Indonesian Journal of Electrical Engineering and Computer Science
Vol. 39, No. 2, August 2025, pp. 1411~1420
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v39.i2.pp1411-1420

Handling missing values and clustering industrial liquid waste using K-medoids

Ratih Hafsarah Maharrani[1], Prih Diantono Abda'u[2], Ganjar Ndaru Ikhtiagung[3], Nur Wahyu Rahadi[2], Zaenurrohman[4]

[1]Cybersecurity Engineering, Department of Computer and Business, Cilacap State Polytechnic, Cilacap, Indonesia
[2]Informatics Engineering, Department of Computer and Business, Cilacap State Polytechnic, Cilacap, Indonesia
[3]Sharia Financial Institution Accounting, Department of Computer and Business, Cilacap State Polytechnic, Cilacap, Indonesia
[4]Electrical Engineering, Department of Electrical Engineering and Mechatronics, Cilacap State Polytechnic, Cilacap, Indonesia

Article Info

Article history:

Received Dec 24, 2024
Revised Mar 30, 2025
Accepted Jul 2, 2025

Keywords:

Industrial liquid waste
K-medoid
KNN imputation
Replace missing values
Z-score normalization

ABSTRACT

The textile industry is a significant contributor to environmental pollution due to its wastewater, which contains hazardous substances such as dyes, heavy metals, and chemicals that can severely harm aquatic ecosystems. Effective management of this wastewater is crucial to mitigate its environmental impact. This study focuses on classifying industrial liquid waste data using the K-medoids clustering method, chosen for its robustness to noise and outliers compared to K-means. To address challenges in wastewater data processing, such as missing values and varying data scales, two approaches are compared: replacing missing values with zero and K-nearest neighbors (KNN) imputation, alongside Z-score normalization for data uniformity. The clustering quality is evaluated using the Davies-Bouldin index (DBI) for cluster variations of k=2, 3, 4, and 5. The results show that the best clustering quality is achieved at k=2, with the smallest DBI values obtained using KNN imputation (0.139) and zero replacement (0.149). The superior performance of KNN imputation highlights its effectiveness in handling missing data. These findings provide valuable insights into the characteristics of textile industry wastewater pollution, offering a robust framework for effective wastewater management. The study concludes with practical recommendations for policymakers and industry stakeholders to adopt advanced data-driven approaches for sustainable wastewater treatment strategies.

Corresponding Author:

Ratih Hafsarah Maharrani
Cybersecurity Engineering, Department of Computer and Business, Cilacap State Polytechnic
Cilacap, Central Java, Indonesia
Email: ratih.hafsarah@pnc.ac.id

1. INTRODUCTION

Effluent, particularly from the textile industry, represents a critical environmental issue with significant global implications. The textile industry produces wastewater containing hazardous substances such as synthetic dyes, heavy metals, and other chemicals, which can severely pollute aquatic ecosystems if not managed properly [1]. This type of waste not only threatens aquatic biodiversity but also poses risks to human health, primarily through the contamination of drinking water and the food chain [2]. According to a United Nations report, more than 80% of industrial wastewater is discharged into rivers and oceans without adequate treatment, leading to environmental pollution and contributing to approximately 50% of global child deaths due to water-related diseases [3]. Consequently, effective effluent management has become a priority to safeguard both the environment and public health.

To address this issue, various approaches have been employed, particularly through the application of data analysis techniques such as data mining and machine learning. These techniques enable the identification of patterns and characteristics in effluent data, which can be used to optimize wastewater treatment and disposal processes. Methods such as clustering and classification have proven effective in categorizing and predicting waste types based on parameters like pH, BOD (biochemical oxygen demand), COD (chemical oxygen demand), and heavy metal concentrations [4]. To address this issue, various approaches have been employed, particularly through the application of data analysis techniques such as data mining and machine learning. These techniques enable the identification of patterns and characteristics in effluent data, which can be used to optimize wastewater treatment and disposal processes. Methods such as clustering and classification have proven effective in categorizing and predicting waste types based on parameters like pH, BOD, COD, and heavy metal concentrations [5]. However, a significant challenge in managing textile industry effluent lies in the incompleteness of data and the high variability of effluent parameters. Effluent datasets often contain missing values and exhibit varying scales, which can compromise the accuracy of analytical results [6].

This study proposes the use of the K-medoids algorithm for clustering wastewater data. Prior to clustering, two methods for handling missing data will be compared: K-nearest neighbor (KNN) imputation and the replacement of missing values with zero. Additionally, data normalization using the Z-Score method will be applied to ensure uniformity in the scale of variables [7]. Several previous studies have developed methodologies for wastewater data analysis. For instance, random forest was employed to predict surface water quality in India, demonstrating superior performance compared to methods such as artificial neural network (ANN) and support vector machine (SVM) [8]. Furthermore, gradient boosting, particularly CatBoost, was identified as the optimal model for regression analysis and classification of wastewater data, achieving a perfect receiver operating characteristic- area under the curve (ROC-AUC) score of 1.0 [9]. In the context of clustering, AGNES (agglomerative nesting) exhibited the highest performance with a purity score of 0.955, while DBSCAN proved effective in handling noisy data [10]. However, these studies did not specifically address the challenges of missing data handling and data normalization in the context of clustering textile industry effluent. Other relevant studies have explored imputation methods for water quality data. For example, a comparative study evaluated various imputation techniques, including inverse distance weighting (IDW), random forest regressor (RFR), and K-nearest neighbors regressor (KNNR), and found that over 76% of imputation results were deemed "satisfactory," with IDW yielding the best outcomes [11]. Similarly, another investigation examined the effectiveness of imputation techniques in water distribution systems, concluding that the success of these techniques heavily depends on the specific case, the chosen method, the chosen method, and the parameters of the dataset [12]. Additionally, a comprehensive review highlighted various approaches to handling missing data in environmental studies, emphasizing the importance of selecting appropriate methods to preserve data integrity [13].

The primary contribution of this study is to compare the effectiveness of KNN imputation and the replacement of missing values with zero in the context of clustering wastewater data. The research employs the K-medoids algorithm to cluster effluent data based on key parameters such as temperature, pH, BOD, and conductivity. The quality of the clustering results is then evaluated using the DBI. The objectives of this study are threefold: (i) to identify the most effective method for handling missing data, (ii) to develop an accurate clustering model, and (iii) to provide actionable recommendations for more efficient effluent management. The potential benefits of this research include enhancing the efficiency of wastewater treatment processes in industrial settings, reducing the environmental impact of pollution, and establishing a foundation for the development of advanced methods for effluent data analysis.

2. METHOD

This research utilizes data obtained from Kaggle in CSV format and processes it using machine learning methods. The dataset consists of 20 attribute columns and 620 rows, although not all attributes are used in the data processing. Only those attributes related to textile industry wastewater data are processed further. As shown in Figure 1, this stage of the process is referred to as data preprocessing. At this stage, irrelevant or unnecessary data will be removed, while relevant data will be selected for use in the next steps. The water quality causal inference dataset is a comprehensive collection of data designed to explore causal relationships between various water quality parameters and the factors influencing them. This dataset typically includes detailed information on water pollution levels, environmental conditions, and the effects of human activities, such as industrial discharge and agricultural runoff, on water quality. It serves as a critical resource for understanding how these variables impact the health and sustainability of aquatic ecosystems.

However, the dataset may contain noise, errors, inconsistencies, exceptions, or missing values. These issues can introduce confusion and inaccuracies during the pattern identification process in data mining. To address missing values, imputation methods were implemented. Imputation is one of the methods employed to handle missing values in the data by providing an accurate estimate of those values [14], [15].

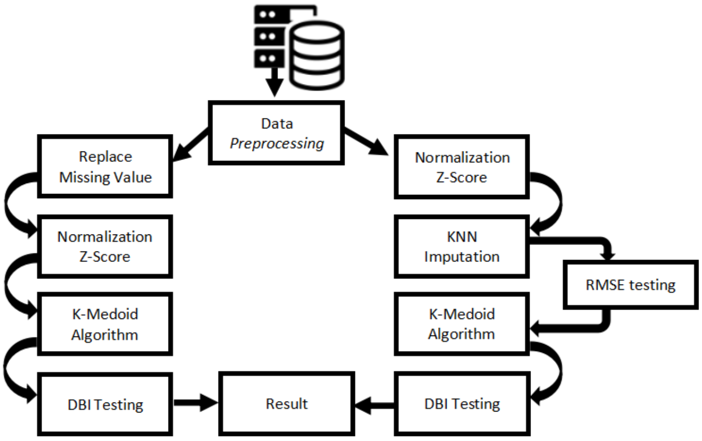


Figure 1. Research stages

2.1. Handling missing data with KNN imputation

In this study, KNN imputation was employed to handle missing values due to its ability to produce accurate estimates by leveraging information from the nearest neighbors in the feature space [14]. This technique fills missing values by calculating the average (or mode) of the k nearest neighbors with complete values, thereby maintaining the integrity of the dataset and improving the performance of analytical models [16]. The process involves three steps: (i) calculating the Euclidean distance between data points with missing values and the rest of the dataset, (ii) selecting the KNN with complete values based on the calculated distance, and (iii) imputing the missing values using the average (or mode) of the selected neighbors. The effectiveness of KNN Imputation was evaluated using root mean square error (RMSE), which measures the accuracy of the imputed values and facilitates comparisons between methods [17], [18]. RMSE serves as a critical metric for assessing the performance of imputation techniques in addressing missing value problems.

2.2. Handling missing data (replacing missing values with zero)

The second method involves replacing missing values with zero, where each missing value in each attribute is substituted with zero without considering the underlying data distribution pattern [19]. While this approach is simpler, it may introduce bias if the missing values are not randomly distributed. This method is particularly useful in scenarios where computational efficiency is prioritized over precision.

2.3. Data normalization

Data normalization was performed to ensure uniformity in the scale of the data, avoid bias in distance calculations, and improve the accuracy of imputation results. The Z-score normalization method was applied, transforming the data into a normal distribution with a mean of 0 and a standard deviation of 1 [20], [21]. The formula for Z-score normalization is as follows:

$$Z \frac{= (X-\mu)}{\sigma} \tag{1}$$

where Z is the normalized value, X is the original value, μ is the mean of the dataset, and σ is the standard deviation of the dataset.

For KNN Imputation, normalization was performed before imputation because KNN relies heavily on distance metrics (e.g., Euclidean or Manhattan) to calculate proximity between data points. If the scales of the variables differ significantly, variables with larger scales may dominate the distance calculation, leading to biased imputation results. In contrast, for the method of replacing missing values with zero, normalization was applied after imputation to ensure all data points (including imputed values) were on the same scale.

2.4. Clustering with K-medoids algorithm

The K-medoids algorithm was selected for clustering the wastewater data due to its robustness in handling noise and outliers [22]. Unlike K-means, which uses centroids, K-medoids selects actual data points (medoids) as cluster representatives. A medoid is the data point within a cluster that has the minimum total distance to all other points in the cluster, making it more representative and less sensitive to outliers [23]–[25].The K-medoids clustering process begins with the initialization step, where K-medoids are randomly selected from the dataset. Next, in the assignment step, each data point is assigned to the nearest medoid using Euclidean distance, which is calculated as:

$$d(x,y) = \ \|x-y\| = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2\ ;1,2,3,\ldots.n} \tag{2}$$

where x and y represent data points, and n is the number of dimensions. After assigning all data points to their nearest medoids, the update medoids step is performed by selecting a new medoid within each cluster that minimizes the total distance to all other points in the cluster. Finally, the algorithm iterates between the assignment and update steps until the medoids no longer change or a predefined stopping criterion is met. This iterative process ensures that the clusters are optimized and representative of the underlying data structure.

2.5. Cluster evaluation using Davies-Bouldin index

The quality of the clustering results was evaluated using the DBI. DBI measures the effectiveness of clustering by calculating the ratio of within-cluster scatter to between-cluster separation. A lower DBI value indicates well-separated and compact clusters, while a higher value suggests overlapping or poorly defined clusters [26]. The calculation of the DBI value can be expressed as follows:

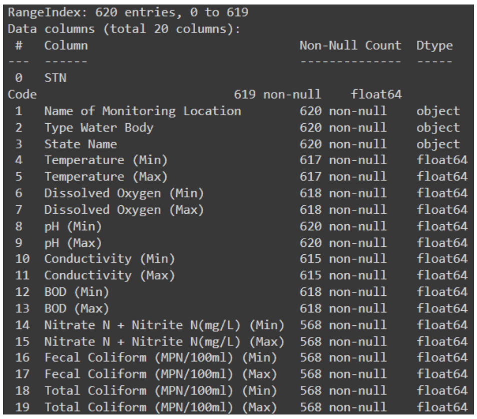
$$\text{DBI}=\frac{\mathbf{1}}{\mathbf{n}}\ \sum_{i=1}^{n}\begin{matrix}max\\ j\neq 1\end{matrix}\left(\frac{Si+Sj}{dij}\right) \tag{3}$$

where: S_i is the size (variance) of cluster i, d_{ij} is the distance between cluster centers I and j, and n is the number of clusters.

3. RESULTS AND DISCUSSION

3.1. Data preprocessing

The initial stage of this study involved preprocessing the raw data to ensure its suitability for further analysis. The dataset, obtained from Kaggle, consists of 620 data points with various attributes relevant to industrial wastewater quality. As illustrated in Figure 2, only specific attributes were selected for analysis, including temperature, dissolved oxygen (DO), pH, conductivity, and BOD. These parameters were chosen due to their significant impact on water quality and environmental health. For instance, temperature influences chemical reactions in wastewater, with higher temperatures indicating thermal pollution [27]. Similarly, DO levels serve as a critical indicator of organic pollution, while pH and conductivity provide insights into the suitability of water for aquatic life [28]. BOD levels, on the other hand, reflect the extent of organic pollution, with higher values indicating stronger pollution due to accelerated bacterial growth and oxygen depletion [29].



```
RangeIndex: 620 entries, 0 to 619
Data columns (total 20 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   STN
Code                           619 non-null    float64
 1   Name of Monitoring Location         620 non-null    object
 2   Type Water Body                     620 non-null    object
 3   State Name                          620 non-null    object
 4   Temperature (Min)                   617 non-null    float64
 5   Temperature (Max)                   617 non-null    float64
 6   Dissolved Oxygen (Min)              618 non-null    float64
 7   Dissolved Oxygen (Max)              618 non-null    float64
 8   pH (Min)                            620 non-null    float64
 9   pH (Max)                            620 non-null    float64
 10  Conductivity (Min)                  615 non-null    float64
 11  Conductivity (Max)                  615 non-null    float64
 12  BOD (Min)                           618 non-null    float64
 13  BOD (Max)                           618 non-null    float64
 14  Nitrate N + Nitrite N(mg/L) (Min)   568 non-null    float64
 15  Nitrate N + Nitrite N(mg/L) (Max)   568 non-null    float64
 16  Fecal Coliform (MPN/100ml) (Min)    568 non-null    float64
 17  Fecal Coliform (MPN/100ml) (Max)    568 non-null    float64
 18  Total Coliform (MPN/100ml) (Min)    568 non-null    float64
 19  Total Coliform (MPN/100ml) (Max)    568 non-null    float64
```

Figure 2. Water quality dataset

3.2. Handling missing values and data normalization

The dataset contained missing values, as illustrated in Table 1, which could affect the accuracy and reliability of the analysis. To address this, two methods were applied: (i) replacing missing values with zero and (ii) KNN imputation. KNN imputation was chosen for its ability to estimate missing values based on the nearest neighbors, preserving the relationships between variables. Additionally, Z-score normalization was applied to rescale the data, ensuring that all attributes had a mean of 0 and a standard deviation of 1 (Table 2). This step was crucial to prevent attributes with larger scales from dominating the analysis, especially in distance-based methods like KNN.

Table 1. Dataset with missing value

No	Temperature (Min)	Temperature (Max)	Dissolved oxygen (Min)	Dissolved oxygen (Max)	pH (Min)	pH (Max)	Conductivity (Min)	Conductivity (Max)	BOD (Min)	BOD (Max)
1	24	27	3.2	6.5	7	8	245	5160	1.6	3.2
2	26	29	3	6.8	6.9	7.8	599	1179	1.9	4.6
3	18	24	4.2	6	7.2	8.2	28000	56900	2.4	2.8
...	...	...	...	...	...	...	...	...	...	...
77	24	32	6	6.8	7.1	7.8	NaN	NaN	8	16
78	21	31	7	7.2	7.1	7.8	NaN	NaN	2	2.8
…	...	...	...	...	...	...	...	...	...	...
110	20	30	5.4	8.2	7.3	8.4	413	1038	1	2.5
111	22	35	2.6	7.7	7.1	8.5	440	1360	8	29
112	NaN	NaN	6.2	7.8	7.2	8	289	812	3.5	4.5
113	NaN	NaN	3.9	7.2	7.6	8.2	329	488	3.2	4.8
114	20.4	32	6.5	7.2	7	7.9	860	1360	4	9
115	NaN	NaN	6.3	8	7	7.8	280	785	3.4	4.6
...	...	...	...	...	...	...	...	...	...	...
617	22.0	32.0	8.8	9.8	7.0	7.6	311.0	422.0	2.7	7.3
618	7.0	16.0	7.1	8.1	6.9	7.6	13.0	29.0	1.0	1.1
619	9.0	24.0	7.2	8.2	6.9	7.6	20.0	36.0	1.0	1.8
620	8.0	18.0	7.1	8.5	6.9	7.9	16.0	73.0	1.0	1.6
616	22.0	31.0	5.2	9.7	7.2	8.5	252.0	826.0	1.8	4.1

Table 2. Normalization Z-score

No	Temperature (Min)	Temperature (Max)	Dissolved oxygen (Min)	Dissolved oxygen (Max)	pH (Min)	pH (Max)	Conductivity (Min)	Conductivity (Max)	BOD (Min)	BOD (Max)
1	0.47	-0.38	-0.26	-0.04	-0.04	-0.06	-0.22	0.48	-0.39	-0.36
2	0.90	0.17	-0.34	0.08	-0.04	-0.47	-0.03	-0.13	-0.32	-0.31
3	-0.82	-1.20	0.18	-0.24	-0.04	0.35	14.30	8.39	-0.19	-0.37
...	...	...	...	...	...	...	...	...	...	...
77	0.47	0.99	0.97	0.08	-0.04	-0.47	NaN	NaN	1.19	0.07
78	-0.17	0.72	1.41	0.24	-0.04	-0.47	NaN	NaN	-0.29	-0.37
...	...	...	...	...	...	...	...	...	...	...
110	-0.39	0.44	0.71	0.64	-0.04	0.75	-0.13	-0.15	-0.54	-0.38
111	0.04	1.82	-0.52	0.44	-0.04	0.96	-0.12	-0.10	1.19	0.51
112	NaN	NaN	1.06	0.48	-0.04	-0.06	-0.20	-0.18	0.08	-0.31
113	NaN	NaN	0.05	0.24	-0.04	0.35	-0.17	-0.23	0.00	-0.30
114	-0.30	0.99	1.19	0.24	-0.04	-0.26	0.10	-0.10	0.20	-0.16
115	NaN	NaN	1.11	0.56	-0.04	-0.47	-0.20	-0.19	0.05	-0.31
116	-1.68	-1.47	2.03	1.25	-0.04	-0.06	-0.25	-0.25	-0.54	-0.43
...	...	...	...	...	...	...	...	...	...	...
617	0.04	-3.39	1.46	0.60	-0.04	-0.87	-0.34	-0.30	-0.54	-0.43
618	-3.12	-1.20	1.50	0.64	-0.04	-0.87	-0.34	-0.30	-0.54	-0.40
619	-2.74	-2.84	1.46	0.76	-0.04	-0.26	-0.34	-0.30	-0.54	-0.41
620	-2.96	0.99	2.20	1.29	-0.04	-0.87	-0.18	-0.24	-0.12	-0.22

Normalization is important, considering that Table 1 contains data with larger value scales. If the data is not normalized before the imputation process, it can negatively impact the resulting RMSE value. Without normalization, the KNN model tends to give more weight to attributes with larger scales, which can lead to less accurate predictions and increased estimation error. After that, the existing dataset continues with processing using KNN imputation. Meanwhile, in the process of replacing missing values with zero, normalization is performed after filling in the value of 0.

3.3. KNN imputation, root mean square error

The presence of missing data is a serious problem, as it can affect the accuracy of data-based analyses and decisions. Therefore, it is urgent to perform missing value imputation using the KNN method. This method not only helps fill in the missing values based on the closest relevant data but also maintains the integrity and relationships between variables in the dataset.

Figure 3 shows the performance of the KNN model measured using RMSE. The model was tested with various values of K, namely 1, 5, 10, 15, and 20, to determine the optimal value of K that provides the lowest prediction error in Table 3. When K=1, the RMSE was 0.884, indicating a high error rate due to the KNN model's sensitivity to noise at low K values. Increasing K to 5 reduced the RMSE to 0.613, improving prediction accuracy. At K=10 and K=15, the RMSE stabilized at 0.618 and 0.607, respectively, with K=15 yielding the lowest RMSE and thus the most accurate predictions. However, increasing K to 20 slightly raised the RMSE to 0.621, showing that too many neighbors can reduce model accuracy. Therefore, K=15 was selected as the optimal value for the next stage.

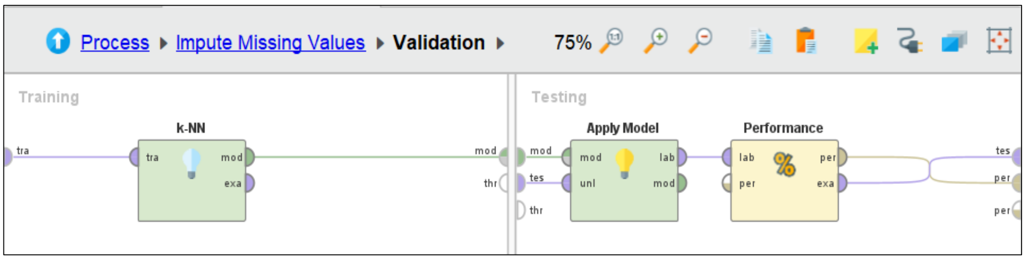


Figure 3. Water quality dataset

Table 3. Value RMSE

Value K	RMSE
1	0.884
5	0.613
10	0.618
15	0.607
20	0.621

3.4. Clustering K-medoids

After the normalization stage, KNN imputation and imputation testing with RMSE are performed. The next step is clustering experiments using the K-medoids method, starting with testing various values of K (the number of groups) for the DBI to determine the optimal number of clusters for grouping industrial waste. DBI is used as an evaluation metric to determine the quality of the resulting clustering. In general, the lower the DBI value, the better the quality of the formed clusters, as it indicates that the clusters are significantly distant from each other and that the data within each cluster is more uniform.

From the results in Tables 4(a) and 4(b), it can be seen that data processing with imputation reduces the resulting DBI index value. This indicates that the KNN Imputation method is superior and can be used in comparison to data without imputation. In a study, imputation methods such as KNN have been shown to provide better results in terms of data representation compared to replacing missing values with fixed numbers, such as zero, which often do not reflect the true condition of the data [30]. The analysis results based on the table show a comparison of the DBI on the test data using two approaches: KNN imputation and replacing missing values with zero. In the KNN imputation approach (Table 4(a)), the best (lowest) DBI value is obtained at K=2, with an index of 0.139, indicating better clustering quality than other K values. Meanwhile, in the approach of replacing missing values with zero (Table 4(b)), the best DBI value is also found at K=2, but with an index of 0.149. Overall, the DBI value at each K for the approach with KNN Imputation is lower than that without imputation. This indicates that using KNN Imputation improves clustering quality, making it more optimal than without the imputation process.

After determining the value of k=2 from the KNN imputation results, the clustering results of the dataset using the K-medoids algorithm can be obtained such as Figure 4. Based on Table 5, the results of the normalization or standardization of the original DO data indicate that higher DO values in water reflect better water quality, as more oxygen is available for living organisms. Therefore, Cluster_0, which has a higher maximum DO range compared to Cluster_1, can be indicated as a cluster with better water quality. Meanwhile, Cluster_1, which has a lower maximum DO value, likely indicates lower water quality or higher levels of organic pollution, as dissolved oxygen tends to deplete during the decomposition of organic materials.

Table 4. DBI testing with (a) KNN imputation and (b) replace missing value with zero

(a)

Value K	DBI index
2	0.139
3	0.192
4	0.161
5	0.156

(b)

Value K	DBI index
2	0.149
3	0.192
4	0.172
5	0.220

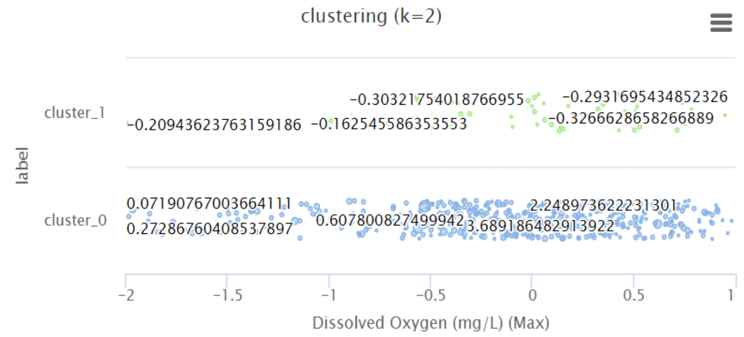


Figure 4. Example of clustering results for the DO attribute

Table 5. K-medoid clustering results

Cluster model	Attribute name	cluster_0	cluster_1
Cluster 0: 553 items	Temperature (Min)	0.042	-2.964
Cluster 1: 67 items	Temperature (Max)	0.993	-2.843
Total number of items: 620	Dissolved Oxygen (Min)	2.202	1.456
	Dissolved Oxygen (Max)	1.286	0.763
	pH (Min)	-0.040	-0.041
	pH (Max)	-0.872	-0.263
	Conductivity (Min)	-0.184	-0.338
	Conductivity (Max)	-0.244	-0.297
	BOD (Min)	-0.120	-0.541
	BOD (Max)	-0.219	-0.410

Based on the existing results, Cluster_0 shows better water quality despite having a higher temperature, with higher DO levels and lower BOD. In contrast, Cluster_1 shows poorer water quality, with low DO and high BOD, indicating greater organic pollution. This difference suggests that Cluster_1 is likely more polluted compared to Cluster_0.

3.5. Implications and future work

The results of this study have several important implications. First, the use of KNN Imputation and Z-score normalization significantly improved the accuracy and reliability of the clustering analysis. Second, the K-Medoids algorithm proved effective in handling noisy data and identifying clusters with distinct water quality characteristics. These findings can inform decision-making in industrial wastewater management, particularly in identifying pollution sources and implementing targeted mitigation strategies.

However, this study has limitations. The dataset, while comprehensive, is limited to specific geographic regions and may not fully represent global wastewater conditions. Future research could expand the dataset to include additional regions and explore other clustering algorithms, such as DBSCAN or hierarchical clustering, to further validate the findings. Additionally, integrating real-time monitoring systems could enhance the practical applicability of the proposed methods.

4. CONCLUSION

This study addresses the critical issue of textile industry wastewater management by comparing two methods for handling missing data replacing missing values with zero and KNN imputation followed by clustering using the K-medoids algorithm. The research demonstrates that KNN imputation, particularly at k=15, achieves the lowest RMSE value (0.607), highlighting its effectiveness in accurately estimating missing values. Furthermore, clustering experiments reveal that KNN Imputation at K=2 yields the optimal DBI value of 0.139, indicating well-separated and uniform clusters. These findings underscore the importance of selecting appropriate parameters, such as the number of neighbors (k) and clusters (K), to ensure optimal clustering quality. By providing a clear separation between clusters and insights into the characteristics of textile industry wastewater, this study offers a robust framework for improving waste management strategies and reducing environmental impact. Future research could explore the integration of real-time monitoring systems and the application of other clustering algorithms, such as DBSCAN or hierarchical clustering, to further validate and enhance these findings. Ultimately, this work contributes to the broader field of environmental science by offering practical solutions for sustainable industrial practices.

ACKNOWLEDGEMENTS

Our gratitude goes to the Research and Community Service Center of Cilacap State Polytechnic in the 2024 internal research financing scheme (DIPA) for assisting in the research process.

FUNDING INFORMATION

The authors declare that the research presented in this article was funded by the DIPA Fund of Politeknik Negeri Cilacap for the year 2024, under Contract No. 160/PL43/AL.04/2024.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ratih Hafsarah Maharrani	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Prih Diantono Abda’u	✓	✓	✓	✓	✓	✓	✓	✓				✓	✓	✓
Ganjar Ndaru Ikhtiagung				✓										✓
Nur Wahyu Rahadi						✓							✓	
Zaenurrohman	✓					✓								

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

The research related to human use has been complied with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY

The data supporting the findings of this study were obtained from Kaggle. Restrictions apply to the availability of these data, which were used under license for this study. The data are accessible at https://www.kaggle.com/code/sasakitetsuya/water-quality-causal-inference-by-lingam/input.

REFERENCES

[1] B. J. Singh, A. Chakraborty, and R. Sehgal, "A systematic review of industrial wastewater management: evaluating challenges and enablers," *Journal of Environmental Management*, vol. 348, pp. 1–34, 2023, doi: 10.1016/j.jenvman.2023.119230.

[2] M. S. I. Afrad, M. B. Monir, M. E. Haque, A. A. Barau, and M. M. Haque, "Impact of industrial effluent on water, soil and rice production in Bangladesh: a case of Turag River Bank," *Journal of Environmental Health Science and Engineering*, vol. 18, no. 2, pp. 825–834, 2020, doi: 10.1007/s40201-020-00506-8.

[3] L. Lin, H. Yang, and X. Xu, "Effects of water pollution on human health and disease heterogeneity: a review," *Frontiers in Environmental Science*, vol. 10, no. June, 2022, doi: 10.3389/fenvs.2022.880246.

[4] R. Shrivastava and N. K. Singh, "Assessment of water quality of textile effluent and its treatment by using coagulants and plant material," *Materials Today: Proceedings*, vol. 43, no. 5, pp. 3318–3321, 2021, doi: 10.1016/j.matpr.2021.02.373.

[5] T. U. Rahman *et al.*, "The advancement in membrane bioreactor (MBR) technology toward sustainable industrial wastewater management," *Membranes*, vol. 13, no. 2, 2023, doi: 10.3390/membranes13020181.

[6] B. Chen *et al.*, "In search of key: protecting human health and the ecosystem from water pollution in China," *Journal of Cleaner Production*, vol. 228, pp. 101–111, 2019, doi: 10.1016/j.jclepro.2019.04.228.

[7] P. K. Jena, S. M. Rahaman, P. K. Das Mohapatra, D. P. Barik, and D. S. Patra, "Surface water quality assessment by random forest," *Water Practice and Technology*, vol. 18, no. 1, pp. 201–214, 2023, doi: 10.2166/wpt.2022.156.

[8] I. Gulshin and O. Kuzina, "Machine learning methods for the prediction of wastewater treatment efficiency and anomaly classification with lack of historical data," *Applied Sciences (Switzerland)*, vol. 14, no. 22, 2024, doi: 10.3390/app142210689.

[9] K. Chen *et al.*, "Using unsupervised learning to classify inlet water for more stable design of water reuse in industrial parks," *Water Science & Technology*, vol. 89, no. 7, pp. 1757–1770, 2024.

[10] Nitya Nand Jha, "Computational machine learning analytics for prediction of water quality," *Communications on Applied Nonlinear Analysis*, vol. 31, no. 4s, pp. 448–465, 2024, doi: 10.52783/cana.v31.942.

[11] R. Rodríguez *et al.*, "Water-quality data imputation with a high percentage of missing values: a machine learning approach," *Sustainability (Switzerland)*, vol. 13, no. 11, pp. 1–17, 2021, doi: 10.3390/su13116318.

[12] M. S. Osman, A. M. Abu-Mahfouz, and P. R. Page, "A survey on data imputation techniques: water distribution system as a use case," *IEEE Access*, vol. 6, pp. 63279–63291, 2018, doi: 10.1109/ACCESS.2018.2877269.

[13] Y. F. Zhang, P. J. Thorburn, M. P. Vilas, and P. Fitch, "Machine learning approaches to improve and predict water quality data," *23rd International Congress on Modelling and Simulation - Supporting Evidence-Based Decision Making: The Role of Modelling and Simulation, MODSIM 2019*, no. December, pp. 491–497, 2019, doi: 10.36334/modsim.2019.d5.zhangyif.

[14] D. M. P. Murti, U. Pujianto, A. P. Wibawa, and M. I. Akbar, "K-nearest neighbor (K-NN) based missing data imputation," in *2019 5th International Conference on Science in Information Technology (ICSITech)*, Oct. 2019, vol. 5, pp. 83–88, doi: 10.1109/ICSITech46713.2019.8987530.

[15] L. Muflikhah, N. Hidayat, and D. J. Hariyanto, "Prediction of hypertention drug therapy response using K-NN imputation and SVM algorithm," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 15, no. 1, pp. 460–467, 2019, doi: 10.11591/ijeecs.v15.i1.pp460-467.

[16] R. M. Syauqi, P. N. Sabrina, and I. Santikarama, "K-means clustering with KNN and mean imputation on CPU Benchmark compilation data," *Journal of Applied Informatics and Computing*, vol. 7, no. 2, pp. 231–239, 2023, doi: 10.30871/jaic.v7i2.6491.

[17] A. Fadlil, Herman, and D. Praseptian M, "K nearest neighbor imputation performance on missing value data graduate user satisfaction," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 570–576, Aug. 2022, doi: 10.29207/resti.v6i4.4173.

[18] K. Seu, M. S. Kang, and H. Lee, "An intelligent missing data imputation techniques: a review," *International Journal on Informatics Visualization*, vol. 6, no. 1–2, pp. 278–283, 2022, doi: 10.30630/joiv.6.1-2.935.

[19] J. A. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn, "Dealing with Zeros and missing values in compositional data sets using nonparametric imputation," *Mathematical Geology*, vol. 35, no. 3, pp. 253–278, 2003, doi: 10.1023/A:1023866030544.

[20] H. Henderi, T. Wahyuningsih, and E. Rahwanto, "Comparison of min-max normalization and Z-score normalization in the K-nearest neighbor (kNN) algorithm to test the accuracy of types of breast cancer," *IJIIS: International Journal of Informatics and Information Systems*, vol. 4, no. 1, pp. 13–20, Mar. 2021, doi: 10.47738/ijiis.v4i1.73.

[21] R. H. Maharrani, P. D. Abda'u, and M. N. Faiz, "Clustering method for criminal crime acts using K-means and principal component analysis," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 34, no. 1, pp. 224–232, 2024, doi: 10.11591/ijeecs.v34.i1.pp224-232.

[22] R. Holloway *et al.*, "Optimal location selection for a distributed hybrid renewable energy system in rural Western Australia: A data mining approach," *Energy Strategy Reviews*, vol. 50, 2023, doi: 10.1016/j.esr.2023.101205.

[24] A. Rezky Pratama, B. Maulana, R. Didho Rianda, and S. El Hasyim, "Comparison of K-means and K-medoids algorithms for grouping video game sales data in North America," *Indonesian Journal of Informatic Research and Software Engineering (IJIRSE)*, vol. 3, no. 2, pp. 111–118, 2023, doi: 10.57152/ijirse.v3i2.898.

[25] S. Febriyanti and J. Nugraha, "Application of K-medoids clustering to increase the 2020 family planning program in Sleman Regency," *Enthusiastic : International Journal of Applied Statistics and Data Science*, vol. 2, no. 1, pp. 10–18, 2022, doi: 10.20885/enthusiastic.vol2.iss1.art2.

[26] A. Idrus, N. Tarihoran, U. Supriatna, A. Tohir, S. Suwarni, and R. Rahim, "Distance analysis measuring for clustering using k-means and davies bouldin index algorithm," *TEM Journal*, vol. 11, no. 4, pp. 1871–1876, Nov. 2022, doi: 10.18421/TEM114-55.

[27] J. A. Awomeso, A. M. Taiwo, A. M. Gbadebo, and J. A. Adenowo, "Studies on the pollution of water body by textile industry effluents in Lagos, Nigeria," *Journal of Applied Science in Environmental Sanitation*, vol. 5, no. 4, pp. 353–359, 2010, [Online]. Available: http://trisanita.org/jases/asespaper2010/ases34v5n4y2010.pdf.

[28] P. R. Kannel, S. Lee, Y. S. Lee, S. R. Kanel, and S. P. Khan, "Application of water quality indices and dissolved oxygen as indicators for river water classification and urban impact assessment," *Environmental Monitoring and Assessment*, vol. 132, no. 1–3, pp. 93–110, 2007, doi: 10.1007/s10661-006-9505-1.

[29] Y. C. Wong, V. Moganaragi, and N. A. Atiqah, "Physico-chemical investigations of semiconductor industrial wastewater," *Oriental Journal of Chemistry*, vol. 29, no. 4, pp. 1421–1428, 2013, doi: 10.13005/ojc/290418.

[30] Z. Zhang, "Missing data imputation: focusing on single imputation," *Annals of Translational Medicine*, vol. 4, no. 1, 2016, doi: 10.3978/j.issn.2305-5839.2015.12.38.

BIOGRAPHIES OF AUTHORS

Ratih Hafsarah Maharrani completed his Bachelor's degree in 2007 at Dian Nuswantoro University, then continued his Master's degree in computer engineering in 2010 at the same university. Currently, she is active as a lecturer at Politeknik Negeri Cilacap, in the Computer and Business Department. The research fields he specializes in include decision support systems, data mining, and artificial intelligence. For communication purposes, she can be contacted at email: ratih.hafsarah@pnc.ac.id.

Prih Diantono Abda'u received the B.Comp.Sc. degree from Universitas Amikom Purwokerto, Indonesia, in 2013, and the M.Comp.Sc. degree from Universitas Amikom Yogyakarta, Indonesia, in 2018. He is currently a Lecturer with the Department of Software Engineering, Politeknik Negeri Cilacap, Indonesia, holding the academic rank of Lektor. He has authored two books, *Pemrograman Web* (2021) and *Interaksi Manusia dan Komputer* (2024). His current research interests include the integration of the internet of things (IoT) and artificial intelligence (AI). He can be contacted at email: abdau@pnc.ac.id.

Ganjar Ndaru Ikhtiagung is a marketing expert with extensive experience in sustainable marketing strategy development, regional potential development, and digital technology application. He has published various scientific studies, including sustainable tourism development models and green marketing concepts to support the environment. The study of marketing performance in potato farming in Wonosobo provides recommendations for improving market efficiency, while the application of extreme programming to the "*Lapak Petani Online*" marketplace facilitates market access for farmers. In the policy field, he was involved in an investment study of an integrated fisheries industrial area in Cilacap, supporting the blue economy, and drafting an academic paper for the RAPERDA on investment. In 2024, formulating regional development policies for Central Java based on vocational education to increase innovation and local employment. With expertise in sustainable marketing, digital marketing, and economic policy analysis, focusing on collaboration and innovative approaches to advance the regional economy and improve community welfare. He can be contacted at email: ganjar@pnc.ac.id

Nur Wahyu Rahadi received the M.Eng. degree. in the field of electrical engineering, concentration in information technology, Gadjah Mada University, Yogyakarta, in 2015. Currently, he is the head of the Information and Communication Technology unit at the Cilacap State Polytechnic and one of the lecturers in the Diploma 3 Informatics Engineering Study Program, Computer and Business Department. His research interests include software engineering, artificial intelligence, cloud computing, and IoT. He can be contacted at email: n.wahyu.r@pnc.ac.id.

Zaenurrohman received a B.Eng. degree in electrical engineering from Wiworotomo College of Engineering, Indonesia, in 2013, and an M.Eng. degree in electrical engineering from Sultan Agung Islamic University, Semarang, Indonesia, in 2018. Currently, he is a Lecturer in the Department of Electronics Engineering, Cilacap State Polytechnic. His area of expertise is embedded systems. His research interests include microcontrollers, the internet of things, agricultural technology, early warning systems, radio frequency communication, control systems, power electronics, and robotics. Currently focusing more on research on LoRa communication in Arduino-based early warning systems. He can be contacted at email: zaenur@pnc.ac.id.