ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1548

Implementation of Random Forest Algorithm for Air Quality Classification: A Case Study of DKI Jakarta's Air Quality Index

Mochammad Junus[1]; Vidorova Nurcahyani[2]; Rachmad Saptono[3]; Nurefa Maulana[4]; Indra Lukmana Putra[5]; Zidan Fahreza[6]

[1]Department of Electrical Engineering, State Polytechnic of Malang, Indonesia
[2]Departement Policy Analysit, State Goverment of Batu, Indonesia
[3]Department of Electrical Engineering, State Polytechnic of Malang, Indonesia
[4]Enha Bena Nusantara Ltd, Batu, Indonesia
[5]Departement Accounting, State Polytechnic of Malang, Indonesia
[6]Department of Electrical Engineering, State Polytechnic of Malang, Indonesia

Publication Date: 2025/04/05

Abstract: Air quality monitoring and classification in urban environments present significant challenges for environmental management and public health policy. This study implements an optimized Random Forest (RF) algorithm to classify air quality levels in DKI Jakarta, Indonesia, using the Air Quality Index (AQI) data from 2021. The analysis incorporates six key pollutants: PM10, PM2.5, NO2, SO2, CO, and O3, with data collected from the Environmental Management Agency of DKI Jakarta. The RF model was developed using 5000 decision trees with optimized parameters (mtry=2) and evaluated through stratified sampling with a 70:30 train-test split. The model achieved an exceptional accuracy of 99.09% with a low Out-of-Bag (OOB) error rate of 2.35%. Feature importance analysis revealed that particulate matter (PM2.5 and PM10) were the most influential factors, collectively accounting for 78.70% of the model's decision-making process. The high performance metrics across all air quality categories (Good, Moderate, and Unhealthy) demonstrate the model's reliability in classification tasks. This research provides insights into environmental monitoring and policymaking, presenting a framework adaptable to other urban settings. The findings highlight the crucial role of particulate matter in air quality assessment and suggest targeted strategies for pollution control.. (Abstract)

Keywords: *Air Quality Classification, Random Forest, Machine Learning, Air Quality Index, Environmental Monitoring, Jakarta.*

How to Cite: Mochammad Junus; Vidorova Nurcahyani; Rachmad Saptono; Nurefa Maulana; Indra Lukmana Putra; Zidan Fahreza (2025). Implementation of Random Forest Algorithm for Air Quality Classification: A Case Study of DKI Jakarta's Air Quality Index. *International Journal of Innovative Science and Research Technology*, 10(3), 2169-2173. https://doi.org/10.38124/ijisrt/25mar1548

I. INTRODUCTION

Air pollution remains one of the most pressing environmental challenges in urban areas, particularly in rapidly expanding megacities across Southeast Asia (Zuo et al., 2019). According to the World Health Organization, approximately 99% of the global population breathes air with elevated pollutant levels, with developing nations bearing the most severe consequences (WHO, 2021). Jakarta, Indonesia's capital, faces significant air quality issues driven by rapid urbanization, increasing vehicle emissions, and industrial activities (Amazing Hope Ekeh et al., 2025).

Rapid urbanization, increased industrial activities, and a surge in vehicular emissions have collectively led to deteriorating air quality in the region (Kusuma et al., 2019; , Syuhada et al., 2023). Several studies have documented that particulate matter (PM2.5) and other pollutants in Jakarta frequently exceed national and international safety thresholds, thereby posing serious health risks to residents (Zulfikri, 2023). The Air Quality Index (AQI) in Jakarta has shown concerning trends, with frequent recordings of unhealthy air quality levels affecting millions of residents (Syuhada et al., 2023).

Given these challenges, enhancing the accuracy and efficiency of air quality monitoring systems is essential for timely policy-making and effective mitigation strategies. Traditional approaches to air quality monitoring and classification often lack the predictive capabilities necessary

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1548

for effective environmental management and public health protection.

Recent advancements in machine learning have offered promising alternatives to traditional statistical methods for environmental data analysis. Machine learning techniques, particularly Random Forest (RF) algorithms, have emerged as powerful tools for environmental data analysis and classification (Beucler et al., 2024). Random Forest has gained significant attention in environmental monitoring due to its ability to handle non-linear relationships, manage high-dimensional data, and provide robust predictions while accounting for variable importance (Amazing Hope Ekeh et al., 2025).

Unlike other black-box methods, Random Forest provides insights into the significance of different predictor variables and supports a more interpretable decision-making process (Idroes et al., 2023). The algorithm's ensemble nature, which integrates multiple decision trees, allows it to effectively capture complex non-linear relationships between meteorological conditions and pollutant concentrations (Jayadi et al., 2024; , Azies, 2023.). This capability is particularly important in urban environments like DKI Jakarta, where multiple factors interact dynamically to influence pollutant levels.

Recent studies have demonstrated the algorithm's effectiveness in air quality prediction and classification across various urban contexts. For example, Natarajan et al. (2024) achieved 95% accuracy in classifying air quality in Delhi, while Rakholia et al. (2024) successfully implemented RF for real-time air quality monitoring in Mexico City.

Despite numerous studies on air quality prediction using various machine learning models, there is still a relative paucity of research applying the Random Forest algorithm to classify the Air Quality Index (AQI) specifically for Jakarta. Prior works have applied techniques such as neural networks and support vector machines; however, they often overlook the advantages of Random Forest in managing imbalanced datasets and providing feature importance analysis (V. Vu et al., 2019). Moreover, investigations into the spatial-temporal variability of air pollutants in Jakarta underscore the need for a more adaptable model that can integrate diverse data sources and yield robust performance in the face of environmental uncertainties (Idroes et al., 2023; Azies, 2023).

This study attempts to fill this gap by developing a Random Forest-based classification from AQI (2021 data) in Jakarta. The research investigates six major pollutants, including: Particulate Matter (PM10 and PM2.5), Nitrogen Dioxide (NO2), Sulfur Dioxide (SO2), Carbon Monoxide (CO), and Ozone (O3). Our approach builds upon previous work by incorporating comprehensive variable importance analysis and optimizing model parameters for Jakarta's specific context.

By addressing the classification of AQI using Random Forest, this research not only advances methodological approaches in air quality analysis but also provides an important tool for local governments and stakeholders. The insights derived from this study will support the design of targeted pollution control policies and eventual improvements in public health outcomes, reaffirming the critical role of machine learning in environmental science (Jayadi et al., 2024; Vu et al., 2019; Azies, 2023).

II. RESEARCH METHOD

A. *Data Collection and Description*

This study utilized air quality monitoring data from DKI Jakarta collected throughout 2021. The dataset comprises 365 daily observations obtained from the Environmental Management Agency of DKI Jakarta. Six air pollutant parameters were measured according to the:

➢ *Input Variables (Air Pollutants):*

- PM10:Particulate matter with diameter ≤ 10 micrometers (μm)
- PM2.5:Fine particulate matter with diameter ≤ 2.5 μm
- NO2: Nitrogen dioxide
- SO2: Sulfur dioxide
- CO: Carbon monoxide
- O3: Ozone

➢ *Output Variable:*

Air quality categories according to AQI standards:

- Good (0-50)
- Moderate (51-100)
- Unhealthy for Sensitive Groups (101-200).
- Unhealthy (151-200)
- Very Unhealthy (201-300)
- Hazardous (301 and higher).

B. *Random Forest Model Development*

➢ *Bootstrap Sampling*

- Each tree uses approximately 2/3 of training data
- Remaining 1/3 used for OOB error estimation

➢ *Node Splitting:*

- √p features randomly selected at each node (where p=6)
- Gini impurity used as splitting criterion
- Minimum samples per leaf = 1

➢ *Hyperparameter Selection Key Parameters were Chosen based on Literature Recommendations:*

- n_estimators: 5000 trees for stable performance
- max_features: 2 (≈√6) following Breiman's recommendation
- class_weight: 'balanced' to handle class imbalances.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1548

III. RESULTS AND DISCUSSION

➢ *Model Performance Analysis*

The Random Forest classifier demonstrated exceptional performance in categorizing air quality levels. Table 1 presents the confusion matrix showing the classification results across different air quality categories.

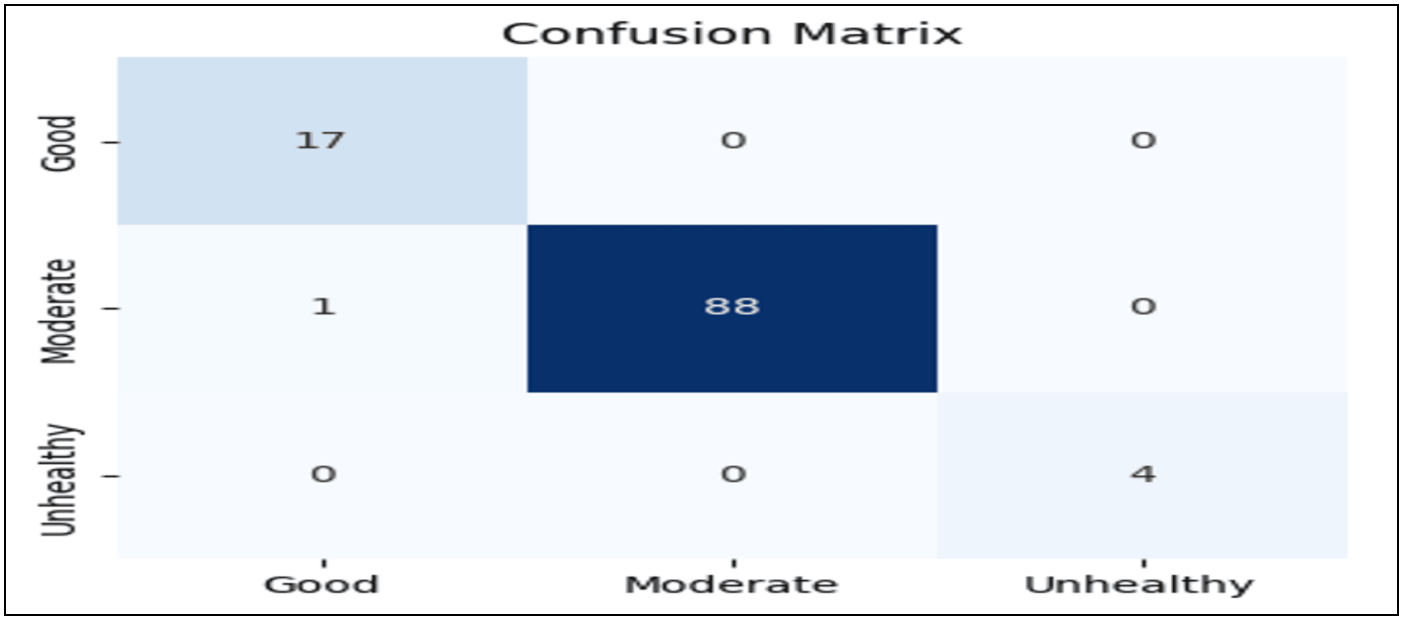

Fig 1 Confusion Matrix of Air Quality Classification

The model achieved an overall accuracy of 99.09% on the test dataset, with an Out-of-Bag (OOB) error rate of 2.35%. Based on findings from similar studies, as reported by (Shaziayani et al., 2022), the Random Forest model achieved an accuracy of 98.37% in classifying air quality in urban environments.The detailed performance metrics for each category are presented in Table 1:

Table 1 Classification Performance Metrics by Category

Category	Precision	Recall	F1-score	Support
Good	0.94	1.00	0.97	17
Moderate	1.00	0.99	0.99	89
Unhealthy	1.00	1.00	1.00	4

➢ *Variable Importance Analysis*

One of the key advantages of Random Forest is its ability to quantify the relative importance of input variables. Figure 2 illustrates the contribution of each pollutant to the classification model:

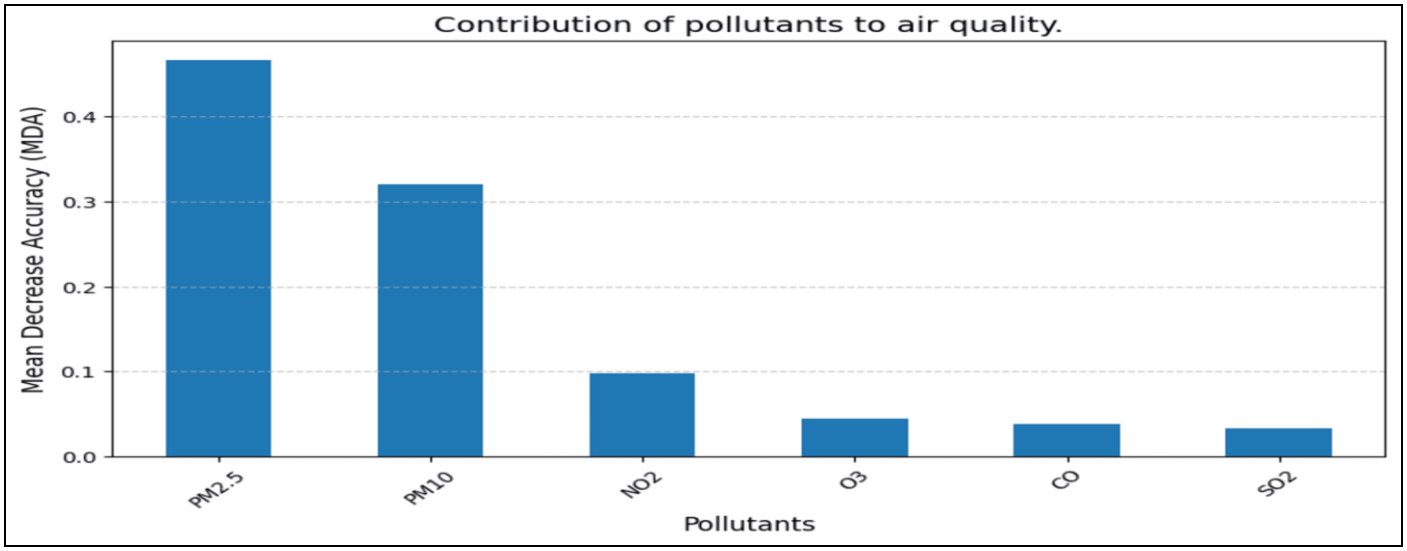

Fig 2 Contribution of each pollutant to air quality classification

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1548

➢ *The Analysis Revealed the Following Hierarchy of Pollutant Importance:*

- PM2.5 (46.62%)
- PM10 (32.08%)
- NO2 (9.77%)
- O3 (4.46%)
- CO (3.76%)
- SO2 (3.32%)

The dominance of particulate matter (PM2.5 and PM10) in the model's decision-making process aligns with findings from recent studies in other Asian megacities (Beucler et al., 2024). This result is particularly significant given that PM2.5 and PM10 are considered the most harmful pollutants to human health (WHO, 2021).

➢ *Model Robustness and Limitations*

While the model demonstrates high accuracy, several considerations should be noted:

- Class Imbalance: The dataset shows an uneven distribution of categories, with moderate conditions being predominant. This was addressed through the use of balanced class weights in the model.
- Spatial Limitations: The current model relies on aggregated data for DKI Jakarta and may not capture localized variations in air quality across different city districts.

IV. CONCLUSIONS

This study successfully implemented a Random Forest algorithm for air quality classification in DKI Jakarta using 2021 monitoring data. The key findings and implications are as follows:

➢ *Model Performance*

- The Random Forest classifier achieved an exceptional accuracy of 99.09%
- The low Out-of-Bag error rate of 2.35% demonstrates the model's robustness
- High precision and recall values across all air quality categories indicate reliable classification performance.

➢ *Pollutant Importance*

- PM2.5 and PM10 emerged as the most influential pollutants, collectively accounting for 78.70% of the model's decision-making process
- Secondary contributions came from NO2 (9.77%) and O3 (4.46%)
- CO and SO2 showed relatively minor influences on air quality classification

REFERENCES

[1]. Amazing Hope Ekeh, Charles Elachi Apeh, Chinekwu Somtochukwu Odionu, & Blessing Austin-Gabriel. (2025). Leveraging machine learning for environmental policy innovation: Advances in Data Analytics to address urban and ecological challenges. *Gulf Journal of Advance Business Research*, *3*(2), 456–482. https://doi.org/10.51594/gjabr.v3i2.92

[2]. Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., Yu, S., Rasp, S., Ahmed, F., O'gorman, P. A., Neelin, J. D., Lutsko, N. J., & Pritchard, M. (2024). Climate-invariant machine learning. In *Sci. Adv* (Vol. 10). https://www.science.org

[3]. Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-54807-1

[4]. Rakholia, R., Le, Q., Vu, K., Ho, B. Q., & Carbajo, R. S. (2024). Accurate PM2.5 urban air pollution forecasting using multivariate ensemble learning Accounting for evolving target distributions. *Chemosphere*, *364*. https://doi.org/10.1016/j.chemosphere.2024.143097

[5]. Shaziayani, W. N., Ul-Saufie, A. Z., Mutalib, S., Mohamad Noor, N., & Zainordin, N. S. (2022). Classification Prediction of PM10 Concentration Using a Tree-Based Machine Learning Approach. *Atmosphere*, *13*(4). https://doi.org/10.3390/atmos13040538

[6]. Syuhada, G., Akbar, A., Hardiawan, D., Pun, V., Darmawan, A., Heryati, S. H. A., Siregar, A. Y. M., Kusuma, R. R., Driejana, R., Ingole, V., Kass, D., & Mehta, S. (2023). Impacts of Air Pollution on Health and Cost of Illness in Jakarta, Indonesia. *International Journal of Environmental Research and Public Health*, *20*(4). https://doi.org/10.3390/ijerph20042916

[7]. WHO. (2021, September 22). *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*.

[8]. Zuo, X., Yang, X., Dou, Z., & Wen, J. R. (2019). RUCIR at TREC 2019: Conversational Assistance Track. *28th Text REtrieval Conference, TREC 2019 - Proceedings*. https://doi.org/10.1145/1122445.1122456

[9]. Azies, H. A. (n.d.). *Air Pollution in Jakarta, Indonesia Under Spotlight: An AI-Assisted Semi-Supervised Learning Approach.*

[10]. Idroes, G. M., Noviandy, T. R., Maulana, A., Zahriah, Z., Suhendrayatna, S., Suhartono, E., Khairan, K., Kusumo, F., Helwani, Z., & Abd Rahman, S. (2023). Urban Air Quality Classification Using Machine Learning Approach to Enhance Environmental Monitoring. *Leuser Journal of Environmental Studies*, *1*(2), 62–68. https://doi.org/10.60084/ljes.v1i2.99

[11]. Jayadi, B. V., Lauro, M. D., Rusdi, Z., Handhayani, T., & Informasi, F. T. (n.d.). *Sistemasi: Jurnal Sistem Informasi Klasifikasi Indeks Standar Pencemaran Udara untuk Data Tidak Seimbang menggunakan Pendekatan Pembelajaran Mesin Air Quality Index Classification for Imbalanced Data Using Machine Learning Approach*. http://sistemasi.ftik.unisi.ac.id

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1548

[12]. Kusuma, W. L., Chih-Da, W., Yu-Ting, Z., Hapsari, H. H., & Muhamad, J. L. (2019). Pm2.5 pollutant in asia—a comparison of metropolis cities in indonesia and taiwan. *International Journal of Environmental Research and Public Health*, *16*(24). https://doi.org/10.3390/ijerph16244924

[13]. Syuhada, G., Akbar, A., Hardiawan, D., Pun, V., Darmawan, A., Heryati, S. H. A., Siregar, A. Y. M., Kusuma, R. R., Driejana, R., Ingole, V., Kass, D., & Mehta, S. (2023). Impacts of Air Pollution on Health and Cost of Illness in Jakarta, Indonesia. *International Journal of Environmental Research and Public Health*, *20*(4). https://doi.org/10.3390/ijerph20042916

[14]. V. Vu, T., Shi, Z., Cheng, J., Zhang, Q., He, K., Wang, S., & M. Harrison, R. (2019). Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. *Atmospheric Chemistry and Physics*, *19*(17), 11303–11314. https://doi.org/10.5194/acp-19-11303-2019

[15]. Zulfikri, A. (2023). Effects of Pollution and Transportation on Public Health in Jakarta. In *West Science Interdisciplinary Studies* (Vol. 1, Issue 04).