Air Quality Prediction

Predicting AQI based on pollutant levels (CO, O3, NO2, SO2, PM10, PM2.5)

Arman Ghaziaskari Naeini

December 2024

What is Air Quality?

Air quality is crucial for public health
Due to its direct impact on respiratory and cardiovascular conditions.

AQI (Air Quality Index):
A common measure used to indicate air pollution levels.

Key Pollutants:
The primary pollutants affecting air quality include

CO (Carbon Monoxide)
O_3 (Ozone)
NO_2 (Nitrogen Dioxide)
SO_2 (Sulfur Dioxide)
PM10 (Particulate Matter 10)
PM2.5 (Particulate Matter 2.5)

Understanding air quality is essential for developing effective pollution management strategies.

Key Features Affecting Air Quality

SO2 (Sulfur Dioxide):

Volcanic gas or sulfufric burning in fossil fuel

PM10 (Particulate Matter 10):

Particles can penetrate the lungs. produced by Dust, wildfires and brush/waste burning, industrial sources.

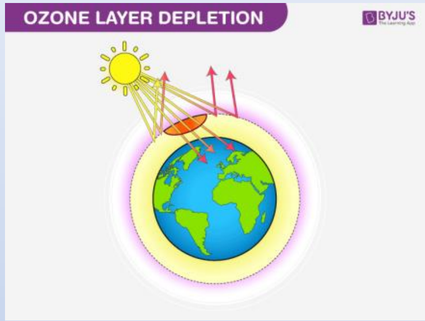
PM2.5 (Particulate Matter 2.5):

Industry and fuel oil, construction site, transboundary, industry, open burning or Traffic

Key Features Affecting Air Quality

Nitrogen Dioxide (NO2)

formed during the process of nitride combustion. Nitride is the product of nitrogen oxide photochemical reaction.



Ozone (O3)

Secondary pollutants from nitrogen oxide, reactive hydrocarbons, or photochemical reaction.

Carbon Monoxide (CO)

Forest fire, methane nitridation, biological activity or incomplete combustion of fuel.

Data Source

- Dataset from Iranian Environmental Organizaion (2017)
- 2123 Air polution stations
- Metrics included:

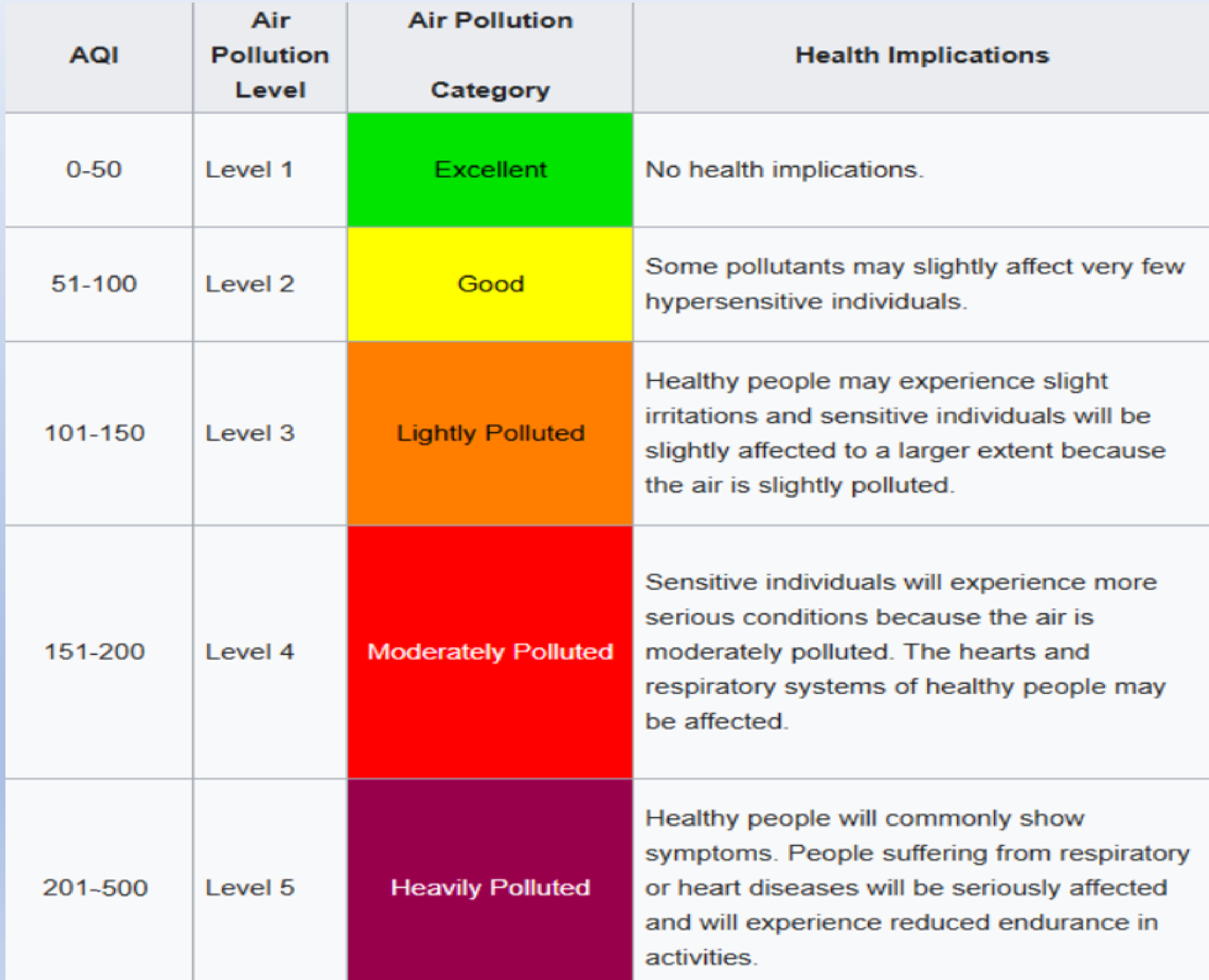
  CO (Carbon Monoxide)
  O3 (Ozone)
  NO2 (Nitrogen Dioxide)
  SO2 (Sulfur Dioxide)
  PM10 (Particulate Matter 10)
  PM2.5 (Particulate Matter 2.5)

Target:

AQI (Air Quality Index)

Creating 2 targets :

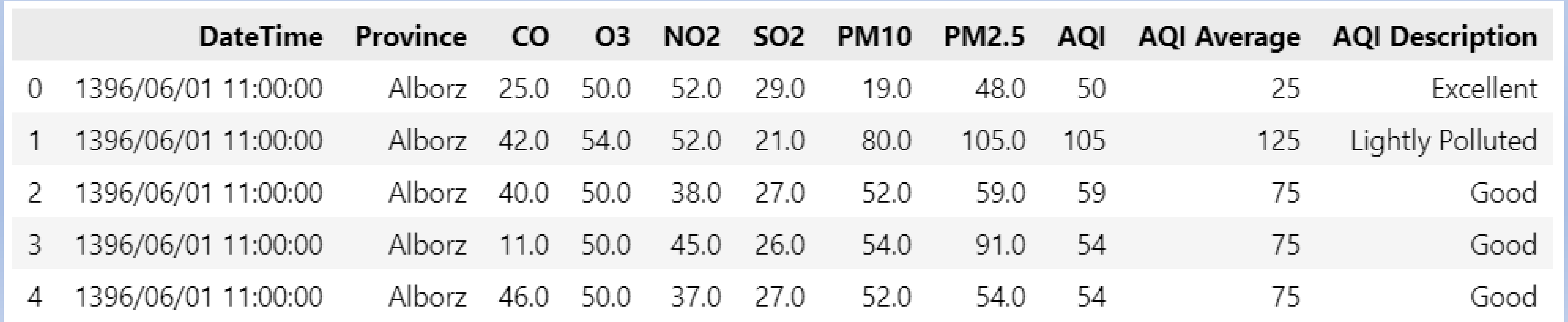
- AQI Description
- AQI Average



AQI	Air Pollution Level	Air Pollution Category	Health Implications
0-50	Level 1	Excellent	No health implications.
51-100	Level 2	Good	Some pollutants may slightly affect very few hypersensitive individuals.
101-150	Level 3	Lightly Polluted	Healthy people may experience slight irritations and sensitive individuals will be slightly affected to a larger extent because the air is slightly polluted.
151-200	Level 4	Moderately Polluted	Sensitive individuals will experience more serious conditions because the air is moderately polluted. The hearts and respiratory systems of healthy people may be affected.
201-500	Level 5	Heavily Polluted	Healthy people will commonly show symptoms. People suffering from respiratory or heart diseases will be seriously affected and will experience reduced endurance in activities.

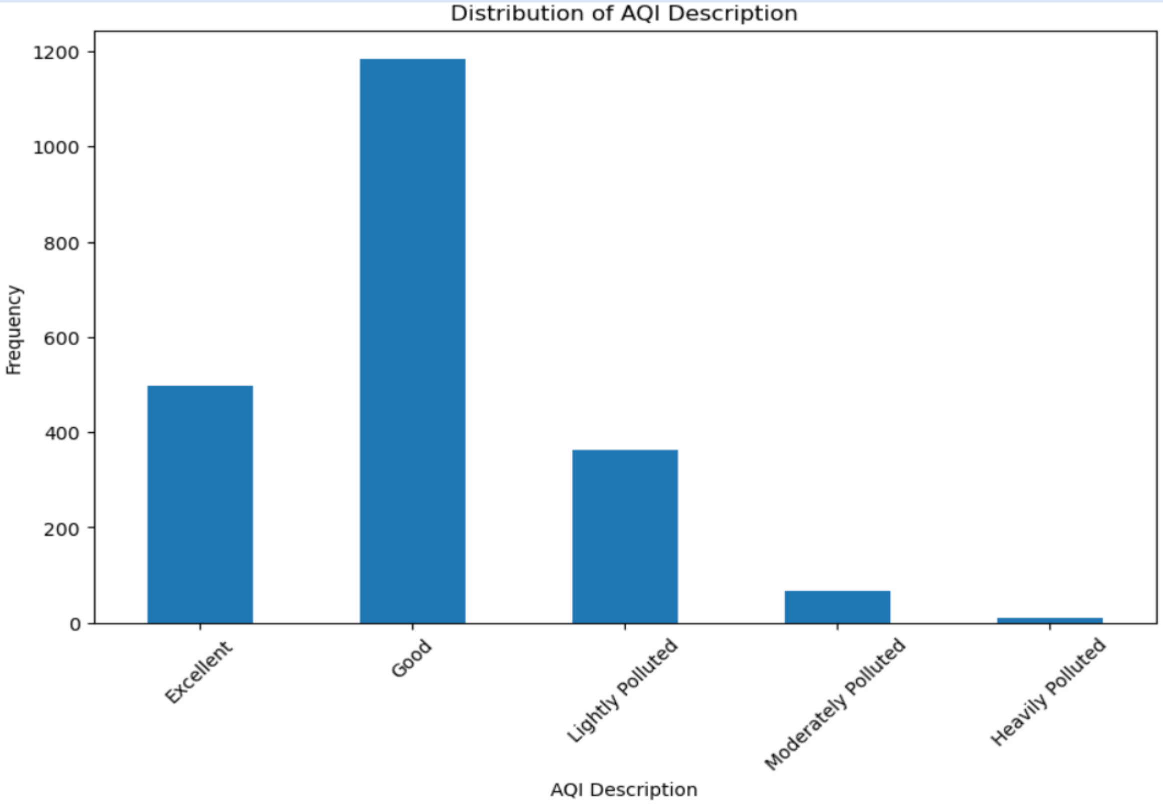
Problems in the data

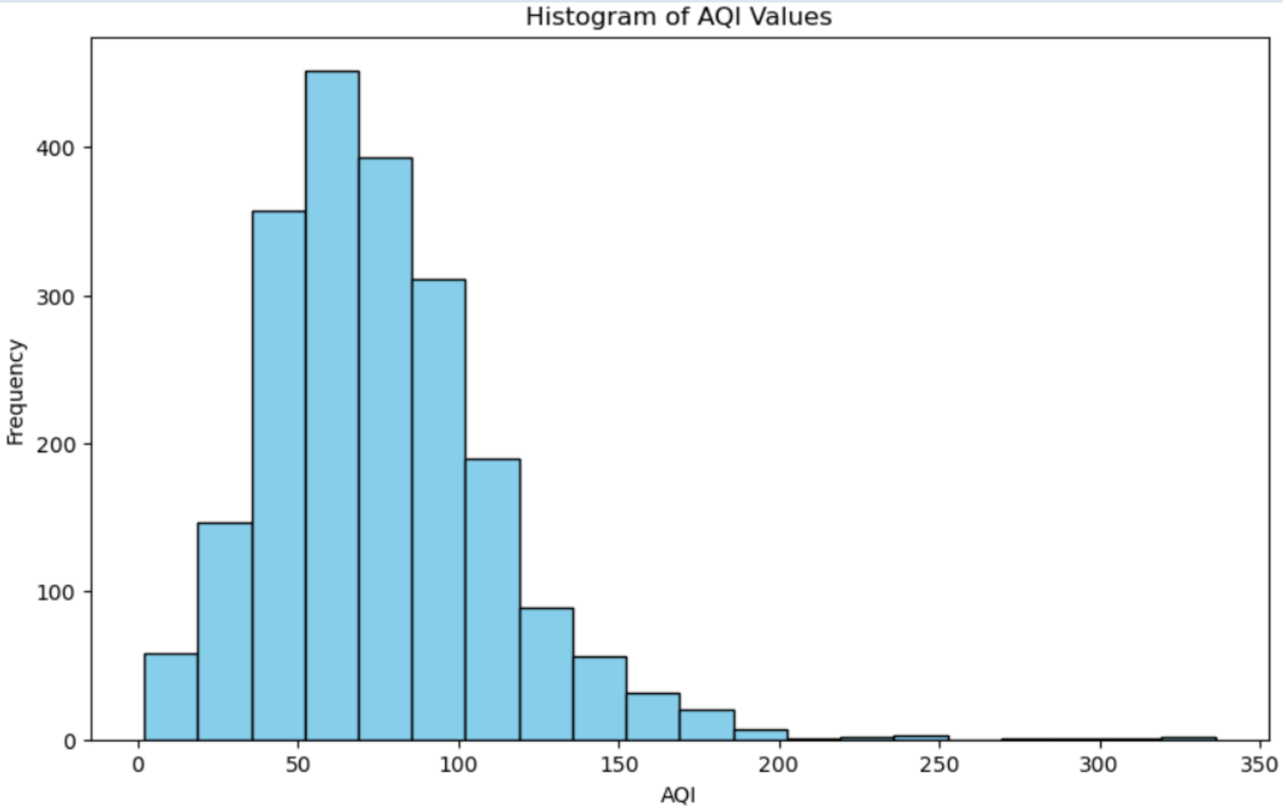
- Missing Values: Resolved using the median.
- Duplicated Rows: Removed.
- Column Names and Contents: Originally written in Persian; converted to English.



	DateTime	Province	CO	O3	NO2	SO2	PM10	PM2.5	AQI	AQI Average	AQI Description
0	1396/06/01 11:00:00	Alborz	25.0	50.0	52.0	29.0	19.0	48.0	50	25	Excellent
1	1396/06/01 11:00:00	Alborz	42.0	54.0	52.0	21.0	80.0	105.0	105	125	Lightly Polluted
2	1396/06/01 11:00:00	Alborz	40.0	50.0	38.0	27.0	52.0	59.0	59	75	Good
3	1396/06/01 11:00:00	Alborz	11.0	50.0	45.0	26.0	54.0	91.0	54	75	Good
4	1396/06/01 11:00:00	Alborz	46.0	50.0	37.0	27.0	52.0	54.0	54	75	Good

AQI

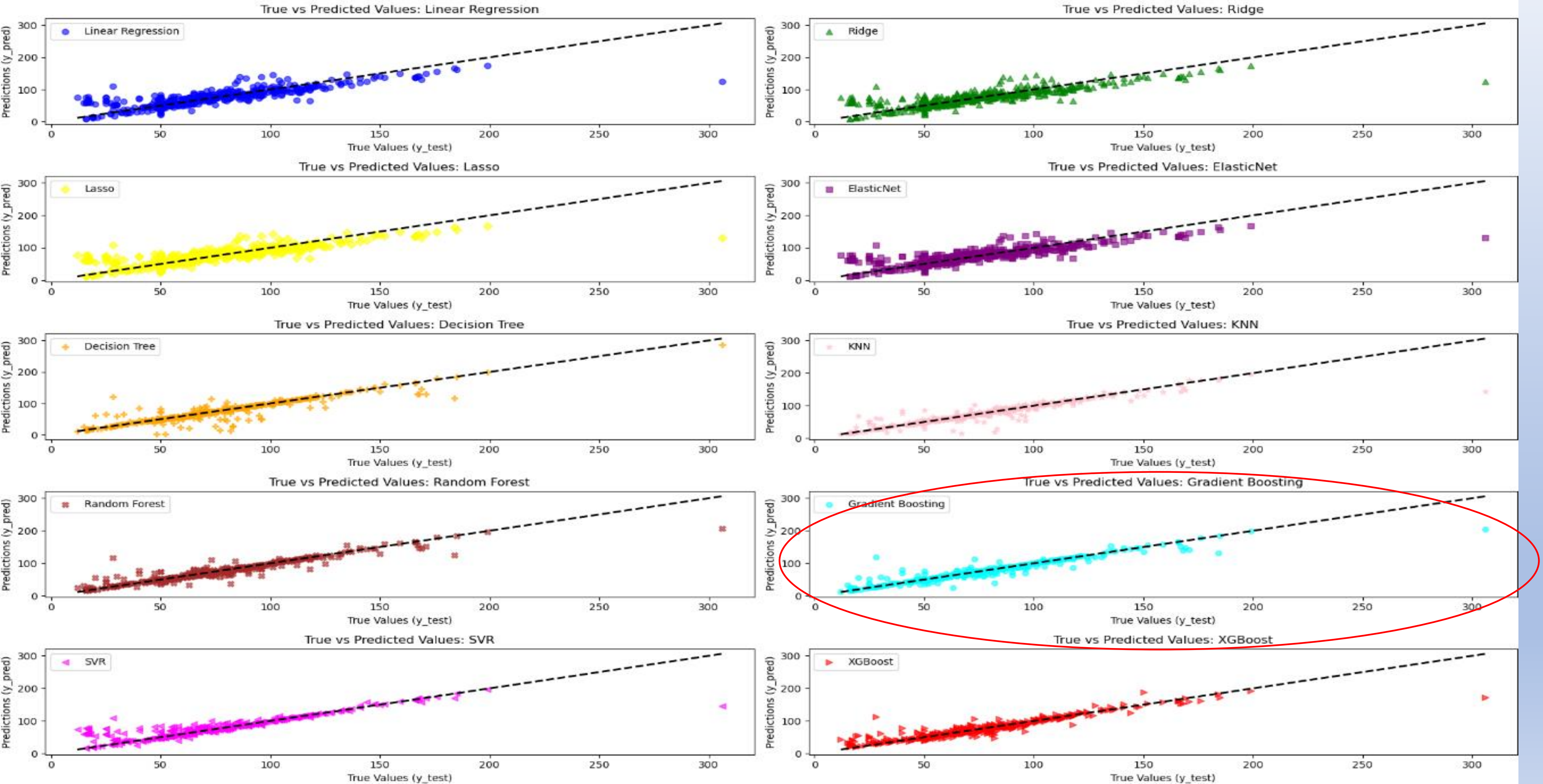

Distribution of AQI Description
Frequency
0
200
400
600
800
1000
1200
Excellent
Good
Lightly Polluted
Moderately Polluted
Heavily Polluted
AQI Description


Histogram of AQI Values
Frequency
0
100
200
300
400
0
50
100
150
200
250
300
350
AQI

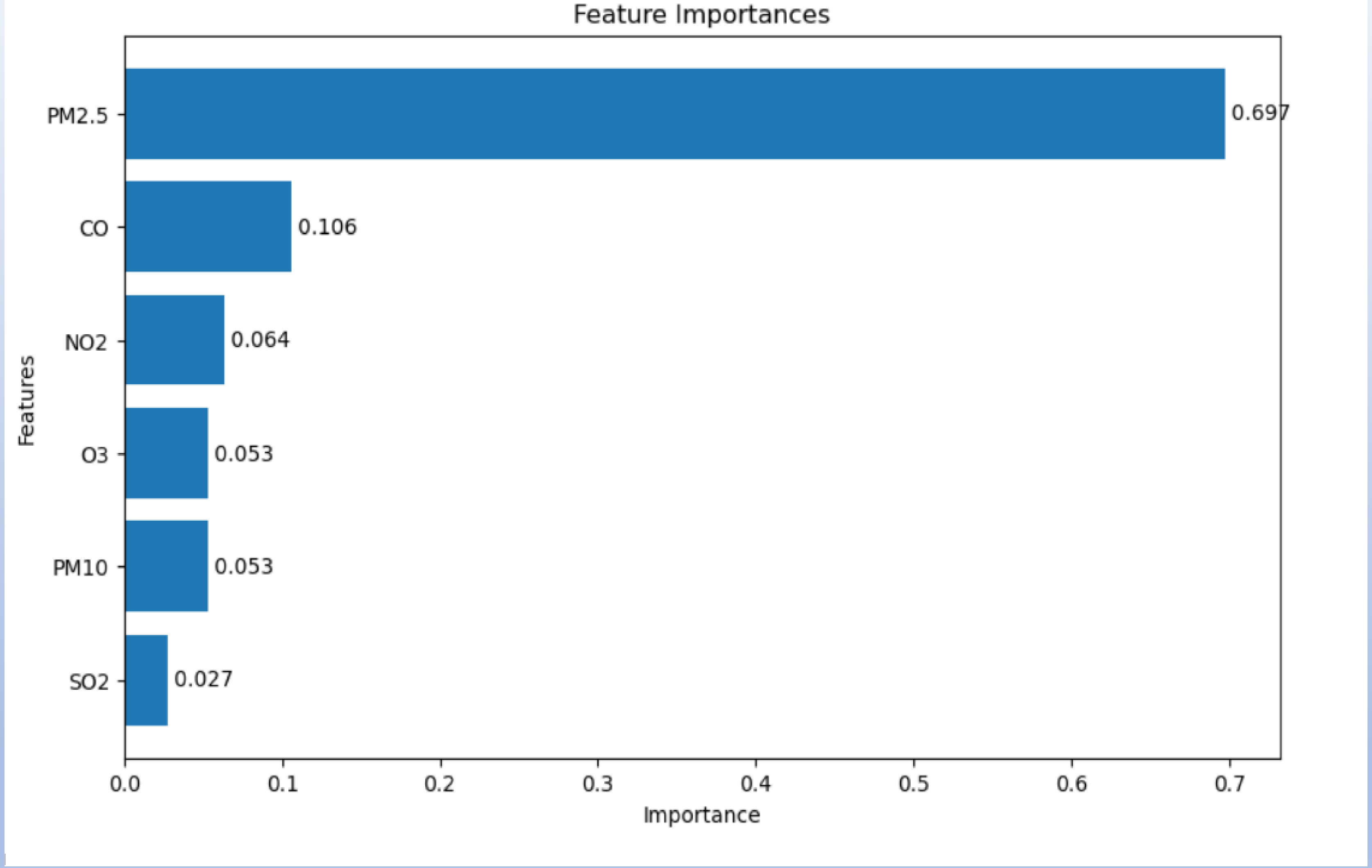
Regression

Rank	Model	MAE	R^2
1	Gradient Boosting	4.576557	0.909186
2	Random Forest	4.947887	0.905484
3	XGBoost	4.962615	0.905852
4	KNN	5.127256	0.873260
5	Decision Tree	5.332536	0.857144
6	SVR	6.478604	0.813697
7	ElasticNet	13.003160	0.693486
8	Lasso	13.003974	0.693382
9	Linear Regression	13.018671	0.691836
10	Ridge	13.018671	0.691836

Regression


True vs Predicted Values: Linear Regression
Linear Regression
True vs Predicted Values: Ridge
Ridge
True vs Predicted Values: Lasso
Lasso
True vs Predicted Values: ElasticNet
ElasticNet
True vs Predicted Values: Decision Tree
Decision Tree
True vs Predicted Values: KNN
KNN
True vs Predicted Values: Random Forest
Random Forest
True vs Predicted Values: Gradient Boosting
Gradient Boosting
True vs Predicted Values: SVR
SVR
True vs Predicted Values: XGBoost
XGBoost
Predictions (y_pred)
True Values (y_test)
0
100
200
300
0
50
100
150
200
250
300

Regression



Gradient Boosting

Controlling PM2.5 emissions should be the primary focus for improving air quality.

CO's contribution highlights the need to monitor traffic emissions and combustion activities more closely.

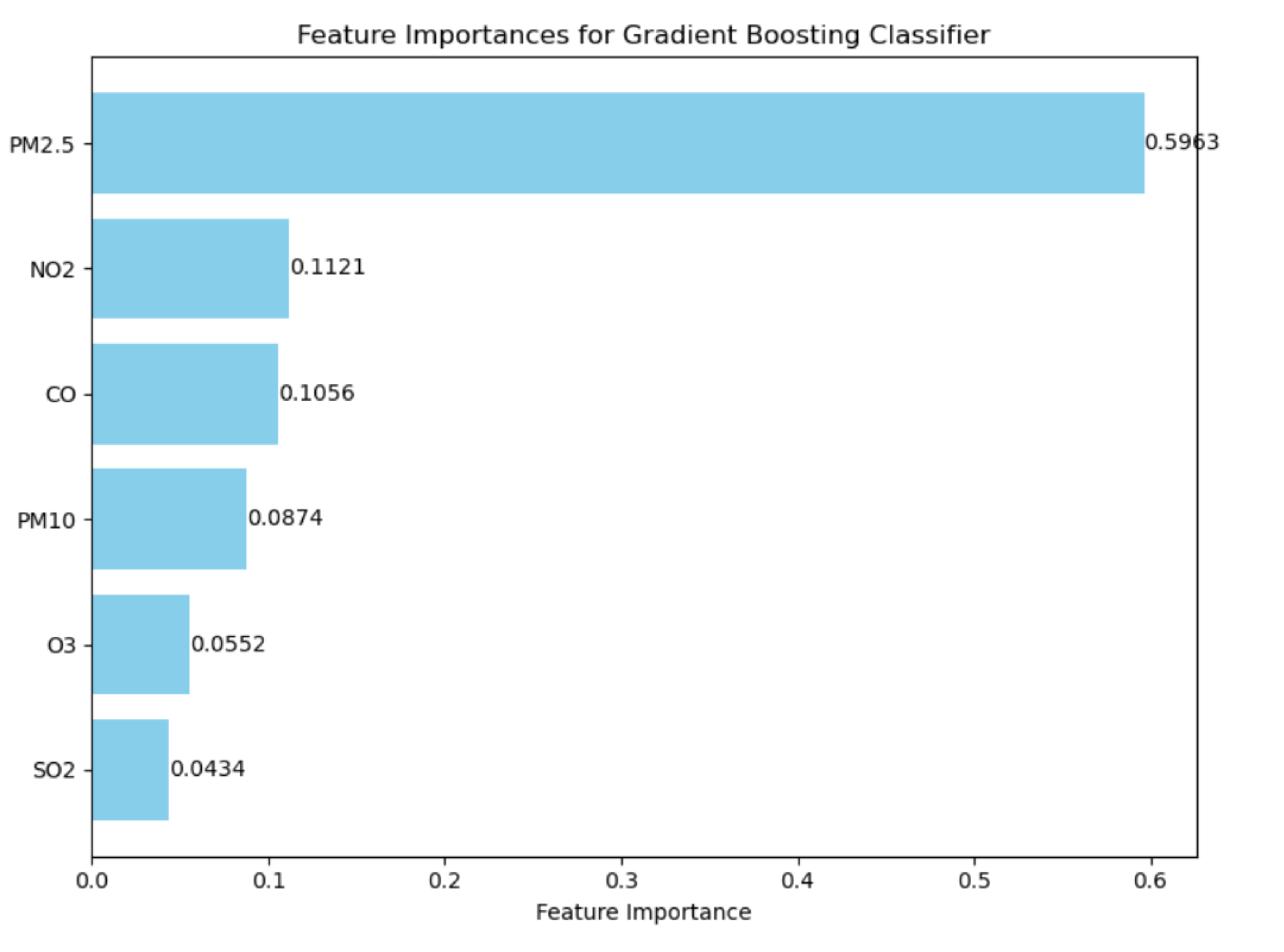
NO2 (Nitrogen Dioxide), O3 (Ozone), and PM10 have relatively low importances. These pollutants still influence air quality but are less significant compared to PM2.5.

Feature	Importance
PM2.5	0.697276
CO	0.105519
NO2	0.063716
O3	0.053219
PM10	0.053100
SO2	0.027170

Classification

Model	Accuracy	F1-score	ROC AUC	Recall
Gradient Boosting	0.9426	0.9840	0.7251	0.7222
XGBoost	0.9330	0.9715	0.7285	0.7266
Random Forest	0.9282	0.9897	0.7219	0.7103
KNN	0.9258	0.9444	0.9268	0.9131
Decision Tree	0.9067	0.8367	0.7081	0.7029
SVC	0.8612	0.9707	0.6777	0.6713
Logistic Regression	0.7727	0.8662	0.6108	0.5854

Classification



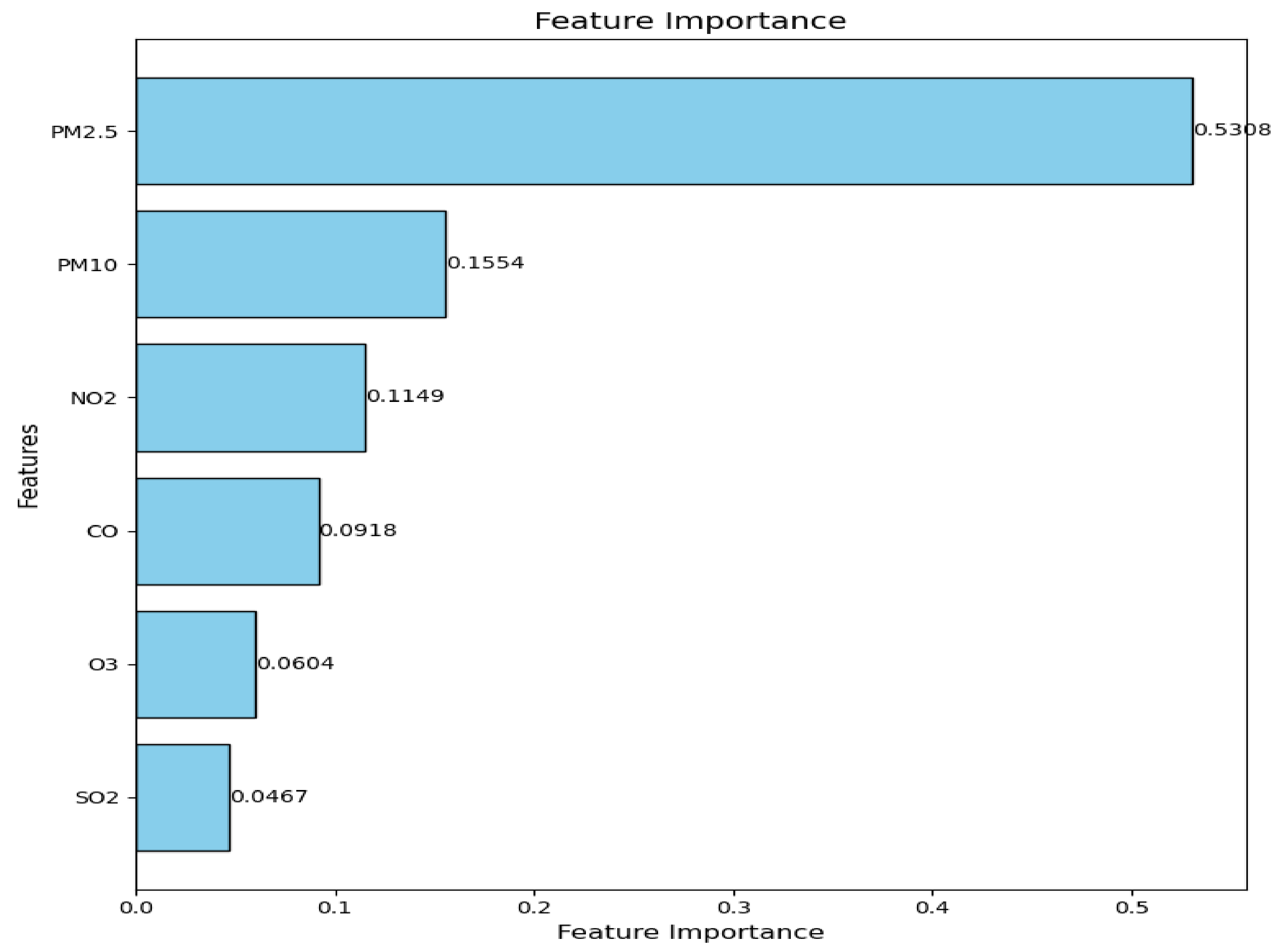
Controlling PM2.5 emissions should be the primary focus for improving air quality.

NO2, CO's contribution highlights the need to monitor traffic emissions and combustion activities more closely.

Feature	Importance
PM2.5	0.596267
NO2	0.112122
CO	0.105613
PM10	0.087409
O3	0.055173
SO2	0.043416

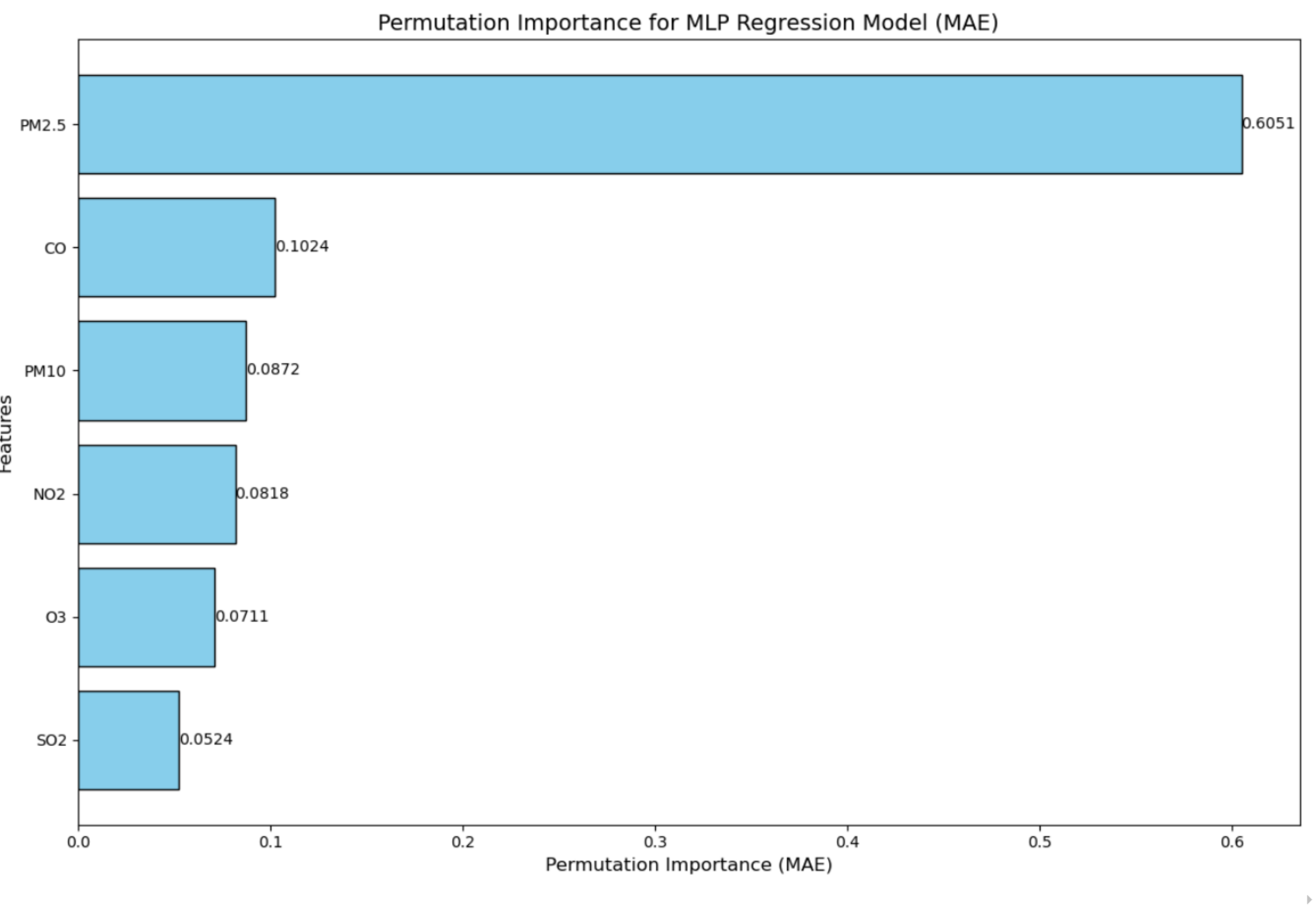
Gradient Boosting

Deep Learning-MPLs-Classification



PM2.5	0.5308
PM10	0.1554
NO2	0.1149
CO	0.0918
O3	0.0604
SO2	0.0467

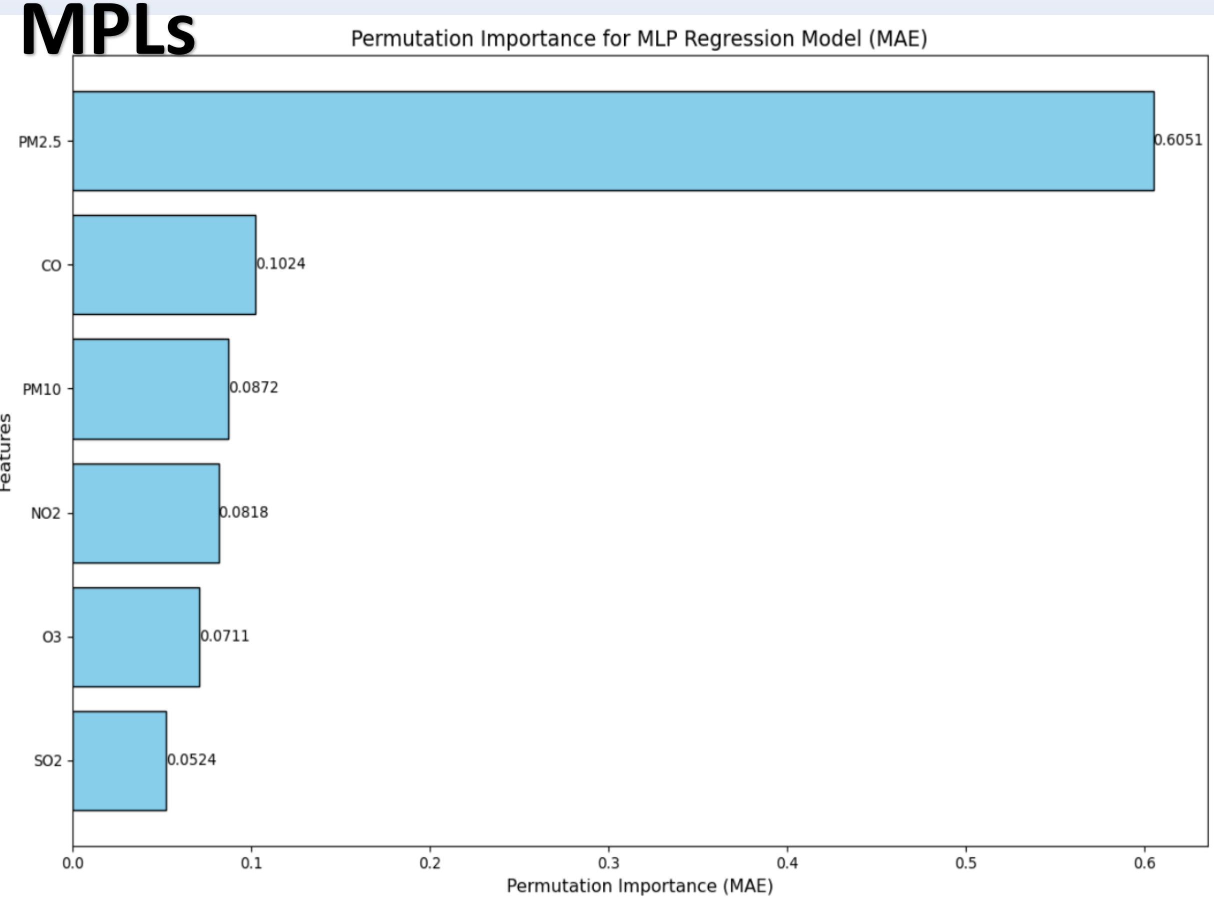
Deep Learning-MPLs-Regressiom



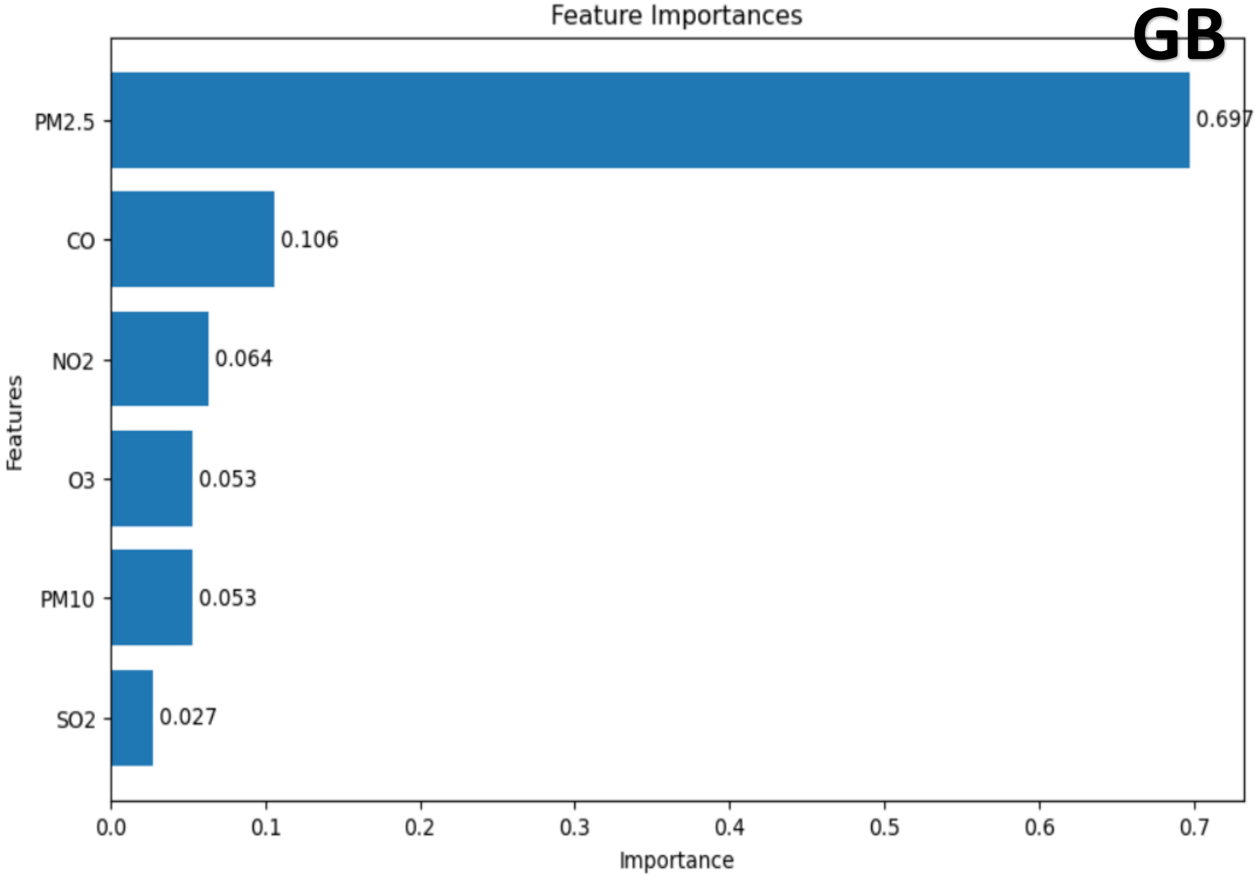
PM2.5	0.6051
CO	0.1024
PM10	0.0872
NO2	0.0818
O3	0.0711
SO2	0.0524

Regression

MPLs



GB

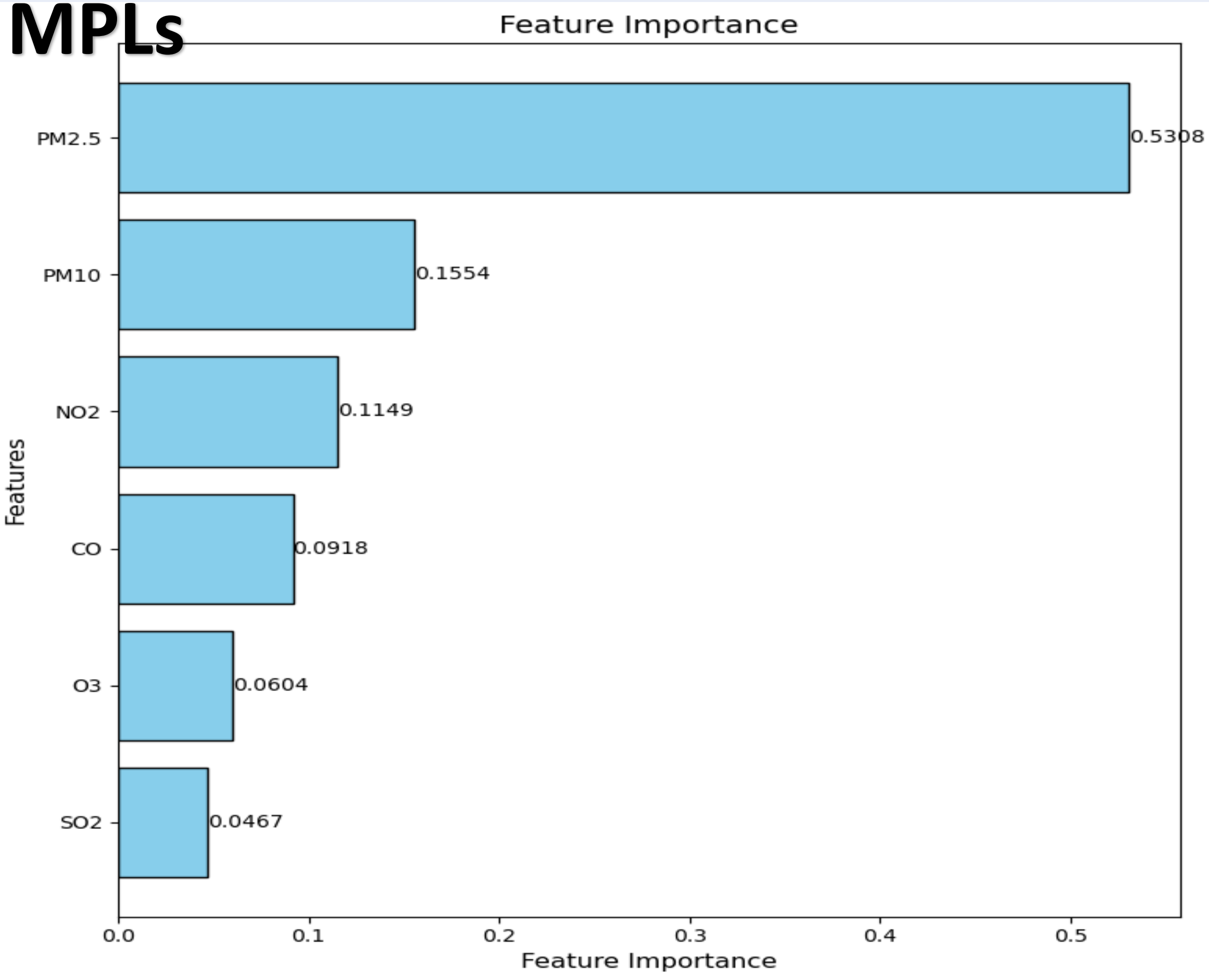


Both models agree on the primary role of PM2.5 in AQI prediction, underscoring its generalizability across methods.

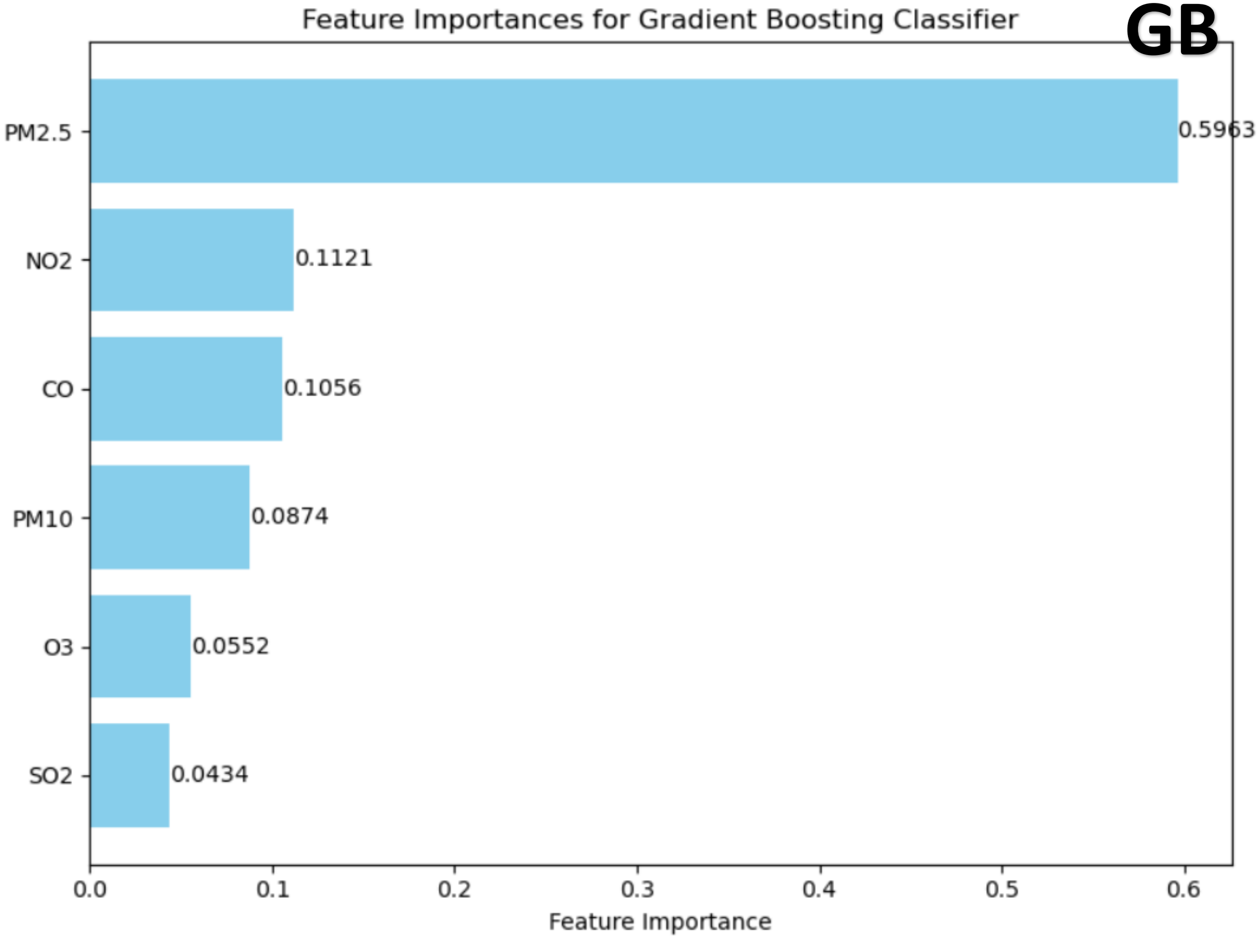
Health and Policy Implications: The prominence of PM2.5, CO, and PM10 aligns with their known adverse health effects, underscoring their importance in air quality management policies.

Classification

MPLs



GB



1. **PM2.5 consistently leads in importance, emphasizing its health and environmental impact.**
2. **Gradient Boosting models tend to assign higher importance to CO and NO2 compared to MLP models.**
3. **MLP models capture a slightly higher role of PM10.**

Conclusion

- **PM2.5 is the Most Important Feature**
- PM2.5 is consistently the top feature in all models, showing its strong impact on air quality and human health**.**

- **Model Performance Highlights**
- Gradient Boosting: Best performance with the lowest MAE (4.576557) in regression and the highest accuracy (94.26%) in classification.
- MLP Models: Showed good results and flexibility in analyzing feature importance.

Future Directions

1. Focus on an Area and Extend the Dataset

   By focusing on a specific area, we can identify localized patterns, such as seasonal trends, traffic density, or industrial pollution.

2. Extend the Dataset to Include Meteorological Factors

   e.g., temperature, humidity, wind speed.

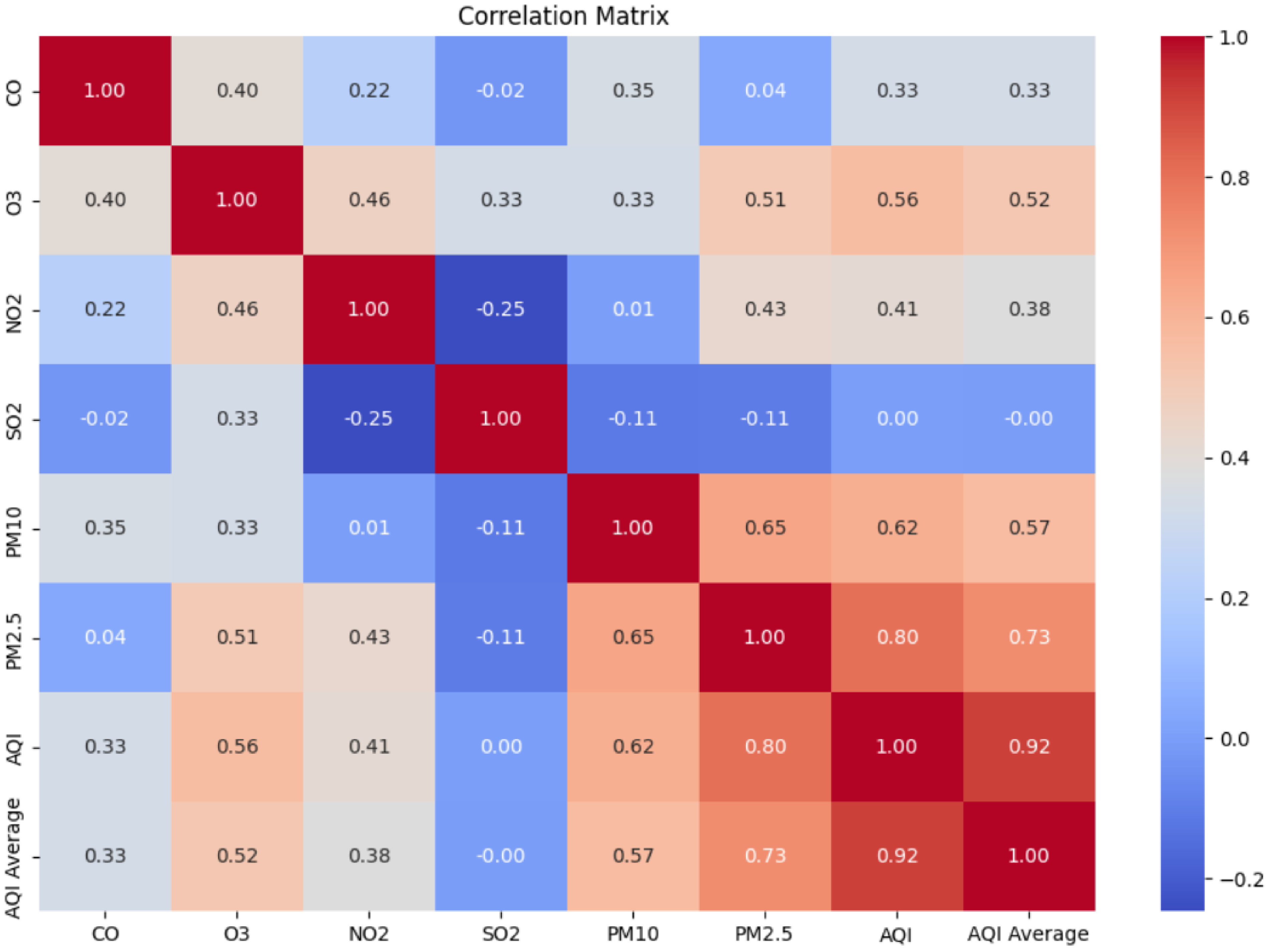
3. Using Time Series Data to Predict AQI Behavior

   To predict AQI both over time and space.

Thanks for your attention

Eskerrik asko zure arretagatik

Heatmap



	CO	O3	NO2	SO2	PM10	PM2.5	AQI	AQI Average
CO	1.00	0.40	0.22	-0.02	0.35	0.04	0.33	0.33
O3	0.40	1.00	0.46	0.33	0.33	0.51	0.56	0.52
NO2	0.22	0.46	1.00	-0.25	0.01	0.43	0.41	0.38
SO2	-0.02	0.33	-0.25	1.00	-0.11	-0.11	0.00	-0.00
PM10	0.35	0.33	0.01	-0.11	1.00	0.65	0.62	0.57
PM2.5	0.04	0.51	0.43	-0.11	0.65	1.00	0.80	0.73
AQI	0.33	0.56	0.41	0.00	0.62	0.80	1.00	0.92
AQI Average	0.33	0.52	0.38	-0.00	0.57	0.73	0.92	1.00

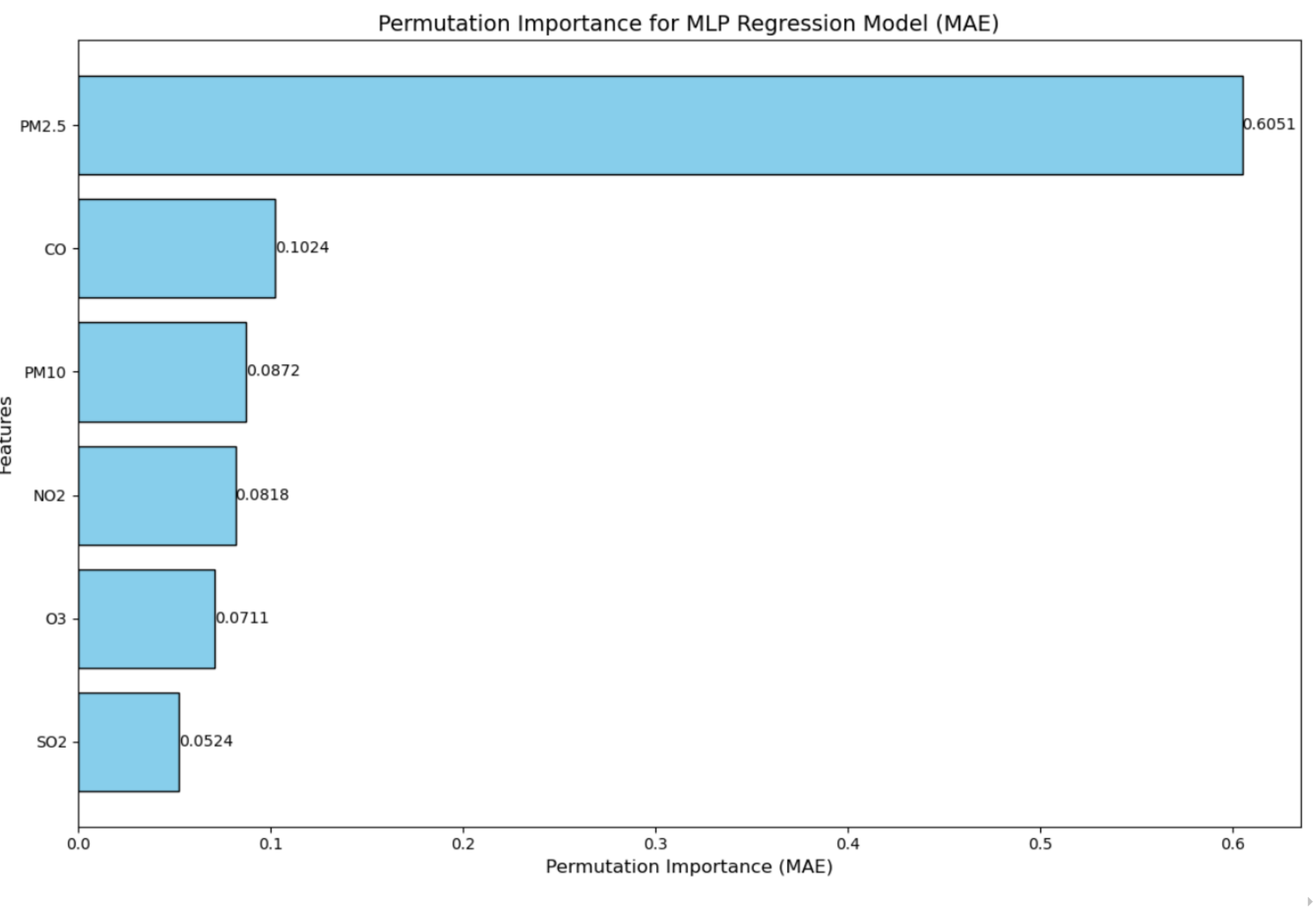
Regression

Rank	Model	MAE	R^2	Hyperparametrs
1	Gradient Boosting	4.576557	0.909186	learning_rate : 0.05, max_depth : 10, min_samples_split : 10, n_estimators : 300, subsample : 0.8
2	Random Forest	4.947887	0.905484	max_depth : None, max_features : None, min_samples_leaf : 1, min_samples_split : 2, n_estimators : 300
3	XGBoost	4.962615	0.905852	learning_rate : 0.1, max_depth : 7, n_estimators : 150, subsample : 0.8
4	KNN	5.127256	0.873260	metric : manhattan , n_neighbors : 2, weights : distance
5	Decision Tree	5.332536	0.857144	criterion : squared_error , max_depth : None, max_features : None, min_samples_leaf : 1, min_samples_split : 2
6	SVR	6.478604	0.813697	C : 100, degree : 1, gamma : auto , kernel : rbf
7	ElasticNet	13.003160	0.693486	alpha : 0.7543120063354622, l1_ratio : 1.
8	Lasso	13.003974	0.693382	alpha : 0.774263682681127
9	Linear Regression	13.018671	0.691836	-

Classification

Model	Accuracy	F1-score	ROC AUC	Recall	Best Params
Gradient Boosting	0.9426	0.9840	0.7251	0.7222	learning_rate : 0.1, max_depth : 7, n_estimators : 100, subsample : 0.8}
XGBoost	0.9330	0.9715	0.7285	0.7266	colsample_bytree : 0.8, learning_rate : 0.1, max_depth : 7, n_estimators : 50, subsample : 0.8
Random Forest	0.9282	0.9897	0.7219	0.7103	max_depth : None, max_features : sqrt , min_samples_leaf : 1, min_samples_split : 2, n_estimators : 200
KNN	0.9258	0.9444	0.9268	0.9131	metric : euclidean , n_neighbors : 1, weights : uniform
Decision Tree	0.9067	0.8367	0.7081	0.7029	criterion : entropy , max_depth : None, min_samples_leaf : 1, min_samples_split : 2}
SVC	0.8612	0.9707	0.6777	0.6713	C : 10, gamma : scale , kernel : rbf
Logistic Regression	0.7727	0.8662	0.6108	0.5854	C : 11.288378916846883, solver : lbfgs

Deep Learning-MPLs-Regressiom

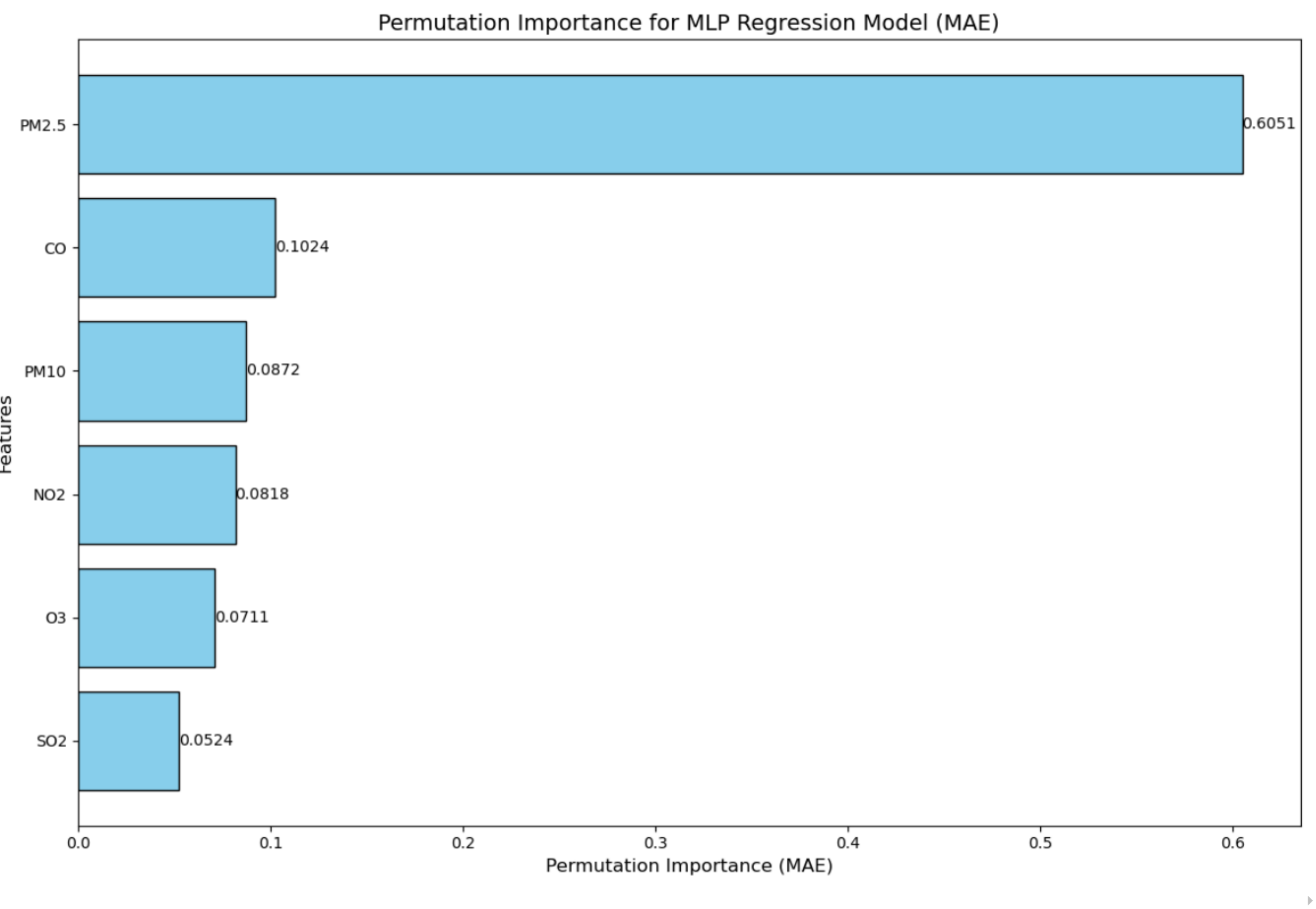


Best Hyperparameters:

- batch_size: 22
- epochs: 381
- dropout_rate: 0.11060969310248417
- learning_rate: 0.0018505523167281056
- neurons: 122

PM2.5	0.6051
CO	0.1024
PM10	0.0872
NO2	0.0818
O3	0.0711
SO2	0.0524

Deep Learning-MPLs-Regressiom



Best Hyperparameters:

- **batch_size: 22**
- **epochs: 381**
- **dropout_rate: 0.11060969310248417**
- **learning_rate: 0.0018505523167281056**
- **neurons: 122**

PM2.5	**0.6051**
CO	**0.1024**
PM10	**0.0872**
NO2	**0.0818**
O3	**0.0711**
SO2	**0.0524**