Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

Implementation of the K-Means Method for Developing an Air Quality Monitoring Information System

Faisal Rifky Nugraha[1], Adiat Pariddudin[2*], Anggra Triawan[3] , Fitria Rachmawati[4]

[1]*Information Systems, Universitas Binaniaga Indonesia, Indonesia*
[2]*Computer Science, Universitas Djuanda, Indonesia*
[3]*Informatics Engineering, Universitas Binaniaga Indonesia, Indonesia*
[4]*Information Systems, Universitas Ibn Khaldun Bogor, Indonesia*
faisalrifky09@gmail.com[1], adiat.pariddudin@unida.ac.id[2*], anggra@unbin.ac.id[3], fitria@uika-bogor.ac.id[4]

Abstract

Air quality in Bogor City is becoming increasingly complex due to the rising number of motor vehicles, small-scale industrial activities, and seasonal dynamics that are difficult to analyze using conventional methods. The Environmental Agency of Bogor City routinely monitors air quality through the Air Quality Monitoring System (AQMS); however, data utilization remains confined to monitoring and reporting, necessitating advanced analysis to achieve a more comprehensive view of air quality patterns. This study aims to classify time periods based on air quality parameters using the K-Means clustering method to identify good and bad air pollution categories. The research data was obtained from the Bogor City Environmental Agency's AQMS for the period from January 2023 to June 2025. The results indicate that the K-Means method successfully clustered the data into two groups: good and bad air quality categories. Good air quality was identified in 2023 (January, February, March, April, November, and December), 2024 (January, February, March, April, October, November, and December), and 2025 (January to May). Conversely, poor air quality occurred in 2023 (May to October), 2024 (May to September), and 2025 (June). The findings of this research are expected to support pollution control strategies and early warning systems based on air quality data.

Keywords: *K-Means, Air Quality, AQMS, Clustering, Bogor City.*

1. Introduction

Air quality in Bogor City is a crucial issue in urban environmental management considering its role as an ecological buffer for the Greater Jakarta metropolitan area (Jabodetabek) [2] [9]. The increasing number of motor vehicles, small-scale industrial activities, and infrastructure expansion have triggered a rise in air pollutant emissions such as PM2.5, PM10, NO_2, SO_2, CO, and O_3 [1] [9]. This phenomenon threatens public health, particularly vulnerable groups such as children and the elderly [1].

In addition to anthropogenic factors, seasonal dynamics also influence variations in air quality through the processes of pollutant dispersion and accumulation in the atmosphere [3] [9]. However, climate anomalies in recent years have caused seasonal patterns to become increasingly unpredictable. As a result, traditional calendar-based approaches are no longer accurate in representing actual air quality fluctuations [3].

The Environmental Agency of Bogor City (DLH) has routinely monitored air quality using the Air Quality Monitoring System (AQMS). Although the collected data are fairly comprehensive, their utilization is still limited to manual monitoring functions and periodic reporting [2]. The potential of these data has not been optimized for long-term pattern analysis or as a decision-making instrument due to the absence of dynamically integrated data processing media [2].

To address these limitations, data mining methods offer a comprehensive solution for environmental data processing. Clustering methods, particularly K-Means, have proven effective in grouping data based on similarities in characteristics without requiring predefined labels (unsupervised learning) [4] [6]. Several previous studies have confirmed that the K-Means algorithm is capable of accurately identifying air quality patterns and seasonal variations based on pollutant parameters [4] [6].

However, data processing using the K-Means algorithm is generally still performed partially using separate computational software. Consequently, policymakers face difficulties in accessing clustering information quickly and in real time [2]. Research specifically integrating multi-parameter K-Means algorithms into a dedicated digital platform for the context of Bogor City remains very limited [2] [6]. Therefore, this study implements the K-Means method integrated directly into the development of an Air Quality Monitoring Information System. This system is designed to automatically classify data in order to identify periods of high and low pollution levels more objectively. The clustering results presented through this information system are expected to serve as a scientific basis for supporting decision-making, formulating air pollution control strategies, and developing more targeted early warning systems in Bogor City [6].

2. Research Methodology

2.1. Theoretical Model of K-Means

This study faces challenges in identifying temporal air quality patterns from large-scale and complex monitoring datasets [2]. Continuous data recording with short intervals generates thousands of data entries each month. This condition complicates manual analysis, thereby requiring a method capable of automatically clustering data based on air pollution levels [7]. Currently, the Environmental Agency of Bogor City does not yet have an automatic clustering system capable of providing rapid, accurate, and efficient air quality mapping to support data-driven policy making [2].
To address these challenges, this study applies the K-Means algorithm as the primary method for clustering air quality data based on the similarity of their characteristics [4] [6]. K-Means is an unsupervised learning algorithm that works by partitioning data into a number of non-overlapping k clusters [4]. Mathematically, this algorithm aims to minimize the objective function in the form of the Sum of Squared Errors (SSE), or the squared distance between data points and their respective cluster centers (centroids) [7], which is formulated as follows.

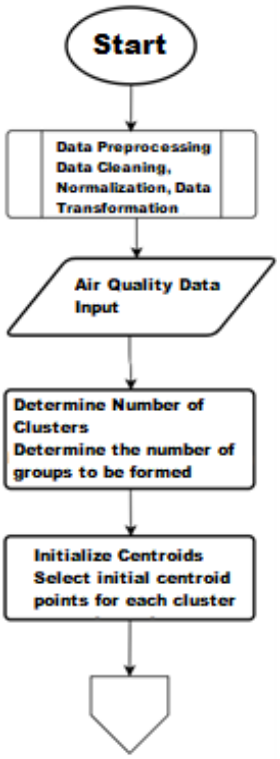
$$\text{SSE} = \sum_{j=1}^{k} \sum_{x \in S_j} ||x - c_j||^2$$

Description:
k = The predetermined number of clusters.
Sj = The set of data belonging to the j-th cluster.
x = The data vector of air quality parameters.
cj = The centroid or center of the j-th cluster.

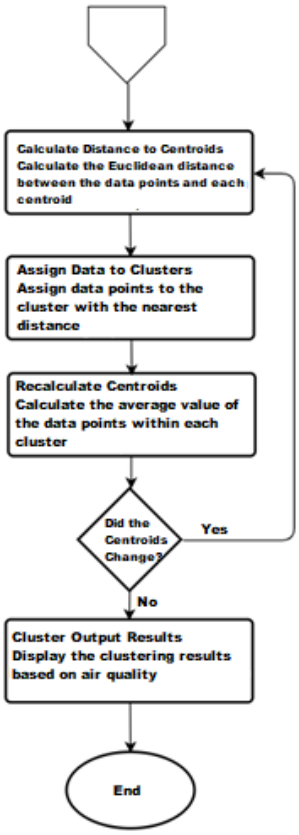
Through this approach, variations in monthly air quality patterns can be identified systematically. The implementation of K-Means in this study includes integrated stages beginning with data preprocessing, determining the optimal number of clusters, centroid calculation iterations, and result evaluation. The entire operational mechanism of this algorithm is visually presented through a flowchart and pseudocode in Table 1 as a technical guide for implementation within the developed information system [6].

Table 1: Flowchart and Pseudocode

Flowchart	Pseudocode

	Start
	# 1. Clean and prepare data
	data = pd.read_csv("kualitas_udara.csv")
	# Remove missing values
	data = data.dropna()
	# 2. Handle missing values using linear interpolation
	data.interpolate(method='linear')
	# Time transformation (month → sine and cosine encoding)
	data['month_sin'] = np.sin(2 * np.pi * data['bulan'] / 12)
	data['month_cos'] = np.cos(2 * np.pi * data['bulan'] / 12)
	# Remove the original 'bulan' column from the dataset
	data = data.drop(columns=['bulan'])
	# 3. Normalize other numerical features
	numerical_features = other parameters
	for feature in numerical_features:
	data[feature] = (data[feature] - data[feature].min()) / (data[feature].max() - data[feature].min())
	# Determine number of clusters (k)
	Select the number of clusters to be formed.
	k = 2
	# Initialize initial centroids
	Randomly select k data points as initial centroids.
	centroids_idx = np.random.choice(range(len(data)), k)
	centroids = dummy_variables[centroids_idx]

Flowchart	Pseudocode

	# Compute distance from each data point to each centroid
	distances = np.linalg.norm(dummy_variables[:, np.newaxis] - centroids, axis=2)
	
	# Assign each data point to the nearest centroid cluster
	labels = np.argmin(distances, axis=1)
	
	# Store previous centroids for convergence check
	old_centroids = centroids.copy()
	
	# Recalculate centroids as the mean of each cluster
	for j in range(k):
	cluster_points = dummy_variables[labels == j]
	if len(cluster_points) > 0:
	centroids[j] = cluster_points.mean(axis=0)
	
	# Check whether centroids have changed
	if np.all(centroids == old_centroids):
	break
	
	End

2.2. Model Prosedural Prototyping

The development of the information system in this study applies the Prototyping method. This approach begins with user requirements analysis, initial system design, and iterative evaluation before the system is finalized and fully implemented [8]. The stages of the Prototyping method implementation are shown in Figure 1 and described as follows.

2.2.1. Listen to Customer

This process is carried out through direct communication between developers and users to define system objectives and identify functional requirements. This stage also includes the collection of air quality parameter data from the Environmental Agency of Bogor City. The data include pollutant concentrations (PM10, PM2.5, SO_2, NO_2) and supporting meteorological data (wind speed, humidity, temperature, and rainfall) [1] [9].

2.2.2. Build/Revise Mockup

This stage focuses on designing an initial prototype that implements the system's core features. These features include the automatic clustering of air pollution levels based on the K-Means algorithm as well as visualization of periods with low and high air quality conditions [4] [6]. The prototype is developed in a modular manner in accordance with the agreed user requirements specification.

2.2.3. Customer Test Drives Mockup

The developed prototype is then tested by system experts (expert judgment) and end users in the field [8]. The test results produce feedback that is used as a basis for system improvement. The refinement and redesign process is carried out iteratively until the system fully meets user requirements for accurately and informatively classifying air quality conditions.

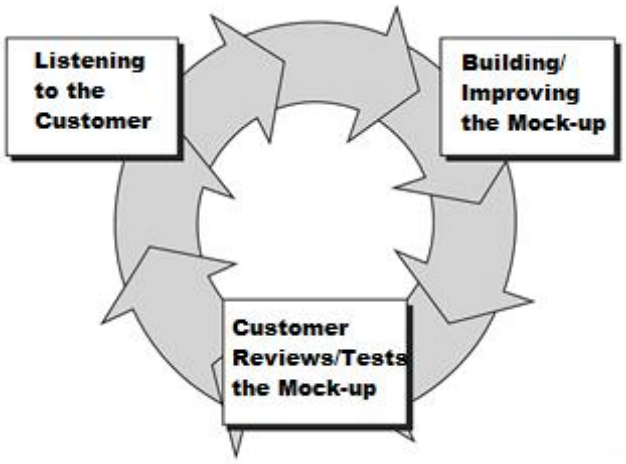


Fig 1: Prototype Procedural Model

3. Results and Discussion

Based on the system requirements analysis, the solution to optimize environmental data clustering is to integrate the K-Means clustering method. This method is used to group time periods based on the similarity of air quality parameters, enabling the automatic identification of periods with low (good) and high (poor) pollution levels. The current system workflow is illustrated in Figure 2.

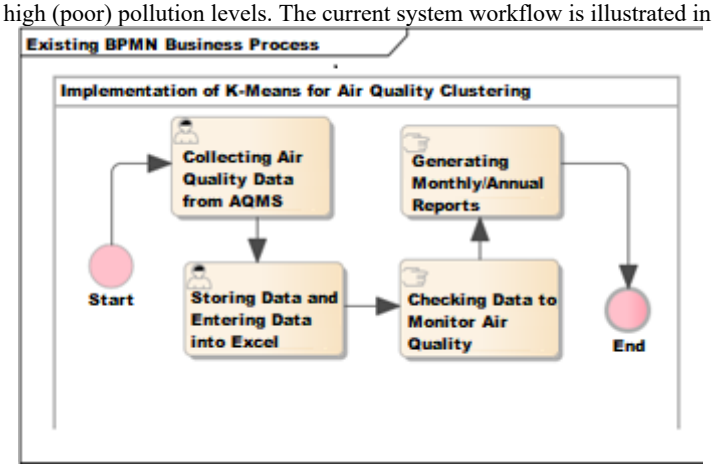


Fig 2: Existing Business Process

Figure 2. Existing Business Process shows that the management of air quality data in the previous system was still performed manually, starting from the recording stage to the preparation of periodic reports. This conventional mechanism caused the processing to be slow, prone to human error, and inefficient. Therefore, an automated system is required to accelerate data processing, improve computational accuracy, and support data-driven decision-making. The proposed new business process flow is presented in Figure 3.

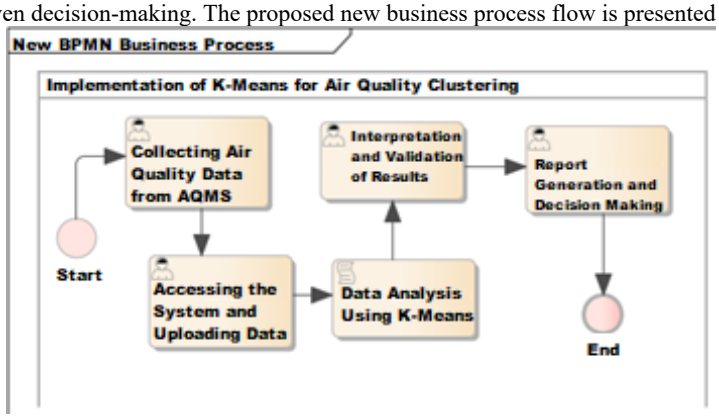


Fig 3: New Business Process

Figure 3. New Business Process begins with the automatic collection of air quality data through the Air Quality Monitoring System (AQMS) stations. The data obtained from AQMS is uploaded into the system and then processed through automated preprocessing stages, including data cleaning, missing value imputation, and normalization. Next, the system performs analysis using the K-Means algorithm to cluster air quality into low (good) and high (poor) categories. The analysis results are visualized in the form of interactive graphs, while reports can be automatically generated by the system. The implementation of this new architecture enables officers to take control measures or formulate environmental policies more quickly, accurately, and efficiently.

3.1. Analysis Results of the Method and Variable Determination

The K-Means algorithm is applied to form temporal clusters based on the similarity of air quality parameter values. The dataset used consists of historical records from the Bogor City AQMS station for the period January 2023 to June 2025, with a total of 40,295 data records, as presented in Table 2.

Table 2: Sample Air Quality Data from January 2023 to June 2025

No	Time	PM10	PM2.5	SO2	NO2	Wind Speed	Humidity	Temperature	Rainfall
1	01/01/2023 00:00	8	7	36	40	7	81	25,1	1,17
2	01/01/2023 00:30	10	8	37	37	7	82	25,1	1,17
...	...	...	...	...	...	...	...	...	...
40295	02/06/2025 23:30	43	52	38	15	0	84	26,1	0

The dataset in Table 2 cannot be directly used as input for the clustering algorithm due to variations in format and the presence of missing values. Therefore, the data must undergo preprocessing stages, including interpolation, transformation, and normalization, to ensure accurate analytical results. Before the clustering process is executed, the research variables are first defined to determine their influence weights on air quality, as presented in Table 3.

Table 3: Definition of Research Variables

No	Variable	Definition and Scientific Importance
1	Time	Represents the measurement period; important for identifying seasonal pattern fluctuations.
2	PM10	Concentration of particulate matter with a diameter less than or equal to 10 micrometers; a key indicator of urban air pollution.
3	PM2.5	Concentration of fine particulate matter with a diameter less than or equal to 2.5 micrometers; the most hazardous pollutant for human respiration.
4	SO2	Sulfur dioxide gas concentration; originates from small-scale industrial emissions and fuel combustion.
5	NO2	Nitrogen dioxide gas concentration; a primary indicator of motor vehicle emission intensity.
6	Wind Speed	Wind speed rate (km/h); influences the dispersion and dilution of atmospheric pollutants.
7	Humidity	Percentage of water vapor (%); affects photochemical reactions of secondary pollutants in the air.
8	Temperature	Air temperature (°C); a meteorological parameter controlling atmospheric layer stability.
9	Rainfall	Rainfall volume (mm); a natural pollutant removal agent through atmospheric washing processes.

3.2. Use Case Diagram

The object modeling of this system is illustrated using a use case diagram to represent the workflow of air quality data clustering using the K-Means method. This diagram serves to explain the interaction between the user (actor) and the system, as well as to map the main available menus. Through this use case diagram, the overall functionality and system workflow can be clearly understood. The use case design of the developed application is presented in Figure 4.

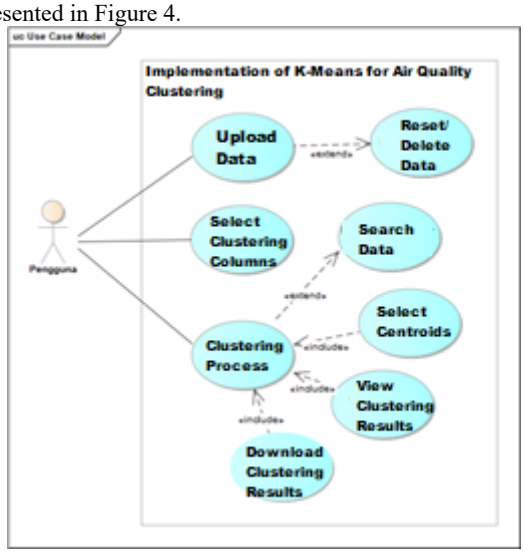


Fig 4: Use Case Diagram

Figure 4 illustrates the interaction between users and the system in the use case diagram. Users can perform several main activities, such as uploading air quality data in Excel or CSV format, selecting parameter columns to be used in the clustering process, and executing the clustering algorithm. In addition, the system provides filtering and data search features based on specific parameters or time ranges to facilitate the determination of initial centroids. After the execution process is completed, users can directly view and download the clustering results processed by the system.

3.3. Implementation of the Information System Interface

Implementation is the realization of previously modeled system functionalities into program code. The main interface of the Bogor City Air Quality Monitoring Information System, which integrates the K-Means algorithm, is presented in Figure 5 (Clustering Results Visualization) and Figure 6 (Rainfall Analysis Results).

Fig 5: Clustering Results Visualization

Fig 6: Rainfall Analysis Results

4. Conclusion and Recommendations

4.1. Conclusion

This collaborative research successfully integrates the K-Means clustering algorithm into a Streamlit-based air quality monitoring information system to automatically cluster historical AQMS data from Bogor City for the period January 2023 to June 2025. Based on the average Air Quality Index (AQI) values, the system successfully classifies the data into two functional clusters, namely Good and Poor categories. The temporal mapping results indicate that air quality categorized as Good is predominantly concentrated from January to April and at the end of the year (November–December) during the 2023–2025 period. Conversely, a decline in air quality reaching the Poor category is consistently identified during the peak dry season, covering May to October 2023, May to September 2024, and June 2025. This information system has been proven capable of summarizing detailed pollution distribution patterns to support environmental control strategies and early warning systems in Bogor City.

4.2. Recommendations

Based on the limitations of this study, several recommended future research directions include expanding the input variables in the system database by adding a wider range of multi-component pollutant gas parameters simultaneously, such as NO_2, O_3, and CO, in order to produce a more comprehensive air quality segmentation process. In addition, improving the data processing granularity in the system backend should also be considered, from a monthly aggregation scale to a daily or hourly fluctuation scale. These technical enhancements are crucial so that the information system can detect anomalies in air quality changes in real time and provide instant early warning functionality for the community in Bogor City.

References

[1.] Auliasari, K., Kertaningtyas, M., & Raya Karanglo Km, J. (2021). Analisis kualitas udara menggunakan algoritma K-means. Jurnal Informatika & Rekayasa Elektronika, 4(2). http://e-journal.stmiklombok.ac.id/index.php/jire
[2.] Carudin, Marisa, Murnawan, Reba, F., Koibur, M. E., Thantawi, A. M., Halim, A., & Wattimena, F. Y. (2024). Buku ajar data mining. PT. Sonpedia Publishing Indonesia
[3.] Hutauruk, C. H., Rahmanto, E., & Pancawati, M. C. (2020). Variasi musiman dan harian PM2.5 di Jakarta periode 2016–2019. Buletin GAW Bariri, 1(1), 20–28. https://doi.org/10.31172/bgb.v1i1.7
[4.] Jayadi, B. V., Handhayani, T., Lauro, D., & Kom, S. (2023). Perbandingan KNN dan SVM untuk klasifikasi kualitas udara di Jakarta.
[5.] Mahajan, T., Singh, G., & Bruns, G. (2021, March). An experimental assessment of treatments for cyclical data. Computer Science Conference for CSU Undergraduates.
[6.] Mahendrasyah, I., Diana, A., Rusdah, & Mahdiana, D. (2024). Penerapan algoritma K-means untuk klasterisasi indeks standar pencemaran udara. Jurnal Teknologi, 14(2), 146–156. https://doi.org/10.26594/teknologi.v14i2.4088
[7.] Rizal, S., Wafdan, R., Hidayat, M. N., Nurhayati, & Iskandar, T. (2025). Belajar matematika dasar dengan R. USK Press
[8.] Shalahuddin, M., & Rosa, A. S. (2011). Rekayasa perangkat lunak. Informatika Bandung
[9.] Sofiati. (2007). Penyebaran polusi udara dan kondisi meteorologinya di Kota Bogor. LAPAN
[10.] Yazid, F., & Affandes, M. (2017). Clustering data polutan udara Kota Pekanbaru dengan menggunakan metode K-means clustering. Jurnal CoreIT, 3(2), 76–81. https://doi.org/10.24014/coreit.v3i2.4419