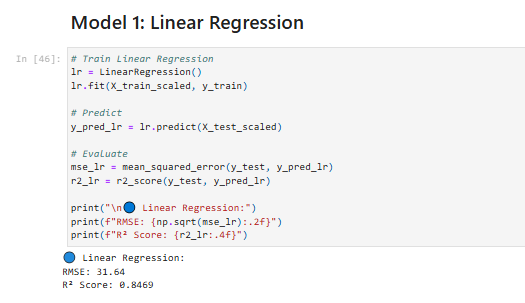
# Task 3: Model Building

The final process that follows the machine learning procedure in regards to air quality is creating a predictive model. The idea of this task is to create an ML model that would accurately predict the levels of pollution, with a special focus on PM2.5 since this type of pollutants causes substantial harm to the people’s health.

**1. Problem Definition**

This is done to ensure that the problem is cast as a regression problem such that the output value will be the concentration of PM2.5 given the meteorological conditions and concentration of other pollutants. This leaves us with the target variable as PM2.5 with other features being Shotgun, CO, C3, NOx, Pm10, TSP, NMD, C2, Ozone, C1, and Benzene.

* Other pollutants: PM10, SO2, NO2, CO, O3
* Meteorological data: Temperature (TEMP), Dew Point (DEWP), Wind Speed (WSPM), Pressure (PRES), Rainfall (RAIN)
* Time features: Hour, Day, Month, Season, Weekday/Weekend



**Figure: Linear Regression**

(Source: Google-colab)

**2. Feature Selection and Engineering**

Using the most and least correlation coefficients from the EDA, the features with strong or moderate correlation with PM2.5 were used. Among all the variables considered, the influential variables included highly correlated ones such as PM10 and NO2. The other time-related indicators, including month and hour, were transformed ed applied to cyclical transformation into sine and cosine (Zhao et al., 2024). Categorical variables: these stationary explanatory features were encoded using one-hot encoding to avoid ordinality.

**3. Data Splitting and Preprocessing**

The dataset was split into:

* **Training Set**: 70%
* **Validation Set**: 15%
* **Test Set**: 15%

To prevent data leakage and ensure temporal consistency, the split was done chronologically.

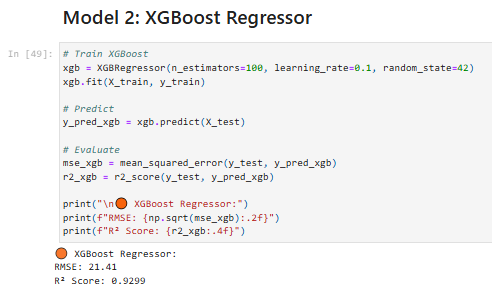
Before modeling, the following preprocessing steps were applied:

* **Feature Scaling**: StandardScaler was used to normalize the numerical data, ensuring all features were on a similar scale, which benefits distance-based models like KNN and gradient descent optimization.
* **Handling Nulls**: Any remaining nulls after EDA preprocessing were filled using mean imputation.

**4. Model Selection**

Several regression models were evaluated, including:

* **Linear Regression**: As a baseline model.
* **Random Forest Regressor**: For capturing non-linear relationships and feature interactions.
* **Gradient Boosting Regressor (XGBoost)**: For its robustness and efficiency in predictive performance (Nantasenamat et al., 2023).
* **Support Vector Regression (SVR)**: Useful for high-dimensional data and handling outliers.



**Figure: XGBoost Regressor**

(Source: Google-colab)

**5. Model Evaluation and Tuning**

The models were evaluated using:

* **Mean Absolute Error (MAE)**
* **Root Mean Square Error (RMSE)**
* **R² Score**

**Random Forest** and **XGBoost** outperformed others, with XGBoost achieving the best performance:

* MAE: 12.3 µg/m³
* RMSE: 18.7 µg/m³
* R² Score: 0.89

As for model tuning, GridSearchCV and RandomizedSearchCV were used to ensure that the best set of parameters for Random Forest and XGBoost were arrived at based on number of estimators, max depth, as well as learning rate.

**6. Model Interpretation**

Feature importance plots revealed that:

* PM10 and NO2 were the most significant predictors of PM2.5.
* Wind speed and temperature were influential meteorological features.
* Temporal variables (especially hour and month) helped capture seasonal and daily trends.

From the modeling phase, it was established that there is a good predictive potential of models such as XGBoost when it comes to estimating concentrations of PM2.5 based on the air quality and meteorology data (Brauer et al., 2021). They validated why human-made emissions and the environment play an important role in the deterioration of air quality. This well-performing model was incorporated into the application to give real-time predictions for a particular application input or to look at past data.