ISSN: 2581-1274

Implementation of Multiple Linear Regression Algorithm to Predict Air Temperature Based on Pollutant Levels in South Tangerang City

Tedja Diah Rani Octavia [a,1], Neny Rosmawarni [a,2*], Ati Zaidiah[a,3], Nunik Destria Arianti [b,4]

[a] Informatics Engineering, Faculty of Computer Science Universitas Pembangunan Nasional Veteran Jakarta;
[b] Universitas Nusa Putra.
raniranoc@gmail.com [1], nenyrosmawarni@upnvj.ac.id[2], atizadiah@upnvj.ac.id [3], nunik@nusaputra.ac.id [4].*

*corresponding author

ARTICLE INFO

Article history:
Published

Keywords:
Air Temperature,
Pollutants,
Multiple Linear Regression

ABSTRACT

Global warming is a phenomenon that has widespread effects, particularly on environmental aspects. Globally, the impact of global warming is evident in the rising temperatures of the Earth. In April 2023, much of South Asia experienced a heatwave with temperatures exceeding 40°C. In Indonesia, the daily maximum temperature recorded reached 37.2°C in South Tangerang City. Global warming is caused by the increasing concentration of greenhouse gases in the Earth's atmosphere. This study proposes a model for predicting air temperature by considering the influence of pollutant levels and daily climate data in South Tangerang City. The prediction modeling in this study uses the Multiple Linear Regression algorithm with an 80% training data and 20% testing data split. Out of 8 trials, the sixth model is the best with a k value of 8 and features including RH_avg, RR, ss, ddd_car, ff_avg, no2, o3, and pm10. The evaluation results of the sixth model yielded an R^2 value of 0.72749, MAE of 0.55593, MSE of 0.50078, and MAPE of 1.99806%.

I. Introduction

Global warming has emerged as one of the most significant challenges faced by humanity today. This phenomenon has widespread impacts, particularly on environmental aspects. Globally, the consequence of global warming is the increase in Earth's temperature [1]. This leads to a domino effect that includes the melting of polar ice, rising sea levels, and increasingly severe heatwaves [2].

In April 2023, a significant portion of South Asian countries experienced an extreme heatwave. Meteorological agencies in countries such as Bangladesh, Myanmar, India, China, Thailand, and Laos reported high temperatures exceeding 40°C, with new records set in several regions. In Indonesia, the daily maximum temperature recorded reached 37.2°C at the BMKG station in Ciputat, with the highest temperatures in various locations ranging between 34°C and 36°C [3].

Global warming and climate change are caused by the increased concentration of greenhouse gases in the atmosphere. The precise mechanism of global warming and climate change remains uncertain. However, research suggests that global warming and climate change are due to increased concentrations of greenhouse gases, particularly CO_2, CH_4, N_2O, CFCs, and ozone (O_3) in the Earth's atmosphere [4].

This study investigates the prediction of average air temperature considering the impact of pollutant levels, such as CO, NO, NO_2, O_3, SO_2, $PM_{2.5}$, PM_{10}, and NH_3. The machine learning modeling in this study employs Multiple Linear Regression, a statistical method used to model the relationship between a dependent variable and independent variables.

II. Literature Review

Global warming is a phenomenon characterized by the rise in Earth's average surface temperature due to increasing emissions of greenhouse gases in the atmosphere [5]. The increase in greenhouse gases, particularly CO_2 [6] and CFCs, which contribute to global warming, can result from human activities [4]

such as industrial activities, methane emissions from agriculture and livestock, vehicle emissions, deforestation, and more.

Global warming leads to various subsequent issues, including climate change, melting polar ice, increased rainfall and flooding, rising sea levels, and the extinction of certain flora and fauna. The most noticeable problem is the increase in Earth's surface temperature. The warming of the surface over recent decades and the anticipated further warming are central to climate change and are the cause of many other problems related to global warming that affect humans directly or indirectly [7].

Pollutants are defined as substances that cause pollution and are harmful to humans and other living organisms when released into the environment. Pollutants can be in the form of solids, liquids, or gases produced in concentrations higher than usual, leading to reduced environmental quality [8]. Air pollutants are harmful elements in the air present in large amounts over extended periods. Common air pollutants include particulate matter (PM) and ground-level ozone. Other examples include dispersed particles, hydrocarbons, CO, CO_2, NO, NO_2, SO_3, and others.

Temperature is something that can be measured with a thermometer [9]. In physics, temperature measures the kinetic energy of particles within a substance. Simply put, the faster the particles move, the hotter the substance. Temperature reflects the average kinetic energy of molecules in an object. It can also be defined as a measure of how hot an object is, usually measured in degrees Celsius (°C), Fahrenheit (°F), or Kelvin (K) [10].

A. Linear Regression

Linear regression is a statistical method used to predict a quantitative response Y based on a single predictor variable X. In linear regression, a linear relationship between X and Y is assumed. Mathematically, linear regression can be formulated as in Equation 1.

$$Y = \beta_0 + \beta_1 X + \epsilon . \quad (1)$$

Where:
Y = Dependent variable
X = Independent variable
β_0 = Intercept
β_1 = Regression coefficient (slope)
ϵ = Error

In this equation, β_0 represents the intercept or the value of Y when X = 0, β_1 represents the slope or the average increase in Y associated with a one-unit increase in X. The error ϵ, represents random variability not explained by the model [11].

B. Multiple Linear Regression

Multiple Linear Regression is a statistical method used to analyze the relationship between one dependent variable and multiple independent variables. It is an extension of simple linear regression, which uses only one predictor variable. Multiple Linear Regression can accommodate multiple predictors simultaneously, helping to understand how these predictors collectively influence the dependent variable. The Multiple Linear Regression model with p predictors can be formulated as in Equation 2.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon . \quad (2)$$

Where:
Y = Dependent variable
β_0 = Intercept
$\beta_1 \ldots \beta_p$ = Regression coefficients
$X_1 \ldots X_p$ = Independent variables
ϵ = Error

In this formula, Y represents the variable to be predicted. X_j represents the j-th predictor variable, and β_j measures the relationship between those variables with responses. β_0 is the intercept representing Y when all X variables are zero. β_j represents the change in Y for a one-unit change in X_j, keeping

other predictors constant [11].

C. Ordinary Least Squares

Ordinary Least Squares (OLS) is a statistical method used to estimate the coefficients in linear regression equations. OLS identifies the relationship between the target variable and one or more predictor variables. It estimates the unknown parameters in the regression model by minimizing the sum of the squared differences between the observed and predicted values [12]. OLS provides a closed-form solution, meaning it directly calculates the optimal coefficients without iterative optimization. The OLS calculation involves three steps: adding an intercept by including a constant column (1) in the feature data X, forming the design matrix X including all features, and calculating the regression coefficients using the OLS formula. Mathematically, OLS is defined as:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3}$$

Where:
$\hat{\beta}$ = Regression coefficient vector $[\beta_0, \beta_1 \ldots, \beta_p]^T$
X = Matrix of independent variables
y = Vector of dependent variables
X^T = Transpose of matrix X
$(X^T X)^{-1}$ = Invers of matrix $X^T X$

D. Coefficient of Determination

The coefficient of determination, or R^2, is a value that shows the extent to which independent variables influence the dependent variable. R^2 represents the proportion of variance in the dependent variable that can be explained by the independent variables. R^2 acts as an evaluation metric in regression to assess how well data points fit around the regression line (Keer et al., 2023). The value of R^2 ranges from 0 to 100 percent. A value of 0 percent means the model does not explain the variability of the response variable around the mean. Conversely, a value of 100 percent means the model explains the variability of the response variable around the mean (. R^2 can be formulated as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}. \tag{4}$$

Where:
y_i = Actual value of the i-th observation
$\bar{y}$ = Mean of the dependent variable
$\hat{y}_i$ = Predicted value of the i-th observation

E. Mean Absolute Error

Mean Absolute Error (MAE) is an evaluation metric that measures the average magnitude of errors in predictions, without considering their sign. MAE is the average of the absolute errors or the differences between predicted values and actual values [12]. MAE is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| . \tag{5}$$

Where:
y_i = Actual value
$\hat{y}_i$ = Predsicted value
n = Number of data samples

F. Mean Squared Error

Mean Squared Error (MSE) evaluates the model by calculating the difference between the predicted

values and the actual values, then squaring these differences to eliminate negative values. The squared differences are summed and divided by the number of data samples . MSE is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\text{y}_\text{i} - \hat{\text{y}}_\text{i})^2 \ . \tag{6}$$

Where:
n = Number of data sample
y_i = Actual value
$\hat{y}_i$ = Predicted value

A regression model or a machine learning is considered better at predicting numerical values if the MSE value is smaller [13]. Conversely, a larger MSE indicates poorer model performance and accuracy in predicting numerical values.

G. `Mean Absolute Percentage Error

Mean Absolute Percentage Error (MAPE) is a metric that measures the average absolute difference between predicted and actual values expressed as a percentage of the actual values. MAPE is used to assess the accuracy of prediction models. It provides an indication of the magnitude of prediction errors relative to the actual values [14]. MAPE is defined as:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\text{y}_\text{i}-\hat{\text{y}}_\text{i}}{\text{y}_\text{i}}\right| \times 100 \ . \tag{7}$$

Where:
n = Number of data samples
y_i = Actual value
$\hat{y}_i$ = Predicted value

MAPE is advantageous as an evaluation metric because it is expressed as a percentage, making it easier to understand. It also allows for comparison of prediction errors across different scales of data. Therefore, MAPE is suitable for describing the percentage error in regression model predictions.

III. Research Methodology

This study will apply the Multiple Linear Regression algorithm to develop a temperature prediction model. The research will begin with problem identification and literature review. Data collection, data preprocessing, data splitting, and model design will follow. The final steps will include testing and evaluating the model.

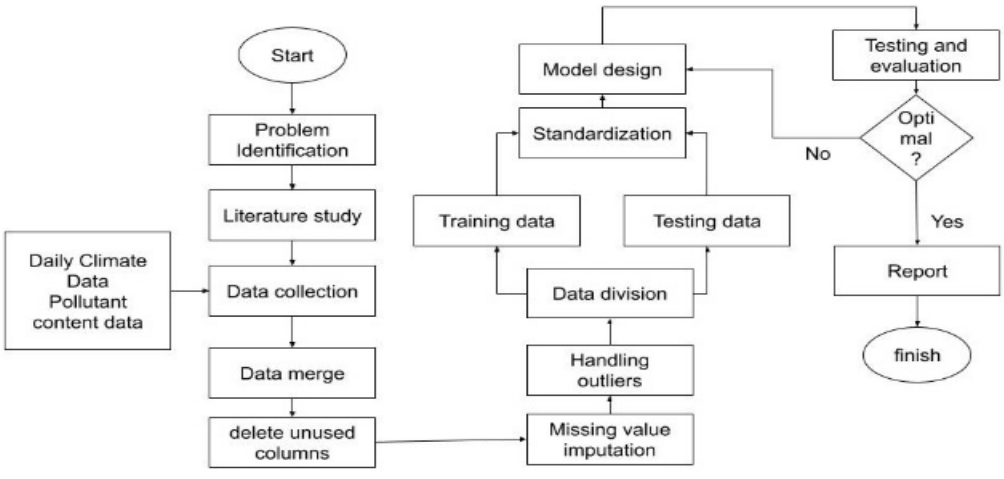

Fig 1. Research Stages

IV. Results and Discussion

A. Data

In this study, two datasets were used: daily climate data from the BMKG website and pollutant concentration data from the OpenWeather website. The daily climate data comprises 1156 records from November 1, 2020, to December 31, 2023. The raw daily climate data includes the date, Tn (minimum temperature (°C)), Tx (maximum temperature (°C)), Tavg (average temperature (°C)), RH_avg (average humidity (%)), RR (rainfall (mm)), ss (sunshine duration (hours)), ff_x (maximum wind speed (m/s)), ddd_x (wind direction at maximum speed (°)), ff_avg (average wind speed (m/s)), and ddd_car (most frequent wind direction (°)).

The pollutant concentration data was collected from the OpenWeather Air Pollution API, consisting of 26,800 hourly records per day from November 25, 2020, to December 31, 2023. The raw pollutant concentration data includes aqi (air quality index on a scale of 1-5), co (carbon monoxide (μg/m³)), no (nitric oxide (μg/m³)), no2 (nitrogen dioxide (μg/m³)), o3 (ozone (μg/m³)), so2 (sulfur dioxide (μg/m³)), pm2_5 (particulate matter ≤2.5 micrometers (μg/m³)), pm10 (particulate matter ≤10 micrometers (μg/m³)), and nh3 (ammonia (μg/m³)).

The daily climate data and pollutant concentration data were merged into a single dataframe named df_combined. The hourly pollutant concentration data was averaged to match the daily climate data. The datetime column in the pollutant concentration data and the Tanggal column in the daily climate data were reformatted to the same date format to facilitate merging based on the date. The final merged dataset contains 1128 rows and 21 columns.

B. Exploratory Data Analysis

The merged data underwent Exploratory Data Analysis (EDA). This phase involved examining correlations between variables using a heatmap, identifying linear/non-linear relationships and detecting outliers with scatter plots, and understanding data distributions to determine imputation techniques using histograms.

1) Heatmap

The heatmap visualization revealed key information: RH_avg and Tavg have a strong negative correlation, indicating that lower average humidity corresponds to higher average daily temperature, and vice versa. Additionally, all pollutant concentrations except o3 showed moderate to strong positive correlations with other pollutant concentrations.

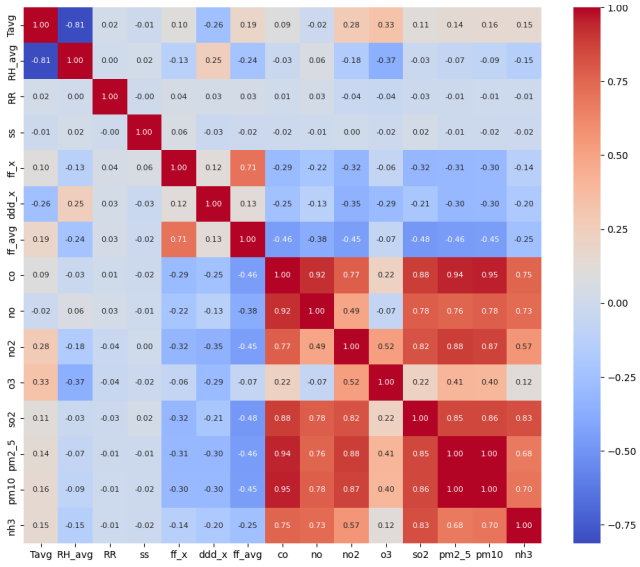

Fig 2. Heatmap or Correlation Map

2) Scatter Plot

The scatter plots indicated that RH_avg has a strong linear relationship with Tavg, suggesting that humidity significantly influences temperature. There were also some outlier data points, such as one outlier each in the o3 vs. Tavg scatter plot and the ss vs. Tavg scatter plot. These outliers need to be addressed in the preprocessing stage.

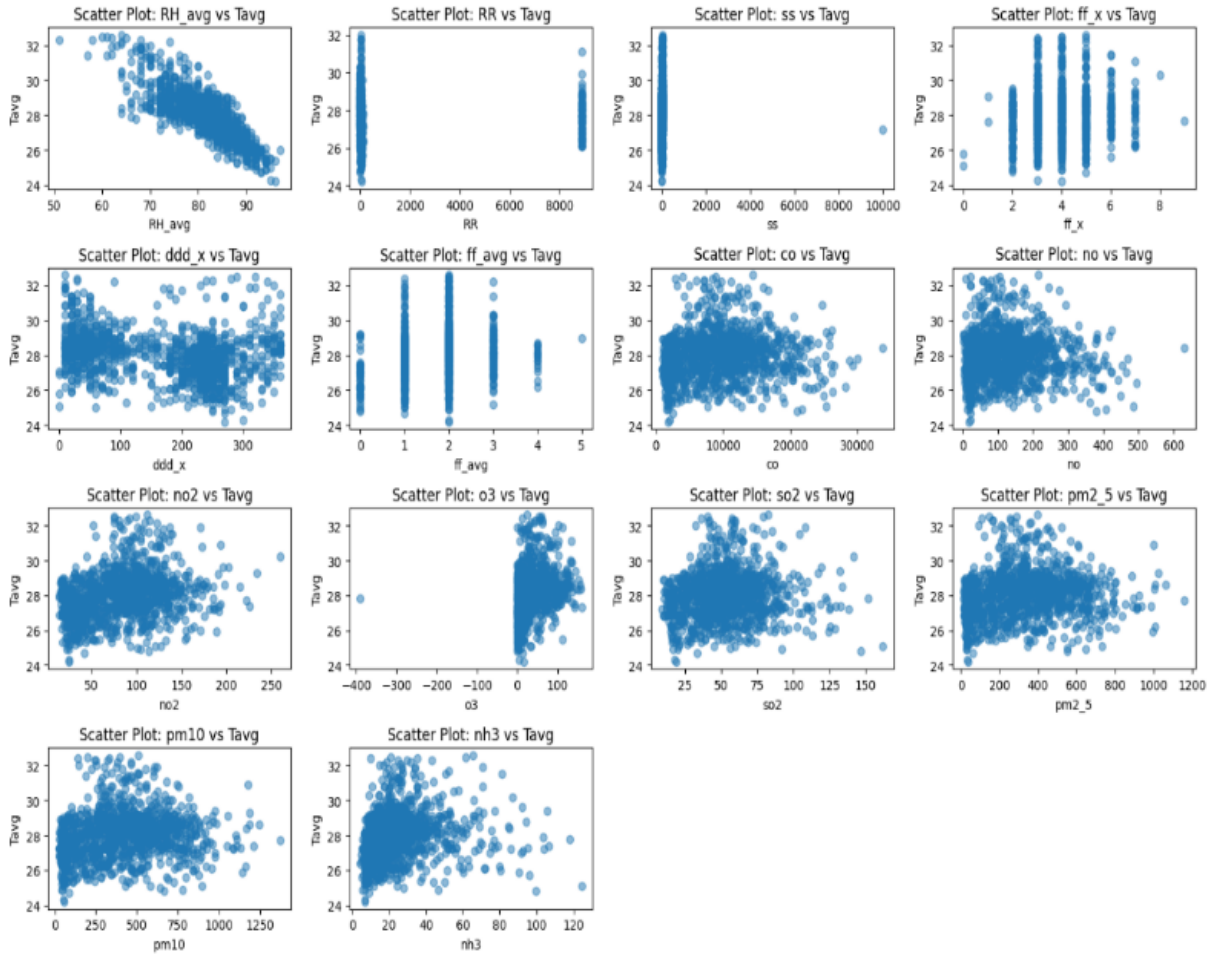

Fig 3. *Scatter Plot*

3) Histogram

The variables RH_avg and Tavg tend to exhibit a normal distribution, indicating that their mean and median values are approximately equivalent. In contrast, the other variables display skewed or non-normal distributions. For instance, the variables RR, CO, NO, NO2, O3, SO2, PM2.5, PM10, and NH3 exhibit rightward skewness (positive skewness). Consequently, the variables RH_avg and Tavg, which have approximately normal distributions, will be imputed with their mean values, whereas the other variables, which exhibit skewed distributions, will be imputed with their median values.

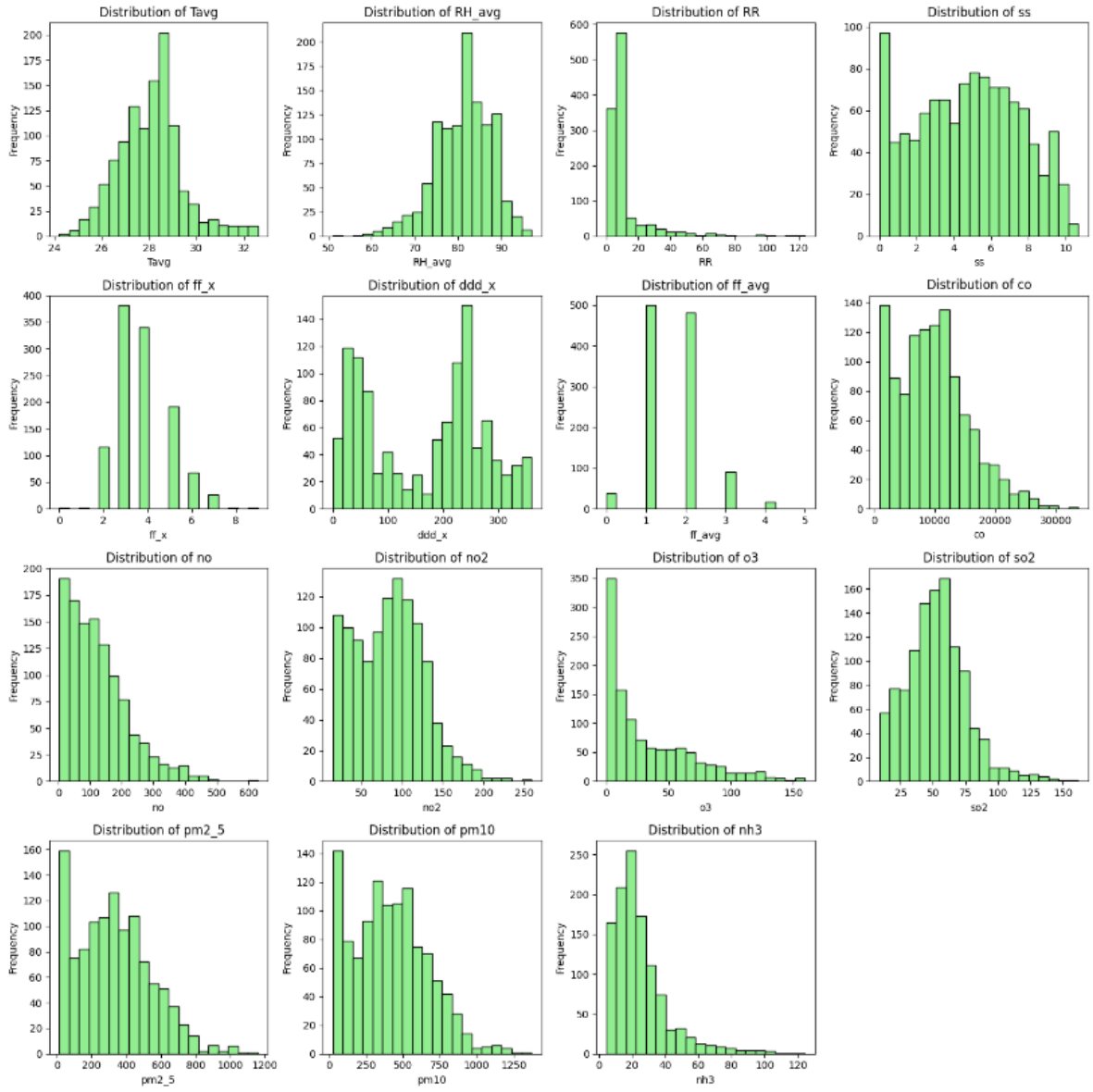

Fig 4. *Histogram* The variables RH_avg and Tavg tend

C. Data Preprocessing

In the dataframe df_combined, several columns were dropped because they were not used in predicting the output and did not provide relevant information. This step was taken to make the code more efficient and to reduce the resource load. The columns dropped at this stage include Tn, Tx, ddd_car, datetime, and aqi.

The next step is imputing missing values. Imputation is performed to fill in missing or null values (NA, NaN, or Null) in variables with substitute values using specific techniques tailored to the data, such as mean values, median values, and so on. In the previously merged data, df_combined, based on information from the BMKG site, if a variable value is 8888 or 9999, it signifies that the data was not measured or is missing (no measurement was taken). Consequently, values of 8888 and 9999 in the dataset are replaced with NaN.

Subsequently, the count of null or missing values across the entire dataset is calculated. This calculation is performed to identify which variables have null values and the extent of these missing values. This step is crucial for determining the appropriate values for imputation. Table 4.4 displays the number of missing values for variables Tavg, RH_avg, RR, ss, ff_x, and ddd_x. For the normally

distributed data Tavg and RH_avg, missing values are imputed with the mean of each respective variable. For other variables, such as RR, ss, ff_x, and ddd_x, which exhibit non-normal distributions, imputation is performed using the median.

Next, outliers are handled. This step is taken as a preventive measure against the negative impacts on statistical analysis and predictive models due to significantly different data values. Based on the scatter plot from the previous EDA stage, variables o3 and ss were identified as containing outliers. In this step, outliers are replaced with the median value based on the distribution analysis from the EDA phase. After preprocessing and outlier handling, the dataset consists of 1128 rows and 16 columns.

The following step involves splitting the data into training and testing sets. Specifically, the data is divided based on variables: X containing independent variables and y containing the dependent variable. Using random subsampling with the train_test_split library, the data is divided into 80% training data and 20% testing data. The divided data is stored in four variables: X_train, X_test, y_train, and y_test. This partitioning is done to assess how well the model can predict unseen data and to avoid overfitting.

After splitting, the data is standardized. Standardization or feature scaling is performed to normalize feature values so that they have a uniform scale. This is achieved by subtracting the mean of each feature and then dividing by the feature's standard deviation. Standardization is carried out using the StandardScaler from the scikit-learn library. This feature scaling is applied only to X_train and X_test. These scaled variables are stored in X_train_scaled and X_test_scaled to ensure that the model performs well when evaluating the test data or unseen data during training.

D. Model Development and Evaluation

This study involves testing eight different model scenarios. The first experiment uses all pollutant concentration features. The second experiment uses only daily climate data features. These experiments are designed to assess the impact of each feature from both datasets individually on model evaluation outcomes. The third experiment models the combined features from both datasets. Experiments four through eight involve feature selection using SelectKBest ANOVA with the highest F-value as the basis for feature selection, varying k from 6 to 10.

The selection of the best model is based on evaluation metrics, namely R^2, MAE, and MSE. he final model chosen as the best predictive model is the one with the highest performance and best metrics. The best-performing model in this study is identified by having the highest R^2 and and the lowest MAE and MSE. The scenarios and their evaluation results are as follows:

Table 1. Evaluation Results of Experiments

The K-th Experiment	K	Fiture	R^2	MAE	MSE
1	-	*co, no, no2, o3, so2, pm2_5, pm10, nh3*	0.30077	0.91385	1.28500
2	-	*RH_avg, RR, ss, ff_x, ddd_x, ff_avg*	0.68902	0.60391	0.57148
3	-	*RH_avg, RR, ss, ff_x, ddd_x, ff_avg, co, no, no2, o3, so2, pm2_5, pm10, nh3*	0.72105	0.55971	0.51262
4	6	*RH_avg, ss, ddd_x, ff_avg, no2, o3*	0.70697	0.57755	0.53850
5	7	*RH_avg, RR, ss, ddd_x, ff_avg, no2, o3*	0.72023	0.56544	0.51413
6	8	*RH_avg, RR, ss, ddd_x, ff_avg, no2, o3, pm10*	0.72749	0.55593	0.50078
7	9	*RH_avg, RR, ss, ddd_x, ff_avg, no2, o3, pm10, nh3*	0.72413	0.55784	0.50697
8	10	*RH_avg, RR, ss, ddd_x, ff_avg, no2, o3, pm2_5, pm10, nh3*	0.72353	0.55748	0.50806

Based on the evaluation metrics R^2, mean absolute error (MAE), and mean squared error (MSE) from the above experiments, most model evaluation results, particularly MAE and MSE, show only minor differences. This indicates that the model errors are relatively consistent, with few large outliers, as evidenced by the generally stable MSE values, which are sensitive to outliers. However, the first experiment, which uses only pollutant concentration data, yields the worst evaluation results, especially with a very high MSE compared to the MAE. This suggests that the pollutant concentration data alone may contain significant outliers affecting the model, or the model built with only pollutant data is less suitable for predicting temperature.

From these experiments, it can be concluded that the best model is the sixth one, which uses features RH_avg, RR, ss, ddd_x, ff_avg, no2, o3, and pm10 selected using SelectKBest ANOVA F-value with k=8. This model has an R^2 of 0.72749, MAE of 0.55593, and MSE of 0.50078, making it the model with the highest R^2 and the lowest MAE and MSE, thus considered the best model from the experiments.

E. Implementation of Multiple Linear Regression

The implementation of the multiple linear regression algorithm on the best model from the sixth experiment, to predict air temperature based on pollutant concentrations, can be simply carried out using the LinearRegression class from the scikit-learn library. The Python implementation is as follows:

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X_train, y_train)
```

The first line of the code imports the LinearRegression class from the scikit-learn library. In the second line, an object of the LinearRegression() class is initialized into a variable named model. The third line trains or fits the model on the training data X_train and target y_train. Internally, the LinearRegression class from scikit-learn uses the ordinary least squares (OLS) method to find the regression coefficients that minimize the sum of squared residuals between the predictions and the actual values. The steps for implementing the OLS method in Python involve calculations as described in the following equations:

1) Forming Matrix X

Matrix X is of size 902×9 where the first column is filled with a constant value of 1 to compute the intercept or β_0. The other columns are filled with the independent variable data $X_1 \; through \; X_8$ in sequence, namely RH_avg, RR, ss, ddd_x, ff_avg, no2, o3, and pm10. The matrix is written as follows:

$$X = \begin{bmatrix} 1 & -0.31929 & \cdots & -0.32116 & 0.69023 \\ 1 & 1.36819 & \cdots & -0.93680 & 0.88881 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & -1.08634 & \cdots & 1.68819 & 2.08893 \\ 1 & 0.14092 & \cdots & -0.72085 & -0.65221 \end{bmatrix}$$

2) Forming Matrix Y

Matrix Y contains the actual values of the dependent variable, which is the daily average air temperature (Tavg). This matrix is of size \(902 \times 1\) and is written as follows:

$$Y = \begin{bmatrix} 28.8 \\ 26.4 \\ \cdots \\ 28.6 \\ 29.2 \end{bmatrix}$$

3) Transpose of Matrix X^T

The previously created matrix X is transposed to form matrix X^T of size \(9 \times 902\). The matrix X^T is written as follows:

$$X^T = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ -0,31929 & 1,36819 & \cdots & -1,08659 & 0,14092 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -0,32116 & -0,93680 & \cdots & 1,68819 & -0,72086 \\ 0,69023 & 0,88881 & \cdots & 2,08893 & -0,65221 \end{bmatrix}$$

4) Calculating X^TX

Next, the result of X^T is multiplied by matrix X to produce matrix X^TX of size \(9 \times 9\). The matrix X^TX is as follows:

$$X^TX = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ -0{,}31929 & 1{,}36819 & \cdots & -1{,}08659 & 0{,}14092 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -0{,}32116 & -0{,}93680 & \cdots & 1{,}68819 & -0{,}72086 \\ 0{,}69023 & 0{,}88881 & \cdots & 2{,}08893 & -0{,}65221 \end{bmatrix} \times \begin{bmatrix} 1 & -0{,}31929 & \cdots & -0{,}32116 & 0{,}69023 \\ 1 & 1{,}36819 & \cdots & -0{,}93680 & 0{,}88881 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & -1{,}08634 & \cdots & 1{,}68819 & 2{,}08893 \\ 1 & 0{,}14092 & \cdots & -0{,}72085 & -0{,}65221 \end{bmatrix}$$

$$= \begin{bmatrix} 902 & -6{,}18 \times 10^{-13} & \cdots & 1{,}14 \times 10^{-14} & 1{,}38 \times 10^{-13} \\ -6{,}18 \times 10^{-13} & 902 & \cdots & -330{,}21850 & -46{,}97333 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1{,}14 \times 10^{-14} & -330{,}21850 & \cdots & 902 & 382{,}56813 \\ 1{,}38 \times 10^{-13} & -46{,}97333 & \cdots & 382{,}56813 & 902 \end{bmatrix}$$

5) Invers of Matrix $(X^TX)^{-1}$

Matrix X^TX is then inverted to obtain matrix $(X^TX)^{-1}$ as follows:

$$(X^TX)^{-1} = \begin{bmatrix} 0{,}001108 & -1{,}1 \times 10^{-18} & \cdots & 4{,}1 \times 10^{-19} & -3{,}03 \times 10^{-19} \\ -1{,}1 \times 10^{-18} & 0{,}001778 & \cdots & 0{,}000390 & -0{,}000224 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 4{,}1 \times 10^{-19} & 0{,}000390 & \cdots & 0{,}001872 & 0{,}000203 \\ -3{,}03 \times 10^{-19} & -0{,}000224 & \cdots & 0{,}000203 & 0{,}004492 \end{bmatrix}$$

6) Calculating X^TY

The previously calculated matrix X^T is then multiplied by matrix Y to obtain the matrix X^TY. The matrix X^TY is as follows:

$$X^TY = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ -0{,}31929 & 1{,}36819 & \cdots & -1{,}08659 & 0{,}14092 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -0{,}32116 & -0{,}93680 & \cdots & 1{,}68819 & -0{,}72086 \\ 0{,}69023 & 0{,}88881 & \cdots & 2{,}08893 & -0{,}65221 \end{bmatrix} \times \begin{bmatrix} 28{,}8 \\ 26{,}4 \\ \cdots \\ 28{,}6 \\ 29{,}2 \end{bmatrix} = \begin{bmatrix} 25339{,}58 \\ -991{,}095 \\ \cdots \\ 406{,}7893 \\ 158{,}1968 \end{bmatrix}$$

7) Calculating $(X^TX)^{-1}X^TY$

Finally, multiply $(X^TX)^{-1}$ with X^TY to obtain the intercept and coefficients. The resulting intercept and coefficients are:

$$\beta = \begin{bmatrix} 0{,}001108 & -1{,}1 \times 10^{-18} & \cdots & 4{,}1 \times 10^{-19} & -3{,}03 \times 10^{-19} \\ -1{,}1 \times 10^{-18} & 0{,}001778 & \cdots & 0{,}000390 & -0{,}000224 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 4{,}1 \times 10^{-19} & 0{,}000390 & \cdots & 0{,}001872 & 0{,}000203 \\ -3{,}03 \times 10^{-19} & -0{,}000224 & \cdots & 0{,}000203 & 0{,}004492 \end{bmatrix} \times \begin{bmatrix} 25339{,}58 \\ -991{,}095 \\ \cdots \\ 406{,}7893 \\ 158{,}1968 \end{bmatrix}$$

$$= \begin{bmatrix} 28{,}09266638 \\ -1{,}0060313 \\ -0{,}05383042 \\ 0{,}08978047 \\ -0{,}020642 \\ 0{,}09961918 \\ 0{,}42136255 \\ -0{,}10340003 \\ -0{,}16588094 \end{bmatrix}$$

The intercept and coefficients resulting from the implementation of the multiple linear regression algorithm using LinearRegression() from the scikit-learn library can be written as follows:

$\beta_0 = 28{,}09266638$ $\beta_1 = -1{,}0060313$ $\beta_2 = -0{,}05383042$
$\beta_3 = 0{,}089780472$ $\beta_4 = -0{,}020642$ $\beta_5 = 0{,}099619183$
$\beta_6 = 0{,}421362546$ $\beta_7 = -0{,}10340003$ $\beta_8 = -0{,}16588094$

Thus, the best multiple linear regression model for the sixth experiment with the features RH_avg, RR, ss, ddd_x, ff_avg, no2, o3, and pm10 can be written as follows.

$$\hat{Y} = 28{,}09266638 + (-1{,}0060313)(X_1) + (-0{,}05383042)(X_2) + 0{,}08978047(X_3) + (-0{,}020642)(X_4) + 0{,}09961918(X_5) + 0{,}42136255(X_6) + (-0{,}10340003)(X_7) + (-0{,}16588094)(X_8)$$

F. Data Testing

In the data testing phase, the independent variables from the test data are input into the multiple linear regression model obtained in the previous stage to display the predicted average daily temperature. The results of the data testing, which compare the predicted values with the actual values, are shown below.

Table 2. Comparison of Predicted Results with Actual Values

Y	$\hat{Y}$	Y-$\hat{Y}$
25,9	26,0363171	-0,136317101
28,9	28,00769037	0,892309628
27	27,37355857	-0,373558573
...	...	...
29	28,86454307	0,135456926
26,6	27,72465038	-1,124650383
Rata-rata		-0,122202245

In Table 2, a negative value or minus sign in the error indicates that the model predicted a higher temperature compared to the actual value. Conversely, a positive value indicates that the model predicted a lower temperature than the actual value. The average error suggests that the model tends to predict temperatures approximately 0.1222°C higher than the actual temperatures.

Using the evaluation metric MAPE, the percentage of error in the daily average temperature prediction model can be calculated with the following equation:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 = \frac{451.56205}{226} = 1{,}99806\%$$

This result indicates that the model has a relatively low error rate of only 1.99806%. Thus, it can be said that the model performs well in predicting the daily average temperature based on pollutant levels.

Figure 5 below shows a graphical visualization of the regression, comparing predicted results with actual values. The blue dots on the graph represent the observed actual data. Each dot represents a pair of actual values (X-axis) and predicted values (Y-axis), while the dashed black line represents the regression line from the model. Based on the graph, the regression model appears to perform well in predicting new data, as the blue dots tend to cluster around the regression line with minimal deviation.

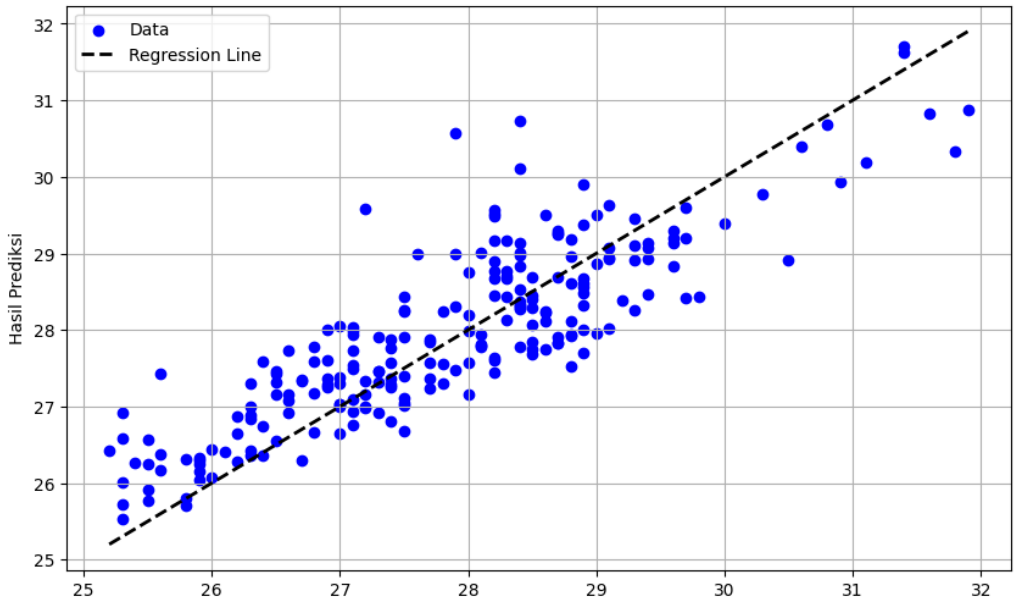

Fig 5. Comparison of Predicted Results with Actual and Real Value

V. Conclusion

The implementation of the multiple linear regression algorithm for predicting air temperature based on pollutant levels begins with data collection, followed by data cleaning and preprocessing, which includes data merging, column dropping, missing value imputation, and outliers handling. The cleaned data is then split into 80% training data and 20% testing data using random subsampling with the train_test_split function from scikit-learn. The data is standardized to ensure equal weighting for each feature. To select the best model, several models were tested using feature selection with SelectKBest ANOVA and the F-value parameter to choose the most relevant k features. The modeling was performed using the LinearRegression() function from scikit-learn, which implements ordinary least squares (OLS). OLS is conducted by calculating the coefficients of the best-fit regression line that minimizes the sum of squared differences between observed and predicted values, using the matrix equation $(X^TX)^{-1}X^TY$. The calculation provided the intercept and coefficients for the multiple linear regression prediction model.

Using the MAPE metric, the model performance achieved an error rate of 1.99806%, indicating a good level of accuracy in predicting daily average air temperature, as the error is relatively low. The model evaluation results from the sixth experiment, utilizing SelectKBest ANOVA F-value with k=8, represent the best evaluation among the eight model schemes, achieving an R^2 value of 0.72749, MAE of 0.5593, and MSE of 0.50078.

Certain pollutants such as nitrogen dioxide (NO_2), ozone (O_3), and particulate matter 10 (PM10) still have an impact on the model and can be used as features in the prediction model. This is evidenced by their selection as predictor variables through SelectKBest ANOVA F-value.

Pollutant levels have a direct but not significant impact on daily average temperature, as indicated by the low correlation values in the correlation matrix between pollutant levels and average temperature, suggesting weak correlation. However, strong correlations among pollutant variables were found.

Daily climate data that most significantly affects model performance and evaluation results includes average humidity (RH_avg), rainfall (RR), sunshine duration (ss), predominant wind direction (ddd_x), average wind speed (ff_avg), nitrogen dioxide (NO_2), ozone (O_3), and particulate matter 10 (PM10).

Future research should explore the strong correlations among pollutant variables, as these might indicate interactions between specific pollutants not directly visible in temperature correlation. Further studies could enhance the understanding of how pollution impacts temperature variability.

Increasing the dataset size, such as using data spanning 10 years or more, could improve model accuracy. Additionally, experimenting with other predictive algorithms and integrating this prediction model with real-time IoT technology for monitoring pollutant levels could be beneficial.

References

[1] H. J. Mukono, *Analisis Kesehatan Lingkungan Akibat Pemanasan Global dan Perubahan Iklim (Tinjauan Kesehatan Masyarakat)*. Surabaya: Airlangga University Press, 2020.

[2] C. Dhea Ulhaq Mardhatillah, F. Permata Jingga, N. Ramadhani, R. Vrika, dan R. Fevria, "Efek Rumah Kaca Pemicu Pemanasan Global dan Upaya Penanggulangannya," *Prosiding Seminar Nasional Biologi*, vol. 2, no. 2, hlm. 328–340, 2022, Diakses: 19 September 2023. [Daring]. Tersedia pada: https://doi.org/10.24036/prosemnasbio/vol2/450

[3] Bmkg, "Gelombang Panas Asia Masih Berlangsung, Namun Tidak Terjadi Di Indonesia: Masyarakat Agar Tidak Panik Dan Tetap Waspada." Diakses: 3 September 2023. [Daring]. Tersedia pada: https://www.bmkg.go.id/press-release/?p=gelombang-panas-asia-masih-berlangsung-namun-tidak-terjadi-di-indonesia-masyarakat-agar-tidak-panik-dan-tetap-waspada&tag=press-release&lang=ID

[4] N. Rahmadania, "Pemanasan Global Penyebab Efek Rumah Kaca dan Penanggulangannya," *Jurnal Ilmu Teknik*, vol. 2, no. 3, 2022.

[5] G. D'amato, C. Akdis, A. H. Speciality, dan " A Cardarelli, "Global Warming, Climate Change, Air Pollution and Allergies," *European Journal of Allergy and Clinical Immunology*, 2020, doi: 10.22541/au.159526736.69654469.

[6] E. Tziperman, *Global Warming Science: A Quantitative Introduction to Climate Change and Its Consequences*. Princeton: Princeton University Press, 2022.

7] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, dan E. Bezirtzoglou, "Environmental and Health Impacts of Air Pollution: A Review," 20 Februari 2020, *Frontiers Media S.A.* doi: 10.3389/fpubh.2020.00014.

[8] E. Bormashenko, "What is temperature? Modern outlook on the concept of temperature," *Entropy*, vol. 22, no. 12, hlm. 1–10, Des 2020, doi: 10.3390/e22121366.

[9] M. Wilson, "Temperature measurement," *Anaesthesia and Intensive Care Medicine*, vol. 22, no. 3, hlm. 202–207, 2021.

[10] G. James, D. Witten, T. Hastie, R. Tibshirani, dan J. Taylor, *An Introduction to Statistical Learning with Applications in Python*, 1 ed. Cham: Springer, 2023. doi: https://doi.org/10.1007/978-3-031-38747-0.

[11] U. A. Ibekwe dan L. A. Ajijola, "Modelling The Naira/U.S. Dollar Currency Exchange Rates Using Decision Tree, Ordinary Least Squares And Random Forest Machine Learning Algorithms," 2022.

[12] P. Schneider dan F. Xhafa, *Anomaly Detection and Complex Event Processing Over IoT Data Streams With Application to eHealth and Patient Data Monitoring*. Barcelona: Mara Conner, 2022. doi: 10.1016/B978-0-12-823818-9.00013-4.

[13] H. H. Nuha, "Mean Squared Error (MSE) dan Penggunaannya," 2023, [Daring]. Tersedia pada: https://ssrn.com/abstract=4420880

[14] I. Nabillah dan I. Ranggadara, "Mean Absolute Percentage Error untuk Evaluasi Hasil Prediksi Komoditas Laut," *JOINS (Journal of Information System)*, vol. 5, no. 2, hlm. 250–255, Nov 2020, doi: 10.33633/joins.v5i2.3900.