Project Motivation
Air quality is a very heated topic for any time. The development of technology is a double-edged sword to the environment. In this project, I try to explore the possibility of predicting carbon monoxide (CO) levels using data from low-cost chemical sensors and environmental readings, offering a scalable solution for real-time, dense air quality estimation.

Data Sources
Data Link: https://archive.ics.uci.edu/dataset/360/air+quality

Variable Name	Role	Type	Description	Units
Date	Feature	Date		
Time	Feature	Categorical		
CO(GT)	Feature	Integer	True hourly averaged concentration CO in mg/m^3 (reference analyzer)	mg/m^3
PT08.S1(CO)	Feature	Categorical	hourly averaged sensor response (nominally CO targeted)	
NMHC(GT)	Feature	Integer	True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)	microg/m^3
C6H6(GT)	Feature	Continuous	True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)	microg/m^3
PT08.S2(NMHC)	Feature	Categorical	hourly averaged sensor response (nominally NMHC targeted)	

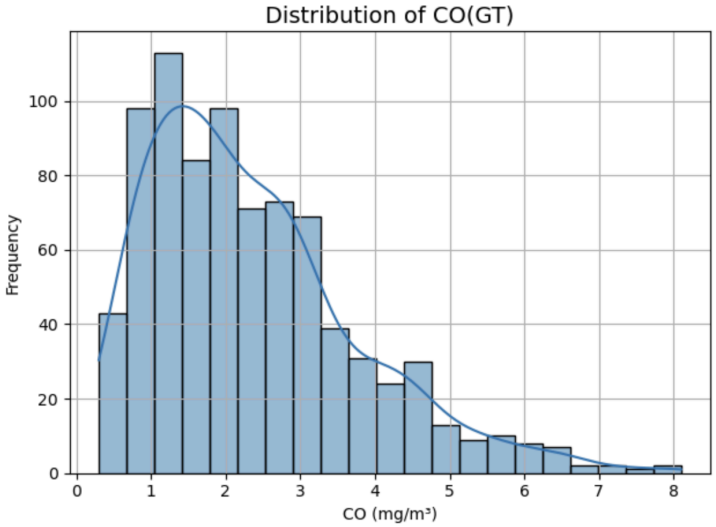
NOx(GT)	Feature	Integer	True hourly averaged NOx concentration in ppb (reference analyzer)	ppb
PT08.S3(NOx)	Feature	Categorical	hourly averaged sensor response (nominally NOx targeted)	
NO2(GT)	Feature	Integer	True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)	microg/m^3
PT08.S4(NO2)	Feature	Categorical	hourly averaged sensor response (nominally NO2 targeted)	
PT08.S5(O3)	Feature	Categorical	hourly averaged sensor response (nominally O3 targeted)	
T	Feature	Continuous	Temperature	°C
RH	Feature	Continuous	Relative Humidity	%
AH	Feature	Continuous	Absolute Humidity	

Methodologies:
Data Cleaning:

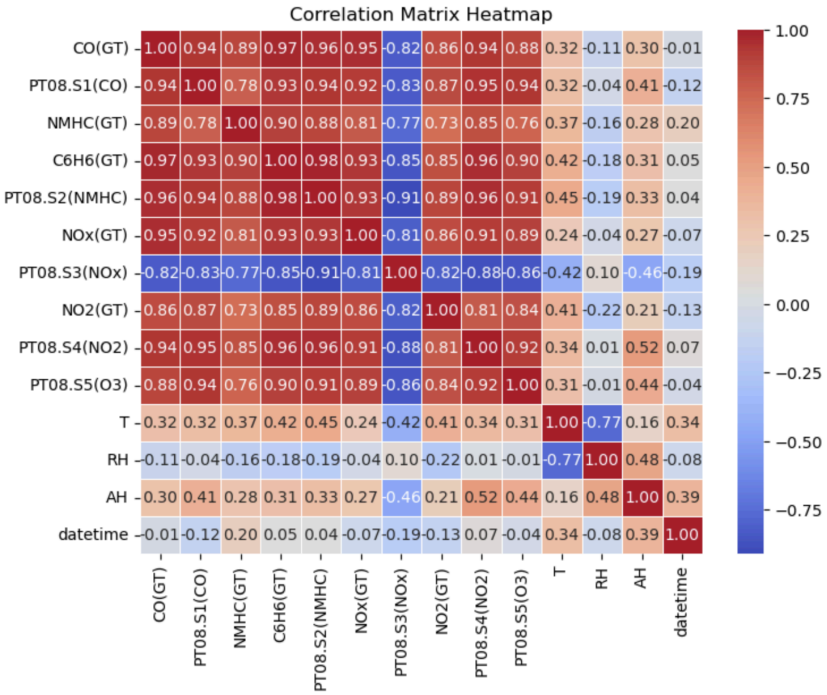
- Convert all numeric data into float data type(originally in object)
- Remove all null values(errors)
- Remove all values with -200 (error value)

Exploratory Data:

- Right-Skewed distribution of CO(GT)



- Correlation heatmap: Strong correlation between CO(GT) and PT08.S1(CO), C6H6(GT), PT08.S2(NMHC).



Modeling

I used Linear Regression, Random Forest, and Random Forest models to get RMSE and R-squared values.

Results

Model name	RMSE	R-Squared value
Linear Regression	0.2876	0.9603
Random Forest	0.2729	0.9642
Random Forest(tuned)	0.3098	0.9539

Insights:

- From feature importance visualization, the visualization shows that PT08.S2(NMHC) is the most important feature in predicting CO levels. It has second highest correlation coefficients with CO(GT)
- Random Forest without tuned model outperformed linear regression.
- T, RH, AH, these 3 environmental features have relatively low correlation coefficients with CO(GT), but have great impact on CO(GT) based on coefficient value based on linear regression.