E-ISSN: XXXX - XXXX
Networked Knowledge and Research in Informatic

Networked Knowledge and Research in Informatic
Vol. 1 No. 1 (2026) : 01-08
Available online at : https://ejournal.unuja.ac.id/index.php/NKRI

Clustering Underprivileged Student Aid Recipients Using Fuzzy C-Means

Wahab Sya'roni [1]
[1] Universitas Nurul Jadid, Indonesia
wahab@unuja.ac.id [1]

Doi
Submitted: 01 Sep 2026 Revision: 15 Sep 2026 Accepted: 25 Sep 2026 Published: 30 Sep 2026

Abstract Social assistance distribution in boarding-school educational settings requires a consistent mapping of prospective recipients so that the verification process does not rely solely on manual assessment. This study aims to cluster the socioeconomic profiles of students as a supporting basis for determining aid verification priorities using Fuzzy C-Means and to present the results through a Streamlit application. The research data consisted of 843 records of female secondary-level students from the 2022 cohort, obtained from the boarding school data management unit. After data cleaning by removing inactive student records, 781 records were processed using the attributes of number of siblings, parents' occupation, and parents' income. Categorical data were transformed into numerical codes and clustered using two clusters, a fuzziness parameter of 2, a maximum of 100 iterations, and a tolerance value of 0.00001. The Fuzzy C-Means process converged at the 28th iteration and produced Cluster 1 with 635 students and Cluster 2 with 146 students. The Silhouette Coefficient value of 0.4800 indicates a moderate level of cluster separation. The Streamlit application displays the initial data, preprocessing results, data transformation, and clustering results interactively. The findings indicate that Fuzzy C-Means can be used as a supporting tool for mapping prospective student aid recipients. However, clustering results should not be used as an automatic decision; aid recommendations must still involve document verification and staff assessment to prevent mistargeting.

Keywords *Aid Recipients; Clustering; Fuzzy C-Means; Pre-Prosperous Students; Streamlit*

1. INTRODUCTION

Social assistance plays an important role in reducing the economic burden of groups that are unable to meet their basic needs. However, the distribution of assistance still faces the risk of mistargeting when the mapping of prospective recipients is conducted manually, uses outdated data, or is not supported by measurable criteria. This condition may cause students who genuinely need assistance to remain unidentified, while those who are less eligible may receive the benefit. Therefore, social assistance administrators require an analytical mechanism that can provide a consistent overview of recipient groups as a basis for further verification (Herdiana, 2020). In boarding school environments, mapping the socioeconomic conditions of students has distinctive characteristics because the available data come from family information, student activity status, and educational records. Previous research involving student data applied K-Means to form four welfare clusters and demonstrated that clustering can help accelerate the management of prospective aid recipient data (Sudriyanto et al., 2023). However, hard clustering methods assign each record strictly to only one group, even though socioeconomic conditions may lie near the boundary between profiles and require a more flexible membership approach.

Fuzzy C-Means is a fuzzy-based clustering method that assigns a degree of membership for each data point to all clusters. This approach allows a data point to have a relative closeness to more than one cluster before it is assigned to the cluster with the highest membership value (Bezdek, 1981). Several studies have applied Fuzzy C-Means for aid recipient mapping, customer segmentation, and socioeconomic profile analysis because the method can separate group characteristics based on cluster centers and membership distributions (Ananda & Monalisa, 2023; Huddin et al., 2023; Supriyanti et al., 2024). Nevertheless, clustering results cannot be directly used as automatic labels of aid eligibility. Many studies stop at cluster formation or only present calculation results through notebooks, meaning that interpretation and verification by administrators are not supported by an accessible interface. In addition, clustering quality should be reported separately from accuracy because clustering is an unsupervised learning method. The Silhouette Coefficient can be used to evaluate the compactness of members within clusters and the separation between clusters, where higher values indicate a clearer cluster structure (Rousseeuw, 1987).

This gap forms the basis of the present study. Unlike previous research, this study applies Fuzzy C-Means to map the profiles of active students based on number of siblings, parents' occupation, and parents' income, followed by the integration of results into a Streamlit web application. The contribution of this study lies in combining fuzzy-based mapping, evaluation using the Silhouette Coefficient, and an interface that can be used by administrators for document review. Streamlit was selected because it enables Python-based analytical code to be presented through an interactive interface without requiring full conventional web application development (Abu, 2023).

This study aims to apply Fuzzy C-Means to form groups of prospective student aid recipients, measure cluster quality using the Silhouette Coefficient, and provide a Streamlit interface for presenting the clustering process and results. The findings are expected to support a more systematic initial identification process, while final

aid distribution decisions remain subject to staff verification based on documents and the actual condition of prospective recipients.

2. RESEARCH METHODS

This study employed a quantitative experimental design using a data mining approach. The research flow consisted of data collection, data selection, data cleaning, attribute transformation, cluster formation using Fuzzy C-Means, evaluation using the Silhouette Coefficient, and Streamlit interface implementation. All findings were presented in aggregate form to reduce the risk of disclosing students' personal identities.

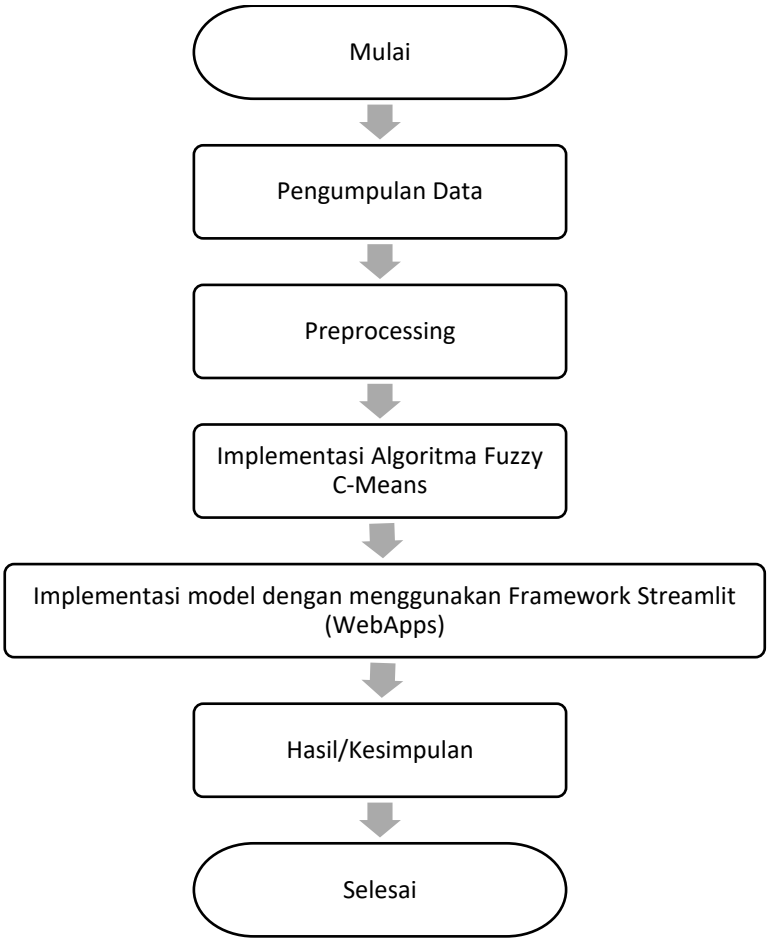

Figure 1. Research Flow Diagram

2.1. Dataset

The dataset was obtained from the boarding school data management unit and contained records of female secondary-level students from the 2022 cohort. The initial dataset consisted of 843 records with attributes including name, number of siblings, enrollment year, student status, institution, major, parents' occupation, and parents' income. Student names were not used in the analytical process. The data were only used for internal mapping purposes, while research results were reported in the form of data totals, cluster centers, and clustering summaries.

Table 1. Dataset Composition

Stage	Number of Records	Description
Initial data	843	All records obtained from the data management unit

Active data after cleaning	781	Active student records processed in the study
Removed data	62	Inactive or withdrawn student records

2.2. Data Preparation

The data selection stage retained three attributes considered relevant to socioeconomic profiles: number of siblings, parents' occupation, and parents' income. Data cleaning was performed by removing records of inactive or withdrawn students. Next, categorical attributes were transformed into numerical codes to enable processing using the Fuzzy C-Means algorithm. These transformations represent operational coding of the source categories and should not be interpreted as absolute economic measurements.

	Jumlah Saudara	Pekerjaan Orang Tua	Penghasilan Orang Tua
0	3	Petani	Rp.1.000.000
1	3	Wiraswasta	Rp.500.000
2	2	Pedagang kecil	Rp.1.500.000
3	2	Pedagang kecil	Rp.2.000.000
4	4	Guru	Rp.1.000.000
...	...	...	...
776	1	Buruh	Rp. 500.000
777	1	Karyawan Swasta	Rp.2.000.000
778	3	Karyawan Swasta	Rp.2.000.000
779	3	Pedagang kecil	Rp.1.000.000
780	2	Pedagang kecil	Rp.1.500.000

781 rows × 3 columns

Figure 2. Student Data Cleaning Process

Table 2. Attributes Used for Clustering

Attribute	Initial Type	Transformation	Role
Number of siblings	Discrete numeric	Code 0–5 for 1–6 siblings	Clustering feature
Parents' occupation	Categorical	Occupational category code	Clustering feature
Parents' income	Ordinal categorical	Code 0–12 from lowest to highest income category	Clustering feature

2.3. Fuzzy C-Means Clustering

Fuzzy C-Means groups data by updating the membership matrix and cluster centers iteratively. The objective function is minimized until the change in the membership matrix is below the specified tolerance value. This study used two clusters to map two primary profiles: a profile with relatively lower economic indicators and a profile with relatively higher economic indicators.

Table 3. Fuzzy C-Means Parameters

Parameter	Value
Number of clusters	2
Fuzziness parameter	2
Maximum iterations	100
Convergence tolerance	0.00001

2.4. Evaluation and Streamlit Deployment

Cluster quality was evaluated using the Silhouette Coefficient. This metric ranges from -1 to 1, where a higher positive value indicates that cluster members are closer to their own cluster than to other clusters. In addition to evaluation, the results were integrated into a Streamlit application that displays the initial data, cleaned data, attribute transformation, Fuzzy C-Means parameters, and clustering summaries. The interface serves as a monitoring tool and does not replace social verification by administrators.

3. RESULTS

3.1. Data Preparation Results

Of the 843 initial records, 62 records with inactive or withdrawn status were removed, resulting in 781 records used for the clustering process. Transformation was applied to ensure that all attributes could be processed numerically. The income codes were arranged from the lowest income category to the highest income category. Therefore, the income centroid values should be interpreted as relative tendencies within the coding scheme rather than actual income values.

3.2. Clustering Results

The Fuzzy C-Means algorithm converged at the 28th iteration, which was below the maximum limit of 100 iterations. The final process produced two clusters with unequal sizes.

Table 4. Final Clustering Results

Cluster	Number of Students	Sibling Centroid	Occupation Centroid	Income Centroid	Relative Profile
Cluster 1	635	2.386	3.295	2.097	Relatively lower income code
Cluster 2	146	2.654	4.729	8.480	Relatively higher income code

The income centroid value of Cluster 1 was lower than that of Cluster 2 within the transformation scheme used in this study. Therefore, Cluster 1 can be used as an initial group for reviewing potential aid needs, while Cluster 2 represents a profile with relatively higher income codes. Final aid priority should not be determined solely based on clustering results because the parents' occupation attribute is categorical, and its numerical representation may influence Euclidean distance calculations. Administrative verification, family condition assessment, and approval from relevant administrators remain necessary before aid distribution.

Figure 3. Fuzzy C-Means Clustering Visualization

3.3. Cluster Quality Evaluation

The evaluation produced a Silhouette Coefficient of 0.48001974885927373. This value indicates a moderate level of cluster separation. In other words, members within each cluster have sufficient internal similarity, but the boundary between clusters is not completely strong. Therefore, the Fuzzy C-Means results are appropriate as supporting information for initial mapping rather than as the sole basis for determining aid eligibility.

3.4. Streamlit Implementation

The model was implemented in a Streamlit application to make the analytical process easier for administrators to review. The interface displays the initial dataset, cleaned data, transformed data, number of clusters, fuzziness parameter, maximum iterations, and error tolerance. When the clustering process button is activated, the application displays clustering summaries interactively. The web-based presentation accelerates access to analytical results. However, operational use of the system should implement access control, minimize the display of personal identities, and record decision-making processes to ensure that students' socioeconomic data remain protected.

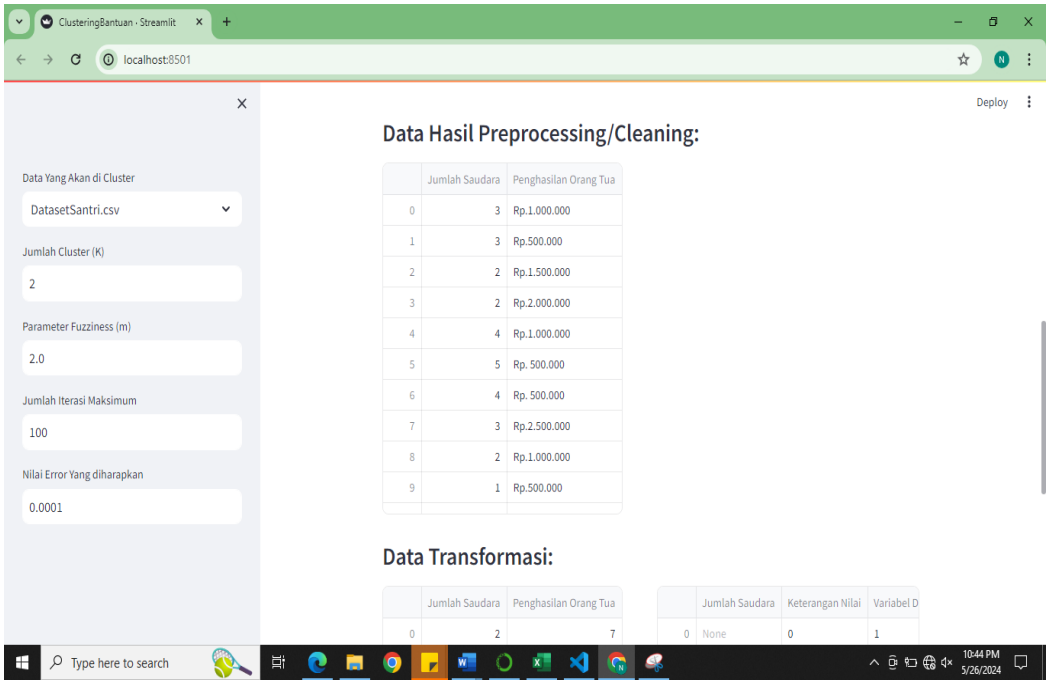

Figure 4. Streamlit Application Interface

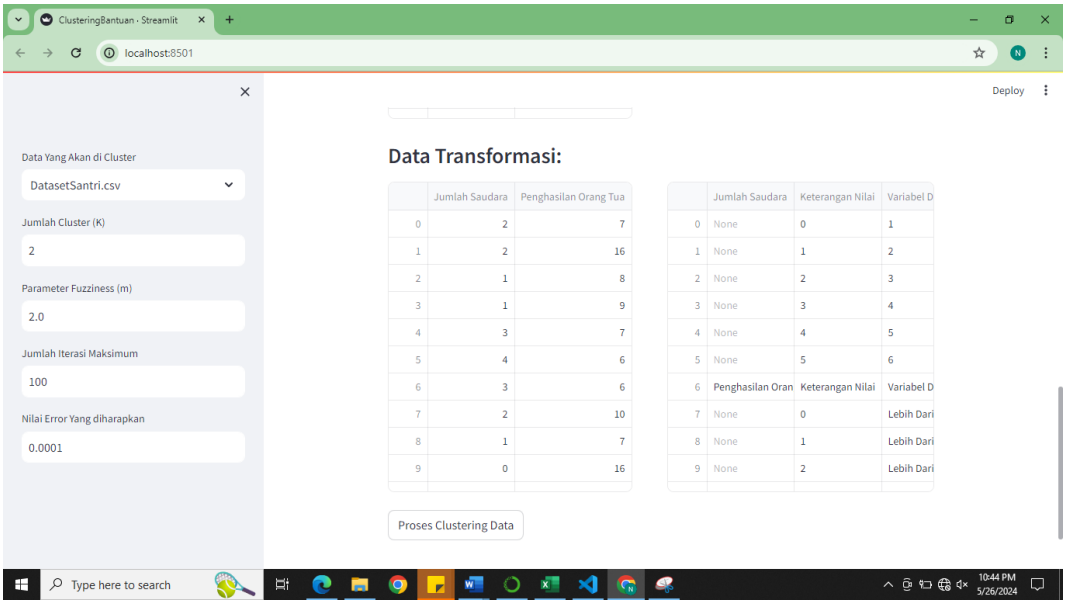

Figure 5. Clustering Results Displayed in the Streamlit Application

4. CONCLUSIONS

Fuzzy C-Means was successfully applied to map 781 active student records based on the number of siblings, parents' occupation, and parents' income. Using two clusters, a fuzziness parameter of 2, and a tolerance value of 0.00001, the clustering process converged at the 28th iteration. Cluster 1 consisted of 635 students with an income-code centroid of 2.097, while Cluster 2 consisted of 146 students with an income-code centroid of 8.480. The Silhouette Coefficient value of 0.4800 indicates a moderate level of cluster separation.

The Streamlit implementation made the data review and clustering process easier to access for administrators. In practice, clustering results can be used to prepare an initial list for verification rather than as an automatic decision regarding aid eligibility. The main limitation of this study is the use of numerical coding for parents' occupation, which is categorical in nature; therefore, distances between occupation categories may not fully represent differences in socioeconomic conditions.

Future research is recommended to use coding schemes validated by domain experts or one-hot encoding, add verifiable attributes such as number of dependents and housing conditions, compare results with other clustering methods, and conduct bias audits and field verification before applying the results to aid decisions.

5. REFERENCES

Abu, T. (2023). Buku referensi implementasi algoritma machine learning berbasis web dengan framework Streamlit. Pustaka Nurja. https://pustakanurja.unuja.ac.id

Alvianta, F. N., Prabowo, A. A., & Komarudin, A. (2021). Pembinaan kesejahteraan keluarga dalam pemberdayaan keluarga prasejahtera. JISIP: Jurnal Ilmu Sosial dan Pendidikan, 5(3), 137–151. https://doi.org/10.36312/jisip.v5i3.2095

Ananda, D. P., & Monalisa, S. (2023). Segmentasi pelanggan B2B dengan model LRFM menggunakan algoritma Fuzzy C-Means pada Rotte Bakery. Jurnal Teknologi Informasi dan Ilmu Komputer, 10(5). https://doi.org/10.25126/jtiik.2023106569

Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algorithms. Plenum Press.

Herdiana, D. (2020). Pengawasan kolaboratif dalam pelaksanaan kebijakan bantuan sosial terdampak COVID-19. JDP: Jurnal Dinamika Pemerintahan, 3(2), 85–99. https://doi.org/10.36341/jdp.v3i2.1323

Huddin, S., Haerani, E., Jasril, J., & Oktavia, L. (2023). Penerapan Fuzzy C-Means pada klasterisasi penerima bantuan pangan non tunai. KLIK: Kajian Ilmiah Informatika dan Komputer, 4(1), 453–461. https://doi.org/10.30865/klik.v4i1.988

Prajetno, J., Setiawan, E. I., & Putra, A. S. (2022). Pembentukan aturan fuzzy untuk pemberian rekomendasi penerima bantuan keluarga berumah tidak layak huni menggunakan K-Means clustering. Information System and Technology, 4(2), 85–92. https://doi.org/10.52985/insyst.v4i2.216

Rahmadina, E. S., Irawan, B., & Bahtiar, A. (2024). Penerapan algoritma Fuzzy C-Means pada data penjualan distro. JATI: Jurnal Mahasiswa Teknik Informatika, 8(2), 2348–2354. https://doi.org/10.36040/jati.v8i2.8467

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sudriyanto, S., Khairi, A., & Hikam, A. S. (2023). Penerapan algoritma K-Means untuk clustering santri pra-sejahtera di Yayasan Bantuan Sosial Az-Zainiyyah Pondok Pesantren Nurul Jadid. NJCA: Nusantara Journal of Computers and Its Applications, 8(1), 22–31.

Supriyanti, K. R., Damiri, B. A., & Ramadhan, W. N. (2024). Pengelompokan faktor-faktor yang mempengaruhi curah hujan di Provinsi Sumatera Utara menggunakan metode Fuzzy C-Means. Jurnal Statistika dan Komputasi, 3(1), 1–10. https://doi.org/10.32665/statkom.v3i1.2623