# Predicting U.S. Annual Average Air Pollution Concentration

with an average error of

**1.72**

**1.85**

**1.92**

The U.S. is raising awareness about health risks posed by air pollution. To monitor air quality, 876 gravimetric monitors using filtration systems to capture fine particulate matter were deployed across 49 states (including D.C., excluding Alaska and Hawaii).

To predict annual average air pollution concentrations in the U.S., we built and trained a machine learning model. The model identifies important features that explain variations in air pollution levels and predicts outcomes for new observations where data is unavailable. The model is optimized through hyperparameter tuning to improve its accuracy on unseen data. Performance is evaluated using metrics RMSE and $R^2$.

**RMSE:** Root Mean Square Error, the average error between predicted and actual values

**ADJUSTED $R^2$:** how much the independent variable explains the variation of dependent variable

## OVERALL features

**1.72** with adjusted $R^2$ of **0.594**

Utilizing all our variables as predictors, 59.4% of our outcome can be explained by the predictors, which gave an error (RMSE) of 1.72.

## TOP 10 best features

**1.85** with adjusted $R^2$ of **0.510**

Utilizing our top ten features as predictors, 51.0% of our outcome can be explained by the predictors, which gave an error (RMSE) of 1.85.

## TOP 5 best features

**1.92** with adjusted $R^2$ of **0.479**

Utilizing our top five features as predictors, 47.9% of our outcome can be explained by the predictors, which gave an error (RMSE) of 1.92.

Although the overall model is the most accurate, our top ten model displays practicality with less predictors to measure while slightly decreasing the accuracy. The top five model offers even a simpler approach with only a slight drop in accuracy compared to top 10, making it viable in resource-limited scenarios.