Application of K-Means Clustering Algorithm for Air Quality Pattern Analysis in Jakarta

Muhammad Arya Fayyadh Razan[a], Nur Juzieatul Alifah[b], Qurrata A'yuni[c], Masna Wati[d], Haviluddin[e]

[a,b,c,d,e,f]*Department of Informatics Engineering, Mulawarman University*

Corresponding Author:

[a]*m.aryafayyadhrazan123@gmail.com*

ABSTRACT

Air pollution in urban areas, particularly in Jakarta, is a significant issue that impacts public health and environmental quality. This study aims to analyze air quality patterns in Jakarta from 2010 to 2023 using the K-Means Clustering method based on Air Pollution Standard Index (ISPU) data. Data processing stages based on the CRISP-DM methodology are applied to process and analyze data systematically. The stages include business understanding, data understanding, data preparation, modeling, and evaluation. The results showed that the data were divided into three distinct clusters: healthy, unhealthy, and moderate. Cluster 0, which includes stations DKI1 and DKI2, shows better air quality, while cluster 2, which consists of stations DKI3, DKI4, and DKI5, shows higher pollution levels. These findings offer valuable insights for policymakers in developing more effective air pollution control strategies. Thus, the results of this study not only contribute to the understanding of air quality in Jakarta but also emphasize the need for data-driven mitigation actions to improve public health and the environment.

Keywords : Air quality, Algorithm, AQI, Clustering, K-Means

INTRODUCTION

The environment is a unit of space that encompasses all forces, objects, conditions, and living things, including humans and their behavior, which collectively affect survival and the natural world. Clean air is a significant factor in determining environmental quality (Adinda et al., 2021). However, the rapid growth of cities and human activities, such as energy use in industry, power generation, and transportation, leads to the burning of fuels that pollute the air.

Air pollution is caused by the accumulation of pollutants from industrial activities, vehicles, and other emission sources (Adityo et al., 2023). Major pollutants include carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), surface ozone (O_3), and dust particles such as PM_{10} and $PM_{2.5}$ (Pertiwi et al., 2024). High concentrations of pollutants have a detrimental impact on both human health and the environment. Jakarta is one of the cities that frequently experiences unhealthy air quality, according to IQAir data (2025), due to its high urbanization activity, large population, and high vehicle density. This condition makes Jakarta vulnerable to air pollution, which is a complex and regionally varied phenomenon. Therefore, a clustering analysis was conducted to determine the pattern of air pollution distribution at various monitoring stations, providing a clearer picture of the severity and characteristics of pollution in each area. This clustering is important as a basis for more targeted environmental policy planning, such as determining priority locations for emission control, spatial improvements, or public awareness campaigns.

To understand air pollution patterns more systematically, data mining clustering techniques can be used. One effective clustering method is the K-Means Clustering algorithm. This method groups regions based on annual pollution characteristics based on ISPU (Salsabila et al., 2024). In their research, Salsabila et al. successfully distinguished areas with high, medium, and low pollution levels more precisely using K-Means. In addition, the study (Dhewayani et al., 2022) applied the Elbow method as an approach to objectively determine the optimal number of clusters by identifying the point at which adding clusters no longer provides a significant reduction in the inertia value. This is important to avoid over-clustering, which can obscure the actual pattern of pollution. The results of this study's analysis can assist the government in designing data-driven air quality management policies and formulating concrete solutions to mitigate the impact of pollution.

LITERATUR REVIEW

Data Mining

Data mining is the process of exploring and analyzing large amounts of data to identify patterns, trends, or relationships that are not immediately apparent in the data (Kusrini & Luthfi, 2021). Some of the important steps in this process include data cleaning, feature selection, data transformation, model building, and result testing. Machine learning, artificial intelligence, and statistics are used in this technique to extract complex data. This research utilizes data mining to analyze Jakarta's air quality data, which includes pollutant parameters such as PM10, NO2, and SO2. Using this data, researchers were able to discover pollution distribution patterns and seasonal trends and categorize regions based on their air quality. This creates a more objective picture of environmental conditions, aiding the data-driven decision-making process.

Algoritma K-Means

K-Means is one of the most commonly used clustering algorithms due to its efficiency in quickly and simply grouping large datasets. This algorithm works by calculating the number of clusters (K), initializing the centroids randomly, grouping data based on the closest distance to the centroid, and then updating the centroid positions until they no longer change or converge (Dharmawan, 2022). This research utilizes K-Means clustering to categorize Jakarta air quality data based on pollutant parameters, including PM10, SO2, and NO2. The purpose of this clustering is to find patterns of distribution of areas with comparable air quality characteristics. However, the limitation of K-Means is that it is sensitive to outliers, and the results are highly dependent on the correct number of clusters selected. Consequently, to find the optimal K value, techniques such as Elbow are required.

Clustering

The goal of unsupervised learning techniques is to group data into various clusters based on the degree of similarity between data elements (Hani et al., 2022). One can use this method to discover natural structures or hidden patterns in a dataset. Clustering is used in air quality research to understand the distribution and variation of pollution across regions or time. Partitioning methods (such as K-Means), hierarchical clustering, and density-based clustering are some standard clustering techniques. Each method has advantages depending on the type of data used and the purpose of the analysis. In this study, clustering is employed to categorize areas in Jakarta based on air quality indicators, including PM10, SO2, and O3. This facilitates the analysis of pollution levels and possible environmental risks in each cluster.

Rapidminer

RapidMiner has emerged as a powerful data analysis platform with a visual interface that facilitates the data mining process, eliminating the need for in-depth programming skills. (Afrahul., 2024) stated that this tool provides an integrated field for the entire data analytics workflow, including pre-processing, model building, evaluation, and visualization results. It also emphasizes the support of data formats, such as databases and Excel files, an extensive library of algorithms, particularly in clustering, including K-Means, and data visualization capabilities in a user-friendly manner. RapidMiner is widely used in university and industrial environments, mainly due to the ease of accelerating model prototyping, even in the free version, compared to the commercial version on several features. Although the free version offers more features, RapidMiner remains the preferred choice for data mining and machine learning implementation due to its ease of integrating various analysis techniques.

Elbow Method

To determine the ideal number of clusters (K value) in clustering algorithms such as K-Means, the radius method is commonly used. The sum of Squared Errors (SSE) values for various numbers of clusters is calculated and then displayed in a graph. The graph shows a decrease in the SSE value as the number of clusters increases; however, at a certain point, the decrease stops and forms an elbow-like angle (Sari, 2021). Since adding clusters does not significantly improve clustering accuracy, the current number of clusters is considered the most efficient. The Elbow Method is employed in this study to determine the optimal number of clusters for air quality data analysis in Jakarta. The clustering results become more accurate and representative, helping to facilitate a relevant interpretation of air pollution conditions.

Silhouette Coefficient

Silhouette Coefficient combines two main concepts of cluster evaluation: cohesion and separation. Cohesion measures how close a data is to other data in a cluster, while separation measures the average distance of a data to other data in the cluster (Kurniawan et al., 2023). The geometric distance formula is usually used to calculate the distance between data. Silhouette values range between -1 and 1, with higher values indicating that the data is well clustered, while lower values indicate possible clustering errors. The Silhouette Coefficient is used to evaluate the quality of the clustering results of air quality data in Jakarta based on parameters such as PM10, SO2, and NO2. This is important to ensure that the clusters formed truly reflect the differences in air pollution characteristics between regions accurately and validly.

METHODS

In this study, the method employed is the Cross Industry Standard Process for Data Mining (CRISP-DM), a standard approach in the data mining process (Wardani, 2021). This method is used to systematically process and analyze data in order to identify patterns or information that can be used in decision-making. By following the CRISP-DM stages, this research ensures that the data exploration, modeling, and evaluation processes are carried out optimally to obtain accurate and relevant results. The CRISP-DM approach has also proven effective in various previous studies that applied K-Means in air quality analysis (Widuri, 2023; Mawaddah & Anjelica, 2023).

In Figure 1, it is shown that Business Understanding, Data Understanding, Data Preparation, and Modeling are the core stages in data analysis, which begin with understanding business objectives, followed by data analysis and preparation, and culminate in model building to produce solutions to identified problems.

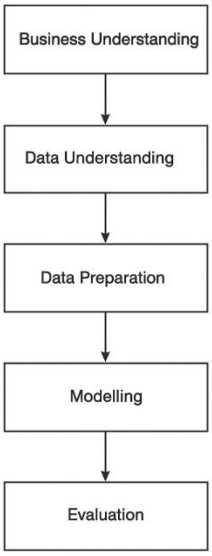


Figure 1. Stages of the Research Method

Business Understanding

This stage involves understanding the needs and objectives from a business perspective. The research started by identifying the main issue, which is the high level of air pollution in the DKI Jakarta area. The primary objective is to develop regional air quality segmentation to support more targeted pollution control policy-making. Therefore, the data mining problem was formulated as a process of clustering areas based on ISPU parameters to reveal spatial patterns of air pollution at each monitoring station. After the objectives and problems are formulated, a data analysis strategy and plan are developed according to the CRISP-DM approach to achieve the desired results (Msy Aulia Hasanah, 2023).

Data Understanding

In this stage, the process begins with the collection of ISPU data for the period 2010-2023 from five air quality monitoring stations in Jakarta (DKI1-DKI5). The dataset includes several main air pollutant parameters, namely PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, and NO_2. The initial exploration stage is done by analyzing the data structure, such as the number of records, the type of data in each column, and the distribution of values for each parameter. Visualizations such as histograms were used to understand distribution patterns and detect the presence of outliers. In addition, missing values were identified. Thus, this stage not only provides an initial understanding of the data, but also becomes the basis for selecting attributes and customizing features that will be used in the next stage (Belia Putri Salsabila, 2023; JSON, 2023).

Data Preparation

This stage aims to compile the final dataset from raw data so that it is ready to be used in the modeling process. The process includes data cleaning, data selection based on relevant records and attributes, and data transformation to fit the needs of the modeling algorithm (Ni Wayan Wardani, 2023). In this research, ISPU data goes through several preparation stages, such as handling missing values, data normalization, selection of key attributes, and conversion of categorical attributes into a numeric format that can be processed by the K-Means algorithm.

Modeling with K-Means Clustering

K-Means algorithm is a clustering method that groups data based on feature similarity. The process begins by determining the desired number of clusters (k), randomly selecting initial centroids, and calculating the distance of the data to each centroid using Euclidean Distance. The data is then inserted into the nearest cluster, and the centroid position is updated based on the average of the points in the cluster. This process is repeated until the centroid position stabilizes or the number of iterations is reached (Adinda Amalia., 2023).

Before clustering, the first step is to determine the optimal number of clusters. This is done with the Elbow Method, where the number of clusters is mapped against the Within-Cluster Sum of Squares (WCSS) value. The optimal point is when the graph starts to slope. In addition, the Silhouette Score is also used to evaluate the extent to which the clusters formed are well separated and have good density. Usually, the optimal K value is in the range of 3 to 5, depending on the data distribution (Adinda Inez Sang., 2023; Adityo Nugroho., 2023). Each data point x_i is calculated the distance to each centroid using the following Euclidean Distance formula:

$$d(x_i\,,\,c_k) = \sqrt{\sum_{j=1}^{n}\left(x_{ij} - c_{kj}\right)^2}$$

Where:
x_i= The kth data
c_k= Centroid of the kth cluster
n = Number of features in the dataset (e.g. PM 2.5, PM 10, CO, NO2)
Once the data has been clustered, we need to update the centroid position based on the average of all points in the cluster:

$$c_k = \frac{1}{N_k}\sum_{i\,=\,1}^{N_k} x_i$$

Where:
c_k = New centroid of the kth cluster
N_k = Number of data points in the kth cluster
x_i = All data points belonging to the kth cluster

The main advantage of the K-Means algorithm is its ability to process large amounts of data quickly, which makes it particularly suitable for air quality analysis. Using K-Means, areas can be clustered based on air pollution levels, such as PM10, CO, and NO2 concentrations. The results of this clustering can provide better insights into zones with high risk of air pollution, ultimately supporting more effective and data-driven health policy making (Kartika Dian Pertiwi., 2023).

Evaluation

This evaluation stage aims to determine the optimal number of clusters. The evaluation is done using the Elbow Method, where the number of clusters is mapped against the Within-Cluster Sum of Squares (WCSS) value. The optimal point is determined when the graph begins to slope, indicating that increasing the number of clusters no longer provides a significant reduction in the WCSS. This evaluation helps to ensure the selection of an effective number of clusters in forming a representative model (Adinda Inez Sang., 2023).

$$WCSS = \sum_{i=1}^{k}\sum_{x\in C_i}\left||x - \mu_i|\right|^2$$

Where:
C_i = The i-th cluster
μ_i = Centroid of the i-th cluster
x = Data points in cluster C_i
|(|x - μ_i |)|^2 = Squared Euclidean distance between x and centroid

To complete the evaluation of the number of clusters, this study also used the Silhouette Coefficient. This method measures how well the data has been grouped by comparing the distance between points in one cluster (cohesion) with the distance to the nearest other cluster (separation). The value of this coefficient is in the range of -1 to 1. The closer to 1, the better the cluster is formed. Conversely, a value close to -1 indicates that the data may be misclustered (Kurniawan et al., 2023).

In general, the Silhouette value for each point is calculated by the formula:

$$\mathrm{sil}(c) = \mathrm{sil}(k)\frac{1}{|k|}\sum_{i=1}^{k}\mathrm{sil}(c_i)$$

Where:
sil(k) = silhouette value of all clusters
|k| = number of clusters k
sil(c_i) = average silhouette value

By combining the Elbow Method and Silhouette Coefficient, the selection of the number of clusters in this study is done more objectively and thoroughly, so that the results of grouping air quality data become more representative and valid.

RESULTS

In this section, the research results are given and explained in accordance with the research methods mentioned earlier. This section explains the results of identifying air quality patterns in Jakarta using the K-Means Clustering method in the context of data mining.

Business Understanding

In an effort to understand the condition of air pollution in Jakarta, this study aims to group monitoring stations based on air quality patterns using the K-Means clustering method. By analyzing ISPU data from 2010 to 2023 without default categories, this research focuses on identifying groups of stations that have similar air pollution characteristics. The clustering results are expected to provide new insights for policy makers and related parties in formulating a more targeted and data-driven air pollution control strategy.

Data Understanding

This study uses public data on air pollution based on Air Quality Index (AQI) or Air Pollution Standard Index (ISPU) measured from 5 Air Quality Monitoring Stations (SPKU) in Jakarta. The dataset used is obtained from a datasets website called Kaggle and can be accessed via the link: https://www.kaggle.com/datasets/senadu34/air-quality-index-in-jakarta-2010-2021/data. This data covers the period from 2010 to 2023 and is used to analyze air quality trends in Jakarta. In the dataset, each station acts as an identifier that distinguishes each line of data, allowing for specific tracking and analysis of air quality based on monitoring location.

Table 1. DKI Jakarta Air Pollution Standard Index Data 2010-2023

No	Date	Station	pm10	pm25	so2	co	o3	no2	max	critical
1	2010-01-01	DKI1	60	-	4	73	27	14	73	CO

		(Bunderan HI)								
2	2010-01-02	DKI1 (Bunderan HI)	32	-	2	16	33	9	33	O3
3	2010-01-03	DKI1 (Bunderan HI)	27	-	2	19	20	9	27	PM10
4	2010-01-04	DKI1 (Bunderan HI)	22	-	2	16	15	6	22	PM10
5	2010-01-05	DKI1 (Bunderan HI)	25	-	2	17	15	8	25	PM10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4443	2023-11-29	DKI4 (Lubang Buaya)	71	105	30	19	22	14	105	PM25
4444	2023-11-30	DK1 (Bunderan HI)	38	67	43	12	34	34	67	PM25

With details of the data attributes as follows:

a. date : Date of air quality measurement.
b. station : Measurement locations at several stations. DKI1 (Bunderan HI), DKI2 (Kelapa Gading), DKI3 (Jagakarsa), DKI4 (Lubang Buaya), and DKI5 (Kebon Jeruk).
c. pm10 : Particulate matter of one of the measured parameters.
d. pm25 : Particulate matter of one of the measured parameters.
e. so2 : Sulfide of one of the measured parameters.
f. co : Carbon Monoxide one of the measured parameters.
g. o3 : Ozone is one of the measured parameters.
h. no2 : Nitrogen dioxide is one of the measured parameters.
i. max : The highest value of all parameters.
j. critical : The parameter with the highest measurement result.

Data Preparation

The data preparation stage is an important step before the modeling process. In this stage, several main processes are carried out, namely data cleaning, attribute selection, and data transformation. The goal is to ensure that the data used is clean, relevant, and ready to be processed analytically.

Data Cleaning

The total dataset from 2010-2023 amounted to 4626 data, then after data cleaning by deleting empty data (missing values), the data obtained became 4444 data. Then some attributes that were deleted were the date, pm25, max, and critical attributes.

Attribute Selection

Other attributes that are not needed for later modeling such as max and critical attributes are removed because these attributes are the maximum values of other parameters. The date attribute was not required for modeling. The pm25 attribute was also removed as it had many missing values and could be represented by another particulate attribute, pm10.

Data Transformation

The numerical data was then transformed using the Z-transformation (standardization) method. This process converts the data values to a standardized scale with a mean of zero and standard deviation of one, in order to improve the model's performance in detecting patterns. In addition, station attributes that were originally in the form of text labels (DKI1 to DKI5) were converted into integer numbers ranging from 0 to 4.

Table 2. Data Preparation Results

No	Station	pm10	so2	co	o3	no2
1	0	-0.269	-1.716	3.098	-1.249	-0.495
2	0	-1.725	-1.860	-1.070	-1.128	-1.079
3	0	-1.986	-1.860	-0.850	-1.391	-1.079
4	0	-2.246	-1.860	-1.070	-1.492	-1.430
5	0	-2.090	-1.860	-0.996	-1.492	-1.196
⋮	⋮	⋮	⋮	⋮	⋮	⋮
4443	3	0.303	0.143	-0.850	-1.350	-0.495
4444	0	-1.413	1.073	-1.362	-1.108	1.842

Table 2 shows the final results after the data preparation process. It can be seen that the data has been cleaned from missing values, irrelevant attributes have been removed, numeric data has been normalized, and station labels have been converted to numeric form. The data from this data preparation will be used for modeling using the K-Means algorithm.

Modelling

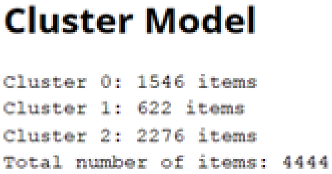
After the data preparation stage, the next stage is modeling clustering using the K-Means algorithm. Performing clustering using the K-Means algorithm has several stages including:

a. Determining the number of clusters to be used. The number of clusters used for this clustering is K=2 to K=7.
b. Determine the initial centroid by initializing three centroids randomly.

c. Calculating the distance between data and centroid using euclidean distance to measure the similarity between data and the closest centroid.

d. The iteration process is carried out until the centroid no longer changes significantly or until the maximum iteration limit of 100 steps is reached (max optimization steps = 100). To ensure stable results, the algorithm was run 10 times with different centroid initialization in each trial (max runs = 10).

After the modeling process using the K-Means algorithm by utilizing rapidminer software, the clustering results show that the data is divided into three clusters that have different characteristics. cluster 0 consists of 1546 items, cluster 1 consists of 622 items, and cluster 2 consists of 2276 items. The total number of clustered data is 4444 items which can be seen in Figure 2.



Cluster Model

```
Cluster 0: 1546 items
Cluster 1: 622 items
Cluster 2: 2276 items
Total number of items: 4444
```

Figure 2. K-Means Clustering Modeling Results.

Then to further understand the characteristics of each cluster, we can analyze the centroid table which contains the average value of each attribute in each cluster formed after the clustering process using the K-Means algorithm as follows:

Table 3. Centroid Table

Cluster	Station	pm10	so2	co	o3	no2
0	0.961	-0.327	-0.455	0.146	-0.159	-0.233
1	2.437	-0.474	1.626	-1.014	-0.818	1.006
2	3.204	0.352	-0.136	0.178	0.331	-0.117

Based on Table 3 centroids, it can be seen that Cluster 0 has values that tend to be lower in pollutant attributes such as PM10, SO2, and NO2, so it can be categorized as a group with healthier air quality (healthy category). Cluster 1 shows higher values on SO2 and NO2 attributes than other clusters, indicating that this group has higher levels of air pollution and potential health risks (unhealthy category). Cluster 2 has varying values with some pollutant attributes higher than Cluster 0, but lower than Cluster 1. This indicates that this group has moderate air quality or is in between the healthy and polluted categories (moderate category).

To observe the pattern of data distribution and determine the tendency of group formation based on air pollutant parameters, an initial clustering process was carried out using a combination of pm10 and co attributes. Visualization of the clustering results with the number of clusters varied with K=2 to K=7 is shown in Figure 3.

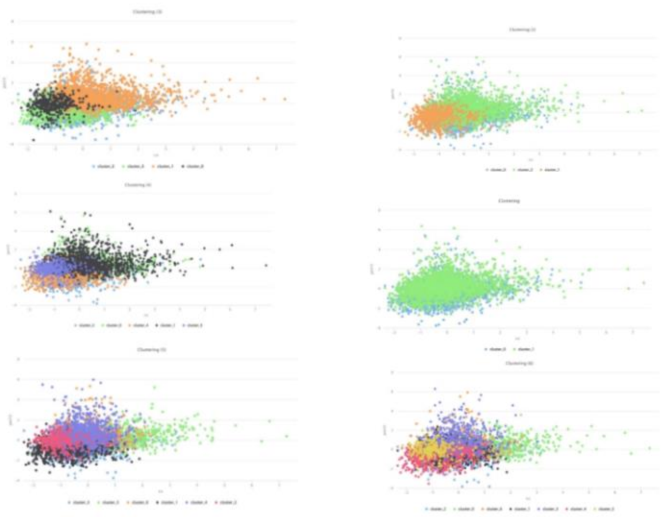


Figure 3. Clustering data based on pm10 and co

Evaluation

In this evaluation stage, the best number of clusters is determined, using one of the methods, namely the Elbow method. By observing the percentage of results obtained when comparing the number of clusters that form the elbow. Then an experiment is conducted to determine the appropriate number of clusters by running the K-Means algorithm with clusters K=2 to K=7 and utilizing the average Cluster Distance Perfomance to compare the values.

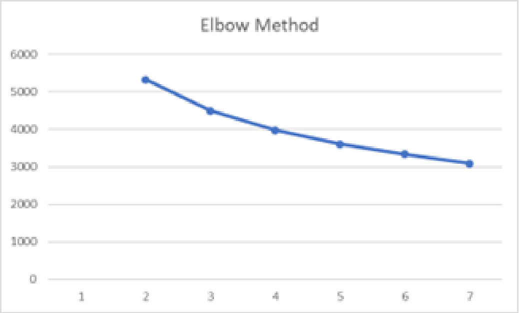


Figure 4. Elbow Method Elbow Method

Based on Figure 4, by observing the percentage of results obtained when comparing the number of clusters that form an elbow, the cluster numbered 3 shows the highest difference value and the sharpest angle on the graph, so it can be concluded that cluster K = 3 is the best number of clusters. Based on Figure 4, by observing the percentage of results obtained when comparing the number of clusters that form an elbow, the cluster numbered 3 shows the highest difference value and the sharpest angle on the graph, so it can be concluded that cluster K = 3 is the best number of clusters.

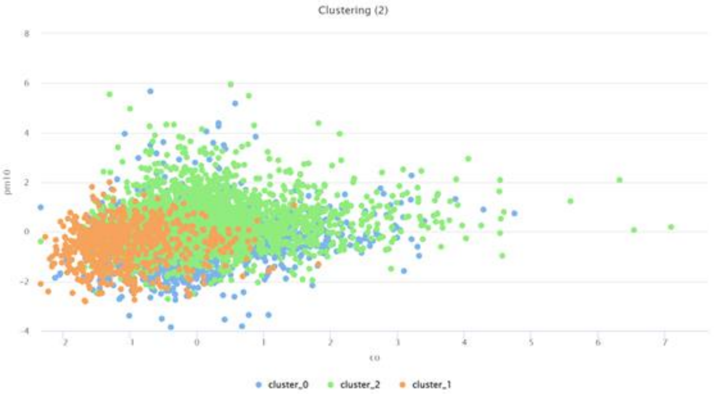


Figure 5. Visualization of data distribution with K=3

Based on the Elbow method that has been carried out, Figure 5 is a visualization of the distribution of cluster data using K = 3. The selection of the K=3 value is based on the elbow point that is clearly visible on the Elbow method graph, which shows that three clusters are the optimal number to group data effectively.

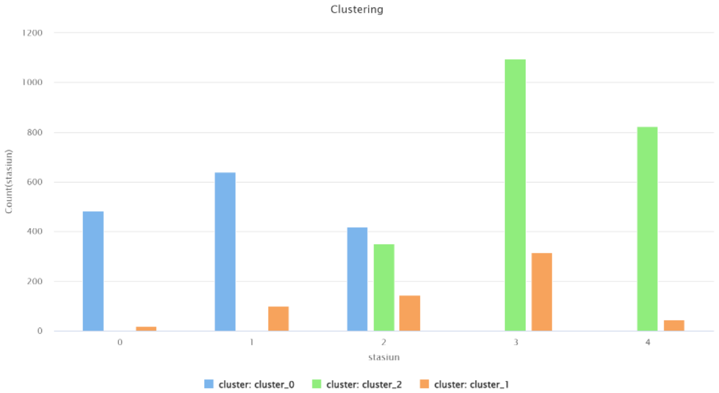


Figure 6. Visualization of Station Distribution Based on K-Means Clustering Results

Figure 6 shows the distribution of stations in each cluster resulting from the modeling process using the K-Means algorithm in rapidminer software. Station 0 is station DKI1 (Bunderan HI), station 1 is station DKI2 (Kelapa Gading), station 2 is station DKI3 (Jagakarsa), station 3 is station DKI4 (Lubang Buaya), and station 4 is station DKI5 (Kebon Jeruk). Based on the visualization of Figure 5 shows that:

a. DKI1 Station (Bundaran HI) falls into the healthy category. Bundaran HI is a city center with high activity, but healthy air quality is likely influenced by the implementation of policies such as the development of public transportation to reduce dependence on private vehicles and the presence of green open spaces that help filter pollutants in the area.

b. Station DKI2 (Kelapa Gading) falls into the healthy category. Although the area is quite congested, the good air quality may be due to effective traffic management and the presence of commercial areas that support the use of public transportation as well as adequate open spaces.

c. DKI3 Station (Jagakarsa) falls into the moderate category. The closer location to dense residential areas and fairly high transportation activities could be the determining factors for this moderate air quality, especially if the lack of green open spaces contributes to the accumulation of pollutants.

d. DKI4 Station (Lubang Buaya) is in the medium category. This condition can be influenced by industrial activities in the vicinity as well as proximity to major highways and toll roads that cause an increase in vehicle emissions.

e. DKI5 Station (Kebon Jeruk) falls into the medium category. This area, which is a mixed area between residential and small industrial areas, is likely to have an impact on moderate air quality, especially if vehicle density factors and local industrial activities are not balanced with pollution control efforts.

Although K-Means succeeded in grouping the data into three clusters that adequately represent air quality conditions, this algorithm has several limitations. These include sensitivity to the initial centroid selection and the assumption that the clusters are spherical and uniform in size, which may not necessarily reflect the real conditions. In addition, K-Means requires determining the number of clusters (K) from the beginning, which can affect the accuracy of the results if not chosen appropriately. To test the validity of the clustering results, the Silhouette Coefficient validity test method was used.

Table 4. Silhouette Coefficient Table

K	Nilai Silhouette Coefficient
2	0.3050
3	0.2129
4	0.2051
5	0.1872
6	0.2104
7	0.2022

Based on Table 4 which shows the results of the validity test using the Silhouette Coefficient for the number of clusters K = 2 to K = 7, it is found that K = 2 has the largest Silhouette Coefficient value, which is 0.3050. This shows that the number of clusters K=2 is the most optimal choice to describe how well the data has been grouped by comparing the distance between points in one cluster.

DISCUSSION

This study aims to categorize air quality in the DKI Jakarta area based on air pollution parameters using the K-Means Clustering algorithm. The data used is secondary data obtained from the official website jakartasatu.jakarta.go.id, covering the time span from 2010 to 2023. The parameters analyzed include PM_{10}, SO_2, CO, O_3, and NO_2, which are the main indicators in air quality assessment according to national and international standards.

The initial stage in this research is data preprocessing to ensure the data is clean, consistent, and can be processed by the algorithm. This process includes removing missing values, normalizing data to have a uniform scale, and selecting relevant attributes for the clustering process.

After the data is ready, the clustering process is carried out using the K-Means Clustering algorithm. Determination of the number of clusters was done using the Elbow method, and it was found that the optimal number of clusters was three (K=3). The clusters represent three categories of air quality, namely:

Cluster 0: Healthy air quality, with relatively low concentrations of pollutants.

1. Cluster 1: Unhealthy air quality, with high concentrations of pollutants in all parameters.
2. Cluster 2: Moderate air quality, with parameter values that vary but do not exceed hazardous thresholds.

Distribusi data pada masing-masing klaster menunjukkan bahwa sebagian besar tahun dengan kualitas udara yang lebih baik berada di klaster 2, sementara beberapa tahun dengan nilai polusi tinggi, khususnya pada parameter PM_{10} dan NO_2, tergolong ke dalam klaster 0. Hal ini menunjukkan bahwa meskipun terdapat peningkatan dalam pengelolaan lingkungan, kualitas udara di DKI Jakarta masih memerlukan perhatian, terutama pada tahun-tahun dengan aktivitas industri dan transportasi yang tinggi.

CONCLUSION

From the results of the research that has been carried out, the following conclusions are obtained:

1. Based on the Elbow method, it is obtained that the best number of clusters is K = 3 for grouping air quality data using the K-Means algorithm with average Cluster Distance Performance and based on Silhouette Coefficient, the best number of clusters is K = 2.
2. Cluster 0 consists of DKI1 and DKI2 areas. This grouping shows relatively good air quality (healthy).
3. Cluster 1 reflects data with poor air quality categories (unhealthy), but is not dominant compared to the other two clusters.
4. Cluster 2 consists of DKI3, DKI4 and DKI5 areas. This grouping indicates moderate air quality.
5. The clustering results provide a systematic and objective picture of the air quality conditions in Jakarta. This information can be utilized by policy makers as a basis for formulating a more targeted air pollution control strategy.
6. Based on the analysis, it shows that the effectiveness of the K-Means Clustering method is highly dependent on the completeness and consistency of the ISPU data.
7. For future research, the addition of pollution parameters such as PM2.5, air temperature, humidity, or distance from pollution sources (e.g. highways or industrial areas), and the AQI index is also highly recommended to obtain a more thorough and accurate representation of air quality.
8. As a follow-up to the limitations of this study, it is recommended that future research consider using quantitative evaluation metrics such as the Davies-Bouldin Index or Calinski-Harabasz Index to objectively measure cluster quality. In addition, the method of determining the number of clusters can be varied with other statistical approaches such as Gap Statistic or agglomerative hierarchy analysis as a comparison.

LIMITATIONS

The results of the implementation of the K-Means Clustering algorithm in this study provide a fairly good picture of air quality grouping, but still have some limitations that need to be considered so that the analysis results can be interpreted appropriately and objectively. The following are the limitations faced, along with explicit suggestions for further research:

1. Validation of cluster results does not involve quantitative evaluation metrics such as the DaviesBouldin Index or Calinski-Harabasz Index, so the quality of clusters is still assessed visually and tends to be subjective. In addition, the determination of the number of clusters only uses the Elbow method which is heuristic and visual. For future research, it is recommended to use these quantitative evaluation metrics to obtain a more objective cluster assessment. The determination of the number of clusters should also be varied with other approaches such as Gap Statistic or agglomerative hierarchy so that the cluster results are more optimal and statistically valid.
2. This research only utilizes five main parameters, namely PM_{10}, SO_2, CO, O_3, and NO_2, according to the availability of data from the sources used. Other important parameters such as PM_{25}, air quality index (AQI), air temperature, humidity, or distance to pollution sources have not been included. Future research is recommended to add these parameters to produce air quality clustering that is more representative and accurate to actual environmental conditions.
3. The analysis did not consider seasonal factors, such as the difference between the rainy and dry seasons, which can significantly affect air pollution levels. For future research, it is strongly recommended to apply a temporal or time series analysis approach, in order to capture the dynamics of air quality changes over time and increase the relevance of clustering results.

REFERENCES

Adiputra, I. N. M. (2021). Clustering Penyakit Dbd Pada Rumah Sakit Dharma Kerti Menggunakan Algoritma K-Means. INSERT: Information System and Emerging Technology Journal, 2(2), 99-105. https://doi.org/10.23887/insert.v2i2.41673

Amalia, A., Zaidiah, A., & Isnainiyah, I. N. (2022). Prediksi Kualitas Udara Menggunakan Algoritma K-Nearest Neighbor. JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika), 7(2), 496–507. https://doi.org/10.29100/jipi.v7i2.2843

Anjelita, M., Windarto, A. P., Wanto, A., & Sudahri, I. (2020). Pengembangan datamining klastering pada kasus pencemaran lingkungan hidup. Semin. Nas. Teknol. Komput. Sains, 1(1), 309-313. https://prosiding.seminar-id.com/index.php/sainteks/article/view/453

Annas, M., & Wahab, S. N. (2023). Data Mining Methods: K-Means Clustering Algorithms. International Journal of Cyber and IT Service Management, 3(1), 40–47. https://doi.org/10.34306/ijcitsm.v3i1.122

Dhewayani, F. N., Amelia, D., Alifah, D. N., Sari, B. N., & Jajuli, M. (2022). Implementasi k-means clustering untuk pengelompokkan daerah rawan bencana kebakaran menggunakan model crisp-dm. Jurnal Teknologi dan Informasi, 12(1), 64-77. https://doi.org/10.34010/jati.v12i1.6674

Hartanti, N. T. (2020). Metode Elbow dan K-Means Guna Mengukur Kesiapan Siswa SMK Dalam Ujian Nasional. J. Nas. Teknol. dan Sist. Inf, 6(2), 82-89. https://doi.org/10.25077/ TEKNOSI.v6i2.2020.82-89

Hasanah, M. A., Soim, S., & Handayani, A. S. (2021). Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. Journal of Applied Informatics and Computing, 5(2), 103–108. https://doi.org/10.30871/jaic.v5i2.3200

Kurniawan, Y. I., Anugrah, P. R., Sugihono, R. M., Abimanyu, F. A., & Afuan, L. (2023). Pengelompokan Prioritas Negara Yang Membutuhkan Bantuan Menggunakan Clustering K-Means dengan Elbow Dan Silhouette. Jurnal Pendidikan Dan Teknologi Indonesia, 3(10), 455-463. https://doi.org/10.52436/1.jpti.343

Marisa, F., Maukar, A. L., & Akhriza, T. M. (2021). Data mining konsep dan penerapannya. Deepublish.

Nugroho, A., Asror, I., & Wibowo, Y. F. A. (2023). Klasifikasi Tingkat Kualitas Udara DKI Jakarta Berdasarkan Open Government Data Menggunakan Algoritma Random Forest. eProceedings of Engineering, 10(2). https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20030/0

Pertiwi, K. D., Lestari, I. P., & Afandi, A. (2024). Analisis Risiko Kesehatan Lingkungan Pajanan Debu PM10 dan PM2. 5 pada Relawan Lalu Lintas di Jalan Diponegoro Ungaran. Pro Health Jurnal Ilmiah Kesehatan, 6(2), 85-91. https://jurnal.unw.ac.id/index.php/PJ/article/view/3351

Prastiwi, H., Pricilia, J., & Rasywir, E. (2022). Implementasi Data Mining Untuk Menentuksn Persediaan Stok Barang Di Mini Market Menggunakan Metode K-Means Clustering. Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM), 2(1), 141-148. https://ejournal.unama.ac.id/index.php/jakakom/article/view/34/58

Salsabila P., B., Belva Cynara Trana Putri, P., Ramadhani, N., & Puspita Sari, A. (2024). Penerapan Algoritma Naive Bayes Terhadap Kualitas Udara Di Jakarta Dan Rekomendasi Aktivitas Masyarakat. JATI (Jurnal Mahasiswa Teknik Informatika), 8(6), 11732–11738. https://doi.org/10.36040/jati.v8i6.11592

Rifqi, A., & Aldisa, R. T. (2023). Penerapan Data Mining Untuk Clustering Kualitas Udara. Jurnal Sistem Komputer Dan Informatika (JSON), 5(2), 289. https://doi.org/10.30865/json.v5i2.7145

Sang, A. I., Sutoyo, E., & Darmawan, I. (2021). Analisis data mining untuk klasifikasi data kualitas udara dki jakarta menggunakan algoritma decision tree dan support vector machine. eProceedings of Engineering, 8(5). https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15900

Siregar, A. H., Sihotang, D. D., Wijaya, B. A., & Siregar, S. D. (2024). Implementasi Algoritma K-Means Menggunakan RapidMiner untuk Klasterisasi Data Obat Pada Rumah Sakit Royal Prima. Jurnal Teknologi Dan Ilmu Komputer Prima (JUTIKOMP), 7(2), 200-211. https://doi.org/10.34012/jutikomp.v7i2.5537

Wardani N. W., Gede, P., Hartono, E., I Wayan Dharma Suryawan, Ayu Manik Dirgayusari, I Wayan Darmadi, & Mahendra, G. S. (2022). Penerapan Data Mining Untuk Klasifikasi Penjualan Barang Terlaris Menggunakan Metode Decision Tree C4.5. Jurnal Teknologi Informasi Dan Komputer, 8(3). https://doi.org/10.36002/jutik.v8i3.2081

Wiranata, A. D., Soleman, S., Irwansyah, I., Sudaryana, I. K., & Rizal, R. (2023). Klasifikasi Data Mining Untuk Menentukan Kualitas Udara Di Provinsi Dki Jakarta Menggunakan Algoritma K-Nearest Neighbors (K-Nn). Infotech: Journal of Technology Information, 9(1), 95-100. https://doi.org/10.37365/jti.v9i1.164