Jurnal Varian
Vol. 9, No. 1, February 2026, pp 77–84
e-ISSN: 2581-2017
DOI: https://doi.org/10.30812/varian.v9i1.6046

Evaluating Random Forest Regression for Air Quality Prediction

Muh. Basyar Izabi, Suwardi Annas, Ansari Saleh Ahmar
Universitas Negeri Makassar, Makassar, Indonesia

Article Info

Article history:

Received : 12-21-2025
Revised : 02-21-2026
Accepted : 02-26-2026

Keywords:

Air Pollution Prediction;
Air Quality;
Environmental Data Analysis;
Machine Learning;
Random Forest.

ABSTRACT

Air pollution is a growing environmental issue in Makassar due to rapid urban development and increasing transportation activity. This study aims to model and predict air pollutant concentrations using the Random Forest (RF) regression method. The data consist of daily $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, and O_3 measurements from September 2024 to September 2025, totaling 395 observations. Missing values (14.05%) were addressed using a hybrid approach combining linear interpolation and multiple linear regression. The RF model was trained under two data-split scenarios (70:30 and 80:20) and evaluated using SMAPE, RMSE, MAE, and R^2. The results show that the 80:20 configuration provides the best predictive accuracy. CO and O_3 yield the most accurate predictions with SMAPE values of 9.75% and 10.87%, and R^2 of 0.973 and 0.964, respectively. $PM_{2.5}$ and PM_{10} also show strong performance, with R^2 values above 0.84. These results indicate that the RF model effectively captures pollutant variability and provides reliable forecasts. Overall, Random Forest has been shown to be a robust and accurate method for predicting air quality in Makassar, supporting environmental monitoring and early warning systems. Despite its strong performance, this study is limited to two data-partition schemes and does not incorporate temporal deep-learning architectures. Future studies may investigate hybrid ensembles or deep learning approaches to determine whether incorporating sequential modeling further enhances predictive stability.

Corresponding Author:

Suwardi Annas,
Department of Statistics, Universitas Negeri Makassar,
Email: suwardi_annas@unm.ac.id

Copyright ©2026 The Authors.
This is an open access article under the CC BY-SA license.

How to Cite:
Izabi, M. B., Annas, S., & Ahmar, A. S. (2026). Evaluating Random Forest Regression for Air Quality Prediction. *Jurnal Varian*, *9*(1), 77–86.
This is an open access article under the CC BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

A. INTRODUCTION

Air quality has become a critical environmental and public health concern worldwide, particularly in rapidly urbanizing cities where industrial activities and transportation emissions contribute to elevated pollutant concentrations. Based on daily monitoring data retrieved from aqicn.org, pollutant levels in Makassar from September 2024 to September 2025 indicate substantial variability, with PM2 $\cdot$ 5 concentrations peaking above 60, PM10 exceeding 100, NO2 rising above 190, and O3 surpassing 200 during high-episode periods. The average concentrations of PM2 $\cdot$ 5 and PM10 throughout the study period remained considerably higher than recommended guideline values, reflecting persistent air quality concerns. These observed fluctuations highlight the urgency of developing accurate predictive models to support environmental monitoring systems and inform data-driven policy decisions.

Over time, modeling approaches have evolved from traditional statistical techniques to advanced machine learning algorithms capable of capturing nonlinear and complex pollutant interactions. In Makassar, Rahmat et al. (2023) applied Support Vector Regression (SVR) to forecast air quality and reported promising predictive performance, although further optimization was required. Farhan et al. (2024) subsequently implemented the Generalized Space-Time Autoregressive (GSTAR) model to account for spatial and temporal dependencies, demonstrating the importance of dynamic modeling in environmental prediction. Other studies have used machine learning approaches, such as Support Vector Machines (SVM), to classify air quality levels based on multiple pollutant parameters and to improve interpretability through data visualization (Lubis et al., 2026), thereby supporting decision-making and

public awareness. These studies indicate ongoing efforts to enhance predictive and analytical capabilities; however, the increasing complexity of pollutant interactions necessitates more robust ensemble-based approaches. These studies indicate ongoing efforts to enhance predictive accuracy; however, the growing complexity of pollutant interactions necessitates more robust ensemble-based approaches.

Recent studies have increasingly emphasized the superiority of ensemble learning methods, particularly Random Forest (RF), in handling nonlinear environmental datasets. J. Yang et al. (2023) demonstrated that RF outperformed Radial Basis Function, Backpropagation Neural Network, and Support Vector Machine models in terms of generalization performance and resistance to overfitting. Hu and Szymczak (2023) further showed that RF effectively modeled pollutant concentrations characterized by irregular fluctuations and nonlinear patterns. Similarly, Li et al. (2022) reported that RF successfully captured spatial and temporal variability in PM2 $\cdot$ 5 concentrations, surpassing several baseline models in predictive accuracy. A more recent urban case study by Alzu'bi et al. (2024) confirmed that RF maintains strong generalization performance under varying pollution conditions. In addition, L. Yang et al. (2020) demonstrated that Random Forest can achieve high predictive accuracy (R^2 above 0.90) in estimating PM2 $\cdot$ 5 concentrations using satellite-based inputs, while Sun et al. (2021) further showed that RF is capable of modeling air pollution dynamics at high temporal resolution with strong performance (R^2 up to 0.95). Collectively, these findings establish RF as a reliable and robust method for environmental prediction; nevertheless, most prior studies focus on specific pollutants, limited geographic regions, or single data-partition schemes.

Despite the proven effectiveness of Random Forest (RF), studies examining its regression performance for simultaneous multi-pollutant prediction in Makassar City remain limited. Previous research in Makassar has applied advanced models, such as encoder–decoder long short-term memory (EDLSTM), to predict AQI and $PM_{2.5}$ with high accuracy (Soedjarwo et al., 2023), indicating the potential of machine learning approaches in this domain. However, this study is limited to specific pollutants and does not explore multiple data-partition scenarios. Therefore, this study addresses these gaps by evaluating the performance of the Random Forest regression model for multi-pollutant air-quality prediction under different data-splitting schemes, using comprehensive evaluation metrics to ensure model stability and generalization.

The objective of this research is to evaluate the predictive performance of the Random Forest regression model for estimating multiple air pollutant concentrations in Makassar City. The model is assessed using SMAPE, RMSE, MAE, and R^2 across two data-partition schemes (70:30 and 80:20) to examine model stability and generalization. The findings are expected to advance ensemble learning applications in environmental monitoring and support data-driven strategies for local air-quality management. Furthermore, this study opens the door to future exploration of advanced Random Forest variants, such as the hedged random forest, which has been shown to further enhance forecasting performance through adaptive tree weighting (Beck & Wolf, 2026).

B. RESEARCH METHOD

The tools used to support the data processing in this research are RStudio and Microsoft Excel. The data used in this study are secondary data obtained from the aqicn.org platform, sourced from the Makassar air-monitoring station managed by the Ministry of Environment and Forestry. The dataset consists of daily air-quality measurements from September 2024 to September 2025, covering six pollutants: $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, and O_3, with a total of 395 daily observations per pollutant. Approximately 14.05% of the dataset contained missing values, mainly in NO2 and $PM_{2.5}$, which were addressed using linear interpolation and multiple linear regression imputation to ensure data completeness. After preprocessing, the dataset was analyzed in RStudio using the randomForest package to fit a Random Forest Regression model. The modeling process was carried out under two data-partition scenarios, namely 70% training–30% testing and 80% training–20% testing, to evaluate the model's predictive performance across different training proportions. The predictive accuracy of the model was assessed using Symmetric Mean Absolute Percentage Error (SMAPE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

1. Random Forest

Random Forest (RF) is a widely used machine learning method for prediction and data analysis, particularly effective in handling nonlinear and complex datasets. The model operates by constructing an ensemble of decision trees and aggregating their outputs to produce more accurate and stable predictions. Each tree is trained using a randomly selected subset of data with replacement (bootstrap sampling), which helps reduce variance and improve model performance (Al-Mahdawi et al., 2023). Jose and Gopakumar (2019) defined the generalization error of an ensemble classifier using a margin function, which quantifies the difference between the average vote for the correct class and the highest vote among the competing classes. Given a random input–output pair (XY), where X represents the feature vector and Y denotes the true class label, each classifier $h_k(X)$ in the

DOI: https://doi.org/10.30812/varian.v9i1.6046

ensemble produces a prediction. The margin is then defined as the difference between the expected vote for the correct class Y and the maximum expected vote across all other classes $j \neq Y$. This formulation provides a mathematical framework for evaluating classifier performance, where a larger margin indicates stronger classification confidence. The margin function is expressed in Equation (1).

$$mg(X,Y) = \frac{\sum_{k=1}^{K} I(h_k(X) = Y)}{K} - \max_{j \neq Y} \left[\frac{\sum_{k=1}^{K} I(h_k(X) = j)}{K} \right] \tag{1}$$

One approach to improving the predictive accuracy of the Random Forest model is by increasing its strength. The strength reflects how well individual decision trees correctly classify the input data. This concept can be mathematically represented as the expected value of the margin function, expressed as Equation (2):

$$s = E_{X,Y} mg(X,Y) \tag{2}$$

where E denotes the expectation operator over the joint distribution of input X and output Y. From this formulation, it can be understood that improving model strength requires increasing the margin value, which can be achieved by enhancing the predictive capability of individual trees within the ensemble. The performance of each tree is influenced by the decisions made at each node, which are determined by impurity measures. One commonly used metric to evaluate node impurity is the Gini Diversity Index, defined as in Equation (3).

$$I_t = 1 - \sum_i p(i)^2 \tag{3}$$

where I_t represents the impurity at node t, i denotes each class, and $P(i)$ is the proportion of observations belonging to class i at that node. The summation is taken over all classes present in node t. If a node contains observations from only one class, its impurity is zero, indicating a pure node; otherwise, it is greater than zero (Jose & Gopakumar, 2019).

2. Symmetric Mean Absolute Percentage Error

sMAPE measures the relative accuracy between predicted and actual values and is calculated as in Equation (4).

$$sMAPE = \frac{2}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)} \times 100\% \tag{4}$$

where A_t denotes the actual observed value at time t, F_t represents the forecasted (predicted) value at time t, n is the total number of observations, and the absolute value symbols $|\cdot|$ represent the magnitude of the difference without considering its direction. sMAPE provides a percentage-based interpretation of prediction error, making it suitable for evaluating model performance across variables with different scales (Yuliyanto et al., 2023).

3. Root Mean Square Error

RMSE measures the square root of the average squared difference between predicted and actual values and is calculated as in Equation (5) (Soedjarwo et al., 2023):

$$RMSE = \sqrt{\sum_{t=1}^{n} \frac{(A_t) - (F_t)^2}{n}} \tag{5}$$

RMSE emphasizes larger errors due to squaring, making it effective for evaluating models where substantial prediction errors require greater penalization.

DOI: https://doi.org/10.30812/varian.v9i1.6046

4. Mean Absolute Error

MAE measures the average magnitude of the absolute difference between predicted and actual values and is calculated as Equation (6):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|A_t - F_t| \tag{6}$$

MAE provides a straightforward interpretation of prediction error, making it useful for evaluating overall model accuracy across different datasets (Benedict, 2022).

5. Coefficient Determination (R^2)

The Coefficient of Determination (R^2) quantifies the proportion of variance in the observed data explained by the model in Equation (7).

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(A_t - F_t)^2}{\sum_{t=1}^{n}(A_t - \bar{A})^2} \tag{7}$$

A higher R^2 value indicates that the model better captures the variance of the actual data, demonstrating its explanatory strength (Indartini & Mutmainah, 2024).

6. Data Imputation

Missing values within the dataset were addressed using two complementary approaches: linear interpolation and multiple linear regression. The selection of these methods was based on the temporal nature of air quality data, where concentration values generally change smoothly over time. For variables with a relatively small proportion of missing values, such as $PM_{2.5}$ and SO_2, the gaps were filled using linear interpolation, which estimates missing values based on adjacent data points. The formula for linear interpolation is expressed in Equation (8) (Mansyur et al., 2024).

$$\hat{y}(x) = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) \tag{8}$$

where $\hat{y}(x)$ denotes the estimated interpolated value at point x, while y_0 and y_1 represent the actual observed values before and after the missing data points, respectively. The terms x_0 and x_1 correspond to the time indices of those preceding and subsequent observations.

Meanwhile, for variables with a higher proportion of missing values, particularly PM_{10} and NO_2, missing values were imputed using multiple linear regression, with correlated pollutant variables as predictors. The general form of the regression model is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n + \varepsilon \tag{9}$$

where Y represents the dependent variable (missing data), X_i are the predictor variables, β_i denote the regression coefficients, and ε is the error term. This hybrid imputation approach ensures the dataset's continuity and consistency while maintaining the temporal and inter-variable relationships critical for accurate air quality prediction.

7. Data Splitting and Data Training

The cleaned dataset was divided into two configurations: 70% for training and 30% for testing, and 80% for training and 20% for testing. The data partition was performed using random sampling with a fixed random seed to ensure reproducibility of the results. Specifically, 70% (or 80%) of the observations were randomly selected as training data, while the remaining observations were used as test data. This approach allows the model to learn from a representative subset of the dataset and evaluate its predictive performance on unseen data. Training data represent the actual observations used by the model to learn existing patterns, while testing data serve to evaluate how well the model performs on unseen data (Anggraini et al., 2019). The proportion between training and testing data is an important factor that influences model accuracy and precision; an inappropriate

ratio may reduce the reliability of the prediction results (Musu et al., 2021). In machine learning, both data partitions play a key role, where the system first learns from training data and subsequently applies the learned patterns to testing data for validation (Irawan, 2021).

The Random Forest regression model was trained separately for each pollutant variable using the *randomForest* package in RStudio. The model employed 500 decision trees ($ntree = 500$) and used the default value for the number of randomly selected variables ($mtry$). This configuration follows Breiman's principle that combining multiple trees in an ensemble reduces overfitting while maintaining high predictive accuracy.

C. RESULT AND DISCUSSION

1. Data Overview

The first step before conducting further analysis is to describe the air quality data for Makassar City during the observation period. Descriptive statistics for each pollutant variable are presented in Table 1.

Table 1. The performance of air quality variables

Variable	Minimum	Median	Mean	Maximum
$PM_{2.5}$	1.29	9.52	10.18	25.48
PM_{10}	0.17	14.88	16.49	60.03
CO	0.01	1.01	0.98	3.72
NO_2	2.12	49.22	51.53	196.52
SO_2	0.02	45.87	47.71	135.85
O_3	0.28	66.92	68.09	214.15

Based on Table 1, the average concentrations of air pollutants in Makassar City vary across parameters. The mean concentration of $PM_{2.5}$ was 10.18 with a median of 9.52, while PM_{10} had an average value of 16.49 and a median of 14.88. The average CO concentration reached 0.98, and NO_2 showed a relatively high mean of 51.53. Meanwhile, SO_2 and O_3 recorded mean concentrations of 47.71 and 68.09, respectively. Overall, the variation between minimum and maximum values indicates dynamic fluctuations in pollutant levels during the observed period.

2. Prediction Performance of Random Forest

To evaluate the Random Forest model's ability to predict air pollutant concentrations, two data partitioning scenarios were implemented: 70% for training and 30% for testing, and 80% for training and 20% for testing. The predicted values were compared with the actual values for each pollutant variable to assess the model's ability to capture temporal variation and trends. The following Figures 1 and 2 illustrate the comparison between actual and predicted values for all pollutant variables under both data-splitting scenarios.

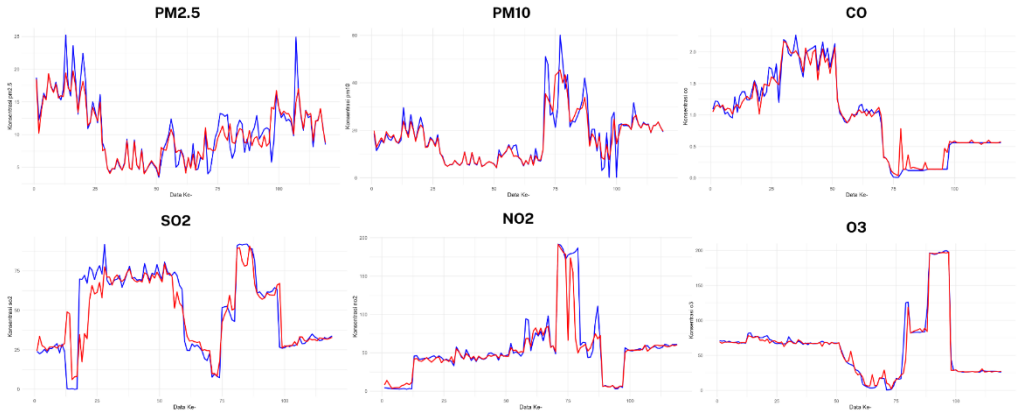

Figure 1. Comparison of actual (blue) and predicted (red) pollutant concentrations using Random Forest with 70:30 data partition

The comparison plots for the 70:30 data partition show that the Random Forest model successfully captured the general

fluctuation patterns of all pollutant variables. The predicted lines (red) closely follow the actual observations (blue), indicating good alignment between the model and the real data. Slight deviations are observed at several peak and trough points, particularly for SO_2 and NO_2, which exhibit higher variability. Nevertheless, the prediction trends remain consistent with the actual dynamics across the observation period. Overall, these results demonstrate that Random Forest can effectively represent the nonlinear and volatile characteristics of air pollutant data, even with a smaller proportion of training samples.

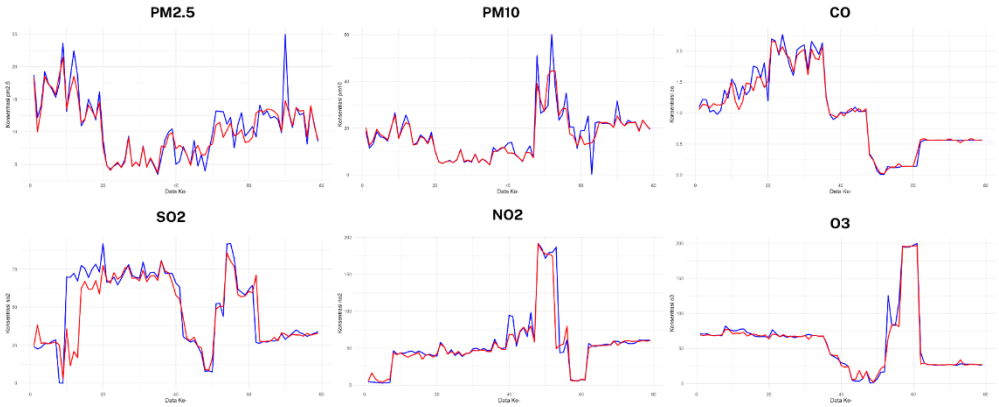

Figure 2. Comparison of actual (blue) and predicted (red) pollutant concentrations using Random Forest with 80:20 data partition

The plots for the 80:20 data partition indicate an improved alignment between predicted and actual pollutant concentrations compared to the previous scenario. The red lines representing the Random Forest predictions appear smoother and more stable, closely following the actual data across nearly all variables. Deviations are relatively minor, particularly for PM2 $\cdot$ 5, PM10, and O_3, while NO_2 and SO_2 still exhibit some fluctuations due to their higher temporal variability. The increased proportion of training data enhances the model's ability to capture complex relationships and generalize patterns more effectively. Overall, these results demonstrate that Random Forest achieves higher prediction accuracy and greater consistency when trained on a larger dataset.

Table 2. Evaluation Metrics of Random Forest Model for Air Quality Prediction

Variable	SMAPE (70:30)	SMAPE (80:20)	RMSE (70:30)	RMSE (80:20)	MAE (70:30)	MAE (80:20)	R^2 (70:30)	R^2 (80:20)
$PM_{2.5}$	10.94%	10.06%	1.949	1.772	1.137	1.064	0.846	0.876
PM_{10}	12.57%	10.74%	4.48	3.917	2.211	1.959	0.835	0.851
CO	13.67%	9.75%	0.12	0.111	0.073	0.069	0.969	0.973
NO_2	16.17%	12.72%	20.44	16.59	6.958	5.155	0.782	0.845
SO_2	19.93%	19.06%	12.03	13.56	6.42	6.691	0.784	0.706
O_3	12.13%	10.87%	9.619	8.618	3.554	3.308	0.959	0.964

Table 2 presents the performance metrics of the Random Forest model across six pollutant variables under two data split scenarios (70:30 and 80:20). Overall, the model demonstrates strong predictive performance, with most variables showing improvements when trained using 80% of the data. The lowest error rates were obtained for CO and O3, indicated by SMAPE values below 11% and R^2 exceeding 0.96, reflecting excellent model fit and stability. $PM_{2.5}$ and PM_{10} also achieved consistent results, with R^2 values above 0.84, suggesting that Random Forest effectively captures their temporal variation.

In contrast, SO_2 exhibited the weakest performance, with the RMSE and MAE slightly increasing under the 80:20 configuration, and R^2 declining to 0.706, suggesting that its concentration pattern may be more irregular or less predictable. Nevertheless, the overall findings confirm that a larger training-to-test split (80:20) enhances model generalization and reduces prediction error, reinforcing the robustness of Random Forest for air quality prediction in Makassar City.

These findings are consistent with previous studies. J. Yang et al. (2023) demonstrated that Random Forest outperformed models such as Radial Basis Function, Backpropagation Neural Network, and Support Vector Machine in predicting traffic accident severity. Similarly, Li et al. (2022) showed that Random Forest effectively captures spatio-temporal variability in $PM_{2.5}$

concentrations with strong predictive accuracy, while Hu and Szymczak (2023) highlighted its ability to handle complex, high-dimensional, and longitudinal data structures.

D. CONCLUSION AND SUGGESTION

This study demonstrates that the Random Forest regression model provides strong predictive performance in estimating air quality parameters in Makassar City. Across six pollutant variables ($PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, and O_3), the model consistently achieved low sMAPE, RMSE, and MAE values, with high R^2 scores exceeding 0.84 for most pollutants, indicating high accuracy and reliability. The 80:20 data split yielded superior performance compared to the 70:30 configuration, emphasizing the benefit of larger training data proportions in improving generalization and stability. Among all pollutants, CO and O_3 yielded the most accurate predictions, whereas SO_2 showed greater variability due to its fluctuating emission patterns. These findings affirm that Random Forest is an effective and robust approach for modeling complex and nonlinear relationships in air quality data.

For future research, it is recommended to extend this work by comparing Random Forest with other ensemble or deep learning algorithms, such as Gradient Boosting, XGBoost, or hybrid models, to further enhance prediction accuracy. Additionally, incorporating meteorological variables and spatiotemporal components could provide a more comprehensive understanding of air pollution dynamics and enhance the robustness of predictive modeling for urban air quality assessment.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Universitas Negeri Makassar for providing the academic environment that supported the completion of this research. Special appreciation is extended to the supervisors (co-author) for their valuable guidance, constructive feedback, and continuous support throughout the research and writing process. The authors also acknowledge the Ministry of Environment and Forestry and the aqicn.org platform for providing access to the air-quality data used in this study.

DECLARATIONS

AUTHOR CONTRIBUTION

According to the authors, this manuscript was prepared through contributions from both parties. The first author carried out the entire research workflow, including conceptualizing the study, processing and analyzing the data, running the software, and writing the full manuscript. The second and the third author contributed by providing academic supervision, guidance, and critical review throughout the research and writing process.

FUNDING STATEMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

COMPETING INTEREST

The authors declare that they have no competing financial or personal interests that could influence the work reported in this article. This manuscript was prepared solely to fulfill the academic requirements for completing the first author's Master's degree program.

REFERENCES

Al-Mahdawi, H. K., Alkattan, H., Subhi, A. A., Al-hadrawi, H. F., Abotaleb, M., Ali, G. K., Mijwil, M. M., Towfeek, A.-S. K., & Helal, A. H. (2023). Analysis and prediction of evaporation rates using random forest models: A case study of Almaty city. *Babylonian Journal of Machine Learning*, *2023*, 55–64. https://doi.org/10.58496/BJML/2023/010

Alzu'bi, F., Al-Rawabdeh, A., & Almagbile, A. (2024). Predicting air quality using random forest: A case study in Amman-Zarqa. *The Egyptian Journal of Remote Sensing and Space Sciences*, *27*(3), 604–613. https://doi.org/10.1016/j.ejrs.2024.07.004

Anggraini, R. A., Widagdo, G., Budi, A. S., & Qomaruddin, M. (2019). Penerapan Data Mining Classification untuk Data Blogger Menggunakan Metode Naïve Bayes. *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, *7*(1), 47. https://doi.org/10.26418/justin.v7i1.30211

Beck, E., & Wolf, M. (2026). Forecasting inflation with the hedged random forest. *Empirical Economics*, *70*(2), 23. https://doi.org/10.1007/s00181-025-02879-x

Benedict, L. (2022). *Prediksi Tingkat Kematian Covid-19 di Indonesia dengan menggunakan Metode Linear Regression*. https://kc.umn.ac.id/id/eprint/22407/

Farhan, M., Sanusi, W., & Ihsan, H. (2024). Pemodelan Pencemaran Udara sebagai Solusi Penurunan Kualitas Udara Menggunakan Generalized Space-Time Autoregressive di Kota Makassar. *Journal of Mathematics, Computations and Statistics*, *7*(2), 258–274. https://doi.org/10.35580/jmathcos.v7i2.4304

Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. *Briefings in Bioinformatics*, *24*(2), bbad002. https://doi.org/10.1093/bib/bbad002

Indartini, M., & Mutmainah, M. (2024). *Analisis Data Kuantitatif: Uji Instrumen, Uji Asumsi Klasik, Uji Korelasi dan Regresi Linier Berganda*. Penerbit Lakeisha. https://doi.org/978-623-119-036-9

Irawan, Y. (2021). Penerapan Algoritma Decision Tree C4.5 untuk Memprediksi Kelayakan Calon Pendonor Melakukan Donor Darah dengan Klasifikasi Data Mining. *JTIM : Jurnal Teknologi Informasi dan Multimedia*, *2*(4), 181–189. https://doi.org/10.35746/jtim.v2i4.75

Jose, C., & Gopakumar, G. (2019). An Improved Random Forest Algorithm for classification in an imbalanced dataset. *2019 URSI Asia-Pacific Radio Science Conference (AP-RASC)*, 1–4. https://doi.org/10.23919/URSIAP-RASC.2019.8738232

Li, X., Li, L., Chen, L., Zhang, T., Xiao, J., & Chen, L. (2022). Random Forest Estimation and Trend Analysis of PM2.5 Concentration over the Huaihai Economic Zone, China (2000–2020). *Sustainability*, *14*(14), 8520. https://doi.org/10.3390/su14148520

Lubis, F. H., Fakhriza, F., & Putri, R. A. (2026). Analysis of Air Pollution Standard Index Using Support Vector Machine Algorithm. *Building of Informatics, Technology and Science (BITS)*, *7*(4), 2761–2770. https://doi.org/10.47065/bits.v7i4.9506

Mansyur, N. N., Arman, A., Gubu, L., Somayasa, W., & Aswani, A. (2024). Penerapan Metode Interpolasi Lagrange dalam Meramalkan Jumlah Pendapatan pada Percetakan (Studi Kasus: Gevira Advertising): Interpolasi Lagrange dalam Meramalkan Jumlah Pendapatan pada Percetakan. *Jurnal Matematika Komputasi dan Statistika*, *4*(1), 540–546. https://doi.org/10.33772/jmks.v4i1.80

Musu, W., Ibrahim, A., & Heriadi, H. (2021). Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4.5. *SISITI : Seminar Ilmiah Sistem Informasi dan Teknologi Informasi*, *10*(1), 186–195. https://doi.org/10.36774/sisiti.v10i1.802

Rahmat, R. W., Annas, S., & Rais, Z. (2023). Analisis Support Vector Regression (SVR) untuk meramalkan Indeks Kualitas Udara di Kota Makassar. *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, *5*(03), 104–117. https://doi.org/10.35580/variansiunm107

Soedjarwo, M., Arifin, B., Tahir, A. M., Farmasiantoro, A., Priyanto, I., & Fauzi, A. (2023). Daily prediction of air quality standard in Makassar city, Indonesia, 040008. https://doi.org/10.1063/5.0181597

Sun, J., Gong, J., & Zhou, J. (2021). Estimating hourly PM2.5 concentrations in Beijing with satellite aerosol optical depth and a random forest approach. *Science of The Total Environment*, *762*, 144502. https://doi.org/10.1016/j.scitotenv.2020.144502

Yang, J., Han, S., & Chen, Y. (2023). Prediction of Traffic Accident Severity Based on Random Forest (I. Ghosh, Ed.). *Journal of Advanced Transportation*, *2023*, 1–8. https://doi.org/10.1155/2023/7641472

Yang, L., Xu, H., & Yu, S. (2020). Estimating PM2.5 concentrations in Yangtze River Delta region of China using random forest model and the Top-of-Atmosphere reflectance. *Journal of Environmental Management*, *272*, 111061. https://doi.org/10.1016/j.jenvman.2020.111061

Yuliyanto, M. R., Wuryandari, T., & Utami, I. T. (2023). Peramalan Pendapatan Bulanan Menggunakan Fuzzy Time Series Chen Orde Tinggi. *Jurnal Gaussian*, *12*(1), 61–70. https://doi.org/10.14710/j.gauss.12.1.61-70