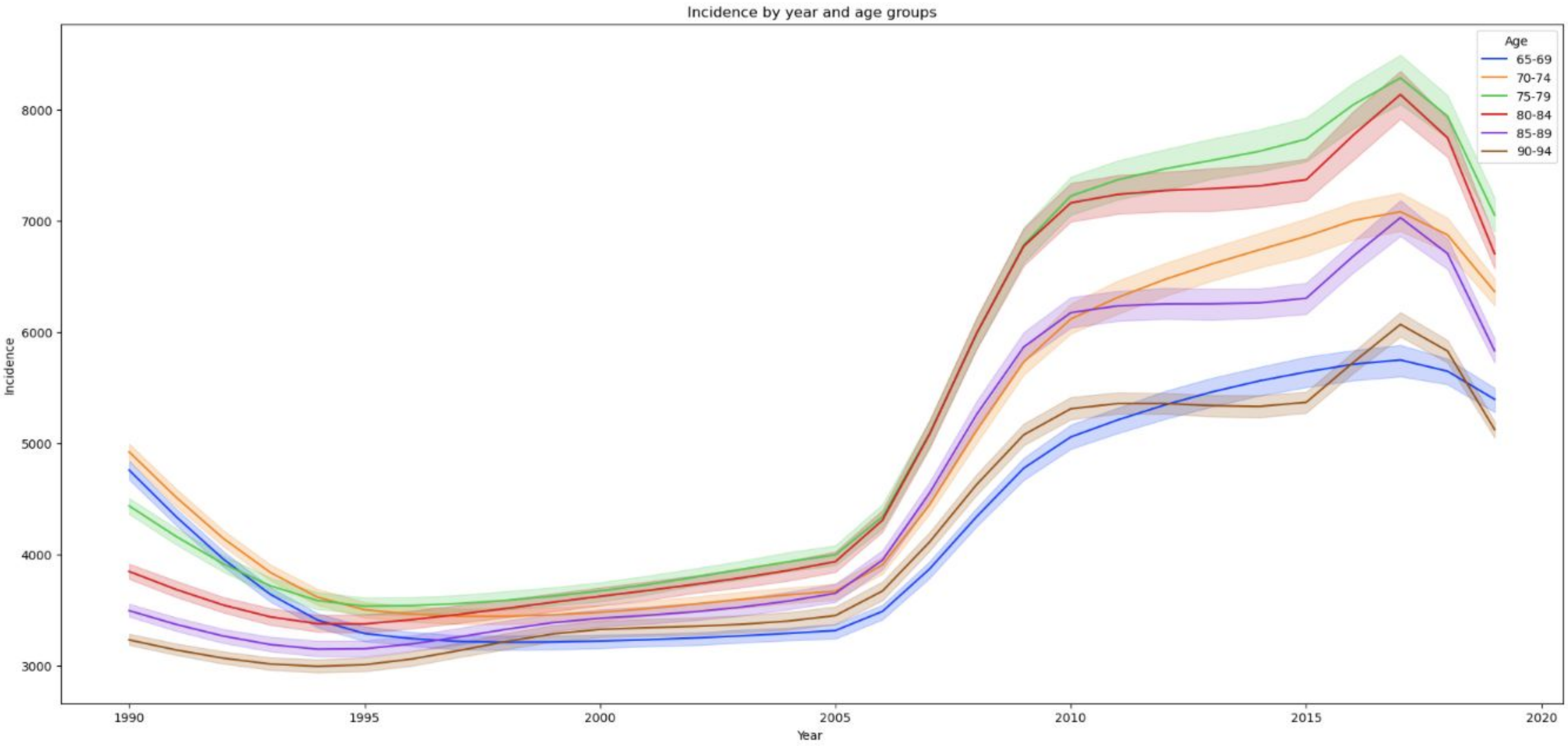
Incidence by age group is different



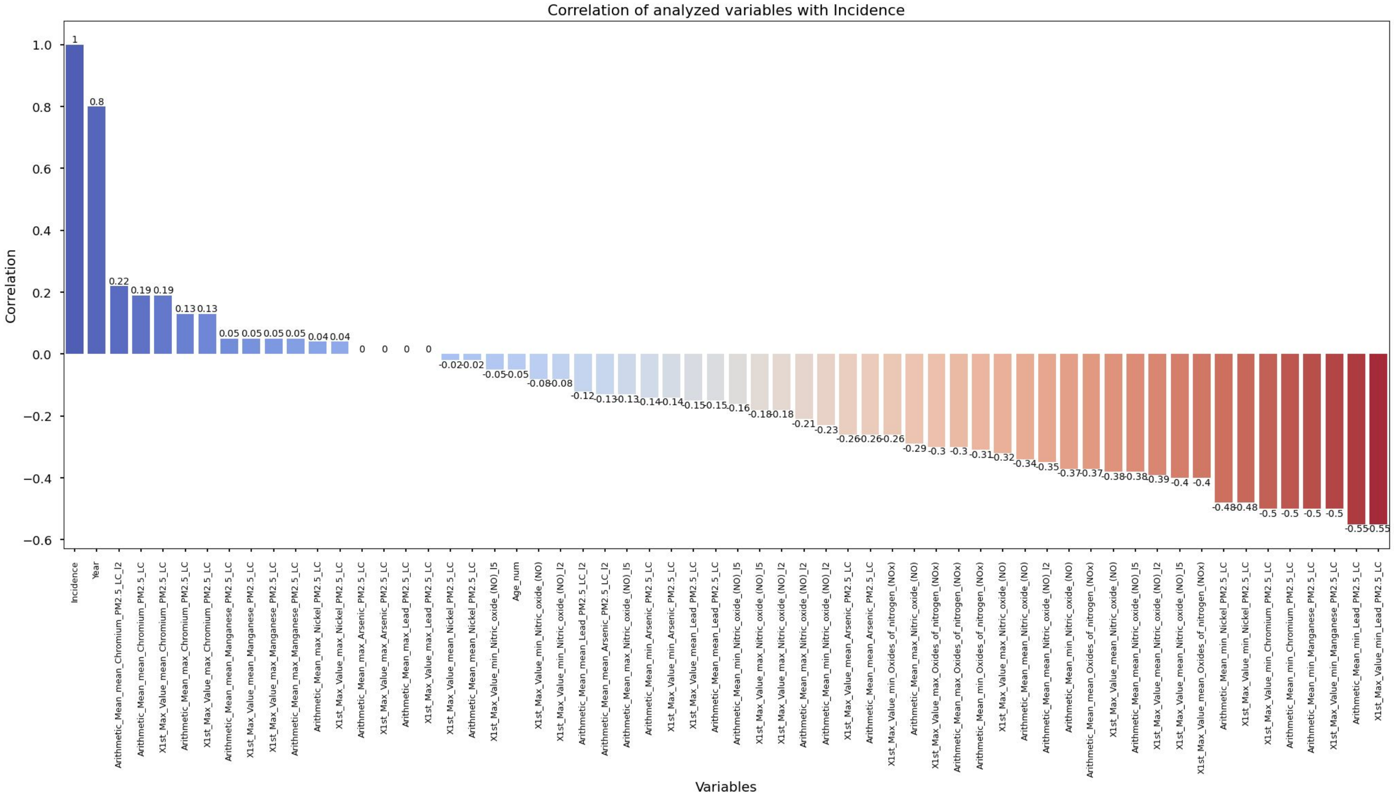




Incidence of respiratory diseases across time



Correlation of the different pollutants with Incidence



Data Imputing methods

- Mean
- Zero
- Iterative imputation with Linear Regression
- Iterative imputation with BayesianRidge
- Knn imputer

Model Selection

- We began exploring less complex models like **Linear Regression** with and without **Lasso regularization**.
However, we knew these models would be outperformed since linear regression tends to overfit and is biased. With Lasso regularization, the models generally are less overfitted but bias increases.

- Later, we implemented models with more complexity like **Random Forest Regressor, Extra Trees Regressor, AdaBoost Regressor** and **Bagging Regressor**. We were interested in these models because they consistently give good predictions and the models are not overfitted.

- Finally, more complex models like **Neural Networks** and **Gradient Boosting** were included.

Metrics

All models were compared to a baseline model of imputed mean, incidence rate evaluated with **R Squared** and **Root Mean Squared Error (RMSE)**.

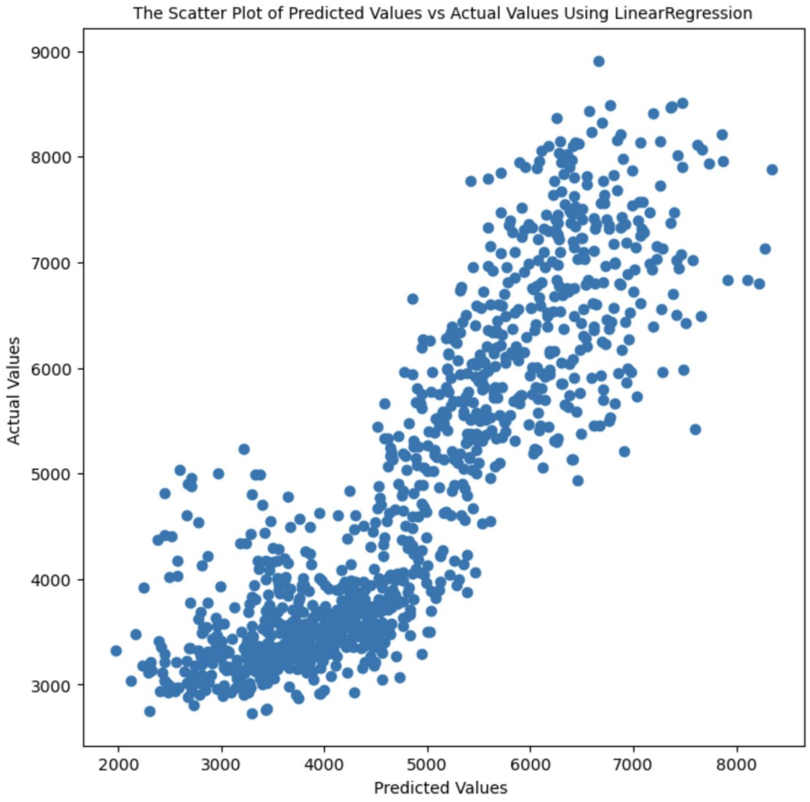
We selected these metrics because R^2 states how much of the variation of Y is explained by the models, and RMSE is useful to have a measure in units of Y of how far in average our predictions are from the observed values.

Iterative Imputer with Linear Regression

Train score : 0.769

Test score: 0.754

RMSE score: 761.718

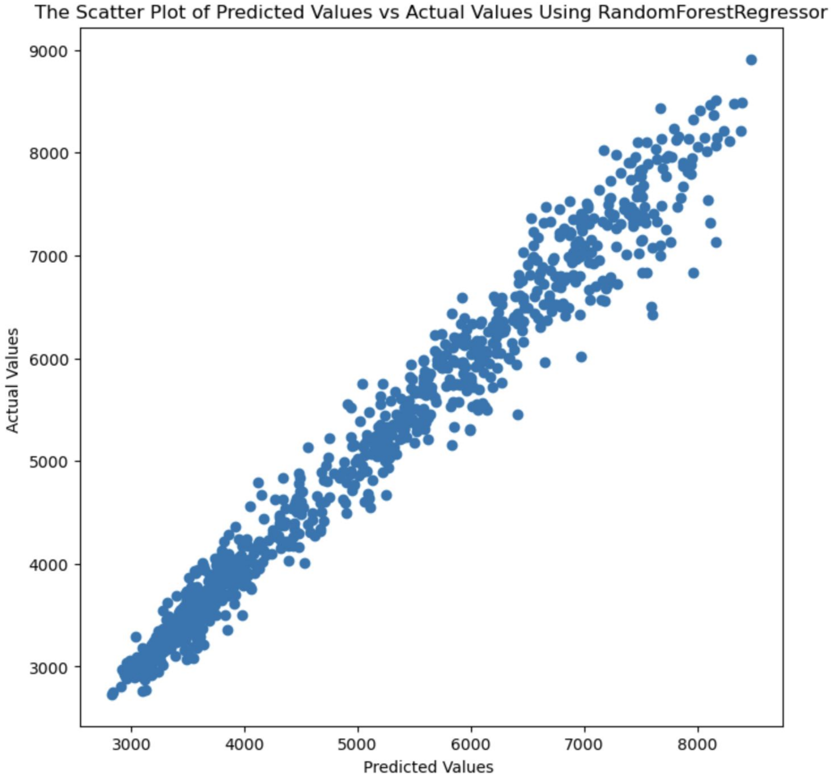


Random Forest Regressor

Train score: 0.997

Test score: 0.976

RMSE: 239.044

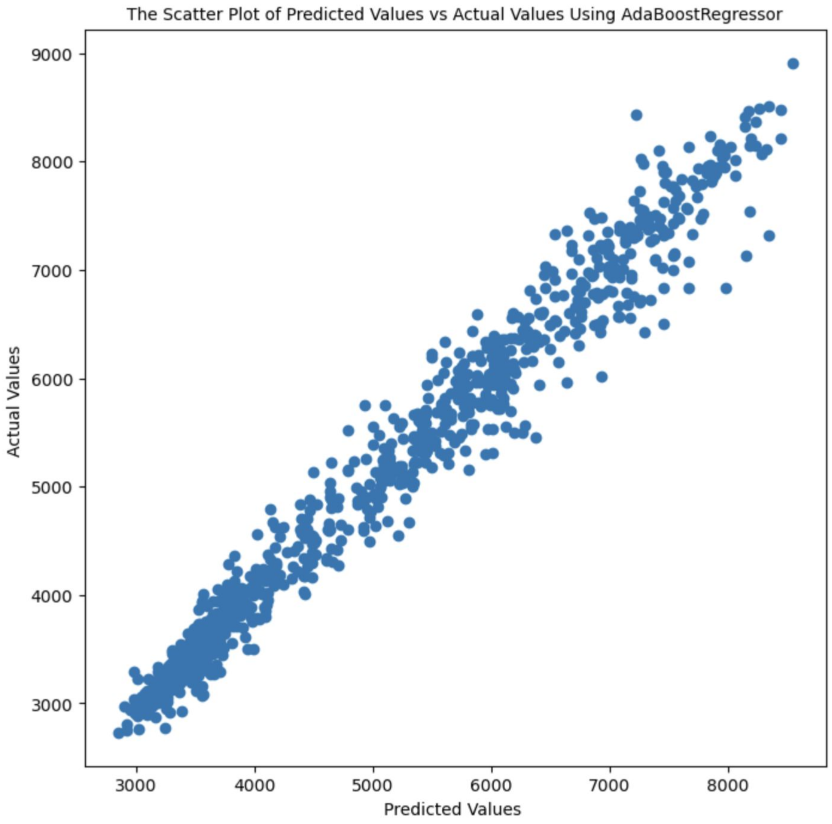


AdaBoost Regressor

Train score: 0.9996

Test score: 0.976

RMSE: 239.181

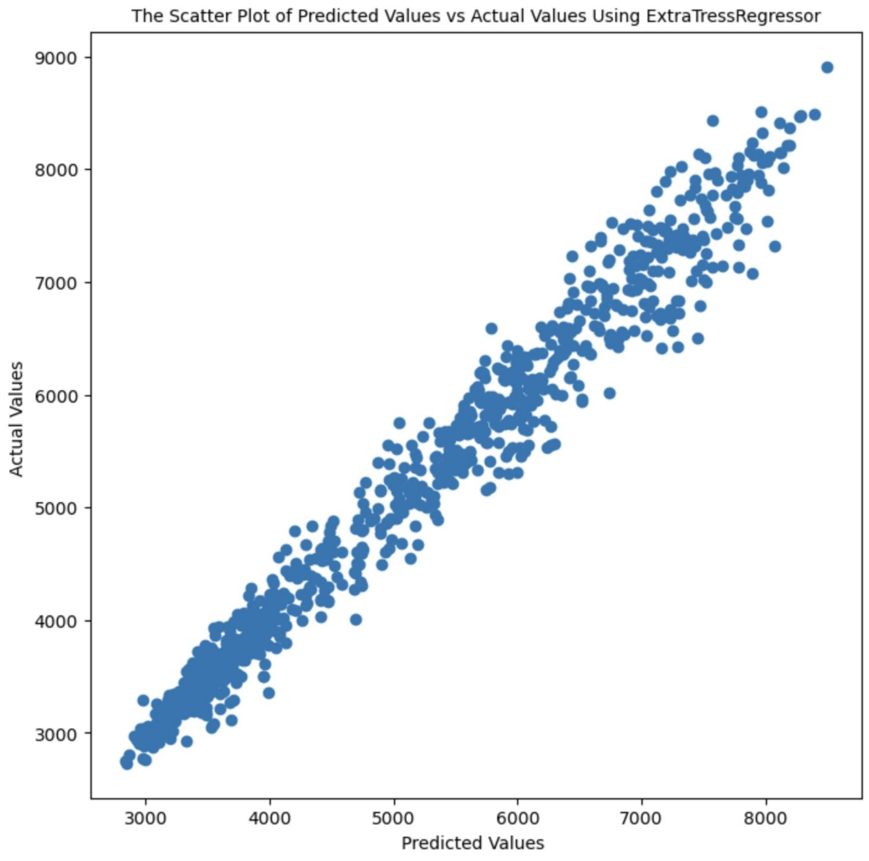


Extra Trees Regressor

Train score: 0.9999

Test score: 0.977

RMSE: 234.437

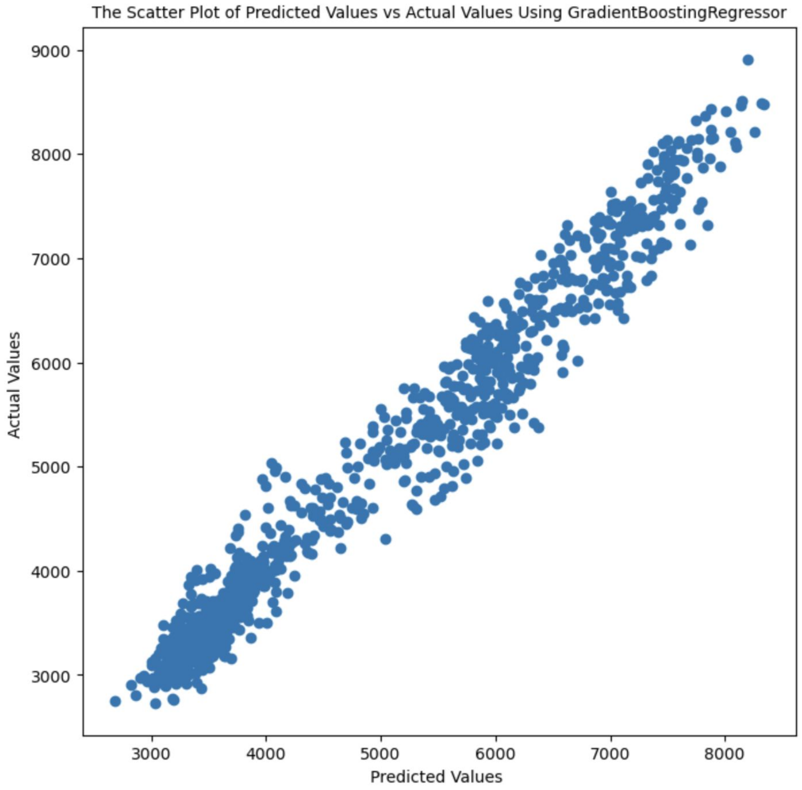


Gradient Boosting Regressor

Train score: 0.976

Test score: 0.969

RMSE: 272.229

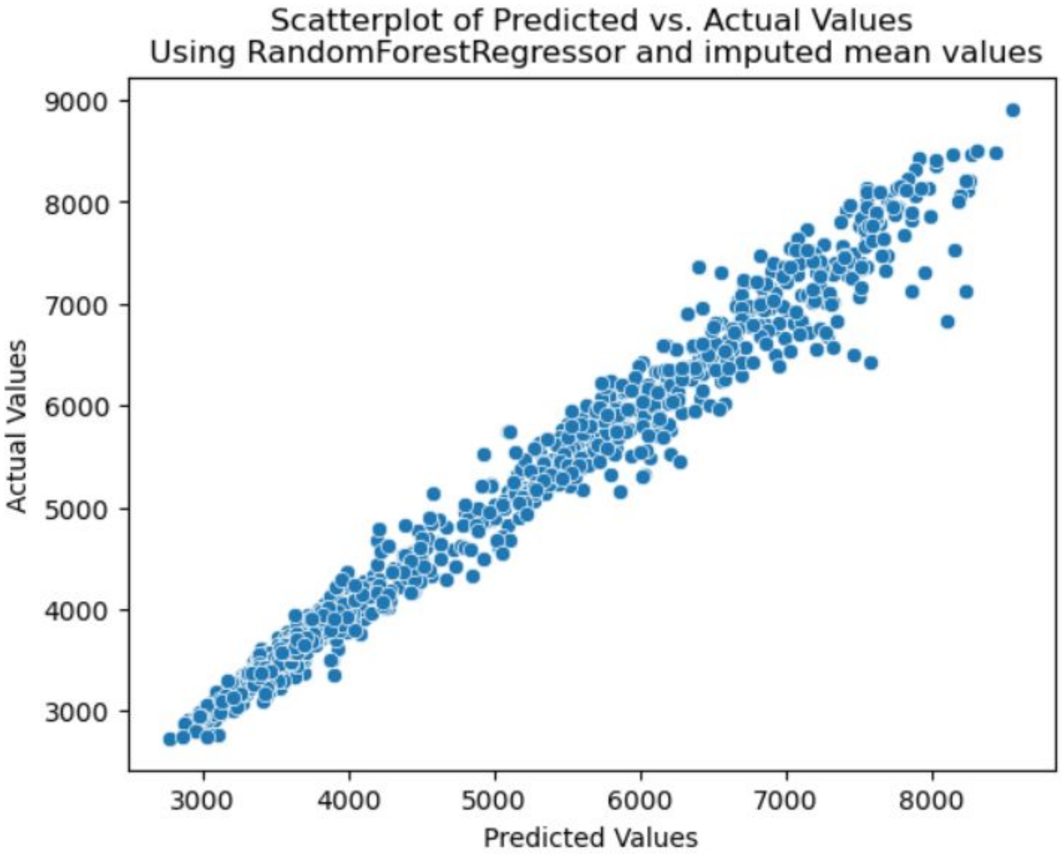


Imputed Mean and RandomForestRegressor

Train score: 0.997

Test score: 0.981

RMSE: 213.523



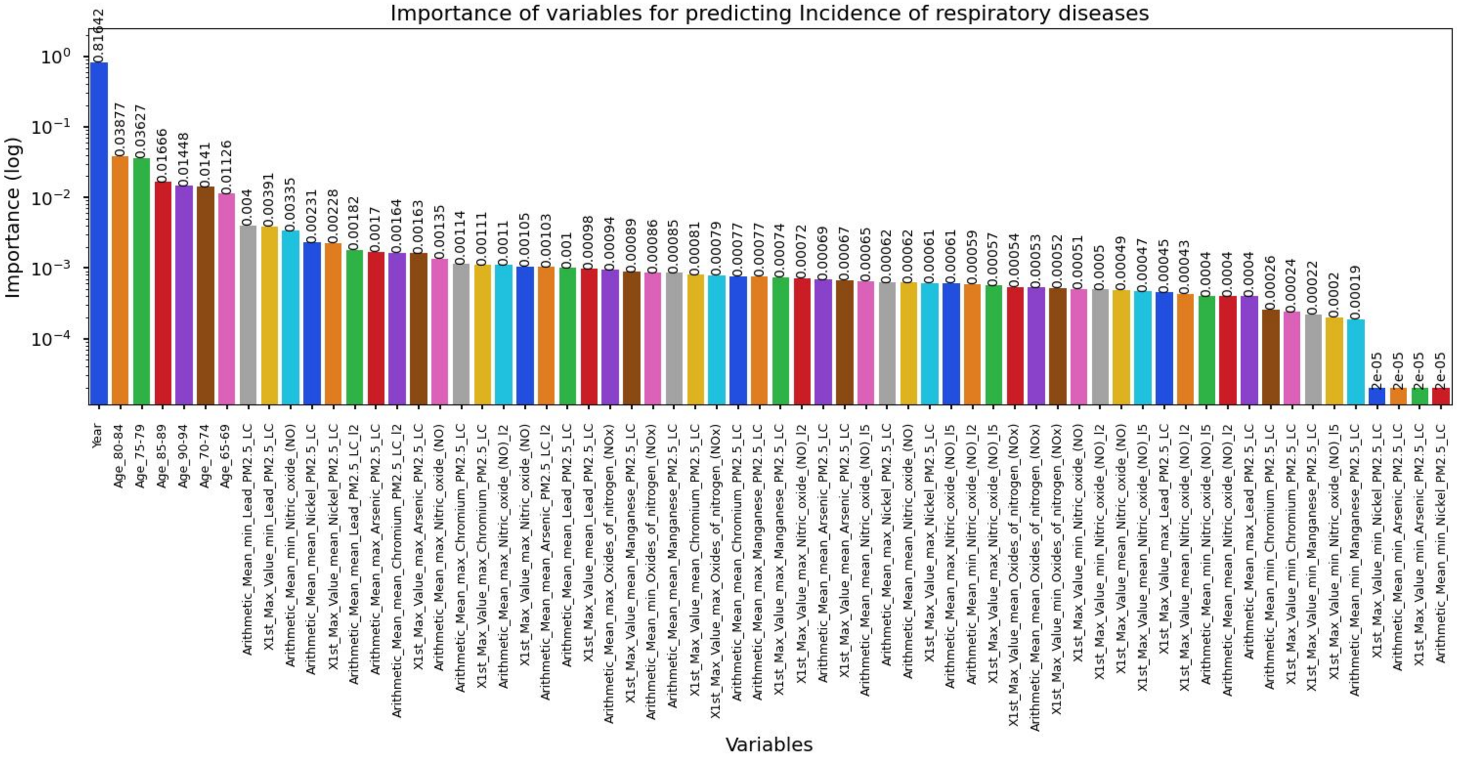
Model results with different data imputing methods

Imputation Method	Model	R2 Score Train	R2 Score Test	RMSE
Baseline	Mean	0	0	1,514.918
Iterative imputer with BayesianRidge	AdaBoostRegressor	0.999	0.976	239.181
	RandomForestRegressor	0.997	0.976	237.638
	ExtraTreeRegressor	0.999	0.977	234.437
	GradientBoostingRegressor	0.976	0.969	272.229
Iterative imputer with Linear Regression	RandomForestRegressor	0.973	0.908	465.222
	Linear Regression	0.769	0.754	761.718
	Lasso Regression	0.773	0.773	769.624
Knn Imputer with scaled data	RandomForestRegressor	0.997	0.979	216.975
	RandomForest Gridsearch (Best Params)	0.997	0.980	213.389
	AdaBoost, DecissionTreeRegressor	0.999	0.977	227.672
	Neural Network	N/A	N/A	1,153.279
Zero	Lasso Regression	0.862	0.849	
Mean	Lasso Regression	0.866	0.853	589.632
	Random Forest	0.997	0.981	213.523

This was the best model in the test data for Kaggle

Feature importances

For the RandomForestRegressor model with best parameters obtained through GridSearch (with KNN imputation)



For visual representation the importance of the variables had to be scaled with logarithm. To see original values see Appendix 1.

Discussion

- Inclusion of Year variable in the models, made them perform better at predicting the incidence of respiratory disease for past years. However, we question whether it is useful to include year in models for making predictions into the future.
- The training dataset includes large, populous states like California and Texas, while the test dataset has smaller states. This likely has an impact on model performance.
- For this specific data, imputing the mean in the missing data gave excellent results because the range for each pollutant is fairly small and close zero.

Conclusions

- Model complexity does not equate to better performance.
- The method used for data imputing has an important effect on model training and performance.
- The ‘best’ performing model in this case was RandomForestRegressor with mean imputed values in the training dataset.
- The minimum of the monthly arithmetic mean for the year of concentration of Lead (PM 2.5 LC) and Nitric Oxide (NO) are the most important pollutants when predicting the occurrence of respiratory diseases.

Future Steps

- Implement models we haven’t tried yet: Pre-made NNs, XGBoost.
- Continue tuning hyperparameters.
- Try different methods for imputing the data.
- Include all possible lag variables.

Appendix

Bibliography

Kaggle competition:
https://www.kaggle.com/competitions/air-toxicity-and-chronic-respiratory-diseases-us/overview

Feature importances

- Feature importances for the RandomForestRegressor model with best parameters obtained through GridSearch (with KNN imputation)
- For a best visual representation in slide 18 the importance of variables was scaled with logarithm

	Variable	Importance
0	Year	0.81642
61	Age_80-84	0.03877
60	Age_75-79	0.03627
62	Age_85-89	0.01666
63	Age_90-94	0.01448
59	Age_70-74	0.01410
58	Age_65-69	0.01126
8	Arithmetic_Mean_min_Lead_PM2.5_LC	0.00400
23	X1st_Max_Value_min_Lead_PM2.5_LC	0.00391
36	Arithmetic_Mean_min_Nitric_oxide_(NO)	0.00335
5	Arithmetic_Mean_mean_Nickel_PM2.5_LC	0.00231
20	X1st_Max_Value_mean_Nickel_PM2.5_LC	0.00228
33	Arithmetic_Mean_mean_Lead_PM2.5_LC_I2	0.00182
11	Arithmetic_Mean_max_Arsenic_PM2.5_LC	0.00170
32	Arithmetic_Mean_mean_Chromium_PM2.5_LC_I2	0.00164
26	X1st_Max_Value_max_Arsenic_PM2.5_LC	0.00163
38	Arithmetic_Mean_max_Nitric_oxide_(NO)	0.00135
12	Arithmetic_Mean_max_Chromium_PM2.5_LC	0.00114
27	X1st_Max_Value_max_Chromium_PM2.5_LC	0.00111
50	Arithmetic_Mean_max_Nitric_oxide_(NO)_I2	0.00110
44	X1st_Max_Value_max_Nitric_oxide_(NO)	0.00105
31	Arithmetic_Mean_mean_Arsenic_PM2.5_LC_I2	0.00103
3	Arithmetic_Mean_mean_Lead_PM2.5_LC	0.00100
18	X1st_Max_Value_mean_Lead_PM2.5_LC	0.00098
39	Arithmetic_Mean_max_Oxides_of_nitrogen_(NOx)	0.00094
19	X1st_Max_Value_mean_Manganese_PM2.5_LC	0.00089
37	Arithmetic_Mean_min_Oxides_of_nitrogen_(NOx)	0.00086
4	Arithmetic_Mean_mean_Manganese_PM2.5_LC	0.00085
17	X1st_Max_Value_mean_Chromium_PM2.5_LC	0.00081
45	X1st_Max_Value_max_Oxides_of_nitrogen_(NOx)	0.00079
2	Arithmetic_Mean_mean_Chromium_PM2.5_LC	0.00077
14	Arithmetic_Mean_max_Manganese_PM2.5_LC	0.00077

	Variable	Importance
29	X1st_Max_Value_max_Manganese_PM2.5_LC	0.00074
56	X1st_Max_Value_max_Nitric_oxide_(NO)_I2	0.00072
1	Arithmetic_Mean_mean_Arsenic_PM2.5_LC	0.00069
16	X1st_Max_Value_mean_Arsenic_PM2.5_LC	0.00067
47	Arithmetic_Mean_mean_Nitric_oxide_(NO)_I5	0.00065
15	Arithmetic_Mean_max_Nickel_PM2.5_LC	0.00062
34	Arithmetic_Mean_mean_Nitric_oxide_(NO)	0.00062
30	X1st_Max_Value_max_Nickel_PM2.5_LC	0.00061
51	Arithmetic_Mean_max_Nitric_oxide_(NO)_I5	0.00061
48	Arithmetic_Mean_min_Nitric_oxide_(NO)_I2	0.00059
57	X1st_Max_Value_max_Nitric_oxide_(NO)_I5	0.00057
41	X1st_Max_Value_mean_Oxides_of_nitrogen_(NOx)	0.00054
35	Arithmetic_Mean_mean_Oxides_of_nitrogen_(NOx)	0.00053
43	X1st_Max_Value_min_Oxides_of_nitrogen_(NOx)	0.00052
42	X1st_Max_Value_min_Nitric_oxide_(NO)	0.00051
54	X1st_Max_Value_min_Nitric_oxide_(NO)_I2	0.00050
40	X1st_Max_Value_mean_Nitric_oxide_(NO)	0.00049
53	X1st_Max_Value_mean_Nitric_oxide_(NO)_I5	0.00047
28	X1st_Max_Value_max_Lead_PM2.5_LC	0.00045
52	X1st_Max_Value_mean_Nitric_oxide_(NO)_I2	0.00043
49	Arithmetic_Mean_min_Nitric_oxide_(NO)_I5	0.00040
46	Arithmetic_Mean_mean_Nitric_oxide_(NO)_I2	0.00040
13	Arithmetic_Mean_max_Lead_PM2.5_LC	0.00040
7	Arithmetic_Mean_min_Chromium_PM2.5_LC	0.00026
22	X1st_Max_Value_min_Chromium_PM2.5_LC	0.00024
24	X1st_Max_Value_min_Manganese_PM2.5_LC	0.00022
55	X1st_Max_Value_min_Nitric_oxide_(NO)_I5	0.00020
9	Arithmetic_Mean_min_Manganese_PM2.5_LC	0.00019
25	X1st_Max_Value_min_Nickel_PM2.5_LC	0.00002
6	Arithmetic_Mean_min_Arsenic_PM2.5_LC	0.00002
21	X1st_Max_Value_min_Arsenic_PM2.5_LC	0.00002
10	Arithmetic_Mean_min_Nickel_PM2.5_LC	0.00002