Published online on the journal's webpage: **http://journal.ittelkom-pwt.ac.id/index.php/dinda**

Journal of Dinda

Data Science, Information Technology, and Data Analytics

Vol. 4 No. 2 (2024) 105 - 114 E-ISSN: 2809-8064

Determining Air Quality Influential Parameters Using Machine Learning Techniques

Evita Fitri[1*], Andi Saryoko[2]

[1*]Information Systems, Faculty of Information Technology, Universitas Nusa Mandiri

[2]Informatic, Faculty of Information Technology, Universitas Nusa Mandiri

[1*]evita.etv@nusamandiri.ac.id, [2]andi.asy@nusamandiri.ac.id

Abstract

Air quality is an important issue in public health and the environment. This research aims to develop an air quality prediction model based on PM_{10} and $PM_{2.5}$ parameters using various regression and machine learning approaches. The dataset used includes air pollutant standard index (ISPU) data from a number of stations in the Jakarta area with an observation period from January to April 2024. The research method includes collecting datasets, reviewing literature and testing several models of machine learning techniques. Furthermore, the handling of outliers was carried out using the numeric outlier's node and data normalization to prepare the data before dividing the training and testing data. The models evaluated include Linear Regression, Random Forest Regression, Gradient Boosted Trees, and Multilayer Perceptron (MLP), with validation using 10 times cross-validation. The results showed that the Random Forest Regression and Gradient Boosted Trees models provided good prediction performance for both PM_{10} and $PM_{2.5}$ parameters. Random Forest Regression showed the lowest RMSE value on testing data for PM_{10} (0.048) and $PM_{2.5}$ (0.037), while Gradient Boosted Trees showed the lowest RMSE value on training data for $PM_{2.5}$ (0.032). The process of handling outliers and normalizing the data successfully improved the prediction accuracy of the model. Suggestions for future research include the exploration of new models, the addition of meteorological and socio-economic variables, and the application of models in real-time air quality monitoring systems.

Keywords: *Air Quality, Data Preprocessing, Gradient Boosted Trees, Machine Learning, Random Forest Regression.*

1. Introduction

Air quality is a crucial aspect of modern life that directly affects human health and the environment. Air pollution, caused by human activities such as industry, transportation, and fossil fuel combustion, has become an urgent global problem to solve [1]-[2]. Increased levels of pollutants such as Particulate Matter ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), carbon monoxide (CO), not only have an impact on human health [3], but also on ecosystems and the lives of other creatures on earth. Based on air pollutant standard index (ISPU) data obtained from the Jakarta Environmental Agency's Stasiun Pemantau Kualitas Udara (SPKU), in 2019, Jakarta experienced 2 days (1%) with good air quality, 172 days (48%) with moderate air quality, 183 days (50%) with unhealthy air quality, and 8 days (2%) with very unhealthy air quality [4].

In this regard, air quality can affect or cause various respiratory diseases such as asthma, bronchitis, and chronic obstructive pulmonary disease, as well as worsen heart conditions and increase the risk of lung cancer [5]. In addition, air pollution resulting from poor air quality also contributes to climate change through the emission of greenhouse gases such as CO_2 and methane (CH_4). Climate change-induced increases in global temperature can worsen air quality by increasing the frequency and intensity of forest fires, which are a major source of harmful particulate and gaseous emissions [6].

Air pollution also has a negative impact on the natural environment. Increased levels of pollutants can result in acid rain, which damages plants, photosynthetic organs, soil, reduces chlorophyll levels and water bodies [7]. In addition, tropospheric ozone, which is formed from chemical reactions between NO_2 and volatile organic compounds (VOC) under sunlight, can cause damage to crops, reduce yields, and disrupt natural ecosystem functions. Therefore, efforts to reduce air pollution

Received: 20-07-2024 | Accepted: 03-08-2024 | Published: 06-08-2024

through strict environmental policies, the use of clean technologies, and increased public awareness are essential for maintaining air quality and global health [8]-[9].

Based on the analysis of the problems above, there are several studies related to air quality prediction and parameters that affect air quality including research [10] which discusses air pollution in Jakarta has exceeded the limit of air quality standards set by WHO, making it a serious problem that needs to be addressed. Research [10] aims to predict the results of air pollution measurements in Jakarta using data from the Jakarta Opendata website. The method used in [10] is Linear Regression, with parameters such as PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, NO_2, date, and location of measurement stations. The results of this study resulted in the prediction and visualization of air quality, providing a clearer picture of the level of pollution in Jakarta and assisting in decision-making to address this problem.

In addition, in [11], this study examines fluctuations in unhealthy air quality in North Jakarta, which is influenced by air pollution, especially PM_{10}. Using a machine learning approach, the study compared the performance of the Random Forest algorithm (Bagging Method) with Catboost and XGBoost (Boosting Method) in predicting PM_{10} concentrations. Meteorological and pollutant factor data from 2017 to 2019 were used for modeling, with the results showing that Random Forest provided higher accuracy in the testing data (R2 = 0.6424) than XGBoost (R2 = 0.6340) and Catboost (R2 = 0.6294).

Furthermore, in [12], the study identified a significant increase in air pollution, particularly $PM_{2.5}$ particle concentrations, as a serious threat to the health of citizens and countries in smart cities around the world. In Malaysia, traffic congestion is a major contributor to air pollution in smart cities such as Kuala Lumpur and Johor Bahru. Although air pollution prediction using machine learning methods has been widely researched globally, very few studies have been conducted in Malaysia to predict air pollution with this approach. This research aims to implement a machine learning algorithm to evaluate the prediction accuracy of $PM_{2.5}$ concentration in air pollution in Malaysian smart cities. In the research test [12], Multi-Layer Perceptron (MLP) and Random Forest were chosen to be compared, using the Malaysian Air Pollution dataset. The results showed that Random Forest provides the best accuracy in predicting $PM_{2.5}$ Particulate Air Pollution Index in Malaysian smart cities compared to MLP.

In addition to the studies [10]-[12], which applied several algorithms in predicting air quality, the study [13] carried out the prediction of air quality in Beijing, China, especially the concentration of $PM_{2.5}$ pollutant, which is getting worse along with the rapid development of industrialization. The proposed method combines various machine learning models to process meteorological and air quality data from the past six days as a set of inputs, totaling 46 days of data. The approach uses multi-model fusion to provide predictions of $PM_{2.5}$ concentration changes within 24 hours across Beijing. The input data is divided into two main feature groups: historical meteorological data and statistical information, date, and polynomial variation. In addition, one million additional data are processed by time-sliding averaging, in which the 500 most crucial features are selected with the Light Gradient Boosting Machine (LightGBM) model and the 300 most important features are selected with the eXtreme Gradient Boosting (XGBoost) model. The experimental results show that the proposed approach, based on model weight fusion, is better than the single modeling scheme with a loss value of 0.4158 under the SMAPE (Symmetric Mean Absolute Percentage Error) index [13].

Based on some of the research reviews above, to overcome problems related to air quality prediction and what factors most affect air quality, machine learning methods are one of the powerful and efficient approaches in predicting related matters. In addition, the use of ensemble techniques such as Random Forest, Gradient Boosting, and XGBoost has also proven effective in describing the complex interactions between various influence factors [10]-[13]. The integration of data from various sources, including air monitoring stations, IoT-based sensor networks, and satellite imagery, enables comprehensive and real-time analysis of air pollution patterns and identification of contributing environmental factors.

This research applies linear regression, multi-layer perceptron (MLP), Random Forest Regression, and Gradient Boosting models to predict and interpret air pollution patterns with greater precision. Overall, this research confirms that machine learning-based approaches, with various models used, are able to provide deep insights and strong predictive capabilities of the most influential parameters in predicting air quality, especially in the Jakarta area.

This research will use several machine learning models to identify key parameters that affect air quality. The data used includes historical air quality data from the air pollution standard index (ISPU) data in Jakarta province. The machine learning models will be trained and tested using cross-validation techniques to evaluate their ability to identify and predict significant air quality patterns.

The main objective of this research is to develop a predictive model that can correctly identify key parameters that affect air quality. By better

understanding these factors, it is expected to provide deeper insights to decision-makers regarding air pollution management and mitigation policies. In addition, this research also aims to compare the performance of various machine learning methods in the context of air quality prediction.

2. Research Methods

This study aims to analyze air quality using various critical pollutant parameters and comprehensive data analysis methods. In this section, explain the materials used and methods that have been carried out in this research, as well as several stages carried out including the stages of data pre-processing, preprocessing, modeling, and model evaluation. The following are details of the materials and methods used including:

2.1 *Dataset Source*

In this study, the dataset used is the air pollution standard index (ISPU) data in the Jakarta province obtained from open-source data on the website https://satudata.jakarta.go.id/ with a time frame of January 2024 to April 2024, the number of instances in this dataset is 605 instances. The 4-month data collection period with 605 records was chosen because it covers a variety of conditions, providing a comprehensive overview of patterns and trends. Despite the short duration, the collected data has undergone rigorous validation. Regarding the limitations of the dataset, a 10-fold cross-validation technique was applied to ensure the model generalizes well. Therefore, this data is sufficient for accurate and reliable detection, and the details of the dataset used can be seen in Table 1.

Table 1. Research Dataset Table

Dataset	Atribute	Number of Instances
ISPU data for Jakarta	Period, date, PM_{10}, $PM_{2.5}$, sulfur_dioxide(SO_2), carbon_monoxide(CO), ozone(O_3), nitrogen_dioxide(NO_2), max, critical_pollutant_parameter, category	605 instance

Table 1 is a detail of the dataset used in this study, there are 11 attributes and 605 data. The attributes used include:

a. Data Period: The dataset covers a specific time period relevant for this study, covering the last few months to get a comprehensive picture of the air quality.
b. Date: Data is collected based on specific dates, with daily information on the measured air quality parameters.
c. Stations: Data were obtained from various air quality monitoring stations scattered in various locations to cover a wide and representative area of air conditions in the region including DKI1 Bundaran Hotel Indonesia (HI) station, DKI2 Kelapa Gading, DKI3 Jagakarsa, DKI4 Lubang Buaya, and DKI5 Kebon Jeruk West Jakarta.
d. Parameters: The dataset includes various critical pollutant parameters such as $PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3, and nitrogen dioxide (NO_2). In addition, the data also includes the daily maximum values of each parameter and air pollution category.

2.2 *Research Stage*

The method in this research uses the implementation of several machine learning models including Linear Regression, Random Forest Regression, Multilayer Perceptron and Gradient Boosting models. Before testing these models, there are several stages used, as for the research stages can be seen in Figure 1.

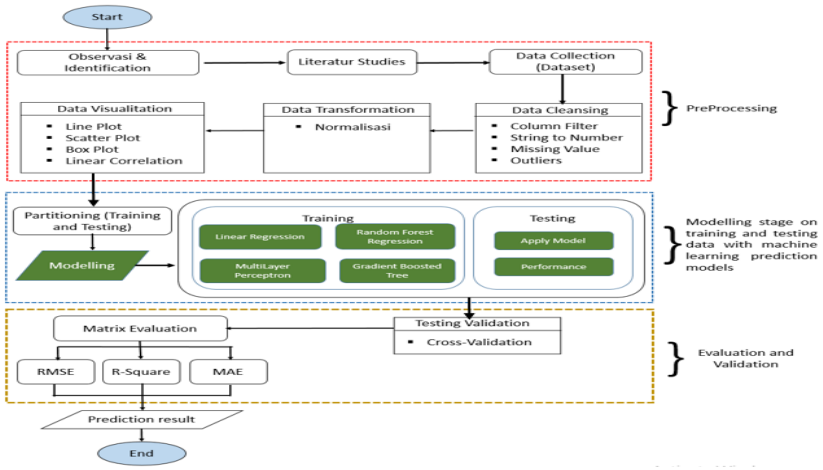

Figure 1. Research Concept Map

Figure 1 shows several stages carried out in this study, including data pre-processing, data preprocessing, modeling, and evaluation.

2.2.1 *Prepocessing Data*

As for some of the stages carried out at this stage including:

a. Observation and Identification: At this initial stage, observations and identification of important parameters related to air quality are made. This identification includes an initial understanding of the various factors that affect air quality and the variables to be analyzed.
b. Literature Study: A comprehensive literature study was conducted to understand the methods that have been used in previous studies related to air quality. This study helped in determining the most suitable approach for this study.
c. Data Collection: Data was collected from available air quality datasets, which include parameters such as $PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3, and nitrogen dioxide (NO_2), as well as daily maximum values and pollution categories.
d. Data Cleansing
   Column Filter: Selection of relevant columns from the dataset based on the parameters under study.
   String to Number: Convert data from string format to numeric format for easy analysis.
   Missing Value: Handling missing values with techniques such as imputation or removal of incomplete data.
e. Data Transformation: Data normalization is performed to ensure consistency and facilitate further analysis.
f. Data Visualization: The processed data is visualized using various types of plots such as line plots, box plots, scatter plots, and linear correlation analysis to understand patterns and trends in the data.

2.2.2 *Modelling Data*

There are tests using 4 prediction algorithms or models in this study, but before testing the model, there is a process of partitioning data or dividing data into training data and testing data with the amount of data division each as much as 80%-20%. This data division process is carried out for the purposes of training and testing models. Modeling Techniques used include:

a. Linear Regression: A statistical method that serves to test the extent of the causal relationship between independent variables and dependent variables [10].
b. Random Forest Regression: It is a popular supervised machine learning method for regression and classification issues. It builds decision trees from a range of data, using the classification vote and the average for regression [14].
c. Multilayer Perceptron (ANN): It is a model that attempts to replicate how the human brain works. Based on a mathematical function to collect data and classify it using a pre-planned architecture, the basic unit of analysis is a neuron. Neural networks are composed of layers of interconnected nodes. Every neural connection has a weight value, which is multiplied by the input value. Furthermore, the activation function of every neuron characterizes its output and is used to introduce some non-linearity into the network model [15].
d. Gradient Boosting: When it learns, it builds a new regression tree by fitting the residuals to lower the loss function until either the number of regression trees exceeds a threshold or the residual is smaller than a threshold [13].

The next process is carried out by conducting validation testing, this process is carried out using cross-validation techniques to ensure that the model can be generalized well to data that has never been seen before.

2.2.3 *Evaluation*

The evaluation in this study was conducted to assess the performance and accuracy of the prediction model that has been built. The purpose of the evaluation is to ensure that the model can predict air quality well and can be used for reliable analysis. The evaluation process involves several different metrics to provide a comprehensive picture of the model's performance, as for the evaluation matrix used in this study refers to several studies including the MAE and R-Square evaluation matrix in research [16] and the RMSE evaluation matrix which is also used in research [17].

a. Root Mean Square Error (RMSE): Measurement of model prediction error by calculating the root of the mean square error [18].
b. R-Square: Determines how well the model explains the variability of the data [18].
c. Mean Absolute Error (MAE): A measurement of the average absolute error of model predictions to give an idea of the model's accuracy [18].

The methods used in this study were designed to ensure that the results obtained are accurate and reliable in providing insight into the factors affecting air quality.

3. Results and Discussion

In this study, a dateset in the form of a history containing Air Pollutant Standard Index (ISPU) data is used. The Air Pollutant Standard Index (ISPU) is a unitless number used to describe ambient air quality conditions at a particular location and is based on the impact on human health, aesthetic value, and other living things. These

data are obtained from measurements through Air Quality Monitoring Stations (SPKU) in DKI Jakarta Province. The following dataset component details can be seen in table 2.

Table 2. Dataset Component Details

Variable	Description
period_data	Data Period Explanation Once a Month
month	is the month of data collection of the Air Pollution Standard Index (ISPU)
date	is the data collection date of the Air Pollution Standard Index (ISPU)
station	is the location where air monitoring equipment is placed
pm_ten (PM_{10})	is the value of the measurement results of the Air Pollutant Standard Index (ISPU) for the PM_{10} parameter (the name of one of the monitored parameter names), namely particulates with a size below 10 microns.
pm_duakomalim ($PM_{2,5}$)	is the measurement value of the Air Pollutant Standard Index (ISPU) for the parameter PM2.5 (the name of one of the monitored parameter names), which is particulate matter with a size below 2.5 microns.
sulfur_dioxide	is the measurement result value of the Air Pollution Standard Index (ISPU) for the parameter sulfur dioxide / SO2 (name one of the monitored parameter names).
carbon_monoxide	is the measurement result value of the Air Pollution Standard Index (ISPU) for the parameter carbon monoxide / CO (name of one of the monitored parameter names).
ozone	is the measurement result value of the Air Pollution Standard Index (ISPU) for the parameter ozone / O3 (name of one of the monitored parameter names)
nitrogen_dioxide	is the measurement result value of the Air Pollution Standard Index (ISPU) for the parameter nitrogen dioxide / NO2 (name one of the monitored parameter names).
max	is the highest value of the Air Pollution Standard Index (ISPU) measurement results of several parameters monitored at a particular Air Quality Monitoring Station (SPKU) and measurement date.
parameter_polluter_critical	is the name of the monitored parameter with the highest Air Pollution Standard Index (ISPU) value at a particular Air Quality Monitoring Station (SPKU) and date.
category	is the category of the measurement results of the Air Pollution Standard Index (ISPU) at the Air Quality Monitoring Station and a certain date.

Table 2 is a detail of the dataset components or variables used in this study, as well as samples of some of the datasets above can be seen in Table 3.

Table 3. Jakarta ISPU Sample Dataset

	date	**station**	**PM_{10}**	**$PM_{2,5}$**	**sulfur_diokside**	**karbon_monokside**	**ozone**	**nitrogen_diokside**	**max**	**parameter_polluter_critical**	**category**
1		DKI1 Bundaran Hotel Indonesia (HI)	64	89	52	9	13	30	89	PM25	sedang
2		DKI1 Bundaran Hotel Indonesia (HI)	53	93	52	4	12	28	93	PM25	sedang
3		DKI1 Bundaran Hotel Indonesia (HI)	52	57	52	6	12	27	57	PM25	sedang
4		.....	...	..	..	...	...	...	...	...	.....
31		DKI1 Bundaran Hotel Indonesia (HI)	25	49	47	19	24	30	49	PM25	baik

In table 3, is a sample dataset used in this study, where the data has not been pre-processed.

In the initial stage of this research, variable selection was carried out first. However, not all variables are used in this study, in the process of selecting variables using column filters, there are seven variables that are finally used to continue the modeling trial. The variables are PM10, PM2.5, sulfur dioxide, carbon monoxide, ozone, nitrogen dioxide and maximum value. Next process, visualization of the air quality dataset was conducted prior to pre-processing. This visualization aims to understand the initial distribution of the data and identify certain anomalies or patterns that may exist. The initial data visualization is done using box plot nodes, where these nodes are used to describe the distribution of data based on five summary numbers (minimum, first quartile, median, third quartile, and maximum) with several functions including identifying outliers or outlier values that might affect the analysis, providing a visual description of the variability and symmetry of the data, facilitating comparison of distributions between different parameters.

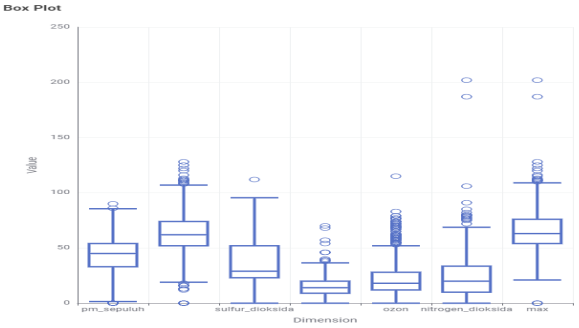

Figure 2. Visualization of Data Pra-PreProcessing Dataset

From the visualization of the dataset in Figure 2 which includes air quality parameters such as $PM_{2.5}$, PM_{10}, NO_2, SO_2, and O_3, it can be seen that there are some columns with incomplete data and some values that do not match. To solve this problem, data preprocessing was performed.

The preprocessing stage begins with the use of column filters to select only relevant columns, so that unnecessary data can be ignored, as for some of the variables used include PM_{10}, $PM_{2.5}$, sulfur_dioxide, carbon_monoxide, ozone, nitrogen_dioxide and max variables. Furthermore, data in string format is converted to numeric using the string to number node, this can support further statistical analysis and calculations, the next pre-processing process is to handle missing values in the data set using the missing value process, either by imputation of missing values or deletion of incomplete data. The missing value technique used is fix value with a value of 1. The results of this preprocessing are then re-visualized to ensure that the data is in an optimal condition for further analysis.

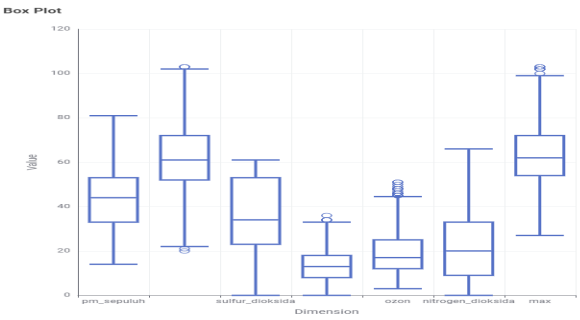

Figure 3. Visualization of Post-Processing Dataset

After preprocessing the data, it can be seen in Figure 3 which is the result of data visualization after the missing value process. However, outliers are still found in the dataset. Therefore, an outlier handling process was carried out to ensure that the data used was free from outlier values that could affect the analysis. The outliers handling process involves identifying and removing or replacing extreme values until the results are free of outliers.

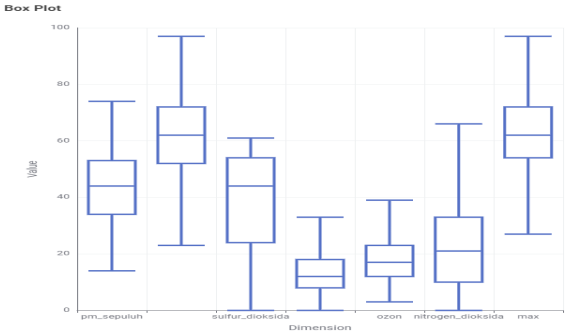

Figure 4. Dataset Visualization Results After Handling Outliers

Figure 4 is the result of visualizing the dataset that has been handled by outliers. The process of handling outliers is done using the numeric outliers node using the interquartile range (IQR) multiplier method k=1.5, Quartile calculation is done using full data estimate using R_4, and the treatment option used is remove outliers rows. This step ensures that data containing outliers is removed from the dataset, so that only clean and relevant data is used for further analysis.

Table 4. Outliers Handling Process Result on Dataset

RowID	Outlier column	Member count	Outlier count	Lower bound	Upper bound
Row0	PM_{10}	466	0	03.00	83.00.00
Row1	$PM_{2,5}$	466	0	22.00	102.00.00
Row2	sulfur_dioksid e	466	0	-21.0	99.00.00
Row3	karbon_monok side	466	0	-7.0	33.00.00
Row4	Ozone	466	0	-4.5	39.05.00
Row5	nitrogen_dioks ide	466	0	-24.5	67.05.00
Row6	max	466	0	27.00.00	99.00.00

It can be seen in table 4 that the number of outliers in each variable used has reached 0, which means that all records in each variable no longer contain outliers. The normalization process is then carried out on the results of the data. Normalization aims to change the scale of the data so that all variables have the same scale, and is important to improve model performance in the next analysis stage.

In the normalization process of this study, researchers used the decimal scaling technique, this technique is one method to change the scale of data so that it is in a certain range. This technique involves moving the decimal point of the attribute values, using decimal scaling, the normalized data will be in the range -1 to 1, depending on the absolute maximum value of the attribute.

Table 5. Dataset Normalization Process Result

Row ID	PM_{10}	$PM_{2,5}$	sulfur_dioksida	karbon_monoksida	ozone	nitrogen_dioksida	max
Row0	0,044	0,062	0,036	0,006	0,009	0,002	0,062
Row1	0,037	0,065	0,036	0,003	0,008	0,019	0,065
Row2	0,036	0,570	0,036	0,004	0,008	0,019	0,570
Row4	0,015	0,038	0,032	0,010	0,007	0,031	0,038
....	.....	....	....	.....	.....	.....	.....
Row466	0,024	0,045	0,009	0,015	0,030	0,024	0,045

Table 5 is the result of the normalization process that has been carried out, after normalization is complete, the data is re-visualized to ensure that the normalization process has run well and the data distribution is optimal. Visualization of data after normalization is done using scatter plot and linear correlation process.

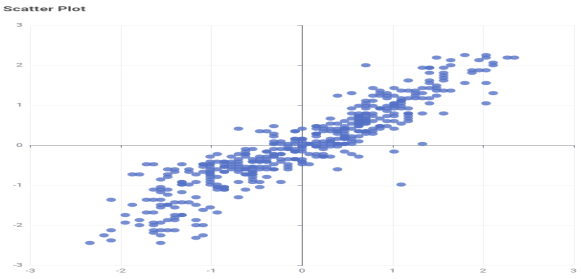

Figure 5. Scatter Plot of Dataset After Normalization

Based on the scatter plot in Figure 5 that has been made, it can be observed that there is a positive relationship between the variables, this is because it can be seen that the distribution of dots tends to rise from the lower left to the upper right, which means that the correlation coefficient value is close to +1.

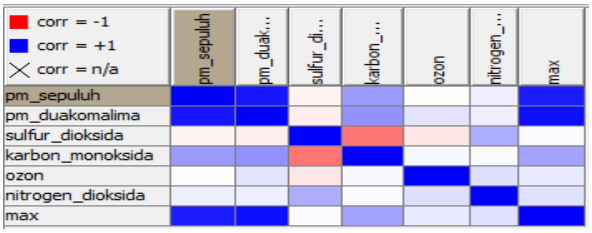

Figure 6. Linear Correlation Results

In Figure 6, the visualization of the resulting linear correlation results uses color to illustrate the direction and strength of the correlation between the analyzed variables. In this case, the linear correlation results show that the relationship between the variables is positive but 2 variables with negative results were found. It can be seen that there are significant relationships between these variables. Next, the dataset was divided into two relevant parts in order to proceed to the evaluation model testing process. The dataset division is 80% as training data and 20% as testing data.

3.1 *Prediction Model Testing and Evaluation*

The next step after the division of training and testing data is model testing. The model used in this research includes several regression and machine learning methods, namely Linear Regression, Random Forest Regression, Gradient Boosted Trees, Multilayer Perceptron.

The cross-validation process is also performed on the training data to ensure that the developed model can be generalized well. Cross-validation using a number of validations = 10 value, helps in assessing the stability and performance of the model on different data by dividing the training data into ten subsets and testing the model on each subset in turn. The results of the modeling that has been done with cross-validation are then tested on the testing data. This test aims to assess the performance of the model on data that has never been seen before, thus providing a more accurate picture of the model's ability to predict the true value.

In testing the prediction model, there are two variables that become the target of the test parameters. The first test uses the PM_{10} target variable, and the second test is conducted using the $PM_{2.5}$ variable. By testing these two target variables, this research can evaluate the performance of the model in predicting different air quality parameters, ensuring that the developed model has good reliability in various conditions and target variables.

3.1.1 *Target Parameters PM_{10}*

Testing the four prediction models with the PM_{10} parameter target can be seen in Table 6.

Table 6. Test Results of LR, RFR, Gradient Boosted Trees and MLP Prediction Models with PM_{10} target parameters.

Matrix Evaluation	LR		RFR		Gradient Boosted Trees		ANN with MLP	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
R^2	0,862	0,814	0,867	0,838	0,86	0,804	0,873	0,807
Mean Absolute Error	0,285	0,313	0,035	0,038	0,037	0,042	0,056	0,067
Mean Squared Error	0,142	0,158	0,002	0,002	0,002	0,003	0,006	0,007
root *Mean Squared Error*	0,377	0,398	0,048	0,048	0,049	0,052	0,077	0,086
mean signed difference	0,001	-0,033	0,001	-0,003	0,001	0	-0,001	-0,004
Mean Absolute Percentage Error	4,644	12,957	0,095	0,097	0,096	0,105	0	0,194

adjusted R^2	0,862	0,814	0,867	0,838	0,86	0,804	0,873	0,807

In table 6, the results of model testing show the performance of each model with the RMSE, R-Square, and MAE metrics, as for the results of table 6 Linear **Regression**: Shows good performance with high R-Square values on training data (0.862) and testing data (0.814). However, the slightly higher RMSE and MAE values in the testing data indicate a slight decrease in accuracy on data that has never been seen before. **Random Forest Regression**: Provided excellent results with very low RMSE values on both training (0.048) and testing data (0.048). The R-Square value was also high in both datasets, indicating this model has strong and stable prediction capabilities. **Gradient Boosted Trees**: Showed good performance with low RMSE and MAE values, as well as high R-Square values on both training (0.860) and testing data (0.804). The model was also able to capture non-linear relationships in the data well. **Multilayer Perceptron (MLP)**: It performed very well with high R-Square values in both training (0.873) and testing data (0.807). Although the RMSE and MAE values were slightly higher in the testing data, the model still showed good predictive ability.

Overall, the Random Forest Regression and Multilayer Perceptron (MLP) models performed best in air quality prediction, with excellent RMSE, R-Square, and MAE values on both datasets. This indicates that both models are highly reliable and can be used for the prediction of PM_{10} air quality parameters.

3.1.2 Target Parameters $PM_{2,5}$

Testing four prediction models with PM2.5 target parameters can be seen in Table 7.

Table 7. Test Results of LR, RFR, Gradient Boosted Trees and MLP Prediction Models with $PM_{2,5}$ target parameters.

Matrix Evaluation	LR		RFR		Gradient Boosted Trees		ANN with MLP	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
R^2	0,937	0,921	0,936	0,93	0,96	0,923	0,939	0,883
Mean Absolute Error	0,172	0,169	0,025	0,024	0,015	0,017	0,039	0,05
Mean Squared Error	0.066	0,063	0,002	0,001	0,001	0,002	0,003	0,004
root *Mean Squared Error*	0,257	0,252	0,041	0,037	0,032	0,039	0,054	0,065
mean signed difference	0	0,014	0,002	0	0,002	0,005	-0,001	0,003
Mean Absolute Percentage Error	0,508	0,5	0,056	0,052	0,037	0,044	0	0,215
adjusted R^2	0,937	0,921	0,936	0,93	0,96	0,923	0,939	0,883

In table 7 the **Linear Regressio**n model: Shows good performance with high R-Square values on training data (0.937) and testing data (0.921). Consistent RMSE and MAE values on both datasets indicate this model has reliable prediction capabilities. Then on the **Random Forest Regression prediction** model: Provides excellent results with very low RMSE values on training data (0.041) and testing data (0.037). The R-Square value is also high in both datasets, indicating this model has strong and stable prediction capabilities. Furthermore, **Gradient Boosted Trees**: Showed the best performance with very low RMSE and MAE values, as well as high R-Square values on the training data (0.960) and testing data (0.923). This model is able to capture non-linear relationships in the data very well. And finally, the **Multilayer Perceptron (MLP)** prediction model: Has a good performance with high R-Square values in training data (0.939) and testing data (0.883). Although the RMSE and MAE values are slightly higher in the testing data, this model still shows good predictive ability.

Overall, the Gradient Boosted Trees model showed the best performance in $PM_{2.5}$ parameter prediction, with excellent RMSE, R-Square, and MAE values on both datasets. This model was followed by Random Forest Regression and Multilayer Perceptron (MLP), which also showed good and reliable performance for air quality prediction.

4. Conclusion

From the research conducted, several things can be concluded including for the PM_{10} parameter, the Random Forest Regression model has the smallest RMSE value both in training data (0.048) and testing data (0.048), so it can be said to be the best prediction model for PM_{10}, while for the PM2.5 parameter, the Random Forest Regression model also has the smallest RMSE value in testing data (0.037), although the smallest value in training data belongs to Gradient Boosted Trees (0.032). However, the RMSE value in testing data is more important because it reflects the performance of the model on data that has never been known before. Followed by the Multilayer Perceptron (MLP) and Linear Regression models which also have quite good performance, although slightly higher than the Random Forest Regression and Gradient Boosted Trees prediction models.

When viewed from the predicted parameters, the results of the study show that the $PM_{2.5}$ parameter is a parameter that is quite strong in influencing air quality compared to the PM_{10} parameter. The application of cross-validation with 10 times validation on the training data

helps ensure that the developed model can be generalized well and avoids overfitting. The results of model testing on the testing data show that the developed model has accurate and reliable prediction capabilities for data that has never been seen before. Likewise, the process of handling outliers using node numeric outliers and data normalization has succeeded in improving the quality of the dataset, so that the model can work better and produce more accurate predictions.

Furthermore, this research obtained several new findings that can be used as scientific contributions. By using various regression and machine learning methods, this research shows that Random Forest Regression and Gradient Boosted Trees models can provide highly accurate air quality predictions. A systematic approach in handling outliers, normalizing data, and using cross-validation helps in improving prediction accuracy and ensuring a robust and reliable model.

As for future research, some suggestions that can be considered are the exploration of other machine learning and deep learning models to see if there are models that can provide even better performance in air quality prediction. Further research can add other independent variables that may affect air quality, such as meteorological variables (temperature, humidity, wind speed), as well as socio-economic and demographic variables, to improve model accuracy. Using real-time air quality data and seeing how the model can handle predictions under dynamic and changing conditions is also recommended. Finally, applying the developed model in an air quality monitoring system that can be used by the government and related agencies to proactively predict air quality and take necessary actions to reduce the negative impacts of air pollution.

References

[1] S. Syihabuddin Azmil Umri, "Analisis Dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara Di Dki Jakarta," *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 2, pp. 98–104, 2021.

[2] A. Khumaidi, R. Raafi'udin, and I. P. Solihin, "Pengujian Algoritma Long Short-Term Memory untuk Prediksi Kualitas Udara dan Suhu Kota Bandung," *J. Telemat.*, vol. 15, no. 1, pp. 13–18, 2020.

[3] D. Septiyana, A. Sukmono, and M. A. Yusuf, "Pemantauan Kualitas Udara Ispu (Pm10, So2, No2) Menggunakan Citra Landsat 8 Dan 9 Untuk Kecamatan Xmijen Selama Pandemi Covid-19," *J. Geod. Undip*, vol. 12, no. 2, pp. 271–280, 2023.

[4] A. Oktaviani and Hustinawati, "Prediksi Rata-Rata Zat Berbahaya Di Dki Jakarta Berdasarkan Indeks Standar Pencemar Udara Menggunakan Metode Long Short-Term Memory," *J. Ilm. Inform. Komput.*, vol. 26, no. 1, pp. 41–55, 2021.

[5] S. Maharani and W. R. Aryanta, "Dampak Buruk Polusi Udara Bagi Kesehatan Dan Cara Meminimalkan Risikonya," *J. Ecocentrism*, vol. 3, no. 2, pp. 47–58, 2023.

[6] I. L. Firmansyah, A. Indah, I. Wati, I. P. Sari, A. M. Syifa, and D. O. Radianto, "Dampak Perubahan Iklim Dapat Meningkatnya Kebakaran Hutan Dan Upaya Pelestarian Lingkungan Politeknik Perkapalan Negeri Surabaya," vol. 2, no. 2, 2024.

[7] A. Dreamyseila, Y. Ningsih, T. Nurita, and A. Salsabila, "Dampak Hujan Asam: Solusi Berkelanjutan Untuk Memperbaiki Ekosistem Atmosfer Dalam Mencapai SDGS," vol. 3, no. 4, pp. 100–111, 2024.

[8] N. S. Ma'rifah, "Upaya Masyarakat dalam Penanggulangan Polusi Udara Akibat Asap Pabrik Geo Dipa Dieng Banjarnegara," *Al-DYAS*, vol. 2, no. 3, pp. 612–622, 2023.

[9] N. P. Handayani *et al.*, "Upaya Pengurangan Polusi Udara di Lingkungan Universitas Negeri Semarang dengan Penanaman Pohon.," *J. Majemuk*, vol. 3.2, no. 2, pp. 256–268, 2024.

[10] M. A. Latief and Y. Karyanti, "Data Mining & Analytic Forecasting Indeks Standar Pencemar Udara Jakarta Menggunakan Metode Linear Regression (Studi Kasus: Dataset Indeks Standar Pencemar Udara Jakarta 2021)," *J. Soc. Res.*, vol. 1, no. 10, pp. 1164–1176, 2022.

[11] E. Elita Rizkiani and D. Brahma Arianto, "Perbandingan Performa Algoritma Metode Bagging dan Boosting pada Prediksi Konsentrasi PM 10 di Jakarta Utara," vol. 01, pp. 74–81, 2024.

[12] M. Rishanti and P. Naveen, "Smart City Air Quality Prediction using Machine Learning," *IEEE 2021 5th Int. Conf. Intell. Comput. Control Syst. Madurai, India*, pp. 1048–1054, 2021.

[13] Y. Zhang *et al.*, "A feature selection and multi-model fusion-based approach of predicting air quality," *ISA Trans.*, vol. 100, no. xxxx, pp. 210–220, 2020.

[14] N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis," *J. Environ. Public Health*, vol. 2023, pp. 1–26, 2023.

[15] S. Simu *et al.*, "Air pollution prediction using machine learning," *2020 IEEE Bombay Sect. Signat. Conf. IBSSC 2020*, pp. 231–236, 2020.

[16] D. Ardiansyah, "Perbandingan Model Prediksi Radiasi Matahari Berbasis Mesin Pembelajaran Pada Stasiun Meteorologi Fatmawati Soekarno Bengkulu," *Megasains*, vol. 14, no. 1, pp. 26–32, 2023.

[17] H. Akbar, M. Ayomi, and Y. Haryanto, "Analisis Perbandingan Model Machine Learningdalam Prediksi Suhu Permukaan Laut Menggunakan Data Model Reanalysis Ecmwf," *Pros. Semin. Nas. Has. Penelit. Kelaut. dan Perikan.*, vol. 5587, pp. 96–100, 2023.

[18] I. Amansyah, J. Indra, E. Nurlaelasari, and A. R. Juwita, "Prediksi Penjualan Kendaraan Menggunakan Regresi Linear : Studi Kasus pada Industri Otomotif di Indonesia," vol. 4, pp. 1199–1216, 2024.