P-ISSN: 2685-8738 E-ISSN: 2685-872X

Variance: Journal of Statistics and Its Applications
Volume 7 Issue 1 | April 2025
10.30598/variancevol7iss1page11-20

PERFORMANCE COMPARISON OF DECISION TREE MODELS FOR PM10 PREDICTION IN JAKARTA

Khairummin Alfi Syahrin[1*], Agung Hari Saputra[2]
[1,2]Undergraduate Program in Applied Meteorology, Indonesia State College of Meteorology Climatology and Geophysics, Tangerang City, Indonesia
Jl. Meteorologi, Tangerang City, 15119, Banten, Indonesia

Corresponding Author's Email: *khairummin19alfisyahrin@gmail.com*

Abstract: *PM10 refers to airborne particulate matter with a diameter of 10 micrometers or less. The potential health hazards associated with PM10 are a growing concern and continue to be the focus of extensive research. This research utilizes PyCaret, a library to accelerate the process of modeling and experimentation in the field of machine learning (ML) and data science. This research compares the performance of three decision tree-based models, Extra Trees, Random Forest, and XGBoost, in predicting PM10 particulate levels, presenting data and visualizations for each model's predictions. The data used is ISPU data at five air quality monitoring stations in Jakarta, with the primary dataset of PM10 in 2021. The forecast results show an increasing graph pattern, with higher fluctuations in XGBoost. The Extra Trees model performs best, with MASE 0.8808, RMSSE 0.8113, MAE 12.6173, RMSE 14.7436, MAPE 0.2433, SMAPE 0.207, and R² -1.2013.*

Keywords: *Extra Trees, PM10, Random Forest, XGBoost*

1. INTRODUCTION

Due to rapid economic development, Indonesia has substantially increased pollutant emissions and atmospheric pollution levels, particularly in major cities. This issue is further exacerbated by frequent biomass burning and urban air pollution, which often occur consecutively in several wildfire-prone[1]. Jakarta, one of Southeast Asia's most densely populated metropolitan areas, is especially affected and faces serious air quality challenges [2].

Air pollution has become a pressing global issue, demanding immediate attention due to its detrimental effects on both public health and environmental sustainability [3]. Harmful atmospheric substances characterize it, including gases, chemicals, and particulate matter. Among the most concerning pollutants is Particulate Matter (PM), which consists of microscopic particles such as dust, smoke, and soot. PM10 refers to particles with a diameter of 10 micrometers or less—small enough to penetrate deep into the respiratory system and cause severe health problems, particularly in densely populated urban environments with elevated pollution levels [4],[5].

Decision tree-based algorithms are often applied in this context to predict PM10 concentrations due to their simplicity, flexibility, and ability to handle both numerical and categorical inputs. [6] Popular examples of these models include Random Forest, Extra Trees, and XGBoost [7]. Random Forest, for instance, consists of an ensemble of decision trees. Each tree is generated from a different subset of the training data and random subsets of features. The final prediction is made by averaging the predictions from all individual trees, making it less prone to overfitting. Random Forest models are particularly useful for time series data, such as predicting PM10 concentrations, because they handle multiple features effectively and can generalize well to new data [8]. Extra Trees, or Extremely Randomized Trees, is a variant of Random Forest. Like Random Forest, it also utilizes an ensemble approach but differs by selecting features and randomly splitting points during tree construction. This approach can reduce overfitting further and improve the model's performance on complex time series data [9]. XGBoost (Extreme Gradient Boosting) is another powerful model in the family of decision trees. Unlike Random Forest and Extra Trees, XGBoost builds trees sequentially, where each new tree corrects the errors of the previous ones. It has proven to be highly effective in regression tasks, such as forecasting PM10 levels, due to its ability to handle complex relationships and its robustness against overfitting [10], [11].

https://ojs3.unpatti.ac.id/index.php/variance/ *Submitted: 2025-01-10* *Accepted: 2025-04-23*

Due to its health and environmental impacts, several studies have applied machine learning to forecast PM10 levels. In Bulgaria, Random Forest was used effectively with meteorological data, achieving high accuracy in short-term PM10 predictions [12]. Similarly, in South Korea, XGBoost and LightGBM were tested, with XGBoost showing slightly better accuracy, while LightGBM offered faster training time [13]. The study of Thiruvananthapuram, India, demonstrated that Extra Trees outperformed both Random Forest and XGBoost in forecasting daily PM10 levels, achieving a high R^2 of 0.945 and an RMSE of 8.174 μg/m³. These findings suggest that Extra Trees can provide valuable insights for air quality management strategies [14].

Although these models show promise, most studies focus on different regions and use varied setups, making direct comparison difficult. They often differ in data sources, input variables, and evaluation methods. There is a lack of studies directly comparing Random Forest, Extra Trees, and XGBoost models under consistent conditions, especially for PM10 forecasting in Jakarta. Additionally, the use of advanced error metrics such as RMSSE, SMAPE, and MASE remains underexplored, which limits model evaluation depth and reliability. The aim of this research is not only to provide an overview of model performance comparison but also to provide forecasts from each model, which related to particulate matter (PM10) concentration forecasts can be utilized. This research can also be a reference for forecasters in predicting PM10 particulates, improving forecast accuracy.

2.2. METHODOLOGY

2.1. Study Area

This study was conducted in DKI Jakarta Province, which is located at geographical coordinates between 5°19'12" S - 6°23'54" S and 106°22'42" E - 106°58'18" E (Figure 1) [15]. The orange strip pattern shows the research area with the city description, and the black box shows the location of Jakarta in Java Island, Indonesia.

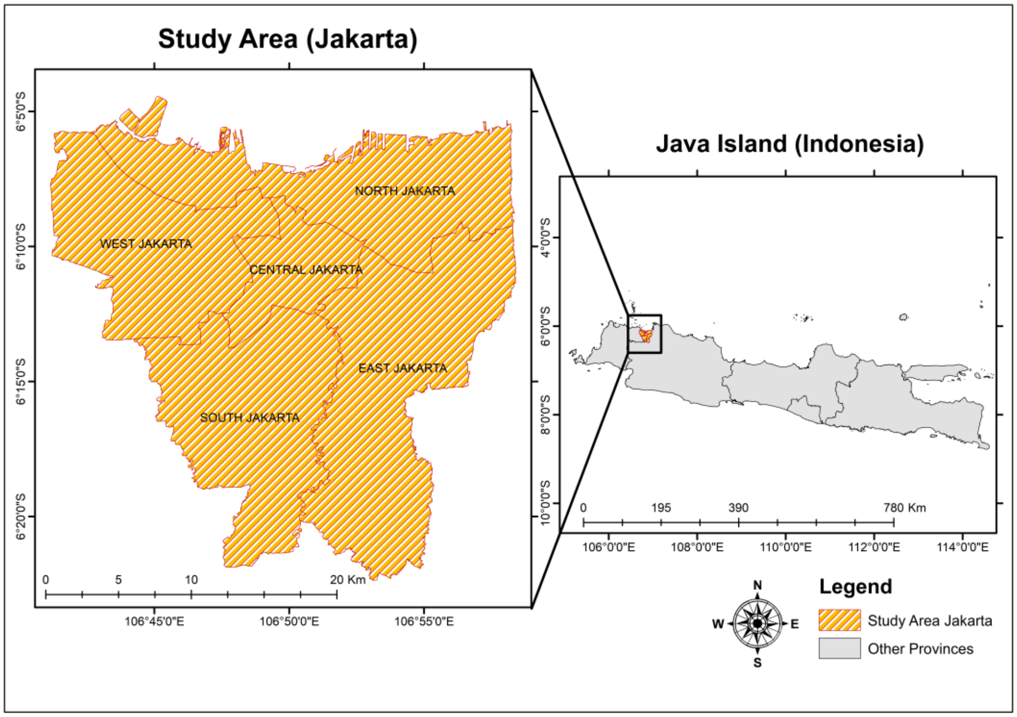


Figure 1. Study Area

2.2. Data Sources

The dataset used in this study was obtained from the Jakarta Open Data portal, specifically from the Air Pollution Standard Index (ISPU) dataset, which includes SPKU data collected from five air quality monitoring stations across DKI Jakarta. These stations are Sta Bundaran HI (Central Jakarta), Sta Kelapa Gading (North Jakarta), Sta Jagakarsa (South Jakarta), Sta Lubang Buaya (East Jakarta), and Sta Kebon Jeruk (West Jakarta) [16], [17]. The dataset covers the year 2021 and was compiled into an Excel format. It represents the overall air quality in Jakarta; this study used the average PM10 readings from all five monitoring stations to generate a single daily concentration value. The data can be accessed at https://data.jakarta.go.id/.

2.3. Analysis Method

The method used in this study is quantitative descriptive, where the data can be directly analyzed statistically to conclude [18]. The research steps started with collecting data in CSV format and processing with an Excel application by compiling tables with time index and particulate value (pm10). The next step will be Data Preprocessing on the Google Collab web to correct missing values in the data; the data is then stored using the system on Google Drive, which will be connected to Google Colab, which already has the pycaret.ary library [19]. Model Processing begins with installing the Pycaret machine learning package and importing data into Google Colab. Data imported on Google Collab will be set up with Hyperparamer to organize the model training space (Table 1) [20]. After that, the performance of the three models is compared by adding the include parameter to the compare_models command.

Table 1. Hyperparameters Syntax

No	Hyperparameters	Syntax	Value
1.	Forecasting Horizon	fh	30
2.	session	session_id	123
3.	seasonal	seasonal_period	7
4.	fold	fold	11
5.	Remove Harmonic	remove_harmonics	True
6.	Fold Strategy	fold_strategy	Sliding
7.	Coverage	coverage	0.7
8.	Verbose	verbose	True

Table 2. Models Syntax

No	Models	Syntax
1.	Extra Trees	et_cds_dt
2.	Random Forest	rf_cds_dt
3.	XGBoost	xgboost_cds_dt

After running the model, each model will obtain the error value in the form of MASE, RMSSE, MAE, RMSE, MAPE, SMAPE, and R2. As a result, the best model can be determined [21]. The equations used to calculate the error are shown below.

1) Root Mean Squared Error (RMSE)

Root Mean Square Error (RMSE) is a metric that measures the average magnitude of prediction errors by calculating the square root of the average squared differences between predicted ($\hat{x}_a$) and observed (x_a) values. RMSE is sensitive to significant errors, making it useful when minimizing them is essential. The range of RMSE is from 0 to infinity, with 0 indicating perfect predictions and higher values reflecting larger errors [22]. The equation is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{a=1}^{n}(x_a - \hat{x}_a)^2}$$

2) Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is another metric used to evaluate model accuracy. Unlike RMSE, it calculates the average of the absolute differences between the predicted values ($\hat{x}_a$) and the observed values (x_a), treating all errors equally. The range of MAE is from 0 to infinity [23]. The equation is:

$$MAE = \frac{1}{n}\sum_{a=1}^{n}|x_a - \hat{x}_a|$$

3) Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) is a metric used to evaluate prediction accuracy by calculating the absolute percentage error for each observation. The absolute error for each period is divided by the observed value, and the result is averaged over all periods. The range of MAPE typically falls between 0% and infinity, where 0% indicates perfect accuracy and larger values reflect greater discrepancies [24]. The equation is:

$$MAPE = \frac{1}{n}\sum_{i=a}^{n}\left|\frac{x_a - \hat{x}_a}{x_a}\right| \times 100$$

4) Root Mean Square Scaled Error (RMSSE)

Root Mean Squared Scaled Error (RMSSE) is an improved version of MASE that avoids the problems found in MAPE and SMAPE. It scales the model's mean squared error using the MSE from a simple reference method that predicts each value based on the previous one, making the metric more stable and reliable. Its value ranges from 0 to infinity, with lower values indicating better accuracy[25].

$$RMSSE = \sqrt{\frac{\frac{1}{n}\sum_{a=1}^{n}(x_a - \hat{x}_a)^2}{\frac{1}{n-1}\sum_{a=1}^{n}(x_a - \hat{x}_{a-1})^2}}$$

where:

x_a : actual observed value at time step a

$\hat{x}_a$: predicted value at time step a

$\hat{x}_{a-1}$: observed value at the previous time step

n : number of observations

5) Symmetric Mean Absolute Percentage Error (SMAPE)

Symmetric Mean Absolute Percentage Error (SMAPE) is an evaluation metric for prediction accuracy in time series models. It improves upon MAPE by normalizing the absolute error using the average of the absolute observed and predicted values. The general range of SMAPE is 0% to 200%, where lower values indicate better model accuracy [26].

$$SMAPE = \frac{100}{n}\sum_{a=1}^{n}\frac{|x_a - \hat{x}_a|}{\frac{|x_a| - |\hat{x}_a|}{2}}$$

6) Mean Absolute Scaled Error (MASE)

Mean Absolute Scaled Error (MASE) is a forecasting accuracy metric that scales the prediction errors using the average absolute error from a simple reference model based on training data. A value less than 1 indicates that the forecasting model performs better than the reference method, while a value greater than 1 means it performs worse [27].

$$MASE = \frac{\frac{1}{n}\sum_{a=1}^{n}|x_a - \hat{x}_a|}{\frac{1}{n-1}\sum_{a=1}^{n}|x_a - x_{a-1}|}$$

where:

x_a : actual observed value at time step a

$\hat{x}_a$: predicted value at time step a

$\hat{x}_{a-1}$: observed value at the previous time step

n : number of observations

7) Coefficient of Determination (R^2)

The coefficient of determination (R^2) is a statistical measure used to evaluate how well a model's predictions match the actual data. It represents the proportion of the total variation in the observed values that the model can explain. R^2 values range from 0 to 1, where 0 indicates that the model does not describe any of the variability in the data, and 1 indicates perfect prediction. In some cases, R^2can be negative, which means the model performs worse than simply using the mean of the actual values [28]. The formula is given as:

$$R^2 = 1 - \frac{\sum_{a=1}^{n}(x_a - \hat{x}_a)^2}{\sum_{a=1}^{n}(x_a - \bar{x})^2}$$

where:

x_a : actual value at position *a*

$\hat{x}_a$: predicted value at position *a*

$\hat{x}_{a-1}$: mean (average) of the actual values

n : number of data points

Furthermore, analysis will be carried out with the create_model command for the three models. Then, predictions for each model are made on the existing data, and the results are plotted with the plot_model command [29]. The last step is the conclusion of the research. The research framework is shown in Figure 2.

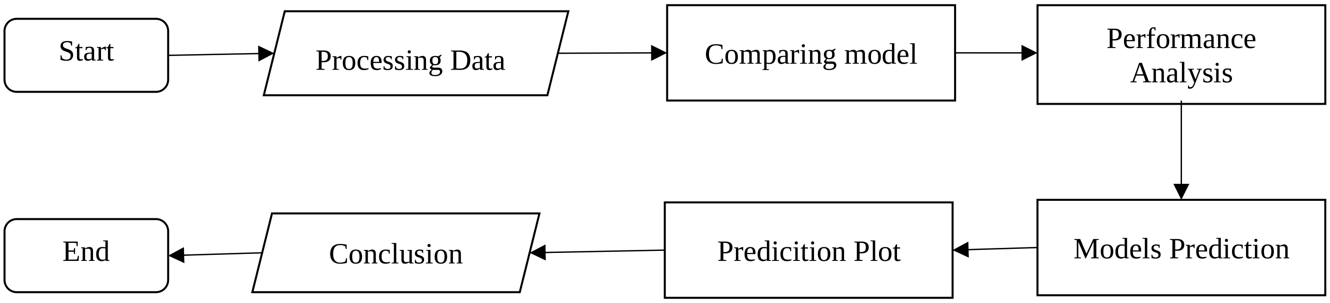


Figure 2. Methodology Framework

3. RESULTS AND DISCUSSION

3.1. Data Processing

The PM10 index dataset (μg/m³) was compiled into a single time series row using the Excel application, allowing for streamlined visualization and analysis. The dataset was then plotted in graphical form to observe the variation in PM10 concentrations throughout 2021. The graph showed that PM10 levels fluctuated over the year with a clear pattern. The highest concentration values were recorded in October and December, while the lowest values occurred in January and December (Figure 3). A decomposition analysis was also performed using a seasonal period of 7, reflecting the weekly pattern identified in the time series. The trend component displayed a rise in values around the middle of the year, and the residual component remained near zero, indicating that the data had minimal noise and was of good quality (Figure 4). This analysis served as a step to confirm data reliability before further steps.

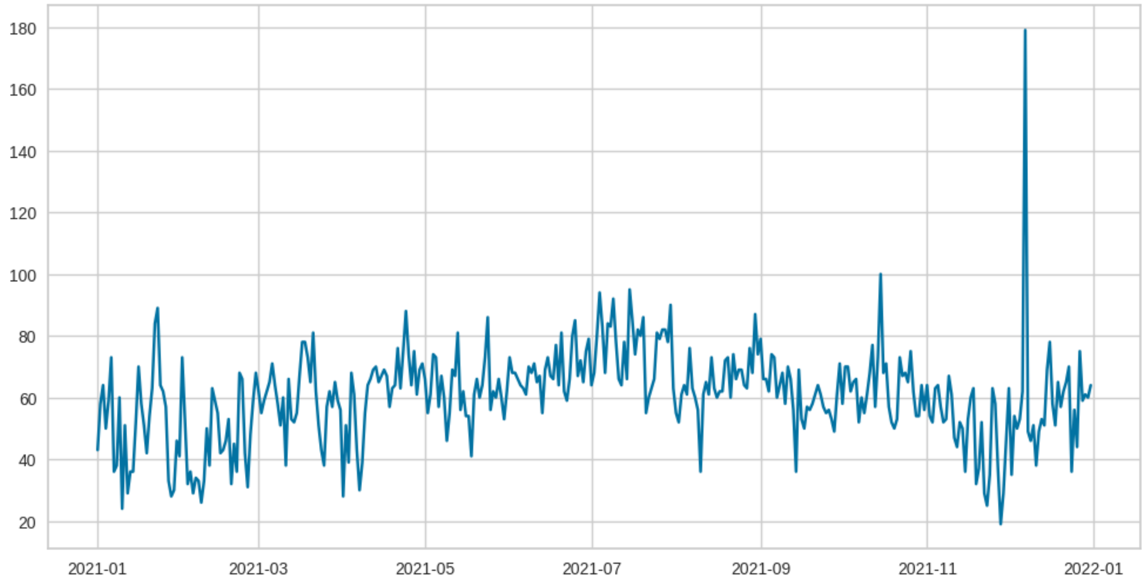


Figure 3. PM10 Timeseries 2021

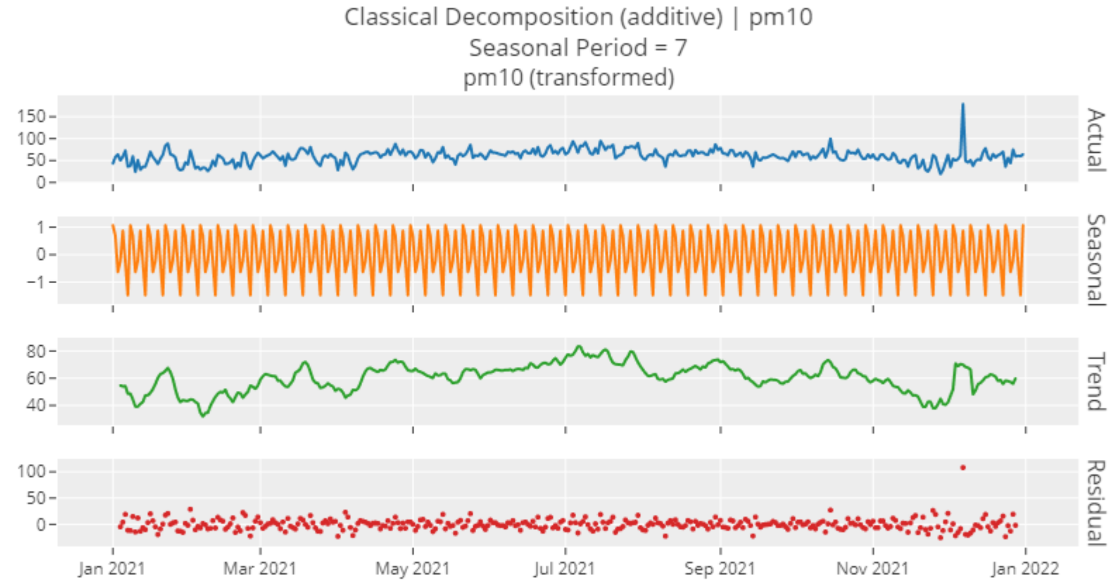


Figure 4. Decomposition Chart

3.2. Models Comparison

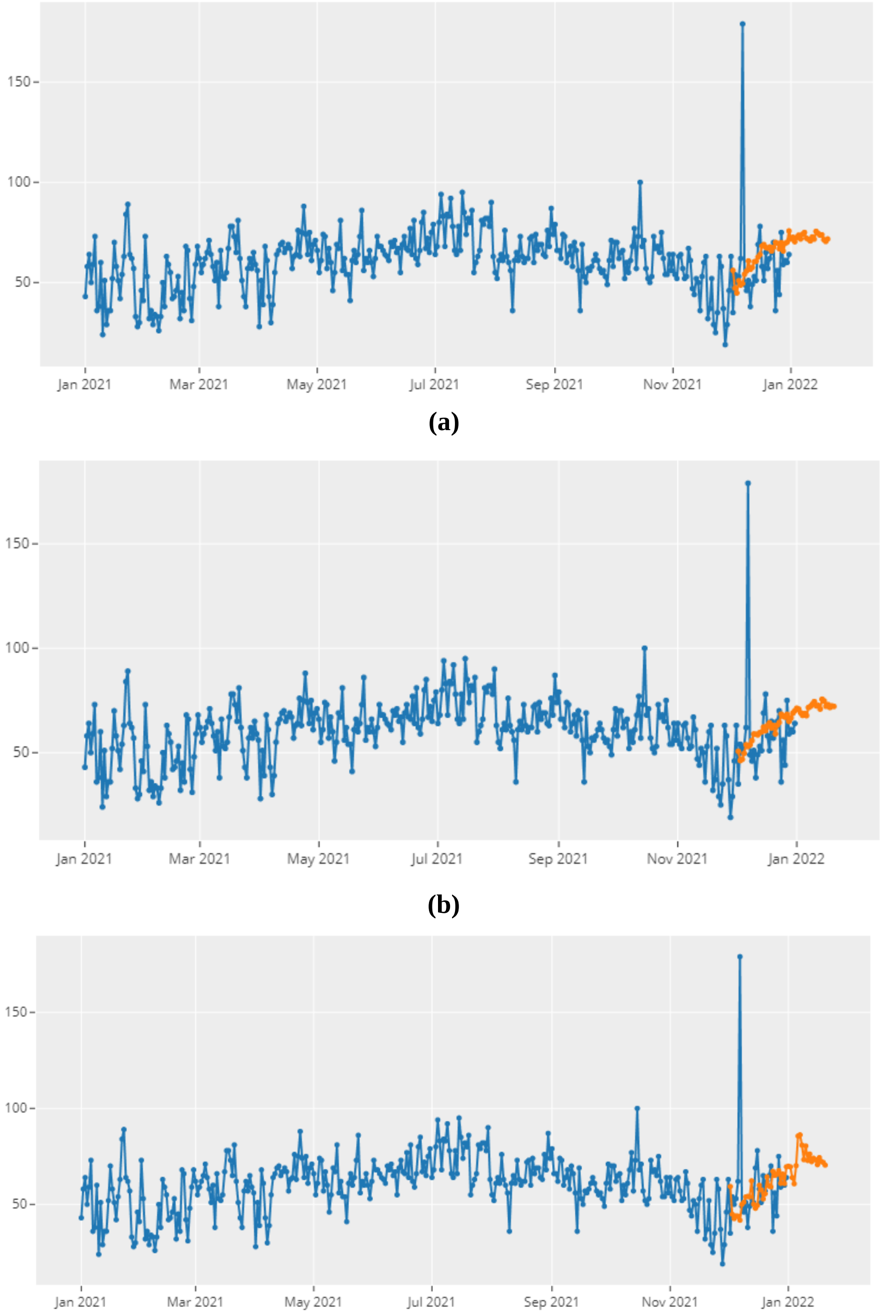
Based on the performance comparison analysis of the three models with the Pycaret library. The smallest error values (marked in blue) of the three models are shown in Table 3. It is evident that among the three decision tree models, Extra Trees demonstrates the best performance, as the analysis results show the smallest error values across all accuracy criteria.

Table 3. The Decision Trees Model from Performance Comparison

No	Models	MASE	RMSSE	MAE	RMSE	MAPE	SMAPE	R^2
1.	Extra Trees	0.8808	0.8113	12.6173	14.7436	0.2433	0.2433	-1.2013
2.	Random Forest	0.9151	0.8485	13.1001	15.4169	0.2505	0.2505	-1.4694
3.	XGBoost	0.9446	0.9106	13.5577	16.6245	0.2622	0.2622	-1.8998

3.3. Models Performance

The three models analyzed for accuracy forecast the data in the next 30 days. Visualization of the predictions can be seen in Figure 5. All three models produce similar graphical patterns. The Extra Trees model predicts values closer to the actual values, the Random Forest predicts slightly larger values and the XGBoost model produces results that appear to fluctuate.



(a)

(b)

(c)

Figure 5. Extra Trees (A) , XGBoost (B), Random Forest (C) (Blue) vs Actual (Orange)

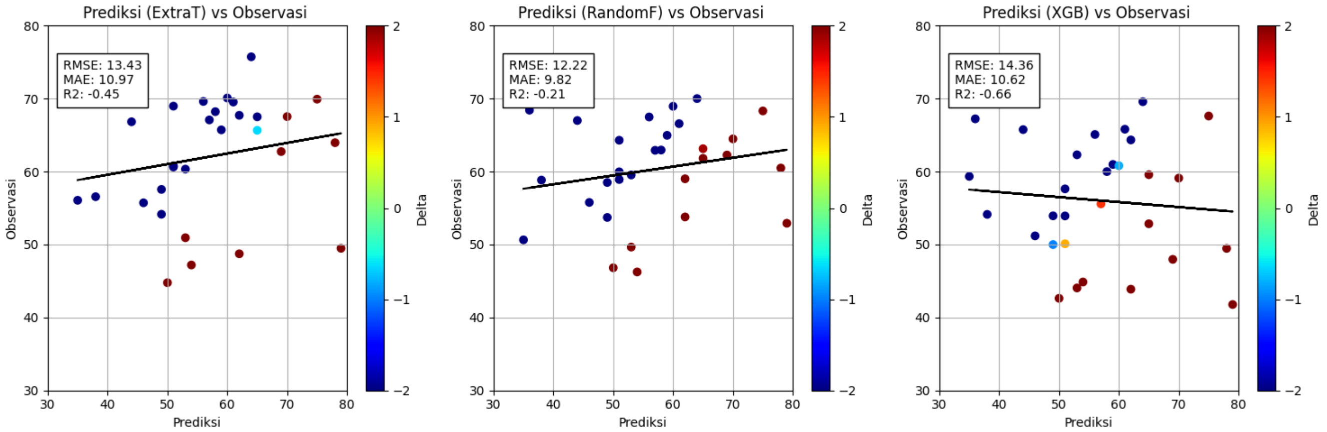


Figure 5. Scatter Plot of Extra Trees , XGBoost, Random Forest

Figure 6 shows the scatter plot of prediction results (abscissa) and observations data (ordinate). The color difference shows how far the prediction value is from the actual value. RMSE, MAE, and R2 are error values between predictions and observations. The three models have slightly different patterns, with the Extra Trees and Random Forest prediction trend showing an increase while the XGBoost model shows a decrease in the prediction trend line.

4. CONCLUSION

The prediction results of the three models show a relatively similar graphical pattern, which is experiencing an increase, but there are higher fluctuations in the XGBoost prediction. According to the performance comparison result of the three models in predicting particulate ISPU data (pm10) by utilizing three Decision tree models, namely Extra Trees, Random Forest, and XGBoost, it is found that the Extra Tree model produces the best performance with MASE 0.8808, RMSSE 0.8113, MAE 12.6173, RMSE 14.7436, MAPE 0.2433, SMAPE 0.207, and R2 -1.2013. It can be concluded that the best model for predicting particulate ISPU data (pm10) is the Extra Trees model.

REFERENCES

[1] S. D. A. Kusumaningtyas, E. Aldrian, M. A. Rahman, and A. Sopaheluwakan, "Aerosol properties in Central Kalimantan due to peatland fire," *Aerosol Air Qual Res*, vol. 16, no. 11, pp. 2757–2767, 2016.

[2] M. Santoso *et al.*, "Assessment of urban air quality in Indonesia," *Aerosol Air Qual Res*, vol. 20, no. 10, pp. 2142–2158, 2020.

[3] M. Brauer *et al.*, "Ambient air pollution exposure estimation for the global burden of disease 2013," *Environ Sci Technol*, vol. 50, no. 1, pp. 79–88, 2016.

[4] E. Kristanti, R. E. Handriyono, M. N. Apsari, and N. R. Abadi, "Evaluasi Monitoring Kualitas Udara Di Pt X (Desa Sedayulawas, Kecamatan Brondong, Kabupaten Lamongan)," in *Prosiding Seminar Teknologi Perencanaan, Perancangan, Lingkungan dan Infrastruktur*, 2021, pp. 406–412. [Online]. Available: http://ejurnal.itats.ac.id/stepplan/article/view/1601

[5] I. Q. A'yun and R. Umaroh, "Polusi Udara dalam Ruangan dan Kondisi Kesehatan: Analisis Rumah Tangga Indonesia," *Jurnal Ekonomi Dan Pembangunan Indonesia*, vol. 22, no. 1, p. 2, 2022, doi: https://doi.org/10.21002/jepi.2022.02.

[6] A. Mosavi, P. Ozturk, and K. Chau, "Flood prediction using machine learning models: Literature review," *Water (Basel)*, vol. 10, no. 11, p. 1536, 2018, doi: https://doi.org/10.3390/w10111536.

[7] A. B. K. Didavi, R. G. Agbokpanzo, and M. Agbomahena, "Comparative study of Decision Tree, Random Forest and XGBoost performance in forecasting the power output of a photovoltaic system," in *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART)*, IEEE, 2021, pp. 1–5. doi: https://doi.org/10.1109/BioSMART54244.2021.9677566.

[8] P. Pothumani and E. S. Reddy, "Original Research Article Network intrusion detection using ensemble weighted voting classifier based honeypot framework," *Journal of Autonomous Intelligence*, vol. 7, no. 3, 2024, doi: https://doi.org/10.32629/jai.v7i3.1081.

[9] Q. C. Doan, C. Chen, S. He, and X. Zhang, "How urban air quality affects land values: Exploring non-linear relationships and its threshold identification using explainable artificial intelligence," *J Clean Prod*, p. 140340, 2023, doi: https://doi.org/10.1016/j.jclepro.2023.140340.

[10] S. Chakraborty and S. Bhattacharya, "Application of XGBoost algorithm as a predictive tool in a CNC turning process," *Reports in Mechanical Engineering*, vol. 2, no. 1, pp. 190–201, 2021, doi: https://doi.org/10.31181/rme2001021901b.

[11] O. Sagi and L. Rokach, "Approximating XGBoost with an interpretable decision tree," *Inf Sci (N Y)*, vol. 572, pp. 522–542, 2021, doi: https://doi.org/10.1016/j.ins.2021.05.055.

[12] A. Ivanov, S. Gocheva-Ilieva, and M. Stoimenova-Minova, "Random forest regression for statistical modeling and forecasting of PM10," in *AIP Conference Proceedings*, AIP Publishing, 2022.

[13] K. Qadeer and M. Jeon, "Prediction of PM10 concentration in South Korea using gradient tree boosting models," in *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*, 2019, pp. 1–6.

[14] S. Babu and B. Thomas, "Daily PM10 prediction of Thiruvananthapuram city and interpretability analysis of influencing factors," *Pollution*, vol. 11, no. 2, pp. 525–537, 2025.

[15] Y. Kristanto, T. Agustin, and F. R. Muhammad, "Pendugaan Karakteristik Awan berdasarkan Data Spektral Citra Satelit Resolusi Spasial Menengah Landsat 8 Oli/Tirs (Studi Kasus: Provinsi Dki Jakarta)," *Jurnal Meteorologi Klimatologi Dan Geofisika*, vol. 4, no. 2, pp. 42–50, 2017, doi: https://doi.org/10.36754/jmkg.v4i2.46.

[16] H. Rachmi, "Klasifikasi Pencemaran Udara Di DKI Jakarta Menggunakan Metode Naïve Bayes," *Jurnal Publikasi Ilmu Komputer dan Multimedia*, vol. 2, no. 2, pp. 86–92, 2023, doi: https://doi.org/10.55606/jupikom.v2i2.2384.

[17] J. W. Simatupang, S. Hamidah, B. Raditya, and F. Hadinegara, "Sistem Monitoring Online Jaringan Sensor Nirkabel: Survei Kualitas Air dan Udara di Daerah Karawang," 2022, doi: https://doi.org/10.32672/jse.v7i2.4210.

[18] A. Lestari, A. Fitrisia, and O. Ofianto, "Metodologi Ilmu Pengetahuan: Kuantitatif Dan Kualitatif Dalam Bentuk Implementasi," *Jurnal Pendidikan Dan Konseling (JPDK)*, vol. 4, no. 6, pp. 8558–8563, 2022, doi: https://doi.org/10.31004/jpdk.v4i6.9710.

[19] F. N. Fajri, A. Tholib, and W. Yuliana, "Application of Machine Learning Algorithm for Determining Elective Courses in Informatics Study Program," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 3, pp. 485–496, 2022, doi: https://doi.org/10.28932/jutisi.v8i3.3990.

[20] M. Vasegh, A. Dehghanbanadaki, and S. Motamedi, "Enhanced soil liquefaction potential estimation using machine learning and web-based platform," 2023, doi: https://doi.org/10.21203/rs.3.rs-2701088/v1.

[21] J. Garcia-Arismendiz, S. Huertas-Zúñiga, C. A. Lizárraga-Portugal, J. C. Quiroz-Flores, and Y. J. Garcia-Lopez, "Improving Demand Forecasting by Implementing Machine Learning in Poultry Production Company," *learning*, vol. 8, p. 9, 2023, doi: https://doi.org/10.14445/22315381/IJETT-V71I2P205.

[22] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Clim Res*, vol. 30, no. 1, pp. 79–82, 2005.

[23] S. Gutmann, C. Maget, M. Spangler, and K. Bogenberger, "Truck parking occupancy prediction: Xgboost-LSTM model fusion," *Frontiers in Future Transportation*, vol. 2, p. 693708, 2021, doi: https://doi.org/10.3389/ffutr.2021.693708.

[24] U. Khair, H. Fahmi, S. Al Hakim, and R. Rahim, "Forecasting error calculation with mean absolute deviation and mean absolute percentage error," in *journal of physics: conference series*, IOP Publishing, 2017, p. 012002.

[25] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Int J Forecast*, vol. 22, no. 4, pp. 679–688, 2006.

[26] P. Goodwin and R. Lawton, "On the asymmetry of the symmetric MAPE," *Int J Forecast*, vol. 15, no. 4, pp. 405–408, 1999.

[27] C. Chen, J. Twycross, and J. M. Garibaldi, "A new accuracy measure based on bounded relative error for time series forecasting," *PLoS One*, vol. 12, no. 3, p. e0174202, 2017, doi: https://doi.org/10.1371/journal.pone.0174202.

[28] N. J. D. Nagelkerke, "A note on a general definition of the coefficient of determination," *Biometrika*, vol. 78, no. 3, pp. 691–692, 1991.

[29] S. R. P. Ariyanto and W. Yustanti, "Prediksi Kenaikan Jabatan Pranata Komputer pada Kementerian X dengan Menggunakan Model Algoritma Klasifikasi Linear Discriminant Analysis (LDA)," *Journal of Emerging Information System and Business Intelligence (JEISBI)*, vol. 4, no. 3, pp. 40–49, 2023, [Online]. Available: https://ejournal.unesa.ac.id/index.php/JEISBI/article/view/54229