

Cluster Analysis of Alumni Job Waiting Times Using the K-Means Algorithm with Elbow and Silhouette Methods

Abdur Ro'uf¹, Hasyim Asy'ari², Maysas Yafi Urrochman³

Department of Information Technology, Faculty of Engineering and Informatics, Institut Teknologi dan Bisnis Widya Gama Lumajang, Indonesia^{1,2,3}

Corresponding Author: Abdur Ro'uf (abdurrouf.ar34@gmail.com)

ARTICLE INFO

Date of entry:

12 January 2026

Revision Date:

06 August 2026

Date Received:

01 October 2026

ABSTRACT

Evaluating graduate employability is a key indicator for assessing the quality of higher education institutions. One effective approach involves analyzing alumni tracer study data using data mining techniques. This study aims to categorize alumni data based on the waiting period for employment after graduation using the K-Means algorithm. To determine the optimal number of clusters, two methods were employed: the Elbow method (Sum of Squared Errors/SSE) and the Silhouette Score. The research process included data preprocessing—comprising data selection, transformation, and normalization—to enhance data quality prior to clustering. Test results indicated that the Elbow method yielded an optimal cluster count of $k=4$, whereas the Silhouette method showed the best result at $k=3$, with a score of 0.62. Consequently, $k=3$ was selected as the optimal number of clusters due to superior cluster quality. The clustering results revealed that the majority of alumni (67.6%) experienced a short waiting period (averaging ± 2 months), while 16.5% faced a long waiting period (± 8 months), and 15.9% fell into the medium category (± 5 months). These findings are expected to serve as a basis for higher education institutions to evaluate and improve graduate quality, ensuring better alignment with the demands of the job market.

Keywords: K-Means; Clustering; Elbow; Silhouette; Tracer Study; Employment Waiting Period



Cite this as: Ro'uf, A., Asy'ari, H., & Urrochman, M. Y. (2026). Cluster Analysis of Alumni Job Waiting Times Using the K-Means Algorithm with Elbow and Silhouette Methods. *Journal of Informatics Development*, 5(1), 7–17. <https://doi.org/10.30741/jid.v5i1.2080>

INTRODUCTION

Higher education institutions play a strategic role in preparing competent human resources capable of meeting the needs of both society and the corporate sector; indeed, a key measure of an institution's success is its ability to produce graduates who meet workforce demands. Information regarding the status of bachelor's degree (S1) alumni is crucial for these institutions, as it serves as a benchmark for assessing the practical application of the knowledge acquired during their undergraduate studies. This serves as a basis for curriculum evaluation and assists stakeholders in formulating workforce policies; mapping graduate employability is a crucial step in understanding the actual status of alumni after graduation. A tracer study is a commonly used method for tracking graduate profiles.

This study introduces an innovation that utilizes clustering, specifically the K-means algorithm, to group alum profile data from the Widya Gama Lumajang Institute of Technology and Business. The clustering aims to identify key information regarding alum distribution, such as employment status, salary levels, and the waiting period following graduation. The resulting data can serve as a basis for mapping outcomes against the targets set for graduates of each study program, enabling the institution to align its curriculum with employers' workforce and expertise requirements.

THEORETICAL FRAMEWORK AND HYPOTHESES

Clustering

Clustering is a data mining technique used to discover patterns or identify groups of objects sharing similar characteristics within a dataset (Dinh et al., 2019). A primary challenge in this process is determining the optimal clustering configuration (Syahfitri et al., 2023); suboptimal determination can lead to irregular clusters that fail to accurately reflect actual conditions (Roring et al., 2025). Consequently, the analysis of alum distribution, employment locations, and post-graduation waiting times is hindered. This study employs the k-means algorithm to group data based on specific attributes (Zhao et al., 2023), including name, workplace, position, salary, and post-graduation waiting time.

Data clustering using the k-means method is crucial as it enables the identification of specific patterns (Eliyanto & Surono, 2022) within the dataset obtained from the Student Affairs Department of the Widya Gama Lumajang Institute of Technology and Business. K-means is a widely used clustering algorithm due to its ability to select more accurate centroids (Suraya et al., 2023). In a previous study conducted by Muhammad Raqib Syahkur et al. (2024) to assess the nutritional status of toddlers and, by extension, public health, the silhouette method identified the optimal number of clusters (Piao Tan & A. Floudas, 2008) at $k=2$, yielding an average silhouette value close to 1; meanwhile, the elbow method indicated an "elbow point" at $k=5$, signifying that the optimal number of clusters (Schubert, 2023) lies at that point.

Silhouette

According to Dewi & Pramita (2019), the silhouette method measures the quality of clustering results by evaluating whether a data point fits its assigned cluster or is a poor fit, meaning it would be better suited to a different cluster. In their study comparing the silhouette and elbow methods for clustering handicraft data in Bali using the Davies-Bouldin Index (DBI) for evaluation, the silhouette method yielded the best result (1.06), compared to 1.10 for the elbow method. The method operates on a straightforward principle: determining cohesion (the distance between members within the same cluster) and separation (the distance to the nearest neighboring cluster).

$$SC = maks_k SI(k)$$

Information:

SC = Silhouette Coefficient

SI = Global Silhouette Index

k = number of clusters

To calculate the SI value for the i -th data point, there are two components: a_i and b_i . The value a_i represents the average distance from the i -th data point to all other data points within the same cluster. In contrast, b_i is obtained by calculating the average distance from the i -th data point to all data points in other clusters, that is, clusters other than the one containing the i -th data point.

Elbow

The elbow method (Alam & Muqem, 2023) helps determine the optimal number of clusters by simplifying the model as a preliminary step before further evaluation. It identifies the optimal

number of clusters by analyzing the decline in the Sum of Squared Errors (SSE), which measures the distance between data points and their respective cluster centroids; a lower SSE indicates greater cluster compactness.

$$SSE = \sum_{k=1}^K \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Information:

K = Number of clusters

C_k = k -th cluster

x_i = i -th data point in the cluster

μ_k = centroid (cluster center)

$(x_i - \mu_k)$ = squared distance between the data point and the centroid

The SSE (Sum of Squared Errors) represents the total distance between data points and the cluster center; the lower the SSE value, the more compact the cluster.

K-means

K-Means is a clustering method (a form of unsupervised learning) used to partition data into several groups (clusters) based on the similarity or proximity of data characteristics (Rouf et al., 2024). It functions by grouping objects or data points that share similarities; consequently, each resulting cluster contains data with largely similar characteristics. The primary objective of the K-Means algorithm is to minimize the distance between data points and their cluster centers (centroids). This process involves partitioning data into clusters by assigning each data point to the nearest cluster—an iterative procedure that continues until the results stabilize.

For each cluster, the centroid is calculated as the average of all data points within that cluster; the formula is provided below.

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

Information:

μ_k = centroid of the k -th cluster

$|C_k|$ = number of data points in the cluster

Using this formula, the average across all clusters can be determined.

Once the centroids in this study have been identified, the next step is to calculate the distances using the Euclidean distance formula.

$$d(x_i, \mu_k) = \sqrt{\sum_{j=1}^n (x_{ij} - \mu_{kj})^2}$$

Information:

d = Distance

x_{ij} = value of the i -th data point for the j -th variable

μ_{kj} = centroid value for the j -th variable

Euclidean distance is a formula used to calculate the distance from data points to a centroid and assign them to the nearest cluster; the centroids are then recalculated, and this process is repeated until the results stabilize.

Flowchart

This study follows a workflow illustrated in Figure 1, which depicts the use of two methods to cluster alums based on data including name, occupation, position, salary, and waiting period.

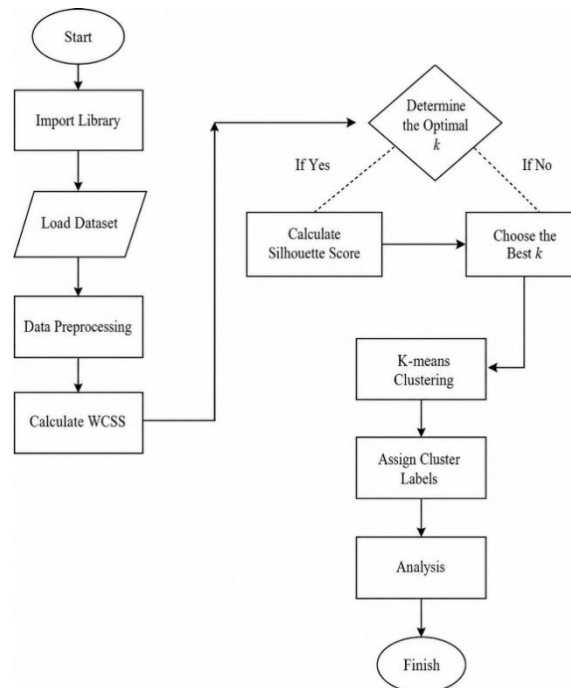


Figure 1. Flowchart

METHODS

The methodology of this research is based on the research workflow or flowchart shown in Figure 1, which comprises the following:

Import Library

In this process, the researcher uses Google Colab to carry out the research stages. The process begins by importing the necessary packages, such as *numpy*, *pandas*, and *matplotlib*. *pyplt*, *import seaboard* and *rn* as *sns*, from *sklearn*. *preprocessing* import *StandardScaler*, from *sklearn*. *cluster* import *KMeans*, from *sklearn*. *metrics* import *silhouette_score* to provide the tools required for the workflow. These libraries facilitate the calculations, visualization (diagrams), and classification tasks.

Load Dataset

This study utilizes structured data obtained through a qualitative approach—specifically, questionnaires and interviews with alumni as part of a tracer study (Prayogi & Irfandi, 2024). The data collected is primary data, meaning it was obtained directly from the sources (Undari Sulung, 2024). The following is an example of the dataset to be used for cluster identification.

```
import io
df = pd.read_csv(io.BytesIO(uploaded["Data import.csv"])), sep=';'
display(df.head())
```

No.	NIM	Nama	Tempat Kerja	Masa Tunggu setelah Lulus (bulan)	Gaji
0	1 218123238	ACHMAD ALFIAN ALI FARHAN	Kantor RSU Permata Lumajang	3	3.100.000
1	2 218123230	ACHMAD MASRURI	Kantor Bepeda	1	2.500.000
2	3 218123113	ADELIA RAHMANISA	Dealer YamahaSIP Semeru	8	2.300.000
3	4 218133188	AFIFATUL FADILAH	Kantor Dokter Ajeng	1	2.700.000
4	5 218123307	AGUS LAKSONO	Kantor Dokter Beny	1	2.700.000

Figure 2. Dataset

Preprocessing

Preprocessing involves selecting the data to be used or removing missing data so that unusable data is discarded (Çetin & Yıldız, 2022); the specific preprocessing stages carried out in this study are as follows:

Data Selection

Data selection is performed to select the variables to be used; the process is illustrated in Figure 3 below.

```
#memilih variabel untuk diproses
X = df[['Masa Tunggu setelah Lulus (bulan)', 'Gaji']].copy()
```

Figure 3. Variable Selection

Data Transformation

Converting data into a specific format so that it can be processed using the k-means algorithm. Data conversion can be performed as shown in Figure 4 below.

```
# membersihkan kolom 'Gaji' dan mengubah ke tipe numerik
X['Gaji'] = X['Gaji'].str.replace('.', '', regex=False).str.replace(',', '.', regex=False)
X['Gaji'] = pd.to_numeric(X['Gaji'])
```

Figure 4. Data Transformation

Normalization

In the data normalization process, data scales are standardized to prevent missing data and to avoid situations where certain data dominate for instance, when there are large numerical disparities between variables, such as values ranging from single digits to hundreds or even thousands. The normalization process can be carried out as illustrated in Figure 5 below.

```
#menghapus missing dan normalisasi
X = X.dropna()
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Figure 5. Normalization

The preprocessing stage is crucial in this study, as it determines the success of the process; without it (Singh et al., 2022), the clustering results could be erroneous. The benefits of not skipping the preprocessing stage include more accurate results, a more stable process, and easier analysis.

Determining the Number of Clusters

Determining the optimal number of clusters (k) involves two methods: the Elbow method (using SSE) and the silhouette score. In this stage, both methods are applied, and clustering is tested with k set to 10. The process is repeated iteratively to improve results, with values approaching the minimum being considered the best outcome. The evaluation metrics for the Elbow method (SSE) are presented in Table 1 below:

Table 1. SSE Value

k	SSE	Information
2	150	
3	90	Sharp
4	70	Still falling
5	65	Starting to level off
6	63	Very small

Based on the table above, the value to use is the lowest obtained across all trials. In contrast to the elbow method (SSE), which looks for the lowest value, evaluating the silhouette method involves identifying the highest value, as shown in Table 2 below:

Table 2. Silhouette Value

<i>k</i>	<i>silhouette</i>	<i>decision</i>
2	0.45	Small
3	0.60	Very High
4	0.55	High
5	0.50	Moderate

Based on the table above, we can select a high value for the silhouette score; both methodological processes are illustrated in Figure 6 below:

```
#menentukan jumlah klaster optimal - metode elbow dan shilhouette
sse = []
silhouette = []
K_range = range(2, 11)
for k in K_range:
    km = KMeans(n_clusters=k, random_state=42)
    km.fit(X_scaled)
    sse.append(km.inertia_)
    silhouette.append(silhouette_score(X_scaled, km.labels_))
```

Figure 6. Determination of the Number of Clusters

Creating A K-Means Model

After determining the optimal number of clusters, having already applied the elbow method (SSE) and silhouette analysis, the normalized data (X_scaled) is grouped. The next step involves forming three clusters based on the best (k) value identified in the previous stage. This process is illustrated in Figure 7 below.

```
#pilih k terbaik misalnya k dibagi menjadi 3 kelompok
k_opt = 3
km = KMeans(n_clusters=k_opt, random_state=42)
cluster = km.fit_predict(X_scaled)
```

Figure 7. Classifier

RESULTS AND DISCUSSION

The testing phase begins by loading the prepared data, followed by selecting variables from the dataset and cleaning the salary column by converting it to numeric format. Subsequently, missing values are removed, and the data is normalized; these steps constitute the preprocessing stage. This stage is crucial for ensuring accurate results, as omitting it would lead to incorrect, inaccurate, and unstable outcomes.

Testing

Before the testing phase, the number of clusters to be formed is determined first. The range established for this determination is between 2 and 11 clusters. The results of this clustering process are illustrated in Figure 8 below:

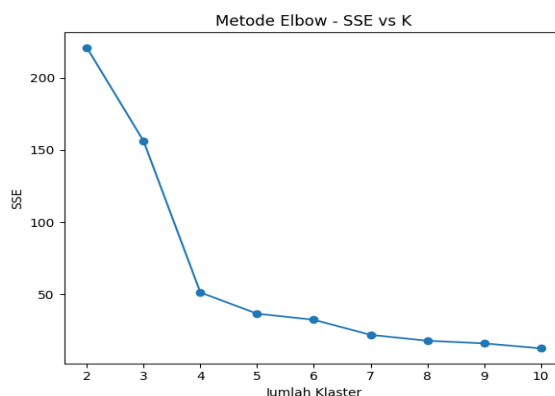


Figure 8. Elbow Method Results

The image above displays the ten groups represented by line plots, along with the SSE scores derived from the elbow-SSE method. The clustering results obtained using the silhouette method are shown in Figure 9 below.

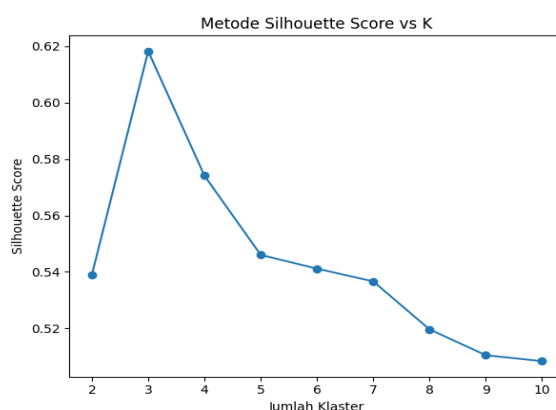


Figure 9. Silhouette Method Results

Figure 9 illustrates that the silhouette method yielded 10 clusters, each with its respective score. The results from the elbow-SSE and silhouette method analyses indicate that the data is suitable for analysis.

Analysis

The data from the diagrams in Figures 8 and 9 can be organized into a table to facilitate analysis. Table 3 presents the results of the analysis obtained using the elbow-SSE method, as shown below: Display the table as below:

Table 3. Analysis of Elbow-SSE Results

<i>k</i>	SSE	Result
2	221	
3	156	Very sharp decrease
4	51.3	Gradual decrease
5	36.7	Very gradual decrease
6	32.4	Gradual decrease
7	22	Gradual decrease
8	18	Gradual decrease
9	16	Gradual decrease
10	13	Gradual decrease

Based on the data derived from the elbow-SSE method, a very sharp decline is observed between $k=2$ and $k=3$, rendering them unsuitable; however, at $k=4$, $k=5$, and $k=6$, the rate of decline begins to level off. Consequently, the "elbow point" appears at $k=4$, representing the optimal value for subsequent clustering-meaning that adding further clusters beyond $k=4$ yields no significant change. Thus, according to the elbow method, the optimal number of clusters is $k = 4$. Furthermore, the silhouette method results shown in Figure 9 can be analyzed using Table 4 below:

Table 4. Silhouette Analysis Results

k	Silhouette	Hasil
2	0.54	Lower than the maximum score
3	0.62	Highest score
4	0.57	Lower than the maximum score
5	0.55	Lower than the maximum score
6	0.54	Lower than the maximum score
7	0.53	Lower than the maximum score
8	0.52	Lower than the maximum score
9	0.51	Lower than the maximum score
10	0.50	Lower than the maximum score

Based on the table above, the optimal silhouette result is observed at $k=3$, with a value of 0.62; this is the highest among the cluster counts. Beyond $k=3$, the values decline, indicating that $k=3$ yields better clustering results than the other values. Both methods have been tested, and each has yielded its own result; the table below compares the outcomes.

Table 5. Comparison of Results

Metode	Result
<i>Elbow</i>	$k=4$
<i>Silhouette</i>	$k=3$

Based on the Elbow method, the optimal number of clusters was determined to be 4, whereas the Silhouette method indicated the best value at $k=3$. Consequently, $k=3$ was selected because it yielded the highest silhouette score, signifying superior cluster quality.

Once the optimal number of clusters ($k=3$) was selected, the next step was to analyze the characteristics of each cluster to address the study's objectives, specifically the average waiting time for employment, average salary, and the number or proportion of alums per cluster. The process began by iterating through the clusters, from cluster 0 up to cluster ($k_{opt}-1$), and calculating and evaluating the resulting cluster statistics: count (number of data points), mean (average), std (standard deviation), min (minimum value), max (maximum value), and quartiles (Q1, median, Q3). The results from the best-performing configuration (as determined by the Silhouette method) are presented in the table below, along with an explanation.

Table 6. Cluster 0

Waiting Period after Graduation (months)	
count	119
mean	2.02
std	1.11
min	1
25%	1
50%	2
75%	3

max	5
%	67.6%

Based on the results above, it can be concluded that for Cluster 0, the average waiting period is approximately 2 months (ranging from 1 to 5 months), with a median of 2 months and a percentage of 67.6%. Further calculations are presented in Table 7 below.

Table 7. Cluster 1

Waiting Period after Graduation (months)	
count	29
mean	8
std	1.01
min	7
25%	7
50%	8
75%	9
max	10
%	16.5%

Regarding the next result, Table 7 shows that Cluster 1 has an average waiting period of approximately 8 months (ranging from 7 to 10 months), a median of 8, and a percentage of 16.5%. Further calculations are shown in Table 8 below.

Table 8. Cluster 2

Waiting Period after Graduation (months)	
count	28
mean	5
std	2.3
min	-6
25%	4
50%	6
75%	6
max	6
%	15.9%

Based on the results above, it can be concluded that for Cluster 2, the average waiting period is approximately 5 months, ranging from less than 6 months up to 6 months, with a median of 6 months and a percentage of 15.9%.

CONCLUSION

The results yielded three clusters, each with specific values. For Cluster 0, the average waiting period is approximately 2 months (ranging from 1 to 5 months), with a median of 2 months and a representation of 67.6%. Cluster 1 shows an average waiting period of approximately 8 months (ranging from 7 to 10 months), with a median of 8 months and a representation of 16.5%. Finally, Cluster 2 has an average waiting period of approximately 5 months (ranging from less than 6 months up to 6 months), with a median of 6 months and a representation of 15.9%. A comparison of the results across all clusters is presented in Table 9 below.

Table 9. Cluster Results

Cluster	average waiting period (months)	%	Category
0	±2	67.6	Fast
1	±8	16.5	Long
2	±5	15.9	Medium

Based on the data above, it can be concluded that Cluster 0 is the best-performing cluster, as it secures employment more quickly with a shorter waiting period of approximately two months; it can be characterized as a cluster that obtains jobs rapidly. Meanwhile, Cluster 1 takes the longest to secure employment, at approximately eight months, and Cluster 2 falls into the intermediate category, securing jobs faster than Cluster 1 with a waiting period of approximately five months.

REFERENCES

- Alam, A., & Muqem, M. (2023). Hybridization of K-means with improved firefly algorithm for automatic clustering in high dimension. *ArXiv*.
- Çetin, V., & Yıldız, O. (2022). A comprehensive review on data preprocessing techniques in data analysis. *Pamukkale University Journal of Engineering Sciences*, 28(2), 299–312. <https://doi.org/10.5505/pajes.2021.62687>
- Dewi, D. A. I. C., & Pramita, D. A. K. (2019). Comparative Analysis of Elbow and Silhouette Methods On K-Medoids Clustering Algorithm in Balinese Craft Production Grouping. *Matrix Journal*, 9(3), 102–109.
- Dinh, D. T., Fujinami, T., & Huynh, V. N. (2019). Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient. *Communications in Computer and Information Science*, 1103 CCIS, 1–17. https://doi.org/10.1007/978-981-15-1209-4_1
- Eliyanto, J., & Surono, S. (2022). The Suitable Distance Function for Fuzzy C-Means Clustering. *AIP Conference Proceedings*, 2578(November). <https://doi.org/10.1063/5.0106185>
- Muhammad Raqib Syahkur, Hartama, D., & Solikhun, S. (2024). Evaluasi Jumlah Cluster pada Algoritma K-Means++ Menggunakan Silhouette dan Elbow dengan Validasi Nilai DBI dalam Mengelompokkan Gizi Balita. *JST (Jurnal Sains Dan Teknologi)*, 13(3), 487–496. <https://doi.org/10.23887/jstundiksha.v13i3.86419>
- Piao Tan, M., & A. Floudas, C. (2008). Determining the Optimal Number of Clusters. *Encyclopedia of Optimization*, 1, 687–694. https://doi.org/10.1007/978-0-387-74759-0_123
- Prayogi, A., & Irfandi, M. A. K. (2024). Jurnal Multidisiplin Ilmu Nasional Pendekatan Kualitatif dan Kuantitatif. *Complex: Jurnal Multidisiplin Ilmu Nasional*, 1(2), 31. <https://ejurnal.faslibsmidia.com/index.php/complex/article/view/7/28>
- Roring, M. A. C., Ginoga, P. A., Ibrahim, R., Rompas, R. C., Yusupa, A., & Paturusi, S. D. E. (2025). E-issn : 2988-1986. *Multidisiplin Saintek*, 8(5), 1–16. <https://cibangsa.com/index.php/kohesi/article/view/6074/5284>
- Rouf, A., Qorittunnadyah, M., Asyari, H., & Urrohman, M. Y. (2024). Clustering of Lecturer Performance Using K-Means. *Journal of Informatics Development*, 3(1), 27–33. <https://doi.org/10.30741/jid.v3i1.1430>
- Schubert, E. (2023). Stop using the elbow criterion for k-means and how to choose the number of clusters instead. *ACM SIGKDD Explorations Newsletter*, 25(1), 36–42. <https://doi.org/10.1145/3606274.3606278>
- Singh, M. P., Gayathri, V., & Chaudhuri, D. (2022). A Simple Data Preprocessing and Postprocessing Techniques for SVM Classifier of Remote Sensing Multispectral Image Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 7248–7262. <https://doi.org/10.1109/JSTARS.2022.3201273>
- Suraya, S., Sholeh, M., & Lestari, U. (2023). Evaluation of Data Clustering Accuracy using K-Means Algorithm. *International Journal of Multidisciplinary Approach Research and Science*, 2(01), 385–396. <https://doi.org/10.59653/ijmars.v2i01.504>

- Syahfitri, N., Budianita, E., Nazir, A., & Afrianty, I. (2023). Pengelompokan Produk Berdasarkan Data Persediaan Barang Menggunakan Metode Elbow dan K-Medoid. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, 4(3), 1668–1675. <https://doi.org/10.30865/klik.v4i3.1525>
- Undari Sulung, M. M. (2024). Jurnal Edu Research Indonesian Institute For Corporate Learning And Studies (IICLS) Page 25. *Jurnal Edu Research : Indonesian Institute For Corporate Learning And Studies (IICLS)*, 5(2), 28–33.
- Zhao, L., Fang, J., Ji, Y., Zhang, Y., Zhou, X., Yin, J., Zhang, M., & Bao, W. (2023). K-means cluster analysis of characteristic patterns of allergen in different ages: Real life study. *Clinical and Translational Allergy*, 13(7). <https://doi.org/10.1002/clt2.12281>