ISSN 2797-7889

Performance of the K-Means Algorithm for Water Quality Clustering

Siska Simamora[1], Paska Marto Hasugian[2]
[1]Program Studi Teknologi Informasi, Universitas Putra Abadi Langkat, Sumatera Utara
[2]Program Studi Sains data, Fakultas Ilmu Komputer, Universitas Katolik Santo Thomas, Sumatera Utara

Clustering is an unsupervised learning technique used to group data based on the degree of similarity among object characteristics. This study aims to analyze the application of a distance formula in cluster formation using the K-Means algorithm on the Water Quality Dataset. The dataset consists of 7,999 observations and 20 columns representing various water quality characteristics. The target column, is_safe, was removed, resulting in 19 features used in the clustering process. The preprocessing stages included checking for duplicate data, handling missing values, converting data into numerical format, and applying Min-Max normalization within the range of [0,1]. Normalization was performed to standardize the scale across features, ensuring that each feature contributed proportionally to the distance calculation. The clustering process was conducted using the K-Means algorithm, with data proximity determined based on the distance formula. The results indicate that data preprocessing and the selection of an appropriate distance formula are important factors in determining proximity patterns among objects and the resulting cluster formation. The use of normalized data can reduce the dominance of features with larger value ranges, thereby enabling the clustering process to represent data characteristics more proportionally. This study demonstrates that distance formula analysis plays an important role in supporting the formation of representative clusters in water quality data.
Keywords: K-Means, Clustering, Distance Formula, Water Quality, Min-Max Normalization.

Corresponding Author:
Siska Simamora
Program Studi Teknologi Informasi, Universitas Putra Abadi Langkat, Sumatera Utara

1. Introduction

Water is one of the most essential natural resources, playing a fundamental role in human life, public health, ecosystem sustainability, and socioeconomic development. Water availability is determined not only by quantity but also by quality. Water quality may change due to natural processes and human activities, including industrial operations, agriculture, mining, urbanization, and domestic waste disposal. These factors can affect the physical, chemical, and biological characteristics of water, making water quality monitoring an essential component of sustainable water resource management [1], [2]. The increasing intensity of human activities and the growing complexity of pollution sources have also created a need for analytical methods capable of processing water quality data more effectively and systematically [3].

Water quality data generally exhibit multivariate characteristics because they consist of multiple parameters measured simultaneously. Parameters such as aluminum, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, and bacteria may have different distributions and value ranges. As the number of observations and attributes increases, identifying patterns using conventional approaches becomes increasingly complex. The development of machine learning-based approaches provides opportunities to analyze large-scale water resource data and identify patterns that may be difficult to detect through conventional analysis [4]. These computational approaches enable the exploration of data structures based on the relationships and characteristics inherent in the data.

One approach that can be used to identify patterns in multivariate data is clustering. Clustering is an unsupervised learning technique that aims to group objects based on their degree of similarity or

dissimilarity without using predefined class labels as a reference for group formation [5]. The primary principle of clustering is to maximize the similarity among objects within the same cluster while increasing the differences between clusters. Through this approach, hidden structures within the data can be identified based on naturally occurring internal patterns. Clustering has become an important method for data exploration and pattern recognition across various types of multidimensional data [5], [6].

K-Means is one of the most widely used clustering algorithms for processing numerical data. The fundamental concept of this algorithm was introduced through the development of partition-based clustering methods that optimize the assignment of objects to cluster centers [7], [8]. K-Means operates by determining a predefined number of cluster centers, or centroids, and subsequently assigning each object to its nearest centroid. Once all objects have been assigned to clusters, the centroid positions are updated based on the mean values of the objects belonging to each cluster. The assignment and centroid update processes are performed iteratively until no significant changes occur or the algorithm reaches convergence. Its relatively simple mechanism and computational efficiency make K-Means particularly suitable for exploring structures in large numerical datasets [9].

Nevertheless, the performance of K-Means is strongly influenced by the characteristics of the data being analyzed. Differences in feature scales may cause attributes with larger value ranges to exert a dominant influence on distance calculations. This condition can affect the assignment of objects to centroids and ultimately alter the resulting cluster structure. Studies examining the effect of feature scaling have demonstrated that scaling is an important step in the implementation of K-Means, particularly when a dataset contains features with different units or value ranges [10]. Therefore, data preprocessing, including normalization, is necessary to standardize feature scales so that each variable can contribute more proportionally to determining the proximity of objects to their respective centroids.

In addition to data characteristics and feature scales, K-Means performance is also influenced by the determination of the number of clusters, the initialization of the initial centroids, and the iterative process used to achieve convergence [9]. Evaluation of clustering results is therefore necessary to determine the extent to which the algorithm can produce compact and well-separated cluster structures. Several indicators can be employed to evaluate clustering quality, including the Sum of Squared Error (SSE), Silhouette Coefficient, and Davies-Bouldin Index [11], [12]. SSE, also referred to as Inertia, represents the total squared distance between each object and the centroid of its corresponding cluster, thereby providing information regarding the internal compactness of the resulting clusters.

Based on these considerations, this study applies the K-Means algorithm to the Water Quality Dataset, which consists of 7,999 observations with multivariate characteristics. The data are initially prepared through cleaning and normalization processes prior to clustering. The analysis focuses on the performance of the algorithm in forming three clusters based on water quality characteristics by considering its convergence capability, the number of iterations required, the SSE value, and the distribution of cluster memberships. The analysis is expected to provide empirical insights into the ability of K-Means to identify meaningful cluster structures in water quality data containing multiple features with diverse characteristics and value ranges.

2. Literature Review

The application of the K-Means algorithm in water quality analysis has been investigated in various studies involving different data characteristics, parameters, and analytical objectives. Clustering approaches provide an opportunity to identify water quality patterns without relying on predefined class labels. In the context of water resource management, the use of machine learning approaches continues to expand due

to their ability to analyze multivariate data and identify complex relationships among parameters [4]. Zubaidah, Karnaningroem, and Slamet [13] applied K-Means to classify river water quality status in Banjarmasin. The parameters used included Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Total Suspended Solid (TSS), and Dissolved Oxygen (DO), with data collected from eight monitoring stations. The study produced four water quality status clusters and identified several stations categorized as heavily polluted. These findings demonstrate that K-Means can be used to identify groups of water quality conditions and locations vulnerable to pollution. However, the analysis primarily focused on interpreting water quality status rather than evaluating the convergence process and internal performance of the K-Means algorithm.

Novianta, Syafrudin, and Warsito [14] applied K-Means to group rivers in the Special Region of Yogyakarta based on water quality parameters. The study used several parameters, including TSS, DO, BOD, COD, phosphate, Fecal Coli, and Total Coliform. The approach demonstrated that K-Means can be used to form groups of rivers based on similarities in their water quality characteristics. This study reinforces the use of clustering as a method for exploring multivariate environmental data, although the evaluation of the number of iterations, convergence process, and SSE value as indicators of algorithm performance was not the primary focus. Arslan, Parim, and Cene [15] compared K-Means and Fuzzy C-Means algorithms for clustering water quality parameters from 17 monitoring stations in the Ergene Basin. A total of 15 water quality parameters were used to form groups representing different water quality conditions. The results showed that clustering approaches were able to distinguish water quality characteristics based on combinations of multiple parameters. Nevertheless, the study placed greater emphasis on comparing the results of K-Means and Fuzzy C-Means than on specifically evaluating the convergence efficiency and internal performance of the K-Means algorithm.

A recent study on water quality patterns in the Bengawan Solo River also applied K-Means to physicochemical and environmental parameter data. The research stages included feature selection, preprocessing, categorical data encoding, Z-score standardization, and testing the number of clusters from K=2 to K=5. The evaluation employed the Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Score, and Inertia. The evaluation results indicated that K=2 provided the best clustering configuration for the dataset used. This study highlights the importance of quantitative evaluation using multiple indicators to determine the quality of the resulting cluster structure.

Based on the synthesis of previous studies, K-Means has demonstrated its ability to identify patterns and group water quality data based on multivariate characteristics. However, each study has a different orientation, including pollution status identification, grouping of monitoring locations, comparison of clustering methods, and determination of the optimal number of clusters. Furthermore, the importance of normalization in K-Means should also be considered because differences in feature scales can significantly affect clustering results [10]. Therefore, the problem statement of this study lies in the need to evaluate the performance of the K-Means algorithm on multivariate water quality data in a more integrated manner by considering data preprocessing and normalization, the convergence process, the number of iterations required, Inertia or SSE as a measure of internal cluster compactness, and the distribution of cluster membership. The evaluation is conducted on the Water Quality Dataset, consisting of 7,999 observations and 19 features after the target column is removed. This approach is expected to provide empirical insights into the performance of K-Means in forming cluster structures in water quality data with a relatively large number of observations and diverse feature characteristics.

3. Method

This study employs an experimental approach to analyze the performance of the K-Means algorithm in clustering water quality data. The research process is conducted through several stages, including data collection and analysis, data preprocessing, K-Means parameter configuration, the clustering process, and evaluation of clustering results. Each stage is carried out systematically to ensure that the data used have the appropriate quality and scale required by the algorithm.

Data Requirements and Analysis

The dataset used in this study is the Water Quality Dataset, consisting of 7,999 observations and 20 attributes representing various water quality characteristics. The dataset includes several parameters, such as aluminum, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, and bacteria, as well as other parameters. In the initial stage, the dataset structure is examined to identify data types, the number of non-null values, and the minimum and maximum value ranges of each attribute. The target column, is_safe, is removed before the clustering process because K-Means is an unsupervised learning algorithm that does not require class label information. After removing the target column, 19 features are used as variables in the clustering process.

Data Preprocessing

The preprocessing stage is performed to ensure that the data are in an appropriate condition before being used in the clustering process. This stage includes checking and removing duplicate data, identifying missing values, converting attributes into numerical format, and normalizing the data. Based on the examination results, no duplicate data or missing values are found; therefore, the number of observations remains at 7,999. Subsequently, all features are normalized using Min-Max Normalization within the range [0,1]. Normalization is required because each feature has a different value range. This process aims to standardize the scale across features so that features with larger value ranges do not dominate the distance calculation process. Mathematically, Min-Max Normalization is formulated as:

$$x' = (x - x\_min) / (x\_max - x\_min)$$

where x' represents the normalized value, x is the original value, while x_min and x_max represent the minimum and maximum values of the corresponding feature, respectively.

K-Means Clustering Configuration and Process

The clustering process is performed using the K-Means algorithm by setting the number of clusters to K = 3, the maximum number of iterations to 100, and the random seed value to 42. The use of a random seed aims to maintain the consistency and reproducibility of the initial centroid initialization process. In the initial stage, centroids are determined as the initial centers of each cluster. Subsequently, each object is assigned to a cluster based on the shortest distance to the centroid. After all objects have been assigned to their respective clusters, the centroids are updated based on the mean values of all objects belonging to each cluster. The object assignment and centroid update processes are repeated iteratively until the centroids reach a stable condition or the number of iterations reaches the predefined maximum limit.

Clustering Performance Evaluation

The performance of the algorithm is evaluated based on algorithm convergence, the number of iterations, Inertia or Sum of Squared Error (SSE), and the distribution of cluster membership. Convergence is used to determine the ability of the algorithm to reach a stable condition before the maximum iteration limit is reached. Meanwhile, SSE is used to measure the compactness of objects within each cluster relative to their respective centroids. Mathematically, SSE is formulated as:

$$SSE = \sum_{j=1}^{K} \sum_{xi \in C_j} \|x_i - \mu_j\|^2$$

where K represents the number of clusters, C_j denotes the j-th cluster, x_i represents an object belonging to the cluster, and μ_j denotes the centroid of the j-th cluster. A lower SSE value indicates that the distances between objects and their respective cluster centroids are smaller, resulting in more compact clusters. In addition, the distribution of the number of members in each cluster is analyzed to identify the object allocation patterns produced by the K-Means algorithm. The overall evaluation results are used to describe the performance of the algorithm in forming cluster structures within the water quality data.

4. Result And Discussion

Data Requirements and Analysis

The dataset used in this study is the Water Quality Dataset, consisting of 7,999 observations with various attributes representing water quality characteristics. The dataset is loaded into the application through a data source link or by using a locally stored CSV file. The target column, is_safe, is removed because the K-Means algorithm is an unsupervised learning method that does not require class labels. Subsequently, an initial analysis is conducted to identify attribute names, data types, the number of non-null values, and the minimum and maximum values of each attribute.

Table 1. Initial Analysis Results of the Water Quality Dataset

No.	Attribute	Data Type	Non-Null	Min	Max
1	aluminium	float64	7,999	0.00000	5.05000
2	ammonia	object	7,999	–	–
3	arsenic	float64	7,999	0.00000	1.05000
4	barium	float64	7,999	0.00000	4.94000
5	cadmium	float64	7,999	0.00000	0.13000
6	chloramine	float64	7,999	0.00000	8.68000
7	chromium	float64	7,999	0.00000	0.90000
8	copper	float64	7,999	0.00000	2.00000
9	fluoride	float64	7,999	0.00000	1.50000
10	bacteria	float64	7,999	0.00000	1.00000

Based on the initial analysis, most attributes are numerical and have the float64 data type, whereas the ammonia attribute is identified as an object data type and therefore needs to be converted into numerical format. In addition, differences in value ranges among attributes may affect distance calculations. Therefore, the data subsequently undergo a preprocessing stage that includes removing duplicate data, handling missing values using the median, converting data into numerical format, and applying Min-Max normalization within the range [0,1]. These stages are performed to produce clean data with a uniform scale before being used in the K-Means clustering process.

Data Preprocessing

During the preprocessing stage, the data are cleaned and transformed before being used in the K-Means clustering process. The procedures include removing duplicate rows, handling missing values using the median, and applying Min-Max normalization within the range [0,1].

Table 2. Data Preprocessing Results

No.	Stage	Result
1	Initial data	7,999 rows × 20 columns
2	Duplicate data removed	0 rows
3	Missing values imputed with median	0 values

 ISSN 2797-7889

No.	Stage	Result
4	Normalization	Min-Max [0,1]
5	Data ready for clustering	7,999 rows × 19 features

Based on the preprocessing results, the initial dataset consists of 7,999 rows and 20 columns. No duplicate rows or missing values are identified; therefore, there is no reduction in the number of observations. After removing the target column, is_safe, 19 features are retained for the clustering process. Subsequently, all features are normalized using Min-Max normalization within the range [0,1] to standardize the value ranges across features, ensuring that each feature contributes proportionally to the distance calculation in the K-Means clustering process.

K-Means Implementation

At this stage, the K-Means algorithm parameters are configured through a Graphical User Interface (GUI) by setting the number of clusters to K = 3, the maximum number of iterations to 100, and the random seed value to 42. The use of a random seed aims to ensure the consistency and reproducibility of the initial centroid initialization process. Once the clustering process is executed, the algorithm iteratively assigns each object to a cluster based on its proximity to the centroid and subsequently updates the centroid position based on the mean of the objects belonging to each cluster. This process continues until no significant changes occur in the centroid positions or the convergence condition is reached. The experimental results show that the algorithm converges at the 32nd iteration out of the maximum 100 iterations specified, with an Inertia or Sum of Squared Error (SSE) value of 9,743.8. The SSE value represents the total squared distance between each object and the centroid of the cluster to which the object is assigned and is used as an indicator of internal cluster compactness. In general, a lower SSE value indicates that cluster members are located closer to their respective centroids. Based on the resulting cluster membership distribution, Cluster 1 (C1) has the largest number of members, comprising 4,035 observations. This indicates that most objects in the dataset have characteristics that are closer to the C1 centroid than to the centroids of the other clusters, as illustrated in Figure 1.

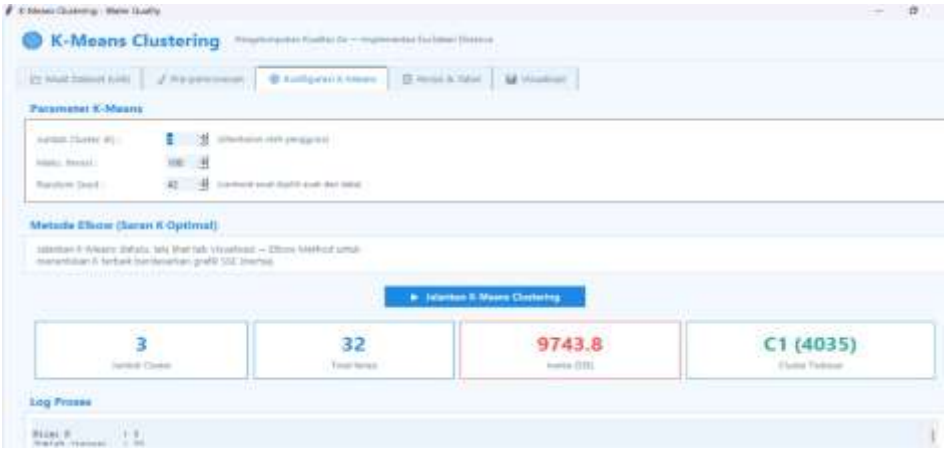

Figure 1. Cluster Formation

Iteration Details and Clustering Table

The application also provides an interactive feature in the “Iteration & Table” tab, allowing users to examine the clustering process at each iteration in detail. Through the “Iteration” input control, users can select a specific iteration number (e.g., the first iteration) to be displayed or click the “Last Iteration” button to directly view the results at the converged iteration (the 32nd iteration). This feature is useful for progressively understanding how the centroid positions and cluster memberships change from one iteration to the next. In the display of Iteration 1 of 32, the initial centroid positions for the first two features are shown as follows:

 ISSN 2797-7889

C1: [0.00594, 0.00000, ...], C2: [0.05347, 0.03810, ...], and C3: [0.04158, 0.15238, ...]. Below the centroid information, a table presents the Euclidean distance of each data point to the three centroids (columns C1, C2, and C3), along with the minimum distance (the Nearest column) and the resulting cluster assignment (the Cluster column). For example, the first data point has distances of 2.30497 to C1, 1.78792 to C2, and 1.97551 to C3. Since the distance to C2 is the smallest, the data point is assigned to Cluster 2. The same mechanism is applied to all data points. Thus, the table directly demonstrates the application of the Euclidean Distance formula and the principle that the shortest distance determines cluster membership, which constitutes the core assignment mechanism of the K-Means algorithm.

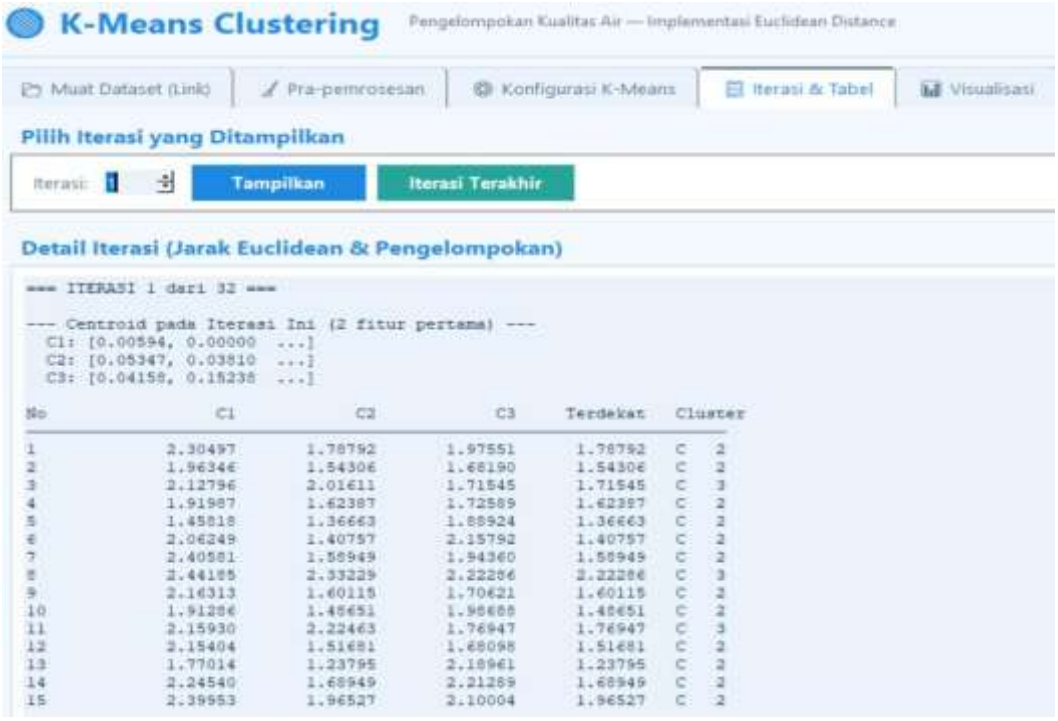

Figure 2. Iteration Details and Clustering Table

This inter-iteration navigation feature provides added value from a learning perspective, as users are not limited to viewing only the final clustering results but can also transparently observe how the algorithm operates at each stage. This includes the distance calculation process, the assignment of data points to the nearest cluster, and the iterative updating of centroid positions until the algorithm ultimately reaches convergence at the 32nd iteration.

Visualization of Clustering Results

The final clustering results are visualized using the Principal Component Analysis (PCA) dimensionality reduction technique, reducing the data to two dimensions (PC1 and PC2), so that the distribution of the 19 data features can be represented in a two-dimensional space that is easier to interpret. In the resulting scatter plot, three groups of data are represented by different colors, namely Cluster 1 (blue), Cluster 2 (green), and Cluster 3 (orange), along with markers indicating the final centroid positions of each cluster (star symbols).

 ISSN 2797-7889

Figure 3. Visualization of Clustering Results

The visualization shows that Cluster 1 (blue) tends to occupy regions with low to moderate PC1 values and is distributed across a relatively wide range of PC2 values, from approximately -0.5 to 1.0. Cluster 2 (green) is located in regions with moderate to high PC1 values and positive PC2 values, whereas Cluster 3 (orange) occupies regions with moderate to high PC1 values but negative PC2 values. The three clusters exhibit relatively clear separation, particularly along the PC2 axis. However, slight overlap is observed at several points near the boundaries between clusters, especially between Cluster 1 and Clusters 2 and 3. This is reasonable because the two-dimensional PCA projection represents only a portion of the total variance contained in the original 19-dimensional data.

Elbow Method

The Elbow Method is used to determine the optimal number of clusters (K) for the K-Means algorithm. The graph illustrates the relationship between the number of clusters (K) and the Sum of Squared Error (SSE), also referred to as Inertia. A lower SSE value generally indicates greater within-cluster compactness. Based on the graph, the SSE decreases substantially up to K = 3, after which the rate of decrease becomes more gradual. This pattern indicates an elbow point at K = 3, suggesting that three clusters provide an appropriate balance between cluster compactness and model complexity. The red point on the graph indicates that the application uses three clusters as the final configuration of the clustering process. Thus, the water quality data are grouped into three clusters without increasing the number of clusters beyond the point at which additional clusters provide only marginal reductions in SSE.

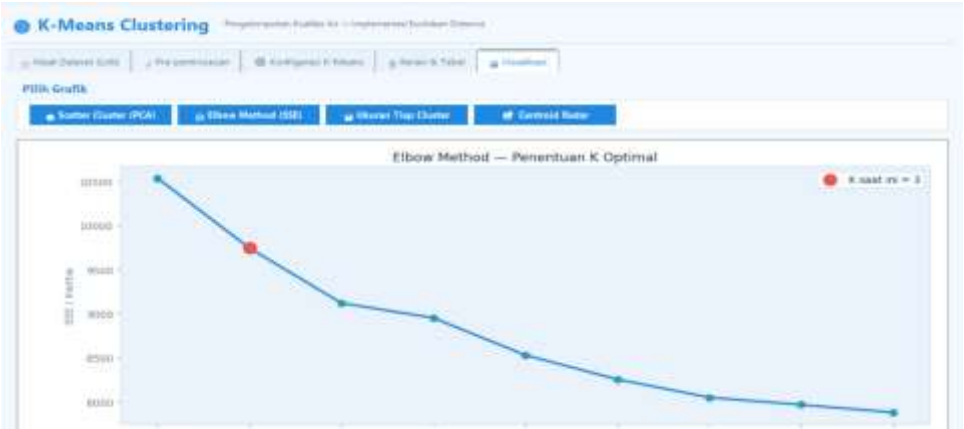

Figure 4. Elbow Method

Cluster Size

The Cluster Size visualization presents the number of observations per cluster and the proportion of each cluster after applying the K-Means algorithm with three clusters. The bar chart displays the number of observations in each cluster, while the pie chart illustrates the percentage contribution of each cluster to the entire dataset. Based on the results, Cluster 1 contains the largest number of observations, with 4,035

 ISSN 2797-7889

data points (50.4%), followed by Cluster 3 with 2,296 data points (28.7%), and Cluster 2 with 1,668 data points (20.9%). These results indicate that approximately half of the water quality observations share characteristics that place them in Cluster 1, while the remaining observations are distributed between Clusters 2 and 3 according to their similarity to the corresponding centroids. Thus, the K-Means algorithm successfully partitions the water quality dataset into three groups with distinct multivariate characteristics.

Figure 5. Number and Proportion of Data per Cluster

Centroid Points

The Centroid Radar chart presents the centroid values, representing the mean normalized value of each water quality attribute for each cluster generated by the K-Means algorithm. Each line represents Cluster 1, Cluster 2, or Cluster 3, while each axis corresponds to a water quality attribute used in the clustering process. Based on the graph, each cluster exhibits a distinct centroid profile. Cluster 1 shows lower values for most attributes, whereas Clusters 2 and 3 exhibit higher values for several attributes, including barium, chloramine, chromium, and fluoride. These differences in centroid values indicate that each cluster represents a distinct water quality profile. Therefore, the Centroid Radar visualization supports the interpretation that the K-Means algorithm partitions the water quality data into three clusters based on similarities and differences across the analyzed attributes.

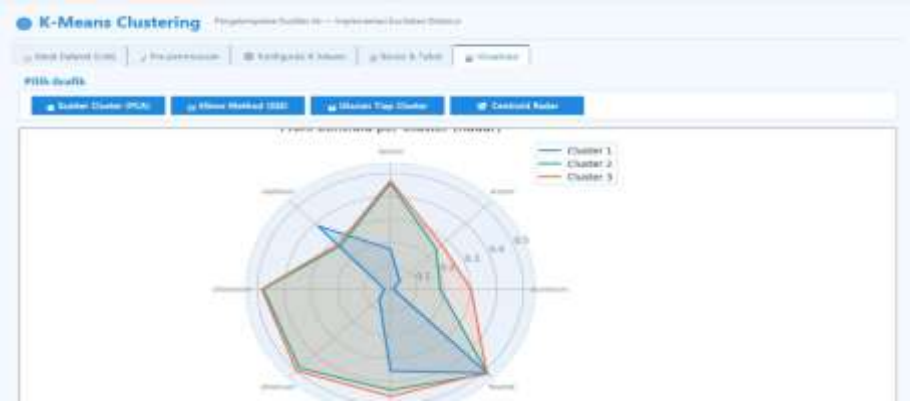

Figure 6. Centroid Radar

5. Conclusion

Based on the analysis conducted in this study, the K-Means clustering algorithm demonstrates its effectiveness in identifying patterns and forming groups within multivariate water quality data. The algorithm remains widely used because of its relatively simple computational mechanism, ease of implementation, and computational efficiency, making it applicable to various domains, including customer analysis, image analysis, bioinformatics, recommendation systems, and environmental data analysis. Nevertheless, conventional K-Means has several limitations, including sensitivity to initial centroid selection,

the requirement to specify the number of clusters before clustering, and dependence on the characteristics and scale of the input data. Outliers and differences in feature scales may also influence distance calculations and potentially affect the quality of the resulting clusters. Therefore, appropriate preprocessing procedures, particularly normalization or standardization, are essential when applying K-Means to multivariate datasets with heterogeneous feature ranges.

In this study, Min-Max normalization was applied to standardize the scale of 19 features before clustering. The K-Means algorithm, configured with K = 3, a maximum of 100 iterations, and a random seed of 42, reached convergence at the 32nd iteration with an Inertia or SSE value of 9,743.8. The resulting cluster distribution consisted of 4,035 observations (50.4%) in Cluster 1, 1,668 observations (20.9%) in Cluster 2, and 2,296 observations (28.7%) in Cluster 3. PCA-based visualization and centroid analysis further indicated differences in the multivariate characteristics represented by the three clusters. These findings demonstrate that K-Means can provide a systematic approach for exploring the underlying structure of water quality data when supported by appropriate preprocessing and quantitative evaluation.

Overall, no single clustering configuration can be considered universally optimal for all datasets and analytical problems. The selection of an appropriate clustering technique and parameter configuration should consider the characteristics of the data, analytical objectives, dataset size, feature distribution, and computational requirements. Future studies may extend this work by comparing K-Means with alternative or enhanced clustering approaches, applying additional internal validation metrics, and examining the stability of the resulting clusters under different centroid initialization strategies and values of K. The findings of this study are expected to provide useful empirical insights for researchers and practitioners seeking to apply clustering techniques to multivariate water quality data more effectively.

6. References

[1] World Health Organization, *Guidelines for Drinking-Water Quality: Fourth Edition Incorporating the First and Second Addenda*. Geneva, Switzerland: World Health Organization, 2022.

[2] World Health Organization and United Nations Human Settlements Programme, *Progress on the Proportion of Domestic and Industrial Wastewater Flows Safely Treated: Mid-term Status of SDG Indicator 6.3.1 and Acceleration Needs*. Geneva, Switzerland: World Health Organization, 2024.

[3] Food and Agriculture Organization of the United Nations, "Water quality," *FAO Land and Water*, Rome, Italy.

[4] F. Ghobadi and D. Kang, "Application of machine learning in water resources management: A systematic literature review," *Water*, vol. 15, no. 4, Art. no. 620, 2023, doi: 10.3390/w15040620.

[5] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[6] R. Xu and D. Wunsch II, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, May 2005.

[7] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.

[8] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[9] J. Blömer, C. Lammersen, M. Schmidt, and C. Sohler, "Theoretical analysis of the k-means algorithm—A survey," in *Algorithm Engineering: Selected Results and Surveys*, Lecture Notes in Computer Science, vol. 9220. Cham, Switzerland: Springer, 2016, pp. 81–116.

[10] C. Wongoutong, "The impact of neglecting feature scaling in k-means clustering," 2024. Penelitian ini membandingkan lima metode penskalaan fitur dan menunjukkan pentingnya *feature scaling* ketika K-Means diterapkan pada fitur dengan satuan yang berbeda.

[11] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[12] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[13] T. Zubaidah, N. Karnaningroem, and A. Slamet, “K-means method for clustering water quality status on the rivers of Banjarmasin, Indonesia,” *ARPN Journal of Engineering and Applied Sciences*, vol. 13, no. 11, pp. 3692–3697, 2018.

[14] M. A. Novianta, S. Syafrudin, and B. Warsito, “K-Means clustering for grouping rivers in DIY based on water quality parameters,” *JUITA: Jurnal Informatika*, vol. 11, no. 1, pp. 155–163, 2023.

[15] G. Arslan, C. Parim, and E. Cene, “Comparison of K-means and Fuzzy C-means clustering algorithms on water quality parameters: Case study of Ergene Basin for 17 stations,” in *Current Debates on Natural and Engineering Sciences 9*, pp. 145–159, 2023.