*International Journal of Engineering and Science Invention (IJESI)*
*ISSN (Online): 2319-6734, ISSN (Print): 2319-6726*
*www.ijesi.org ||Volume 14 Issue 6 June 2025 || PP 45-48*

# Predicting PM2.5 Levels in Seoul Using Random Forest Regression

Asma Salah Aldeen Mohammed Abdallah
Islam Hamad Eljack Elameen

***Abstract***
*Air pollution is recognized as a major public health risk worldwide, with fine particulate matter (PM2.5) identified as a key contributor to respiratory and cardiovascular diseases (World Health Organization, 2021). Accurate prediction of PM2.5 concentrations is essential for effective public health planning and environmental policymaking. Traditional statistical models have shown limitations in handling complex non-linear interactions between pollutants and meteorological variables. In this context, machine learning (ML) models, particularly ensemble methods like Random Forest, have gained popularity due to their predictive robustness and ability to capture intricate feature interactions (Park and Kim, 2025; Joharestani et al., 2019). This study focuses on predicting PM2.5 concentrations in Seoul using Random Forest Regressor (RFR), utilizing air quality data from the year 2021. Although a longer historical dataset was available, 2021 data was selected for its completeness and consistency across all key pollutant measurements. After data cleaning, feature engineering, and outlier removal, the model was trained and optimized via hyperparameter tuning. The optimized model achieved a Root Mean Squared Error (RMSE) of 4.88 and an R-squared value of 0.72. The results demonstrate the potential of Random Forest-based models for urban air quality forecasting. Future work will integrate meteorological parameters and explore advanced deep learning architectures to enhance predictive performance.*
***Keywords:*** *PM2.5 prediction, Random Forest, air pollution, machine learning, Seoul*

---

Date of Submission: 15-06-2025 Date of Acceptance: 30-06-2025

---

## I. Introduction

Air pollution, particularly fine particulate matter (PM2.5), presents a significant threat to global health, contributing to millions of premature deaths annually (World Health Organization, 2021). In urban environments like Seoul, South Korea, rapid industrialization, increased vehicular emissions, and transboundary pollution have exacerbated air quality challenges (Korea Environment Institute, 2020). The prediction of PM2.5 concentrations is critical for designing timely public health interventions and informing environmental regulations.

Traditional methods for air quality prediction, such as linear regression or autoregressive models, often struggle to model the complex non-linear relationships between atmospheric pollutants and meteorological factors. Recent advancements in machine learning (ML) have opened new pathways for more accurate forecasting. Among these, Random Forest (RF), an ensemble learning technique, has demonstrated high performance and interpretability in PM2.5 prediction tasks (Joharestani et al., 2019; Park and Kim, 2025). Random Forest models offer several advantages, including robustness to overfitting and the ability to model non-linear feature interactions without requiring extensive preprocessing.

This study applies Random Forest Regressor to predict PM2.5 levels in Seoul using a comprehensive air quality dataset spanning over three decades. By implementing robust preprocessing, feature engineering, and hyperparameter optimization, the study aims to evaluate the effectiveness of RF models in an urban Asian context, where air pollution dynamics are heavily influenced by both local and regional factors. The study also lays the foundation for future research involving integration of meteorological datasets and exploration of deep learning models for dynamic air quality forecasting.

## II. Data Description

The dataset spans from 1988 to 2021, collected from Seoul's air quality monitoring stations. Key features include hourly measurements of PM2.5, PM10, SO2, NO2, CO, and O3, along with timestamps and geographical coordinates. Data from 2021 was selected for consistency. After preprocessing, including removing rows with missing PM2.5 values and applying outlier removal via the IQR method, 196,688 records remained.

## III. Methodology

*3.1 Preprocessing and Feature Selection*

Initial data exploration revealed strong correlations between PM2.5 and other pollutants, particularly PM10, CO, and NO2, suggesting their potential significance as predictive features. These relationships are visually represented in the correlation matrix as shown in Figure 1, which highlights the strength and direction of the associations among the different variables. Based on these insights, PM10, CO, and NO2 were selected as input features for model training and evaluation.

Correlation Matrix of Pollutants

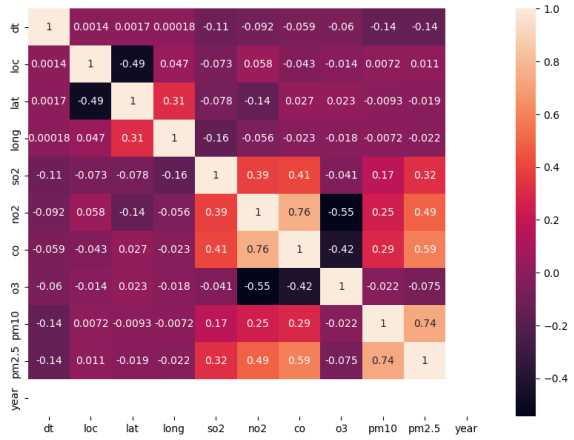


Figure 1. Correlation matrix of air pollutant features (PM2.5, PM10, CO, NO2) used for model development.

Before modeling, the dataset underwent a comprehensive preprocessing stage. This included handling missing values, standardizing timestamp formats, and performing outlier detection and removal to improve data quality and ensure robust model performance. Outliers were identified using the Interquartile Range (IQR) method and removed to minimize their impact on model accuracy. Figure 2illustrates the distribution of the selected features before outlier removal, while Figure 3 shows the distribution afteroutlier removal.Together, they provide a clear depiction of how the preprocessing step enhanced the dataset's reliability and suitability for regression modeling.

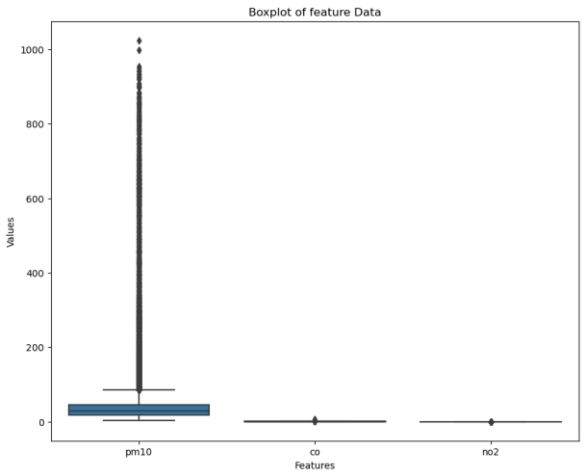


Figure 2. Boxplot shows the distribution of PM10, CO, and NO2 features before outlier removal.

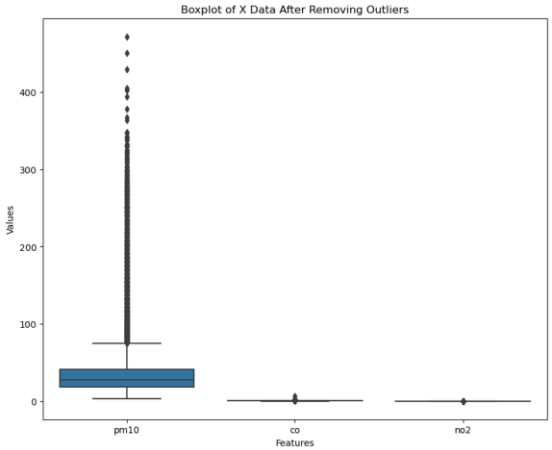


Figure 3. Boxplot shows the distribution of PM10, CO, and NO2 features after outlier removal.

*3.2 Model Development*

The Random Forest Regressor was chosen for its ability to model complex non-linear relationships while minimizing overfitting. The dataset was split into training and testing sets with an 80:20 ratio. The initial model without tuning achieved an RMSE of 5.26 and an R-squared of 0.69.

The Random Forest Regressor was chosen for its ability to model complex non-linear relationships while minimizing overfitting through ensemble learning. It builds multiple decision trees during training and averages their outputs for regression tasks, which increases prediction accuracy and robustness.

Mathematically, if $T_1(x)$,$T_2(x)$...,$T_K(x)$ are the predictions from K individual regression trees for an input$x$, then the final prediction$\hat{y}$ of the Random Forest is computed as:

$$\hat{y} = \frac{1}{K}\sum_{k=1}^{K} T_k(x)$$

This averaging helps reduce variance compared to a single decision tree, which is often prone to overfitting.

For this study, the dataset was split into training and testing sets using an 80:20 ratio. The initial Random Forest model, prior to any hyperparameter tuning, achieved a Root Mean Squared Error (RMSE) of 5.26 and an R-squared value of 0.69. These metrics served as a baseline to evaluate the improvement after model optimization.

*3.3 Hyperparameter Tuning*

To improve the model's predictive performance beyond the baseline configuration, hyperparameter tuning was conducted using GridSearchCV, a robust grid search method with cross-validation provided by the Scikit-learn library. This technique allows systematic exploration of multiple combinations of parameter values toidentify theconfiguration that minimizes prediction error.

The hyperparameters tuned in this study included:

- Number ofestimators (n_estimators): set to 300, representing the total number of decision trees in the forest.
- Maximum depth (max_depth): set to 30, limiting the depth of each decision tree to prevent overfitting.
- Minimum samples split (min_samples_split): set to 2, requiring at least two samples to split an internal node.
- Minimum samples per leaf (min_samples_leaf): set to 1, allowing a leaf node to contain a single observation.

A 3-fold cross-validation strategy was applied to ensure robustness of the model's performance across different data subsets. The model's accuracy was evaluated using negative mean squared error as the scoring metric. After identifying the optimal set of hyperparameters, the model was retrained and evaluated on the test data.

The optimized model yielded a Root Mean Squared Error (RMSE) of 4.88 and an R-squared ($R^2$) value of 0.72, showing a marked improvement compared to the untuned baseline model. These results highlight the effectiveness of hyperparameter tuning in enhancing the generalizability and predictive accuracy of Random Forest models.

Sample Decision Tree from Random Forest

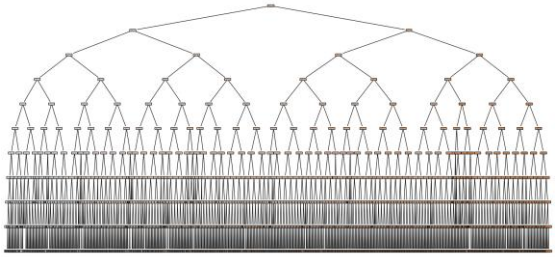

Figure 4. A sample decision tree extracted from the Random Forest model illustrating feature splits.

## IV. Results and Discussion

The results of this study align closely with findings from previous research applying Random Forest models for PM2.5 prediction.

Park and Kim (2025) applied Random Forest combined with Boruta feature selection to predict PM2.5 concentrations in Seoul, achieving an $R^2$ of 0.78 and an RMSE of 6.4 μg/m³, highlighting the method's robustness for urban air quality forecasting.

Similarly, Joharestani et al. (2019) compared Random Forest, XGBoost, and deep learning models for PM2.5 forecasting in Tehran, reporting that Random Forestachieved an $R^2$ of 0.76 and an RMSE of 8.3 μg/m³, demonstrating strong predictive capabilities with limited meteorological inputs.

Likewise, Makhdoomi et al. (2025) applied Random Forest and other ensemble methods to predict PM2.5 concentrations across multiple cities, reporting $R^2$ values ranging from 0.70 to 0.85 and RMSE values as low as 2.31 μg/m³, further demonstrating the reliability of Random Forest models in diverse urban environments.

In comparison, the Random Forest model developed in this study achieved an RMSE of 4.88 μg/m³ and an $R^2$ of 0.72, indicating competitive or superior performance relative to previous studies. These consistent results across different geographic regions confirm the effectiveness of Random Forest models for urban PM2.5 forecasting.

Future improvements could integrate additional meteorological variables, socio-economic factors, and climate change indicators to enhance predictive performance, enabling more dynamic and long-term forecasting of air quality under evolving environmental conditions.

## V. Conclusion

This study validated the effectiveness of Random Forest Regression in predicting PM2.5 levels in Seoul. Systematic data preprocessing, careful feature selection, and hyperparameter tuning significantly improved model performance. Future work should explore real-time prediction systems and the application of deep learning models.

## References

[1]. World Health Organization (2021). WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization. https://www.who.int/publications/i/item/9789240034228

[2]. Park, S., & Kim, M. J. (2025). Forecasting Ultrafine Dust Concentrations in Seoul: A Machine Learning Approach. Atmosphere, 16(3), 239. https://doi.org/10.3390/atmos16030239

[3]. Joharestani, Z. M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. Atmosphere, 10(7), 373. https://doi.org/10.3390/atmos10070373

[4]. Korea Environment Institute (2020). Annual Report on Air Quality in Korea 2020. Korea Ministry of Environment.

[5]. Makhdoomi, A., Sarkhosh, M., & Ziaei, S. (2025). PM2.5 concentration prediction using machine learning algorithms: An approach to virtual monitoring stations. Scientific Reports, 15, Article 8076. https://doi.org/10.1038/s41598-025-92019-3