

Predicting Air Quality: From Correlation to Machine Learning

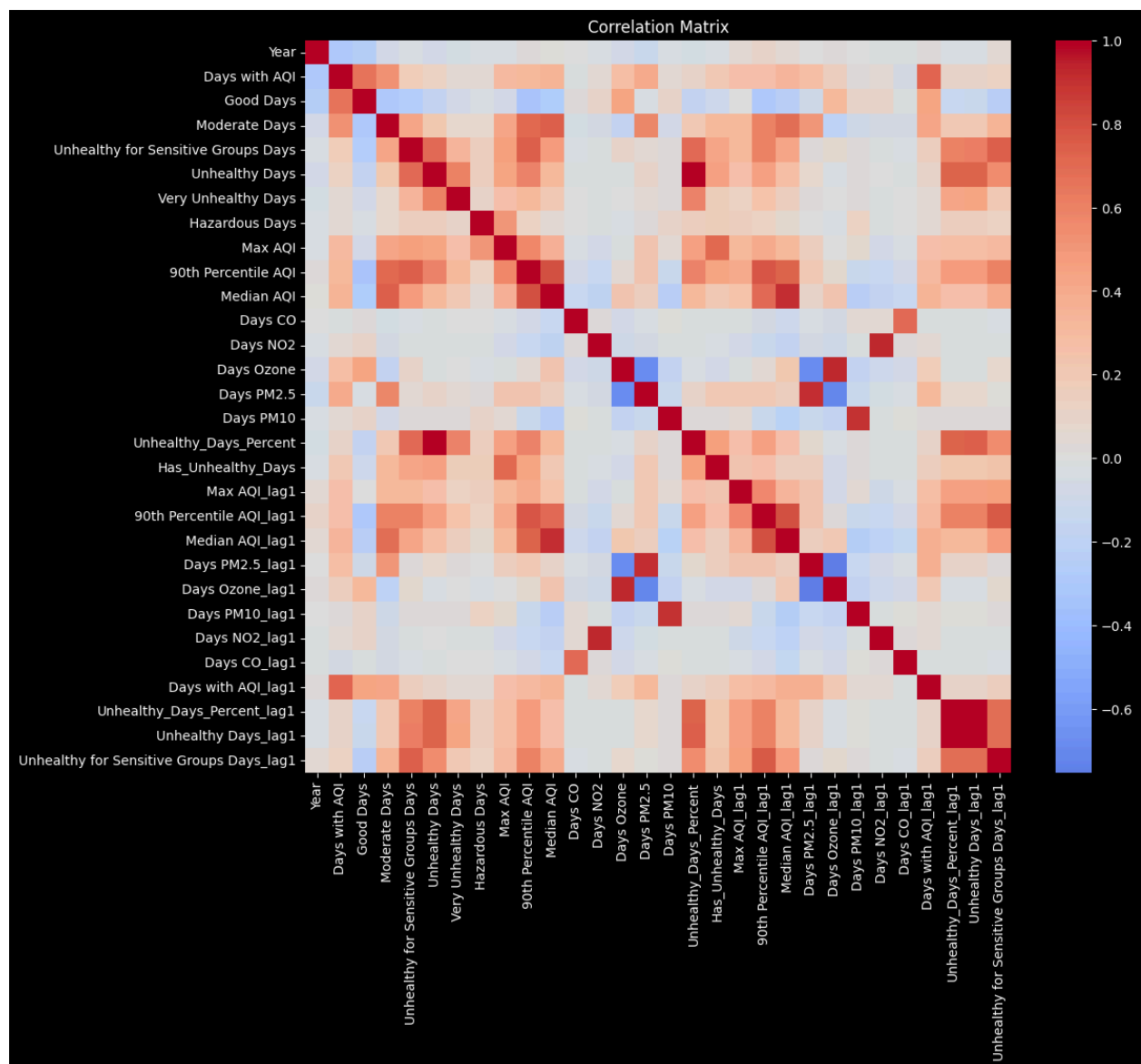
Our Approach

We built two machine learning models to predict the number of unhealthy air quality days for counties in the next year, starting with correlation analysis to identify the best features.

Step 1: Correlation Analysis - Finding What Matters

Process:

- Calculated correlation matrix for all 15 numeric features
- Identified features most correlated with "Unhealthy Days"
- Checked for multicollinearity (features too similar to each other)



Findings:

| Feature | Correlation | Notes |
|-------------------------------------|-------------|--|
| Unhealthy for Sensitive Groups Days | +0.699 | Strong predictor, however, excluded to prevent data leakage risk |
| 90th Percentile AQI | +0.593 | Strong predictor but presented multicollinearity challenges when used with Median AQI |
| Very Unhealthy Days | +0.591 | Strong predictor |
| Max AQI | +0.442 | Moderate |
| Median AQI | +0.326 | Strong predictor but presented multicollinearity challenges when used with 90th Percentile AQI |
| Good Days | -0.176 | Weak predictor |

Step 2: Data Preparation

Data Set:

- Excluded outliers, duplicates & N/A values
- 5,902 county-year observations (2020-2025)
- Target Variable: Number of unhealthy air quality days (continuous: 0-69)

Train/Test split:

- 70% Training Data: 4,131 observations
- 30% Test Data: 1,771 observations

Step 3: Model 1 - Linear Regression

Why Start with Linear Regression?

- Simple, interpretable baseline
- Correlation suggests linear relationships exist
- Fast to train and evaluate

Process:

- Standardized features (correlation works better with scaled data)
- Trained linear model using OLS (Ordinary Least Squares)
- Predicted on test set

Results:

- R^2 Score: 0.55 (55% of variance explained)
- MAE: 1.05 days (average error ~1 day)
- Model captures moderate patterns from correlation
- Limited by linearity - can't capture complex interactions
- Predictions off by about 1 day on average

Step 4: Model 2 - Random Forest Regression

Why Random Forest?

Correlation analysis revealed:

- Interactions between features (PM2.5 \times Ozone matters)
- Need for complex relationships beyond simple correlation
- Non-linear patterns

Process:

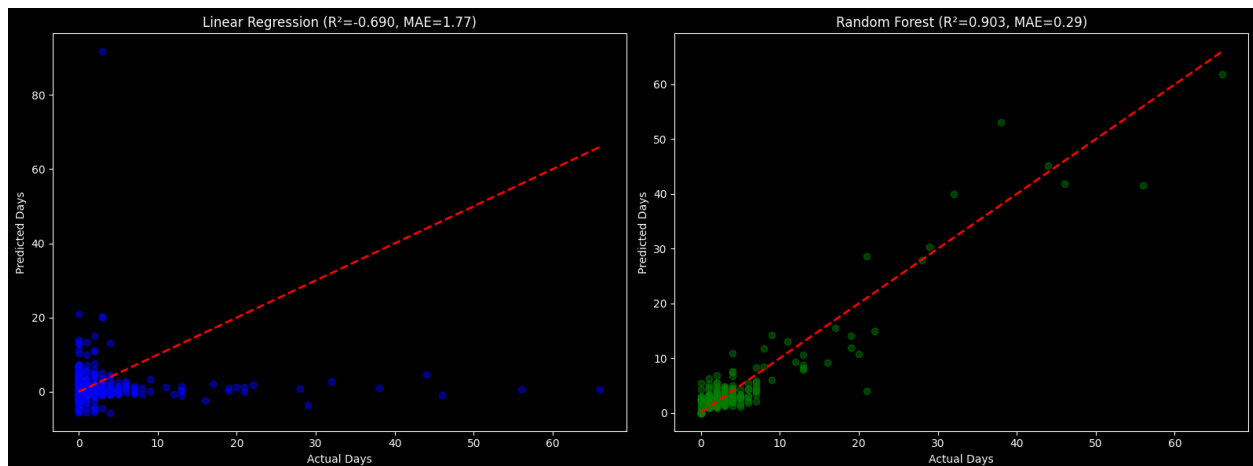
- Trained ensemble of 300 decision trees
- Each tree learns from random feature subsets
- Final prediction = average of all trees

The Value of Our Predictive Model

Our Random Forest model, achieving 90% accuracy in predicting unhealthy air quality days, represents significant value:

- Proactive Alerts: Counties predicted to have >10 unhealthy days can issue advance health warnings to vulnerable populations (children, elderly, asthma patients)
- Healthcare Resource Allocation: Hospitals can stock inhalers, prepare respiratory units, and schedule additional staff based on predicted unhealthy days
- Resource Distribution: Limited public health budgets can be allocated to counties with highest predicted risk, not just highest population

Conclusion:



From Linear Regression:

- Linear relationships explain 55% of variance

- Simple model achieves 1.05 day MAE
- Room for improvement with non-linear methods

From Random Forest:

- Non-linear approach captures 90% of variance
- AQI Volatility emerges as 2nd most important
- 0.29 day MAE - production-ready accuracy

Result: Data-driven predictions accurate within ± 0.3 days, enabling proactive public health responses.

The correlation analysis was essential - it guided feature selection, revealed non-linearity, and validated our Random Forest's feature importance rankings.