Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.5900

# Implementation of K-Means Clustering for Social Assistance Recipients with Silhouette Score Evaluation

**Herfia Rhomadhona[1*], Wiwik Kusrini[2], Winda Aprianti[3], Jaka Permadi[4]**
[1,2,3,4]Politeknik Negeri Tanah Laut, Indonesia
[1]herfia.rhomadhona@politala.ac.id, [2]wiwik.kusrini@politala.ac.id, [3]winda@politala.ac.id,
[4]jakapermadi.88@politala.ac.id

open access

***Corresponding Author**

**Article History:**
Submitted: 14-05-2025
Accepted: 19-05-2025
**Published: 28-05-2025**

**Keywords:**
Data Mining; K-Means Clustering; Silhouette Coefficient; Social Assistance.



**ABSTRACT**

The distribution of direct social assistance continues to face several challenges, particularly regarding inaccurate targeting and unequal allocation. One of the main causes of this issue is the lack of transparency in the distribution process, where assistance is often granted to individuals with familial ties to local committee members or even government officials. As a result, the groups most in need frequently do not receive the aid they deserve. This condition is also evident in Tanjung Village, Bajuin Subdistrict, Tanah Laut Regency. The manual process of grouping prospective aid recipients contributes to inaccuracies in targeting, which in turn leads to public dissatisfaction. To address this issue, this study applies the K-Means Clustering method to group potential social assistance recipients using data from 150 individuals and three main attributes: age, occupation, and income. The method clusters the data based on the similarity of characteristics, thus supporting a more equitable and efficient identification process. The evaluation is conducted using the Silhouette Coefficient to assess the quality of clustering. The results indicate that the highest Silhouette Score is achieved at k=2k = 2k=2, with a value of 0.8278, suggesting that dividing the data into two clusters provides the most optimal configuration. The Silhouette Score tends to decrease as the number of clusters increases, confirming that adding more clusters does not necessarily improve the quality of separation.

## INTRODUCTION

The distribution of social assistance is a key government strategy to reduce poverty levels and improve the welfare of economically disadvantaged and vulnerable groups. Despite its strategic nature, implementation still faces numerous significant challenges, particularly regarding targeting accuracy and equitable distribution. The Ministry of Social Affairs of the Republic of Indonesia reported that there are still beneficiaries who do not meet the established eligibility criteria, resulting in public dissatisfaction and potential social conflict. These issues are further supported by findings from the Indonesian Ombudsman, which revealed that inaccuracies in the distribution of social assistance not only exacerbate social inequality but have also led to substantial state losses.

This problem is not only present at the national level but also evident in local contexts. A concrete example can be observed in Tanjung Village, Bajuin District, Tanah Laut Regency, which has a population of 4,092 (BPS, 2023). In this village, the process of identifying prospective aid recipients is still conducted manually and heavily relies on the subjective judgment of village officials. This condition creates room for bias, whereby families with close ties to village authorities are more likely to receive assistance regardless of their actual socioeconomic status. Such practices undermine the effectiveness of social aid programs and intensify social tensions at the community level.

To address these issues, a data-driven approach is required to improve objectivity in the recipient selection process. One relevant and promising method is the K-Means clustering algorithm. This technique, a form of unsupervised machine learning, is capable of grouping individuals into clusters based on similarities in socioeconomic attributes, such as income, number of dependents, housing conditions, and education level. By applying this algorithm, the selection process can be carried out systematically and based on evidence, thereby reducing subjective judgment and potential abuse of authority at the local level (Shahapure & Nicholas, 2020). Furthermore, cluster quality can be evaluated using internal metrics such as the Silhouette Score, which assesses how well each data point fits into its assigned cluster without requiring ground truth labels (Aprianti & Permadi, 2018; Kasim, Bahri, & Amir, 2021; Nugraha, Laxmi, & Riana, 2024).

Considering the complexity of the problem and the potential of technology-based solutions, this study aims to implement the K-Means algorithm for clustering prospective social assistance recipients based on socioeconomic attributes. This study also evaluates various cluster configurations (k-values) using the Silhouette Score to objectively determine the optimal number of clusters. The results of this study are expected to contribute to more accurate targeting

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.5900

of social assistance and support decision-making processes that are more transparent, fair, and data-driven, particularly in regions where manual selection methods remain vulnerable to subjectivity.

## LITERATURE REVIEW

The K-Means clustering algorithm is widely used to segment prospective recipients of social assistance based on poverty characteristics and socioeconomic indicators. The primary objective is to ensure objective identification of eligible and non-eligible individuals, thereby enhancing the accuracy of aid targeting. Several studies have implemented K-Means for this purpose.

For instance, several studies clustered 400 prospective beneficiaries of the Family Hope Program (PKH) in Lemberang Village using the K-Means method; however, these studies did not include cluster validation analysis such as the Silhouette Score (Sari & Utamajaya, 2022). Likewise, researchers who compared the K-Means and K-Medoids algorithms to map stunting-prone areas focused solely on classification accuracy without validating the internal quality of the clusters (Nur & Abdurakhman, 2024). In line with previous studies, research on the implementation of K-Means Clustering to classify families based on economic status for the distribution of the Kartu Indonesia Pintar assistance program also shows limitations, namely a focus on result accuracy without including cluster quality evaluation such as the Silhouette Score, thus the validity of the clustering still requires further review (Fammaldo & Hakim, 2018; Sriaji, Zaenah, Istiawan, Prayogi, & Mahiruna, 2021).

The Silhouette Score is a well-recognized internal metric used to evaluate the consistency and separation of clusters generated by algorithms like K-Means (Shahapure & Nicholas, 2020). It measures how appropriately each data point fits within its own cluster relative to others, with higher values closer to 1 indicating better clustering performance. Researchers have combined the Silhouette Score with other internal evaluation methods, such as the Elbow Method and the Davies–Bouldin Index, to improve clustering outcomes in various domains, including employee performance evaluation and student dropout prediction (Amrulloh, Pudjiantoro, Sabrina, & Hadiana, 2022; Singh, Mittal, Malhotra, & Srivastava, 2020). These internal metrics are particularly valuable for unsupervised learning tasks, as they assess clustering quality without requiring labeled data (Mulyani, Setiawan, & Fathi, 2023; Rhomadhona & Permadi, 2019).

## METHOD

The implementation of K-Means is carried out through the stages of the data mining process, namely:

### Data Collection

Data collection was carried out through an interview with the Secretary of Tanjung Village, Bajuin Subdistrict, regarding social assistance recipients in 2024. Through this process, several criteria were identified as the basis for determining social assistance eligibility. A detailed overview of the criteria used is presented in Table 1.

Table 1. The Significance of The Relationship in The Model

| No. | Attribute Name | Data Type | Description |
|---|---|---|---|
| 1 | Age | Numeric (Integer) | The age of the individual in years |
| 2 | Occupation | Cattegorical | Type of job or source of livelihood of the respondent |
| 3 | Income | Numeric (Integer) | Monthly income of the individual in Indonesian Rupiah |

### Data Preprocessing

Data preprocessing is an essential step in the data mining process, as the quality of analysis depends heavily on the quality of the data (Joshi & Patel, 2020; Saraswat & Raj, 2022). In this study, preprocessing was conducted to ensure that the social assistance recipient data had a consistent format and was ready for processing using the K-Means algorithm. The steps included Data cleaning, to handle missing or invalid values, Data transformation, by converting categorical attributes such as occupation and income into numerical values, and Normalization, to scale all attributes uniformly using the Min-Max Scaling method.

### Implementation of Data Mining

The data mining method employed in this study is the K-Means Clustering algorithm. The K-Means algorithm can be illustrated using the flowchart shown in Figure 1.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 5, Number 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.5900

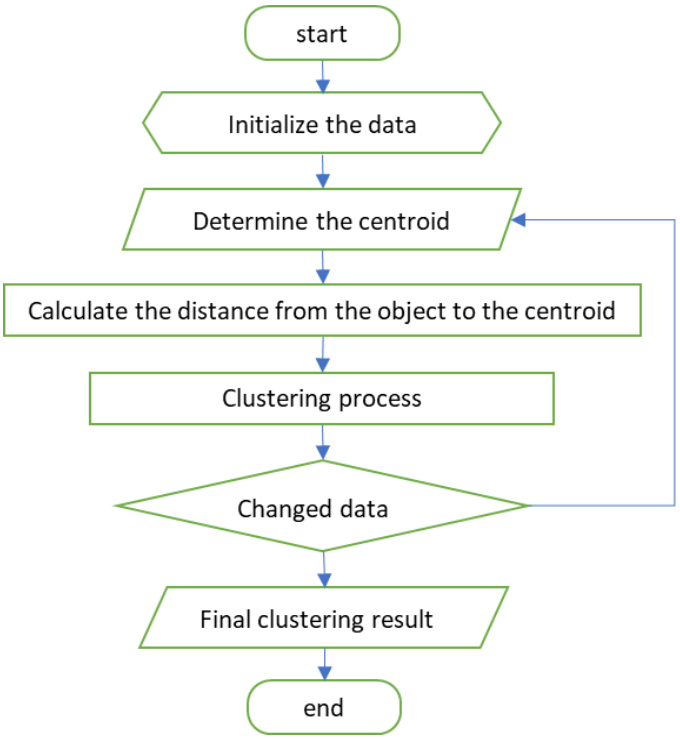


Figure 1. Flowchart of K-Means Clustering algorithm implementation

The implementation of the K-Means algorithm follows the stages outlined, which include:

1. determining the number of clusters
2. initializing the centroids
3. assigning each data point to the nearest centroid based on a distance metric

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2}$$

4. updating the centroid positions by calculating the mean of the assigned points
5. iteratively repeating the assignment and update steps until convergence is achieved

**Evaluation of the Number of Clusters**

The method used to evaluate the quality of clustering results and to determine the optimal number of clusters in the application of the K-Means algorithm is the Silhouette Coefficient. The Silhouette score is a widely used metric for evaluating cluster quality in K-means and other clustering algorithms (Shahapure & Nicholas, 2020). It measures how well data points fit within their assigned clusters compared to other clusters. Higher Silhouette scores indicate better clustering, with values closer to 1 being optimal. The method can help determine the optimal number of clusters by comparing scores across different k values (Shahapure & Nicholas, 2020). The Silhouette value is calculated using the following formula:

$$S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

where:

$a(i)$ : the average distance from data point iii to all other points within the same cluster (intra-cluster distance).

$b(i)$ : the minimum average distance from data point iii to all points in the nearest neighboring cluster (nearest-cluster distance).

$S(i)$ : yields a value between -1 and +1:

$S(i) \approx +1$ : the point is appropriately clustered

$S(i) \approx +0$ : the point lies on the boundary between two clusters

$S(i) \approx -1$ : the point might have been assigned to the wrong cluster

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, Number 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.5900

## RESULT

### Data Collection

Table 2 presents the raw data collected from 150 residents of Tanjung Village, Bajuin District, as the primary dataset for this study. Each row represents an individual respondent, with attributes including age, occupation, and estimated monthly income. The data was obtained through field interviews and structured questionnaires conducted in collaboration with village authorities. The majority of respondents are engaged in farming or are housewives with relatively low income levels, falling within the range of IDR 0 to 1,500,000 per month. These socioeconomic indicators were selected as input variables for clustering, as they are directly relevant to the eligibility criteria for social assistance programs. The data was subsequently preprocessed and normalized prior to analysis using the K-Means clustering algorithm.

Table 2. Dataset of social assistance beneficiaries

| Candidate | Age | Occupation | Income |
|---|---|---|---|
| 1 | 69 | Farmers | 500.000 - 1.500.000 |
| 2 | 52 | Farmers | 500.000 - 1.500.000 |
| 3 | 43 | Housewife | 0 - 500.000 |
| 4 | 42 | Housewife | 0 - 500.000 |
| 5 | 32 | Housewife | 0 - 500.000 |
| … | … | … | … |
| 146 | 53 | Farmers | 500.000 - 1.500.000 |
| 147 | 39 | Housewife | 0 - 500.000 |
| 148 | 55 | Farmers | 500.000 - 1.500.000 |
| 149 | 67 | Housewife | 0 - 500.000 |
| 150 | 45 | Farmers | 500.000 - 1.500.000 |

### Data Preprocessing

The preprocessing stage begins with data cleaning, where the data set is checked to ensure there are no missing values, duplicate records, or invalid entries. In this study, all records were found to be complete and consistent; therefore, no deletion or imputation procedures were required.

Next, data transformation was performed to convert categorical attributes into numeric values, since the K-Means algorithm requires numeric input for its distance calculation. Specifically, the Occupation attribute was coded as 1 for Farmer and 0 for Housewife. Meanwhile, the Income attribute was transformed based on the category range, where “0 - 500,000” is represented by 250,000 and “500,000 - 1,500,000” is represented by 1,000,000. This numeric mapping provides a meaningful representation of the categorical variables, which allows for accurate clustering during the modeling process.

After transformation, the dataset underwent normalization using the Min-Max method to scale all numerical attributes Age, Occupation, and Income to a standard range of [0, 1]. This step was crucial to ensure that no single attribute would dominate the clustering process due to differences in magnitude or scaleThe results of the Preprocessing and Normalization are presented in Table 3.

Table 3. The results of preprocessing and normalization

| Candidate | Age | Occupation | Income | Age | Occupation | Income |
|---|---|---|---|---|---|---|
| 1 | 69 | 1 | 1.000.000 | 0.770833 | 1.0 | 1.0 |
| 2 | 52 | 1 | 1.000.000 | 0.416667 | 1.0 | 1.0 |
| 3 | 43 | 0 | 250.000 | 0.229167 | 0.0 | 0.0 |
| 4 | 42 | 0 | 250.000 | 0.208333 | 0.0 | 0.0 |
| 5 | 32 | 0 | 250.000 | 0.000000 | 0.0 | 0.0 |
| … | … | … | … | … | … | … |
| 146 | 53 | 1 | 1.000.000 | 0.437500 | 1.0 | 1.0 |
| 147 | 39 | 0 | 250.000 | 0.145833 | 0.0 | 0.0 |
| 148 | 55 | 1 | 1.000.000 | 0.479167 | 1.0 | 1.0 |
| 149 | 67 | 0 | 250.000 | 0.729167 | 0.0 | 0.0 |
| 150 | 45 | 1 | 1.000.000 | 0.270833 | 1.0 | 1.0 |

This study utilized a sample of 15 individuals who are prospective recipients of social assistance. Before conducting the clustering process, a preprocessing stage was carried out to ensure that all attributes were in a format suitable for analysis. Preprocessing involves transforming categorical attributes, namely Occupation and Income, into numerical values, followed by normalizing all attributes (Age, Occupation, and Income) using the Min-Max Scaling method. The results of preprocessing and normalization are presented in Table 2.

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.5900

**Implementation of K-Means Clustering**

After preprocessing the social assistance data, the K-Means Clustering algorithm was applied to group individuals based on the attributes of age, occupation, and income. At this stage, the clustering results were compared across different values of k, ranging from 2 to 10. The results of the comparison are presented in Table 4.

Table 4. Transformation and grouping of potential aid recipients

| Candidate | Transformation Data | | | Cluster Order | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Age | Occupation | Income | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster … | Cluster 9 | Cluster 10 |
| 1 | 69 | 1 | 1.000.000 | 0 | 0 | 3 | 1 | … | 7 | 5 |
| 2 | 52 | 1 | 1.000.000 | 0 | 0 | 1 | 3 | … | 3 | 7 |
| 3 | 43 | 0 | 250.000 | 1 | 1 | 0 | 4 | … | 5 | 6 |
| 4 | 42 | 0 | 250.000 | 1 | 1 | 0 | 4 | … | 5 | 6 |
| 5 | 32 | 0 | 250.000 | 1 | 2 | 0 | 4 | … | 5 | 2 |
| … | … | … | … | … | … | … | … | … | … | … |
| 146 | 53 | 1 | 1.000.000 | 0 | 0 | 1 | 3 | … | 3 | 7 |
| 147 | 39 | 0 | 250.000 | 1 | 1 | 0 | 4 | … | 5 | 2 |
| 148 | 55 | 1 | 1.000.000 | 0 | 0 | 1 | 3 | … | 3 | 7 |
| 149 | 67 | 0 | 250.000 | 1 | 2 | 2 | 0 | … | 4 | 0 |
| 150 | 45 | 1 | 1.000.000 | 0 | 0 | 1 | 3 | … | 6 | 3 |

Table 4 presents the transformation and clustering results of individuals identified as prospective social assistance recipients using the K-Means Clustering algorithm, with varying numbers of clusters from *k=2 to k=10*. From this table, it can be observed that some individuals exhibit relatively stable clustering patterns, particularly at lower values of *k*. For example, individuals characterized as housewives with low income tend to remain in the same cluster from *k=2 to k=4*, indicating consistency in the grouping of the lower socioeconomic segment. However, as the number of clusters increases, a shift in cluster assignment occurs for several individuals.

**Evaluation**

After the clustering process using the K-Means algorithm was completed, the next step was to evaluate the quality and cohesion of the resulting clusters through Silhouette Coefficient analysis. The results of this evaluation are presented in Table 5.

Table 5. Silhouette Coefficient Analysis

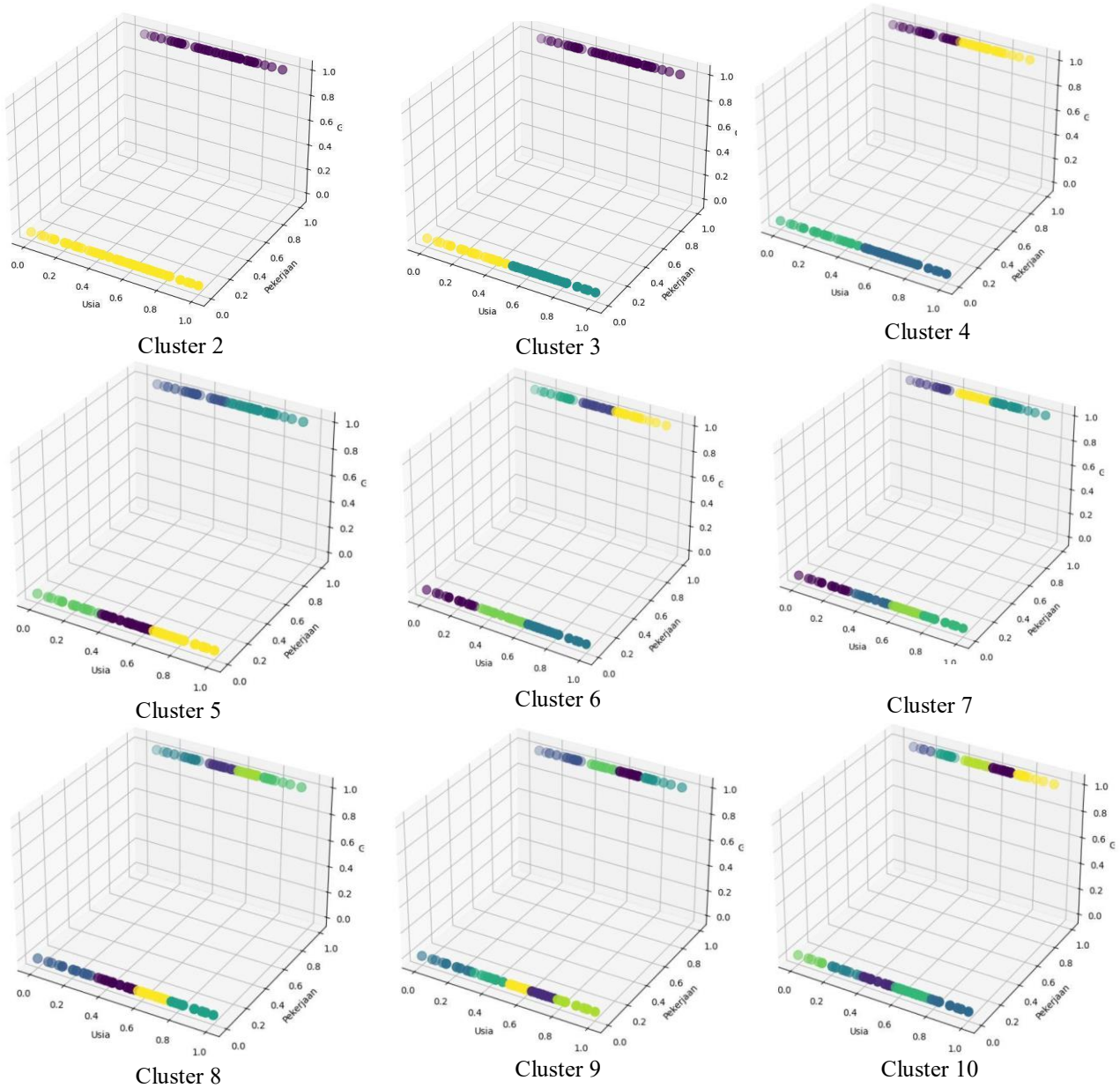
| Cluster | Sillhouette Coeficients |
|---|---|
| 2 | 0.8278 |
| 3 | 0.7265 |
| 4 | 0.6176 |
| 5 | 0.5937 |
| 6 | 0.5783 |
| 7 | 0.5828 |
| 8 | 0.5872 |
| 9 | 0.5851 |
| 10 | 0.5868 |

## DISCUSSION

To provide a clearer picture of the data distribution patterns based on the clustering results, a 3D visualization was performed for three different cluster scenarios: 2 until 10 clusters. This visualization considers three main normalized attributes, namely age, occupation, and income. The following figure shows the data grouping results in cluster form.

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.5900




Figrue 2. Grouping the number of clusters k-2 to k10

The clustering results with the number of clusters ranging from *k=2 to k=10* are visualized side by side in 3D plots, as shown in Figure 2. Each plot illustrates the distribution of the population data based on the normalized attributes of age, occupation, and income. The colors in the plots represent the cluster assignments produced by the K-Means algorithm. From the visualizations, it is evident that at *k=2* the cluster separation is relatively clear with two dominant groups distinguished by occupation and income level. As the value of *k* increases, the clustering becomes more granular. However, the boundaries between clusters become increasingly blurred in several configurations, especially from *k=5* onwards. This indicates that increasing the number of clusters does not necessarily result in more meaningful separations.

To further validate that *k=2* represents the most optimal cluster division, an evaluation was conducted using the Silhouette Score for cluster counts ranging from *k=2 to k=10*. The Silhouette Score was chosen because it effectively measures clustering quality by considering intra-cluster cohesion and inter-cluster separation. This score ranges from -1 to 1, where values closer to 1 indicate better clustering quality. The following figure illustrates the Silhouette Score graph across nine cluster configurations, based on the data presented in Figure 3.

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.5900

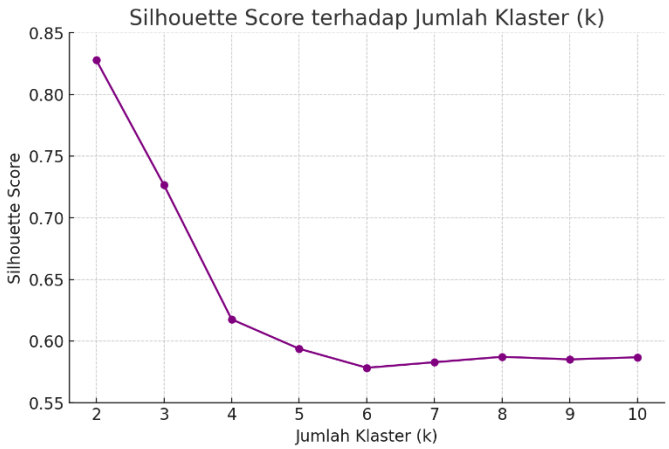



Figure 3. Silhouette Score

Based on the evaluation results visualized in Figure 3, the highest Silhouette Score is achieved at *k=2*, with a value of 0.8278. This indicates that the data structure is most optimal when divided into two clusters. Increasing the number of clusters beyond *k=2* leads to a significant decrease in the Silhouette Score, suggesting that the separation between clusters becomes less distinct and tends to result in over clustering.

## CONCLUSION

Based on the implementation of the K-Means algorithm on the data of prospective social assistance recipients, as well as the visualization of clustering results in the form of 3D graphs, a clear distinction can be observed between groups based on the attributes of age, occupation, and income. The grid visualization for cluster counts ranging k=2 to k=10 indicates that the most representative separation occurs when k=2, where two main groups consistently emerge: one comprising housewives with low income, and the other consisting of farmers with higher income. As the value of k increases, the distribution of clusters becomes more dispersed, yet tends to result in divisions that are less meaningful and split groups that are otherwise homogeneous**.**

This observation is consistent with the evaluation results using the Silhouette Score, which reached its highest value at k=2 specifically 0.8278, indicating that two clusters are the most optimal number for this dataset. The Silhouette Score tends to decrease as the number of clusters increases, confirming that adding more clusters does not necessarily improve the quality of separation.

Therefore, it can be concluded that the K-Means algorithm is effective in grouping prospective social assistance recipients into two primary categories, and this approach can serve as a basis for more objective decision making in targeting aid distribution more accurately.

## REFERENCES

Amrulloh, K., Pudjiantoro, T. H., Sabrina, P. N., & Hadiana, A. I. (2022). Comparison Between Davies-Bouldin Index and Silhouette Coefficient Evaluation Methods in Retail Store Sales Transaction Data Clusterization Using K-Medoids Algorithm. *In Proceedings of the 3rd South American International Industrial Engineering and Operations Management Conference, Asuncion*, (pp. 1952-1961). Paraguay.

Aprianti, W., & Permadi, J. (2018). K-means clustering untuk data kecelakaan lalu lintas jalan raya di Kecamatan Pelaihari. *Jurnal Teknologi Informasi dan Ilmu Komputer, 5*(5), 613-620.

BPS. (2023). *Kabupaten Tanah Laut dalam Angka.* Pelaihari: Badan Pusat Statistik Kabupaten Tanah Laut. Retrieved from https://tanahlautkab.bps.go.id.

Fammaldo, E., & Hakim, L. (2018). Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Tingkat Kesejahteraan Keluarga Untuk Program Kartu Indonesia Pintar. *Jurnal Ilmiah Teknologi Infomasi Terapan, 5*(1), 23-31.

Joshi, A. P., & Patel, B. V. (2020). Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process. *Orient.J. Comp. Sci. and Technol, 13*(2-3), 78–81.

Kasim, R. J., Bahri, S., & Amir, S. (2021). Implementasi Metode K-Means Untuk Clustering Data Penduduk Miskin Dengan Systematic Random Sampling. *Prosiding SISFOTEK. 5*, pp. 95-101. Padang: Ikatan Ahli Informatika Indonesia (IAII).

Mulyani, H., Setiawan, R. A., & Fathi, H. (2023). Optimization Of K Value In Clustering Using Silhouette Score (Case Study: Mall Customers Data). *Journal Of Information Technology And Its Utilization, 6*(2), 45-50.

Nugraha, R. P., Laxmi, G. F., & Riana, F. (2024). Penerapan K-Means++ Untuk Pengelompokan Mahasiswa Berpotensi

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.5900

Drop Out Studi Kasus: Universitas Ibn Khaldun Bogor. *JATI (Jurnal Mahasiswa Teknik Informatika), 8*(3), 3493-3500.

Nur, I. M., & Abdurakhman, A. (2024). Analysis of Social Vulnerability in Java Island using K-Medoids Algorithm with Variation of Distance Measurements (Euclidean, Manhattan, Minkowski). *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), 7*(2), 467-475.

Rhomadhona, H., & Permadi, J. (2019). Klasifikasi Berita Kriminal Menggunakan Naive Bayes Classifier (NBC) dengan Pengujian K-Fold Cross Validation. *Jurnal Sains dan Informatika, 5*(2), 108-117.

Saraswat , P., & Raj, S. (2022). Data Pre-processing Techniques in Data Mining: A Review. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST), 10*(1), 122-125.

Sari, S., & Utamajaya, J. N. (2022). Sistem Pendukung Keputusan Penerima Bantuan Langsung Tunai Dana Desa Menggunakan Metode Algoritma K-Means Clustering. *Jurnal JUPITER, 14*(1), 150-160.

Shahapure, K. R., & Nicholas, C. (2020). Cluster Quality Analysis Using Silhouette Score. *International Journal of Computer Applications. 182*, pp. 1-5. Sydney: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA).

Singh, A. K., Mittal, S., Malhotra, P., & Srivastava, Y. V. (2020). Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, (pp. 306-310). Erode, India.

Sriaji, C. D., Zaenah, Z., Istiawan, D., Prayogi, S. Y., & Mahiruna, A. (2021). Pengelompokan Tingkat Kemiskinan Kabupaten/Kota Di Provinsi Jawa Timur Mengunakan Algoritma K-Means++. *Journal of Applied Statistics and Data Mining, 2*(2), 29-40.