https://jurnal.pustakagalerimandiri.co.id/index.php/pustakaai
DOI : https://doi.org/10.55382/jurnalpustakaai.v5i2.1139

JURNAL PUSTAKA AI
JURNAL PUSAT AKSES KAJIAN TEKNOLOGI ARTIFICIAL INTELLIGENCE
Pustaka Galeri Mandiri
Vol. 5 No. 2 (2025)193– 203
E ISSN: 2809-4069

Perbandingan Metode *Machine Learning* untuk Klastering Penerima Bantuan Pendidikan Siswa

M. Rafly Gusmansyah[1], Rahmaddeni[2], Rohid[3], Suandi Daulay[4], Ahmad Rivaldi[5]
[1]Program Studi Teknik Informatika, Universitas Sains dan Teknologi Indonesia
[2]Program Studi Teknik Informatika, Universitas Sains dan Teknologi Indonesia
[3]Program Studi Teknik Informatika, Universitas Sains dan Teknologi Indonesia
[4]Program Studi Sistem Informasi, Sekolah Tinggi Teknologi Pekanbaru
[5]Program Studi Teknik Informatika, Universitas Sains dan Teknologi Indonesia

[1]raflygusmansyah5@gmail.com, [2]rahmaddeni@usti.ac.id, [3]rohidpku01@gmail.com , [4]suwandidaulay90@gmail.com, [5]ahmadrivaldi222@gmail.com

Abstract

The selection of educational assistance recipients in educational institutions is often carried out manually and subjectively. It leads to inaccuracies in identifying students who genuinely need financial support. This condition results the potential misallocation of aid and social injustice. This study aims to identify the best clustering algorithm to do grouping the recipients of educational assistance based on socioeconomic data. It supports to more objective and accurate decision-making system. The method used is comparative approach by three clustering algorithms: Fuzzy C-Means, K-Medoids, and Hierarchical Clustering. The data include 1,200 students from MTsN 1 Kota Pekanbaru with attributes such as parental income, number of dependents, and parental status. To reduce data complexity, Principal Component Analysis (PCA) is applied as a dimensionality reduction technique. Performance evaluation is conducted using the Silhouette Coefficient and Davies-Bouldin Index. The results indicate that the Fuzzy C-Means algorithm outperforms the others with a Silhouette score of 0.449 and a DBI of 0.779, followed by K-Medoids and Hierarchical Clustering. The cluster visualization also shows that Fuzzy C-Means provides the most balanced grouping for the categories "Highly Eligible," "Eligible," and "Not Eligible." The main contribution of this research is the development of a machine learning-based decision support system to enhance the objectivity of educational aid distribution. It is concluded that Fuzzy C-Means effectively handles the complexity of socioeconomic data and is recommended for implementation in future educational aid selection systems to promote fairness and precision.

Keywords: machine learning, clustering, hierarchical, fuzzy c-means, k-medoids.

Abstrak

Proses seleksi penerima bantuan pendidikan di institusi pendidikan sering kali dilakukan secara manual dan subjektif, sehingga menimbulkan ketidakakuratan dalam penentuan siswa yang benar-benar membutuhkan bantuan. Hal ini menyebabkan ketidaktepatan alokasi bantuan dan potensi ketidakadilan dalam penyaluran. Penelitian ini bertujuan untuk mengidentifikasi algoritma *clustering* terbaik guna mengelompokkan siswa calon penerima bantuan pendidikan berdasarkan data sosial ekonomi, sehingga mendukung sistem pengambilan keputusan yang lebih objektif dan akurat. Metode yang digunakan adalah pendekatan komparatif terhadap tiga algoritma *clustering*, yaitu *Fuzzy C-Means*, *K-Medoids*, dan *Hierarchical Clustering*. Data penelitian terdiri dari 1.200 entri siswa MTsN 1 Kota Pekanbaru dengan atribut penghasilan orang tua, jumlah tanggungan, dan status orang tua. Untuk mengurangi kompleksitas data, digunakan *Principal Component Analysis (PCA)* sebagai teknik reduksi dimensi. Evaluasi performa dilakukan menggunakan *Silhouette Coefficient* dan *Davies-Bouldin Index*. Hasil penelitian menunjukkan bahwa algoritma *Fuzzy C-Means* memberikan performa terbaik dengan nilai *Silhouette* 0,449 dan *DBI* 0,779, diikuti oleh *K-Medoids* dan *Hierarchical Clustering*. Visualisasi hasil klaster

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

menunjukkan bahwa *Fuzzy C-Means* menghasilkan distribusi kategori "Sangat Layak", "Layak", dan "Tidak Layak" yang paling proporsional. Kontribusi penelitian ini adalah pada pengembangan sistem pendukung keputusan berbasis *machine learning* untuk seleksi bantuan pendidikan yang lebih objektif dan terukur. Penelitian ini menyimpulkan bahwa *Fuzzy C-Means* efektif dalam menangani kompleksitas data sosial ekonomi dan direkomendasikan untuk implementasi sistem seleksi bantuan pendidikan yang lebih adil dan tepat sasaran di masa mendatang.

Kata kunci: machine learning, clustering, hierarchical, fuzzy c-means, k-medoids.

1. Pendahuluan

Pemerintah daerah dan lembaga zakat seperti Badan Amil Zakat Nasional (BAZNAS) terus berupaya menyalurkan bantuan pendidikan kepada siswa dari keluarga kurang mampu, sebagai bagian dari strategi pengurangan kesenjangan pendidikan dan angka putus sekolah. Namun, proses identifikasi calon penerima masih dilakukan secara manual oleh pihak sekolah sehingga rawan subjektivitas dan kesalahan sasaran. Pada penelitian [1] telah menunjukkan bahwa pendekatan *data-driven* mampu meningkatkan akurasi seleksi penerima bantuan dengan memanfaatkan atribut sosial ekonomi siswa sebagai indikator objektif.

Penerapan metode *unsupervised learning* berbasis *clustering* mulai banyak diteliti dalam konteks sosial dan pendidikan karena kemampuannya dalam mengelompokkan data tanpa label. *Fuzzy C-Means* (*FCM*), *K-Medoids*, dan *Hierarchical Clustering* adalah tiga pendekatan yang umum digunakan. Penelitian oleh [2] mengindikasikan bahwa kombinasi metode ini dapat mengidentifikasi pola performa akademik berdasarkan latar belakang sosial dengan cukup baik. Di sisi lain, studi oleh [3] menekankan bahwa metode *FCM* cenderung lebih akurat dalam menangani data sosial dengan ketidakpastian dan tumpang tindih antar kategori.

Dalam konteks sosial lokal di Indonesia, metode *clustering* telah diterapkan dalam pemetaan indikator kesejahteraan maupun pendidikan. Misalnya, [4] menggunakan *K-Medoids* untuk menganalisis kerentanan sosial di Jawa Tengah. Penelitian serupa dilakukan oleh [5] yang membandingkan *FCM*, *K-Medoids*, dan *K-Means* untuk memetakan kualitas pendidikan di wilayah Indonesia Timur. Hasil-hasil tersebut mengonfirmasi bahwa setiap metode memiliki keunggulan spesifik, namun studi komparatif berbasis data siswa madrasah di bawah naungan lembaga zakat seperti BAZNAS masih sangat terbatas dalam literatur ilmiah terkini.

Kompleksitas data sosial siswa seperti tingkat pendidikan orang tua, pekerjaan, penghasilan, hingga kepemilikan aset, membutuhkan proses *dimensionality reduction* untuk efisiensi pengolahan. *Principal Component Analysis (PCA)* telah terbukti efektif menyederhanakan representasi data tanpa kehilangan informasi penting [6]. Penggunaan *PCA* sebelum *clustering* juga diadopsi dalam penelitian oleh [7] dalam kerangka pendidikan berbasis data.

Evaluasi kualitas klaster dilakukan dengan metrik internal seperti *Silhouette Coefficient* dan *Davies-Bouldin Index*. Validasi ini digunakan untuk menilai keakuratan struktur klaster dan kemampuan metode dalam membedakan karakteristik antar kelompok. Penelitian oleh [8] menunjukkan bahwa kombinasi metrik validasi dapat meningkatkan keandalan sistem rekomendasi berbasis data sosial. Dengan demikian, penelitian ini tidak hanya membandingkan efektivitas metode *clustering* tetapi juga memberikan kontribusi praktis dalam pengembangan sistem pendukung keputusan lembaga zakat berbasis kecerdasan buatan.

2. Metode Penelitian

Penelitian ini menerapkan pendekatan kuantitatif eksploratif dengan tujuan untuk mengelompokkan siswa berdasarkan data sosial sebagai dasar dalam seleksi penerima bantuan pendidikan oleh lembaga zakat. Data dikumpulkan dari MTsN 1 Kota Pekanbaru, mencakup sekitar 1.200 entri. Setiap entri terdiri dari atribut yang merepresentasikan kondisi sosial ekonomi siswa, seperti pekerjaan orang tua, penghasilan keluarga, jumlah tanggungan, dan kepemilikan kendaraan. Pada studi [9] menunjukkan bahwa pemanfaatan data sosial secara sistematis mampu memberikan gambaran akurat dalam penyusunan kebijakan berbasis kebutuhan sosial masyarakat.

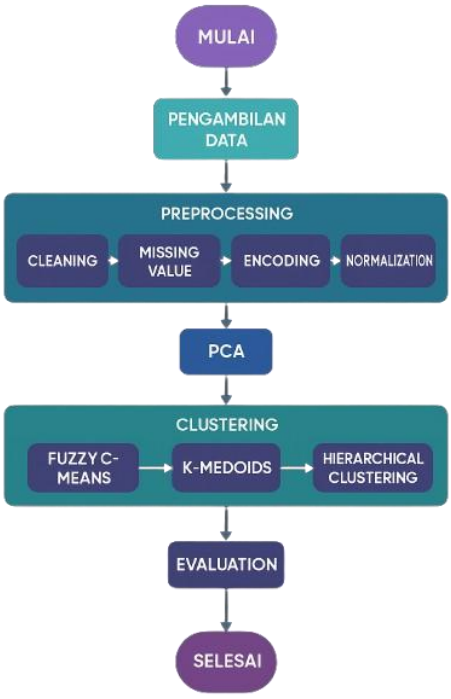
Untuk mengatasi kompleksitas data multivariabel yang tinggi, digunakan teknik reduksi dimensi *Principal Component Analysis (PCA)*. Metode ini berfungsi menyederhanakan representasi data dengan mempertahankan informasi penting, sehingga efisien dalam mempercepat proses pemodelan klasifikasi. Penelitian oleh [10] menunjukkan bahwa *PCA* mampu meningkatkan akurasi identifikasi pola dalam data sosial-ekologis kompleks, yang relevan diterapkan pada konteks klasifikasi sosial siswa. Penggunaan *PCA* juga efektif dalam menghindari redundansi antar atribut data yang sering kali terjadi dalam data survei sosial.

Tahap berikutnya adalah proses pengelompokan data menggunakan tiga algoritma *clustering* berbeda, yaitu *Fuzzy C-Means, K-Medoids, dan Hierarchical Clustering*. Masing-masing metode memiliki keunggulan berbeda, *Fuzzy C-Means* memungkinkan partisipasi data dalam lebih dari satu klaster, *K-Medoids* lebih stabil terhadap outlier, sementara *Hierarchical Clustering* mempermudah visualisasi struktur relasi antar klaster. Pendekatan komparatif ini sejalan dengan studi [11] yang menunjukkan bahwa pemilihan algoritma klasterisasi harus disesuaikan dengan karakteristik spesifik data sosial.

Efektivitas masing-masing metode diuji menggunakan dua metrik evaluasi internal, yaitu *Silhouette Coefficient dan Davies-Bouldin Index*. *Silhouette Coefficient* digunakan untuk mengukur kohesi dalam klaster dan separasi antar klaster, sedangkan *Davies-Bouldin Index* mengukur efisiensi pemisahan antar kelompok. Studi oleh [12] menunjukkan bahwa kombinasi dua metrik ini memberikan hasil evaluatif yang lebih komprehensif dalam pemodelan sosial berbasis klaster.

Metode yang digunakan dalam penelitian ini menawarkan pendekatan sistematis berbasis data sosial siswa dan menghadirkan evaluasi objektif terhadap algoritma klasterisasi. Berbeda dengan pendekatan sebelumnya yang hanya mengandalkan satu algoritma atau tidak melakukan validasi metrik ganda, penelitian ini memanfaatkan tiga algoritma *clustering* dan dua metrik evaluatif sekaligus. Hasil penelitian ini diharapkan dapat memberikan landasan yang kuat dalam pengembangan sistem rekomendasi penerima bantuan pendidikan oleh lembaga seperti BAZNAS, agar lebih tepat sasaran dan adil.

PCA digunakan untuk mereduksi dimensi data dan menangkap variabel-variabel yang memiliki kontribusi paling besar terhadap varian data. *PCA* mentransformasikan data asli ke dalam komponen utama yang tidak saling berkorelasi.



Gambar 1. Alur Penelitian

Pada Gambar 1, alur penelitian diatas dapat dijelaskan sebagai berikut ;

2.1. Pengambilan Data

Data diperoleh dari MTsN 1 Kota Pekanbaru melalui sistem EMIS (Education Management Information System) madrasah. Atribut yang digunakan meliputi data sosial siswa seperti penghasilan bulanan orang tua, pekerjaan, status rumah tinggal dan jumlah tanggungan. Seluruh data diekspor dalam format Excel seperti yang ditampilkan pada gambar 2, Agar dapat diolah menggunakan *Google Collab*.

1	Nama Siswa	Pendidikan Terakhir Ayah	Pekerjaan Ayah	Penghasilan per Bulan Ayah	Masih Hidup / Sudah Meninggal Dunia Tahun (Ayah)	Pendidikan Terakhir Ibu	Pekerjaan Ibu	Penghasilan per Bulan Ibu	Masih Hidup / Sudah Meninggal Dunia Tahun (Ibu)	Transportasi dari Rumah
1209	KHANSA AQILLAH SANTOSO	S1	Pegawai Swasta Con	25000000	Masih Hidup	D3	Entrepreneur	15000000	Masih Hidup	Diantar Jemput Orang
1210	NABILA CHAIRUNNISA DIANDRA	S2	Notaris	8000000	Masih Hidup	S2	Karyawan Sv	7000000	Masih Hidup	Diantar Jemput Orang
1211	JENYTA RAFANIA SUSKA	S1	Jasa Konstruksi Pelak	5000000	Masih Hidup	S1	Wirausaha (p	4000000	Masih Hidup	Diantar Jemput Orang
1212	KHANSA BALQIS AURORA	SLTA	Karyawan Swasta	3000000	Masih Hidup	SLTA	Ibu Rumah T	0	Masih Hidup	Diantar Jemput Orang
1213	MUHAMMAD IQBAL	S2	Guru Honor	3000000	Masih Hidup	S1	Wiraswasta,	1500000	Masih Hidup	Diantar Jemput Orang
1214	NAJWA MIFTAH AL0ZAHRA	D3	Pedagang	2500000	Masih Hidup	S1	Ibu Rumah T	0	Masih Hidup	Diantar Jemput Orang
1215	AMELIA SAFIRA	S1	ASN	5000000	Masih Hidup	SLTA	Ibu Rumah T	0	Masih Hidup	Diantar Jemput Orang

Gambar 2. Dataset Siswa MTsN 1 Kota Pekanbaru

2.2. *Preprocessing*

Pada tahapan ini, dilakukan proses data preprocessing, yaitu serangkaian teknik untuk mempersiapkan data sosial siswa agar dapat digunakan dalam proses klasterisasi secara optimal. Preprocessing merupakan langkah krusial dalam analisis data mining untuk membersihkan data dari kesalahan, menyamakan skala antar variabel, serta mengkonversi format data agar dapat diolah oleh algoritma klasterisasi seperti *PCA* dan Fuzzy C-Means [13]. Beberapa langkah yang diterapkan dalam preprocessing data pada penelitian ini antara lain adalah:

2.2.1. *Data Cleaning*

Pada tahap ini, dilakukan penghapusan data duplikat dan penanganan nilai kosong (missing values) yang ditemukan pada atribut seperti penghasilan, pekerjaan orang tua. Jika nilai kosong sedikit, maka baris dihapus; jika signifikan, dilakukan imputasi dengan rata-rata atau modus.

2.2.2. *Missing Value*

Beberapa atribut ditemukan memiliki nilai kosong, terutama pada variabel pekerjaan dan pendidikan orang tua. Penanganan nilai hilang dilakukan dengan dua pendekatan:

Untuk atribut kategorik, nilai hilang diisi menggunakan modus (nilai terbanyak dalam kolom tersebut).

Untuk atribut numerik, nilai kosong diisi menggunakan nilai median, karena lebih tahan terhadap pencilan dibanding mean.

2.2.3. *Encoding*

Untuk atribut kategorik seperti pekerjaan orang tua dan status kepemilikan rumah, dilakukan konversi ke dalam bentuk numerik menggunakan teknik *one-hot encoding* agar dapat digunakan dalam analisis numerik seperti *PCA* dan *FCM*.

2.2.4. *Normalization*

Semua atribut numerik seperti penghasilan orang tua, jumlah tanggungan, dinormalisasi menggunakan metode min-max normalization agar nilainya berada pada rentang 0–1. Ini dilakukan untuk menyamakan skala antar fitur agar tidak terjadi bias terhadap fitur dengan nilai besar [14].

2.3. Reduksi Dimensi Dengan *PCA*

PCA digunakan untuk mereduksi dimensi data dan menangkap variabel-variabel yang memiliki kontribusi paling besar terhadap varian data. *PCA* mentransformasikan data asli ke dalam komponen utama yang tidak saling berkorelasi. Pada penelitian ini, *PCA* bertujuan untuk menyaring atribut yang paling berpengaruh terhadap kondisi sosial ekonomi siswa dengan cara mereduksi variabel-variabel yang bersifat redundan atau memiliki korelasi tinggi. Proses ini penting karena data sosial siswa mencakup banyak atribut seperti pekerjaan orang tua, penghasilan, jumlah tanggungan, dan kepemilikan aset yang bisa saling berkorelasi. Dengan mengurangi jumlah dimensi data tanpa mengorbankan informasi utama, proses *clustering* menjadi lebih efisien dan hasil klasterisasi menjadi lebih representatif.

Rumus *PCA* :

$z = XW$ (1)

di mana:

X = data awal,

W = *eigenvector* dari matriks kovarians data,

Z = data hasil transformasi ke komponen utama [15].

2.4. Pengelompokan Dengan Fuzzy C-Means

Setelah reduksi dimensi, dilakukan pengelompokan data menggunakan algoritma *Fuzzy C-Means* (*FCM*). *FCM* memberikan fleksibilitas dalam klasifikasi karena satu data dapat memiliki derajat keanggotaan terhadap lebih dari satu klaster.

Fungsi Objektif *FCM* :

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} \left\| x_1 - c_j \right\|^2 \quad (2)$$

Dengan :

u_{ij} = derajat keanggotaan data I pada cluster j,

c_j = pusat klaster ke – j,

m = nilai *fuzziness* (umumnya m = 2)[16].

2.5. Pengelompokan Dengan *K-Medoids*

Setelah proses reduksi dimensi selesai, data diklasterkan menggunakan algoritma *K-Medoids*, yaitu metode klasterisasi berbasis partisi yang menggunakan medoid (data aktual) sebagai pusat klaster. Pemilihan K-Medoids dilakukan karena kemampuannya dalam menangani data yang mengandung nilai pencilan (outlier) dan distribusi yang tidak merata, sehingga cocok diterapkan pada data sosial siswa.

Fungsi objektif dari *K-Medoids* adalah meminimalkan total jarak antara anggota klaster dengan pusatnya, yang dirumuskan sebagai:

Total Cost = $\sum_{i=1}^{k} \sum_{x \in Ci} d(x, m_i)$ (3)

Dengan:

Ci = adalah himpunan anggota klaster ke-i

mi = adalah medoid klaster ke-i (dipilih dari data)

d(x,mi) = adalah jarak antara data x dan medoid mi, biasanya menggunakan jarak Manhattan atau *Euclidean*.

Proses iteratif dilakukan untuk menukar medoid secara dinamis hingga diperoleh konfigurasi klaster yang menghasilkan total jarak minimum antar anggota klaster terhadap medoidnya. Metode ini dinilai unggul untuk data sosial karena tidak sensitif terhadap data ekstrem dan menghasilkan pemisahan klaster yang lebih stabil [17].

2.6. Pengelompokan Dengan *Hierarchical Clustering*

Algoritma Hierarchical *Clustering* digunakan untuk membentuk struktur klaster secara bertahap dalam bentuk hierarki. Metode ini tidak memerlukan jumlah klaster yang ditentukan sejak awal dan bekerja dengan menggabungkan atau memisahkan data berdasarkan tingkat kemiripan antar objek. Proses pengelompokan divisualisasikan dalam bentuk dendrogram, yang memungkinkan analisis visual terhadap struktur relasional antar individu dalam dataset.

Terdapat dua pendekatan umum: Agglomerative (bottom-up) yang memulai dengan setiap data sebagai klaster tunggal lalu digabungkan, dan Divisive (top-down) yang memulai dari satu klaster besar kemudian dipecah. Penentuan kedekatan antar klaster dapat menggunakan beberapa linkage criteria, seperti single linkage (jarak

terdekat antar elemen), complete linkage (jarak terjauh), dan average linkage (rata-rata jarak). Metode ini sangat sesuai untuk mengungkap pola berjenjang dalam konteks sosial-ekonomi siswa karena fleksibel terhadap struktur klaster yang tidak diketahui di awal [18].

2.7. Evaluasi Klaster

Evaluasi klaster dilakukan dengan dua metrik validasi internal, yakni *Silhouette Coefficient* dan *Davies-Bouldin Index*. *Silhouette Coefficient* digunakan untuk mengukur konsistensi suatu data terhadap klaster yang ditempatinya, dengan membandingkan kedekatan data terhadap klaster lain terdekat. Nilainya berada antara -1 hingga 1, di mana nilai mendekati 1 menandakan klasterisasi yang optimal. Sementara itu, *Davies-Bouldin Index* mengukur rasio antara kedekatan intra-klaster dan jarak antar klaster, yang memberikan gambaran seberapa efisien klaster dipisahkan secara struktural.

Hasil evaluasi ini didukung oleh pendekatan yang digunakan dalam penelitian [19], di mana kedua metrik ini berhasil menunjukkan kualitas struktur klaster pada data pendidikan di Kabupaten Karawang. Selain metrik kuantitatif, visualisasi hasil klaster dilakukan menggunakan scatter plot dua dimensi dari hasil *PCA*, yang memberikan pandangan visual apakah ketiga kelompok Sangat Layak, Layak, dan Tidak Layak dapat dipisahkan secara jelas dalam ruang data tereduksi.

2.8. Interpretasi Hasil

Setelah proses evaluasi menunjukkan kualitas klaster yang baik berdasarkan nilai Silhouette Coefficient dan Davies-Bouldin Index, interpretasi dilakukan untuk memahami karakteristik utama dari masing-masing klaster yang terbentuk. Proses ini penting sebagai dasar bagi lembaga seperti BAZNAS atau pihak sekolah untuk membuat keputusan berbasis data dalam menyalurkan bantuan pendidikan.

Tiga klaster yang terbentuk secara umum dapat ditafsirkan sebagai berikut:

Klaster 1 (Sangat Layak): Kelompok ini umumnya terdiri dari siswa dengan penghasilan orang tua rendah, jumlah tanggungan tinggi, tidak memiliki aset produktif, dan dalam banyak kasus memiliki pencapaian akademik yang baik. Karakteristik ini menjadikan kelompok ini sebagai prioritas utama penerima bantuan karena menunjukkan kebutuhan tinggi namun potensi yang menjanjikan.

Klaster 2 (Layak): Kelompok ini berada dalam kategori menengah, dengan pendapatan keluarga yang terbatas namun tidak ekstrem, jumlah tanggungan sedang, dan kondisi sosial-ekonomi yang cukup stabil. Kelompok ini direkomendasikan sebagai penerima bantuan bersyarat atau sesuai kuota.

Klaster 3 (Tidak Layak): Siswa dalam kelompok ini cenderung memiliki pendapatan orang tua di atas rata-rata, jumlah tanggungan rendah, dan kondisi sosial-ekonomi yang mapan, sehingga dinilai tidak masuk prioritas penerima bantuan pada skema berbasis kebutuhan.

Interpretasi ini diperoleh melalui eksplorasi terhadap nilai rata-rata tiap atribut dalam klaster serta didukung visualisasi klaster pada ruang *PCA* dua dimensi. Hasil ini menunjukkan bahwa pendekatan *clustering* tidak hanya mampu mengelompokkan siswa secara statistik, tetapi juga mampu mencerminkan struktur kebutuhan sosial secara nyata yang dapat diterjemahkan ke dalam kebijakan pemberian bantuan yang lebih akurat dan adil.

3. Hasil dan Pembahasan

3.1. *Dataset*

Dataset yang digunakan dalam penelitian ini berasal dari data internal MTsN 1 Kota Pekanbaru, yang terdiri dari sekitar 1.200 entri siswa. Data ini bersifat primer dan dikumpulkan secara tertutup melalui kerja sama dengan pihak sekolah, dengan tetap menjaga kerahasiaan identitas siswa. Dataset mencakup sejumlah atribut sosial dan ekonomi yang dianggap relevan dalam penentuan kelayakan penerima bantuan pendidikan dari lembaga zakat.

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

Tabel 1. Atribut Data Set

No	Atribut	Jenis Data	Keterangan
1	Penghasilan Ayah	Numerik	Estimasi pendapatan bulanan
2	Penghasilan Ibu	Numerik	Estimasi pendapatan bulanan
3	Pekerjaan Ayah	Kategorik	Pekerjaan Atau Usaha
4	Pekerjaan Ibu	Kategorik	Pekerjaan Atau Usaha
5	Status Ayah (Masih Hidup/Meninggal)	Kategorik	Status Ayah Sudah meninggal atau masih hidup

3.2. *Preprocessing*

Cleaning, handling missing values, encoding, dan normalisasi merupakan langkah awal dalam tahap *preprocessing*. Data siswa dibersihkan dari duplikasi, nilai kosong ditangani menggunakan modus dan median, variabel kategorik dikodekan secara numerik, dan data numerik dinormalisasi ke dalam skala yang seragam. Proses ini memastikan data dalam kondisi siap untuk tahap reduksi dimensi dan pengelompokan selanjutnya. Gambar 3 di bawah menunjukkan hasil akhir dari proses *preprocessing* yang telah dilakukan.

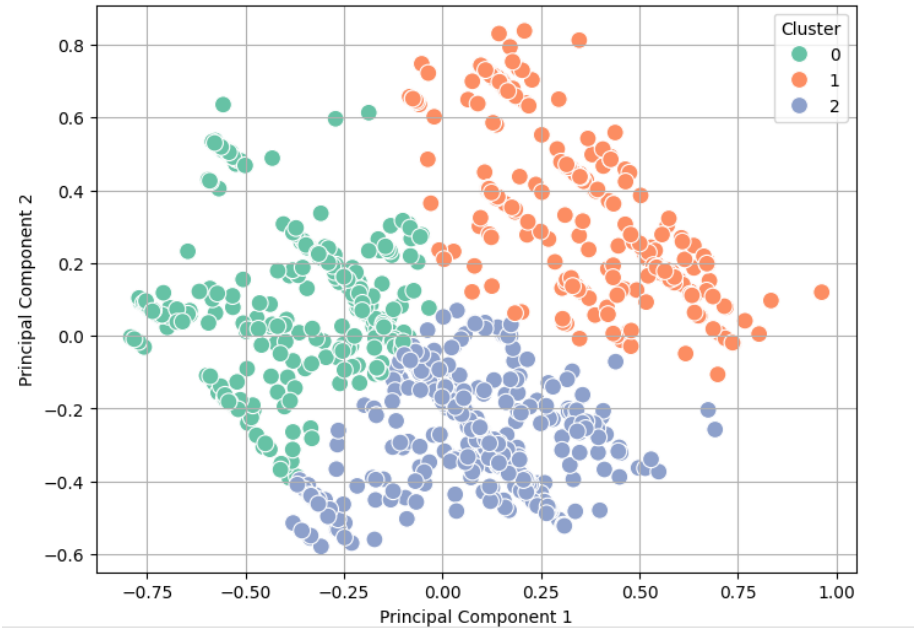
	Pendidikan Terakhir Ayah	Pekerjaan Ayah	Pendidikan Terakhir Ibu	Pekerjaan Ibu	Transportasi dari Rumah	Total Penghasilan
0	0.363636	0.782051	0.846154	0.889503	0.750	0.04375
1	0.363636	0.685897	0.461538	0.000000	0.750	0.08750
2	0.363636	0.403846	0.846154	0.889503	0.750	0.09375
3	0.363636	0.403846	0.461538	0.889503	0.750	0.09375
4	0.363636	0.782051	0.461538	0.889503	0.750	0.04375
5	0.727273	0.003205	0.846154	0.889503	0.750	0.09375
6	0.272727	0.403846	0.230769	0.889503	0.750	0.09375
7	0.727273	0.782051	0.230769	0.624309	0.750	0.08750
8	0.363636	0.403846	0.846154	0.889503	0.875	0.09375
9	0.363636	0.003205	0.153846	0.900552	0.875	0.13750

Gambar 3. Hasil PreProcessing

3.3. Reduksi Dimensi Menggunakan *PCA*

Sebelum proses pengelompokan dilakukan, data sosial siswa yang terdiri dari berbagai atribut numerik dan kategorik dikonversi serta dinormalisasi, kemudian direduksi dimensinya menggunakan metode *Principal Component Analysis* (*PCA*). *PCA* digunakan untuk menyederhanakan representasi data multivariat ke dalam dimensi yang lebih rendah, tanpa kehilangan informasi penting dari varian data asli. Dalam penelitian ini, *PCA* mengekstraksi dua komponen utama yang secara kumulatif mampu menjelaskan proporsi besar dari total variansi data awal.

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025



Gambar 4. Visualisasi Hasil Reduksi Dimensi

Pada Gambar 4 menunjukkan hasil visualisasi scatter plot dua dimensi berdasarkan dua komponen utama dari *PCA.* Titik-titik pada grafik merepresentasikan masing-masing siswa, sementara warna berbeda menandakan hasil pengelompokan dari salah satu algoritma *clustering* (misalnya *Fuzzy C-Means, K-Medoids, atau Hierarchical*).

Terlihat bahwa data berhasil dikelompokkan menjadi tiga klaster yang relatif terpisah, menunjukkan bahwa struktur data sosial memiliki kecenderungan untuk membentuk segmen-segmen homogen. Visualisasi ini mendukung validitas hasil klaster yang telah diperoleh, serta menunjukkan bahwa kombinasi *PCA* dan algoritma *clustering* dapat mengungkap pola dalam data sosial siswa secara efektif.

3.4. Hasil Evaluasi *Cluster*

Setelah proses pengelompokan dilakukan dengan tiga algoritma (*Fuzzy C-Means, K-Medoids,* dan *Hierarchical Clustering*), hasil klasterisasi dievaluasi menggunakan dua metrik validasi internal: *Silhouette Coefficient* dan *Davies-Bouldin Index* (*DBI*). Evaluasi ini bertujuan untuk mengukur kualitas pemisahan klaster dan konsistensi data dalam masing-masing klaster.

Silhouette Coefficient mengukur seberapa dekat suatu data dengan anggota dalam klasternya sendiri dibandingkan dengan klaster lain. Nilai *Silhouette* berada dalam rentang -1 hingga 1. Semakin mendekati 1, semakin baik klasterisasi.

Davies-Bouldin Index mengukur seberapa besar dispersi dalam klaster dibandingkan dengan jarak antar pusat klaster. Nilai *DBI* yang lebih kecil menunjukkan pemisahan klaster yang lebih efisien dan kompak[20].

Hasil evaluasi masing-masing metode ditampilkan pada Tabel 2 berikut:

Tabel 2. Hasil Evaluasi *Cluster*

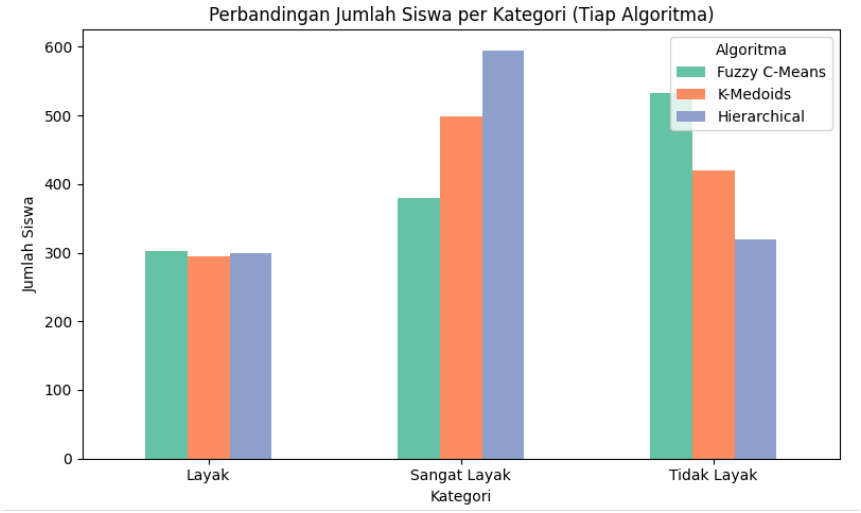
Algoritma Clustering	Silhouette Coefficient	Davies-Bouldin Index
Fuzzy C- Means	0.449	0.779
K-Medoids	0.436	0.787
Hierarchical Clustering	0.405	0.790

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

Berdasarkan hasil evaluasi yang ditampilkan pada Tabel 2, algoritma *Fuzzy C-Means* memberikan performa terbaik dengan nilai *Silhouette Coefficient* tertinggi sebesar 0.449 dan *Davies-Bouldin Index* terendah sebesar 0.779. Hal ini menunjukkan bahwa *Fuzzy C-Means* menghasilkan struktur klaster yang paling kompak dan terpisah dengan baik pada dataset sosial siswa. Meskipun algoritma *K-Medoids* menunjukkan nilai *Silhouette* yang juga tinggi (0.436), nilai *DBI*-nya sedikit lebih tinggi (0.787), menandakan bahwa kepadatan klaster sedikit kurang stabil dibanding *FCM*. Di sisi lain, *Hierarchical Clustering* menghasilkan performa terendah dengan *Silhouette Coefficient* sebesar 0.405 dan *DBI* tertinggi (0.790), yang menunjukkan bahwa struktur klaster yang terbentuk kurang optimal secara statistik.

Temuan ini memperkuat bahwa efektivitas algoritma *clustering* sangat tergantung pada karakteristik distribusi data yang digunakan. Sejalan dengan studi yang dilakukan oleh [18] dan [19] dalam pemilihan metode klasterisasi yang tepat harus mempertimbangkan hasil kuantitatif dari metrik evaluasi internal, bukan hanya kemudahan implementasi atau visualisasi semata.

3.5. Perbandingan Hasil Klasifikasi Antar Algoritma



Gambar 5. Perbandingan Hasil Klasifikasi

Gambar 5 menunjukkan distribusi jumlah siswa dalam tiga kategori penerima bantuan Sangat Layak**,** Layak, dan Tidak Layak berdasarkan hasil klasifikasi dari tiga algoritma *clustering*. Terlihat bahwa ketiganya memberikan hasil pengelompokan yang berbeda secara signifikan, terutama pada kategori Sangat Layak dan Tidak Layak**.**

Algoritma *Hierarchical Clustering* cenderung mengelompokkan lebih banyak siswa ke dalam kategori Sangat Layak, mencapai hampir 600 siswa, sementara proporsi Tidak Layak justru paling sedikit, yaitu sekitar 320 siswa. Hal ini dapat menunjukkan bahwa metode hierarkis bersifat inklusif terhadap atribut sosial yang memiliki kecenderungan kebutuhan tinggi. Sebaliknya, *K-Medoids* memberikan distribusi yang relatif lebih seimbang, dengan 500 siswa masuk dalam kategori Sangat Layak dan sekitar 420 siswa pada Tidak Layak, menunjukkan kemampuannya dalam membentuk klaster berdasarkan jarak aktual antar data. Sementara itu, *Fuzzy C-Means* menghasilkan klasifikasi yang paling merata di antara ketiga kategori, dengan jumlah siswa dalam tiap klaster tidak berbeda secara mencolok, mencerminkan fleksibilitas metode ini dalam menangani ambiguitas karakteristik sosial.

Hasil penelitian ini menunjukkan bahwa pemilihan algoritma *clustering* tidak hanya berdampak pada aspek teknis seperti nilai *Silhouette Coefficient* atau *Davies-Bouldin Index*, tetapi juga berimplikasi langsung terhadap keputusan penyaluran bantuan pendidikan. Perbedaan hasil pengelompokan antar algoritma dapat menyebabkan siswa yang sangat membutuhkan justru tidak teridentifikasi sebagai penerima, sementara siswa dengan kondisi ekonomi lebih baik bisa saja terklasifikasi sebaliknya. Hal ini berpotensi menimbulkan ketidaktepatan sasaran, ketimpangan distribusi bantuan, serta persepsi ketidakadilan di lingkungan sekolah dan masyarakat. Oleh karena itu, pemilihan metode yang paling tepat harus mempertimbangkan tidak hanya

akurasi teknis, tetapi juga dampaknya terhadap keadilan sosial dan efektivitas program bantuan secara keseluruhan.

4. Kesimpulan

Penelitian ini bertujuan untuk mengelompokkan siswa MTsN 1 Kota Pekanbaru sebagai calon penerima bantuan pendidikan dengan memanfaatkan pendekatan *unsupervised learning* berbasis *clustering*. Tiga algoritma yang dibandingkan adalah *Fuzzy C-Means, K-Medoids, dan Hierarchical Clustering*, dengan dukungan *Principal Component Analysis* (*PCA*) sebagai metode reduksi dimensi.

Hasil evaluasi menunjukkan bahwa algoritma *Fuzzy C-Means* memberikan performa paling optimal dengan *Silhouette Coefficient* tertinggi sebesar 0.449 dan *Davies-Bouldin Index* terendah sebesar 0.779, mengindikasikan pemisahan dan kohesi klaster yang paling baik. Algoritma *K-Medoids* juga menunjukkan performa yang stabil, meskipun dengan nilai *DBI* sedikit lebih tinggi (0.787). Sementara itu, *Hierarchical Clustering* memiliki performa terendah dalam evaluasi numerik dan menghasilkan klasifikasi kategori siswa yang cenderung kurang proporsional.

Perbandingan distribusi hasil klasifikasi menunjukkan bahwa *Fuzzy C-Means* menghasilkan kelompok yang paling seimbang, sedangkan *Hierarchical Clustering* cenderung mengelompokkan lebih banyak siswa ke dalam kategori "Sangat Layak". Hal ini menegaskan bahwa setiap algoritma memiliki sensitivitas yang berbeda terhadap karakteristik sosial yang digunakan.

Dengan demikian, *Fuzzy C-Means* direkomendasikan sebagai algoritma paling tepat untuk mendukung sistem pendukung keputusan dalam penyaluran bantuan pendidikan berbasis data sosial. Implementasi metode ini dapat membantu lembaga zakat seperti BAZNAS atau instansi pendidikan dalam menyalurkan bantuan secara lebih objektif, transparan, dan berbasis data.

Daftar Rujukan

[1] R. Jayasree and N. A. S. Selvakumari, "Analyzing Student Performance using Fuzzy Possibilistic C-Means Clustering Algorithm," *Indian J. Sci. Technol.*, vol. 16, no. 38, pp. 3230–3235, 2023, doi: 10.17485/ijst/v16i38.226.

[2] A. F. Mohamed Nafuri, N. S. Sani, N. F. A. Zainudin, A. H. A. Rahman, and M. Aliff, "Clustering Analysis for Classifying Student Academic Performance in Higher Education," *Appl. Sci.*, vol. 12, no. 19, 2022, doi: 10.3390/app12199467.

[3] O. N. Kenger, Z. D. Kenger, E. Ozceylan, and B. Mrugalska, "Clustering of Cities Based on Their Smart Performances: A Comparative Approach of Fuzzy C-Means, K-Means, and K-Medoids," *IEEE Access*, vol. 11, no. October, pp. 134446–134459, 2023, doi: 10.1109/ACCESS.2023.3333753.

[4] Fadlurohman, A., & Nur, I. M. (2023). Pengelompokkan Provinsi di Indonesia Berdasarkan Indikator Varians Menggunakan Algoritma K-Medoids. *Variance: Journal of Statistics and Its Applications*, 2(1), 1–13.

[5] Aji, B. B., & Udin, E. (2024). *Clustering Educational Data in Eastern Indonesia Using FCM, K-Medoids, and K-Means*. Cakrawala Pendidikan.

[6] Y. He, Z. Xu, and N. Liu, "Research on K-medoids Algorithm with Probabilistic-based Expressions and Its Applications," *Appl. Intell.*, vol. 52, no. 10, pp. 12016–12033, 2022, doi: 10.1007/s10489-021-02937-8.

[7] A. Peña-Ayala, *Educational data science: essentials, approaches, and tendencies : proactive education based on empirical big data evidence*. 2023.

[8] G. C. Messakh, M. N. Hayati, and S. Sifriyani, "Comparison K-Means and Fuzzy C-Means In Regencies/Cities Grouping Based on Educational Indicators," *J. Varian*, vol. 7, no. 1, pp. 99–114, 2023, doi: 10.30812/varian.v7i1.2879.

[9] M. A. H, W. Widiatmaka, K. Nirmala, S. Pertiwi, and W. Ambarwulan, "Analisis Kualitas Lingkungan Dan Produktivitas Tambak Budidaya Udang Windu Sistem Teknologi Tradisional Di Kabupaten Bulungan," *Saintek Perikan. Indones. J. Fish. Sci. Technol.*, vol. 18, no. 2, pp. 93–104, 2022, doi: 10.14710/ijfst.18.2.93-104.

[10] S. K. Purba, G. S. Indrawan, and Y. Suteja, "Kondisi Makrozoobentos Kaitannya dengan Ekosistem Mangrove di Kawasan Mangrove Estuari Perancak, Jembrana, Bali," *Bul. Oseanografi Mar.*, vol. 14, no. 1, pp. 1–12, 2025, doi: 10.14710/buloma.v14i1.64699.

[11] A. I. Gunawan, F. Amalia, W. Senalasari, and V. Gaffar, "Pengukuran Aktivitas Pemasaran pada Media Sosial Instagram," *J. Adm. Bisnis*, vol. 10, no. 2, pp. 133–142, 2021, doi: 10.14710/jab.v10i2.35768.

[12] D. Anggraini, Y. Wardi, A. Abror, and V. Dwita4, "Electronic Word of Mouth (Ewom) Dan Sosial Media Marketing Untuk Layanan Transportasi Online: Tinjauan Literatur Sistematis," *J. Sains Pemasar. ...*, vol. 22, no. 2, pp. 57–72, 2023, [Online]. Available:

https://ejournal.undip.ac.id/index.php/jspi/article/view/55684

[13] A. Armansyah and R. K. Ramli, "Model Prediksi Kelulusan Mahasiswa Tepat Waktu dengan Metode Naïve Bayes," *Edumatic J. Pendidik. Inform.*, vol. 6, no. 1, pp. 1–10, 2022, doi: 10.29408/edumatic.v6i1.4789.

[14] Yusuf, M. A., & Hakim, A. (2022). *Perbandingan Normalisasi Min-Max dan Z-Score terhadap Akurasi Klasifikasi Naïve Bayes*. Jurnal Teknologi dan Sistem Komputer, 10(4), 595–602. https://doi.org/10.14710/jtsiskom.2022.3065

[15] A. M. R. Armaya, "Pengaruh Feature Selection Dan Feature Extraction Dalam Peningkatan Akurasi Klasifikasi Kebakaran Hutan," *JuTI "Jurnal Teknol. Informasi,"* vol. 3, no. 1, p. 13, 2024, doi: 10.26798/juti.v3i1.1039.

[16] W. Andriyani, R. Kurniawan, and Y. Arie Wijaya, "Analisis Data Penerimaan Peserta Didik Baru Menggunakan Cross Validation Dan Algoritma Decision Tree Di Sma Negeri 1 Bandung," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 3, pp. 2951–2956, 2024, doi: 10.36040/jati.v8i3.9603.

[17] A. Puri, D. Solihudin, S. Anwar, D. Pratama, and E. Wahyudin, "Analisis Klaster K-Medoid Untuk Pengelompokan Dan Pemetaan Provinsi Di Indonesia Berdasarkan Nilai Ujian Nasional," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 918–924, 2024, doi: 10.36040/jati.v8i1.8653.

[18] A. A. R. Mulyana, A. R. Juwita, A. M. Siregar, and A. Fauzi, "Penerapan Algoritma K-means Clustering dan Hierarchical Clustering dalam Mengelompokkan Data Pengangguran di Karawang," *J. Algoritm.*, vol. 21, no. 2, pp. 209–220, 2024, doi: 10.33364/algoritma/v.21-2.2155.

[19] Abdussalam Amrullah, Intam Purnamasari, Betha Nurina Sari, Garno, and Apriade Voutama, "Analisis Cluster Faktor Penunjang Pendidikan Menggunakan Algoritma K-Means (Studi Kasus: Kabupaten Karawang)," *J. Inform. dan Rekayasa Elektron.*, vol. 5, no. 2, pp. 244–252, 2022, doi: 10.36595/jire.v5i2.701.

[20] S. Suraya, M. Sholeh, and U. Lestari, "Evaluation of Data Clustering Accuracy using K-Means Algorithm," *Int. J. Multidiscip. Approach Res. Sci.*, vol. 2, no. 01, pp. 385–396, 2023, doi: 10.59653/ijmars.v2i01.504.