The Prediction of Air Pollution in Jakarta Using Machine Learning Approach

Yonatan Sie1,a), Benny Chandra Setiawan1,b), Yohanes Deo Bilkhan2,a), Afdhal Kurniawan1,d), and Karli Eka Setiawan1,e)

1 Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

2 International Relations Department, Faculty of Humanities, Bina Nusantara University, Jakarta, Indonesia

a) Corresponding author: <yonatan.sie@binus.ac.id>

b) [benny.setiawan@binus.ac.id](benny.setiawan@binus.ac.id%20)

c) <yohanes.bilkhan@binus.ac.id>

d) <afdhal.kurniawan@binus.ac.id>

e) <karli.setiawan@binus.ac.id>

**Abstract.**  This study aims to predict the future concentrations of key air pollutants—namely PM10, PM2.5, SO2, CO, O3, and NO2—in Jakarta using various machine learning techniques. Linear regression, polynomial regression, and decision tree regression were applied to analyze historical pollution data and forecast future trends. The models' performance was evaluated using Mean Squared Error (MSE) and R-squared values to assess predictive accuracy. Results demonstrated that linear regression outperformed both polynomial regression and decision tree regression, with the lowest MSE and highest R-squared scores. This indicates that linear regression is the most effective approach for predicting air pollution levels in Jakarta, among the tested methods. The findings offer valuable insights for urban planners and policymakers, underscoring the urgent need to address air quality issues. By providing data-driven strategies, this research seeks to raise awareness and inform efforts to mitigate air pollution, ultimately improving public health and safeguarding Jakarta's global standing.

**Keywords:** air pollution, linear regression, polynomial regression, decision tree regression, public health

# INTRODUCTION

Jakarta, the capital of Indonesia, is recognized as one of the cities with the highest levels of air pollution globally. According to data from IQAir, Jakarta ranks 6th out of 119 countries as one of the cities contributing the most pollution worldwide, following Delhi, India [1]. Additionally, recent data indicates that the air quality index in Jakarta has consistently been categorized as unhealthy and far from safe, with pollutant concentrations ranging from 0 to 9.0 µg/m³ [2]. The significant level of air pollution in Jakarta is attributed to several factors, including high vehicle emissions, wind patterns, industrial activities, and the dry season [3]. These factors contribute to severe pollution levels, which lead to various health issues such as asthma, acute respiratory infections (ARI), bronchitis, dermatitis, lung cancer, and more [4]. Beyond the impact on public health, this also tarnishes Indonesia’s global reputation, considering Jakarta ranks second in terms of tourist visits, just after Bali [5].

This research focuses on analyzing pollutants in Jakarta over the past few years, particularly PM10, PM2.5, SO2, CO, O3, and NO2. The primary goal is to predict the concentrations of these pollutants in the coming years. PM10 refers to particulate matter with a diameter of 10 micrometers or less, while PM2.5 consists of finer particles with diameters of 2.5 micrometers or less. SO2 (sulfur dioxide) is a gas generated by volcanic eruptions and industrial processes, especially those involving the burning of fossil fuels like coal and oil. CO (carbon monoxide) is a colorless, odorless gas formed through the incomplete combustion of carbon-based fuels. O3 (ozone) is a molecule composed of three oxygen atoms, present both at ground level and in the earth's upper atmosphere. NO2 (nitrogen dioxide) is a reddish-brown gas with a sharp, biting odor, and is another major pollutant. These particles are small enough to be inhaled deeply into the lungs, posing serious health risks [6]. Long-term exposure to such pollutants has been associated with increased mortality rates and negative effects on lung development, especially in children [7].

In this study, we employ machine learning techniques, including linear regression, polynomial regression, and decision tree regression, to predict the future concentrations of PM10, PM2.5, SO2, CO, O3, and NO2 in Jakarta. This research aims to raise awareness among the government and all stakeholders regarding the high pollution levels in Jakarta and to provide actionable insights that can assist urban planners and policymakers in developing strategies to mitigate air pollution and improve public health.

# LITERATURE REVIEW

The prediction of air pollution in Jakarta using machine learning approaches has emerged as an important area of research in recent years. Several studies have been conducted, providing valuable insights into the relationships between traffic density, industrial emissions, and air quality in urban environments. These studies also explore how machine learning models can be applied to predict future pollution levels, offering actionable information for environmental management and policy formulation.

Conducted a study that reviewed the global application of land use regression (LUR) models, focusing on methodologies, air sampling techniques, and model transferability. Their work highlighted the success of the ESCAPE project, widely used in Europe. The study found that while alternative air sampling techniques can improve pollution estimates, LUR models generally do not transfer well across regions unless they share similar geographic and pollution characteristics. The review emphasized the growing integration of LUR models with machine learning techniques to improve prediction accuracy [8].

Examined ways to improve traffic-related air pollution (TRAP) forecasting by incorporating unconventional data such as terrain, background pollutant concentrations, and emission factors into traditional models. Their multi-target regression models, particularly the Fastai framework, showed better performance in predicting concentrations of NO2, PM2.5, and PM10. The study identified traffic count, speed, weather conditions, and emission factors as critical features affecting prediction accuracy. Their findings suggest further investigation into localized pollution scenarios to enhance model performance for specific pollutants [9].

Evaluated machine learning models' effectiveness in predicting PM2.5 concentrations using data from the Taiwan Air Quality Monitoring Network (2012-2017). Their study found that machine learning models, particularly gradient boosting regressors, outperformed traditional models, as indicated by statistical measures such as MAE, MSE, RMSE, and R². The close alignment of predicted and actual values suggests that machine learning models offer significant potential for enhancing air quality forecasts and addressing urban pollution challenges [10].

Explored the role of machine learning models in managing PM2.5 levels, particularly in densely populated and developing countries. Their study employed logistic regression for pollution detection and autoregression for forecasting future PM2.5 levels. The study demonstrated the efficacy of these models in air quality monitoring, providing timely data to meteorological authorities and supporting proactive interventions. This research also underscored the importance of generating localized data sources, often overlooked compared to larger urban centers [11].

Applied time-series modeling to forecast air pollutants (CO, SO2, O3, NO2) in Tehran. Their study addressed the limitations of traditional linear models by using an Adaptive Neuro-Fuzzy Inference System (ANFIS), which outperformed semi-experimental nonlinear regression models. The ANFIS model showed higher accuracy with coefficients of determination of 0.8686, 0.8011, 0.8350, and 0.7640 for the respective pollutants. This advanced modeling technique supports the formulation of more effective environmental policies to protect public health and ecosystems [12].

# METHODOLOGY

## DATASET

The dataset used in this study is sourced from katalog.data.go.id, titled "Data Indeks Standar Pencemaran Udara (ISPU) Data in DKI Jakarta Province 2023." The dataset consists of 2000 samples with 13 features, including ID, periode\_data, tanggal, stasiun, pm\_sepuluh, pm\_duakomalima, sulfur\_dioksida, karbon\_monoksida, ozon, nitrogen\_dioksida, max, parameter\_pencemar\_kritis, dan kategori.

# Data Preprocessing

Data preprocessing was conducted to ensure the dataset is clean and ready for analysis. First, missing values in numerical features such as ‘pm\_sepuluh`, `pm\_duakomalima`, `sulfur\_dioksida`, `karbon\_monoksida`, `ozon`, and `nitrogen\_dioksida` were filled with the median values, while categorical features like `parameter\_pencemar\_kritis` and `kategori` were filled with the mode. Numerical data were normalized using Min-Max Scaling, and categorical features were converted into numerical representations using one-hot encoding.

# The Linear Regression

By fitting a linear equation to the observed data, linear regression is a crucial statistical technique used to examine the connection between a dependent variable and one or more independent variables [8]. The following equation represents the linear regression model:

|  |  |
| --- | --- |
|  | (1) |

where:

* is the dependent variable,
* ​,​…, are the independent variables,
* denotes the intercept,
* , ,..., are the coefficients indicating the relationship between the dependent variable and each independent variable,
* ϵ represents the error term.

This method assumes a linear correlation between the dependent and independent variables, which simplifies interpretation but may not adequately address complex patterns in the data.

# The Polynomial Regression

By representing the relationship between the dependent and independent variables as a n-th degree polynomial, polynomial regression expands on linear regression [9]. The equation for polynomial regression is:

|  |  |
| --- | --- |
|  | (2) |

This approach introduces polynomial terms of the independent variables, enabling the model to fit more flexible curves compared to linear regression. It can capture non-linear relationships through higher-order terms, though excessive polynomial degrees may result in overfitting.

# The DECISION TREE Regression

Decision Tree Regression is a non-parametric technique that models the relationship between the dependent and independent variables by partitioning the data into subsets based on feature values. The model is structured as a tree where:

* The root node represents the entire dataset.
* Each internal node makes a decision based on a feature.
* Each branch corresponds to the outcome of that decision.
* Each leaf node signifies the predicted value.

The construction of the decision tree involves:

* Choosing the feature that best splits the data to minimize a chosen criterion, such as mean squared error.
* Dividing the data into subsets based on this feature.
* Recursively applying the process to each subset until a stopping criterion is met (e.g., maximum tree depth or minimum samples per node).

While decision tree regression effectively models complex interactions and relationships between features, it is susceptible to overfitting and may require techniques like pruning to enhance its generalization capability.

# RESULTS AND DISCUSSION

In Linear Regression, Polynomial Regression, and Decision Tree Regression, the models produce values as shown in the table below by combining all sample features and making predictions simultaneously. Based on the data, we will determine which model performs better in terms of prediction accuracy. The **TABLE 1** is as follows:

**TABLE 1**. Testing models

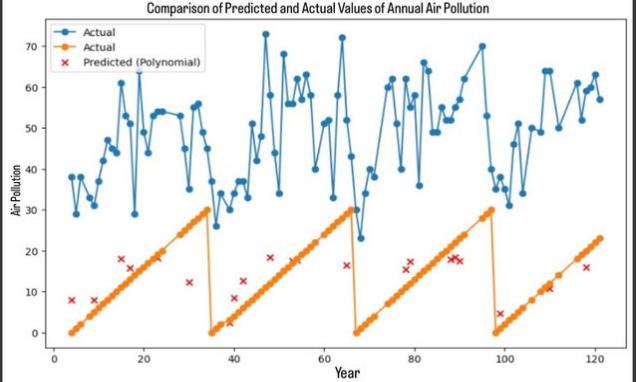
|  |  |  |
| --- | --- | --- |
| **Models** | **MSE** |  |
| Linear Regression | 54.413360 | 0.187942 |
| Polynomial Regression | 33.882651 | 0.187942 |
| Decision Tree Regression | 51.946379 | 0.224759 |

A graph showing the value of air pollution

Description automatically generated

**FIGURE 1**. Graph of Linear Regression

This **FIGURE 1** predicts air pollution levels for the coming days. The blue dots represent the actual values on the Y-axis, while the orange dots represent the predicted values on the X-axis. Based on the MSE and R-squared values, it can be concluded that the model does not provide good predictions. The high MSE of 54.41 indicates that the average difference between the predicted and actual values is quite large. The low R-squared value of 0.188 indicates that the model explains only about 18.8% of the variance in the data.



**FIGURE 2**. Graph of Polynomial Regression

From this **FIGURE 2** graph and the values of MSE and R-squared, it is evident that the high MSE of 33.88 indicates that the model is not very accurate in predicting pollution levels. This is supported by the low R-squared value of 0.188, which indicates that the model explains only about 18% of the variance. Overall, the Polynomial Regression model is not well-suited for predicting annual air pollution levels and needs further improvement using a more accurate model.

A diagram of a tree regression

Description automatically generated with medium confidence

**FIGURE 3**. Graph of Decision Tree Regression

The graph in **FIGURE 3** indicates that the Decision Tree Regression model provides a relatively good prediction. An R-squared value of 0.22 shows that about 22% of the variability in “hari” can be explained by “parameter\_pencemar\_kritis” using this model, suggesting that other factors not included in the model may also play a role. The red dots (predicted) represent the days predicted by the model. The closer the red dots are to the blue dots (actual), the more accurate the model is in predicting future values. Conversely, the greater the distance between these points, the larger the prediction error of the model. By observing the distribution of these points, we can visually assess how well the model captures the relationship between the features in making predictions.

Comparison of Models Overall, Linear Regression, Polynomial Regression, and Decision Tree Regression demonstrate limited predictive capabilities for air pollution levels in Jakarta, with low R-squared values across all models. While Polynomial Regression slightly reduces the MSE, it does not provide a significant improvement over Linear Regression in terms of R-squared. Decision Tree Regression shows a marginally better performance, but it still lacks sufficient explanatory power.

# CONCLUSIONS

The aim of this study was to forecast air pollution levels in Jakarta using machine learning techniques. We utilized multiple datasets and employed three different methods—Linear Regression, Polynomial Regression, and Decision Tree Regression—to maken predictions. Each model’s performance was evaluated based on Mean Squared Error (MSE) and R-squared values, as detailed in the previous results. The analysis revealed that Linear Regression achieved the lowest MSE and the highest R-squared value, indicating superior predictive accuracy compared to Polynomial Regression and Decision Tree Regression. While these results demonstrate Linear Regression’s effectiveness in this context, it is essential to recognize that model performance can vary based on dataset characteristics and parameter settings. In summary, Linear Regression has proven to be the most effective model for predicting air pollution levels in Jakarta among the methods tested.

# References

1. S. Suroshe, S. V. Dharpal, and N. W. Ingole, "Prediction of Air Quality Index Using Regression Models," *GIS science journal,* vol. 9, no. 8, pp. 576-591, 2022.
2. L. Kumar and G. Kumar, "A Prediction Model for Air Pollution using Artificial Neural Network and Multiple Linear Regression," 2024.
3. S. Pandya, H. Ghyvat, K. Kotecha, and P. Gope, "Linear regression and artificial neural network (ANN)-based approaches to predict air pollution," 2023.
4. K. S. Harishkumar, K. M. Yogesh, and I. Gad, "Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models," *Procedia Computer Science,* vol. 171, pp. 2057-2066, 2020.
5. C. R. Aditya, C. R. Deshmukh, D. K. Nayana, and P. G. Vidyavastu, "Detection and prediction of air pollution using machine learning models," *International journal of engineering trends and technology (IJETT),* vol. 59, no. 4, pp. 204-207, 2018.
6. M. Zeinalnezhad, A. G. Chofreh, F. A. Goni, and J. J. Klemeš, "Air pollution prediction using semi-experimental regression model and Adaptive Neuro-Fuzzy Inference System," *Journal of Cleaner Production,* vol. 261, p. 121218, 2020.
7. T. D. Akinosho, M. Bilal, E. T. Hayes, A. Ajayi, A. Ahmed, and Z. Khan, "Deep learning-based multi-target regression for traffic-related air pollution forecasting," *Machine Learning with Applications,* vol. 12, p. 100474, 2023.
8. W. N. F. W. Azmi, T. R. Pillai, M. T. Latif, S. Koshy, and R. Shaharudin, "Application of land use regression model to assess outdoor air pollution exposure: A review," *Environmental Advances,* vol. 11, p. 100353, 2023.