Implementation of the K-Means Algorithm to Determine the Classification of River Water Quality in Jakarta Based on Chemical Parameters

Irpan Kusyadi*, Mayang Anglingsari Putri, Mochamad Bagoes Satria, Denisha Trihapningsari

Universitas Terbuka, Information Systems Study Program, South Tangerang, Banten, Indonesia, 15437

Abstract - This study aims to implement the K-Means algorithm to classify river water quality in Jakarta based on key chemical parameters: Biochemical Oxygen Demand (BOD), nitrate, and nitrite levels. Water quality is a critical issue in Jakarta due to industrial activities and domestic waste contributing to pollution, which poses risks to public health and ecosystems. Data were collected from various monitoring points along the rivers, focusing on the mentioned parameters. The K-Means algorithm was applied to classify the water samples into categories: good, moderate, and poor quality. Results showed that high BOD levels were strongly associated with poor water quality, indicating organic pollution. Elevated nitrate and nitrite levels also contribute to water degradation, reflecting impacts from agricultural runoff and wastewater. The clustering results revealed that the water quality in Jakarta's rivers is predominantly poor, especially in areas with high BOD levels, which indicates organic pollution from domestic and industrial waste. The study demonstrates the K-Means algorithm's effectiveness in analyzing water quality data and suggests its potential as a valuable tool for environmental monitoring. The findings highlight the need for enhanced water quality management in Jakarta and provide a foundation for future research to integrate more parameters and time-based data to better understand trends and support decision-making in pollution control.

Keywords: environmental monitoring, K-Means algorithm, river water quality, water pollution.

1 Introduction

The quality of river water in Jakarta continues to deteriorate along with increasing industrial activities and poorly managed domestic waste. Rivers in Jakarta, which used to be a source of water for the community, are now polluted by hazardous chemicals such as Biochemical Oxygen Demand (BOD), nitrate, and nitrite. This condition poses a serious threat to public health, especially for residents living around riverbanks. In addition, river water pollution has a negative impact on local ecosystems, disrupting environmental balance, and worsening the clean water crisis faced by the city of Jakarta. Therefore, the urgency of this research is very high, because it requires regular monitoring and classification of water quality to assess the level of pollution and develop better control strategies [1].

* Corresponding author: irpan.kusyadi@ecampus.ut.ac.id.

The study by Yuli Handayani et al. [2] applied the K-Means algorithm to cluster rivers in Surakarta based on pollution levels. Among 12 parameters, Total Coliform and Fecal Coliform were key differentiators. Jenes Hilir River showed the poorest water quality due to the highest microbial pollutant levels. This study aims to implement the K-Means algorithm in classifying river water quality in Jakarta based on the main chemical parameters, namely BOD, nitrate, and nitrite. By using this algorithm, the study will group water samples into three categories: good, moderate, and poor quality. This categorization will help map the areas most affected by pollution, as well as identify the main sources of pollution. In addition, this study will also provide an overview of water pollution trends in the Jakarta river basin, which can be the basis for necessary preventive or rehabilitation actions. The use of the K-Means algorithm in the context of water quality monitoring is also expected to be an effective and efficient solution in analyzing complex environmental data.

The outputs of this study include several important things. First, the developed K-Means-based classification model can be used as a tool for periodic water quality monitoring by related agencies. Second, this study will produce a map of river water quality in Jakarta, which shows water categories based on pollution levels. Third, the results of the analysis will provide recommendations that can be used as references by the government in formulating policies related to water management and pollution control. The main contribution of this study is to provide a data-based methodology to support better decision-making related to environmental management, especially in river water quality management in Jakarta. Thus, this study not only contributes academically, but also has a direct impact on sustainable public policies in the future.

2 Materials and methods

2.1 Methods

This research method aims to cluster water quality data using the K-Means algorithm, starting with data collection from sample points that include the sample_point and measurement_result columns. After the data is collected, the next step is to clean and normalize the data to ensure consistency. Determining the optimal number of clusters is done through the Elbow method, before applying K-Means which includes initializing cluster centers, grouping data, and updating cluster centers until convergence is achieved. The clustering results are analyzed through cluster visualization and description to identify water quality patterns, providing insights that can be used in better water resource management.

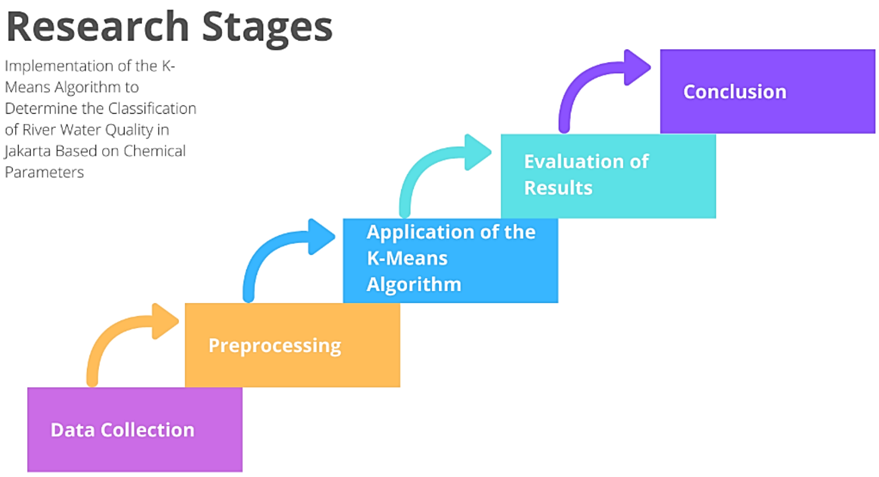

Research Stages
Implementation of the K-Means Algorithm to Determine the Classification of River Water Quality in Jakarta Based on Chemical Parameters
Conclusion
Evaluation of Results
Application of the K-Means Algorithm
Preprocessing
Data Collection

Fig. 1 Research Method

2.2 Water Quality

Water quality refers to the physical, chemical, and biological properties of water that affect its suitability for various purposes, including human health, ecosystems, and industrial activities [3]. In this study, water quality was measured through chemical parameters such as Biochemical Oxygen Demand (BOD), nitrate (Derived from the perfect oxidation process of nitrogen and ammonia compounds in water), and nitrite which is an important indicator of water pollution. BOD reflects the level of organic contamination, while the presence of nitrate and nitrite can indicate the impact of agricultural and domestic waste [4], [5]. Accurate water quality assessment is essential to identify sources of pollution and assist decision making in water resource management, especially in areas affected by pollution, such as Jakarta. Through the application of the K-Means algorithm, this study aims to classify water quality based on the data collected, so that it can provide a clearer picture of the condition of water quality in Jakarta's rivers.

2.3 K-Means

The K-Means algorithm is a data clustering method used in data analysis and machine learning to divide a set of data into groups (clusters) based on similar characteristics [6]. This algorithm works by finding the centroid (center point) of each cluster and grouping data into clusters based on the closest distance between the data and the centroid [7]. The following are the general steps in the K-Means algorithm:

a. Initialization: Determine the desired number of clusters (K) and choose K random points as initial centroids [8].
b. Cluster Assignment: For each data in the dataset, calculate the distance to each centroid and assign the data to the cluster whose centroid is closest.
c. Update Centroid: After all data is assigned to clusters, recalculate the centroid positions by taking the average of all data points in that cluster.
d. Iteration: Repeat the centroid assignment and update steps until the centroid positions do not change significantly or the maximum number of iterations is reached.

K-Means is widely used in various applications, such as market segmentation, document clustering, and image analysis, due to its ability to cluster data efficiently and easily [9]. However, this algorithm has several limitations, such as dependence on the initial centroid selection and the difficulty of determining the optimal number of clusters [10].

3 Results and discussion

Sample Data Set of River Water Quality that can be used for analysis with the K-Means algorithm comes fromhttps://satudata.jakarta.go.id/ [11]. This dataset includes several parameters that measure river water quality, which can be used to determine patterns or groups based on measured water quality conditions. The parameters used in the dataset focus on the type of chemical parameters with BOD, Nitrate and Nitrite parameters.

Table 1. Sample Table of Water Quality in Jakarta

Sample Point	River	Address	Parameter	Measurement Results (mg L)
CLW2-3	Ciliwung	PLN Warehouse	BOD	4.78
CLW2-3	Ciliwung	PLN Warehouse	Nitrate NO3	1.53
CLW2-3	Ciliwung	PLN Warehouse	Nitrite	0.17

Sample Point	River	Address	Parameter	Measurement Results (mg L)
MKV-1	Mookervart	Kalideres Star Beer	BOD	8.18
MKV-1	Mookervart	Kalideres Star Beer	Nitrate NO3	0.08
MKV-1	Mookervart	Kalideres Star Beer	Nitrite	0.05
STR-8	Sunter	Bogasari Koja	BOD	25.46
STR-8	Sunter	Bogasari Koja	Nitrate NO3	0.13
STR-8	Sunter	Bogasari Koja	Nitrite	0.06
KLB-4	West Kalibaru	In front of Pasar Minggu Terminal	BOD	15.86
KLB-4	West Kalibaru	In front of Pasar Minggu Terminal	Nitrate NO3	0.81
KLB-4	West Kalibaru	In front of Pasar Minggu Terminal	Nitrite	0.25
KLT-7	East Kalibaru	Line	BOD	25.01
KLT-7	East Kalibaru	Line	Nitrate NO3	0.25
KLT-7	East Kalibaru	Line	Nitrite	0.04
KNT-1	East Channel	BKT Cipinang Indah Housing Complex	BOD	12.58
KNT-1	East Channel	BKT Cipinang Indah Housing Complex	Nitrate NO3	0.31
KNT-1	East Channel	BKT Cipinang Indah Housing Complex	Nitrite	0.06
KNT-2	East Channel	BKT Jl.	BOD	9.55
KNT-2	East Channel	BKT Jl.	Nitrate NO3	0.16
KNT-2	East Channel	BKT Jl.	Nitrite	0.06
KNT-3	East Channel	BKT Marunda Water Gate	BOD	15.55
KNT-3	East Channel	BKT Marunda Water Gate	Nitrate NO3	0.4
KNT-3	East Channel	BKT Marunda Water Gate	Nitrite	0.009
CDG-3	The Cider	Front of GD Allianz KPK	BOD	45.31
CDG-3	The Cider	Front of GD Allianz KPK	Nitrate NO3	1.12
CDG-3	The Cider	Front of GD Allianz KPK	Nitrite	0.03
MPG-2	Mampang	Near the exit of Pasar Minggu Regional Hospital	BOD	14.69
MPG-2	Mampang	Near the exit of Pasar Minggu Regional Hospital	Nitrate NO3	0.33
MPG-2	Mampang	Near the exit of Pasar Minggu Regional Hospital	Nitrite	0.04

K-Means is a clustering algorithm that divides data into groups based on similarities. The steps in this algorithm include:

1. Specify Number of Clusters (): Specifies how many clusters you want to form k

From the data above, the author first determines the optimal number of clusters. At this stage, the author uses the Elbow method. Based on the Elbow method, the following results can be obtained.

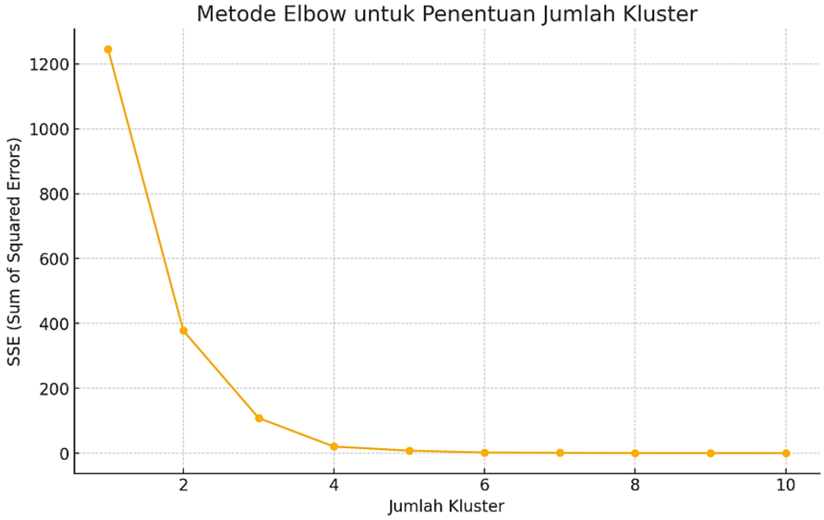


Fig. 2. Elbow Method to Determine Number of Cluster

From this graph, the optimal number of clusters for this water quality data appears to be 3, as this is the point where the curve begins to flatten out, indicating that we are reaching a point where adding further clusters is inefficient in reducing the variability of the data. This is a good number of clusters to strike a balance between model complexity and the ability to distinguish different groups in the data.

2. Centroid Initialization: Choose an initial position for the centroid of each cluster. This can be done randomly from the existing data points.
3. Calculating Distance: Calculate the distance of each data point from the centroid using the Euclidean distance formula:

$$Jarak = \sqrt{(x_i - c_j)^2} \quad (1)$$

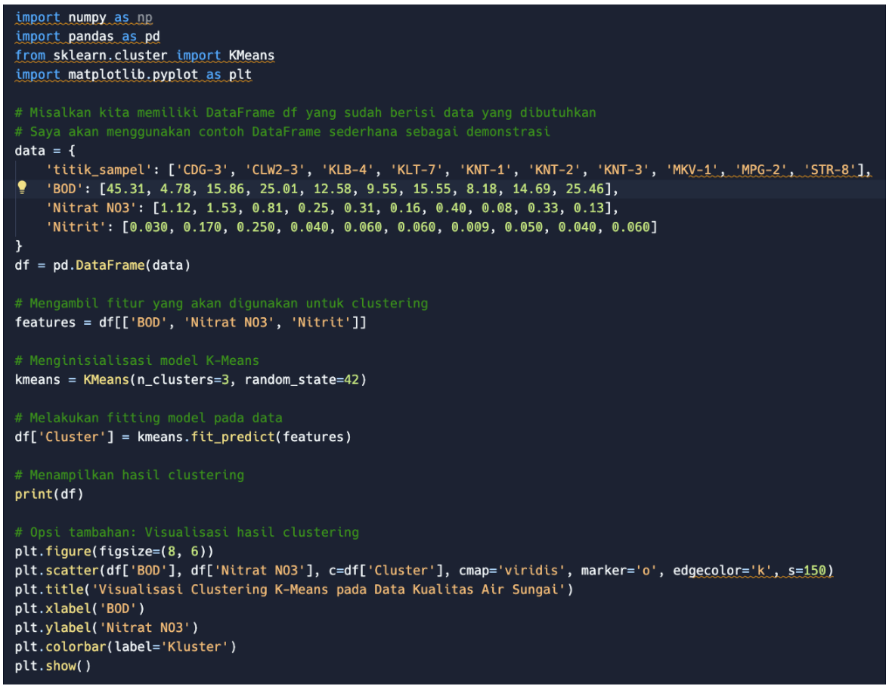
where is the value of the data point and is the value of the centroid of the th cluster.$x_i c_j$ *j*.

4. Data Clustering: After calculating the distance, each data point will be grouped into the cluster whose centroid is closest.
5. Updating Centroids: After clustering, the centroids are updated by taking the average of all data points in the cluster:

$$\mathrm{c_j} = \frac{1}{N_j}\sum_{i=1}^{N_j} x_1 \quad (2)$$

where is the number of data points in the th cluster $N_j j$.

From the formula above, the author implements the K-Means algorithm, which is used to group data into several clusters based on the proximity between data points. The process begins by determining the desired number of clusters, where each cluster is represented by a randomly generated centroid. This algorithm iterates with the following steps: first, each data point will be assigned to the nearest cluster based on the Euclidean distance from the centroid. After all data points are classified, a new centroid is calculated as the average of all points in the cluster. This process is repeated until the centroid does not change significantly, indicating that convergence has been achieved. The end result of this algorithm is a well-organized group of data points, providing better insight into the data structure. This implementation allows users to understand patterns in the data more clearly and provides a basis for further analysis.



```python
import numpy as np
import pandas as pd
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# Misalkan kita memiliki DataFrame df yang sudah berisi data yang dibutuhkan
# Saya akan menggunakan contoh DataFrame sederhana sebagai demonstrasi
data = {
    'titik_sampel': ['CDG-3', 'CLW2-3', 'KLB-4', 'KLT-7', 'KNT-1', 'KNT-2', 'KNT-3', 'MKV-1', 'MPG-2', 'STR-8'],
    'BOD': [45.31, 4.78, 15.86, 25.01, 12.58, 9.55, 15.55, 8.18, 14.69, 25.46],
    'Nitrat NO3': [1.12, 1.53, 0.81, 0.25, 0.31, 0.16, 0.40, 0.08, 0.33, 0.13],
    'Nitrit': [0.030, 0.170, 0.250, 0.040, 0.060, 0.060, 0.009, 0.050, 0.040, 0.060]
}
df = pd.DataFrame(data)

# Mengambil fitur yang akan digunakan untuk clustering
features = df[['BOD', 'Nitrat NO3', 'Nitrit']]

# Menginisialisasi model K-Means
kmeans = KMeans(n_clusters=3, random_state=42)

# Melakukan fitting model pada data
df['Cluster'] = kmeans.fit_predict(features)

# Menampilkan hasil clustering
print(df)

# Opsi tambahan: Visualisasi hasil clustering
plt.figure(figsize=(8, 6))
plt.scatter(df['BOD'], df['Nitrat NO3'], c=df['Cluster'], cmap='viridis', marker='o', edgecolor='k', s=150)
plt.title('Visualisasi Clustering K-Means pada Data Kualitas Air Sungai')
plt.xlabel('BOD')
plt.ylabel('Nitrat NO3')
plt.colorbar(label='Kluster')
plt.show()
```

Fig. 3 Program Code using K Means

Data in Table 2 is the result of the code in Fig. 3.

Table 2. K-Means table with 2 clusters

Sample Point	BOD	Nitrate NO3	Nitrite	Cluster
CDG-3	45.31	1.12	0.030	1
CLW2-3	4.78	1.53	0.170	0
KLB-4	15.86	0.81	0.250	0
KLT-7	25.01	0.25	0.040	2
KNT-1	12.58	0.31	0.060	0
KNT-2	9.55	0.16	0.060	0
KNT-3	15.55	0.40	0.009	0
MKV-1	8.18	0.08	0.050	0
MPG-2	14.69	0.33	0.040	0
STR-8	25.46	0.13	0.060	2

The data in Table is then poured into a scatter plot graph with the following results.

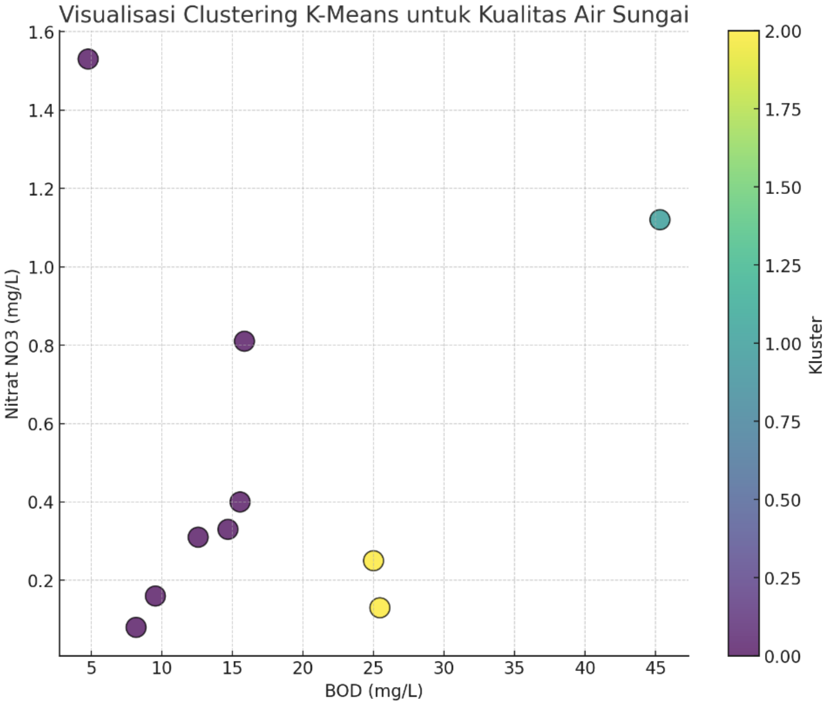


Fig. 4. Visualization of K-Means Clustering for River Water Quality

This image illustrates the visualization of clustering results using the K-Means algorithm to evaluate river water quality based on two parameters: Biochemical Oxygen Demand (BOD) in mg/L and Nitrate (NO3) in mg/L. The points on the graph represent data samples grouped into three distinct clusters, marked by different colors (purple, yellow, and cyan-green). The horizontal axis (X-axis) shows the BOD values, while the vertical axis (Y-axis) displays the Nitrate values. The color scale on the right side of the graph indicates the clusters formed during the clustering process.

4 Conclusion

This study successfully implemented the K-Means algorithm to classify river water quality in Jakarta based on key chemical parameters, namely Biochemical Oxygen Demand (BOD), nitrate, and nitrite levels. The clustering results revealed that the water quality in Jakarta's rivers is predominantly poor, especially in areas with high BOD levels, which indicates organic pollution from domestic and industrial waste. Furthermore, elevated nitrate and nitrite levels also contribute to water quality degradation, reflecting agricultural runoff and other pollution sources.

The K-Means algorithm proved effective in clustering water quality data and providing insights into the most affected areas. The classification of water into three categories (good, moderate, and poor) serves as an important tool for policymakers and relevant agencies in monitoring river conditions and designing strategies to mitigate pollution. This study highlights the urgent need for improved water quality management in Jakarta and suggests that future research should incorporate more parameters and time-based data to achieve a more comprehensive understanding of water pollution trends. Overall, this research provides a significant contribution by utilizing a data-driven approach to support decision-making in water resource management and pollution control efforts in Jakarta.

References

[1] S. Yudo, "Kondisi Kualitas Air Sungai Ciliwung di Wilayah DKI Jakarta Ditinjau dari Parameter Organik, Amoniak, Fosfat, Detergen dan Bakteri Coli," JAI, vol. 6, no. 1, pp. 34-44, 2010.

[2] Y. Handayani, S. Syafrudin, dan S. Suherman, "Analisis Klaster K-Means Terhadap Kualitas Air Sungai di Kota Surakarta, Jawa Tengah" (K-Means Cluster Analysis on River Water Quality in Surakarta City, Central Java), 2021.

[3] Z. W. Zhao and F. Y. Cui, "Multivariate Statistical Analysis for the Surface Water Quality of the Luan River, China," Journal of Zhejiang University: Science A, vol. 10, no. 1, pp. 142–148, 2009. https://doi.org/10.1631/jzus.A0850094

[4] D. Dimri, A. Daverey, A. Kumar, and A. Sharma, "Monitoring Water Quality of River Ganga Using Multivariate Techniques and WQI (Water Quality Index) in Western Himalayan Region of Uttarakhand, India," Environmental Nanotechnology, Monitoring and Management, vol. 15, pp. 100375, July 2020. https://doi.org/10.1016/j.enmm.2020.100375

[5] J. N. Landhiani, "Analisis Kualitas Air Sungai Pepe dan Sungai Anyar yang Melintas di Kota Surakarta," Universitas Sebelas Maret, 2016.

[6] S. Javadi, S. Hashemy, K. Mohammadi, K. W. Howard, and A. Neshat, "Classification of Aquifer Vulnerability Using K-means Cluster Analysis," Journal of Hydrology, vol. 549, pp. 27–37, 2017.

[7] R. P. Rahmadina, Y. Arien, and Sutikno, "Analisis Cluster Virus Corona (COVID-19) di Indonesia pada 2 Maret 2020 – 12 April 2020 dengan Metode K-Means Clustering," 2020.

[8] H. Suryono and D. W. Triscowati, "Pengelompokan Provinsi di Indonesia Berdasarkan Resiko Covid-19 dan Ketahanan Pangan," Seminar Nasional Official Statistics 2020: Statistics in New Normal: A Challenge of Big Data and Official Statistics, 2020.

[9] H. Jatnika, M. Huda, R. R. Amelia, M. A. Manuhutu, A. P. Windarto, P. Sumantrie, and A. Waluyo, "Analysis of Data Mining in the Group of Water Pollution Areas Using the K-Means Method in Indonesia," Journal of Physics: Conference Series

[10]D. T. Sukmawati, "Penentuan Kategori Kualitas Air Sungai dengan Menggunakan Metode K-Means," Universitas Islam Majapahit, 2019.

[11]Satu Data, "Sample Data Set Kualitas Air Sungai," diakses dari https://satudata.jakarta.go.id/ pada 1 Oktober 2024.