ICDSOS
International Conference
on Data Science and Official Statistics

“Harnessing Innovation in Data Science and Official Statistics to Address Global Challenges towards the Sustainable Development Goals”

Air Pollution in Jakarta, Indonesia Under Spotlight: An AI-Assisted Semi-Supervised Learning Approach

H A Azies[1,2,*]

[1]Study Program in Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, 50131, Semarang, Indonesia
[2]Research Center for Materials Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, 50131, Semarang, Indonesia

*Corresponding author’s e-mail: harun.alazies@dsn.dinus.ac.id

Abstract. The air quality in the Jakarta area is examined in this study using artificial intelligence (AI) to assist a semi-supervised learning technique. The clustering approach is used in this article to separate air pollution into three main categories moderate, low, and high levels. This clustering helps identify shared characteristics among measures like particulates (PM10 and PM2.5), sulfur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), and ozone (O3), even when air quality labels are not always accessible. Using the Random Forest method, the air quality will be categorized in this experiment with an accuracy rate of 93%. Additionally, the results of variable significance analysis are examined on this article to identify the variables with the biggest effects on air quality, notably PM10, SO2, and NO2. This study demonstrates the enormous potential of applying machine learning techniques, particularly semi-supervised learning approaches, to assist sustainable environmental regulations while also monitoring and enhancing Jakarta's air quality. We describe the experimental procedures, the findings, and the implications of our research for comprehending and addressing urban air pollution in this article.

1. Introduction

Deteriorating air quality has become one of the most significant environmental issues in major cities, including Jakarta, Indonesia's capital city. Rapid population growth, the inevitability of urbanization, and high mobility have resulted in increased emissions of dangerous air pollutants such as PM2.5 particles, NO2, SO2, and O3 [1]–[3]. This air pollution has a negative influence on human health, the environment, and the economy of the city [4]. A previous study in this area attempted to analyze Jakarta's air pollution using several methodologies [5]–[8]. However, there is still a knowledge vacuum that must be filled to have a better understanding of the patterns and causes that drive air pollution in increasingly complex metropolitan contexts. Seeing the need for a deeper understanding, this study provides a novel technique for assessing air pollution in Jakarta by employing artificial intelligence (AI) technology, specifically semi-supervised machine learning methods.

As a result, this study will investigate the problem of air pollution in Jakarta in depth utilizing AI methodology and semi-supervised learning methods. AI models can use labeling data intelligently thanks to semi-supervised learning approaches [9]–[11]. Data from the previous iteration's labeling by the model can be utilized to train the model in the next iteration [12]–[14]. As a result, the model can learn from its prediction results and gradually increase its capacity to accurately diagnose air pollution. The application of the semi-supervised learning method to air pollution analysis is innovative in this study. This method is likely to produce more precise and detailed results for determining Jakarta's air pollution trends.

The aim of this study is to create a model that can properly and efficiently assess air pollution data and provide a greater understanding of air pollution patterns in Jakarta. This research intends to provide a solution to Jakarta's air pollution problems by combining AI technology and semi-supervised learning approaches. The findings of this study are expected to be valuable to policymakers in establishing more effective and sustainable air pollution management programs. Furthermore, the application of AI technology to address environmental challenges is an excellent example of how innovation may be leveraged to more intelligently respond to global issues.

2. Related Work

Several studies on air pollution employing a machine learning approach are pertinent to this subject. Kaya et all, for example, used short-term memory (LSTM) machine learning to predict air quality in Jakarta during the COVID-19 outbreak. They believed that the large-scale social restrictions (PSBB) imposed during the epidemic resulted in a significant reduction in Jakarta's air pollution. The root mean square error (RMSE) is utilized to evaluate the LSTM model in this study, and the results suggest that the Adam optimizer can bring the prediction results closer to the dataset used [15]. In addition, Marviola Hardini and colleagues proposed the use of convolutional neural networks (CNN) and image-based machine learning to estimate air quality. They took feature information from landscape photographs to evaluate air quality levels. Data from a network of air quality sensors throughout the city is used in this study. The results show that this approach can provide accurate predictions of air quality compared to traditional methods [16]. Furthermore, research by Wan Yun Hong and his team focuses on the statistical analysis and prediction of air pollution in Labuan, Malaysia. They used exponential triple smoothing (ETS) and seasonal autoregressive integrated moving average (SARIMA) forecasting methods to analyze and predict various air pollutants and the air pollution index (API). These models are used for various air pollutants such as PM10, CO, SO2, NO2, and O3. The results show that the ETS and SARIMA models can provide accurate estimates of air pollutant concentrations in Labuan [17].

Andri and his colleagues' study, on the other hand, highlight the importance of coping with missing data and class imbalances in data analysis and machine learning. They suggested and tested a preprocessing approach that incorporates Multiple Imputation by Chained Equations (MICE) and Synthetic Minority Oversampling Technique (SMOTE), as well as three machine learning algorithms, including Random Forest, Support Vector Machine, and K-Nearest Neighbour. The results show that the g-mean measure is getting better at dealing with class imbalance and missing values in air pollution datasets [18]. Finally, Suhartono et all created a hybrid model that predicts PM10 in Surabaya, Indonesia, by combining Time Series Regression (TSR) as a statistical method and Feedforward Neural Network (FFNN) or Long Short-Term Memory (LSTM) as machine learning. This study illustrates the differences between these models and demonstrates that PM10 in Surabaya has a nonlinear pattern. This demonstrates that combining TSR and FFNN or LSTM can produce better predicts than standalone models[19]. The primary difference in this study is that it focuses on monitoring and assessing air quality in Jakarta areas using a semi-supervised learning approach that allows categorizing air pollution data into groups based on similarities. This is a significant contribution to monitoring and improving Jakarta's air quality, and it demonstrates the enormous potential of employing machine learning approaches to support long-term environmental strategies.

3. Materials and Methodology

The main aim of this study is to analyze Jakarta's air pollution data using Open Data Jakarta information for the year 2021. This dataset contains 1517 data points gathered from five air quality monitoring stations (SPKU) in the DKI Jakarta Province. Six major variables are employed as features in this analysis:

a. **Particulates (PM10 and PM2.5):** PM10 particulates are tiny particles with a diameter of fewer than 10 micrometers, while PM2.5 particulates have a diameter of fewer than 2.5 micrometers.

These particulates can come from a variety of sources, including burning fuel, road dust, industry, and other pollution. They can harm human health since they can be breathed and enter the lungs.

b. **Carbon Monoxide (CO):** Carbon monoxide is a poisonous gas created by the combustion of fossil fuels such as petrol, oil, and natural gas. High CO exposure can impair the ability of the blood to carry oxygen, which can have a severe influence on human health.

c. **Sulfur Dioxide (SO2):** Sulphur dioxide is a gas that is created when sulfur-containing fuels, such as coal and oil, are burned. SO2 can irritate the eyes, nose, and throat, as well as contribute to respiratory difficulties.

d. **Nitrogen Dioxide (NO2):** Nitrogen dioxide is a gas created by the combustion of gasoline in automobiles and power plants. NO2 exposure can harm the human respiratory tract and aggravate chronic respiratory illnesses.

e. **Ozone (O3):** Ozone is a gas that forms a protective layer in the atmosphere, but it can be air pollution at the Earth's surface. Surface ozone is a component of air pollution that can irritate the respiratory system and have negative health effects on humans and the environment.

Figure 1 depicts the overall structure of this study approach. Data preprocessing was done in the early stages of this research, which included data normalization to offset discrepancies in variable scales and screening for missing data [20], [21]. Outliers are also discovered and handled to ensure data integrity[20], [22]. In this study, semi-supervised learning is used from unlabeled data (Step 1). The data is separated into two sections: training data and pseudo-labelling data. The data will be divided into multiple clusters based on feature similarity using the clustering technique. The K-Medoids approach applies a semi-supervised learning strategy during the clustering stage. The semi-supervised learning method allows us to recognize patterns and structures that traditional grouping methods cannot. The silhouette score approach is used to estimate the optimal number of clusters [23]. The K-Medoids technique, which was used in this study, is a clustering algorithm that identifies more stable cluster centers based on actual data points in the cluster and is less sensitive to outliers or extreme data points [24], [25].

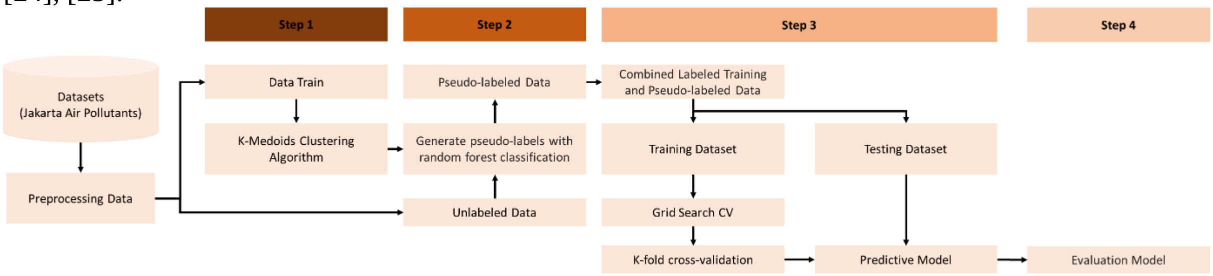

Figure 1. Semi-supervised learning research framework for Jakarta air pollution

Step 2 involves training the model using the Random Forest technique, with the cluster outputs acting as labels (targets). The pseudo-labeling strategy is used to predict labels on the pseudo-dataset based on the outcomes of the trained model constructed using the Random Forest method (Step 2)[26]. Pseudo-labelling (or pseudo-supervised learning) is a machine learning strategy that uses unlabeled or less reliable labels to increase model performance [27]. Clustered data with labels and pseudo-label data are integrated into one data frame. The Random Forest model is then changed using a combination of data, specifically data with labels (cluster discoveries in stage 1) and data with pseudo-labels (in stage 2). This is done in Step 3 to allow the model to learn on data with pseudo-labels. This study's classification model was developed utilizing the semi-supervised learning principle, which employs data clustering to uncover cluster features and improve classification performance. The model evaluation stage is carried out after the classification model has been trained to measure the model's performance in classifying air quality[28]. This assessment includes parameters such as accuracy, recall, and F1-score[29]. To confirm the model's reliability, it is also validated using the cross-validation approach. Furthermore, the results of the data clustering analysis are used to understand the features of each cluster. This provides a more in-depth understanding of the degree of air pollution in various areas in Jakarta. This research's final results include a better understanding of Jakarta's air quality, the contribution of semi-supervised learning methods with artificial intelligence support, and the potential use of analysis results to improve air quality and more sustainable environmental policies.

In this study, we address a typical problem in the field of artificial intelligence: the constraints of labeled data. The use of the semi-supervised learning idea in this study has a solid foundation. Fully labeled data is sometimes limited, especially in the context of air quality study, and requires a large effort to obtain and analyze. As a result, we opted to use increasingly plentiful but unlabeled data as a valuable resource. We may use the semi-supervised learning approach to efficiently use current data, even unlabeled data, to develop models that can deliver more accurate results in predicting air pollution levels depending on certain factors. In this sense, we understand the tremendous potential for using unlabeled data to increase the accuracy and relevance of our study findings. This approach allows us to create higher-quality and more relevant information in air quality analysis, which can help society and stakeholders concerned about environmental issues.

4. Result and Discussion

4.1. The Cluster Algorithm's Experimental Results

Semi-supervised learning is a machine learning approach in which models are trained to utilize both labelled and unlabelled data[12]. In the context of air pollution data clustering, the initial step is to cluster data and group data into groups with comparable characteristics. The silhouette approach is one technique for determining the ideal number of clusters[23], [30].

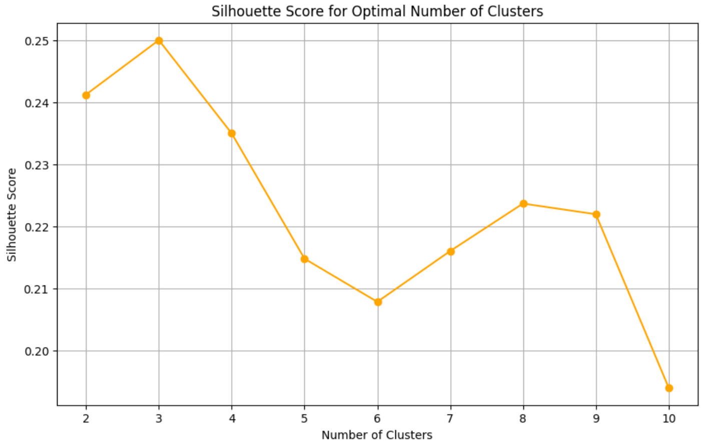

Figure 2. Silhouette Score for Optimal Number of Clusters

The silhouette technique (or silhouette score) is an evaluation metric used to determine how similar each data point in one cluster is to data points in other clusters in a data cluster [23]. This method is used in clustering to identify the appropriate number of clusters. The silhouette score, which runs from -1 to 1, is the result of this procedure. A good score shows that entities in one cluster share comparable traits and are well separated from entities in other clusters, whereas a negative score suggests the inverse. Based on the experimental data, this study generated a graph (Figure 2) displaying the silhouette values for various cluster counts. The findings of this experiment reveal that the number of clusters with the maximum silhouette value is three. In other words, air pollution data is best classified into three groups based on similarities. This optimal number of clusters serves as the foundation for the following steps in this investigation. In the context of this research, clustering experiments use the k-medoid method [25]. K-medoids is a clustering approach similar to K-means but with a significant change in the selection of cluster centres[31]. K-medoids select the cluster centre as the arithmetic average of all data points in the cluster, whereas K-means selects the cluster centre as the medoids or representative of the cluster [24]. K-medoids are employed in this study to organize air pollution data into clusters with comparable features, assisting in the first understanding of the patterns that exist in the data. According to the findings of this study, the best number of clusters for air pollution data is three. Cluster composition describes the amount of people in each group.

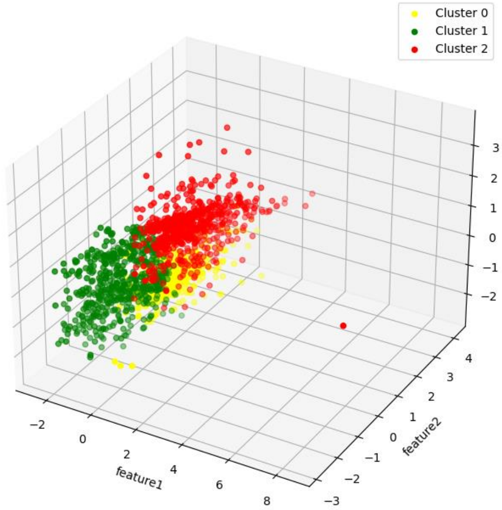

Figure 3. The composition of the members of each cluster as a result of the k-medoids clustering algorithm

The study results (Figure 3) demonstrate that Cluster 0 has 365 members, Cluster 1 has 530 members, and Cluster 2 has 622 members. For this study, the next phase in data analysis is cluster profiling. Based on the characteristics of air pollution data, this study identified three distinct clusters. The researcher will next delve deeper into each cluster, attempting to comprehend the distinct characteristics of each group. The distinctive qualities of each group are used to give cluster names based on available data. Table 1 depicts the characteristics of each cluster based on pollution chemicals.

Table 1. Characteristics of each cluster resulting from the k-medoids clustering algorithm

pm10	pm25	so2	co	o3	no2	Cluster
38.292	56.942	29.906	9.092	27.451	13.687	Low Pollution
56.847	83.912	27.181	16.008	25.088	31.474	Moderate Pollution
62.970	93.056	45.687	11.916	38.318	19.679	High Pollution

Clustering analysis using an unsupervised learning method based on k-medoids is used to identify air pollution categories based on data similarities that are not always obvious or easy to interpret when looking at raw data. This approach enables objective data grouping based on similar air pollution parameters. Identifying groups according to characteristics such as "Low Pollution,", " Moderate Pollution," and "High Pollution" is a good approach because it provides a clearer and more understandable understanding for those without a strong scientific background, while also reflecting relevant information about the level of air pollution in each group.

Moderate Pollution: This cluster is named "Moderate Pollution" because its degree of air pollution is in the center of the other two. We detect moderate amounts of PM10 and PM2.5 within this cluster, indicating the presence of solid contaminants in the air. The amounts of SO2, CO, O3, and NO2 are also not too high or low. As a result, we consider it to be of moderate contamination.

Low Pollution: The second cluster is designated as "Low Pollution" due to its low amount of air pollution. There is clear evidence that PM10, PM2.5, SO2, and CO levels are lower in this cluster than in others. This means that solid and gaseous pollution levels in the air are fairly low. As a result, we consider it to have a low degree of air pollution.

High Pollution: The third cluster is designated as "High Pollution" due to significant levels of air pollution. This cluster has elevated levels of PM10, PM2.5, SO2, CO, O3, and NO2. This shows that there are a lot of solid and gaseous contaminants in the air, which means there's a lot of pollution.

Aside from that, the semi-supervised learning concept proposed in this paper is a critical strategy with the potential to yield significant advances. An efficient technique to use existing data is to start with unlabeled data and label it with clustering results. Clustering data will be used as labels in the development of a supervised learning model with cluster naming as the target label. The constructed model is used to predict pseudo-labels on labeled data. Based on the clustering results, this will aid in the construction of models that can forecast or categorize new data into appropriate groupings.

4.2. The Classification Algorithm's Experimental Results

The labels generated by clustering are critical in the next stage, which is air quality classification. Based on the cluster labels, the data is divided into three groups. This stage is critical because it enables the development of a more precise categorization model that takes into consideration the unique characteristics of each location. The machine learning model is trained in the classification stage utilizing training data that already includes cluster labels as targets. This is a semi-supervised strategy in which some data has labels while others do not. This classification machine will learn patterns in the data, including actual levels of air pollution, and will subsequently be able to predict future data. This study used a classification system known as Random Forest to carry out the air quality predict step in various regions of Jakarta. Random forest is an extremely effective machine learning algorithm that can handle a variety of categorization obstacles [32]. Random Forest employs several randomly generated decision trees [33]. Each of these trees is a model that learns from training data and can predict air quality based on the properties it has discovered. The Random Forest's key advantage is its ability to avoid overfitting, which occurs when the model "memorizes" the training data and cannot generalize successfully to new data[34], [35].

The first stage of this research is hyperparameter tuning to optimize the performance of the Random Forest model[36]. This stage is critical because ideal hyperparameters enable the model to make accurate predictions. This study creates a parameter grid that accepts multiple values for the four primary hyperparameters in the Random Forest model. These combinations will be investigated in the search for hyperparameters, which will also include *n* estimators or the number of decision trees in the model. The researchers experimented with values of 100, 200, and 300[37]. The following parameter is max_depth, which is the maximum depth of each decision tree with alternatives such as 0, 10, and 20 [38]. Min_samples_split is the smallest number of samples required to divide the tree's nodes[39]. The values under consideration are 2, 5, and 10. Finally, the min_samples_leaf parameter specifies the minimum amounts of samples necessary in each tree leaf[40]. There are three possible combinations: 1, 2, and 4. GridSearchCV will then attempt all possible hyperparameter combinations from the provided grid[41], [42]. The hyperparameter search yields hyperparameter combinations with the highest accuracy score. The researcher in this example has determined the ten greatest choices based on the highest accuracy ratings.

Table 2. The results of the hyperparameter combination experiment

Experiment	max_depth	min_samples_leaf	min_samples_split	n_estimators	Accuracy
31[a]	10	1	5	200	0.917624
34	10	1	10	200	0.916973
41	10	2	5	300	0.916966
29	10	1	2	300	0.916966
4	0	1	5	200	0.916964
58	20	1	5	200	0.916964
32	10	1	5	300	0.916308
42	10	2	10	100	0.916304
3	0	1	5	100	0.916304
57	20	1	5	100	0.916304

[a] Experiment with best performance

Table 2 displays the experimental findings of numerous hyperparameter combinations that were evaluated, as well as the accuracy gained for each combination. The first hyperparameter combination, which is the result of the 31st experiment in the list of experimental results, has the maximum accuracy of 0.9176 with a max depth of 10, min samples of leaf 1, min samples of split 5, and n_estimators of 200. As a result, this hyperparameter combination was selected as the best and will be used to train an air quality classification model in Jakarta cities. An air quality classification model will be trained using the set of hyperparameters that was determined to be the best one in Jakarta. Additionally, the 5-fold technique and cross-validation will be used to test this model[43]. Five distinct subsets of the data will be used in this cross-validation process. The other four subsets will be utilized as training data, and each subset will be used as testing data alternatively. Five repetitions will be needed to complete this process, resulting in five separate subsets of test results. The accuracy outcomes from each iteration will be used to gauge how well and consistently the trained model can generate predictions.

Table 3. Results of the cross-validation process using the K-fold method

Fold	Accuracy
Fold 1	0.9275
Fold 2	0.9126
Fold 3	0.9176
Fold 4	0.9152
Fold 5	0.9086

The outcome of the 5-fold cross-validation process is shown in Table 3. These findings show that the trained classification algorithm consistently predicts air quality in Jakarta areas. The average precision of these five folds is approximately 91.63%, demonstrating the model's effectiveness in categorising air quality. The fold with the highest accuracy, according to the findings of cross-validation with five folds, is "Fold 1," with a precision of 0.9275. As a result, "Fold 1" was determined to be the best fold and will be assessed using test data. The model that was developed using "Fold 1" will be regarded as the ultimate model that will be applied to predict air quality during testing.

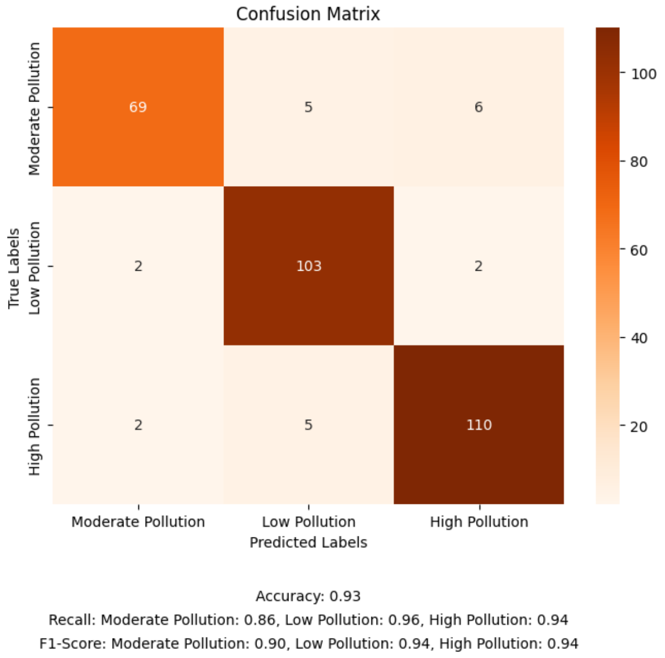

Figure 4. Confusion matrix of model prediction results using the random forest algorithm

The confusion matrix in Figure 4 is used to evaluate the classification model's performance on test data. This confusion matrix is divided into three categories or labels: "moderate pollution," "low pollution," and "high pollution." The model correctly predicted 69 of the 69 samples in the "Moderate Pollution" category. This demonstrates that the model can correctly identify instances of this form of air pollution. However, six samples were incorrectly labelled as "high pollution" and five samples as "low pollution." This shows that adding "moderate pollution" as a new category was a mistake. The program identified 103 samples in the "Low Pollution" category correctly. These results demonstrate the model's ability to recognize clean air pollution conditions. Two samples were incorrectly labelled as "moderate pollution" and two samples as "high pollution." This implies a minor misunderstanding of the term "low pollution." The program correctly predicted the presence of 110 samples in the "high pollution" category. This proves that the model can identify extraordinarily high amounts of air pollution. There were just two samples that were incorrectly labelled "moderate pollution" and five samples that were labelled "low pollution." This illustrates that the error in classifying "high pollution" is also minor.

Overall, the confusion matrix findings suggest that the classification model performs well in classifying air pollution into three categories. Despite a few mistakes, most of the predicts were true. Further analysis using metrics such as accuracy, recall, and F1-score may provide a more comprehensive insight into this model's performance [29]. The findings of the air quality classification model's performance evaluation in the Jakarta area reveal that this model has a decent ability to categorize air pollution into three categories, namely "moderate pollution," "low pollution," and "high pollution." This model has an accuracy of roughly 93%, which implies that the model correctly predicts the majority of the time.

In the context of air quality, there are two key measures to consider: "recall" and "F1-score." The recall assesses the model's ability to properly recognize air pollution, whereas the F1-score represents the model's ability to identify and classify air pollution types. The model has a recall of roughly 86% for the "Moderate Pollution" category. This means that the program can identify that around 86% of all actual air pollution cases fall into the "moderate pollution" category. The F1-score for this category is 0.9, indicating a reasonable balance between the model's ability to distinguish and categorize moderate pollution. The model has a very high recall for the "Low Pollution" category, which is over 96%. This demonstrates that the algorithm is quite good at detecting clean air pollution, recognizing around 96% of all situations of "low pollution." This category has an F1-score of 0.94, indicating extremely strong performance. Meanwhile, the model has a recall of around 94% for the "High Pollution" category, indicating a good capacity to distinguish extremely high levels of air pollution. This category's F1-score is also 0.94, indicating that this model performs well in classifying high pollution. Overall, this classification model is capable of categorizing air pollution into three groups with excellent accuracy, sensitivity, and a good balance of F1-score and recall. As a result, this model may be depended on to accurately estimate air quality in the Jakarta area.

4.3. Pollutants' contribution to Jakarta's air pollution

In the framework of this study, it is critical to identify what elements have the greatest influence on air quality in the Jakarta area. The categorization model's variable importance (Figure 5) analysis results demonstrate that the concentration of various types of air pollutants has a substantial impact on determining air quality. Several important conclusions are obtained based on the variable's importance:

a. **Particulate Matter 10 (PM10)**, or airborne particles having a diameter of fewer than 10 micrometers [44], has the largest impact on Jakarta's air quality. The high significance number shows that the PM10 level has a considerable impact on the region's level of air pollution. As a result, controlling PM10 emissions is critical in attempts to improve air quality.

b. **Sulfur dioxide (SO2)** has a significant impact on air quality. SO2 is often produced by industry and the combustion of fossil fuels[45]. Controlling SO2 emissions must be a major priority to reduce air pollution.

c. **Nitrogen dioxide (NO2)**, which is frequently emitted by motor vehicles, also has a substantial impact[46]. This demonstrates the significance of reducing car emissions to preserve excellent air quality.

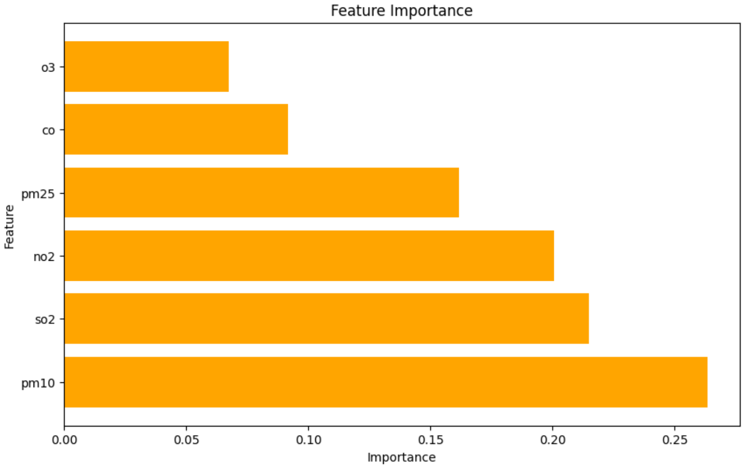

Figure 5. The contribution of the importance variable to air pollution in Jakarta

d. **Particulate matter 2.5 (PM2.5)**, while less essential than PM10, is still an important factor in determining air quality[47]. PM2.5 is a very small particulate matter that, if ingested over time, can have major health consequences.

e. **Carbon monoxide (CO)**, which is formed during fuel burning, also has an impact on air quality, though its importance is lesser than that of other causes[48].

f. **Ozone (O3)** has the least relevance yet still has an impact on air quality[49]. If a chemical reaction occurs in the open air, ozone can generate air pollution.

Decisions about emission control policies can be more directed and successful if these aspects are understood. Efforts to minimize PM10, SO2, NO2, and PM2.5 emissions, as well as monitor CO and O3 levels, will be critical in preserving and improving Jakarta's healthy air quality.

5. Conclusion

This study contributes significantly to our understanding of air quality in Jakarta areas by combining the capability of AI with a semi-supervised learning approach. The clustering method was used to divide air pollution data into three categories: moderate pollution, low pollution, and high pollution. This phase is to comprehend the similarities between metrics such as PM10, SO2, NO2, and others, even in the absence of a clear air quality label. This contribution is a vital first step towards improving and sustaining Jakarta's air quality. To estimate air quality in these places, the Random Forest classification model has been adjusted with the optimal parameters. The model has an accuracy rate of roughly 93%, proving AI's capacity to discern complicated patterns in unstructured air data. Furthermore, variables are necessary to comprehend the impact of each component on air quality. The findings emphasize the importance of PM10, SO2, and NO2 in influencing air quality. This is an example of how AI can aid in the investigation of intricate interactions between multiple environmental data. This study demonstrates the enormous potential of AI, particularly semi-supervised learning approaches, in understanding and managing air quality in places such as Jakarta. We intend to make a greater contribution to sustainable environmental policies and a healthier environment for Jakarta's inhabitants by utilizing this technology.

References

[1] W.-Y. Su, D.-W. Wu, S.-C. Chen, and L. Aleya, “Impact of Different Air Pollutants (PM10, PM2.5, NO2, and Bacterial Aerosols) on COVID-19 Cases in Gliwice, Southern Poland,” *International*

Journal of Environmental Research and Public Health 2022, Vol. 19, Page 14181, vol. 19, no. 21, p. 14181, Oct. 2022, doi: 10.3390/IJERPH192114181.

[2] W. Y. Su, D. W. Wu, S. C. Chen, C. H. Hung, and C. H. Kuo, "Association between air pollutants with calcaneus ultrasound T-score change in a large Taiwanese population follow-up study," *Environmental Science and Pollution Research*, vol. 30, no. 28, pp. 72607–72616, Jun. 2023, doi: 10.1007/S11356-023-27368-5/TABLES/4.

[3] R. S. Kumar, A. Arulanandham, and S. Arumugam, "Air quality index analysis of Bengaluru city air pollutants using Expectation Maximization clustering," *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation, ICAECA 2021*, 2021, doi: 10.1109/ICAECA52838.2021.9675669.

[4] S. Kaur and S. Sharma, "Internet of Things Enabled Framework for Sustainable Mobility and Clean Environment in Smart Cities," pp. 285–298, 2023, doi: 10.1007/978-3-031-37303-9_21.

[5] D. Mage *et al.*, "Urban air pollution in megacities of the world," *Atmos Environ*, vol. 30, no. 5, pp. 681–686, Mar. 1996, doi: 10.1016/1352-2310(95)00219-7.

[6] S. Roy, M. Saha, B. Dhar, S. Pandit, and R. Nasrin, "Geospatial analysis of COVID-19 lockdown effects on air quality in the South and Southeast Asian region," *Science of The Total Environment*, vol. 756, p. 144009, Feb. 2021, doi: 10.1016/J.SCITOTENV.2020.144009.

[7] P. Lestari, M. K. Arrohman, S. Damayanti, and Z. Klimont, "Emissions and spatial distribution of air pollutants from anthropogenic sources in Jakarta," *Atmos Pollut Res*, vol. 13, no. 9, p. 101521, Sep. 2022, doi: 10.1016/J.APR.2022.101521.

[8] A. A. Yusuf and B. P. Resosudarmo, "Does clean air matter in developing countries' megacities? A hedonic price analysis of the Jakarta housing market, Indonesia," *Ecological Economics*, vol. 68, no. 5, pp. 1398–1407, Mar. 2009, doi: 10.1016/J.ECOLECON.2008.09.011.

[9] M. Moradi, A. Broer, J. Chiachío, R. Benedictus, T. H. Loutas, and D. Zarouchas, "Intelligent health indicator construction for prognostics of composite structures utilizing a semi-supervised deep neural network and SHM data," *Eng Appl Artif Intell*, vol. 117, p. 105502, Jan. 2023, doi: 10.1016/J.ENGAPPAI.2022.105502.

[10] X. Li, Y. Li, K. Yan, H. Shao, and J. (Jing) Lin, "Intelligent fault diagnosis of bevel gearboxes using semi-supervised probability support matrix machine and infrared imaging," *Reliab Eng Syst Saf*, vol. 230, p. 108921, Feb. 2023, doi: 10.1016/J.RESS.2022.108921.

[11] K. Yu, T. R. Lin, H. Ma, X. Li, and X. Li, "A multi-stage semi-supervised learning approach for intelligent fault diagnosis of rolling bearing using data augmentation and metric learning," *Mech Syst Signal Process*, vol. 146, p. 107043, Jan. 2021, doi: 10.1016/J.YMSSP.2020.107043.

[12] Y. Ouali, C. Hudelot, and M. Tami, "An Overview of Deep Semi-Supervised Learning," Jun. 2020, Accessed: Sep. 03, 2023. [Online]. Available: https://arxiv.org/abs/2006.05278v2

[13] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach Learn*, vol. 109, no. 2, pp. 373–440, Feb. 2020, doi: 10.1007/S10994-019-05855-6/FIGURES/5.

[14] A. Ligthart, C. Catal, and B. Tekinerdogan, "Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification," *Appl Soft Comput*, vol. 101, p. 107023, Mar. 2021, doi: 10.1016/J.ASOC.2020.107023.

[15] Y. Kaya, Z. Yiner, M. Kaya, and F. W. Wibowo, "Prediction of air quality in Jakarta during the COVID-19 outbreak using long short-term memory machine learning," *IOP Conf Ser Earth Environ Sci*, vol. 704, no. 1, p. 012046, Mar. 2021, doi: 10.1088/1755-1315/704/1/012046.

[16] M. Hardini *et al.*, "Image-based Air Quality Prediction using Convolutional Neural Networks and Machine Learning," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 5, no. 1Sp, pp. 109–123, Aug. 2023, doi: 10.34306/ATT.V5I1SP.337.

[17] W. Y. Hong, D. Koh, A. A. A. Mohtar, and M. T. Latif, "Statistical Analysis and Predictive Modelling of Air Pollutants Using Advanced Machine Learning Approaches," *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2020*, Dec. 2020, doi: 10.1109/CSDE50874.2020.9411636.

[18] A. A. Dharmasaputro, N. M. Fauzan, M. Kallista, I. P. D. Wibawa, and P. D. Kusuma, "Handling Missing and Imbalanced Data to Improve Generalization Performance of Machine Learning Classifier," *2021 International Seminar on Machine Learning, Optimization, and Data Science, ISMODE 2021*, pp. 140–145, 2022, doi: 10.1109/ISMODE53584.2022.9743022.

[19] Suhartono, H. Prabowo, D. D. Prastyo, and M. H. Lee, “New Hybrid Statistical Method and Machine Learning for PM10 Prediction,” *Communications in Computer and Information Science*, vol. 1100, pp. 142–155, 2019, doi: 10.1007/978-981-15-0399-3_12/COVER.

[20] S. García, J. Luengo, and F. Herrera, “Data Preprocessing in Data Mining,” 2015.

[21] J. Sun and Y. Xia, “Pretreating and normalizing metabolomics data for statistical analysis,” *Genes Dis*, Jul. 2023, doi: 10.1016/J.GENDIS.2023.04.018.

[22] Y. Zhang, N. Meratnia, and P. Havinga, “Outlier detection techniques for wireless sensor networks: A survey,” *IEEE Communications Surveys and Tutorials*, vol. 12, no. 2, pp. 159–170, Jun. 2010, doi: 10.1109/SURV.2010.021510.00088.

[23] K. R. Shahapure and C. Nicholas, “Cluster quality analysis using silhouette score,” *Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*, pp. 747–748, Oct. 2020, doi: 10.1109/DSAA49011.2020.00096.

[24] B. W. Otok, A. Suharsono, Purhadi, R. E. Standsyah, and H. Al Azies, “Partitional Clustering of Underdeveloped Area Infrastructure with Unsupervised Learning Approach: A Case Study in the Island of Java, Indonesia,” *Journal of Regional and City Planning*, vol. 33, no. 2, pp. 177–196, Aug. 2022, doi: 10.5614/JPWK.2022.33.2.3.

[25] H. S. Park and C. H. Jun, “A simple and fast algorithm for K-medoids clustering,” *Expert Syst Appl*, vol. 36, no. 2, pp. 3336–3341, Mar. 2009, doi: 10.1016/ J.ESWA.2008.01.039.

[26] Y. Wang *et al.*, “Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels.” pp. 4248–4257, 2022. Accessed: Oct. 10, 2023. [Online]. Available: https://haochen-wang409.github.io/U2PL.

[27] E. Arazo, Di. Ortego, P. Albert, N. E. O’Connor, and K. McGuinness, “Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning,” *Proceedings of the International Joint Conference on Neural Networks*, Jul. 2020, doi: 10.1109/IJCNN48605.2020.9207304.

[28] Y. Jiao and P. Du, “Performance measures in evaluating machine learning based bioinformatics predictors for classifications,” *Quantitative Biology*, vol. 4, no. 4, pp. 320–330, Dec. 2016, doi: 10.1007/S40484-016-0081-2/METRICS.

[29] R. Yacouby Amazon Alexa and D. Axman Amazon Alexa, “Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models,” pp. 79–91, Nov. 2020, doi: 10.18653/V1/2020.EVAL4NLP-1.9.

[30] T. Gupta and S. P. Panda, “Clustering Validation of CLARA and K-Means Using Silhouette DUNN Measures on Iris Dataset,” *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, pp. 10–13, Feb. 2019, doi: 10.1109/COMITCON.2019.8862199.

[31] I. D. Ratih *et al.*, “Mapping the health quality in SUMENEP using k-medoids algorithm,” *AIP Conf Proc*, vol. 2668, no. 1, Oct. 2022, doi: 10.1063/5.0111821/2832232.

[32] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, “Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review,” *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 13, pp. 6308–6325, 2020, doi: 10.1109/JSTARS.2020.3026724.

[33] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, “Random Forests and Decision Trees,” 2012, Accessed: Sep. 03, 2023. [Online]. Available: www.IJCSI.org

[34] P. Martínez-Santos and P. Renard, “Mapping Groundwater Potential Through an Ensemble of Big Data Methods,” *Groundwater*, vol. 58, no. 4, pp. 583–597, Jul. 2020, doi: 10.1111/GWAT.12939.

[35] P. Doupe, J. Faghmous, and S. Basu, “Machine Learning for Health Services Researchers,” *Value in Health*, vol. 22, no. 7, pp. 808–815, Jul. 2019, doi: 10.1016/J.JVAL.2019.02.012.

[36] D. M. Ge, L. C. Zhao, and M. Esmaeili-Falak, “Estimation of rapid chloride permeability of SCC using hyperparameters optimized random forest models,” *https://doi.org/10.1080/21650373.2022.2093291*, vol. 12, no. 5, pp. 542–560, 2022, doi: 10.1080/21650373.2022.2093291.

[37] W. Zhang, C. Wu, H. Zhong, Y. Li, and L. Wang, “Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization,” *Geoscience Frontiers*, vol. 12, no. 1, pp. 469–477, Jan. 2021, doi: 10.1016/J.GSF.2020.03.007.

[38] Q. Zhou, W. Lan, Y. Zhou, and G. Mo, “Effectiveness Evaluation of Anti-bird Devices based on Random Forest Algorithm,” *2020 7th International Conference on Information, Cybernetics, and Computational Social Systems, ICCSS 2020*, pp. 743–748, Nov. 2020, doi: 10.1109/ICCSS52145.2020.9336891.

[39] A. K. Nasution *et al.*, “Prediction of Potential Natural Antibiotics Plants Based on Jamu Formula Using Random Forest Classifier,” *Antibiotics 2022, Vol. 11, Page 1199*, vol. 11, no. 9, p. 1199, Sep. 2022, doi: 10.3390/ANTIBIOTICS11091199.

[40] Y. Hu, Z. Sun, L. Pei, W. Li, and Y. Li, “Evaluation of pavement surface roughness performance under multi-features conditions based on optimized random forest,” *Proceedings - 2021 9th International Conference on Advanced Cloud and Big Data, CBD 2021*, pp. 133–138, 2022, doi: 10.1109/CBD54617.2021.00031.

[41] S. George and B. Sumathi, “Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction,” *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 11, no. 9, 2020, Accessed: Sep. 03, 2023. [Online]. Available: www.ijacsa.thesai.org

[42] Q. Liang, E. Vanem, K. E. Knutsen, and H. Zhang, “Data-Driven Prediction of Ship Propulsion Power Using Spark Parallel Random Forest on Comprehensive Ship Operation Data,” *IEEE International Conference on Control and Automation, ICCA*, vol. 2022-June, pp. 303–308, 2022, doi: 10.1109/ICCA54724.2022.9831854.

[43] K. Pal and B. V. Patel, “Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques,” *Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, pp. 83–87, Mar. 2020, doi: 10.1109/ ICCMC48092.2020.ICCMC-00016.

[44] P. Fermo *et al.*, “Improving indoor air quality through an air purifier able to reduce aerosol particulate matter (PM) and volatile organic compounds (VOCs): Experimental results,” *Environ Res*, vol. 197, p. 111131, Jun. 2021, doi: 10.1016/J.ENVRES.2021.111131.

[45] J. Likus-Cieślik, J. Socha, P. Gruba, and M. Pietrzykowski, “The current state of environmental pollution with sulfur dioxide (SO2) in Poland based on sulfur concentration in Scots pine needles,” *Environmental Pollution*, vol. 258, p. 113559, Mar. 2020, doi: 10.1016/J.ENVPOL.2019.113559.

[46] D. C. Carslaw, N. J. Farren, A. R. Vaughan, W. S. Drysdale, S. Young, and J. D. Lee, “The diminishing importance of nitrogen dioxide emissions from road vehicle exhaust,” *Atmos Environ X*, vol. 1, p. 100002, Jan. 2019, doi: 10.1016/J.AEAOA.2018.100002.

[47] Z. Fan, Q. Zhan, C. Yang, H. Liu, and M. Zhan, “How Did Distribution Patterns of Particulate Matter Air Pollution (PM2.5 and PM10) Change in China during the COVID-19 Outbreak: A Spatiotemporal Investigation at Chinese City-Level,” *International Journal of Environmental Research and Public Health 2020, Vol. 17, Page 6274*, vol. 17, no. 17, p. 6274, Aug. 2020, doi: 10.3390/IJERPH17176274.

[48] O. Kanat *et al.*, “Do natural gas, oil, and coal consumption ameliorate environmental quality? Empirical evidence from Russia,” *Environmental Science and Pollution Research 2021 29:3*, vol. 29, no. 3, pp. 4540–4556, Aug. 2021, doi: 10.1007/S11356-021-15989-7.

[49] M. Brancher, “Increased ozone pollution alongside reduced nitrogen dioxide concentrations during Vienna’s first COVID-19 lockdown: Significance for air quality management,” *Environmental Pollution*, vol. 284, p. 117153, Sep. 2021, doi: 10.1016/J.ENVPOL.2021.117153.