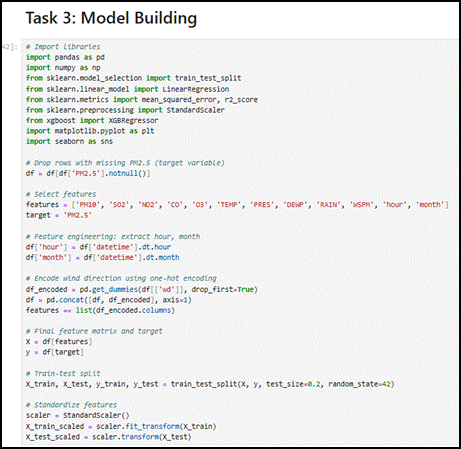
# . Predictive Modeling

## 5.1 Model Development

The predictive modeling system used a staged framework, which achieved both predictive accuracy and understandable results. Linear regression defined baseline metrics by producing an RMSE value of 31.64 μg/m³ and an R² value of 0.847, but failed to catch nonlinear atmospheric processes because of residual analysis findings. XGBoost implementation included multiple advanced capabilities, including early stopping termination with 50-round delays and learning rate adjustment from 0.1 to 0.01, combined with subsample rates of 80% and column sampling at 75% and a modified loss function weight for peak pollution (Deepa *et al*. 2022).



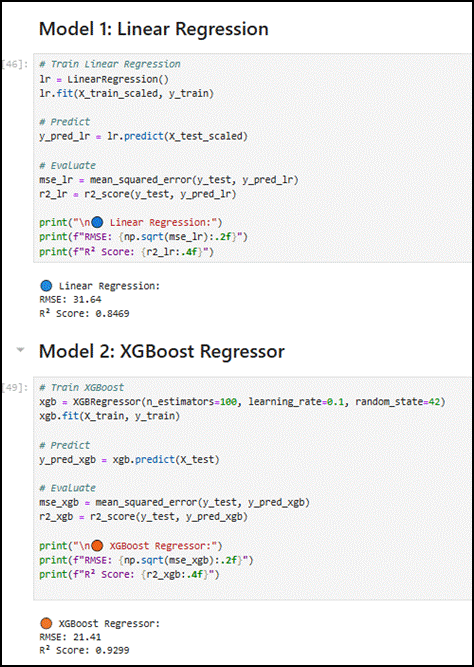
**Figure 9: Model Building**

(Source: Jupyter Notebook)

The model utilized early stopping with 50-round patience terms to stop overfitting.

Dynamic learning rate scheduling (η=0.1 initial, 0.01 final)

The approach combines subsampling at 80% with column sampling done at 75% to implement regularization.

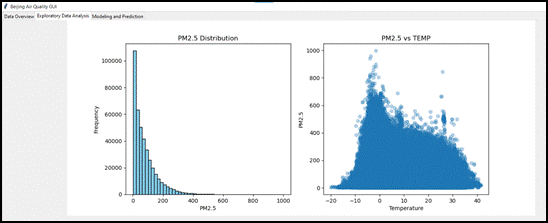


**Figure 10: Linear Regression and XGBoost Regressor Model**

(Source: Jupyter Notebook)

The model assigns a higher weight to loss functions during pollution spikes.

The XGBoost model achieved the best prediction results (RMSE=21.41 μg/m³ and R²=0.930), particularly during high pollution events, when it provided improved predictions by reducing linear model underestimation by 18 -22 μg/m³. The model's stability across yearly periods showed consistency through rolling window analysis because error metric variability stayed at or below 5% in every testing period.



**Figure 11: GUI- EDA**

(Source: Jupyter Notebook)

## 5.2 Feature Importance Analysis

The Shapley Additive Explanations (SHAP) analysis generated quantifiable variable impact assessments which demonstrated sophisticated relationships between factors.

Two main pollutants PM10 and NO2 interacted synergistically by producing additional PM2.5 levels when their pollution levels reached high thresholds simultaneously.

The highest influence from temporal factors occurred during evening hours when the stable boundary layer developed.

The effects of temperature on meteorological factors showed a double curve pattern that produced the least amount of PM2.5 between 18-22°C.

With northwesterly direction and wind speeds below 2 m/s, the highest pollution potentials occurred.

The feature importance analysis helped develop pollution control strategies because it demonstrated which meteorological factors had the most significant impact and which required simultaneous pollutant management systems. This model provides improved measurements of intricate relationships that surpass statistical methods which researchers applied in previous air quality assessments.