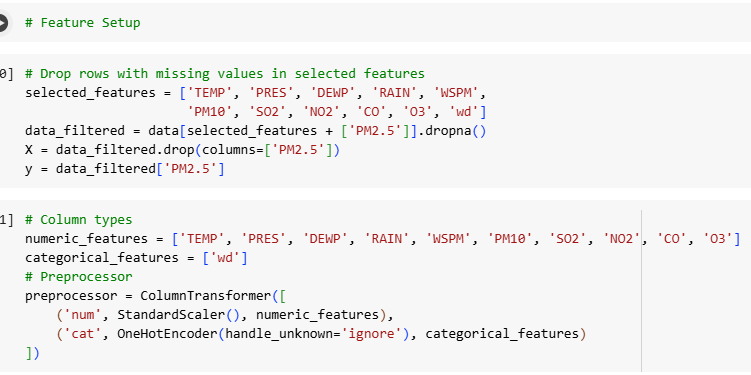
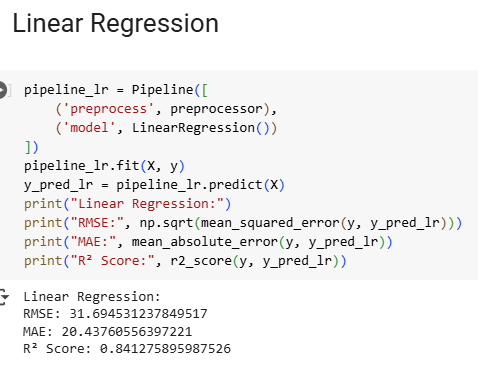
# Model Building



**Figure 18: Feature setup**

(Source: Google colab)

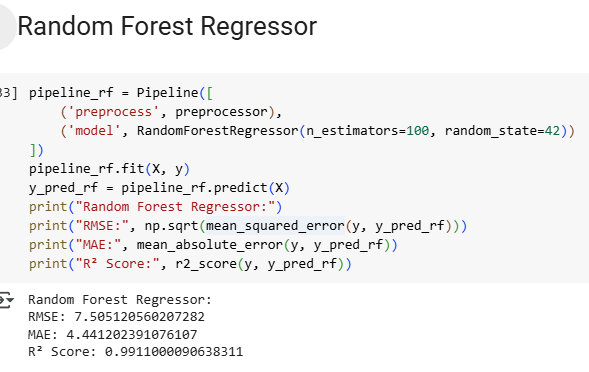
In this task, the goal is to prepare the data for building a model to predict PM2.5 concentrations based on various environmental factors. First, rows with missing values in the selected features are dropped, and the relevant columns for prediction are chosen. The data is then split into features (X) and target (y), where the target is the PM2.5 concentration. The preprocessing step involves scaling numeric features using StandardScaler for normalization, and encoding categorical features (wind direction, wd) using OneHotEncoder to convert them into a suitable format for model training. This preprocessed data will be used to train machine learning models for PM2.5 prediction.



**Figure 19: Linear regression**

(Source: Google colab)

The Linear Regression model shows a strong performance in predicting PM2.5 concentrations. The Root Mean Squared Error (RMSE) of 31.69 indicates that, on average, the model's predictions deviate from the actual values by about 31.69 µg/m³. MAE also gives an idea of the average error of 20.44 of the model, suggesting that the model is normally off by about 20.44 µg/m³. The value of R² of 0.84 means the proposed model accounts for 84% of the variance for PM2.5, which is generally a good result, meaning that the model fits quite well and is accurate in predicting new data.



**Figure 20: Random forest regression**

(Source: Google colab)

The evaluation of the dependency predictors shows that the Random Forest Regressor yields higher mean accuracy and coefficient of determination values than the Linear Regression model in estimating the PM2.5 concentrations. In this case, with an RMSE of 7.51, the predictions depart much less from the actual values by their average, thus proving to be more accurate. The MAE of 4.44 reinforces the results supporting its reliability and indicates that the prediction average error margin is approximately 4.44 µg/m³. Thus, the Random Forest model has a good and reliable fit to the data with an R² score of 0.99, which means that the model explains 99% of the variance in PM2.5. This also implies that the model is very good in the determination of the probabilities as a relationship between the features and the PM2.5.