Model Evaluation: Random Forest Regressor

# Objective

To evaluate the performance of a Random Forest Regressor for forecasting PM2.5 concentration levels, with the aim of improving accuracy compared to the baseline linear regression and XGBoost models.

# Model Configuration

- Model Type: Random Forest Regressor (`sklearn.ensemble.RandomForestRegressor`)

- Hyperparameters:

* • n\_estimators = 100
* • max\_depth = 5
* • random\_state = 42

- Feature Set:

* • Lag features: pm25\_lag\_1, pm25\_lag\_2, pm25\_lag\_3
* • Time-based features: year, month, sin\_month, cos\_month

# Evaluation Metrics

Mean Absolute Error (MAE): 1.0454 µg/m³

Root Mean Squared Error (RMSE): 1.0920 µg/m³

# Test Set Predictions

|  |  |  |
| --- | --- | --- |
| Date | Actual (PM2.5 µg/m³) | Predicted (PM2.5 µg/m³) |
| 2021-12-01 | 7.40 | 8.04 |
| 2022-06-01 | 6.99 | 8.34 |
| 2022-12-01 | 6.92 | 8.28 |
| 2023-06-01 | 9.27 | 8.43 |

# Conclusion

The Random Forest model performed comparably to the baseline linear regression model, with RMSE nearly identical and MAE slightly higher. Although it demonstrated the ability to capture general trends, the model exhibited a tendency to predict toward the mean, resulting in a loss of accuracy during higher or lower PM2.5 periods. This confirms the difficulty of outperforming linear models on small, regular, and clean time series data unless more predictive features or data points are introduced.